A full waveform tomography algorithm for teleseismic body and surface waves in 2.5 dimensions

B. Baker and S. Roecker
Rensselaer Polytechnic Institute, Troy, NY 12180, USA. E-mail: bakerb845@gmail.com

SUMMARY
We describe a 2.5-D, frequency domain, viscoelastic waveform tomography algorithm for imaging with seismograms of teleseismic body and surface waves recorded by quasi-linear arrays. The equations of motion are discretized with p-adaptive finite elements that allow for geometric flexibility and accurate solutions as a function of wavelength. Artificial forces are introduced into the media by specifying a known wavefield along the model edges and solving for the corresponding scattered field. Because of the relatively low frequency content of teleseismic data, regional scale tectonic settings can be parametrized with a modest number of variables and perturbations can be determined directly from a regularized Gauss–Newton system of equations. Waveforms generated by the forward problem compare well with analytic solutions for simple 1-D and 2-D media. Tests of different approaches to the inverse problem show that the use of an approximate Hessian serves to properly focus the scattered field. We also find that while full waveform inversion can provide significantly better resolution than standard techniques for both body and surface wave tomography modelled individually, joint inversion both enhances resolution and mitigates potential artefacts.

Key words: Numerical solutions; Inverse theory; Seismic tomography; Computational seismology.

1 INTRODUCTION
A common and cost effective strategy in seismic imaging investigations is to acquire teleseismic data along a quasi linear array that spans an area of interest. The relative abundance, signal strength, and total medium sampling makes earthquakes recorded at teleseismic distances excellent probes for structure beneath local seismic networks. Estimates of properties such as elastic wave velocities or density can be made by interpretation and/or inversion of the collected data. Standard tools for analysis of body waves include arrival time tomography and receiver function analysis (e.g. Aki et al. 1977; Vinnik 1977; Langston 1979). Though typically 1-D in principle, receiver functions under dense arrays can be migrated to map sharp, subhorizontal interfaces below receivers and infer a 2-D or 3-D structure for a given model. Alternatively, teleseismic arrival time tomography, which is commonly performed in two or three dimensions (e.g. Li et al. 2009), can robustly estimate lateral variations in material properties, but typically resolution is poor in depth and ultimately limited by the Fresnel zone. Standard surface wave analysis (e.g. Ritzwoller et al. 2002; Priestley et al. 2006) employs a form of arrival time tomography in the generation of phase and/or group velocity maps that are then used to construct 3-D models through the juxtaposition of 1-D inversions of localized dispersion curves. While benefiting from a sensitivity to absolute shear wave speeds and variations with depth, this type of imaging often suffers from low lateral resolution.

The implicit combination of arrival time tomography and migration through full waveform inversion (FWI) has long been known to possess superior resolving power since the pioneering work of Lailly (1983) and Tarantola (1984). Impressive results have been obtained for passive source experiments at regional scales by inverting only selected phases in the waveform with time domain adjoint tomographic methods (e.g. Fichtner et al. 2009; Tape et al. 2010; Zhu et al. 2012). While the selection of phases likely produces more linear objective functions, the computational expense remains prohibitive for many researchers. Alternatively, the frequency domain analogue for active source FWI in 2-D is much more accessible (Pratt 1990; Pratt & Worthington 1990). This advantage was recently exploited for analysis of teleseismic data by Roecker et al. (2010) who overcame the 2-D restriction to in-plane sources and receivers by parametrizing the medium in 2.5-D, which is defined as a 3-D medium with a materially invariant direction. Pageot et al. (2013) performed resolution analysis of the lithospheric imaging problem by plane waves in 2-D using a more sophisticated inversion scheme (Broyden–Fletcher–Goldfarb–Shanno or l-BFGS method) and demonstrated that targets of geologic interest could be sufficiently well reconstructed for receiver spacings between
5 and 10 km. Similar results for 10 km receiver spacing were obtained by Tong et al. (2014) with a 2-D time domain hybrid spectral element method frequency wavenumber solver using a conjugate gradient inversion.

Comprehensive overviews of FWI techniques can be found elsewhere (e.g. Virieux & Operto 2009; Liu & Gu 2012); we provide a brief review here. As with many inverse procedures, FWI consists of a forward problem and an inverse problem. The forward problem in this case involves the generation of synthetic waveforms. The inverse problem consists of estimating perturbations to model parameters that will better explain the data ‘wiggle for wiggle’. To pair the forward and inverse problems for local optimization schemes requiring information from derivatives, the adjoint state method allows efficient gradient computations (e.g. Lions 1968; Chavent 1974; Plessix 2006). The gradient can then be used in a Newton optimization scheme to estimate parameter updates that match the estimated waveforms to the observed waveforms.

Unlike forward modelling, methodologies for inversion are not as straightforward as simply incorporating a more realistic physical description. Consequently, there are many strategies for inversion of seismic data that inherently include some tradeoff between robustness, resolving power, and computational feasibility. Given the complexity of the forward problem and number of inversion variables involved in a typical application we are limited to Newton type methods as opposed to global search strategies such as grid search or Monte Carlo methods. Even for this subdomain of inversion schemes a wealth of strategies exist for the approximation of a Newton step (see Nocedal & Wright 2006, for a thorough review). Methods not requiring knowledge of the Hessian matrix include steepest descent (e.g. Roecker et al. 2010), and conjugate gradients (e.g. Tape et al. 2010) and are computationally efficient for large problems. At the same time, it is well known that more accurate results can be obtained in fewer iterations by incorporating estimates of the Hessian. Techniques for doing so include the limited memory l-BFGS method that provides for a low rank approximation to the Hessian through finite differencing (e.g. Nocedal 1980; Sourbier et al. 2009), and truncated Gauss–Newton/Newton methods that resolve the Gauss–Newton/Newton equations in a matrix free fashion with iterative solvers such as conjugate gradients (e.g. Nash 2000; Métivier et al. 2012). All these methods benefit from the application of a suitable pre-conditioner such as an inverse diagonal approximate Hessian (Shin et al. 2001). As computer memory and processing power continually expand and become less expensive, we are increasingly capable of avoiding these approximations and solving the Gauss–Newton system of equations directly, and can explicitly include non-diagonal regularization schemes in formulations similar to those described in Tarantola (2005).

In this paper we introduce a finite element strategy for discretization of the viscoelastic wave equations in 2.5 dimensions. Compared to the finite difference formulation of Roecker et al. (2010), a finite element approach makes it much easier to take topography into account, and also allows for greater flexibility in model parameterization. Within the finite element we include polynomial order, or p adaptivity, so that we can match element size to expected inversion resolution. We also review and implement a strategy for generating an appropriate force distribution, local to the model, for teleseismic body and surface waves. We then describe strategies for full waveform inversion and discuss the necessary steps for calculation of Jacobian matrices. These strategies are then illustrated by a series of examples.

2 FORWARD MODELLING

The accurate computation of synthetic seismograms for a sourceless media involves two key components: a viable discretization scheme and a specification of an artificial force distribution that generates teleseismic body or surface waves. If the ultimate objective of forward modelling is inversion then our discretization scheme must be accurate at the level of resolution. Additionally, in many applications the forward solver must account for topographic effect, especially if the waveforms are used for imaging (e.g. Bleibinhaus & Rondenay 2012). An appropriate balance of accuracy and geometric flexibility can be achieved with a p-adaptive finite element method (e.g. Babuska & Guo 1992). Such a strategy can substantially reduce the number of variables in the inversion, as we need only generate an image at anchor nodes while adjusting the polynomial order of the basis functions to ensure accurate computation of the forward problem. While conceptually straightforward, the computation of numerical solutions to the full 3-D elastic wave equation at scales of tectonic interest requires substantial computational resources. Roecker et al. (2010) argue that there are many geological structures (e.g. mountain ranges, subduction zones and many features associated with quasi-linear plate boundaries) that are approximately invariant for at least a few tens of km along an azimuth and so could be usefully parametrized in 2.5-D. The computational advantage of 2.5-D is the reduction in the number of variables from O(n³) to O(n²) which makes this approach feasible for modest computers. Similarly motivated, we begin with the frequency domain 2.5-D elastic wave equation in terms of displacement u, angular frequency ω, compressional wave speed α, shear wave speed β and slowness in the invariant direction p:

\[ \omega^2 \left( 1 - p^2 \beta^2 \right) u_1 + \frac{\partial}{\partial x} \left[ \alpha^2 \frac{\partial u_1}{\partial x} + \left( \alpha^2 - 2 \beta^2 \right) \frac{\partial u_3}{\partial z} \right] + \frac{\partial}{\partial z} \left[ \beta^2 \frac{\partial u_3}{\partial z} + \frac{\partial u_1}{\partial x} \right] - i \omega \rho \left[ \frac{\partial}{\partial x} \left[ \alpha^2 \frac{\partial u_1}{\partial x} + \left( \alpha^2 - 2 \beta^2 \right) u_2 \right] + \beta^2 \frac{\partial u_3}{\partial x} \right] = 0 \]  

