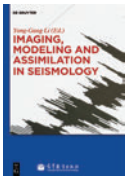
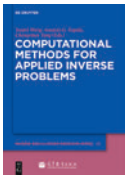
Yong-Gang Li (Ed.)
Seismic Imaging, Fault Damage and Heal

Also of Interest

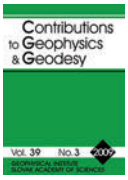


Imaging, Modeling and Assimilation in Seismology
Yong-Gang Li (Ed.), 2012
ISBN 978-3-11-025902-5, e-ISBN 978-3-11-025903-2,
Set-ISBN 978-3-11-220440-5



Computational Methods for Applied Inverse Problems
Yanfei Wang, Anatoly G. Yagola, Changchun Yang (Eds.)
ISBN 978-3-11-025904-9, e-ISBN 978-3-11-025905-6,
Set-ISBN 978-3-11-220441-2

Direct and Inverse Problems in Wave Propagation and Applications
Ivan Graham, Ulrich Langer, Jens Melenk, Mourad Sini (Eds.), 2013
ISBN 978-3-11-028223-8, e-ISBN 978-3-11-028228-3,
Set-ISBN 978-3-11-028229-0



Contributions to Geophysics and Geodesy Online
The Journal of Geophysical Institute of Slovak Academy of Sciences,
4 issues/year
ISSN 1338-0540

Seismic Imaging, Fault Damage and Heal

Edited by
Yong-Gang Li

DE GRUYTER
Higher Education Press

Physics and Astronomy Classification 2010
91.30.Ab, 91.30.Bi, 91.30.Jk, 91.30.pd, 93.85.Rt

ISBN 978-3-11-032991-9
e-ISBN 978-3-11-032995-7
Set-ISBN 978-3-11-032996-4

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Cover image: SteffenHuebner/iStock/Thinkstock
Printing and binding: CPI buch bücher.de GmbH, Birkach
♾Printed on acid-free paper
Printed in Germany

www.degruyter.com

Preface

This book is the second monograph of the earth science specializing in computational, observational and interpretational seismology and geophysics, containing the full-3D waveform tomography method and its application; beamlets and curvelets method for wavefield representation, propagation and imaging; two-way coupling of solid-fluid with discrete element model and lattice Boltzmann model; fault-zone trapped wave observations and 3-D finite-difference synthetics for high-resolution imaging subsurface rupture zone segmentation and bifurcation; fault rock damage and heal associated with earthquakes in California and New Zealand; characterization of pre-shock accelerating moment release with careful considerations in processing and analysis of seismicity using earthquake catalogues; and statistical modeling of earthquake occurrences based on the ultra-low frequency ground electric signals. Each chapter in this book includes the detailed discussion of the state-of-the-art method and technique with their applications in case study. The editor approaches this as a broad interdisciplinary effort, with well-balanced observational, metrological and numerical modeling aspects. Linked with these topics, the book highlights the importance for imaging the crustal complex structures and internal fault-zone rock damage at seismic depths that are closely related to earthquake occurrence and physics.

Researchers and graduate students in geosciences will broaden their horizons about advanced methodology and technique applied in seismology, geophysics and earthquake science. This book can be taken as an expand of the first book in the series, and covers multi-disciplinary topics to allow readers to grasp the new methods and skills used in data processing and analysis as well as numerical modeling for structural, physical and mechanical interpretation of earthquake phenomena, and to strengthen their understanding of earthquake occurrence and hazards, thus helping readers to evaluate potential earthquake risk in seismogenic regions globally. Readers of this book can make full use of the present knowledge and techniques to serve the reduction of earthquake disasters.

Contents

Seismic Imaging, Fault Damage and Heal: An Overview

Yong-Gang Li

This book presents state-of-the-art methods and technique in observational, computational and analytical seismology for earthquake science. Authors from global institutions present multi-disciplinary topics with case studies to illuminate high-resolution imaging of complex crustal structures and earthquake-borne fault zones by the full-3D waveform tomography, beamlets and curvles of localized waves, discrete element model for fully-coupled solid-fluid, and 3-D finite-difference simulation of fault zone trapped (guided) waves observed at recent rupture zones in California and New Zealand. In addition, authors discuss the significance in characterization of the pre-shock moment release using cataloged seismicity, and statistical modeling of earthquake occurrence based on the ultra-low frequency ground electric signals. All topics in this book help further understanding earthquake physics and hazard assessment in global seismogenic regions.

The detailed crustal structure and physical properties of fault network are of great interest because of the factors that control the occurrence and dynamic rupture in earthquake. Observations suggest that the crustal complexity may segment fault zones (Aki, 1984; Malin *et al.*, 1989; Ellsworth, 1990; Beck and Christensen, 1991) or control the timing of moment release in earthquakes (Harris and Day, 1993; Wald and Heaton, 1994). Rupture models have been proposed that involved variations in fluid pressure over the earthquake cycle (Hickman *et al.*, 1995; Blanpied *et al.*, 1992). Geometrical, structural, and rheological fault discontinuities, caused by the spatial variations in strength and stress, will affect the earthquake rupture (e.g., Wesson and Ellsworth, 1973; Das and Aki, 1977; Rice, 1980; Day, 1984; Duan, 2012). Rupture segmentation is often related to fault bends, step-overs, branches, and terminations that have been recognized by surface mapping (e.g., Sieh *et al.*, 1993; Johnson *et al.*, 1994), exhumation (e.g., Chester *et al.*, 1993), and seismic profiling and tomography

(e.g., Lees and Malin, 1990; Thurber *et al.*, 2004). In order to relate present-day crustal stresses and fault motions to the geological structures formed by previous ruptures, we must understand the evolution of fault systems on many spatial and temporal scales in the complex earth crust.

Because the fault plane is thought to be a weakness plane in the earth crust, it facilitates slip to occur under the prevailing stress orientation. As suggested by laboratory experiments, shear faulting is highly resisted in brittle material and proceeds as re-activated faults along surfaces which have already encountered considerable damage (e.g., Dieterich, 1997; Marone, 1998). Field evidence shows that the rupture plane of slip on a mature fault occurs at a more restricted position, the edge of damage zone at the plane of contact with the intact wall rock (Chester *et al.*, 1993; Chester and Chester, 1998). Assuming that this is an actual picture of rupture preparation on the major faults, high-resolution defining the crustal complex and internal damage structure of faults as well as their temporal variations in physical property are challenging work in earthquake science.

Monitoring seismic events and other physical field related to the principal rupture plane would be crucial for earthquake prediction. The slip of these events in series with the main fault is most likely to load the principal slip plane to a point of a major through-going rupture. In these circumstances, it is important to image where the principal fault plane is accompanied with damage zone at depth. Detailing the crustal structure and local variations in seismic velocities has implications for near-fault hazards and expected ground shaking. Greater amplitude shaking is expected near faults due to both proximity to the fault and localized amplification in damaged material. Examining the geometry and physical properties of fault zones as well as the crustal complex structure will help us understand the origin of spatial and temporal variations in rock damage and the evolution of heterogeneities in stress and strain in a seismogenic region.

Other geophysical parameters, such as signals from the ultra-low frequency ground electric field, can be applied for modeling earthquake occurrence. For instance, the version of Ogata's Lin-Lin algorithm (Ogata, 1988) presented in this book is useful for examining the influence of an explanatory signal on the occurrence of earthquakes in a stochastic point process. The statistical models based on observations of these signals allow to forecast earthquakes in its associated circle.

In this book, we introduce the new methodology and technology used in data assimilation for defining subsurface complexity, seismically imaging the multiscale crustal heterogeneity and fault zone geometry, characterizing fault damage magnitude and heal progression, and its physical properties with high-resolution. We also introduce a sophisticated discrete element model with solid-fluid coupling mechanics for earthquake fracture zone rheological simulation, and the pre-

shock accelerating moment release (AMR) model related to the critical-point-like behavior of earthquake preparation. This book includes seven chapters.

Chapter 1: "Applications of Full-Wave Seismic Data Assimilation (FWSDA)" by Dawei Mu, En-Jui Lee and Po Chen.

In the first volume of this book series, Po Chen (2012) introduced theoretical background and recent advances of full-waveform seismic data assimilation (FWSDA) as well as its mathematical formulations in the framework of the various data assimilation theories. In this chapter, Mu *et al.* further discuss the full-wave seismological inverse, as a weakly constrained generalized inverse problem, in which the seismic wave equation with its initial and boundary conditions, the structural and source parameters and the waveform misfit measurements are all allowed to contain errors. The issues related to the applications of FWSDA in realistic seismological inverse problems are also discussed in detail.

Authors present the recent development of FWSDA that can potentially improve the efficiency of some numerical algorithms used for solving acoustic and visco-elastic seismic wave equations. To fully take advantage of the newly emerging computing hardware, algorithmic changes are needed. For the earth structure models in 3-D with highly irregular surface topography and fault structures, the efficiency and the accuracy of the wave equation solver are highly important in solving the problem in a realistic amount of time. In some of the recent successful full-3D waveform tomography applications, the waveform misfit measurements were made on selected wave packets on the seismograms. In order to achieve successful full-3D waveform tomography applications with a large amount of seismic data, the waveform selection process needs to be automated to a certain extent. Authors provide some of the latest developments in numerical solutions of the forward problem and their implementation and optimization on modern CPU-GPU hybrid parallel computing platforms. A realistic full-3D, full-wave tomography for the crustal structure in Southern California is used to illustrate the various components of FWSDA.

Chapter 2: "Wavefield Representation, Propagation and Imaging Using Localized Waves: Beamlet, Curvelet and Dreamlet" by Ru-Shan Wu and Jinghuai Gao.

In this chapter, authors review phase-space localization, mainly along the line of time-frequency localization, and then phase-space localization using generalized wavelet transform applied to wave field and one-way propagator decompositions. Physically the phase-space localized propagators are beamlet or wavepacket propagators which are propagator matrices for short-range iterative propagation. When asymptotic solutions are applied to the beamlet for long-range propagation, beamlets evolve into global beams. Various asymptotic beam propagation methods have been developed in the past, such as the Gaussian beam, complex ray, coherent state, and more recently the curvelet methods. Local perturbation method for propagation in strongly heterogeneous media is

also briefly described in this chapter. Finally, authors review the development of curvelet transform and its application to propagation and imaging in comparison with the beamlet approach.

