Geophys. J. Int. (2007) **168,** 1105–1129 doi: 10.1111/j.1365-246X.2006.03191.x

Finite-frequency tomography using adjoint methods—Methodology and examples using membrane surface waves

Carl Tape, Qinya Liu and Jeroen Tromp

Seismological Laboratory, California Institute of Technology, Pasadena, CA 91125*, USA. E-mail: carltape@gps.caltech.edu*

Accepted 2006 August 17. Received 2006 August 2; in original form 2006 March 17

SUMMARY

We employ adjoint methods in a series of synthetic seismic tomography experiments to recover surface wave phase-speed models of southern California. Our approach involves computing the Fréchet derivative for tomographic inversions via the interaction between a forward wavefield, propagating from the source to the receivers, and an 'adjoint' wavefield, propagating from the receivers back to the source. The forward wavefield is computed using a 2-D spectral-element method (SEM) and a phase-speed model for southern California. A 'target' phase-speed model is used to generate the 'data' at the receivers. We specify an objective or misfit function that defines a measure of misfit between data and synthetics. For a given receiver, the remaining differences between data and synthetics are time-reversed and used as the source of the adjoint wavefield. For each earthquake, the interaction between the regular and adjoint wavefields is used to construct finite-frequency sensitivity kernels, which we call *event kernels*. An event kernel may be thought of as a weighted sum of phase-specific (e.g. *P*) *banana–doughnut kernels*, with weights determined by the measurements. The overall sensitivity is simply the sum of event kernels, which defines the *misfit kernel*. The misfit kernel is multiplied by convenient orthonormal basis functions that are embedded in the SEM code, resulting in the gradient of the misfit function, that is, the Fréchet derivative. A non-linear conjugate gradient algorithm is used to iteratively improve the model while reducing the misfit function. We illustrate the construction of the gradient and the minimization algorithm, and consider various tomographic experiments, including source inversions, structural inversions and joint source-structure inversions. Finally, we draw connections between classical Hessian-based tomography and gradient-based adjoint tomography.

Key words: adjoint methods, inverse problem, seismic tomography, spectral-element method, wave propagation.

1 INTRODUCTION

Seismic tomography is in a state of transition from ray-based inversions using 1-D reference models towards finite-frequency-kernel-based inversions using 3-D reference models (Akçelik *et al.* 2003; Zhao *et al.* 2005). The transition from ray- to kernel-based inversions has been motivated in part by the pioneering studies of Marquering *et al.* (1999), Zhao *et al.* (2000) and Dahlen *et al.* (2000), which were based upon 1-D reference models but showed that seismological measurements are sensitive to structure away from the ray path and are affected by wave front healing. The transition from 1-D to 3-D reference models has been motivated by computational advances coupled with success in modelling the forward problem of seismic wave propagation in complex media (e.g. Komatitsch & Vilotte 1998; Komatitsch *et al.* 2002; Capdeville *et al.* 2003).

The purpose of this paper is to illustrate an approach for '3D–3D' seismic tomography, by which we mean seismic tomography based upon a 3-D reference model, 3-D numerical simulations of the complete seismic wavefield and finite-frequency sensitivity kernels. The success of 3D–3D tomography depends largely on two factors.

(1) The accuracy and efficiency of the technique used to generate 3-D synthetic seismograms.
(2) The efficiency of the inversion algorithm.

We have implemented numerical methods—the spectral-element method (SEM)—on parallel computers to simulate 3-D seismic wave propagation at regional and global scales (e.g. Komatitsch & Tromp 1999; Komatitsch *et al.* 2004; Komatitsch & Tromp 2002a,b).

The inverse problem can be cast as a minimization problem, where the objective or misfit function measures some difference between data and synthetic seismograms computed from a 3-D model. Our approach to the inverse problem utilizes adjoint methods (Tarantola 1984; Talagrand & Courtier 1987), which provide the gradient of the misfit function but not its second derivatives, that is, the Hessian.

The efficiency of the inverse algorithm is controlled by the computation of the gradient, which requires only two 3-D simulations per earthquake (i.e. the gradient is independent of the number of receivers or the number of measurements), as well as an effectively chosen gradient method.

The framework for 3D–3D tomographic inversions using adjoint methods was developed in exploration geophysics (e.g. Tarantola 1984; Gauthier *et al.* 1986; Mora 1987; Pratt *et al.* 1998; Pratt 1999). These studies illustrated the computation of the gradient and the related inversion technique using 2-D heterogeneous models and 2-D numerical algorithms. Applications of 3D–3D tomographic techniques are presented in Bijwaard & Spakman (2000), Zhao *et al.* (2005), Capdeville *et al.* (2005) and Akçelik *et al.* (2003), among others. Bijwaard & Spakman (2000) performed 3-D ray tracing though 3-D models to iteratively improve a global *P*-wave model. Zhao *et al.* (2005) used fully numerical methods (finite differencing) to compute traveltime misfit function gradients for 3-D models of the greater Los Angeles area. Capdeville *et al.* (2005), using synthetic data, demonstrated a technique of stacking synthetic records that limits the number of forward simulations to one per event (per model iteration); however, the technique requires modification when the data set is incomplete, as is generally the case. Akçelik *et al.* (2003), using synthetic data, illustrated a tomographic inversion using a finite-element method together with an adjoint approach within a conjugate gradient framework. They also addressed multiscale approaches to the inverse problem in an attempt to avoid reaching local minima during the inversion.

This paper is an extension of Tromp *et al.* (2005), which synthesized the work on adjoint methods with studies in finite-frequency tomography (Marquering *et al.* 1999; Dahlen *et al.* 2000; Zhao *et al.* 2000) and time-reversal imaging (Fink *et al.* 1989; Fink 1992, 1997). In Tromp *et al.* (2005) we illustrated how the computation of a sensitivity kernel for a particular model and a particular type of measurement could be achieved via the interaction of two wavefields, one constituting the 'regular' wavefield travelling from source to receiver, and the other constituting the 'adjoint' wavefield travelling from receiver to source, constructed by a suitable time-reversed synthetic seismogram recorded at the receiver. We performed a simple source inversion to illustrate the conjugate gradient algorithm, whereby only the gradient of the misfit function is used to iteratively invert for the source parameters. In this paper, we use the conjugate gradient approach to illustrate wave speed inversions, source inversions and joint (source and structure) inversions. In each example, the 'observed' seismograms are computed for a 'target' model and the synthetic seismograms are computed from a current model that iteratively improves towards the target model over the course of the inversion. All of the simulations illustrated in this paper were performed on a single Linux PC.

We begin by highlighting the differences between classical and adjoint tomography in the context of a minimization problem. We define *classical tomography* as a Newton inversion scheme that computes model sensitivities for each measurement by constructing the gradient and Hessian of the misfit function (Section 3) (e.g. Woodhouse & Dziewonski 1984; Ritsema *et al.* 1999). In *adjoint tomography* only the gradient is computed, and it is computed via adjoint methods (e.g. Gauthier *et al.* 1986; Akçelik *et al.* 2003). In Section 5, we illustrate the construction of a *misfit kernel*, which can be thought of as the gradient of the misfit function. In Section 6, we show how this gradient is used in the conjugate gradient algorithm to iteratively improve the model. We finish by showing several tomographic experiments, including simultaneous source-structure inversions, as well as a comparison between ray- and kernel-based classical inversions and adjoint tomography.

2 GENERAL FORMULATION OF THE INVERSE PROBLEM

Our objective will be to minimize a measure of the misfit between a set of data, for example waveforms or traveltimes, and a complementary set of synthetics. The synthetics are generated based upon a model $\mathbf{m}$, for example a set of structural and source parameters, and our aim is to reduce the misfit between the data and the synthetics by making (successive) model corrections $\delta\mathbf{m}$. We define the *misfit function* $\chi(\mathbf{m})$ to be a measure of misfit between the data and synthetics computed for model $\mathbf{m}$. The function χ is alternatively called an 'objective' or 'cost' function. For example, χ could represent least-squares measures of waveform or traveltime differences.

Let us suppose we have a particular model $\mathbf{m}$, and we wish to obtain an updated model $\mathbf{m}+\delta\mathbf{m}$ that brings us closer to a minimum of the misfit function χ (Nolet 1987; Tarantola 2005, appendix 6.22). We make a quadratic Taylor expansion of $\chi(\mathbf{m}+\delta\mathbf{m})$

$$\chi(\mathbf{m}+\delta\mathbf{m}) \approx \chi(\mathbf{m}) + \mathbf{g}(\mathbf{m})^T\delta\mathbf{m} + \frac{1}{2}\delta\mathbf{m}^T\mathbf{H}(\mathbf{m})\,\delta\mathbf{m}, \tag{1}$$

where the gradient vector $\mathbf{g}(\mathbf{m})$ is defined in terms of the first derivative of the misfit function (also known as the Fréchet derivative) by

$$\mathbf{g}(\mathbf{m}) = \left.\frac{\partial\chi}{\partial\mathbf{m}}\right|_{\mathbf{m}}, \tag{2}$$

and the Hessian matrix $\mathbf{H}(\mathbf{m})$ is defined in terms of the second derivatives of the misfit function by

$$\mathbf{H}(\mathbf{m}) = \left.\frac{\partial^2\chi}{\partial\mathbf{m}\partial\mathbf{m}}\right|_{\mathbf{m}}. \tag{3}$$

The '$|_{\mathbf{m}}$' dependence is used to emphasize that the preceding variable is evaluated at model $\mathbf{m}$.

The gradient of (1) with respect to $\delta\mathbf{m}$ is given by

$$\mathbf{g}(\mathbf{m}+\delta\mathbf{m}) \approx \mathbf{g}(\mathbf{m}) + \mathbf{H}(\mathbf{m})\,\delta\mathbf{m}, \tag{4}$$

which can be set equal to zero to obtain the (local) minimum of (1):

$$\mathbf{H}(\mathbf{m})\,\delta\mathbf{m} = -\mathbf{g}(\mathbf{m}). \tag{5}$$

An updated model $\mathbf{m}+\delta\mathbf{m}$ may be obtained with or without the Hessian $\mathbf{H}$. If the gradient and Hessian (or approximate Hessian) are both available, then the inverse approach is known as a *Newton method*; if only the gradient is available, then it is a *gradient method* (e.g. steepest descent, conjugate gradient). In classical traveltime tomography, one generally has access to both the gradient $\mathbf{g}$ and the Hessian $\mathbf{H}$ of the misfit function, in which case the model update $\delta\mathbf{m}$ may be obtained based upon (5). For complex, heterogeneous models, computation of the gradient is generally still feasible, but computation of the Hessian is not. In the absence of the Hessian, one can minimize the misfit function using only the gradient (2) based upon iterative methods.

3 CLASSICAL TOMOGRAPHY

We begin by investigating 2-D surface wave traveltime tomography based upon either ray or finite-frequency sensitivity kernels. These classical inversions, which involve access to both the gradient and the Hessian of the misfit function, serve as a reference and standard

for subsequent iterative inversions based upon only the gradient (Section 6). In particular, we will investigate how many iterations of the conjugate-gradient adjoint approach are required to obtain a similar misfit to the data as an inversion based upon knowledge of the gradient and Hessian. Of course our ultimate goal is to use the adjoint approach to address inverse problems for fully 3-D reference models, when the calculation of the Hessian is generally not feasible, and the experiments in this paper serve as a guide to the implementation and convergence of such iterative inversions.

3.1 Theory

The traveltime misfit function may be expressed as

$$\chi(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{N}\left[T_i^{\text{obs}} - T_i(\mathbf{m})\right]^2, \tag{6}$$

where T_i^{obs} denotes the observed traveltime for the ith source–receiver combination, $T_i(\mathbf{m})$ the predicted traveltime based upon the current model $\mathbf{m}$, and N the number of traveltime measurements. The variation of the misfit function (6) is given by

$$\delta\chi = -\sum_{i=1}^{N}\Delta T_i\ \delta T_i, \tag{7}$$

where δT_i is the theoretical traveltime perturbation and

$$\Delta T_i = T_i^{\text{obs}} - T_i(\mathbf{m}) \tag{8}$$

denotes the traveltime anomaly. The sign convention for the traveltime anomaly follows that of Dahlen *et al.* (2000) and Dahlen & Baig (2002), such that a negative traveltime indicates a delay in the synthetic arrival relative to the recorded arrival. Throughout this paper, an upper-case delta, Δ, will denote a differential measurement, and a lower-case delta, δ, will denote a mathematical perturbation.

In ray-based tomography, the predicted traveltime anomaly δT_i along the ith ray path may be related to fractional wave speed perturbations $\delta \ln c = \delta c/c$ based upon the relationship

$$\delta T_i = -\int_{\text{ray}_i} c^{-1}\,\delta \ln c\ \mathrm{d}s, \tag{9}$$

where ds denotes a segment of the ith ray.

Taking into account finite-frequency effects, Marquering *et al.* (1999), Zhao *et al.* (2000) and Dahlen *et al.* (2000) demonstrate that the traveltime anomaly may alternatively be related to relative wave speed perturbations based upon a finite-frequency sensitivity kernel $K_i(\mathbf{x})$ for the ith source–receiver combination by

$$\delta T_i = \int_V K_i \delta \ln c\ \mathrm{d}^3\mathbf{x}. \tag{10}$$

Marquering *et al.* (1999) dubbed these finite-frequency kernels 'banana–doughnut kernels' on account of their shape in smooth, spherically symmetric earth models for cross-correlation traveltime measurements. These kernels are also referred to as 'sensitivity', 'finite-frequency' or 'Born' kernels. For our purposes, the key point is that a banana–doughnut kernel does not incorporate the traveltime measurement, whereas the event and misfit kernels discussed in Section 5 do incorporate measurements.

Unlike the ray-theoretical expression (9), eq. (10) relates the traveltime anomaly to 3-D heterogeneity $\delta \ln c$ throughout the entire earth model, as seen through the kernel K_i. The relations (9) and (10) are valid for any model. Frequently the model is chosen to be 1-D because this makes the ray and finite-frequency kernel calculations much simpler, but this is not required (Zhao *et al.* 2005).

Substituting (10) into (7), we express the variation of the traveltime misfit function for finite-frequency tomography as

$$\delta\chi = \int_V K\,\delta \ln c\ \mathrm{d}^3\mathbf{x}, \tag{11}$$

where the traveltime *misfit kernel* $K(\mathbf{x})$ is a weighted sum of the kernels $K_i(\mathbf{x})$:

$$K(\mathbf{x}) = -\sum_{i=1}^{N}\Delta T_i\ K_i(\mathbf{x}), \tag{12}$$

such that the weight associated with the kernel for the ith source–receiver combination K_i is the corresponding traveltime anomaly ΔT_i. It is important to note the distinction that misfit kernels $K(\mathbf{x})$ depend upon the data, whereas the banana–doughnut kernels $K_i(\mathbf{x})$ are data-independent.

To make the tomographic inversions practical, we need to choose a finite set of basis functions in which to expand our model. Let $B_k(\mathbf{x})$, $k = 1, \ldots, M$, denote a set of M basis functions. We expand our fractional phase-speed perturbations, $\delta \ln c(\mathbf{x})$, into these basis functions:

$$\delta \ln c(\mathbf{x}) = \sum_{k=1}^{M}\delta m_k\ B_k(\mathbf{x}), \tag{13}$$

where δm_k, $k = 1, \ldots, M$, represent the perturbed model coefficients, which are determined in terms of the gradient $\mathbf{g}$ and Hessian $\mathbf{H}$ of the misfit function by (5).

Next, we determine $\mathbf{g}$ and $\mathbf{H}$ for this classical traveltime tomography problem. Substituting (13) into (9) and (10), respectively, we obtain

$$\delta T_i = \sum_{k=1}^{M}\delta m_k\ G_{ik}, \tag{14}$$

where for ray theory

$$G_{ik} \equiv \left.\frac{\partial T_i}{\partial m_k}\right|_{\mathbf{m}} = -\int_{\text{ray}_i} c^{-1} B_k\ \mathrm{d}s, \tag{15}$$

whereas for finite-frequency tomography

$$G_{ik} \equiv \left.\frac{\partial T_i}{\partial m_k}\right|_{\text{m}} = \int_V K_i B_k\ \mathrm{d}^3\mathbf{x}. \tag{16}$$

We note that in either case G_{ik} will depend on the source–receiver geometry (index i), the choice of basis functions (index k), and the choice of reference model ($\mathbf{m}$).

Using (13) we express the variation in the misfit function (11) as

$$\delta\chi = \sum_{k=1}^{M}\int_V K B_k\ \mathrm{d}^3\mathbf{x}\,\delta m_k. \tag{17}$$

Upon comparing this result to

$$\delta\chi = \mathbf{g}\cdot\delta\mathbf{m} = \sum_{k=1}^{M} g_k\,\delta m_k, \tag{18}$$

we deduce that the elements of the gradient vector, g_k, are determined by

$$g_k = \frac{\partial\chi}{\partial m_k} = \int_V K B_k\ \mathrm{d}^3\mathbf{x}, \qquad k = 1, \ldots, M. \tag{19}$$

This highlights the simple relationship between the misfit kernel and the gradient of the misfit function. Substituting (12) into (19),

we obtain

$$g_k = -\sum_{i=1}^{N} \int_V K_i B_k \, \mathrm{d}^3\mathbf{x} \, \Delta T_i$$
$$= -\sum_{i=1}^{N} G_{ik} \, \Delta T_i, \qquad k = 1, \ldots, M, \tag{20}$$

which in matrix notation becomes

$$\mathbf{g} = -\mathbf{G}^T \mathbf{d}. \tag{21}$$

Here $\mathbf{G}$ is the $N \times M$ design matrix constructed using (15) for rays or (16) for finite-frequency kernels, a superscript T denotes the transpose, and $\mathbf{d}$ is defined as an N-dimensional data vector of cross-correlation traveltime measurements:

$$\mathbf{d} = (\Delta T_1, \ldots, \Delta T_i, \ldots, \Delta T_N)^T. \tag{22}$$

Note that the data vector depends on model $\mathbf{m}$ through the synthetics.

The second derivatives of the misfit function are given by (3), and thus the elements of the Hessian $\mathbf{H}$ are given by

$$H_{kk'} = \left. \frac{\partial^2 \chi}{\partial m_k \partial m_{k'}} \right|_{\mathbf{m}} = \left. \frac{\partial g_k}{\partial m_{k'}} \right|_{\mathbf{m}}$$
$$= \sum_{i=1}^{N} \left(G_{ik'} G_{ik} + \Delta T_i \left. \frac{\partial^2 T_i}{\partial m_k \partial m_{k'}} \right|_{\mathbf{m}} \right), \tag{23}$$

where G_{ik} is defined in (16). We introduce an *approximate Hessian* $\tilde{\mathbf{H}}$ by ignoring the second-order terms:

$$\tilde{H}_{kk'} \equiv \sum_{i=1}^{N} G_{ik} G_{ik'}, \qquad k, k' = 1, \ldots, M, \tag{24}$$

which in matrix notation is

$$\tilde{\mathbf{H}} \equiv \mathbf{G}^T \mathbf{G}. \tag{25}$$

Henceforth, we will refer to $\tilde{\mathbf{H}}$ as the Hessian. This approximation, $\tilde{\mathbf{H}} \approx \mathbf{H}$, characterizes the *Gauss–Newton method* and is exact if the model perturbations are linearly related to the traveltime measurements.

Having established the gradient (21) and Hessian (25), the model correction $\delta\mathbf{m}$ is determined by (5):

$$\mathbf{G}^T \mathbf{G} \, \delta\mathbf{m} = \mathbf{G}^T \mathbf{d}, \tag{26}$$

where $\delta\mathbf{m}$ is defined in (13), $\mathbf{d}$ is defined in (22), and $\mathbf{G}$ is defined according to (15) or (16).