(1a)

\[ \omega^2 \left( 1 - p^2 \alpha^2 \right) u_2 + \frac{\partial}{\partial x} \left[ \beta^2 \frac{\partial u_2}{\partial x} + \frac{\partial u_3}{\partial z} \right] + \frac{\partial}{\partial z} \left[ \beta^2 \frac{\partial u_3}{\partial z} + \frac{\partial u_1}{\partial x} \right] - i \omega \rho \left[ \left( \alpha^2 - 2 \beta^2 \right) \frac{\partial u_1}{\partial x} + \frac{\partial u_3}{\partial z} \right] + \alpha^2 \left( \alpha^2 - 2 \beta^2 \right) \frac{\partial u_2}{\partial x} + \beta^2 \frac{\partial u_3}{\partial z} = 0 \]  

(1b)

\[ \omega^2 \left( 1 - p^2 \beta^2 \right) u_3 + \frac{\partial}{\partial z} \left[ \alpha^2 \frac{\partial u_1}{\partial z} + \left( \alpha^2 - 2 \beta^2 \right) \frac{\partial u_1}{\partial x} \right] + \frac{\partial}{\partial x} \left[ \beta^2 \frac{\partial u_2}{\partial x} + \frac{\partial u_3}{\partial z} \right] - i \omega \rho \left[ \left( \alpha^2 - 2 \beta^2 \right) u_2 + \beta^2 \frac{\partial u_3}{\partial x} \right] = 0 \]  

(1c)

To explicitly satisfy the free surface boundary condition, σ, n = 0, and thereby provide for geometric flexibility we seek the weak form corresponding to (1a)–(1c). Let \textbf{u} be a trial function that satisfies all the Dirichlet boundary conditions. We multiply (1a)–(1c) by a weight
function \( \mathbf{u} \) which vanishes on the Dirichlet boundaries. We then integrate by parts and note that because the Neumann condition at the free surface is such that \( \sigma / n = 0 \), we are left with no surface integrals. The corresponding variational form is

\[
-\omega^2 \int_\Omega \left( 1 - p_i^2 \beta_i^2 \right) w_i u_i \, dx \, dz + \int_\Omega \frac{\partial w_i}{\partial x} \frac{\partial u_i}{\partial x} + \frac{\partial w_i}{\partial z} \frac{\partial u_i}{\partial z} \, dx \, dz + \int_\Omega \frac{\partial w_i}{\partial x} (\alpha^2 - 2 \beta^2) \frac{\partial u_i}{\partial x} + \frac{\partial w_i}{\partial z} \beta^2 \frac{\partial u_i}{\partial z} \, dx \, dz = 0
\]

\[
-\omega^2 \int_\Omega \left( 1 - p_i^2 \alpha^2 \right) w_i u_i \, dx \, dz + \int_\Omega \frac{\partial w_i}{\partial x} \beta_i \frac{\partial u_i}{\partial x} + \frac{\partial w_i}{\partial z} \beta_i \frac{\partial u_i}{\partial z} \, dx \, dz = 0
\]

\[
-\omega^2 \int_\Omega \left( 1 - p_i^2 \beta_i^2 \right) w_i u_i \, dx \, dz + \int_\Omega \frac{\partial w_i}{\partial x} \beta_i \frac{\partial u_i}{\partial x} + \frac{\partial w_i}{\partial z} \beta_i \frac{\partial u_i}{\partial z} \, dx \, dz - \omega^2 \int_\Omega \left( 1 - p_i^2 \alpha^2 \right) w_i u_i \, dx \, dz = 0
\]

We discretize (2a)–(2c) with the finite element method and in the discussion below adopt the notation of Hughes (2000). The domain is first decomposed into a patchwork of \( n_{\text{elem}} \) conforming quadrilaterals or elements such that \( \Omega = \bigcup_{i=1}^{n_{\text{elem}}} \Omega_i \). Next, the trial \( u_i \) and weight functions \( w_i \) are approximated by the finite dimensional polynomial functions \( u_i^h \) and \( w_i^h \). These approximate trial and weight functions are made from an expansion of \( p \)th order Lagrange polynomial basis functions \( \phi_i \), where the number of element nodes, \( n_{\text{en}} = (p + 1)^2 \):

\[
w_i \approx u_i^h = \sum_{i=1}^{n_{\text{en}}} c_i^j \phi_i(x, z)
\]

\[
u_i \approx u_i^h = \sum_{i=1}^{n_{\text{en}}} d_i^j \phi_i(x, z).
\]

To minimize the condition number of the element matrices we choose the interpolation points for the element nodes to coincide with the Gauss–Lobatto quadrature grid. Note, that this is not a spectral element method since our mass matrix will be consistent, meaning that the basis functions are fully integrated. Ignoring the superscript \( a \) and noting that each coefficient \( c_i^j \) will multiply each linear equation and hence can be eliminated, the finite element approximation to (2a)–(2c) is given by

\[
-d_i^j \omega^2 \int_{\Omega_i} \left( 1 - p_i^2 \beta_i^2 \right) \phi_i \phi_i \partial \Omega_i + d_i^j \int_{\Omega_i} \alpha^2 \phi_i \phi_i \partial \Omega_i + \beta_i^2 \phi_i \phi_i \partial \Omega_i + d_i^j \int_{\Omega_i} \beta_i \phi_i \phi_i \partial \Omega_i + \omega^2 \int_{\Omega_i} \left( 1 - p_i^2 \alpha^2 \right) \phi_i \phi_i \partial \Omega_i = 0
\]

\[
-\omega^2 \int_{\Omega_i} \left( 1 - p_i^2 \alpha^2 \right) \phi_i \phi_i \partial \Omega_i + d_i^j \int_{\Omega_i} \beta_i \phi_i \phi_i \partial \Omega_i + \omega^2 \int_{\Omega_i} \left( 1 - p_i^2 \beta_i^2 \right) \phi_i \phi_i \partial \Omega_i = 0
\]

\[
-d_i^j \omega^2 \int_{\Omega_i} \left( 1 - p_i^2 \beta_i^2 \right) \phi_i \phi_i \partial \Omega_i + d_i^j \int_{\Omega_i} \alpha^2 \phi_i \phi_i \partial \Omega_i + \beta_i^2 \phi_i \phi_i \partial \Omega_i + d_i^j \int_{\Omega_i} \beta_i \phi_i \phi_i \partial \Omega_i + \omega^2 \int_{\Omega_i} \left( 1 - p_i^2 \alpha^2 \right) \phi_i \phi_i \partial \Omega_i = 0
\]

The 3 \( \times \) 3 element matrices that multiply the \( d_i^j \) vector, as well as strategies for evaluating the integrals above, are given in Appendix A. Eqs (4a)–(4c) can then be assembled into a global impedance matrix \( S \) with \( n_{\text{def}} \) degrees of freedom or linearly independent equations. The number of degrees of freedom is a function of the number of nodal points in the mesh. As noted above, with 2.5-D we reduce our problem from 3-D to a 2-D slice, thereby reducing the number of linear equations from \( O(n^3) \) complexity to \( O(n^2) \) and making this scheme tractable on small computers.
2.1 Absorbing boundaries and attenuation

The boundary conditions for this application are a traction free surface, zero displacement at the other three boundaries, and the requirement that no energy is back scattered from outside the model. The last condition is implemented by specifying a perfectly matched layer (PML).

To make the computational domain finite we follow Zheng & Huang (2002) and introduce a stretched coordinate system from \( x_i \rightarrow \tilde{x}_i \)

\[
\tilde{x}_i = x_i^0 + \int_{x_i^0}^{x_i} \gamma(x')\,dx',
\]

where \( x_i^0 \) denotes the edge of the PML. When \( x_i < x_i^0 \) we are in the computational domain and the only viscous damping would be that introduced by attenuation. The effect on the spatial derivatives is that

\[
\frac{\partial}{\partial \tilde{x}_i} = \frac{1}{\gamma(x_i)} \frac{\partial}{\partial x_i}.
\]

The complex valued damping function \( \gamma \), is related to that of Collino & Tsogka (2001) by

\[
\gamma(x_i) = 1 + \frac{d(x_i)}{i\omega}.
\]

where \( d(x_i) \) is a real valued damping profile defined by eq. (8b). Our numerical tests indicate that this PML formulation can be insufficient for long period surface waves (generally periods longer than 50 s) unless attenuation is added to the medium. To mitigate this meshing difficulty we instead adopt a convolutional PML approach (Komatitsch & Martin 2007) and define

\[
\gamma(x_i) = \kappa(x_i) + \frac{d(x_i)}{\nu(x_i) + i\omega}.
\]

\[
d(x_i) = -(N+1)\alpha \log(R) \left(\frac{x_i}{L}\right)^N
\]

\[
\kappa(x_i) = 1 + (k_{\text{max}} - 1) \left(\frac{x_i}{L}\right)^2
\]

\[k_{\text{max}} \geq 1\]

\[
\nu(x_i) = \pi f_0 \left(1 - \frac{x_i}{L}\right)
\]

the effect of which is to shift the pole in \( \gamma(x_i) \) at \( \omega = 0 \) off the real axis. In the equations above, \( R \) is a reflection coefficient generally taken to be 0.001, \( L \) is the width of the PML and \( f_0 \) is the dominant frequency in the modelled waveform. In practice it is equivalent to either derive the discretized weak forms for complex stretched coordinates by taking care to separate the real and complex parts of the strong form and integrate or to substitute the complex stretched derivatives of eq. (6) into the previously discretized weak form.

