On the use of sensitivity tests in seismic tomography

N. Rawlinson[1] and W. Spakman[2,3]

[1]*School of Geosciences, University of Aberdeen, Aberdeen, Scotland AB24 3UE. E-mail:* nrawlinson@abdn.ac.uk
[2]*Department of Earth Sciences, Faculty of Geosciences, Utrecht University, PO BOX 80115, 3584 TC, Utrecht, The Netherlands*
[3]*Centre of Earth Evolution and Dynamics (CEED), University of Oslo,* 0316 *Oslo, Norway*

SUMMARY

Sensitivity analysis with synthetic models is widely used in seismic tomography as a means for assessing the spatial resolution of solutions produced by, in most cases, linear or iterative nonlinear inversion schemes. The most common type of synthetic reconstruction test is the so-called checkerboard resolution test in which the synthetic model comprises an alternating pattern of higher and lower wave speed (or some other seismic property such as attenuation) in 2-D or 3-D. Although originally introduced for application to large inverse problems for which formal resolution and covariance could not be computed, these tests have achieved popularity, even when resolution and covariance can be computed, by virtue of being simple to implement and providing rapid and intuitive insight into the reliability of the recovered model. However, checkerboard tests have a number of potential drawbacks, including (1) only providing indirect evidence of quantitative measures of reliability such as resolution and uncertainty, (2) giving a potentially misleading impression of the range of scale-lengths that can be resolved, and (3) not giving a true picture of the structural distortion or smearing that can be caused by the data coverage. The widespread use of synthetic reconstruction tests in seismic tomography is likely to continue for some time yet, so it is important to implement best practice where possible. The goal of this paper is to develop the underlying theory and carry out a series of numerical experiments in order to establish best practice and identify some common pitfalls. Based on our findings, we recommend (1) the use of a discrete spike test involving a sparse distribution of spikes, rather than the use of the conventional tightly spaced checkerboard; (2) using data coverage (e.g. ray-path geometry) inherited from the model constrained by the observations (i.e. the same forward operator or matrix), rather than the data coverage obtained by solving the forward problem through the synthetic model; (3) carrying out multiple tests using structures of different scale length; (4) taking special care with regard to what can be inferred when using synthetic structures that closely mimic what has been recovered in the observation-based model; (5) investigating the range of structural wavelengths that can be recovered using realistic levels of imposed data noise; and (6) where feasible, assessing the influence of model parametrization error, which arises from making a choice as to how structure is to be represented.

Key words: Inverse theory; Seismic tomography.

1 INTRODUCTION

The effective assessment of model reliability is still a major challenge in seismic tomography despite over four decades of development. The ill-posed nature of the tomographic inverse problem means that multiple data-satisfying solutions exist, and these solutions tend to be unstable with respect to small changes in data noise and initial conditions unless regularization is applied (Rawlinson *et al.* 2014). The combination of implicit (e.g. via choice of a model parametrization) and explicit (e.g. damping and smooth-ing) regularization, a poor knowledge of data noise, uncertainty in prior information, and simplifying assumptions made in the forward and inverse step (e.g. using the ray approximation, ignoring anisotropy, linearization) actually make it extremely difficult for any method to provide a reliable assessment of model robustness.

The pioneers of geophysical inverse problems (Backus & Gilbert 1967, 1968, 1970; Franklin 1970; Wiggins 1972) and seismic tomography (Aki & Lee 1976; Aki *et al.* 1977) clearly recognized the importance of assessing solution robustness and the need to try and quantify uncertainty. For example, in both the local earthquake

tomography study of Aki & Lee (1976) and the teleseismic tomography study of Aki *et al.* (1977), formal estimates of posterior covariance and resolution are made for all model unknowns. Despite the massive increases in computing power and ongoing theoretical developments that have taken place since these seminal works, progress in the assessment of model robustness has been limited. As well as estimates of posterior covariance and uncertainty from linear theory, other methods that have been tried include synthetic reconstruction tests, jackknife and bootstrap methods and sampling strategies (for a review see Rawlinson *et al.* 2014).

For small to medium-sized linear or linearizable inverse problems, the calculation of posterior covariance and resolution is relatively straightforward, and has been frequently used as a measure of solution reliability in seismic tomography (e.g. Aki *et al.* 1977; White 1989; Zelt & Smith 1992; Steck *et al.* 1998). For larger problems, approximation methods have been applied (e.g. Yao *et al.* 1999; Zhang & Thurber 2007), although care must be taken when iterative solution methods like LSQR, which only explore a restricted subspace of the full model space when the number of iteration steps is limited, are used (Nolet *et al.* 1999). However, while the attainment of quantitative estimates of model uncertainty and spatial resolution is attractive, the general limitations associated with the tomographic inverse problem, as described earlier, means that at the very least the absolute value of these quantities are poorly constrained. Moreover, their validity decreases as the nonlinearity of the inverse problem increases.

An alternative to directly computing resolution or posterior covariance is to instead target a proxy such as resolution length. For example, Fichtner & Trampert (2011) develop a method for full waveform inversion problems, based on low-rank approximations of the Hessian operator that can determine the 3-D distribution of direction-dependent resolution lengths with potentially greater computational efficiency than a synthetic test. This approach can be viewed as a generalization of the ray density tensor (Kissling 1988), which quantifies space-dependent azimuthal coverage. More recently, Fichtner & van Leeuwen (2015) develop and apply a random probing technique for resolution analysis that avoids the algorithmic complexity and computational requirements of the Fichtner & Trampert (2011) method. In addition, it can be applied to any tomographic technique, including full waveform inversion and linearized ray tomography. A thorough comparison between this technique and more traditional methods of resolution analysis is yet to be carried out.

Sensitivity analyses, or synthetic reconstruction tests, are the most common method for assessing solution robustness in seismic tomography, and were originally introduced to investigate spatial resolution in inverse problems for which formal resolution cannot be computed due to the large model size (Spakman & Nolet 1988). They involve the formulation of a heterogeneous synthetic model through which the forward problem is solved using the identical data coverage as the observational data set. The inversion method is then applied to the synthetic data set in an effort to recover the test model. Differences between the true and recovered model provide a basis for assessing the reliability of the solution. Input structures range from single discrete spikes (Walck & Clayton 1987; Rasmussen & Humphreys 1988), to widely spaced spikes (Grand 1987; Spakman & Nolet 1988), to tightly spaced checkerboards (Inoue *et al.* 1990; Fukao *et al.* 1992; Glahn & Granet 1993; Zhao *et al.* 1996; Zelt & Barton 1998; Gorbatov *et al.* 2000; Fishwick *et al.* 2005; Chen & Jordan 2007; Yang *et al.* 2009; Rawlinson *et al.* 2011) to structures designed to investigate particular features such as subduction zones (Spakman *et al.* 1989; Eberhart-Phillips &

Reyners 1997; Bijwaard *et al.* 1998; Graeber & Asch 1999; Wolfe *et al.* 2002). Of these, the tightly spaced checkerboard pattern of alternating positive and negative anomalies (relative to some reference structure) is by far the most common, but also the most criticized. For example, Lévêque *et al.* (1993) clearly show that it is possible to devise data geometries for which small-scale structures (as is commonly used in a checkerboard test) are well retrieved while larger scale structures are poorly retrieved. More generally, their results show - unsurprisingly - that the recovery has a dependence on the choice of the synthetic model.

Standard statistical methods of error assessment such as jackknife and bootstrap tests have seen limited use in seismic tomography (Lees & Crosson 1989, 1990; Su & Dziewonski 1997; Zelt 1999; Gung & Romanowicz 2004). Both methods rely on running repeat inversions with various subsets of the data set (via resampling in the case of bootstrap, and omission in the case of jackknife) and interrogating the ensemble of solutions that are produced for summary information. One of the main limitations of both methods is that they require an overdetermined inverse problem in order to work effectively (Nolet *et al.* 1999), which rarely occurs in practice.

The idea of sampling regions of model space in which the data are satisfied in order to produce an ensemble of solutions is gradually becoming more common, thanks in part to ongoing increases in computing power. Within a linear or weakly nonlinear framework, a number of different techniques have been tried, including multiple starting models (Vasco *et al.* 1996), the so-called 'null space shuttle' (Deal & Nolet 1996; de Wit *et al.* 2012), regularized extremal bounds analysis (Meju 2009), Lie group methods (Vasco 2007) and the dynamic objective function scheme (Rawlinson & Kennett 2008). However, it is in the realm of fully nonlinear sampling where the greatest strides are currently being made. For example, Bayesian trans-dimensional tomography, in which the number of unknowns is an unknown itself, and the parametrization is adaptive, is starting to become increasingly popular. It has been shown to be computationally tractable for most 2-D and some 3-D problems (Bodin & Sambridge 2009; Bodin *et al.* 2012; Young *et al.* 2013; Galetti *et al.* 2015; Piana Agostinetti *et al.* 2015; Hawkins & Sambridge 2015), and generates a large ensemble of solutions which can be used to quantitatively assess solution reliability. In general, these methods have an intrinsic parsimony which results in a (variable) spatial resolution that contains only structure that is 'required' by the data; the uncertainty associated with structure at this spatial resolution can be estimated by taking the standard deviation of the solution ensemble.

In this paper, we focus on establishing best practice for the implementation of sensitivity analysis, which has become the standard approach for assessing solution quality in seismic tomography, even when the resolution matrix can actually be computed (in which case, the recovery of any test structure can be rapidly obtained; see Section 2). As well as results and findings from previous studies, these guidelines are founded in theory and informed by a series of numerical tests that use both synthetic data (see Section 3) and data recorded in the field (see Supporting Information). Note that we only consider tomographic inverse problems that are linear or weakly nonlinear. As such, it is assumed that the starting model is sufficiently close to the required minimum of the objective function that the assumption of local linearization will allow the solution to enter a region of acceptable data-fit (whether by a single step or iteratively) within the neighbourhood of this minimum. If this is not the case, then the problem may require a fully nonlinear solver and consequently a different approach to assessing solution robustness.

2 THEORETICAL CONSIDERATIONS

Our aim in this section is to provide a theoretical framework for the proper use of spatial resolution and noise analysis with synthetic models in tomographic inverse problems that solve a linear(ized) system of equations. In addition, we address how implicit formulations of the inverse problem that focus on minimizing a cost function may adhere to this theory.

2.1 The forward problem

The general nonlinear forward problem of noise-free data $\mathbf{d}_E$ (e.g. traveltimes, waveforms, some portion of the seismogram) can be written as $\mathbf{d}_E = \mathbf{g}(\mathbf{m}_E) + \boldsymbol{\epsilon}_t$, where $\mathbf{m}_E$ represents the distribution of some true-Earth seismic property (e.g. *P*-wave velocity) and $\mathbf{g}$ is a nonlinear integral operator. The error $\boldsymbol{\epsilon}_t$ represents the discrepancy between $\mathbf{d}_E$ and its theoretical prediction $\mathbf{g}(\mathbf{m}_E)$ and the magnitude of $\boldsymbol{\epsilon}_t$ is a function of the approximations made in solving the forward problem. For example, the tomographic inverse problem in its simplest form, using ray theory, is nonlinear and the finite frequency traveltime $\mathbf{d}_E$ may be inconsistent with the theoretical ray prediction $\mathbf{g}(\mathbf{m}_E)$.

If we now add observational noise $\boldsymbol{\epsilon}_d$ (e.g. picking error), such that $\mathbf{d} = \mathbf{d}_E + \boldsymbol{\epsilon}_d$, then

$$\mathbf{d} = \mathbf{g}(\mathbf{m}_E) + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_d. \tag{1}$$

Next, assuming a suitable model parametrization that projects $\mathbf{m}_E$ on the model vector $\mathbf{m}_p$, this leads to

$$\mathbf{d} = \mathbf{g}(\mathbf{m}_p) + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_p. \tag{2}$$

where $\boldsymbol{\epsilon}_p$ is the implied parametrization error. In the explicit formulation that we follow here, the next step is defining the matrix representation of the integral equations as a locally linear representation of $\mathbf{g}(\mathbf{m}_p)$ about the true model $\mathbf{m}_p$, so that

$$\mathbf{d}_p = \mathbf{G}_p \mathbf{m}_p \tag{3}$$

where $\mathbf{d}_p = \mathbf{g}(\mathbf{m}_p)$ and $\mathbf{G}_p$ is the observation matrix relating the data $\mathbf{d}_p$ to the true Earth model $\mathbf{m}$. We assume that this relationship is exact, so that substitution into eq. (2) yields

$$\mathbf{d} = \mathbf{G}_p \mathbf{m}_p + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_p. \tag{4}$$

Note that from the inversion viewpoint, eq. (4) is still a nonlinear relationship between observations $\mathbf{d}$ and $\mathbf{m}_p$ because $\mathbf{G}_p$ and $\mathbf{m}_p$ are both unknown and in addition $\mathbf{G}_p$ generally depends on $\mathbf{m}_p$. For instance, in ray-based tomography, $\mathbf{G}_p$ is obtained from integration along the 3-D geometry of the true ray paths that depends on the true model $\mathbf{m}_p$.

For a linear(ized) inversion of eq. (4) one assumes an approximation $\mathbf{G} = \mathbf{G}_p$ where $\mathbf{G}$ is a known matrix obtained from integration over a background reference model. This leads to an additional inconsistency between the data and model predictions, which we denote with the linearization error $\boldsymbol{\epsilon}_l$. For later use we formalize this as follows

$$\mathbf{G}_p \mathbf{m}_p = (\mathbf{G} - \mathbf{E})\mathbf{m}_p = \mathbf{G}\mathbf{m}_p + \boldsymbol{\epsilon}_l \tag{5}$$

where $\mathbf{E} = (\mathbf{G} - \mathbf{G}_p)$ is the linearization error in the observation matrix.

The forward equation of a linearized inversion thus becomes $\mathbf{d} = \mathbf{G}\mathbf{m}_p + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_l + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_p$. Although the observational errors $\boldsymbol{\epsilon}_d$ can be random, the implicit data errors may have a systematic component, which would likely manifest more prominently in the tomographic solution. In tomographic formulations

of the inverse problem, data errors $\boldsymbol{\epsilon}$ are often represented by the data covariance matrix $\mathbf{C}_d$, which in theory can accommodate all variances and error correlations.

Due to a lack of data and the presence of data errors, the tomographic inverse problem is generally underdetermined or mixed-determined (Tarantola 1987; Menke 1989). Consequently, a null space exists in which modifications can be made to $\mathbf{m}_p$ without changing the data $\mathbf{d}$. As a result, we cannot expect to retrieve $\mathbf{m}_p$ from the inversion of $\mathbf{d} = \mathbf{G}\mathbf{m}_p + \boldsymbol{\epsilon}$, as it is only one of many data satisfying solutions. To make this explicit, $\mathbf{m}_p$ is replaced by $\mathbf{m}$, which denotes any model that can fit the data equally as well as $\mathbf{m}_p$ based on an adopted measure of data fit. This leads to the well-known form of the linearized forward equation

$$\mathbf{d} = \mathbf{G}\mathbf{m} + \boldsymbol{\epsilon} \tag{6}$$

where we make explicit that $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_l + \boldsymbol{\epsilon}_p$ comprises all sources of observation error and errors introduced by the various assumptions that we have outlined above.

2.2 The linearized inverse problem and the resolution matrix

Here we take a generalized and practical approach to tomographic inversion which is largely independent of the number of model parameters and volume of data. The two main features are (1) solution of a linear/linearized inverse problem, and (2) assessment of solution robustness either via computation of the resolution matrix or sensitivity analysis. In general, sensitivity analysis is the only way to assess solution characteristics for problems involving a very large number of inversion parameters.

Formulation of the tomographic inverse problem starts with the design of a data misfit function, which is usually extended by the addition of one or more regularization terms which penalizes model attributes, such as model amplitude, flatness and/or smoothness (Menke 1989). Notions about prior model covariance are also used to precondition the inverse problem (e.g. Rawlinson *et al.* 2014). The penalty term serves to deal with the underdetermined nature of the inverse problem, usually by trying to exclude models which exhibit levels of detail not required by the data. A basic example is the cost function associated with the linearized and regularized least squares inversion (Rawlinson *et al.* 2010):

$$S(\mathbf{m}) = (\mathbf{d} - \mathbf{G}\mathbf{m})^{\mathrm{T}}\mathbf{C}_d^{-1}(\mathbf{d} - \mathbf{G}\mathbf{m}) + \alpha^2 \mathbf{m}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{m} \tag{7}$$

The first term is the quadratic data misfit scaled by the prior data covariance $\mathbf{C}_d$ and the second term defines the penalties on model attributes through a pre-designed damping matrix $\alpha\mathbf{D}$. The tuning parameter α regulates the trade-off between fitting the data and satisfying the model penalties.

In order to solve the inverse problem, the next step is to find the model $\hat{\mathbf{m}}$ which minimizes the adopted cost function $S(\mathbf{m})$. For linear problems, the solution can be formally written as

$$\hat{\mathbf{m}} = \mathbf{G}^{-g}\mathbf{d} \tag{8}$$

where $\mathbf{G}^{-g}$ is the so-called generalized inverse (Backus & Gilbert 1970). For eq. (7), the generalized inverse has the form (e.g. Rawlinson *et al.* 2014)

$$\mathbf{G}^{-g} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{G} + \alpha^2 \mathbf{D}^{\mathrm{T}}\mathbf{D}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1} \tag{9}$$

This illustrates that generally $\mathbf{G}^{-g}$ is not unique and is dependent on $\mathbf{C}_d$, representing all data errors $\boldsymbol{\epsilon}_d$, and on the adopted regularization operator $\alpha\mathbf{D}$.

If we now substitute the relationship between the observations and the true Earth model of eq. (4) into eq. (8), we obtain an instructive equation that shows how the true Earth model and actual data errors map into the tomographic solution

$$\hat{\mathbf{m}} = \mathbf{R}_p \mathbf{m}_p + \mathbf{G}^{-g}(\boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_p) \tag{10}$$

where

$$\mathbf{R}_p = \mathbf{G}^{-g}\mathbf{G}_p \tag{11}$$

is the true spatial resolution matrix and defines the linear dependence of each model parameter $\hat{m}_i$ on components of the true model $\mathbf{m}_p$. In effect, it measures the amount of 'blurring' produced by the inverse operator. If $\mathbf{R} = \mathbf{I}$, then there is no blurring and the least squares solution has perfect resolution.