In general, the Hessian matrix (25) is not full rank, which means that its inverse does not exist. To stabilize the inverse problem, one introduces a damping matrix $\mathbf{D}$ typically involving the norm, gradient, or second derivative of the wave speed perturbations, and a damping parameter γ:

$$\tilde{\mathbf{H}}_\gamma = \mathbf{G}^T \mathbf{G} + \gamma^2 \mathbf{D}. \tag{27}$$

The damping parameter γ is chosen in a subjective manner, generally by inspecting a graph that trades-off misfit of the solution against complexity of the model. Having stabilized the inverse of the Hessian, the solution to (26) may now be expressed as

$$\delta\mathbf{m} = (\mathbf{G}^T \mathbf{G} + \gamma^2 \mathbf{D})^{-1} \mathbf{G}^T \mathbf{d}, \tag{28}$$

from which the updated model, $\mathbf{m} + \delta\mathbf{m}$, may be obtained. In Appendix A, we show how (28) is obtained by adding a regularization term to the misfit function. More generally, for non-linear inverse problems one uses an iterative Gauss–Newton method to minimize the misfit function. In that case (28) is replaced by an iterative expression that relates model $k + 1$ to model k and the initial model (e.g. Tarantola 2005).

3.2 Experimental set-up

We simulate 2-D elastic wave propagation using a SEM, which combines the flexible spatial parametrization of finite-element methods with the accuracy of pseudo-spectral methods (e.g. Komatitsch & Vilotte 1998; Komatitsch & Tromp 1999). For simplicity, we consider 'membrane waves' (Tanimoto 1990; Peter *et al.* 2006) travelling in the x–y plane with a vertical (z) component of motion. The elastic wave equation for the vertical component of displacement $s(x, y, t)$ is given by

$$\rho \, \partial_t^2 s = \partial_x(\mu \, \partial_x s) + \partial_y(\mu \, \partial_y s) + f, \tag{29}$$

where $\rho(x, y)$ denotes the density distribution and $\mu(x, y)$ the shear modulus. The source $f(x, y, t)$ is given by

$$f(x, y, t) = h(t) \, \delta(x - x_s) \, \delta(y - y_s), \tag{30}$$

where $h(t)$ denotes the source-time function and (x_s, y_s) the source location. All four membrane edges are absorbing, and attenuation and anisotropy are not incorporated. The relationship between membrane-wave phase-speed, c, and rigidity is $\mu = \rho c^2$.

We take southern California as our region of interest (Fig. 1) in anticipation of eventually improving the present 3-D reference wave speed models (Hauksson 2000; Magistrale *et al.* 2000; Süss & Shaw 2003). The modelled region is 480 km by 480 km. The numerical simulations are carried out on a planar grid with $N_{\mathrm{glob}} = 25\,921$ gridpoints. The source-time function of the point source (30) used in the simulations is a Gaussian of the form

$$h(t) = (-2\alpha^3/\sqrt{\pi}) \, (t - t_s) \, \exp[-\alpha^2 (t - t_s)^2], \tag{31}$$

where $\alpha = 2\tau_0/\tau$, $\tau_0 = 2.628$ s, $\tau = 20.0$ s is the duration of $h(t)$, and $t_s = 48.0$ s is the origin time (e.g. Fig. 6a). The duration of each simulation is $T = 240$ s unless otherwise noted.

The synthetic records are computed using source locations of actual events ($M \geq 4$) recorded in southern California between 1990 and 2005 (Fig. 1). The initial set of synthetics is computed using a model with homogeneous phase-speed c. In general, the synthetics in our experiments are generated from a laterally varying model, while the data are generated from what is designated as the 'target' model. Computationally, the model correction is expressed as a fractional perturbation, $\delta c/c = \delta \ln c$, with current phase-speed c. In the figures, however, each phase-speed model is plotted as a percent perturbation from the phase-speed value for the initial model. In Section 8, we allow for additional perturbations in the source parameters, so that in general the synthetics are computed from a model with perturbed sources and perturbed structure.

3.3 2-D tomographic example

To illustrate a classical tomographic inversion, we begin by choosing a set of basis functions, $B_k(\mathbf{x})$, in which to expand the fractional wave speed perturbations $\delta \ln c(\mathbf{x})$ (13). We use spherical spline basis functions (Wang & Dahlen 1995; Wang *et al.* 1998), which are well suited for regional models where multiscale parametrization is desired because of non-uniform path coverage (e.g. Boschi *et al.* 2004). (We do not exploit the multiscale aspects here.) An example of a spherical spline basis function is plotted in Fig. 2(b). We choose $M = 286$ spherical spline basis functions to cover the southern California region.

The data are computed using the phase-speed model in Fig. 1(b), and the synthetics are computed for a homogeneous phase-speed model with $c = 3.78$ km s^{-1}. We make cross-correlation traveltime measurements between data and synthetics to obtain the data

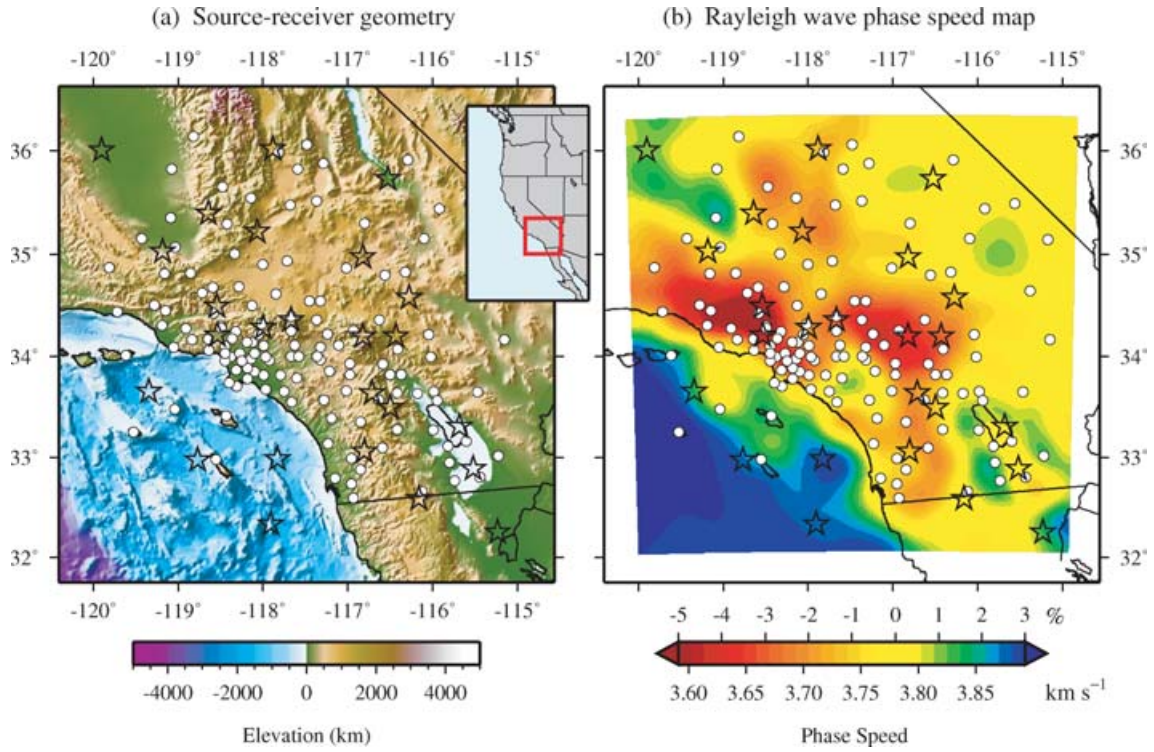

Figure 1. Source–receiver geometry for the numerical experiments in this study. The ☆ symbols denote the locations of 25 earthquakes (each has a $M \geq 4.0$ and occurred between 1990 and 2005); the ○ symbols denote the locations of 132 broad-band receivers in the SCSN. The earthquakes are selected to obtain relatively uniform coverage; all SCSN receivers in the area are included. (a) Topography and bathymetry in the region. (b) Phase-speed map for 20 s Rayleigh waves, based on the regional model of Hauksson (2000), modified with the Moho map of Zhu & Kanamori (2000). This phase-speed map is used to generate synthetic data used in some of the inversion experiments.

vector $\mathbf{d}$ (22). The total number of measurements is $N = N_{\text{events}} \times N_{\text{receivers}} = 25 \times 132 = 3300$.

We illustrate the classical tomographic approach using both rays and banana–doughnut kernels to represent the sensitivities of the measurements to the model parameters. Thus, we compute two $N \times M$ design matrices, $\mathbf{G}^{\text{ray}}$ and $\mathbf{G}^{\text{ker}}$, respectively. Figs 2(a)–(c) shows the computation of a single G_{ik}^{ray} element, and Figs 2(d)–(f) shows an example for G_{ik}^{ker}. Fig. 2 illustrates why the choice between kernels or rays may be moot, depending on the resolution of the basis functions. The infinitesimally thin ray path is smeared out by the relatively smooth basis functions. Thus, in our example, $\mathbf{G}^{\text{ray}} \approx \mathbf{G}^{\text{ker}}$, and we will simply use a generic $\mathbf{G}$ to denote either the ray or kernel design matrix.

The (approximate) Hessian matrix, $\tilde{\mathbf{H}} = \mathbf{G}^T\mathbf{G}$, and the gradient vector, $\mathbf{g} = -\mathbf{G}^T\mathbf{d}$, are visualized in Fig. 3. The Hessian and gradient are determined by the source–receiver geometry and the banana–doughnut kernels (or ray paths), but only the gradient is controlled by the data.

Fig. 4 shows the model recovery using classical tomography based upon a single iteration of the Gauss–Newton method. The recovered model is strongly dependent on the damping parameter γ. When $\gamma \approx 0$, the inverse is unstable and structure is artificially introduced into regions where there is no coverage, that is, the edges of the domain and the oceans (Fig. 4a). When $\gamma \to \infty$, the recovered model is simply the initial model (Fig. 4f), although the spatial pattern is that of the gradient (e.g. compare Fig. 3c with Fig. 4g). The reason for this is that for large values of the damping parameter γ the damped Hessian (27) is dominated by the damping matrix $\mathbf{D}$, which in our case is the identity matrix $\mathbf{I}$. In this case, the solution to the inverse problem given by (28) is effectively a scaled version of the gradient $\mathbf{g}$. For the example in Fig. 4, the L-curve suggests that $\gamma = 10.0$ is a reasonable model selection; this model is shown in Figs 4(c) and 20(c).

4 COMPUTATION OF THE GRADIENT AND HESSIAN

Obtaining the Hessian involves computing banana–doughnut kernels K_i for each source–receiver combination. Thus, the cost of computing the Hessian is the cost of computing all the kernels. For a problem involving N_{events} earthquakes, $N_{\text{receivers}}$ stations, $N_{\text{comp}} = 3$ component seismograms and N_{picks} measurements per seismogram one would need to calculate $N_{\text{events}} \times N_{\text{receivers}} \times N_{\text{comp}} \times N_{\text{picks}}$ kernels.

In adjoint tomography one computes a misfit kernel K from which only the gradient is obtained. One of the primary benefits of adjoint tomography is that the misfit kernel need not be computed by summing over individual banana–doughnut kernels for each source–receiver pair, as in (12). Instead, the measurements, ΔT_i, are incorporated into the adjoint source, which is used to compute the misfit kernel (Section 5). This kernel is constructed via the interaction between a forward wavefield and an adjoint wavefield, requiring only two simulations per earthquake (Tromp *et al.* 2005). So if our inverse problem involves N_{events} earthquakes, obtaining the gradient of the misfit function involves $2N_{\text{events}}$ numerical simulations, that is, this calculation is *independent of the number of receivers, components and picks*. The main drawback of adjoint tomography is that the Hessian is not available, which means that iterative techniques must be used to determine the minimum of the objective function.

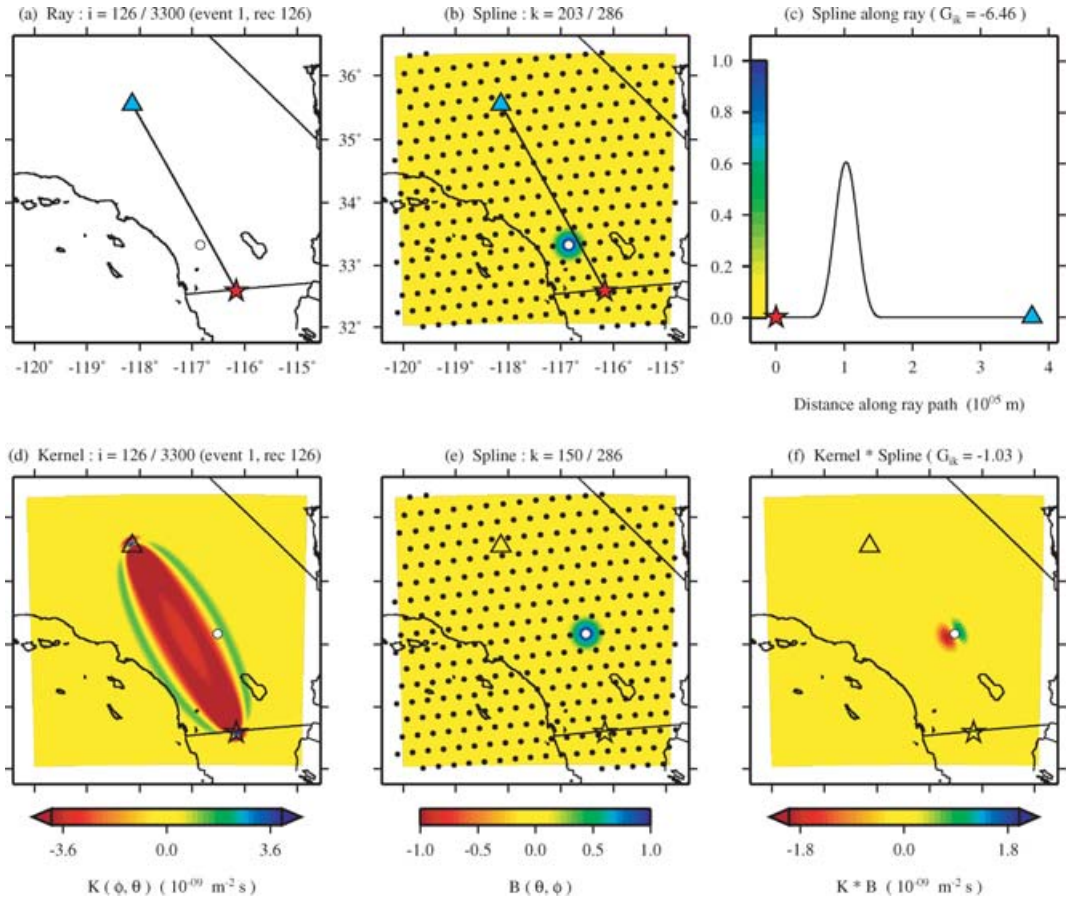

Figure 2. Example computation for an element, G_{ik}, of the design matrix $\mathbf{G}$, using rays (a–c) and finite-frequency kernels (d–f). The row index i is the source–receiver combination, the column index k is the basis function index. The source is denoted by the ☆, the receiver is denoted by the △, and the ○ shows the centre-point of the spherical spline in (b) or (e). (a) Ray path for event number 1 and receiver number 126 (Fig. 1), corresponding to the $i = 126$ index of the $N = 3300$ ray paths. (b) $B_{203}(\mathbf{x})$, the spherical spline basis function for index $k = 203$. Also shown are the centre-points of the $M = 286$ spherical splines. (c) Spline B_{203} evaluated along the ray path. The value of the phase speed for the reference model is constant, so $G_{ik} = (-1/c) \int_{\text{ray}_i} B_k \, \mathrm{d}s$ (eq. 15). In this example $G_{ik} = -1/(3780\,\mathrm{m\,s^{-1}})(2.45 \times 10^4\,\mathrm{m}) = -6.46$ s. (d) Cross-correlation traveltime sensitivity kernel for event number 1 and receiver number 126 (Fig. 1), corresponding to the $i = 126$ index of the $N = 3300$ kernels. (e) $B_{150}(\mathbf{x})$, the spherical spline basis function for index $k = 150$. Also shown are the centre-points of the $M = 286$ spherical splines. (f) The function $K_{126}(\mathbf{x})\, B_{150}(\mathbf{x})$. The integral of this function gives the value $G_{ik} = \int_\Omega K_{126}\, B_{150}\, \mathrm{d}^2\mathbf{x} = -1.03$ s (see Section 2).

Thus, a fundamental distinction between classical and adjoint tomography is whether or not individual banana–doughnut kernels are computed. In the context of classical tomography, there are several ways to compute the kernels. For 1-D earth models, they may be calculated cheaply and rapidly, in particular if approximate expressions are used (Dahlen *et al.* 2000). Using normal modes, Zhao & Jordan (2006) computed global finite-frequency kernels for spherically symmetric models. The kernels may be used to construct the design matrix $\mathbf{G}$, which has $N_{\text{events}} \times N_{\text{receivers}} \times N_{\text{comp}} \times N_{\text{picks}} \times M$ elements. The parametrization of the model (13) must be carefully considered, since M scales $\mathbf{G}$. Once $\mathbf{G}$ is obtained, the Hessian follows from (25).

The computation of the kernels K_i for 3-D models may be accomplished in two ways:

(i) We may perform an adjoint simulation for every single measurement, which requires a total of $2N_{\text{events}} \times N_{\text{receivers}} \times N_{\text{comp}} \times N_{\text{picks}}$ simulations (two for each measurement). For 3-D models the numerical cost is prohibitive.

(ii) Alternatively, we may invoke source–receiver reciprocity and for every source and receiver calculate and store Green's functions as a function of both space and time. This requires one to perform and store $N_{\text{events}} + 3N_{\text{receivers}}$ simulations: one simulation for each event and one simulation for each receiver component. For realistic 3-D simulations the storage requirements are formidable, although for small problems the approach is feasible, as demonstrated by Zhao *et al.* (2005).

Our goal is to improve fully 3-D reference models. Therefore, to make the inverse problem tractable, we are forced to consider an approach based upon knowledge of the value of the misfit function $\chi(\mathbf{m})$, its gradient $\mathbf{g}$, but not its Hessian $\tilde{\mathbf{H}}$. Minimization of the misfit function based upon this information may be accomplished using a non-linear conjugate gradient method, as discussed in Section 6. However, first we demonstrate how we compute the gradient using adjoint methods.

