Geophysical Journal International

Geophys. J. Int. (2018) **212,** 522–542
Advance Access publication 2017 October 9
GJI Seismology

doi: 10.1093/gji/ggx428

Low frequency full waveform seismic inversion within a tree based Bayesian framework

Anandaroop Ray, Sam Kaplan, John Washbourne and Uwe Albertin

Chevron Energy Technology Company, Houston, TX 77002*, USA. E-mail: a.ray@chevron.com*

Accepted 2017 October 5. Received 2017 October 3; in original form 2017 July 6

SUMMARY

Limited illumination, insufficient offset, noisy data and poor starting models can pose challenges for seismic full waveform inversion. We present an application of a tree based Bayesian inversion scheme which attempts to mitigate these problems by accounting for data uncertainty while using a mildly informative prior about subsurface structure. We sample the resulting posterior model distribution of compressional velocity using a trans-dimensional (trans-D) or Reversible Jump Markov chain Monte Carlo method in the wavelet transform domain of velocity. This allows us to attain rapid convergence to a stationary distribution of posterior models while requiring a limited number of wavelet coefficients to define a sampled model. Two synthetic, low frequency, noisy data examples are provided. The first example is a simple reflection + transmission inverse problem, and the second uses a scaled version of the Marmousi velocity model, dominated by reflections. Both examples are initially started from a semi-infinite half-space with incorrect background velocity. We find that the trans-D tree based approach together with parallel tempering for navigating rugged likelihood (i.e. misfit) topography provides a promising, easily generalized method for solving large-scale geophysical inverse problems which are difficult to optimize, but where the true model contains a hierarchy of features at multiple scales.

Key words: Inverse theory; Probability distributions; Wave propagation.

1 INTRODUCTION

The active source seismic full waveform inversion (FWI) method is in principle, a simple idea. With minimal processing or manual intervention, it aims to provide not just an image of the subsurface, but a velocity model which when put though a forward operator, 'closely' matches the observed seismic field (Tarantola 1984). This entails the solution of an inverse problem, with the forward physics governed by the seismic wave equation. However, such inverse problems with limited receiver coverage as well as frequency bandwidth are extremely nonlinear and thus very challenging to solve. Further, the presence of noise at inopportune frequencies confounds many optimization methods, and complicated earth models make for a very high dimensional model space that is difficult to work with in a computationally efficient manner. The nonlinearity alluded to manifests as local misfit minima, leading to models that are not optimally converged or are 'cycle skipped' in FWI parlance. A thorough review of seismic FWI and the challenges involved can be found in Virieux & Operto (2009). Active efforts to tackle these challenges are documented in special journal sections edited by Routh *et al.* (2016) and Tanis & Behura (2017). Various promising methods to improve convergence exist, such as the estimation of time shifts to minimize the kinematic differences between initially modelled and observed data, the use of extended model domains and/or non-local wave physics (Biondi & Almonin 2014; Vigh *et al.* 2016; Fu & Symes 2017). Another approach is to solve a sequence of constrained, locally convex subproblems (e.g. Esser *et al.* 2016). Yet other methods seek to improve the convexity of the misfit function through the use of an optimal transport distance as in Métivier *et al.* (2016), via the addition of artificial low frequencies to data (Choi & Alkhalifah 2016), the iterative use of Wiener filters (Warner & Guasch 2016), or the use of quadratic penalty methods (Van Leeuwen & Herrmann 2015). One commonality of all these methods is an effort to make the misfit topography easier for optimization algorithms to navigate. To varying degrees, all of these methods work well under different circumstances, but cannot guarantee convergence as argued by Fichtner & Trampert (2011). Further, given the various steps involved, these methods are not easily amenable to solution appraisal or uncertainty estimation. In this paper, we attempt to quantify the credibility (in the Bayesian sense e.g. Jaynes 2003) with which we provide solutions to the FWI problem, when such solutions themselves are not easy to find. Further, the algorithm automatically selects and operates with a limited set of discrete wavelet transform (Mallat 1989) coefficients of the velocity model. This leads to a reduced number of unknowns than cells in the forward modelling finite difference grid, thus allowing for tractable uncertainty estimation in 2-D and potentially 3-D FWI with minimal assumptions being made *a priori*.

2 MODEL SELECTION AND TREE BASED PARAMETERIZATION

2.1 The need for model selection

In most conventional schemes for geophysical inversion, the model grid geometry is fixed, that is, the size of the cells and their number is not allowed to vary during inversion. Traditionally, solutions have focused on minimizing the following objective function:

$$\arg\min \phi(\mathbf{m}) = ||\mathbf{W}(\mathbf{d} - \mathbf{f}(\mathbf{m}))||_2^2 + \lambda^2 ||\mathbf{R}\mathbf{m}||_p^p, \quad (1)$$

where **m** is a model vector, **d** is the observed data and **f(m)** provides the forward modelled prediction due to **m**. λ^2 is the regularization parameter, **R** is any operator which once applied to **m**, produces a measure of length in the p norm that is deemed sensible to keep small. The first term in (1) is the data misfit (weighted by the data precision **W**), and the second is the regularization term designed to keep the model (or deviations of the model from a preferred model) small. The trade-off between the two is controlled by the so called Tikhonov regularization parameter λ^2 (Tikhonov 1963). This is akin to the statistical technique of ridge regression, that is, depending on the value of λ^2, for a linear problem and the $p = 2$ norm, the solution to (1) lies on the 'ridge' between the minimum of the data misfit term and the minimum of the model length term in order to simultaneously minimize both. Clearly, choices need to be made regarding the operator **R**, the weight given to the model length, and the selection of model norm. Nonlinear least squares FWI solutions involving gradient descent or the use of the Jacobian matrix (or its inverse) in conjunction with Tikhonov regularization are easy enough to conceptualize, well understood, but notoriously slow to converge if **R** is poorly chosen. Choosing a smaller number of parameters, or using a $p = 1$ norm in conjunction with a sparse model 'frame' does away with some of this hyper-parameter selection, as illustrated by Aravkin *et al.* (2011) and Xue & Zhu (2015).

Of course, the use of sparse model representations with small measures of length not only aid FWI convergence to a suitable solution of (1), there is another observation which can be made regarding parsimonious model parametrizations—simpler theories (models in our case) offer clearer insight. This is the approach used by Occam's inversion, which aims to produce the smoothest model (Constable *et al.* 1987) or the sparsest model (Wheelock & Parker 2013) compatible with the data noise. However, these models are extremal models, and should not be looked at as being truly representative of the earth. To wit, we should consider models which are suitably simple, but also fit the data appropriately. Statistically, this is known as the model selection problem. The goal is to avoid producing simple models which have low variance but high bias, or complicated models with high variance but low bias. Lever *et al.* (2016) discusses this fundamental problem when selecting regression models to fit data. Ideally for geophysical inversion, we should be sampling over not one, but a range of models compatible with our data as well as our prior notions of the earth *and* its complexity.

2.2 A Bayesian solution

In the methods outlined so far, the goal has been to find a minimum of (1), with the hope that it is a global minimum. As we have mentioned, no such convergence guarantee exists (Fichtner & Trampert 2011). Further, even if a global minimum were to be found, it would not preclude the existence of other models with similar misfits which fit within the data noise. These models will likely exhibit very different velocity structure, typical of a nonlinear problem. Continuing with the geophysical ideal mentioned at the end of the last section, we would like to sample with a range of hyper-parameters, a range of models such that the models themselves are of an appropriate complexity, with seismic velocities that conform to log data, outcrops, and laboratory experiments while being compatible with the noisy seismic observations. We could try to do this manually by trial and error but would quickly realize the need to use a systematic approach. We would still need to quantitatively weight the outcomes due to each combination of hyper-parameters and inverted models.

Fortunately, a mathematically sound method of accomplishing this task exists—we can re-examine (1) in a Bayesian sense. For every sampled model **m**, loosely speaking, the misfit term provides a measure of the likelihood of the model, while the length of the model vector encapsulates our prior knowledge about the model, including its complexity (e.g. Fukuda & Johnson 2010). More rigorously speaking, a Bayesian formulation is

$$p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m})p(\mathbf{m}), \quad (2)$$

which for our purposes is better read from right to left as follows. $p(\mathbf{m})$ is the prior probability of **m**, which we know independent of the observations **d**. We re-assess our prior notion of **m** by carrying out a seismic experiment which shows us how likely it is that **m** fits the observations. This weight is given by the likelihood function $p(\mathbf{d}|\mathbf{m})$. The result of re-weighting or updating our prior notion by the likelihood provides the *posterior* probability of observing the model **m**. The posterior probability is represented by the term $p(\mathbf{m}|\mathbf{d})$. We then repeat this process for various models **m** admissible by our prior notions until we obtain an ensemble of models represented by the probability density function or PDF $p(\mathbf{m}|\mathbf{d})$. We can thus turn the optimization problem (1) with many possible solutions, into a sampling problem (2). Astute readers will note that (2) is missing a normalization constant which ensures it integrates to unity, and thus is not truly a probability density function. Indeed, (2) is more representative of a multidimensional histogram until we normalize it by integrating over all models on the right-hand side:

$$p(\mathbf{d}) = \int_{\mathbf{m}} p(\mathbf{d}|\mathbf{m})p(\mathbf{m})\,\mathrm{d}\mathbf{m}, \quad (3)$$

where $p(\mathbf{d})$ is known as the *evidence*. However for model appraisal, we are only interested in the relative probabilities of various models. We can thus sample up to a constant of proportionality using (2) for our purposes. It is important to note that our prior in (2) includes a specification over various levels of complexity (including parametrizations with different numbers of variables) and $p(\mathbf{d})$ is therefore the 'total' evidence (see Sambridge *et al.* 2006).