An attenuating medium can easily be simulated in the spectral domain by introducing a complex velocity (e.g. Aki & Richards 2002; Pratt 1990) in terms of a reference frequency \( \omega_0 \) and compressional and shear quality factors \( Q_\alpha \) and \( Q_\beta \) as

\[
\tilde{v}(\omega_{\alpha,\beta}) = v_{\alpha,\beta} \left[1 + \frac{1}{\pi Q_{\alpha,\beta}} \ln \left(\frac{\omega}{\omega_0}\right) - \frac{i}{2Q_{\alpha,\beta}} \right].
\]

Generally, we choose \( \omega_0 \) to be a frequency significantly higher than any we are modelling.

2.2 The Bielak boundary

Modelling teleseismic surface or body waves requires an effective force representation of an excitation that originates outside the model. A useful way to specify an effective force distribution is to recognize that the external excitation is independent of the medium response, and hence can be specified in terms of the known response of a ‘background’ medium to this excitation. For example, Roecker et al. (2010) proposed calculating the displacement \( u \) in a medium characterized by an impedance matrix \( S \) by first specifying a medium \( S_0 \) for which the displacement \( u^0 \) from an equivalent excitation could be computed analytically. Hence

\[
Su = S_0u^0 = f.
\]

The displacement field can then be computed from

\[
u = S^{-1}S_0u^0.
\]
and to obtain the total field. The latter form has the advantage of introducing effective force terms only at ‘scatterers’, that is locations where the two media differ. A potential disadvantage of this approach is that $S_o$ and $u^0$ need to be specified everywhere in the modelled region, which usually means that the background media must be simple.

A conceptually similar approach was proposed by Bielak et al. (2003), but in this case the numerical domain is divided into internal ($i$), external ($e$) and boundary ($b$) regions (Fig. 1) and the effective force $f$ in eq. (10) initially is relegated the external region. If we require that $S = S_o$ in the external and boundary regions, and hence that $u^0$ is known in these two regions, Bielak et al. (2003) show that $f$ can be represented by a force distribution located solely within the boundary region (referred to here as the Bielak layer). Adding a region ($a$) to allow for an attenuating (PML) boundary layer, and again specifying the total field $u = \{u, u_a, u_b, u_e\}$ in each of the four regions as the sum of background $u^0 = \{u^0_i, u^0_b, u^0_e\}$ and scattered $w = \{w, w_b, w_e, w_a\}$ fields, the forward problem can be written analogously to eq. (12) as

$$
\begin{bmatrix}
S_{ii} & S_{ib} & 0 & 0 \\
S_{bi} & S_{bb} & S_{be} & 0 \\
0 & S_{eb} & S_{ee} & S_{eo} \\
0 & 0 & S_{ae} & S_{ao}
\end{bmatrix}
\begin{bmatrix}
u_i \\
u_b \\
w_e \\
w_a
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
-S_o u^0_f \\
S_{ob} u^0_b \\
0
\end{bmatrix}.
$$

Here $S$ has been partitioned so that $S_o$ is the piece of $S$ corresponding to the interior domain, $S_{ib}, S_{be}, S_{ob}$ and $S_{eo}$ are small pieces along the interior and exterior boundary, $S_{eo}$ is the piece along the exterior PML boundary, and $S_{ao}$ is the piece within the PML. $S_{ob}$ and $S_{eo}$ correspond to the degrees of freedom contained entirely within the Bielak layer and those contained entirely in the external domain, respectively.

The advantage of this approach is that, unlike eq. (12), the background wavefield does not need to be known everywhere; it only needs to be specified within those regions that project onto $S_{ib}$ and $S_{eo}$, or, in other words, only within the Bielak layer. Hence any region for which analytic forms are difficult to generate (for example because of topography or other forms of heterogeneity) can be confined to the interior ‘model’ region.

A typical implementation of this approach is shown in Fig. 1. The external domain includes the PML, with impedance matrix row/column pairs $aa$ and $ae$, and the domains with row/column pairs $ea$, $ee$ and $eit$ where we specify the known background wavefield $u^0_i$. The interior domain contains the model heterogeneity and topography and includes row/column pairs $ib$ and $ii$. Note that the entire wavefield can be obtained after the solution of eq. (13), but generally we will be interested only in those parts within the interior domain (i.e. $u_i$).

We calculate expressions for body waves in a background layered media (i.e. $u^0_i$ and $u^0_e$) using the analytic methods of Haskell (1960, 1962), modified for the viscoelastic case in a manner similar to Silva (1976). For surface waves, we use the locked mode technique of Gomberg & Masters (1988). We note that wavefields in the exterior domain could also be generated with finite difference or boundary element methods when modelling regional earthquakes.

**Figure 1.** Top panel: an example of the mesh decomposition in eq. (13). The seven elements in the green area represent the PML, the two blue elements are the Bielak layer, and the red area is the computational domain. We specify an analytic solution in the blue layer and along the red-blue boundary. The matrix multiplication for the effective force distribution occurs only for elements in contact with the red-blue boundary, but we specify an extra layer to move potential sources away from the PML. Bottom panel: an example of graph partitioning for the mesh. The four colours correspond to the four processes for this process group. Each process assembles and factors the distributed impedance matrix.
2.3 Programming strategies

We conclude this section with a discussion of programming paradigms that significantly speed up the forward problem. As computational hardware has evolved it is common to own or have access to computer resources with many physical cores. Consequently, it is useful to identify operations that can be efficiently and simply split between processes or groups of processes through message passing (e.g. OpenMPI). As each impedance matrix is frequency dependent, a straightforward parallelization strategy is to assign subgroups of processes to simultaneously resolve the system of equations $\mathbf{S}\mathbf{u} = \mathbf{f}$ by first decomposing $\mathbf{S}$ into lower (L) and upper (U) matrix factors then solving $LU\mathbf{u} = \mathbf{f}$. To reduce storage and operations in the LU factorization we reorder entries of the impedance matrix with a nested dissection scheme (George & Liu 1981) as implemented in the Metis package (Karypis & Kumar 1998). The reordered matrix can be factored with the MULTIFrontal Massively Parallel sparse direct Solver MUMPS (Amestoy et al. 2000). Since MUMPS accepts matrices in a distributed form we can address the computationally expensive task of element assembly. To load balance element integration and assembly we use a graph partition from Metis’s PartGraphKWay algorithm (Fig. 1). The number of partitions is dictated by when partition-induced communication bottlenecks in MUMPS begin to counteract gains in parallelism. The optimal number of processes per frequency group will be a function of machine architecture and model size and is generally determined empirically. In our experience, it is best to use only a few processes per group and make as many frequency groups as possible. We note that additional computational gains may be made on shared memory processors by enabling multithreading for Basic Linear Algebra Subprograms Level 3 (BLAS3) operations.

3 INVERSION

Given a solution to the forward problem, we next seek an estimate of a new set of $m$ model parameters $\mathbf{m}$ such that the calculated wavefields extracted at $n_{\text{rec}}$ receiver locations better match the observations $\mathbf{d}$ at these receivers by minimizing an objective function $E(\mathbf{m})$ such as

$$E(\mathbf{m}) = \frac{1}{2}(\mathbf{d} - \mathbf{A}\mathbf{u})^{T}C_{D}^{-1}(\mathbf{d} - \mathbf{A}\mathbf{u}).$$

We define $n_{\text{obs}}$ as the number of observations for a source–frequency pair; $n_{\text{obs}}$ is at most $3 \cdot n_{\text{rec}}$. The $[n_{\text{obs}} \times n_{\text{dof}}]$ matrix $\mathbf{A}$ extracts values of the wavefield only at the receivers (e.g. Bae et al. 2012). Since the receivers are located at interpolation nodes in the finite element scheme, $\mathbf{A}$ contains at most $3 \cdot n_{\text{rec}}$ non-zero columns. The $[n_{\text{obs}} \times n_{\text{dof}}]$ matrix $C_{D}^{-1} = \Sigma^{-1}L\Sigma^{-1}$ is an inverted data covariance matrix composed of data weights, denoted by diagonal matrix $\Sigma^{-1}$, and potentially a roughening or coupling operator denoted by $L$.

To minimize $E(\mathbf{m})$ one can retain first order derivative terms in the Taylor expansion (e.g. Schuster 2009) to obtain the Gauss–Newton system of equations

$$\Re\{[\mathbf{J}^{T}C_{D}^{-1}\mathbf{J}]\mathbf{p}\} = -\Re\{[\mathbf{J}^{T}C_{D}^{-1}\delta\mathbf{d}]\},$$

where the $[n_{\text{dof}} \times m]$ matrix $\mathbf{J}$ is the wavefield Jacobian, the vector $\mathbf{p}$ denotes the search direction and $\delta\mathbf{d} = \mathbf{d} - \mathbf{A}\mathbf{u}$ are the residuals. $\Re\{}$ denotes the real part of the complex quantity inside the brackets.

The Jacobian matrix $\mathbf{J}$ retains its standard form (e.g. Pratt et al. 1998)

$$\mathbf{J} = \frac{\partial \mathbf{u}}{\partial \mathbf{m}} = -S^{-1} \left[ \frac{\partial S}{\partial m_{1}}, \ldots, \frac{\partial S}{\partial m_{m}} \right] = -S^{-1} F,$$

where $F$ is commonly referred to as the matrix of virtual forces and $S$ is the inverted impedance matrix in eq. (13). Calculation of the derivatives of $S$ with respect to compressional velocity is discussed in Appendix B; derivatives with respect to shear velocity are analogous. We note that model derivatives could be expressed equally well in terms of other medium properties such as the Lamé parameters (e.g. Operto et al. 2013). Physically speaking, each column of $F$ contains the wavefield scaled by a local medium perturbation. As the size of a finite element comprising $S$ should not influence the weight of any virtual force $\frac{\partial S}{\partial m_{i}}$, each element is normalized by its area when calculating the columns of $F$. Note that $F$ inherits sparsity from the impedance matrix and can be effectively stored in a compressed sparse column format.

