Title

Global Full-Waveform Tomography Using the Spectral Element Method: New Constraints on the Structure of Earth's Interior

Author

French, Scott Winfield

Publication Date

2014

Global Full-Waveform Tomography Using the Spectral Element Method: New Constraints on the Structure of Earth's Interior

by

Scott Winfield French

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Earth and Planetary Science
and the Designated Emphasis
in
Computational Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Barbara Romanowicz, Chair
Professor Bruce Buffett
Professor James Demmel

Spring 2014

Global Full-Waveform Tomography Using the Spectral Element Method: New Constraints on the Structure of Earth's Interior

Abstract

Global Full-Waveform Tomography Using the Spectral Element Method: New Constraints on the Structure of Earth's Interior

by

Scott Winfield French

Doctor of Philosophy in Earth and Planetary Science
and the Designated Emphasis
in
Computational Science and Engineering

University of California, Berkeley

Professor Barbara Romanowicz, Chair

The radially anisotropic shear-velocity structure of the earth's mantle provides a critical window on the interior dynamics of the planet, with isotropic variations interpretable in terms of thermal and compositional heterogeneity and anisotropy in terms of flow. Indeed, more than 30 years after the advent of global seismic tomography, many open questions remain regarding the dual roles of temperature and composition in shaping mantle convection, as well as interactions between different dominant scales of convective phenomena. To this end, we use full-waveform inversion of the long-period teleseismic wavefield to image radially anisotropic shear-wave velocity at the scale of the entire globe. In particular, we use a technique which we have termed the "hybrid" waveform-inversion approach, which combines the accuracy and generality of the spectral finite element method (SEM) for *forward* modeling of the global wavefield, with non-linear asymptotic coupling theory for efficient *inverse* modeling. This hybrid technique helps considerably in making SEM-based global waveform inversion tractable, as it allows for the use of a rapidly converging Gauss-Newton scheme for optimization of the underlying seismic model. We take additional steps to reduce the cost of these inversions using novel techniques for treatment of the earth's crust. Namely, naive modeling of thinly-layered crustal structure can lead to an overly restrictive time-stability condition in the SEM, which in turn drives up the cost these simulations. Instead, taking advantage of the physics of long-period wave propagation, we introduce alternative parameterizations of crustal structure which appear identical to the wavefield, but relax these constraints on stability.

We approach this imaging problem in an iterative fashion, hoping to learn something about the earth's interior at each step. First, we present our work focused on the upper mantle and transition zone ($\leq$ 800km depth) in the form of the global model SEMum2, discussing both its development and general properties. Second, we take a detailed look at

novel structures in SEMum2 – namely, never-before-seen low-velocity structures in the upper mantle beneath the ocean basins, showing intriguing correlations with other geophysical observables (e.g. absolute plate motions and the geoid). Third, we move on to imaging of the whole mantle, examining the relationships between different scales of convective phenomena in the upper and lower mantle, and particularly those previously seen in SEMum2. Finally, we discuss novel computational aspects of our work, focusing specifically on applications of the partitioned global address-space model of parallel computation to the large-scale data-driven calculations underlying our hybrid inversion approach.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, Barbara Romanowicz. Her guidance and support over the last five years has been invaluable, and I sincerely appreciate the freedom she has granted me to focus on the problems I find most exciting. Of course, I would also like to thank the other two members of my dissertation committee, Bruce Buffett and Jim Demmel (EECS) – both of whom provided valuable feedback not only on the material in this thesis, but also as members of my qualifying exam committee and in their courses that I've had the pleasure of taking.

I would like to thank the collaborators with whom I've had the opportunity to work over the past five years – particularly Ved Lekic and Huaiyu Yuan for their insights and for many hours of stimulating conversation. In addition, I would like to thank Katherine Yelick (EECS / LBL) and Yili Zheng (LBL) for the opportunity to collaborate on some of the computational aspects featured in this work, and whose feedback benefited the latter considerably. I would also like to more broadly acknowledge past and present members of the Global Seismology Group, including Yann Capdeville for providing us with the coupled spectral element code used extensively in this work.

I cannot imagine having completed this journey without Natalia Bilenko, with whom I've been immeasurably lucky to share my life for the past eight years and whose support throughout has been limitless. I would also like to thank my family, whose continual support and encouragement, even from 3000 miles away, has also been fundamental to making this journey possible. I would like to acknowledge my fellow students in EPS and the BSL, as all have contributed to making this such a great environment to work and learn. I can say the same for the BSL and EPS support staff, for their assistance with everything from administrative issues to keeping our computing resources up and running.

I would like to thank the National Science Foundation for their support through the Graduate Research Fellowship Program, as well as additional financial support through NSF grant EAR-0738284. In addition, I would like to acknowledge support from the BSL through the Tocher-Byerly Fellowship. Finally, much of the work presented in this thesis would not have been possible without access to the computational resources at the National Energy Research Scientific Computing Center (NERSC). Thus, I would like to thank both the Department of Energy Office of Science for granting us access to these resources, as well as NERSC for maintaining them and providing user support.

Chapter 1

Introduction

Global seismic tomography provides some of the most valuable constraints on the interior structure and dynamics of the earth, and for more than 30 years (e.g. Dziewonski et al., 1977), has yielded images of the seismic velocity structure with progressively finer resolution (e.g. Thurber and Ritsema, 2007). These gains have been made using a diverse range of methods and theories (ray theory, normal-mode perturbation theory, etc.), as well as associated types of seismological data (travel times of discrete body-wave phases, surface-wave dispersion, long-period waveforms, etc.), but largely agree in the *isotropic* structure of the earth's mantle at long wavelengths (Dziewonski, 2005). However, as noted by numerous recent studies (e.g. Lekić and Romanowicz, 2011a; Ritsema et al., 2011), significant differences persist at shorter wavelengths, and particularly in the *anisotropic* structure (e.g. Kustowski et al., 2008a).

These discrepancies may arise from a number of sources, but chief among them is the "null-space" associated with different methods for processing of seismic data (and in turn the type of misfit function used in seismic tomography), as well as the theoretical treatment of wave propagation. For example, while delay times of well-isolated seismic phases are easily measured (e.g. by cross-correlation: Su and Dziewonski, 1992), they implicitly disregard information contained in the scattered wavefield and, further, are not applicable in the case of interfering phases. Similarly, ray-theoretical treatment of seismic delay times derived from broadband waveform data, where finite frequency effects must be considered, ignores diffractive phenomena such as wavefront healing (Nolet and Dahlen, 2000) – often leading to systematic under-estimation of the amplitudes of heterogeneity in seismic velocities.

Fortunately, these shortcomings are not universal to global seismic tomography and may be remedied with more advanced techniques. Namely, waveform inversion, which constrains seismic structure by optimizing the fit between observed and predicted seismic recordings (requiring simulation of the physics of wave propagation in order to deliver these predictions), can readily exploit the information content of the full wavefield, including scattered and complex interfering phases, as well as amplitudes. We note that these advantages were recognized early on, and waveform inversion has been used effectively at the global scale for nearly 30 years, though typically in the context of normal-mode perturbation theory under progressively more advanced approximations (e.g. Woodhouse and Dziewonski, 1984; Li and

Romanowicz, 1995). Further, within the past decade, the spectral finite element method (SEM) has become the tool of choice for numerical simulation of the seismic wavefield at the scale of the entire globe (e.g. Komatitsch and Tromp, 2002a,b; Chaljub et al., 2003) – solving the elastodynamic equations governing global wave propagation directly, without the asymptotic approximations that have historically been applied. Owing to the rapid expansion in high-performance computing resources available to earth science researchers, SEM-based waveform inversion at the global scale has recently become possible (e.g. upper-mantle model SEMum: Lekić and Romanowicz, 2011a), but remains computationally formidable.

In this dissertation, we model radially anisotropic shear-wave velocity structure in the earth's mantle using the particular combination of advanced methods outlined above: namely, global-scale full-waveform inversion based on the SEM. We do this in an iterative fashion, progressing from inversion in the upper-mantle and transition-zone only ($\leq$ 800 km depth) to that at the scale of the entire mantle (2891 km depth), with the overall goal of learning something new about the earth's interior structure at each step. Concurrently, we explore new methods for fast and accurate seismic modeling at the global scale, as well as techniques to overcome computational hurdles linked to the problem sizes considered in the case of whole-mantle modeling – both of which have proven critical to rendering the present work tractable.

In Chapter 2, we present SEMum2, a radially anisotropic shear-velocity model of the upper mantle and transition zone. We discuss the particular methods used in developing SEMum2, namely the *hybrid* waveform inversion technique (Lekić and Romanowicz, 2011a; French et al., 2013), and focus specifically on our novel treatment of crustal structure and the computational savings that it enables. We also present an extensive discussion of SEMum2 model structure, primarily motivated by comparisons with other global *and* regional-scale seismic models. Importantly, we find that SEMum2 is characterized by larger, more realistic amplitudes of shear-velocity heterogeneity than typically seen in global tomographic models, commensurate with nominally higher-resolution regional-scale inversions.

We continue our discussion of SEMum2 in Chapter 3, though now focused on intriguing features in the oceanic upper mantle: finger-like bands of significantly lower than average shear-wave velocities that extend for thousands of kilometers aligned with the direction of absolute plate motion. These features are most prominent in the 200-350 km depth range, and exhibit a nearly regular periodicity of $\sim$ 2000 km, which correlates well with long-wavelength undulations observed in the earth's geoid. In contrast to these horizontally elongate patterns, shear velocity structure in SEMum2 abruptly changes below $\sim$ 400 km depth, and is characterized by quasi-vertical conduit-like features originating from the lower mantle. Motivated by these observations, we discuss possible causative dynamics, including secondary convection and viscous fingering phenomena. To support these discussions, we also present an extensive analysis of model uncertainties, considerably more thorough than those typically accompanying published tomographic models.

In Chapter 4, we extend the hybrid inversion technique employed in Chapters 2 and 3 to imaging of the entire mantle. We discuss the necessary modifications to our approach to enable whole-mantle modeling, and introduce particular computational issues that arise, returning to the latter in greater detail in Chapter 5. We present preliminary results for our whole-mantle inversion, focusing in particular on exploring the relationships between upper-mantle and transition-zone structure originally seen in SEMum2 and that now resolved in the lower mantle. More broadly, we discuss implications for the dominant scales of mantle convection and their associated depth-extent, as well as the morphology of the quasi-vertical low-velocity conduits and the nature of their interactions with the upper 1000 km. Further, we present a detailed overview of our ongoing analysis, which (like SEMum2) is largely focused on uncertainty estimation in order to qualify our inferences.

In Chapter 5, we present the design and evaluation of a new distributed matrix data structure which allows our hybrid inversion technique to be employed for whole-mantle imaging – namely, by enabling scalable distributed assembly of the waveform Hessian matrix using non-linear asymptotic coupling theory (NACT: Li and Romanowicz, 1995). We developed our novel solution using UPC++ (Zheng et al., 2014), an experimental extension to the C++ language which introduces support for parallel computations under the partitioned global address space (PGAS) model. We present scaling results for NACT Hessian estimation on up to 12,288 cores of NERSC *Edison*. We also draw comparisons with an alternative implementation based on MPI one-sided remote memory access (RMA) operations, and provide a detailed discussion of challenges encountered during development – particularly in reasoning about performance and progress.

Finally, in Chapter 6, we conclude with an integrative summary of the work presented here, as well as future directions for the next generation of global tomographic models developed using these techniques.

Chapter 2

SEMum2: A second-generation global upper-mantle radially anisotropic shear-velocity model constructed using the Spectral Element Method

2.1 Introduction

Seismic tomography provides some of the clearest constraints on the present day dynamics of the Earth's interior. The earliest global models of three-dimensional seismic velocity structure (Dziewonski et al., 1977; Woodhouse and Dziewonski, 1984) were the first to deliver images of long-wavelength heterogeneity (> 5000km), focusing on the lower and upper mantle respectively. Since then, several generations of models have been developed with steadily improving resolution. Recent global studies (e.g. Panning and Romanowicz, 2006; Simmons et al., 2006; Houser et al., 2008; Kustowski et al., 2008a; Simmons et al., 2010; Ritsema et al., 2011; Lekić and Romanowicz, 2011a), some claiming lateral resolution on the order of 1000km, have extended the interpretable spectrum of model structure, while further confirming the long-wavelength features seen in earlier studies, which show remarkable consistency between models (e.g. Lekić et al., 2012). At the same time, it was recognized that shorter-wavelength features (< 2500km) do not, in general, correlate well across published models (Becker and Boschi, 2002; Dziewonski, 2005) – an observation that remains true today (e.g. Ritsema et al., 2011) and implies considerable uncertainty even at purportedly resolved scales.

Despite short-wavelength discrepancies, the underlying trend of steadily improving resolution persists, and may be broadly attributed to (e.g. Romanowicz, 2008): (1) Improved quantity, quality, and accessibility of global seismic data; (2) Theoretical advances in the treatment of wave propagation in realistic media; and (3) The availability of computational resources capable of supporting these endeavors. There is no consensus regarding the relative importance of these factors, and indeed this topic can prove controversial (e.g. de Hoop and van der Hilst, 2005; Dahlen and Nolet, 2005). Regardless, there are two fundamental concerns faced by ongoing and future tomographic-modeling efforts with the goal of reducing uncertainty at small scales. First, the surface area of the Earth accessible to seismic instrumentation is limited – almost wholly to the continents and ocean islands, and predominantly the northern hemisphere – as is the distribution of global seismic sources. Given this sub-optimal acquisition geometry, quantity and quality of seismic data must be complemented by attempts to fully exploit information contained in seismic waveforms to constrain earth structure. Second, "exact" numerical solution of the elastodynamic equations governing global wave propagation, which are now possible with the spectral element method (SEM: e.g. Komatitsch and Vilotte, 1998), has become a viable replacement for approximate wavefield modeling traditionally used in tomographic inversions. A promising approach to address both of these concerns is to use the SEM to invert full seismic waveforms (e.g. Capdeville et al., 2005). Fully numerical inversions combining the SEM and adjoint-state methods (Tarantola, 1984; Tromp et al., 2005) have been achieved so far at local (Tape et al., 2010) and regional (Fichtner et al., 2009; Zhu et al., 2012) scales. However, adjoint inversions at the global scale remain extremely computationally heavy, and have not yet been fully realized. Meanwhile, we have been exploring somewhat less costly alternatives.

Indeed, the SEMum model of (Lekić and Romanowicz, 2011a) was the first global upper-mantle shear-velocity model developed through waveform inversion of entire long period seismograms based on spectral-element forward modeling. Inversion of full seismic waveforms, as opposed to travel-time or phase-velocity measurements of easily separable phases, allows for more effective use of the rich information content of the complete wavefield (i.e. complex interfering phases, amplitudes). Further, SEM-based modeling avoids certain classes of errors incurred by asymptotic alternatives, such as that induced by wavefront-healing in the case of geometric ray theory (Nolet and Dahlen, 2000). The SEMum model was developed using a *hybrid* inversion scheme, combining the accuracy of SEM forward-modeling with efficient normal-mode based sensitivity kernels (Li and Romanowicz, 1995), thus posing a considerably reduced computational load over adjoint approaches. SEMum is characterized by stronger lateral variations of shear velocity in the upper 250km of the mantle relative to other recent global models developed using approximate forward modeling theories (e.g. Panning and Romanowicz, 2006; Kustowski et al., 2008a; Simmons et al., 2010; Ritsema et al., 2011), and more accurately captures the true lateral variations in velocity inferred from local/regional studies. Indeed, SEMum exhibits intriguing patterns of structure – particularly in the oceans – not observed in earlier models developed using approximate forward treatments of wave propagation (Lekić and Romanowicz, 2011b; French et al., 2013). These latter features will be revisited in detail in Chapter 3.

To further render the SEM-based hybrid waveform inversion tractable at global scales, SEMum was developed using a smooth anisotropic crustal layer of uniform 60km thickness. This smooth crustal model, designed to fit global surface-wave dispersion data, significantly reduced the cost of SEM computation over direct implementation of a "realistic" thinly layered crust (Section 2.2.3). However, while such a compromise with respect to crustal-layer thickness is justifiable in the continents, it has the potential to complicate interpretation of shallow (<80km depth) upper-mantle structure in the oceans. To address this concern, we now present an updated version of SEMum (hereafter: SEMum2), which has been obtained using a smooth crustal model that is conceptually similar, but with a more geologically plausible *laterally varying* thickness. In Section 2, we discuss our inversion methodology, with an emphasis on what differentiates SEMum2 from SEMum – particularly the treatment of crustal structure. In Section 3, we quantify model performance and assess resolution. Finally, in Section 4, we present the updated model and discuss broad characteristics of the 3D structure obtained.

2.2 Methods

Our approach to data processing and inversion largely mirrors that developed in (Lekić and Romanowicz, 2011a), as well as previous Berkeley tomographic efforts (e.g. Li and Romanowicz, 1996; Mégnin and Romanowicz, 2000; Panning and Romanowicz, 2006). Here, we provide an overview of the inversion procedure, referring the reader to Lekić and Romanowicz (2011a) for more detailed discussion, and instead focus on the elements that distinguish SEMum2 from SEMum.

2.2.1 Parameterization and Data

Like SEMum, we seek 3D perturbations to Voigt-average shear-wave velocity V_S and the shear-wave radial-anisotropy parameter $\xi = V_{SH}^2/V_{SV}^2$. Perturbations to the remaining four coefficients of a radially anisotropic medium (V_P, $\phi = V_{PV}^2/V_{PH}^2$, η, and ρ) are scaled based on empirical relationships as in Montagner and Anderson (1989). Though we used SEMum as our starting model, we first chose to adjust the 1D reference model of SEMum in order to: (1) remove the Moho associated with the 60km crustal layer; and (2) adopt the transition-zone structure of the STW105 reference model of Kustowski et al. (2008a). This latter step has the effect of simplifying SEM mesh design relative to the previous thermally and compositionally constrained reference structure (Cammarano and Romanowicz, 2007). Following the above modifications, the resulting reference model was validated against a compilation of normal-mode eigenfrequency measurements available from the Reference Earth Model project (`http://igppweb.ucsd.edu/~gabi/rem.html`). In Figure 2.1, we show profiles of 1D reference V_S structure – isotropic and radially anisotropic parts – from SEMum, the SEMum2 starting model (modified as described above), and the final-iteration SEMum2 model. We also show PREM (Dziewonski and Anderson, 1981) for comparison. For a detailed discussion of the design of the SEMum 1D reference model, see Lekić and Romanowicz (2011a). From Figure 2.1, it is clear that all three SEMum-series reference models differ significantly from PREM, reflecting the influence of thermal and compositional constraints in their initial construction and the removal of the 220km discontinuity which is present in PREM, but is not a global feature (e.g. Gu et al., 2001). Additionally, there are noticeable differences between the SEMum2 initial and final reference models, reflecting progressive evolution of (1D) mean model structure with additional inversion iterations.

Parameterization of SEMum2 is also similar to that of SEMum (Figure 2.2), with 20 cubic b-spline basis functions in depth between the Moho and CMB (e.g. Mégnin and Romanowicz, 2000), and spherical splines laterally (Wang and Dahlen, 1995). Also like SEMum, we invert only for upper-mantle and transition-zone structure (colored b-splines in Figure 2.2), leaving the lower mantle below 800 km fixed to that of the model SAW24B16 of Mégnin and Romanowicz (2000), a 3D long wavelength model that was developed using full-waveform inversion, albeit with a theoretical formalism based on normal mode perturbation theory (Li and Romanowicz, 1995). A finer lateral parameterization is used for perturbations in V_S than in ξ, reflecting the expected disparity in resolution of these parameters: 10,242 spherical-spline nodes for $\delta \ln\ V_S$ ($< 2°$ lateral spacing) and 642 for $\delta \ln\ \xi$ ($< 8°$ lateral spacing), equivalent to spherical harmonic angular orders 96 and 24, respectively. The choice of 10242 spline nodes for $\delta \ln\ V_S$ differs from that used in SEMum (2562 nodes $\simeq$ angular order 48). We justify this refinement as follows: (1) we determine model smoothness purely through the estimated prior covariance operator, not the parameterization; (2) we have confidence in the resolution of isotropic structure that is possible with our dataset (Section 2.3.3); and (3) we have developed the ability to factorize larger dense linear systems governing the model update (Section 2.2.2) – as large as 120,000 × 120,000.

Given our choice of parameterization, we may express continuous perturbations to the parameter $X \in \{V_S, \xi\}$ as a function of radius, colatitude, and longitude in terms of the associated set of discrete spline coefficients c^X_{pq} as

$$\delta \ln X(r,\theta,\phi) = \sum_{p}^{N_p} \sum_{q}^{N_q} c^X_{pq} \beta_p(\theta,\phi) \nu_q(r) \tag{2.1}$$

where $\beta_p(\theta,\phi)$ and $\nu_q(r)$ represent the lateral spherical-spline and radial cubic b-spline basis functions, respectively. Our model may be represented by the concatenated vector $\mathbf{m}$ of discrete coefficients $c^{V_S}_{pq}$ and c^{ξ}_{pq}.

Our dataset is identical to that used in the development of SEMum, consisting of long-period (60 - 400s) three-component accelerograms of 203 well-distributed global earthquakes ($6.0 \leq \mathrm{Mw} \leq 6.9$), as well as global group-velocity dispersion maps at periods between 25 and 150s (M. Ritzwoller, personal communication 2009). The group-velocity maps are those of Ritzwoller et al. (2002) and a discussion of uncertainties associated with the underlying dispersion measurements may be found in Shapiro and Ritzwoller (2002). Dispersion at periods $\leq$60s serves to constrain our homogenized crustal model (see Section 2.2.3), while, for consistency, the entire period range is included in our inversion for mantle structure. We note that the resolving power of the dispersion dataset is rather weak at depths below roughly 100km and that upper-mantle structure is primarily constrained by the waveform data, which includes longer-period fundamental-mode surface waves as well as surface-wave overtones. Prior to inversion, the full traces that comprise the waveform dataset are first bandpass filtered using a cosine taper with cutoff periods at 60 and 400s and corners at 80 and 250s, and then partitioned into wave-packets containing trains of phases with similar amplitude (Li and Romanowicz, 1996). This wavepacket-based preprocessing scheme allows for direct inversion of the waveform data, without bias toward fitting large-amplitude phases (e.g. fundamental-mode surface waves), and further permits per-wavepacket weights reflecting uniqueness of constraint on earth structure. These are indeed necessary conditions for attaining resolution beyond 300km depth using overtone surface waveforms (Mégnin and Romanowicz, 1999). A breakdown of our dataset by wavepacket classification (e.g. fundamental modes, overtones, mixed) and component can be found in Table 2.1. Source mechanisms from the Global CMT Project (`http://www.globalcmt.org`) are assumed and remain static throughout the inversion. While it would be desirable to invert for updated sources as the mantle model evolves, this represents a significant computational investment over a purely structural SEM-based inversion. Further, we note that: (1) near-nodal paths, strongly affected by source error in addition to amplitude-focusing effects, are largely rejected during picking; and (2) preliminary source inversion for a subset of the full 203 events indicate that changes in source mechanism are in general quite small, as are changes in source depth ($\leq$3km on average).

2.2.2 Hybrid waveform inversion

Waveform inversion is a non-linear problem, which we solve iteratively following the generalized least-squares formalism of Tarantola and Valette (1982). We define a discrete L_2 misfit functional

$$2\Phi(\mathbf{m}_k) = \left[\mathbf{d} - \mathbf{g}\left(\mathbf{m}_k\right)\right]^T \mathbf{C}_d^{-1} \left[\mathbf{d} - \mathbf{g}\left(\mathbf{m}_k\right)\right] + \left[\mathbf{m}_p - \mathbf{m}_k\right]^T \mathbf{C}_m^{-1} \left[\mathbf{m}_p - \mathbf{m}_k\right] \tag{2.2}$$

where $\mathbf{m}_k$ represents the k^{th} iterative model estimate, $\mathbf{d}$ and $\mathbf{g}\left(\mathbf{m}_k\right)$ are the observations and model predictions (time-discretized waveforms *and* period-discretized group velocities), $\mathbf{m}_p$ is the model prior (generally taken to be zero perturbation), and $\mathbf{C}_m$ and $\mathbf{C}_d$ represent *a priori* model and data covariance operators. The data covariance matrix $\mathbf{C}_d$ is diagonal, with entries reflecting data quality and uniqueness; see Li and Romanowicz (1996) for a detailed discussion of its particular form. The model covariance matrix $\mathbf{C}_m$ is banded in structure and based upon prescribed model-parameter variances and correlation-lengths, with the latter spatially tuned according to data-coverage density and quality. As in Lekić and Romanowicz (2011a), the element of $\mathbf{C}_m$ jointly describing the state of *a priori* information on model coefficients i and j is given by

$$\left(\mathbf{C}_m\right)_{ij} = \sigma_{ij} \;\; \exp\left(\frac{\cos\Delta_{ij} - 1}{h_{ij}^2}\right) \;\; \exp\left(-\frac{2d_{ij}}{v_{ij}^2}\right) \tag{2.3}$$

where Δ_{ij} is the minor-arc distance separating i and j, d_{ij} their radial separation, and v_{ij} and h_{ij} the averages of their associated radial and lateral correlation lengths, with h_{ij} (km) normalized to the range $0 \leq |\cos\Delta_{ij} - 1| \leq 1$. Note that, because of the locality properties of the spline basis functions, we have made the approximation that model coefficients may be treated directly in the above formulation, not model values. Further, we do not correlate V_S and ξ structure *a priori*, and generally prescribe a single (near-unit) variance for each.

We prescribe radial correlation lengths with a fixed depth dependence, consistent with the expected resolving power of our dataset: model coefficients associated with b-spline basis functions centered above ∼300km depth are assigned correlation lengths of 50km, while 100km is used for those below. Lateral correlation lengths are allowed to vary throughout the model volume, based upon a simple estimator for relative sensitivity. In particular, we take the diagonal entries of the Gauss-Newton Hessian $\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G}$ associated with the data misfit term in eq. 2.2, where $\mathbf{G}$ is the model Jacobian matrix (partial derivatives of the data with respect to the model parameters $\mathbf{m}$), to represent aggregate sensitivity of the data to each model coefficient. Here, contributions from different data are weighted according to the quality and uncertainty estimates appearing in $\mathbf{C}_d$. These values are, in turn, used to select *local* correlation-length estimates by scaling to the interval between prescribed minimum and maximum values for a given model parameter (V_S, ξ). For the purpose of our inversion, we adopted lateral correlation length ranges of 400-800km for V_S and 1200-2400km for ξ. Given the form of eq. (2.3) above, we expect significant correlation between neighboring model coefficients at up to twice these distances. These correlation-length ranges are compatible with the expected resolving power of the dataset (Section 2.3.3).

We solve for successive updates to the current model estimate $\mathbf{m}_k \rightarrow \mathbf{m}_{k+1}$ following the update equation that minimizes eq. (2.2), namely:

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \left(\mathbf{C}_m\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{I}\right)^{-1}\left(\mathbf{C}_m\mathbf{G}^T\mathbf{C}_d^{-1}\left[\mathbf{d} - \mathbf{g}\left(\mathbf{m}_k\right)\right] + \mathbf{m}_p - \mathbf{m}_k\right) \tag{2.4}$$

where we note that the model Jacobian matrix $\mathbf{G}$ has been evaluated *at* the current model estimate $\mathbf{m}_k$ (i.e. at each step, we linearize around $\mathbf{m}_k$ in order to estimate $\mathbf{m}_{k+1}$). We employ a hybrid waveform inversion scheme, in which "exact" forward modeling of the global wavefield (i.e. the waveform elements of $\mathbf{g}\left(\mathbf{m}_k\right)$) is achieved using the spectral element method (SEM: e.g. Komatitsch and Vilotte, 1998; Komatitsch and Tromp, 2002a,b), while approximate sensitivity kernels (i.e. waveform elements of $\mathbf{G}$) are calculated using non-linear asymptotic coupling theory (NACT: Li and Romanowicz, 1995). Group-velocity partial derivatives with respect to model structure are obtained following a non-linear, finite-difference based approach detailed below (see Section 2.2.3). SEM-based forward modeling distinguishes our work from previous generations of global studies, while the inclusion of second-orbit phases, critical to coverage of the southern hemisphere (Section 2.3.1), separates SEMum2 from other ongoing global tomographic efforts using the SEM (e.g. Bozdağ et al., ting).

The hybrid inversion scheme represents a compromise between speed and accuracy, wherein the waveform misfit functional is evaluated exactly, though at considerable cost, while waveform partial derivatives computed using NACT are inexpensive and approximate. Despite the approximate nature of NACT, we note two critical properties of this mode perturbation-theory based approach: the resulting sensitivity kernels (1) accommodate non-linearity due to multiple forward scattering (Romanowicz et al., 2008); and (2) include finite-frequency effects in the 2D great-circle plane, critical to modeling surface-wave overtones and long-period body waves (Li and Romanowicz, 1995). Further, it may be shown that, while erroneous perturbations in the computation of $\mathbf{g}\left(\mathbf{m}_k\right)$ are indeed first-order in eq. (2.4) above, error in the elements of $\mathbf{G}$ is second-order (Lekić and Romanowicz, 2011a). The choice to invest in evaluation of $\mathbf{g}(\mathbf{m}_k)$ using the SEM is thus advantageous, while the normal mode formalism used in the inverse step primarily limits us in the type of model parametrization that we can use. Clearly, our future efforts will strive to remove this limitation. Meanwhile, this compromise allows for (at least) a factor of three savings in computation over fully numerical alternatives (e.g. Tarantola, 1984; Tromp et al., 2005), and the observed convergence of our scheme implies that NACT kernels are sufficiently accurate to justify this choice. Further, NACT yields waveform partial derivatives (elements of $\mathbf{G}$), not the gradient of the misfit functional (eq. 2.2), thereby allowing for quickly converging (Gauss-)Newton-like model optimization schemes, such as that arising above from the generalized least-squares formalism (eq. 2.4). We largely credit this hybrid approach for rendering the inverse problem computationally tractable, in concert with: (1) a modest choice of period range ($T \geq 60$s); and (2) a coupled-mode global SEM implementation (cSEM: Capdeville et al., 2003) combined with a novel treatment of crustal structure. The latter concern - modeling of crustal effects – represents much of our recent effort in developing SEMum2.

2.2.3 Homogenized crustal model

2.2.3.1 Motivation

In the spectral element method, solution accuracy is strongly controlled by fidelity to: (1) the shortest wavelengths present in the wavefield; and (2) spatial variation of the underlying earth model (cf. Komatitsch and Tromp, 2002a,b; Chaljub et al., 2003). The latter concern requires that discontinuities in material properties or their depth-derivatives must be coincident with element faces. Using an explicit time-integration scheme, stability of the SEM is determined by the Courant-Friedrichs-Lewy (CFL) condition – characterizing the maximum stable time step in terms of the minimum ratio between spatial discretization and wave speed. The time-step size in turn determines the overall cost of time integration, given a desired synthetic-record duration (e.g. 10,000s in order to capture most second-orbit fundamental-mode Rayleigh waves). In a pure global SEM, such as `SPECFEM3D_GLOBE` (Tromp et al., 2008), the CFL condition is often dominated by high compressional velocities in the core. If the core is replaced with a pseudo-analytical modal solution such as used here (Capdeville et al., 2003), then small spatial discretization in the thin oceanic crust dominates instead.

The fact that crustal thickness controls our CFL condition motivates us to implement a homogenization scheme (e.g. Capdeville and Marigo, 2007), in which a set of thin discontinuous layers are replaced with a single, smoothly varying anisotropic layer that has an equivalent seismic response at long periods (i.e. the zeroth-order scheme of Backus (1962), in the terminology of Capdeville and Marigo (2007)). Like its predecessor, SEMum2 was developed using a crustal model based on a "hidden" homogenization, designed to fit *observed* surface-wave group-velocity dispersion directly, instead of homogenization of an intermediate or *a priori* crustal model (i.e. "hidden" in that we have no knowledge of the true crust, only its effect on the wavefield measured through surface-wave dispersion). The rationale for this choice is that the available crustal models, i.e. Crust2.0 (Bassin et al., 2000), are known to not be accurate in many parts of the world, where they are based on extrapolations from other regions. In order to permit relatively large time steps in the SEM, SEMum employed a uniform 60km thickness for the crustal layer (Lekić and Romanowicz, 2011a). Here, we adopt a more geologically plausible laterally varying thickness, at the expense of a more restrictive CFL condition. Starting from Crust2.0 Moho depth (Bassin et al., 2000), we restrict crustal thickness to the interval between 30 and 60km and filter this fictitious Moho surface to twice the lateral resolution of our SEM mesh. The filtering step, valid at long periods, prevents spatial aliasing of Moho topography in the SEM. Given this prescribed geometry, we use a two-step model-space sampling approach to find a radially anisotropic crustal-layer structure that fits the dispersion data. Crustal structure is parameterized in depth using degree-4 Lagrange polynomials with Gauss-Lobatto-Legendre interpolation points (hereafter: GLL) of the same type used in the SEM – a similar approach to that taken by Fichtner and Igel (2008) in their alternative implementation of a smooth crustal layer.

2.2.3.2 Development

The crustal model is generated following a two-step procedure. A high-level overview of this process, as well as its role in the broader mantle inversion, appears in Figure 2.3. As shown, our crustal model implementation is *iterative* in nature - receiving periodic updates in order to remain consistent with the evolving mantle model below.

Step 1: Crustal layer initialization

We begin by defining a space of admissible *isotropic* GLL-parameterized 1D crustal models, characterized by uniform distributions over

1. *GLL-nodal shear-velocity*: $3.0 \leq V_S \leq 4.6$ km/s
2. *Restricted Moho depth*: $30 \leq h_m \leq 60$ km
3. *Realistic bathymetry*: $0 \leq h_s \leq 6$ km

The above range of V_S nodal values was chosen based upon examination of realistic (e.g. Crust2.0) crustal layers overlying typical upper-mantle structure at depths $\leq$ 60km, which have in turn been vertically smoothed with a $\sim$ 10km sliding window. V_P and ρ nodal values for the crustal layer are scaled from V_S following the relations of Brocher (2005), leading to a family of 1D models of the crustal layer parameterized as

$$V_S(r) = \sum_{n=0}^{4} \hat{V}_S^n \ell^n \left(x(r)\right) \tag{2.5}$$

$$V_P(r) = \sum_{n=0}^{4} \hat{V}_P^n \ell^n \left(x(r)\right) \tag{2.6}$$

$$\rho(r) = \sum_{n=0}^{4} \hat{\rho}^n \ell^n \left(x(r)\right) \tag{2.7}$$

where $\{\hat{V}_S^n, \hat{V}_P^n, \hat{\rho}^n : n = 0, \ldots, 4\}$ represent nodal values, r is earth radius within the crustal layer, $\ell^n(x)$ are the GLL-Lagrange basis functions, and $x(r)$ maps crustal radii onto the reference interval $x \in [-1, 1]$. Crustal Q is taken from QL6 (Durek and Ekström, 1996), while ρ, V_P, and V_S are held constant in the water layer, which is not parameterized using the polynomial basis – consistent with the fact that the ocean is modeled as a load in the SEM and is not internally meshed (e.g. Komatitsch and Tromp, 2002b). In a grid-search fashion, we draw approximately 2×10^5 realizations of smooth structure from these *a priori* distributions, and splice the resulting crustal models to our 1D-reference mantle structure below the Moho. For each model realization, we calculate fundamental-mode Love and Rayleigh wave group-velocity dispersion.

Group-velocity dispersion maps (25-60s period) are resampled on the same set of 10,242 nodes against which the V_S spherical splines are registered (equivalent to $\sim 2^\circ$ lateral spacing). At each node, we select the model realization that best fits the resampled Rayleigh and Love wave dispersion simultaneously, while honoring local crustal thickness from the filtered Moho surface introduced above and the bathymetry from a similarly filtered ETOPO2 (`http://www.ngdc.noaa.gov`). We are able to perform this matching step because of manner in which we parameterized the above sampling. In particular, for each realization of *volumetric* crustal-layer structure (i.e. velocity and density), letting $\mathbf{c}_i$ denote the i^{th} realization, we have calculated synthetic dispersion curves over a regular grid of Moho depth h_m and bathymetry h_s – thus defining a discrete group-velocity "surface" $U_{L,R}^{syn}[\mathbf{c}_i](h_m, h_s)$ at any modeled period. We may therefore obtain an accurate estimate of $U_{L,R}^{syn}[\mathbf{c}_i]$ at arbitrary (h_m, h_s), within the bounds of the sampling, using bicubic interpolation. Once all crustal-layer realizations are interpolated to exactly the (h_m, h_s) point associated with a given model node location, we can easily select the best-fitting realization – a process that is repeated for all 10,242 nodes. Fit between synthetic and observed group velocities is measured in the L_1-norm to reduce sensitivity to outlier measurements common at short periods, with additional weighting by estimated measurement uncertainty (Shapiro and Ritzwoller, 2002); e.g.

$$\Phi(\mathbf{c}_i) = \frac{N_p(\sigma_R + \sigma_L)}{\sigma_R \sigma_L} \sum_{c \in \{R,L\}} \sum_{j=1}^{N_p} \frac{\left| U_c^{syn}[\mathbf{c}_i](T_j) - U_c^{obs}(T_j) \right|}{\sigma_c} \tag{2.8}$$

where $\mathbf{c}_i$ again represents the i^{th} crustal-layer realization, $\{T_j : j = 1, \ldots, N_p\}$ is the range of periods considered, $U^{syn}_{R,L}[\mathbf{c}_i](T_j)$ and $U^{obs}_{R,L}(T_j)$ are synthetic and observed Rayleigh and Love wave group velocities, and σ_R and σ_L are measurement uncertainties assigned to Rayleigh and Love wave data (40m/s and 50m/s respectively – see Shapiro and Ritzwoller (2002)). The resulting isotropic initial model is thus parameterized radially in GLL-Lagrange interpolants and laterally by linear barycentric interpolation on simplices defined by triples of neighboring model nodes.

Step 2: Iterative inversion for the anisotropic crustal layer

We next perform an iterative inversion for *radially anisotropic* crustal-layer structure in a neighborhood near the 3D initial model. The introduction of anisotropy is required in order to simultaneously fit the Rayleigh and Love wave dispersion more consistently than is possible with the purely isotropic initial model (e.g. Backus, 1962). Importantly, the non-linear, model-space enumeration approach of the first step considerably reduces the required number of inversion iterations relative to a linearized scheme alone.

We follow a generalized least-squares approach similar to that employed in the upper-mantle inversion (eq. (2.4) above). Lateral smoothness is again enforced through the *a priori* model covariance matrix $\mathbf{C}_m$ – here assigning fixed 500 and 1000km correlation lengths to crustal V_S and ξ, respectively. The data covariance matrix $\mathbf{C}_d$ is assumed diagonal, with elements equal to the corresponding σ_R and σ_L values introduced above. As in the first step, group velocities between 25 and 60s are considered.

Unique group-velocity partial derivatives are used at each model node, which are calculated in – and therefore appropriate for – the current crustal *and* mantle-model estimates. Partial derivatives are calculated following a straightforward finite-difference approach, conceptually similar to the “brute force” scheme employed by Rodi et al. (1975) to validate their technique for deriving group-velocity partial derivatives from those for phase velocity. The crustal model at node n may be represented by a vector $\mathbf{c}_n$ containing GLL-nodal V_S and ξ values. The elements of $\mathbf{c}_n$ are successively perturbed by small quantities ϵ and forward-modeled, yielding finite-difference approximations to the desired partial derivatives:

$$\left(\frac{\partial U_c}{\partial \mathbf{c}_n}\right)_{ij} \simeq \frac{U_c^{syn}[\mathbf{c}_n + \epsilon\ \mathbf{e}_j](T_i) - U_c^{syn}[\mathbf{c}_n](T_i)}{\epsilon} \tag{2.9}$$

where c again denotes evaluation for either Rayleigh or Love waves, i is an index over measurement periods, and $\mathbf{e}_j$ denotes the unit vector in the direction of the j^{th} model-vector component. The underlying mantle structure, denoted $\mathbf{m}(\theta_n, \phi_n)$, that comprises cubic b-spline coefficients evaluated beneath node n, i.e.

$$\left(\mathbf{m}(\theta_n, \phi_n)\right)_q = \sum_p^{N_p} c^X_{pq} \beta_p(\theta_n, \phi_n) \text{ where } X \in \{V_S, \xi\}, \tag{2.10}$$

is held fixed and, as mentioned above, is included in the computation of $U_c^{syn}[\cdot]$. Given a local crustal model of N nodal-valued components, the associated group-velocity Jacobians

$$(\mathbf{G}_c)_{ij} = \left(\frac{\partial U_c}{\partial \mathbf{c}_n}\right)_{ij} \tag{2.11}$$

may be recovered in $2(N+1)$ dispersion calculations (including both Rayleigh and Love waves), which are in fact trivial to evaluate in parallel. We find that after 3 iterative crustal updates, the mean absolute misfit over all periods and model nodes falls into the target range of < 50m/s. New crustal models are sought as the inversion for SEMum2 proceeds and underlying upper-mantle structure evolves. We further note that a similar approach is used in estimating partial derivatives with respect to the mantle model, in which case, the crustal model $\mathbf{c}_n$ is held fixed, while the elements of $\mathbf{m}(\theta_n, \phi_n)$ are perturbed and the range of periods T_i extended to 150s instead of 60s. Partial derivatives for the model coefficients of $\mathbf{m}$ themselves may then easily be recovered in a manner analogous to eq. (2.9).

2.2.3.3 Contamination of mantle structure

It is widely recognized that inaccurate treatment of crustal structure can bias mantle models obtained by tomographic inversion, particularly regarding lateral variations in radial anisotropy (e.g. Bozdağ and Trampert, 2008; Ferreira et al., 2010; Lekić et al., 2010). Given a hypothetical "true" crustal model that honors Earth's crust exactly, this effect can be attributed to modelization error: i.e. failure of approximate wave-propagation schemes to reconstruct the strong non-linear effect of realistic crustal structure on the wavefield, especially critical for fundamental-mode surface waves. In the case of SEM-based modeling, assuming the SEM mesh is properly designed (Section 2.2.3.1), modelization error is not a significant concern. Conversely, in the absence of modelization error, inaccurate knowledge of the Earth's crust becomes an impediment to properly modeling its effect on the wavefield. As our smooth crustal layer is not based upon a "realistic" *a priori* model, and by construction cannot honor the thinly layered nature of the crust, one might ask: how confident are we that our smooth crustal layer is equivalent (at long periods) to the true crust?

First, the crustal model is based on *actual* measurements of fundamental-mode surface-wave dispersion, which capture the integrated effect of wave propagation in the Earth's crust on the portion of our waveform dataset most strongly affected by crustal structure. While the underlying dispersion maps do indeed have associated uncertainties (e.g. Barmin et al., 2001; Shapiro and Ritzwoller, 2002), we find this preferable to, for example, unquantifiably uncertain *a priori* models that rely on analogy where constraint on local structure is incomplete (e.g. Crust2.0: Bassin et al., 2000). While constraining the effect of the crustal layer directly from data is theoretically pleasing, questions remain regarding implementation. In particular, our reliance on surface-wave data, combined with our discontinuous parameterization between the crust and mantle models, makes it difficult to constrain the velocity jump across the fictitious Moho. Thus, tradeoffs across the interface could hypothetically bias mantle structure immediately underlying the Moho. One potential solution is to remove the jump entirely, allowing the structure of the homogenized layer to blend smoothly into the mantle (e.g. Fichtner and Igel, 2008). Such a choice may be suitable for the oceans, where the 30km layer is consistently below realistic Moho depth, but the same is not true for the continents, where we have maintained a plausible, albeit laterally smooth, Moho surface. Another option, which we adopt and discuss below (Section 2.3.2), is to introduce independent constraints on mantle-side structure at the interface. Ultimately, given the locality properties of our radial b-spline parameterization, our choice of radial correlation length, and the introduction of independent constraints on mantle-side structure, we conservatively estimate the mantle model to be interpretable below 50km depth in the oceans.

More important, however, is the potential effect on deeper mantle structure – particularly in the oceans, where a 30km crustal-layer thickness does not on first consideration appear realistic. To investigate this, we performed a series of synthetic tests, in which we modeled surface-wave dispersion in realistic oceanic crustal structures similar to Crust2.0 (crustal thickness $\simeq 7 \pm 1$km) atop oceanic mantle structures sampled from SEMum2. As SEMum2 is defined only below 30km, mantle structure is smoothly continued from below using a cubic spline – "filling" the gap between 30km and the oceanic Moho. Admittedly, this choice is not entirely realistic, but represents an effort to maintain consistent radial derivatives in the upper-most portion of the test structure. The resulting synthetic dispersion data were then fit using the above crustal-layer matching scheme, and we compared long-period ($\geq$60s) synthetic waveforms and surface-wave dispersion derived from the "input" synthetic structure with those from the "output" containing a 30km crustal layer. In general, the results confirmed that introduction of the smooth crust should have negligible effect on longer-period fundamental-mode surface waves, either dispersion or waveforms, as illustrated in Figure 2.4. For the subset of test cases in which long-period surface-wave dispersion was adversely affected, the results were biased toward *reduced* phase and group velocities for fundamental-mode Rayleigh and, to a lesser-extent, Love waves. This effect is illustrated in Figure 2.5 for a representative example, where we observe phase-velocity reductions ranging from 4m/s (3m/s) for 400s period Rayleigh (Love) waves to 16m/s (11m/s) at 80s period. Indeed, if this phenomenon were systematic throughout the SEMum2 model, we should expect to see compensatory suspiciously high velocities (relative to other global models) at depths primarily constrained by fundamental modes. However, this is not what we see at all (Section 2.4, Figure 2.13) – instead finding that SEMum2 exhibits stronger low-velocity anomalies throughout, especially above ~ 250km depth. Further, and equally importantly, our dataset is not limited to fundamental modes, but instead has a sizeable overtone surface-wave component. In fact, at depths below ~ 250km, where the sensitivity of fundamental modes begins to wane, our model is primarily constrained by overtones. Upon examination of Figures 2.4 and 2.5, we see that overtone waveforms are unaffected by the presence of the crustal layer, regardless of the effect upon fundamental modes. Thus, we maintain that our use of a homogenized crustal layer should not bias our resulting mantle model at depth, because: (1) we do not see the expected type of systematic compensation in velocity structure; and (2) our extensive use of overtone waveforms should not be susceptible to tradeoffs induced by the crustal layer.

2.2.4 Crustal structure in NACT

Although the SEM may accurately treat arbitrary 3D crustal structure, care must be taken to properly account for crustal effects in our NACT-based sensitivity kernel calculations. Because the SEMum model was derived using a uniform-thickness crustal layer, crust-induced path effects in NACT were easily handled by accounting for volumetric structure alone (perturbations to V_S and ξ). In other words, the strongly non-linear effect of a laterally varying crustal thickness – which is generally more difficult to accurately model with perturbation theory – was not an issue. The variable Moho depth in the SEMum2 model, on the other hand, necessitates accurately treating the non-linear effects of crustal structure on the NACT-based kernels. We adopt a two-part approach to ensure accurate NACT-based modeling: (1) specially formulated corrections for crust-induced path effects that mimic the non-linear response of the crust considerably better than “standard” linear corrections (Lekić et al., 2010); and (2) mantle-structure sensitivity kernels modified to reflect overlying crustal thickness variations. We relegate the detailed description of both schemes to Appendix A.

2.3 Inversion and fits

SEMum2 is the product of three iterative updates to SEMum mantle structure, following the introduction of the new 1D-reference model (Section 2.2.1) and variable-thickness crustal layer (Section 2.2.3), as well as refinement of the model mesh during the final iteration (Section 2.2.1). The required four rounds of SEM forward modeling comprise over 489,000 CPU-hours. This estimate does not include the cost of NACT and group-velocity partial derivatives, nor the parallel dense-matrix factorizations required to solve for the model update (eq. 2.4).

2.3.1 Model performance

In Table 2.1, we summarize performance of SEMum2 in terms of waveform variance reduction VR, which we define as one minus the squared 2-norm of the waveform residual normalized by that of the data (i.e. the residual variance RV):

$$VR\left(\mathbf{m}_k\right) = 1 - RV\left(\mathbf{m}_k\right) = 1 - ||\mathbf{d} - \mathbf{g}\left(\mathbf{m}_k\right)||_2^2/||\mathbf{d}||_2^2 \ . \tag{2.12}$$

VR values are presented for each waveform component (L,T,Z) and wavepacket classification (fundamental, overtone, and mixed mode) separately, with equivalent SEMum values shown for comparison. Further, only wavepackets with RV below a prescribed threshold are admitted into each iteration, allowing the effective dataset size – also shown in Table 2.1 – to grow as the model converges. We note consistent improvement in fits across all data types, while nearly all exhibit modest dataset growth. Additionally, in Figures 2.6 and 2.7 we show waveforms for a typical event in our dataset (28 August, 2000: Mw 6.8 Banda Sea, 15km depth), as well as synthetics predicted from both SEMum2 and the 1D starting model of SEMum (1D mantle combined with Crust2.0; see Lekić and Romanowicz (2011a)).

	L		T		Z	
	VR (%)	N_{wp}	VR (%)	N_{wp}	VR (%)	N_{wp}
F	68.1 [62.1]	7964 [7968]	77.3 [59.1]	13438 [13192]	69.9 [63.7]	13562 [13525]
H	80.9 [67.2]	14475 [14403]	71.2 [62.9]	14640 [14478]	79.5 [69.7]	22296 [22185]
M	74.9 [67.2]	3434 [3423]	82.2 [68.3]	4596 [4579]	77.6 [70.1]	4951 [4930]

Table 2.1: Final-iteration waveform variance reduction (%) and total wavepacket counts for SEMum2 and [SEMum], listed by component and data type (fundamental-, higher-, and mixed-mode).

In order to investigate how even data coverage is for the waveform data admissible in the final iteration, Figure 2.8 (upper panel) shows ray-path density for fundamental-mode Rayleigh waves, including both minor and major-arc paths. Note that, in practice, the per-wavepacket weighting scheme (Section 2.2.1) further improves the homogeneity of coverage over that shown in Figure 2.8. As mentioned previously, the choice to include major-arc phases in our SEM-based inversion is a costly one, given the length of time integration required (at least 10,000s of synthetic record). However, to illustrate the utility of including major-arc data, especially in constraining structure in the southern hemisphere, Figure 2.8 (lower panel) shows ray-path density of *only* minor-arc fundamental-mode Rayleigh waves. The difference between the two panels is quite clear: particularly striking is the poor coverage of the southern hemisphere without major-arc paths. Also shown in Figure 2.8 are rose diagrams, again for fundamental-mode Rayleigh waves, illustrating the locally uniform azimuthal coverage of the dataset – necessary to avoid mistakenly mapping azimuthal anisotropy into our radially anisotropic inversion.

In addition to waveform fits, it is also useful to quantify the performance of our model in terms of surface-wave dispersion. As in Section 2.2.3.2, we predict dispersion maps by solving the normal-mode eigenproblem in 1D earth models sampled from SEMum2 on a global grid, leading to discrete estimates of spatial variation in dispersion. In the 25-150s period band, we observe a mean absolute misfit of 67.5m/s for SEMum2-predicted group-velocity dispersion maps relative to those included in the inversion (M. Ritzwoller, personal communication 2009; hereafter, the "MRPC" maps). For the longer-period 60-150s band, primarily sensitive to mantle structure, mean misfit falls to 51.5m/s – close to the level of uncertainty reported for the MRPC maps (Shapiro and Ritzwoller, 2002). We also validate our model against dispersion maps *not* included in our inversion, such as those predicted by the GDM52 dispersion model of Ekström (2011). In this case, we observe mean misfits of 89.7m/s and 58.3m/s for the 25-150s and 60-150s period bands, respectively. We note that the GDM52 misfit values are close to those given above for the MRPC maps – particularly at longer-periods, where the maps are expected to be most compatible.

Ultimately, the MRPC and GDM52 dispersion maps are themselves the product of tomographic inversions. Therefore, they reflect path-average constraints on phase or group slowness, modified by effects of parameterization and regularization. Because of this, direct comparison between these “observed” (MRPC or GDM52) and SEMum2-predicted dispersion maps is likely to be affected by different *a priori* assumptions regarding model smoothness, and it is perhaps more useful to compare *path-average* dispersion estimates. Indeed, these values more-closely reflect the underlying observational constraints upon which dispersion maps are based. In Figure 2.9, we show differences between predicted (SEMum2) and “observed” path-average dispersion for a number of primarily oceanic paths. As the GDM52 dispersion model yields estimates of phase velocities as well, and indeed is derived from phase-velocity measurements, we also compare predicted SEMum2 path-averages for these quantities. Here, dispersion is estimated from great-circle path integrations over phase or group slowness, sampled from the corresponding dispersion maps, and is consistent with the manner in which the “observed” maps were constructed. In general, path-average predictions for oceanic phase and group velocities, traversing a geophysically diverse range of environments, agree quite well – even compared to GDM52.

At the same time, we find that path-average dispersion estimates in purely continental settings, such as in Figure 2.9e, do not agree well with those derived from either set of “observed” dispersion maps. Referring back to Figures 2.6 and 2.7, however, we see that waveform fits are indeed quite good for source-aligned station pairs corresponding to paths nearly identical to that in Figure 2.9e (MAKZ or AAK to RUE). We believe that this disparity results from off-great-circle-path sensitivity and reflects shortcomings of the simplistic ray-theoretical forward modeling described above in regions characterized by strong lateral heterogeneity (continental), relative to settings with more laterally contiguous, albeit strongly layered, structure (oceanic). Indeed, the consistent way in which to validate dispersion along individual paths in SEMum2 would be to forward-model using the SEM and subsequently measure dispersion directly from the synthetic waveforms. However, as such a scheme is computationally very costly, we instead note that: (1) where the simple analysis above should be accurate (i.e. oceanic settings), predicted SEMum2 path-average dispersion is compatible with that obtained from both the MRPC and GDM52 maps; and (2) we prefer to rely primarily upon the measurements of waveform misfit for the most general assessment of model performance (i.e. Table 2.1).