2.3 Model complexity and Reversible Jump Markov chains

For the optimization problem (1), as applicable to any geophysical problem, model regularization is necessary from a number of different viewpoints, be it for improving the stability of a matrix inverse, for keeping model fluctuations (or the update) small, or for keeping the model (or the update) close to a preferred model. However, the number of parameters with which to describe the model, a measure of the complexity of the model, can also be treated as an unknown to sample, without explicitly requiring regularization. With this approach, we consider not only the simplest or smoothest model with which to describe the data, but a collection of models with a *different number* of parameters which are compatible with the observed data. The trans-dimensional (Sambridge *et al.* 2006)

inversion method based on birth/death Monte Carlo (Geyer & Møller 1994) and the more general Reversible Jump Markov chain Monte Carlo (RJ-McMC) method (Green 1995) accomplishes just this task. For a 1-D model, this would mean sampling over a variable number of layers (Malinverno & Leaney 2000; Minsley 2011; Bodin *et al.* 2012b; Dettmer *et al.* 2015). For 2-D models, Voronoi cells with different numbers of cells have been widely used (e.g. Bodin & Sambridge 2009; Dettmer *et al.* 2014; Ray *et al.* 2014; Galetti *et al.* 2015; Saygin *et al.* 2016). In effect, the trans-D algorithm via Bayes' theorem performs the task of model selection, with regard to the complexity of the model. The fact that models are neither overfit nor underfit is based on the idea of Bayesian parsimony, as brought to the attention of the geoscientific community by Malinverno & Leaney (2000) and Malinverno (2002). An 'Occam factor' which penalizes overly complicated models, is built into the framework of Bayes' theorem when formulated appropriately (MacKay 2003). To examine this argument, we note that a trans-D model vector is defined as $\mathbf{m} = [\mathbf{m}_k, k]$, where $\mathbf{m}_k$ is a model with k parameters that describe compressional velocity (for the FWI application in this study). As shown in equation 5 of Ray *et al.* (2016), we can derive from the joint probability of the data and models, that

$$p(k|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m}_k, k)}{p(\mathbf{d})}\left[\frac{p(\mathbf{m}_k|k)p(k)}{p(\mathbf{m}_k|k, \mathbf{d})}\right]. \quad (4)$$

Treating the total evidence $p(\mathbf{d})$ as a constant, we get

$$p(k|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m}_k, k)\left[\frac{p(\mathbf{m}_k|k)p(k)}{p(\mathbf{m}_k|k, \mathbf{d})}\right]. \quad (5)$$

The term on the left-hand side of (5) is the posterior probability (after performing the experiment), on inferring the number of parameters k. The first term on the right is the likelihood of k parameters fitting the data adequately. To examine the bracketed, second term on the right, we first note from the definition of joint and conditional probability that $p(\mathbf{m}_k, k) = p(\mathbf{m}_k|k)p(k)$. Therefore, the bracketed term on the right-hand side is the ratio of prior model probability to posterior probability for a k-parameter model. The more number of parameters k there are, the more thinly spread (i.e. lower) the prior probability is, since the prior PDF needs to integrate to 1 over a larger volume. Since acceptable k-parameter models occupy *a posteriori* a tiny amount of the prior space, the k-parameter posterior probability is generally higher (i.e. peakier) than the prior. The more parameters k are used, the less therefore is the bracketed fraction. However, the likelihood of the k-parameter fit increases the more number of parameters k we use. In a trans-D formulation, the bracketed factor and the data likelihood trade off, *automatically* providing a solution akin to regularization, depending largely on the *data*. With a uniform probability for $p(k)$, and some simplifying assumptions discussed in (Ray *et al.* 2016), the bracketed fraction can be interpreted as the ratio of the posterior accessible volume to the prior accessible volume, sometimes known as the 'Occam Factor.' This formulation allows an inversion to self-parametrize to good effect, providing higher model resolution in areas with better data coverage and low noise (Bodin *et al.* 2009).

Now that we have interpreted the trans-D formulation, lest it appear that the right-hand side of (5) depends on $\mathbf{m}_k$ while the left does not, we can simply use the definition of conditional probabilities again, to verify that the right-hand side of (5) equals $p(k, \mathbf{d})$. This is entirely consistent with (4), since by definition, $p(k|\mathbf{d}) = \frac{p(k,\mathbf{d})}{p(\mathbf{d})}$. It was demonstrated by Bodin *et al.* (2012a) that trans-D outperforms inversion based on subspace transformations using B-splines, for a seismic surface wave tomography application. Alternatives to a trans-D formulation based on evaluating the evidence for different parametrizations via the *marginal likelihood* $p(\mathbf{d}|k)$, or, the evidence for a given hypothesis (in our case a k-parameter model) can be found in Kass & Raftery (1995). However, this involves the computationally prohibitive task of finding the evidence for each k-parametrization, and is only feasible for certain kinds of geophysical inversion (e.g. Dettmer *et al.* 2010; Brunetti *et al.* 2017).

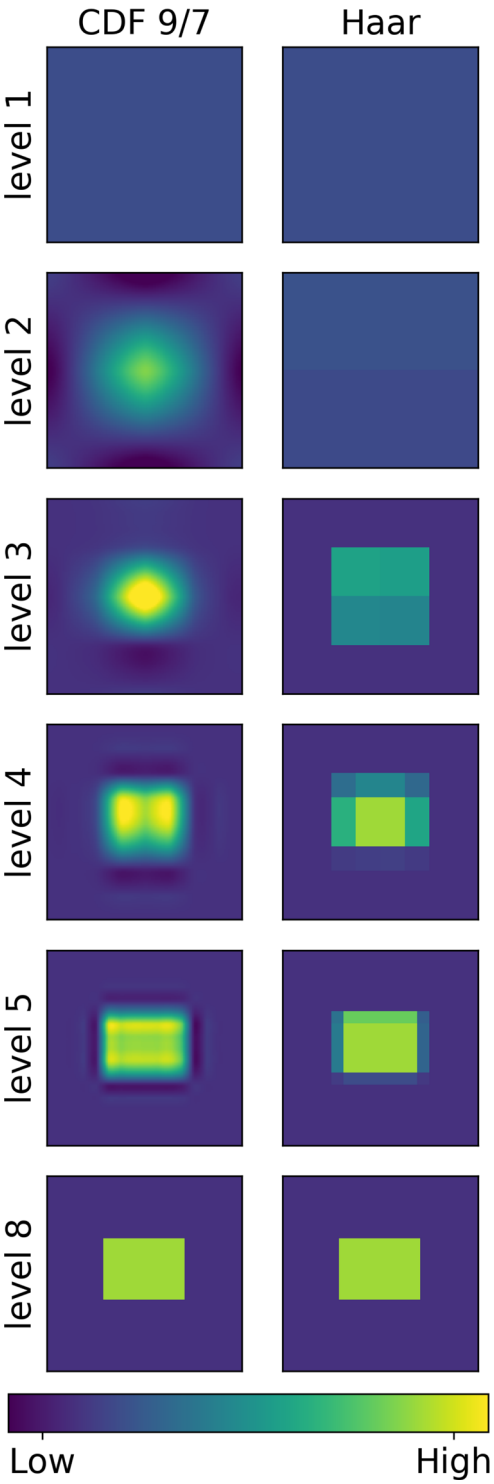

Figure 1. Inverse discrete wavelet transforms at five levels of truncation in the transform domain for the same image. Level 8 corresponds to no truncation and the true 128×128 image. Left: using the CDF 9/7 basis. Right: using the Haar basis functions.

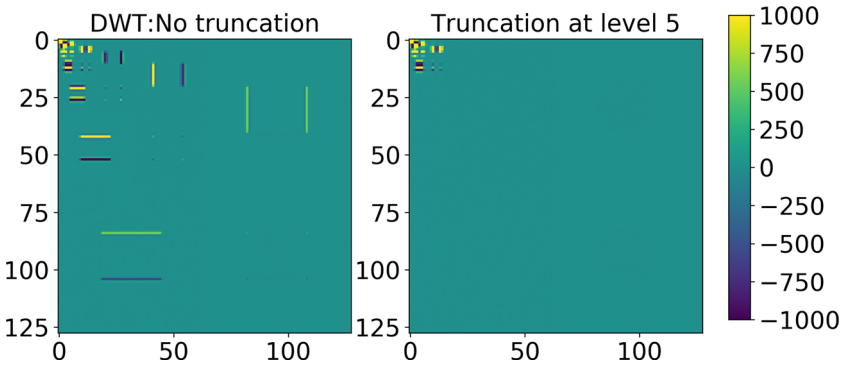

Figure 2. Left: transform domain view of the DWT with the CDF 9/7 basis of the image in the last row of Fig. 1. Lower order coefficients, corresponding to coarser, low wavenumber features are seen to the top left. Most of the wavelet coefficients are near zero (green). Right: inverse transforming this truncated DWT produces the level 5 image in Fig. 1.

For the exploration seismic FWI problem, solutions to characterize the full nonlinear uncertainty have only recently been put forward, owing to the huge computational cost of a forward evaluation. Fang *et al.* (2014) present a Bayesian solution based on randomized source subsampling but make use of a fixed parametrization while assuming a Gaussian distribution about the maximum a posteriori (MAP) model, similar to the pioneering work of Scales & Gouveia (1998). Mazzotti *et al.* (2016) use a genetic algorithm (GA) in conjunction with model resampling using the neighbourhood algorithm (NA; Sambridge 1999) followed again by Gibbs sampling (GS; Geman & Geman 1984). They use a two grid approach, coarse for the inverse model, and fine for the forward model. However, the data do not determine the coarseness of the inverse model grid, and the quality of the estimated uncertainty also depends on the input ensemble from the GA to the NA+GS algorithm (see Sambridge 1999). Stuart *et al.* (2016) present a two grid approach which involves operator upscaling though the inverse model grid is fixed. All of these methods are promising efforts to quantify seismic FWI uncertainty, but do not address the model selection problem. The only efforts we are aware of which have attempted this with trans-D inversions are Malinverno & Leaney (2005) for the vertical seismic profile (VSP) inversion problem, and Ray *et al.* (2016) for the elastic FWI problem, but both assume a laterally invariant earth model. Dettmer & Dosso (2013) use 2-D Voronoi cells for the related range-dependent geoacoustic sounding problem. In theory, the Bayesian model selection principles demonstrated for 1-D and 2-D earth models are equally applicable for 3-D inversion. However, as pointed out by Hawkins & Sambridge (2015), computationally efficient parametrizations for trans-D problems in 3-D (e.g. Piana Agostinetti *et al.* 2015; Burdick & Lekić 2017) are not easy to construct, and the inclusion of prior knowledge about geometric structure is difficult.

2.4 Tree based model representations

The recent work of Hawkins & Sambridge (2015) has demonstrated that any basis function set which can be represented by a tree based structure, can be used as a valid model representation for trans-D inversion. A major advantage of using this formulation is that from both a theoretical and practical efficiency point of view, it is agnostic to the spatial dimensionality of the earth model, be it 1-D, 2-D or 3-D. In this work, we specifically use wavelet basis functions (e.g. Daubechies 1992) and the discrete wavelet transform (DWT) Mallat (1989), which is readily amenable to a hierarchical tree based representation. Wavelet transforms with a suitable basis set (e.g. CDF 9/7, Cohen *et al.* 1992) are routinely used to compress image information (e.g. JPEG 2000, Taubman & Marcellin 2002). This makes the transform domain attractive for parsimonious geophysical model representations, as we will demonstrate with our synthetic examples. As we have mentioned in the introduction, curvelet or wavelet basis sets have been used by Aravkin *et al.* (2011) and Lin *et al.* (2012) respectively for exploration seismic FWI, but in an optimization set up. As discussed by Hawkins & Sambridge (2015), a valid wavelet tree which is incompletely filled can represent a hierarchy of features from low to high spatial wavenumbers. In conjunction with the trans-D algorithm, this provides a multiresolution approach which adaptively parametrizes according to the observed data. Adaptive inversion grid meshing has been carried out by Sambridge & Faletič (2003) and Plattner *et al.* (2012), but these used fixed criterions for the adaptation rather than sample over a range of parametrizations where model complexity is dictated by the data. Successful recent applications of such a trans-D tree based approach can be found in Hawkins *et al.* (2017) for airborne electromagnetic inversion, Dettmer *et al.* (2016) to quantify uncertainty for tsunami sea surface displacement, Hawkins & Sambridge (2015) for 2-D and 3-D seismic tomography, and our work is the first usage we are aware of for seismic FWI.