For a linear inverse problem $\mathbf{G}_p = \mathbf{G}$, and hence the resolution matrix can formally be computed. Instead, for a linearized problem, which is the usual case in seismic tomography, we do not know $\mathbf{G}_p$. However, by inserting eq. (5) into eq. (10), we can arrive at a useful equation which allows for a practical implementation of eqs (10) and (11) and explicitly shows the role of the linearization error, which is not present in eq. (10):

$$\hat{\mathbf{m}} = \mathbf{R}\mathbf{m}_p - \mathbf{G}^{-g}\mathbf{E}\mathbf{m}_p + \mathbf{G}^{-g}(\boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_p) \tag{12}$$

or

$$\hat{\mathbf{m}} = \mathbf{R}\mathbf{m}_p + \mathbf{G}^{-g}\boldsymbol{\epsilon} \tag{13}$$

where $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$ is the more commonly used resolution matrix (e.g. Nolet 2008). Eq. (12) demonstrates that $\mathbf{R}$ can only be considered to be an approximation to the true resolution matrix $\mathbf{R}_p$ in case of a linearized inverse problem, that is, $\mathbf{R}_p = \mathbf{R} - \mathbf{G}^{-g}\mathbf{E}$. This is an important observation that has an analogy when considering the spatial resolution of the final model of a nonlinear inversion (see below).

The error term $\mathbf{G}^{-g}\boldsymbol{\epsilon}$ in eq. (13) includes the linearization error and explicitly accounts for the propagation of all data noise into the solution. As this noise is unknown, it cannot be computed except in controlled synthetic experiments. As an approximation, the effect of data errors $\boldsymbol{\epsilon}$ on the solution is estimated from the propagation of the prescribed data covariance $\mathbf{C}_d$ into the posterior model covariance $\mathbf{C}_M$. Since the covariance of any linear combination $\mathbf{Ap}$ of a Gaussian distributed random variable $\mathbf{p}$ is $\text{Cov}(\mathbf{Ap}) = \mathbf{A}\text{Cov}(\mathbf{p})\mathbf{A}^T$, then the posterior covariance of the model parameters is (using eq. (8))

$$\mathbf{C}_M = \mathbf{G}^{-g}\mathbf{C}_d(\mathbf{G}^{-g})^T \tag{14}$$

Thus, the model covariance for a least squares inverse problem depends on the data errors and not the data itself (since $\mathbf{G}^{-g}$ is not a function of $\mathbf{d}$), so any assumptions used in the construction of $\mathbf{C}_d$ will also determine the quality of the tomographic solution.

2.3 Sensitivity analysis with synthetic velocity models

For tomographic inverse problems of the class described above (linear/linearized least squares), it has long been recognized that the computation of $\mathbf{R}$ may not be practical when large numbers of model parameters are involved. Motivated by the form of eq. (13), this has led to the use of sensitivity analysis with synthetic models and synthetic data (e.g. Spakman & Nolet 1988). This is carried out by constructing a synthetic model $\mathbf{m}_s$ (e.g. a checkerboard or

spike model), from which synthetic data are computed by forward solution with $\mathbf{G}$, and by optionally adding synthetic noise yielding

$$\mathbf{d}_s = \mathbf{G}\mathbf{m}_s + \boldsymbol{\epsilon}_s. \tag{15}$$

Solving $\mathbf{d}_s = \mathbf{G}\mathbf{m} + \boldsymbol{\epsilon}_s$ in the same way as eq. (6) leads to the solution model $\hat{\mathbf{m}}_s$, and analogous to eq. (13), $\hat{\mathbf{m}}_s$ relates to the synthetic model $\mathbf{m}_s$ as

$$\hat{\mathbf{m}}_s = \mathbf{R}\mathbf{m}_s + \mathbf{G}^{-g}\boldsymbol{\epsilon}_s \tag{16}$$

Because the resolution matrix $\mathbf{R}$ in eqs (16) and (13) are identical, sensitivity analysis with noise-free synthetic data constitutes a correct basis for inferring properties of $\mathbf{R}$ from the comparison of the synthetic model $\mathbf{m}_s$ with the tomographic solution $\hat{\mathbf{m}}_s$. Experiments that are restricted to inverting synthetic noise, in effect by taking $\mathbf{m}_s = 0$ and inverting a synthetic noise vector, can be useful to assess the importance of the second term in eq. (13). Inversion of noisy synthetic data is used to examine the combined effects of lack of resolution and noise propagation on the tomographic solution (see Rawlinson *et al.* 2014, for more details). However, it is worth noting that the full information content of $\mathbf{R}$ cannot be retrieved from a single synthetic test, even in the absence of imposed data noise, because $\mathbf{R}$ is not uniquely constrained by the relationship $\hat{\mathbf{m}}_s = \mathbf{R}\mathbf{m}_s$. If the full blurring effects are to be understood, then a spike test for each model parameter would need to be run independently, and the results collated. A single spike test of this sort would retrieve the information equivalent of only one column or row of $\mathbf{R}$.

2.4 Nonlinear inversion and sensitivity analysis

In nonlinear inversion, a set of observations $\mathbf{d}$ of the nonlinear integral functional $\mathbf{g}(\mathbf{m}_p)$ is used for developing a model space search for the model $\mathbf{m}_p$. As before, we write $\mathbf{d} = \mathbf{d}_p + \boldsymbol{\epsilon}$ such that the exact equation is $\mathbf{d}_p = \mathbf{g}(\mathbf{m}_p)$. Note that $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_p$ and does not include the linearization error $\boldsymbol{\epsilon}_l$. By defining a weighted cost function $||\mathbf{d} - \mathbf{g}(\mathbf{m})||_w$ a sequence of models $\mathbf{m}^k, k = 1, \ldots, K$ can be created (usually under the assumption of weak nonlinearity), that aims at sequentially minimizing $||\mathbf{d} - \mathbf{g}(\mathbf{m})||_w$ such that at convergence we have found a model $\hat{\mathbf{m}} \approx \mathbf{m}^K$ with an acceptable data misfit $||\mathbf{d} - \mathbf{g}(\hat{\mathbf{m}})||_w \approx ||\boldsymbol{\epsilon}||_w$. An example of the objective or cost function is

$$S(\mathbf{m}) = (\mathbf{d} - \mathbf{g}(\mathbf{m}))^T \mathbf{C}_d^{-1}(\mathbf{d} - \mathbf{g}(\mathbf{m})) + \alpha^2 P(\mathbf{m}) \tag{17}$$

where $P(\mathbf{m})$ stands for any combination of regularization and prior model information.

If $\mathbf{d}^k$ is the prediction of $\mathbf{d}$ after step k, the nonlinear equation $\mathbf{d}^k = \mathbf{g}(\mathbf{m}^k)$ is formally represented by $\mathbf{d}^k = \mathbf{G}^k\mathbf{m}^k$ where $\mathbf{G}^k$ is the observation matrix associated with $\mathbf{m}^k$. Similarly, we can write $\mathbf{d} = \mathbf{G}\hat{\mathbf{m}} + \hat{\boldsymbol{\epsilon}}$ for the final solution $\hat{\mathbf{m}} = \mathbf{m}^K$ and $\mathbf{d}_p = \mathbf{G}_p\mathbf{m}_p$ for the true solution. The different data errors ($\boldsymbol{\epsilon}$ and $\hat{\boldsymbol{\epsilon}}$) reflect the risk of irrecoverably mapping part of the original errors into the solution while stepping through model space. A nonlinear inversion scheme particularly aims to improve the solution of the forward problem such that $\mathbf{G}^k \rightarrow \mathbf{G}_p$ and consequently $\mathbf{m}^k \rightarrow \mathbf{m}_p$. For weakly nonlinear problems, $\mathbf{G}^k$ is created explicitly as a result of step-wise linearization. Strongly nonlinear tomographic problems are often solved via a direct model-space search and do not necessarily involve an explicit formulation of $\mathbf{G}^k$. This class of solution strategy can also address pure least squares formulations of the inverse problem ($\alpha = 0$ in eq. 17), but in this case ad hoc decisions are usually required to choose what can be regarded as acceptable models, for example, selecting the minimum norm solution.

In a linearized inverse problem, we implicitly accept that $\mathbf{G} \approx \mathbf{G}_p$, or $\mathbf{G} = \mathbf{G}_p + \mathbf{E}$ (see eq. 5). In nonlinear inversion we want to improve on this approximation, but the attempted convergence $\mathbf{G}^k \to \mathbf{G}_p$ generally gives rise to the problem of nonlinear error propagation. For example, in ray-based tomography a step-wise linearization of the inverse problem requires 3-D ray tracing at step k through model $\mathbf{m}^k$, which is required for determination of the next observation matrix $\mathbf{G}^{k+1}$. Model errors in $\mathbf{m}^k$ (i.e. lack of spatial resolution and model amplitude error) cause errors in the 3-D ray geometries that propagate into $\mathbf{G}^{k+1}$, which is subsequently used to invert for $\mathbf{m}^{k+1}$. Lack of spatial resolution at each step may prevent ray geometry errors from being corrected in subsequent steps and can even accumulate during iteration. This affects the convergence to the final model in an intractable way and may force convergence to a local minimum of the misfit function $||\mathbf{d} - \mathbf{g}(\mathbf{m}^k)||_w$. Hence, as a result of data errors, of insufficient data constraints and possibly of the model search strategy used, the model $\mathbf{m}^{k+1}$ depends on previous models and contains model errors as a result of nonlinear errors that propagated into $\mathbf{G}^{k+1}$, which was used for determining $\mathbf{m}^{k+1}$. This generally affects the convergence $\mathbf{G}^k \to \mathbf{G}_p$ such that $\mathbf{G}$ does not necessarily converge to $\mathbf{G}_p$ and consequently $\mathbf{m}^k$ does not necessarily converge to $\mathbf{m}_p$. In analogy with the linearized inversion (Section 2.2), we express this error as $\mathbf{E} = \mathbf{G} - \mathbf{G}_p$ for the accepted final model $\hat{\mathbf{m}} = \mathbf{m}^k$.

The above general description allows for the creation of a theoretical starting point for the application of sensitivity analysis to the solution of nonlinear inverse problems. For the final solution $\hat{\mathbf{m}}$ we can formally write $\hat{\mathbf{m}} = \mathbf{G}^{-g}\mathbf{d}$ (see eq. 8). Here, the generalized inverse $\mathbf{G}^{-g}$ stands for a repeatable inversion process to obtain the solution $\hat{\mathbf{m}}$, or for an explicit operator obtained from the last inversion step, and incorporates the propagated forward problem error $\mathbf{E}$. Thus we can write

$$\hat{\mathbf{m}} = \mathbf{G}^{-g}(\mathbf{d}_p + \boldsymbol{\epsilon}) = \mathbf{G}^{-g}(\mathbf{G}_p\mathbf{m}_p + \boldsymbol{\epsilon}) = \mathbf{G}^{-g}((\mathbf{G} - \mathbf{E})\mathbf{m}_p + \boldsymbol{\epsilon})$$
$$= \mathbf{R}\mathbf{m}_p - \mathbf{G}^{-g}(\mathbf{E}\mathbf{m}_p - \boldsymbol{\epsilon}). \tag{18}$$

This equation is analogous to eq. (13) that was obtained for a one-step linearization inversion where we also had $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$. A similar equation holds at each step k of the nonlinear inversion scheme. The term $\mathbf{G}^{-g}\mathbf{E}\mathbf{m}_p$ accounts for the data error that results from convergence of the nonlinear solution scheme to $\hat{\mathbf{m}}$ instead of $\mathbf{m}_p$, which is due to the nonlinear propagation of data and model errors and includes linearization errors. Note that in eq. (13) the original data error occurs and that by implication the error $\mathbf{E}$ in $\mathbf{G}$ absorbs all the nonlinear error propagation. As for linearized inversion, the true resolution matrix is $\mathbf{R}_p = \mathbf{G}^{-g}\mathbf{G}_p = \mathbf{G}^{-g}(\mathbf{G} - \mathbf{E})$, but in this case $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$ cannot exactly describe the spatial resolution matrix for the final result of a nonlinear inversion due to nonlinear error propagation. In practice, however, $\mathbf{R}$ will be our best guess of the resolution matrix because we can compute $\mathbf{G}$ but not $\mathbf{G}_p$. Therefore, synthetic data $\mathbf{d}_s$ for sensitivity analysis of the solution $\hat{\mathbf{m}}$ can only be generated by $\mathbf{G}\mathbf{m}_s = \mathbf{d}_s$.

The use of eq. (18) for sensitivity analysis aimed at retrieving aspects of $\mathbf{R}_p$ is therefore approximate, as was the case for a one-step linearized inverse problem where the difference matrix $\mathbf{E}$ accounted for only the linearization error (see Section 2.2). On a practical note: similar to a one-step linearized inverse problem, the matrix $\mathbf{G}$ with which the final solution $\hat{\mathbf{m}}$ is computed could still be improved by using $\hat{\mathbf{m}}$ if convergence has not been achieved. For instance, in ray-based tomography, one could perform ray tracing through $\hat{\mathbf{m}}$ and construct an improved observation matrix $\mathbf{G}$. The improved $\mathbf{G}$ can

then be used in eq. (18) which would further reduce the linearization error $\mathbf{E}$ and may improve the approximation of $\mathbf{R}$ to $\mathbf{R}_p$.

In summary, sensitivity analysis applied to the last stage of an iterative nonlinear inversion scheme for solution $\hat{\mathbf{m}}$ leads to the inference $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$, which is an approximation of the true resolution matrix $\mathbf{R}_p = \mathbf{G}^{-g}(\mathbf{G} - \mathbf{E})$. The size of the propagated nonlinear resolution error $\mathbf{G}^{-g}(\mathbf{E})$ will determine the accuracy of the approximation $\mathbf{R} \approx \mathbf{R}_p$.

In sensitivity tests we effectively assume $\mathbf{E} = \mathbf{0}$, which implies the same strengths and limitations of sensitivity analysis as in a purely linear inversion, albeit with the additional uncertainty of the second error propagation term $-\mathbf{G}^{-g}\mathbf{E}\mathbf{m}_p$. Eq. (18) does, however, provide a general means to investigate effects of nonlinear error propagation by explicitly computing this term in a designed synthetic experiment that starts with creating $\mathbf{d}_s$ and $\mathbf{G}_s$ for a synthetic model $\mathbf{m}_s$ such that $\mathbf{d}_s = \mathbf{G}_s\mathbf{m}_s$. Next, a nonlinear inversion can be started from some model $\mathbf{m}_s^1$ that converges to the equation $\hat{\mathbf{d}}_s = \mathbf{G}\hat{\mathbf{m}}_s$. This allows for the evaluation of $\mathbf{E} = \mathbf{G} - \mathbf{G}_s$. Similarly, given a particular source and receiver network, one can use eq. (18) as a basis for investigating if certain perceived Earth structure (e.g. slab geometry; gap in slab) can be resolved by the nonlinear inversion scheme in use. Note that all such experiments fundamentally depart from testing the spatial resolution (i.e. the algebraic relationship between model parameters) of the real data experiment, that is, belonging to $\hat{\mathbf{m}}$, as both $\mathbf{G}^{-g}$ and $\mathbf{G}$ will generally be different in synthetic experiments that solve the forward problem for $\mathbf{m}_s$, for example, by 3-D ray tracing through $\mathbf{m}_s$.

We note that if $\mathbf{G}^{-g}$ can be reproduced independent of $\mathbf{m}_s$ then the designed amplitudes of $\mathbf{m}_s$ do not influence sensitivity analysis because we are performing a strictly linear inversion for $\hat{\mathbf{m}}_s$ aimed at estimating algebraic properties of $\mathbf{R}$. However, if this is not the case and the entire nonlinear scheme needs to be repeated, then an approximate alternative is to superimpose $\mathbf{m}_s$ as a synthetic pattern on the data-generated solution $\hat{\mathbf{m}}$, such that the synthetic data are computed as $\mathbf{G}^1(\hat{\mathbf{m}} + \mathbf{m}_s)$ as the starting data set for nonlinear inversion (Zelt 1998). This may mimic a comparable convergence sequence $\mathbf{G}^k \to \mathbf{G}$ if the amplitude of $\mathbf{m}_s$ are a sufficiently small fraction of $\hat{\mathbf{m}}_s$ such that they do not lead to additional nonlinearity and error propagation.

2.5 Theoretical implications for sensitivity analysis

For proper application and interpretation of sensitivity analysis, the following implications that arise from the theory outlined in Sections 2.1–2.4 are now discussed.

(i) Sensitivity analysis only applies to linear systems like eqs (6) and (15) and implicitly assesses the linear dependence between model parameters as defined by $\mathbf{R}$ and explicitly assesses noise propagation into the solution via $\mathbf{G}^{-g}\boldsymbol{\epsilon}_s$. However, it does not assess the effects of all assumptions and approximations that were made to convert the nonlinear inverse problem $\mathbf{d} = \mathbf{g}(\mathbf{m}_E) + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_d$ into the linearized inverse problem $\mathbf{d} = \mathbf{G}\mathbf{m} + \boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_l + \boldsymbol{\epsilon}_p$. This is not a specific weakness of sensitivity analysis, because it also applies when $\mathbf{R}$ is explicitly computed. However, one advantage of sensitivity analysis is that it does allow the influence of the error term in eq. (13) to be analysed (see Sections 3.4, 3.5, 3.7 and 3.9 for examples).