Further, we also do not think it is likely that the use of a smooth crustal layer is responsible for the discrepancy in predicted dispersion for continental settings. Indeed, the crustal layer was not only calibrated directly from the dispersion maps, but also was used in calculating the example SEM waveforms discussed above, which did not appear compatible with poor dispersion fits along the corridor in Figure 2.9e. In addition, Fichtner and Igel (2008) proposed, and later employed (Fichtner et al., 2009), a scheme for smooth crustal layer development that also relies on dispersion matching (as well as the GLL-Lagrange basis for crustal-model parameterization). Because the authors used the scheme to match a known *a priori* crustal model (Crust2.0, Bassin et al., 2000), they were able to perform a more complete analysis of its accuracy, and further, did not report systematic poor performance in either continental or oceanic settings.

We return to this topic – the apparent discrepancy in predicted dispersion observed in continental settings – in our extensive validation of the SEMum2 model in Chapter 3. There, we demonstrate using the scheme suggested above (measuring dispersion directly from held-out waveform data and SEM-based synthetic waveforms) that path-average dispersion in SEMum2 fits the observed dispersion very well along both continental and oceanic paths (Figure 3.7).

2.3.2 Evolution of the crustal model

As noted above, the 25-60s subset of dispersion data used to calibrate the crustal model exhibits non-negligible sensitivity to the underlying mantle. Thus, after each iterative update to the SEMum2 mantle model, we invert for *new* anisotropic velocity structure within the smooth crustal layer, again using the two-step procedure outlined in Section 2.2.3. Further, by including the full 25-150s dispersion dataset in the mantle update, we are able to enforce a degree of consistency between the independently parameterized crust and mantle portions of the model – i.e. they are no longer constrained by wholly disjoint datasets. However, a staggered-update scheme of this type is not without pitfalls: particularly, potential tradeoffs in velocity structure across the fictitious 30km Moho in the oceans, leading to unrealistic velocity contrasts across the interface that progressively worsen. We previously touched on this issue in Section 2.2.3.3, and discussed options for mitigation.

In particular, we chose to introduce independent constraints on mantle-side V_S below the artificial Moho during the second iterative update to mantle structure. We did so by damping oceanic sub-Moho V_S toward values sampled from the global upper-mantle model CUB2 (Shapiro and Ritzwoller, 2002) at the same depths and locations. The CUB2 model was chosen because it not only shows impressive waveform fits when independently validated with regional SEM simulations (H. Yuan, personal communication 2011), but also because it was developed using the same underlying dispersion data as we have used in our modeling. We view this approach as a prudent preventive measure, which addresses perceived potential limitations of the homogenized-layer technique.

2.3.3 Resolution

It is widely acknowledged that the standard resolution analysis employed in many tomographic studies is strictly valid only for linear problems (Tarantola, 2005). Further, even when restricted to this limited domain of applicability, counterintuitive results may be obtained under common implementations of the analysis (Lévêque et al., 1993), and care must be taken in assessing model quality. Though waveform inversion is strongly non-linear, the argument can be made that it behaves pseudo-linearly under certain conditions, such as near the optimal model (Tarantola, 2005). Fully numerical alternatives to the linear analysis, relying upon second-order adjoint state methods for assessment of posterior uncertainty (Fichtner and Trampert, 2011), are an area of ongoing research. However, such approaches are not yet production-ready and pose prohibitive computational cost at global scales. Thus, while acknowledging the above qualifications, the linear analysis still provides the best currently available tool for probing certain characteristics of our model that result from data quality/coverage and the influence of *a priori* information.

Under the standard analysis, the resolution operator $\mathbf{R} = \mathbf{I} - \mathbf{C}'_m \mathbf{C}_m^{-1}$, with the posterior covariance operator $\mathbf{C}'_m = \left(\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \mathbf{C}_m^{-1} \right)^{-1}$, is applied to a model perturbation $\mathbf{m}'$, such as the commonly used checker-board pattern. The difference between $\mathbf{m}'$ and $\mathbf{Rm}'$ in some sense characterizes the quality of model recovery for the problem considered. Intuitively, it is clear that if the posterior state of uncertainty is equal to the prior state, meaning the data supplies no constraint on the model, then $\mathbf{C}_m = \mathbf{C}'_m$ and $\mathbf{R} = \mathbf{0}$, i.e. nothing is resolved. Further examination of the expressions above makes clear that the resolution operator depends solely upon: (1) the data and model prior covariance operators; and (2) the model Jacobian $\mathbf{G}$. Therefore, $\mathbf{R}$ reflects only the *a priori* estimates of data noise and model uncertainty / smoothness, as well as spatial variation in data coverage characterized by $\mathbf{G}$. *Note that any reflection upon the accuracy of the modeling theory is conspicuously absent.* The resolution operator may thus be viewed as a spatially variable smoothing operator: a filter applied to a hypothetical model perturbation in a way that gives guidance as to what scales of solution structure *might* be interpretable, and highlights pathological conditions such as strongly non-uniform data coverage (resulting in smearing). We further note that the underlying linearization of the forward problem, inherent in development of the analysis, is not consistent with our inversion methodology. In particular, the approximation:

$$\begin{aligned} & \mathbf{d} - \mathbf{g}(\mathbf{m}_k + \mathbf{m}') \\ & \quad \simeq \quad \mathbf{d} - \mathbf{g}(\mathbf{m}_k) - \mathbf{G}\mathbf{m}' \\ & \quad \simeq \quad -\mathbf{G}\mathbf{m}' \text{ as } ||\mathbf{d} - \mathbf{g}(\mathbf{m}_k)|| \ll ||\mathbf{G}\mathbf{m}'|| \text{ near the solution} \end{aligned} \tag{2.13}$$

is not compatible with our "exact" spectral-element implementation of $\mathbf{g}(\cdot)$. However, SEM-based resolution analysis – in which we directly model $\mathbf{g}(\mathbf{m}_k + \mathbf{m}')$ for the entire dataset – is prohibitively expensive. Thus, while the linear analysis is useful for validating recovery of the morphology of the test structure (i.e. against smearing), the *amplitudes* of model recovery obtained under the analysis should under-predict those expected for the hybrid scheme.

Keeping these limitations in mind, we perform the standard analysis for a suite of test patterns in order to determine rough bounds on the scales of interpretable structure in SEMum2. These estimates may in turn be used to validate the *a priori* correlation lengths introduced in Section 2.2.2. In Figures 2.10 and 2.11, we present checkerboard resolution tests at different depths, using blocks of 800-4000km maximum width in both equatorial and polar orientations. As can be seen in Figure 2.10, recovery of the morphology of V_S input structure is quite good, showing no evidence of systematic ray-like smearing or gaps in sensitivity, even at the poles, while very little cross-contamination with anisotropic structure is observed. Turning our attention to Figure 2.11, it can be seen that model recovery for ξ is less impressive. Indeed, while the 2200km-scale performance at shallow depths is adequate, again indicative of fairly homogeneous data coverage and negligible cross-contamination, recovery at greater depth is limited. At 470km, we estimate robust recovery of structure only at 4000km-scales, with small, but non-negligible tradeoffs with V_S. This is commensurate with the resolution estimates for anisotropic structure obtained for the original SEMum (Lekić and Romanowicz, 2011a), which is not surprising, as we have employed an identical dataset. Ultimately, we recover structure at length scales commensurate with the estimated minimum distances of correlation for V_S and ξ respectively, though limited to shallow depths for the latter. Further, tradeoffs between V_S and ξ are small at shallow depths ($<$ 300km), though increase in the transition zone. Thus, we expect shallow V_S structure to be interpretable on 800km length scales and quite near that at greater depth - certainly at smaller scales than the well-recovered 2200km transition-zone case. However, we expect only shallow ($\leq$200km depth) ξ to be robustly resolved at 2200km scales, with noticeably reduced resolution below; we caution against interpreting ξ structure at scales $<$ 4000km at transition-zone depths.

Also, in Figure 2.12, we show results obtained for isolated input perturbations in order to assess vertical smearing. In particular, we compare two representative locations: Hawaii, with unusually dense data coverage, and south-Pacific Superswell roughly equidistant from Tahiti and Pitcairn, with potentially poorer coverage. In both locations we see that while the amplitude of the input perturbation is not accurately retrieved, in agreement with our expectations for linearization of the forward problem, the morphology with depth is. Thus, we expect our model to reflect accurate depth-distribution of V_S anomalies down to 400km, with only a small amount of vertical smearing below that.

2.4 Model structure

In Figure 2.13, we show a comparison of global upper-mantle and transition-zone V_S structure at a range of depths for both SEMum and SEMum2, as well as two recent global models: S40RTS (Ritsema et al., 2011) and S362ANI (Kustowski et al., 2008a). The first and least surprising observation is that SEMum2 and its predecessor share many common features, and agree quite well overall, while acknowledging that smaller scales of heterogeneity were permitted in the inversion for SEMum2. Second, and equally important, while all four models appear to correlate well at long wavelengths, there are considerable differences in shorter-wavelength structure at all depths between the SEMum models and either S40RTS or S362ANI. Interestingly, the SEMum and SEMum2 models systematically exhibit stronger lateral heterogeneity at depths above 250km than either of the models from other groups. This is particularly evident at the shallowest depths, where the strong low-velocity signal of the mid-ocean ridge system dominates at the global scale. We shall return to this point shortly. Turning to Figure 2.14, we perform a similar comparison, now focusing on global ξ structure, again showing SEMum and SEMum2 as well as two recent whole-mantle radially anisotropic models: SAW642ANb of Panning et al. (2010) and S362WMANI, also from the study of Kustowski et al. (2008a). What is most apparent from visual inspection of Figure 2.14 is that while shallow ξ structure appears broadly compatible across models, albeit at very long wavelengths, there is clearly significant disagreement below $\sim$ 250km depth. One might ask whether this can be attributed to fundamental differences in the 1D mean ξ of these three models. However, as seen in Figure 2.15, their 1D ξ structure is indeed quite compatible – particularly between SEMum2 and S362WMANI.

To further assess the lack of consistency in 3D ξ structure, we first estimate its relative contribution to total waveform variance reduction in SEMum2. In particular, we compute SEM synthetics for a representative set of 14 events, both with and without 3D ξ – leaving the radially anisotropic crust, 1D mantle ξ, and 3D mantle V_S intact. In the absence of 3D ξ, we find that VR values for minor and major-arc Rayleigh waves fall by 4% and 6%, respectively, while analogous values for Love waves fall by 3% and 4%. At the same time, VR values for overtone surface waves (Rayleigh or Love) consistently decrease by $< 0.5\%$, which is comparatively quite small. Taken together, SEMum2 ξ structure corresponds to 3-4% of VR for the complete fundamental-mode dataset, which is compatible with values estimated for the SEMum model by Lekić and Romanowicz (2011a). Further, because Rayleigh and Love wave fits improve nearly uniformly with the addition of 3D ξ, it would appear that our data-weighting scheme is effective at integrating constraints from both so as not to bias the resulting ξ model. This latter property is necessary if *3D* ξ structure is to be considered robust, at least in the upper ~ 300km where fundamental modes are most sensitive.

Next, we quantitatively examine differences in ξ structure across models by calculating long-wavelength correlation (up to spherical harmonic degree $\ell = 8$) between the SEMum2 ξ model and those of SEMum, SAW642ANb, and S362WMANI, as a function of depth – shown in Figure 2.16a. We find that ξ structure is reasonably well-correlated across models above ~ 300km depth, confirming our intuition based on the maps presented in Figure 2.14, with the exception of a narrow band of poor correlation just above ~ 100km. In order to further explore this latter observation, we plot radial auto-correlations for the same suite of ξ models (Figure 2.16b) and find that SEMum2, SEMum, and S362WMANI all possess a shallow ξ layer that is markedly decorrelated with deeper structure, and further varies in depth-extent and sharpness across models – a likely source of the poor cross-model correlation observed in this depth range. We also note that this shallow decorrelated layer is largely absent from SAW642ANb, potentially responsible for the slight improvement in correlation with SEMum2 above 100km depth relative to S362WMANI in Figure 2.16a. Additionally, the S362WMANI, SEMum, and SEMum2 1D-mean ξ models (Figure 2.15) all exhibit strong negative gradients in this depth range, while the latter two also contain nearby ξ local minima (60-80km depth) – evidence of rapid variation in depth that further contributes to the decorrelation noted above. An explanation for the existence of this decorrelated layer – apparently a feature common to three of the four models discussed here – is not immediately obvious, and a detailed investigation of the significance of this observation is beyond the scope of this paper. Ultimately, we emphasize that ξ structure above ~ 300km depth is reasonably consistent at long wavelengths across the four models discussed here, while considerable differences at greater depths imply an elevated level of uncertainty in global upper-mantle ξ tomography at the present time.

Returning to the isotropic V_S model, one might ask how well the amplitudes of the pronounced low-velocity anomalies seen in the upper 250km of SEMum2 correspond to those seen in higher-resolution local-scale tomographic studies. As noted above, the strong low velocities of the mid-ocean ridge system are a prominent feature of SEMum2 (as well as

SEMum) – especially those beneath the fast-spreading East Pacific Rise (EPR), where we observe reductions in V_S in excess of 9%. In Figure 2.17a, we compare 1D profiles of V_S obtained from SEMum2, S362ANI (Kustowski et al., 2008a), and S40RTS (Ritsema et al., 2011), for a point along the EPR coincident with the OBS-based tomographic study of Harmon et al. (2009). We also show a 1D-mean profile from the central portion of the Harmon et al. (2009) V_S model, averaged over length scales consistent with the minimum lateral correlation length used in our inversion (Section 2.2.2). While the structure at depths shallower than 50 km in SEMum2 cannot be interpreted due to the introduction of the homogenized 30 km crust, we see that SEMum2 recovers both the strength and depth of low-velocity zone underlying the EPR, as imaged by Harmon et al. (2009), far-more closely than either of the other two global models. The implication of this result is that we can, with some degree of confidence, infer realistic amplitude recovery in areas that lack local high-resolution modeling – i.e. most of the Earth. Further, in Figure 2.17b, we see that the amplitude of SEMum2 V_S anomaly observed in the Harmon et al. study region extends for a considerable distance along the EPR and suggests that such pronounced low velocities: are (a) ubiquitous beneath the EPR at these latitudes and (b) consistently imaged in SEMum2, where present.

Recovery of concentrated low velocities in oceanic settings is not limited to the ridges in SEMum2. Indeed, we systematically observe a pronounced low-velocity zone (LVZ) underlying the seismically fast oceanic lithosphere – reflecting one of the earliest observations regarding the seismic structure of the oceans (e.g. Dorman et al., 1960). In Figure 2.18, we see two types of oceanic low-velocity structure, both of which were previously seen in SEMum: (1) a concentrated sub-lithospheric LVZ, progressively weakening and deepening in a manner consistent with a plate cooling with age, though with some scatter due to local structure; and (2) separate, deeper low-velocity anomalies, suggesting varying degrees of interaction with the overlying LVZ. Further, it is also interesting to examine the pattern of ξ structure shown both along (Figure 2.17b) and away from (Figure 2.18) the ridges. We observe: (1) strong $\xi > 1$ ($V_{SH} > V_{SV}$) within and beneath the oceanic lithosphere, consistent with the signal of dominantly horizontal shear concentrated above $\sim$ 200km depth (either ongoing or frozen-in); and (2) significantly reduced ξ, and locally $\xi < 1$ ($V_{SV} > V_{SH}$), beneath the ridges, consistent with the signal expected to result from dominantly vertical flow and lattice-preferred orientation resulting from A, D, or E-type olivine fabric (e.g. Table 2, Karato et al., 2008). These long-wavelength patterns of shallow ξ beneath the oceans are broadly consistent with, and therefore validate, earlier observations (e.g. Nishimura and Forsyth, 1989; Ekström and Dziewonski, 1998), though the SEMum2 ξ model differs in detail from even the most recent studies, as noted above.

2.4.1 Regional structure

We now turn our attention to the discussion of SEMum2 V_S structure in the context of regional-scale tomographic studies, which we consider a valuable exercise for assessing our expectations of model resolution.

2.4.1.1 Africa

In Figure 2.19, all four major African cratons (Western African, Congo, Tanzanian, and Kalahari) are visible as strong fast anomalies at 150km, consistent with SEMum and a range of continental-scale studies (Sebai et al., 2006; Pasyanos and Nyblade, 2007; Priestley et al., 2008). Between 150 and 200km, the signal of the Tanzanian craton is the first to vanish, in agreement with the studies of Weeraratne et al. (2003) and Sebai et al. (2006). The remaining three cratons persist to 200-250km depths, with the Kalahari lost first – as in Pasyanos and Nyblade (2007) and Priestley et al. (2008) – and the Western African showing the greatest depth extent (to at least 250km, consistent with SEMum). These observations of the morphology and relative depth extent of the African cratons are confirmed upon examining Figure 2.20, where cross-sections through the African upper mantle traverse all four.

The most prominent low-velocity anomalies in Figure 2.19 are those associated with the Red Sea and East African Rift (EAR) system: progressively shifting SSW from the northern EAR and Red Sea (75-150km) to the southern EAR and Tanzania ($\sim$250km), consistent with SEMum and similar to Sebai et al. (2006) and Pasyanos and Nyblade (2007). Of particular interest are the strong low velocities beneath the anomalously thin Tanzanian craton – also observed by Weeraratne et al. (2003) and interpreted as a mantle upwelling – as well as the overall morphology of the system of low velocities seen in cross-section view in Figure 2.20. The complex, bifurcated appearance of these structures, concentrated at depth beneath Tanzania and Afar, suggests disparate causative processes, consistent with observations of melt source-signature variation along the EAR (e.g. Rogers et al., 2000). Also present in Figure 2.19, we observe a low-velocity band trending SW from the Tibesti hotspot through the Cameroon Line, again consistent with SEMum and previous regional studies (e.g. Pasyanos and Nyblade, 2007). Finally, the strength of lateral heterogeneity shown in Figure 2.20 is itself of note, reaching $> 17\%$ peak-to-peak in fairly close juxtaposition. This magnitude of lateral variation would seemingly preclude a purely thermal origin, instead likely requiring some degree of partial melting beneath the rift.

2.4.1.2 Eastern Eurasia

Turning our attention to the eastern portion of Eurasia (Figure 2.21), we see that SEMum2 again exhibits structure largely consistent SEMum. At 75km, notable features that persist, and indeed strengthen in SEMum2, include: (1) pronounced small-scale fast anomalies beneath the Tarim and Sichuan basins, consistent with the studies of Priestley et al. (2006) and Friederich (2003); (2) a localized low-velocity anomaly beneath the Altai Mountains, as previously noted by Priestley et al. (2006) and Kustowski et al. (2008b); (3) extensive low-velocity anomalies tracing the subduction zones in the east of the region (e.g. Friederich, 2003; Kustowski et al., 2008b); and (4) a band of low velocities stretching from Tibet to the Anatolian Convergence Zone, similar to that seen by Kustowski et al. (2008b) and also partially visible in the recent European adjoint tomographic model of Zhu et al. (2012).

SEMum2 at 75km differs from SEMum most clearly beneath Tibet, where the strong broad low-velocity anomaly seen in the latter is now reduced in both amplitude and lateral extent. One potential explanation for this change would be improved decoupling between the crust and mantle portions of our model following the new crustal-layer implementation.

By 150km, we note the appearance of high-velocity anomalies beneath the whole of Tibet, consistent with SEMum as well as numerous previous studies (Friederich, 2003; Priestley et al., 2006; Kustowski et al., 2008a; Panning et al., 2012). Although less-pronounced than in SEMum2, the Eastern European and Siberian cratons (EEC and SC) again appear as separate high-velocity bodies at 150km depth and below, unlike, for example, the model of Kustowski et al. (2008b). Further, the boundaries of these high-velocity anomalies have sharpened in SEMum2, such as in the case of the western margin of the EEC, notable for its sharpness in the adjoint tomography of Zhu et al. (2012) as well. In the 200-250km depth range, we see a pronounced reduction in the high-velocity signal beneath the SC, while those beneath the EEC, the Arabian craton, and Tibet persist – with the latter taking on a more band-like morphology, similar to SEMum and Kustowski et al. (2008b). By 350km, the fast anomaly beneath Tibet has weakened dramatically, as has that of EEC. At greater depths, we see images dominated by the fast signature of subducted material along the convergent margins to the east, as well as stretching from Italy, along the Hellenic Arc, to Iran – also seen in SEMum and Kustowski et al. (2008b), and partially visible in the model of Zhu et al. (2012).

2.4.1.3 North America

Consistent with the earliest (e.g. Romanowicz, 1979) as well as more-recent regional studies (e.g. Marone et al., 2007; Nettles and Dziewonski, 2008; Bedle and van der Lee, 2009; Yuan et al., 2011), we see that shallow SEMum2 V_S structure beneath North America (Figure 2.22) is dominated by the duality between strong low velocities of the tectonically active west and high velocities of the stable continental platform in the east. Similar to SEMum, this sharp contrast persists to at least 200km depth. However, SEMum2 appears to recover smaller-scale features embedded within these two domains in greater detail.

Near 75km depth, concentrated anomalies exist within the eastern domain of high velocities – particularly beneath the Wopmay orogen and Slave craton to the northwest, the Archean Superior and Rae provinces to the north, and beneath Greenland to the northeast. Further, we still observe "tongues" of high velocities extending into the Atlantic Ocean off Newfoundland and the Southeast U.S., as well as the Gulf of Mexico. Within the low-velocity domain beneath western North America, we image a less-slow band of V_S extending from the California coast out into the Pacific. This feature, coincident with the transform plate boundary joining active rifting in the Gulf of California to that at the Gorda Ridge, was also present in SEMum and is consistent with colder oceanic lithosphere progressively isolated over the past 30Ma from ongoing rifting to the north and south. At 150km depth, the eastern high-velocity anomalies largely merge - except beneath Greenland, which becomes more distinct – and further concentrate beneath the Slave, Rae, and Superior provinces near

200km. By 250km, strong high velocities are expressed only beneath the Rae province, and vanish entirely shortly thereafter. To the west, we again find that low velocities underlying the Basin and Range are lost near 200km depth, while neighboring anomalies beneath the Pacific remain. Interestingly, we now see a modest high-velocity signature of the subducted Juan de Fuca slab in the Pacific Northwest at 200-250km depth, similar to that in the SAWum_NA2 model of Yuan et al. (2011), but previously absent in SEMum.

At transition-zone depths, we find that the overall pattern of V_S structure beneath North America is compatible with that in SEMum, though smaller-scale details have continued to evolve. For example, near 575km we still see the northwest-to-southeast trending band of high velocities previously attributed to the subducted Farallon slab and broadly consistent with that seen by Sigloch et al. (2008). At the same time, this feature is less contiguous is SEMum2: indeed showing along-trend segmentation, particularly in the Pacific Northwest, more consistent with the work of Sigloch et al. (2008) or more recently Obrebski et al. (2011). Similarly, we observe that the broad low-velocity anomalies seen in SEMum off the east and west coasts of North America have been replaced by distributed groups of smaller-scale features generally consistent with those seen in SAWum_NA2 (Yuan et al., 2011) at similar depths.

2.5 Conclusions and future directions

Over the past decade, the spectral element method has become an indispensable tool for accurate modeling of wave propagation at global scales (e.g. Komatitsch and Tromp, 2002a,b; Chaljub et al., 2003). However, only recently has it become computationally feasible for the SEM to supplant approximate forward-modeling methods in the context of global tomographic studies (Lekić and Romanowicz, 2011a). At these scales, fully numerical inversions (e.g. Tarantola, 1984; Tromp et al., 2005) remain extremely computationally heavy – a problem we mitigate by progressively integrating SEM-based modeling into existing approaches, such as in the hybrid inversion scheme. Here, we have presented an updated version of the SEMum model, SEMum2, obtained using a more realistic geometry for the smooth crustal layer and a finer-scale mantle-model parameterization, all at a manageable increase in cost. We find that many of the distinguishing characteristics of the SEMum model persist in SEMum2, including the strength of lateral heterogeneity above 250km depth – particularly for low-velocity anomalies. This consistency between structures resolved in SEMum and those retrieved in SEMum2 demonstrates that reservations about the use of a 60km thick homogenized crustal layer in SEMum were, in retrospect, unjustified. Even at regional scales, the general character of SEMum2 model structure remains remarkably similar to that of its predecessor (Section 2.4.1). Notable characteristics of the model are: (1) stronger lateral variations in velocity, accompanied by accurate estimation of low-velocity minima and their depths in the oceans, which allows us to explore oceanic upper-mantle structure with confidence even in the absence of local arrays of stations; and (2) details in continental structure compatible with collocated regional studies, where they exist, and at a finer scale of resolution than is typically possible in global tomographic studies. Further, the more accurate retrieval of oceanic low-velocity structure in SEMum2 has enabled detailed investigation of novel model features observed beneath the ocean basins (French et al., 2013).

In addition to providing new constraints on global variations of V_S and ξ, SEMum2 can also serve as a jumping-off point for further tomographic modeling efforts. In particular, in combination with a long-wavelength model of the earth's lower mantle, SEMum2 would serve as an excellent starting model for a whole-mantle SEM-based inverison. We return to this ambitious next phase of modeling in Chapter 4. Further, correcting for propagation and source-excitation effects using SEMum2 should improve the robustness of global-scale seismic-source inversions. We intend to explore this further after obtaining our preliminary whole-mantle model – a necessary first step in order for shorter-period body-wave phases to provide meaningful constraints on source mechanism. Additionally, the more accurate modeling of (de)focusing of seismic energy by the strong lateral-velocity gradients present in SEMum2 should allow for the construction of higher-resolution models of *anelastic* (Q) mantle structure by reducing contamination from purely elastic effects. Indeed, a hybrid anelastic inversion, using SEM-based forward modeling and SEMum2 as the elastic starting model, is ongoing and will be reported in a forthcoming publication. Finally, the SEM-based development and validation of SEMum2 provides a desirable starting point for high-resolution regional inversions that incorporate higher frequency data and/or rely on the fully numerical adjoint-state inversion schemes referred to above – both of which require high-quality starting models in order for the inversion to remain well-posed and attain satisfactory convergence rates.

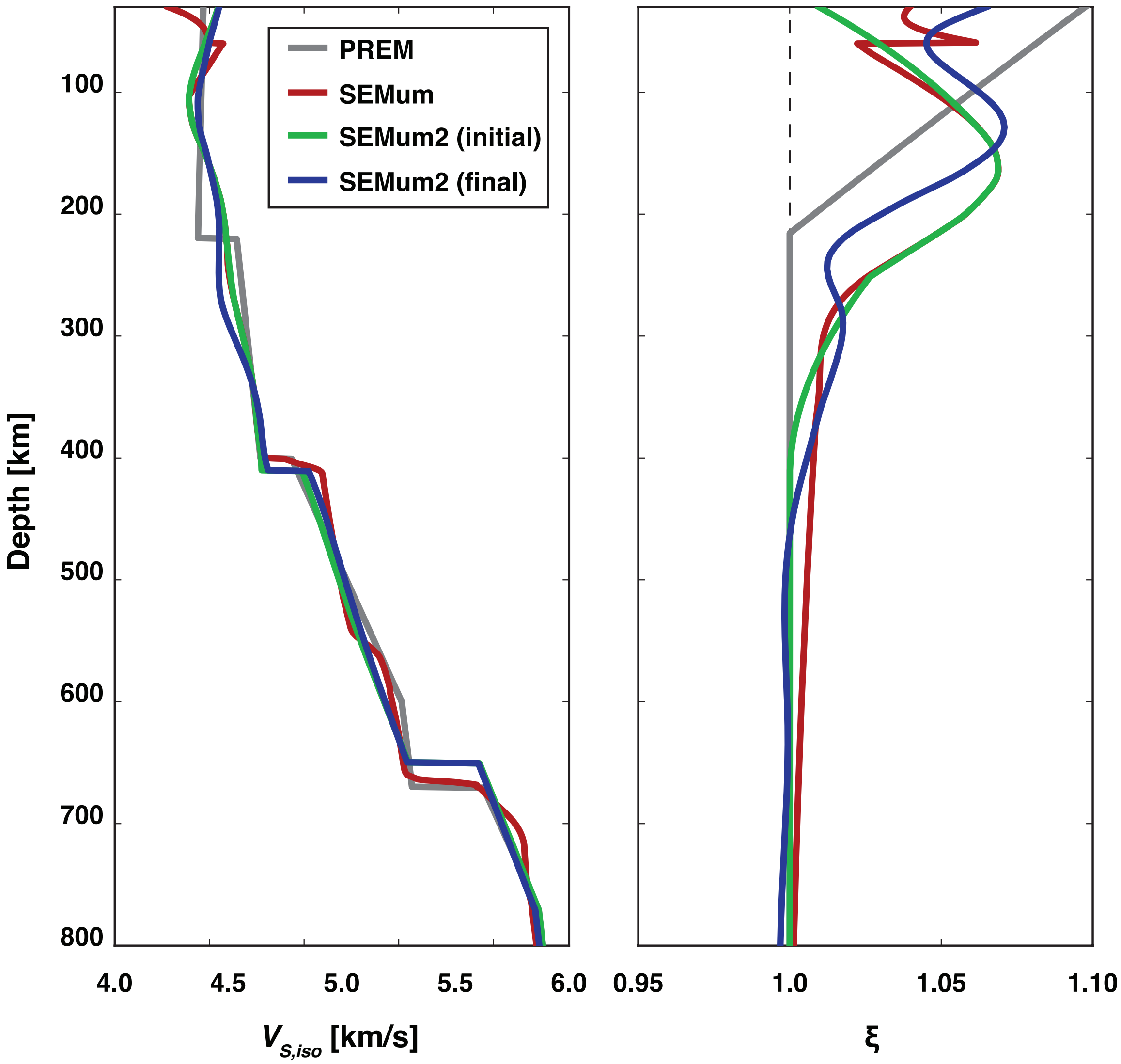


Figure 2.1: Profiles of 1D reference structure associated with SEMum, SEMum2 (starting), SEMum2 (final), and PREM (Dziewonski and Anderson, 1981) for comparison. SEMum2 1D structure differs from SEMum primarily due to: (1) removal of the 60km crustal layer; and (2) introduction of transition-zone structure from STW105 (Kustowski et al., 2008a) (see text). Left panel: Voigt-average V_S; Right panel: ξ.

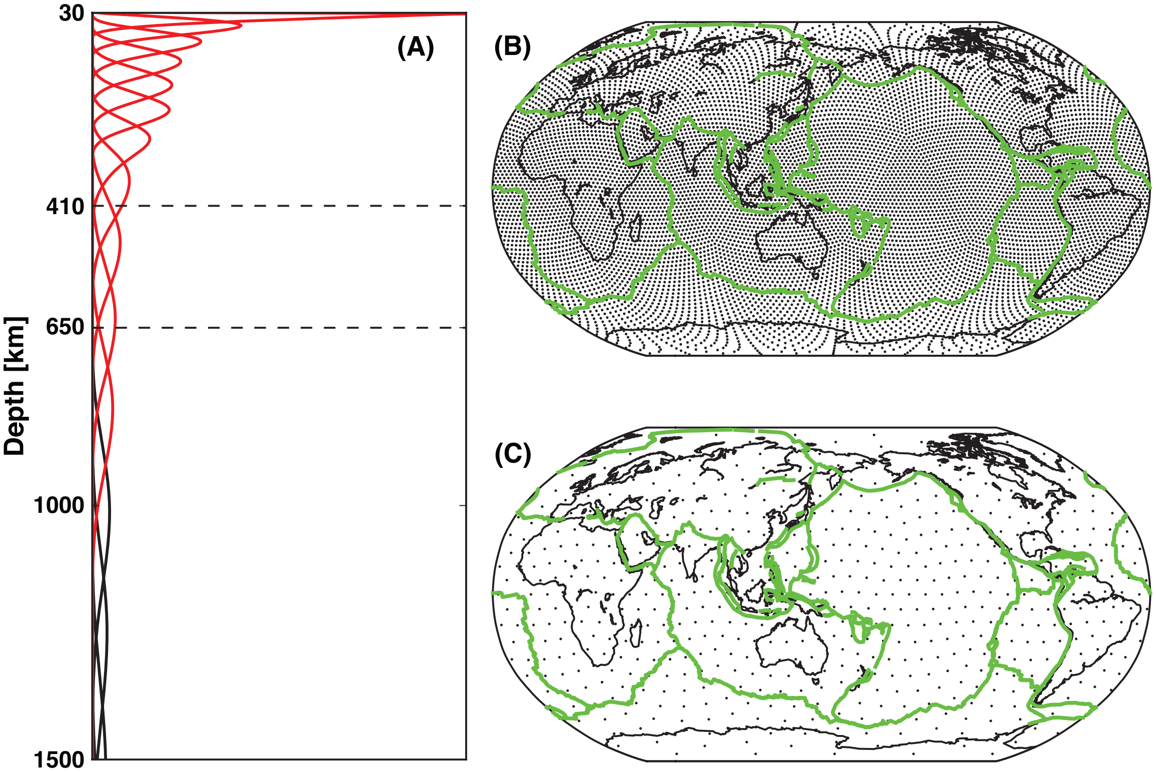


Figure 2.2: SEMum2 model parameterization: (a) radial b-spline basis functions (Mégnin and Romanowicz, 2000) exhibiting variable spacing consistent with expected radial resolution – red splines are inverted for; (b) and (c) spherical-spline knots (Wang and Dahlen, 1995) for the V_S (10,242 knots) and ξ (642 knots) models, respectively.

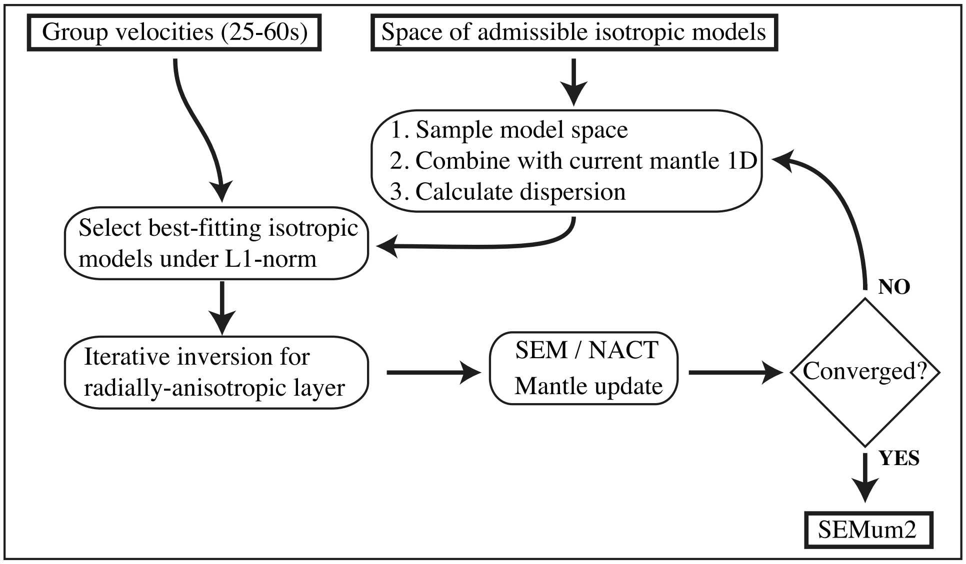


Figure 2.3: A high-level overview of the development of the smooth crustal layer, including its overall context in the mantle-model inversion.

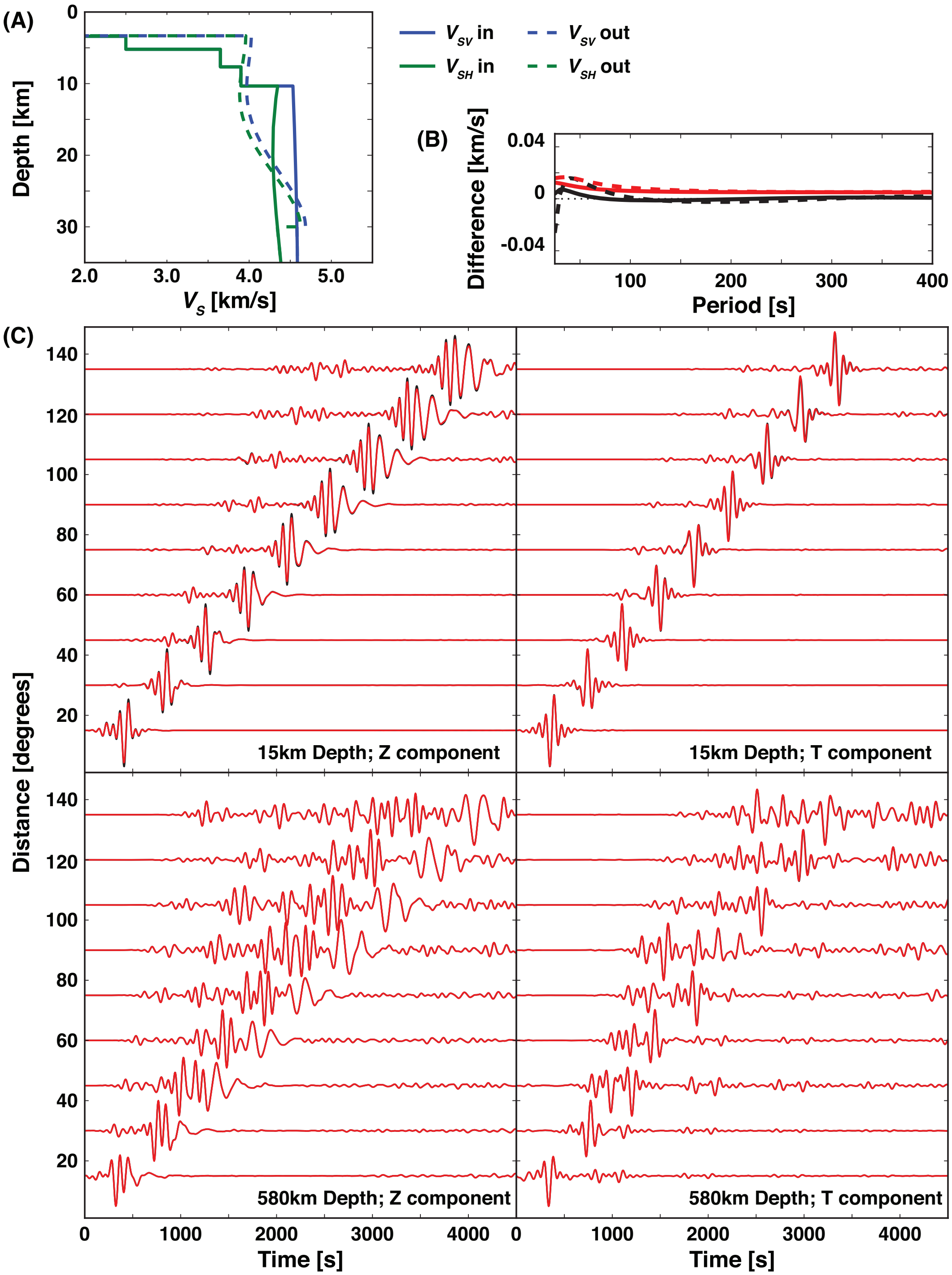


Figure 2.4: A smooth 30km crustal layer derived for a synthetic "oceanic" input structure using the two-step matching scheme (see text), for which long-period surface-wave dispersion is unaffected by this process. (Continued on the following page.)

Figure 2.4: (Continued) Panels: (a) layered synthetic input and smooth 30km output structures; (b) long-period Love (red) and Rayleigh (black) fundamental-mode phase (solid) and group (dashed) velocity misfit between layered input and smooth output models; (c) vertical and transverse component long-period ($> 60s$) waveforms for shallow and deep test events calculated in the layered input (black) and smooth output (red) models.

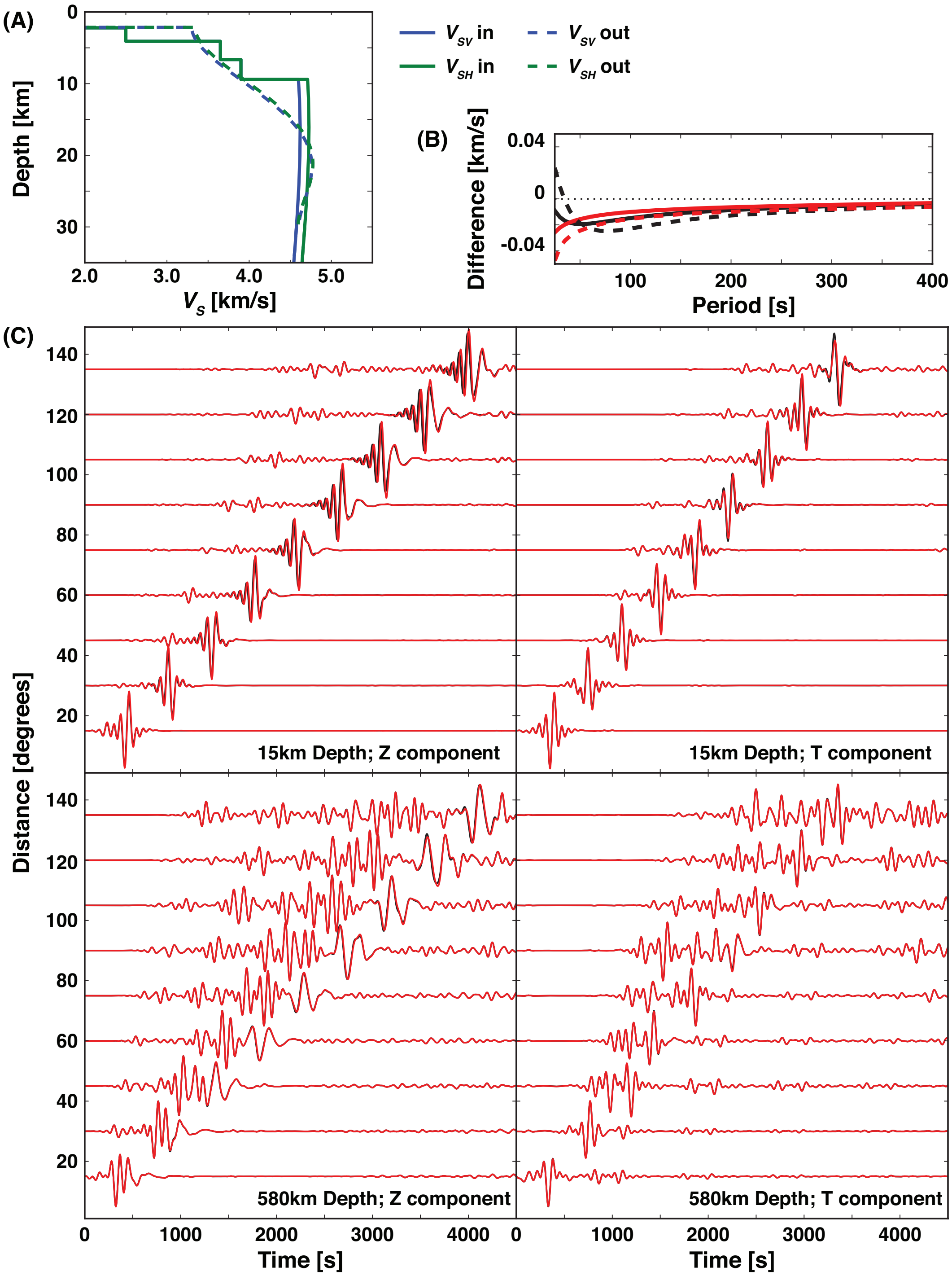


Figure 2.5: Similar to Figure 2.4, but instead showing a pathological example – exhibiting characteristic depressed phase/group velocities at long periods and slightly poorer fundamental-mode waveform fits. Note that overtone waveforms are practically unaffected.

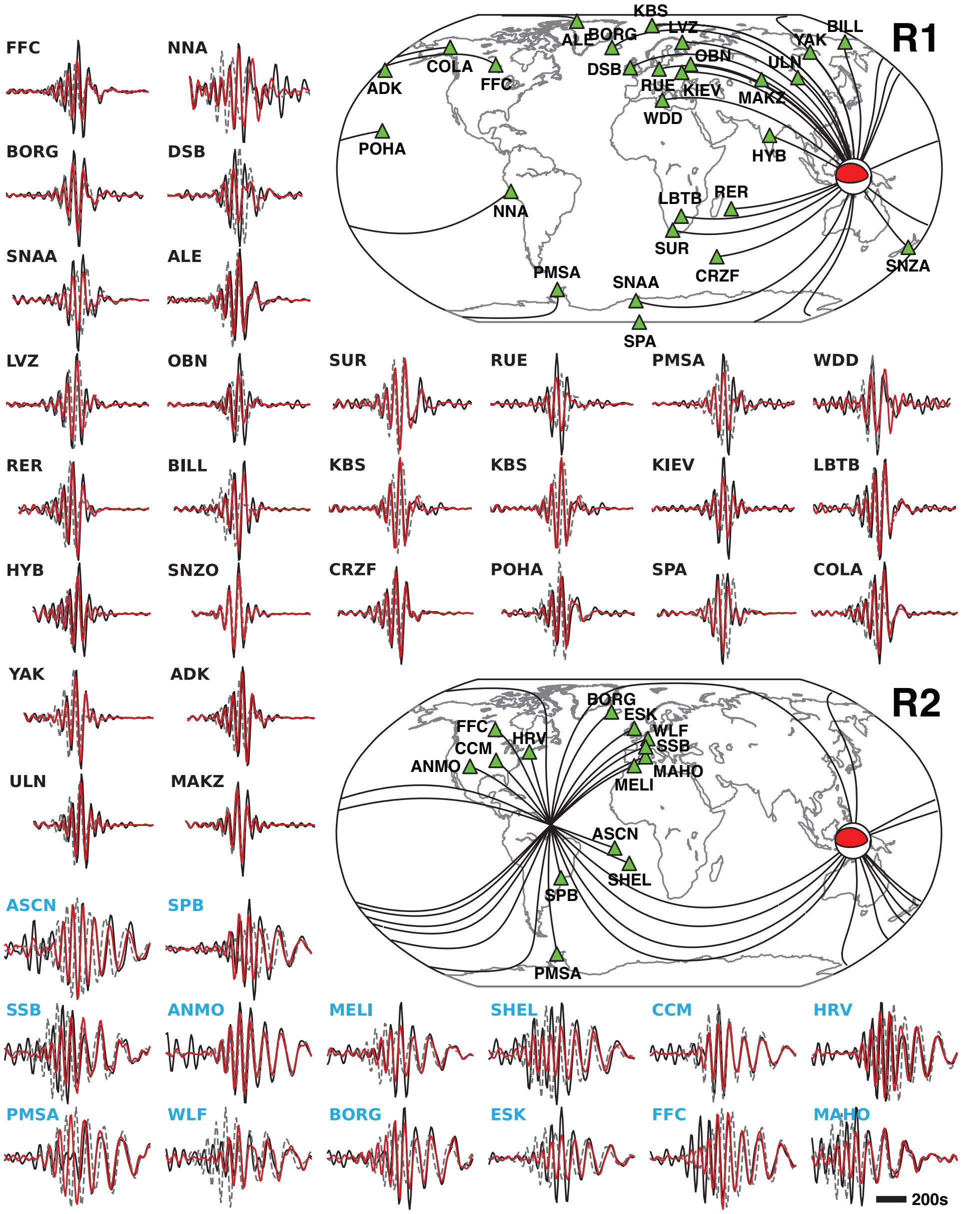


Figure 2.6: Comparisons between observed (black) minor or major-arc fundamental-mode Rayleigh waves (Z component) and those predicted from both the SEMum initial-model (grey dashed) and SEMum2 (red) for a representative event in our dataset (28 August, 2000: Mw 6.8 Banda Sea, 15km depth). Blue trace labels denote major-arc phases.

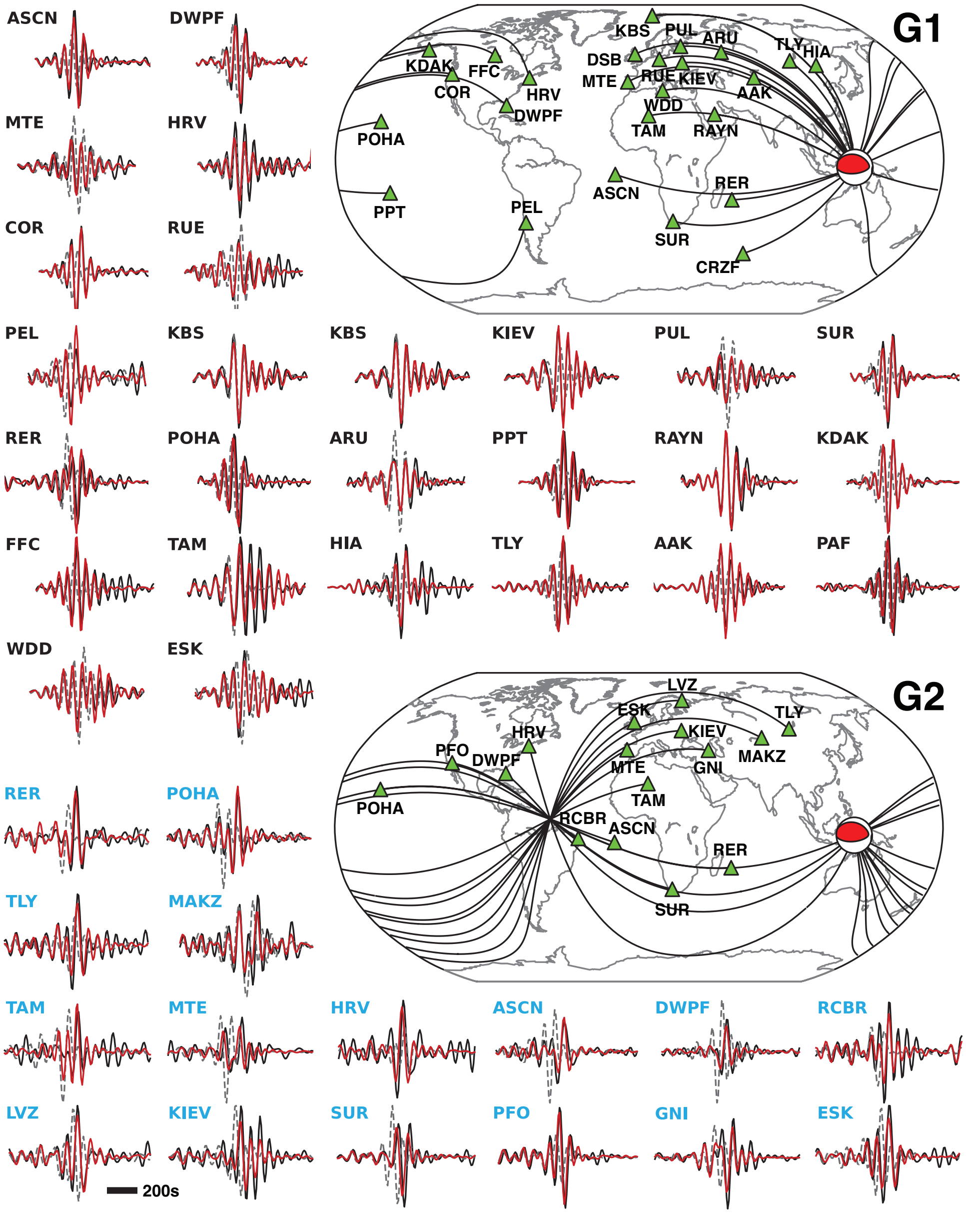


Figure 2.7: Same as Figure 2.6, but for fundamental-mode Love waves (T component).

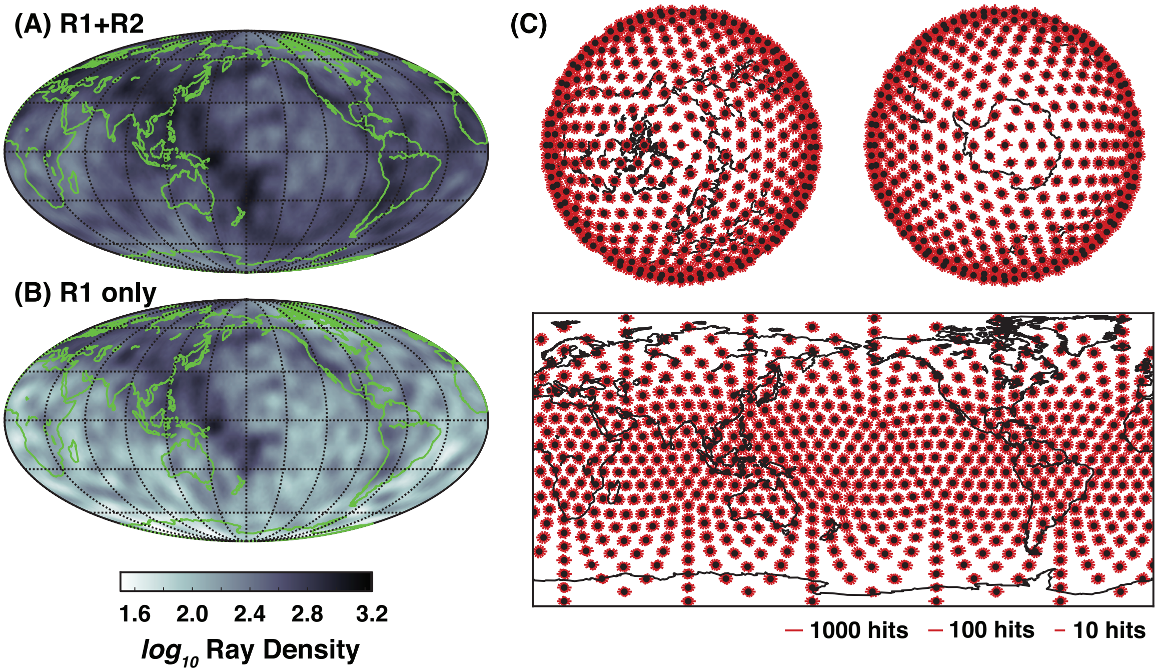


Figure 2.8: Measures of dataset quality: (a) and (b) *unweighted* logarithmic ray-path density for fundamental-mode (Z component) Rayleigh waves, including minor *and* major-arc phases (a) and minor-arc *only* (b); (c) (logarithmic) rose diagrams, for the same dataset as in panel (a), illustrating homogeneity of azimuthal coverage, which is crucial for reducing trade-offs between azimuthal anisotropy and isotropic velocity variations.