3 METHODS

3.1 Model parametrization

For 1-D, 2-D and 3-D models, the tree representation requires use of *modified* binary tree, quaternary tree and octree structures respectively. For all these representations in the wavelet transform domain, the first node coefficient (which is at the top level of the tree) represents the average value of velocities in the model (to be presented to the finite difference operator). This node branches into 1, 3 and 7 nodes (again, for 1-D, 2-D and 3-D models respectively) at the second level, with coefficients at this level representing the strength of basis functions with wavelengths of roughly half the length scale of the total model. From this level downwards, each node branches into a pure binary tree, quadtree or octree where each child has 2, 4 and 8 children exactly. The tree depth is restricted by the size of the forward model grid. Each successive depth level (in the inverse wavelet transform domain) represents finer scaled features in the velocity model. In all the work presented here, we use the modified restricted quaternary trees as we are working in 2-D. Illustrations of this tree structure are presented in detail in the applications Section 4.1 and Appendix.

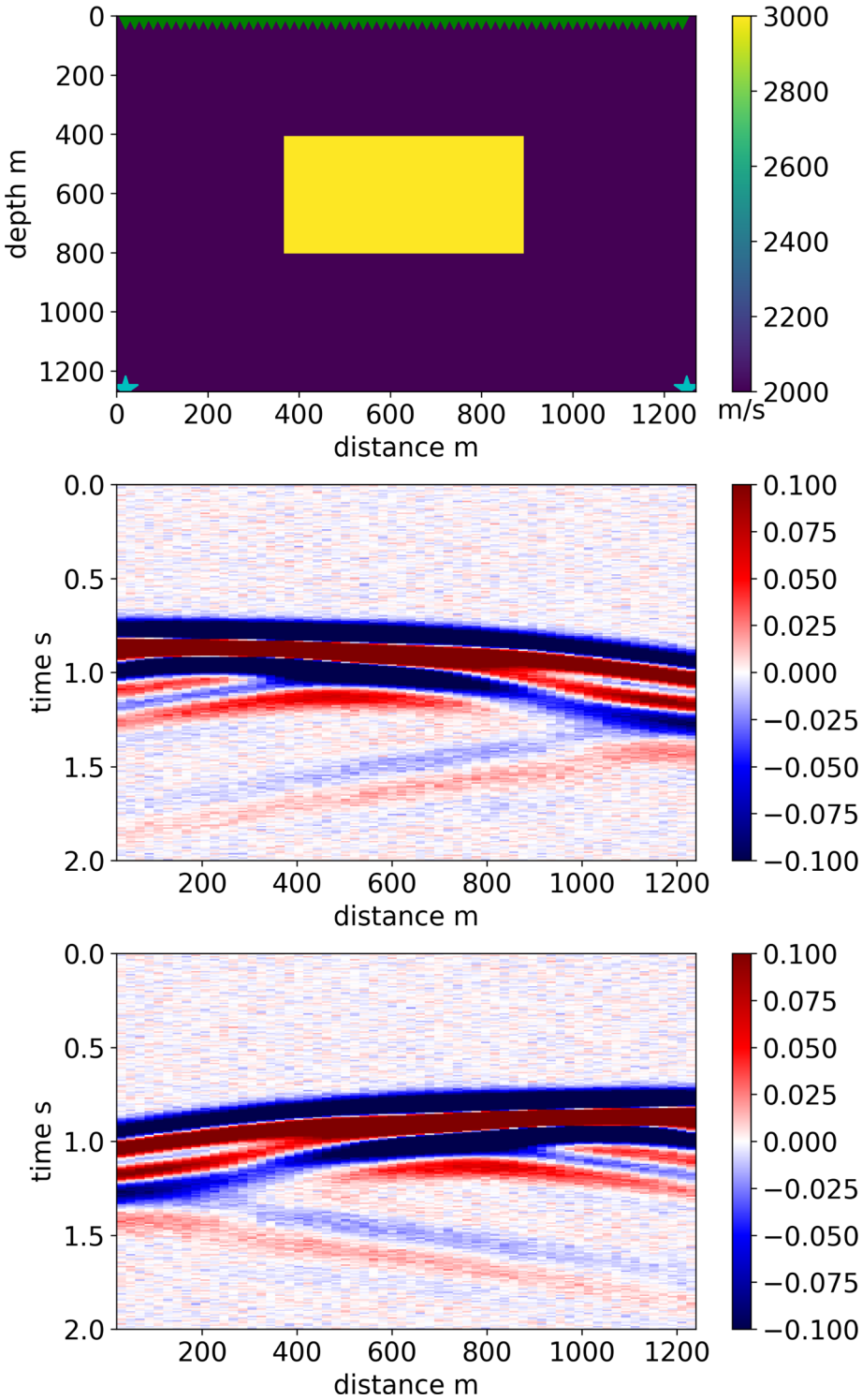

Figure 3. Model and noisy data for the transmission dominated study. The two shots (blue stars) are at depth near the edges of the model with a carpet of receivers (green triangles) at the surface. The anomalous body is a velocity spike of 3 km s^{-1} within a homogeneous background of 2 km s^{-1}. The density, not considered an unknown goes from 2.0 to 2.1 gm cc^{-1}.

Another advantage of working with the tree based wavelet transform representation is that different wavelet bases can be used, depending on the problem at hand. For transmission dominated problems, smooth basis functions such as CDF 9/7 may be appropriate. For reflection dominated problems, sharp basis functions such as the Haar wavelets (Haar 1910) could be used. Fig. 1 shows a 128 × 128 pixel image, at five levels of truncation in the transform domain, inverse transformed back to the image domain using these two kinds of basis functions. Level 8 corresponds to no truncation for a 128 × 128 square image, as $2^{\text{level}-1} = 128$. A limitation of using the DWT is that all dimensions must be a power of two. While we use square velocity models in this work, Hawkins & Sambridge

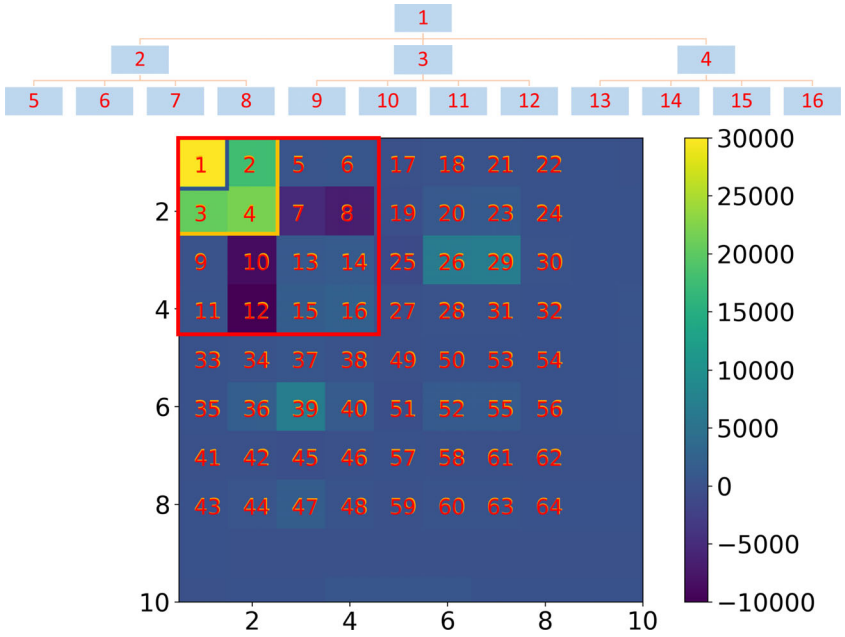

Figure 4. Top: the tree structure and its node numbering. Bottom: the corresponding node locations in the DWT domain. Different levels of the tree have been marked out on the DWT image. For example, level 2 corresponds to nodes numbered 2, 3 and 4. Level 3 corresponds to nodes 5–16. Each node corresponds to a basis function, the strength of which is represented by coefficient values that can be read from the colour scale. As previously mentioned, lower level nodes correspond to basis functions with low wavenumbers that represent coarser features.

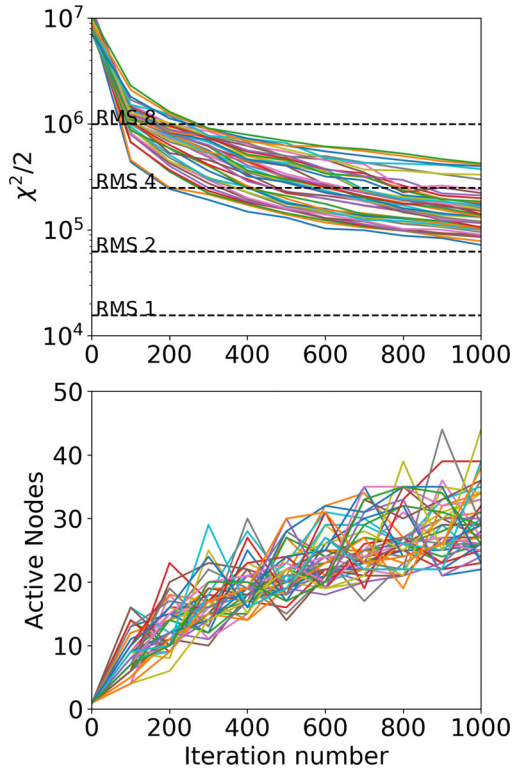

Figure 5. Initial McMC sampling progress. Top: each colour corresponds to a different McMC chain in Parallel Tempering. If the data is fit within noise, the expected RMS value attained should be 1. Bottom: the number of active nodes needed to achieve lower misfits increases with iteration number.

(2015) have shown how to use the DWT for rectangular model domains, by using more than one root node for the wavelet tree model. Fig. 2 shows a comparison in the wavelet transform domain using the CDF 9/7 basis, between the full wavelet transform and the truncated version at level 5. The level 5 representation requires a handful from a maximum of 16 × 16 coefficients to be non-zero, while providing the approximation in the 5th row, 1st column of Fig. 1.