(ii) It follows from $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$ that if the synthetic model $\mathbf{m}_s$ lies in the null space of $\mathbf{G}$, then $\mathbf{R}\mathbf{m}_s = \mathbf{0} = \hat{\mathbf{m}}_s$ and there will be no recovery in a synthetic experiment. Conversely, any model $\mathbf{m}_s$ that is entirely in the row space of $\mathbf{G}$ (i.e. maps only in the

range of **G**) can be fully recovered ($\mathbf{m}_s = \hat{\mathbf{m}}_s$). Sensitivity analysis may lead to an incomplete interpretation of **R** when applied to the two end-member models. Lévêque *et al.* (1993) demonstrated this in an idealized experiment that showed perfect recovery for a spatially detailed checkerboard model, while a coarse checkerboard was almost entirely in the null space of **G**, despite identical path coverage. By implication, *sensitivity tests can only be used to expose a lack of resolution*. Resolution artefacts should be inferred from sensitivity tests using a wide variety of synthetic models in an attempt to explore all null space components (e.g. Bijwaard *et al.* 1998). Spatial scale dependence of sensitivity tests is illustrated with experiments in Section 3.5.

(iii) A synthetic model that consists of only one value for model parameter *i* (a spike) and zeros elsewhere leads to the determination of the *i*th row and column of the resolution matrix (Spakman & Nolet 1988). Spakman & Nolet (1988) specifically devised the 'spike model' comprising a regular grid of well-separated spikes with zeros in between as a more economic way of obtaining a mapping of many columns of **R** in one inversion. Still, the interpretation pitfall exposed by Lévêque *et al.* (1993) may theoretically occur such that a particular choice of synthetic model may not sufficiently detect lack of resolution. Therefore, as noted before, sensitivity analysis should be conducted with a variety of different synthetic models covering all scale-lengths to optimize the *detection of lack of resolution* (see Section 3.5).

(iv) Note that eq. (16) requires that sensitivity tests should use in eq. (15) the matrix **G** of eq. (6) for computation of synthetic data and, in addition, the $\mathbf{G}^{-g}$ used in $\hat{\mathbf{m}} = \mathbf{G}^{-g}\mathbf{d}$ for inversion. If $\mathbf{G}^{-g}$ is not explicitly available then the synthetic inversion must mimic the real data inversion to the extent that the use of the same $\mathbf{G}^{-g}$ is implied (this would apply equally to finite frequency tomography for example, where one should use the finite frequency kernels built for the real data paths in any sensitivity test involving a synthetic model). If we instead construct a matrix **G** in eq. (15) from 3-D ray tracing (for example) through the synthetic model $\mathbf{m}_s$, then this leads to a different **G** than in eq. (6), so the resolution matrix $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$ in eq. (16) is not the same as in eq. (10). This is illustrated in Section 3.3. Also, the synthetic model $\mathbf{m}_s$ should be defined on the same model parametrization as used for the construction of **G** otherwise model projection noise $\boldsymbol{\epsilon}_p$ (see eq. 6) will enter the analysis though the second term in eq. (10). This is illustrated in Section 3.9.

(v) Eq. (16) implies that noise-free synthetic data should be used if the sensitivity analysis focuses only on assessing the properties of **R**. As mentioned in the previous section, inversion of only synthetic noise assesses the importance of the second term in eq. (6), while the inversion of noisy synthetic data explores the combination of the two terms. Synthetic noise can be generated that has a distribution consistent with that assumed for the real data by means of the data covariance (e.g. Gaussian). The signal-to-noise ratio can be tuned such that the same data misfit is obtained as in the real-data inversion (Spakman & Nolet 1988). This, however, is still approximate, as we do not know the real data errors. Section 3.8 explores the effect of data noise in the recovery of structure.

(vi) The second term in eq. (16), $\mathbf{G}^{-g}\boldsymbol{\epsilon}_s$, opens the possibility to investigate how a wide variety of random or systematic data noise can propagate into the solution. For instance, tests can be devised for the influence on a tomographic model of the projection error due to model parametrization (as is done in Section 3.9). Propagation of random noise can be studied to detect noise-sensitive regions of the model. A special case is the 'permuted data test' (Spakman &

Nolet 1988; Spakman 1991), in which the data vector in eq. (6) is randomly permuted prior to inversion. The application of this technique to global traveltime tomography led to a data-derived estimate of average model amplitude error (Bijwaard *et al.* 1998).

(vii) The linear nature of eq. (16) implies that the inversion response to any synthetic model can be obtained from a superposition of single-spike responses (ignoring the effect of $\boldsymbol{\epsilon}_s$), that is,

$$\hat{\mathbf{m}}_s = \mathbf{R}\mathbf{m}_s = \mathbf{R}\sum_n a_n \mathbf{m}_n^{\text{spike}} = \sum_n a_n \hat{\mathbf{m}}_n^{\text{spike}} \quad (19)$$

where $a_i, i = 1, \ldots, n$ is a scaling coefficient. In particular, checkerboard models can be seen as a superposition of synthetic models with a regular distribution of spikes. This suggests that inferring resolution artefacts from checkerboard models can be more complex than for each of the separate spike models. This will be investigated in Sections 3.1 and 3.2.

(viii) Care should be taken when conducting sensitivity experiments with the observational model, that is, the model obtained from inverting the real data. When using it as a synthetic model, one can expect an almost identical output. The part of the model that is built from the row space of **G** (i.e. data constrained), will be completely recovered. Merged with this is a null space component that is shaped and constrained by the applied regularization. This leads to linear dependencies between model parameters that are fully described by **R** and which, hence, describes their mapping in the observational model. From a theoretical point of view the observational model is defined by $\hat{\mathbf{m}} = \mathbf{R}\mathbf{m}_p$; if we take $\hat{\mathbf{m}}$ as our synthetic model, then $\hat{\mathbf{m}}_s = \mathbf{R}\mathbf{R}\mathbf{m}_p$. If **R** is an idempotent matrix ($\mathbf{R}\mathbf{R} = \mathbf{R}$), then $\hat{\mathbf{m}}_s = \hat{\mathbf{m}}$ and the input and output models should be identical. While this may not be true in general, our tests seem to indicate that for our chosen objective function, the corresponding resolution matrix is likely to be at least quasi-idempotent ($\mathbf{R}\mathbf{R} \approx \mathbf{R}$), as demonstrated in Section 3.6.

What can be done successfully is to construct synthetic models by removing a certain feature from the observational model, for example, a slab in a mantle model, and test if it fully reappears in $\hat{\mathbf{m}}_s$, in which case it is a resolution artifact. In the end-member case of setting the amplitude of only one model parameter to zero in the observational model, that is, applying an 'anti-spike', a correct estimate is obtained in the recovered model of how the ambient model contributes to amplitude and sign imaged for this parameter. This exposes the summed contribution of all off-diagonal elements in the pertinent row of **R** in determining the value (see Supplementary Information for an example). These are very powerful ways of testing specific attributes of the observational model.

3 NUMERICAL EXPERIMENTS

In order to conduct our numerical experiments to provide additional validation and insight with respect to the theory of Section 2, we use the 2-D spherical shell tomography code described in Rawlinson *et al.* (2008). It is designed to solve weakly nonlinear traveltime tomography problems using an iterative approach that involves repeated application of a forward and linearized inverse solver. In this case, the Fast Marching Method or FMM (Sethian 1996; Rawlinson & Sambridge 2003) is used to solve the forward problem of computing traveltimes between sources and receivers through a heterogeneous velocity medium, and a subspace inversion method (Kennett *et al.* 1988) is used to invert source–receiver traveltimes for velocity structure. Velocity is represented on a regular grid in latitude and longitude, with cubic B-splines used to achieve a smooth

continuous medium. This code is best suited to the inversion of surface wave traveltimes for group or phase velocity maps at a specified period, and has been most frequently used for ambient noise tomography (e.g. Saygin & Kennett 2010; Stankiewicz *et al.* 2010; Young *et al.* 2011). Although the code is 2-D and assumes that the high frequency assumption is valid, the implications of our results should largely hold for any class of linear or iterative nonlinear tomography.

The use of a subspace scheme to search model space for a data-fitting model means that we do not directly solve eq. (8). Rather, we attempt to iteratively minimize a cost function of the form given by eq. (17) under the assumption that we solve a weakly nonlinear inverse problem. In this case, the specific cost function is defined by:

$$S(\mathbf{m}) = \left(\mathbf{d} - \mathbf{g}(\mathbf{m})^{\mathrm{T}} \mathbf{C}_{\mathrm{d}}^{-1}(\mathbf{d} - \mathbf{g}(\mathbf{m}))\right) + \epsilon(\mathbf{m} - \mathbf{m}_0)^{\mathrm{T}} \mathbf{C}_{\mathrm{m}}^{-1}(\mathbf{m} - \mathbf{m}_0)$$
$$+ \eta \mathbf{m}^{\mathrm{T}} \mathbf{D}^{\mathrm{T}} \mathbf{D} \mathbf{m}, \quad (20)$$

where $\mathbf{m}_0$ is the reference model, $\mathbf{C}_{\mathrm{m}}$ the *a priori* model covariance matrix and $\mathbf{D}$ the second derivative smoothing operator. The constant terms ϵ (≥ 0) and η (≥ 0) are the damping and smoothing parameters respectively, which govern the trade-off between data fit, model smoothness, and model perturbation relative to the starting model. The model perturbation given by a single iteration of the subspace method is defined by (see Rawlinson *et al.* 2006, for more details):

$$\delta \mathbf{m} = -\mathbf{A} \left[\mathbf{A}^{\mathrm{T}} \left(\mathbf{G}^{\mathrm{T}} \mathbf{C}_{\mathrm{d}}^{-1} \mathbf{G} + \epsilon \mathbf{C}_{\mathrm{m}}^{-1} + \eta \mathbf{D}^{\mathrm{T}} \mathbf{D} \right) \mathbf{A} \right]^{-1} \mathbf{A}^{\mathrm{T}} \hat{\boldsymbol{\gamma}} \quad (21)$$

where $\mathbf{A} = [\mathbf{a}^j]$ is the $M \times n$ projection matrix (where M is the number of unknowns and n is the subspace dimension) and $\hat{\boldsymbol{\gamma}}$ is the gradient vector ($\hat{\boldsymbol{\gamma}} = \partial S / \partial \mathbf{m}$). The basis vectors that define the projection matrix are based on the gradient vector in model space $\boldsymbol{\gamma} = \mathbf{C}_{\mathrm{m}} \hat{\boldsymbol{\gamma}}$ and the model space Hessian $\mathbf{H} = \mathbf{C}_{\mathrm{m}} \hat{\mathbf{H}}$, where $\hat{\mathbf{H}} = \partial^2 S / \partial \mathbf{m}^2$. In this application, we use a 20-D subspace, but this is dynamically reduced by the application of Singular Value Decomposition (SVD) if linear dependence between the different $\mathbf{a}^j$ becomes an issue. This approach is still valid for sensitivity analysis using synthetic structures.

Below, we run a series of experiments using a purely synthetic data set; a similar set of experiments applied to ambient noise data recorded in Australia can be found in the Supplementary Information. The synthetic data set used here is based on a test model, source–receiver geometry and ray-path configuration shown in Fig. 1. The pattern of velocity anomalies is randomly generated and exhibits a Gaussian distribution between peak values of ±20 per cent. Although this may seem large, there are many examples of ambient noise surface wave tomography that recover peak perturbations in excess of 25 per cent relative to a background average (e.g. Saygin & Kennett 2010, and Supplementary Information). The design of the source–receiver array is such that there is good path coverage in the NW region of the model, but it gradually degrades to the SE, where it becomes more unidirectional. This forms a good basis for assessing the characteristics of various synthetic recovery tests.

Fig. 1(c) shows a reconstructed model based on inverting the true noise-free traveltimes associated with the paths shown in Fig. 1(b). 20 iterations of the forward and inverse step were applied to obtained the solution, which corresponds to an RMS reduction in traveltime residuals from 18.4 to 0.35 s, at which point convergence is well and truly achieved (subsequent iterations had no appreciable effect on the traveltime misfit). Minimal damping and smoothing was applied in this case as the implicit regularization imposed by the choice of basis function (cubic B-splines) was sufficient to stabilize

the inversion. A constant velocity model, set to the average velocity of the synthetic model, was used as the starting or initial model. Clearly, the reconstruction is best in the NW of Fig. 1(c), but gradually degrades to the SE, where significant streaking effects can be observed, which can be attributed to the source–receiver geometry.

Below, we use the experimental set-up of Fig. 1 to examine the characteristics of a variety of synthetic reconstruction tests. In each test, the reconstruction is carried out using the path coverage shown in Fig. 1(d) (i.e. the path coverage associated with the output model), unless stated otherwise. Section 3.3 investigates what happens if a nonlinear method, in which rays are re-traced after every iteration, is used instead.

3.1 Checkerboard test

Fig. 2(a) shows a checkerboard model comprising a fine-scale pattern of positive and negative anomalies, with peak values equal to those of the 'observational' model (Fig. 1c). The sizes of the anomalies are approximately equal to the minimum scale-length of the structures present in Fig. 1(a). If we use this checkerboard model as input, and compute a synthetic source–receiver data set by integrating along the path geometries of Fig. 1(d), then we can solve the linear inverse problem to obtain the result shown in Fig. 2(b); perhaps unsurprisingly, the checkerboard pattern is retrieved within the triangular region defined by the source geometry in the NW of the model but not elsewhere. A conventional interpretation of this figure might be that the NW sector of the model is quite well resolved, while the remainder of the model is poorly resolved, with a widespread tendency to smear structure in the NW–SE direction. Is this seemingly straightforward conclusion really borne out by the resolving power of the ray coverage?

3.2 Spike test

Fig. 3(a) shows a spike model (or sparse checkerboard model) which has the same scale length and amplitude of anomalies as Fig. 2(a), but is much more widely spaced. Fig. 3(b) shows the result of applying exactly the same inversion method that was used to derive Fig. 2(b). Like before, the anomalies are most accurately reconstructed in the NW of the model, and generally become poorer to the southeast. However, when comparing Figs 2(b) and 3(b), there are a number of important differences. First, there is clear evidence of smearing towards the edges of the model for the most accurately recovered anomalies in the NW sector of Fig. 3(b); in Fig. 2(b), the distortion in the shape of the anomalies is much less apparent. Presumably, in the checkerboard model, this is due to the close proximity of adjacent anomalies, which tends to mask smearing effects that are not very major. Furthermore, in the southeast region of the spike model (Fig. 3b), there is not the same dominance of NW-SE smearing, and it is now much clearer from the spike model how the anomalies are distorted as a result of the path coverage. In fact, we can now see that most anomalies experience some degree of recovery, even if their aspect ratios have changed from 1:1 in the input model to as much as 5-6:1 in the recovered model. This is useful information, not only on the level of distortion, but also the direction, and is entirely absent from the checkerboard result of Fig. 2(b).

A comparison of the recovered anomalies in Figs 2(b) and 3(b) with the associated path coverage is shown in Figs 4(a) and (b) respectively. The close proximity and regular pattern of the anomalies in Fig. 2(a) mean that when path coverage is diagonally dominant,
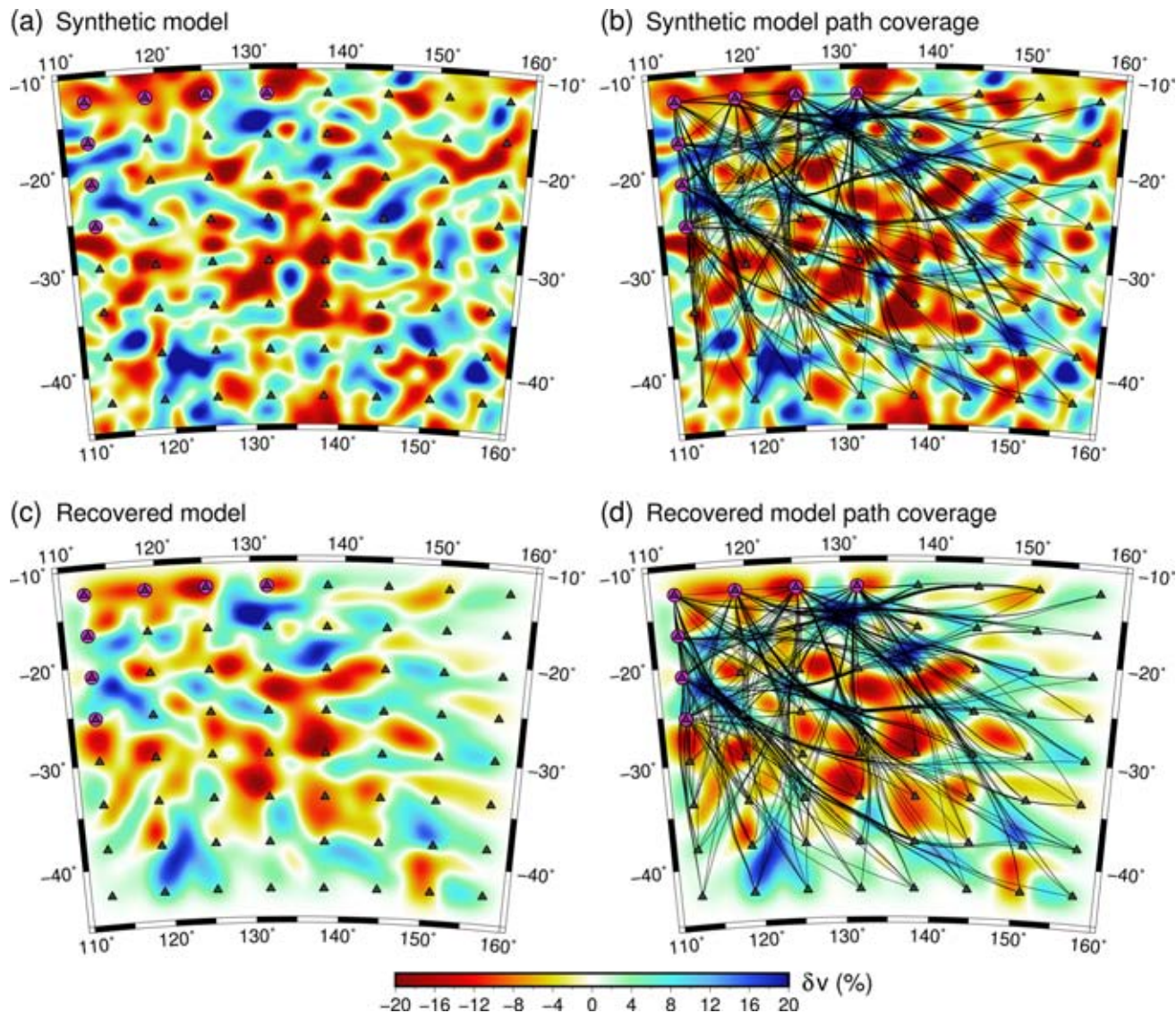
Figure 1. (a) Synthetic test model and source–receiver geometry used for numerical experiments; (b) ray-path geometry through the synthetic model; (c) reconstruction of (a) obtained after application of the iterative nonlinear inversion scheme; (d) ray-path geometry through the recovered model. Dark grey triangles denote receivers and magenta circles denote sources.

the recovered anomalies do a reasonable job at fitting the traveltimes by smearing out in the diagonal direction, which is the direction in which we do not have an oscillation from positive to negative anomalies in the input model. By contrast, the spike test is able to better reveal the (lack of) resolving power of the ray illumination, and the recovery is hence more useful for interpreting directionally dependent solution smearing. For example, we see that the negative anomaly at approximately (151°E, 21°S), which sits between receivers immediately to the west and to the east in Fig. 3(a), is recovered as a horizontally smeared anomaly in Fig. 3(b). The path coverage map in Fig. 4(b) shows that this anomaly is crossed by a single horizontal ray, but just to the west, there is crossing path coverage, so the anomaly is represented by its path average approximation between the two stations. Another example is the negative anomaly at approximately (119°E, 34°S), which achieves a degree of recovery in Fig. 3(b), but there is no evidence of recovery in the same region in Fig. 2(b). Although the experiment set-up here is seems clear that the discrete spike test provides more useful information on spatial resolution from the constrain-

ing power of the path coverage than the equivalent checkerboard test.