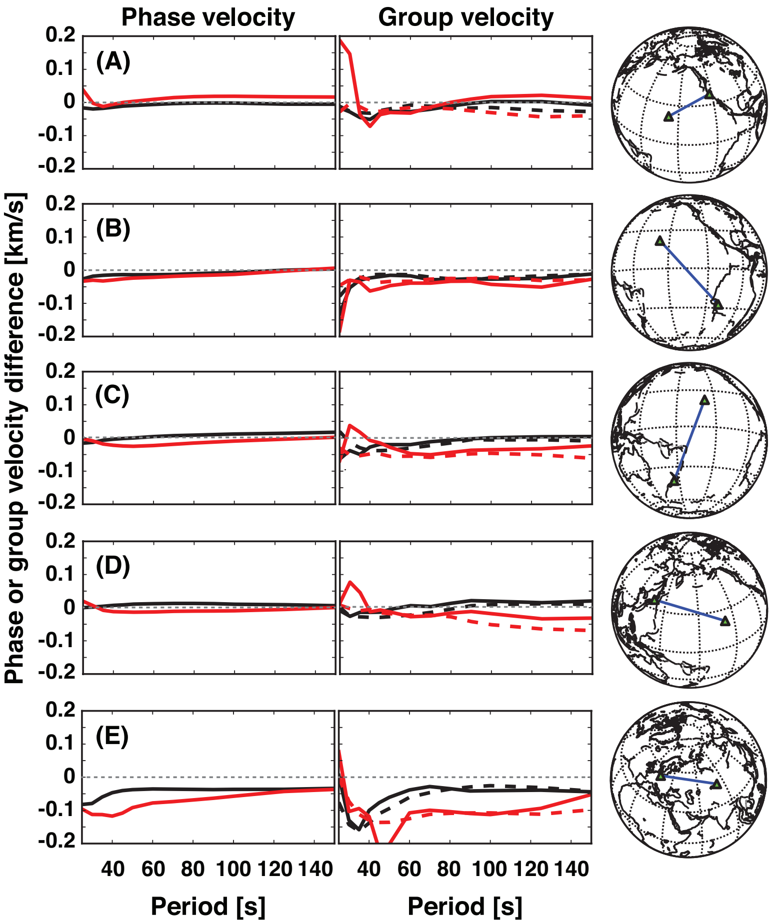


Figure 2.9: Differences in path-average Rayleigh (black) and Love (red) phase and group-velocity dispersion (left and right panels, respectively), between that predicted from SEMum2 (see text) and that from the MRPC (dashed) and GDM52 (Ekström, 2011) (solid) dispersion maps. Note: because the MRPC maps are for group-velocity, only GDM52 misfits appear in the left (phase velocity) panels.

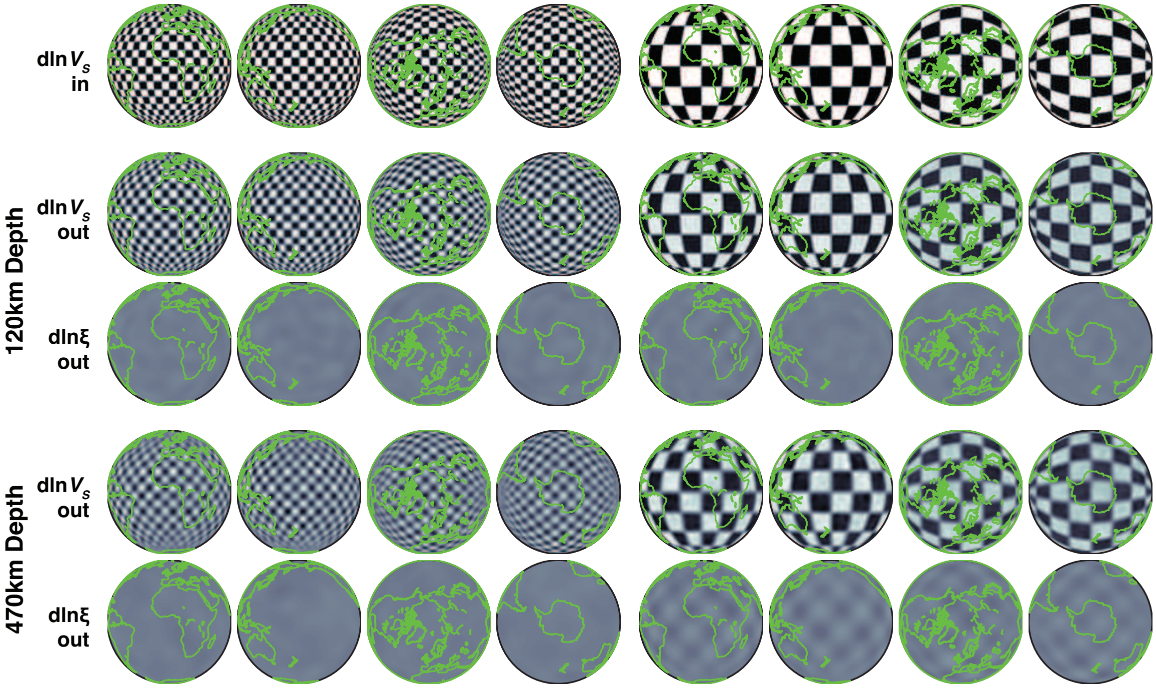


Figure 2.10: Linear resolution analysis for the isotropic V_S model. Upper panel: input test structure of ∼ 800km and ∼ 2200km width in polar and equatorial orientations (input ξ structure is zero). Middle panel: output V_S and ξ structure for input placed at 120km depth. Lower panel: output V_S and ξ structure for input placed at 470km depth. See text for discussion.

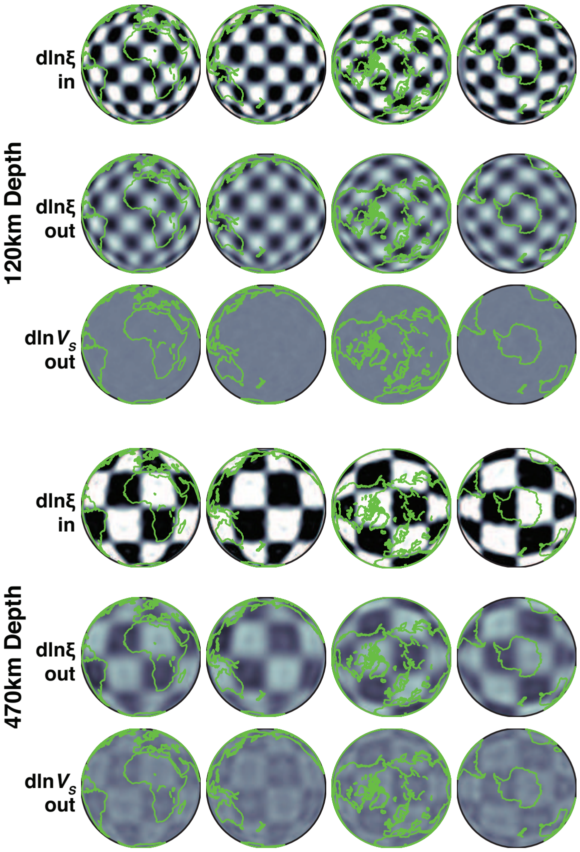


Figure 2.11: Linear resolution analysis for the anisotropic ξ model. Upper panel: $\sim$ 2200km width input structure at 120km in polar and equatorial orientations (input V_S structure is zero), as well as the resulting output ξ and V_S structure. Lower panel: $\sim$ 4000km width input structure at 470km in polar and equatorial orientations (input V_S structure is zero), as well as the resulting output ξ and V_S structure. See text for discussion.

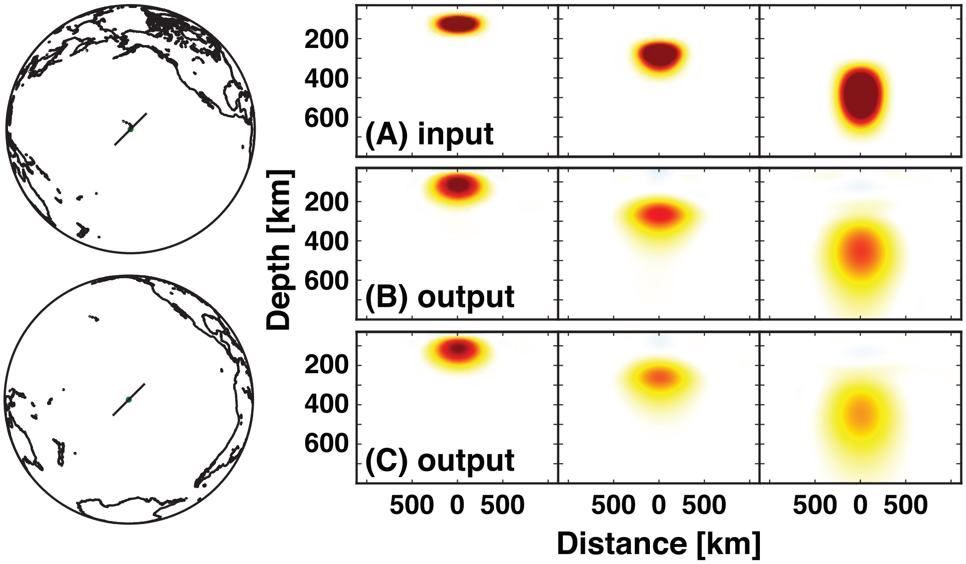


Figure 2.12: Linear resolution analysis based on isolated V_S perturbations at a range of depths for assessment of vertical smearing: (a) input perturbation structure used in locations (b) and (c); (b) output structure for input centered on Hawaii, characterized by unusually dense data coverage; and (c) output structure for input centered on the South-Pacific Superswell (roughly equidistant from Tahiti and Pitcairn) with potentially poorer coverage. In general, we note satisfactory recovery of the morphology of input structures at both locations.

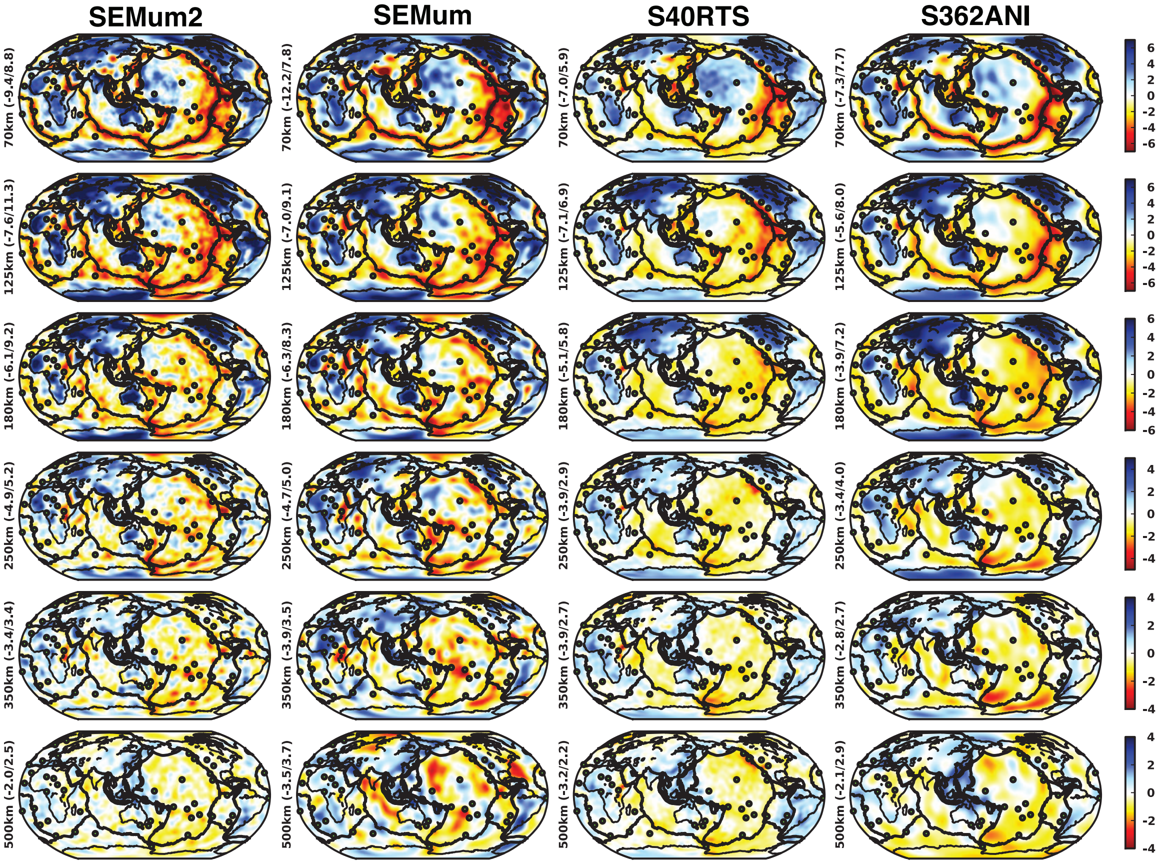


Figure 2.13: Maps of global upper-mantle and transition-zone V_S structure (% deviation) from SEMum2 (this study), SEMum (Lekić and Romanowicz, 2011a), S40RTS (Ritsema et al., 2011), and S362ANI (Kustowski et al., 2008a). Green circles denote hotspot locations of Steinberger (2000).

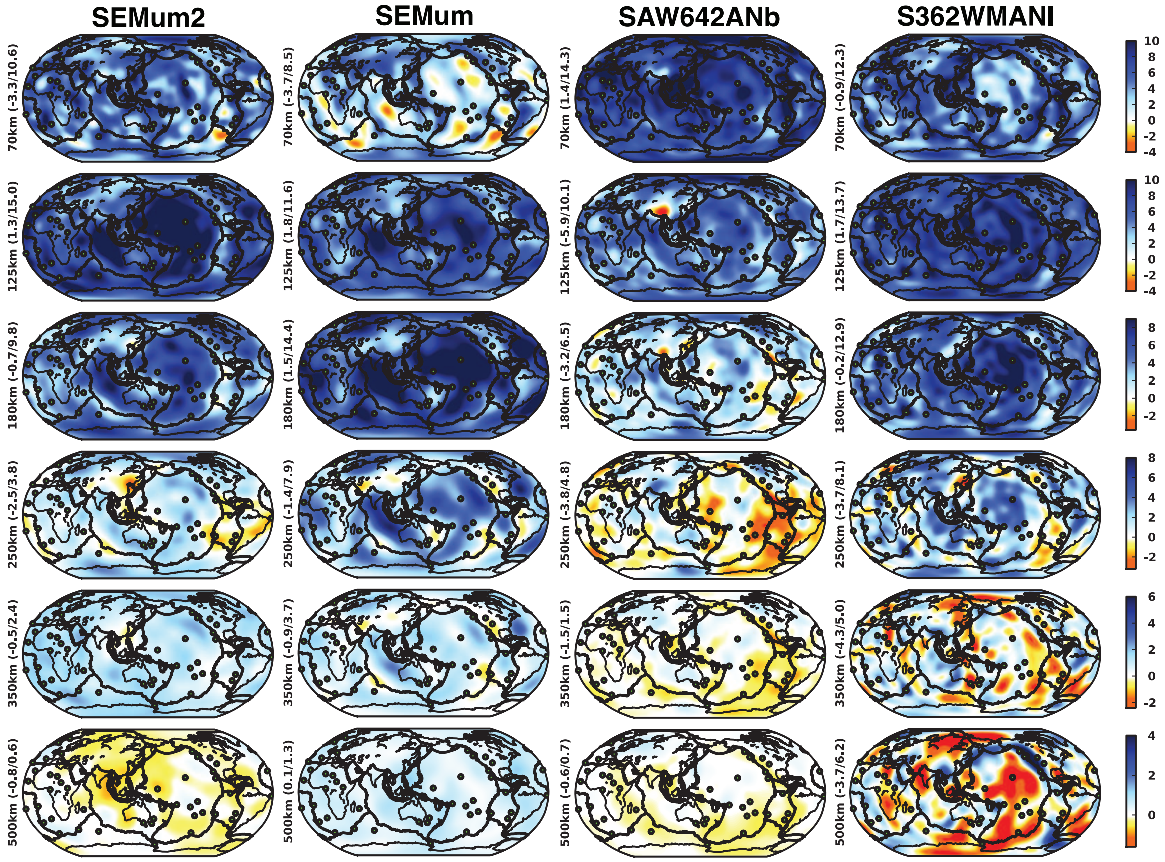


Figure 2.14: Maps of global upper-mantle and transition-zone ξ structure (% deviation from isotropy) from SEMum2 (this study), SEMum (Lekić and Romanowicz, 2011a), SAW642ANb (Panning et al., 2010), and S362WMANI (Kustowski et al., 2008a). Note the asymmetry of the color scale. Green circles denote hotspot locations of Steinberger (2000).

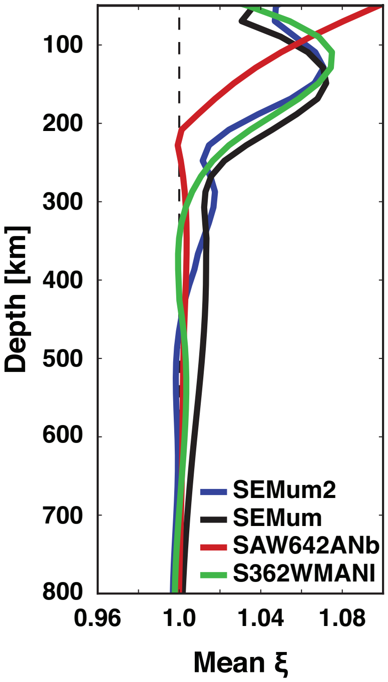


Figure 2.15: Mean (1D) profiles of global upper-mantle and transition-zone ξ structure from SEMum2, SEMum (Lekić and Romanowicz, 2011a), SAW642ANb (Panning et al., 2010), and S362WMANI (Kustowski et al., 2008a).

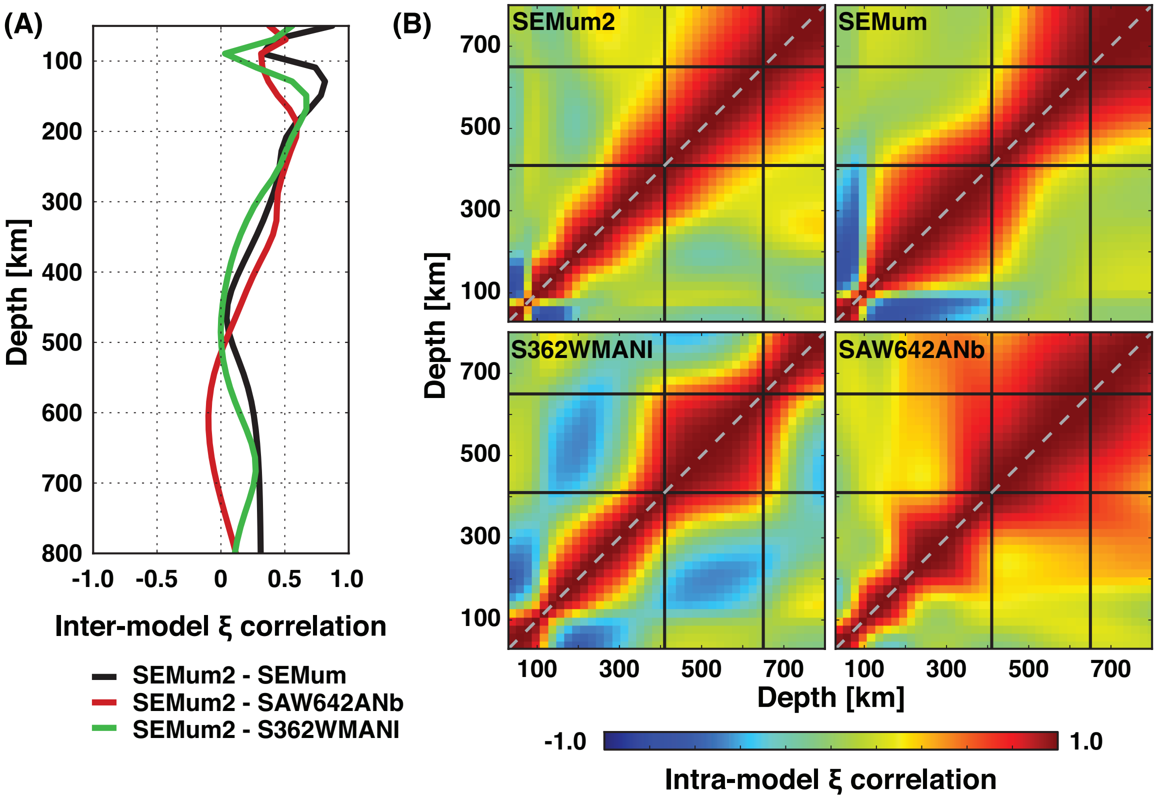


Figure 2.16: Inter and intra-model ξ correlations: (a) inter-model ξ correlation, as a function of radius, computed between SEMum2 and each of SEMum (Lekić and Romanowicz, 2011a), SAW642ANb (Panning et al., 2010), and S362WMANI (Kustowski et al., 2008a); (b) intra-model radial ξ (auto)correlation, shown for SEMum2, SEMum, SAW642ANb, and S362WMANI.

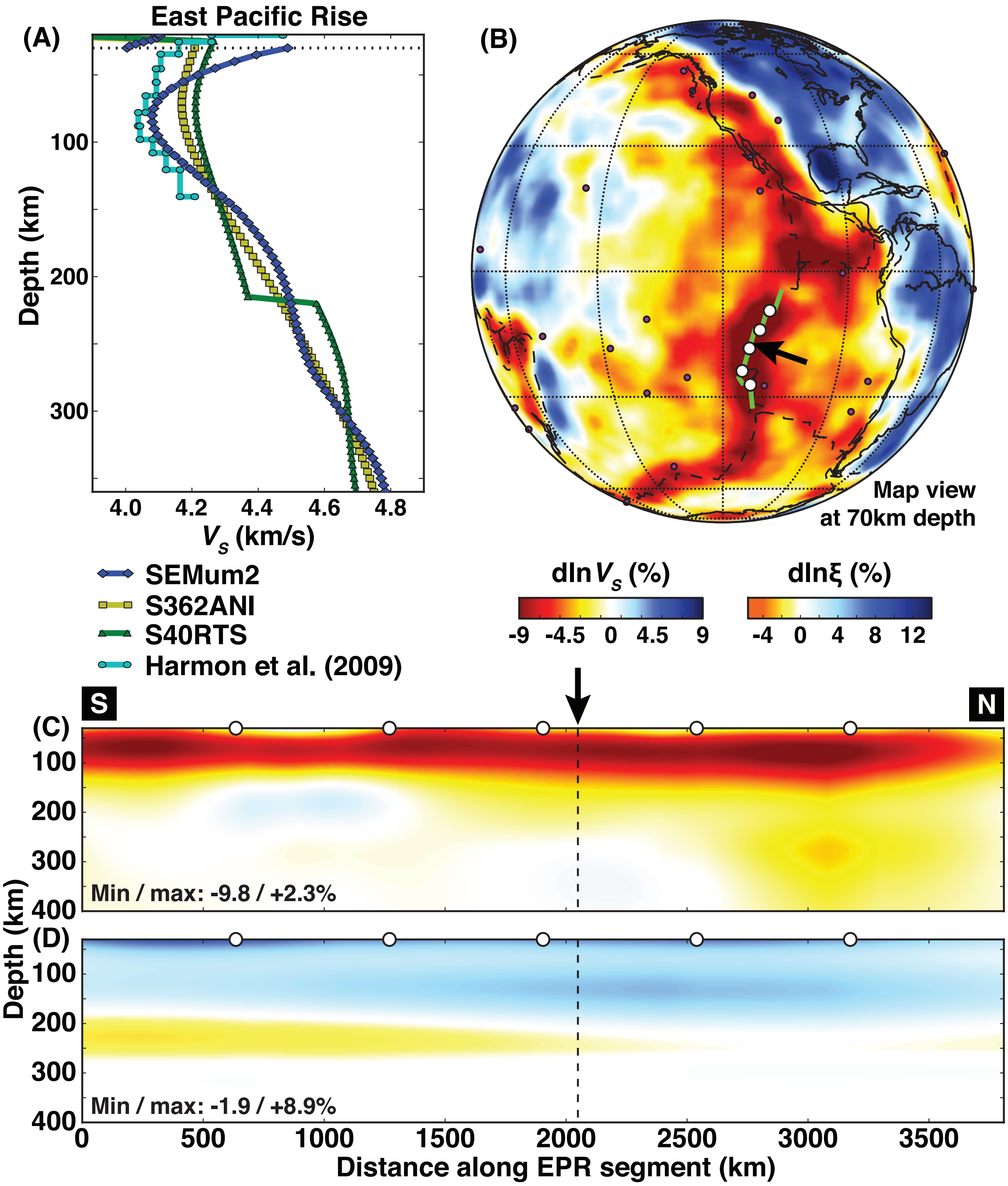


Figure 2.17: Examination of SEMum2 structure beneath a portion of the East Pacific Rise (EPR). (Contined on the following page.)

Figure 2.17: (Continued) Panels: (a) comparison of 1D V_S from global models SEMum2, S40RTS (Ritsema et al., 2011), and S362ANI (Kustowski et al., 2008a), as well as the local OBS-based V_S model of Harmon et al. (2009) (see text); (b) SEMum2 V_S and ξ cross sections following a $\sim$ 3750km segment of the EPR, traced in green in map view (showing background V_S at 70km depth). Note that in (b): the white diamond corresponds approximately to the Harmon et al. study region, the line of section is not straight, and the maximum V_S reduction is on the order of 10%.

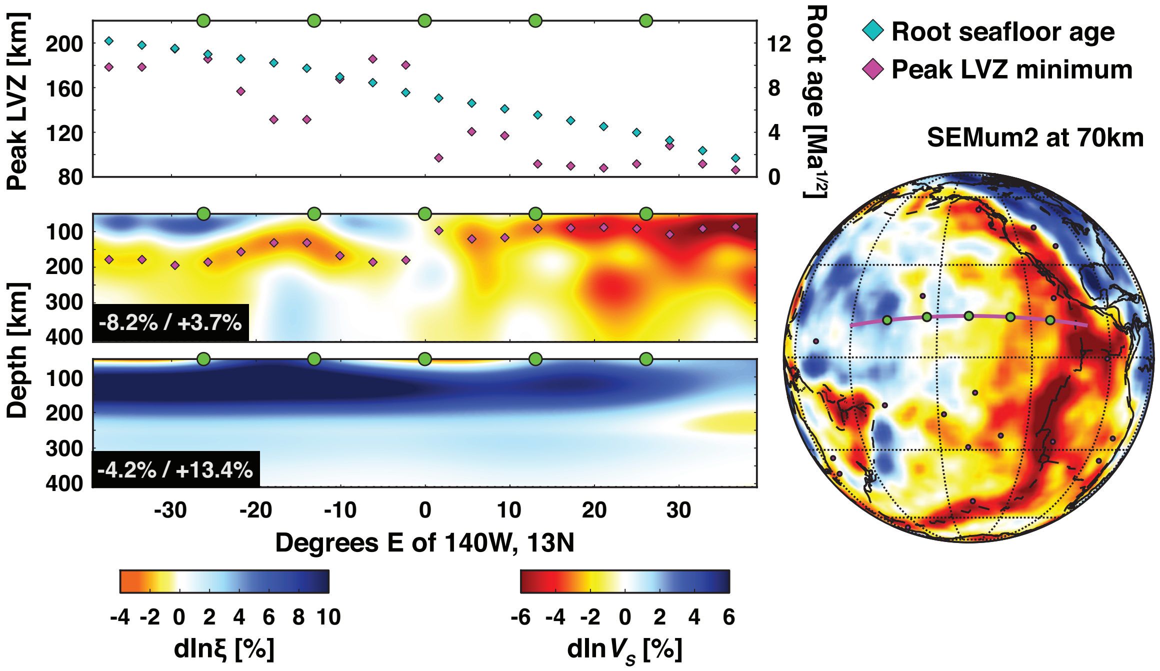


Figure 2.18: SEMum2 V_S and ξ structure along a representative cross section though the Pacific upper mantle. Also shown: depth to the peak of the low-velocity zone underlying the Pacific plate, as well as its relationship with overlying plate age.

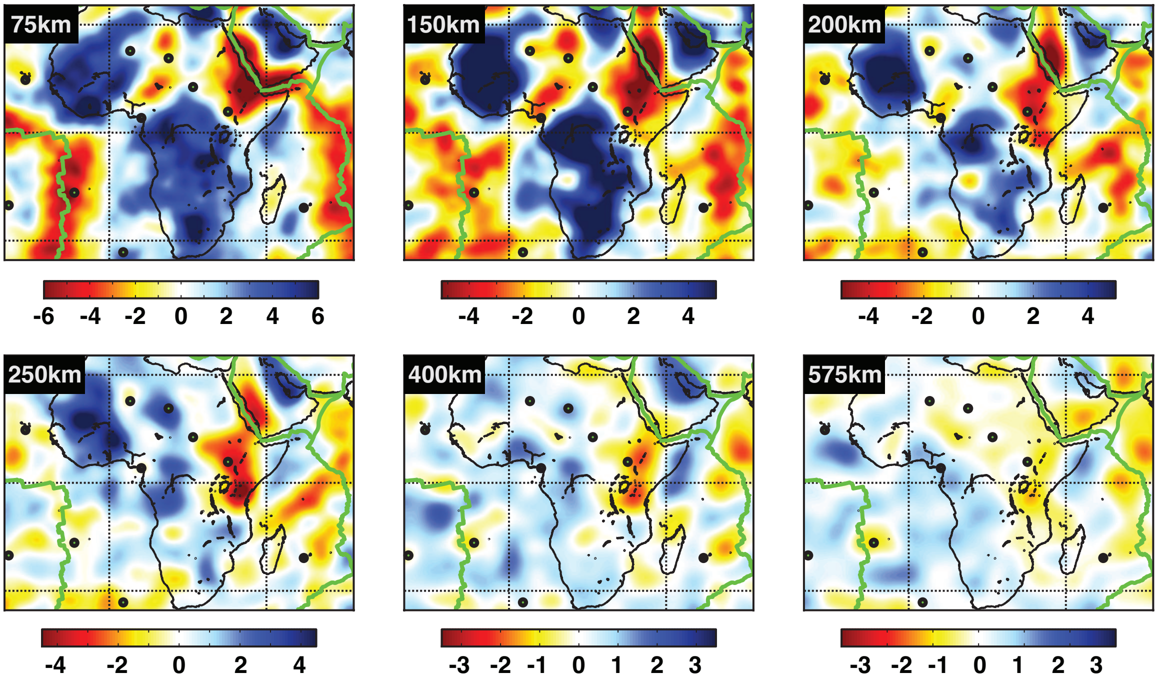


Figure 2.19: Lateral variation in V_S structure (%) at a number of upper-mantle and transition-zone depths beneath a region centered on Africa. Circles denote hotspots of Steinberger (2000). Note that the colorscale changes with depth and partially saturates in some locations.

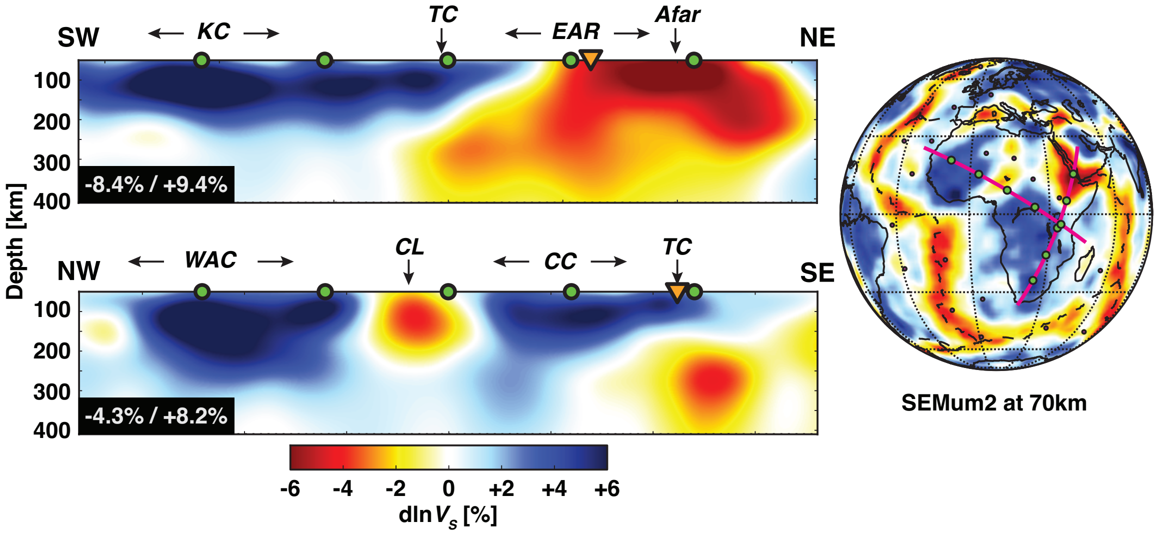


Figure 2.20: Cross-sections showing upper-mantle V_S structure beneath Africa, traversing a broad range of tectonic settings, including the East African Rift and a number of major African cratons (see text for discussion; KC: Kalahari Craton; TC: Tanzanian Craton; WAC: Western African Craton; CC: Congo Craton; CL: Cameroon Line; EAR: East African Rift). Note extremal perturbation values (inset), indicative of strong lateral heterogeneity. Inverted orange triangles denote crossing lines of section.

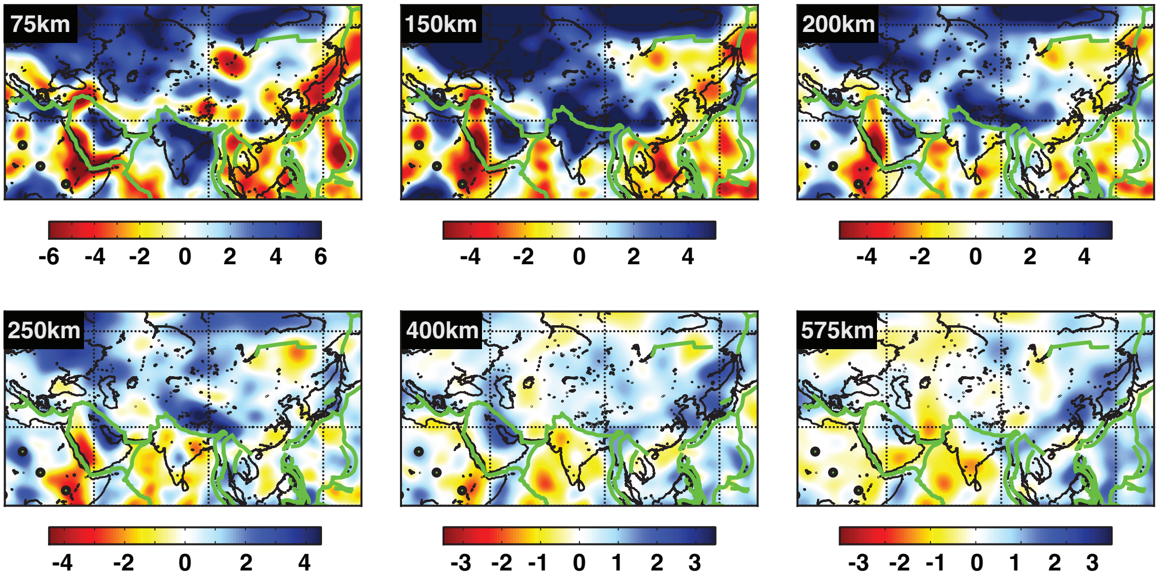


Figure 2.21: Lateral variation in V_S structure (%) at a number of upper-mantle and transition-zone depths beneath a region centered on the eastern portion of Eurasia. Circles denote hotspots of Steinberger (2000). Note that the colorscale changes with depth and partially saturates in some locations.

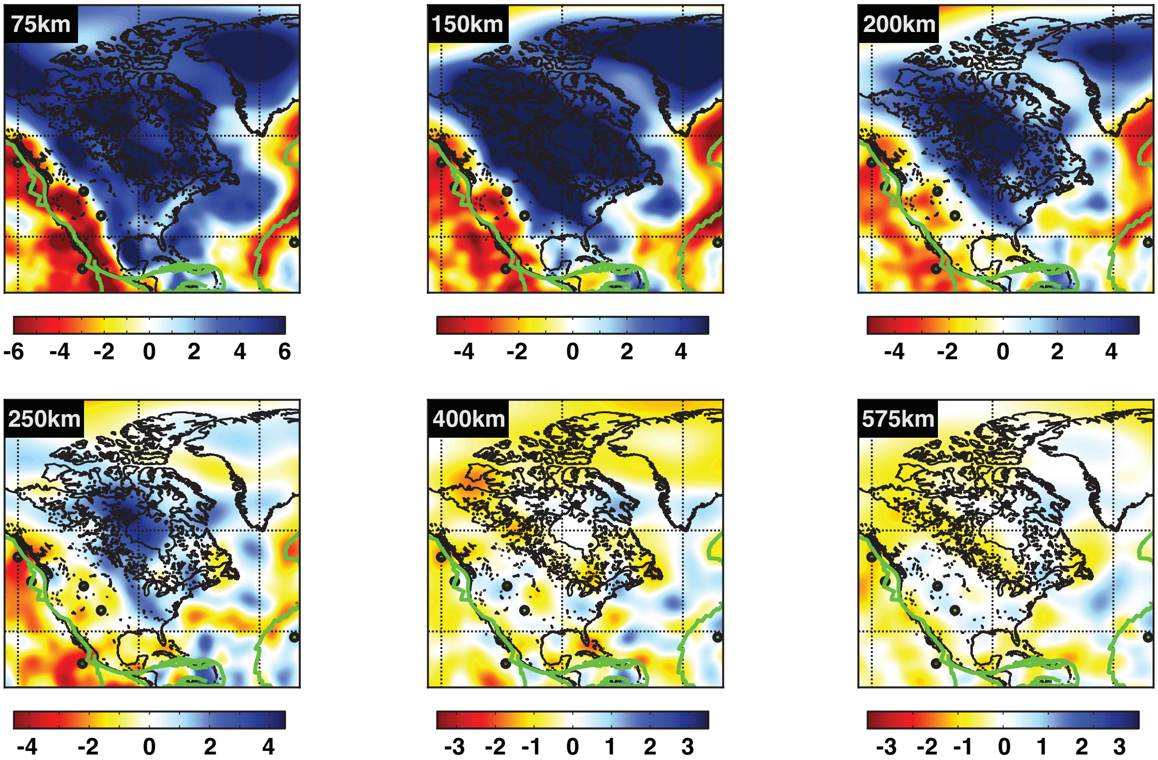


Figure 2.22: Lateral variation in V_S structure (%) at a number of upper-mantle and transition-zone depths beneath a region centered on North America. Circles denote hotspots of Steinberger (2000). Note that the colorscale changes with depth and partially saturates in some locations.

Chapter 3

Waveform tomography reveals channeled flow at the base of the oceanic asthenosphere

This chapter has previously been published in *Science* (French et al., 2013) under the title: "Waveform Tomography Reveals Channeled Flow at the Base of the Oceanic Asthenosphere." The associated online supplementary information appears in Section 3.4. Both are presented here in modified form so as to fit more cohesively into this work.

3.1 Summary

Understanding the relationship between different scales of convection that drive plate motions and hotspot volcanism still eludes geophysicists. Using full-waveform seismic tomography, we imaged a pattern of horizontally elongated bands of low shear velocity, most prominent between 200 and 350 kilometers depth, which extends below the well-developed low-velocity zone. These quasi-periodic fingerlike structures of wavelength ∼2000 kilometers align parallel to the direction of absolute plate motion for thousands of kilometers. Below 400 kilometers depth, velocity structure is organized into fewer, undulating but vertically coherent, low-velocity plume-like features, which appear rooted in the lower mantle. This suggests the presence of a dynamic interplay between plate-driven flow in the low-velocity zone and active influx of low-rigidity material from deep mantle sources deflected horizontally beneath the moving top boundary layer.

3.2 Introduction

Mantle convection is responsible for driving plate motions on Earth, but the detailed morphology of convection patterns remains unresolved. Because seismic velocities are affected by temperature, and seismic anisotropy is affected by alignment of crystals, seismic tomography can be used to map the patterns of flow in the earth's mantle. Global seismic mantle tomography has provided important constraints on the long-wavelength shear-velocity structure, highlighting in particular the correlation of velocity patterns in the top 200 km with surface tectonics and documenting the widespread presence of the low-velocity zone (LVZ) under ocean basins.

Likewise, the presence of two antipodal large low shear-velocity provinces (LLSVPs) at the base of the mantle under the central Pacific and Africa is a robust feature of all tomographic models (Lekić et al., 2012). Hotspots appear to be located preferentially above the LLSVPs (Richards and Engebretson, 1992) or on their borders (Davaille et al., 2005). There is also a striking correlation at long wavelengths between the location of the LLSVPs and high attenuation in the mantle transition zone (Romanowicz and Gung, 2002). However, plume conduits (Morgan, 1971; Rickers et al., 2012) and roll-like secondary convection patterns (Richter and Parsons, 1975) remain difficult to image tomographically.

3.3 Results and Discussion

We used full-waveform inversion, coupled with synthetic seismogram computation using the Spectral Element Method, to image global radially anisotropic shear-velocity (V_S) structure at upper-mantle and transition-zone depths. This approach is well suited to remedy the known limitations of classical tomographic techniques (Nolet and Dahlen, 2000), as already demonstrated at the local (Tape et al., 2010) and regional (Fichtner et al., 2009) scales. Our second-generation global model, SEMum2 (Chapter 2), refines an earlier one developed by our group (Lekić and Romanowicz, 2011a) and in particular includes a more realistic crust (Section 2.2.3 and supplementary Figs. 3.5 to 3.7). Compared with other global shear-velocity models (Section 2.4, Fig. 2.17 and supplementary Figs. 3.8 to 3.10), SEMum2 more accurately recovers both the depth and strength of the low-velocity minimum under ridges. It also shows stronger velocity minima in the LVZ, a more continuous signature of fast velocities in subduction zones, and stronger, clearly defined, low-velocity "conduits" under the Pacific Superswell (McNutt and Fischer, 1987) while confirming the robust long-wavelength structure imaged in previous studies (Section 2.4, Section 3.4.3.3, and supplementary Figs. 3.9 and 3.10), such as the progressive weakening and deepening of the oceanic LVZ with overlying plate age.

Cluster analysis (Lekić and Romanowicz, 2011b) of V_S profiles in the depth range 30 to 350 km in SEMum2 (Section 3.4.4) provides an objective way to analyze the model and isolates an anomalously low-velocity region – most prominent in the depth range 200 to 350 km although also reflected in the overlying LVZ (Fig. 3.1, A and B, and supplementary Fig. 3.11), organized in elongated bands, and clearest on the Pacific plate (Fig. 3.1A), where it spans from ∼100 million-year-old ocean floor to the East Pacific Rise (EPR). In a map view of SEMum2 at a depth of 250 km (Fig. 3.2A), these prominent structures appear as fingerlike zones of significantly slower-than-average V_S (∼3 to 4%). They are also present under other plates: off west Antarctica, in some parts of the North and South Atlantic and western

Indian Oceans, and possibly in the southwestern part of the Australian plate (Fig. 3.2A and supplementary Fig. 3.12). These finger-like structures are not only well-resolved in the SEMum2 model but also are robust with respect to estimated model uncertainties, are compatible with independent waveform data, and cannot be explained by unmapped azimuthal anisotropy in our inversion (Section 3.4.5 and supplementary Figs. 3.15 to 3.18).

We find that these low-velocity fingers (LVFs) are oriented subparallel to the direction of absolute plate motion (APM) (Fig. 3.2B and Section 3.4.6.1) (Kreemer, 2009). Perpendicular to the APM, the alternating zones of very low and somewhat higher-than-average V_S have a wavelength of $\sim$2000 km, as illustrated in depth cross-sections on the Pacific Plate (Fig. 3.3, B and C). This wavelength corresponds to a peak in power in the geoid, as determined with directional wavelet analysis, which is also aligned with the direction of the APM (supplementary Fig. 3.13 and Section 3.4.6.3) (Hayn et al., 2012). In cross-section parallel to the APM, the contrast in structure at depth within and adjacent to the LVFs is very clear (Fig. 3.3, D and E). The LVFs extend for many thousands of kilometers and reach beneath the conventional LVZ, which bottoms at an approximately constant depth of $\sim$150 to 200 km (Fig. 3.3D). Below 200 km, velocities are as low within the LVFs as they are between fingers in the LVZ, despite the greater depth (Fig. 3.3). In contrast, the EPR itself is a shallow feature in V_S (e.g. Fig. 2.17 of Chapter 2) but stands out in radial anisotropy as a zone where $V_{SV} > V_{SH}$ in the depth range 150 to 300 km (supplementary Fig. 3.14C). This indicates that dominantly horizontal flow in the LVFs away from the ridge transitions to dominantly vertical flow under the ridge. Although local minima in the LVZ are associated with the LVFs, the strongest minima in the LVZ appear under ridges (Fig. 3.3 and supplementary Fig. 3.14).

Such alternating zones of high and low velocities have previously been found along the Fiji-Hawaii corridor (Katzman et al., 1998), and an elongate band of low velocities, within a similar depth range to the LVFs, has recently been imaged in the south Atlantic (Colli et al., 2013). At a smaller scale, tomographic maps based on the PLUME experiment (Wolfe et al., 2011; Laske et al., 2011) show a zone of fast velocities surrounding Hawaii, particularly strong in V_S to the southwest near 300 km depth. In our model, this corresponds to a domain of higher velocities between LVFs. Our study thus ties together these isolated observations, suggesting that they are manifestations of a single, consistent, large-scale pattern of LVFs (Figs. 3.2 and 3.3) aligned with the APM, present in the oceans worldwide, and extending in a narrow depth range below the LVZ.

At the global scale, many of the fingers underlie regions associated with hotspot tracks or seamount chains: for example, in the northwest Atlantic, the New England seamount chain, in the South Atlantic, the Walvis ridge, or the Cape Verde track, and in the Indian Ocean, portions of the Reunion hotspot track (Fig. 3.2 and supplementary Fig. 3.12). In contrast to the top 300 km, deeper V_S structure in the region spanning from the Pacific Superswell to Hawaii is characterized by vertically elongated plumelike conduits (Fig. 3.4). Not all LVFs are connected to the conduits below, and the latter are not straight, but meander with depth and appear to be rooted in the lower mantle. The main hotspots in the central Pacific are located generally in the vicinity of the deep conduits but not immediately above them

(Fig. 3.4D).

Although the resolution of our modeling enables the detection of the stronger mantle upwellings, such as those beneath Hawaii and the Superswell, the actual plume conduits could be narrower, and other, weaker ones, may not yet be resolved and will necessitate modeling at shorter periods (Rickers et al., 2012). The absence of a direct vertical correspondence between hotspot locations and the imaged plumes suggests a complex interaction of the upwelling flow with the lithosphere (O'Connor et al., 2012). Above ∼350 km depth, two interacting structural patterns appear to dominate: (i) the increasing depth and decreasing strength of the LVZ as a function of age in the depth range 50 to 200 km (Fig. 3.1B) and (ii) the difference in velocity – and therefore likely temperature and/or composition, as well as viscosity – within and outside of LVFs (Figs. 3.1B and 3.3 and Section 3.4.6.2). In some locations, the LVFs appear to feed from the quasi-vertical conduits, suggesting deflection and channeling in the asthenosphere of active upwelling from low-viscosity plumes - similar to viscous fingering experiments in which a low-viscosity fluid is injected between two rigid horizontal plates or stratified, higher-viscosity fluids (Snyder and Tait, 1998). This horizontally deflected flow then aligns in the direction of plate motion, driven by a combination of asthenospheric return flow (Morgan et al., 1995; Harmon et al., 2011) and upwelling-induced flow directed toward pressure minima at ridges (Toomey et al., 2002; Hillier and Watts, 2004; Ballmer et al., 2013). The pattern of radial anisotropy in the vicinity of the ridge (supplementary Fig. 3.14C) further supports active ridge-ward flow in these channels. Active influx from deep upwellings deflected toward the ridge may be enhanced by flow in a narrow low-viscosity layer (Hoink et al., 2012). The absence of any distinct deeper low-velocity structure beneath ridges (Fig. 2.17 of Chapter 2), and the fact that some of the LVFs terminate at ridges (such as the Antarctic plate), confirms the passive nature of mantle upwelling beneath ridges.

Whether or not these observations can be explained by viscous fingering and channelization alone or in combination with other phenomena, such as secondary convection, is unknown. Other studies have described evidence for viscous fingering on the Pacific plate, aligned with the plate motion, albeit at an order-of-magnitude smaller spatial scale than seen here (Harmon et al., 2011). The width of the fingers we observed (∼1000 km) is large compared with the thickness of the channel (up to 350 km), whereas typical scaling in laboratory or numerical fingering experiments obtain a width-to-thickness ratio of ∼2 (Snyder and Tait, 1998). Secondary convection in the form of Richter rolls occurs with a horizontal-to-vertical scaling of 1 [albeit in a constant-viscosity fluid (Richter and Parsons, 1975), a condition quite different from that in the Earth] but has previously been sought in the upper mantle at smaller scales than seen here (Ballmer et al., 2009). The LVFs are observed both below the fast spreading Pacific plate, where roll-like secondary convection may be expected, and below slow-moving plates [for example, the Antarctic plate (Fig. 3.2 and supplementary Fig. 3.19) and the Atlantic Ocean (supplementary Figs. 3.12 and 3.20)], where Richter rolls are unlikely to form. Comparison of the LVFs and Pacific geoid undulations with the same orientation and wavelength (supplementary Fig. 3.13) (Hayn et al., 2012) may also provide insight into causative dynamics. Indeed, in a simple Richter-like secondary convection scenario one would expect the bands of quasi-APM orientation in the geoid to be aligned vertically with the up- (LVF) and down-welling (inter-LVF) limbs. Instead, the LVFs fall at the edges or between these features – an observation more consistent with the presence of channelized flow (Section 3.4.6.3). A further clue as to the nature and origin of the global pattern of LVFs that we document here might come from the geochemistry of mid-ocean ridge basalts, whose long-wavelength isotopic anomalies fluctuate with a similar pattern along the mid-Atlantic Ridge (Hofmann, 2007).

3.4 Detailed analysis of uncertainties

3.4.1 Motivation

These materials are intended to provide support to inferences appearing in the previous section. In particular, we describe the provenance of the updated SEMum (Lekić and Romanowicz, 2011a) model, SEMum2 (Chapter 2), upon which our analysis is based (Section 3.4.2). We also present a systematic series of tests that quantify model performance (Section 3.4.3) as well as resolution and uncertainties (Section 3.4.5). We analyze the effects of three different sources of error on our model: 1. Uncertainty in the surface-wave dispersion measurements used to constrain crustal structure; 2. Potential bias due to parameterization of crustal structure; and 3. Uncertainty due to noise in the waveform dataset. Further, we carry out four different types of error analysis to quantify the reliability of our tomographic model: 1. Standard linear resolution analysis using a resolution operator; 2. Bootstrap analysis of stability of model structure; 3. Independent validation against waveform data that (a) were not used in the construction of the tomographic model and (b) extend to shorter periods than used in the inversion; and 4. Independent validation against structures inferred by regional / smaller-scale studies. Additionally, we present further discussion of: (a) the oceanic upper-mantle cluster analysis of SEMum2 (Section 3.4.4), (b) the model of absolute plate-motion cited in the main text (Section 3.4.6.1), (c) thermal implications of imaged V_S structures (Section 3.4.6.2), and (d) comparisons with long-wavelength geoid undulations (Section 3.4.6.3).

3.4.2 The SEMum and SEMum2 models

The SEMum model (Lekić and Romanowicz, 2011a) was the first global upper-mantle shear-velocity model developed using full-waveform inversion combined with spectral-element forward modeling (SEM: e.g. Komatitsch and Vilotte, 1998). SEMum2, introduced in detail in Chapter 2, is the successor to SEMum – derived using similar methods (namely, the "hybrid" waveform inverison approach of Section 2.2.2), but with two key improvements aimed at enhanced recovery of oceanic upper-mantle structure: 1. A new implementation of the crustal modeling scheme used in SEMum (Section 2.2.3); and 2. Inversion for shorter-wavelength mantle structure, supported by a change in parameterization and resolution tests (Sections 2.2.1 and 2.3.3). In this section (3.4.2), we focus specifically on these two areas of improvement in the context of a broader discussion of model uncertainties and resolution, referring the reader to the relevant sections of Chapter 2 for background on the inversion procedure as appropriate.

3.4.2.1 Crustal modeling

To render SEM computations tractable, SEMum used long-period waveform data (60-400s) combined with a 60 km-thick crustal layer, whose seismic response is equivalent to that of the earth's true crust in the frequency range of interest. In order to improve recovery of shallow oceanic upper-mantle structure, we introduced a new crustal modeling technique, enabling more realistic lateral variation in crustal thickness: approximately honoring that of the Crust2.0 model (Bassin et al., 2000) in continental regions, while fixed at 30 km in the oceans. For a detailed discussion of the development of the crustal model, we refer the reader to Section 2.2.3 of Chapter 2. Here, we discuss two particular sources of uncertainties associated with our treatment of the crust, specifically in the context of potential contamination of our mantle model.