3.2 Sampling the posterior velocity using trans-D trees

Sampling the posterior model PDF (2) is done via the trans-D McMC algorithm, details of which are provided in Appendix A. In particular, we sample different wavelet trees, adding nodes, deleting nodes or modifying node coefficients according to a prior specification and the likelihood function.

3.2.1 $p(\mathbf{d}|\mathbf{m})$ or the model likelihood $\mathcal{L}(\mathbf{m})$

In particular, for additive noise, which by central limiting is asymptotically Gaussian (especially in the frequency domain, as shown in Ray *et al.* 2016), we define the likelihood function $\mathcal{L}(\mathbf{m})$ as

$$\mathcal{L}(\mathbf{m}) = p(\mathbf{d}|\mathbf{m}) = \exp\left(-\frac{1}{2}[\mathbf{f}(\mathbf{m}) - \mathbf{d}]^{\mathrm{t}}\mathbf{C}_{\mathbf{d}}^{-1}\Big[\mathbf{f}(\mathbf{m}) - \mathbf{d}\Big]\right), \qquad (6)$$

where $\mathbf{C_d}$ is the covariance matrix of data errors. Since the DWT is a linear transformation, we can write

$$\mathbf{f}(\mathbf{m}) = \mathbf{F}(\mathbf{H}\mathbf{m}), \qquad (7)$$

where $\mathbf{F}$ is the seismic forward operator, $\mathbf{H}$ is the inverse DWT operator and $\mathbf{m}$ is a model vector represented by coefficient values on a wavelet tree. In other words, $\mathbf{Hm}$ is the 2-D velocity model fed to a variable density, acoustic and isotropic finite difference engine. The source signature is assumed known, or it can be derived as a maximum likelihood estimate as a function of the model, as shown in (Virieux & Operto 2009; Dettmer *et al.* 2015; Ray *et al.* 2016).

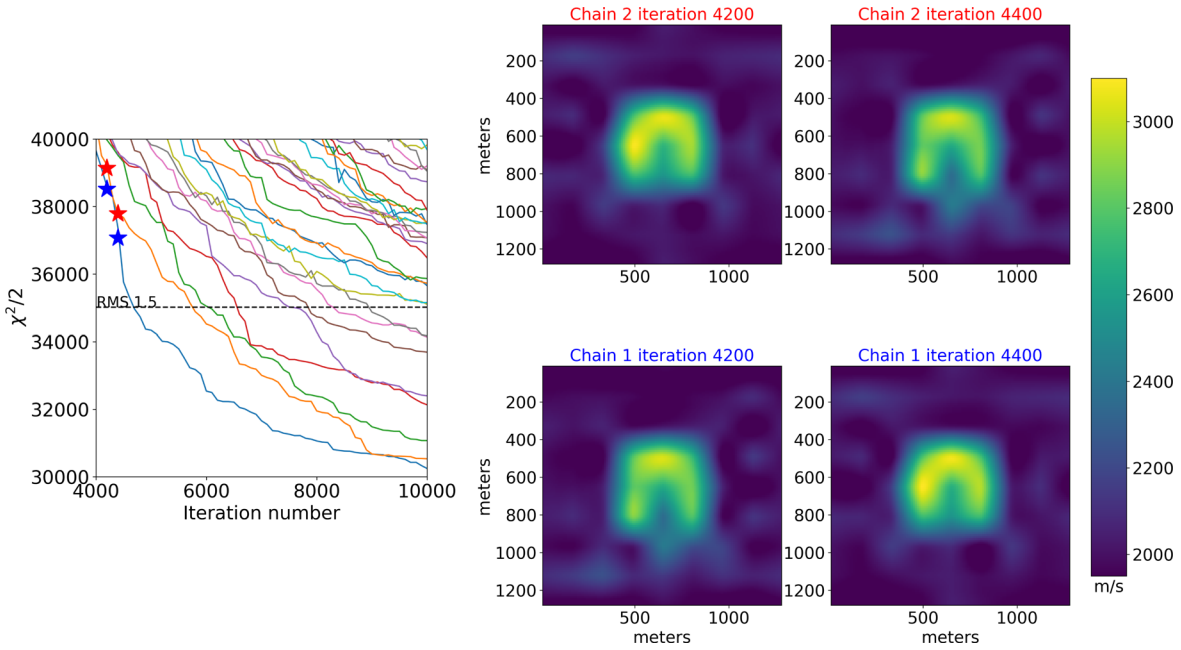

Figure 6. Parallel Tempering in action. Left: models from the red chain shown as red stars are exchanged with those from the chain at the lowest temperature shown in blue. Right: whenever an McMC chain samples better fitting models than chains at temperatures below it, a model exchange is proposed. The better fitting model at iteration 4200, initially at a higher temperature, is exchanged with a worse fitting model in the 'target' chain at $\mathcal{T} = 1$. At later steps, the target chain has the better fitting model. Since the higher temperature chains are better able to navigate local minima, this provides the system of McMC chains a rigorous way to navigate steep misfit topography.

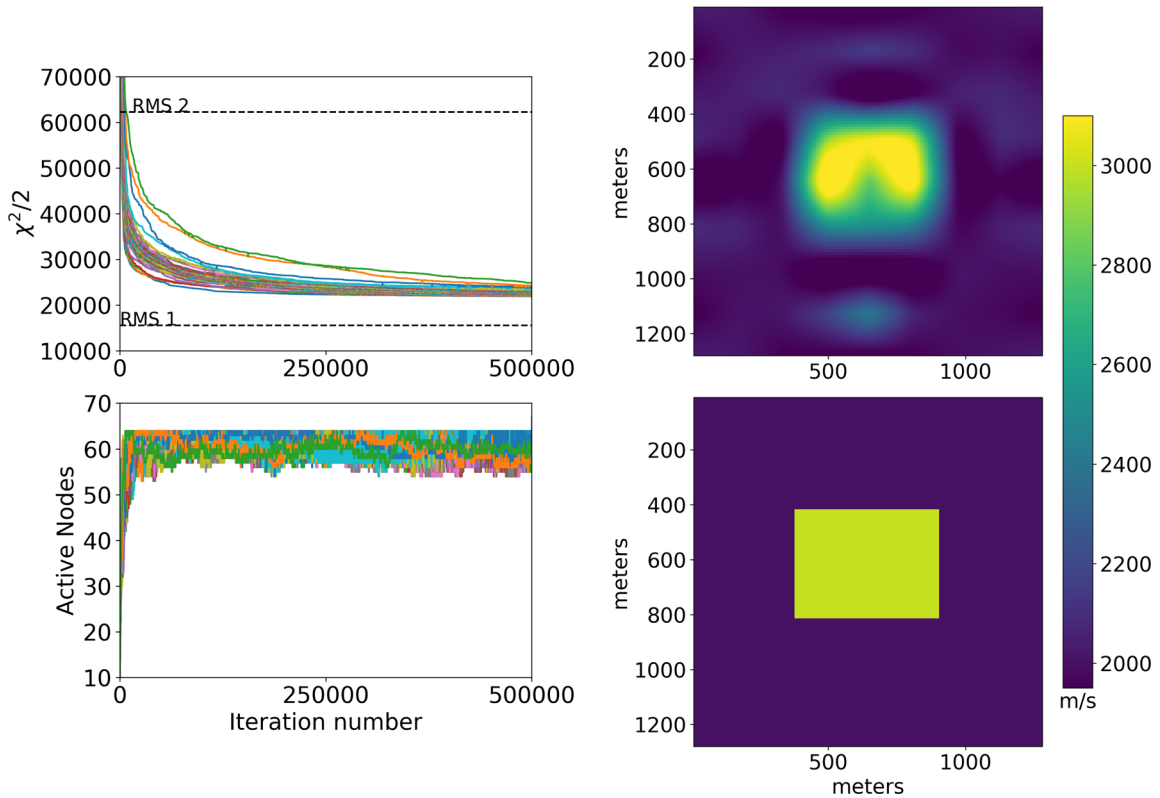

Figure 7. Left: sampling statistics when the maximum number of nodes is 64 (level 4). Right: true model and mean velocity from samples 250 000 to 500 000.

3.2.2 Prior formulation $p(\mathbf{m})$

In this work, we only concern ourselves with changes in velocity in the earth, assuming that density changes are known or that there are no changes in density. This is not a limitation of the method, which easily generalizes to more variables. The prior models need to specify the probabilities of nodes on a tree. Hence we can write

$$p(\mathbf{m}) = p(\mathbf{v}, T, k), \tag{8}$$

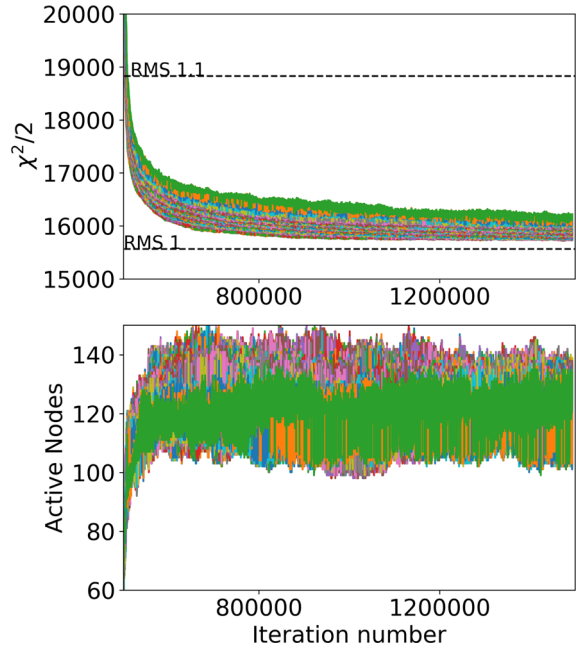

Figure 8. Sampling statistics when level 5 of the DWT is made accessible to sampled models. Though the maximum number of nodes allowed is 256, this number of active nodes is never required to fit the data within noise. The χ^2 misfit drops to very close to the theoretical expected value at RMS 1.

where $\mathbf{v}$ is a vector of velocities (in our application) in the *wavelet transform domain*, which is a point to note, that makes the tree based formulation different from layer or cell based trans-D. T is a particular type of wavelet tree (modified restricted quaternary trees for our 2-D application) and k is the number of active nodes representing a valid tree structure. Using the chain rule of probabilities, we can write:

$$p(\mathbf{v}, T, k) = p(\mathbf{v}|T, k)p(T|k)p(k),$$
$$p(\mathbf{v}, T, k) = p(T|k)p(k)\prod_{i=1}^{k} p(v_i|T, k). \quad (9)$$