For applications in teleseismic tomography the approximate Hessian is typically singular to working precision so we introduce a regularization term, here taken as an inverse model covariance matrix $C_{M}^{-1}$

$$\Re\{[\mathbf{J}^{T}C_{D}^{-1}\mathbf{J}]\mathbf{p}_{k} = -\eta_{k}\Re\{[\mathbf{J}^{T}C_{D}^{-1}\delta\mathbf{d}] + C_{M}^{-1}(m_{k} - m_{0}) \},$$

where the $k$ subscript refers to the current iteration. To conform with Tarantola (2005) we have included the use of an a priori model $m_{0}$. Technically this formulation changes our objective function in (14) so that the solution of (17) corresponds to a maximum likelihood model (e.g. Aster et al. 2005). We observe in practice that our search directions are typically poorly scaled and obtain an $\eta$ heuristically or through a line search such as that proposed by Moré & Thuente (1994). This line search requires a gradient as well as an objective function evaluation. During the line search, when no Jacobian is required, it is faster to calculate a gradient through an adjoint state (Appendix C). Eq. (17) is still demanding on memory usage for a single processor, but can be efficiently solved with distributed memory, for example with the linear algebra package ScalAPACK (Blackford et al. 1997).

Since our search direction is calculated directly from the Gauss–Newton equations we must calculate the Jacobian explicitly. To do so, we first define $G \equiv -AS^{-1}$, where the $[n_{\text{dof}} \times 3 \cdot n_{\text{rec}}]$ matrix $G$ denotes the Green’s functions at a receiver from every point in the medium,
that is the scaled response of any point in the medium to a source collocated at the receiver. If we post multiply by \( S \) and then transpose the corresponding equation we obtain

\[
S^T G^T = (LU)^T G^T = -A^T
\]

(18)

which can easily be solved for a modest number of right-hand sides. After transposition \( G \) can then multiply the sparse matrix of virtual forces to obtain the requisite \([3 \cdot n_{\text{rec}} \times m]\) weighted Jacobian

\[
J = GF = -AS^{-1} F = AJ.
\]

(19)

As with the adjoint method we recall that the \( LU \) factorization of \( S \) is a remnant of the forward problem and so is readily available for the \( 3 \cdot n_{\text{rec}} + 1 \) solutions required to form eq. (19) (the +1 equation represents the forward wavefield solution used when forming \( F \)). Note that the Gauss–Newton systems need not be explicitly formed and instead can be solved in an implicit iterative fashion via the truncated Gauss–Newton method (Métivier et al. 2012). However since our inversions generally involve only a few thousand variables it is more efficient to solve the Gauss–Newton system in a direct fashion than run multiple forward problems as may be required by the truncated Gauss–Newton method.

If we consider the gradient \( g = \nabla E = \Re \{ F^T G (d - Au) \} \) with regards to eq. (15) we see the right-hand side is the backpropagation of weighted residuals cross-correlated with the forward scattered field. Here \( G^T \) corresponds to the temporal and spatial reciprocity, and \( G^T \) purely spatial reciprocity, in the Green’s functions (e.g. Aki & Richards 2002). Elements of the gradient \( g \) for a particular frequency, receiver, and source are

\[
g = \left\{ \begin{array}{l}
u^T \left[ \left( \frac{\partial G}{\partial m} \right)^T (d^w_w - u^*_w) \right] \\
u^T \left[ \left( \frac{\partial G}{\partial w} \right)^T (d^w_w - u^*_w) \right]
\end{array} \right\}.
\]

(20)

Here \( d^w_w \) is the conjugated data, and \( u^*_w \) the conjugated estimate wavefield, at a receiver backpropagated into the wavefield space. Both are of dimension \([n_{\text{dof}} \times 1]\). Using the imaging principle of Claerbout (1971), the gradient is the difference between data migrated image and estimate migrated image with the inclusion of a weighting from the differential operator. As Tarantola (1984) observed, migration is the first step in inversion and consequently the two are fundamentally linked. Since the Jacobian is acting on the residual, we see that the Jacobian is the migration operator (Kim et al. 2011). Since backpropagation involves geometric spreading the residual wavefield will be relatively small close to the receiver. A diagonal approximation to the Hessian then acts to restore the true model perturbations associated with small amplitude data which result from geometric spreading (Shin et al. 2001). However, because the data is band limited, Hessians will decay away from the main diagonals in a form analogous to a smoothing matrix. The utility of the inverse of the approximate Hessian matrix applied to the gradient is to then act as a sharpening filter that refocuses the gradient (e.g. Pratt et al. 1998).

4 EXAMPLES

In this section, we discuss examples illustrating the forward and inverse algorithms. The first examples are simple assessments of the forward problem where we compare waveform computations for body and surface waves in media for which we have analytic solutions. To demonstrate the potential effects of topography on the teleseismic wavefield, we show an example of an undulating surface on top of a 1-D media. The inverse examples illustrate the advantages of different approaches in the cases of body waves only, surface waves only, and joint inversions for heterogeneous media. We also examine the effects of solving for an unknown source time function on the images produced.

To mimic actual data analysis scenarios we adopt topographies and recording geometries from existing deployments in areas where this type of technique may be usefully applied. Specifically, we use network distributions from the CASC93 (NaBelek et al. 1993) and MANAS (e.g. Li et al. 2009) deployments to represent subduction zone and continental collision environments, respectively.

While the specification of independent shear (\( \beta \)) and compressional (\( \alpha \)) wave speeds is conceptually straightforward, for reasons of expediency we explicitly fix the \( \alpha/\beta \) ratio to \( \sqrt{3} \) in these examples and show results only for the shear wave speed \( \beta \). We note that the types of examples we discuss in this paper are analogous to \( P \) receiver function and surface wave inversions, which are relatively insensitive to the compressional wave speeds. Hence, even though we could make them independent, solving for both wave speeds is often not useful for these types of data sets. Similarly, as these types of inversions depend weakly on density, we couple the density model to the current \( P \) wave speeds using an empirical relationship described in Roecker et al. (2004).

In general, the element size is based on the velocity and frequencies modelled so that the anchor node spacing varies at the scale of the expected resolution of the FWI. Analogous to the finite difference case, we set the maximum anchor node spacing to a quarter of the shortest modelled wavelength. Numerical tests indicate that at this spacing, polynomial orders greater than four produce negligible improvement in the solution.

\[1\] The wavefield \( u \), unlike its backpropagated counterparts, will approximate a plane wave in teleseismic applications.
Figure 2. Tests of the forward problem in 1-D and 2-D media. Top panel: model used for generating synthetic seismograms. Green and red regions correspond to the external and internal regions of the model, respectively, with $\alpha = 8000 \text{ m s}^{-1}$. The blue slab in the internal region has $\alpha = 8800 \text{ m s}^{-1}$. In all examples, $\alpha/\beta = \sqrt{3}$. In the 1-D test, the (infinite) blue slab extends across the entire width of the model (0–500 km); in the 2-D test, the (finite) slab is confined to the interior region as shown in this figure. Bottom panel: synthetic seismograms calculated for a $P$ wave incident at the bottom of the medium at $20^\circ$ from vertical, as shown by the back lines at the base of the interior region, and recorded at a seismic station located by the yellow triangle at the surface. Solid black lines (exact) correspond to those calculated for the 1-D infinite slab using the Thomson–Haskell algorithm, red dashed lines to a numerical representation of the infinite slab, and blue dashed lines to a numerical representation of the finite slab. Horizontal and vertical components of displacement are plotted at the same scale; the difference between both the infinite and finite numerical seismograms with the exact seismograms are plotted beneath the corresponding components.

All computations were performed on two nodes of Penguin on Demand’s MT1 cloud server. Each node consists of dual 6 core Intel X5650 3.1 GHz processors and 48 Gb of RAM. Simulations were run with 12 processes per node and 4 processes per frequency group. For typical global impedance matrices with 180 000–230 000 unknowns, wall-clock times for forward simulations were less than 5 min. Thus, each simulation used less than 3 core hours or a cost of about US $ 0.36 per simulation.

4.1 Body waves in 1-D and 2-D media

As a simple test of the accuracy of our forward problem, we repeat an example from Roecker et al. (2010) and compute teleseismic body waves in a 20-km-thick horizontal slab that has $P$ and $S$ wave speeds 10 per cent higher than the background (Fig. 2). The model space is 500 $\times$ 148 km with a 400 $\times$ 100 km internal region. Anchor nodes are spaced every 4 km in both $x$ and $z$, and we use a fourth order polynomial for interpolation. The frequency band consists of 80 frequencies equally spaced between 0.01 and 0.80 Hz. By extending the high wave speed slab laterally across the entire model space, we obtain a 1-D model with which we can compare numerical seismograms directly with an analytical response computed using Thomson–Haskell (Haskell 1960, 1962). In this case the background wavefield excitation in the Bielak layer is everywhere the same as the total wavefield. As a test in 2-D media, the vertical edges of the (finite) slab are moved 20 km into the internal region. In this case the background wavefield excitation corresponds to that of a half-space, and the slab generates a scattered field only from the interior. While this model is no longer 1-D, if the algorithm is working properly it should be the case that the seismograms at stations located far from the edges resemble the 1-D analytic waveforms. A comparison of seismograms (Fig. 2) shows very good agreement between the analytical 1D, numerical 1-D, and numerical 2-D results. The mismatches are due mostly to slight phase shifts in either the primary or scattered fields. We note that the scattered fields in the 1-D and 2-D numerical seismograms shift identically, while the main phase in the 2-D seismogram is shifted relative to the analytical and numerical 1-D seismograms.