Uncertainty due to group-velocity dispersion data

As noted in Section 2.2.3, the crustal layer is calibrated to match surface-wave group-velocity dispersion maps. We find it preferable to match these dispersion data directly, reflecting the integrated effect of Earth's crust on the wavefield, rather than matching dispersion predicted by an a priori model with poorly quantified uncertainties. Still, the dispersion maps also have associated uncertainties (Shapiro and Ritzwoller, 2002), so it is reasonable to ask whether error in the maps, absorbed into the crustal layer, can significantly affect the long-period wavefield and in turn bias our mantle model. To answer this question, we built an ensemble of crustal models following the 2-step procedure described in Section 2.2.3.2, again derived from 25-60s dispersion maps, but now with added Gaussian noise (σ=40m/s, similar to uncertainties reported in (Shapiro and Ritzwoller, 2002)). We calculated SEM synthetics in the period band of our mantle inversion (60-400s) for each crustal-layer realization and examined variation in the resulting waveforms. As seen in Fig. 3.5 for a representative event from our dataset, the ensembles of synthetic traces clearly do not differ significantly. Thus, we are confident that variation in crustal-layer velocity structure due to uncertainties in the dispersion maps cannot appreciably bias our mantle model.

Uncertainty due to crustal parameterization

A second potential source of model uncertainty is systematic error due to parameterization (Section 2.2.3.2). In particular, we ask whether our choice of Lagrange interpolants or scaling relations for ρ and V_P (Brocher, 2005) can adversely affect the 60-400s waveforms in our mantle inversion – especially in the case of oceans, where the 30 km crustal-layer thickness should necessitate larger velocity and radial-anisotropy variations than required in continents in order to match the observed dispersion. To answer this question quantitatively, return to the earlier analysis performed in Chapter 2. In particular, in Section 2.2.3.3, we calculated synthetic 25-60s dispersion data for an ensemble of realistic thinly layered oceanic crustal structures (Crust2.0 (Bassin et al., 2000) with randomly-perturbed layer thicknesses) atop oceanic upper-mantle structure sampled from SEMum2. We fit the synthetic data using the same 2-step scheme, and compared full long-period waveforms (60-400s) calculated for the input layered models and the output models with a 30 km crust. Illustrated in Figs. 2.4 and 2.5, the results confirm that parameterization – which is the only source of errors in these noise-free tests – has a negligible effect on long-period waveforms. Both fundamental-mode and overtone wavetrains generally show excellent fits between the layered and 30 km smooth-crust models, with slightly degraded fits exhibited *only* by fundamental modes in the pathological test case shown.

3.4.2.2 SEMum2 mantle-model update

The SEMum2 model reflects two additional iterations of SEM-based hybrid waveform inversion following the introduction of the new crustal layer, as well as one further iteration following a change in parameterization of the mantle model. The first two iterations were necessary in order for shallow upper-most mantle structure (particularly in the oceans) to adjust to the new crustal modeling approach. Here, we focus on the final iteration, which enabled us to take full advantage of the constraints on earth structure supplied by our waveform dataset and improve the lateral resolution of our model.

As discussed in Section 2.2.1, both the SEMum and SEMum2 mantle models are parameterized laterally in spherical splines (Wang and Dahlen, 1995) and radially in cubic b-splines (Mégnin and Romanowicz, 2000). Similarity between the long-wavelength structure of SEMum and that obtained with the improved crustal model (and two additional iterations) of SEMum2 suggested that the iteratively linearized tomographic inversion had converged. Combined with the excellent recovery of small-scale structures apparent in the resolution operator analysis of SEMum (see Lekić and Romanowicz (2011a)), this observation motivated us to refine the lateral parameterization of SEMum2, which would allow us to recover more detailed images of the particular low-velocity structures discussed in this chapter. We halved the spherical-spline knot spacing used for V_S, from 4° to 2°, and performing an additional waveform-inversion iteration, reducing the minimum a priori correlation length for V_S permitted in our formulation from 800 km to 400 km (see Section 2.2.2 for details).

While we could not rule out ahead of time the possibility that no additional recoverable information was present in the waveform data and that we would fit only noise, we subsequently verified that this was not the case in two ways. First, we performed an additional resolution analysis, which demonstrated that structures smaller than those permitted by the earlier lateral parameterization could indeed be retrieved (this statement should be interpreted with the usual caveats regarding checkerboard tests, see Section 3.4.5.1 as well as the earlier discussion in Section 2.2.2). Second, the inversion with the finer parameterization resulted in improved fits for data that were *not* originally included in the inversion; among this held-out data, 168 additional waveform windows (fundamental and/or overtone-mode wavetrains), previously excluded from the inversion due to large residuals, became admissible. Thus, together with the demonstrations of model performance in Section 3.4.3, we are confident that our progression to shorter-scale structure is justified.

3.4.3 Model performance and validation

3.4.3.1 Data and variance reduction

In Table 3.1, we summarize SEMum2 performance in terms of waveform variance reduction, listed for each component and windowed-wavetrain type separately, in addition to the corresponding numbers of windows and data points contained therein. In total, our dataset contains more than 99,000 individual windowed wavetrains, corresponding to over 5,200,000 data points (sampled at 30s – the Nyquist limit for our 60s minimum-period – implying maximal independence between neighboring points without information loss). We note that our fundamental-mode and overtone waveform dataset is considerably larger than that used in previous global models from our group (see Table 1 of Panning and Romanowicz (2006), for comparison). Further, considering total number of data points alone, 5.2M surpasses previous UCB whole-mantle models even when including their extensive body-waveform datasets (Panning and Romanowicz, 2006).

3.4.3.2 Independent validation

Waveform fits for earthquakes not in our dataset, and at periods shorter than used in our inversion, provide another way to evaluate model performance. We demonstrate this ability to generalize in Fig. 3.6, where we observe close fits between data and synthetics, particularly in phase but also in amplitude, now calculated down to 40s instead of the 60s cut-off for waveforms used in the inversion, and for 3 events not previously included in our modeling. Performance may also be quantified in terms of surface-wave dispersion. In Fig. 3.7 we compare phase velocities measured from both the observed data and SEMum2 synthetics, again for an event not included in our inversion, using the multitaper technique of Laske and Masters (1997). We find that dispersion fits are quite good – in general within 10m/s where the measurements remain stable – and further represent a conservative estimate of model quality, given the sensitivity of single-station phase measurements to source uncertainty (at least for observed data – sources used in the SEM are well defined). Further, Fig. 3.7 also includes phase velocities inferred from the single-station phase-anomaly dataset of Ekström (2011) (where available), which are in general compatible with our measurements to within associated uncertainties.

3.4.3.3 Comparison to other models

Global V_S structure in the SEMum2 model is shown in Fig. 3.8 at a number of upper-mantle and transition-zone depths, accompanied by structure from the recent global models of Kustowski et al. (2008a) and (Ritsema et al., 2011). While the three models agree well at long wavelengths, it is clear that SEMum2 is characterized by stronger, more-focused anomalies. Local- and regional-scale tomography provides an additional avenue for model validation: especially important given the unusually large-amplitude heterogeneity in SEMum2 (e.g. low-velocity anomalies beneath the East Pacific Rise consistently in excess of 9%). Indeed, in Fig. 2.17 of Chapter 2, we saw 1D V_S profiles from SEMum2 and the same two recent global models, collocated at the EPR with the OBS-based tomographic study of Harmon et al. (2009) for which we show a 1D-mean centered on the ridge axis (averaging over approximately one SEMum2 correlation length; Section 3.4.2.2). While structure above 50 km cannot be interpreted due to the 30 km crustal layer (Section 3.4.2.1), SEMum2 recovers both the strength and depth of the low-velocity zone beneath the EPR imaged by Harmon et al. (2009) far more closely than either of the other global models. With this local validation of amplitudes in mind, we compare SEMum2 Pacific upper-mantle and transition-zone cross-sections to these same two global models in Figs. 3.9 and 3.10. We see that SEMum2 amplitudes are systematically stronger, particularly for low velocities, while morphology of structure is generally consistent at long wavelengths. Notably, SEMum2 structure appears more concentrated and contiguous, particularly in the images of subducted slabs and columnar low-velocity features discussed in the main text (e.g. beneath the Pacific Superswell (McNutt and Fischer, 1987)).

3.4.4 Cluster analysis

Given an ensemble of data points distributed among a set of clusters – defined by their means, which are not known a priori – as well as an assumed number of clusters, the k-means algorithm (MacQueen, 1967) provides an objective estimate of both the unknown means and data-point partitioning. As shown in Lekić and Romanowicz (2011b), this technique can be applied to V_S profiles sampled from tomographic models to discover objective classes of model structure – both their lateral extent and appearance. We applied this analysis to oceanic upper-mantle (30-350 km) structure in SEMum2, and feature a pertinent subset of the results in Fig. 3.1 of the main text. For completeness, the full results can be found in Fig. 3.11, showing global distribution of the structural classification, as well as more-detailed characterization of intra-cluster variation. Further, to estimate uncertainties in the cluster centroids, we repeat the cluster analysis for each of the bootstrap models obtained in Section 3.4.5.2 below, and show the resulting ensembles of cluster-centroid profiles in panel (B) (analogous to the SEMum2 cluster centroids shown in panel (B) of Fig. 3.1). We note that the 2σ uncertainty characterizing each ensemble of profiles is indistinguishable from the width of the ensemble itself at this plotting scale. Indeed, it is clear that the different classes of SEMum2 structure revealed by the cluster analysis remain distinct at the 2σ level, much like the underlying V_S anomalies associated with the LVFs, as discussed below in Section 3.4.5.2 and demonstrated in Fig. 3.16.

3.4.5 Resolution and uncertainties

3.4.5.1 Resolution analysis

Pitfalls of the "standard" resolution analyses employed in many tomographic studies are well known, including strict validity only for linear problems (Tarantola, 2005) and potential for misleading results Lévêque et al. (1993) (see (Lekić and Romanowicz, 2011a) and Section 2.3.3 of Chapter 2 for discussion in the context of waveform inversion). Further, such analyses cannot be used to address uncertainties arising from inaccuracy of the modeling theory. Therefore, standard resolution analysis is likely to overestimate the resolution of models constructed using approximate wave propagation approaches compared to those constructed using more accurate wave propagation calculations, such as SEMum and SEMum2. Still, these analyses can provide useful insight on model parameterization and data coverage. Indeed, poor fidelity to a test model (e.g. checkerboard) upon projection onto the model basis clearly implies that structure of similar sharpness or scale cannot be imaged, while data coverage that is too sparse relative to a priori constraints on model smoothness (e.g. correlation length) can result in poor test-model recovery, showing gaps or ray-like smearing – again indicative that small scales may not be well resolved.

Keeping these points in mind, we present resolution analyses for V_S at LVF depths in Fig. 3.15, using the common checkerboard pattern at global scales, as well as isolated point and band-like structures in the Pacific basin motivated by the columnar anomalies and LVFs. As such, these analyses are meant to complement those reported in Section 2.3.3 of Chapter 2. In panel (A), we see that checkerboard inputs are very well retrieved – for both large (2200 km) and small-scale (800 km) structures placed at LVF-relevant depths in equatorial and polar orientations - with no evidence of ray-like smearing or gaps in sensitivity, even at the poles. As such, we expect V_S to be well resolved at 800 km scales in the upper mantle, coincident with the minimum distance of significant correlation permitted in our inversion (2 x 400 km), and only slight degradation in the transition zone - certainly at scales shorter than the well-recovered 2200 km case. In panel (B), we show results for isolated input perturbations at two representative locations to assess vertical smearing: Hawaii, with fairly dense data coverage, and the south-Pacific Superswell (equidistant from Tahiti and Pitcairn) with possibly sparser coverage. Though input amplitudes are not necessarily recovered, morphology with depth is, and we expect accurate distribution of V_S anomalies down to at least 400 km, with only slight vertical smearing below. Finally, in panel (C), we show results for band-like test structures (450 and 900 km widths), concentrated in the 200-350 km depth range, and oriented both normal and parallel to Pacific APM (Kreemer, 2009). Both scales and orientations of bands are well retrieved – demonstrating that: (a) the $\sim$2000 km APM-normal periodicity of the Pacific LVFs cannot be an artifact of our model parameterization or data coverage; and (b) the widths of the Pacific LVFs are well resolved.

3.4.5.2 Uncertainty estimation

While resolution analysis provides a useful assessment of certain aspects of model quality, quantitative estimates of model uncertainties are clearly desirable. One efficient uncertainty-estimation technique is the bootstrap (Efron and Tibishirani, 1991), in which an ensemble of models is obtained by inverting randomly resampled data and ensemble variation is taken as a proxy for uncertainty (e.g. Panning and Romanowicz, 2006). Here, for simplicity, we partition our data into 6 subsets (event month modulo 6) and construct resampled data by selecting 4 subsets without and 2 with replacement, such that at most 50% of each resampled set is a repeated subset and 33% of available subsets are missing. For 20 resampled-data realizations, we re-run the final iteration of our inversion (a Gauss-Newton update scheme forces the model to adjust very quickly to changes in the data). Further, we do not reapply our path-weighting scheme (Li and Romanowicz, 1996) to the resampled data, increasing the likelihood of large model variations. In Fig. 3.16, we show map views at 250 km depth of both the resulting 2σ uncertainties and SEMum2 V_S structure with regions below the local 2σ masked. Assuming our uncertainty estimates are valid, it is clear that the LVFs are significant at the 2σ level.

3.4.5.3 Stability of model structure

Model structure may further be considered robust if it remains stable following the introduction of new independent data. With this in mind, we performed one additional iteration of the hybrid inversion – a hypothetical update to the SEMum2 model – using a new, 144-event waveform dataset that is wholly disjoint from that in our modeling. As noted above (Section 3.4.5.2), our use of a Gauss-Newton update scheme implies that model structure should adjust very quickly if required by the new data. In Fig. 3.17, we show V_S structure at 250 km depth in SEMum2 (also shown in Fig. 3.2 of the main text) alongside that from the updated model described here. It is clear that the LVF structures remain stable following application of an independent waveform dataset. Indeed, SEMum2 model structure and that retrieved in the hypothetical update remain correlated at $R > 0.95$ when considering up to spherical-harmonic degree 12 and $R > 0.9$ to degree 48 throughout the upper mantle. This test again emphasizes the robust nature of the LVF structures seen in the SEMum2 upper mantle.

3.4.5.4 LVF orientations and possible artifacts due to azimuthal sampling

Finally, we wish to explain why the orientations of the LVFs cannot be due to poor azimuthal sampling – either in the form of ray-like smearing or as artifacts of un-modeled azimuthal anisotropy. As noted previously, LVFs tend to align preferentially to APM (Kreemer, 2009), which in turn corresponds to streamlines of solid rotation on the surface of the sphere. Except in the degenerate case arising 90 from the pole of rotation, these paths cannot be fit with a great circle – precluding correspondence between LVFs and ray paths. This point is demonstrated in practice in Fig. 3.18, where we compare LVF orientations from Fig. 3.2B in the Pacific and West Antarctic to a sparse, but representative, sampling of rays that traverse these regions from Western-Pacific and Arctic sources to Western-Hemisphere and ocean-island receivers (the distribution of rays when transposing the source and receiver sides would be similar). As expected from the geometric argument above, it is immediately clear from examination of Fig. 3.18 that LVFs cannot be fit by, and cannot correspond to, ray paths. Indeed, this result complements those of our resolution analysis (Section 3.4.5.1), which previously ruled-out ray-like smearing. Further, it is extremely unlikely that the LVF structures are artifacts due to un-modeled azimuthal anisotropy, because of two reasons. First, the azimuthal coverage of our dataset is very homogeneous, particularly in the central Pacific, where the LVFs are most clearly developed, as seen in the rose diagrams presented in Fig. 2.8 of Chapter 2. Second, the bootstrap resampling analysis, which estimates model uncertainties by repeatedly simulating artificially poor data coverage, should yield much larger uncertainties for the LVF structures if our dataset were characterized by uneven azimuthal sampling (Section 3.4.5.2, Fig. 3.16).

3.4.6 Plate motion model, temperature, and the geoid

3.4.6.1 Absolute plate motions

In the main text, we present comparisons between LVF orientations and the absolute plate motion (APM) model of Kreemer (2009). This APM model combines a comprehensive geodetic dataset, used to constrain relative plate motions, with SKS splitting measurements that constrain the orientation of plate motion relative to the deep mantle – thereby providing an alternative to hotspot-based reference frames, which may be affected by uncertainties in assumed hotspot motion / fixity. Still, this model is consistent with those based on the hotspot frame for the Pacific plate, such as HS3-NUVEL1A (Gripp and Gordon, 2002), while showing marked differences in APM direction on several slow moving plates – in particular the Antarctic and African plates (it also reflects different assumptions regarding deformation of the African plate – see Kreemer (2009)). Because SEMum2 was constructed using completely independent data and methods from Kreemer's APM model, the striking similarity between plate-motion and LVF orientation in all oceans provides a validation of both models and supports the use of SKS splitting measurements as constraints on the net rotation of the lithosphere with respect to the deep mantle.

3.4.6.2 Temperature variations

Assuming a purely thermal origin for lateral V_S variations in the oceanic upper mantle of SEMum2, we estimate corresponding temperature differences to be on the order of 200°C. This estimate is derived using a model of anharmonic V_S temperature dependence for realistic upper-mantle compositions developed by Stixrude and Lithgow-Bertelloni (2005), and corrected for anelastic dispersion using an ensemble of physically-plausible attenuation models in the manner of Cammarano et al. (2009). The latter correction is parameterized in terms of homologous temperature, which in turn relied on upper-mantle solidus temperatures derived by Hirschmann (2000). Together with absolute V_S from SEMum2, these models form a non-linear system of equations, which may easily be solved numerically – leading to the above estimates of spatial variations in temperature. Importantly, because these estimates assume a purely thermal origin for lateral variations in V_S – ignoring the profound effects of water or partial melt, for example, on shear-wave velocities – they represent upper bounds on the associated temperature variations. Future efforts toward SEM-based attenuation tomography will allow for better characterization of the anelastic structure associated with these V_S anomalies, and thus may provide insight on the roles of temperature vs. composition.

3.4.6.3 Comparison with Pacific geoid undulations

As noted in the main text, the ∼2000 km wavelength of the LVFs corresponds to a peak in power in Pacific geoid undulations observed by Hayn et al. (2012), which are also preferentially aligned with absolute plate motion. This geoid signal provides strong evidence for patterns of upper-mantle heterogeneity consistent with the LVFs in both scale and orientation, although it cannot be used to isolate a specific depth range for such features. In Fig. 3.13, we compare the locations and orientations of the LVFs to the distribution of preferential orientation seen in ∼1850 km scale Pacific geoid undulations (Hayn et al., 2012). Interestingly, the paths associated with the LVFs do not correspond exactly to the broad bands of quasi-APM orientation seen in the geoid: in general, the LVFs fall either between or at the edges of these regions, where more complex variation in preferential orientation of the geoid pattern is observed. In the main text, we note that this observation is not compatible with simple secondary convection in the form of Richter rolls (Richter and Parsons, 1975), where one would expect regular bands of quasi-APM orientation in the geoid to be aligned with the up- (LVF-centered) and down-welling (inter-LVF) limbs. Instead, we note that the observed pattern of preferential geoid orientation may be more compatible with channelization and fingering of plume-fed material in the upper mantle. In this latter case, smaller-scale convective phenomena would be expected to arise within the propagating finger itself (Ballmer et al., 2013), resulting in more complex local variation in geoid signature. This suggestion could be supported by the imaged variability of shear velocity within the LVFs, which despite being continuous and consistently low-velocity (as demonstrated in the cluster and bootstrap resampling analyses), still exhibit internal V_S minima. Conversely, large expanses of ambient mantle structure between LVF channels are expected to be relatively laterally uniform, and are also APM-aligned – indeed they result from the presence of the through-going LVFs. Therefore, in this scenario, the most laterally coherent regions of APM-parallel preferential orientation in the geoid are actually associated with the comparatively fast bands of ambient mantle, while the system as whole (the ∼2000 km periodicity in particular) is controlled by the LVFs.

	L			**T**			**Z**		
	VR	N_{wp}	N_{d}	VR	N_{wp}	N_{d}	VR	N_{wp}	N_{d}
F	68.1	7,964	428,700	77.3	13,438	509,014	69.9	13,562	731,590
H	80.9	14,475	913,434	71.2	14,640	511,837	79.5	22,296	1,341,729
M	74.9	3,434	219,187	82.2	4,596	232,885	77.6	4,951	313,140

Table 3.1: Final-iteration waveform variance-reduction for SEMum2, and associated numbers of accepted waveform windows N_{wp} and data points N_{d}, listed by component (L,T,Z) and windowed-wavetrain type (fundamental, higher, and mixed-mode). Over 99,000 waveform windows, corresponding to more than 5,200,000 data points, are accepted into the final inversion iteration. Here, variance reduction is defined in percent of one minus the squared 2-norm of the residual between the data and synthetics normalized by that of the data: $100 \times [1 - \|\mathbf{d} - \mathbf{g}(\mathbf{m})\|_2^2 / \|\mathbf{d}\|_2^2]$.

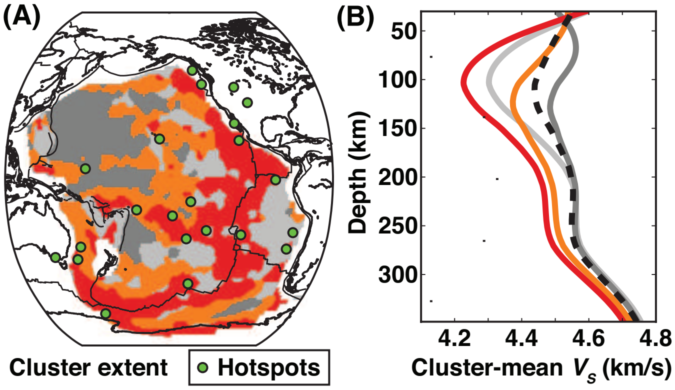


Figure 3.1: Cluster analysis of oceanic upper-mantle shear velocity. (A) Pacific portion of a global cluster analysis (e.g. Lekić and Romanowicz, 2011a) of model SEMum2 assuming four regions in the oceans (see Section 3.4.4). (B) Corresponding average depth profiles in each of the four regions (global average also shown, black dashed line).

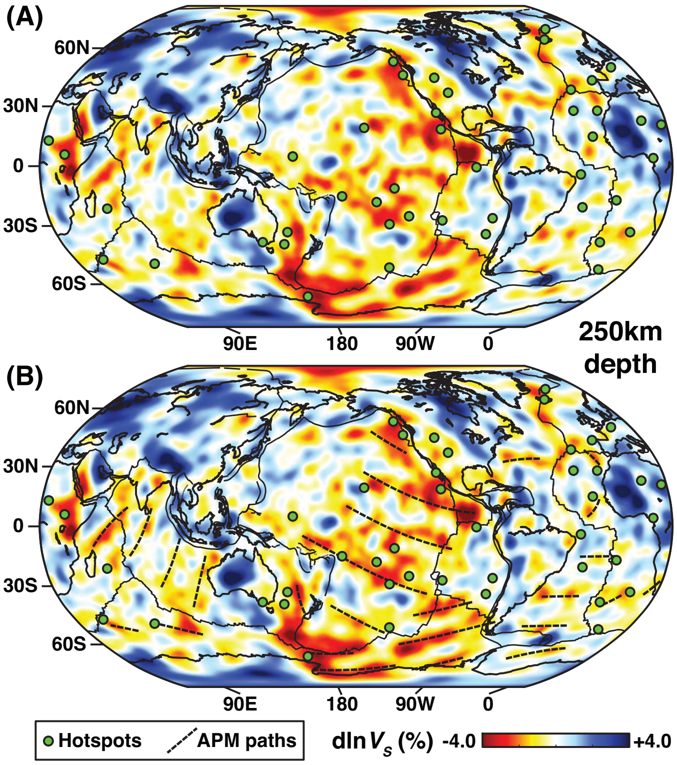


Figure 3.2: Shear-velocity structure and absolute plate motion. (A) Map view of relative V_S variations (percent) from the mean at 250 km. (B) Identical to (A) but with plate-motion streamlines (broken lines) from the APM model of Kreemer (2009). Green circles denote hotspot locations from Steinberger (2000).

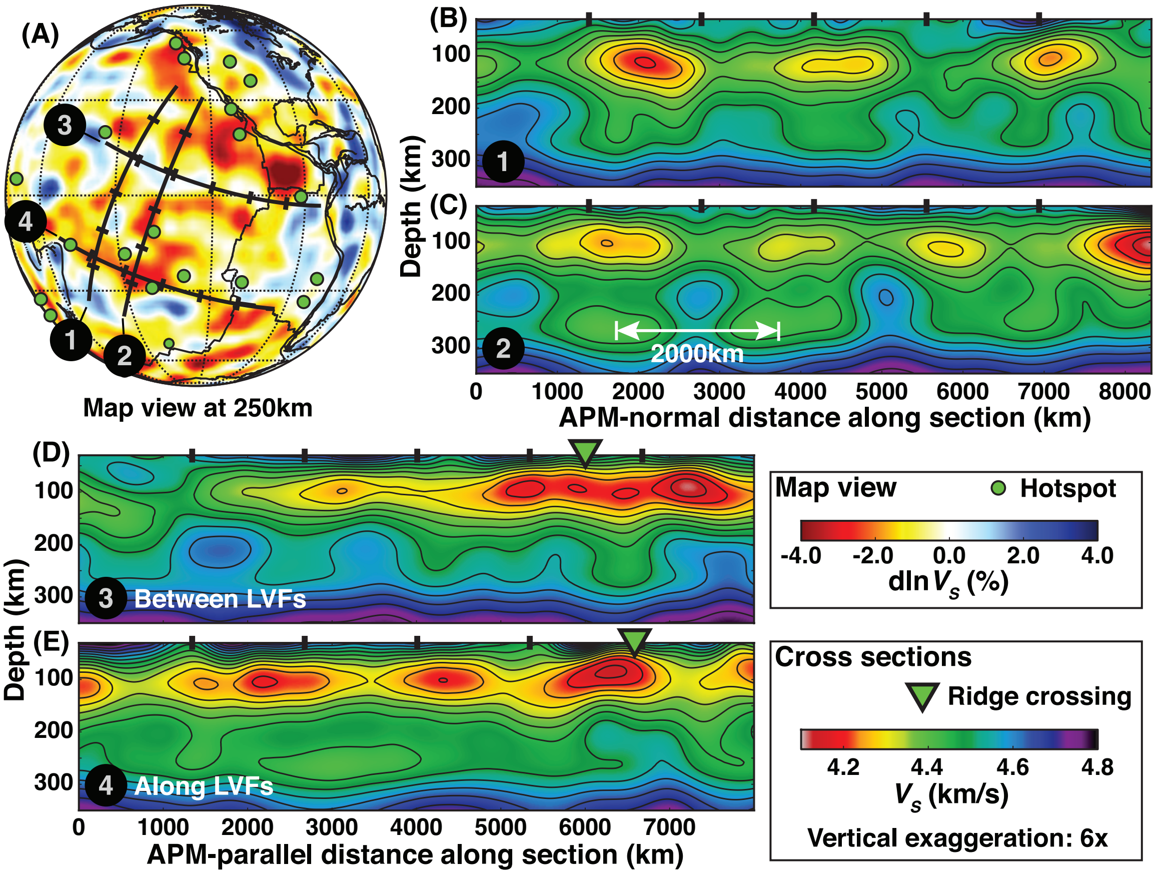


Figure 3.3: Low-velocity fingers in cross-section. (A) Map view of relative V_S variations at 250 km depth, with location of depth profiles shown in (B) to (E). (B and C) Absolute variations in V_S along two profiles perpendicular to the APM (Kreemer, 2009). (D and E) Same as (B) and (C) for profiles parallel to the APM: (D) between two LVFs and (E) along an LVF. Green circles in (A) denote hotspots of Steinberger (2000).

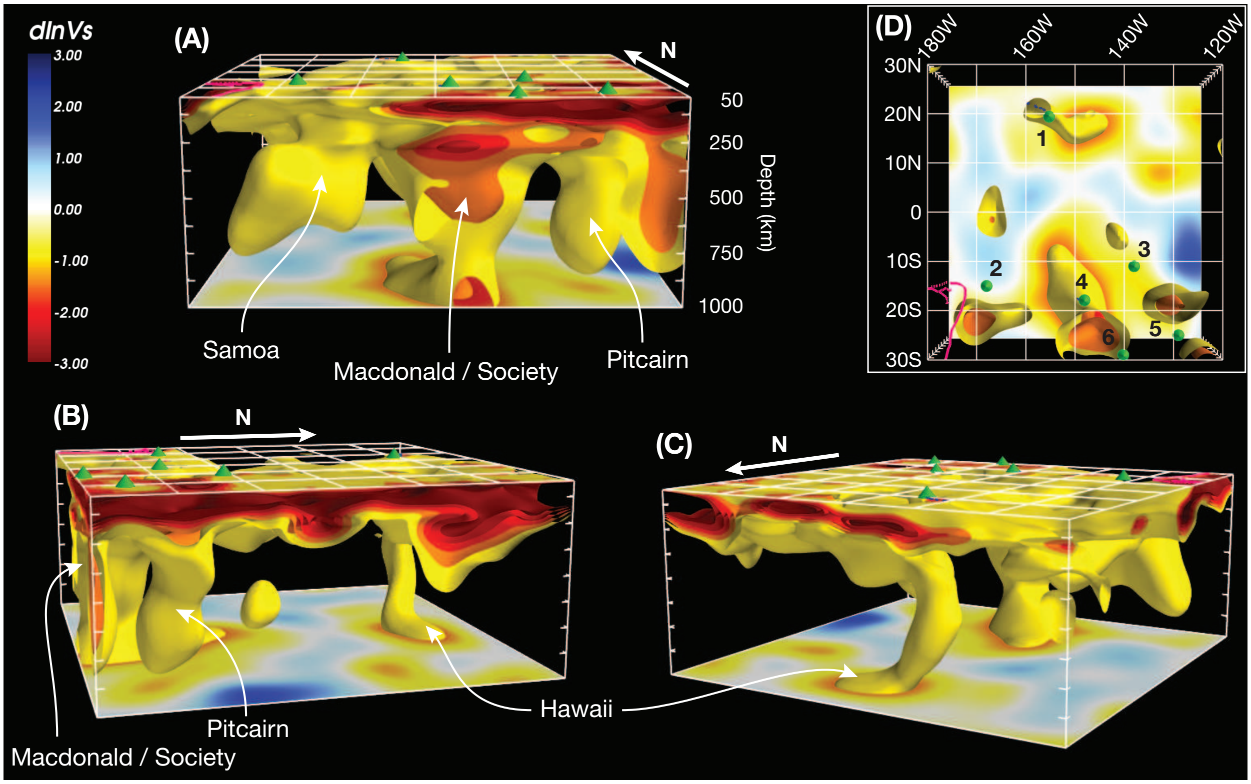


Figure 3.4: Three-dimensional rendering of SEMum2 shear-velocity structure. Relative V_S variations in the central Pacific region (from the Superswell north to Hawaii) viewed from (A) the south, (B) the east and (C) the north. Minimum and maximum isosurface levels are -3 and -1%, respectively. In (A), the LVZ becomes thinner to the west, and the LVF disappears at the Tonga-Fiji subduction zone. In (B), the LVFs appear clearly separated from one another in the direction perpendicular to the APM. The absence of pronounced horizontally elongated low velocities below 200 km depth between fingers is visible in (C). Below 300 to 400 km, the low velocities are organized into predominantly vertical plume-like features. In particular, the Hawaiian "plume" appears east of Hawaii at the bottom of our model (1000 km depth) then turns to the northwest, before being deflected eastward again just below the LVF [(B) and (C)]. (D) View from the top shows the geographic location of the box and major hotspots in relation to the low-velocity conduits, rendered here at 500 km depth: 1, Hawaii; 2, Samoa; 3, Marquesas; 4, Tahiti 5, Pitcairn; and 6, Macdonald. The magenta outline indicates the location of the Tonga-Fiji subduction zone. Hotspot locations are those of Steinberger (2000).

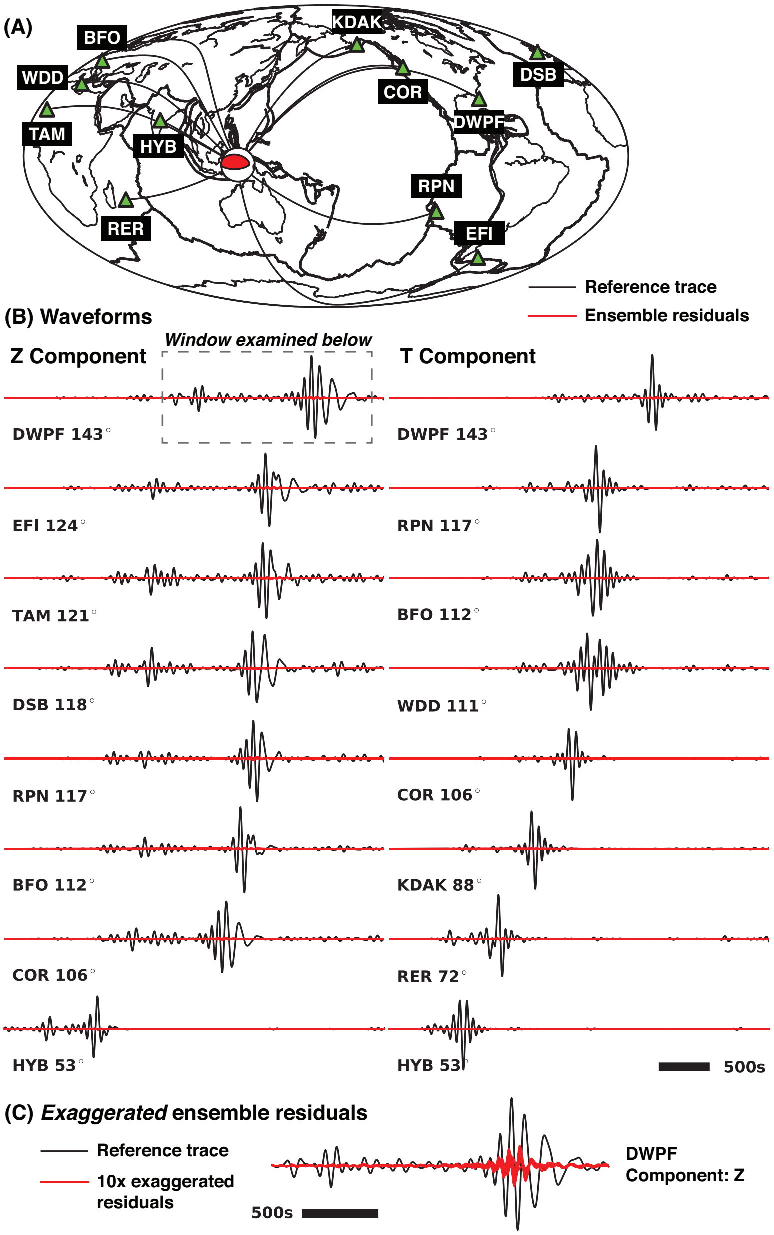
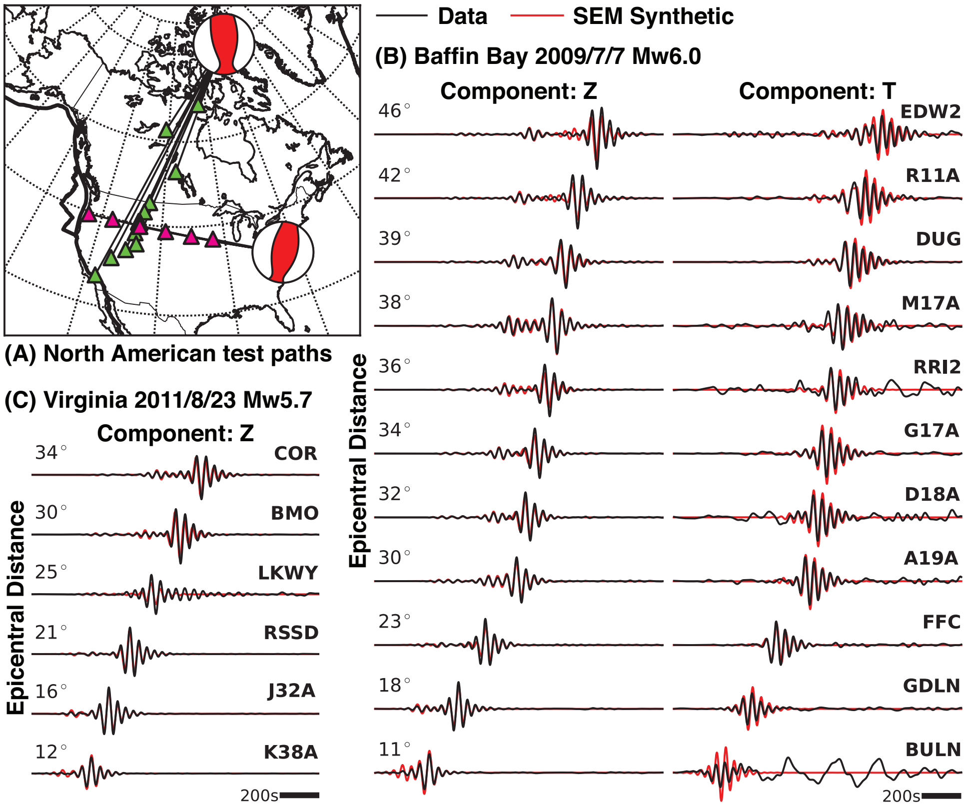


Figure 3.5: Variation in long-period (60-400s) SEM synthetics due to the introduction of realistic-amplitude (Shapiro and Ritzwoller, 2002) artificial errors into the dispersion maps that form the basis of our crustal-modeling scheme (Section 3.4.2.1). For each path in panel (A) selected to sample oceanic, continental, and mixed settings panel (B) shows SEM synthetics calculated in a reference crustal layer (black) for Z (left) and T (right) components, as well as residuals between the reference trace and those calculated in the ensemble of crustal layers based on data with added errors (red). (Continued on the following page.)

Figure 3.5: (Continued) It is immediately apparent that the residuals are very small. Thus, we also show exaggerated (10x) residuals in panel (C), corresponding to the longest path in (B) traversing both continental and oceanic settings again illustrating that the apparent variation is not significant. With these results in mind, it is clear that uncertainties in the dispersion data on the order of those reported by Shapiro and Ritzwoller (2002) cannot significantly affect our waveform modeling through contamination of the crustal layer and in turn bias our mantle model.


Data
SEM Synthetic
(A) North American test paths
(B) Baffin Bay 2009/7/7 Mw6.0
Component: Z
Component: T
Epicentral Distance
46°
EDW2
42°
R11A
39°
DUG
38°
M17A
36°
RRI2
34°
G17A
32°
D18A
30°
A19A
23°
FFC
18°
GDLN
11°
BULN
200s
(C) Virginia 2011/8/23 Mw5.7
Component: Z
Epicentral Distance
34°
COR
30°
BMO
25°
LKWY
21°
RSSD
16°
J32A
12°
K38A
200s

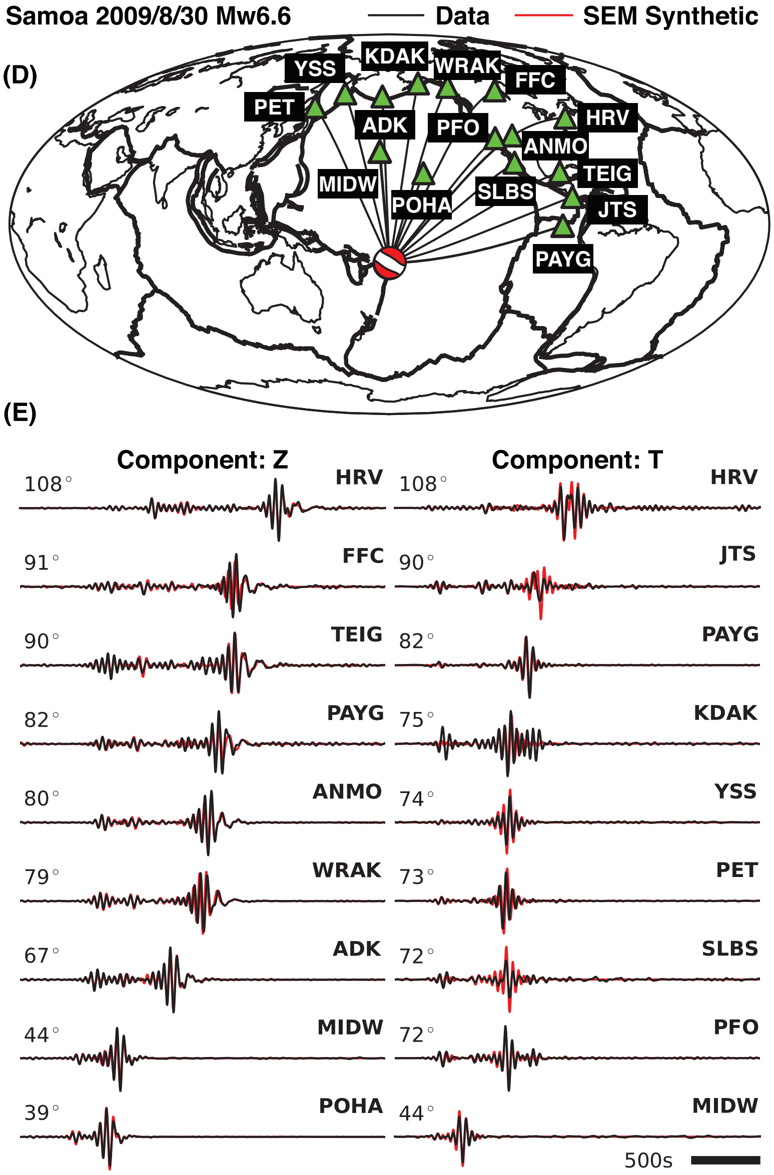


Figure 3.6: Examples of waveform fits for events not included in our dataset and at periods shorter than used in our modeling (40s minimum period instead of 60s). Continental paths are featured in panels (A)-(C), while oceanic paths are shown in (D) and (E). In general, we observe very close fits between data and SEM synthetics in both settings, for both fundamental-mode and overtone wavetrains on vertical and transverse components (with the exception of station BULN, which exhibits high noise levels on the transverse component in panel (B)). These results supply an important independent validation of the SEMum2 model.

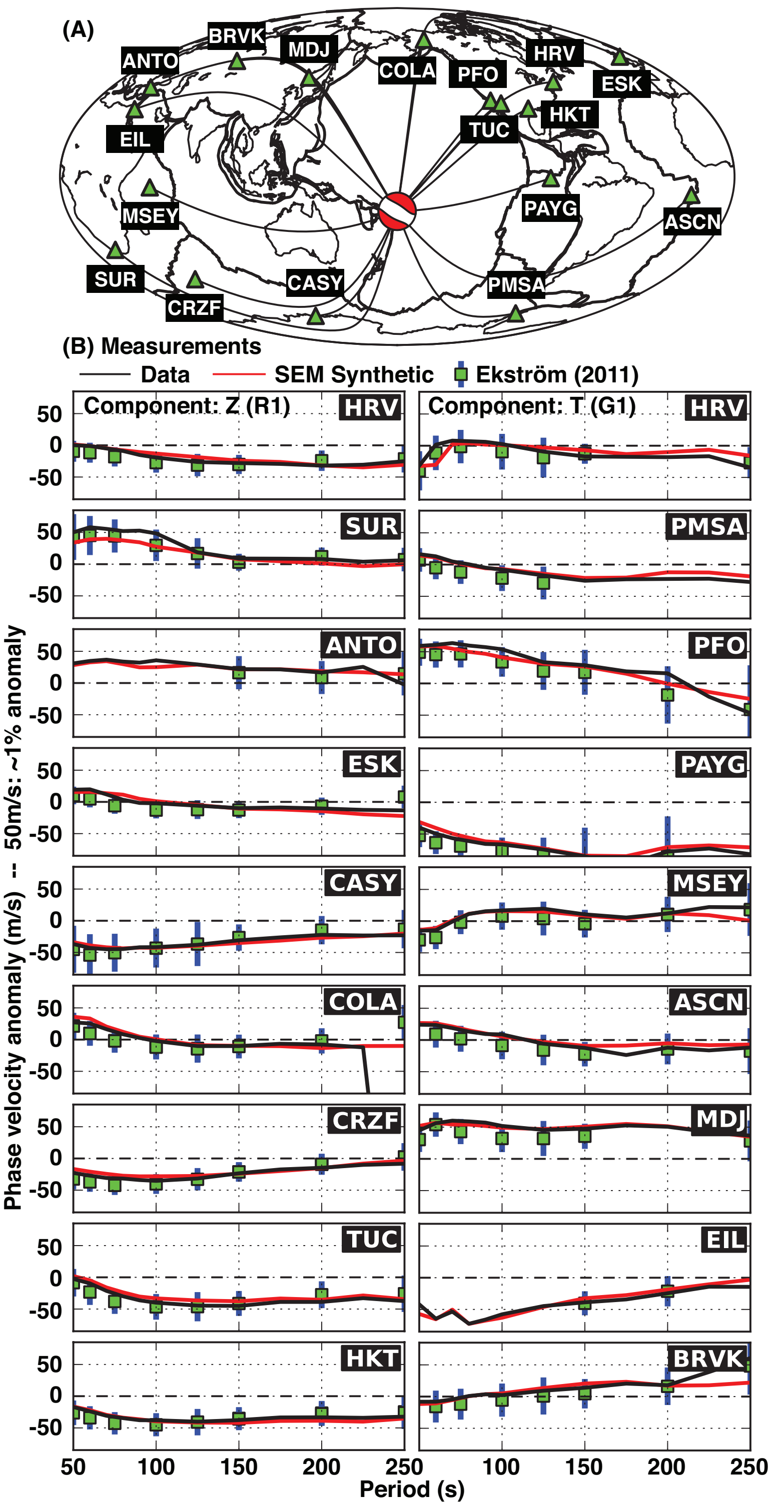


Figure 3.7: Surface-wave dispersion measurements from waveform data and SEMum2 SEM synthetics. (Continued on the following page.)

Figure 3.7: (Continued) Comparisons between fundamental-mode phase velocities measured from waveform data and SEM synthetics for SEMum2 using the multitaper technique of Laske and Masters (1997) again for an event not included in our inversion (Samoa 2009/8/30 Mw6.6), along a range of oceanic, continental, and mixed paths. Absolute phase-velocity anomalies are displayed with respect to the global 1D-mean reference model associated with SEMum2, for which 50m/s corresponds to an anomaly of 1-1.3% in the period range considered. Overall, we find excellent fits between measurements performed on data and SEM synthetics consistent with the quality of waveform fits observed for SEMum2. Further, we also show phase velocities for the same event and paths inferred from the single-station phase-anomaly dataset of Ekström (2011) (used to constrain the GDM52 dispersion model discussed therein). Where available, phase-velocity anomalies from Ekström (2011) are in general compatible with our own measurements to within their published uncertainties. Panel (A): paths for which phase velocity measurements are available. Panel (B): phase-velocity measurements performed on data (black) and SEM synthetics (red) using the technique of Laske and Masters (1997), as well as those inferred from Ekström (2011) with associated uncertainties (green squares, blue 2σ error-bars).

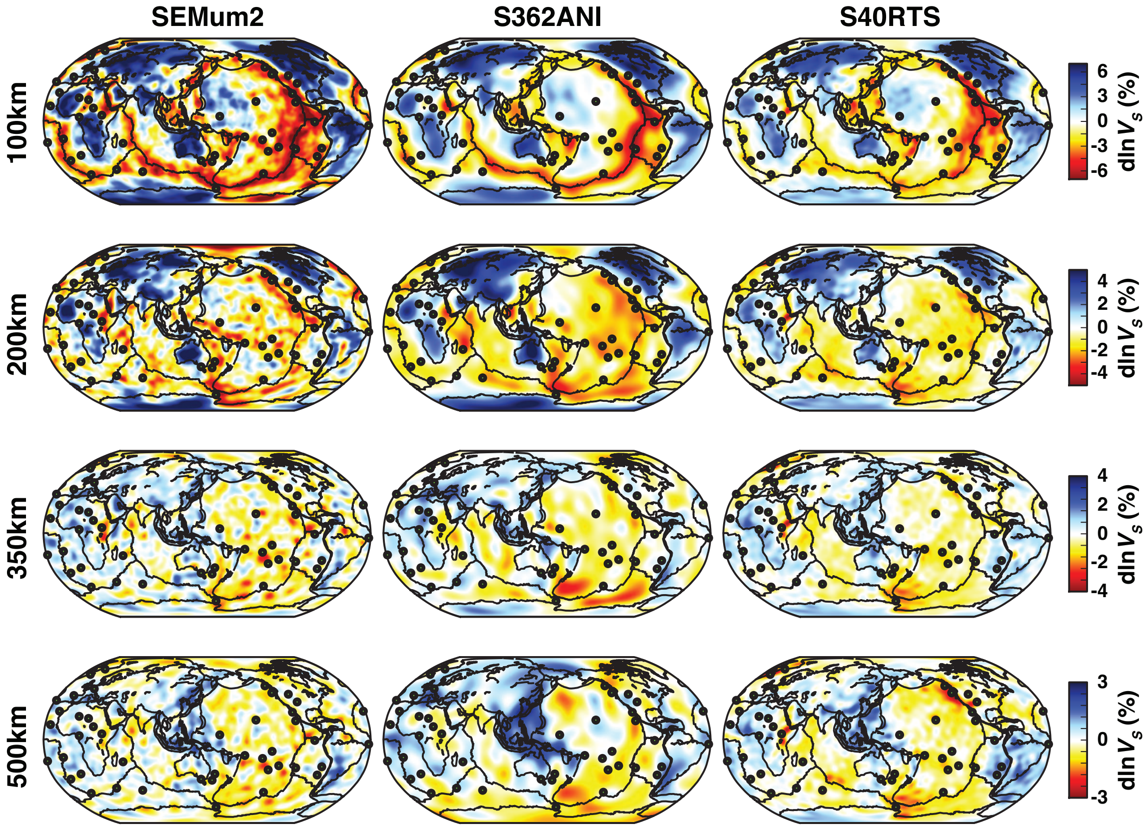


Figure 3.8: Global isotropic V_S structure in the SEMum2 model, as well as two other recent global models S362ANI (Kustowski et al., 2008a) and S40RTS (Ritsema et al., 2011) at a range of upper-mantle and transition zone depths. Model structure is plotted as variations (%) with respect to the 1D reference associated with each model. Circles denote hotspots of Steinberger (2000). All three models agree well at long wavelengths, while SEMum2 in general exhibits stronger, more concentrated anomalies.

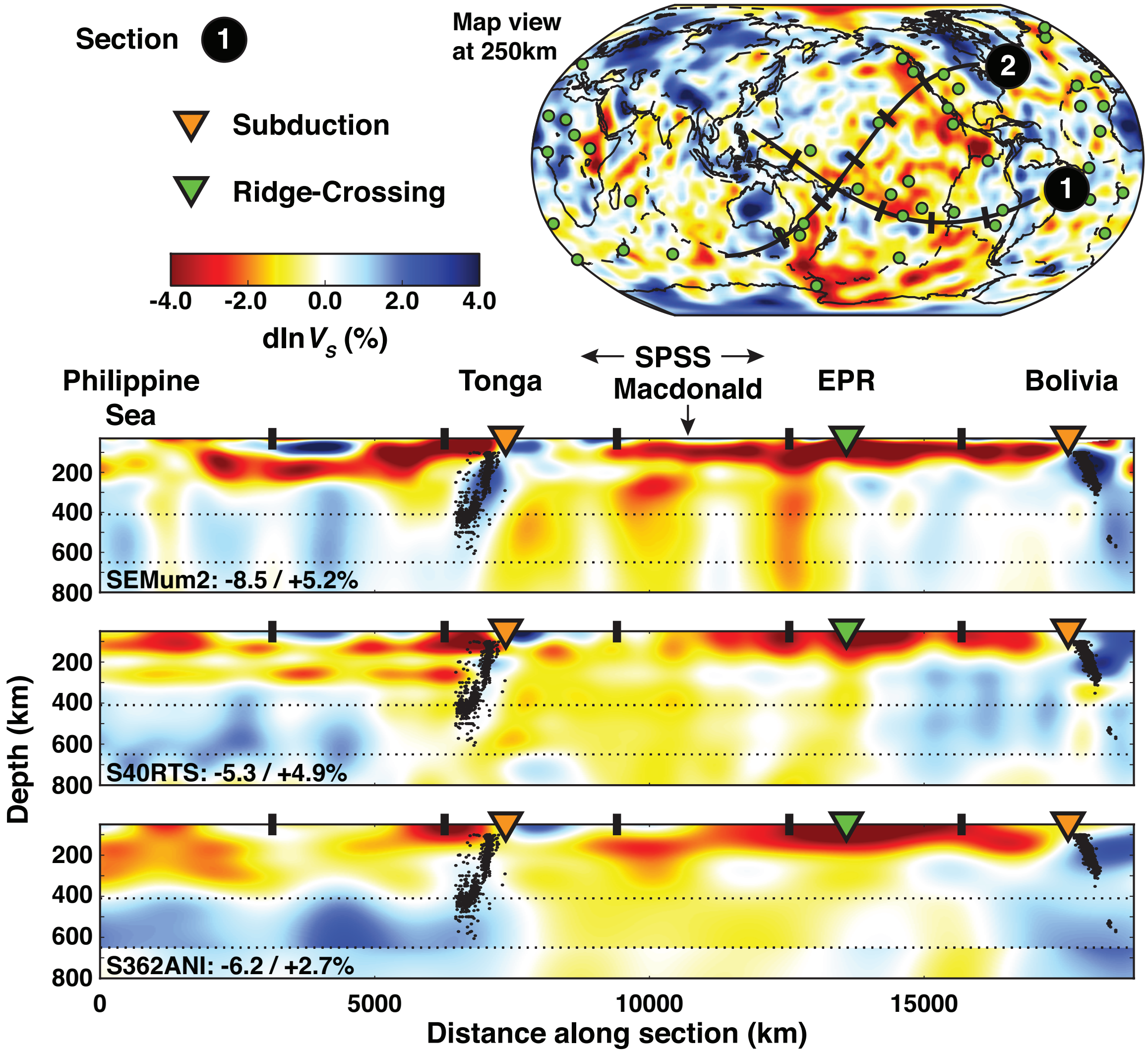


Figure 3.9: Cross sections showing relative variations in V_S structure across the Pacific, corresponding to path (1) (WNW-to-ESE) on the accompanying map. Upper section: SEMum2; middle section: S40RTS (Ritsema et al., 2011); bottom section: S362ANI (Kustowski et al., 2008a). Inverted triangles in green denote ridges crossing the line of section, while those in orange correspond to subduction zones (earthquake hypocenters also shown, highlighting subducted slabs). Maximum positive and negative V_S perturbation for each model (relative to their own respective 1D reference) is also indicated (inset). Large-scale patterns of V_S structure are clearly compatible between the three models, though SEMum2 recovers stronger lateral heterogeneity particularly in the LVZ as well as a more-continuous signature of high velocities associated with subducted slabs and low-velocity “conduits” beneath the South-Pacific Superswell (McNutt and Fischer, 1987). Abbreviations: SPSS Pacific Superswell. Green circles in map view correspond to hotspot locations of Steinberger (2000).