The last term under the product assumes that the wavelet coefficients at each node, given k active nodes for the specific tree type T, are independent of each other. Hawkins & Sambridge (2015) and Dettmer *et al.* (2016) simply use wide uniform bounds at each node position. However, as can be seen in Fig. 2, these coefficient values span many orders of magnitude, but at a particular depth level the values are all within a limited span. To elaborate, for most naturally occurring images, values are generally more extreme at the top levels of the tree (representing coarser features) than values at depth levels that are farther from the origin (representing finer features). This is exactly analogous to a Fourier spectrum of most natural images containing stronger low wavenumber content as opposed to high wavenumber content. $p(k)$ is simply a prior on the number of nodes. It could be constant (i.e. uniform) or a Jeffrey's prior (Jeffreys 1939) inversely proportional to the number of active nodes (Hawkins & Sambridge 2015). The crucial development by Hawkins & Sambridge (2015) was the definition of $p(T|k)$. If we assume that given k active nodes, all valid tree structures with active nodes from the root node at the top down to a specified maximum depth are equally probable—a reasonable assumption since it would imply

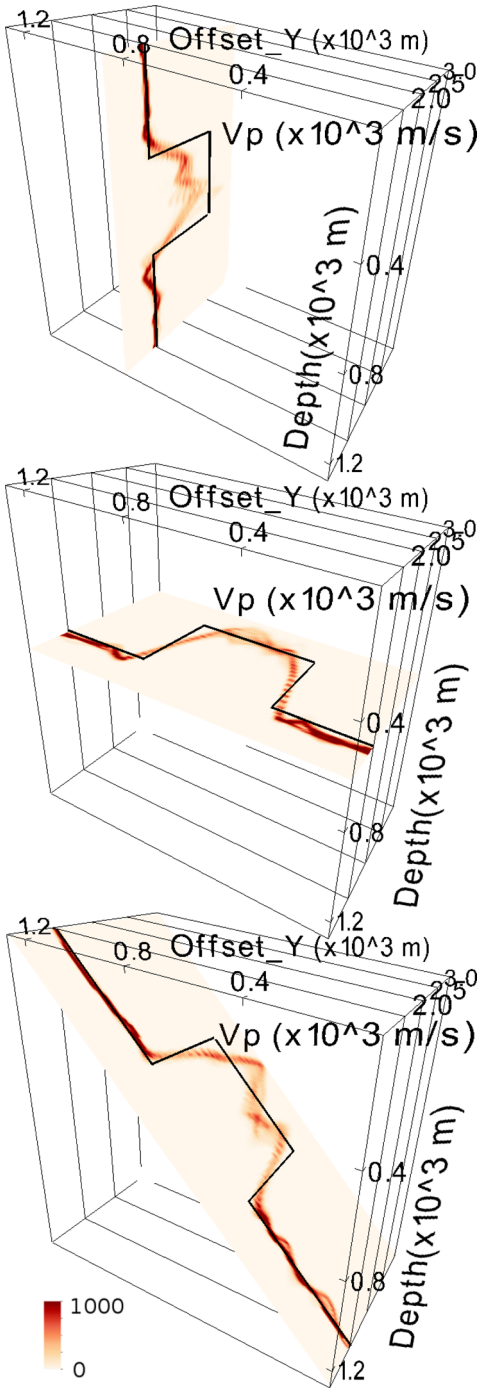

Figure 9. Slices through the marginal probability density function $p(\mathbf{m}|\mathbf{d})$ of velocity at every point in the model. Darker colours correspond to more probable values, with the true model given by the black line. These plots show that the background velocities and anomaly edges are fairly well inferred. However, the anomaly centre velocities are quite variable, probably due to the lack of illumination from directly beneath or above the anomaly. Trade-offs and multimodal distributions of velocity can also be clearly seen.

that any velocity model will possess features at coarse as well as fine scales—then to define this probability, we need to count the number of arrangements of such valid trees $\mathcal{N}_k$. The probability is simply given by

$$p(T|k) = \frac{1}{\mathcal{N}_k}. \quad (10)$$

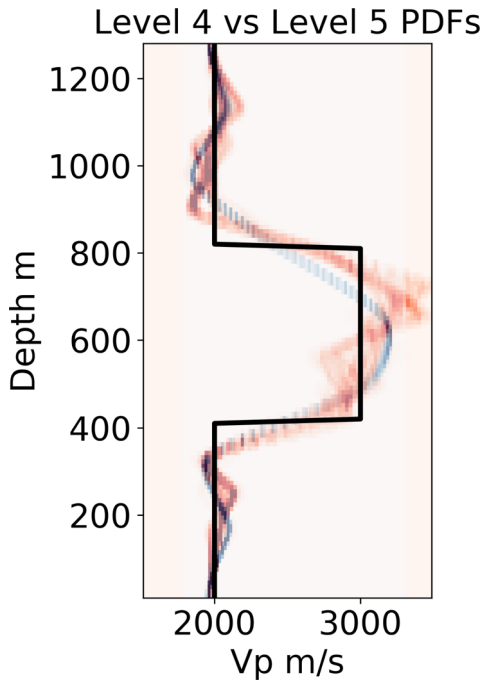

Figure 10. A comparison of posterior uncertainties three-quarters of the total offset away from the origin, at levels 4 and 5 of the DWT. Level 4 uncertainties shown in blue are smoother though they fit the data slightly worse than the wigglier, level 5 uncertainties in red. The level 5 uncertainties are more representative of the limited amount of data coverage available to form the model likelihood.

It would appear that there is no formulaic method of counting the number of arrangements of a modified, restricted tree. For general restricted trees, there is an efficient recursive method to calculate $\mathcal{N}_k$, presented in Hawkins & Sambridge (2015). In this manuscript, we provide a less general, probably easier to implement, efficient recursive pseudo-code for the 2-D wavelet tree structure (Appendix). It can be modified easily for the 1-D or 3-D wavelet trees for the DWT.

3.2.3 *Sampling* $p(\mathbf{m}|\mathbf{d})$

Finally, we state that obtaining the posterior model PDF requires sampling (2) using the Metropolis–Hastings–Green algorithm (Green 1995; Hastie & Green 2012). The criterion to accept or reject a model proposal is given by the probability

$$\alpha(\mathbf{m} \rightarrow \mathbf{m}') = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})}\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}\left(\frac{\mathcal{L}(\mathbf{m}')}{\mathcal{L}(\mathbf{m})}\right)^{\frac{1}{\mathcal{T}}}|\mathbf{J}|\right], \qquad (11)$$

where $q(\mathbf{m}'|\mathbf{m})$ is the proposal probability of stepping from model $\mathbf{m}$ to $\mathbf{m}'$ and $|\mathbf{J}|$ is the determinant of the Jacobian of transformation of variables while changing dimension. It computes to unity for the Birth–Death algorithm (Bodin *et al.* 2009; Sen & Biswas 2017) used in this case. To escape local misfit minima (likelihood maxima), various *interacting* McMC chains are run in parallel at different 'temperatures' $\mathcal{T}$ using the Parallel Tempering algorithm (Swendsen & Wang 1987; Geyer 1991; Dettmer & Dosso 2012; Sambridge 2013; Ray *et al.* 2013). Posterior inference is carried out using the unbiased $\mathcal{T} = 1$ chain. Details of the sampling methodology and model proposals are provided in Appendix.

4 A TRANSMISSION DOMINATED APPLICATION

The model and noisy synthetic data are shown in Fig. 3. 62 receivers were placed on the surface, at a spacing of 20 m, with two sources placed at a depth of 1260 m at the edges of the model. The model is 128 × 128 cells with a grid spacing of 10 m. The source is a Ricker wavelet centred at 5 Hz. Uncorrelated Gaussian noise at 0.5 per cent of the maximum shot amplitude was added to all the traces. The presence of correlated noise for real-world bandpassed time domain data, not considered in our examples, will require the use of a modified likelihood in (6), with off diagonal terms in the data covariance (Malinverno & Briggs 2004).

4.1 Prior specification

A CDF 9/7 basis was chosen for the inversion as it provided a lower χ^2 misfit at level 5 truncation than the Haar basis. (see Fig. 2). Prior bounds for $p(v_i|T, k)$ were set to be bounded uniform following Hawkins & Sambridge (2015). We are careful not to overspecify the bounds—as we explain in this section. Referring to Fig. 4, for a 2-D image, level 1 corresponds to the root node of the tree, with one coefficient numbered 1. Level 2 has three children nodes (of the root) numbered 2–4. From level 2 on, the tree follows a quarternary structure, with each of the nodes having 4 children each. Therefore, level 3 contains the nodes numbered 5–16. Finally, level 4 contains each of the 4 children of all nodes in level 3, numbered 17–64. The minimum and maximum wavelet coefficients of the true model were found at every level, and the bounds for *all* coefficients at this level were set to be 2 per cent less than as well as greater than the extremal values. As with all Bayesian methods, the necessity of prior specification can be viewed as both a blessing and a curse. If one knows absolutely nothing about earth structure and likely velocity variations in the earth, this method will not be of much use, but all geophysical inverse problems require some constraining assumptions to be made and this is not unique a limitation of our approach. However, if we have some idea of what structure could be, we could indeed quantify this interpretive aspect via setting prior bounds in this manner (see also Pasquale & Linde 2017). Example prior model realizations using our method are shown for the second synthetic example. We think that the transform domain provides a very elegant method of specifying compressed sampling bounds, for conceptual geological models (images) in the inverse transform domain. The inverse transform domain is the domain in which we are used to thinking. The nodes can be conveniently represented with a linear tree array numbered in the so called 'Z' or 'Morton' order (Morton 1966) which is equally applicable for 3-D compression. The array index values follow a self-similar Z pattern. Binary interleaving translates the linear indices to their appropriate row and column position in the transform domain image. A word of caution is necessary here—inverse transformed images from wavelet tree models can contain unphysical features such as negative velocities, so it is important to set the prior probabilities of these models to zero. A stochastic approach with a mechanism for navigating difficult topography is a must for the use of this method, as iterative optimization methods may get stuck in the objective function landscape between two zones separated by infinitely high ridges (i.e. zero posterior probability). Our solution has been to use Parallel Tempering for this purpose.