5 THE GRADIENT: CONSTRUCTION OF A MISFIT KERNEL

In this section, we demonstrate how we compute the gradient of the misfit function, $\mathbf{g} = \partial\chi/\partial\mathbf{m}$, using adjoint methods. The gradient

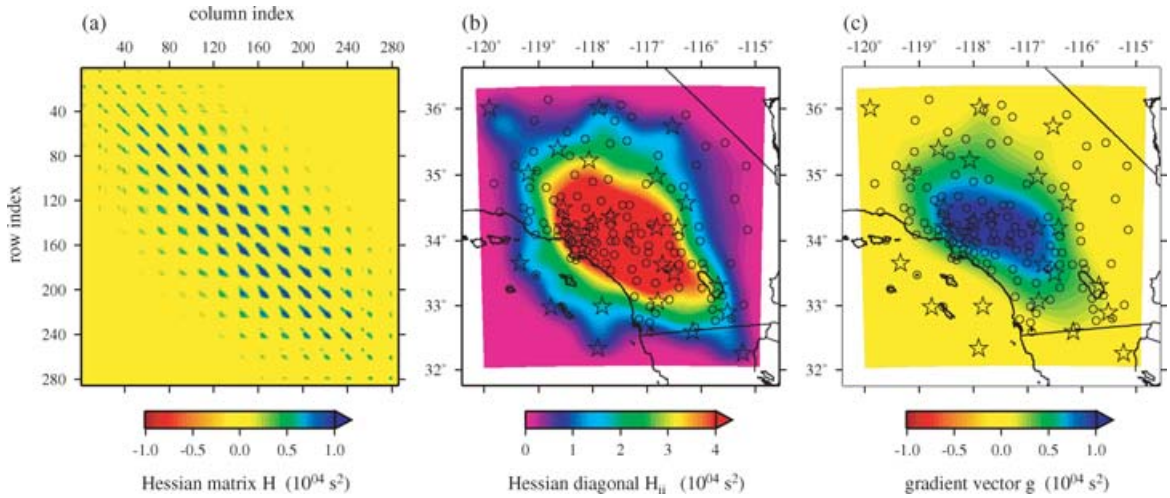

Figure 3. The Hessian matrix and gradient vector for a classical tomography inversion. (a) The Hessian matrix, $\tilde{\mathbf{H}} = \mathbf{G}^T\mathbf{G}$, for the source–receiver geometry shown in Fig. 1, using finite-frequency kernels based upon cross-correlation traveltime measurements. Each element of $\mathbf{G}$ is constructed as shown in 2(d)–(f). The Hessian matrix computed using rays, as shown in 2(a)–(c), is indistinguishable from the one shown in this figure. In practice, a damping matrix is added to the Hessian to stabilize the inversion (27). (b) The diagonal elements of the Hessian matrix, $\tilde{H}_{ii}$, expanded in the spherical spline basis functions to illuminate the spatial pattern (e.g. Zhou *et al.* 2005, fig. 3). This map is a proxy for spatial coverage and depends on the source–receiver geometry, the basis functions, and the sensitivity kernels. (c) The gradient vector, $\mathbf{g} = -\mathbf{G}^T\mathbf{d}$, expanded in the spherical spline basis functions, whereby $\mathbf{d}$ includes cross-correlation traveltime measurements between data computed for the target phase-speed model in Fig. 1(b) and synthetics computed for a homogeneous phase-speed model ($c = 3.78$ km s^{-1}). The ☆ symbols denote the sources, and the ○ symbols denote the receivers.

of the misfit function is obtained from (19):

$$g_k = \int_\Omega K B_k \, \mathrm{d}^2\mathbf{x}, \tag{32}$$

where for the 2-D examples in this paper the integration is over the model surface Ω. Given the misfit kernel, K, and the basis functions, B_k, we can readily compute the gradient of the misfit function. The misfit kernel can also be thought of as a sum of *event kernels*, which we discuss next.

5.1 Event kernels

Tromp *et al.* (2005, Fig. 3) illustrated the construction of a data-independent banana–doughnut kernel based upon adjoint methods. In this paper, the kernels we show are *misfit* kernels, whereby the adjoint source is constructed based in part on a set of measurements between data and synthetics.

The construction of misfit kernels based on cross-correlation traveltime measurements is outlined in Tromp *et al.* (2005, Section 4). For membrane waves, motion is restricted to the vertical direction, and the source functions and wavefields are scalar quantities. The source for the adjoint wavefield for a particular event is given by (Tromp *et al.* 2005, eq. 57)

$$f^\dagger(x, y, t) = -\sum_{r=1}^{N_r} \Delta T_r \, \frac{1}{M_r} \, w_r(T-t) \, \partial_t s(x_r, y_r, T-t) \times \delta(x - x_r)\, \delta(y - y_r), \tag{33}$$

where r is the receiver index, N_r is the number of receivers, ΔT_r is the cross-correlation traveltime measurement over a time window $w_r(t)$, $s(x, y, t)$ is the forward wavefield determined by (29), (x_r, y_r) is the location of the receiver, T is the length of the time-series, and M_r is a normalization factor. The key point is that the adjoint force comprises time-reversed velocity seismograms, input *at the location of the receivers* and weighted by the traveltime measurement associated with each receiver.

For a given earthquake (event), the interaction between the adjoint wavefield and the forward wavefield gives rise to the membrane *event kernel*

$$K(x, y) = -2\mu(x, y) \int_0^T \left[\partial_x s^\dagger(x, y, T-t)\partial_x s(x, y, t) + \partial_y s^\dagger(x, y, T-t)\partial_y s(x, y, t)\right] \mathrm{d}t. \tag{34}$$

Note that the misfit between the data and synthetics is incorporated into the adjoint source (33), which gives rise to the adjoint wavefield $s^\dagger$. Eq. (34) is obtained from the expression for an SH β-kernel in Tromp *et al.* (2005), which contains a product of the adjoint and regular deviatoric strain tensors. In the case of the SH (or membrane) waves, there are four non-zero components (two unique) of each deviatoric strain tensor, which leads to (34).

Figs 5 and 6 show the construction of an event kernel for a single source-receiver pair for a cross-correlation traveltime measurement. The source–receiver geometry and forward wavefield are shown in the left column of Fig. 5. The synthetics are computed for a homogeneous reference model ($c = 3.50$ km s^{-1}), and the data are computed for a uniformly perturbed 'target' model with $\delta \ln c = 0.1$, that is, $c(1 + \delta \ln c) = 3.85$ km s^{-1}. The cross-correlation traveltime measurement at the receiver is $\Delta T = -9.72$ s, indicating a late arrival of the synthetics with respect to the data. The adjoint source function is constructed by time-reversing the synthetic velocity recorded at the receiver and multiplying by ΔT (Fig. 6; eq. 33).

We now replace the homogeneous target model with the checkerboard target model in Fig. 7(a). Fig. 8 shows the construction of an event kernel for this target model for multiple receivers, thereby incorporating multiple measurements. Just as in Fig. 5, the event kernel that forms in Fig. 8 highlights the regions of the current model that give rise to the (cross-correlation traveltime) discrepancies between the data and synthetics. However, in Fig. 8 this is more obvious since the model used to generate the data is not simply a homogeneous perturbation but rather a large-scale checker pattern. The event kernel in Fig. 8 looks qualitatively similar to the phase-speed model in Fig. 7, except with the opposite sign, which

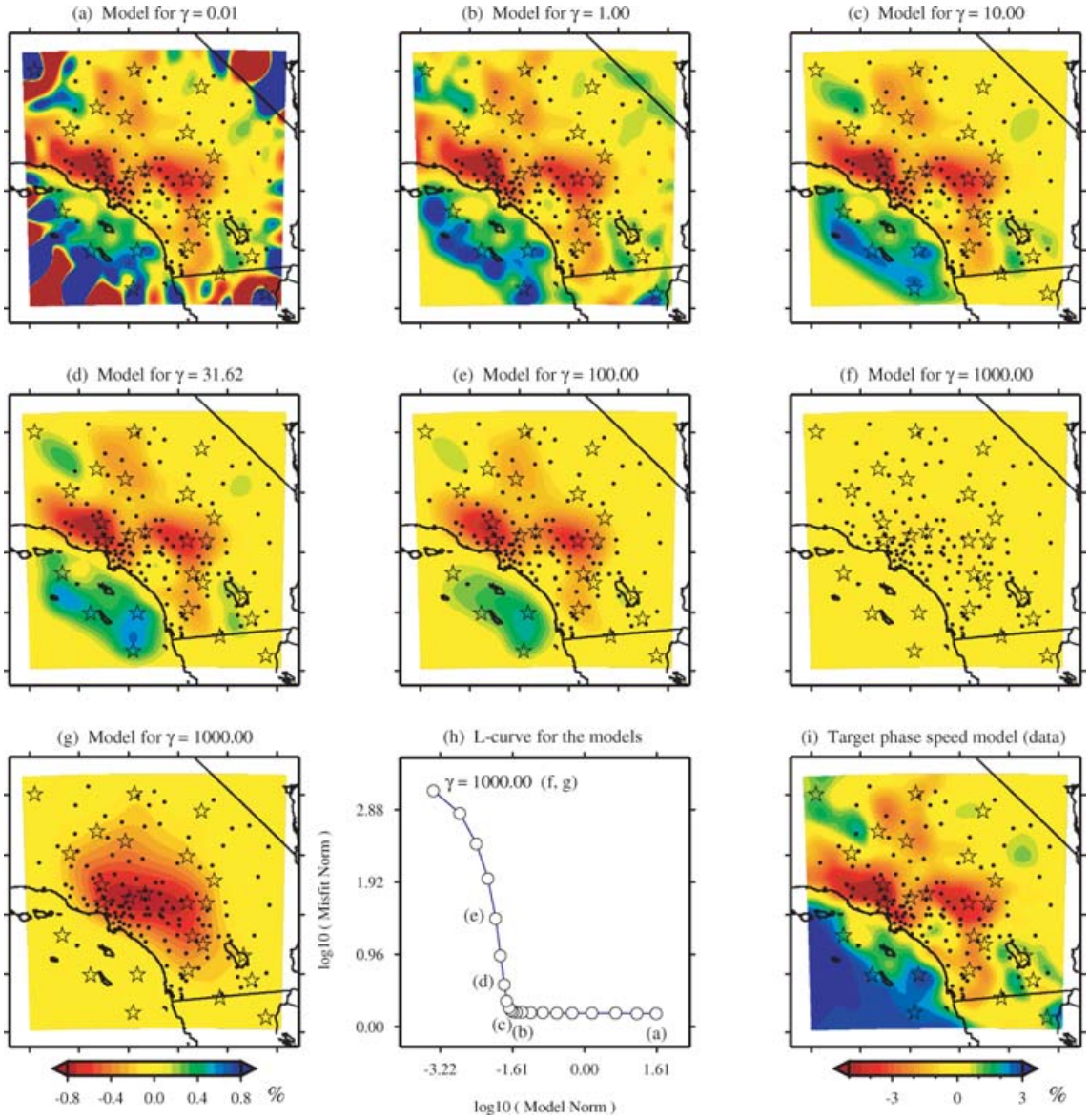

Figure 4. Model recovery and damping in classical tomography, illustrated for an inversion using 3300 banana–doughnut kernels. Each model is computed via $\mathbf{m} = -\tilde{\mathbf{H}}_\gamma^{-1}\mathbf{g}$, where $\tilde{\mathbf{H}}_\gamma = \mathbf{G}^T\mathbf{G} + \gamma^2\mathbf{I}$ is the Hessian matrix with damping parameter γ, and $\mathbf{g} = -\mathbf{G}^T\mathbf{d}$ is the gradient vector (Fig. 3c). The undamped Hessian matrix, $\tilde{\mathbf{H}}_0$, is shown in Fig. 3(a)–(b). (a)–(f) Recovered phase-speed models for various values of γ. The colour scale for each model is shown below (i). (g) Same as (f), only with a more saturated colour scale to show its resemblance to the gradient (Fig. 3c). (h) *L*-curve illustrating the trade-off between misfit norm and model norm, that is, $||\mathbf{G}\,\mathbf{m} - \mathbf{d}||_2$ versus $||\mathbf{m}||_2$. Note that this measure of misfit is not the same as $\mathbf{d}^T\mathbf{d}$, the misfit function in (6). The γ values for the model-points are spaced by uniform $\log_{10}$ increments. (i) Target phase-speed model used to generate the data (Fig. 1b). The ☆ symbols denote the sources, and the • symbols denote the receivers (see Section 3.3).

is consistent with (11): for the variation of the misfit function to be negative, we invoke a fast, positive (blue) structural perturbation where the kernel is negative (red), and a slow, negative structural perturbation where the kernel is positive.

As shown in (33), the amplitude of the adjoint source at a particular receiver, r, is determined in part by the traveltime measurement ΔT_r. Changing the values of ΔT_r changes the weights of the corresponding individual banana–doughnut kernels that comprise the event kernel, something that is explicit in the classical case (12). It is possible to incorporate some measure of weighting at the stage of constructing the adjoint source in order to account for uneven coverage (Fig. 1), as demonstrated in Takeuchi & Kobayashi (2004). Another option is to weight the adjoint sources according to realistic uncertainties associated with each measurement (Tarantola 1984): a measurement with a high uncertainty will have a small amplitude weight, and thus a relatively weak contribution to the event kernel.

5.2 Misfit kernels and damping

We define the *misfit kernel* as the sum of the event kernels for a particular model. Thus, the gradient of the misfit function, $\mathbf{g}$, is obtained as in (32) using the misfit kernel $K(\mathbf{x})$. Fig. 9 shows the construction of a misfit kernel for 25 events. Note that features of each event kernel are very different, even for the simple checkerboard model in this example (Fig. 7). Only after summing the event kernels does

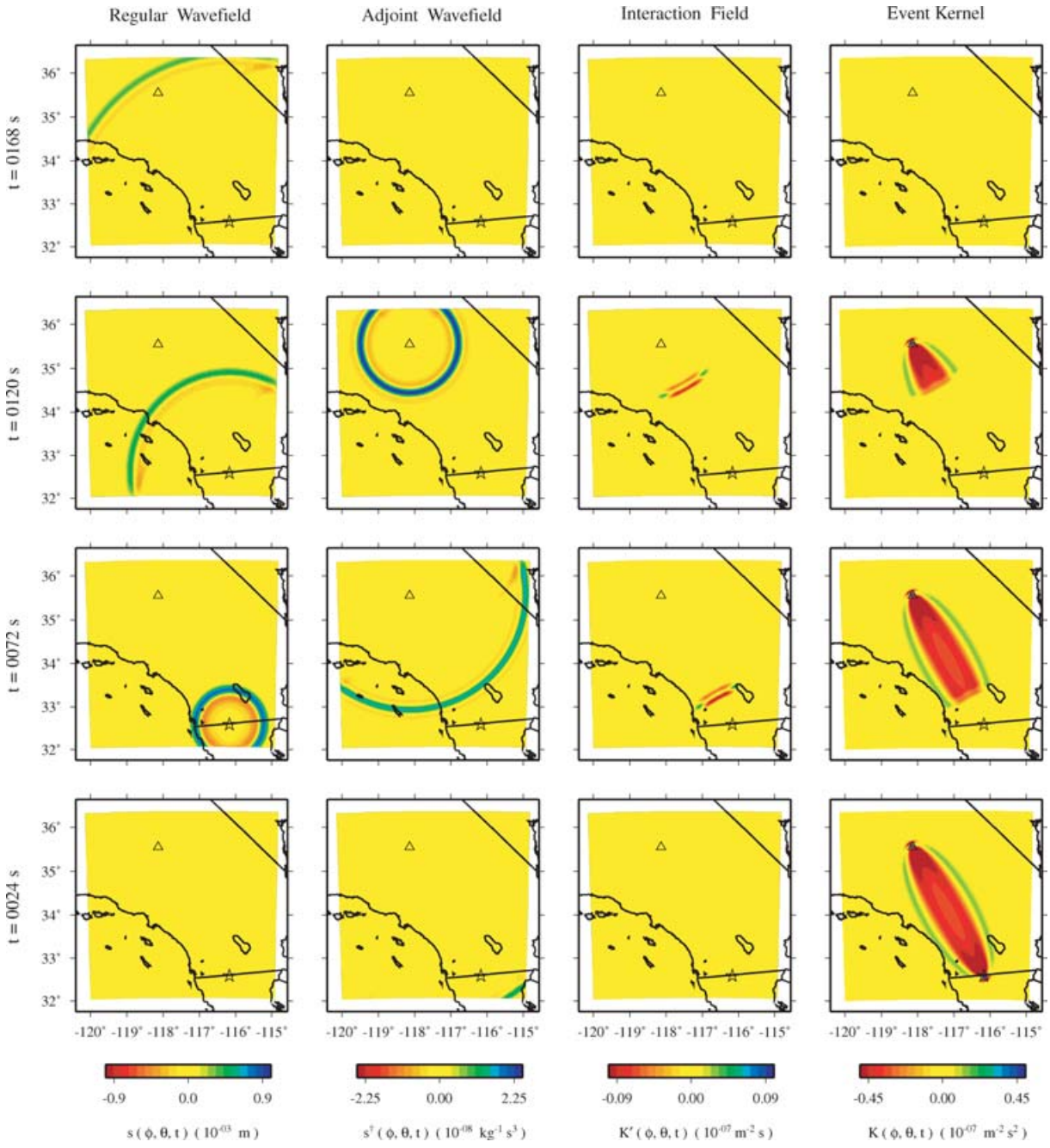

Figure 5. Sequence of interactions between the regular and adjoint wavefields during the construction of a traveltime cross-correlation event kernel $K(\mathbf{x})$. The ☆ symbol denotes the source, and the Δ symbol denotes the receiver. Each row represents the time-step indicated on the left. In this case, with only a single receiver and a uniform model perturbation, the event kernel resembles a banana–doughnut kernel $K_i(\mathbf{x})$. The event kernel is constructed via the interaction between the forward wavefield (first column) and the adjoint wavefield (second column). The interaction field (third column) is the instantaneous product of the two wavefields, which is integrated to form the event kernel (fourth column). The event kernel shows the region of the current model that gives rise to the discrepancy between the data and the synthetics. The regular source function and adjoint source function are shown in Fig. 6.

the pattern (Fig. 9h) begin to resemble the model used to generate the data (Fig. 9i).

We apply a smoothing operator to the misfit kernels in order to remove spurious amplitudes in the immediate vicinity of the sources and receivers (Fig. 10). This is accomplished by convolving (in 2-D) the unsmoothed misfit kernel with a Gaussian of the form

$$G(x, y) = \frac{4}{\pi \Gamma^2} e^{-4(x^2+y^2)/\Gamma^2}, \tag{35}$$

where Γ is the full-width of the Gaussian, defined such that at a (polar) distance $r = \Gamma/2$, the Gaussian has amplitude $G(r) = G(0)e^{-1}$; thus Γ is the scalelength of smoothing (Fig. 10). The choice of Γ is somewhat analogous to the choice of damping parameter γ for the inversion of the Hessian (eq. 27), which involves a degree of subjectivity. In the adjoint method, subjectivity may be removed by selecting Γ according to the shortest wavelengths of the waves. It seems sensible to smooth the kernels using scalelengths somewhat less than the wavelengths of the seismic waves resolved in the numerical simulation.

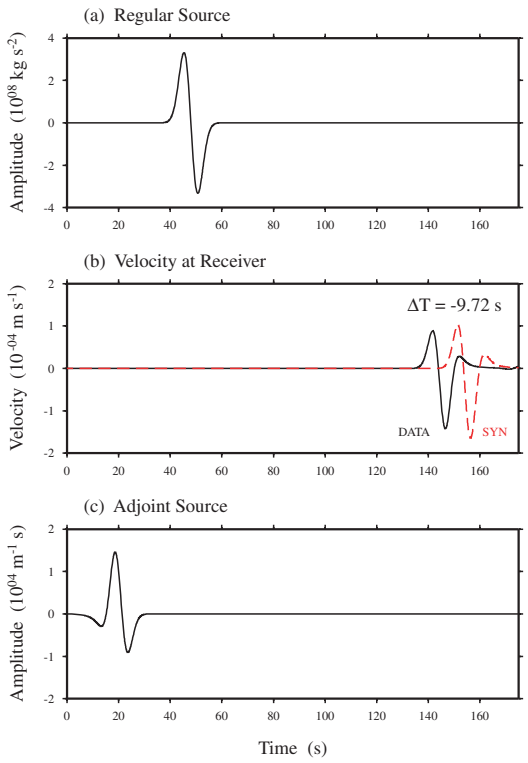

Figure 6. Construction of an adjoint source function used in calculating the membrane surface wave event kernel in Fig. 5. The traveltime sign convention is shown in (8), such that $\Delta T < 0$ represents a delay of the synthetics with respect to the data. The duration of the simulation is $T = 175$ s.