For wavefield decomposition, both beamlet and curvelet transforms have elementary functions of directional wavelets. Beamlet is a type of physical wavelet, representing an elementary wave in various wavefield decomposition schemes using localized building elements, such as coherent state, Gabor atom, Gabor-Daubechies frame vector, local trigonometric basis function. Curvelet transform is a specifically defined mathematical transform, characterized by the parabolic scaling. Its generalization *width* is similar to the beam-aperture requirement for asymptotic beam solution: the beamwidth must be smaller than the scale of heterogeneity and much greater than the wavelength. Optimal beamwidth is reached by balancing the beam geometric spreading and the beam-front distortion. Using optimal beamwidth, beamlet or curvelet propagator will be sparse in smooth media for short-range propagation. For strong and rough heterogeneities, beamlet or curvelet scattering will occur and asymptotic propagator may not work well. In this case, the local perturbation method can be applied, in which the propagator is decomposed into a background propagator and a perturbation operator for each forward marching step. Numerical examples demonstrate the validity of the approach in this chapter.

Chapter 3: "Two-way Coupling of Solid-fluid with Discrete Element Model and Lattice Boltzmann Model" by Yucang Wang, Sheng Xue and Jun Xie.

This chapter presents a fully coupled solid-fluid code using Discrete Element Method (DEM) and Lattice Boltzmann Method (LBM). The new and distinctive features of this coupled approach compared with the existing coupled DEM-LBM models include the permission of bonded DEM particles, the capability to simulate explicitly fracturing events by the breakage of bonds, simulation of Darcy flow, free flow, and turbulent flow with the same integrated code, adoption of a more stable and efficient moving boundary condition, and a unified parallel algorithm for both codes based on MPI libraries, which allows larger scale parallel computing using super computers in the future. Two widely used open source codes, the Esys-Particle and OpenLB, are integrated as both of the codes are written using C++ and paralleled with MPI library. Recently, LBM has made a significant progress as a new method into numerical modeling of fluid dynamics. In contrast to the conventional computational fluid dynamics (CFD) techniques that solve macroscopic Navier-Stokes equations, LBM is built on a mesoscopic scale in which fluid is described by a group of discrete particles that propagate along a regular lattice and collide with each other. The use of LBM instead of CFD also eliminates severe mesh distortion due to frequent mesh geometry adaptation required in CFD. Because of its Eulerian grids, LBM is particularly suitable for modeling fluid-solid interaction problems, and a large number of solid particles can easily be accommodated.

Authors present three simple preliminary numerical results to assess the performance of the coupled DEM-LBM approach. The small scaled models are used as a qualitative display to demonstrate the capability and potential of the coupled approach. Some preliminary 2-D simulations, such as particles moving in the fluid, fluid flow in a narrow tunnel or crack and hydraulic fracture induced by the injection of fluid into a borehole, are carried out to validate the integrated code. These results show that the new method is capable of simulating solid particle flow in fluid, fluid flow inside narrow fracture, and hydraulic fracture by injection of fluid. The validation of large-scale simulations in 3-D and detailed comparisons with physical experiments are under development.

Chapter 4: "Co-seismic Damage and Post-Mainshock Healing of Fault Rocks at Landers, Hector Mine and Parkfield, California Viewed by Fault-Zone Trapped Waves" by Yong-Gang Li.

This chapter reviews fault rock co-seismic damage and post-mainshock healing progressions associated with the 1992 $M7.4$ Landers, the 1999 $M7.1$ Hector Mine, and the 2004 $M6.0$ Parkfield earthquakes in California through observations and 3-D finite-difference modeling of fault-zone trapped waves (FZTWs) generated by explosions and aftershocks, and recorded at linear seismic arrays deployed across and along the rupture zones (Li *et al.*, 1990, and further references). Because FZTWs arise from coherent multiple reflections at the boundaries between the low-velocity fault zone and the high-velocity surrounding rock, their amplitudes, frequencies and dispersive waveforms strongly depend on the fault geometry and physical properties, these waves enable to insight the internal structure and physical properties of fault zones at seismogenic depths with a higher resolution than ever before. The author with his colleagues from multiple institutions (see acknowledgement and references of Chapter 4) have used FZTWs to delineate the studied rupture zones being a low velocity waveguide about 100 to 250 m wide, in which S velocities are reduced by 40%–50% from wall-rock velocities and Q values are 10–50, which is interpreted as a remnant of process zone where inelastic deformation occurs around the propagating crack tip during dynamic rupture in the mainshocks. The width of the fault zone waveguide scales to the rupture length as predicted in published dynamic rupture models (e.g., Scholz, 1990). FZTWs also show the rupture segmentation and bifurcation associated with these earthquakes.

The strength of the low-velocity anomalies along the fault might vary over the earthquake cycle (e.g., Vidale *et al.*, 1994; Marone, 1998). Repeated seismic experiments conducted at the Landers rupture zone showed fault healing with recovery of seismic velocity by approximate 2% between 1994 and 1998. The survey in 1998 showed a reduction of the healing rate by a factor of two between 1994–1996 and 1996–1998. The ratio of the rates of P-wave and S-wave speed recovery is consistent with healing caused by closure of cracks that are partially fluid-filled. A similar experiment at Hector Mine has confirmed that healing is

not unique to Landers and shows that there is variability in healing rates among the fault segments that we have measured. However, the healing at the Landers rupture was interrupted in 1999 by the $M7.1$ Hector Mine earthquake rupture, which occurred 20–30 km away. The Hector Mine earthquake both strongly shook and permanently strained the Landers fault, adding damage discernible as a temporary reversal of the healing process. The fault has since resumed the trend of strength recovery that it showed after the Landers earthquake. These observations suggest that fault damage caused by strong seismic waves may help to explain earthquake clustering and seismicity triggering by shaking, and may be involved in friction reduction during faulting. At Parkfield, repeated surveys reveal an approximately 2.5% co-seismic decrease in seismic velocity within the San Andreas fault (SAF), due to the co-seismic damage of fault-zone rocks at seismogenic depths during dynamic rupture in the 2004 $M6$ Parkfield earthquake. Seismic velocities then increased by an approximate 1.2% in the following ∼4 months, indicating that the rock damaged in the $M6$ mainshock recovers rigidity through time. These observations lead us to speculate that fault damage caused by strong seismic waves may help to explain earthquake clustering and seismicity triggering by shaking, and may be involved in friction reduction during faulting.

Chapter 5: "Subsurface Rupture Structure of the $M7.1$ Darfield and $M6.3$ Christchurch Earthquake Sequence Viewed with Fault-Zone Trapped Waves" by Yong-Gang Li, Gregory De Pascale, Mark Quigley and Darren Gravely.

In this chapter, Li *et al.* present the subsurface fault rock damage structure along the Greendale fault (GF) and Port Hills fault (PHF) that ruptured in the 2010 $M7.1$ Darfield and 2011 $M6.3$ Christchurch earthquake sequence using fault-zone trapped waves (FZTWs) generated by aftershocks recorded at a linear seismic array installed across the surface rupture along the GF. FZTWs were identified for aftershocks occurring on both the GF and the PHF. The post-S duration of these FZTWs increases as focal depths and epicentral distances from the array increase, showing an effective low-velocity waveguide formed by severely damaged rocks existing along the GF and PHF at seismogenic depths. Locations of aftershocks generating prominent FZTWs delineate the subsurface GF rupture extending eastward as bifurcating blind fault segments an additional ∼5–8 km beyond the mapped ∼30 km surface rupture into a zone with comparably low seismic moment release west of the PHF rupture. The propagation of FZTW through the intervening 'gap' indicates moderate GF-PHF structural connectivity. This zone is interpreted as a fracture mesh reflecting the interplay between basement faults and stress-aligned microcracks that enable the propagation of PHF-sourced FZTWs into the GF damage zone.

Combined with previous rupture models for slip distributions in the Canterbury earthquake sequence (Quigley *et al.*, 2012; Barnhart *et al.*, 2011; Beavan *et al.*, 2012; Elliott *et al.*, 2012), authors construct a plausible model of subsurface

rupture zones associated with the Darfield-Christchurch earthquakes. Velocities of basement rocks in this model are constrained by the existing regional velocity models in Canterbury Plains (e.g., Smith *et al.*, 1995; Eberhart-Phillips and Bannister, 2002; Kaiser *et al.*, 2012). The 3-D finite-difference simulations of observed FZTWs suggest that the GF rupture zone is ∼200–250-m wide, consistent with the surface deformation widths, in which velocities are reduced by 35%–55% with the maximum reduction in the ∼100-m wide damage core zone corresponding to surface and shallow subsurface evidence for discrete fracturing. The damage zone delineated by FZTWs indicates an effective low-velocity waveguide extending ∼65 km along the GF and PHF under the Canterbury Plains while the waveguide varies in its velocity and geometry along multiple rupture segments viewed by FZTWs, and penetrates down to the depth of ∼8 km or deeper, consistent with hypocentral locations and geodetically-derived fault models. Their experiment also illuminates a potential approach to image the buried part of a rupture zone using FZTWs recorded at seismic array deployed at the surface-exposed part of the rupture zone.

Authors have examined the possible temporal change in wave velocity for repeated aftershock occurring just before and after the large aftershocks to find the additional co-seismic damage in rocks associated with these large aftershocks. We measured ∼2% decrease of seismic velocity with fault rocks due to co-seismic damage by an $M5.3$ aftershock. This value is in general consistent with observations of fault rock damage and healing at the San Andreas fault associated with the 2004 $M6$ Parkfield earthquake (Li *et al.*, 2007, 2006).

Chapter 6: "Characterizing Pre-shock (Accelerating) Moment Release: A Few Notes on the Analysis of Seismicity" by Changsheng Jiang and Zhongliang Wu.

Understanding of seismicity is one of the frontiers in the modern seismology. Careful considerations in processing and analysis of seismicity using earthquake catalogues are necessary. in this chapter, Jiang and Wu demonstrate some useful tactics in analysis of earthquake catalog data and make notes on the existing methods used for careful analysis of seismicity in terms of (1) interfering events and the eclipse method, (2) the Bayesian information criterion, (3) the spatio-temporal scales for the sampling of seismic events, and (4) removal of aftershocks and the de-clustered Benioff strain method.