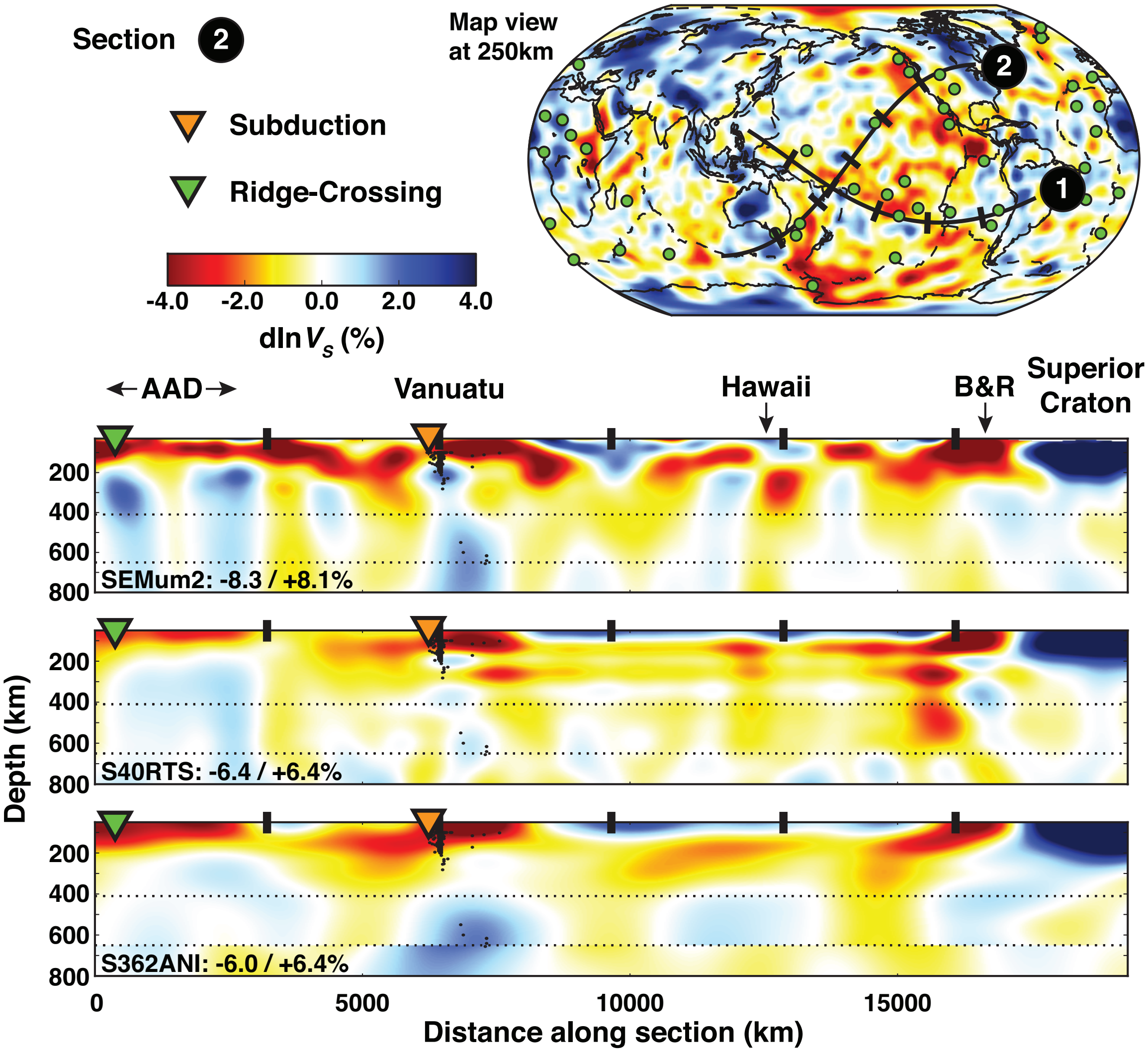


Figure 3.10: Similar to Fig. 3.9, but now showing relative variations in V_S corresponding to path (2) (SW-to-NE). Again, large-scale patterns of structure are consistent across models, while SEMum2 exhibits stronger heterogeneity. Abbreviations: AAD Australian-Antarctic Discordance (Christie et al., 1998); B&R Basin and Range.

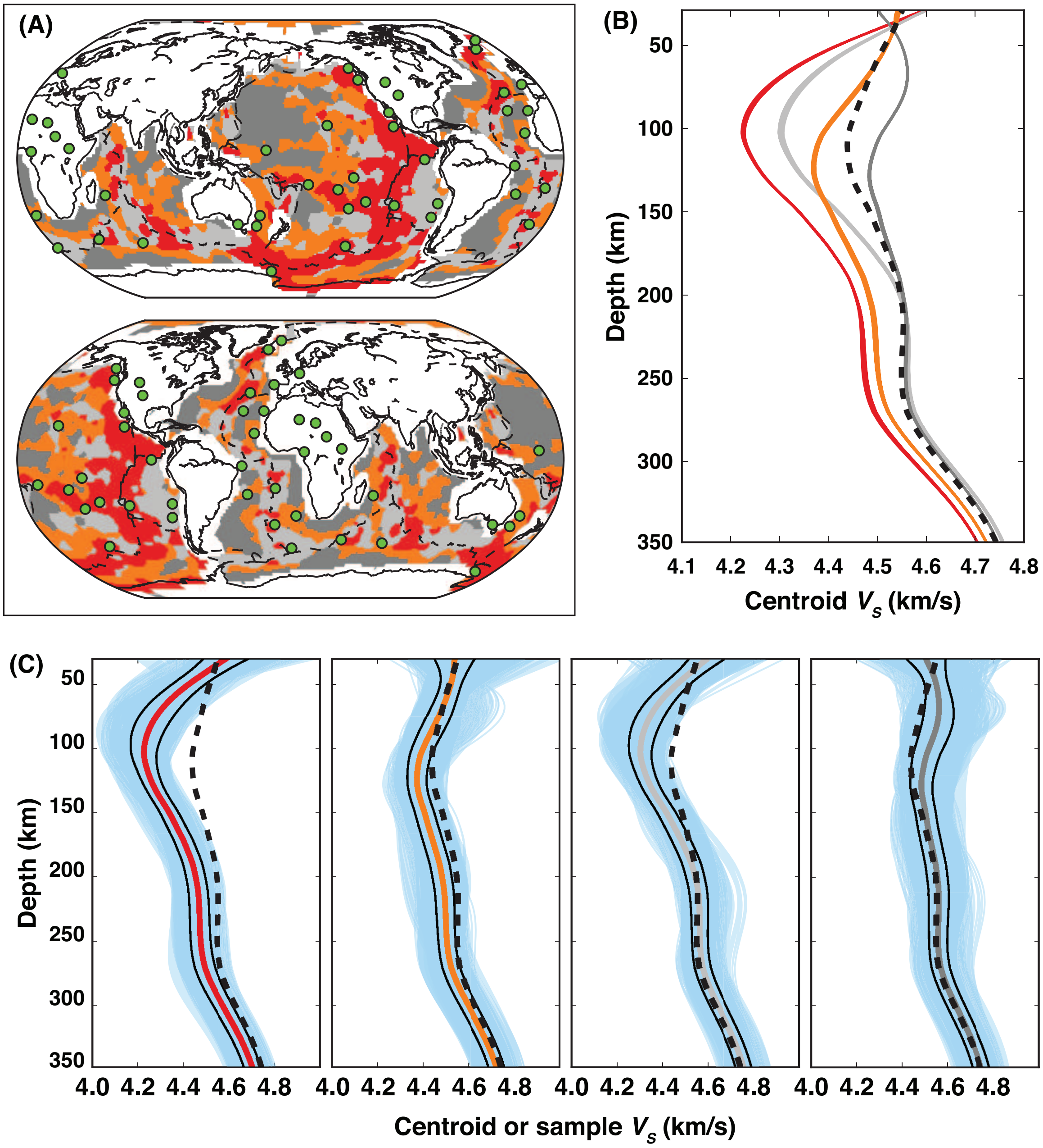


Figure 3.11: Global cluster analysis of SEMum2 oceanic V_S structure. (Continued on the following page.)

Figure 3.11: (Continued) Detailed cluster analysis of SEMum2 V_S structure in the oceanic upper mantle (30-350 km depth) using the k-means algorithm (MacQueen, 1967) (4 clusters). Panel (A): lateral extent of the resulting 4 classes of oceanic V_S structure in both Pacific and Atlantic Ocean-centered views. Green circles denote hotspots of Steinberger (2000). Panel (B): ensembles of mean 1D V_S models for the 4 classes shown in (A) using an identical color scheme, along with the global 1D-mean for comparison (black, dashed). Ensembles of means reflect model uncertainties obtained in the bootstrap resampling analysis (Section 3.4.5.2). For the mean profile associated with each class of structure, we again note the presence of a strong V_S gradient (150-200 km) above almost constant V_S (200-350 km), as found in early models of the oceanic upper-mantle (Montagner and Jobert, 1981). Panel (C): populations of 1D V_S models sampled from SEMum2 that fall into each of the 4 clusters (light blue, background), overlain by the corresponding cluster-mean profiles and the global mean (black, dashed) black lines correspond to $\pm 1\sigma$ variation around the cluster mean within each population.

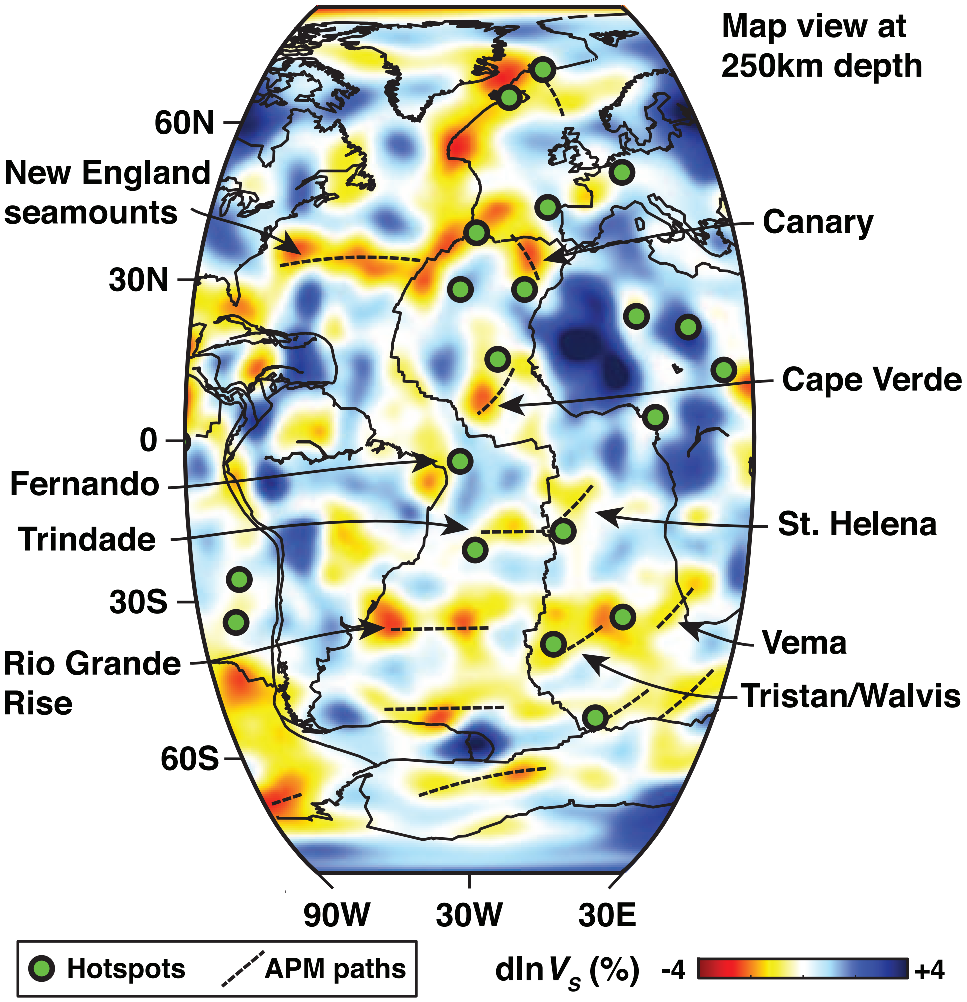


Figure 3.12: Atlantic Ocean centered view of SEMum2 V_S variations (w.r.t. the global mean) at 250 km depth, demonstrating preferential alignment between Atlantic LVFs and the APM of Kreemer (2009) (broken lines). LVF orientations often correlate with Atlantic hotspot tracks and seamount chains, as labeled, including those referred to in the main text. Further, we note impressive recovery of the high-velocity signature of the South-Sandwich Islands subduction zone at ∼60°S. Green circles denote hotspots of Steinberger (2000).

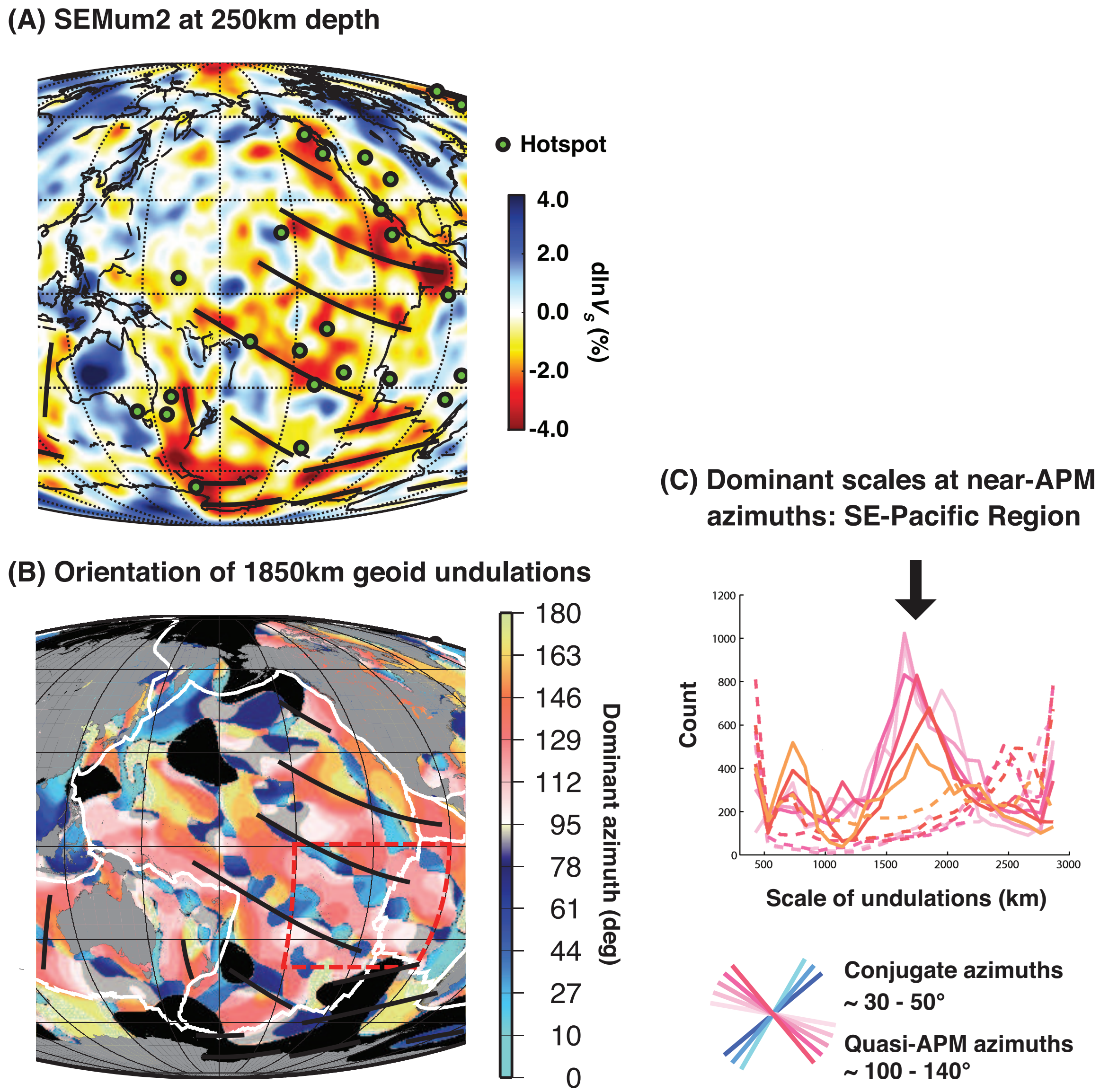


Figure 3.13: Relationship between SEMum2 Pacific LVFs and the geoid analysis of Hayn et al. (2012). (Continued on the following page.)

Figure 3.13: (Continued) Comparison between SEMum2 Pacific LVFs and the directional wavelet analysis of the EGM2008 geoid by Hayn et al. (2012). Panel (A): SEMum2 V_S structure at 250 km depth, focused on the Pacific, with APM paths (Kreemer, 2009) overlain as in Fig. 3.2. We note that the approximate 2000 km spacing between the Pacific LVFs is closer to ∼1900 km when limited specifically to the SE Pacific region. Green circles hotspots (Steinberger, 2000). Panel (B): modified from Hayn et al. (2012, reproduced with permission of the authors) azimuth of dominant orientation of geoid undulations at 1850 km wavelength. As noted, pink colors correspond to quasi-APM azimuths in the central/eastern Pacific, while blue colors correspond to the conjugate direction. Intriguingly, the LVFs appear to fall between the APM-parallel lineations in the geoid at the 1850 km scale an observation that we expand upon in Section 3.4.6.3. Panel (C): modified from Hayn et al. (2012, reproduced with permission of the authors) histograms of dominant wavelengths of geoid undulations (solid lines) and predicted bathymetric effect on the geoid (dashed lines) for the SE Pacific region (red dashed box in Panel (B)) at quasi-APM azimuths. A clear peak (black arrow) is seen at a mean wavelength of ∼1900 km that appears independent of the bathymetric signal (Hayn et al., 2012).

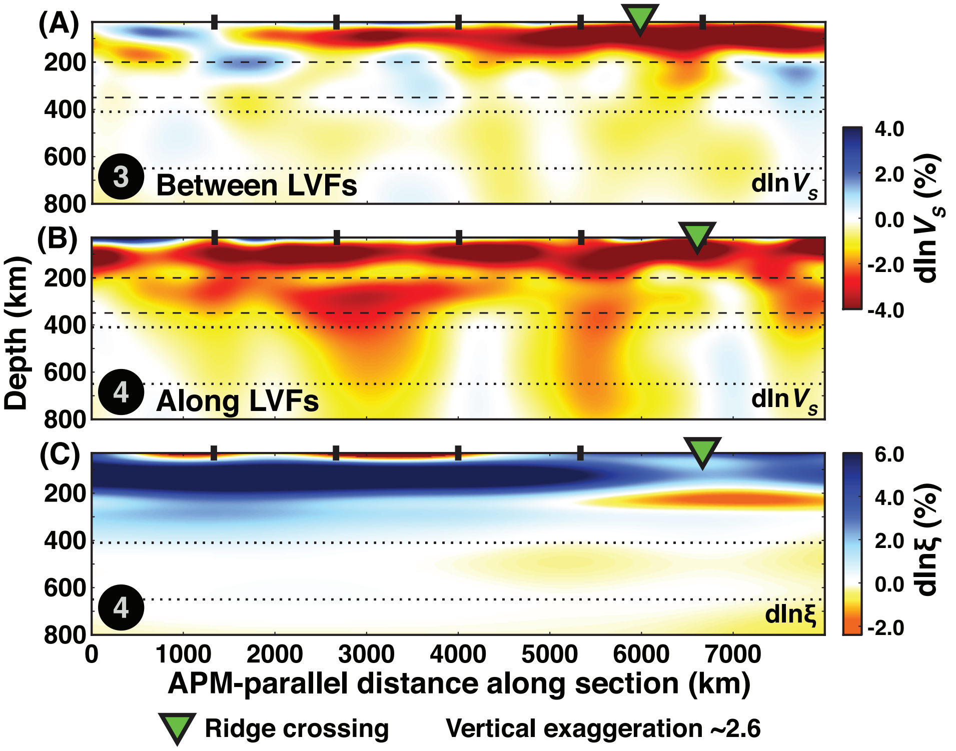


Figure 3.14: Depth cross-sections through SEMum2, along profiles parallel to APM (Kreemer, 2009) that are identical to those in Fig. 3.3 (profiles 3 and 4), but now extending to 800 km depth. Refer to Fig. 3.3A for profile locations. Panels (A) and (B): relative variations in V_S along profiles 3 and 4; Panel (C): variations in the radial anisotropy parameter $\xi = V_{SH}^2/V_{SV}^2$ along profile 4 the same profile as in panel (B). Profile (A) is located between LVFs, while profile (B) is within an LVF.

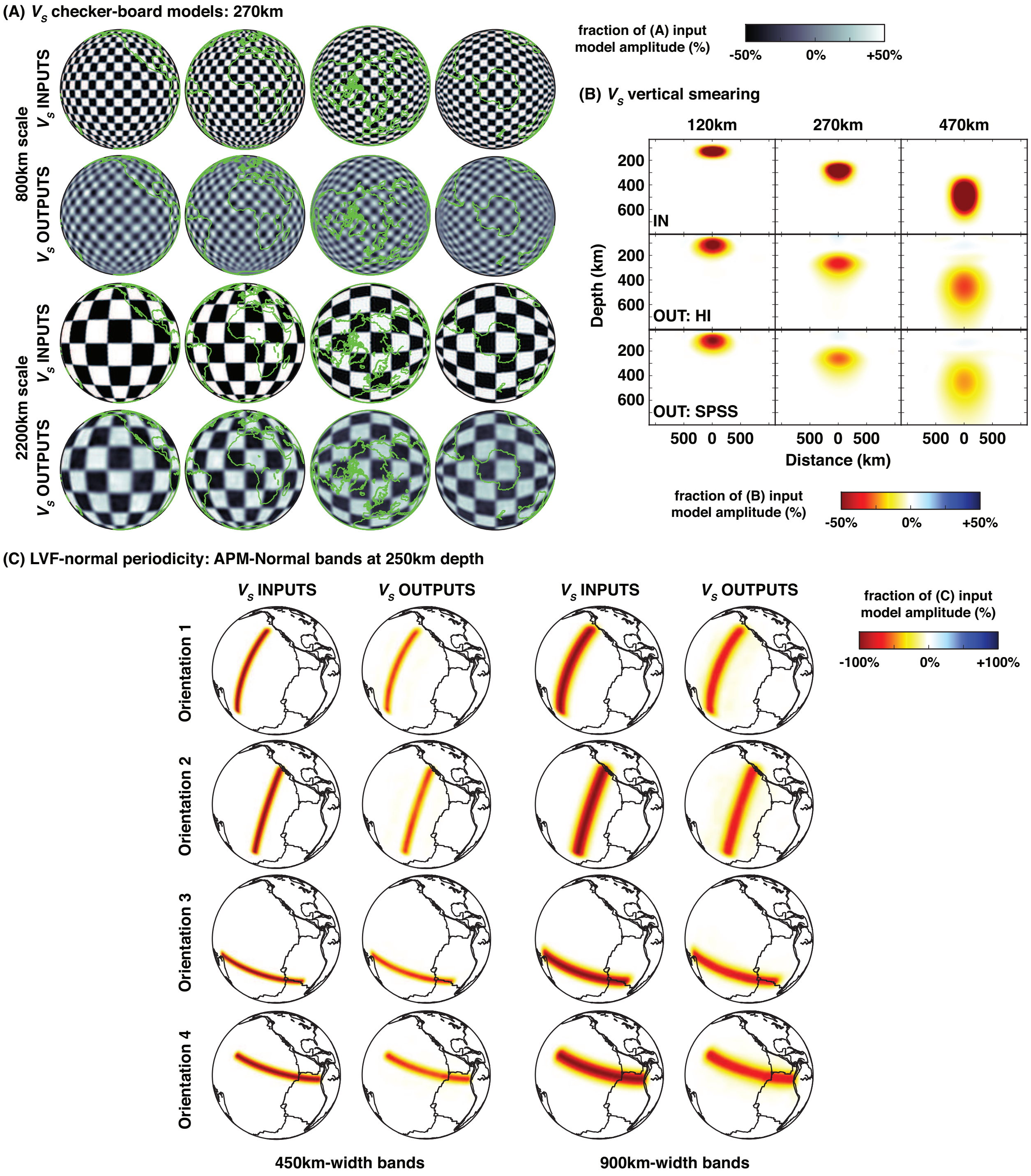


Figure 3.15: Resolution analysis of SEMum2 V_S structure. (Continued on the following page.)

Figure 3.15: (Continued) Linear resolution analysis for the V_S portion of the SEMum2 model, focused on both global and central/eastern Pacific structure. Panel (A): global checkerboard test models at 270 km depth with ∼800 km and ∼2200 km maximum block widths for both equatorial and polar orientations. Both scales are well retrieved, with no indication of ray-like smearing or gaps in sensitivity, even at the poles. Panel (B): isolated point perturbations (∼2° width) at 120, 270, and 470 km depth inputs shown in the top row, structure retrieved with inputs located beneath Hawaii (high-density data coverage) and the Pacific Superswell (lower-density) shown in the middle and bottom row, respectively. Though amplitudes are not honored, we note good recovery of the morphology of input structures at both locations, with little evidence of vertical smearing. Panel (C): band-like input structures (200-350 km depth) with ∼450 km and ∼900 km widths oriented normal (1,2) and parallel (3,4) to APM (Kreemer, 2009). The homogeneous nature of the input bands is honored in the retrieved APM-normal structure, indicative that the ∼2000 km APM-normal periodicity of Pacific LVFs is not the result of poor data coverage or model parameterization. Both scales of APM-parallel input bands are retrieved, indicating that the widths of the Pacific LVFs are well resolved. Color scales chosen to highlight deficiencies in the resolution of our model: panels (A) and (B) saturate at 50% of the input level in order to emphasize low-amplitude smearing artifacts (lateral or vertical, respectively), while panel (C) saturates at 100% of the input level to emphasize gaps in coverage that could artificially give rise to the APM-normal periodicity or lead to poor recovery of the narrow APM-parallel test structures (otherwise difficult to detect if over saturated).

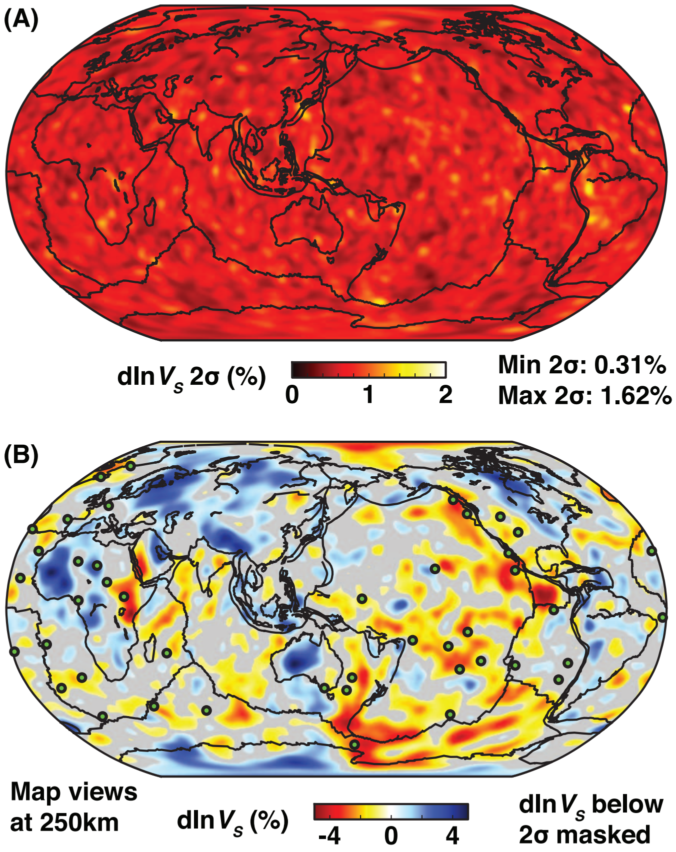


Figure 3.16: Results of the bootstrap resampling analysis (Section 3.4.5.2), here featured at LVF depths (250 km). Panel (A): 2σ uncertainties (% $\delta \ln V_S$) from 20 bootstrap iterations, shown in global map view. Panel (B): relative variations in V_S structure at 250 km, shown in the same view, but with perturbation amplitudes below local 2σ in (A) masked. Given the uncertainty estimates obtained here, it is clear that the LVF structures remain significant at the 2σ level.

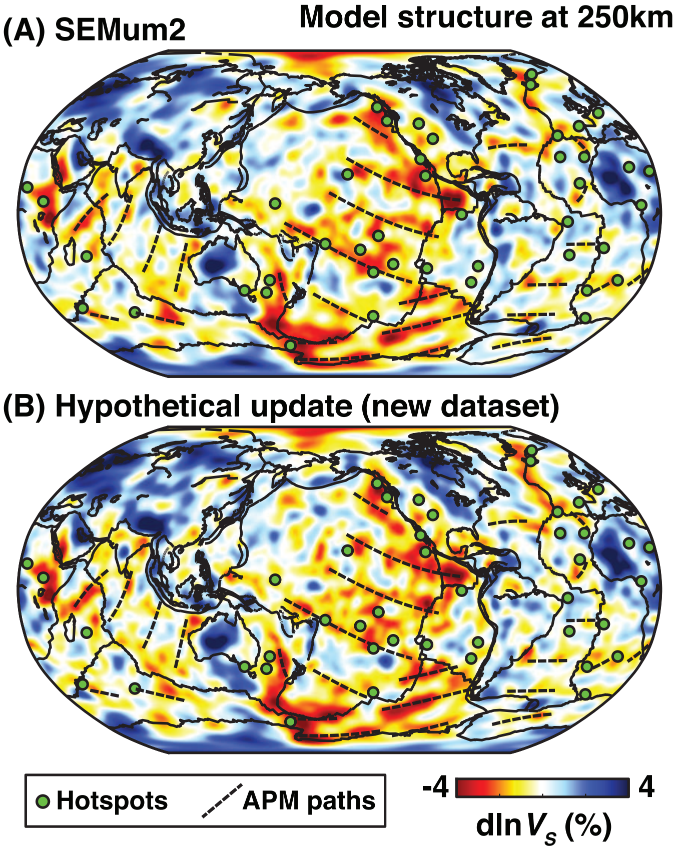


Figure 3.17: Comparison between SEMum2 model structure at LVF depths (250 km) similar to Fig. 3.2 shown in panel (A), and that obtained from a hypothetical model update using a new independent waveform dataset (Section 3.4.5.3), in panel (B). As noted in the text, the model update scheme requires structure to adjust rapidly, if required by the data. We observe that SEMum2 structure at 250 km depth remains stable following the update step, with the LVFs remaining a prominent feature. Further, both models correlate at $R > 0.9$ (up to spherical harmonic degree 48) at all depths throughout the upper mantle. Also shown: hotspot locations (green circles) (Steinberger, 2000); APM (black dashed) (Kreemer, 2009).

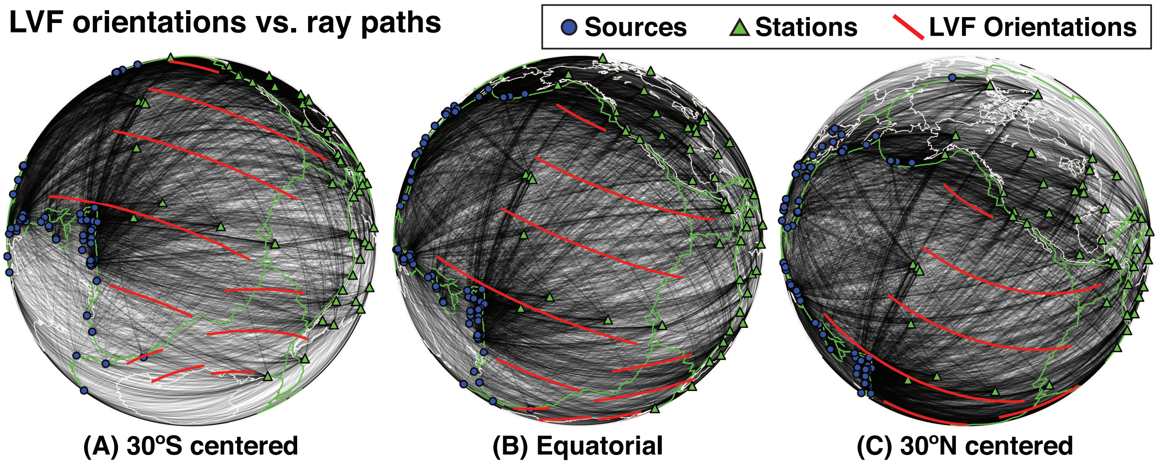


Figure 3.18: Comparisons between Pacific and West-Antarctic LVF orientations (APM paths from Fig. 3.2) and a sparse, but representative subset of rays traversing the region from Western-Pacific and Arctic sources to Western Hemisphere and ocean-island receivers (Section 3.4.5.4. Maps shown with 30S (A), equatorial (B), and 30N centered (C) views; sources and stations shown as blue circles and green triangles, respectively. It is immediately clear that the LVFs cannot correspond to ray paths, confirming in practice that the alignment of the LVFs with APM is not an artifact of the path coverage.

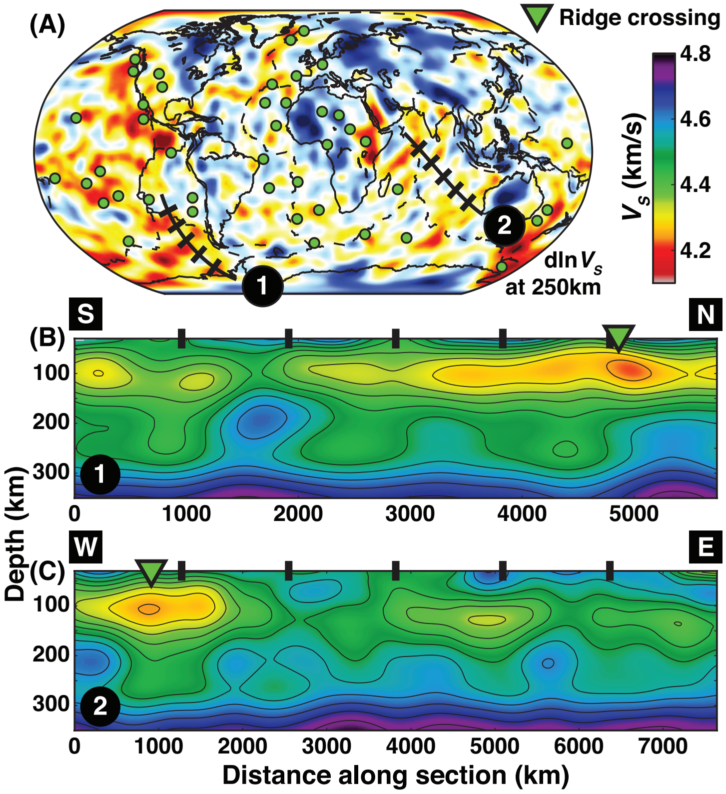


Figure 3.19: Absolute V_S cross-sections perpendicular to APM (Kreemer, 2009) in the upper 350 km of the mantle for paths shown in map view in panel (A): panel (B) beneath the Antarctic Plate; panel (C) beneath the Indian Plate. As in Fig. 3.3, the vertical velocity gradient at the base of the LVZ is considerably stronger in regions between LVFs than where LVFs are present, even beneath the slow-moving Antarctic plate. Overlain contour interval: 50m/s.

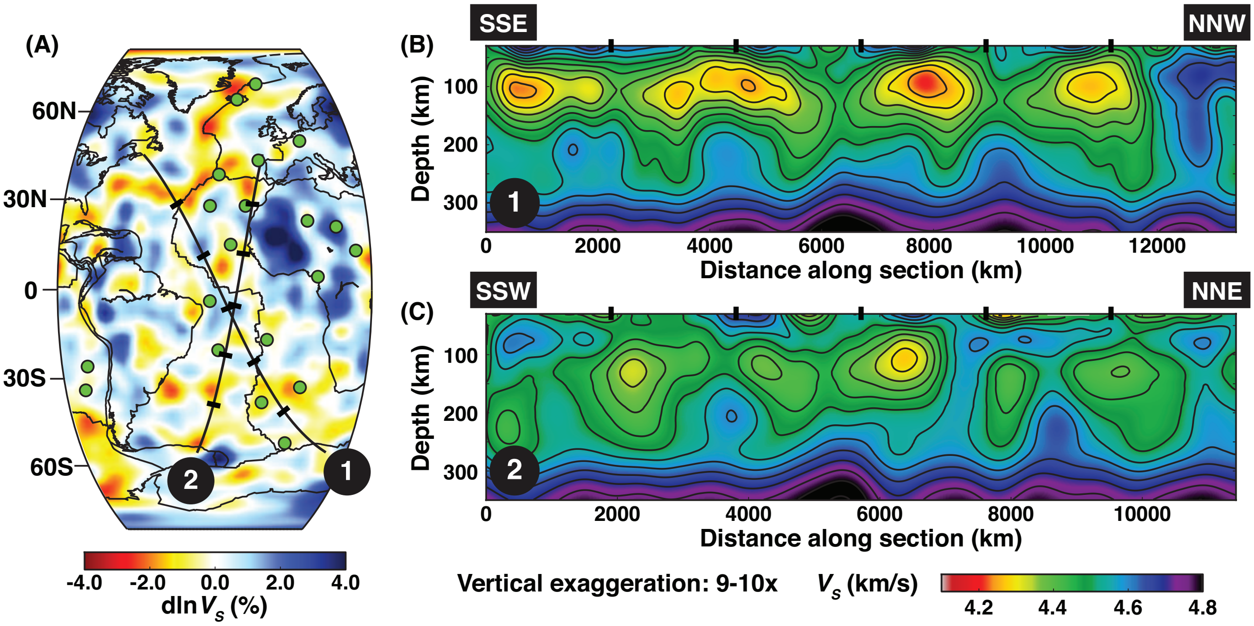


Figure 3.20: Absolute V_S cross-sections in the upper 350 km of the Atlantic upper mantle for paths shown in map view in panel (A) ($\delta \ln\ V_S$ at 250 km depth). The nature of Atlantic APM (e.g. Fig. 3.12 makes it difficult to consistently remain quasi-perpendicular to APM on such long profiles, however those shown above attempt to do so on average (path (1) in panel (B) for its entire length, path (2) in panel (C) until ∼6500 km along section). Especially panel (B), but also in the quasi-APM-normal portion of panel (C), the LVF structures are immediately apparent particularly their ∼2000 km periodicity and depth extent beyond the base of the classical LVZ. Overlain contour interval: 50m/s.

Chapter 4

Radially anisotropic shear-velocity structure of the whole mantle from SEM-based full-waveform tomography

Here, we present the preliminary results of our whole-mantle SEM-based inversion, and detail both the methods used in developing the model and our ongoing efforts to evaluate it. This work is currently in preparation for publication.

4.1 Introduction

Global seismic tomography has made considerable progress over the past 30 years in identifying robust, large-scale features in the seismic velocity structure of the earth's mantle, such as the large low shear-velocity provinces (LLSVPs) in the deep mantle (e.g. Dziewonski et al., 1977; Lekić et al., 2012) and high-velocity anomalies associated with subducted slabs (e.g. van der Hilst et al., 1997) (although their depth distribution is still debated, e.g. Fukao and Obayashi (2013)). Still, many open questions remain regarding the nature of mantle convection which are critical to lending context to these observations, including: the anticipated scales and depth extents of convective phenomena (e.g. Morgan et al., 1995; Korenaga and Jordan, 2004), the roles of the LLSVPs in shaping mantle convection patterns (e.g. Richards and Engebretson, 1992; Davaille et al., 2005) and their implications for compositional heterogeneity (e.g. Davaille, 1999; McNamara and Zhong, 2005), and controls on the surface distribution of hotspots due to large-scale convection patterns (e.g. Husson and Conrad, 2012).

In conjunction with geodynamic studies implementing more realistic mantle convection simulations or laboratory analogue experiments (e.g. incorporating more complex rheological effects or stronger viscosity variations), higher-resolution global-scale tomographic models, delivering more detailed images of seismic structure, provide a key avenue for constraining which convective phenomena seen in simulation or the lab are likely to manifest in the mantle. These advances should be possible by using advanced seismic modeling techniques, such as full-waveform inversion based on accurate numerical wavefield simulations – similar to the approach taken by the SEMum (Lekić and Romanowicz, 2011a) and SEMum2 (French et al., 2013) global models. Indeed, both models exhibit novel features not seen in previous generations of global tomography based on approximate modeling techniques. These structures are imaged particularly clearly in SEMum2 and are discussed extensively in Chapter 3. However, the limited depth range of both models (namely, inverting only for structure above 800 km depth) makes it difficult to fully assess both the implications of the novel features contained therein, and how these structures relate to the dominant convection patterns in the lower mantle.

In order to address these questions, and motivated by the successful application of our inversion technique for upper-mantle and transition-zone imaging, we here present the first whole-mantle tomographic model derived using full-waveform tomography based in the spectral finite element method (SEM: e.g. Komatitsch and Vilotte, 1998). This new model of whole-mantle structure allows us to reexamine a number of the intriguing features in SEMum2 – for example, the detailed morphology and deep origins of the plume-like conduits previously observed in the transition zone (below approximately 400 km depth), including the relationships thereof with the LLSVPs. We are also able to revisit the depth extent of 2000 km periodicity in upper-mantle structure reported in SEMum2, which now appears to extend into the transition zone, as well as interactions with the dominant scales of convection in the lower mantle. Further, we can examine changes in the character of seismic structure near 1000 km depth, often noted as a horizon associated with slab stagnation (in addition to the transition zone).

The present work proceeds in three parts: In Section 4.2, we revisit the “hybrid” waveform tomographic technique used previously in developing the SEMum and SEMum2 models, focusing in particular on the extension of this technique to whole-mantle imaging. Next, in Section 4.3, we discuss the waveform dataset used in our inversion, which now includes body-wave seismograms down to 32 s period. In Section 4.4, we lay out the specific steps involved in our inversion, with an emphasis on progressive incorporation of more data to shorter period, and present our tomographic results along with a discussion motivated by the questions posed above. Finally, in Section 4.5, we discuss our ongoing efforts to evaluate the model, focused primarily on uncertainty estimation and motivated by the style of analysis employed in Chapter 3.

4.2 Methodology

Here, we detail the particular inversion approach used in developing our whole-mantle model, with an emphasis on what differentiates our techniques from those used in previous and ongoing global modeling efforts.

4.2.1 Hybrid full-waveform inversion

We use the *hybrid* full-waveform inversion technique previously employed by our group in developing the SEMum (Lekić and Romanowicz, 2011a) and SEMum2 (French et al., 2013) global models. This technique combines the accuracy and generality of the spectral finite element method (SEM: e.g. Komatitsch and Vilotte, 1998) for wavefield forward modeling, with efficient sensitivity kernel calculation using non-linear asymptotic coupling theory (NACT: Li and Romanowicz, 1995). We combine these two approaches in the context of the generalized least-squares formalism of Tarantola and Valette (1982). Here, we provide a brief overview of the particular features that drove the adoption of this technique, and refer the reader to Chapter 2 for a more extensive discussion of the hybrid approach.

Using the SEM comes with numerous advantages, particularly in global-scale modeling (Komatitsch and Tromp, 2002a,b; Chaljub et al., 2003), for instance: (1) accurate simulation of wave propagation in complex anisotropic and anelastic media, including treatment of phenomena often ignored by approximate modeling techniques (e.g. wavefront healing, Nolet and Dahlen, 2000); (2) excellent numerical dispersion properties (Seriani and Oliveira, 2008), critical for modeling surface waves at long propagation times (i.e. second-orbit fundamental modes); and (3) natural treatment of the free-surface boundary. Further, unlike finite difference methods, the SEM is able to mesh complex geometries characterized by discontinuous material properties, such as the earth's crust (Komatitsch and Vilotte, 1998).

NACT, based on normal-mode perturbation theory, yields finite-frequency waveform sensitivity kernels in the source-receiver great-circle plane, at little cost relative to SEM simulation (Li and Romanowicz, 1995). This method includes coupling along *and* across mode dispersion branches, necessary for accurate modeling of the sensitivity of body waves and overtone surface waves (Li and Romanowicz, 1996; Mégnin and Romanowicz, 1999). Unlike most approximate techniques for sensitivity kernel calculation, NACT is *non-linear*, evolving as the inversion progresses and the earth model iteratively evolves, and includes the effect of multiple forward scattering (Romanowicz et al., 2008). For a more detailed discussion of normal-mode perturbation theory in waveform inversion, see Romanowicz et al. (2008).

One of the primary benefits of the hybrid approach over fully SEM-based techniques using first-order adjoint-state methods (e.g. Tarantola, 1984; Tromp et al., 2005) is convergence rate. Direct access to the waveform Jacobian under NACT allows a rapidly converging Gauss-Newton optimization scheme to be used, in contrast to the gradient-based approaches required when using first-order adjoint. By converging on a model more quickly, thus reducing the number of inversion iterations, we are able to perform significantly fewer SEM simulations (which dominate the cost of the inversion). Though our Gauss-Newton scheme has computational drawbacks of its own, both in distributed construction of the Hessian via NACT (see Chapter 5) and subsequent factorization, we have found that the overall savings outweighs the complexity of addressing these issues.

4.2.2 Crustal model

To further reduce the cost of SEM simulation, we use a smooth anisotropic crustal layer in lieu of a more geologically plausible thinly layered model (e.g. Crust2.0: Bassin et al., 2000). As discussed at length in Chapter 2, this approach enables us to relax constraints on SEM time stability due to the crustal model (namely, in conjunction with the coupled SEM: Capdeville et al., 2003), allowing larger and fewer time-integration steps. Further, as discussed in Sections 2.2.3.3 and 3.4.2.1, this approach leads to no significant loss of accuracy for long-period waveform modeling. Here, we use the same technique for parameterization and calibration of the crustal layer as in SEMum2 (Section 2.2.3), which is characterized by a minimum thickness of 30 km and is based on matching surface-wave dispersion measurements down to 25 s period. We note that this is still well below the minimum period considered here (see discussion of data in Section 4.3), particularly for our surface-wave dataset, which exhibits the greatest sensitivity to crustal structure.

4.2.3 Model parameterization and prior information

Our whole-mantle model adopts the same parameterization as that used in SEMum2 (Section 2.2.1). We invert for 3D variations in Voigt-average isotropic shear-wave velocity V_S and the radially anisotropic parameter $\xi = V_{SH}^2/V_{SV}^2$ with respect to the 1D-mean reference model derived in the SEMum2 inversion (Figure 2.1). From these perturbations, we use the empirical scaling relationships of Montagner and Anderson (1989) in order to estimate variations in the remaining four parameters of a radially anisotropic medium (V_P, $\phi = V_{PV}^2/V_{PH}^2$, η, and ρ), as in previous whole-mantle studies from our group (Panning and Romanowicz, 2006). Such scalings are required, as our dataset lacks sufficient sensitivity to invert for these parameters independently – even in the case of V_P, despite our addition of P-SV waveform data down to 32 s (see Section 4.3). That said, these scalings are only technically valid in the upper-mantle, and it is widely recognized that scaling relations for the deep mantle may differ significantly, such as the V_S to V_P perturbation scaling changing from ~ 2 in the upper and mid-mantle to ~ 3 or more in the lower-most mantle (e.g. Robertson and Woodhouse, 1996; Romanowicz, 2001). We return to this point in Section 4.5.3.1, where we discuss uncertainties in scaling factors and conclude that potential bias in the resulting mantle model is likely not significant (based on inversion experiments with plausible depth-variable scalings).

We express perturbations to V_S and ξ in cubic b-splines radially (e.g. Mégnin and Romanowicz, 2000) and in spherical splines laterally (Wang and Dahlen, 1995). The knots associated with our spherical-spline basis exhibit average lateral spacings of less than 2° for V_S and 8° for ξ, while our radial b-spline basis uses 20 knots with variable spacing between the core-mantle boundary and 30 km depth – clustered more tightly near the top and bottom, where better radial resolution is expected (Figure 4.1). Together, our whole-mantle radially anisotropic model $\mathbf{m}$ is comprised of 2.2×10^5 spline coefficients.

One of the issues we encountered in developing our whole-mantle inversion was the size of the NACT-based Hessian matrix, which has the same (square) dimension as $\mathbf{m}$ and thus requires nearly 100GB at single precision to store just the upper or lower triangular part (as the Hessian is symmetric). Linear algebra operations requiring the Hessian to be stored in full form double this storage requirement. While neither I/O (easily tuned to take advantage of collective buffering on high-performance parallel filesystems) nor factorization of the Hessian (using ScaLAPACK: Blackford et al., 1997), proved problematic at this new problem size, distributed assembly of the Hessian using parallel NACT calculations did – particularly as the full Hessian can no longer fit in memory on a single machine. After careful consideration, we designed a distributed-memory solution based on the partitioned global address space (PGAS) model of parallel computations (Chapter 5), which enables us to tackle problem sizes beyond our previous capabilities.

Finally, we also include *a priori* information in order to keep the inverse problem well-posed – namely, through the data and model covariance operators appearing in the generalized least-squares misfit functional (see Section 2.2.2 of Chapter 2). The data covariance operator is configured to account for data quality and redundancy, as detailed in Li and Romanowicz (1996), while the model covariance operator is used to limit the permissible length-scales of heterogeneity in our model to those that should be well-resolved by the data. As in SEMum2 and detailed in Section 2.2.2, we adopt a scheme for adaptive lateral correlation length based on inferred data sensitivity (measured from the diagonal of the NACT-based Hessian). By analogy with our upper-mantle inversion ($\leq$ 800 km depth), which limited these ranges of correlation lengths to 400-800 km and 1200-2400 km for V_S and ξ, respectively, we selected 400-1200 km and 1200-3600 km for the whole mantle inversion. Based on the sensitivity profile associated with our data, this configuration leads to a similar distribution of correlation lengths as before in the SEMum2 depth-range, while permitting relatively long correlation in the mid-mantle where aggregate sensitivity is poorest, and again shorter-wavelength structure at the base of the mantle where sensitivity improves.

4.3 Waveform dataset

While previous SEM-based global tomographic models from our group (Lekić and Romanowicz, 2011a; French et al., 2013) focused primarily on upper-mantle and transition zone structure, and thus included only fundamental and overtone mode surface-wave waveforms at long periods ($T \geq 60$s), our current whole-mantle modeling must now include body-wave data. To this end, we use a dataset comprised of full, three-component teleseismic waveforms, filtered in multiple passbands, allowing us to incrementally incorporate higher-frequency body-wave data as the inversion progresses:

- **Surface-wave passband:** cutoff at 400 and 60 s (corners at 250 and 80 s)
- **Body-wave passband (Filter I):** cutoff at 300 and 36 s (corners at 180 and 45 s)
- **Body-wave passband (Filter II):** cutoff at 300 and 32 s (corners at 180 and 38 s)

Table 4.1: Final-iteration waveform-window counts N_{wp} and aggregate dataset sizes (number of datapoints N_d, sampled at the corresponding Nyquist rate), organized by receiver component and waveform data type / passband.

	L		**T**		**Z**	
	N_{wp}	N_d	N_{wp}	N_d	N_{wp}	N_d
Body (32 s)	83,374	1,928,538	62,619	959,878	106,497	2,527,759
Fundamental (60 s)	17,688	921,443	26,643	990,836	30,974	1,635,960
Overtone (60 s)	25,932	1,603,959	25,452	883,297	42,297	2,502,012
Mixed (60 s)	6,703	422,053	9,411	469,645	10,253	645,579

We note that this surface-wave passband is identical to that used in constructing SEMum and SEMum2. In addition, we expand our dataset as the inversion progresses by incorporating additional events: starting from the same 203 events used in developing the latter two models, and increasing to 273 events total. The particular set of new events were chosen to be spatially distributed in a complementary manner to the earlier set of 203 (to maximize independent constraint on structure). The complete set of sources and receivers used in our inversion is shown in Figure 4.2.

Waveform data is processed using a windowing approach (Li and Romanowicz, 1996) that groups trains of phases by amplitude and constraint on earth structure (Figure 4.3). Using this technique, we are able to invert long-period waveforms in the time domain – making use of both phase and amplitude information – while preventing large-amplitude phases (e.g. the SS body-wave phase, sensitive to upper and mid-mantle structure) from dominating comparatively low-amplitude ones (e.g. S_{diff}, critical to resolving lower-most mantle structure). Further, these windows may in turn be weighted according to data uncertainty and / or noise estimates, as well as redundancy in sensitivity – see Li and Romanowicz (1996) (Appendix A) for a more detailed discussion of the weighting scheme. Indeed, these weights are the basis for construction of the data covariance operator referred to in Section 4.2.3 above. Window selection is performed based on similarity to SEM synthetic seismograms, computed in the previous iteration of the 3D mantle model – an operation that is performed at *each* iteration of our inversion. By the end of the inversion, our dataset is comprised of over 447,800 waveform windows, corresponding to nearly 15.5M datapoints sampled at their corresponding Nyquist rates (Table 4.1).