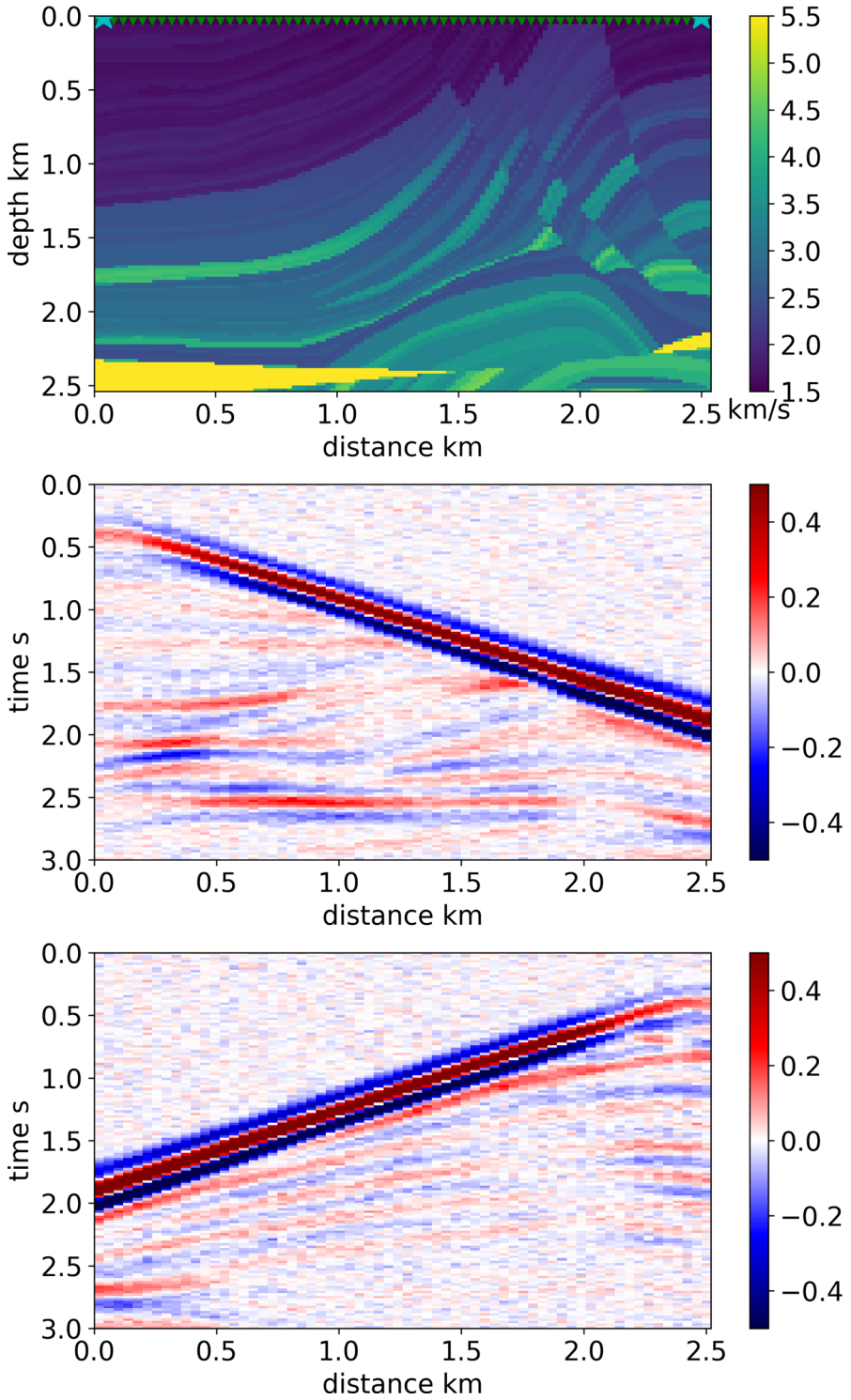

Figure 11. Model and noisy data for a 2-shot surface reflection experiment on a scaled version of the Marmousi model. The sources are represented by the blue stars and the receivers by the green triangles. The direct wave has been removed for clarity of illustration. Besides the free surface, all multiples have been modelled.

4.2 Sampling the posterior velocities

The algorithm very quickly reduces misfit till it reaches RMS (root mean square) misfits close to 2, within just 400 iterations (Fig. 5). The model complexity is also seen to increase as misfit decreases. Since we are using parallel tempering, each Markov chain at a different temperature is represented by a different colour. Posterior inference is carried out only from the chain at $\mathcal{T} = 1$. By construction, parallel tempering ensures that the lower temperature chains always contains the lowest misfits, while higher temperature chains escape less likely (i.e. higher) misfits to escape local misfit minima (Sambridge 2013; Ray *et al.* 2016) as illustrated in Fig. 6. Forty-three parallel, interacting McMC chains were run with log-spaced temperatures between 1 and 2.5 to temper the

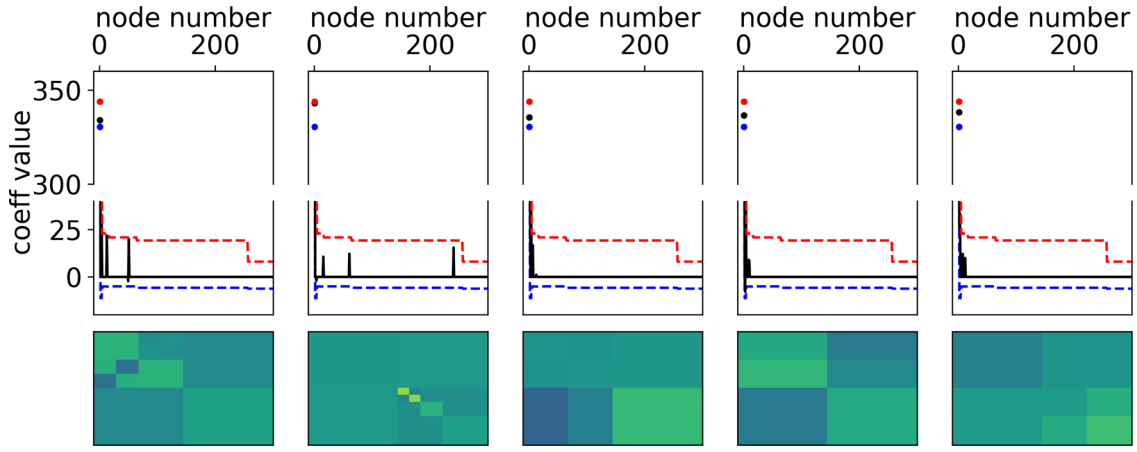

Figure 12. Top: each column corresponds to a 5 node wavelet tree model for the reflection example. Wavelet coefficient values are shown in black, and their prior bounds are shown in red and blue. Note the discontinuity in the y-axis to accommodate the root node value. As mentioned earlier, all nodes at a particular tree level are assigned the same high and low coefficient bounds, obtained from the wavelet transform of some conceptual image. Bottom: the velocity models obtained from inverse wavelet transforming the top row. The velocity colour scale is the same as for Fig. 11.

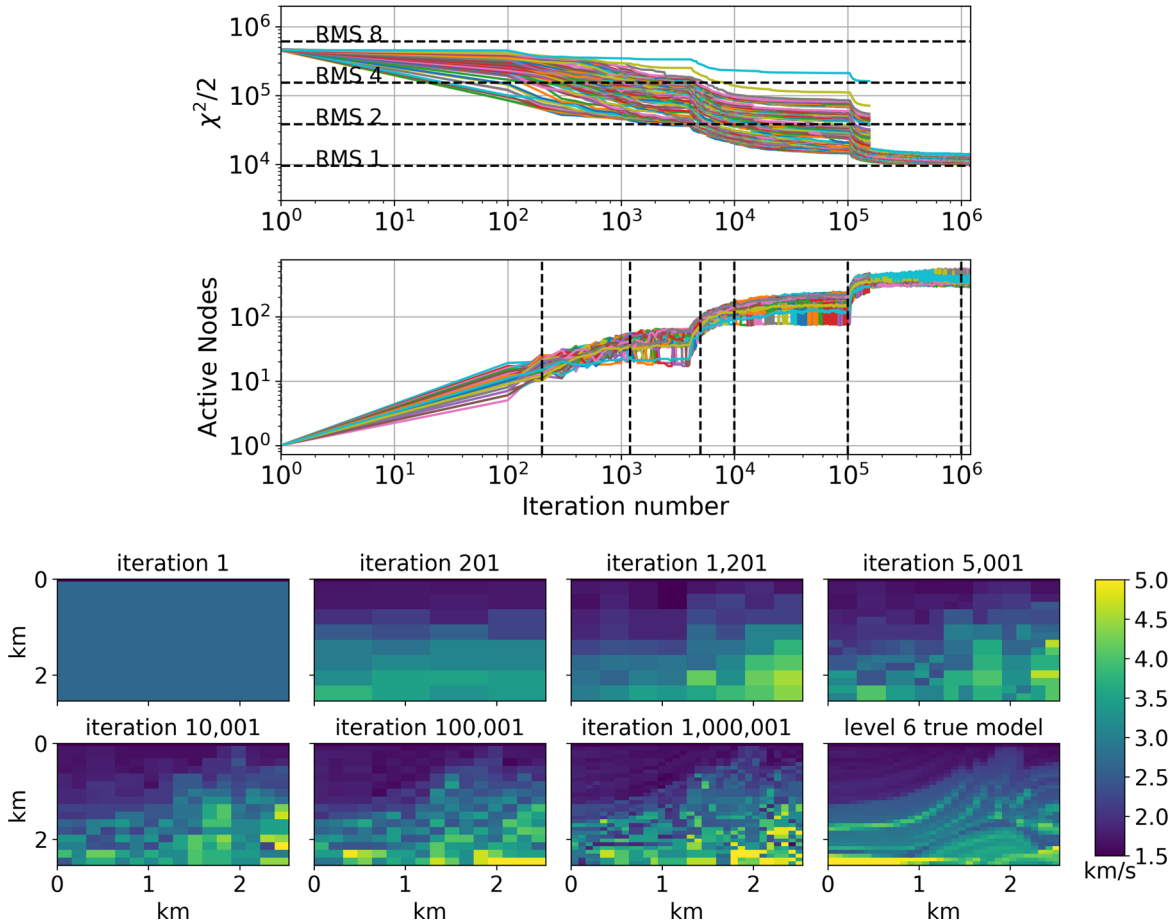

Figure 13. Progress of trans-D McMC sampling with parallel tempering for the reflection experiment. Top: as before, each colour corresponds to an McMC chain at a different temperature, with the lower temperature chains at lower misfits. Rapid model exchanges can be seen, indicative of healthy sampling, by which higher temperature chains help the system escape local minima. Once we were confident that local minima have been escaped, un-necessary chains were discontinued at 200 000 iterations. Bottom: inverse transformed velocity models at particular iterations marked in the top figure. Depending on acceptable levels of misfit, models approximating the background are seen in as few as 201 iterations.

likelihood function in (11). Though good-fitting models were obtained fairly quickly as evidenced from the misfit decrease and models sampled (Figs 5 and 6), to obtain an estimate of model uncertainty we needed to sample longer, and the RMS misfit stayed around 1.18 ($\chi^2/2$ of around 20 000), by most measures a good indication of convergence (Fig. 7). From this figure, we can see that the mean sampled model is quite close to the true velocity model. However, the number of active nodes frequently hits 64, the