In Section 2.3, the theory behind spike tests and their association with the resolution matrix was discussed. In effect, the input model of Fig. 3(a) represents $\mathbf{m}_s$ in eq. (16), while the output model of Fig. 3(b) represents $\hat{\mathbf{m}}_s$, with the noise term set to zero, that is, $\boldsymbol{\epsilon}_s = \mathbf{0}$. If the resolution matrix $\mathbf{R}$ was available, then any input model could be transformed into an output model. For this example, there are 3844 unknowns, which means that it is computationally feasible to calculate $\mathbf{R}$. In general, this may not be the case, particularly for large 3-D problems. We solve $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$ using Cholesky decomposition, which produces the complete 3844×3844 resolution matrix. The compute time is approximately 12 min on a workstation equipped with an Intel Xeon 3.1 GHz E5-2687W processor, 128 Gb of RAM and running Ubuntu 14.04. Fig. 5 compares the spike test result of Fig. 3(b) with $\hat{\mathbf{m}}_s^* = \mathbf{R}\mathbf{m}_s^*$, where $\mathbf{m}_s^*$ is a selection of two anomalies from Fig. 3(a). These two anomalies were chosen to be far apart so that any interference between their reconstruction is unlikely. It is clear that Fig. 5(b) is essentially a subset
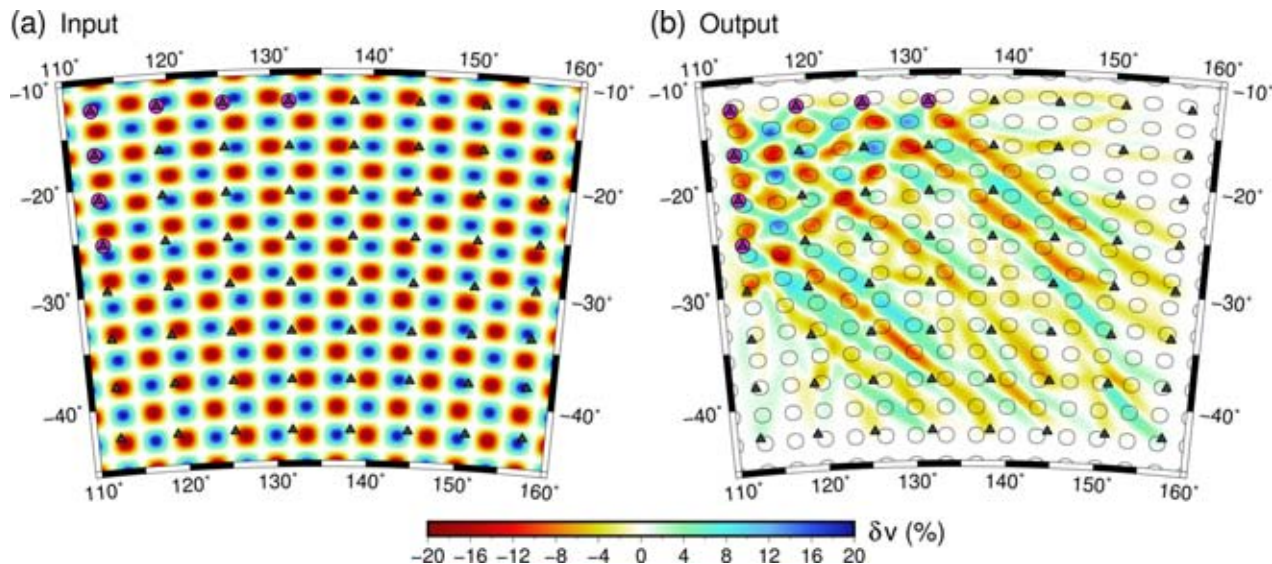
Figure 2. (a) Input checkerboard model; (b) recovered model using the path geometry of Fig. 1(d). Black contour lines represent the ±10 per cent contour interval of the input checkerboard.
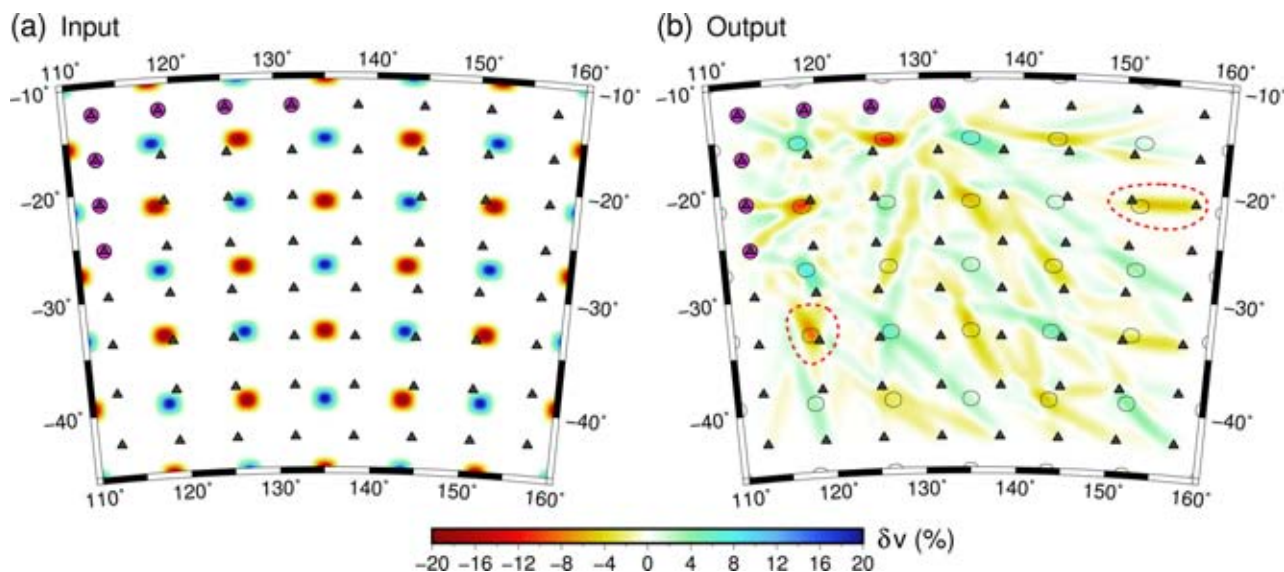

Figure 3. (a) Input spike model; (b) recovered model using the path geometry of Fig. 1(d). Black contour lines represent the ±10 per cent contour interval of the input spikes. Closed dashed red lines highlight the locations of features discussed in the text.

of Fig. 5(a); the very minor differences in recovery can be attributed to the fact that there is a small amount of smearing overlap between adjacent spikes. However, it is insignificant enough to ignore for interpretation purposes. The compute time needed to perform the Fig. 5(a) spike test on the same computer was approximately 8 s, which is two orders of magnitude faster than computing **R**. This example demonstrates the power of well-chosen sensitivity tests: practical information on model resolution without the need to compute **R**. The checkerboard result of Fig. 2(b) takes the same compute time as the spike test, but the overprinting of smearing effects from adjacent spikes makes it less useful measure of model resolution.

3.3 Nonlinearity and the proper application of sensitivity analysis

In the above tests, the ray-path geometry (Fig. 1d) is inherited from the observational model of Fig. 1(c); this ensures that the synthetic

test provides information on spatial resolution that is consistent with the path coverage through the solution model. This is the same philosophy that is used to compute posterior model covariance and resolution for weakly nonlinear inverse problems (e.g. Tarantola 1987; Rawlinson *et al.* 2010).

If instead the synthetic data set is computed by solving the full forward problem (ray tracing in this case) through the input model used for the synthetic test, and the input model is then recovered using an iterative nonlinear method which updates the ray paths, the result of the reconstruction can be quite different. Fig. 6 compares the true path coverage through the checkerboard and spike models of Figs 2 and 3 with the path coverage inherited from Fig. 1(c). The differences are significant; in the case of the checkerboard model (Fig. 6c) the paths are attracted to high velocity regions since only first arrivals are used. This is also true of the spike model (Fig. 6d), although the path coverage is more even due to the increased sparsity of model anomalies. The corresponding inversion results when
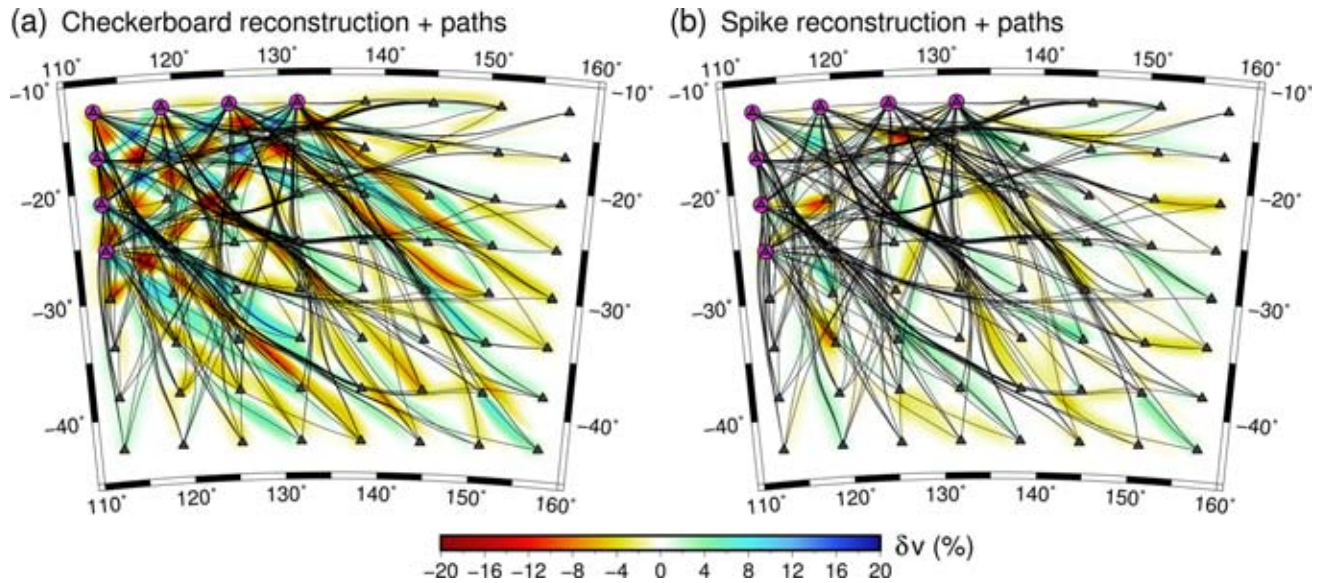
Figure 4. Comparison between path coverage and recovered structure for (a) checkerboard recovery of Fig. 2(b); (b) spike recovery of Fig. 3(b).
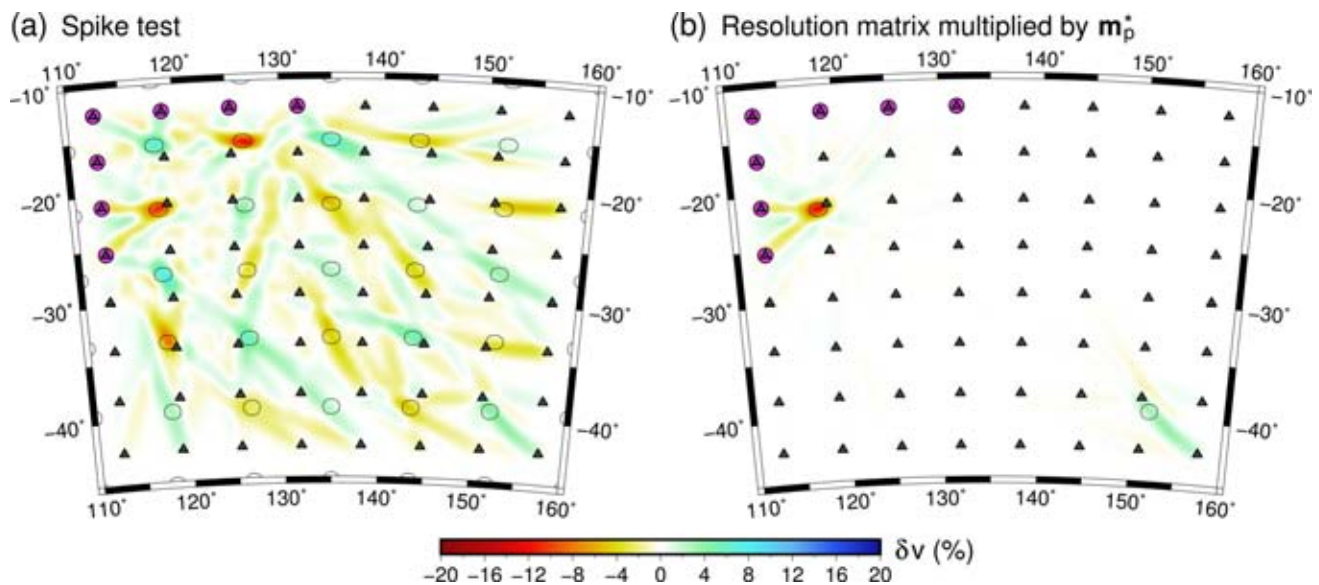

Figure 5. Comparison between sensitivity test and resolution matrix. (a) Spike test output from Fig. 3(b); (b) resolution matrix multiplied by the true model $\mathbf{m}_s^*$, where $\mathbf{m}_s^*$ consists of two spikes taken from Fig. 3(a). In both (a) and (b), black contour lines represent the ± 10 per cent contour interval of the input spikes.

using these different sets of ray paths is clearly revealed in Fig. 7. In the case of Figs 7(c) and (d), the synthetic data set is created by integrating along the ray paths shown in Figs 6(c) and (d) respectively. The output models are obtained by running the iterative nonlinear procedure until convergence is achieved. The checkerboard recovery in this case has far greater diagonal smearing compared to the linear inversion result of Fig. 7(a), which uses the rays from Fig. 6(a). This is partly due to the rays favouring the high velocity regions, which have an NW–SE orientation. There is also some hint of this in Fig. 7(d), although it is less pronounced. The clear difference in results produced by this example means that it is important to use rays that are inherited from the original observational model as required by the theory (see point iv of Section 2.5), rather than rays traced through the synthetic test model, when the nonlinearity of the inverse problem is taken into account.