4.4 Results

4.4.1 Inversion and fits

The starting model for our whole-mantle inversion is comprised of SEMum2 in the upper mantle and transition zone ($\leq$ 800 km depth) and the model SAW24B16 of Mégnin and Romanowicz (2000) below (lower-mantle structure was previously fixed to SAW24B16 during the SEMum and SEMum2 inversions). Throughout our iterative inversion, we assume seismic source parameters for each event consistent with those reported by the Global CMT project (`http://www.globalcmt.org`) – we return to the question of seismic source *inversion* in Section 4.5.1. The structural inversion is comprised of three phases, allowing us to incrementally enlarge our dataset and incorporate shorter-period waveform data:

- **Phase I** In the first phase, we performed one iteration of inversion for whole-mantle structure using the 60 s surface-wave and 36 s body-wave (Filter I) datasets, picked from the 203 events used in SEMum and SEMum2. We found that the upper-mantle structure in the model changed very little, aside from slightly larger amplitudes following the introduction of the body-wave data. Motivated by this, we chose only to invert for structure at $\geq$ 300 km depth in the remaining iterations of the inversion. However, since upper-mantle amplitudes had changed slightly, we first performed one last recalibration of the crustal model using the two-step inversion scheme described in Chapter 2 (Section 2.2.3.2).

- **Phase II** Next, we introduced 70 new events with moment-magnitudes between 5.8 and 7.3, selected so as to be maximally complementary to the distribution of the 203 earlier events and primarily motivated by enlarging our body-wave dataset. We picked the new-event data and reprocessed the older-event data using the SEM synthetics from the previous iteration. Thereafter, we performed another inversion iteration, again using the 60 s and 36 s filter passbands, but now for structure below 300 km depth only. We chose to include both overtone and fundamental mode waveform data, as their sensitivity is non-negligible in the depth range considered.

- **Phase III** In the final phase, we reprocessed the data from the 273 events using a new shorter-period body-wave passband (Filter II). We then inverted one last time for structure below 300 km depth using the 60 s and now 32 s data passbands.

In Table 4.2, we summarize the final-iteration waveform fits for the different filter passbands and receiver components used in our inversion. To ensure that our inversion was converging, we inspected at each iteration whether more waveform windows were selected in the next data reprocessing round than would have been using synthetics from the previous model. By the final iteration, we found only small gains in the numbers of selected windows, indicative that the inversion had likely converged (for the particular passbands considered). This assessment of convergence may also be supplemented by testing fits to held-out data – namely, waveforms from events not included in the inversion – to ensure we are not over-fitting the dataset at the expense of generality (i.e. fitting noise). For example, this latter approach was used to confirm that our final iteration was indeed warranted (see Section 4.5.2).

Surface-wave variance reduction (VR) values in Table 4.2 are quite similar to those obtained for SEMum2, which is not surprising given that upper-mantle and transition-zone structure has remained largely the same, while body-wave VR is in general higher than that reported for previous Berkeley whole-mantle models (typically $< 50\%$, e.g. Mégnin and

Table 4.2: Final-iteration waveform variance reduction, defined as one minus the squared 2-norm of the waveform residual normalized by that of the data, expressed in percent as $VR = 100\times[1 - \|\mathbf{d} - \mathbf{g}(\mathbf{m})\|_2^2/\|\mathbf{d}\|_2^2]$, and organized by component and data type / passband.

	L	**T**	**Z**
Body (32 s)	54.6 %	56.8 %	51.5 %
Fundamental (60 s)	69.1 %	76.6 %	71.9 %
Overtone (60 s)	78.2 %	68.2 %	79.4 %
Mixed (60 s)	75.4 %	79.6 %	78.9 %

Romanowicz, 2000; Panning and Romanowicz, 2006). The latter comparison is not straightforward, however, as VR for these earlier models was evaluated using approximate waveform modeling techniques (i.e. not SEM). Unexplained variance may be attributable to a number of factors, including: unmodeled structure (for example, below the limit of our parameterization), noise or minor instrument errors (though data is rejected when either is particularly severe), and uncertainty in seismic source parameters (we will return to this latter point in Section 4.5.1).

Overall, four rounds of SEM simulation were required to complete these three phases of inversion – one for the earlier 203-event dataset, and three for the enlarged 273-event dataset – in addition to ancillary simulations needed for validation (e.g. ensuring convergence against held-out data, numerical experiments for tuning the weights assigned to our waveform datasets, etc.). Omitting our NACT-based Hessian-estimation calculations and parallel linear algebra solves for computing model updates under the Gauss-Newton scheme, the present study required approximately 3M CPU hours for SEM simulations alone. These latter simulations were performed on *Hopper*, a Cray XE6 at the National Energy Research Scientific Computing Center (NERSC), while the NACT-based Hessian estimation and Gauss-Newton model update computations were performed on NERSC *Edison*, a Cray XC30.

4.4.2 Whole-mantle shear-velocity structure: Preliminary results

4.4.2.1 Global isotropic structure

In Figure 4.4, we show global maps of isotropic V_S variation at a range of depths in our model. For comparison, we also show V_S structure from two other recent whole-mantle studies: S40RTS of Ritsema et al. (2011) and S362ANI of Kustowski et al. (2008a). From these maps, it is immediately clear that all three models tend to agree at long wavelengths – consistent with our earlier observations for the SEMum2 model at upper-mantle and transition-zone depths (Chapter 2). One of the clear differences across models is again amplitudes: peak-to-peak variation in V_S is larger in our model, particularly within the top and bottom boundary layers of the mantle where heterogeneity is expected to be strong, while comparable variation is seen in the mid-mantle. This effect is especially pronounced at shallow depth, with our model showing nearly 19% peak-to-peak variation at 70 km. Indeed, similar relative amplitudes were seen for SEMum2 and our upper-mantle structure has not changed significantly (as noted in Section 4.4.1). As mentioned in previous chapters, a likely candidate for this difference in retrieved amplitudes is the use of SEM-based forward modeling, which easily accounts for wave-propagation phenomena that can otherwise lead to systematic underestimation of amplitudes if neglected, such as wavefront healing (e.g. Nolet and Dahlen, 2000).

In contrast to our earlier observation regarding long-wavelength agreement across models, there are obviously clear differences at shorter length scales. For example, near the base of the mantle (2800 km), the large low shear-velocity provinces (LLSVPs) in our model appear to be composed of distinct low-velocity anomalies distributed within a long-wavelength degree-2 background component, as opposed to the relative lateral continuity of the LLSVPs seen in the other two models. Indeed, as seen in Figure 4.5, comparing spherical-harmonic power spectral density across the models shown in Figure 4.4, the well-known degree-2 signal, found both at the base of the mantle and in the transition zone, is still quite prominent in our model (though in both depth ranges, our model also has significant power out to higher degrees). At the same time, there are smaller-scale features at this depth which consistently appear in all three models, such as the low-velocity anomaly approximately located beneath Perm, Russia (Lekić et al., 2012).

As the upper mantle remains very similar to SEMum2, we instead choose to focus on depths in and below the transition zone for the remainder of this discussion, as this is where our model has evolved the most. We examine this point in Figure 4.6, where we compare V_S structure at 1000 km depth in our final model with that in our initial model – the latter based on model SAW24B16 (Mégnin and Romanowicz, 2000) at this depth. We note that V_S structure is considerably sharper in our new whole-mantle model relative to SAW24B16, including narrower *distinct* low-velocity features beneath the South-Pacific Superswell (McNutt and Fischer, 1987) in the vicinity of major hotspots, as well as more concentrated fast anomalies associated with subducted slabs (e.g. Tonga-Kermadec, South / Central America). Narrow low-velocity anomalies have also now appeared beneath Tanzania and Iceland, which were not present in our starting model. We will reexamine a number of these features in cross section shortly (Section 4.4.2.3). There are also a number of large, isolated high-velocity anomalies previously appearing at this depth in SAW24B16 that are now missing – for example, in the Southeast Pacific. This contrast between the starting and final models, achieved with only three inversion iterations, also serves as an illustration of the power of our hybrid inversion approach, which allows us to take large steps in the model space and rapidly converge on our final result.

4.4.2.2 Global anisotropic structure

Next, in Figure 4.7, we present maps of global ξ structure (expressed in perturbation away from isotropy, i.e. $\xi = 1$) in our model, SAW642ANb of Panning et al. (2010) (a revised version of SAW642AN from Panning and Romanowicz (2006), using a new treatment of the crust), and S362WMANI of Kustowski et al. (2008a) (the whole-mantle radially anisotropic variant of S362ANI). Examining our model first, we see the expected pattern of strong radial anisotropy at the top and bottom boundary layers of the mantle, where lateral shear is expected to be strongest, and weak anisotropy at mid-mantle depths. In the upper 250 km, we see strong $\xi > 1$ beneath the ocean basins, consistent with strong horizontal shear (i.e. flow), accompanied by $\xi \leq 1$ beneath the mid-ocean ridge system, typically interpreted as a transition to dominantly vertical flow.

At the base of the mantle, we see structure decomposed into contiguous domains of strong $\xi > 1$ and $\xi < 1$. Intriguingly, these patterns match the long wavelength background (i.e. degree-2) pattern seen in V_S – with $\xi > 1$ in high-velocity regions often hypothesized to be dominated by horizontal shear due to subducted slabs impinging upon the core-mantle boundary (CMB), and $\xi < 1$ associated with the low-velocity LLSVPs, potentially areas characterized by broad upwelling. Of course, the exact mechanism for the development of these anisotropic textures is not immediately clear – particularly the roles of shape vs. lattice preferred orientation (e.g. McNamara et al., 2002), as well as compositional controls on anisotropy development, such as the presence of post-perovskite in the D" region (e.g. Murakami et al., 2004). In addition, we cannot rule out that the observed anisotropic pattern may be affected by our data coverage: namely, differential lateral distribution in sensitivity of the SH and P-SV portions of our body-wave dataset.

In comparison with the ξ structure of S362WMANI, we find that our model is in general smoother and weaker than the latter – with the exception of 125 km depth, where peak-to-peak variation is slightly stronger in our model. Indeed, mid-mantle radial anisotropy (for example, at 1900 km) is considerably stronger in S362WMANI than in our model, and corresponds to a depth range long expected to be nearly isotropic (e.g. Dziewonski and Anderson, 1981). Conversely, we find that the amplitudes of ξ structure in SAW642ANb exhibit similar depth variation to our model, with the exception of 125 km where peak-to-peak heterogeneity in the former is stronger. Further, we find that our ξ model is again smoother than SAW642ANb in general. While lateral variation in ξ agrees relatively well between all three models at the longest wavelengths, we find that local relative variation is in some locations anti-correlated across models. For example, there is the local ξ maximum seen in S362WMANI at 250 km beneath Hawaii, which corresponds to a local minimum in our model. Similarly, the strong ξ minimum beneath Tibet in SAW642ANb at 125 km is missing from our model, which exhibits locally high ξ. In addition, there are numerous small-scale variations seen at the base of the mantle which are not consistent between models. Indeed, the level of disagreement between ξ structure seen in our model and the others is not surprising – whole-mantle radial anisotropy is notoriously difficult to constrain, typically explaining only a small fraction of data variance, and is in general poorly correlated across published models (e.g. Becker et al., 2008). In light of this, and given limits imposed by the sensitivity of their dataset, Kustowski et al. (2008a) chose to feature S362ANI, which only inverted for anisotropy above 400 km depth.

4.4.2.3 Structure in cross-section

To gain further insight into the structure of our model, we now turn to cross sections, focused specifically on the newly developed features in the lower mantle.

First, in Figure 4.8, we examine whole-mantle V_S structure in cross-section beneath the Pacific along two profiles normal to absolute plate motion (APM) (Kreemer, 2009), showing both our model and the whole-mantle model S40RTS of Ritsema et al. (2011) to facilitate comparison. At first glance, both models show similar long-wavelength patterns of alternating high and low velocities. Focusing on our model in particular, these images are dominated by large conduit-like low-velocity features in the lower mantle. These features appear to be associated with large-scale low-velocity zones at the base of the mantle within the Pacific LLSVP, and reach the upper mantle in the approximate vicinity of known hotspots – though not strictly collocated with the latter. These features confirm the existence of conduit-like anomalies previously inferred from the deep structure of the SEMum2 model (as discussed in detail in Chapter 3), but as seen earlier in map-view at 1000 km (Figure 4.6), they now appear significantly sharper. In addition, these low-velocity conduits are not the only ones readily associated with Pacific hotspots in our model, with similar structures seen beneath the Caroline, Easter Island, and Macdonald hotspots, for example (though not featured in Figure 4.8 due to space constraints). Further, we note that the morphology of these conduit features changes rapidly at ∼1000 km depth, above which they take on a more undulatory appearance as they extend into the transition zone. This is particularly true for the comparatively thin conduit beneath Hawaii, which is strongly perturbed in this depth range after traversing the lower mantle from what appears to be a semi-isolated low-velocity anomaly near the CMB. Intriguingly, this isolated “source” anomaly to the WSW of Hawaii is approximately collocated with the unusually large ultra-low velocity zone reported by Cottaar and Romanowicz (2012).

Further, upon closely examining the Pitcairn-Guadelupe section in particular, we note that the ∼2000 km APM-normal periodicity previously seen in the SEMum2 upper mantle may in fact reach into and perhaps through the transition zone to ∼1000 km depth. This observation would imply that the 1000 km half-width of this periodicity, i.e. the distance between alternating high and low-velocity bands in map view (250 km depth panel of Figure 4.4), is nearly identical to the depth extent of the associated anomalies. Interestingly, such features may be consistent with the coexistence of multiple scales of convection, dominant over different depth ranges of the mantle: namely, APM-aligned roll-like secondary convection above ∼1000 km depth (with 1000 km separation between the up and downwelling limbs) and larger-scale convection in the mantle below (e.g. Korenaga and Jordan, 2004). Arguments for the requisite stratification of the mantle could be strengthed by observations of slab stagnation below the transition zone near 1000 km (e.g. Fukao and Obayashi, 2013). Indeed, a different dominant scale of convection above 1000 km would help to explain the undulations observed in the low-velocity conduits, similar to the concept of overall convection patterns helping to guide plume-like upwellings, as suggested by Husson and Conrad (2012). In comparing our model with S40RTS along the Pacific cross-sections shown in Figure 4.8, it is immediately apparent that the two are broadly compatible at long wavelengths along the Samoa-Hawaii corridor. However, the same is not true for the Pitcairn-Guadelupe section, where our conduit-like anomaly appearing in the vicinity of Pitcairn is considerably stronger, exhibits a more pronounced skew to the NNE in above 1000 km, and appears to originate directly from a large low-velocity anomaly within the LLSVP. Further, in this same section, we note that the periodicity referred to above is simply not present in S40RTS.

Next, in Figure 4.9, we examine whole-mantle structure beneath Africa following two cross-sections: one running S to N, including the East African Rift (EAR), and the other conjugate to that, running W to E. The desire to examine deeper structure beneath the EAR stems from our earlier observations of a potential "plume" conduit at the southern end of the EAR in SEMum2 (Section 2.4.1 and Figure 2.20). Indeed, our new model confirms the presence of a distinct low-velocity conduit beneath Tanzania, as seen in SEMum2 and previously noted by Weeraratne et al. (2003) in their regional-scale tomographic modeling and also supported by geochemical arguments (e.g. Rogers et al., 2000). Further, similar to the conduit-like features seen in our model beneath the Pacific, the anomaly beneath Tanzania is particularly strong above 1000 km depth and extending into the transition zone (though with some degree of undulation) – a phenomenon that is also apparent in the model S40RTS of Ritsema et al. (2011) (shown in Figure 4.9 as well). In addition, unlike Hawaii, for example, it is more difficult to trace the source of this low-velocity conduit to a deeper anomaly at the base of the mantle (this is also also true for S40RTS), though we note that there *is* a collocated low-velocity feature seen at the base of the mantle in our model – perhaps indicative of some episodic phenomenon that might have causally linked the two. In the conjugate W-to-E plane, lower-mantle structure in our model is dominated by a large-scale slow anomaly extending to mid-mantle depths (we catch the edge of this feature in the earlier EAR section, suggesting that it may be slightly tilted to the east). The apparently sharp top and sides of this anomaly in the mid-mantle, well as the potential eastward trend, are consistent with observations of SKS multi-pathing and waveform modeling reported for the same region of the mantle by Ni et al. (2002).

Finally, in Figure 4.10, we examine whole-mantle cross-sections beneath Iceland, finding strong evidence for a low-velocity conduit extending to an associated anomaly at the CMB. This is in contrast, for example, to what is seen in S40RTS (also shown in Figure 4.10), where a direct connection to lower-most mantle structure is less clear. The two models also differ in the details of upper-mantle structure associated with the Iceland hotspot. Namely, our model finds low-velocity structure above 410 km to be considerably more subtle than that in S40RTS – similar to our earlier observations of upper-mantle structure beneath Pacific hotspots in SEMum2 (Chapter 3) and perhaps indicative of complex flow in the comparatively low-viscosity upper-mantle or other associated change in "plume" morphology. We also note that this upper-mantle structure is more consistent with that seen in the recent regional-scale full-waveform model of Rickers et al. (2013). Further, both global models show an "arm" of low velocities reaching to the ESE below 1000 km depth, toward Eiffel hotspot, which is also seen in the model of Rickers et al. (2013) (e.g. Figure 10a therein).

4.5 Ongoing model evaluation efforts

Here, we detail our ongoing efforts both to evaluate our whole-mantle model, primarily by: (a) probing the cause of remaining waveform misfit and validating with held-out waveform data; and (b) assessing model robustness, largely focused on uncertainty estimation.

4.5.1 Inversion for seismic sources

As proposed at the end of Chapter 2, one of our next steps after obtaining an initial whole-mantle model is to attempt inversion for seismic sources. Earlier tests involving source inversion at the 60 s surface-wave passband, using the SEMum2 model to estimate mantle structure, determined that sensitivity to source parameters was quite small. These earlier experiments led only to very slight changes in source-depth and moment-tensor elements (parameters to which our surface-wave dataset should be most sensitive) relative to source parameters published by the Global CMT project, and indicative that shorter-period modeling (including body waves) would be necessary.

Having inverted for whole-mantle structure, we can now revisit the source inversion including body-wave data in the 32 s passband, in hopes of attaining further waveform variance reduction and improved estimates of source parameters. These forthcoming experiments will use either a hybrid inversion approach analogous to our structural inversion, combining SEM-based forward modeling with source-parameter sensitivities from perturbation theory (the latter were used, for example, in Panning and Romanowicz, 2006), or a fully SEM-based approach similar to that of Liu et al. (2004). We may also explore both techniques for a limited number of events to ensure that they yield similar results.

4.5.2 Independent validation

One key metric for evaluation of tomographic models is examination of predicted waveform fits for held-out data (i.e. waveforms not included in the inversion). We used this technique to validate SEMum2 (discussed in Section 3.4.3), by examining SEM-predicted waveforms for held-out events at *shorter* periods than used in developing the model. Because upper-mantle structure has not changed significantly, it is likely that similarly impressive validation results, again focused on surface-wave data at 40 s period, will be obtained for our new whole-mantle model. However, this needs to be confirmed quantitatively and such tests are ongoing.

More important, however, is evaluation of *lower-mantle* structure, which requires validation at body-wave appropriate periods (down to at least 32 s). During the course of our inversion, we used held-out waveform data to confirm that our model was indeed converging (and not simply over-fitting the inversion dataset). As noted above (Section 4.4.1), our tests using held-out events at the 32 s body-wave inversion passband did indeed confirm that our model progressively improved fits to these data. This was quantified by the number of waveform windows selected by our automated picking procedure (applied to the held-out data, but without incorporating it into our inversion dataset) – allowing, for example, ∼100 new lower-mantle sensitive windows to be selected from 10 held-out events using SEM synthetics computed in our final model relative to the previous iteration. However, by analogy with our tests for SEMum2, what would be even more informative is (a) examination of shorter periods and (b) comparison against SEM-predictions from other global tomographic models, such as those shown in Figure 4.4. These latter tests are ongoing.

4.5.3 Uncertainty

4.5.3.1 Scaling factors

As noted in Section 4.2.3, the scaling factors that we use to estimate model perturbations to the remaining four parameters of a radially anisotropic medium are only strictly valid in the upper mantle (Montagner and Anderson, 1989). While our sensitivity to ρ is anticipated to be quite small throughout the mantle, and anisotropy is expected to be largely limited to the boundary layers, our sensitivity to Voigt-average isotropic V_P may be non-negligible (particularly as we include shorter-period waveform data). Potential depth dependence of the scaling factor between V_S and V_P is well known, and potentially quite strong in the lower-most mantle (e.g. Su and Dziewonski, 1997; Masters et al., 2000; Romanowicz, 2001), leading some studies to incorporate depth-dependent scaling factors into their inversions (e.g. Ritsema et al., 2011).

Motivated by these observations, we sought to assess the effect of "realistic" (i.e. as implied by previous studies) variation in lower-most mantle V_S to V_P perturbation scaling. Namely, we repeated our final-iteration inversion for two different configurations of the $\delta \ln V_P / \delta \ln V_S$ scaling factor: using a ratio of 0.5 throughout the mantle, except for the bottom ∼300 km, where we used ratios of either 0.33 or 0.25. We found that the two models obtained were nearly identical to our preferred model, which used 0.5 everywhere (to within a fraction of a percent V_S perturbation). A similar result was obtained by Kustowski et al. (2008a) in their tests of scaling-factor sensitivity. One implication of this result is that our aggregate sensitivity to V_P, even at 32 s period, remains quite small. In future work incorporating even shorter-period waveform data, however, these scaling factors will have to be reexamined in greater detail.

4.5.3.2 Linear resolution analysis

Traditional linear resolution analysis (e.g. Tarantola, 2005) comes with numerous caveats (Lévêque et al., 1993) – particularly for strongly non-linear problems (where it is expected only to be valid near the optimal model). We previously discussed these shortcomings in Sections 2.3.3 and 3.4.5.1, and refer the reader to those sections for more details. Importantly, it must be recognized that the standard linear analysis in no way reflects the accuracy of the underlying theoretical approach used in forward modeling nor in sensitivity kernel calculation (using SEM and NACT respectively in our case). Instead, given a particular choice of test structure, the analysis may only be used to assess: (a) admissibility of the structure under the chosen model basis and (b) sub-optimal distribution of sensitivity, typically characterized by gaps or ray-like smearing in the recovered test structure.

All of this said, the latter two capabilities of the analysis are still of use. Not only for the commonly used “checkerboard” test structures, which can yield misleading results (Lévêque et al., 1993), but for examining the robustness of certain specific features of interest using appropriately chosen input structures. For example, the plume-like conduits in our lower-mantle model are wider than those typically expected in a mantle heated only from below, which has important geodynamic implications. Whether this is a robust result may partially be assessed via resolution analysis using a range of thinner, hypothetical plume-like test structures. Similarly, our model shows evidence for apparent lateral translation of “plume” material in the lower mantle near the 1000 or 650 km depth horizons, with some hotspots appearing linked directly to deep sources (for example, Hawaii and Iceland) while others do not appear so (for example, Tanzania) (e.g. Courtillot et al., 2003). Further resolution analysis using plume-like test structures may be a valuable tool for assessing (though not ruling out) whether such lateral translation may be attributed to “ray-like” smearing, and whether small scale conduits in the mid-mantle may be present but remain invisible to our inversion. These tests are ongoing.

4.5.3.3 Resampling techniques

Another way to probe uncertainty in the resulting model is through statistical resampling techniques, such as the deleted jackknife (e.g. Efron and Stein, 1981) or bootstrap (e.g. Efron and Tibishirani, 1991). For example, we found bootstrap resampling to be a valuable tool for uncertainty estimation in our supplementary analysis of the SEMum2 model in Chapter 3 (Section 3.4.5.2). Here, we take a similar resampling approach, based on repeating our final inversion iteration for many different realizations of our waveform dataset (again owing to the large changes to the model that are possible in a single step of the Gauss-Newton scheme, should the data warrant it). Previously, Panning and Romanowicz (2006) used both the bootstrap and jackknife techniques in the analysis of their global model SAW642AN, and found that the jackknife rapidly converged to the same uncertainty estimate as the bootstrap, but required significantly fewer samples. This latter point is an advantageous property, given that each sample requires a significant amount of I/O (reading the Gauss-Newton Hessian) and subsequent factorization, and drove us specifically to consider the jackknife technique.

Here, we follow a similar approach to that in Panning and Romanowicz (2006). Namely, we separate our waveform dataset into 12 bins by month of earthquake occurrence, and use these bins as aggregate data (i.e. our dataset contains 12 observations). We then, in turn, use the deleted jackknife technique, to estimate variance of our model $\sigma^2_{\text{jackknife}}$ via

$$\sigma^2_{\text{jackknife}} = \frac{n-d}{d \cdot C} \sum_{i=1}^{C} \left[\theta_{(i)} - \hat{\theta}\right]^2 \quad \text{where} \quad \hat{\theta} = \frac{1}{C} \sum_{i=1}^{C} \theta_{(i)} \quad \text{and} \quad C = \frac{n!}{d!(n-d)!}$$

where $\theta_{(i)}$ is the model estimate produced by the i^{th} deleted-jackknife realization of our dataset, the number of data n is 12, and the number of deletions d per dataset realization is 1 or 2. For the $d = 2$ case, this approach requires only 66 solves (each producing a model sample), while the month-based binning scheme requires pre-computation and storage of only 12 Hessian matrices which may be read in different combinations at solution time (occupying approximately 2.2TB if stored in full form, chiefly to simplify collective I/O prior to factorization, or 1.1TB if storing only the upper triangular part). We further note that the uncertainty estimates recovered in this manner will be *conservative over-estimates* thereof, in that we do not re-optimize the waveform-window weights for each jackknife sample (which our inversion realistically does prior to solution, allowing us to account to some extent for uneven data coverage).

In Figure 4.11, we show preliminary results obtained at three lower-mantle depths for $d = 2$ (the $d = 1$ case is very similar), where we have masked V_S model structure that is below the local 2-σ uncertainty estimates determined by resampling. At all three depths, we find that the primary structures interpreted above in Section 4.4.1 remain robust – including: (a) high-velocity features associated with subducted slabs extending beneath the transition zone; (b) concentrated low-velocity features at mid-mantle depths interpreted as plume-like conduits (beneath Hawaii, the Superswell, Tanzania, Iceland, etc.); and (c) low-velocity features at the base of the mantle, such as the isolated anomalies embedded within the LLSVPs, as well as the Perm anomaly. In our ongoing work, we intend to examine more closely the spatial variation of the uncertainty estimates, specifically focusing on how to interpret these results in terms of the robustness of long-wavelength, low-amplitude background structures (i.e. the degree-2 pattern). We will also apply the same approach to examine the robustness of our ξ model.

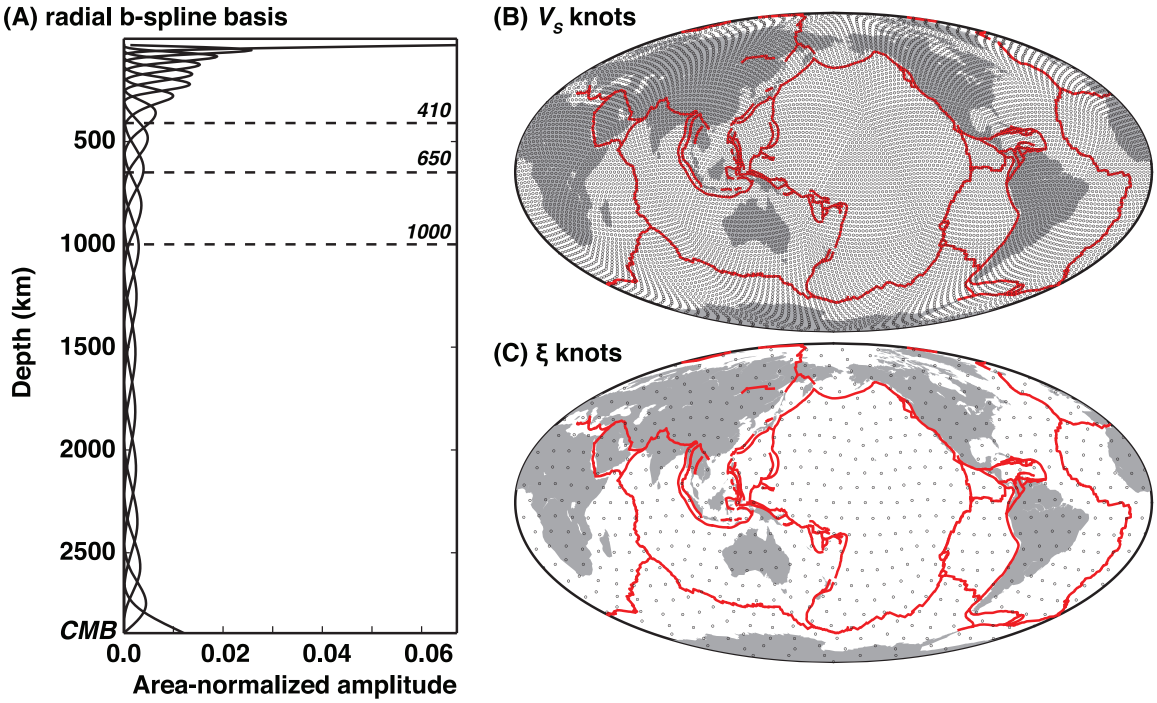


Figure 4.1: Parameterization of our whole-mantle model. Panel (A), 20 radial b-spline basis functions (e.g. Mégnin and Romanowicz, 2000), distributed between the CMB and the shallowest Moho in our smooth crustal model (30 km) according to expected resolution. Panels (B) and (C), the distribution of knots supporting our spherical spline (Wang and Dahlen, 1995) basis for V_S and ξ (with average spacing of $< 2°$ and $< 8°$, respectively).

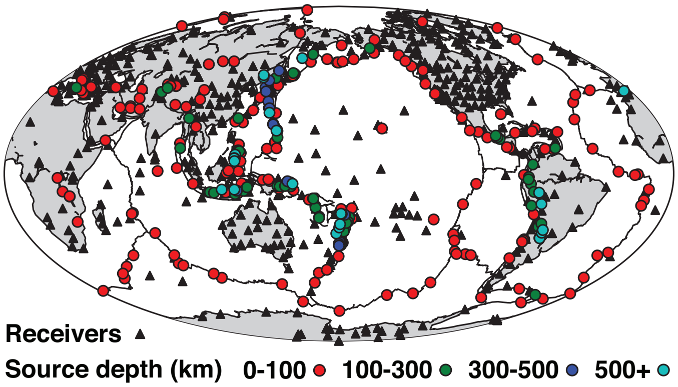


Figure 4.2: Distribution of the 273 events and over 500 seismic stations used in this study.

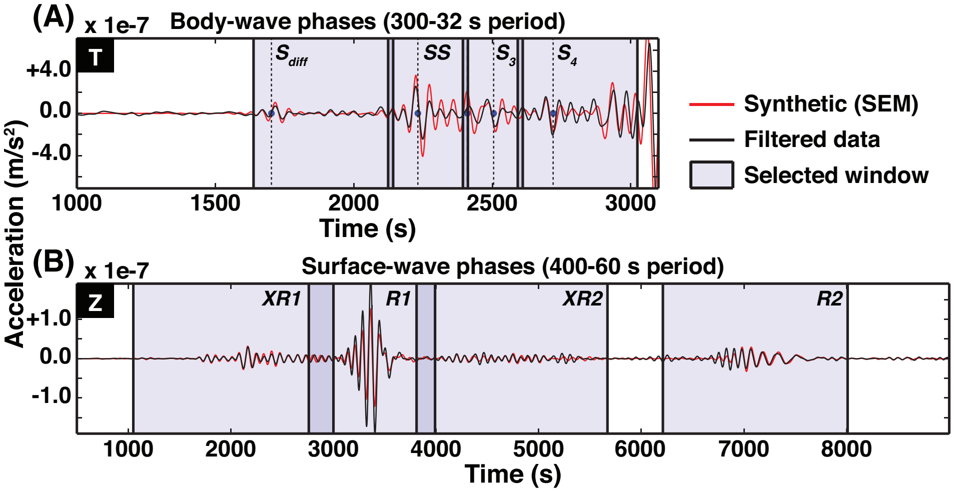


Figure 4.3: Example windowed body-wave (A) and surface-wave (B) waveforms using the processing scheme described in the text (Li and Romanowicz, 1996). Data windows are selected based on similarity to SEM synthetics computed in the most recent iteration of the mantle model (i.e. the dataset is reprocessed at each iteration).

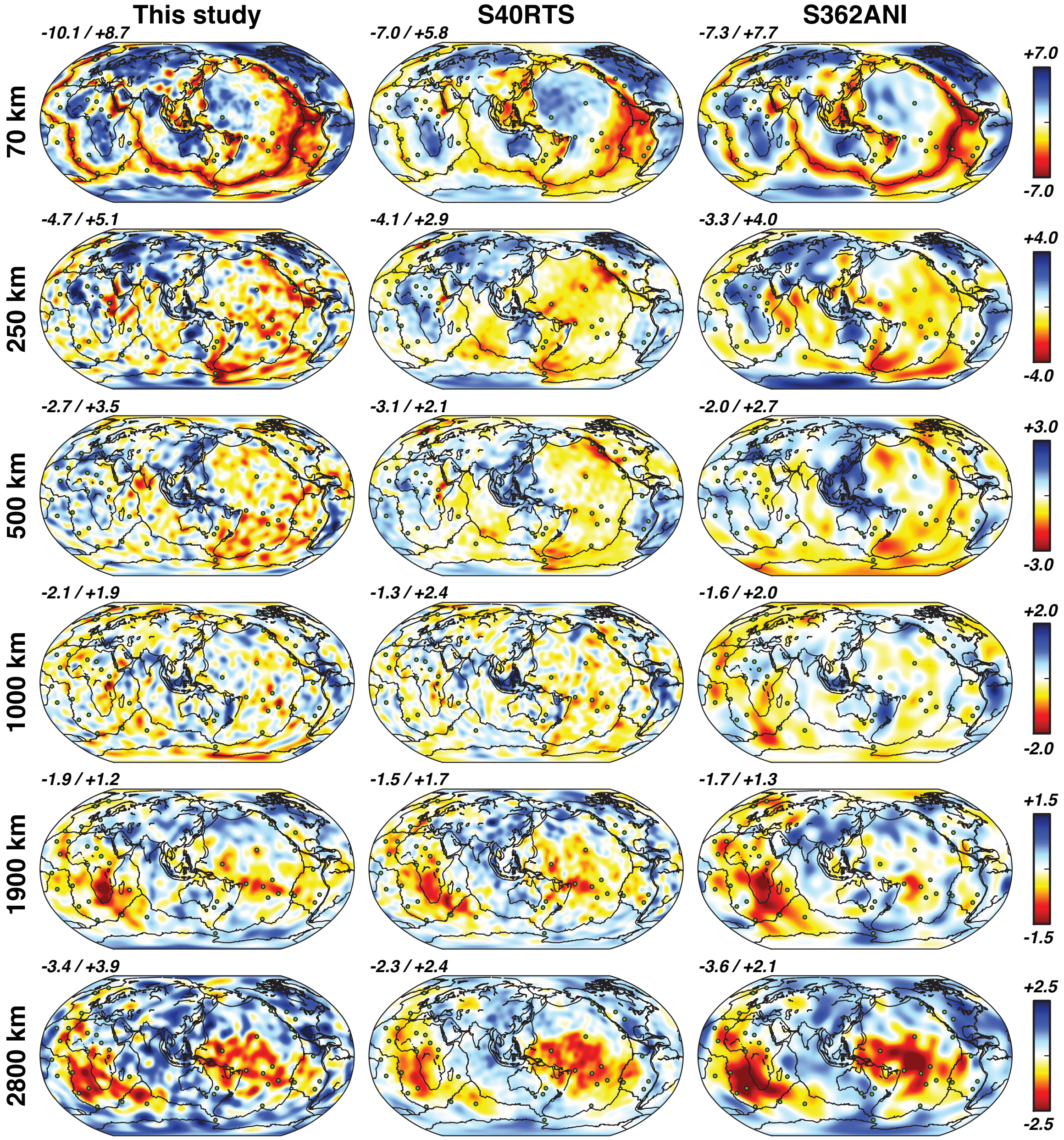


Figure 4.4: Map views of global V_S variation at a range of depths throughout the mantle for the model obtained in this study, as well as recent whole-mantle models S40RTS of Ritsema et al. (2011) and S362ANI of Kustowski et al. (2008a). Variations are plotted in percent with respect to the global mean at each depth, with the exception of 2800 km, plotted with respect to the 1D model PREM (Dziewonski and Anderson, 1981). Inset values (upper-left corner of each panel) represent maximum peak-to-peak variation for each model at the corresponding depth. Circles denote hotspot locations from Steinberger (2000).

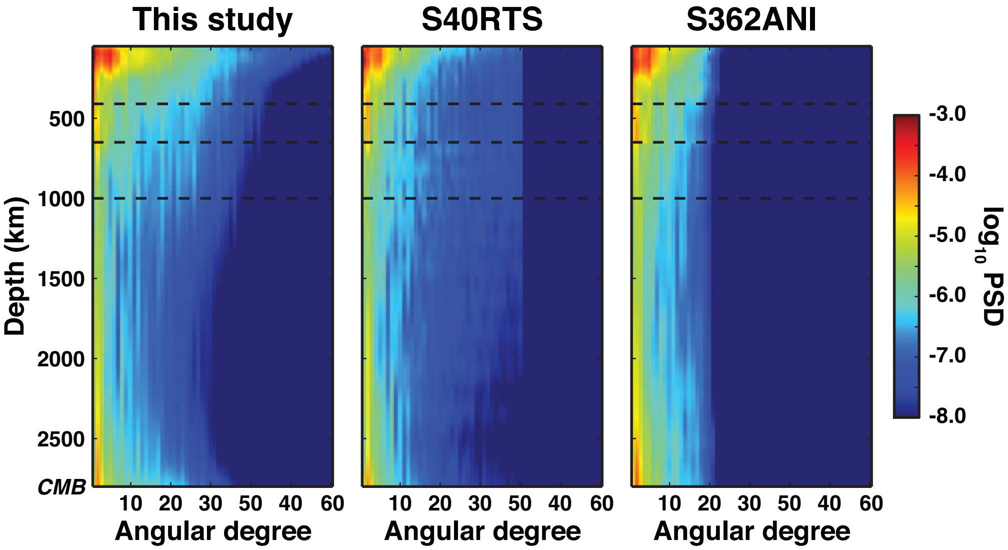


Figure 4.5: Spherical-harmonic power spectral densities, computed for the three global models shown in Figure 4.4 as a function of depth (note: color-scale is logarithmic). Horizontal dashed lines correspond to 410, 650, and 1000 km depth (from top to bottom).

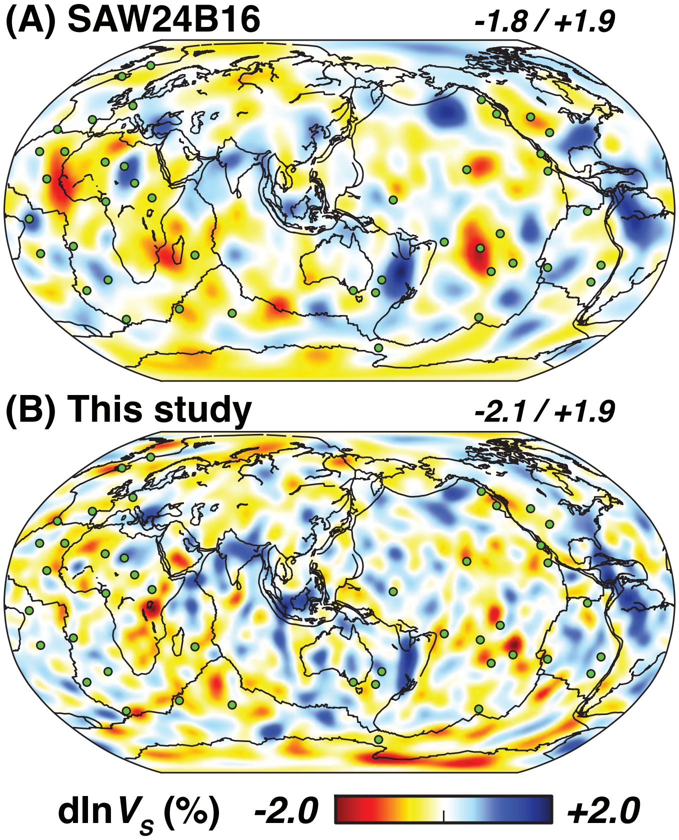


Figure 4.6: A comparison between starting and final-iteration model V_S structure at 1000 km depth (plotted as percent variation with respect to the global mean at that depth). At 1000 km, the starting model is SAW24B16 of Mégnin and Romanowicz (2000). Inset values (upper-right corner of each panel) represent maximum peak-to-peak variation for each model at 1000 km. Circles denote hotspot locations from Steinberger (2000).

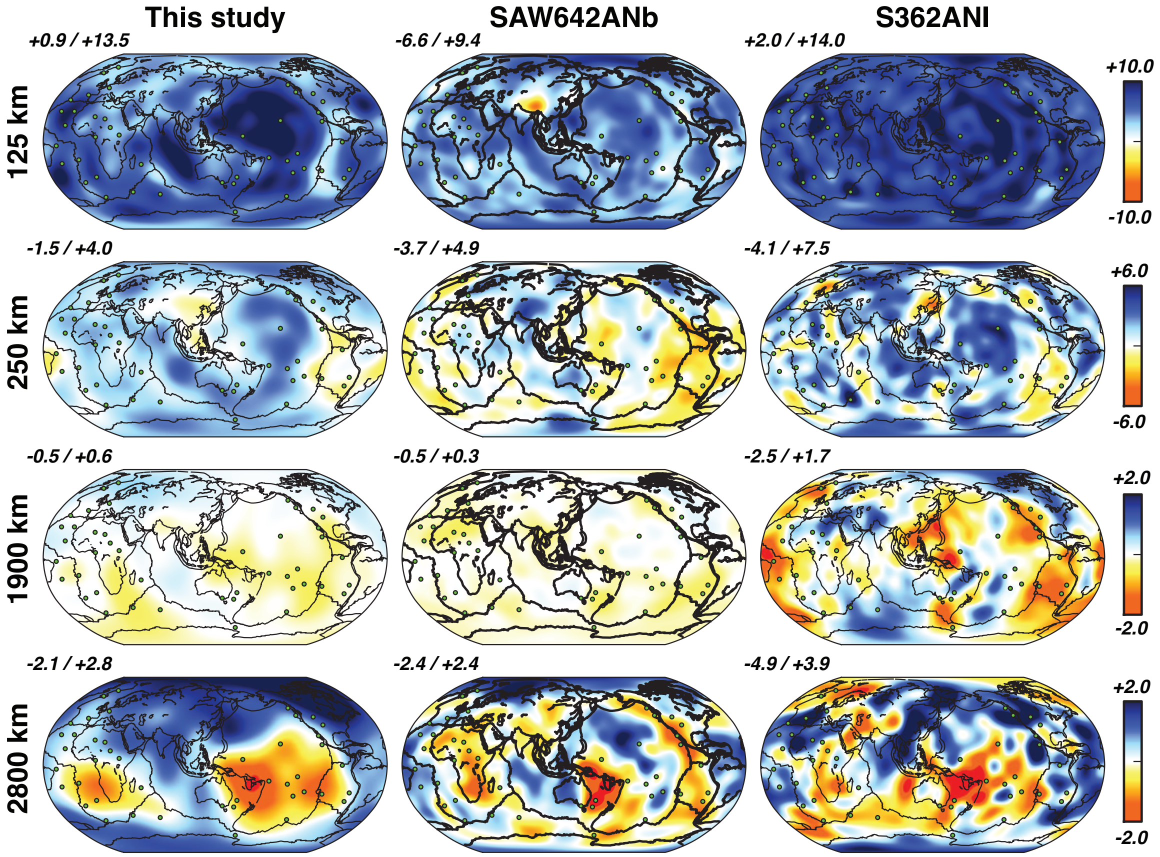


Figure 4.7: Map views of global ξ variation, expressed in percent relative to isotropy ($\xi = 1$), at a range of depths throughout the mantle. We show ξ structure obtained in this study (left), as well as that from the models SAW642ANb (middle) of Panning et al. (2010) and S362WMANI (right) of Kustowski et al. (2008a). SAW642ANb is a revised version of SAW642AN of Panning and Romanowicz (2006), obtained using a different treatment of crustal structure, while S362WMANI is the whole-mantle anisotropic variant of S362ANI (from the same study). Inset values (upper-left corner of each panel) represent maximum peak-to-peak variation for each model at the corresponding depth. Circles denote hotspot locations from Steinberger (2000).

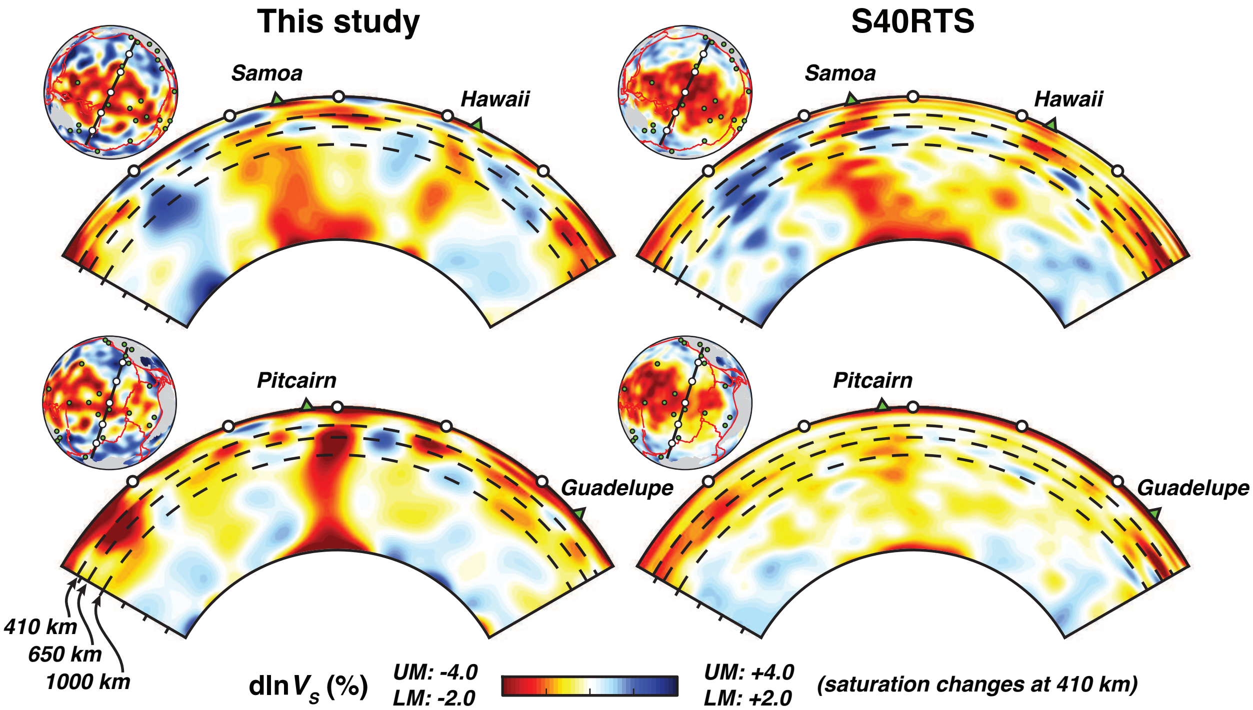


Figure 4.8: Cross sections showing V_S variation (%) throughout the mantle beneath the Pacific, including both the present study (left panels) and model S40RTS (right panels) of Ritsema et al. (2011) (section width: 120°; white circles denote waypoints spaced every 20°). Lines of section are oriented normal to absolute plate motion (Kreemer, 2009). Inset map shows cross-section location, with background V_S structure at 2800 km depth. Note: Saturation level of the colormap changes at 410 km depth, from 2% below to 4% above (more appropriate for upper-mantle structure).

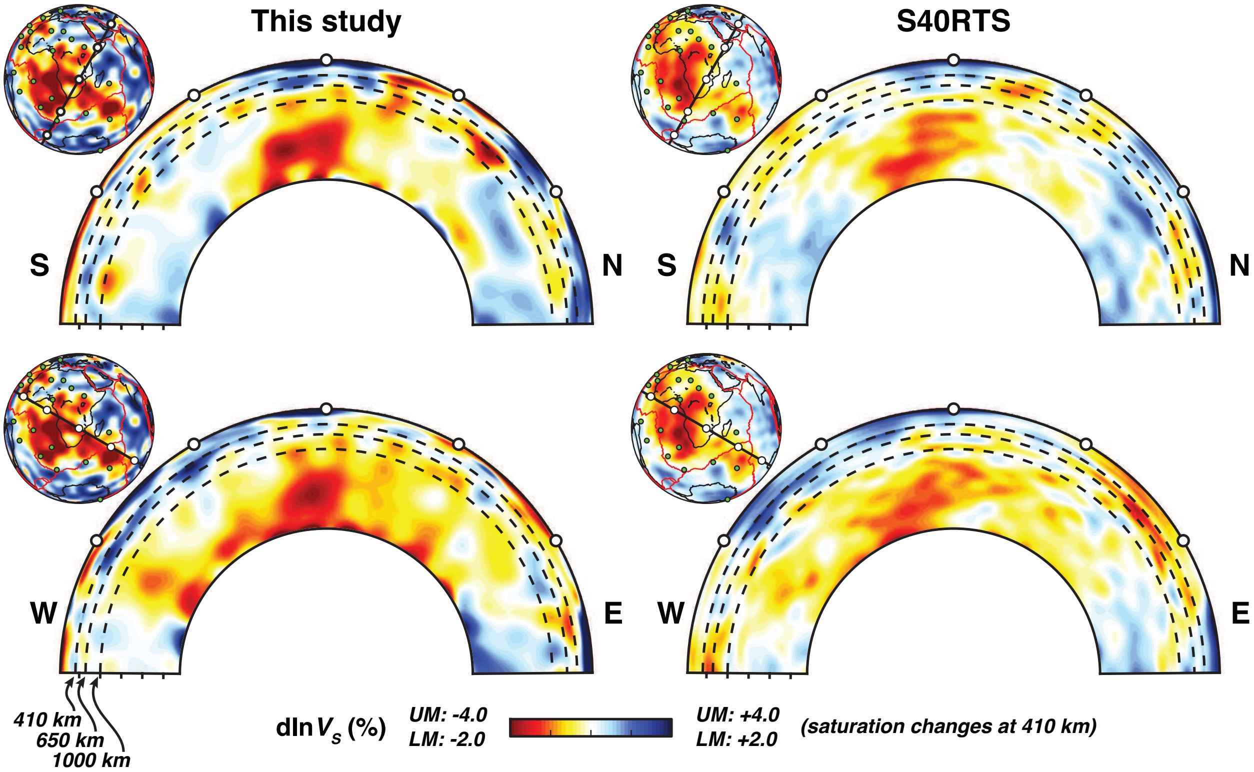


Figure 4.9: Cross sections showing V_S variation (%) throughout the mantle beneath the African region, including the East African Rift (width: 180°; waypoint spacing: 30°), including the present study (left panels) and S40RTS (right panels) of Ritsema et al. (2011). Inset map shows cross-section location, with background V_S structure at 2800 km depth. Note: Saturation level of the colormap changes at 410 km depth, from 2% below to 4% above (more appropriate for upper-mantle structure).

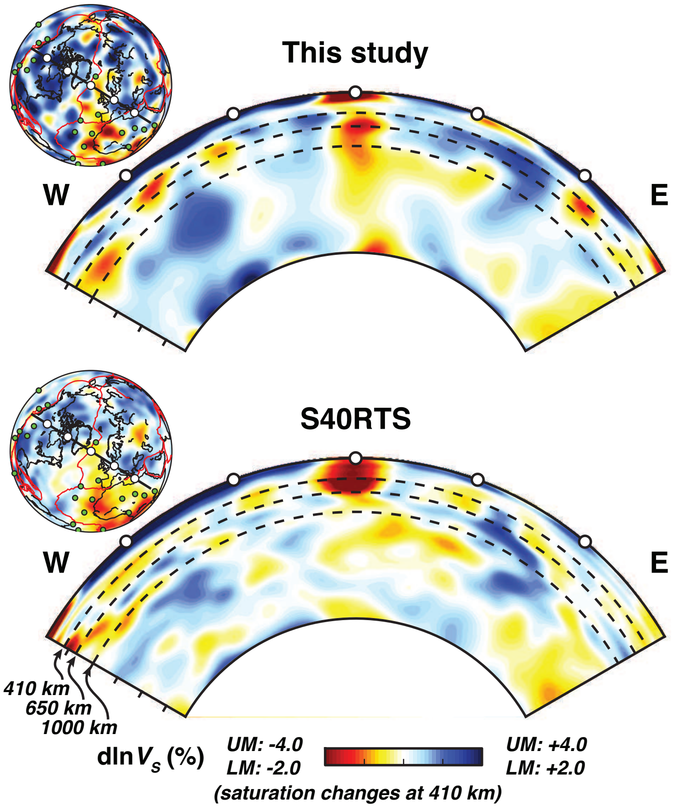


Figure 4.10: Cross sections showing V_S variation (%) throughout the mantle beneath Iceland (width: 120°; waypoint spacing: 20°), including the present study (upper panel) and S40RTS (lower panel) of Ritsema et al. (2011). Inset map shows cross-section location, with background V_S structure at 2800 km depth. Note: Saturation level of the colormap changes at 410 km depth, from 2% below to 4% above (more appropriate for upper-mantle structure).