There will exist short-scalelength features and fringes in kernels based upon more complicated 2-D or 3-D models, such as the fringes shown in Tromp *et al.* (2005, Fig. 9) for the *P-SV* wavefield or in Zhou *et al.* (2004, Fig. 13b). The smoothing operation will tend to remove these subresolution features from the kernel. An alternative approach to smoothing the inversion is to add an explicit damping term to the misfit function (e.g. Akçelik *et al.* 2002, 2003), as outlined in Appendix A. This approach leads to an additional term in the expression for the gradient, which represents the desire to obtain a smooth model. We prefer to convolve the misfit kernel with a simple Gaussian that represents the resolution of the simulation, and this is the approach we will take in this paper.

5.3 Basis functions

As shown in (32), the calculation of the gradient of the misfit function requires a choice of model parametrization. Which basis functions should one use? In the classical tomographic example discussed in Section 3.3 we used $M = 286$ spherical spline basis functions to parametrize the model (see Fig. 2). In adjoint tomography, where the wavefields and kernels are represented on discretized grids, we can use the basis functions embedded in the numerical method itself, for example for the SEM we use Lagrange polynomials (Komatitsch & Tromp 1999). This has the advantage that no restrictions are placed on the wavelengths of the model, other than that they need to be resolvable by the waves used in the inversion. This approach increases the number of model parameters dramatically compared to a classical inversion, but because we do not need to invert a Hessian in the adjoint approach this is of no consequence.

Any smooth function $f(\mathbf{x})$, where $\mathbf{x} = (x, y)$, that is sufficiently resolved by the SEM mesh can be expressed in discrete form as

$$f(\mathbf{x}) = \sum_{k=1}^{N_{\text{glob}}} f_k L_k(\mathbf{x}), \tag{36}$$

where $k = 1, \ldots, N_{\text{glob}}$ is the index of the N_{glob} global node points, $f_k = f(\mathbf{x}_k)$ is the functional value at global node $\mathbf{x}_k$, and $L_k(\mathbf{x})$ is a global function defined by

$$L_k(\mathbf{x}) = \begin{cases} l_\alpha(\xi(x,y))\, l_\beta(\eta(x,y)) & \text{if } \mathbf{x}_k \in \Omega_e \text{ and } k|_{\Omega_e} = (\alpha, \beta), \\ 0 & \text{if } \mathbf{x}_k \notin \Omega_e. \end{cases} \tag{37}$$

Here l_α and l_β are Lagrange polynomials of degree α and β, respectively. We use degree-4 polynomials, that is,

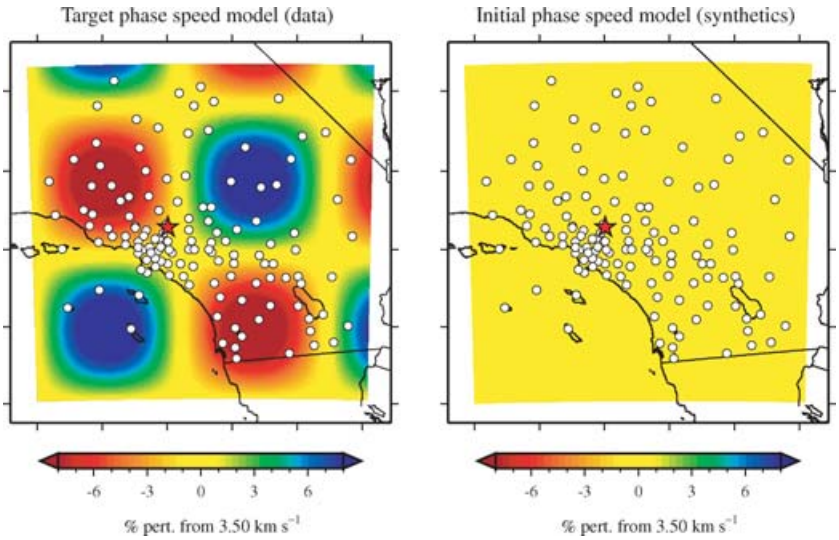

Figure 7. Experimental set-up for the event kernel shown in Fig. 8. The data are computed using the target phase-speed model, and the synthetics are computed using the initial model. The minimum and maximum per cent perturbations in the target model are ±10 per cent. The red star is the event location, and the circles denote the 132 receivers. For plotting purposes, the gridpoints are converted to longitude–latitude points, which results in the non-rectangular appearance of the boundary of the grid.

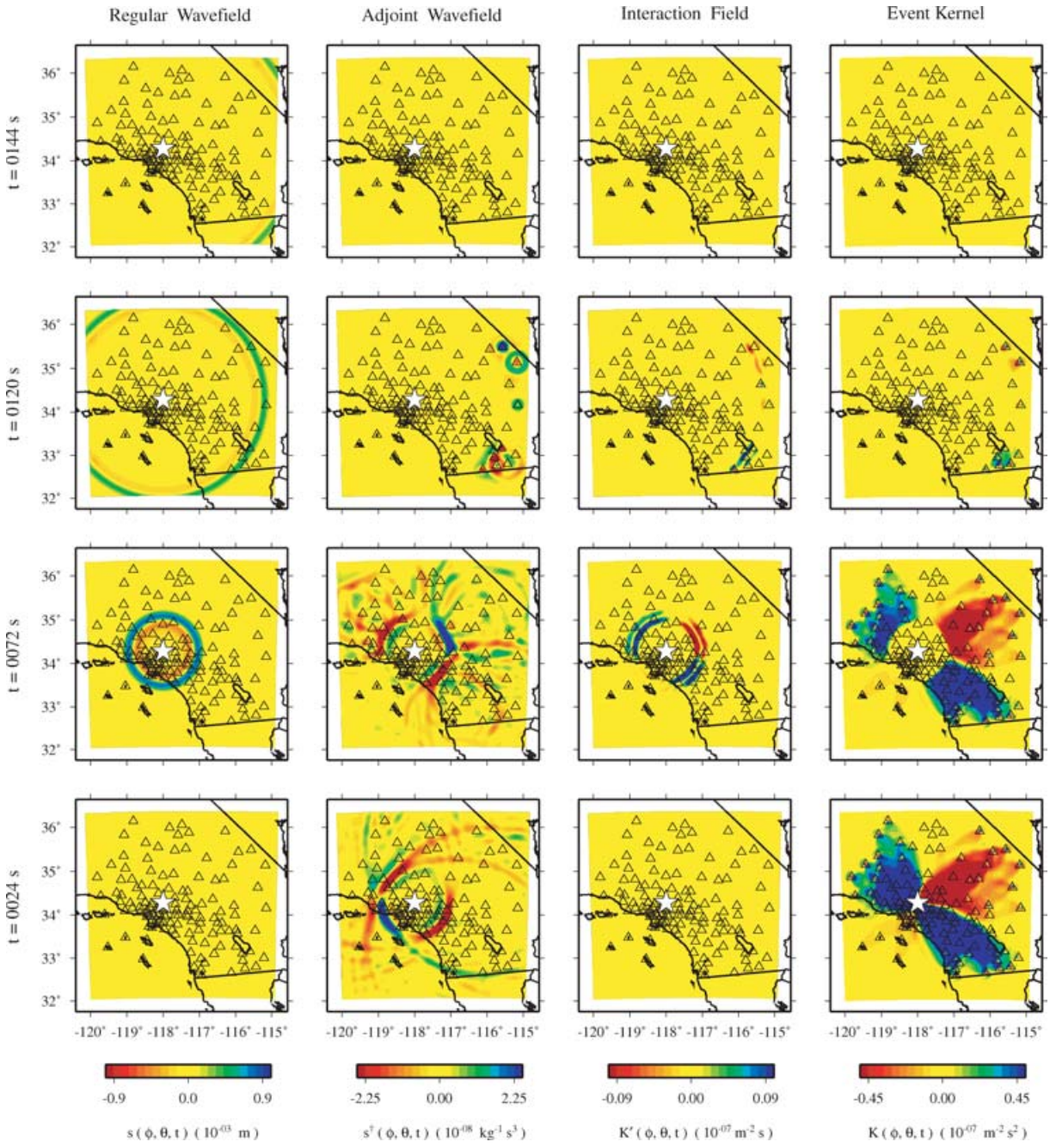

Figure 8. Formation of an event kernel for multiple receivers. The phase-speed models used to generate the data and synthetics are shown in Fig. 7. See Fig. 5 for details. In comparison with Fig. 5, here the event is in a different location, there are 132 receivers instead of one, and the data are generated from a checkerboard model, not a uniformly perturbed model.

5 Gauss–Lobatto–Legendre points, in the 2-D simulations presented in this article. The invertible mapping from the reference square with points (ξ, η), with $-1 \leq \xi \leq 1$ and $-1 \leq \eta \leq 1$, to the deformed quadrilateral spectral-element Ω_e with points (x, y) may be written in the form $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ (e.g. Komatitsch & Vilotte 1998; Komatitsch & Tromp 1999). Note that functions $L_k(\mathbf{x})$ corresponding to global gridpoints $\mathbf{x}_k$ located on the edges or corners of elements have non-zero contributions from all elements that share the global point. At the kth node,

$$L_k(\mathbf{x}_k) = 1, \tag{38}$$

in accordance with (36).

The functions $L_k(\mathbf{x})$ are orthogonal but not orthonormal. We may obtain a set of orthonormal basis functions $B_k(\mathbf{x})$ based upon the definition

$$B_k(\mathbf{x}) = L_k(\mathbf{x})/A_k, \tag{39}$$

where A_k is the square-root-area associated with the kth node:

$$A_k^2 = \int_\Omega L_k^2(\mathbf{x})\, \mathrm{d}^2\mathbf{x}. \tag{40}$$

The B_k are orthonormal in the sense that

$$\int_\Omega B_k(\mathbf{x}) B_{k'}(\mathbf{x})\, \mathrm{d}^2\mathbf{x} = \delta_{kk'}, \tag{41}$$

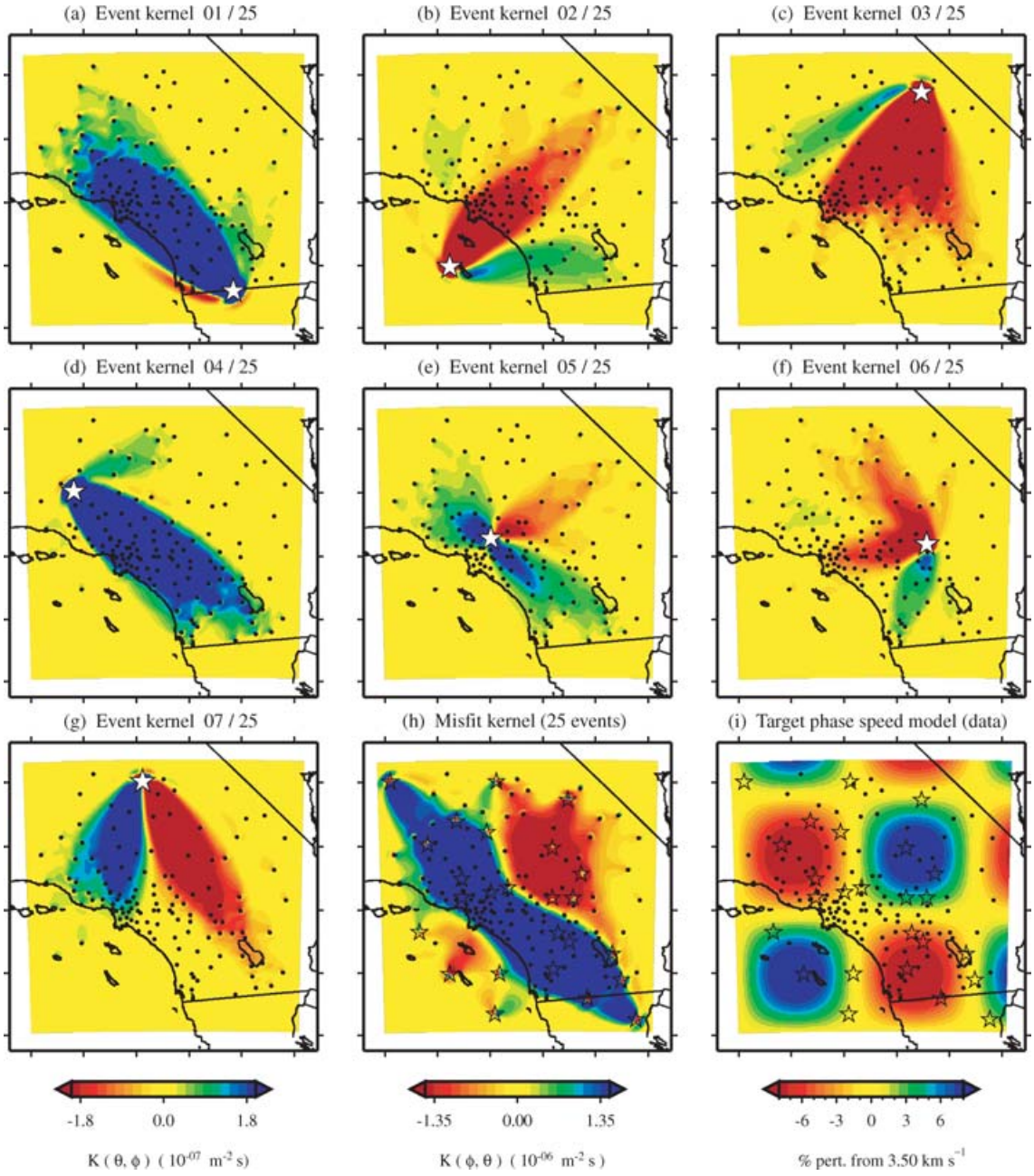

Figure 9. Construction of a misfit kernel. (a)–(g) Individual event kernels, each constructed via the method shown in Fig. 8 (which shows Event 5). The colour scale for each event kernel is shown beneath (g). (h) The misfit kernel is simply the sum of the 25 event kernels. (i) The source–receiver geometry and target phase-speed model. There are a total of $N = 25 \times 132 = 3300$ measurements that are used in constructing the misfit kernel (see Section 5).

and any function can be expanded in terms of these basis functions. For example, we may expand the misfit kernel $K(\mathbf{x})$ in terms of the basis functions $B_k(\mathbf{x})$ as

$$K(\mathbf{x}) = \sum_{k=1}^{N_{\text{glob}}} \tilde{K}_k B_k(\mathbf{x}). \tag{42}$$

The expansion coefficients $\tilde{K}_k$ are determined by

$$\begin{aligned} \tilde{K}_k &= \int_\Omega K(\mathbf{x}) B_k(\mathbf{x})\, \mathrm{d}^2\mathbf{x} = \int_\Omega \sum_{k'} K_{k'} L_{k'}(\mathbf{x}) B_k(\mathbf{x})\, \mathrm{d}^2\mathbf{x} \\ &= \sum_{k'} K_{k'} A_{k'} \int_\Omega B_{k'}(\mathbf{x}) B_k(\mathbf{x})\, \mathrm{d}^2\mathbf{x} = K_k A_k, \end{aligned} \tag{43}$$

where $K_k = K(\mathbf{x}_k)$ is the value of the misfit kernel at a global gridpoint, and we have used (36) and the orthonormality relation (41).

Now let us assume we have computed a misfit kernel $K(\mathbf{x})$. In discrete form, we can write $K(\mathbf{x}_k) = K_k$, since K is defined on the $N_{\text{glob}} = 25\,921$ global nodes of the SEM mesh. Upon comparing (32) with (43), we see that, using the basis functions (39), the gradient of the misfit function is simply

$$g_k = K_k A_k. \tag{44}$$

This provides a trivial step from the discretized kernel to the gradient. Using the $M = N_{\text{glob}}$ basis functions in (39), the model

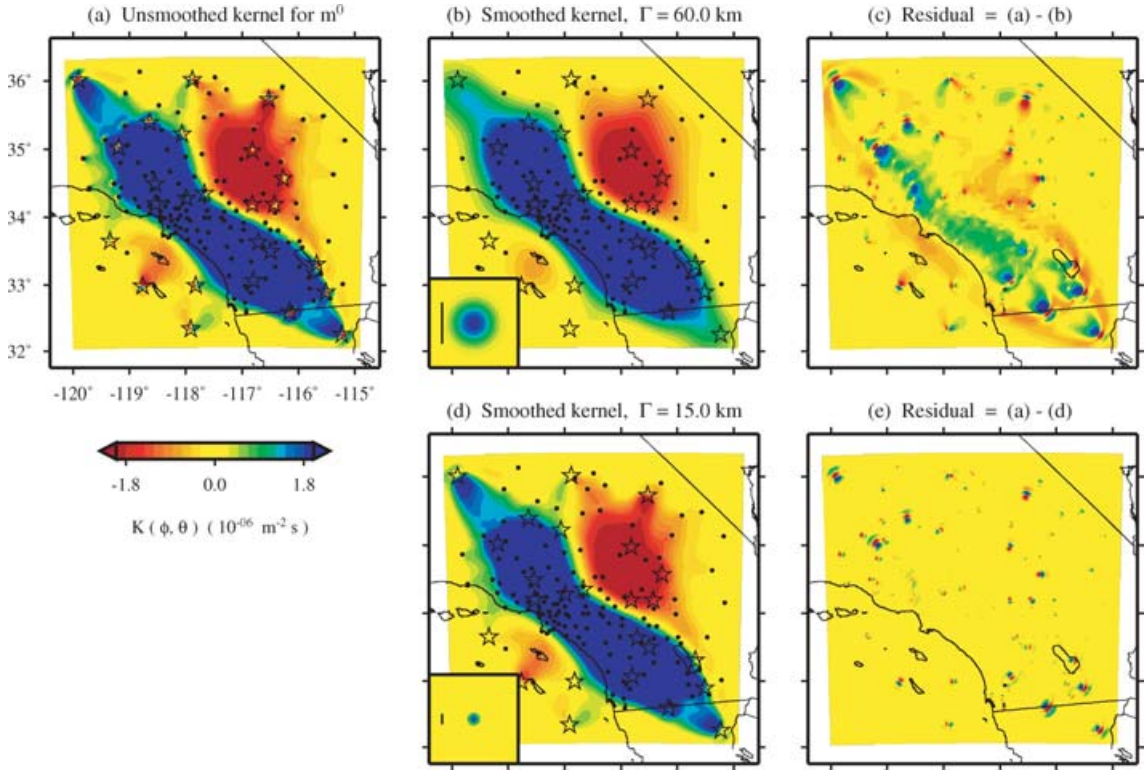

Figure 10. Smoothing the misfit kernel. (a) Unsmoothed misfit kernel (Fig. 9h). (b)–(c) Smoothed misfit kernel, with residual, obtained via convolution of a Gaussian function (bottom left inset) with (a). The parameter $\Gamma = 60$ km controls the width of the Gaussian and, thus, the degree of smoothing; its value is plotted as a line next to the Gaussian. (d)–(e) Same as (b)–(c), only for less smoothing ($\Gamma = 15$ km). Note that the source and receiver labels are not plotted in the residual plots (see Section 5.2).

parameters (13) are, therefore,

$$\delta m_k = \delta \ln c_k A_k, \tag{45}$$

where $\delta \ln c_k$ is the discrete version of $\delta \ln c(\mathbf{x})$.