3.4 Nonlinear error propagation

In the previous section, we established the correct procedure for sensitivity analysis in the presence of weakly nonlinear inverse problems. However, it does not account for the effects of nonlinear error propagation, which remains an unsolved problem in synthetic testing. However, as was noted in Section 2.4, we can investigate nonlinear error propagation for a purely synthetic experiment by evaluating $\mathbf{E} = \mathbf{G} - \mathbf{G}_s$, where $\mathbf{G}_s$ is computed from the true (or input) model and $\mathbf{G}$ is computed from the reconstructed (or output) model. This allows the error propagation term $\mathbf{G}^{-g}\mathbf{E}\mathbf{m}_p$ in eq. (18) to be computed, thus providing quantitative insight into the influence of nonlinearity. Fig. 8 shows the result of this analysis in the presence of our default test model (shown in Figs 1a and 8a), and a model which exhibits an identical pattern of anomalies but with
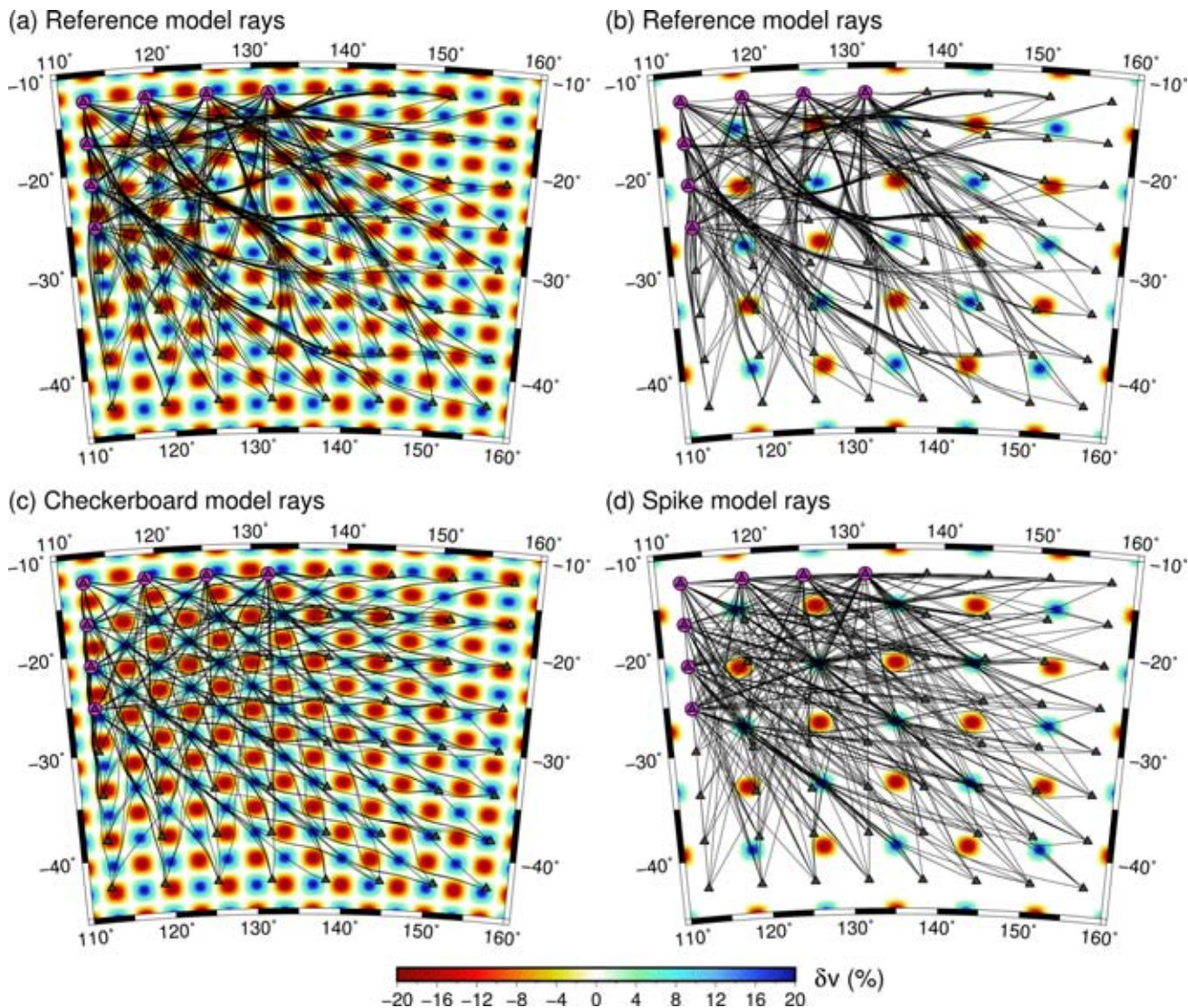
Figure 6. Path coverage through checkerboard (left column) and spike (right columns) models from Figs 2 and 3, respectively. (a), (b) show the path coverage inherited from the recovered model shown in Fig. 1(d), while (c), (d) show the path coverage obtained by solving the forward problem through the checkerboard and spike model respectively.

reduced amplitude (Fig. 8d). In the case of the original test model, the magnitude of the error is smallest in the NW portion of the model where there is both good path coverage and relatively short paths in comparison to the dominant wavelength of the velocity anomalies. The errors tend to peak in the central and SE region of the model, where path density is still moderate, but paths are on average longer and angular coverage is poorer. Overall, the magnitude of the error is smaller than the amplitudes of the recovered anomalies (*cf.* Figs 8b and c), particularly in the well resolved region in the NW.

If we now decrease the amplitudes of the input anomalies by 50 per cent (Fig. 8d), the effect of the nonlinear error propagation is much reduced (Fig. 8f), which is consistent with solving a more linear inverse problem. Our results indicate that even in the presence of sizable anomalies (up to 20 per cent) the weakly nonlinear assumption can be valid (*cf.* Figs 8a and b), and in regions of good angular path coverage, the nonlinear propagation error can be small relative to the amplitude of the recovered anomalies.

3.5 Using various test models with difference scale lengths

One problem with using a single checkerboard or spike test with only one scale length of structure is that it does not necessarily reveal resolution artefacts over multiple scales. Moreover, as demonstrated by Lévêque *et al.* (1993) (see Section 2.5, point ii), there are certain specific circumstances in which the common expectation that the region of good recovery increases in size as the scale length of structure increases is not always a given. Therefore, it may ultimately be misleading to present the result of a single spike or checkerboard test to assess lack of resolution. We illustrate scale dependence in Fig. 9, which shows the results of performing the same inversions as Figs 2 and 3, but now with much broader anomalies. In this case, both the checkerboard and spike anomalies have much improved recovery (in shape and amplitude) compared to the previous test. This is consistent with the Fig. 1 test, in which the broader scale anomalies are more accurately retrieved than the smaller scale anomalies. These results illustrate that, at least in this case, larger structures
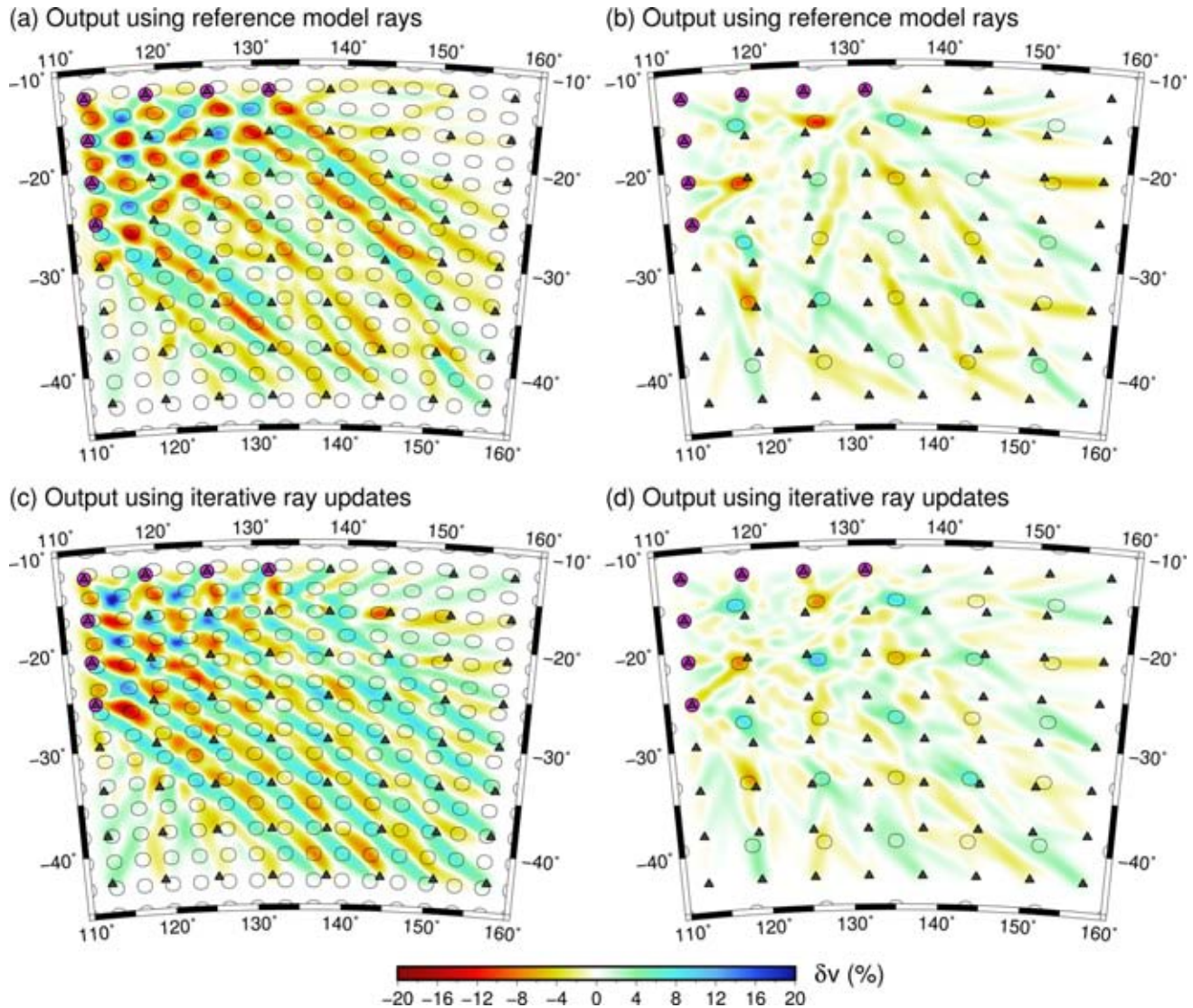
Figure 7. Checkerboard (left column) and spike (right column) output models based on different path coverage. (a), (b) are the result of a linear inversion using rays from the Fig. 1(d) model. (c), (d) are the result of an iterative nonlinear inversion in which rays are updated for the model obtained after each iteration.

are more easily resolved than smaller structures with the same path coverage.

Fig. 9 also helps to reinforce the point that spike tests are more useful than checkerboard tests in providing insight into the ability of a rayset to resolve structure and reveal directional dependence in resolution artefacts. Even though, as a result of the presence of longer wavelength anomalies, the recovery of equivalent anomalies between Figs 9(c) and 8(d) show greater similarity than between Figs 2(b) and 3(b), the full extent of smearing is more successfully revealed in Fig. 9. For example, the positive anomaly at approximately (116°E, 33°S) shows much less distortion in the N–S direction in Fig. 9(c) than in Fig. 9(d), and the positive anomaly at approximately (143°E,14°S) shows much less E–W distortion in the output checkerboard model. Both of these effects can be attributed to the close proximity of anomalies of opposite sign.

3.6 Synthetic test based on observational model

As well as synthetic tests involving anomalies of a fixed scale-length, it is relatively common to find studies that use a test struc-

ture that bears some resemblance to a geological feature such as a subducting slab (e.g. Spakman *et al.* 1989; Eberhart-Phillips & Reyners 1997; Bijwaard *et al.* 1998; Graeber & Asch 1999; Wolfe *et al.* 2002). While this kind of hypothesis testing is reasonable, it is important that the synthetic model does not too closely resemble what has been recovered from the observational data set, as was pointed out in Section 2.5, point vii. Fig. 10 compares the result of the recovery test shown in Fig. 1 with a new test that uses the model shown in Fig. 1(c) as the input test model (i.e. the output model of Fig. 1a). It is clear that the model recovered in this case is almost identical to the input model (see Fig. 10f), confirming what is expected from theoretical considerations (Section 2.5, point vii). This result stands in stark contrast to Fig. 10(e), which shows the difference between the original synthetic model and its reconstruction. Although this is a somewhat extreme example, it does demonstrate the point that care needs to be taken in the choice of synthetic test model used.

Tests based on the observational model should concentrate on detecting lack of resolution e.g. by setting parts of the model to zero perturbation and inferring from the sensitivity tests what is being
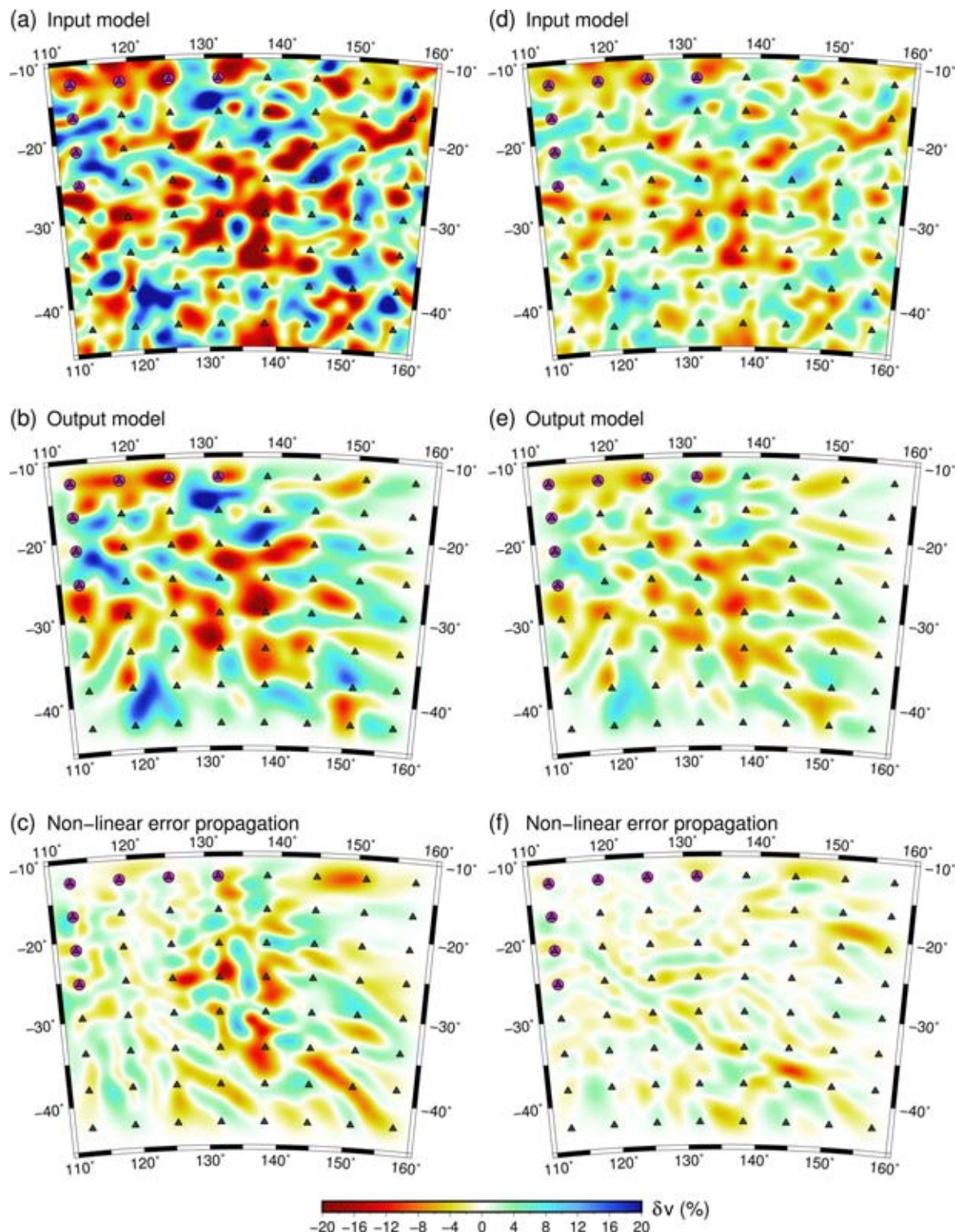
Figure 8. Test results illustrating the influence of nonlinear error propagation in the solution. Left column uses the same input model as in Fig. 1, while the input model for the right column has an identical pattern of anomalies but with a 50 per cent reduction in amplitude. (a), (d) Input model; (b), (e) output model from iterative nonlinear inversion; (c), (f) estimate of nonlinear error propagation as given by the term $\mathbf{G}^{-g}\mathbf{Em}_p$ in eq. (18).

mapped in these regions due to a lack of resolution. For example, will a short slab produce a long slab (Spakman *et al.* 1989); will a layer-cake slab produce a continuous slab (Bijwaard *et al.* 1998); will a layer-cake plume produce a continuous plume (Bijwaard & Spakman 1999)?

3.7 Testing the influence of explicit regularization

Explicit regularization in the form of damping and/or smoothing will invariably influence the recovery of structure, as shown in Fig. 11. Clearly, in performing a synthetic test, one should use the
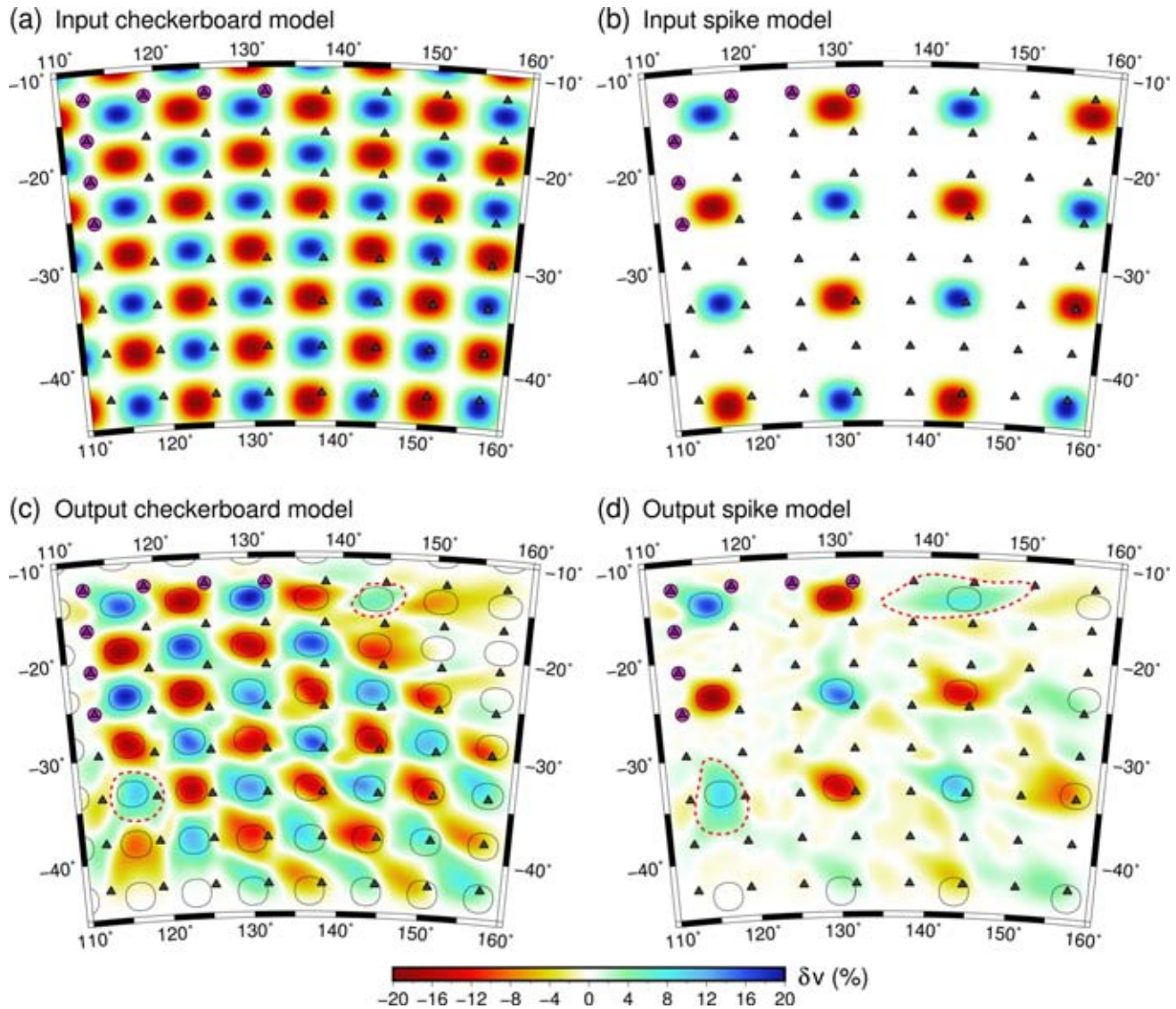
Figure 9. Synthetic recovery test involving larger anomalies. (a) Input checkerboard model; (b) input spike model; (c) output checkerboard model; (d) output spike model. Compared to the equivalent tests displayed in Figs 2 and 3 respectively, the anomalies are approximately four times the size. Black contour lines represent the ±10 per cent contour interval of the input anomalies. Closed red dashed lines denote the location of anomalies discussed in the text.

identical regularization that is chosen for the observational data. The sensitivity of the solution to model regularization can also be tested using sensitivity analysis with the aim of converging to optimal settings of regularization and model parametrization. This would involve tests similar to those used for assessing resolution and noise propagation for the final observational model: (i) noise-free tests covering a range of length scales combined with (ii) various noise tests such as the permuted data test (Section 2.5, point vi). In addition, these tests can be conducted for various levels of detail in the model parametrization. If one converts the outcome of such tests into a scalar measure of local resolution this can be used as input for optimizing strategies for the design of spatially variable model parametrization (e.g. Spakman & Bijwaard 2001) that are adapted to the expected resolution.