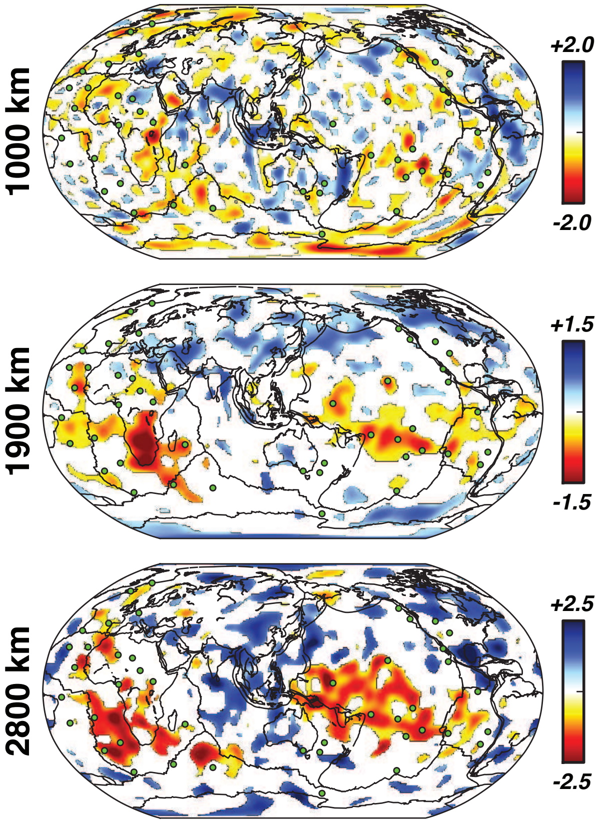


Figure 4.11: Maps of global V_S variation at three lower-mantle depths similar to Figure 4.4, but now with V_S perturbations below local 2-σ uncertainties, as determined by our preliminary deleted-jackknife resampling, masked ($d = 2$ case, see text).

Chapter 5

Parallel Hessian estimation for full-waveform seismic inversion using the PGAS model

An abridged version of this chapter has previously been submitted to *Supercomputing* 2014 as an application-track paper, under the title: “Parallel Hessian Estimation for Full-Waveform Seismic Inversion using the PGAS Model.” This version has been modified in order to fit more cohesively into the present work.

5.1 Summary

We present the design and evaluation of a distributed matrix-assembly abstraction for large-scale inverse problems in HPC environments: namely, parallel physics-based Hessian estimation in full-waveform seismic inversion at the scale of the entire globe. Our solution is based on UPC++, a new PGAS extension to the C++ language, leveraging its novel features to implement relaxed asynchronous updates to distributed matrix elements.

Our evaluation includes scaling results for Hessian estimation on up to 12,288 cores, representative of current production scientific runs and next-generation global-scale inversions. We also draw comparisons to a one-sided MPI-based reference implementation, focusing on performance and developer productivity. We find that interoperability between UPC++ and other parallel programming tools (e.g. OpenMP, MPI), allows for incremental adoption of the PGAS model where most beneficial. Further, we note that this particular model of asynchronous distributed matrix assembly can easily generalize to other application domains.

5.2 Introduction

Full-waveform seismic inversion broadly refers to a class of techniques used in developing models of the earth’s interior structure and properties. The unifying characteristic of all full-waveform approaches lies in the underlying model optimization problem, which attempts to minimize the difference between predicted and observed seismic recordings (seismograms) – either of naturally occurring earthquakes or controlled sources (explosions). The primary motivation for inverting complete waveforms, in lieu of commonly used discrete measures (e.g. arrival times of seismic phases), is to fully exploit the information content of the wavefield. High-quality models of the earth's interior properties, as seen by seismic waves, have a diverse range of applications, including basic science (interior structure and composition), geophysical exploration (energy production), environmental monitoring (carbon sequestration), and Comprehensive Nuclear-Test-Ban Treaty verification.

Over the past 30 years, seismic inversion at the global scale has made substantial progress in refining our understanding of the earth's interior (Thurber and Ritsema, 2007). Further, global seismology has a long history of employing waveform inversion techniques, using increasingly accurate approximations for the physics of wave propagation (Woodhouse and Dziewonski, 1984; Li and Romanowicz, 1996). More recently, the spectral finite element method has become a popular tool for numerical simulation of the seismic wavefield at the regional and global scale (Komatitsch and Tromp, 2002a; Chaljub et al., 2003; Capdeville et al., 2003). Concurrently, the adjoint-state method for iterative model optimization (Tarantola, 1984) – long popular in exploration geophysics – has been adopted by the wider seismology community (Tromp et al., 2005) and applied at both the regional (Tape et al., 2009) and continental (Fichtner et al., 2009) scale, with considerable success. However, application of these techniques for full-waveform inversion at the global scale remains computationally formidable.

One key practical issue facing full-waveform inversion based on adjoint methods is convergence rate. As typically implemented (first-order adjoint), this technique yields only the gradient of the underlying misfit function being optimized – requiring a slowly converging gradient-based optimization scheme, and leading to large numbers of numerical wavefield simulations. Newton-like methods, with enhanced convergence rates, are a popular topic in the exploration literature, particularly the Newton-Krylov (Epanomeritakis et al., 2008) and truncated Newton methods (Métivier et al., 2013). However, such schemes rely on second-order adjoint state methods to calculate the necessary Hessian-vector products, and thus require additional *inner* wavefield simulations.

An alternative to these *matrix-free* Newton schemes can be implemented if an inexpensive, but sufficiently accurate physics-based estimate of the Hessian matrix is available. Such a scheme would eliminate the expensive inner wavefield simulations (at the expense of introducing matrix factorization) while still attaining Newton-like convergence rates. We have found this approach to be effective at rendering global-scale full-waveform inversion tractable (Lekić and Romanowicz, 2011a; French et al., 2013). One inherent issue with this technique is that the Hessian matrix is not only fully dense, but also can be very large: its dimension is equal to the number of model parameters, which rapidly increases as models of the earth's interior attain higher resolution. Thus, assembly of the Hessian – typically the aggregate result of *many* distributed computations – becomes difficult in practice, due to both its size and impediments to achieving high-performance random-access mutations of matrix elements in a distributed memory setting.

Here, we present our solution to this problem: a dense-matrix abstraction for *scalable* distributed matrix assembly in HPC environments. This is achieved using UPC++ (Zheng et al., 2014), a new partitioned global address-space (PGAS) extension to C++ that incorporates and builds upon numerous popular features from other well-established PGAS languages. We exploit this novel combination of capabilities to enable fully concurrent communication and (physics) computations, while requiring minimal synchronization. Namely, the combination of UPC++ support for one-sided bulk communication and remote memory management operations, with asynchronous remote task execution for application of matrix updates with fully user-defined logic, proved fundamental to the success of this solution. In addition, interoperability of UPC++ with other parallel programming systems or libraries, as required by OpenMP and MPI-based components of our application (Hessian computation and collective IO, respectively), allowed for incremental adoption of the PGAS model where most appropriate. We further note that this approach is broadly translatable to HPC applications in other data-driven domains requiring similar distributed matrix assembly operations.

The remainder of this Chapter is structured as follows: In Section 5.3, we discuss the particular methods employed in global full-waveform inversions using the scheme described above, providing context for the underlying computational problem and motivating our solution. In Section 5.4, we focus on the design and implementation of our distributed matrix abstraction based on UPC++. Evaluation of our solution follows in Section 5.5, including scaling analyses for present- and anticipated next-generation scales of seismic inversion, as well as comparisons to alternative implementation strategies using other technologies. Finally, we present new scientific results obtained using these techniques in Section 5.6, and conclude with Section 5.7, where we summarize the contributions of the present work and discuss future directions.

5.3 Computational problem

5.3.1 Hybrid full-waveform inversion

Full-waveform inversion may be viewed as an optimization problem, in which we seek a model $\mathbf{m}$ of the earth's interior properties that minimizes the difference between *observations* of the seismic wavefield (seismograms) and *predictions* thereof given $\mathbf{m}$. Here, we adopt the generalized least-squares criterion (Tarantola and Valette, 1982) and define a misfit function $\chi(\mathbf{m})$ of the form:

$$2\chi(\mathbf{m}) = \|\mathbf{d} - \mathbf{g}(\mathbf{m})\|^2_{\mathbf{C}_\mathbf{d}^{-1}} + \|\mathbf{m} - \mathbf{m}_p\|^2_{\mathbf{C}_\mathbf{m}^{-1}} \tag{5.1}$$

where $\mathbf{d}$ are the observed data, $\mathbf{g}(\cdot)$ is the forward operator which predicts $\mathbf{d}$ given $\mathbf{m}$, $\mathbf{C}_\mathbf{d}$ reflects uncertainty in the data space, and $\mathbf{m}_p$ and $\mathbf{C}_\mathbf{m}$ characterize an assumed prior distribution in the model space ($\mathbf{m} \sim \mathcal{N}(\mathbf{m}_p, \mathbf{C}_\mathbf{m})$).

Because $\mathbf{g}(\mathbf{m})$ is non-linear in $\mathbf{m}$, $\chi(\mathbf{m})$ must be minimized with either stochastic model-space sampling or iterative optimization. Due to the expense of evaluating $\mathbf{g}(\mathbf{m})$ in full-waveform imaging, an iterative technique is invariably chosen. This may be achieved using the Gauss-Newton scheme following naturally from eqn 5.1, namely:

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \left(\mathbf{C_m}\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \mathbf{I}\right)^{-1} \left(\mathbf{C_m}\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\left[\mathbf{d} - \mathbf{g}(\mathbf{m}_k)\right] - \mathbf{m}_k + \mathbf{m}_p\right) \tag{5.2}$$

where $\mathbf{m}_k$ and $\mathbf{m}_{k+1}$ represent the k^{th} iterative model estimate and its successor, while $\mathbf{G}$ is the wavefield Jacobian matrix, i.e. the partial derivatives of the forward operator (and thus also the data) with respect to the model parameters, evaluated at $\mathbf{m}_k$:

$$\{\mathbf{G}\}_{ij} = \partial\{\mathbf{g}(\mathbf{m})\}_i \;/\; \partial\{\mathbf{m}\}_j \;|_{\mathbf{m}=\mathbf{m}_k}$$

where $\{\cdot\}_q$ represents the q^{th} element of a vector or matrix (and q is a linear index or tuple, respectively). The Gauss-Newton approximation arises due to the inherent linearization of $\mathbf{g}(\mathbf{m})$, yielding the leading-term Hessian estimate $\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}$. As noted in Section 5.2, one advantage of a Newton-like scheme over purely gradient-based optimization methods (e.g. non-linear CG (Fletcher and Reeves, 1964)) is rapid convergence, thereby cutting down on the total number of iterations of eqn 5.2 and expensive computations of $\mathbf{g}(\mathbf{m})$. The Gauss-Newton scheme in particular is appropriate for cases where: (a) the total iteration count is also too small for quasi-Newton methods to build up a sufficiently accurate inverse-Hessian estimate; and (b) the Jacobian $\mathbf{G}$, or a sufficiently accurate estimate, may be calculated cheaply. We refer to our approach as *hybrid*, because it combines computationally heavy but highly accurate numerical simulations for the forward computation $\mathbf{g}(\mathbf{m})$ and light-weight physics-based estimates of the Jacobian $\mathbf{G}$ to obtain the Gauss-Newton Hessian $\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}$ (as well as the (negative) misfit gradient $\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\left[\mathbf{d} - \mathbf{g}(\mathbf{m})\right]$).

5.3.2 Estimation of the Hessian

5.3.2.1 Perturbation theory

To derive a light-weight estimate of the wavefield Jacobian, we use non-linear asymptotic coupling theory (NACT) (Li and Romanowicz, 1996). This formalism is based on perturbation theory of the earth's normal modes – the eigensolutions of the elastodynamic equations governing wave propagation in a self-gravitating earth. Namely, in a *reference* earth that is spherically symmetric, non-rotating, elastic, and isotropic, the set of normal modes $\{\mathbf{u}_k\}$, with associated frequencies of oscillation $\{\omega_k\}$, form a complete basis, and the wavefield excited by an earthquake-like body force may be represented by summation over modes:

$$\mathbf{u}^{(0)}(\mathbf{x}, t) = \mathbf{Re}\sum_k \alpha_k \exp\left[i\omega_k t\right]\mathbf{u}_k(\mathbf{x}) \tag{5.3}$$

where α_k are coefficients due to the particular excitation. To first order in small perturbations away from this reference state (e.g. hydrostatic ellipticity, lateral heterogeneity in elastic properties, etc.), the elastodynamic equations may be linearized, and an expression analogous to eqn 5.3 for the wavefield in the *perturbed* earth can be derived, though now including coupling between pairs of modes (k, k'):

$$\mathbf{u}^{(1)}(\mathbf{x}, t) = \mathbf{Re} \sum_{k,k'} \alpha_{k'} \exp \left[i \left(\omega_k \delta_{kk'} + \frac{Z_{kk'}}{\omega_k + \omega_{k'}} \right) t \right] \mathbf{u}_k(\mathbf{x}) \tag{5.4}$$

where $Z_{kk'}$ now embodies the coupling effect between mode pairs and requires a computationally heavy integration over the entire earth volume.

To render eqn 5.4 usable for time-domain waveform inversion, an additional linearization is applied, often in the context of the Born (single-scattering) approximation, such that the perturbed wavefield may be expressed as:

$$\mathbf{u}^{(1)}(\mathbf{x}, t) = \mathbf{u}^{(0)}(\mathbf{x}, t) + \delta \mathbf{u}(\mathbf{x}, t) \tag{5.5}$$

where the term $\delta \mathbf{u}(\mathbf{x}, t)$ depends linearly on the perturbations away from the reference state. Indeed, this set of perturbations comprise our image of the earth's interior $\mathbf{m}$, and the partial derivative terms $\partial \mathbf{u}^{(1)}(\mathbf{x}, t_i)/\partial\{\mathbf{m}\}_j$ are the wavefield Jacobian elements $\{\mathbf{G}\}_{ij}$ that we seek. Unfortunately, the time-domain Born approximation is not stable for realistically large amplitudes of perturbations.

NACT addresses this shortcoming by retaining a portion of coupling effect, namely that associated with multiple forward scattering, in the exponential term (e.g. eqn 5.4). This property also implies that estimates of Jacobian elements obtained using NACT depend non-linearly on $\mathbf{m}$, and must be re-calculated as the iterative inversion proceeds and $\mathbf{m}$ evolves. Since the Jacobian estimate must be recalculated regularly, NACT takes additional steps to reduce computational cost by collapsing the expensive coupling integration over the entire earth onto the great-circle plane joining each earthquake source and seismic receiver using the stationary phase approximation.

Thus, the underlying computational kernel of our Jacobian estimation reduces to a series of path integrations: one integration for each choice of mode pair (k, k') and source-receiver path (corresponding to a single seismogram used in the inversion). The overall cost scales as

$$O\left(N_{\mathbf{m}_r} \cdot \sqrt{N_{\mathbf{m}_{\theta\phi}}} \cdot N_{SR} \cdot f^4\right)$$

where $N_{\mathbf{m}_r}$ and $N_{\mathbf{m}_{\theta\phi}}$ correspond to the radial (depth) and lateral (latitude-longitude) dimensions of $\mathbf{m}$ (i.e. the overall dimension of $\mathbf{m}$ is $N_{\mathbf{m}} = N_{\mathbf{m}_r} N_{\mathbf{m}_{\theta\phi}}$), N_{SR} is the number of source-receiver paths, and f is the maximum frequency considered in the wavefield (the number of coupling-mode pairs (k, k') grows as f^4). A more complete review of normal-mode perturbation theory in the context of waveform inversion may be found in (Romanowicz et al., 2008).

5.3.2.2 Practical considerations

When considering realistically large numbers of data $N_{\mathbf{d}} = \dim \mathbf{d}$, where $N_{\mathbf{d}} \gg N_{\mathbf{m}}$, the Jacobian $\mathbf{G}$ is in general too large to form explicitly. Instead, we form the $N_{\mathbf{m}} \times N_{\mathbf{m}}$ Hessian estimate $\mathbf{G}^T\mathbf{G}$ and the (negative) misfit gradient vector $\mathbf{G}^T[\mathbf{d} - \mathbf{g}(\mathbf{m})]$ directly (here, we have absorbed $\mathbf{C}_\mathbf{d}^{-1/2}$ into $\mathbf{G}$ and $\mathbf{d} - \mathbf{g}(\mathbf{m})$ for notational convenience). Typically, for each datum i, corresponding to a particular source-receiver path and recorded seismogram, NACT (Section 5.3.2.1) yields a single column-strided panel of the Jacobian $\mathbf{G}$, which we denote $\mathbf{G}_{(i)}$. Similarly, we can denote the associated observed seismogram and model prediction as $\mathbf{d}_{(i)}$ and $\mathbf{g}(\mathbf{m})_{(i)}$. The particular striding pattern associated with the trailing index of $\mathbf{G}_{(i)}$ arises from the application of the stationary phase approximation, which limits the non-zero elements of $\mathbf{G}$ to model parameters encountered along the source-receiver great circle, and thus depends entirely on the source-receiver geometry. For each datum i of size k (the number of time samples in $\mathbf{d}_{(i)}$), $\mathbf{G}_{(i)}$ is $k \times n$, where n is typically an order of magnitude smaller than $N_\mathbf{m}$ (namely, $n \propto N_{\mathbf{m}_r}\sqrt{N_{\mathbf{m}_{\theta\phi}}}$), while k varies independently from n and is at least an order of magnitude smaller in practice.

Thus, for each datum i, there is an $n \times n$ symmetric Hessian update $\mathbf{G}_{(i)}^T\mathbf{G}_{(i)}$ that must be *merged* into the full Hessian $\mathbf{G}^T\mathbf{G}$. The particular merge operation is simply addition – namely, the additive "augmented assignment" operator `+=` – and the mapping between elements is given by a strided slicing operation; or, in pseudocode: `GtG[ix,ix] += GtG_i[:,:]` where `ix` is a suitable indexing array (Fig. 5.1). Updates to the (negative) misfit gradient vector for each datum i, namely $\mathbf{G}_{(i)}^T\left[\mathbf{d}_{(i)} - \mathbf{g}(\mathbf{m})_{(i)}\right]$, follow a similar pattern, though clearly only in one dimension: `grad[ix] += grad_i[:]`.

5.3.2.3 Parallel implementation with replication

Each NACT calculation, corresponding to one particular datum, is wholly independent of every other. Thus, NACT-based Hessian and gradient estimation is data-parallel and proceeds in two phases:

1. A map operation over the waveform data $\mathbf{d}$ and corresponding predictions from numerical simulations $\mathbf{g}(\mathbf{m})$, with each datum i yielding a Hessian update $\mathbf{G}_{(i)}^T\mathbf{G}_{(i)}$ as discussed in Section 5.3.2.2 (as well as a gradient update $\mathbf{G}_{(i)}^T\left[\mathbf{d}_{(i)} - \mathbf{g}(\mathbf{m})_{(i)}\right]$); and
2. A parallel reduction operation, merging the individual updates $\mathbf{G}_{(i)}^T\mathbf{G}_{(i)}$ to yield a single estimate of the full Hessian (individual gradient updates are treated similarly).

Our particular implementation adopts a mixed OpenMP/MPI programming model, appropriate for modern multi-core HPC platforms. In particular, the outermost level of parallelism corresponds to MPI tasks, typically distributed among the available compute resources one-to-one with NUMA domains, with additional efforts to ensure strict memory containment and therefore enhance locality when supported. All MPI tasks are equivalent (executing in a SPMD fashion), with the exception that a designated root task spawns a separate coordinator Pthread responsible for work distribution (assigning data to the pool of MPI tasks).

Work is assigned to MPI tasks in blocks (more than one datum) reflecting locality of the underlying data (observations and simulation output) on disk. Each block is processed in parallel by the OpenMP thread team associated with each MPI task (one datum per thread) and occupying the remaining available CPU cores. Up to the limit that the full Hessian estimate $\mathbf{G}^T\mathbf{G}$ can fit in memory (or at least its upper or lower triangular part), we adopt a replicated approach to the reduction of updates. This approach is motivated by the assumption that the underlying merge operation (addition) can be considered associative and commutative for our purposes (not strictly true for floating-point arithmetic). Each MPI task maintains its own copy of the Hessian and gradient, to which the OpenMP thread team applies per-datum updates (protected by a mutex). Once all blocks have been processed, the second level of reduction proceeds by summing all replicated Hessian and gradient copies across MPI tasks. Thereafter, the results are saved to disk, either by a single root task or a collective write via MPI-IO if large enough to warrant it (with collective buffering for improved aggregate throughput on parallel filesystems).

5.3.3 Hybrid inversion in practice

5.3.3.1 Inversion setup and workflow

In our recent global-scale imaging efforts (Lekić and Romanowicz, 2011a; French et al., 2013), the dataset is comprised of tens of thousands of time-discretized seismograms recorded from hundreds of earthquakes distributed around the globe (Fig. 5.2), typically yielding a $\mathbf{d}$ of dimension $\sim 10^7$. The model $\mathbf{m}$ characterizes 3D variations of seismic shear-wave velocity in the earth's mantle, which is expressed in a spline basis of $10^4 - 10^5$ free parameters (see (Lekić and Romanowicz, 2011a; French et al., 2013) for details on model parameterization).

Given an iterative model estimate $\mathbf{m}_k$, we use a spectral finite-element method to compute $\mathbf{g}(\mathbf{m}_k)$ – chosen for its excellent numerical-dispersion behavior and natural treatment of the free-surface boundary condition, among other attractive properties (Komatitsch and Tromp, 2002a; Capdeville et al., 2003). While small in isolation, each occupying 200-300 CPU cores under typical production configurations, these spectral element simulations are numerous: requiring an independent simulation for each earthquake and iterative model estimate. This stands in contrast to the data assimilation and Hessian computation described in Section 5.3.2.3, which is cheap in comparison, is run only once per model iteration, and scales trivially. As noted in Section 5.2, the use of a quickly converging Newton-like optimization scheme leads to a small total number of iterations - typically ~ 10 in practice. For reference, the overall workflow for iterative optimization of $\mathbf{m}$ is summarized in Fig. 5.3.

5.3.3.2 Memory considerations

Because we construct the Hessian estimate $\mathbf{G}^T\mathbf{G}$ directly, the overall space complexity of the replicated approach described in Section 5.3.2.3 is independent of the number of data. Instead, this approach requires $O(N_{\mathbf{m}}^2)$ space per replica, where $N_{\mathbf{m}}$ is in turn dictated by the resolution of the model (hypothetically increasing by a factor of 8 if the nominal resolution is doubled in all three spatial dimensions).

In our most recent work, where $N_{\mathbf{m}} \simeq 2.2 \times 10^5$, the Hessian has become difficult to manipulate – exceeding 90GB in size at single precision (or 180GB at double precision), even when storing only the upper or lower triangular part. As such, the Hessian already exceeds memory resources available on a typical shared-memory HPC compute node, rendering the replicated approach detailed in Section 5.3.2.3 infeasible (though not for the assembly of the misfit gradient vector $\mathbf{G}^T\left[\mathbf{d} - \mathbf{g}(\mathbf{m})\right]$, which requires only $O(N_{\mathbf{m}})$ space per replica). Further, in next-generation global-scale inversions currently in the planning stage, $N_{\mathbf{m}}$ is expected to grow by approximately a factor of 4, chiefly due to doubling resolution in the lateral (rather than radial) dimensions. Concretely, this will yield an $N_{\mathbf{m}}$ exceeding 8.2×10^5, requiring over 1.2TB to store the Hessian (again at single precision) and clearly motivating the development of a scalable *distributed* dense matrix assembly abstraction tailored to this application.

5.4 Abstraction design

Here, we discuss the design and implementation of a distributed dense matrix abstraction, motivated by the discussion in Section 5.3.3.

5.4.1 Requirements

Prior to implementing the distributed abstraction, we carefully enumerated the requirements intrinsic to our application, while also prescribing additional requirements desirable from a usability and scalability perspective. Namely, the resulting abstraction should: **(1)** support distribution schemes common in parallel dense linear algebra (e.g. block-cyclic); **(2)** support updates on strided slices of the distributed matrix, parameterized by associative-commutative operations; **(3)** provide clear consistency guarantees for updates: namely, that each update should appear to take place atomically with respect to all others (as if performed in isolation); **(4)** attain scalability through overlapped communication and computation with minimal synchronization; **(5)** provide a collective, barrier-like "commit" operation, after which all preceding updates are guaranteed to have been applied; and **(6)** readily interoperate with existing OpenMP/MPI codes. While many of these requirements are constrained by the structure of our current production code (e.g. interoperation with MPI and OpenMP), others are based on anticipated functionality. For example, the requirement on underlying distribution scheme reflects a desire to perform in situ parallel linear algebra operations with the resulting Hessian, such as with the PBLAS/ScaLAPACK (Blackford et al., 1997) (avoiding an intermediate phase of parallel I/O before attempting solution of the linear system in eqn 5.2).

5.4.2 Implementation

5.4.2.1 UPC++ – PGAS extensions to C++

Given the requirement to perform remote updates on distributed matrix elements in a scalable manner – namely, with minimal synchronization – we focus on implementation languages supporting the PGAS model. Integration with existing scientific applications that use mixed OpenMP/MPI parallelism, as is the case in our application, makes either UPC (UPC Language Specifications, 2005) or UPC++ (Zheng et al., 2014) an attractive choice, given that both readily interoperate with the former two. Additionally, object orientation in C++ provides a convenient way to build the abstraction, while providing a simplified but expressive interface to the user.

Finally, two particularly novel features of UPC++ make it an especially attractive choice of implementation language. First, UPC++ provides functionality for both one-sided allocation of memory on a remote target, and one-sided (possibly asynchronous) bulk copies – convenient building blocks for synchronization-free transfers of update data. Second, the asynchronous remote task feature of UPC++ provides a novel implementation strategy for encapsulating bulk updates of remote matrix elements in a manner that can be guaranteed to be performed in isolation (asynchronous tasks in UPC++ *can* be configured to executed serially on the target, though this is not an intrinsic limitation imposed by the language). We note that there is overlap between *some* of the high-level functionality described here and a hypothetical implementation based on C++ and one-sided MPI operations supporting associative-commutative merge operators (e.g. `MPI_Accumulate`). However, the level of generality supported by UPC++ stands in stark contrast, permitting the application programmer to fully specify the update logic and then offload it for execution on the target-side in an application-specific manner (addressing specific consistency guarantees, taking advantage of specific optimizations, etc.). We will return to this in Section 5.5.2.

5.4.2.2 Structure and interface

Our abstraction is implemented as a C++ class, with one instance per UPC++ process, and will hereafter be referred to as `ConvergentMatrix` (so-named for the ability to "converge" to its final state in an asynchronous manner). For the use case described above, typical configurations would associate a single UPC++ process with some set of physical CPU cores (generally a NUMA domain to enhance memory locality), responsible for the multi-threaded computations that produce matrix updates, which `ConvergentMatrix` in turn applies globally. The dense-matrix distribution scheme in `ConvergentMatrix` is largely modeled on that of the PBLAS (column-major storage), and parameters thereof (block size, process grid dimensions, leading dimension of local storage, etc.) are passed to `ConvergentMatrix` as template parameters (to encourage elimination of common arithmetic subexpressions). The data type stored by a `ConvergentMatrix` object is also a template parameter, though typically assumed to be `float` or `double`. Thus, the constructor takes the form:

```
ConvergentMatrix<typename T,
                 long NPROW, long NPCOL,
                 long MB, long NB,
                 long LLD>( long m, long n );
```

where template parameter `T` is the matrix data type, `NPROW` and `NPCOL` are the dimensions of the process grid (in PBLAS parlance), `MB` and `NB` are the distribution blocking factors, and `LLD` is the local leading dimension of matrix storage on all processes. In addition, the constructor arguments `m` and `n` are the global dimensions of the distributed matrix. Finally, we also note that the constructor is itself a collective call, during which local storage arrays are initialized and processes exchange configuration information needed in later asynchronous updates (such as remote memory references to each other's local storage, in the form of UPC++ `global_ptr<T>` objects). The public interface to `ConvergentMatrix` consists primarily of three methods:

```
void update( LocalMatrix<T> *Mat,
             long *ix, long *jx );
```

Arguments: `Mat` is a pointer to a `LocalMatrix<T>` object, a light-weight abstraction for non-distributed matrices, and `ix` and `jx` are indexing arrays which map the leading and trailing dimensions of `Mat` into the global matrix.
Behavior: a strided slice update contained in `Mat` will be merged (`+=`) into the distributed matrix asynchronously and atomically at some point after the time of invocation and before a subsequent collective call to `commit` returns.
Implementation: matrix elements contained in `Mat` are binned by target UPC++ process under the given distribution scheme (storing both the elemental update and associated linear indexing on the target). Once a bin becomes sufficiently large (a runtime-tuneable parameter), the bin is *flushed*, whereby the stored updates are applied using an asynchronous remote task scheduled to execute on the target process (see Section 5.4.2.3). To ensure progress on enqueued update tasks, after a tuneable number of bin-flushes, each `ConvergentMatrix` instance will pause to drain its own UPC++ task queue (update tasks for which it is the target).

```
void commit();
```

Behavior: a blocking collective operation that ensures all preceding asynchronous updates to the global matrix have been applied.
Implementation: UPC++ processes first synchronize using the `upcxx::barrier()` collective, at which point each process ensures that all locally initiated remote task invocation have been delivered. Thereafter, each process drains its local queue of remotely initiated asynchronous tasks, and initiates a second `upcxx::barrier()` call before returning, which ensures the above specified behavior that all updates have been applied.

```
T *get_local_data();
```

Behavior: returns a pointer to the local storage associated with the calling UPC++ process under the chosen distribution scheme, appropriate for use with parallel linear algebra libraries or MPI collective IO via a suitable `darray` type.
Implementation: simply returns raw pointer to the underlying local storage array, with the caveat that the underlying data will continue to mutate if further calls to `update` are made while the array is in use.

Together with a small number of methods for run-time tuning (such as the bin size), this interface provides a simple and concise abstraction for the problem considered.

5.4.2.3 The asynchronous update

Possible implementations
As mentioned previously, asynchronous remote tasks are used to perform matrix updates on the target UPC++ process in an isolated manner. Remote tasks in UPC++ may be implemented using the `upcxx::async` function:

```
void async( int t_id,
            event *e )( func, arg1, ... );
```

where `t_id` is the id of the target UPC++ process, `e` is an optional `upcxx::event` object for monitoring completion of asynchronous tasks (supporting a test / wait interface), `func` is the user-defined function to execute (i.e. the update logic), and the remaining arguments are those taken by `func`. In our case, `func` contains logic for initiating or cleaning up after any remote data transfer operations on which it depends, as well as application of the update (optimized loops, etc.), while its arguments consist of references to target-side memory accessed by the underlying update (instance fields of the `ConvergentMatrix` object on the target are not available inside the scope of `func`). Further, all remote updates are registered to a `upcxx::event` object, which is used as an additional check on completion during `commit`. Please see (Zheng et al., 2014) for a more detailed discussion of the functionality and underlying implementation of `upcxx::async`.

Having chosen to use asynchronous remote tasks to perform updates in an isolated manner, one key question is how update information (data and indexing) should be conveyed to the target process. This point deserves careful consideration, as it will have a significant impact on the execution time of the remote tasks. Two illustrative scenarios are described below, both of which employ the same functionality for data transfer to remote processes, but *differ* in how remote memory is managed. Specifically, UPC++ supports bulk one-sided copy functionality via the `upcxx::copy` function:

```
int copy( global_ptr<T> src, global_ptr<T> dst,
          size_t count );
```

where one of `src` or `dst` must be a reference to a local memory location and `count` is the number of elements of type `T` transferred. This call is blocking (though UPC++ supports a non-blocking variant), where a successful return implies the transfer has completed. In addition, there are the `upcxx::allocate` and `upcxx::deallocate` functions for remote allocation and deallocation of remotely accessible memory:

```
global_ptr<T> allocate( int t_id,
                        size_t count );

int deallocate( global_ptr<T> p );
```

where `t_id` is the id of the target process on which `count` elements of type `T` will be allocated. Both calls are blocking and *effectively* one-sided (in contrast to `upcxx::copy`, their implementation requires logic on the target side, though without explicit intervention – we will return to this shortly). With these semantics in mind, the two example scenarios are:

- **Push**: The initiating process could "*push*" the data to the destination process before calling `upcxx::async` to enqueue the update task. This can be achieved with successive calls to `upcxx::allocate` and `upcxx::copy`. After the latter call returns, temporary buffers containing update data on the initiating side may immediately be freed or reused. In this case, `upcxx::global_ptr` objects referring to the already transferred update data will need to be passed as arguments to the asynchronous task.
- **Pull**: The initiating process could store the update data to local (but still remotely addressable) temporary buffers and pass an associated `upcxx::global_ptr` reference to the asynchronous task. The latter must then "*pull*" (again via `upcxx::copy`) the update data *at execution* onto the target process. Importantly, the asynchronous task is then responsible for subsequently calling `upcxx::deallocate` in order to free the temporary buffers on the initiating process.

Evaluation of implementations

There are two metrics by which these proposed communication mechanisms can be easily appraised: safety guarantees regarding memory usage and the "weight" of the asynchronous update task. However, before discussing these points in depth, it is valuable to briefly visit the underlying communications runtime on which UPC++ is implemented, namely GASNet (Bonachea, 2002). GASNet provides a suite of portable communications primitives, able to exploit advanced hardware capabilities of high-performance interconnects (e.g. remote DMA), and is specifically tailored for use by the runtime libraries of PGAS languages (or, more directly, as a compilation target). UPC++ uses GASNet for both its remote memory access capabilities, as well as the GASNet active messages (AM) API for high-performance interrupt or polling-based triggers for one-sided operations that require specific logic to execute on the target (implemented via callback handlers). For example, `upcxx::copy` is based on the one-sided `gasnet_put` and `gasnet_get` primitives, while `upcxx::allocate` or `upcxx::deallocate` require AM-based handlers (to execute memory management operations on the target), as does `upcxx::async` (to enqueue the associated task for later execution on the target). A more detailed discussion of the set of communications primitives supported by GASNet can be found in (Bonachea, 2002). With this distinction between fully one-sided and AM-based operations in mind, a more detailed discussion of the proposed asynchronous update implementations is warranted.

First, while the *pull* procedure requires no two-sided operations within the `update` call on the initiating process, the *push* procedure will block in `upcxx::allocate` until the associated active-messages (AM) handler runs on the target and returns a `upcxx::global_ptr` object to the receiving buffer. Either approach requires at least one immediate duplication of the storage associated with the binned-updates: either a copy retained on the initiating process for asynchronous retrieval by the update task (*pull*), or a copy performed over the network (*push*). Importantly, due to the one-sided nature of `upcxx::copy`, there is no risk in the *push* case that a pause in AM handler progress / network polling on the target will prevent the next step completing: namely the binned-update storage being freed or marked for reuse. Conversely, under the *pull* case, the `upcxx::deallocate` call is remote (freeing memory on the initiating side), and must be serviced by AM. Thus, the window during which multiple copies of the update data can consume memory resources in *push* is determined only by the throughput of `upcxx::copy`, while in *pull* this window could potentially be much larger (in the case of delayed AM handler execution). Second, in order to reason about the cost of pausing to drain the task queue following the bin-flush operation (as described in Section 5.4.2.2), the asynchronous update tasks should be as light-weight as possible. This requirement is clearly satisfied more closely in the *push* case, where no communication is performed within the update task. Thus, the `update` method was designed to follow the *push* procedure.

Finally, from the above discussion it is important to note that UPC++ grants the application programmer not only fine control over the update logic (the task itself), but also the windows of execution during which resources are utilized. This stands in contrast, for example, to MPI, in which case these considerations are often wholly internal, and may be implementation or even vendor dependent.

5.4.2.4 Integration into the production application

In Fig. 5.4, we broadly summarize the new *distributed* implementation of our production application (in contrast to the earlier replicated approach detailed in Section 5.3.2.3). As before, work (partitions of the dataset reflecting filesystem locality and load balance constraints) is distributed among many OpenMP thread teams, generally associated with NUMA domains. Previously, the stream of Jacobian panels produced by the thread team were consumed by a single collocated MPI process, responsible for applying the resulting Hessian (or gradient) updates to the local replicated Hessian (or gradient) instance. This MPI-based consumer has now been replaced with a UPC++ process, responsible for managing an associated instance of `ConvergentMatrix` to which updates are applied (a replicated approach is still used for the gradient vector).

As before, MPI is still used for distribution of work partitions (where a two-sided, request-response coordinator model makes sense), as well as for high-performance collective IO on parallel filesystems. This latter functionality has become increasingly important as the distributed Hessian estimate has grown (currently hundreds of GB, and soon multiple TB – see discussion of problem size in Section 5.5). Further, we note that MPI may also be used, albeit indirectly, through subsequent calls to the PBLAS / ScaLAPACK, in which case it may be possible to avoid an intermediate phase of IO before attempting to solve eqn. 5.2 (this approach is not currently used, however).

5.4.3 Challenges

Here, we discuss two challenges that arose during the development and deployment of `ConvergentMatrix`, along with the particular solutions we adopted.

5.4.3.1 Reasoning about progress

One of the more fundamental challenges encountered in developing `ConvergentMatrix` is reasoning about progress: in terms of both execution of the asynchronous update tasks and remote memory management.

Execution of asynchronous tasks

As noted above, asynchronous task invocation and remote memory management operations in UPC++ both require GASNet to poll the network for new messages on the target side and execute the associated AM handlers. While GASNet implicitly calls `gasnet_AMPoll` (which services the network, triggering the associated handlers) in numerous message-sending operations (Bonachea, 2002), reasoning about where, when, and if at all, additional calls to `gasnet_AMPoll` are necessary is non-trivial (in addition to the limited number thereof internal to UPC++). Indeed, GASNet is specifically designed for network-polling and message-servicing operations to occur at the implementation level of the supported PGAS language, not at the application programmer level.

Further, even when AM handlers for asynchronous tasks are run on the destination process, UPC++ only enqueues these tasks to run. As described above in Section 5.4.2.2, each participating process must periodically ensure that update tasks enqueued by remote processes are executed. Initially, there was no support in UPC++ for querying or draining the local task queue: only `upcxx::progress`, which calls `gasnet_AMPoll` and subsequently executes a single task from the queue. This functionality was added to UPC++ as part of the development of `ConvergentMatrix` and has subsequently been merged into the former as the `peek` and `drain` functions.

Implications for memory management

Pausing to make progress on the enqueued update tasks has additional implications for memory overhead, as the update tasks are responsible for freeing their own yet-to-be-applied update data (see Section 5.4.2.3). Failure to periodically free these buffers can lead to runtime failures on calls to `upcxx::allocate` due to memory exhaustion on the target process. One potential solution would be the addition of a backpressure mechanism: for example, whereby `upcxx::allocate` would be permitted to return a failure code, indicating that it should be retried later on. During the pause, the calling process could spin in `upcxx::progress` to ensure execution of AM handlers and enqueued tasks, though this may not be necessary to ensure progress if other measure are taken (see ***Aiding progress***). This functionality (failure in `upcxx::allocate` without aborting execution) has not been introduced into UPC++, but could easily be in the future.

Aiding progress

We explored two approaches for mitigating the issues raised above. The first approach is simply to set the interval between `upcxx::drain` calls internal to the `update` method to one (described in Section 5.4.2.2), requiring that every round of asynchronous update tasks is accompanied by a call to `upcxx::drain` on the initiating side. Assuming approximate load-balance between `ConvergentMatrix` instances, this should ensure that progress is made on enqueued tasks (and AM handlers) at roughly the same rate they are initiated.

However, this assumption is at odds with the asynchronous design of the abstraction, as well as numerous real-world considerations (imperfect load balance, non-determinism in IO rates, etc.). Though additional `upcxx::drain` calls may be invoked while a UPC++ process waits for new Hessian updates from the OpenMP thread team, there are numerous other operations where such calls cannot easily be interleaved. To this end, we introduced an additional "progress" thread, responsible for periodically invoking `upcxx::drain`. While this solution requires locks to prevent concurrent calls to UPC++ routines that alter the task queue, this critical section of calls within the `update` method is compact, and the additional code complexity is minimal (less than 30 SLOC). We found this approach to be effective at ensuring progress despite the asynchronous nature of update operations, thereby enabling both high update throughput (rapid execution of update tasks) and efficient memory management (requiring a smaller reserved fast segment for GASNet, as well as less chance of memory exhaustion due to other operations using the heap).

5.4.3.2 Unintended synchronicity

Another intriguing challenge was only encountered during benchmark tests (Section 5.5) at high levels of concurrency: namely, network congestion due to large numbers of simultaneous updates (specifically, the `upcxx::copy` phase) against the same target. This particular case can arise when update load and rate are near perfectly balanced, as is possible in the synthetic benchmarks discussed below. In particular, such tests were designed to establish a lower bound on performance by constructing a worst-case scenario – one in which the typically unstructured / asynchronous patterns of updates distributed across the application become synchronous (more like all-to-all operations). A simple and effective strategy to mitigate this issue is randomized target ordering – ensuring that (at least with high probability) no two `ConvergentMatrix` instances will initiate updates (Section 5.4.2.2) on the same set of targets in the same order. Thus, even in the case of unintended synchronicity, the likelihood of a performance bottleneck due to simultaneous updates against a single target is low. An analogous approach has been used to optimize PGAS-based parallel FFT (Nishtala et al., 2009; Almási et al., 2005).

5.5 Evaluation

5.5.1 Scaling

Here, we present scaling results for a range of problem sizes representative of present-day and planned next-generation inversions. To this end, we developed a framework for performing synthetic benchmark experiments modeled closely after run-time behavior of the production application. This tool is a drop-in replacement for the OpenMP thread team in Fig. 5.4, yielding streams of artificial updates with realistically distributed sizes, access patterns (indexing into the distributed Hessian), and production rates. As noted above (Section 5.4.3.2), we configure the framework to produce updates at a nearly uniform rate, thus inducing the worst case simultaneous communication volume and placing a lower-bound on performance (namely, we use the *mean* measured rate for each problem size considered).

These benchmarks are performed on *Edison*, a Cray XC30 at the Department of Energy National Energy Research Scientific Computing Center and our primary production platform. Each *Edison* compute node has 64 GB of memory among two NUMA domains, each associated with a 12-core Intel "Ivy Bridge" processor. There are 5,576 compute nodes in total, linked via a Cray Aries high-speed interconnect, yielding a theoretical peak performance of 2.57 PFLOPS. In all of our scaling experiments, we mimic the distribution of processes / threads seen in the production application: one UPC++ process (and `ConvergentMatrix` instance) per NUMA domain and 8 OpenMP threads performing simulated work (in practice the remaining cores perform separate work distribution or IO tasks). Similar to our production application, we use the GNU Compilers (4.8.2) in all of our tests (optimization: `-O3`). We store all matrix data in single-precision 32-bit `float`, again mimicking the production application (limited by the precision of the seismic data, stored as `float` for compact representation on disk).

5.5.1.1 Strong Scaling

We examine three fixed problem sizes: two borrowed from recent inversions ($N_{\mathbf{m}} = 1.1 \times 10^5$ and 2.2×10^5) and one from a planned next-generation inversion ($N_{\mathbf{m}} = 8.2 \times 10^5$) motivated by doubling the lateral resolution of the former two. These runs are configured to use a 2D block-cyclic distribution scheme (64 × 64 block size), occupying P UPC++ processes (`ConvergentMatrix` instances associated with NUMA domains) for $\sqrt{P} \in \{2, 4, 8, 16, 32\}$, and are representative of production calculations on up 1024 NUMA domains, or 12,288 cores, of *Edison*. We quantify strong scaling in terms of relative parallel efficiency:

$$E_R(P) = \frac{T(P_{min})}{T(P) \cdot P/P_{min}}$$

where $T(P)$ is the time to solution using P processes (elapsed time from thread-team start to when `commit()` returns) and P_{min} corresponds to the *reference* run: the smallest P in the set above at which the problem size considered can be solved (due to memory limitations). We hold the total number of updates N_{up} initiated across all processes fixed at a range of values $N_{up} \in \{4096, 32768, 65536\}$, reflecting present-day and anticipated next-generation inversions, while also allowing us to measure relative efficiencies across three orders of magnitude in core counts by extrapolation. Namely, T scales quasi-linearly with the number of updates initiated by each `ConvergentMatrix` instance. This observation allows us, for example, to infer $T(P = 4, N_{up} = 32768)$ from timing runs completed at $N_{up} = 4096$ by assuming that the former should take approximately 8 times longer than the latter. The ability to extrapolate in this manner is highly valuable, as the $N_{up} = 32768$ case would in reality take prohibitively long to measure for $P = 4$.

To elaborate, our application is partially parallel pipelined: the thread team produces updates in parallel, which are then buffered and consumed by `ConvergentMatrix` for parallel application. Because of this pipelining, there is a fixed non-zero spin-up time (independent of P) at the beginning of each run, during which the thread team is working but has not yet produced work for `ConvergentMatrix`. For small P and fixed N_{up} (many updates per instance), the fraction of T spent in spin-up will be smaller than for larger P (where there are fewer updates per instance, but the spin-up time remains the same). For example, we might expect that the $T(P, N_{up})$ ratio

$$R(P) = \frac{T(P, 32768)}{T(P, 4096)}$$

will be approximately 8 for $P = 4$, but will likely be *less* for $P = 64$ (due to the larger spin-up fraction relative to the time spent performing updates). Therefore, in this example, we can use $R(P = 64)$ to extrapolate a *lower bound* on $T(4, 32768)$ from $T(4, 4096)$ (again, we have not calculated the former, as it would take prohibitively long). Because extrapolation to larger N_{up} in this manner yields an under-estimate of T (when performed with $R(P)$ measured at larger P), this extrapolated value may be used as the reference time $T(P_{min})$ for computing a *lower bound* on $E_R(P)$ at larger P, where test runs for the larger N_{up} are actually feasible.

In Tables 5.1 through 5.3, we present time to solution and $E_R(P)$ for the test runs described above. For all problem sizes and values of P considered, we observe impressive relative speedup and find that $E_R(P)$ remains consistently above 85%. In our particular application, $N_{\mathbf{m}}$ is constrained *a priori* by the physics of wave propagation (namely, the attainable resolution) and held fixed for numerous inversion iterations. Thus, we note that strong scaling, in terms of time to solution, is a critically important axis of evaluation for our application. Further, these tests clearly demonstrate that `ConvergentMatrix` readily scales to anticipated next-generation problem sizes.

Table 5.1: Strong scaling for $N_{\mathbf{m}} = 1.1 \times 10^5$ on up to 12,288 cores of NERSC *Edison*. Values in red are extrapolated (using $R(P = 64) = 7.88$, see text). $E_R(P)$ values are computed relative to the lowest P for each N_{up} (which may have been extrapolated, in which case $E_R(P)$ is a lower bound).

P	**Cores**	N_{up}	$T(P)$ s	$E_R(P)$	N_{up}	$T(P)$ s	$E_R(P)$
4	48	4096	5070.59	100.0%	32768	39948.20	100.0%
16	192	4096	1271.40	99.7%	32768	10016.61	99.7%
64	768	4096	322.24	98.3%	32768	2538.74	98.3%
256	3072	-	-	-	32768	640.96	97.4%
1024	12288	-	-	-	32768	171.68	90.9%

Table 5.2: Strong scaling for $N_{\mathbf{m}} = 2.2 \times 10^5$ on up to 12,288 cores of NERSC *Edison*. Values in red are extrapolated (using $R(P = 64) = 7.80$, see text). $E_R(P)$ values are computed relative to the lowest P for each N_{up} (which may have been extrapolated, in which case $E_R(P)$ is a lower bound).

P	**Cores**	N_{up}	$T(P)$ s	$E_R(P)$	N_{up}	$T(P)$ s	$E_R(P)$
16	192	4096	2318.57	100.0%	32768	18079.84	100.0%
64	768	4096	592.80	97.8%	32768	4622.56	97.8%
256	3072	-	-	-	32768	1173.27	96.3%
1024	12288	-	-	-	32768	321.92	87.7%

Table 5.3: Strong scaling for $N_{\mathbf{m}} = 8.2 \times 10^5$ on up to 12,288 cores of NERSC *Edison*. No extrapolation has been performed. $E_R(P)$ values are computed relative to the lowest P for each N_{up}.

P	**Cores**	N_{up}	$T(P)$ s	$E_R(P)$	N_{up}	$T(P)$ s	$E_R(P)$
256	3072	32768	2399.96	100.0% 16	65536	4703.16	100.0%
1024	12288	32768	703.72	85.3% 32	65536	1279.66	91.9%

5.5.1.2 Weak scaling

For our application, it is difficult to define a meaningful notion of weak scaling, tied to a nominal fixed problem size *per process* while scaling global problem size by enlarging the number of processes. Two natural axes along which to scale global problem size are matrix dimension $N_{\mathbf{m}}$ and total quantity of data N_{up}. Growing $N_{\mathbf{m}}$ while retaining a fixed-size partition of the distributed matrix per process does not retain a fixed per-process problem size, as the dimension of each update must grow accordingly (Section 5.3.2.2). Holding $N_{\mathbf{m}}$ fixed while scaling N_{up} (adding more processes, each performing the same number of updates), does not maintain the same distributed-matrix partition size per process, but does maintain the same update dimension and per-update communication volume.

Among these two options, we believe the second (scaling N_{up}) may be more informative.

Table 5.4: Weak scaling for $N_{\mathbf{m}} = 2.2 \times 10^5$ on up to 12,288 cores of NERSC *Edison*: UPC++ and MPI-based implementations.

			UPC++		**MPI**	
P	**Cores**	N_{up}	$T(P)$ s	$\epsilon(P)$	$T(P)$ s	$\epsilon(P)$
16	192	1024	591.18	1.000	fail	-
64	768	4096	592.50	1.002	1452.24	1.000
256	3072	16384	597.24	1.010	1620.22	1.116
1024	12288	65536	609.96	1.032	3560.28	2.452

Importantly, though the per-update problem size is fixed, the total volume of concurrent communication increases with P, as does the cost of the binning / flushing operation. Further, unlike the fixed total N_{up} runs used in assessing strong scaling, these experiments are comparatively insensitive to the effect of spin-up time fraction (which is the same for all P). Thus, variation in T observed in these runs is primarily sensitive to communication overhead inherent to our implementation, therefore yielding a more informative notion of problem size for evaluating the particular communication model used in `ConvergentMatrix`. In Table 5.4, we illustrate weak scaling for an $N_{\mathbf{m}}$ of 2.2×10^5 in terms of the ratio $\epsilon(P) = T(P)/T(P_{min})$ for a range P and a fixed number of updates per-process. We find that $\epsilon(P)$ stays very close to unity for our UPC++-based solution over a wide range of core counts (192 to 12288) and corresponding total numbers of updates (1024 to 65535) – reaching only 3% communication-induced overhead at the highest level of concurrency / dataset problem size, and indicative of near-total overlap of computation and communication.

5.5.2 Comparison with MPI

For comparison, we designed a second implementation, based on MPI-3 one-sided remote memory access (RMA) operations. While the shortcomings of MPI-2 for emulating PGAS-like functionality are well known (Bonachea and Duell, 2004), MPI-3 largely addresses these issues. The particular semantics of `MPI_Accumulate` fit well with our requirements: concurrent, element-wise atomic updates to remote memory using predefined commutative merge operations (e.g. `MPI_SUM`). These are *weaker* atomicity guarantees than the UPC++ version (which applies the *entire* update atomically), that we hoped could lead to performance advantages when using MPI-3 RMA.

In our re-implementation of `ConvergentMatrix`, we use per-target MPI window objects, allowing fine control over access epochs for distributed-matrix storage arrays. Access is managed with the passive `MPI_Win_lock` / `unlock` pattern, requiring minimal synchronization and no target-side exposure management. Locks are acquired for each accumulate call in exclusive mode, after initial tests found performance advantages over shared locks, possibly due to implicit coordination between concurrent updates and poor locally when updates from distinct origins are interleaved. We also tried a single passive access epoch with shared-mode locks, i.e. opened by the first `update` call and closed by `commit`, in conjunction with per-accumulate `MPI_Win_flush_local` calls, but again found performance poorer than the exclusive-lock approach. Finally, because individual binned updates are arbitrarily structured, we define per-update MPI *indexed* derived types.

As noted in Section 5.5.1.2, our weak-scaling tests are particularly sensitive to the volume of concurrent communication, and thus provide a useful framework for assessing different communication models. We repeated these tests for the MPI-based implementation, using the same compiler configuration and Cray MPT 6.2.0 (based on MPI-3 compliant MPICH 3.0.3). These *partial* results are shown in Table 5.4. Namely, the use of a 32-bit `int` for window offsets (indexing), defined in the MPI standard, severely limits window size and places a lower bound on P for a given $N_{\mathbf{m}}$ (here, the $P = 16$ case leads to overflow). This is not a problem for the UPC++ implementation, as element indices may be parameterized in any integer type (default: 64-bit). In general, we find that time to solution is larger than that seen for UPC++ and weak scaling is comparatively poor, with 11% performance degradation between the $P = 64$ and 256 cases, and further degradation to 145% of the $P = 64$ case by $P = 1024$. It is clear that there are both functionality and scalability / performance advantages to the UPC++-based implementation.

We further believe that there is little advantage to choosing MPI from a programmer productivity perspective. In terms of code complexity, UPC++ and MPI require analogous initialization steps (exchanging `upcxx::global_ptr` references vs. collective window creation), and similar quantites of code are needed to implement updates with `upcxx::async` vs. MPI RMA (92 and 75 SLOC, respectively). While UPC++ required additional care in reasoning about progress (Section 5.4.3), it was considerably simpler to reason about performance (relative to MPI, where details critical to debugging comparatively poor performance are hidden in the runtime). Indeed, examining the `MPI_Accumulate` implementation in foMPI (Gerstenberger et al., 2013), a highly optimized RDMA-aware MPI-3 implementation, the style of bulk floating-point accumulate that we require involves multiple phases of data-movement (target-window lock, RDMA-get, origin-side accumulate, RMDA-put, unlock), in contrast to our *single* phase. Further, this limitation has its roots in design of the API itself, which encourages a truly passive target (e.g. avoiding extensive target-side buffering in the runtime, which can lead to memory contention). Thus, it is clear that for certain types of accumulate-like operations (such as our remote matrix updates) the parallel programming model exposed by UPC++ has additional advantages owing to its generality.