4.2 Surface waves in a layer over a half-space model

As a test of the surface wave part of the forward problem, we construct a model with a 40-km-thick crust ($\alpha = 6500 \text{ m s}^{-1}$) over a half-space mantle ($\alpha = 8000 \text{ m s}^{-1}$) using the same mesh as that described above for the body wave example (Fig. 2). The source is a 10 km deep earthquake located about 1100 km to the left of the mesh. The frequency band consists of 88 frequencies evenly spaced between 0.0065 and

http://www.penguincomputing.com/services/hpc-cloud/pod
2.5-D full waveform teleseismic tomography

Figure 3. Synthetic seismograms computed for the vertical component of fundamental mode Rayleigh waves in a 40-km-thick layer with \(a = 6500 \text{ m s}^{-1}\) over a half-space with \(a = 8000 \text{ m s}^{-1}\) \((\alpha/\beta = \sqrt{3})\). The model has the same vertical and lateral dimensions as those shown in Fig. 2, and the seismogram is recorded at the same station (yellow triangle in Fig. 2). The horizontal axis is traveltime from an event located about 1100 km to the left of the model. The solid black line is a seismogram computed using the locked mode algorithm of Gomberg & Masters (1988), the red line is the numerical seismogram, and the blue line is the difference between the two, all plotted at the same scale. Note that the phase at about 550 s in the numerical seismogram is a reflection from the right (far) side of the model.

0.05 Hz. For attenuation (eq. 9) we choose a reference frequency \(\omega_0 = 20\pi\), and compressional and shear quality factors \(Q_\alpha = 600\) and \(Q_\beta = 300\).

The numerical estimate of the vertical component of the fundamental mode Rayleigh wave (Fig. 3) agrees well with the analytical solution, especially in phase; most of the misfit results from a slight numerical overestimate of the amplitudes in the main part of the wave train between about 250 and 500 s. At the same, there is a late arriving artefact at about 550 s due to reflection from the far (right) side of the model that suggests the PML is not entirely effective. We incorporated a convolutional PML (eq. 8a) with \(\kappa = 1.5\) and \(f_0 = 0.035\) Hz, and after some trial and error testing, set the damping function to \(d(x_i) = 2(x_i^2/L^2)\). Tests show that while the convolutional PML performs better than its non-convolutional counterpart, it does not eliminate small amplitude numerical reflections entirely. Increasing the size of the PML from 10 to 13 anchor nodes reduced the amplitude of the reflections only marginally. Hence, while the contribution of this artefact to the overall wave train is relatively minor, one might be advised to keep recording stations away from the boundaries so that such reflections can be easily identified and windowed before inverting.

Finally, because surface wave inversions often presume smooth gradients from crust to mantle, we repeated this test with wave speeds linearly increasing from 36 to 60 km depth. The results obtained were virtually identical to the layer over a half-space model described here, showing that either type of parametrization is feasible.

4.3 Effects of topography

As an example of the potential effects of topography on scattered waves generated by a teleseismic body wave, we compute wavefields in a 1-D model with a flat surface and the same model bounded by an irregular free surface (Fig. 4). In this case we take a topographic profile that is representative of what one might find in the vicinity of a subduction zone. A record section showing the horizontal component of motion (Fig. 4) shows that the effects on the scattered field by topography are comparable to that from heterogeneity. As may be expected, the largest effects are associated with areas of high relief, which tend to generate surface waves.

4.4 A body wave inversion example

As an example of inversion of teleseismic body waves, we attempt to retrieve the model in Fig. 5 (top panel) from a starting half-space model with topography. As before, \(\alpha/\beta\) is fixed at \(\sqrt{3}\) and density is updated using the heuristic formula of Roecker et al. (2004). We calculated synthetics for 4 teleseismic events, two of which approach from the right side of the model at azimuths of 162° and 188°, and two from the left at azimuths of 11° and 26°. The corresponding angles of incidence are 20°, 25°, 20° and 30°. All source time functions are Ricker wavelets with dominant frequencies varying from 0.16 to 0.19 Hz. The inversion solves for phase and amplitude for 12 frequencies evenly spaced between 0.11 and 0.22 Hz. The frequency interval in each example is chosen so that the corresponding time window captures all of the modelled energy without wraparound. It may be the case that the information provided to an inverse problem by such a dense band is redundant, but including all relevant frequencies in an inversion of real data can act as a hedge against noise.

In the Gauss–Newton optimization scheme we incorporate a Gaussian model covariance matrix (eq. 17) defined as

\[
C(m)_{ij} = v^2w_iw_j \exp \left[ \frac{-(x_j-x_i)^2 + (z_j-z_i)^2}{2\sigma^2} \right],
\]

(21)
where $\nu^2$ is defined as the expected variance in wave speed, $w$ are weights $(0, 1]$ at each node $(i,j)$, and $\sigma$ is a spatial correlation length. In this implementation the Gaussian covariance matrix is explicitly formed and inverted. Because the elements of this matrix are large at the inversion nodes and gradually taper with distance from them, they reflect the structure of $J^T J$ arising from the use of bandlimited data. For this example we heuristically set $\sigma = 2.5$ km, and chose all weights $w_i$ to be 1. The spatial correlation window is the same scale as the mesh, so that in effect it serves as a nearest neighbour averaging scheme. The variance $\nu$ is chosen so that the associated singular values of $C^{-1} \nu^2$ are of the same order as the Gauss–Newton matrix, resulting in a smooth, low spatial frequency model update. We note, however, that the covariance matrix becomes numerically indefinite for larger windows. Since the resulting search and gradient directions are both poorly scaled, we estimate the initial step length so that the maximum velocity perturbation matches the true model perturbation of 10 per cent. Although this is clearly biased, we could have obtained this perturbation by overestimating it and then backtracking during the line search. Alternatively, we could achieve a similar result with multiple iterations of smaller perturbations.

Comparison of results from the Gauss–Newton scheme with those from the gradient method (Fig. 5), shows that while the gradient method performs well it would lead one to conclude that there is significant structure near the receivers. This artefact most likely is generated by a combination of multiples, surface wave conversions, and geometric spreading associated with the backpropagation step. In contrast, the search direction from the solution to the normal equations (Fig. 5, bottom panel) refocuses primary scattered energy on the true deeper target. This effect is most notable for the faster region on the left-hand side of the model. Additionally, below the slow perturbation the gradient method generates an artefact at depth similar to those found in single scattering schemes for receiver function migration. Once again, however, this artefact is mitigated by the approximate Hessian correctly focusing energy at depth.
Despite the non-linear nature of the inversion, we achieved a 90 per cent reduction in the objective function after one iteration with this search direction. The results show that we can resolve the targets well both laterally and vertically, giving the full waveform inversion of teleseismic body wave data a significant advantage over other imaging methods such as traveltime tomography. The FWI advantage lies in its exploiting the information in P–S mode conversions. As a test we repeated this inversion using data windowed just around the onset, and found that the search directions had poor vertical resolution in a manner reminiscent of traveltime tomography.

4.4.1 Effect of the source time function

A seismogram is a convolution of media Green’s functions with an instrument response and a source time function (STF). While the instrument response typically is known and can be deconvolved from a seismogram, the STF, which essentially accounts for all the signal modification outside of the modelled region, is generally unknown and must be either eliminated or somehow estimated from the data. A typical approach in receiver function analysis is to eliminate the STF by deconvolution of one component of motion with respect to another at a given station. One could adapt that approach to the current scheme by modelling the deconvolved component of motion as a ratio of Green’s functions. Alternatively, we can follow Pratt (1990) and Roecker et al. (2010), and estimate the STF \( s(\omega) \) from

\[
s(\omega) = \frac{\hat{g}(\omega)\hat{d}(\omega)}{\hat{g}(\omega)\hat{g}(\omega)}
\]

where \( \hat{g} \) is a \( 3 \times n_{\text{rec}} \) length vector representing the estimated Green’s function responses on all receivers and components of motion for a given source and \( \hat{d} \) corresponds to the corresponding observations for the same source. Note that the operation in eq. (22) is essentially a deconvolution of the observed wavefield by the Green’s functions, so is somewhat analogous to receiver functions. However, it takes advantage of the commonality of the STF to all components at all stations that record a given event. Hence one would expect a more robust accounting of source effects with this approach than may be expected from single station receiver function deconvolution.