3.8 The influence of data noise

Data noise is ubiquitous to all seismic data sets, but in most cases the magnitude and distribution of this noise is poorly understood,

as noted previously. Synthetic recovery tests are either noise free to assess optimal spatial resolution or include noise with a Gaussian distribution and standard deviation equal to that of noise estimates obtained from the data such as picking error (Rawlinson *et al.* 2014). Spakman & Nolet (1988) advocate the addition of synthetic noise such that a data fit is obtained similar to that of the real data inversion. Although both approaches appear to be reasonable, estimating data uncertainty is notoriously subjective, and there is often little evidence that the actual noise distribution is Gaussian. To partially mitigate this, one can apply permuted data tests, which randomize the data vector prior to inversion in order to investigate the propagation of more realistic noise (Spakman & Nolet 1988). Another potential issue is that the use of an L2 norm in most inversion schemes is not robust in the presence of noise with a non-Gaussian distribution (Parker 1994), which may also cloud the results.

Sensitivity tests can be used to study the effects of various noise levels on the solution. The Fig. 12 example is the same as the Fig. 3 example but now includes varying levels of Gaussian noise. As
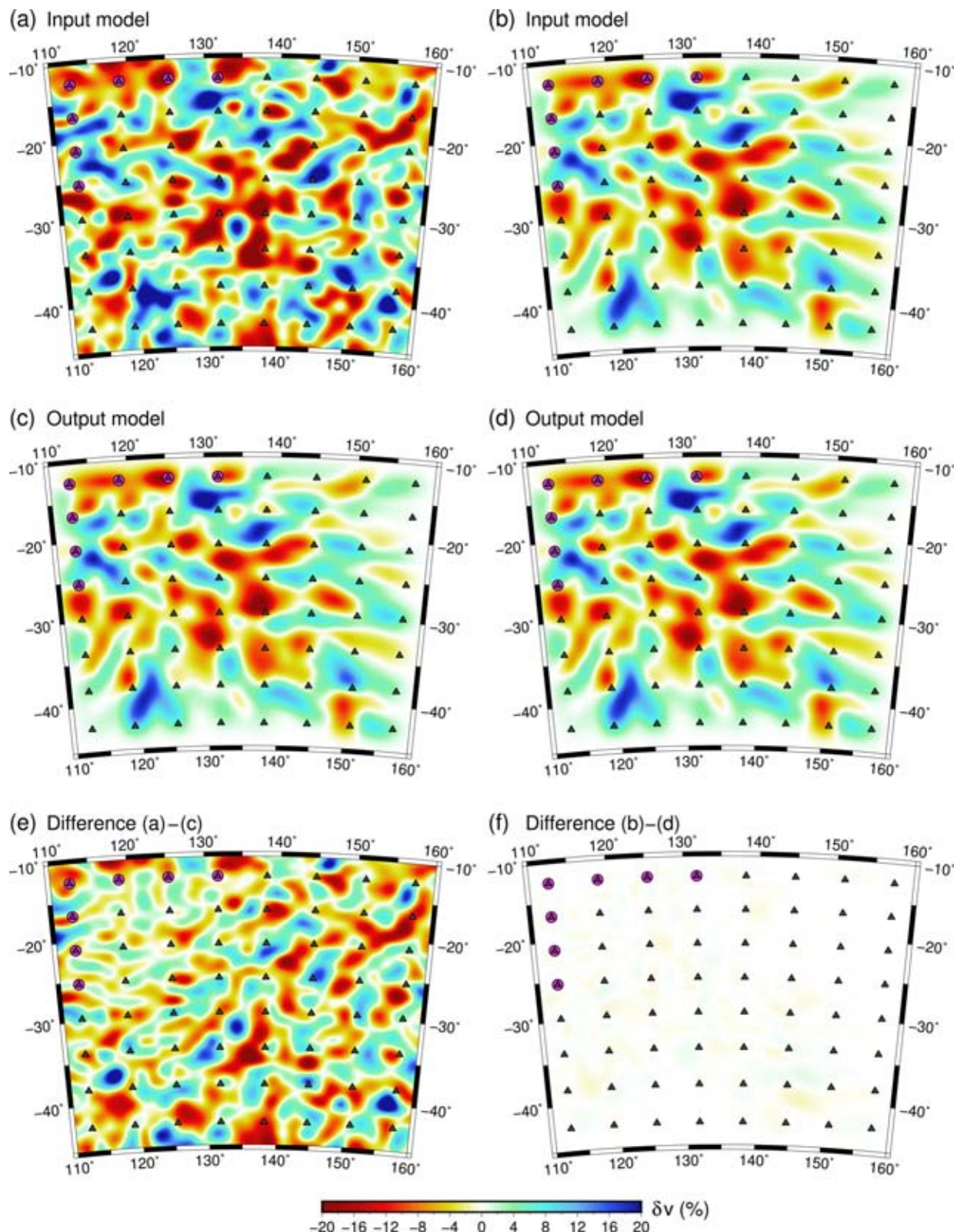
Figure 10. Left column: synthetic recovery test from Fig. 1; right column: synthetic test which uses the recovered model (c) of the left column test as the input model for a new test. (e) and (f) show the difference between the input and output models for both tests.

expected, increasing the noise level degrades the quality of the reconstruction. However, as Fig. 12(c) shows, if explicit regularization is not applied, then the inversion will attempt to overfit the data and spurious structure is introduced. If instead the simplest model (obtained using smoothing and damping) is found that fits the data, then the more robust elements of the model are recovered. One of the challenges in using the standard deviation of the noise as a measure of fit (as in the χ^2 test) is that this value is often poorly constrained.
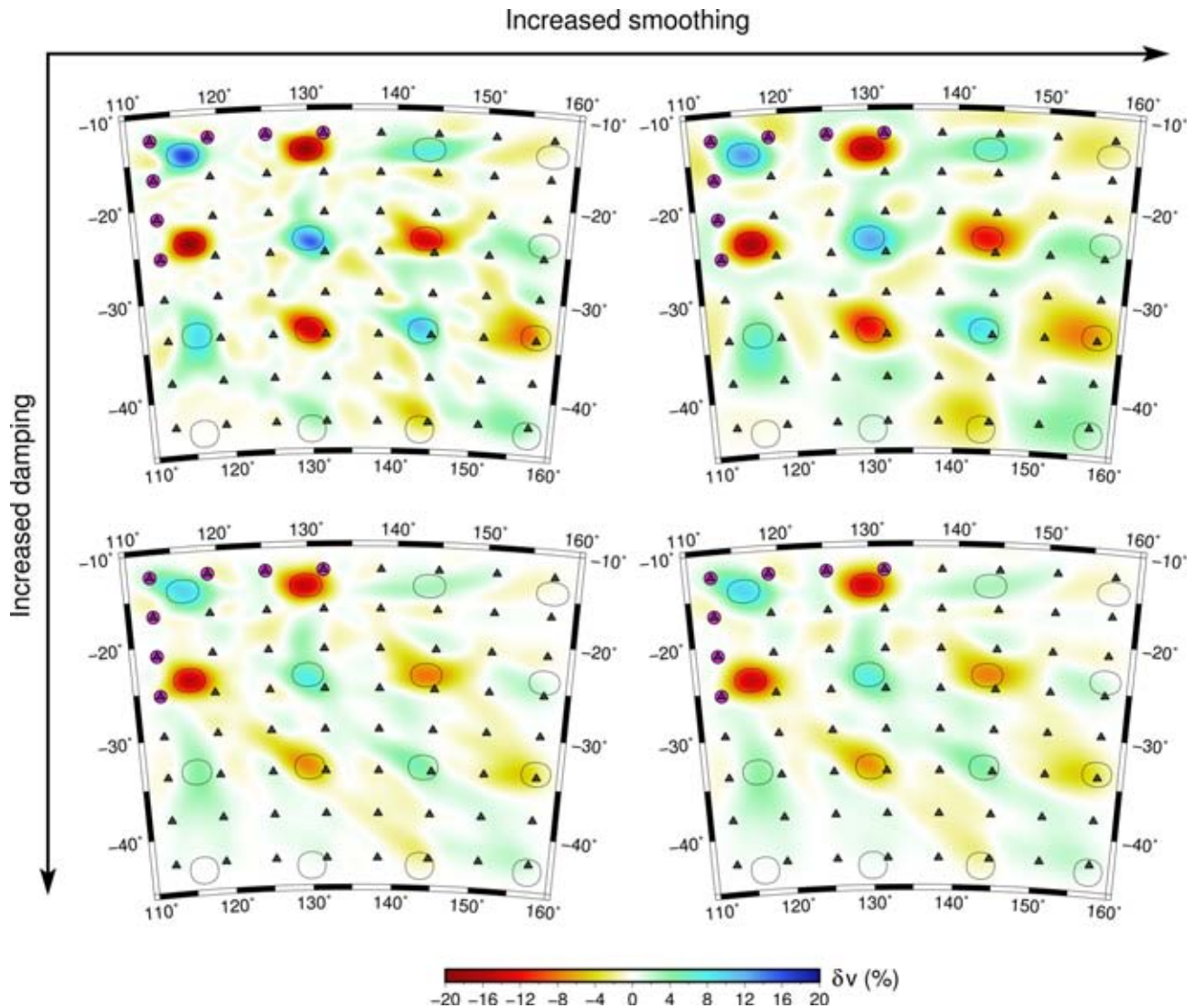
Figure 11. Synthetic spike test (based on the Fig. 9 example) showing the effects of varying damping and smoothing.

A good example of this is ambient noise surface wave tomography, where little attention is often given to data noise for the simple fact that it is hard to measure. The use of hierarchical Bayesian trans-dimensional tomography schemes, which include the level of noise as an unknown, help to address this issue (Bodin *et al.* 2012).

3.9 Preconditioning and model projection errors

Preconditioning refers to setting up an inversion that favours the reconstruction of certain features. In this sense, the use of explicit regularization is a form of preconditioning, because it provides control over the wavelength and amplitude of structure that is recovered, particularly in regions poorly constrained by the data The effects of explicit regularization on the spatial resolution can be assessed with sensitivity analysis. However, it is implicit regularization in the form of the choice of basis function used to represent structure that often plays a less defined role in recovery tests. As required by the theory, the same parametrization used to represent structure for the real data inversion should be used in the synthetic data inversion. Although this seems to be at variance with the inversion

of observations, where we know that our choice of parametrization is unlikely to be able to faithfully represent the range of structures present in the real Earth, leading to a projection error ϵ_p (see eq. 6), this does not stand in the way of proper sensitivity analysis for resolution artefacts, as this should be done with noise-free data. The effects of implicit projection errors can be tested and is illustrated in Fig. 13, where we see that the two examples of 'mismatching' parametrizations (Figs 13c and d) give a poorer result to the exact match (Figs 13a and b). In this case, the size of the anomaly is the same for the two tests but the shape and magnitude of the velocity gradient is different. This is accomplished by using a finer-scale parametrization for Fig. 13(c) compared to Fig. 13(a) (but still cubic-splines) in order to achieve sharper velocity gradients that cannot be represented by the coarser parametrization of Fig. 13(a). The output of model 2 uses the same parametrization as the model 1 test, but is unable to exactly represent the model 2 input anomalies. This leads to implicit data inconsistencies (eq. 6) that brings the second term of eq. (16) into play.

Of course, one could argue that using a fine parametrization and applying smoothing and damping might allow for a greater range of
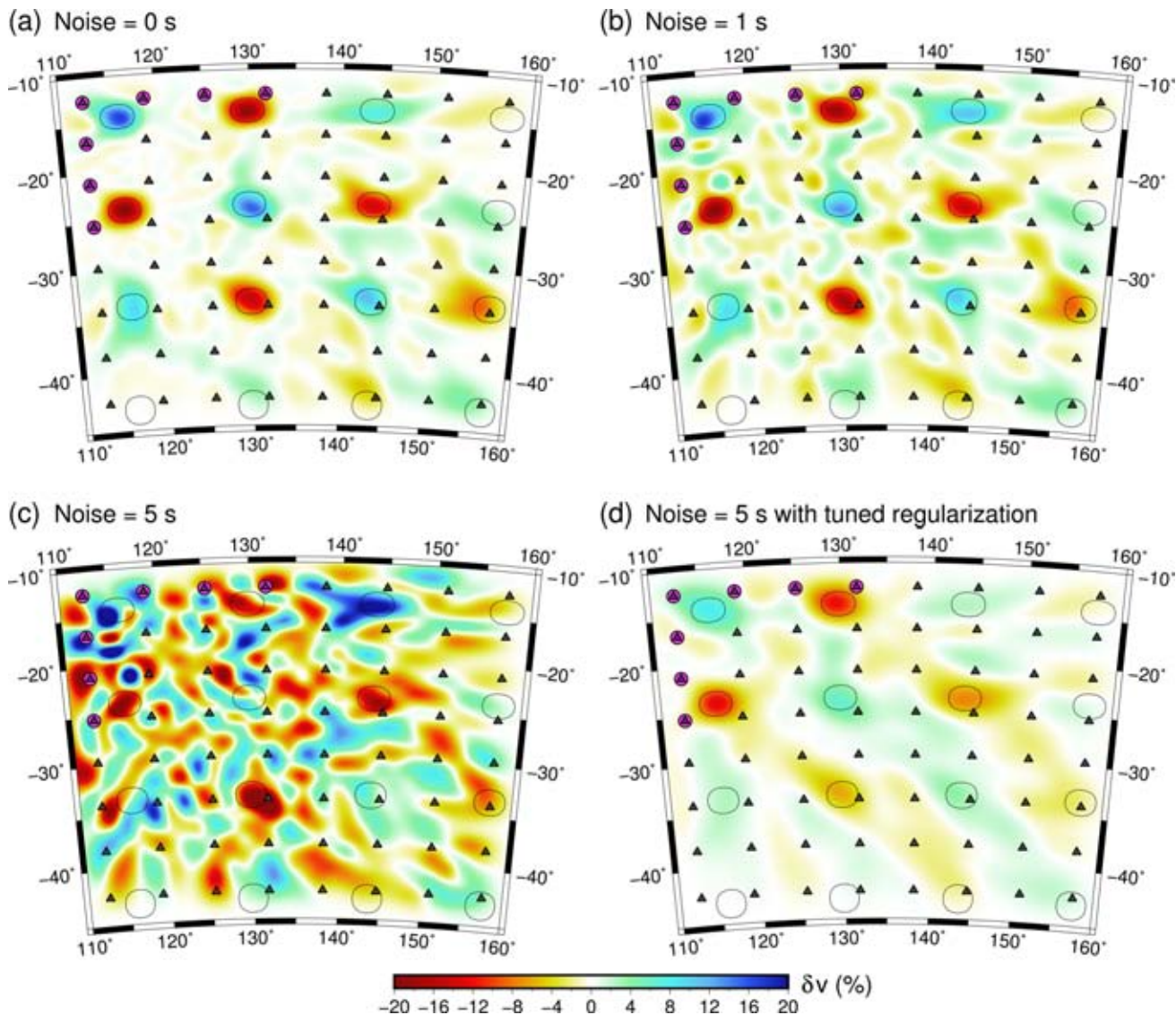
Figure 12. Synthetic spike test (based on the Fig. 9 example) showing the effects on the recovery of structure of the introduction of Gaussian noise with a standard deviation of (a) 0 s; (b) 1 s; (c) 5 s. In (d), smoothing and damping has been applied to the example from (c) in an attempt to find the simplest model that satisfies the data.

structures to be represented, but regularization is normally applied evenly without accounting for variations in the information content of the data. It is true that the sensitivity of the model unknowns to the data fit may vary, in which case the influence of the regularization will also vary spatially, but this is without regard to data resolution (i.e. a poorly constrained parameter can still have a large impact on data fit). Data adaptive parametrizations (e.g. Bijwaard *et al.* 1998; Spakman & Bijwaard 2001; Bodin & Sambridge 2009; Piana Agostinetti *et al.* 2015) attempt to minimize this effect, but most tomography is still carried out using static regular parametrizations. Other types of preconditioning include approximations in the solution of the forward and inverse problem. For example, ray theory is assumed to be valid in all numerical experiments undertaken in this study, so finite frequency effects are ignored; this will degrade the quality of the reconstruction as a result of the modelling error ϵ_t (see eq. 6), yet the degree to which this might do so is hard to quantify without running sophisticated full wave equation solvers.