Authors use the pre-shock accelerating moment release (AMR) model (Bufe *et al.*, 1994; Brehm and Braile, 1998; Bowman and King, 2001) related to the critical-point-like behavior of earthquake preparation (Sornette and Sammis, 1995; Bowman *et al.*, 1998; Jaumé and Sykes, 1999; Rundle *et al.*, 2000). They explore whether the claimed and controversial pre-shock acceleration have a firm statistical (and seismological) basis by retrospective investigation in which they focus on the scaling exponent with the failure time fixed to the origin time of the 'target' earthquake so that the fitting can be stabilized by reducing one free

parameter (origin time). Borrowing from the concept of modern astronomy for analyzing remote planets, they use an 'eclipse method' for screening out the seismicity in the neighboring active fault zones as shown in analysis of seismicity for the 2008 $M8$ Wenchuan earthquake catalog data. The Bayesian Information Criterion (BIC) consideration provides a useful aid to judge whether the apparent 'accelerating' trend is statistically significant. The BIC criterion may be able to reveal more clues regarding the accelerating/quiescence behavior in the seismic moment release. To de-cluster an earthquake catalogue, previous works on AMR tended to use simple schemes (e.g., Robinson, 2005; Jiang and Wu, 2010), an alternative approach is to use the 'Epidemic-Type Aftershock Sequences' (ETAS) model (Ogata, 1988; Zhuang *et al.*, 2002; Zhuang and Ogata, 2006), in which a stochastic de-clustering scheme is proposed no longer determine whether an earthquake is a 'background event' or if it is triggered by another. To check the accelerating behavior objectively, authors also try to map the scaling coefficient calculated for different spatio-temporal windows, with different cutoff magnitude of the catalog (Jiang and Wu, 2005, 2010). The method extends a manifestation of the Gutenberg-Richter's law. Deviation from the G-R power-law relation can be used for judging the completeness of an earthquake catalogue. Quantitatively, the goodness of fit between a power law fit to the data and the observed frequency-magnitude distribution as a function of a lower cutoff of the magnitude can be used (Wiemer and Wyss, 2000). Finally, they provide the case study in seismicity analysis using real catalog data: (1) 'crack-like' spatial window for the 2008 $M8.0$ Wenchuan earthquake, (2) a finite earthquake rupture of the 2004 $M9.1$ Sumatra-Andaman earthquake, and (3) seismic moment tensors to investigate the moment release before the 2011 $M9.0$ Tohoku earthquake.

Chapter 7: "Statistical Modeling of Earthquake Occurrences Based on External Geophysical Observations: With an Illustrative Application to the Ultra-low Frequency Ground Electric Signals Observed in the Beijing Region" by Jiancang Zhuang, Yosihiko Ogata, David Vere-Jones, Li Ma and Huaping Guan.

In this chapter, authors present the idea on developing models for earthquake probability forecasts based on the precursor data from observations of the ultra-low frequency components of the underground electric signals used as an example to illustrate the modeling strategies. In the study case, signals from 4 stations in the vicinity of Beijing are used to monitor the variations in ultra-low frequency components electric field for forecasting the occurrence of $M \geqslant 4$ earthquakes within a 300-km circle centered in Beijing. The model used is a version of Ogata's Lin-Lin algorithm for examining the influence of an explanatory signal on the occurrence of events in a stochastic point process, which is highly significant, and greatly superior to the explanatory effect of the same signals applied to a randomized version of the earthquake data. The results from all four stations show significant explanatory power although in combination the two most effective tend to dominate the forecasts. The predictions appear to

be most effective for events with $M \geqslant 5$, for which probability gains are up to 3–4 over the simple Poisson process, and for the events closer to the observing stations. Some smaller events appear to produce detectable signals at distances of over 100 km from the source.

The probability modeling framework adapted in this chapter is extended to the development of probability forecasts, which can be assessed directly, and in their turn can form the basis for a variety of decision procedures (e.g., Vere-Jones, 1995, and further references). Authors present a brief discussion of the performance of probability forecasts based on the best Lin-Lin model, which provides a strong confirmation of the reality of the explanatory power of the electric signals. They also carefully examine the effect of changes in background seismicity. Results show that the Lin-Lin model based on the electrical signals still out-performs the two-stage Poisson model.

The purpose of this book is to introduce the new approaches in solid-earth geophysics research with case studies. The following new methods and results presented in this book will be of particular interest to the readers:

- The full-3D waveform tomography method, and beamlets and curvelets methods for imaging complex subsurface structure.
- Observations and 3-D finite-difference simulations of fault-zone trapped wave for high-resolution delineation of fault internal structure and physical properties.
- Co-seismic rock damage and post-mainshock heal in major earthquakes.
- Discrete element method for solid-fluid coupling mechanics in earthquake fracture modeling.
- Pre-shock accelerating moment release with analysis of seismicity for earthquake risk assessment.
- Ultra-low frequency ground electric signals for statistical modeling of earthquake occurrences.

This book is a self-contained volume starting with an overview of the subject then explores each topic with in depth detail. Extensive reference lists and cross references with other volumes to facilitate further research. Full-color figures and tables support the text and aid the readers in understanding. Content is suited for both the senior researchers and graduate students in geosciences who will broaden their horizons about observational, computational and applied seismology and earthquake sciences. This book covers multi-disciplinary topics to allow readers to gasp the new methods and techniques used in data analysis and numerical modeling for structural, physical and mechanical interpretation of earthquake phenomena, to aid the understanding of earthquake processes and hazards, and thus helps readers to evaluate potential earthquake risk in seismogenic regions globally.

Part of articles in the preceded book (Book 1) edited by Li (2012) and this book (Book 2) came out of International Symposium on Earthquake Seismology and Earthquake Predictability (ISESEP) held in Beijing, China, 2009, sponsored by Institute of Geophysics in China Earthquake Administration (CEA), co-sponsored by the Asian Seismological Commission (ASC) of the International Association of Seismology and Physics of the Earth's Interior (IASPEI) and supported by the International Union for Geodesy and Geophysics (IUGG). The meeting included two special sessions: I. "Wenchuan Earthquake: One Year After" and II. "Keiiti Aki Workshop on Earthquake Physics and Earthquake Predictability". The meeting highlights the importance for an international discussion on the seismology, geology, and geodynamics of strong to great earthquakes, their predictability, and how to make full use of the present knowledge and techniques to reduce earthquake disasters. Chapter 3 by Yong-Gang Li, Peter E. Malin, and Elizabeth S. Cochran; Chapter 6 by Xiang-Chu Yin, Yue Liu, Lang-Ping Zhang, and Shuai Yuan in Book 1, and Chapter 6 by Changsheng Jiang and Zhongliang Wu; Chapter 7 by Jiancang Zhuang, Yosihiko Ogata, David Vere-Jones, Li Ma and Huaping Guan in Book 2 came from representations in the 2009 ISESEP meeting.

The editor of this book series wishes to thank reviewers who contributed to referee articles in Volume 1 (Chen, 2012; Wu *et al.*, 2012; Li *et al.*, 2012a,b; Duan, 2012; Yin *et al.*, 2012; Wang *et al.*, 2012) and the present Volume. In addition to many chapter authors, reviewers include Zhengxi Ge (PKU), Elizabeth Cochran (UCR), En-Jui Lee (UOW), David Oglesby (UCR), Martha Savage (VUOW), Yushen Sun (MIT), Xiao-Bi Xie (UCSC), Xiangzu Yin, and Yingcai Zheng (MIT). We are grateful to many organizations and individuals, including HEP Director Bingxiang Li and Editors Zhengxiong Chen and Yan Guan , who help to make both books possible. This article was completed partly during the Author's (YGL) visit as Honorary Professor in Chinese Academy of Geological Science, Beijing, China.

Key Words: Data assimilation, Full-3D waveform tomography, Beamlets and curvelets methods, Fault-zone trapped waves, Rock damage and heal, Two-way coupling of solid-fluid, Discrete element model and lattice Boltzmann model, Pre-shock moment release, Earthquake catalogues, Relocation of the Wenchuan earthquake, Statistical modeling of earthquake occurrences, Ultra-low frequency ground electric signals.

References

Aki, K., 1984. Asperities, barriers, characteristic earthquakes, and strong motion prediction. *J. Geophys. Res.*, 89, 5867–5872.

Barnhart, W. D., M. J. Willis, R. B. Lohman, and A. K. Melkonian, 2011. InSAR and optical constraints on fault slip during the 2010–2011 New Zealand earthquake sequence. *Seismological Research Letters*, 82 (6), 815–823.

Beck, S. L. and D. H. Christensen, 1991. Rupture process of the February 4, 1965, Rat Islands earthquake. *J. Geophys. Res.*, 96, 2205–2221.

Beavan J., M. Motagh, E. Fielding, N. Donnelly, and D. Collett, 2012. Fault slip models of the 2010–2011 Canterbury, New Zealand, earthquakes from geodetic data, and observations of post-setismic ground deformation. *New Zealand Journal of Geology and Geophysics*, 55, doi: 10.1080/00288306.2012.697472.

Blanpied, M. L., D. A. Lockner, and J. D. Byerlee, 1992. An earthquake mechanism based on rapid sealing of faults. *Nature*, 359, 574–576.

Bowman, D. D. and G. C. P. King, 2001. Accelerating seismicity and stress accumulation before large earthquakes. *Geophys. Res. Lett.*, 28: 4039–4042.

Bowman, D. D., G. Ouillon, C. G. Sammis, A. Sornette, and D. Sornette, 1998. An observational test of the critical earthquake concept. *J. Geophys. Res.*, 103, 24359–24372.

Brehm, D. J. and L. W. Braile, 1998. Intermediate-term earthquake prediction using precursory events in the New Madrid seismic zone. *Bull. Seism. Soc. Am.*, 88, 564–580.

Bufe, C. G., S. P. Nishenko, and D. J. Varnes, 1994. Seismicity trends and potential for large earthquake in the Alaska-Aleutian region. *PAGEOPH*, 142, 83–99.

Chen, P., 2012. Full-wave seismic data assimilation: A unified methodology for seismic waveform inversion. In: Li, Y. G. (Ed.). *Imaging, Modeling and Assimilation in Seismology.* Higher Education Press, Beijing, China, De Gruyter, Boston, USA, 19–63.

Chester, F. M., J. P. Evans, and R. L. Biegel, 1993. Internal structure and weakening mechanisms of the San Andreas fault. *J. Geophys. Res.*, 98, 771–786.

Chester, F. M. and J. S. Chester, 1998. Ultracataclasite structure and friction processes of the San Andreas fault. *Tectonophysics*, 295, 199–221.