In each of the examples discussed here we compared models generated using a known STF with one determined by first solving for the STF first and using it as an input to the forward problem. In each case we found that the recovered STF was close to the correct value; an example from the inversion discussed in the previous section is shown in Fig. 6, and an example of recovering a surface wave and body wave STFs (from the joint inversion example discussed below) is shown in Fig. 7. As we are able to recover the true STF, the images we retrieve
Figure 6. Comparison of the true and recovered source time function (STF) in the body wave inversion example. Top panel: the true STF (black), and estimates prior to (blue) and after (red) model perturbation. Bottom panel: the difference between true and estimated STF prior to (blue) and after (red) model perturbation.

Figure 7. The true (black), recovered (red) and residual (blue) waveforms for the (top panel) surface wave and (bottom panel) body way source time functions after the second iteration of the joint surface wave/body wave inversion. There is no discernible difference in the source time function updates from one iteration to the next.

will be quite similar. These results suggest that as long as an event is recorded by multiple components at several stations, the solution for the STF may not be strongly coupled to perturbations to the wave speed model.

4.5 A surface wave inversion example

As an example of a pure surface wave inversion we use the model and receiver distribution shown in Fig. 8(top panel), which is based on the topography and receiver distribution from the Middle AsiaN Active Source (MANAS) project in the Tien Shan (e.g. Li et al. 2009).

Figure 8. Top panel: true shear ($\beta$) wave speed model for the surface wave inversion example. The background velocity is 4400 m s$^{-1}$ and the perturbations are $\pm 7.5$ percent. Bottom panel: the nodal weights used in defining eq. (21) to reduce the influence of small singular values near the edges of the network. Black triangles in both plots indicate the 40 receiver positions.
2.5-D full waveform teleseismic tomography

Figure 9. Top panel: the gradient for the second iteration in the phase and amplitude surface wave inversion normalized to ±1. The gradient has been negated to conform with the search direction. Middle panel: the search direction for the second iteration in the phase and amplitude inversion normalized to ±1. Although the gradient is sensitive to both the slow target on the right and the fast target on the left, their lateral locations are poorly resolved. In contrast, the Gauss–Newton search direction provides much better focusing of the residuals, even though some smearing results from sources approaching exclusively from the right. The gradient indicates nonexistent perturbations at the ends of the model, while the search direction, largely because of the weights in the model covariance matrix, has eliminated such artefacts. Bottom panel: recovered shear wave speed model after two iterations. The recovered model shows good vertical resolution at shallow depths but increasingly degrades at depth. This is likely the result of the reliance on longer periods and the decay of surface wave sensitivities as a function of depth. Smearing of the slow target is likely the result of sources approaching the model exclusively from the right. Note that the deeper artefacts in the second Gauss–Newton search direction are not prominent in the recovered image because significantly smaller perturbations resulted from the second step.

The data consists of 6 fundamental mode Rayleigh wave sources approaching from the right side of the model. The source time functions are Ricker wavelets with dominant frequencies of 0.013 Hz, and the inversion band consists of 24 frequencies evenly spaced between 0.022 and 0.0335 Hz or approximately 50–30 s period. Numerical experiments indicate that areas near the end or outside of the network can be associated with small singular values that can produce large fluctuations during inversion. To mitigate this we decrease the a priori certainty in the initial model via weighting in the model covariance matrix in eq. (21) as shown graphically in Fig. 8 (bottom panel).

We performed two iterations of the surface wave inversion for both the gradient and Gauss–Newton methods. Additional iterations do not result in any significant improvement to the objective function for these synthetic data sets. As with the body wave example, comparison of the gradient and the Gauss–Newton search directions (Fig. 9) after the second iteration clearly demonstrates the superior focusing of the Gauss–Newton solution. While the gradient is sensitive to the model perturbations, its lateral resolution is poor. The Gauss–Newton search direction shows some lateral ambiguity, but given that all of the illumination is from one side of the model, the result is surprisingly good. Vertical resolution of the recovered model decreases with depth, which can be expected of fundamental mode surface waves because of the frequency dependent decay of amplitudes. Sensitivity at depth requires longer periods, which will tend to smooth any heterogeneity.

4.6 A joint inversion example

As a final example we invert the model in Fig. 10 (top panel) using both body and surface waves generated by four events at azimuths of 11°, 13°, 184° and 185°. All body waves have a 25° angle of incidence and an inversion frequency band consisting of 28 frequencies evenly spaced between 0.09 and 0.28 Hz. The surface waves are all fundamental mode Rayleigh waves with an inversion frequency band of 16 frequencies evenly spaced between 0.022 and 0.0295 Hz.

Because spectral domain inversions tend to be more stable when they are done with incrementally increasing frequency bands, and because surface waves generally are recorded at longer periods than body waves, we start by modelling the surface waves first. In addition to constructing a gradient (Fig. 11), and calculating a Gauss–Newton search direction using raw residuals as was done in the previous example, we also attempt to increase resolution by inverting differential residuals. Surface waves integrate phase differences as they traverse a heterogeneous medium; hence a lack of change in residual reflects a lack of heterogeneity. Moreover, because the initial phase of the STF is common to all receivers, differencing phase residuals reduces dependence of the result on the STF. We implement this approach by applying a roughening operator $L$ built from the finite element discretization of the 1-D Laplace operator $\frac{\partial^2}{\partial x^2}$, where each element node coincides with the receiver offset location and the basis functions are hat shape functions. From eq. (14), we can define an inverse data covariance matrix
Figure 10. Top panel: true shear wave speed model used in the joint inversion example. The background velocity is 4400 m s$^{-1}$ and the perturbations are ±5 percent. Bottom panel: the nodal weights used in defining the covariance in eq. (21). Black triangles in both plots locate the 40 receiver positions.

Figure 11. Top panel: the gradient normalized to $[-1, 1]$ and negated for the first iteration. The gradient is sensitive to a fast perturbation overlying a slow perturbation in the middle but also incorrectly backpropagates the residual near the model edges. Middle panel: the search direction normalized to $[-1, 1]$. Surprisingly, the deeper target is well resolved vertically while the shallow target migrates upwards. Artefacts near the model edges have been mitigated with the data covariance matrix. Bottom panel: the search direction resulting from application of a roughening operator to the data. Note the improved lateral resolution and the suppression of artefacts within the model.

as $C_0^{-1} = \Sigma^{-1}L \Sigma^{-1}$, where for unity data weights, $\Sigma^{-1} = I$. We design $L$ to be a differencing matrix such that $L\delta d$ emphasizes receivers with quickly varying phase residuals. Comparison of models generated by these two approaches (Fig. 11) show that while the images are similar, the edges of the heterogeneous bodies are noticeably better defined when phase differencing is used. We would suggest applying this approach to phase residuals at an early stage of imaging to generate more realistic starting models.

Using the image generated with the surface waves, we then jointly inverted the original four surface waves with the four body ($P$) waves. Relative weights, applied to prevent one wave type from dominating the search direction, were optimized by trial and error so that the amplitude spectrum of the body waves were one fourth that of the surface waves. Note that because of the fast over slow structure in the true model, the phase residuals on the vertical (primary) component for body waves largely cancel at lower frequencies (effects of mode conversions become evident only above about 0.4 Hz), but are strong from the $P$–$S$ mode conversions on the horizontal components. Indeed, misfits to these converted phases clearly dominate the search direction and shift the location of the fast region down to its true position.
2.5-D full waveform teleseismic tomography

Figure 12. The model update after iteration two for the joint surface and body wave inversion. Note that body wave mode conversions have located the upper perturbation to the correct depth.

Figure 13. Top panel: the gradient normalized to $[-1, 1]$ and negated for the first iteration starting from a half-space model with body waves only. The gradient is sensitive to a fast perturbation overlying a slow perturbation in the middle but also incorrectly backpropagates the residual near the model edges. Bottom panel: the corresponding search direction normalized to $[-1, 1]$. While this search direction can identify the fast perturbation it appears to mislocate the lower slower region above the fast region. However, this is most likely a result of mode conversions being more sensitive to gradients in wave speed than their absolute values. In any event, for shallow velocity inversions it will be advantageous to first use surface waves, then incorporate the receiver function aspect of body wave inversions to resolve vertical interfaces.

We note that this was an unexpected result. Joint body and surface wave inversion schemes typically are performed to take advantage of complimentary resolution strengths, with body waves usually being better at lateral resolution and surface waves at vertical resolution. In this case, the mode conversions from the body waves significantly enhanced the vertical resolution of the surface waves.

To further demonstrate the advantages of joint inversion, we repeated this example using only the body waves. In this case, while the faster wave speed body is recovered reasonably well, the deeper, slower body is almost completely undetected. The apparent low wave speed anomaly appearing above the high wave speed body most likely is a result of mode conversions being more sensitive to gradients in wave speed than their absolute values. Hence, while the body wave only image is still informative, it is clearly inferior to that obtained from joint inversion with surface waves (Fig. 13).

5 CONCLUSION

We have introduced a variational method for the accurate modelling of teleseismic body and surface waves along quasi linear arrays in 2.5 dimensions. This approach is flexible enough to allow for realistic data collection environments (e.g. off axis sources and receivers) while still being tractable for modest computational facilities. The forward modelling algorithm was tested against analytic solutions for body and surface waves. We also explored the effects of topography on teleseismic body waves, and found that even for relatively low frequency bands a modest free surface variation can produce significant scattered waves, principally in the form of surface waves. While we examined only a single type of teleseismic surface wave (fundamental mode Rayleigh waves) in this paper, application to other types (e.g. higher mode Rayleigh or Love waves) is straightforward. Note that the 2.5-D approach would also allow for direct modelling of multipathing by summing contributions from different azimuths. Similarly, while we examined only the response to teleseismic compressional body wave excitation in this paper, specification of an incident shear wave is straightforward.