5.6 Scientific results: Whole-mantle imaging

Our new abstraction for scalable distributed matrix assembly has enabled us to solve problems that could not have been attempted using our earlier replicated implementation described in Section 5.3.2.3. In our earlier work (French et al., 2013), we focused on imaging seismic shear-wave velocity structure of the earth's upper mantle and transition zone (≤ 800 km below the surface) – indeed, at the limits of feasibility under the replicated approach. Our waveform inversion technique revealed never before seen "fingers" of low seismic veloc-

Table 5.5: Time breakdowns for a single whole-mantle inversion iteration (Section 5.6), showing the order of computational phases (Fig. 5.3), as well as the number of runs per phase, number of cores per run, typical wallclock times, total CPU hours, and production platform at NERSC. Runs within each phase or sub-phase (i.e. Ia and Ib) may happen concurrently. Phases: *SEM*. forward modeling of the wavefield using the spectral element method (two frequency ranges / meshes); *NACT*. Hessian estimation using NACT, with distributed assembly using `ConvergentMatrix` (12 data sub-sets in two frequency ranges); *GN*. Gauss-Newton update system assembly and factorization (eqn. 5.2), yielding model updates (additional runs for different configurations of the prior).

Phase	**Runs**	**Cores**	**Wallclock**	**CPU Hours**	**Platform**
Ia *SEM*	273	144	8 hr	314,496	XE6 *Hopper*
Ib *SEM*	273	288	5 hr	396,000	XE6 *Hopper*
IIa *NACT*	9	3,072	0.5 - 1 hr	17,664	XC30 *Edison*
IIb *NACT*	3	3,072	3 hr	21,504	XC30 *Edison*
III *GN*	$\geq$2	2,304	1 hr	$\geq$4,608	XC30 *Edison*

ities in the upper-mantle beneath the world's ocean basins (Fig. 5.5a). While the images also suggested a connection between these structures and columnar low-velocity features extending into the lower mantle, the depth range of our inversion was limited.

To more fully examine the interactions between different scales of convective phenomena in the upper and lower mantle, with implications for the dynamics of the system as a whole, we have now moved on to whole-mantle imaging (to the core-mantle boundary at 2891 km depth). Thus far, we have performed three inversion iterations (eqn. 5.2) at the whole-mantle scale ($N_{\mathbf{m}} = 2.2 \times 10^5$) using a growing dataset of surface-wave and higher-frequency body-wave waveforms to attain better resolution (Fig. 5.2).

In Table 5.5, we show time breakdowns for the major computational phases (Fig. 5.3) of a single inversion iteration at this scale. The most expensive phase is clearly SEM simulation: requiring over 700,000 CPU hours per iteration and highlighting why a quickly converging optimization scheme is so critically important (to reduce the total number of iterations). Thus, while Hessian estimation is not the major computational bottleneck for our inversion (and accounts for only $\sim$5% of the total CPU time), the *feasibility* of the inversion is directly tied to our ability to construct the distributed Hessian in a scalable manner from parallel computations. Further, this functionaly – which is enabled by the `ConvergentMatrix` abstraction – will only become more important as our inversion progresses to higher resolution and the dimension of the Hessian grows commensurately.

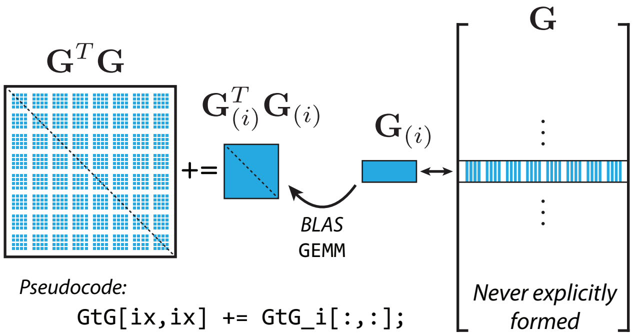
The current whole-mantle inversion is progressing quickly, owing largely to the rapid convergence of our Newton-like model optimization scheme. Our preliminary results are already yielding intriguing new images of coupling between different scales of convection in the earth's upper and lower mantle, as illustrated in Fig. 5.5b. We will soon complete our initial whole-mantle inversion, and will prepare our results for publication. Concurrently, we will start exploring next-generation, higher-resolution imaging characterized by the largest scales of problem size discussed in Section 5.5.1.1.

5.7 Conclusions

Here, we presented the design and implementation of a distributed matrix assembly abstraction for our physics-based Hessian estimation code. This approach not only allows us to tackle previously intractable current-generation inversions, but also scales to the anticipated next generation of problem size. Our solution was made possible by the specific combination of PGAS functionality provided by UPC++, particularly remote memory management and asynchronous task execution, all while available from a "familiar" language (C++) and readily interoperable with MPI and OpenMP-based subsystems of the production code. We observed impressive scaling behavior based on synthetic benchmark experiments under realistic, scientifically meaningful configurations. We also found that our solution based on UPC++ has clear performance and functionality advantages over one based on MPI-2 one-sided, while introducing little additional complexity. Finally, we again note that this particular approach is quite general, and should be applicable to any similarly structured dense-matrix assembly problem arising in an HPC environment.

In the immediate future, our primary focus is the completion of our current whole-mantle imaging work, followed in the longer term by higher-resolution global inversions (similar in scale to the largest scaling experiments discussed above). We will also explore further stability and usability improvements to `ConvergentMatrix` as our use-case evolves. For example, the addition of a *backpressure* mechanism for remote memory allocation (as posed in Section 5.4.3) would allow `ConvergentMatrix` to handle memory pressure gracefully via a fail-wait-retry pattern. This could be achieved either by minor modifications to the behavior of `upcxx::allocate` in the UPC++ runtime, or though globally visible (in the PGAS sense) accounting of available memory resources within `ConvergentMatrix`, consulted before attempting remote allocation. In addition, *compaction* (summation) of repeatedly "hit" elements in the binned updates may allow us to reduce communication volume (as updates are assumed associative). For speed, update bins are currently implemented as immutable (append-only) indexed arrays, but a pre-communication sort / compaction phase, for example, may be valuable if repeated hits become more common in our workloads.

The C++ code associated with the `ConvergentMatrix` abstraction, which remains under active development, is publicly available from https://github.com/swfrench/convergent-matrix.



```
GtG[ix,ix] += GtG_i[:,:];
```

Figure 5.1: A schematic illustration of the strided-slice update operation described in the text.

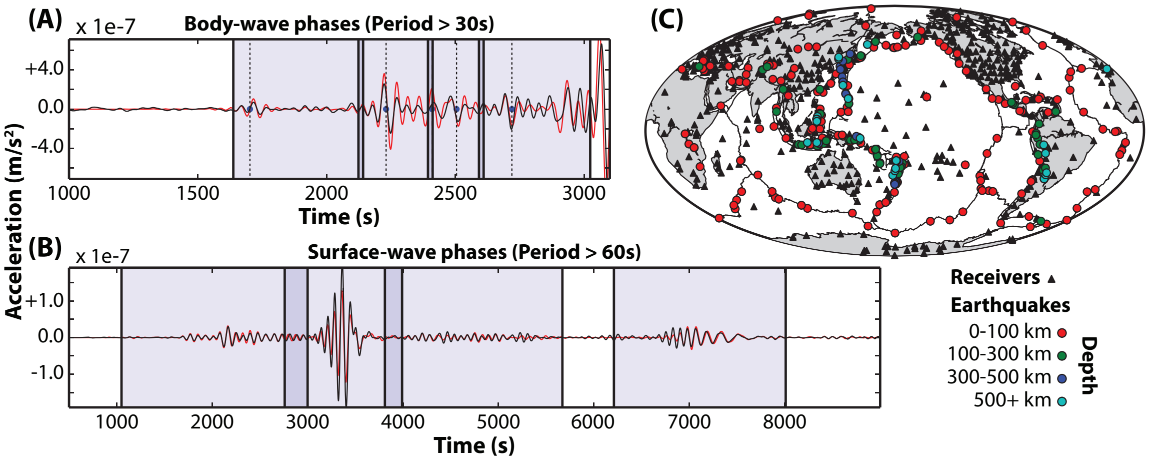


Figure 5.2: An illustration of the types and quantities of data used in our global inversion. Continuous windows (blue shading) are selected from time-domain seismic waveforms in different frequency bands (A and B), allowing us optimize our sensitivity to the earth's interior structure (observations are shown in black, spectral element predictions from the seismic model SEMum2 (French et al., 2013) are shown in red). We also select seismic sources and receivers (C) so as to provide as uniform illumination of the interior as possible (here, showing our dataset of 273 earthquakes and over 500 seismic receivers).

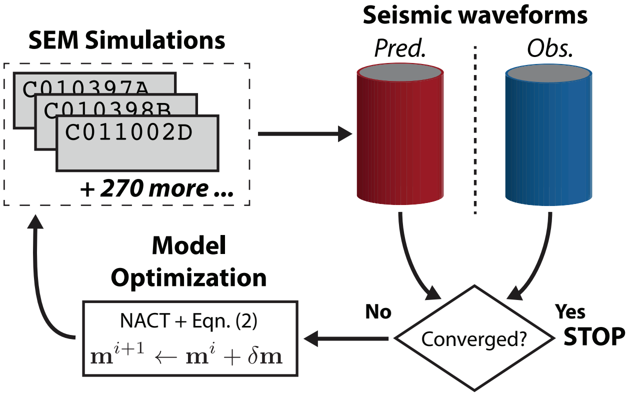
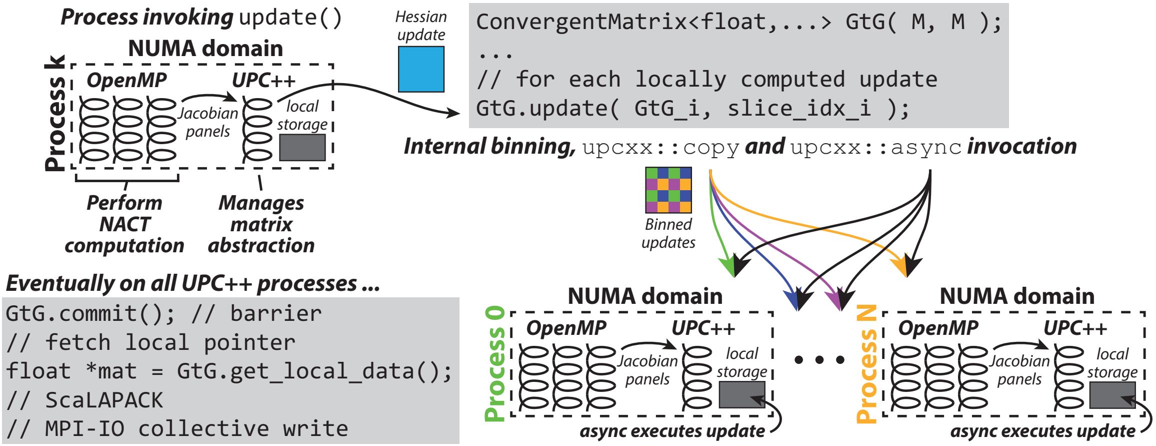


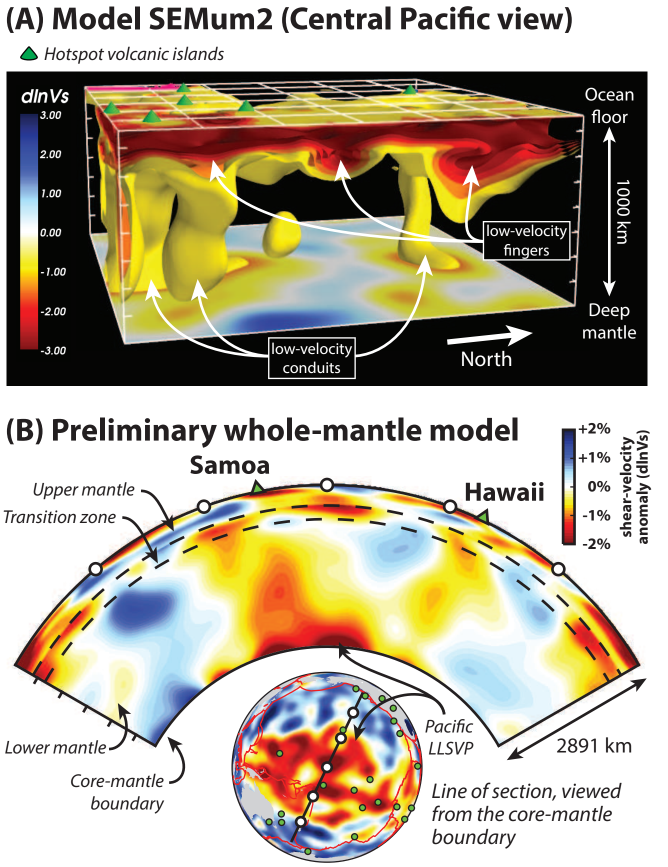
Figure 5.3: An overview of the iterative waveform inversion procedure described in the text. Spectral element (SEM) simulations are used to predict waveforms for the current model estimate $\mathbf{m}_i$, which are in turned compared to the corresponding observed data. If the waveforms agree sufficiently well, the model is considered to have converged with respect to the chosen dataset. Otherwise, the model is updated by solving eqn. 5.2 and the process repeats.



```
ConvergentMatrix<float,...> GtG( M, M );
...
// for each locally computed update
GtG.update( GtG_i, slice_idx_i );
```

```
GtG.commit(); // barrier
// fetch local pointer
float *mat = GtG.get_local_data();
// ScaLAPACK
// MPI-IO collective write
```

Figure 5.4: A schematic illustration of how the `ConvergentMatrix` abstraction is used in our production application, focused on the path taken by a single Hessian update and highlighting the roles of different coexisting parallel programming models / tools (UPC++, OpenMP, and MPI).



(A) Model SEMum2 (Central Pacific view)

(B) Preliminary whole-mantle model

Figure 5.5: Full-waveform images of shear-wave velocity (V_S) structure in the earth's mantle. Our previous model SEMum2 (French et al., 2013) revealed finger-like low V_S anomalies in the upper mantle, accompanied by conduit-like structures extending below (A). Enabled by the matrix assembly abstraction described here, our preliminary whole-mantle model (B) sheds further light on the origins of these structures, and more broadly the interactions between different scales of convective phenomena – linking surface expression of these processes (hotspot volcanic islands like Samoa and Hawaii), to the well known large low shear-wave velocity provinces (LLSVPs) in the deep mantle.

Chapter 6

Conclusions

6.1 Summary of contributions

Here, we have adapted and extended the hybrid waveform inversion technique, which combines the accuracy of the SEM with the efficiency and rapid convergence of an NACT-based Gauss-Newton optimization scheme, to image whole-mantle radially anisotropic shear-velocity structure. Our imaging work progressed in three phases, intended to facilitate progressive interpretation of model structure as it developed.

1. We began in Chapter 2 by presenting SEMum2, a second-generation SEM-based model focused on the upper mantle and transition zone, and described both its development and general properties. We detailed the novel crustal modeling technique that we used to reduce the cost of our SEM simulations and how it fits into the hybrid inversion framework. Further, we drew comparisons between SEMum2 structure and that seen in regional-scale models, and discovered an impressive correspondence in the amplitude of shear-velocity heterogeneity as well as its distribution.

2. In Chapter 3, we took a more detailed look at SEMum2 structure beneath the ocean basins – namely, the low-velocity "fingers" (LVFs) and quasi-vertical conduit-like features. We discussed the correspondence between the LVFs and other geophysical observations, including absolute plate motions and the geoid, as well as potential causative dynamics. In addition, we provided an extensive quantification of model uncertainties, particularly geared toward evaluating the robustness of the LVF and conduit-like structures.

3. Finally, in Chapter 4, we presented our model of whole-mantle radially anisotropic shear-velocity structure – notably, the first ever derived using SEM-based waveform inversion. We drew connections between the structures seen here and in the earlier model SEMum2, and discussed implications for the nature of mantle convection. Further, we provided a detailed road-map for our ongoing model analysis, again largely

focused on uncertainty estimation. The latter efforts will be completed soon, and these results submitted for publication.

In addition to communicating the results of our tomographic imaging, we also featured some of the computational aspects of our work. Namely, while extending our inversion approach to the whole-mantle case, we encountered computational hurdles associated with distributed assembly of the NACT-based Hessian matrix. We presented our solution, which takes advantage of the PGAS-model functionality provided by UPC++, in Chapter 5 in the form of a novel matrix abstraction geared toward distributed assembly in HPC environments. While motivated by our particular application, this model of asynchronous distributed matrix assembly should readily translate to other problem domains with similar requirements.

Finally, it is important to emphasize that our tomographic imaging results have already found use in other studies. For example, Yuan et al. (2014) adopted both the tools developed here for efficient crustal modeling in hybrid inversions, as well as the SEMum2 model itself as a starting model, for their radially anisotropic full-waveform model of North America. These results were recently published in *Earth and Planetary Science Letters*. The SEMum2 model is also featured, again as a starting model, in preliminary work constraining global attenuation (Q) structure – an application area where a well-constrained elastic model is critical to interpreting the amplitudes of seismic waveforms. Finally, SEMum2 has also found use in recent efforts to characterize the spectral properties of seismic heterogeneity from tomographic models, and compares favorably with recent high-resolution full-waveform (adjoint state) models of the European region (M. Meschede, *personal communication*).

6.2 Future directions

Our emphasis throughout this work has been on a progressive approach – not only to keep the inverse problem well-posed and avoid becoming trapped in local minima, but also to use available computational resources wisely and learn something about the earth in each phase of the inversion. This consideration motivated us to take great care in selecting the frequency bands used in our work: first, extending only to 60 s period for upper-mantle and transition-zone imaging, and later down to 32 s for the whole mantle. In this vein, the logical next step in the continuation of this work is extension to shorter-period modeling. Not only will higher-frequency seismic data enable future inversions to attain higher resolution, it will also allow them to constrain V_P structure. Importantly, the ability to invert independently for V_P may enable inferences regarding the presence of compositional heterogeneity (based upon the manner in which V_P and V_S co-vary, or fail to).

In order to take full advantage of the potential for higher resolution imaging, one must also adopt a refined model parameterization. As already shown in Chapter 5, the NACT-based Hessian-estimation phase of our hybrid inversion technique is ready to scale to this next generation of problem size. Thus, in the near term, it may be advantageous to refine the parameterization, while progressively relaxing the permissible resolution through the model covariance operator as warranted, similar to the technique proposed by Tarantola and Nercessian (1984). As discussed in Chapter 2, over-parameterization in this manner also helps to protect the inversion from Gibbs-like distortion that would otherwise be encountered when approaching the spectral limits of the spline basis.

Another potential avenue for continued work is to revisit the upper mantle with the goal of inverting for global *azimuthal* anisotropy structure. While SEM-based modeling of azimuthal anisotropy has been performed successfully in the European region (Zhu and Tromp, 2013), it has not yet been applied at the global scale. Further, such an inversion would also likely call for the use of shorter-period surface-wave data (for example, initially down to 40 s in lieu of the current 60 s), in order to examine fine depth-layering of azimuthal anisotropy (e.g. Yuan and Romanowicz, 2010).

Finally, and perhaps one of the most ambitious among future directions, is the possibility of merging the hybrid and adjoint inversion techniques – namely, by combining the waveform-misfit gradient calculated to great accuracy with the adjoint state method and the Hessian computed using NACT. Importantly, while this approach would require at least 2x the number of SEM simulations per inversion iteration (as is typical of adjoint inversions) relative to the previous hybrid approach, it would still maintain the Newton-like convergence rate of the latter. We note that this novel hybrid-adjoint technique is currently ready for test application at the regional scale – again focused on North America and using SEMum2 as its start model.

Bibliography

Almási, G., Archer, C., Erway, C., Heidelberger, P., Martorell, X., Moreira, J., Steinmacher-Burow, B., and Zheng, Y. (2005). Optimization of MPI Collective Communication on BlueGene/L Systems. In *Proceedings of the 19th Annual International Conference on Supercomputing*.

Backus, G. (1962). Long-wave elastic anisotropy produced by horizontal layering. *Journal of Geophysical Research*, 67(11):4427–4440.

Ballmer, M. D., Conrad, C. P., Smith, E. I., and Harmon, N. (2013). Non-hotspot volcano chains produced by migration of shear-driven upwelling toward the East Pacific Rise. *Geology*, 41:479482.

Ballmer, M. D., van Hunen, J., Ito, G., Bianco, T. A., and Tackley, P. J. (2009). Intraplate volcanism with complex age-distance patterns: A case for small-scale sublithospheric convection. *Geochem. Geophys. Geosyst.*, 10:GC002386.

Barmin, M., Ritzwoller, M., and Levshin, A. (2001). A fast and reliable method for surface wave tomography. In *Monitoring the Comprehensive Nuclear-Test-Ban Treaty: Surface Waves*, pages 1351–1375. Springer.

Bassin, C., Laske, G., and Masters, G. (2000). The current limits of resolution for surface wave tomography in North America. *EOS. Trans. Am. geophys. Un.*, 81:1351–1375.

Becker, T. and Boschi, L. (2002). A comparison of tomographic and geodynamic mantle models. *Geochem. Geophys. Geosyst.*, 3(1):1351–1375.

Becker, T., Kustowski, B., and Ekström, G. (2008). Radial seismic anisotropy as a constraint for upper mantle rheology. *Earth Planet. Sci. Lett.*, 267(1):213–227.

Bedle, H. and van der Lee, S. (2009). S velocity variations beneath North America. *J. Geophys. Res.*, 114(B7):B07308.

Blackford, L. S., Choi, J., Cleary, A., D'Azevedo, E., Demmel, J., Dhillon, I., Dongarra, J., Hammarling, S., Henry, G., Petitet, A., Stanley, K., Walker, D., and Whaley, R. C. (1997). *ScaLAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Bonachea, D. (2002). GASNet specification, v1.1. Technical Report UCB/CSD-02-1207, Computer Science Division (EECS), UC Berkeley.

Bonachea, D. and Duell, J. (2004). Problems with using MPI 1.1 and 2.0 as compilation targets for parallel language implementations. *Int. J. High Performance Computing and Networking*, 1:91–99.

Bozdağ, E. and Trampert, J. (2008). On crustal corrections in surface wave tomography. *Geophys. J. Int.*, 172:1066–1082.

Bozdağ, E., Zhu, H., Peter, D., and Tromp, J. (2011 Fall Meeting). Towards Global Adjoint Tomography, Abstract Abstract S41A-2167. San Francisco, CA. AGU.

Brocher, T. (2005). Empirical Relations between Elastic Wavespeeds and Density in the Earth's Crust. *Bull. seism. Soc. Am.*, 95:2081–2092.

Cammarano, F. and Romanowicz, B. (2007). Insights into the nature of the transition zone from physically constrained inversion of long-period seismic data. *Proc. Natl. Acad. Sci.*, 104:9139–9144.

Cammarano, F., Romanowicz, B., Stixrude, L., Lithgow-Bertelloni, C., and Xu, W. (2009). Inferring the thermochemical structure of the upper mantle from seismic data. *Geophys. J. Int.*, 179:11691185.

Capdeville, Y., Chaljub, E., Vilotte, J.-P., and Montagner, J.-P. (2003). Coupling the spectral element method with a modal solution for elastic wave propagation in global earth models. *Geophys. J. Int.*, 152:34–67.

Capdeville, Y., Gung, Y., and Romanowicz, B. (2005). Towards global earth tomography using the spectral element method: a technique based on source stacking. *Geophys. J. Int.*, 162:541–554.

Capdeville, Y. and Marigo, J.-J. (2007). Second order homogenization of the elastic wave equation for non-periodic layered media. *Geophys. J. Int.*, 170:823–838.

Chaljub, E., Capdeville, Y., and Vilotte, J.-P. (2003). Solving elastodynamics in a fluid-solid heterogeneous sphere: a parallel spectral element approximation on non-conforming grids. *J. Comp. Phys.*, 187:457–491.

Christie, D. M., West, B. P., Pyle, D. G., and Hanan, B. B. (1998). Chaotic topography, mantle flow and mantle migration in the AustralianAntarctic discordance. *Nature*, 394:637644.

Colli, L., Fichtner, A., and Bunge, H.-P. (2013). Full waveform tomography of the upper mantle in the South Atlantic region: Imaging a westward fluxing shallow asthenosphere? *Tectonophysics*, Published online 27 June 2013.

Cottaar, S. and Romanowicz, B. (2012). An unsually large ULVZ at the base of the mantle near Hawaii. *Earth Planet. Sci. Lett.*, 355:213–222.

Courtillot, V., Davaille, A., Besse, J., and Stock, J. (2003). Three distinct types of hotspots in the Earths mantle. *Earth Planet. Sci. Lett.*, 205(3):295–308.

Dahlen, F. and Nolet, G. (2005). Comment on 'On sensitivity kernels for 'wave-equation' transmission tomography' by de Hoop and van der Hilst. *Geophys. J. Int.*, 163:949–951.

Davaille, A. (1999). Simultaneous generation of hotspots and superswells by convection in a heterogeneous planetary mantle. *Nature*, 402(6763):756–760.

Davaille, A., Stutzmann, E., Silveira, G., Besse, J., and Courtillot, V. (2005). Convective patterns under the Indo-Atlantic "box". *Earth Planet. Sci. Lett.*, 239:233–252.

de Hoop, M. and van der Hilst, R. (2005). On sensitivity kernels for 'wave-equation' tomography. *Geophys. J. Int.*, 160:621–633.

Dorman, J., Ewing, M., and Oliver, J. (1960). Study of shear-velocity distribution in the upper mantle by mantle Rayleigh waves. *Bull. seism. Soc. Am.*, 50:87–115.

Durek, J. J. and Ekström, G. (1996). A Radial Model of Anelasticity Consistent with Long-Period Surface-Wave Attenuation. *Bull. seism. Soc. Am.*, 86:144–158.

Dziewonski, A. (2005). The robust aspects of global seismic tomography. *Plates. Plumes. and Paradigms*, 86:144–158.

Dziewonski, A. and Anderson, D. (1981). Preliminary reference Earth model. *Phys. Earth Planet. Inter.*, 25:297–356.

Dziewonski, A., Hager, B., and O'Connell, R. (1977). Large-scale heterogeneities in the lower mantle. *J. Geophys. Res.*, 82:239–255.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Stat.*, pages 586–596.

Efron, B. and Tibishirani, R. J. (1991). *An Introduction to Bootstrap.* Chapman and Hall, New York.

Ekström, G. (2011). A global model of Love and Rayleigh surface wave dispersion and anisotropy. *Geophys. J. Int.*, 187:25–250.

Ekström, G. and Dziewonski, A. (1998). The unique anisotropy of the Pacific upper mantle. *Nature*, 394:168–172.

Epanomeritakis, I., Akcelik, V., Ghattas, O., and Bielak, J. (2008). A Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion. *Inverse Problems*, 24:034015.

Ferreira, A., Woodhouse, J., Visser, K., and Trampert, J. (2010). On the robustness of global radially anisotropic surface wave tomography. *J. Geophys. Res.*, 115:168–172.

Fichtner, A. and Igel, H. (2008). Efficient numerical surface wave propagation through the optimization of discrete crustal models - a technique based on non-linear dispersion curve matching (DCM). *Geophys. J. Int.*, 173:519–533.

Fichtner, A., Kennett, B., Igel, H., and Bunge, H. (2009). Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods. *Geophys. J. Int.*, 179:1703–1725.

Fichtner, A. and Trampert, J. (2011). Resolution analysis in full waveform inversion. *Geophys. J. Int.*, 187:1604–1624.

Fletcher, R. and Reeves, C. (1964). Function minimization by conjugate gradients. *Comput. J.*, 7:149–154.

French, S., Lekić, V., and Romanowicz, B. (2013). Waveform Tomography Reveals Channeled Flow at the Base of the Oceanic Asthenosphere. *Science*, 342:227–230.

Friederich, W. (2003). The S-velocity structure of the East Asian mantle from inversion of shear and surface waveforms. *Geophys. J. Int.*, 153:88–102.

Fukao, Y. and Obayashi, M. (2013). Subducted slabs stagnant above, penetrating through, and trapped below the 660 km discontinuity. *J. Geophys. Res.*, 118(11):5920–5938.

Gerstenberger, R., Besta, M., and Hoefler, T. (2013). Enabling highly-scalable remote memory access programming with MPI-3 one sided. In *Proceedings of SC13: International Conference for High Performance Computing, Networking, Storage and Analysis*, page 53. ACM.

Gripp, A. E. and Gordon, R. G. (2002). Young tracks of hotspots and current plate velocities. *Geophys. J. Int.*, 150:321–361.

Gu, Y., Dziewonski, A., and Ekstrom, G. (2001). Preferential detection of the Lehmann discontinuity beneath continents. *Geophys. Res. Lett.*, 28:4655–4658.

Harmon, N., Forsyth, D., and Weeraratne, D. (2009). Thickening of young Pacific lithosphere from high-resolution Rayleigh wave tomography: a test of the conductive cooling model. *Earth Planet. Sci. Lett.*, 278:96–106.

Harmon, N., Forsyth, D. W., Weeraratne, D. S., Yang, Y., and Webb, S. C. (2011). Mantle heterogeneity and off axis volcanism on young Pacific lithosphere. *Earth Planet. Sci. Lett.*, 311:306–315.

Hayn, M., Panet, I., Diament, M., Holschneider, M., Mandea, M., and Davaille, A. (2012). Wavelet-based directional analysis of the gravity field: evidence for large-scale undulations. *Geophys. J. Int.*, 189:14301456.

Hillier, J. K. and Watts, A. B. (2004). "Plate-like" subsidence of the East Pacific RiseSouth Pacific superswell system. *J. Geophys. Res.*, 109:B10102.

Hirschmann, M. M. (2000). Mantle solidus: Experimental constraints and the effects of peridotite composition. *Geochem. Geophys. Geosyst.*, 1:GC000070.

Hofmann, A. W. (2007). Sampling Mantle Heterogeneity through Oceanic Basalts: Isotopes and Trace Elements. In Holland, H. D. and Turekian, K. K., editors, *Treatise on Geochemistry*, pages 1–44. Elsevier, New York.

Hoink, T., Lenardic, A., and Richards, M. (2012). Depth-dependent viscosity and mantle stress amplification: implications for the role of the asthenosphere in maintaining plate tectonics. *Geophys. J. Int.*, 191:30–41.

Houser, C., Masters, G., Shearer, P., and Laske, G. (2008). Shear and compressional velocity models of the mantle from cluster analysis of long-period waveforms. *Geophys. J. Int.*, 174:195–212.

Husson, L. and Conrad, C. (2012). On the location of hotspots in the framework of mantle convection. *Geophys. Res. Lett.*, 39(17).

Karato, S., Jung, H., Katayama, I., and Skemer, P. (2008). Geodynamic significance of seismic anisotropy of the upper mantle: new insights from laboratory studies. *Annu.Rev.Earth planet Sci.*, 36:195–212.

Katzman, R., Zhao, L., and Jordan, T. H. (1998). High-resolution, two-dimensional vertical tomography of the central Pacific mantle using ScS reverberations and frequency-dependent travel times. *J. Geophys. Res.*, 103:1793317971.

Komatitsch, D. and Tromp, J. (2002a). Spectral-element simulations of global seismic wave propagation - I. Validation. *Geophys. J. Int.*, 149:390–412.

Komatitsch, D. and Tromp, J. (2002b). Spectral-element simulations of global seismic wave propagation - II. Three-dimensional models, oceans, rotation and self-gravitation. *Geophys. J. Int.*, 150:303–318.

Komatitsch, D. and Vilotte, J.-P. (1998). The Spectral Element Method: An Efficient Tool to Simulate the Seismic Response of 2D and 3D Geological Structures. *Bull. seism. Soc. Am.*, 88:368–392.

Korenaga, J. and Jordan, T. (2004). Physics of multiscale convection in earth's mantle: Evolution of sublithospheric convection. *J. Geophys. Res.*, 109(B01):405.

Kreemer, C. (2009). Absolute plate motions constrained by shear wave splitting orientations with implications for hot spot motions and mantle flow. *J. Geophys. Res.*, 114:B10405.

Kustowski, B., Ekström, G., and Dziewonski, A. (2008a). Anisotropic shear-wave velocity structure of the Earth's mantle: a global model. *J. Geophys. Res.*, 113:303–318.

Kustowski, B., Ekström, G., and Dziewonski, A. (2008b). The shear-wave velocity structure in the upper mantle beneath Eurasia. *Geophys. J. Int.*, 174:978–992.

Laske, G., Markee, A., Orcutt, J. A., Wolfe, C. J., Collins, J. A., Solomon, S. C., Detrick, R. S., Bercovici, D., and Hauri, E. H. (2011). Asymmetric shallow mantle structure beneath the Hawaiian Swell-evidence from Rayleigh waves recorded by the PLUME network. *Geophys. J. Int.*, 187:17251742.

Laske, G. and Masters, G. (1997). Constraints on global phase velocity maps from long-period polarization data. *J. Geophys. Res.*, 101:16,05916,075.

Lekić, V., Cottaar, S., Dziewonski, A., and Romanowicz, B. (2012). Cluster analysis of global lower mantle heterogeneity: A new class of structure and implications for chemical heterogeneity. *Earth Planet. Sci. Lett.*, 182:265–278.

Lekić, V., Panning, M., and Romanowicz, B. (2010). A simple method for improving crustal corrections in waveform tomography. *Geophys. J. Int.*, 182:265–278.

Lekić, V. and Romanowicz, B. (2011a). Inferring upper-mantle structure by full waveform tomography with the spectral element method. *Geophys. J. Int.*, 185:799–831.

Lekić, V. and Romanowicz, B. (2011b). Tectonic regionalization without a priori information: A cluster analysis of upper mantle tomography. *Earth Planet. Sci. Lett.*, 308:151–160.

Lévêque, J., Rivera, L., and Wittlinger, G. (1993). On the use of the checker-board test to assess the resolution of tomographic inversions. *Geophys. J. Int.*, 115:313–318.

Li, X. and Romanowicz, B. (1995). Comparison of global waveform inversions with and without considering cross-branch modal coupling. *Geophys. J. Int.*, 121:695–709.

Li, X. and Romanowicz, B. (1996). Global mantle shear velocity model developed using nonlinear asymptotic coupling theory. *J. Geophys. Res.*, 101:245–22.

Liu, Q., Polet, J., Komatitsch, D., and Tromp, J. (2004). Spectral-element moment tensor inversions for earthquakes in southern California. *Bull, Seism. Soc. Am.*, 94(5):1748–1761.

MacQueen, J. (1967). *in Proc. Fifth Berkeley Symp. on Math. Statist. and Prob*, 1:281297.

Marone, F., Gung, Y., and Romanowicz, B. (2007). Three-dimensional radial anisotropic structure of the North American upper mantle from inversion of surface waveform data. *Geophys. J. Int.*, 171:206–222.

Marone, F. and Romanowicz, B. (2007). Non-linear crustal corrections in high-resolution regional waveform seismic tomography. *Geophys. J. Int.*, 170:460–467.

Masters, G., Laske, G., Bolton, H., and Dziewonski, A. (2000). *The Relative Behavior of Shear Velocity, Bulk Sound Speed, and Compressional Velocity in the Mantle: Implications for Chemical and Thermal Structure*, pages 63–87. American Geophysical Union.

McNamara, A. and Zhong, S. (2005). Thermochemical structures beneath africa and the pacific ocean. *Nature*, 437(7062):1136–1139.

McNamara, A. K., van Keken, P. E., and Karato, S.-I. (2002). Development of anisotropic structure in the Earth's lower mantle by solid-state convection. *Nature*, 416(6878):310–314.

McNutt, M. K. and Fischer, K. M. (1987). The south Pacific superswell. In Keating, B. H., editor, *Islands and Atolls*, Geophys. Monogr., pages 25–34. AGU, Washington D.C.

Mégnin, C. and Romanowicz, B. (1999). The effects of the theoretical formalism and data selection on mantle models derived from waveform tomography. *Geophys. J. Int.*, 138:366–380.

Mégnin, C. and Romanowicz, B. (2000). The three-dimensional shear velocity structure of the mantle from the inversion of body, surface and higher-mode waveforms. *Geophys. J. Int.*, 143:709–728.

Métivier, L., Brossier, R., Virieux, J., and Operto, S. (2013). Full Waveform Inversion and the Truncated Newton Method. *SIAM J. Sci. Comput.*, 35:B401–B437.

Montagner, J. and Anderson, D. (1989). Petrological constraints on seismic anisotropy. *Phys. Earth Planet. Inter.*, 54:82–105.

Montagner, J. P. and Jobert, N. (1981). Investigation of upper mantle structure under young regions of the southeast Pacific using long-period Rayleigh waves. *Phys. Earth Planet. Inter.*, 27:206222.

Morgan, J. P., Morgan, W. J., Zhang, Y.-S., and Smith, W. H. F. (1995). Observational hints for a plume-fed, suboceanic asthenosphere and its role in mantle convection. *J. Geophys. Res.*, 100(B7):12,75312,767.

Morgan, W. J. (1971). Convection Plumes in the Lower Mantle. *Nature*, 230:42–43.

Murakami, M., Hirose, K., Kawamura, K., Sata, N., and Ohishi, Y. (2004). Post-perovskite phase transition in MgSiO3. *Science*, 304(5672):855–858.

Nettles, M. and Dziewonski, A. (2008). Radially anisotropic shear velocity structure of the upper mantle globally and beneath North America. *J. Geophys. Res.*, 113:B02303.

Ni, S., Tan, E., Gurnis, M., and Helmberger, D. (2002). Sharp Sides to the African Superplume. *Science*, 296(5574):1850–1852.

Nishimura, C. and Forsyth, D. (1989). The anisotropic structure of the upper mantle in the Pacific. *Geophys. J. Int.*, 96:203–229.

Nishtala, R., Hargrove, P., Bonachea, D., and Yelick, K. (2009). Scaling Communication-intensive Applications on BlueGene/P Using One-sided Communication and Overlap. In *Proceedings of the 2009 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*.

Nolet, G. and Dahlen, F. (2000). Wave front healing and the evolution of seismic delay times. *J. Geophys. Res.*, 105:043–19.

Obrebski, M., Allen, R., Pollitz, F., and Hung, S.-H. (2011). Lithosphere-asthenosphere interaction beneath the western United States from the joint inversion of body-wave traveltimes and surface-wave phase velocities. *Geophys. J. Int.*, 185:1003–1021.

O'Connor, J. M., Jokat, W., le Roex, A. P., Class, C., Wijbrans, J. R., Keling, S., Kuiper, K. F., and Nebel, O. (2012). Hotspot trails in the South Atlantic controlled by plume and plate tectonic processes. *Nat. Geosci.*, 5:735738.

Panning, M., Cao, A., Kim, A., and Romanowicz, B. (2012). Non-linear 3-D Born shear waveform tomography in Southeast Asia. *Geophys. J. Int.*, 190:463–475.

Panning, M., Lekić, V., and Romanowicz, B. (2010). Importance of crustal corrections in the development of a new global model of radial anisotropy. *J. Geophys. Res.*, 115:361–379.

Panning, M. and Romanowicz, B. (2006). A three-dimensional radially anisotropic model of shear velocity in the whole mantle. *Geophys. J. Int.*, 167:361–379.

Pasyanos, M. and Nyblade, A. (2007). A top to bottom lithospheric study of Africa and Arabia. *Tectonophysics*, 444:1–4.

Priestley, K., Debayle, E., McKenzie, D., and Pilidou, S. (2006). Upper mantle structure of eastern Asia from multimode surface waveform tomography. *J. Geophys. Res.*, 111:B10304.

Priestley, K., McKenzie, D., Debayle, E., and Pilidou, S. (2008). The African upper mantle and its relationship to tectonics and surface geology. *Geophys. J. Int.*, 175:1108–1126.

Richards, M. and Engebretson, D. (1992). Large-scale mantle convection and the history of subduction. *Nature*, 355:437440.

Richter, F. M. and Parsons, B. (1975). On the interaction of two scales of convection in the mantle. *J. Geophys. Res.*, 80:2529–2541.

Rickers, F., Fichtner, A., and Trampert, J. (2012). Imaging mantle plumes with instantaneous phase measurements of diffracted waves. *Geophys. J. Int.*, 190:650–664.

Rickers, F., Fichtner, A., and Trampert, J. (2013). The Iceland–Jan Mayen plume system and its impact on mantle dynamics in the North Atlantic region: Evidence from full-waveform inversion. *Earth Planet. Sci. Lett.*, 367:39–51.

Ritsema, J., Deuss, A., van Heijst, H., and Woodhouse, J. (2011). S40RTS: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltime and normal-mode splitting function measurements. *Geophys. J. Int.*, 184:1223–1236.

Ritzwoller, M., Shapiro, N., Barmin, M., and Levshin, A. (2002). Global surface wave diffraction tomography. *J. Geophys. Res.*, 107(B12):2335.

Robertson, G. and Woodhouse, J. (1996). Ratio of relative S to P velocity heterogeneity in the lower mantle. *J. Geophys. Res.*, 101(B9):20041–20052.

Rodi, W., Glover, P., Li, T., and Alexander, S. (1975). A fast, accurate method for computing group-velocity partial derivatives for Rayleigh and Love modes. *Bull. seism. Soc. Am.*, 65:1105–1114.

Rogers, N., Macdonald, R., Fitton, J., George, R., Smith, M., and Barreiro, B. (2000). Two mantle plumes beneath the East African rift system: Sr, Nd, and Pb isotope evidence from Kenya Rift basalts. *Earth Planet. Sci. Lett.*, 176:387–400.

Romanowicz, B. (1979). Seismic structure of the upper mantle beneath the United States, by three-dimensional inversion of body wave arrival times. *Geophys. J. R. astr. Soc.*, 57:479–506.

Romanowicz, B. (2001). Can we resolve 3d density heterogeneity in the lower mantle? *Geophys. Res. Lett.*, 28(6):1107–1110.

Romanowicz, B. (2008). Using seismic waves to image Earth's internal structure. *Nature*, 451:266–268.

Romanowicz, B. and Gung, Y. C. (2002). Superplumes from the core-mantle boundary to the lithosphere: implications for heat flux. *Science*, 296:513–516.

Romanowicz, B., Panning, M., Gung, Y., and Capdeville, Y. (2008). On the computation of long period seismograms in a 3-D earth using normal mode based approximations. *Geophys. J. Int.*, 175:520–536.

Sebai, A., Stutzmann, E., Montagner, J., Sicilia, D., and Beucler, E. (2006). Anisotropic structure of the African upper mantle from Rayleigh and Love wave tomography. *Phys. Earth Planet. Inter.*, 155:1–2.

Seriani, G. and Oliveira, S. (2008). Dispersion analysis of spectral element methods for elastic wave propagation. *Wave Motion*, 45(6):729–744.

Shapiro, N. and Ritzwoller, M. (2002). Monte-Carlo inversion for a global shear-velocity model of the crust and upper mantle. *Geophys. J. Int.*, 151:88–105.

Sigloch, K., McQuarrie, N., and Nolet, G. (2008). Two-stage subduction history under North America inferred from multiple-frequency tomography. *Nat. Geosci.*, 1:458–462.

Simmons, N., Forte, A., Boschi, L., and Grand, S. (2010). GyPSuM: A joint tomographic model of mantle density and seismic wave speeds. *J. Geophys. Res.*, 115:109–124.

Simmons, N., Forte, A., and Grand, S. (2006). Constraining mantle flow with seismic and geodynamic data: a joint approach. *Earth Planet. Sci. Lett.*, 246:109–124.

Snyder, D. and Tait, S. (1998). A flowfront instability in viscous gravity currents. *J. Fluid Mech.*, 369:121.

Steinberger, B. (2000). Plumes in a convecting mantle: models and observations for individual hotspots. *J. Geophys. Res.*, 105:127–11.

Stixrude, L. and Lithgow-Bertelloni, C. (2005). Mineralogy and elasticity of the oceanic upper mantle: Origin of the low-velocity zone. *J. Geophys. Res.*, 110:B03204.

Su, W.-J. and Dziewonski, A. (1992). On the scale of mantle heterogeneity. *Phys. Earth Planet. Inter.*, 74(1):29–54.

Su, W.-J. and Dziewonski, A. (1997). Simultaneous inversion for 3-d variations in shear and bulk velocity in the mantle. *Phys. Earth. Planet. Inter.*, 100(1):135–156.

Tape, C., Liu, M., Maggi, A., and Tromp, J. (2009). Adjoint Tomography of the Southern California Crust. *Science*, 325:988–992.

Tape, C., Liu, Q., Maggi, A., and Tromp, J. (2010). Seismic tomography of the southern California crust based on spectral-element and adjoint methods. *Geophys. J. Int.*, 180:433–462.

Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49:1259–1266.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation.* Society for Industrial and Applied Mathematics, Philadelphia.

Tarantola, A. and Nercessian, A. (1984). Three-dimensional inversion without blocks. *Geophys. J. R. astr. Soc.*, 76(2):299–306.

Tarantola, A. and Valette, B. (1982). Generalized Nonlinear Inverse Problems Solved Using the Least Squares Criterion. *Rev. Geophys.*, 20:219–232.

Thurber, C. and Ritsema, J. (2007). Seismic Tomography and Inverse Methods. In Schubert, G., editor, *Treatise on Geophysics*, pages 323 – 360. Elsevier, Amsterdam.

Toomey, D. R., Wilcock, W. S. D., Conder, J. A., Forsyth, D. W., Blundy, J. D., Parmentier, E. M., and Hammond, W. C. (2002). Asymmetric mantle dynamics in the MELT region of the East Pacific Rise. *Earth Planet. Sci. Lett.*, 200:287–295.

Tromp, J., Komatitsch, D., and Liu, Q. (2008). Spectral-Element and Adjoint Methods in Seismology. *Comm. Comp. Phys.*, 3:1–32.

Tromp, J., Tape, C., and Liu, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophys. J. Int.*, 160:195–216.

UPC Language Specifications (2005). UPC Language Specifications, v1.2. Technical Report LBNL-59208, Lawrence Berkeley National Lab.

van der Hilst, R., Widiyantoro, S., and Engdahl, E. (1997). Evidence for deep mantle circulation from global tomography. *Nature*, 386(6625):578–584.

Wang, Z. and Dahlen, F. A. (1995). Spherical-spline parameterization of three-dimensional Earth models. *Geophys. Res. Lett.*, 22:3099–3102.

Weeraratne, D., Forsyth, D., Fischer, K., and Nyblade, A. (2003). Evidence for an upper mantle plume beneath the Tanzanian craton from Rayleigh wave tomography. *J. Geophys. Res.*, 108:3099–3102.

Wolfe, C. J., Solomon, S. C., Laske, G., Collins, J. A., Detrick, R. S., Orcutt, J. A., Bercovici, D., and Hauri, E. H. (2011). Mantle P-wave velocity structure beneath the Hawaiian hotspot. *Earth Planet. Sci. Lett.*, 303:267280.

Woodhouse, J. (1980). The coupling and attenuation of nearly resonant multiplets in the Earth's free oscillation spectrum. *Geophys. J. R. astr. Soc.*, 61:261–283.

Woodhouse, J. and Dziewonski, A. (1984). Mapping the upper mantle: three dimensional modeling of Earth structure by inversion of seismic waveforms. *J. Geophys. Res.*, 89:5953–5986.

Yuan, H., French, S., Cupillard, P., and Romanowicz, B. (2014). Lithospheric expression of geological units in central and eastern North America from full waveform tomography. *Earth Planet. Sci. Lett.*

Yuan, H. and Romanowicz, B. (2010). Lithospheric layering in the North American craton. *Nature*, 466(7310):1063–1068.

Yuan, H., Romanowicz, B., Fischer, K., and Abt, D. (2011). 3-D shear wave radially and azimuthally anisotropic velocity model of the North American upper mantle. *Geophys. J. Int.*, 184:1237–1260.

Zheng, Y., Kamil, A., Driscoll, M., Shan, H., and Yelick, K. (2014). UPC++: a PGAS extension for C++. In *28th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*.

Zhu, H., Bozdağ, E., Peter, D., and Tromp, J. (2012). Structure of the European upper mantle revealed by adjoint tomography. *Nat. Geosci.*, 5:493–498.

Zhu, H. and Tromp, J. (2013). Mapping Tectonic Deformation in the Crust and Upper Mantle Beneath Europe and the North Atlantic Ocean. *Science*, 341(6148):871–875.

Appendix A

Crustal structure in NACT

A.0.1 Background: NACT-based modeling

NACT is a method for waveform and sensitivity-kernel calculation using normal-mode perturbation theory, and we refer the reader to Li & Romanowicz 1995 for a detailed introduction. Under NACT, along-branch mode coupling is treated in a non-linear fashion, thus including the effect of multiple forward scattering, while cross-branch coupling allows for accurate modeling of overtone surface-waves Romanowicz et al. (2008). To avoid coupling calculations over the whole sphere, sensitivity kernels are limited to the 2D plane containing the source and receiver. Volumetric model perturbations $\mathbf{m}(r, \theta, \phi)$ and perturbations to discontinuity topography $\{h_d : d = 1, \ldots, N_d\}$, enter into the NACT-predicted accelerogram $u(t)$ (and sensitivity kernels $\partial u/\partial \mathbf{m}$) through source-receiver minor-arc and great-circle path integrations over the local frequency shift $\delta\omega_{kk'}(\theta, \phi)$ Woodhouse (1980). This quantity corresponds to the shift in average degenerate eigenfrequency of the multiplet pair (k, k') due to coupling induced by structural heterogeneity beneath point (θ, ϕ):

$$\begin{aligned} &\delta\omega_{kk'}(\theta, \phi) = \\ &\quad \frac{1}{\omega_k + \omega_{k'}} \int_0^a \mathbf{m}(r, \theta, \phi) \cdot \mathbf{M}_{kk'}(r) r^2 \, \mathrm{d}r \\ &\quad - \sum_d^{N_d} r_d^2 h_d(\theta, \phi) H_{kk'}^d \end{aligned} \tag{A.1}$$

where a is earth radius, r_d is the unperturbed radius of discontinuity d, and $\mathbf{M}_{kk'}(r)$ and $H_{kk'}^d$ are eigenfrequency partial-derivatives with respect to volumetric and discontinuity-topography perturbations (e.g. Romanowicz, 1987).

Given the (linearized) form of eq. (A.1), two particular concerns arise regarding treatment of the smooth crustal layer (Section 2.2.3): (1) fidelity to the true non-linear response of $\delta\omega_{kk'}$ to crustal-thickness variation; and (2) recovery of sensitivity kernels reflecting the split parameterization between crust and mantle - i.e. $\mathbf{m}(r, \theta, \phi)$ is only strictly defined at mantle depths. Point (2) may easily be treated by ensuring that the upper limit of integration on the radial inner product in eq. (A.1) honors the local Moho depth $h_M(\theta, \phi)$, i.e.

$$\int_0^{h_M(\theta,\phi)} \mathbf{m}(r,\theta,\phi) \cdot \mathbf{M}_{kk'}(r) r^2 \, \mathrm{d}r. \tag{A.2}$$

However, treatment of point (1) is more involved, and we now discuss our solution at greater length.

A.0.2 Non-regionalized modified-linear crustal corrections

We adopt an extended implementation of the modified linear crustal correction (MLC) scheme of Lekić et al. 2010, extended in that we do not limit ourselves to a small set of "regionalized" crustal models. Instead, we obtain continuously-parameterized corrections, directly reflecting lateral variation of the crustal model.

Given a 1D reference earth-model m_{ref} and structure associated with each crustal-model node $\{\mathbf{c}_n : n = 1, \dots, N_k\}$, we calculate normal-mode eigenfrequencies for both m_{ref} itself and a set of N_k composite 1D models, denoted m^n_{ref}, that combine crustal structure from $\mathbf{c}_n$ with mantle structure from m_{ref}. For a given normal-mode multiplet k, we refer to

$$\delta\omega^n_k = \omega_k \left[m^n_{ref}\right] - \omega_k \left[m_{ref}\right] \tag{A.3}$$

as the non-linear local eigenfrequency shift induced by crustal structure at node n - non-linear because it is recovered through direct evaluation of eigenfrequency dispersion in the perturbed the model, not through the linearized approach in eq. (A.1) above. Further, given m_{ref}, we calculate eigenfrequency partial derivatives due to self-coupling ($k = k'$) for perturbations in Moho and seafloor topography, as well as average crustal shear-wave velocity: here denoted H^M_{kk}, H^S_{kk}, and $\bar{M}^c_{kk}$, respectively. For each crustal model node n, we seek linear perturbations to Moho depth (h_M), topography/bathymetry (h_S), and crustal average V_S ($\bar{V}^c_{S,iso}$):

$$\delta\mathbf{x}_n = \left(\delta h_M, \delta h_S, \delta \ln \bar{V}^c_{S,iso}\right)^T \tag{A.4}$$

such that the over-determined linear system

$$\begin{pmatrix} H^M_{k_1k_1} & H^S_{k_1k_1} & \bar{M}^c_{k_1k_1} \\ H^M_{k_2k_2} & H^S_{k_2k_2} & \bar{M}^c_{k_2k_2} \\ & \vdots & \\ & & \end{pmatrix} \delta\mathbf{x}_n = \begin{pmatrix} \delta\omega^n_{k_1} \\ \delta\omega^n_{k_2} \\ \vdots \\ \end{pmatrix} \tag{A.5}$$

is satisfied in a least-squares sense for multiplets $\{k_1, k_2, \dots\}$ within the period range of interest (e.g. 60 - 400s). The set of three-factor correction coefficients $\{\delta\mathbf{x}_n : n = 1, \dots, N_k\}$, mimicking non-linearity induced by local crustal structure, are expanded laterally in spherical splines and used to calculate path-dependent corrections to integrations over $\delta\omega_{kk'}$ in NACT.

Note that, in deriving our correction factors, we have limited our attention to the self-coupling case only. As noted by Lekić et al. 2010, satisfying the self-coupling case alone should be sufficient for the phases most affected by crustal structure at long periods (e.g. fundamental-mode surface waves). The resulting three-factor MLCs are sought for spheroidal and toroidal, and fundamental and overtone modes separately. This scheme allows us to use a single set of eigenfrequency partial derivatives (those for m_{ref}) in our NACT calculations, reducing file I/O and memory overhead relative to previous regionalized non-linear schemes Marone and Romanowicz (2007), while achieving similar fidelity to crustal effects on the long-period wavefield. The accuracy of this approach relative to both "traditional" crustal corrections and SEM benchmarks has previously been demonstrated by Lekić et al. 2010, and we refer the reader to their results.