Das, S. and K. Aki, 1997. Fault plane with barriers: A versatile earthquake model. *J. Geophys. Res.*, 82, 5658–5670.

Day, S. M., 1984. Three-dimensional simulation of spontaneous rupture: The effect of nonuniform prestress. *Bull. Seismol. Soc. Am.*, 72, 1881–1902.

Dieterich, J. H., 1997. Modeling of rock friction. 1. Experimental results and constitutive equations. *J. Geophys. Res.*, 84, 2169–2175.

Duan, B. C., 2012. Ground-motion simulations with dynamic source characterization and parallel computing. In: Li, Y. G. (Ed.). *Imaging, Modeling and Assimilation in Seismology.* Higher Education Press, Beijing, China, De Gruyter, Boston, USA, 199–218.

Eberhart-Phillips, D. and S. Bannister, 2002. Three-dimensional crustal structure in the Southern Alps region of New Zealand from inversion of local earthquake and active source data. *J. Geophys. Res.*, 107, doi:10.1029/2011JB000567.

Elliott J. R., E. K. Nissen, P. C. England, J. A. Jackson, S. Lamb, Z. Li, M. Oehlers, and B. Parsons, 2012. Slip in the 2010–2011 Canterbury earthquakes, New Zealand. *J. Geophys. Res.*, 117, B03401, 1–36.

Ellsworth, W. L., 1990. Earthquake history, 1769–1989. In: Wallace, R. E. (Ed.). *The San Andreas Fault System, California.* U. S. Geol. Surv. Prof. Pap., 1515, 153–187.

Harris, R. A. and S. M. Day, 1993. Dynamics of fault interaction: Parallel strike-slip faults. *J. Geophys. Res.*, 98, 4461–4472.

Hickman, S., R., Sibson, and R. Bruhn, 1995. Introduction to special section: Mechanical involvement of fluids in faulting. *J. Geophys. Res.*, 100, 12831–12840.

Jaumé, S. C. and I. R. Sykes, 1999. Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquake. *PAGEOPH*, 155, 279–306.

Jiang, C. S. and Z. L. Wu, 2005. Test of the preshock accelerating moment release (AMR) in the case of the 26 December 2004 $M_{\mathrm{W}}9.0$ Indonesia earthquake. *Bull. Seism. Soc. Am.*, 95, 2016–2025.

Jiang, C. S. and Z. L. Wu, 2010. Seismic moment release before the May 12, 2008, Wenchuan earthquake in Sichuan of Southwest China. *Concurrency Computat.: Pract. Exper.*, 22, 1784–1795.

Johnson, A. M., R. W. Fleming, and K. M. Cruikshank, 1994. Shear zones formed along long straight traces of fault zones during the 28 June 1992 Landers, California, earthquake. *Bull. Seism. Soc. Am.*, 84, 499–510.

Kaiser, A., C. Holden, J. Beavan, D. Beetham, R. Benites, A. Celentano, D. Collett, J. Cousins, M. Cubrinovski, G. Dellow, P. Denys, E. Fielding, B.Fry, M. Gerstenberger, R.Langridge, C. Massey, M. Motagh, N. Pondard, G. McVerry, J. Ristau, M. Stirling, J. Thomas, S. R. Uma, and J. Zhao, 2012. The $M_{\mathrm{W}}6.2$ Christchurch earthquake of February 2011: Preliminary report. *New Zealand Journal of Geology and Geophysics*, 55 (1), 67–90.

Lees, J. M. and P. E. Malin, 1990. Tomographic images of P wave velocity variation at Parkfield, California. *J. Geophys. Res.*, 95, 21793–21804.

Li, Y. G. and P. C. Leary, 1990. Fault zone trapped seismic waves. *Bull. Seism. Soc. Am.*, 80, 1245–1271.

Li, Y. G., P. C. Leary, K. Aki, and P. E. Malin, 1990. Seismic trapped modes in the Oroville and San Andreas fault zones. *Science*, 249, 763–766.

Li, Y. G., K. Aki, D. Adams, A. Hasemi, W. H. K. Lee, 1994. Seismic guided waves trapped in the fault zone of the Landers, California, earthquake of 1992. *J. Geophys. Res.*, 99, 11705–11722.

Li, Y. G., J. E. Vidale, K. Aki, F. Xu, T. Burdette, 1998. Evidence of shallow fault zone strengthening after the 1992 $M7.5$ Landers, California, earthquake. *Science*, 279, 217–219.

Li, Y. G., P. Chen, E. S. Cochran, J. E. Vidale, and T. Burdette, 2006. Seismic evidence for rock damage and healing on the San Andreas fault associated with the 2004 $M6$ Parkfield earthquake. Special issue for Parkfield $M6$ earthquake. *Bull. Seism. Soc. Am.*, 96(4), S1-15, doi:10.1785/0120050803.

Li, Y. G., J. E. Vidale, K. Aki, and F. Xu, 2000. Depth-dependent structure of the Landers fault zone from trapped waves generated by aftershocks. *J. Geophys. Res.*, 105, 6237–6254.

Li, Y. G., J. E. Vidale, S. M. Day, and D. Oglesby, 2002. Study of the $M7.1$ Hector Mine, California, earthquake fault plan by fault-zone trapped waves. Hector Mine Earthquake Special Issue. *Bull. Seism. Soc. Am.*, 92, 1318–1332.

Li, Y. G., J. E., Vidale, and S. E. Cochran, 2004. Low-velocity damaged structure of the San Andreas fault at Parkfield from fault-zone trapped waves. *Geophy. Res. Lett.*, 31, L12S06.

Li, Y. G., P. Chen, E. S. Cochran, and J. E. Vidale, 2007. Seismic velocity variations on the San Andreas Fault caused by the 2004 $M6$ Parkfield earthquake and their implications. *Eearth and Planate Science*, 59, 21–31.

Li, Y. G. and P. E. Malin, 2008. San Andreas Fault damage at SAFOD viewed with fault-guided waves. *Geophys. Res. Lett.*, 35, L08304, doi:10.1029/2007GL032924.

Li, Y. G., 2012. *Imaging, Modeling and Assimilation in Seismology.* Higher Education Press, Beijing, China, De Gruyter, Boston, USA, 1–262.

Li, Y. G., P. Malin, and E. Cochran, 2012a. Fault-zone trapped waves: High-resolution characterization of the damage zone on the Parkfield San Andreas fault at depth. In: Li, Y. G. (Ed.). *Imaging, Modeling and Assimilation in Seismology.* Higher Education Press, Beijing, China, De Gruyter, Boston, USA, 108–150.

Li, Y. G., J. Y. Sue, and T. C. Chen, 2012b. Fault-zone trapped waves at a dip fault: Documentation of rock damage on the thrusting Longmen-Shan fault ruptured in the 2008 $M8$ Wenchuan earthquake. In: Li, Y. G. (Ed.). *Imaging, Modeling and Assimilation in Seismology.* Higher Education Press, Beijing, China, De Gruyter, Boston, USA, 151–198.

Malin, P. E., S. N. Blakeslee, M. G. Alvarez, and A. J. Martin, 1989. Microearthquake imaging of the Parkfield asperity. *Science*, 244, 557–559.

Marone, C., 1998. Laboratory-derived friction laws and their application to seismic faulting. *Annu. Rev. Earth Planet. Sci.*, 26, 643–696.

Ogata, Y., 1988. Likelihood analysis of point processes and its application to seismological data. *Bulletin of the International Statistical Institute*, 50, 943–961.

Quigley, M., R. Van Dissen, N. Litchfield, P. Villamor, D. Barrell, T. Stahl, E. Bilderback, D. Noble, 2012. Surface rupture during the 2010 $M_{\rm W}7$ Darfield (Canterbury) earthquake: Implications for fault rupture dynamics and seismic hazard analysis. *Geology*, 40 (1), 55–58.

Rice, J. R., 1980. The mechanics of earthquake rupture. In: Dziewonski, A. M. and E. Boschi. (Eds.). *Physics of the Earth's Interior.* Amsterdam, 555–649.

Robinson, R., S. Y. Zhou, S. Johnston, and D. Vere-Jones, 2005. Precursory accelerating seismic moment release (AMR) in a synthetic seismicity catalog: A preliminary study. *Geophys. Res. Lett.*, 32, L07309, doi:10.1029/2005GL022576.

Rundle, J. B., W. Klein, D. L. Turcotte, and B. D. Malamud, 2000. Precursory seismic activation and critical-point phenomena. *PAGEOPH*, 157, 2165–2182.

Scholz, C.H., 1990. Wear and gouge formation in brittle faulting. *Geology*, 15, 493–495.

Sieh, K., et al., 1993. Near-field investigations of the Landers earthquake sequence, April to July 1992. *Science*, 260, 171–176.

Smith E. G., T. Stern, and B. O'Brien, 1995. A seismic velocity profile across the central South Island, New Zealand, from explosion data. *New Zealand Journal of Geology and Geophysics*, 38, 565–570.

Sornette, D. and C. G. Sammis, 1995. Critical exponents from renomalization group theory of earthquakes: Implications for earthquake prediction. *J. Phys. I.*, 5: 607–619.

Thurber, C., S. Roecker, H. Zhang, S. Baher, and W. Ellsworth, 2004. Fine-scale structure of the San Andreas fault zone and location of the SAFOD target earthquakes. *Geophys. Res. Letter*, 31, L12S02, doi:10.1029/2003GL019398.

Vere-Jones, D., 1995. Forecasting earthquakes and earthquake risk. *International Journal of Forecasting*, 11, 503–538.

Vidale, J. E., W. L. Ellsworth, A. Cole, and C. Marone, 1994. Rupture variation with recurrence interval in eighteen cycles of a small earthquake. *Nature*, 368, 624–626.

Vidale, J. E. and Y. G. Li, 2003. Damage to the shallow Landers fault from the nearby Hector Mine earthquake. *Nature*, 421, 524–526.

Wald, D. J. and T. H. Heaton, 1994. Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bull. Seism. Soc. Am.*, 84, 668–691.

Wang, Y. C., S. Xue, and J. Xie, 2012. Discrete element method and its applications in earthquake and rock fracture modeling. In: Li, Y. G. (Ed.). *Imaging, Modeling and Assimilation in Seismology.* Higher Education Press, Beijing, China, De Gruyter, Boston, USA, 235–262.