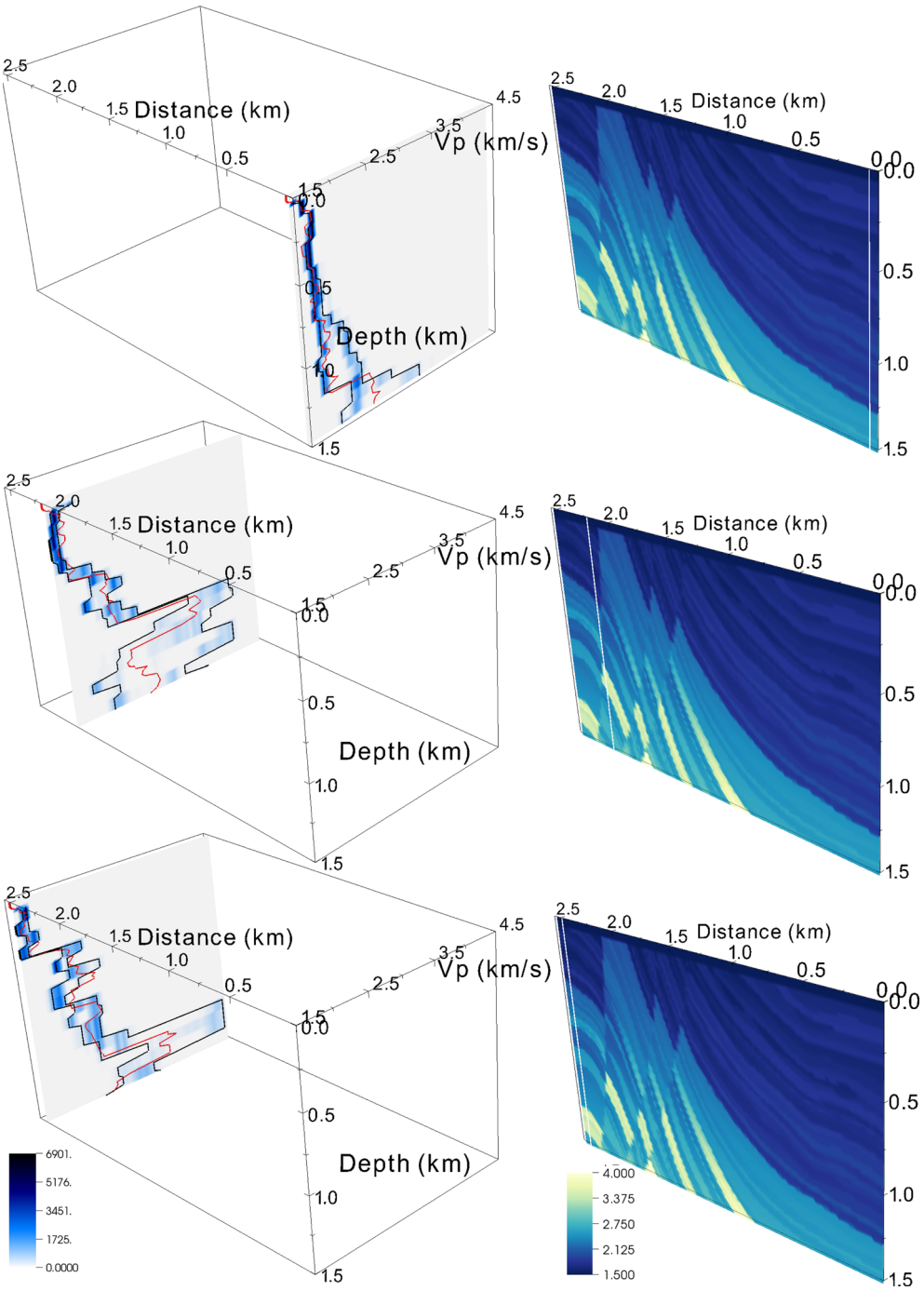

Figure 14. Similar to Fig. 9, marginal distributions of posterior velocity. Left: generally narrow distributions of velocity, particularly shallower, indicate good resolution near the illumination sources. The true model is shown with a red line, and the posterior 5 per cent and 95 per cent quantiles are overlain on the marginal distributions. As before, darker colours correspond to higher probability. Right: locations of the marginal velocity distributions have been overlain on the true velocity model with a white line.

maximum number allowed for a tree depth restricted to level 4. This implies that the data demand a more complicated parametrization.

When we allowed the wavelet tree models to occupy level 5, for a total of 256 possible active nodes, we sample for far longer and arrive at the situation described in Fig. 8. The RMS drops down from 1.18 to 1.004, and the number of coefficients sampled now goes up to 140, though never exceeding 150. We can find the velocity models corresponding to each of these tree models, and instead of computing the mean as we did in Fig. 6, we can now compute the marginal PDFs of velocity at every subsurface location, and display these as 'probability cubes' (e.g. Ray *et al.* 2014) in Fig. 9. The true velocity model is shown in black, coincident with slices through the probability cube, where darker colours at any location are indicative of a high probability of the position and velocity at that point in the cube. The edges of the anomaly seem to be well resolved, with velocities neatly clustered together, but the centre of the anomalous region is not,

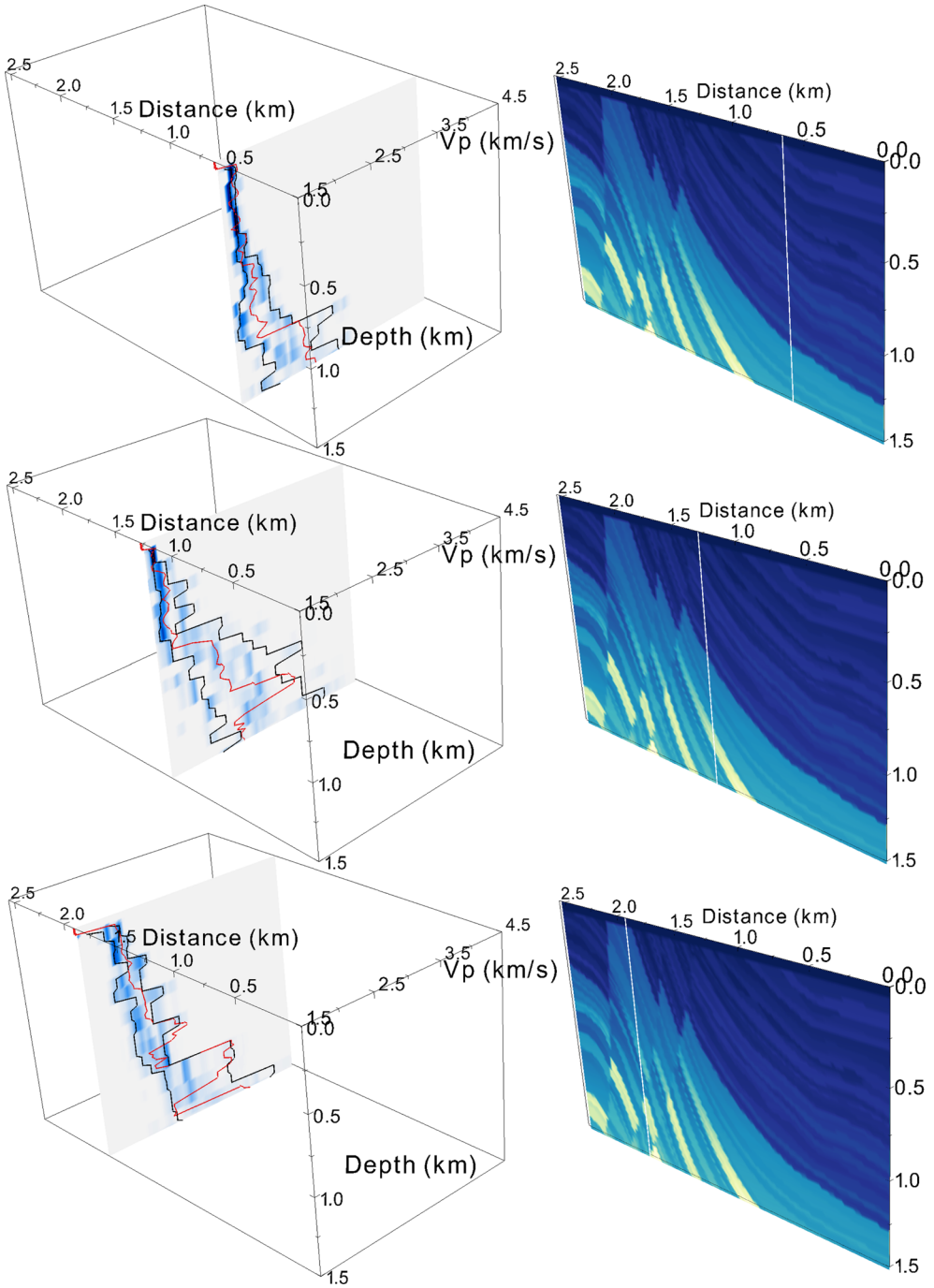

Figure 15. Same as previous figure, at different locations away from the sources. Clearly, the inferred velocities are less certain further away from the illumination sources, though within the 90 per cent credible intervals. It is important to note that the tree models parametrized adaptively to provide this level of resolution.

probably because of the lack of illumination at that point. Also note the multimodality at certain spatial locations, where more than one velocity is possible. Velocity trade-offs are also visible with lower velocities along a propagation path leading to higher velocities at a different location along the propagation path. Had we decided to end our sampling at Level 4, we would have obtained a more optimistic picture of uncertainty, though with slightly worse data fit. A comparison of uncertainty at the two levels is provided in Fig. 10. This figure illustrates again how choices made during inversion affects our conclusions about the subsurface.

With the lessons learned in this example, we proceed to a slightly more complicated, reflection dominated example.