6 OPTIMIZATION: ITERATIVE IMPROVEMENT OF THE MODEL

In the previous section we showed how to compute the gradient of the misfit function by summing event kernels (Fig. 9) and subsequently multiplying by the basis functions of the model (32). In this section we illustrate how iterative improvements to the model may be determined based upon a non-linear conjugate gradient algorithm (Fletcher & Reeves 1964). We demonstrated this algorithm for a simple source inversion in Tromp *et al.* (2005, Section 8.1). In Section 6.2, we consider a 2-D tomographic example.

6.1 Conjugate gradient algorithm

The algorithm we use may be summarized as follows: given an initial model $\mathbf{m}^0$, calculate $\chi(\mathbf{m}^0)$, $\mathbf{g}^0 = \partial\chi/\partial\mathbf{m}\,(\mathbf{m}^0)$, and set the initial conjugate gradient search direction equal to minus the initial gradient of the misfit function, $\mathbf{p}^0 = -\mathbf{g}^0$. If $||\mathbf{p}^0|| < \epsilon$, where ϵ is a suitably small number, then $\mathbf{m}^0$ is the model we seek to determine, otherwise:

(i) Perform a line search to obtain the scalar ν^k that minimizes the function $\tilde{\chi}^k(\nu)$, where

$$\tilde{\chi}^k(\nu) = \chi(\mathbf{m}^k + \nu\mathbf{p}^k)$$

$$\tilde{g}^k(\nu) = \frac{\partial\tilde{\chi}^k}{\partial\nu} = \frac{\partial\chi}{\partial\mathbf{m}}\left(\mathbf{m}^k + \nu\mathbf{p}^k\right)\cdot\mathbf{p}^k$$

- Choose a test parameter $\nu_t^k = -2\tilde{\chi}^k(0)/\tilde{g}^k(0)$, based on quadratic extrapolation.
- Calculate the test model $\mathbf{m}_t^k = \mathbf{m}^k + \nu_t^k\mathbf{p}^k$.
- Calculate $\chi(\mathbf{m}_t^k)$ and, for cubic interpolation, $\mathbf{g}_t^k = \mathbf{g}\,(\mathbf{m}_t^k)$.
- Interpolate the function $\tilde{\chi}^k(\nu)$ by a quadratic or cubic polynomial and obtain the ν^k that gives the (analytical) minimum value of this polynomial.

(ii) Update the model : $\mathbf{m}^{k+1} = \mathbf{m}^k + \nu^k\mathbf{p}^k$, then calculate $\mathbf{g}^{k+1} = \partial\chi/\partial\mathbf{m}\,(\mathbf{m}^{k+1})$.

(iii) Update the conjugate gradient search direction: $\mathbf{p}^{k+1} = -\mathbf{g}^{k+1} + \beta^{k+1}\mathbf{p}^k$, where $\beta^{k+1} = \mathbf{g}^{k+1}\cdot(\mathbf{g}^{k+1} - \mathbf{g}^k)/(\mathbf{g}^k\cdot\mathbf{g}^k)$.

(iv) If $||\mathbf{p}^{k+1}|| < \epsilon$, then $\mathbf{m}^{k+1}$ is the desired model; otherwise replace k with $k+1$ and restart from (i).

A key decision is the choice of the test parameter, ν_t^k, which determines how far one should go in the direction of the search direction (initially the negative gradient) to obtain the test model. We assume a quadratic form of the misfit function and determine ν_t^k based upon this assumption. Computation of $\chi(\mathbf{m})$ (misfit) and $\mathbf{g}\,(\mathbf{m})$ (misfit kernel) are expensive in the tomographic problem, and thus we must limit the number of computations as much as possible. Some of these aspects are addressed in Section 6.3.

6.2 2-D tomographic example

Using (6), we can define the average traveltime anomaly for a particular model:

$$\overline{\Delta T} = \sqrt{2\,\chi(\mathbf{m})/N}. \tag{46}$$

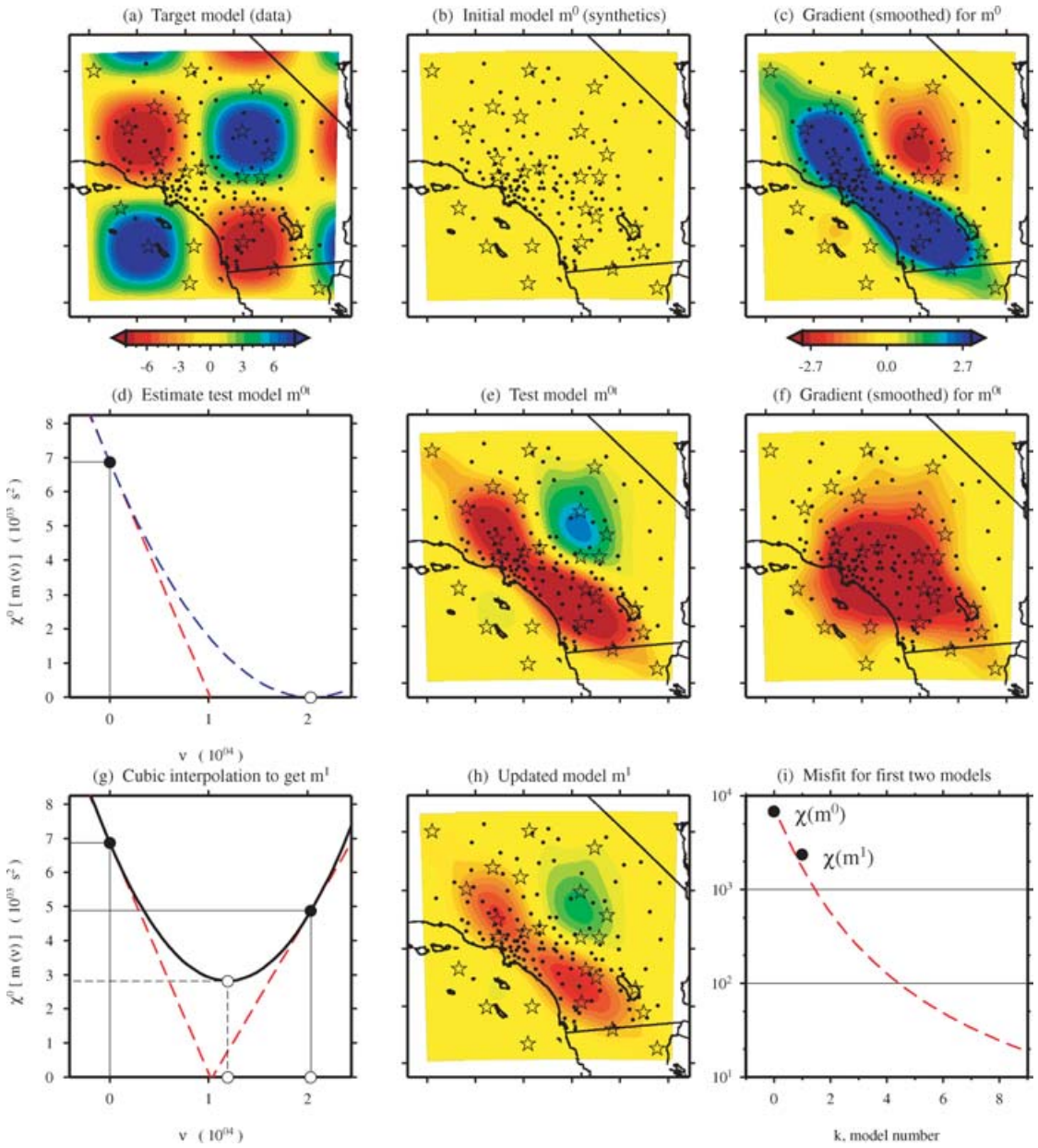

Figure 11. The conjugate gradient algorithm applied to a 2-D tomographic example. An extended explanation can be found in Section 6.2. The algorithm is repeated to obtain the models in Fig. 12. (a) Target phase-speed model used to generate the data. (b) Phase-speed model used to generate the initial synthetics. The period of the source is $\tau \approx 20$ s, the reference wave speed is $c = 3.50$ km s^{-1}, and thus the reference wavelength is $\lambda \approx 70$ km. (c) Misfit kernel—corresponding to the gradient of the misfit function—constructed as illustrated in Figs 8–10, with smoothing parameter $\Gamma = 30$ km. This kernel highlights the regions of model (b) that need to be improved to reduce the misfit between data and synthetics. (d) Representation of the misfit of the initial model (b) and the initial gradient (c) in the conjugate gradient algorithm. The misfit is denoted by the •, and the gradient is denoted by the red dashed line. The white circle indicates the test model obtained by quadratic extrapolation of the gradient through $\chi = 0$. (e) Test model $\mathbf{m}_t^0$ corresponding to the white circle in (d). (f) Gradient associated with the test model in (e). (g) Cubic interpolation of two misfit values, $\chi(\mathbf{m}^0)$ and $\chi(\mathbf{m}_t^0)$, and two gradients, shown in (c) and (f). The analytical minimum provides ν^0, the distance away from $\mathbf{m}^0$ (b) in the direction of (c) that is taken to obtain the first updated model, $\mathbf{m}^1$. (h) First updated model, $\mathbf{m}^1$, corresponding to the white circles in (g) with $\nu^0 = 1.24 \times 10^4$. (i) Misfit values for the first two models. The red dashed curve is taken from Fig. 12(f).

This gives some physical meaning to the χ-values in the plots in this section. Figs 11(a)–(i) shows one cycle of the conjugate gradient algorithm for the 2-D tomographic example. Part (a) shows the phase-speed model used to generate the data (the 'target' model) and (b) shows the initial phase-speed model, $\mathbf{m}^0$, used to generate the initial synthetics. The phase-speed of the initial model is $c = 3.50$ km s^{-1}, the period of the source in the simulations is $\tau = 20$ s, and thus the reference wavelength is approximately $\lambda = c\tau = 70$ km.

Fig. 11(c) shows the (smoothed) gradient for this model. The gradient is represented by the slope $\tilde{g}^0(0)$ of a line passing through $[0,\ \tilde{\chi}(0)]$ (Fig. 11d). Quadratic extrapolation with a parabolic minimum at $(\nu_t^0, 0)$ gives the ν-value for a new test model (Fig. 11d, Appendix B). Fig. 11(e) shows the test model, $\mathbf{m}_t^0$, for which we

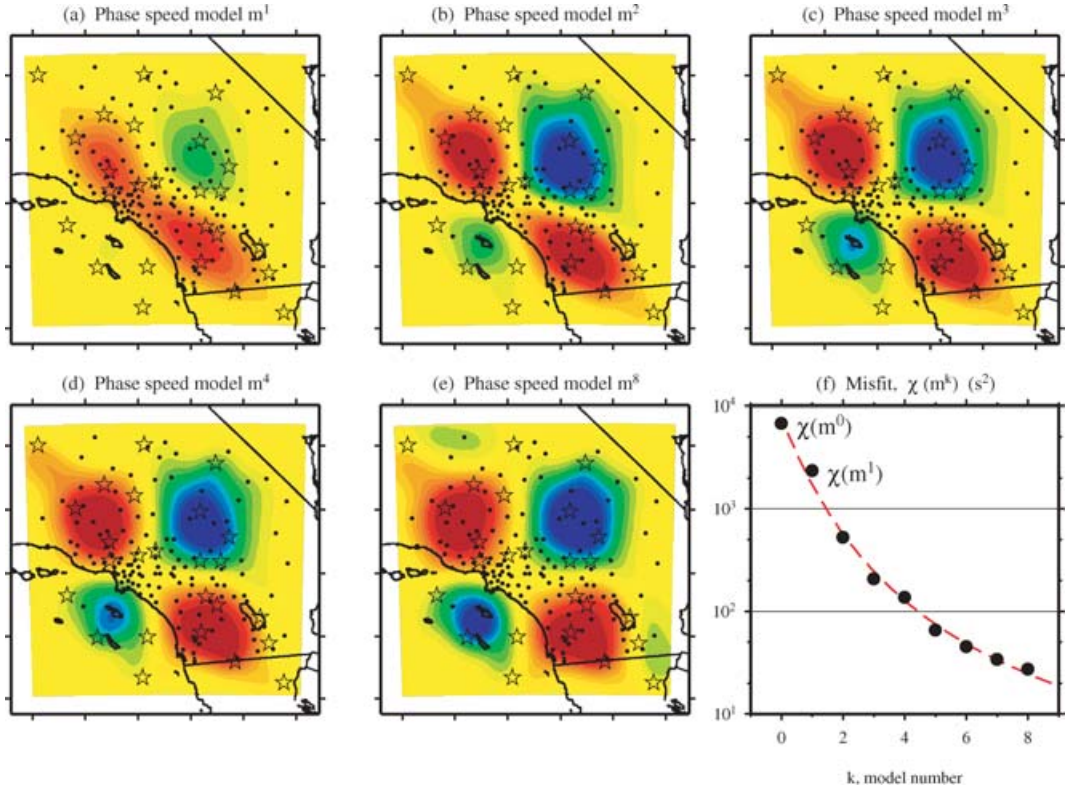

Figure 12. Iterative improvement of the reference phase-speed model using the conjugate gradient algorithm illustrated in Fig. 11. An extended explanation can be found in Section 6.2. The first iteration in Fig. 11 produces $\mathbf{m}^1$ (a), which becomes the current model, from which we obtain $\mathbf{m}^2$ (b), and so on. The red dashed hyperbolic curve in (f) is drawn to accentuate the reduction in misfit.

compute the gradient via the process shown in Figs 7–10, only now the model is no longer homogeneous. The gradient, shown in Fig. 11(f), is then depicted as the slope of a line passing through $[\nu_t^0,\ \tilde{\chi}(\nu_t^0)]$ (Fig. 11g).

Next, in Fig. 11(g) we approximate $\tilde{\chi}^0(\nu)$ by a cubic polynomial, $P^0(\nu)$, passing through two points, $[0,\ \tilde{\chi}^0(0)]$ and $[\nu_t^0,\ \tilde{\chi}(\nu_t^0)]$, and having slopes at these points corresponding to the respective gradients. In other words, six values are needed to obtain an analytical minimum of the cubic function: the two models (represented by $\nu = 0$ and ν_t^0), the misfits of these models, and the derivatives at these points (see Appendix B). The minimum, $[\nu^0, P^0(\nu^0)]$, indicates the expected value of the misfit for the updated model given by $\mathbf{m}^1 = \mathbf{m}^0 - \nu^0\mathbf{g}^0$, which is shown in Fig. 11(h) and represented by the point $(\nu^0 = 1.2 \times 10^4, 0)$ in Fig. 11(g). Fig. 11(i) shows the decrease in the misfit function going from $\mathbf{m}^0$ to $\mathbf{m}^1$. The dashed curve is determined based upon the nine iteration points in Fig. 12(f).

Fig. 11 thus constitutes one iteration of the conjugate gradient algorithm. The process is repeated, and the results are shown in Fig. 12. Each iteration produces a model that looks qualitatively more similar to the target model shown in Fig. 11(a), and generates a lower value of the misfit function (6). We draw a best-fitting hyperbola to the $\log_{10}$ values to highlight the convergence.

We next use the seismologically more interesting Rayleigh wave phase-speed model in Fig. 1. In comparison with Fig. 11(a), this model has variable scalelength and lower amplitude perturbations. The weaker perturbations result in a lower initial misfit, $\chi(\mathbf{m}^0) = 1182.0\ \mathrm{s}^2$. Fig. 13 shows the recovery of an interior portion of the model, where path coverage is good. The basic features in the target phase-speed model (Fig. 20a) are recovered by the third iteration (Fig. 13d). The two sets of points in the Fig. 13(f) are discussed in the next section. The model obtained after the first iteration, model $\mathbf{m}^1$ shown in Fig. 13(b), looks very similar to the model obtained based upon a classical Hessian-based inversion with heavy damping shown in Fig. 4(g). This reflects the fact that in the conjugate gradient approach one is effectively working with an initial approximation to the Hessian that is the identity matrix.

6.3 Variations on the conjugate gradient algorithm

Based on the conjugate gradient algorithm outlined in Section 6.1, we require $4N_{\text{events}}$ numerical simulations for each iterative improvement of the model: synthetics for $\mathbf{m}^0$, the gradient for $\mathbf{m}^0$, synthetics for test model $\mathbf{m}_t^0$, and the gradient for $\mathbf{m}_t^0$. This information is used to compute the analytical minimum for a cubic polynomial. An alternative approach is to perform $3N_{\text{events}}$ numerical simulations per iteration by neglecting the gradient of the test model and using a quadratic polynomial to compute an analytic minimum (Appendix B).

A comparison of these two approaches is shown in Fig. 20(f). The initial model for both cases has a misfit of $\chi(\mathbf{m}^0) = 1182.0\ \mathrm{s}^2$ for $N = 3300$ seismograms, corresponding to an average traveltime anomaly of $\overline{\Delta T}(\mathbf{m}^0) = 0.85\mathrm{s}$ (eq. 46). Using $4N_{\text{events}}$ simulations with a cubic polynomial, we obtain a negligible advantage in terms of a better convergence of $\chi(\mathbf{m})$: for example, $\chi(\mathbf{m}^8_{\text{cubic}}) = 3.52\ \mathrm{s}^2$ whereas $\chi(\mathbf{m}^8_{\text{quad}}) = 3.75\ \mathrm{s}^2$ (Fig. 20). To the eye, the recovered models $\mathbf{m}^k_{\text{cubic}}$ and $\mathbf{m}^k_{\text{quad}}$ are indistinguishable.