Wiemer, S. and M. Wyss, 2000. Minimum magnitude of complete reporting in earthquake catalogs: Examples from Alaska, the Western United States, and Japan. *Bull. Seism. Soc. Am.*, 90, 859–869.

Wesson, R. L. and W. L. Ellsworth, 1973. Seismicity preceding moderate earthquakes in California. *J. Geophys. Res.*, 78, 8527–8545.

Wu, R. S., X. B. Xie, and S. W. Jin, 2012. One-return propagators and the applications in modeling and imaging. In: Li, Y. G. (Ed.). *Imaging, Modeling and Assimilation in Seismology.* Higher Education Press, Beijing, China, De Gruyter, Boston, USA, 65–105.

Yin, X. C., Y. Liu, S. A. Yuan, L. P. Zhang, 2011. LURR and its new progress. In: Li, Y. G. (Ed.). *Imaging, Modeling and Assimilation in Seismology.* Higher Education Press, Beijing, China, De Gruyter, Boston, USA, 219–234.

Zhuang, J. and Y. Ogata, 2006. Properties of the probability distribution associated with the largest event in an earthquake cluster and their implications to foreshocks. *Phys. Rev. E.*, 73, 046134, doi: 10.1103/PhysRevE.73.046134.

Zhuang, J., Y. Ogata, and D. Vere-Jones, 2002. Stochastic declustering of space-time earthquake occurrences. *J. Amer. Stat. Assoc.*, 97, 369–380.

Author Information

Yong-Gang Li
Department of Earth Sciences, University of Southern California, Los Angeles, CA 90089, USA.
E-mail: ygli@usc.edu

Chapter 1
Applications of Full-Wave Seismic Data Assimilation (FWSDA)

Dawei Mu, En-Jui Lee, and Po Chen

In the first volume of this book series, we introduced the concept of full-wave seismic data assimilation (FWSDA) and its mathematical formulations in the framework of the various data assimilation theories (Chen, 2010). The full-wave seismological inverse problem, which aims at estimating earth structure parameters and seismic source parameters using observed waveform data and the seismic wave equation, can be formulated as a weakly constrained generalized inverse, in which the seismic wave equation (with its initial and boundary conditions), the structural and source parameters and the waveform misfit measurements are all allowed to contain errors. FWSDA provides a unified framework for solving seismological inverse problems and for estimating uncertainties associated with the nonlinear inversion process. Both the adjoint-wavefield (AW) method and the scattering-integral (SI) method can be derived from FWSDA as special cases. In this chapter, we will discuss issues related to the applications of FWSDA in realistic seismological inverse problems. In FWSDA, the seismic wave equation and its adjoint system, if the AW method is adopted, or the receiver-side Green's tensors (RGTs), if the SI method is adopted, need to be solved many times. For three-dimensional earth structure models with highly irregular surface topography or fault structures, the efficiency and the accuracy of the wave equation solver are highly important in solving the problem in a realistic amount of time. In this chapter, we will review and discuss some of the latest developments in numerical solutions of the forward problem and their implementation and optimization on modern CPU-GPU hybrid parallel computing platforms. In some of the recent successful full-3D waveform tomography applications, the waveform misfit measurements were made on selected wave packets on the seismograms. For realistic inversions involving a large amount of seismic data, this waveform selection process needs to be automated to a certain extent. We will discuss

some recent developments in automating seismic waveform data processing and selection. A realistic full-3D, full-wave tomography for the crustal structure in Southern California will be used to illustrate the various components of FWSDA.

Key Words: Data assimilation, Full-wave tomography, Full-3D inversion, Earthquake source parameters, Discontinuous Galerkin, Adjoint method and scattering-integral methods, Finite-difference, Discontinuous mesh, GPU, Waveform selection.

1.1 Numerical Solutions of Seismic Wave Equations

Computer simulations of seismic wavefields have played an important role in seismology in the past few decades. However, the accurate and computationally efficient numerical solution of the three-dimensional (visco)elastic seismic wave equation is still a very challenging task, especially when the material properties are complex and the modeling geometry, such as surface topography and subsurface fault structures, is irregular. In the past, several numerical schemes have been developed to solve the elastic seismic wave equation. The finite-difference (FD) method was introduced to simulate SH and P-SV waves on regular, staggered-grid, two-dimensional meshes in Madariaga (1976) and Virieux (1984, 1986). The FD method was later extended to three spatial dimensions and to account for anisotropic, viscoelastic material properties (e.g., Mora 1989; Igel *et al.*, 1995; Tessmer, 1995; Graves, 1996; Moczo *et al.*, 2002). The spatial accuracy of the FD method is mainly controlled by the number of grid points required to accurately sample the wavelength. The pseudo-spectral (PS) method with Chebychev or Legendre polynomials (e.g., Carcione, 1994; Tessmer and Kosloff, 1994; Igel, 1999) partially overcomes some limitations of the FD method and allows for highly accurate computations of spatial derivatives. However, due to the global character of its derivative operators, it is relatively cumbersome to account for irregular modeling geometry and efficient and scalable parallelization on distributed-memory computer clusters is not as straightforward as in the FD method. Another possibility is to consider the weak (i.e., variational) form of the seismic wave equation. The finite-element (FE) method (e.g., Lysmer and Drake, 1972; Bao *et al.*, 1998) and the spectral-element (SE) method (e.g., Komatitsch and Vilotte, 1998; Komatitsch and Tromp, 1999, 2002) are based on the weak form. An important advantage of such methods is that the free-surface boundary condition is naturally accounted for even when the surface topography is highly irregular. And in the SE method, high-order polynomials (e.g.,

Lagrange polynomials defined on Gauss-Lobatto-Legendre points) are used for approximation, which provides a significant improvement in spatial accuracy and computational efficiency.

The arbitrary high-order discontinuous Galerkin (ADER-DG) method on unstructured meshes was introduced to solve two-dimensional isotropic elastic seismic wave equation in Käser and Dumbser (2006). It was later extended to three-dimensional isotropic elastic case in Dumbser and Käser (2006) and to account for viscoelastic attenuation (Käser *et al.*, 2007), anisotropy (la Puente *et al.*, 2007) and poroelasticity (la Puente *et al.*, 2009). The p-adaptivity (i.e., the polynomial degrees of the spatial basis functions can vary from element to element) and locally varying time steps were addressed in Dumbser *et al.* (2007). Unlike conventional numerical schemes, which usually adopt a relatively low-order time-stepping method such as the Newmark scheme (Hughes, 1987) and the 4^{th}-order Runge-Kutta scheme (e.g., Igel, 1999), the ADER-DG method achieves high-order accuracy in both space and time by using the arbitrary high-order derivatives (ADER), which was originally introduced in Titarev and Toro (2002) in the finite-volume framework. The ADER scheme performs high-order explicit time integration in a single step without any intermediate stages. In three dimensions, the ADER-DG scheme achieves high-order accuracy on unstructured tetrahedral meshes, which allows for automated mesh generation even when the modeling geometry is highly complex. Furthermore, the majority of the operators in the ADER-DG method are applied in an element-local way, with weak element-to-element coupling based on numerical flux functions, which results in strong locality in memory access patterns. And the high-order nature of this method lets it require fewer data points, therefore fewer memory fetches, in exchange for higher arithmetic intensity. These characteristics of the ADER-DG method make it well suited to run on massively parallel graphic processing units (GPUs).

In the following sections, we will discuss some recent developments in the finite-difference method, in particular, its extensions to non-uniform and discontinuous meshes, and the ADER-DG method in more detail. It is likely that the literature cited in the following is incomplete. However, some of the key references are included and readers who are interested in studying these topics in depth can use them as a starting point for further investigation. This is a highly active research area with many new ideas and implementations emerging rapidly. The advance in computing architecture certainly plays an important role and many new implementations and optimizations are facilitated by innovations in computer sciences.

1.1.1 Stable Finite-Difference Solutions on Non-Uniform, Discontinuous Meshes

The finite-difference method for solving acoustic and (visco)elastic seismic wave equations has been used extensively in seismology because its numerical efficiency is high both on commodity desktops and on modern distributed-memory parallel computing platforms and it is relatively easy to program and use. In conventional uniform-mesh finite-difference method, the grid space and time step length are determined based on the maximum desired frequency of the resulting synthetic seismograms and the CFL (Courant-Friedrichs-Levy) stability condition, i.e.,

$$\frac{\alpha_{\max}\Delta t}{h} < 0.5 \tag{1.1}$$

where $\alpha_{\max}$ is the maximum P-wave speed, Δt is the time-step length and h is the grid space. Using our tomography in Southern California as an example, the maximum desired frequency of the synthetic seismograms is 0.2 Hz and the minimum S-wave speed in our three-dimensional starting model is 900 m/s, which gives a minimum wavelength of 4,500 m. If we choose a grid space of 500 m, we can guarantee 9 grid points per minimum wavelength in our three-dimensional 4^{th}-order staggered-grid finite-difference simulations. In a 4^{th}-order finite-difference scheme, 5.5–6 grid points per minimum wavelength are usually sufficient to ensure accuracy of the synthetic seismograms. We are using 9 grid points per minimum wavelength in the starting model because the minimum S-wave speed in our structure model may reduce when we update our velocity model during the iterative tomographic inversion process. The maximum P-wave speed in the simulation volume is 8,223 m/s, considering Equation (1.1), the time-step length must be smaller than 0.0304 s for the simulation to be stable. For a simulation volume that is 900 km long, 450 km wide and 50 km deep, the total number of grid points is 162 million. If the desired length of the synthetic seismograms is 180 s and the time-step length is around 0.03 s, the total number of time steps is about 6,000. On the latest IBM Blue Gene/Q system, it takes 2,048 cores in about 15 minutes of wall-time to complete one simulation.

For many earth structure models, the minimum S-wave speed close to the surface of the earth can be much smaller than that at greater depths. If this is the case, using a discontinuous mesh with finer grid in the upper part of the model and a coarser grid in the lower part of the model may significantly improve computational efficiency without scarifying simulation accuracy. Considering our Southern California example, the minimum S-wave speed increases from around 900 m/s at 250 m depth to around 3,000 m/s at around 5 km depth. If we adopt a finer grid with 500 m grid space for the modeling volume above 5 km depth and a coarser grid with 1,500 m grid space for the volume below 5 km

depth, the total number of grid points is 21.6 million, a reduction of about 87% compared with the uniform mesh configuration, which can be directly translated into a significant amount of savings in either the wall-time or the core count or both.