3.10 Colour representation

The visual representation of structure in seismic tomography is usually done via colour or greyscale variations. It is well known in seismic tomography that the choice of colour scale can influence the way in which the results are interpreted, and the same applies to the synthetic test. In general, using a colour scale that exhibits a linear variation and constant intensity in colour is preferred, although there is no 'correct' choice. Fig. 14 shows the spike test reconstruction of Fig. 9(d) using two contrasting colour scales, which may give different impressions of the resolving power of the data.

4 DISCUSSION

No matter what data or method is used, the seismic tomography problem usually boils down to a large under or mixed determined nonlinear inverse problem, with the challenge of quantifying the range of data-satisfying solutions exacerbated by poorly constrained
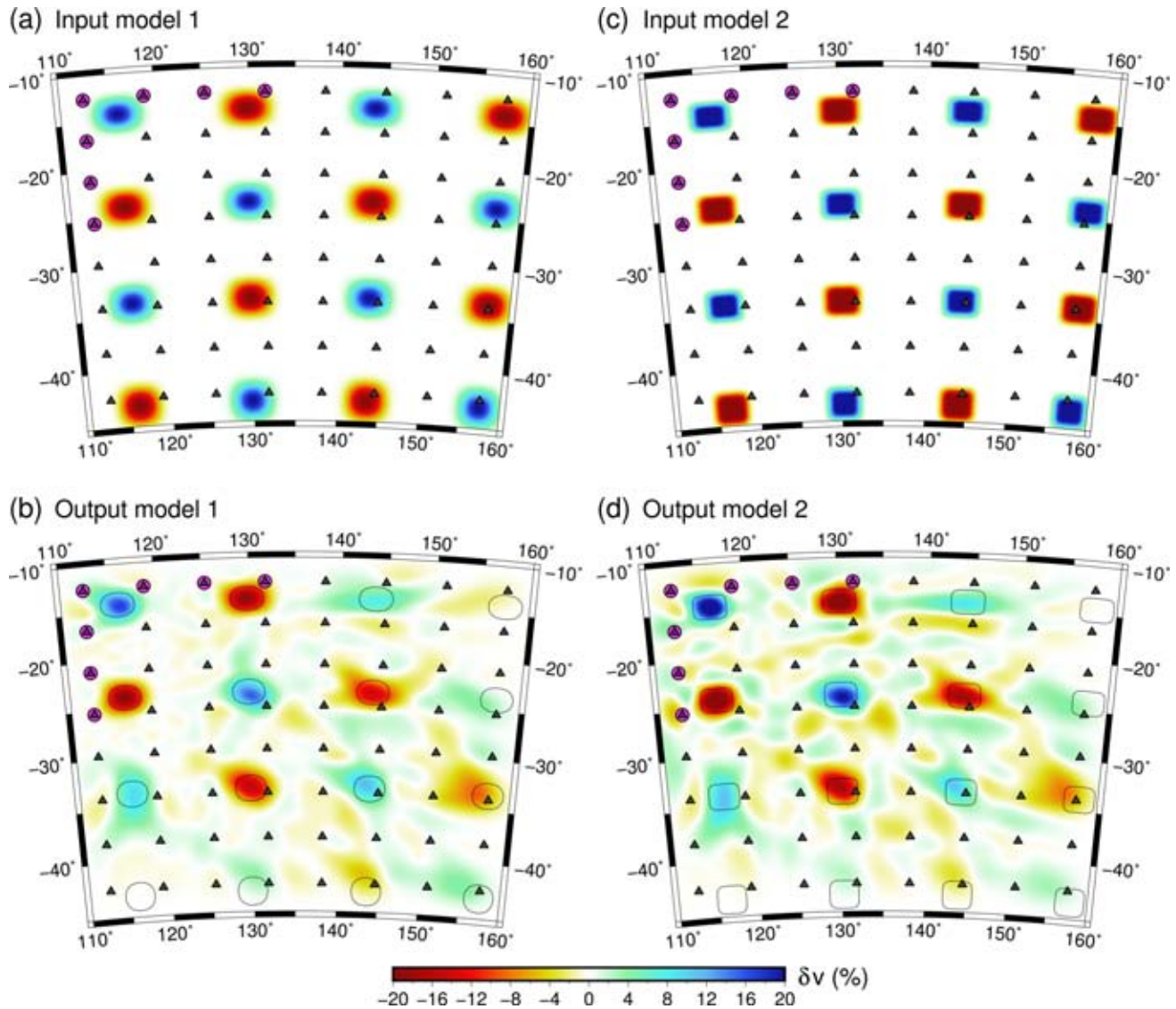
Figure 13. Synthetic spike test illustrating the effect of structural preconditioning. Input model 1 (a) and output model 1 (b) use identical parametrization, but input model 2 (c) uses a different parametrization to its corresponding output model (d).

data noise (explicit data errors ϵ_d) and a variety of assumptions, including the extent of prior knowledge, parametrization choice (ϵ_p), approximations in the physics of wave propagation (implicit data errors ϵ_t) and the linearization of the inverse problem (ϵ_l). The combined effect of all of this on solution uncertainty is almost impossible to measure.

In the last two–three decades there has essentially been no major advancement in the way that model accuracy (e.g. spatial resolution and model covariance) is assessed, either by direct computation or implicitly by sensitivity analysis. This is despite rapid gains in computing power, and is out of balance with the significant theoretical and computational development of physically elaborate forward models leading to modern tomographic techniques such as full waveform inversion. This is part of the reason why, in most studies that use synthetic recovery tests to assess model robustness, the results are generally interpreted qualitatively (Rawlinson *et al.* 2014); this can often be as simple as the qualitative description of detected resolution artefacts from a visual correlation between the input and output anomalies. The other reason is that sensitivity analysis prin-

cipally leads to an implicit assessment of spatial resolution unless only one spike is used (Spakman & Nolet 1988).

Of course, methods that quantify solution uncertainty explicitly are more attractive and should be preferred, but are limited in application because of computational demands. For example, posterior covariance, resolution and correlation (Rawlinson *et al.* 2014) can be readily computed for linear or weakly nonlinear problems, but can be expensive to obtain and difficult to interpret for large problems. For linear inverse problems we have demonstrated the theoretical links between the resolution matrix **R**, which can be computed explicitly for some problems, and sensitivity analysis. This provides a theoretical foundation for sensitivity analysis and an alternative means to retrieving information on the resolution matrix. Sensitivity analysis can also focus on the propagation of any kind of noise into the solution. If the resolution matrix can be computed, sensitivity analysis becomes even more powerful as a complementary means to assess how lack of resolution affects the observational model. With **R** available, spike tests on all scales can be rapidly computed to expose lack of resolution more insightfully than inspection of **R**
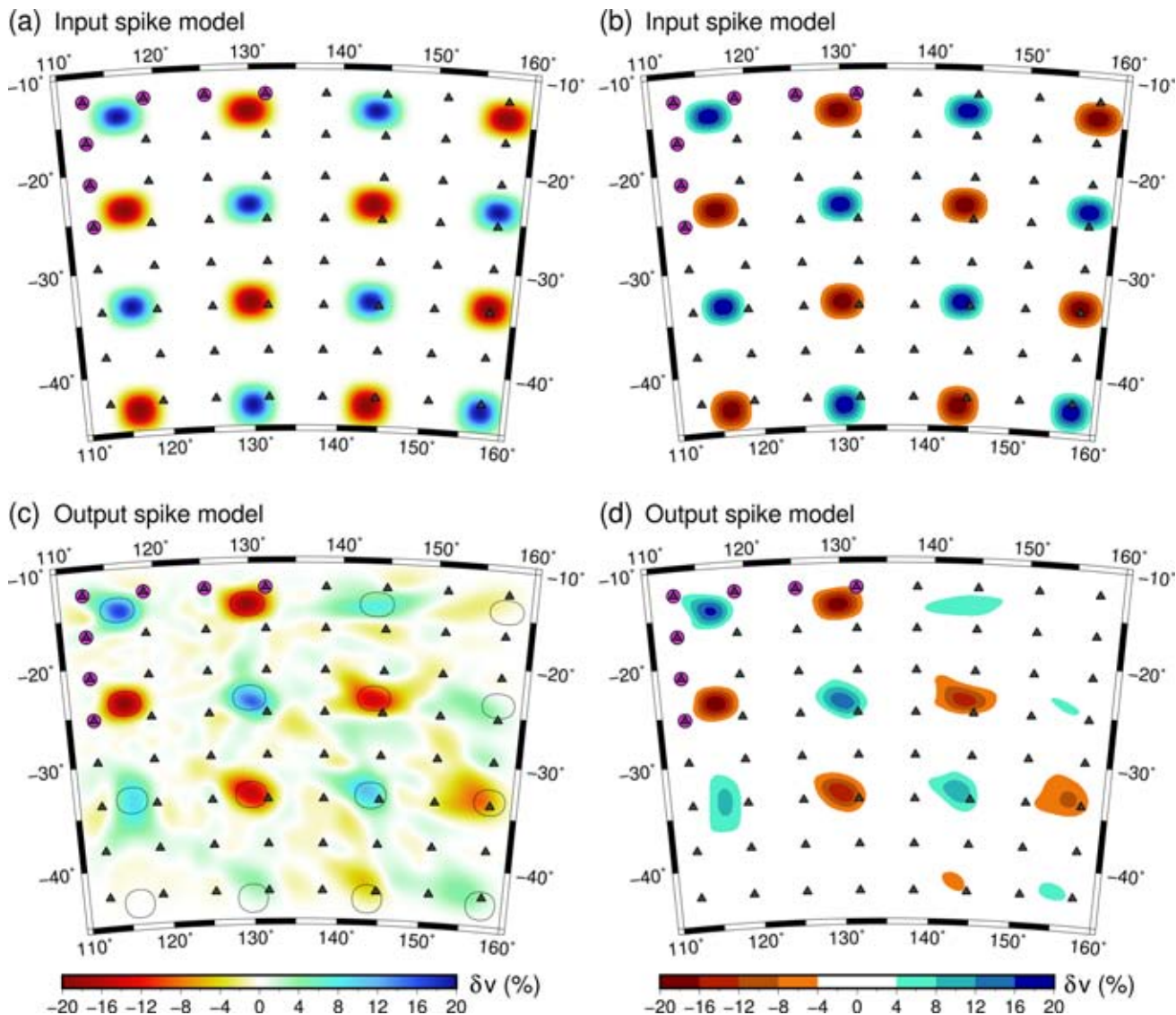
Figure 14. Synthetic spike test which uses different colour scales to produce different impressions of the resolving power of the data set. The inputs (a) and (b) are identical, as are the outputs (c) and (d), except for the choice of colour scale.

in isolation. In addition, special tests on the observational model can be conducted equally fast and may give meaning to subjective resolution proxies such as resolution length.

The series of numerical experiments carried out in this study is designed to illustrate some of the theoretical implications outlined in Section 2.5, and allows us to highlight a number of pitfalls that should be avoided. One of our main messages is that sensitivity analysis should be used to detect *lack of resolution* (not the degree of model recovery) and that checkerboard tests are less useful for this purpose than sparse spike tests. The reason is simple: resolution artefacts in checkerboard models are a superposition of the resolution artefacts exposed in the constituent spike models. The use of spikes allows the true blurring of structure to be more accurately assessed, and can be more closely associated with the rows of the resolution matrix (Spakman & Nolet 1988).

Another important confirmation of our numerical experiments pertains to the proper use of synthetic tests when the inverse problem is treated as nonlinear. Figs 6 and 7 clearly illustrate that path coverage through a checkerboard or spike model can be very different from that through a model obtained by inversion of the actual observations. The wrong implementation of sensitivity tests is to use the paths and traveltimes associated with the synthetic test model and perform a nonlinear inversion that includes updating data predictions and ray paths through the model. In this case, the recovered structure can be quite different compared to a linear inversion using rays inherited from the observational model. This is highlighted by two tests, illustrated in Fig. 7 and Supporting Information Fig. S4. For weaker velocity heterogeneity, such as the few percent velocity perturbations typically encountered in teleseismic tomography (e.g. Rawlinson & Fishwick 2012), the differences in outcome will be much smaller. Given that our observational model is the best estimate we have available for the true model, it makes sense, and is in agreement with the theory, to use rays through this model as a basis for assessing model robustness, rather than paths through some arbitrary model which likely bears no resemblance to the Earth that the data samples. Of course, nonlinear error propagation cannot be ignored if velocity heterogeneities are large, but this is hard to assess in the case of real data. However, the underlying theory does allow a potentially useful avenue to investigate the effects of nonlinearity, as was demonstrated in Fig. 8.

While it may be tempting to set up a synthetic recovery test that only uses the smallest anomalies that the data are capable of resolving, this is inadvisable for several reasons. First, as has been demonstrated by Lévêque *et al.* (1993), this does not necessarily mean that when recovery of anomalies of this size shows little evidence for lack of resolution, the same will apply to larger anomalies. Second, the multi-scale nature of Earth structure means that regions of good recovery will at least vary according to the wavelength of anomalies (see Fig. 9 and Supporting Information Fig. S5). Thus, performing multiple tests using structures of different size is important to detect lack of resolution across a range of plausible scales. The need for this approach has long been recognized, with the whole Earth studies of Inoue *et al.* (1990) and Fukao *et al.* (1992) implementing some of the first synthetic tests of this kind, albeit with a checkerboard structure, or by Bijwaard *et al.* (1998) using a whole range of spike sensitivity tests and test models derived from the observational model. Zelt (1998) advocates an approach in which a range of checkerboard tests are carried out, with anomaly position, size and orientation being the three key variables. The semblance between the true and recovered model is measured in each case, and the smallest well-resolved cell at each model node is used as a spatially dependent resolution estimate.

The inclusion of structural features derived from the inversion of the observational data set as input structure for a synthetic test needs to be done with care, because we have demonstrated that taking the observation model as the synthetic test model leads to almost complete recovery (Fig. 10), which indicates that such an approach is insensitive to resolution artefacts. For example, if the inversion of local earthquake data produces an elongated structure that resembles a subduction zone, using this feature of the model as input into a synthetic test will not necessarily produce useful constraints on how well it is resolved if significant ray smearing is responsible for its original appearance. Downdip smearing can, however, be detected correctly by deriving from the observational model a special 'layer-cake' model in which complete layers in the observational model have been set to zero (Bijwaard *et al.* 1998). An alternative suggested by Bezada *et al.* (2014) is the so-called 'squeezing test', in which the inverse problem is preconditioned to find the best-fitting model that excludes a certain feature of interest. The residuals that result from the subtraction of the data predictions through this model from the observational data are then inverted to see if the data requires this feature to be present.

Synthetic tests in which features of the model recovered from the observational data set are used as input, are distinct from the class of so-called hypothesis tests in which one seeks to understand whether a particular source–receiver distribution can recover a speculative structure such as a gap in a subduction zone. In this case, incorporating nonlinearity in the forward problem is worthwhile. For example, the recovery test in Fig. 1 could be regarded as a hypothesis test, but it does not examine the resolution of the resulting model, which is the focus of this paper. One could easily conceive a hypothesis test in which the speculative structure is an intrusive dyke with laterally elevated velocities. Placing two stations at either end of the dyke and applying ambient noise surface wave tomography (for example) may well result in excellent recovery of the dyke, but ultimately it is likely to be a misleading test if the true structure does not exactly conform with this hypothesis.

The effect of explicit regularization on the recovery of structure can be significant, as illustrated in Fig. 11. The regularization used in a synthetic test for assessing lack of resolution in the observational model should be inherited from its observational counterpart (as it is part of the generalized inverse; see Section 2), so testing different

damping and smoothing (for example) parameters in a synthetic recovery test is only relevant if one wants to tune the regularization for use in the inversion of the real data. Since the choice of damping and smoothing is usually fairly *ad hoc* (the popular L-curve approach, in which one examines the rate of change of data fit with respect to the regularization term, is a case in point; see Rawlinson *et al.* 2014), it is important to understand its influence. This is particularly the case when one considers the noise level in the data (Fig. 12 and Supporting Information Fig. S6), which is often poorly constrained. Increasing the level of noise increases the presence of spurious small-scale structure, but this effect can be greatly reduced by applying regularization to distill out longer wavelength features. The effect of different levels of noise on the recovery of structure across a range of scale-lengths is important, as it provides insights on how noise propagates into the observational model by detecting noise-sensitive model regions for a given degree of regularization.

The preconditioning of synthetic resolution tests by using the same parametrization for the recovered model as the input model is common and is required by the theory. However, we showed that tests can be designed to assess the influence of implicit data errors due to model approximation in the solution. This kind of implicit noise is less of a problem for data adaptive parametrizations (e.g. Spakman & Bijwaard 2001; Bodin *et al.* 2012), but the vast majority of tomography is still undertaken using regular basis functions to represent structure. The tests illustrated in Fig. 13 and Supporting Information Fig. S7 clearly demonstrate that even seemingly minor differences in structure (sharp velocity gradients at the edge of anomalies instead of smooth ones) can result in a significant degrading of the result if the chosen parametrization is unable to accurately represent such a feature, implicitly leading to data noise that activates the second term of eq. (16). While in practice there is no one or even small ensemble of tests that could be applied to properly explore this issue with a particular choice of parametrization and particular data set, it nonetheless remains an important consideration.