An additional part of the conjugate gradient algorithm that can be adjusted is the selection of the test model, which we discuss in Appendix B. Finally, we note that entrapment into local minima is common in the conjugate gradient method, as addressed in Akçelik

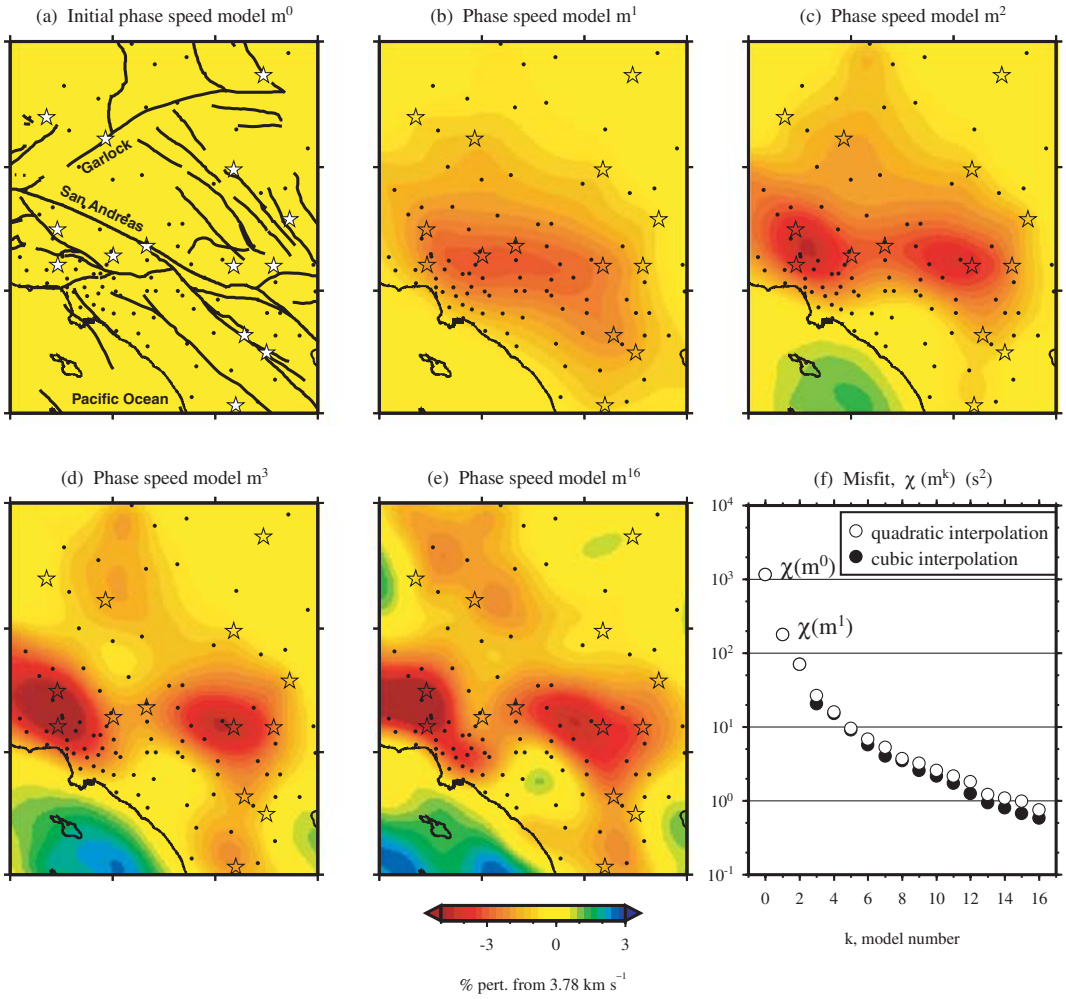

Figure 13. Adjoint tomography recovery of a Rayleigh wave phase-speed model (Fig. 20a). Here we show an interior portion of the southern California region with sufficient path coverage. The colour scale for each model is shown below (e). (a) Initial phase-speed model $\mathbf{m}^0$. Faults of Jennings (1994) are drawn only for scale. (b)–(e) Iterations $\mathbf{m}^1$, $\mathbf{m}^2$, $\mathbf{m}^3$ and $\mathbf{m}^{16}$. (f) Reduction in the misfit function (6) using cubic interpolation (•) versus quadratic interpolation (○) in the conjugate gradient algorithm (Section 6.3).

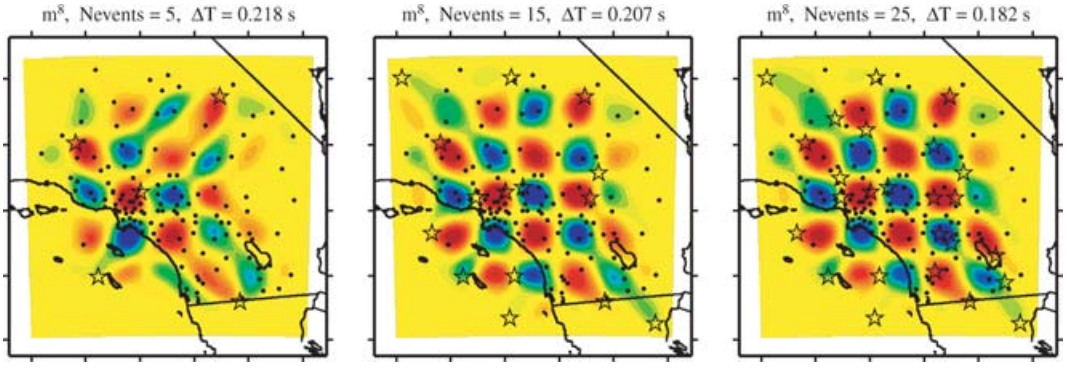

Figure 14. Effect of the number of events on the recovery of the phase-speed models. Data are generated from the phase-speed model in Fig. 15(i). The average traveltime anomaly, $\overline{\Delta T}$, is computed from the misfit function value, $\chi(\mathbf{m}^8)$, using (46). As expected, $\overline{\Delta T}$ decreases as we increase the number of events.

et al. (2002, 2003). Such local minima may be avoided by using multiscale methods (Bunks *et al.* 1995). Alternatively, by starting at longer periods, which constrain the long wavelength heterogeneity, and gradually moving to shorter periods, which constrain smaller scale structures, one can also try to avoid local minima.

7 TOMOGRAPHIC EXPERIMENTS

The greater the number of events used in the inversion, the better the recovery of the model. Fig. 14 shows the recovery of the model in Fig. 15(i) using 5, 15 and 25 sources, respectively.

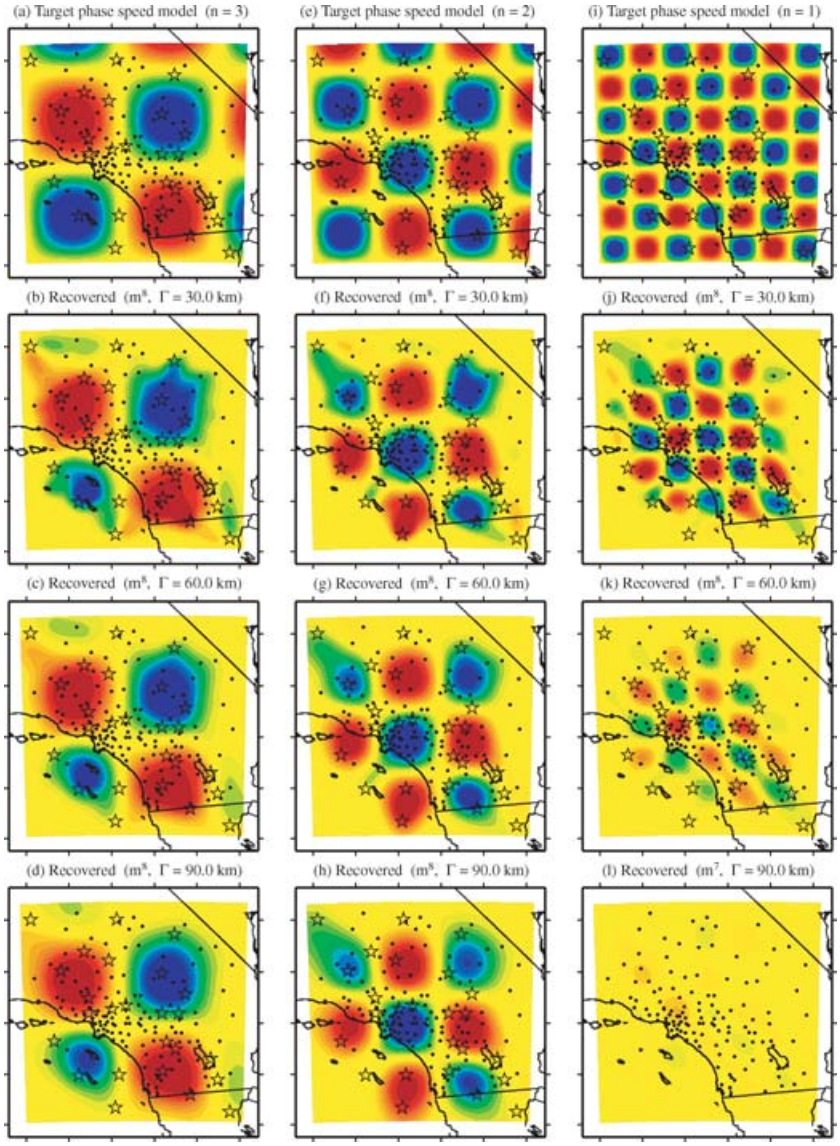

Figure 15. Effect of the degree of smoothing and scalelength of heterogeneity on the recovery of the phase-speed models. The factor n is given by $\Lambda = n\lambda$, where Λ is the scalelength of heterogeneity and $\lambda = c\tau = 70$ km is the reference wavelength. The smoothing parameter, Γ, is constant for each row (see Fig. 10 and Section 7).

Fig. 15 examines the effect of the smoothing parameter, Γ, on the recovery of three different phase-speed models, each having a scalelength of structural heterogeneity that is proportional to the reference wavelength. Using a smaller Γ we resolve shorter-scalelength structures, whether they are in the target phase-speed model or not. When the scalelength of the smoothing exceeds that of the structure ($\Gamma > \Lambda$), the structure is smoothed out, as expected (Fig. 15l).

The introduction of random errors into the cross-correlation traveltime measurements, ΔT_i, has essentially no impact on model recovery in our examples. For example, we denote a 50 per cent error in the measurements by $\Delta T_i' = \Delta T_i(r + 0.5)$, where $r \in [0, 1]$ is a random number, ΔT_i is the 'actual' measurement, and $\Delta T_i'$ is the randomized measurement used in the inversion. In terms of the adjoint method, the introduction of random errors has the effect of changing the amplitude of the various banana–doughnut kernels that comprise the event kernel. Because the coverage in this example is very good, several similar kernels are 'stacked' in constructing the event kernel, and thus the random errors effectively cancel.

8 SOURCE, STRUCTURE AND JOINT INVERSIONS

The traveltime differences between data and synthetics may be due to an inaccurate structural model, inaccurate source models, or some combination of both. In this section we illustrate the simultaneous inversion for structural and source parameters using adjoint methods and the conjugate gradient algorithm. We first describe and illustrate the basic source inversion and then address the joint inversion.

8.1 Basic source inversion

A perturbation of the point source (30) may be written in the form

$$\delta f(x, y, t) = -\dot{h}(t)\delta t_s\, \delta(x - x_s)\, \delta(y - y_s) + h(t)(\delta x_s\, \partial_{x_s} + \delta y_s\, \partial_{y_s})[\delta(x - x_s)\, \delta(y - y_s)], \quad (47)$$

where δt_s denotes a perturbation in the origin time, $(\delta x_s, \delta y_s)$ a perturbation in the source location, and $\dot{h}(t) = \partial h/\partial t = -\partial h/\partial t_s$.

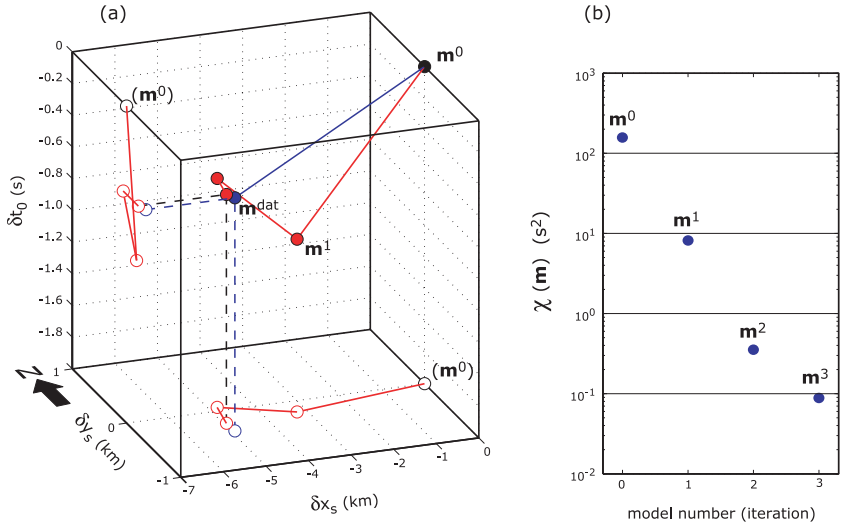

Figure 16. Basic source inversion: source recovery using an unperturbed (fixed) wave speed structure. The model vector, $\mathbf{m} = (\delta x_s, \delta y_s, \delta t_s)$, represents the source with respect to the initial model, where (x_s, y_s) is the location and t_s is the origin time. The data are generated using the target source model $\mathbf{m}^{\text{dat}}$. The initial source model for the synthetics, $\mathbf{m}^0$, initiates 0.53 s late with respect to the data and is mislocated by 4.93 km at an azimuth of N85.8°E with respect to the data. The initial source parameters are randomly chosen from a mislocation 'patch' with a 5 km radius and a timing error range of ±1 s. (a) Iterative improvement of the source model towards the target source model. White circles show the projections of the model points onto horizontal and vertical planes, respectively; these are shown to aid in the perspective. (b) Reduction in the traveltime cross-correlation misfit (6) for the source models shown in (a). (See also Fig. 18.)

Based on the theory outlined in Tromp *et al.* (2005, Section 8), a change in the traveltime misfit function (7) due to a change in the point source is given by

$$\delta\chi = \int_0^T \int_\Omega \delta f(x, y, t)\, s^\dagger(x, y, T-t)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}t, \tag{48}$$

where $s^\dagger$ denotes the adjoint wavefield, whose sources are time-reversed, measurement-weighted seismograms, injected at the receivers, just as in the case of the structure inversions (33). (Here the traveltime measurement is affected by source perturbations only.) Upon substituting (47), we obtain

$$\begin{aligned}\delta\chi = &-\delta t_s \int_0^T \dot{h}(t)\, s^\dagger(x_s, y_s, T-t)\, \mathrm{d}t \\ &+ (\delta x_s\, \partial_{x_s} + \delta y_s\, \partial_{y_s}) \int_0^T h(t)\, s^\dagger(x_s, y_s, T-t)\, \mathrm{d}t.\end{aligned} \tag{49}$$

We may express (49) in terms of the gradient as $\delta\chi = \mathbf{g}\cdot\delta\mathbf{m}$, where

$$\delta\mathbf{m} = \left[\left(x_s^k - x_s^0\right)/\lambda,\ \left(y_s^k - y_s^0\right)/\lambda,\ \left(t_s^k - t_s^0\right)/\tau\right], \tag{50}$$

$$\begin{aligned}\mathbf{g} = \Bigg[&\lambda \int_0^T h(t)\, \partial_{x_s} s^\dagger(x_s, y_s, T-t)\, \mathrm{d}t, \\ &\lambda \int_0^T h(t)\, \partial_{y_s} s^\dagger(x_s, y_s, T-t)\, \mathrm{d}t, \\ &-\tau \int_0^T \dot{h}(t)\, s^\dagger(x_s, y_s, T-t)\, \mathrm{d}t\Bigg].\end{aligned} \tag{51}$$

Here $\delta\mathbf{m}$ is a three-parameter non-dimensionalized model vector describing the source perturbation. The source origin time t_s is scaled by the reference period τ, and the source coordinates are scaled by the reference wavelength $\lambda = c\,\tau$, where c is the reference phase speed. The gradient vector, $\mathbf{g}$, depends on the model $\mathbf{m}$ through the adjoint wavefield $s^\dagger$: by perturbing the source, the measurement between data and synthetics changes, and thus the adjoint wavefield changes correspondingly.

In the experiments in Section 6, the sources for the data and synthetics were identical, whereas the structure was not. We now consider the effects of source perturbations, where the point sources for the initial synthetics are mislocated and initiate at an incorrect time.

Tromp *et al.* (2005, Fig. 12) demonstrated a two-parameter source inversion based upon an adjoint method and the conjugate gradient algorithm. In that example, the two source parameters described the magnitude vector of the point source. In Fig. 16 we illustrate a three-parameter source inversion for $\delta\mathbf{m} = (\delta x_s, \delta y_s, \delta t_s)$. The structural models for the data and synthetics are identical. We use the adjoint method to compute the gradient (51) of the misfit function (6). Using the conjugate gradient algorithm, we recover the source by the third iteration.

Finally, we emphasize that all of the equations in this section apply generally for any measurement, for example waveforms or amplitudes. The computed values for the expressions will differ, however, because the adjoint source $f^\dagger$ (and the corresponding adjoint wavefield $s^\dagger$) will vary for each measurement.

8.2 Joint inversions

In a joint inversion the sources and structure are initially different from the 'target' sources and structure, and we seek to determine both. If we consider the three-parameter source inversion in Fig. 16, then the model vector for the joint inversion is $\delta\mathbf{m} = [\delta\mathbf{m}_{\text{str}}\,; \delta\mathbf{m}_{\text{src}}]$ with dimension $N_{\text{structure}} + 3\,N_{\text{event}}$. The misfit function is given by (6). We adjust the gradient of the misfit function at each iteration according to a constant, F, computed from the initial gradient:

$$\mathbf{g}^k = \left[F\mathbf{g}^k_{\text{str}}\,; \mathbf{g}^k_{\text{src}}\right], \tag{52}$$

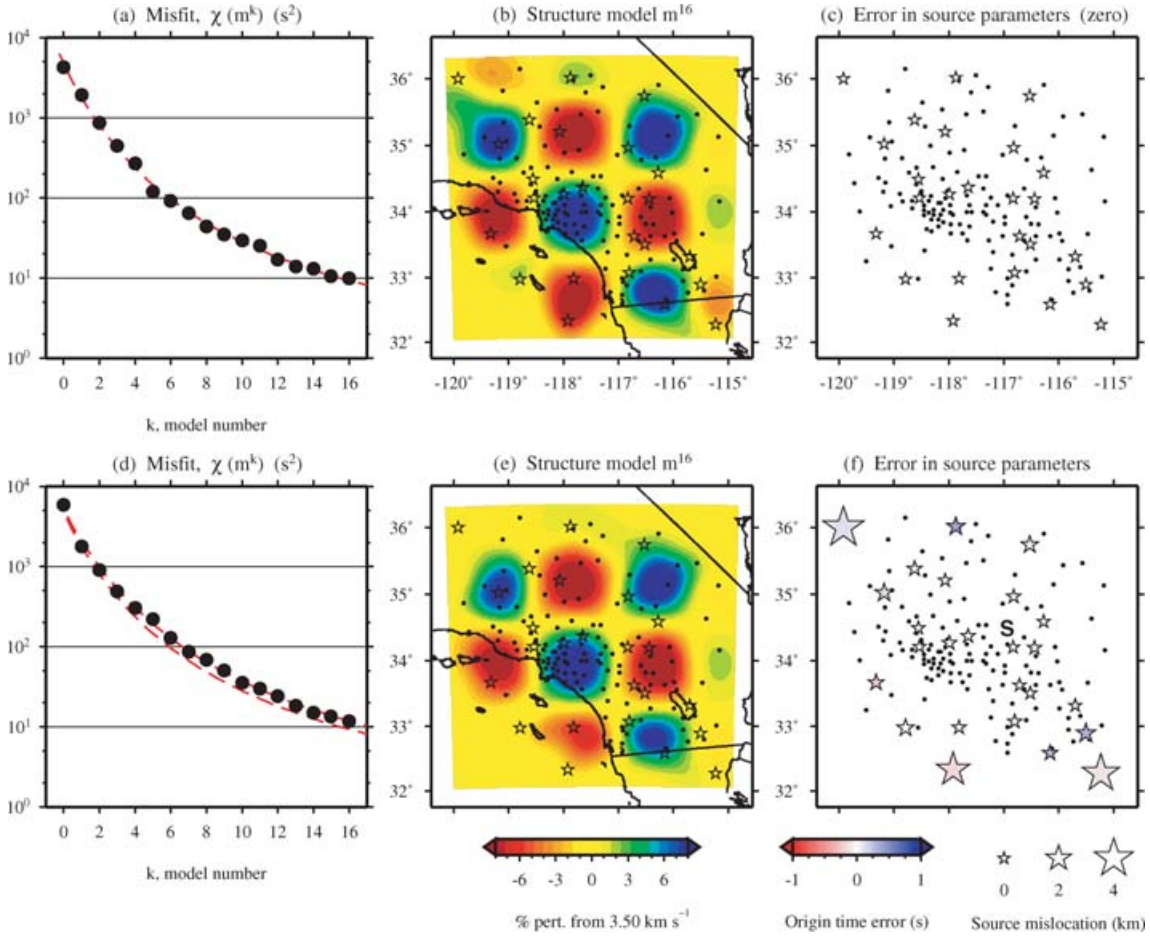

Figure 17. Joint inversion for source and structural parameters. The initial structural model is homogeneous. The traveltime cross-correlation misfit function values in (a) and (d) are computed from (6). The data are generated using Fig. 15(e). (a) Reduction in misfit for a basic structure inversion, that is, one in which the structure of the initial model differs from that of the data, but the sources are always identical to those that generated the data. (b) Recovered model $\mathbf{m}^{16}$. Colour scale is shown in (e). (c) Error in recovered source parameters. In the basic source inversion, the sources for the data and synthetics are identical and hence there is no error. Key is shown in (f). (d) Reduction in misfit for a joint inversion. The lower dashed curve is the basic structure inversion in (a). (e) Recovered model $\mathbf{m}^{16}$. Subtle differences from (b) can be seen near the edges. (f) Error in recovered source parameters. The initial error in the source parameters is shown in Fig. 19(c). Sources near the edges have the largest remaining error. The recovery of the source parameters for the event labelled **S** is shown in Fig. 18.

denoting a concatenation of the structure gradient $\mathbf{g}_{\text{str}}$ computed via (32) and the source gradient $\mathbf{g}_{\text{src}}$ computed via (51). The scaling factor F is given by

$$F = \|\mathbf{g}^0_{\text{src}}\|_2 \,/\, \|\mathbf{g}^0_{\text{str}}\|_2, \tag{53}$$

where $||\cdot||_2$ denotes the L2-norm of the enclosed vector. The motivation behind (52) is that we want the source parameters and structural parameters to have about the same contribution in the gradient in the conjugate gradient algorithm. The exact choice of F, for example, L1- versus L2-norm, is not important. Note that the factor F is chosen once and for all based upon the *initial* structural and source gradients. Also, the gradients ($\mathbf{g}^k$, $\mathbf{g}^k_{\text{str}}$ and $\mathbf{g}^k_{\text{src}}$) are all with respect to the misfit function (6), evaluated at model $\mathbf{m}^k$.