An important challenge in implementing finite-difference methods on discontinuous meshes is how to reduce the instability caused by the numerical noise generated at the interface between the finer and the coarser grids. On this interface, in order to compute the spatial derivatives of the field variables (e.g., velocity and stress) at the finer-grid boundary we need access to the field variables at grid positions that do not exist at the coarser-grid boundary. Some type of interpolation scheme is needed to obtain the field variables at those missing grid positions. The existing finite-difference implementations on discontinuous meshes can be categorized based on their interpolation approaches for reducing the instability. For two-dimensional acoustic wave equations, Jastram and Behle (1992) used trigonometric interpolation in the horizontal direction to obtain the pressure at those missing grid positions at the boundary of the coarser grid. The trigonometric interpolation scheme is closely related to Fourier spectral methods, which have been shown to be highly accurate in computing spatial derivatives of the field variables. This interpolation scheme allows arbitrary integer ratio of the coarser grid space $H$and the finer grid space h, although intuitively one can expect that the larger is the grid ratio H/h, the higher is the possibility of generating numerical instability. The same methodology was extended to two-dimensional P-SV elastic wave equation using a staggered grid in Jastram and Tessmer (1994). An interpolation scheme that is closely related to trigonometric interpolation is the interpolation in the wavenumber domain, which is adopted in Wang and Schuster (1996) to solve three-dimensional acoustic and elastic wave equations. The same technique was extended to the viscoelastic wave equation in Wang *et al.* (2001). Simple linear or bilinear interpolation schemes have also been adopted in both two-dimensional (e.g., Hayashi *et al.*, 2001) and three-dimensional (e.g., Aoi and Fujiwara, 1999) finite-difference simulations. In Aoi and Fujiwara (1999), numerical evidences have shown that when the grid space ratio $H/h = 3$ and the number of grid points per wavelength is larger than 10, the error introduced by a linear interpolation scheme is less than 2.2%, which is sufficiently accurate for the 2^{nd}-order staggered-grid finite-difference scheme used in their simulations.

A different issue that is also related to the instability problem is how to downsample the field variables from the finer grid to the coarser grid on the interface. Theoretical considerations (e.g., Kristek *et al.*, 2010) and some numerical experiments (e.g., Hayashi *et al.*, 2001; Kristek *et al.*, 2010) have shown that one cannot simply take the field variable values in the finer grid to replace those coarser-grid field variable values at the coarser grid positions that coincide with the grid points in the finer grid when computing spatial derivatives of the field

variables in the coarser grid. From a theoretical point of view (e.g., Kristek *et al.*, 2010), the minimum wavelength supported by the finer grid λ_h is smaller than the minimum wavelength supported by the coarser grid λ_H for a given frequency. When the wave-field enters the coarser grid from the finer grid at the interface, waves with wavelength larger than λ_h but smaller than λ_H will introduce aliasing effect into the coarser grid and a filtering process that removes waves with wavelength smaller than λ_H is needed at the interface to ensure numerical stability. In Hayashi *et al.* (2001), a one-dimensional five-point averaging formula was used to improve the stability of their two-dimensional P-SV viscoelastic finite-difference scheme. In Kristek *et al.* (2010), the Lanczos down-sampling filter was used to improve the stability of their three-dimensional 4^{th}-order staggered-grid finite-difference scheme. The Lanczos filter is a windowed sinc function in space and provides a good approximation to a boxcar in the wavenumber space. It can be implemented efficiently using a weighted averaging formula on the interface (Kristek *et al.*, 2010).

If a single time step is used for the discontinuous spatial mesh, this time step may become unnecessarily small for some spatial grid points. To further improve numerical efficiency, a straightforward extension is to use a locally varying time step that is adapted to the stability condition, Equation (1.1), in each submesh. This type of local-time-step, discontinuous-grid finite-difference method was implemented in Kang and Baag (2004). In their implementation, a simple linear interpolation scheme was adopted for both the temporal and the spatial interpolations of the field variables on the mesh interface and the 4^{th}-order staggered-grid finite-difference scheme is used for all interior grid points. The efficient implementation of such local-time-step, discontinuous-grid finite-difference schemes on modern distributed-memory parallel computing platforms is still a very challenging issue. If the spatial mesh is distributed evenly among all processors using a simple domain decomposition approach, the processors that are mainly occupied by the coarser grid will likely be idle for a significant amount of time because the field variables on the coarser grid are updated less frequently than those on the finer grid, which is a serious load-balancing problem. Another possibility is to evenly distribute the finer grid and the coarser grid separately so that each processor owns an equal number of finer grids, as well as an equal number of coarser grids. In such a case, every processor will always have some work to do at every time step, but the spatial decomposition of the finer and the coarser grids may no longer conform to simple boundaries and may introduce additional complexity in exchanging boundary field variables among processors.

Instead of using a discontinuous mesh, one can also try to adapt the mesh to the velocity model using a non-uniform but continuous mesh. In a non-uniform mesh, the number of grid points in each direction does not change; therefore one does not need to interpolate field variables. However, the grid space can vary in accordance with the velocity model and avoid oversampling in regions with

high velocity. Following Pitarka (1999), the 4^{th}-order difference operator D_x on a field variable $g(x)$ at location x_i can be expressed as

$$D_x g(x_i) = c_1 g(x_i + \Delta_1) + c_2 g(x_i - \Delta_2) + c_3 g(x_i + \Delta_3) + c_4 g(x_i - \Delta_4) \quad (1.2)$$

where c_i are 4 coefficients to be determined and Δ_i are spatial increments on both sides of x_i and can be expressed in terms of the non-uniform grid spaces. Transforming Equation (1.2) into the Fourier domain, we obtain an equation in terms of the wavenumber k,

$$ik = c_1 \exp(ik\Delta_1) + c_2 \exp(-ik\Delta_2) + c_3 \exp(ik\Delta_3) + c_4 \exp(-ik\Delta_4) \quad (1.3)$$

The exponentials in Equation (1.3) can be expanded into Taylor series and we can truncate the Taylor expansion to 4^{th}-order. Using the first term on the right-hand-side as an example, we have,

$$\exp(ik\Delta_1) \approx 1 + ik\Delta_1 - \frac{k^2\Delta_1^2}{2} - i\frac{k^3\Delta_1^3}{6} \quad (1.4)$$

Bringing Equation (1.4) into Equation (1.3) and collecting the terms according to the order of k, we obtain

$$\begin{aligned} ik = {} & (c_1 + c_2 + c_3 + c_4) + ik(c_1\Delta_1 - c_2\Delta_2 + c_3\Delta_3 - c_4\Delta_4) \\ & + \frac{k^2}{2}(-c_1\Delta_1^2 - c_2\Delta_2^2 - c_3\Delta_3^2 - c_4\Delta_4^2) \\ & + i\frac{k^3}{6}(-c_1\Delta_1^3 + c_2\Delta_2^3 - c_3\Delta_3^3 - c_4\Delta_4^3) \end{aligned} \quad (1.5)$$

Equation (1.5) can be expressed in a matrix form as

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ \Delta_1 & -\Delta_2 & \Delta_3 & -\Delta_4 \\ -\Delta_1^2 & -\Delta_2^2 & -\Delta_3^2 & -\Delta_4^2 \\ -\Delta_1^3 & \Delta_2^3 & -\Delta_3^3 & \Delta_4^3 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad (1.6)$$

which can be solved for the coefficients c_i. The same analysis can also be performed on the y- and z-axis. Explicit expressions for c_i in terms of Δ_i can be obtained by solving Equation (1.6) using a computer algebra system such as Maple and Mathematica. Once the non-uniform mesh has been set up, the spatial increments Δ_i are known and the coefficients c_i only need to be computed once and stored on disk. For a staggered-grid mesh, two sets of c_i need to be computed for field variables located on the grid points and those located on positions shifted by half the grid space.

Perhaps an even more efficient implementation would be a combination of a discontinuous mesh with a non-uniform mesh. In Liu and Archuleta (2002), the

mesh is allowed to be discontinuous in the vertical direction with a grid space ratio $H/h = 3$ and also non-uniform in all three spatial dimensions. The perfectly-matched-layer (PML) boundary condition is implemented for all boundaries of the modeling volume except for the free-surface and the 4^{th}-order staggered-grid finite-difference scheme is adopted for all interior grid points. This code has been parallelized using the message-passing-interface (MPI). It is used in some of our own modeling and inversion studies in which the effects of surface topography and the curvature of the Earth do not need to be considered. The improvement in computational efficiency is really astonishing compared with a uniform-mesh 4^{th}-order staggered-grid finite-difference code. In cases where irregular surface topography and/or subsurface fault structures need to be accounted for, we use the ADER-DG method for solving the seismic wave equation. More discussions about our ADER-DG implementation are presented in Sections 1.1.3 and 1.1.4.

1.1.2 Accelerating Finite-Difference Methods Using GPUs

In the past four decades, the development in the computing chip industry has roughly followed the Moore's law. Many of the performance improvements were due to increased clock speeds and sophisticated instruction scheduling in a single core. As the transistor density keeps increasing, the industry is now facing a number of engineering difficulties with using a large number of transistors efficiently in individual cores (e.g., power consumption, power dissipation). The effect is that clock speeds are staying relatively constant and core architecture is expected to become simpler. As a consequence, when we consider future platforms for high-performance scientific computing, there are some inevitable trends, for instance, the increase in the number of cores in general-purpose CPUs, and the adoption of many-core accelerators (e.g., Field Programmable Gate Array, Graphic Processing Unit, Cell Broadband Engine) due to their footprints smaller and power consumptions per flop lower than general-purpose CPUs. The users who want to once again experience substantial performance improvements as before need to learn how to exploit multiple/many cores.

The graphic processing unit (GPU) has become an attractive many-core coprocessor for general-purpose scientific computing in the past few years. In the conventional CPU architecture, a large amount of transistors are dedicated for caches, prediction and speculation, which is mainly to battle the memory bottleneck caused by bandwidth limitations and memory-fetch latency. Unlike in a conventional CPU, in a typical GPU, many more transistors are dedicated for arithmetic calculations rather than data caching and flow control. The abundance of cheap computing power on a GPU allows us to effectively hide memory-access latencies with massive parallelism. In particular, on a GPU one

can launch a large number of threads and the thread scheduler can effectively overlap memory transactions for some threads with arithmetic calculations on other threads. Such a massive parallelism offered by GPUs is particularly well suited for addressing data-parallel calculations such as those used in solving seismic wave equations. In fact, most of the numerical algorithms used for solving seismic wave equations can be expressed in terms of simple local-scale operations applied in parallel on many different pieces of distributed data with limited or no interdependence, i.e., single-instruction-multiple-data (SIMD) style.