We obtained a full waveform inversion scheme through the discretization and optimization approach that is characteristic of discrete adjoint methods. We tested the Gauss-Newton scheme and observed that reasonable results can be obtained for a modest number of frequencies,
sources, and iterations as well as a conservative frequency band. The Gauss–Newton approach is clearly superior to the gradient method in focusing residuals. Additionally, we observed how features in the Bayesian inversion formulation can be used as tools not only to regularize the Gauss–Newton system but to mitigate common artefacts. Finally, we observe that the combination of long period surface waves and short period body waves can be an effective tool in resolving structure both laterally and horizontally beneath an approximately 2-D receiver network.

ACKNOWLEDGEMENTS

This work was supported at various stages by National Science Foundation grants CMG-0327634, EAR-0838384 and EAR-1114147. The surface wave part of this work started while one of the authors (Roecker) was a Leverhulme Visiting Professor at Cambridge University and he thanks the Leverhulme Trust for their support and Keith Priestly for his advice in developing this part of the algorithm. We also thank Penguin Computing’s cloud computing program and technical support which greatly expedited development of the code. Finally, we received very useful critical feedback on this manuscript from Stephane Operto, Jean Virieux and an anonymous reviewer.

REFERENCES

APPENDIX A: THE 2.5-D ELASTIC FINITE ELEMENT MATRICES AND ISOPARAMETRIC ELEMENTS

In this appendix, we present a brief overview of the element matrices and a way towards parametrizing and integrating through isoparametric elements.

A1 Element matrices

The element matrices corresponding to the discretization in eqs (4a)–(4c) with PML factors described in eqs (8a) are

\[ m^{e}_{11} = -\omega^2 \int_{\Omega_e} \left( 1 - p^2 \beta^2 \right) \phi \phi_{\Omega} \partial \Omega_e \] (A1a)

\[ m^{e}_{22} = -\omega^2 \int_{\Omega_e} \left( 1 - p^2 \alpha^2 \right) \phi \phi_{\Omega} \partial \Omega_e \] (A1b)

\[ m^{e}_{33} = -\omega^2 \int_{\Omega_e} \left( 1 - p^2 \beta^2 \right) \phi \phi_{\Omega} \partial \Omega_e \] (A1c)

\[ k^{e}_{11} = \int_{\Omega_e} \alpha^2 \frac{\beta^2}{\gamma_{\alpha}^2} \frac{\partial \phi_{\Omega}}{\partial x} \frac{\partial \phi_{\Omega}}{\partial x} + \beta^2 \frac{1}{\gamma_{\beta}^2} \frac{\partial \phi_{\Omega}}{\partial z} \frac{\partial \phi_{\Omega}}{\partial z} \partial \Omega_e \] (A1d)

\[ k^{e}_{22} = \int_{\Omega_e} \beta^2 \left( \frac{1}{\gamma_{\beta}^2} \frac{\partial \phi_{\Omega}}{\partial x} \frac{\partial \phi_{\Omega}}{\partial x} + \frac{1}{\gamma_{\beta}^2} \frac{\partial \phi_{\Omega}}{\partial z} \frac{\partial \phi_{\Omega}}{\partial z} \right) \partial \Omega_e \] (A1f)

\[ k^{e}_{33} = \int_{\Omega_e} \alpha^2 \frac{\beta^2}{\gamma_{\alpha}^2} \frac{\partial \phi_{\Omega}}{\partial z} \frac{\partial \phi_{\Omega}}{\partial z} + \beta^2 \frac{1}{\gamma_{\beta}^2} \frac{\partial \phi_{\Omega}}{\partial x} \frac{\partial \phi_{\Omega}}{\partial x} \partial \Omega_e \] (A1g)

\[ k^{e}_{12} = -i \omega p \int_{\Omega_e} \left( \alpha^2 - 2 \beta^2 \right) \frac{1}{\gamma_{\alpha}} \frac{\partial \phi_{\Omega}}{\partial x} - \beta^2 \frac{1}{\gamma_{\beta}} \frac{\partial \phi_{\Omega}}{\partial z} \partial \Omega_e \] (A1i)

\[ k^{e}_{21} = i \omega p \int_{\Omega_e} \left( \alpha^2 - 2 \beta^2 \right) \frac{1}{\gamma_{\alpha}} \frac{\partial \phi_{\Omega}}{\partial x} - \beta^2 \frac{1}{\gamma_{\beta}} \frac{\partial \phi_{\Omega}}{\partial z} \partial \Omega_e \] (A1j)
A2 Isoparametric elements

Since we do not have access to the global derivatives in eqs (A1a)–(A1l) we obtain them with the use of isoparametric elements (Hughes 2000). The idea is to create a map from the unit square $(\xi, \eta) \in [-1, 1] \times [-1, 1]$ to the parent domain. We express $(x, z)$ in terms of $(\xi, \eta)$ by linear interpolation

\[ x(\xi, \eta) = \sum_{a=1}^{4} N_a(\xi, \eta) x_a^e \]  
\[ z(\xi, \eta) = \sum_{a=1}^{4} N_a(\xi, \eta) z_a^e, \]  

where $N_a(\xi, \eta)$ are the ‘hat’ shape functions

\[ N_a(\xi, \eta) = \frac{(1 + \xi_a)(1 + \eta_a)}{4} \]  

for \( \{ (\xi_1, \eta_1), (\xi_2, \eta_2), (\xi_3, \eta_3), (\xi_4, \eta_4) \} = \{ (-1, -1), (1, -1), (1, 1), (-1, 1) \} \). With the chain rule we can write the global shape function derivatives as

\[ \begin{bmatrix} \frac{\partial \phi_a}{\partial x} & \frac{\partial \phi_a}{\partial z} \end{bmatrix} = \begin{bmatrix} \frac{\partial \phi_a}{\partial \xi} & \frac{\partial \phi_a}{\partial \eta} \end{bmatrix} J^{-1} \begin{bmatrix} \frac{\partial z}{\partial \xi} & \frac{\partial z}{\partial \eta} \end{bmatrix}, \]  

where the last term arises from a matrix inversion. The the inversion to exist we need the Jacobian, $j$,

\[ j = \frac{\partial x}{\partial \xi} \frac{\partial z}{\partial \eta} - \frac{\partial x}{\partial \eta} \frac{\partial z}{\partial \xi} > 0 \]  

to be greater than 0. This is satisfied if all interior angles of the quadrilaterals are less than 180°. The terms on the right can be calculated by differentiating (A2a) and (A2b) and evaluating at the global anchor node locations $x_a$ and $z_a$ from the finite element mesh.

Finally, we can integrate the element terms approximately with a Gauss–Lobatto quadrature rule that is accurate for polynomial orders $p = 2n_m - 3$ for $n_m$ the number of integration points using

\[ \int_{\Omega_i} f(x, y) d\Omega_e = \int_{-1}^{1} \int_{-1}^{1} f[x(\xi, \eta), z(\xi, \eta)] j(\xi, \eta) d\xi d\eta \approx \sum_{k=1}^{n_{2m}} \sum_{l=1}^{n_{2m}} f[x(\xi_l, \eta_k), z(\xi_l, \eta_k)] W_l W_k. \]  

**APPENDIX B: DERIVATIVES OF ELEMENT MATRICES**

Here we give the derivatives of the element matrices with respect to compressional wave speed. The matrices can be derived by finite element formulation of elastic wave equation in terms of compressional wave speed where the shear wave speed is fixed by a proportionality constant $R$. For a Poisson solid $R$ is $\sqrt{3}$. One can then differentiate the dynamic equation of motion with respect to compressional wave speed, discretize with finite elements, and obtain the following derivative element matrices.

\[ \frac{\partial m_{ij}}{\partial \alpha} = 2 \omega^2 \rho_i^2 \int_{\Omega_i} \frac{\alpha}{R^2} \phi_i \phi_j d\Omega_e \]  
\[ \frac{\partial m_{ij}}{\partial \alpha} = 2 \omega^2 \rho_i^2 \int_{\Omega_i} \alpha \phi_i \phi_j d\Omega_e \]  
\[ \frac{\partial \delta_{ij}}{\partial \alpha} = 2 \int_{\Omega_i} \alpha \frac{\partial \phi_i}{\partial x} \frac{\partial \phi_j}{\partial x} + \alpha \frac{\partial \phi_i}{\partial z} \frac{\partial \phi_j}{\partial z} d\Omega_e \]  
\[ \frac{\partial m_{ij}}{\partial \alpha} = 2 \omega^2 \rho_i^2 \int_{\Omega_i} \frac{\alpha}{R^2} \phi_i \phi_j d\Omega_e \]  
\[ \frac{\partial \delta_{ij}}{\partial \alpha} = 2 \int_{\Omega_i} \alpha \frac{\partial \phi_i}{\partial x} \frac{\partial \phi_j}{\partial x} + \alpha \frac{\partial \phi_i}{\partial z} \frac{\partial \phi_j}{\partial z} d\Omega_e \]
\[
\frac{\partial k_{ii}}{\partial \alpha} = 2 \int_{\Omega_e} \alpha \frac{\partial \phi_0}{\partial x} \frac{\partial \phi_0}{\partial z} + \frac{\alpha}{R^2} \left( \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial x} - 2 \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial x} \right) d\Omega_e. \tag{B1e}
\]