Fig. 17 compares a basic source inversion with a joint inversion. In the joint inversion the initial structural model is homogeneous, and the sources are mislocated randomly within 5 km of the target source and have an inaccurate origin time within the range ± 1 s. The two misfit curves in Fig. 17(d) show that the joint inversion does almost as well as the basic structure inversion; in fact, it lags the misfit by only one or two iterations. In the final model only the sources on the edges of the grid contain location and timing errors (Fig. 19c to Fig. 17f), which is expected since there are few, if any, paths to constrain the structure.

Fig. 18 shows the recovery of a single source during the joint inversion. It takes approximately 16 iterations to fully recover the source (instead of the three iterations in Fig. 16 for the basic source inversion), although most of the source location is still recovered in the first few iterations. This increase is, of course, due to the gradual improvement of the structural parameters, which initially differ by up to 10 per cent from the target structure.

In an inversion with real data, the initial model is bound to be deficient both in terms of structure and sources. Thus, a joint inversion is a logical approach. Fig. 19 shows the consequences of neglecting either source or structure in the inversion. In Figs 19(a)–(c), we invert for structure and assume that the sources are accurate, when in fact they are perturbed as shown in (c). The misfit curve in (a) shows that the conjugate gradient approach appears to be working: the misfit decreases as the structure iterates to updated models. However, it is clear that Fig. 19(b) does not represent the true structure, since we know the target model we are trying to recover, as well as its associated misfit curve for the basic structure inversion. This

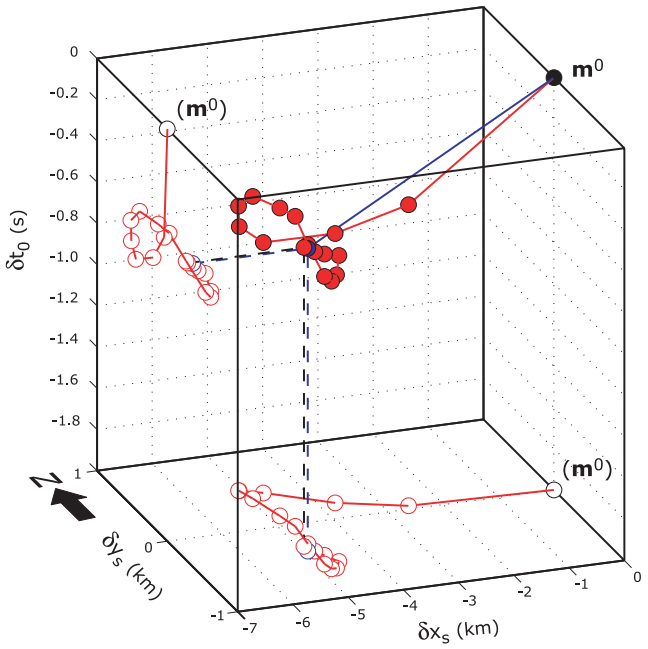

Figure 18. Source recovery of a particular event during the joint inversion shown in Fig. 17(d)–(f). The source location is denoted by the **S** in Fig. 17(f). By the sixteenth iteration, the source is nearly identical to the source used to generate the synthetics. The recovered structure is shown in Fig. 17(e). Compare with Fig. 16(a), which is the same source perturbation, (4.93 km, −0.53 s), but for a basic source inversion.

illustrates how (fixed) errors in the source parameters are mapped into errors in the structural parameters. Figs 19(d)–(f) shows the opposite scenario: the structure is fixed and assumed to be accurate, and we allow the source parameters to be perturbed to reduce the traveltime misfit. The source parameters adjust themselves from Fig. 19(c) (initial) to Fig. 19(f) (final) while reducing the misfit.

9 DISCUSSION

9.1 Three types of sensitivity kernels

We have designated three types of sensitivity kernels: 'banana–doughnut kernels', 'event kernels' and 'misfit kernels'. A banana–doughnut kernel (e.g. Marquering *et al.* 1999) is a phase-specific (e.g. *P*) kernel for an individual source–receiver combination; for our purposes, the key point is that this kernel does not incorporate the measurement. Alternative names include 'finite-frequency', 'Born' and 'sensitivity' kernel. An event kernel can be thought of as a sum of individual banana–doughnut kernels, such that each kernel in the sum is weighted by its corresponding measurement. Using the adjoint approach, however, the event kernel is not computed by summing banana–doughnut kernels, but rather in one single simulation through the interaction between the forward wavefield and an adjoint wavefield generated by simultaneous fictitious sources for all available arrivals at all available stations (Section 5.1). A misfit kernel is simply the sum of event kernels, and may be thought of as a graphical representation of the gradient of the misfit function. In classical tomography, the banana–doughnut kernels are used to compute the gradient and (approximate) Hessian of the misfit function for the Newton approach to the inverse problem. In adjoint tomography, only the misfit kernels are used in the inverse problem.

9.2 Classical tomography versus adjoint tomography

In this paper, 'classical tomography' refers to Hessian-based inversions, whereby the Hessian is constructed from individual source–receiver paths, either in terms of rays or finite-frequency kernels. The Hessian matrix, with a damping parameter γ, can be inverted to obtain structural models. We compute the traveltime anomalies, and thus χ, via (6), and then compare these values with those obtained from gradient-derived models.

Fig. 20 shows a comparison among models produced using classical tomography, $\mathbf{m}^{\text{ray}}$ (ray-based inversion) and $\mathbf{m}^{\text{ker}}$ (kernel-based inversion), and the model produced using adjoint tomography, $\mathbf{m}^{16}$ (16 conjugate gradient iterations). All three models are only subtly different from the target model used to generate the data (Fig. 20a). The misfit for each approach is summarized in Fig. 20(e). The misfit values for the classical models, $\chi(\mathbf{m}^{\text{ray}}) = 5.26\ \text{s}^2$ and $\chi(\mathbf{m}^{\text{ker}}) = 4.90\ \text{s}^2$, correspond to average traveltime anomalies of $\overline{\Delta T}(\mathbf{m}^{\text{ray}}) = 0.056\,\text{s}$ and $\overline{\Delta T}(\mathbf{m}^{\text{ker}}) = 0.055\,\text{s}$ (eq. 46), indicating that each recovered model explains almost all of the traveltime differences between a homogeneous model and the target model in Fig. 1(b). Two points regarding the two χ-values are important.

(1) They are essentially the same, which is expected, since the Hessian used in each inversion is very similar (Fig. 3a).

(2) They are met by the conjugate gradient approach by the seventh conjugate gradient iteration.

In other words, $\chi(\mathbf{m}^{\text{ray}}) \approx \chi(\mathbf{m}^{\text{ker}}) \approx \chi(\mathbf{m}^{7})$; after seven conjugate gradient iterations, we recover a model equivalent to what could be recovered by having the (ray- or kernel-based) Hessian. It is important to note that $\mathbf{m}^{\text{ray}}$, $\mathbf{m}^{\text{ker}}$, and $\mathbf{m}^{1}$ are based upon the homogeneous reference model $\mathbf{m}^{0}$, but for $k > 1$, the adjoint tomography models $\mathbf{m}^{k}$ are based on heterogeneous models.

Fig. 20 might suggest that classical tomography 'does pretty well' in comparison with adjoint tomography. This is more or less true for the simple examples in this paper. However, seismic tomography is transitioning from simple 1-D reference models to fully 3-D reference models. The calculation of a Hessian for 3-D reference models is generally not an option, and thus one must resort to iterative, gradient-based algorithms. The results in this paper illustrate that for the problems considered here, such iterative techniques work quite efficiently and converge quickly.

The main advantages of the adjoint tomography approach are fivefold. First, all the complexities that are considered in the forward problem (e.g. Komatitsch & Tromp 2002a,b) can be considered in the inversion. For example, in this paper we have shown finite-frequency sensitivity kernels based on heterogeneous models. However, one could also consider fully anisotropic earth models with 21 elastic parameters for essentially the same numerical cost as an isotropic simulation involving just two parameters. Secondly, the style of tomography—traveltime, amplitude, waveform—is determined by the choice of the misfit function (Tromp *et al.* 2005). Given the choice of measurement, one simply determines the associated adjoint source that gives rise to the corresponding kernel. Thirdly, any time segment where the data and synthetics match reasonably well is suitable for a measurement. One does not need to label a particular phase, like *P* or *SS*, because the adjoint simulation will reveal how this particular measurement (or time window) 'sees' the earth model, and the resulting sensitivity kernel will reflect this view. Fourth, the model parametrization is trivial (43) and requires only a conservative level of smoothing to remove

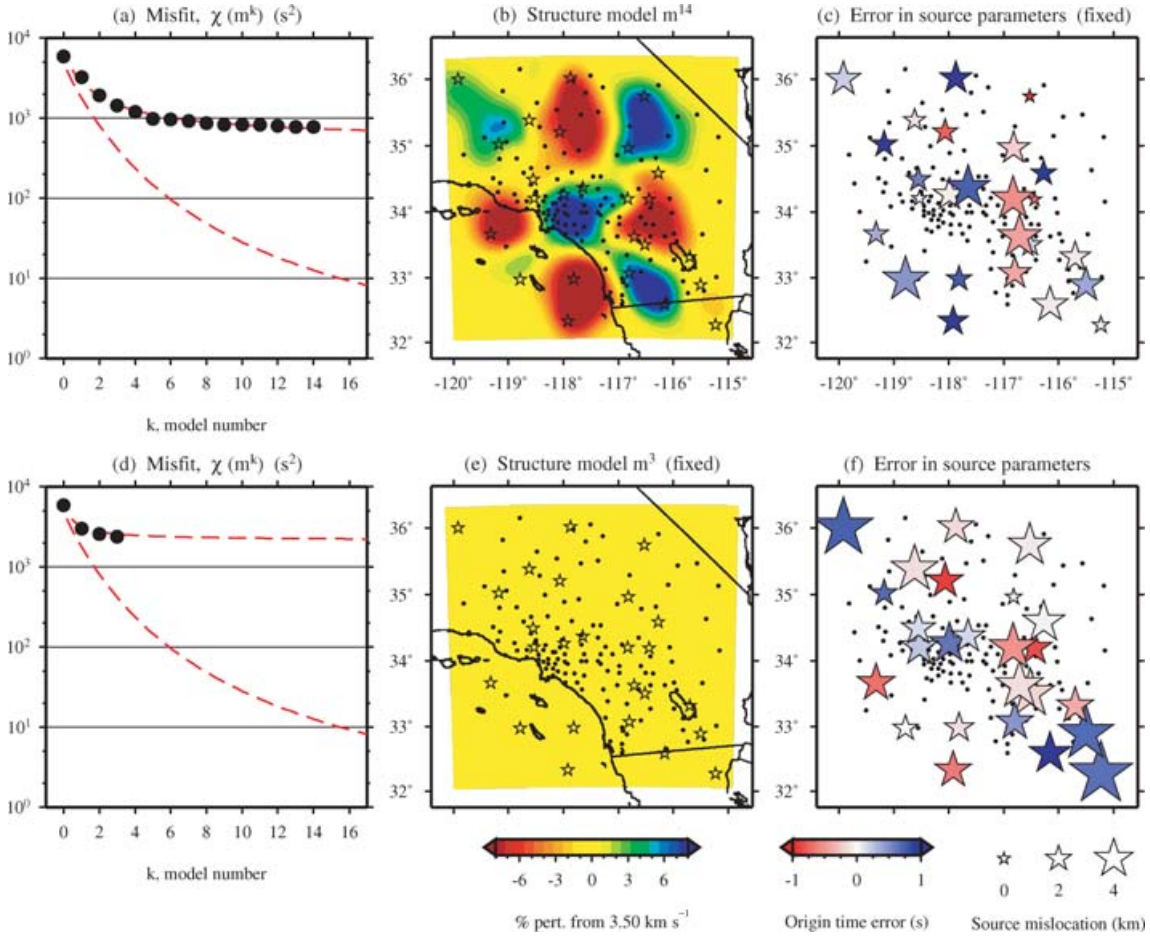

Figure 19. Mapping source errors onto structure and vice versa. The initial source and structural model parameters are different from the target source and structural parameters. The traveltime cross-correlation misfit function values in (a) and (d) are computed from (6); the number of values is <17, because the stopping criterion for the conjugate gradient algorithm was reached. The data are generated using Fig. 15(e). (a) Reduction in misfit for a structure inversion, whereby the source errors are fixed. The lower dashed curve is the basic structure inversion in Fig. 17(a). (b) Recovered model $\mathbf{m}^{16}$. Colour scale is shown in (e). Note the discrepancy with Fig. 17(b). (c) Error in source parameters used in the inversion. Key is shown in (f). (d) Reduction in misfit for a source inversion, whereby the structure errors are fixed. The lower dashed curve is the basic structure inversion in Fig. 17(a). (e) Structure model used in the inversion. (f) Error in recovered source parameters. The initial error in the source parameters is shown in (c).

numerical artefacts in the kernels near the sources and receivers (Section 5.2). Furthermore, structure can only be introduced in regions where the kernel (or gradient) is non-zero. This is in contrast to classical tomography, where both the selection of basis functions and the choice of damping involve a certain degree of undesired subjectivity. Finally, the approach scales linearly with the number of earthquakes but is independent of the number of receivers and the number of arrivals that are used in the inversion.

With southern California in mind, say we have $N_{\text{events}} = 150$ earthquakes, $N_{\text{receivers}} = 150$ Southern California Seismic Network (SCSN) stations, $N_{\text{comp}} = 3$ components per seismogram and $N_{\text{picks}} = 4$ time-windowed measurements per component, for a total of $N_{\text{events}} N_{\text{receivers}} N_{\text{comp}} N_{\text{picks}} = 270\,000$ measurements. An adjoint approach would require $2N_{\text{events}} = 300$ simulations to compute one misfit kernel. A complete seven iteration conjugate gradient inversion based upon cubic interpolation would require $7 \times 3N_{\text{events}} = 3150$ total simulations. By comparison, a Hessian-based inversion would require individual kernels for the 270 000 measurements, which, for 3-D models, is neither computationally feasible nor practical.

9.3 Feasibility of 3D–3D tomography

This paper is a step towards '3D–3D tomography', denoting 3-D heterogeneity within the reference models and a 3-D physical domain for the model, from which we compute finite-frequency sensitivity kernels. (Based upon this labelling, the classical tomographic examples in this paper are 0D–2D, whereas the adjoint tomographic examples are 2D–2D.) Presently, our SEM codes are set up to compute 3D–3D sensitivity kernels on both regional and global scales (Liu & Tromp 2006, 2007). In this paper, we have highlighted some aspects of the inversion process that will be key to limiting the number of wavefield simulations required in the inversion.

Let us estimate the computational cost of a regional-scale tomographic inversion. As discussed, for 150 earthquakes we require 3150 total simulations for a seven iteration inversion. Each simulation takes approximately 35 min on 72 nodes (144 processors). Thus we can perform 40 runs per day on 72 nodes, and more than 500 runs per day on a 1000 node machine. Therefore, on this kind of hardware the whole inversion can theoretically be completed in about 1 week.

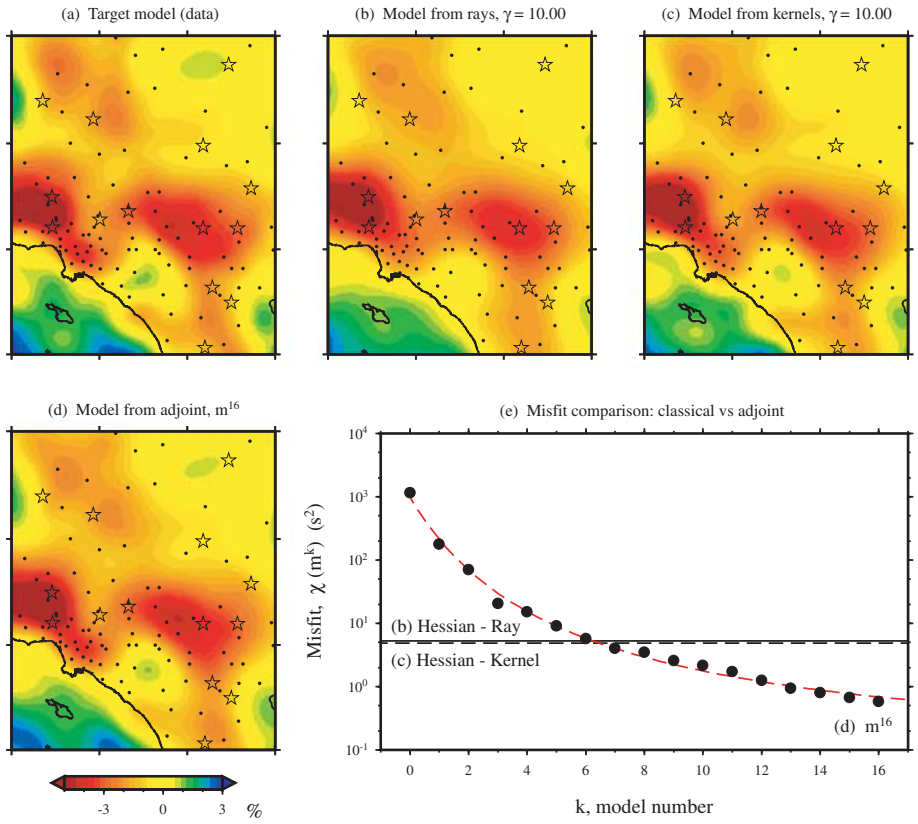

Figure 20. Comparison of recovered phase-speed models for classical and adjoint tomography. (a) Target model used to generate the data; expanded version is shown in Fig. 1(b). (b) Recovered model $\mathbf{m}^{\mathrm{ray}}$ using classical, ray-based inversion. The damping parameter γ is defined in (27). (c) Recovered model $\mathbf{m}^{\mathrm{ker}}$ using classical, kernel-based inversion (Fig. 4c). (d) Recovered model $\mathbf{m}^{16}$ using the adjoint method and a conjugate gradient algorithm (Fig. 13e). (e) Misfit comparison for the three approaches (eq. 6). The horizontal lines denote the misfit computed for the ray- and kernel-based models shown in (b) and (c) (see Section 9.2 for details).