In the past, programming on GPUs was difficult and different from that on CPUs because of the significant barriers to recast scientific algorithms into unfamiliar graphic programming frameworks. Recent efforts by GPU vendors, in particular, NVIDIA's CUDA (Compute Unified Device Architecture) programming model, the OpenCL (Open Computing Language) framework and the OpenACC compiler directives and APIs, have significantly increased the programmability of commodity GPUs. Using these tools, a programmer can directly issue and manage data-parallel computations on GPUs using high-level instructions without the need to map them into a set of graphic-processing instructions. For readers who are not familiar with CUDA or GPU programming, we give a very brief introduction about the programming model in the following section.

1.1.2.1 CUDA programming model

The CUDA software stack is composed of several layers, including a hardware driver, an application programming interface (API) and its runtime environment. There are also two high-level, extensively optimized CUDA mathematical libraries, the fast Fourier transform library (CUFFT) and the basic linear algebra subprograms (CUBLAS), which are distributed together with the software stack. The CUDA API comprises an extension to the C programming language for a minimum learning curve. The complete CUDA programming toolkit is distributed free of charge and is regularly maintained and updated by NVIDIA.

A CUDA program is essentially a C program with multiple subroutines (i.e., functions). Some of the subroutines may run on the "host" (i.e., the CPU) and others may run on the "device" (the GPU). The subroutines that run on the device are called CUDA "kernels". A CUDA kernel is typically executed on a very large number of threads to exploit data parallelism, which is essentially a type of SIMD operation. Unlike on CPUs where thread generation and scheduling usually takes thousands of clock cycles, GPU threads are extremely "light-weight" and cost very few cycles to generate and manage. The very large amounts of threads are organized into many "thread blocks". The threads within a block are executed in groups of 16, called a "half-warp", by the "multiprocessors" (a

type of vector processor), each of which executes in parallel with the others. A multiprocessor can have a number of "stream processors", which are sometimes called "cores". A high-end Fermi GPU has 16 multiprocessors and each multiprocessor has two groups of 16 stream processors, which amounts to 512 processing cores.

The memory on a GPU is organized in a hierarchical structure. Each thread has access to its own register, which is very fast, but the amount is very limited. The threads within the same block have access to a small pool of low-latency "shared memory". The total amount of registers and shared memory available on a GPU restricts the maximum number of active warps on a multiprocessor (i.e., the "occupancy"), depending upon the amount of registers and shared memory used by each warp. To maximize occupancy, one should minimize the usage of registers and shared memory in the kernel. The most abundant memory type on a GPU is the "global memory", however, accesses to the global memory have much higher latency. To hide the latency, one needs to launch a large number of thread blocks so that the thread scheduler can effectively overlap the global memory transactions for some blocks with the arithmetic calculations on other blocks. To reduce the total number of global memory transactions, each access needs to be "coalesced" (i.e., consecutive threads accessing consecutive memory addresses), otherwise the access will be "serialized" (i.e., separated into multiple transactions), which may heavily impact the performance of the code.

In addition to data-parallelism, GPUs are also capable of task-parallelism, which is implemented as "streams" in CUDA. Different tasks can be placed in different streams and the tasks will proceed in parallel despite the fact that they may have nothing in common. Currently task parallelism on GPUs is not yet as flexible as on CPUs. Current-generation NVIDIA GPUs now support simultaneous kernel executions and memory copies either to or from the device.

1.1.2.2 CUDA implementations of finite-difference methods

With the rapid development of the GPU programming tools, various numerical algorithms have been successfully ported to GPUs and GPU-CPU hybrid computing platforms and substantial speedups, compared with pure-CPU implementations, have been achieved for applications in different disciplines. In the area of acoustic/elastic seismic wave propagation simulations, finite-difference methods (e.g., Abdelkhalek *et al.*, 2009; Michéa and Komatitsch, 2010; Okamoto *et al.*, 2010; Wang *et al.*, 2010; Unat *et al.*, 2012; Zhou *et al.*, 2012), the spectral-element method (e.g., Komatitsch *et al.*, 2009; Komatitsch *et al.*, 2010) and the ADER-DG method (Mu *et al.*, 2013) have been successfully ported to GPUs using the CUDA programming model. The speedup obtained varies from around

20-fold to around 60-fold depending on several factors, e.g., whether a particular calculation is amenable to GPU acceleration, how well the reference CPU code is optimized, the particular CPU and GPU architectures used in the comparisons and the specific compilers, as well as the compiler options, used for generating the binary codes. In this section, we discuss CUDA implementations of finite-difference methods. In Section 1.4, we will discuss the CUDA implementation of the ADER-DG method.

In most of the finite-difference methods, the majority of the calculations involve a central point and a set of neighboring points in space. The spatial derivatives of the field variable at the central point are approximated using a weighted average of field variables at neighboring points. This neighborhood in space is often referred to as the *stencil*. The stencil operator applied to every point is the same, except for possible differences in the weights in non-uniform meshes. In a typical C-language implementation, the computation is implemented as nested for-loops, in which the loop indices sweep through every grid point in the mesh and update the field variables in place. A straightforward parallelization scheme is to use one thread to handle one central point. For a 4^{th}-order finite-difference scheme, each stencil is composed of 13 grid points. If all the field variables are stored in the global memory, to compute the spatial derivatives of the field variable at the central point, each thread will need 13 accesses to the global memory, which will result in very poor performance since the global memory has the highest access latency. But in practice, it is not necessary for each thread to carry out all 13 accesses because neighboring stencils share many grid points and the field variables on those shared grid points can be fetched from the high-latency global memory and stored in the low-latency shared memory. If the number of threads in a thread block is large enough, on average each thread will only need one access to the global memory to fetch the field variable located at its own central point and store it into the shared memory and the rest 12 fetches will be from the low-latency shared memory. For the few threads located at the boundary of a thread block, more fetches from the global memory are needed because different thread blocks cannot share data directly on current-generation GPUs.

The use of the shared memory improves the performance of the CUDA code significantly by removing most of the redundant accesses to the high-latency global memory. To further improve performance of the code, we need to make sure that the remaining accesses to the global memory are coalesced. In our case, this problem involves understanding two different issues, i.e., how a three-dimensional array is laid out in the global memory and what is the indexing scheme for a three-dimensional thread block. In both C and CUDA, a three-dimensional field variable, say vz[NY][NX][NZ], is laid out linearly in memory as, vz[0][0][0], vz[0][0][1], vz[0][0][2], ..., vz[0][0][NZ-1], vz[0][1][0], vz[0][1][1], ..., which is known as "row-major ordering". For this particular example, we often

say that "the z-axis is the fastest direction and the y-axis is the slowest direction for array vz". In a three-dimensional thread block, each thread is indexed using three integers, threadIdx.x, threadIdx.y and threadIdx.z and in CUDA thread topology, threadIdx.x is the fastest direction and the threadIdx.z is the slowest direction. To ensure that consecutive threads are accessing consecutive addresses in the global memory, for our example, one would like to match threadIdx.x with the z-axis of the array vz and threadIdx.y with the x-axis of vz and threadIdx.z with the y-axis of vz. In practice, one often uses a two-dimensional thread topology. In such a case, threadIdx.x should be matched with the z-axis of array vz, threadIdx.y should be matched with the x-axis of vz and each thread corresponds to one point in the x-z plane of vz and has to loop through the entire y-axis of vz.

For a two-dimensional thread block, to compute spatial derivatives of field variables in the y-axis one can store multiple x-z planes of field variables in shared memory if the amount of shared memory on the GPU is large enough. For a 4^{th}-order finite-difference scheme, one widely used algorithm is to store 4 consecutive x-z planes in shared memory. Then for each iteration in the loop direction (y-axis in our example), 3 out of the 4 x-z planes from the previous iteration can be re-used and we only need to fetch one x-z plane from the global memory. The 4 x-z planes in shared memory are constantly updated during the loop with one old plane being discarded and one new plane being added. This rotation process reduces about 75% of global memory accesses after all the threads in the thread block loop through the entire y-axis.

There are also other issues need to be considered to fully take advantage of the computing capability of the GPU. Certain directive-based C-to-CUDA translation software, such as mint (Unat *et al.*, 2012), can be used to facilitate this process for finite-difference calculations. We have successfully ported the discontinuous, non-uniform mesh, finite-difference code of Liu and Archuleta (2002) to GPU. On the latest Kepler K20 GPU, we obtained a speedup of around 15-fold when compared with a single Intel Nehalem 2.4 GHz CPU with 4 cores.

1.1.3 The ADER-DG Method

The ADER-DG method for solving the seismic wave equation is both flexible and robust. It allows unstructured meshes and easy control of accuracy without compromising simulation stability. Like the SE method, the solution inside each element is approximated using a set of orthogonal basis functions, which leads to diagonal mass matrices. These types of basis functions exist for a wide range of element types. Unlike the SE or typical FE schemes, the solution is allowed to be discontinuous across element boundaries. The discontinuity is treated us-

ing well-established ideas of numerical flux functions from the high-order finite-volume framework. The spatial approximation accuracy can be easily adjusted by changing the order of the polynomial basis functions within each element (i.e., p-adaptivity). The ADER time-stepping scheme is composed of three major ingredients, a Taylor expansion of the degree-of-freedoms (DOFs, i.e., the coefficients of the polynomial basis functions in each element) in time, the solution of the Derivative Riemann Problem (DRP) (Toro and Titarev, 2002) that approximates the space derivatives at the element boundaries and the Cauchy-Kovalewski procedure for replacing the temporal derivatives in the Taylor series with spatial derivatives. We summarize major equations of the ADER-DG method for solving the three-dimensional isotropic elastic wave equation on unstructured tetrahedral meshes in the following. Please refer to Dumbser and Käser (2006) for details of the numerical scheme.