\[
\frac{\partial k_{ij}}{\partial \alpha} = 2 \int_{\Omega_e} \frac{\alpha}{R^2} \left( \frac{\partial \phi_0}{\partial x} \frac{\partial \phi_0}{\partial z} + \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial x} - 2 \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial x} \right) d\Omega_e. \tag{B1f}
\]

\[
\frac{\partial k_{ii}}{\partial \alpha} = 2 \int_{\Omega_e} \alpha \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial z} + \frac{\alpha}{R^2} \left( \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial z} - 2 \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial z} \right) d\Omega_e. \tag{B1g}
\]

\[
\frac{\partial k_{ij}}{\partial \alpha} = 2 \int_{\Omega_e} \frac{\alpha}{R^2} \left( \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial z} + \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial z} - 2 \frac{\partial \phi_0}{\partial z} \frac{\partial \phi_0}{\partial z} \right) d\Omega_e. \tag{B1h}
\]

\[
\frac{\partial k_{ii}}{\partial \alpha} = -2i \alpha \int_{\Omega_e} \frac{\partial \phi_0}{\partial x} \phi_0 - \frac{\alpha}{R^2} \left( 2 \frac{\partial \phi_0}{\partial x} \phi_0 + \phi_0 \frac{\partial \phi_0}{\partial x} \right) d\Omega_e. \tag{B1i}
\]

\[
\frac{\partial k_{ij}}{\partial \alpha} = 2i \alpha \int_{\Omega_e} \frac{\alpha}{R^2} \left( 2 \frac{\partial \phi_0}{\partial z} \phi_0 + \phi_0 \frac{\partial \phi_0}{\partial z} \right) d\Omega_e. \tag{B1j}
\]

\[
\frac{\partial k_{ii}}{\partial \alpha} = 2 \int_{\Omega_e} \frac{\alpha}{R^2} \left( 2 \frac{\partial \phi_0}{\partial z} \phi_0 + \phi_0 \frac{\partial \phi_0}{\partial z} \right) d\Omega_e. \tag{B1k}
\]

\[
\frac{\partial k_{ij}}{\partial \alpha} = -2i \alpha \int_{\Omega_e} \frac{\partial \phi_0}{\partial z} \phi_0 - \frac{\alpha}{R^2} \left( 2 \frac{\partial \phi_0}{\partial z} \phi_0 + \phi_0 \frac{\partial \phi_0}{\partial z} \right) d\Omega_e. \tag{B1l}
\]

In calculating the element matrices the anchor nodes act as Kronecker delta functions so that \(\frac{\partial \phi_0}{\partial x}\) will only include elements attached to anchor node \(i\). The wave speeds over the element are then linearly interpolated at all integration abscissas and finite element integration and assembly would continue as usual.

**APPENDIX C: A REGULARIZED GRADIENT COMPUTATION THROUGH ADJOINT STATES**

Here we follow Plessix (2006) and formulate the inverse problem in the context of constrained optimization. In practice this will allow us to obtain a gradient during the line search without the need to obtain a Jacobian. In practice, three steps are required for an adjoint method.

First, build the associated Lagrangian \(\mathcal{L}\):

\[
\mathcal{L}(\mathbf{m}, \hat{u}, \hat{\lambda}) = h(\mathbf{m}, \hat{u}) - \eta \{ F(\mathbf{m}, \hat{u}) \}^\dagger \hat{\lambda}, \tag{C1}
\]

where \(\mathbf{m}\) is a real \(n\)-dimensional vector and denotes the model, \(\hat{u}\) is a complex \(n_{\text{def}}\)-dimensional vector in the wavefield space \(\mathbb{W}\) and \(\hat{\lambda}\) is a complex \(n_{\text{def}}\)-dimensional vector in the dual wavefield space \(\mathbb{W}^\ast\). The function \(h\) is related to the objective function of eq. (14) and \(F\) is the forward modelling constraint in eq. (13). The real part is taken to conform to \(h\).

Secondly, calculate the adjoint state equations with derivatives evaluated at \([u(\mathbf{m}), \lambda(\mathbf{m})]\). Here, \(u(\mathbf{m})\) is a physical realization of the forward problem and therefore is model dependent. Likewise, \(\lambda(\mathbf{m})\) is the adjoint field corresponding to the physical realization \(u(\mathbf{m})\). The adjoint equations can be thought of as locating the stationary points of the associated Lagrangian, that is,

\[
\frac{\partial \mathcal{L}(\mathbf{m}, \hat{u}, \hat{\lambda})}{\partial \hat{u}} \bigg|_{[u(\mathbf{m}), \lambda(\mathbf{m})]} = \frac{\partial \mathcal{L}[\mathbf{m}, u(\mathbf{m}), \lambda(\mathbf{m})]}{\partial \hat{u}} = 0
\]

which, after taking the argument of the \(\eta\), gives

\[
\left[ \frac{\partial \mathcal{F}[\mathbf{m}, u(\mathbf{m})]}{\partial \hat{u}} \right] \lambda(\mathbf{m}) = \frac{\partial h[u(\mathbf{m})]}{\partial \hat{u}}.	ag{C2}\]

Thirdly, compute the gradient of \(E(\mathbf{m})\) by differentiating the augmented functional \(\mathcal{L}(\mathbf{m}, \hat{u}, \hat{\lambda})\) with respect to \(\mathbf{m}\) and evaluating at the stationary points to obtain

\[
\frac{\partial E}{\partial \mathbf{m}} = \frac{\partial h[u(\mathbf{m})]}{\partial \mathbf{m}} - \eta \left\{ \left[ \frac{\partial \mathcal{F}[\mathbf{m}, u(\mathbf{m})]}{\partial \mathbf{m}} \right] \lambda(\mathbf{m}) \right\}. \tag{C3}\]

The real is taken since \(E\) is real function.
To apply this method, let us first qualitatively describe our goal. We desire a wavefield \( \tilde{u} \) that, when extracted at the receiver locations, ‘best’ matches the observations. Furthermore, we impose a constraint that this wavefield must satisfy our seismic modelling operation. Using the \( L_2 \) norm as an objective definition of ‘best’, the previous sentences are written as

\[
\min \frac{1}{m} (d - A\tilde{u})^T C_D^{-1} (d - A\tilde{u}) + \frac{1}{2} (m - m_0)^T C_M^{-1} (m - m_0)
\]

subject to \( S(m) \tilde{u} = f \).

First write associated functional in (C1)

\[
\mathcal{L}(m, \tilde{u}, \lambda) = \frac{1}{2} (d - A\tilde{u})^T C_D^{-1} (d - A\tilde{u}) + \frac{1}{2} (m - m_0)^T C_M^{-1} (m - m_0) - \Re \{ \{ S(m) \tilde{u} - f \} \lambda \}.
\]

Next, calculate the adjoint equations

\[
\frac{\partial \mathcal{L}}{\partial \tilde{u}} = \frac{\partial}{\partial \tilde{u}} \left[ \frac{1}{2} (d - A\tilde{u})^T C_D^{-1} (d - A\tilde{u}) - \frac{1}{2} (m - m_0)^T C_M^{-1} (m - m_0) \right] - \frac{\partial}{\partial \tilde{u}} \left[ \Re \{ \{ S(m) \tilde{u} - f \} \lambda \} \right]
= -\Re \{ A^T C_D^{-1} (d - A\tilde{u}) + S(m)^\dagger \lambda \}.
\]

Since the above is 0 when evaluating at the stationary points we simply extract the argument to obtain the adjoint equation

\[
S(m)^\dagger \lambda(m) = -A^T C_D^{-1} [d - Au(m)]
\]

which, as seen before, says the adjoint wavefield is the backpropagated residual wavefield.

The gradient is then readily extracted from

\[
\frac{\partial E}{\partial m} = \frac{\partial}{\partial m} \left[ \frac{1}{2} (d - A\tilde{u})^T C_D^{-1} (d - A\tilde{u}) + \frac{1}{2} (m - m_0)^T C_M^{-1} (m - m_0) \right] - \Re \left[ \left\{ \frac{\partial}{\partial m} \right\} \{ S(m) \tilde{u} - f \} \lambda \right]
= -\Re \left[ \left\{ \frac{\partial S}{\partial m} \tilde{u} \right\} \lambda \right] + C_M^{-1} (m - m_0).
\]

Evaluating at the stationary points gives the gradient

\[
\nabla E(m) = -\Re \left[ \left\{ \frac{\partial S}{\partial m} u(m) \right\} \lambda(m) \right] + C_M^{-1} (m - m_0).
\]

We know \( \lambda(m) \) from eq. (C6), so inserting (C6) into (C7) gives the final result for the gradient

\[
\nabla E(m) = \Re \left[ \left\{ \frac{\partial S}{\partial m_1} u(m), \ldots, \frac{\partial S}{\partial m_m} u(m) \right\} \left( S^\dagger \right)^{-1} A^T C_D^{-1} (d - Au(m)) \right] + C_M^{-1} (m - m_0)
\]

as discussed in the main text. The regularization term is a matrix vector multiply and therefore of trivial expense after computing the inverse of the covariance matrix. Once again, the inverse of the impedance matrix is merely notational.