As shown in Fig. 14, the choice of colour scale may play a role in how one assesses the resolving power of a data set. Again, this is quite well known, and also applies to the way tomographic models are represented in general. Since colour scales are unavoidably subjective, one cannot be prescriptive in their use. However, it is advisable to avoid strongly varying the colour gradient (rate of change from one colour to another) or intensity (brightness) with respect to the measured quantity, as this makes interpretation more difficult.

5 CONCLUSIONS

Theory is presented and a series of numerical experiments have been performed in order to help establish best practice for the proper implementation of synthetic reconstruction tests. The objective of this study is to demonstrate the theoretical links with spatial resolution and, in recognition of their broad popularity in seismic tomography, to try and highlight potential pitfalls and provide recommendations as to what kinds of synthetic tests may be the most useful. Our findings include the following:

(i) As for formal resolution analysis, sensitivity tests only strictly apply to linear tomographic problems. However, they can provide useful insight in the presence of weakly nonlinear inverse problems.

(ii) Sensitivity analysis can be theoretically coupled to the formal resolution matrix. For example, a discrete spike recovery test is akin to retrieving a column or row of the resolution matrix.

(iii) Theoretical prediction errors (e.g. the use of approximate forward theory like ray tracing) are ignored in sensitivity analysis, yet it is conceivable that they may influence the results significantly.

(iv) Sensitivity tests are only useful for the detection of *lack of resolution* and not the detection of model recovery. Good recovery of one particular test model is relatively meaningless.

(v) Discrete spike tests are more useful for assessing the resolving power of the data set to recover structure compared to traditional checkerboards, which feature a tight oscillatory pattern of positive and negative anomalies.

(vi) For nonlinear problems, the path coverage through the final observational model (i.e. the model produced by inversion of the recorded data) should be used in the synthetic test, not the path coverage obtained by solving the forward problem through the synthetic model.

(vii) Synthetic experiments should test lack of resolution across at least the same range of scale lengths that are found or interpreted in the observational model. For synthetic models containing only one wavelength of structure, multiple tests involving different-sized anomalies should be used.

(viii) Input structures that closely resemble the output structure from the observational model should not be used in synthetic tests, as they cannot detect lack of resolution.

(ix) The effects of noise and its interplay with the imposed regularization can be explored with sensitivity tests by investigating the range of structural wavelengths that can be recovered using different levels of imposed noise.

(x) Experiments that test the influence of the imposed parameterization on the accuracy of the reconstruction (structural preconditioning) may be important, although they can be time consuming and ultimately difficult to draw firm conclusions from.

(xi) It is important to use sensible colour scales that avoid large fluctuations in intensity and gradient.

ACKNOWLEDGEMENTS

REFERENCES

Aki, K. & Lee, W.H.K., 1976. Determination of the three-dimensional velocity anomalies under a seismic arraynusing first *P* arrival times from local earthquakes: 1. A homogeneous inital model, *J. geophys. Res.,* **81,** 4381–4399.

Aki, K., Christoffersson, A. & Husebye, E.S., 1977. Determination of the three-dimensional seismic structure of the lithosphere, *J. geophys. Res.,* **82,** 277–296.

Backus, G. & Gilbert, F., 1970. Uniqueness in the inversion of inaccurate gross Earth data, *Phil. Trans. R. Soc. Lond., A,* **266,** 123–192.

Backus, G.E. & Gilbert, J.F., 1967. Numerical applications of a formalism for geophysical inverse problems, *Geophys. J. R. astr. Soc.,* **13,** 247–276.

Backus, G.E. & Gilbert, J.F., 1968. The resolving power of gross earth data, *Geophys. J. R. astr. Soc.,* **16,** 169–205.

Bezada, M., Humphreys, E., Davila, J.M., Carbonell, R., Harnafi, M. & Palomeras, I., 2014. Piecewise delamination of Moroccan lithosphere from beneath the Atlas Mountains, *Geochem. Geophys. Geosyst.,* **15,** 975–985.

Bijwaard, H. & Spakman, W., 1999. Tomographic evidence for a narrow whole mantle plume below Iceland, *Earth planet. Sci. Lett.,* **166,** 121–126.

Bijwaard, H., Spakman, W. & Engdahl, E.R., 1998. Closing the gap between regional and global travel time tomography, *J. geophys. Res.,* **103,** 30 055–30 078.

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *Geophys. J. Int.,* **178,** 1411–1436.

Bodin, T., Sambridge, M., Rawlinson, N. & Arroocau, P., 2012. Transdimensional tomography with unknown data noise, *Geophys. J. Int.,* **189,** 1536–1556.

Chen, P., Zhao, L. & Jordan, T.H., 2007. Full 3d tomography for the crustal structure of the los angeles region, *Bull. seism. Soc. Am.,* **97,** 1094–1120.

Deal, M.M. & Nolet, G., 1996. Nullspace shuttles, *Geophys. J. Int.,* **124,** 372–380.

de Wit, R.W.L., Trampert, J. & van der Hilst, R.D., 2012. Toward quantifying uncertainty in travel time tomography using the null-space shuttle, *J. geophys. Res.,* **117,** B03301, doi:10.1029/2011JB008754.

Eberhart-Phillips, D. & Reyners, M., 1997. Continental subduction and three-dimensional crustal structure: the northern South Island, New Zealand, *J. geophys. Res.,* **102,** 11 848–11 861.

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

Fichtner, A. & van Leeuwen, T., 2015. Resolution analysis by random probing, *J. geophys. Res.,* **120,** 5549–5573.

Fishwick, S., Kennett, B.L.N. & Reading, A.M., 2005. Contrasts in lithospheric structure within the Australian craton - insights from surface wave tomography, *Earth planet. Sci. Lett.,* **231,** 163–176.

Franklin, J.N., 1970. Well-posed stochastic extensions of ill-posed linear problems, *J. Math. Anal. Appl.,* **31,** 682–716.

Fukao, Y., Obayashi, M., Inoue, H. & Nebai, M., 1992. Subducting slabs stagnant in the mantle transition zone, *J. geophys. Res.,* **97,** 4809–4822.

Galetti, E., Curtis, A., Baptie, B. & Meles, G., 2015. Uncertainty loops in seismic tomography, *Phys. Rev. Lett.,* **114,** doi:10.1103/PhysRevLett.114.148501.

Glahn, A. & Granet, M., 1993. Southern Rhine Graben: small-wavelength tomographic study and implications for the dynamic evolution of the graben, *Geophys. J. Int.,* **113,** 399–418.

Gorbatov, A., Widiyantoro, S., Fukao, Y. & Gordeev, E., 2000. Signature of remnant slabs in the North Pacific from *P*-wave tomography, *Geophys. J. Int.,* **142,** 27–36.

Graeber, F.M. & Asch, G., 1999. Three-dimensional models of *P* wave velocity and *P*-to-*S* velocity ratio in the southern central Andes by simultaneous inversion of local earthquake data, *J. geophys. Res.,* **104,** 20 237–20 256.

Grand, S.P., 1987. Tomographic inversion for shear velocity beneath the North American plate, *J. geophys. Res.,* **92,** 14 065–14 090.

Gung, Y. & Romanowicz, B., 2004. Q tomography of the upper mantle using three-component long-period waveforms, *Geophys. J. Int.,* **157,** 813–830.

Hawkins, R. & Sambridge, M., 2015. Geophysical imaging using transdimensional trees, *Geophys. J. Int.,* **203,** 972–1000.

Inoue, H., Fukao, Y., Tanabe, K. & Ogata, Y., 1990. Whole mantle *P*-wave travel time tomography, *Phys. Earth planet. Inter.,* **59,** 294–328.

Kennett, B.L.N., Sambridge, M.S. & Williamson, P.R., 1988. Subspace methods for large scale inverse problems involving multiple parameter classes, *Geophys. J.,* **94,** 237–247.

Kissling, E., 1988. Geotomography with local earthquake data, *Rev. Geophys.,* **26,** 659–698.

Lees, J.M. & Crosson, R.S., 1989. Tomographic inversion for three-dimensional velocity structure at Mount St. Helens using earthquake data, *J. geophys. Res.,* **94,** 5716–5728.

Lees, J.M. & Crosson, R.S., 1990. Tomographic imaging of local earthquake delay times for three-dimensional velocity variation in western washington, *J. geophys. Res.,* **95,** 4763–4776.

Lévêque, J.J., Rivera, L. & Wittlinger, G., 1993. On the use of the checkerboard test to assess the resolution of tomographic inversions, *Geophys. J. Int.,* **115,** 313–318.

Meju, M.A., 2009. Regularized extremal bounds analysis (REBA): an approach to quantifying uncertainty in nonlinear geophysical inverse problems, *Geophys. Res. Lett.,* **36,** L03304, doi:10.1029/2008GL036407.

Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory,* Academic Press.

Nolet, G., 2008. *A Breviary of Seismic Tomography: Imaging the Interior of the Earth and Sun,* Cambridge Univ. Press.

Nolet, G., Montelli, R. & Virieux, J., 1999. Explicit, approximate expressions for the resolution and a posteriori covariance of massive tomographic systems, *Geophys. J. Int.,* **138,** 36–44.

Parker, R.L., 1994. *Geophysical Inverse Theory,* Princeton Univ. Press.

Piana Agostinetti, N., Giacomuzzi, G. & Malinverno, A., 2015. Local three-dimensional earthquake tomography by trans-dimensional monte carlo sampling, *Geophys. J. Int.,* **201,** 1598–1617.

Rasmussen, J. & Humphreys, E., 1988. Tomographic image of the Juan de Fuca plate beneath Washington and western Oregen using teleseismic P-wave travel times, *Geophys. Res. Lett.,* **15,** 1417–1420.

Rawlinson, N. & Fishwick, S., 2012. Seismic structure of the southeast Australian lithosphere from surface and body wave tomography, *Tectonophysics,* **572,** 111–122.

Rawlinson, N. & Kennett, B.L.N., 2008. Teleseismic tomography of the upper mantle beneath the southern Lachan Orogen, Australia, *Phys. Earth planet. Inter.,* **167,** 84–97.

Rawlinson, N. & Sambridge, M., 2003. Seismic traveltime tomography of the crust and lithosphere, *Adv. Geophys.,* **46,** 81–198.

Rawlinson, N., Reading, A.M. & Kennett, B.L.N., 2006. Lithospheric structure of Tasmania from a novel form of teleseismic tomography, *J. geophys. Res.,* **111,** B02301, doi:10.1029/2005JB003803.

Rawlinson, N., Sambridge, M. & Saygin, E., 2008. A dynamic objective function technique for generating multiple solution models in seismic tomography, *Geophys. J. Int.,* **174,** 295–308.

Rawlinson, N., Pozgay, S. & Fishwick, S., 2010. Seismic tomography: a window into deep Earth, *Phys. Earth planet. Inter.,* **178,** 101–135.

Rawlinson, N., Kennett, B., Vanacore, E., Glen, R. & Fishwick, S., 2011. The structure of the upper mantle beneath the Delamerian and Lachlan orogens from simultaneous inversion of multiple teleseismic datasets, *Gondwana Res.,* **19,** 788–799.

Rawlinson, N., Fichtner, A., Sambridge, M. & Young, M.K., 2014. Seismic tomography and the assessment of uncertainty, *Adv. Geophysics,* **55,** 1–76.

Saygin, E. & Kennett, B.L.N., 2010. Ambient seismic noise tomography of Australian continent, *Tectonophysics,* **481,** 116–125.

Sethian, J.A., 1996. A fast marching level set method for monotonically advancing fronts, *Proc. Natl. Acad. Sci.USA,* **93,** 1591–1595.

Spakman, W., 1991. Delay-time tomography of the upper mantle below Europe, the Mediterranean and Asia Minor, *Geophys. J. Int.,* **107,** 309–332.

Spakman, W. & Bijwaard, H., 2001. Optimization of cell parameterizations for tomographic inverse problems, *Pure appl. Geophys.,* **158,** 1401–1423.

Spakman, W. & Nolet, G., 1988. Imaging algorithms, accuracy and resolution in delay time tomography, in *Mathematical Geophysics: A Survey of Recent Developments in Seismology and Geodynamics,* pp. 155–187, eds Vlaar, N.J., Nolet, G., Wortel, M.J.R. & Cloetingh, S.A.P., Springer.

Spakman, W., Stein, S., van der Hilst, R. & Wortel, R., 1989. Resolution experiments for nw pacific subduction zone tomography, *Geophys. Res. Lett.,* **16,** 1097–1100.

Stankiewicz, J., Ryberg, T. & Haberland, C., 2010. Lake Toba volcano magma chamber imaged by ambient seismic noise tomography, *Geophys. Res. Lett.,* **37,** L17306, doi:10.1029/2010GL044211.

Steck, L.K., Thurber, C.H., Fehler, M., Lutter, W.J., Roberts, P.M., Baldridge, W.S., Stafford, D.G. & Sessions, R., 1998. Crust and upper mantle *P* wave velocity structure beneath Valles caldera, New Mexico: results from the Jemez teleseismic tomography experiment, *J. geophys. Res.,* **103,** 24 301–24 320.

Su, W.-J. & Dziewonski, A.M., 1997. Simultaneous inversion for 3-D variations in shear and bulk velocity in the mantle, *Phys. Earth planet. Inter.,* **100,** 135–156.

Tarantola, A., 1987. *Inverse Problem Theory,* Elsevier.

Vasco, D.W., 2007. Invariance, groups, and non-uniqueness: the discrete case, *Geophys. J. Int.,* **168,** 473–490.

Vasco, D.W., Peterson, J.E. & Majer, E.L., 1996. Nonuniqueness in travel-time tomography: ensemble inference and cluster analysis, *Geophysics,* **61,** 1209–1227.

Walck, M.C. & Clayton, R.W., 1987. *P* wave velocity variations in the Coso region, California, derived from local earthquake travel times, *J. geophys. Res.,* **92,** 393–405.

White, D.J., 1989. Two-dimensional seismic refraction tomography, *Geophys. J.,* **97,** 223–245.

Wiggins, R.A., 1972. The general linear inverse problem: implication of surface waves and free oscillations for earth structure, *J. geophys. Res.,* **10,** 251–285.

Wolfe, C.J., Solomon, S.C., Silver, P.G., VanDecar, J.C. & Russo, R.M., 2002. Inversion of body-wave delay times for mantle structure beneath the Hawaiian islands: results from the PELENET experiment, *Earth planet. Sci. Lett.,* **198,** 129–145.

Yang, T., Grand, S.P., Wilson, D., Guzman-Speziale, M., Gomez-Gonzalez, J., Dominiguez-Teyes, T. & Ni, J., 2009. Seismic structure beneath the Rivera subduction zone from finite-frequency seismic tomography, *J. geophys. Res.,* **114,** B01302, doi:10.1029/2008JB005830.

Yao, Z.S., Roberts, R.G. & Tryggvason, A., 1999. Calculating resolution and covariance matrices for seismic tomography with the LSQR method, *Geophys. J. Int.,* **138,** 886–894.

Young, M.K., Rawlinson, N., Arroucau, P., Reading, A.M. & Tkalčić, H., 2011. High-frequency ambient noise tomography of southeast Australia: new constraints on Tasmania's tectonic past, *Geophys. Res. Lett.,* **38,** L13313, doi:10.1029/2011GL047971.

Young, M.K., Cayley, R.A., Mclean, M.A., Rawlinson, N., Arroucau, P. & Salmon, M., 2013. Crustal Structure of the east Gondwana margin in southeast Australia revealed by transdimensional ambient seismic noise tomography, *Geophys. Res. Lett.,* **40,** 4266–4271.

Zelt, C.A. & Smith, R.B., 1992. Seismic traveltime inversion for 2-D crustal velocity structure, *Geophys. J. Int.,* **108,** 16–34.

Zelt, C.A. & Barton, P.J., 1998. Three-dimensional seismic refraction tomography: a comparison of two methods applied to data from the Faeroe Basin, *J. geophys. Res.,* **103,** 7187–7210.

Zelt, C.A., 1998. Lateral velocity resolution from three-dimensional seismic refraction data, *Geophys. J. Int.,* **135,** 1101–1112.

Zelt, C.A., 1999. Modelling strategies and model assessment for wide-angle seismic traveltime data, *Geophys. J. Int.,* **139,** 183–204.

Zhang, H. & Thurber, C.H., 2007. Estimating the model resolution matrix for large seismic tomography problems based on Lanczos bidiagonalization with partial reorthogonalization, *Geophys. J. Int.,* **170,** 337–345.

Zhao, D., Kanamori, H. & Humphreys, E., 1996. Simultaneous inversion of local and teleseismic data for the crust and mantle structure of southern California, *Phys. Earth planet. Inter.,* **93,** 191–214.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this paper:

Figure S1. Australian ambient noise Rayleigh wave tomography model, showing variations in 6.7 s period group velocities (a) and the associated path coverage (b). Small black triangles denote station locations.

Figure S2. Comparison of a checkerboard test (left column) with a spike test (right column) for the Australian ambient noise data set.

Figure S3. Comparison of spike test from Fig. S2 (left column) with an anti-spike test (right column) for the Australian ambient noise data set.

Figure S4. Checkerboard (left column) and spike (right column) test results from using different path geometries. (a), (b) result from linear inversions using the path geometry of Fig. S1b; (c), (d) result

from iterative non-linear inversion with ray path updates. Closed red dashed lines denote features discussed in the text.

Figure S5. Checkerboard and spike tests similar to Fig. S2 but with longer wavelength anomalies. In this case, only the output models are illustrated.

Figure S6. Synthetic spike tests illustrating the role of data noise in the accuracy of reconstruction. In each case, Gaussian noise with a standard deviation of (a) 0 s, (b) 2 s, (c) 5 s and (d) 5 s is used. In the case of (d) regularisation is applied to locate a minimum structure model. Initial RMS and Final RMS refer to the root mean square of the data fit for the initial and recovered model respectively.

Figure S7. Synthetic spike test which illustrates the effect of a preconditioned parameterization. Spike test 1 (left column) uses identical parametrization for the input and output models, while spike test 2 (right column) does not.
(http://gji.oxfordjournals.org/lookup/suppl/doi:10.1093/gji/ggw084/-/DC1).

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.