The three-dimensional elastic wave equation for an isotropic medium can be expressed using a first-order velocity-stress formulation and written in a compact form as

$$\partial_t Q_p + A_{pq}\partial_x Q_q + B_{pq}\partial_y Q_q + C_{pq}\partial_z Q_q = 0 \tag{1.7}$$

where $\boldsymbol{Q}$ is a 9-vector consisting of the 6 independent components of the symmetric stress tensor and the velocity vector $\boldsymbol{Q} = (\sigma_{xx}, \sigma_{yy}, \sigma_{zz}, \sigma_{xy}, \sigma_{yz}, \sigma_{xz}, u, v, w)^{\mathrm{T}}$ and A_{pq}, B_{pq} and C_{pq} are space-dependent 9×9 sparse matrices with the nonzero elements given by the space-dependent Lamé parameters and the buoyancy (i.e., the inverse of the density). Summation for all repeated indices is implied in all equations. The seismic source and the free-surface and absorbing boundary conditions can be considered separately as shown in Käser and Dumbser (2006) and Dumbser and Käser (2006).

Inside each tetrahedral element $T^{(m)}$, the numerical solution Q_h can be expressed as a linear combination of space-dependent and time-independent polynomial basis functions $\Phi_l(\xi,\eta,\zeta)$ of degree N with support on $T^{(m)}$,

$$[Q_h^{(m)}]_p(\xi,\eta,\zeta,t) = \widehat{Q}_{pl}^{(m)}(t)\,\Phi_l(\xi,\eta,\zeta) \tag{1.8}$$

where $\widehat{Q}_{pl}^{(m)}(t)$ are time-dependent DOFs and ξ,η,ζ are coordinates in the reference element T_E. Explicit expressions for the orthogonal basis functions $\Phi_l(\xi,\eta,\zeta)$ on a reference tetrahedral element are given in Cockburn *et al.* (2000) and the appendix A of Käser *et al.* (2007). Bringing Equation (1.8) into Equation (1.7), multiplying both sides with a test function Φ_k, integrating over an element $T^{(m)}$ and then applying integration by parts, we obtain

$$\begin{aligned}&\int_{T^{(m)}} \mathrm{d}V(\Phi_k \partial_t Q_p) + \int_{\partial T^{(m)}} \mathrm{d}S(\Phi_k F_p^h)\\ &- \int_{T^{(m)}} \mathrm{d}V(\partial_x \Phi_k A_{pq} Q_q + \partial_y \Phi_k B_{pq} Q_q + \partial_z \Phi_k C_{pq} Q_q) = 0\end{aligned} \tag{1.9}$$

The numerical flux F_p^h between the element $T^{(m)}$ and one of its neighboring elements, $T^{(m_j)}, j = 1, 2, 3, 4$, can be computed from an exact Riemann solver,

$$\begin{aligned} F_p^h = & \frac{1}{2} T_{pq}^j \left(A_{qr}^{(m)} + \left| A_{qr}^{(m)} \right| \right) \left(T_{rs}^j \right)^{-1} \widehat{Q}_{sl}^{(m)} \Phi_l^{(m)} \\ & + \frac{1}{2} T_{pq}^j \left(A_{qr}^{(m)} - \left| A_{qr}^{(m)} \right| \right) \left(T_{rs}^j \right)^{-1} \widehat{Q}_{sl}^{(m_j)} \Phi_l^{(m_j)} \end{aligned} \tag{1.10}$$

where T_{pq}^j is the rotation matrix that transforms the vector $\boldsymbol{Q}$ from the global Cartesian coordinate to a local normal coordinate that is aligned with the boundary face between the element $T^{(m)}$ and its neighbor element $T^{(m_j)}$. Bringing Equation (1.10) into Equation (1.9) and converting all the integrals from the global xyz-system to the $\xi\eta\zeta$-system in the reference element T_E through a coordinate transformation, we obtain the semi-discrete discontinuous Galerkin formulation,

$$\begin{aligned} & |J| \partial_t \widehat{Q}_{pl}^{(m)} M_{kl} - |J| \left(A_{pq}^* \widehat{Q}_{ql}^{(m)} K_{kl}^{\xi} + B_{pq}^* \widehat{Q}_{ql}^{(m)} K_{kl}^{\eta} + C_{pq}^* \widehat{Q}_{ql}^{(m)} K_{kl}^{\zeta} \right) \\ & + \frac{1}{2} \sum_{j=1}^{4} |S_j| T_{pq}^j \left(A_{qr}^{(m)} + \left| A_{qr}^{(m)} \right| \right) \left(T_{rs}^j \right)^{-1} \widehat{Q}_{sl}^{(m)} F_{kl}^{-,j} \\ & + \frac{1}{2} \sum_{j=1}^{4} |S_j| T_{pq}^j \left(A_{qr}^{(m)} - \left| A_{qr}^{(m)} \right| \right) \left(T_{rs}^j \right)^{-1} \widehat{Q}_{sl}^{(m_j)} F_{kl}^{+,j,i,h} = 0 \end{aligned} \tag{1.11}$$

where $|J|$ is the determinant of the Jacobian matrix of the coordinate transformation being equal to 6 times the volume of the tetrahedron, $|S_j|$ is the area of face j between the element $T^{(m)}$ and its neighbor element $T^{(m_j)}$, A_{pq}^*, B_{pq}^* and C_{pq}^* are linear combinations of A_{pq}, B_{pq} and C_{pq} with the coefficients given by the Jacobian of the coordinate transformation, $M_{kl}, K_{kl}^{\xi}, K_{kl}^{\eta}$ and K_{kl}^{ζ} are the mass and stiffness matrices and the flux matrices are given by

$$F_{kl}^{-,j} = \int_{\partial (T_E)_j} \left[\Phi_k \left(\xi^{(j)}(\chi, \tau) \right) \Phi_l \left(\xi^{(j)}(\chi, \tau) \right) \right] \mathrm{d}\chi \mathrm{d}\tau, \quad \forall 1 \leqslant j \leqslant 4 \tag{1.12}$$

$$\begin{aligned} F_{kl}^{+,j,i,h} = & \int_{\partial (T_E)_j} \left[\Phi_k \left(\xi^{(j)}(\chi, \tau) \right) \Phi_l \left(\xi^{(i)} \left(\widetilde{\chi}^{(h)}(\chi, \tau), \widetilde{\tau}^{(h)}(\chi, \tau) \right) \right) \right] \mathrm{d}\chi \mathrm{d}\tau, \\ & \forall 1 \leqslant i \leqslant 4, \forall 1 \leqslant h \leqslant 3 \end{aligned} \tag{1.13}$$

The mass, stiffness and flux matrices are all computed on the reference element, which means that they can be evaluated analytically beforehand using a computer algebra system (e.g., Maple, Mathematica) and stored on disk.

If we project Equation (1.7) onto the DG spatial basis functions, the temporal derivative of the DOF can be expressed as

$$\partial_t \widehat{Q}_{pn}(t) = (-M_{nk}^{-1} K_{lk}^{\xi} A_{pq}^* - M_{nk}^{-1} K_{lk}^{\eta} B_{pq}^* - M_{nk}^{-1} K_{lk}^{\zeta} C_{pq}^*) \widehat{Q}_{ql}(t)$$

and the m-th temporal derivative can be determined recursively as

$$\partial_t^m \widehat{Q}_{pn}(t) = (-M_{nk}^{-1} K_{lk}^{\xi} A_{pq}^* - M_{nk}^{-1} K_{lk}^{\eta} B_{pq}^* - M_{nk}^{-1} K_{lk}^{\zeta} C_{pq}^*) \partial_t^{m-1} \widehat{Q}_{ql}(t) \tag{1.14}$$

The Taylor expansion of the DOF at time t^n is

$$\widehat{Q}_{pn}(t) = \sum_{m=0}^{N} \frac{(t-t^n)^m}{m!} \partial_t^m \widehat{Q}_{pn}(t^n)$$

which can be integrated from t^n to t^{n+1},

$$I_{pnql}(\Delta t)\widehat{Q}_{ql}(t^n) \equiv \int_{t^n}^{t^{n+1}} \widehat{Q}_{pn}(t)\mathrm{d}t = \sum_{m=0}^{N} \frac{\Delta t^{m+1}}{(m+1)!} \partial_t^m \widehat{Q}_{pn}(t^n) \tag{1.15}$$

where $\Delta t = t^{n+1} - t^n$, and $\partial_t^m \widehat{Q}_{pn}(t^n)$ can be computed recursively using Equation (1.8).

Considering Equation (1.15), the fully discretized system can then be obtained by integrating the semi-discrete system, Equation (1.11), from t^n to t^{n+1},

$$\begin{aligned}
&|J| \left[\left(\widehat{Q}_{pl}^{(m)} \right)^{n+1} - \left(\widehat{Q}_{pl}^{(m)} \right)^{n} \right] M_{kl} \\
&= |J|(A_{pq}^* K_{kl}^{\xi} + B_{pq}^* K_{kl}^{\eta} + C_{pq}^* K_{kl}^{\zeta}) I_{qlmn}(\Delta t)(\widehat{Q}_{mn}^{(m)})^n \\
&\quad -\frac{1}{2} \sum_{j=1}^{4} |S_j| T_{pq}^j (A_{qr}^{(m)} + |A_{qr}^{(m)}|)(T_{rs}^j)^{-1} F_{kl}^{-,j} I_{slmn}(\Delta t)(\widehat{Q}_{mn}^{(m)})^n \\
&\quad -\frac{1}{2} \sum_{j=1}^{4} |S_j| T_{pq}^j (A_{qr}^{(m)} - |A_{qr}^{(m)}|)(T_{rs}^j)^{-1} F_{kl}^{+,j,i,h} I_{slmn}(\Delta t)(\widehat{Q}_{mn}^{(m_j)})^n
\end{aligned} \tag{1.16}$$

Equation (1.16), together with Equations (1.14) and (1.15), provides the mathematical foundation for our GPU implementation and optimization.

1.1.4 Accelerating the ADER-DG Method Using GPUs

The implementation and optimization of the ADER-DG method on a single GPU was documented in Mu *et al.* (2013). Extending the implementation to a cluster of GPUs is relatively straightforward. We give a brief summary in the following and demonstrate the performance of our multi-GPU CUDA-MPI code using specific examples.

Prior to running our wave-equation solver, a tetrahedral mesh for the entire modeling domain was generated on a CPU using the commercial mesh generation software "GAMBIT". The mesh generation process is fully automated and the generated tetrahedral mesh conforms to all discontinuities built into the modeling geometry, including irregular surface topography and subsurface fault structures. The entire mesh was then split into subdomains, one per GPU, using