To avoid reaching a local minimum in the optimization procedure, we intend to start by using longer-period waveforms, which existing 3-D models fit reasonably well, and work our way towards shorter periods. As we improve the model and increase the frequency contents of the waveforms, we expect to not only improve the fit to the current data used to constrain the model, but also to steadily increase the number of picks that is used in the inversion, that is, more and more parts of the seismograms are expected to be used and matched in the iterative inversion process. Unlike Akçelik *et al.* (2003), our emphasis will be on matching targeted, frequency-dependent body-wave traveltimes and surface-wave phase anomalies, rather than entire waveforms. Waveform tomography is largely controlled by amplitude differences, which are notoriously difficult to fit in seismology. Traveltime or phase, on the other hand, is a robust measure of misfit that has been used for decades to constrain local, regional and global earth models. From our perspective, the progression from ray-based traveltime tomography to finite-frequency 'banana–doughnut' tomography to frequency-dependent adjoint tomography is very natural.

ACKNOWLEDGMENTS

We thank Yann Capdeville and an anonymous reviewer for detailed comments and suggestions, which helped improve the manuscript. All figures except Figs 16 and 18 were produced using GMT (Wessel & Smith 1991). This research was supported by the National Science Foundation under grant EAR-0309576, and by the Southern California Earthquake Center (SCEC). SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 1009. This is contribution No. 9138 of the Division of Geological & Planetary Sciences, California Institute of Technology.

REFERENCES

Akçelik, V., Biros, G. & Ghattas, O., 2002. Parallel multiscale Gauss–Newton–Krylov methods for inverse wave propagation, *Proc. ACM/IEEE Supercomputing SC'2002 Conference,* published on CD-ROM and at www.sc-conference.org/sc2002.

Akçelik, V. et al., 2003. High resolution forward and inverse earthquake modeling on terascale computers, *Proc. ACM/IEEE Supercomputing SC'2003 Conference,* published on CD-ROM and at www.sc-conference.org/sc2003.

Bijwaard, H. & Spakman, W., 2000. Nonlinear global *P*-wave tomography by iterated linearized inversion, *Geophys. J. Int.,* **141,** 71–82.

Boschi, L., Ekström, G. & Kustowski, B., 2004. Multiple resolution surface wave tomography: the Mediterranean basin, *Geophys. J. Int.,* **157,** 293–304.

Bunks, C., Saleck, F.M., Zaleski, S. & Chavent, G., 1995. Multiscale seismic waveform inversion, *Geophysics,* **60,** 1457–1473.

Capdeville, Y., Chaljub, E., Vilotte, J.P. & 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. & Romanowicz, B., 2005. Towards global earth tomography using the spectral element method: a technique based on source stacking, *Geophys. J. Int.,* **162,** 541–554.

Dahlen, F.A. & Baig, A.M., 2002. Fréchet kernels for body-wave amplitudes, *Geophys. J. Int.,* **150,** 440–446.

Dahlen, F.A., Hung, S.-H. & Nolet, G., 2000. Fréchet kernels for finite-frequency traveltimes—I. Theory, *Geophys. J. Int.,* **141,** 157–174.

Fink, M., 1992. Time reversal of ultrasonic fields—Part I: Basic principles, *IEEE Trans. Ultras. Ferroelec. Freq. Contr.,* **39**(5), 555–566.

Fink, M., 1997. Time reversed acoustics, *Phys. Today,* **50,** 34–40.

Fink, M., Prada, C., Wu, F. & Cassereau, D., 1989. Self focusing in inhomogeneous media with "time reversal" acoustic mirrors, in *Proc. IEEE Ultrasonic. Symp. 1989,* Vol. 2, pp. 681–686.

Fletcher, R. & Reeves, C.M., 1964. Function minimization by conjugate gradients, *Comp. J.,* **7,** 149–154.

Gauthier, O., Virieux, J. & Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: Numerical results, *Geophysics,* **51,** 1387–1403.

Hansen, P.C., 1998. *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, Penn.

Hauksson, E., 2000. Crustal structure and seismicity distribution adjacent to the Pacific and North America plate boundary in southern California, *J. geophys. Res.,* **105,** 13 875–13 903.

Jennings, C.W., 1994. Fault activity map of California and adjacent areas, with locations and ages of recent volcanic eruptions, Calif. Div. Mines and Geology, Geologic Data Map No. 6, map scale 1:750,000.

Komatitsch, D. & Tromp, J., 1999. Introduction to the spectral element method for three-dimensional seismic wave propagation, *Geophys. J. Int.,* **139,** 806–822.

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

Komatitsch, D. & Tromp, J., 2002b. Spectral-element simulations of global seismic wave propagation—II. Three-dimensional models, oceans, rotation and self-gravitation, *Geophys. J. Int.,* **150,** 308–318.

Komatitsch, D. & 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.

Komatitsch, D., Ritsema, J. & Tromp, J., 2002. The spectral-element method, Beowulf computing, and global seismology, *Science,* **298,** 1737–1742.

Komatitsch, D., Liu, Q., Tromp, J., Süss, P., Stidham, C. & Shaw, J.H., 2004. Simulations of ground motion in the Los Angeles basin based upon the spectral-element method, *Bull. seism. Soc. Am.,* **94,** 187–206.

Liu, Q. & Tromp, J., 2006. Finite-frequency kernels based upon adjoint methods, *Bull. seism. Soc. Am.,* **96,** 2383–2397.

Liu, Q. & Tromp, J., 2007. Finite-frequency sensitivity kernels for global seismic wave propagation based upon adjoint methods, *Geophys. J. Int.,* in preparation.

Magistrale, H., Day, S., Clayton, R.W. & Graves, R., 2000. The SCEC Southern California reference three-dimensional velocity model Version 2, *Bull. seism. Soc. Am.,* **90**(6B), S65–S76.

Marquering, H., Dahlen, F.A. & Nolet, G., 1999. Three-dimensional sensitivity kernels for finite-frequency traveltimes: the banana-doughnut paradox, *Geophys. J. Int.,* **137,** 805–815.

Mora, P., 1987. Nonlinear two-dimensional elastic inversion of multioffset seismic data, *Geophysics,* **52,** 1211–1228.

Nolet, G., 1987. Waveform tomography, in *Seismic Tomography: With Applications in Global Seismology and Exploration Geophysics*, pp. 301–322, ed. Nolet, G., Reidel Publishing, Dordrecht, The Netherlands.

Peter, D., Tape, C., Boschi, L. & Woodhouse, J., 2006. Surface wave tomography: global membrane waves and adjoint methods, *Geophys. J. Int.,* submitted.

Pratt, R.G., 1999. Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model, *Geophysics,* **64,** 888–901.

Pratt, R.G., Shin, C.S. & Hicks, G.J., 1998. Gauss–Newton and full Newton methods in frequency–space seismic waveform inversion, *Geophys. J. Int.,* **133,** 341–362.

Ritsema, J., van Heijst, H.J. & Woodhouse, J.H., 1999. Complex shear velocity structure imaged beneath Africa and Iceland, *Science,* **286,** 1925–1928.

Süss, M.P. & Shaw, J.H., 2003. P-wave seismic velocity structure derived from sonic logs and industry reflection data in the Los Angeles basin, California, *J. geophys. Res.,* **108**(B3), 2170, doi:10.1029/2001JB001628.

Takeuchi, N. & Kobayashi, M., 2004. Improvement of seismological earth models by using data weighting in waveform inversion, *Geophys. J. Int.,* **158,** 681–694.

Talagrand, O. & Courtier, P., 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory, *Q. J. R. Meteorol. Soc.,* **113,** 1311–1328.

Tanimoto, T., 1990. Modelling curved surface wave paths: membrane surface wave synthetics, *Geophys. J. Int.,* **102,** 89–100.

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*, SIAM, Philadelphia, Penn.

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

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

Wang, Z., Tromp, J. & Ekström, G., 1998. Global and regional surface-wave inversion: a spherical-spline parameterization, *Geophys. Res. Lett.,* **25,** 207–210.

Wessel, P. & Smith, W.H.F., 1991. Free software helps map and display data, *EOS, Trans. Am. geophys. Un.,* **72**(41), 441 ff.

Woodhouse, J.H. & Dziewonski, A.M., 1984. Mapping the upper mantle: Three-dimensional modeling of Earth structure by inversion of seismic waveforms, *J. geophys. Res.,* **89,** 5953–5986.

Zhao, L. & Jordan, T.H., 2006. Structural sensitivities of finite-frequency seismic waves: a full-wave approach, *Geophys. J. Int.,* **165,** 981–990.

Zhao, L., Jordan, T.H. & Chapman, C.H., 2000. Three-dimensional Fréchet differential kernels for seismic delay times, *Geophys. J. Int.,* **141,** 558–576.

Zhao, L., Jordan, T.H., Olsen, K.B. & Chen, P., 2005. Fréchet kernels for imaging regional earth structure based on three-dimensional reference models, *Bull. seism. Soc. Am.,* **95,** 2066–2080.

Zhou, Y., Dahlen, F.A. & Nolet, G., 2004. Three-dimensional sensitivity kernels for surface wave observables, *Geophys. J. Int.,* **158,** 142–168.

Zhou, Y., Dahlen, F.A., Nolet, G. & Laske, G., 2005. Finite-frequency effects in global surface-wave tomography, *Geophys. J. Int.,* **163,** 1087–1111.

Zhu, L. & Kanamori, H., 2000. Moho depth variation in southern California from teleseismic receiver functions, *J. geophys. Res.,* **105** (B2), 2969–2980.

APPENDIX A: REGULARIZATION

Here we review the fact that stabilizing the Hessian matrix (as in eq. 5) via damping is equivalent to adding a regularization term R to the misfit function (6):

$$\chi_R(\mathbf{m}) = \chi(\mathbf{m}) + R(\mathbf{m}), \tag{A1}$$

whose gradient is, using (20),

$$\frac{\partial \chi_R}{\partial m_k} = \frac{\partial \chi}{\partial m_k} + \frac{\partial R}{\partial m_k} = -\sum_{i=1}^{N} G_{ik}\,\Delta T_i + \frac{\partial R}{\partial m_k}. \tag{A2}$$

There are many options for regularization. For illustrative purposes, we consider regularization according to the wave speed model itself:

$$R = \frac{1}{2}\gamma^2 \int_V (\delta \ln c)^2 \,\mathrm{d}^3\mathbf{x}, \tag{A3}$$

where γ is the damping parameter. Substituting (13), and then differentiating with respect to the kth model parameter, we obtain

$$R = \frac{1}{2}\gamma^2 \sum_{k=1}^{M} \delta m_k \sum_{k'=1}^{M} \delta m_{k'} D_{kk'}, \tag{A4}$$

$$\frac{\partial R}{\partial m_k} = \gamma^2 \sum_{k'=1}^{M} \delta m_{k'} D_{kk'}, \tag{A5}$$

where the $M \times M$ damping matrix $\mathbf{D}$ is given by

$$D_{kk'} = \int_V B_k \, B_{k'} \, \mathrm{d}^3\mathbf{x}. \tag{A6}$$

If the basis functions are orthonormal, then $\mathbf{D} = \mathbf{I}$, the identity matrix. Substituting (A5) into (A2), we obtain

$$\frac{\partial \chi_R}{\partial \mathbf{m}} = -\mathbf{G}^T\mathbf{d} + \gamma^2 \mathbf{D}\delta\mathbf{m}, \tag{A7}$$

where $\mathbf{D}$ is, for example, (A6) or (A10). Substituting this for $\mathbf{g}(\mathbf{m})$ into (5), with $\mathbf{H} = \mathbf{G}^T\mathbf{G}$, we obtain

$$\left(\mathbf{G}^T\mathbf{G} + \gamma^2\mathbf{D}\right)\delta\mathbf{m} = \mathbf{G}^T\mathbf{d}, \tag{A8}$$

which leads to (28). Eq. (A8) is known as 'Tikhonov' regularization or 'ridge regression', and is based on minimizing an L2-norm measure of $\mathbf{D}\,\delta\mathbf{m}$ (e.g. Hansen 1998, Ch. 5). (Typically these two labels refer to the case $\mathbf{D} = \mathbf{I}$.)

Instead, if we regularize using the gradient of the wave speed model (e.g. Akçelik *et al.* 2003), we obtain

$$R = \frac{1}{2}\gamma^2 \int_V \boldsymbol{\nabla}(\delta \ln c) \cdot \boldsymbol{\nabla}(\delta \ln c) \, \mathrm{d}^3\mathbf{x}, \tag{A9}$$

then the damping matrix is

$$D_{kk'} = \int_V (\boldsymbol{\nabla} B_k) \cdot (\boldsymbol{\nabla} B_{k'}) \, \mathrm{d}^3\mathbf{x}. \tag{A10}$$

Regularization according to the *roughness* of the model (e.g. Zhou *et al.* 2005) leads to

$$R = \frac{1}{2}\gamma^2 \int_V \left(\nabla^2 \delta \ln c\right)^2 \, \mathrm{d}^3\mathbf{x}, \tag{A11}$$

$$D_{kk'} = \int_V \left(\nabla^2 B_k\right)\left(\nabla^2 B_{k'}\right) \mathrm{d}^3\mathbf{x}. \tag{A12}$$

Different norms or constraints, as well as combinations of constraints (resulting in multiple damping parameters), may be used in forming $R(\mathbf{m})$. For example, Akçelik *et al.* (2002) advocated the use of L1-based, 'total variation' regularization, which avoids smoothing of sharp gradients in material properties. Akçelik *et al.* (2003) applied L1 regularization for the structure gradients and L2 regularization for the source gradients.

APPENDIX B: DETAILS OF THE CONJUGATE GRADIENT ALGORITHM

The computation of the misfit value $\chi(\mathbf{m})$ and gradient $\mathbf{g}$ takes $2N_{\text{events}}$ simulations. Because each simulation is expensive, it is important to limit the number of simulations in the inverse problem, which we approach using a conjugate gradient algorithm (Section 6.1). Two possible areas to aid in this are the selection of the trial step ν_t and the choice of the polynomial (quadratic or cubic) to use in the interpolation. In this section, we have omitted the superscript k on quantities to avoid clutter.

B1 Selection of the trial step

The trial step, or test parameter, ν_t, determines how far away from the current model to go in the search direction in order to obtain a test model (and, possibly, test gradient). Given the misfit value, $\chi(\mathbf{m})$, and the gradient, $\mathbf{g}(\mathbf{m})$, for the current model, the user is faced with determining how far to step in the search direction away from the current model to obtain a test model, for which an additional misfit value will be computed. The gradient vector $\mathbf{g}$ is represented in the conjugate gradient algorithm as a slope, $\tilde{g}(0)$, and the misfit function in the search direction by $\tilde{\chi}(\nu)$. In the algorithm, we select the test parameter by interpolating $\tilde{\chi}(\nu)$ using a quadratic polynomial, $R(\nu)$:

$$R(\nu) = a\nu^2 + b\nu + c, \tag{B1}$$

where a, b and c are determined using the value (r_1) and slope (g_1) for the current model, and a test model location such that the value (r_2) and slope (g_2) of $R(\nu)$ at ν_t are both zero (see Fig. 11d). The four values are given by

$$r_1 \equiv R(0) = \tilde{\chi}(0) = \chi(\mathbf{m}),$$
$$g_1 \equiv R'(0) = \tilde{g}(0),$$
$$r_2 \equiv R(\nu_t) = 0,$$
$$g_2 \equiv R'(\nu_t) = 0.$$

These equations can be used to determine the coefficients of $R(\nu)$:

$$a = -g_1/(2\nu_t) = g_1^2/(4r_1),$$
$$b = g_1,$$
$$c = r_1,$$

as well as the test parameter

$$\nu_t = \frac{-2r_1}{g_1} = \frac{-2\chi(\mathbf{m})}{\tilde{g}(0)}, \tag{B2}$$

which is the value used in the algorithm discussed in Section 6.1.

The 'test model parabola' $R(\nu)$ is chosen such that its vertex lies on $\chi = 0$; however, one could choose the vertex at some $\chi > 0$ that is determined based upon the change in misfit from a previous step. The quadratic extrapolation through $\chi = 0$ is perhaps too conservative, and computational savings—in the form of better convergence—could be obtained by exploring the choice of the initial step.

B2 Quadratic versus cubic interpolation

As discussed in Section 6.3, the tomographer is faced with the choice of using a second- or third-order polynomial, $P(\nu)$, in the interpolation scheme within the conjugate gradient algorithm. Here we outline the formulas required to compute an analytical minimum, $\nu_{\min}$, using each interpolation scheme.

With an order-3 polynomial, one must have four quantities in addition to the test parameter ν_t: the misfit and gradient for the current model—$\chi(\mathbf{m})$ and $\mathbf{g}$—and the misfit and gradient for the test model—$\chi(\mathbf{m}_t)$ and $\mathbf{g}_t$. These values are converted into scalar values for an interpolating polynomial $P(\nu)$:

$$\nu_1 = 0,$$
$$\nu_2 = \nu_t,$$
$$p_1 \equiv P(\nu_1) = \chi(\mathrm{m}),$$
$$g_1 \equiv P'(\nu_1) = \tilde{g}(0),$$
$$p_2 \equiv P(\nu_2) = \chi(\mathrm{m}_t),$$
$$g_2 \equiv P'(\nu_2) = \tilde{g}(\nu_t).$$

The cubic polynomial can be written in terms of these quantities as

$$P(\nu) = a(\nu - \nu_1)^3 + b(\nu - \nu_1)^2 + c(\nu - \nu_1) + d, \tag{B3}$$

where

$$a = [-2(p_2 - p_1) + (g_1 + g_2)(\nu_2 - \nu_1)] / (\nu_2 - \nu_1)^3,$$
$$b = [3(p_2 - p_1) - (2g_1 + g_2)(\nu_2 - \nu_1)] / (\nu_2 - \nu_1)^2,$$
$$c = g_1,$$
$$d = p_1.$$

An analytical minimum of $P(\nu)$ can be obtained when $|c| > 0$:

$$\nu_{\min} = \begin{cases} \nu_1 + [-b + (b^2 - 3ac)^{1/2}] / (3a), & a \neq 0 \text{ and } b^2 - 3ac > 0 \\ -c/(2b), & a = 0 \text{ and } b \neq 0;\ b^2 - 3ac < 0. \end{cases} \tag{B4}$$

With an order-2 polynomial, the gradient of the test model—$\mathbf{g}_t$ or g_2—is not required. In this case, the quadratic polynomial can be written in terms of (B3) as

$$P(\nu) = a(\nu - \nu_1)^2 + b(\nu - \nu_1) + c, \tag{B5}$$

where

$$a = [(p_2 - p_1) - g_1(\nu_2 - \nu_1)] / (\nu_2^2 - \nu_1^2),$$
$$b = g_1,$$
$$c = p_1 - a\nu_1^2 - b\nu_1.$$

The analytical minimum of $P(\nu)$ is simply

$$\nu_{\min} = -b/(2a). \tag{B6}$$

Based on our experiments, the quadratic interpolation is preferred over the cubic interpolation, since it costs $3N_{\text{events}}$ per iteration (versus $4N_{\text{events}}$) and performs only slightly worse (Fig. 13f).