5 A SURFACE REFLECTION APPLICATION

This example is based on a scaled version of the Marmousi model (Versteeg 1994). It is 128 × 128 pixels, with a grid spacing of 20 m. The source wavelet, assumed known, is a Ricker with peak frequency at 3.75 Hz. Two shots were modelled, with uncorrelated

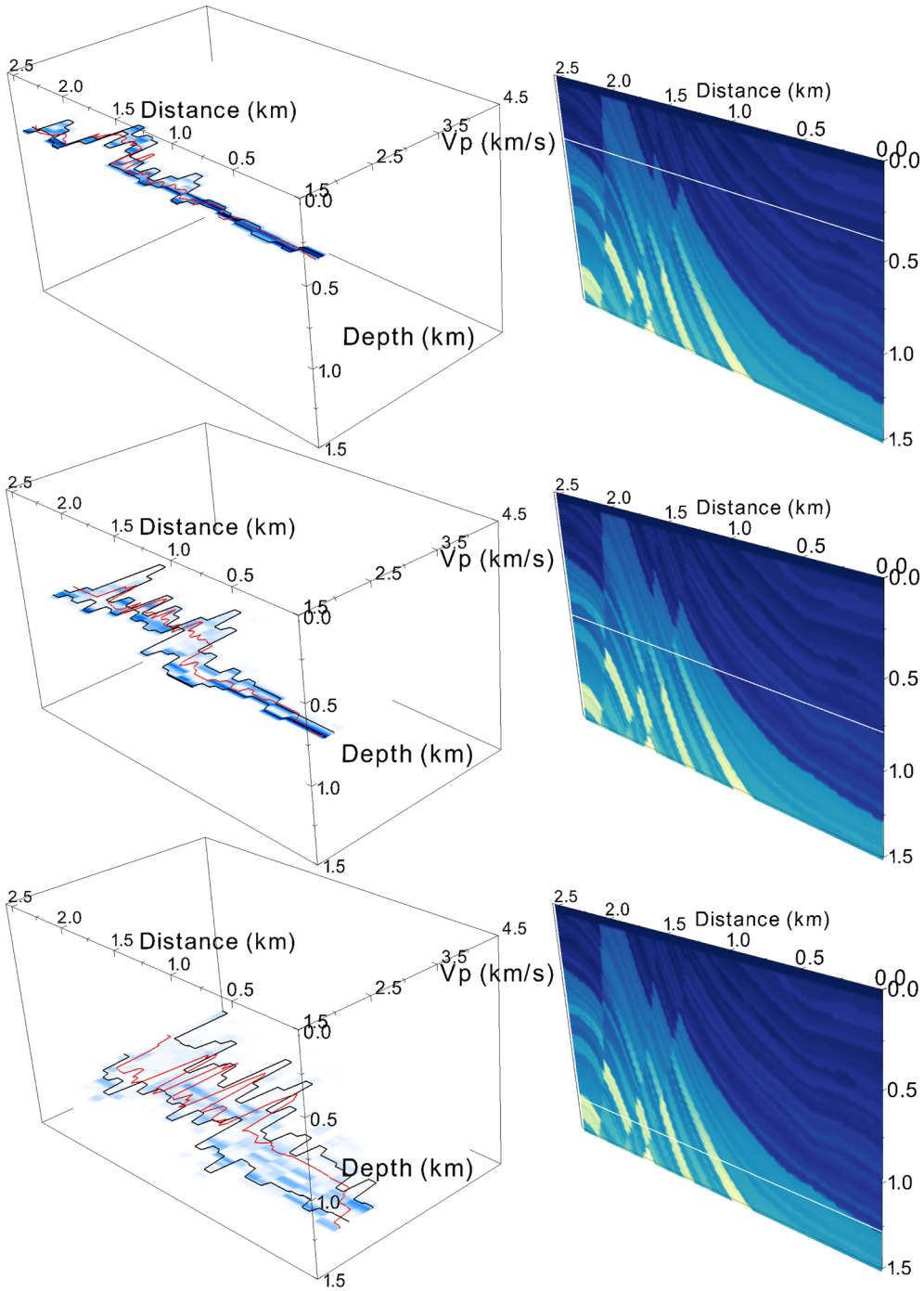

Figure 16. Same as previous figure, but this time showing resolution with depth. We should note from this plot and Figs 15 and 14 that though the marginal posterior velocities were computed from individual models, the model resolution itself is spatially variable and dictated primarily by the physics of the experiment, acquisition geometry and noise.

Gaussian noise at 0.2 per cent of the maximum amplitude added to all traces. The model and noisy data (minus the direct wave) are shown in Fig. 11.

Similar to the previous example, prior bounds were obtained by finding the minimum and maximum DWT coefficients at each level, and going above and below these bounds by 2 per cent of the value. Fig. 12 shows a few 5-node realizations from the prior. We used the Haar basis set for this example, as the smooth CDF 9/7 basis did not work satisfactorily in our trials—we conjecture this was because reflections required sharper edges than the CDF wavelet coefficients at lower levels were able to provide. Bayesian parsimony will not encourage the sampling of more complicated trees if misfit is not substantially reduced by the addition of more active nodes. With the Haar basis, we obtained quick convergence to models resembling the background velocity from within 200 to 10 000 models (RMS 2.3 to 1.44) depending on the notion of an 'acceptable misfit'. We should mention here that a naive implementation of Gauss–Newton with water velocity fixed to the true value and a constant background velocity of 2.6 km s^{-1} was simply not able to provide an update. The progress of sampling and models sampled in the

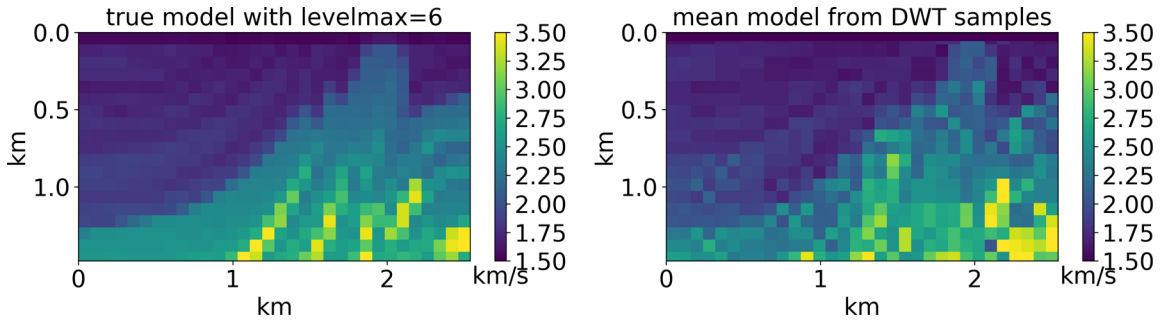

Figure 17. A comparison of the true model at the maximum allowed DWT truncation level and the mean posterior model. While informative, the mean is no substitute for interrogating the posterior ensemble like we have done in Figs 14–16, particularly in the presence of multimodality.

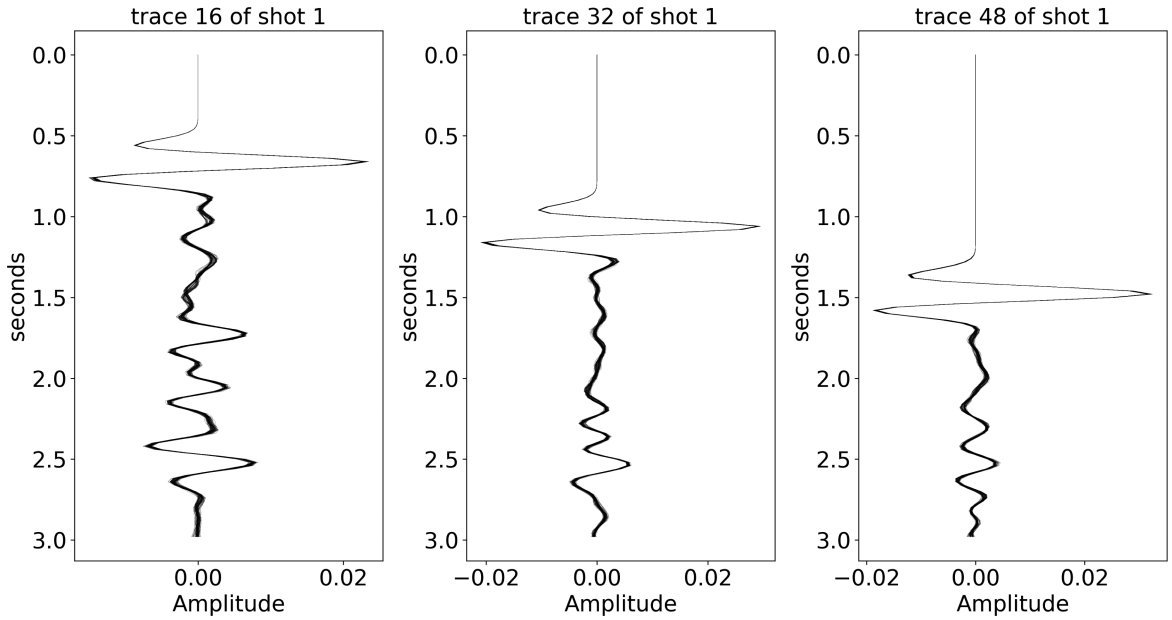

Figure 18. One hundred model responses from randomly selected posterior velocity models, shown at a few receivers, with the direct wave removed. The spread in responses is one of the contributing factors to uncertainty in velocity.

target chain ($\mathcal{T} = 1$) at select iterations is shown in Fig. 13. 80 parallel tempering McMC chains were used initially with log-spaced temperatures between 1 and 5. After 200 000 iterations we were reasonably confident that local minima (likelihood maxima) had been escaped, and only the first 40 chains (temperatures from 1 to 2.213) were used to sample the posterior model PDF. The misfit level asymptoted to RMS 2.0 after 1000 iterations, with the allowed tree depth maximum set to level 4 (64 maximum nodes). After the algorithm was allowed to access level 5 (256 active nodes maximum) the misfit asymptoted again at about an RMS of 1.37, close to the expected value of 1. However, the number of nodes sampled was close to 256, and it was evident that if RMS 1.0 was to be reached, at least the next depth level had to be made available to the algorithm. When level 6 with 1024 maximum nodes was made accessible to the models, an RMS very close to 1 was reached around 200 000 iterations. Sampling was then allowed to go on for another 1 million iterations, and no model required more than 468 active nodes.

For posterior inference, we used only the last 700 000 iterations to obtain samples from a stationary Markov chain unbiased by poorly fitting models. Only the target chain was used for posterior inference. Similar to the previous example, we can create probability cubes with marginal PDFs of velocity at every subsurface location, and the results are shown in Figs 14–16. Again, in the left column, darker colours are representative of higher probability of velocity at a particular point in the cube. The true velocity profile is shown with a red line, and the 90 per cent credible interval at every depth is between the two black lines. The best velocity resolution appears to be near the illumination sources (Fig. 14), getting worse towards the centre (Fig. 15). As expected, resolution is better shallower (Fig. 16). Beyond 1.5 km depth, the PDFs of velocity are too diffuse to provide meaningful information. It is heartening that in most cases, the true velocity lies within the 5 per cent and 95 per cent credible intervals and velocity changes can be inferred when the PDFs of velocity change *en masse* with distance and depth. The picture of resolution which emerges *a posteriori* is consistent with the acquisition setup with two shots at the edges, surface recording and the high levels of noise. We should note that the sampled posterior models parametrize adaptively to provide this picture of resolution—resulting in fine detail only where the data are able to provide it. We can also provide posterior statistics such as the mean (Fig. 17) and quantile velocities at every pixel, but displays of the marginal posterior PDF of velocity (Figs 14–16) with depth are more useful, in our opinion.

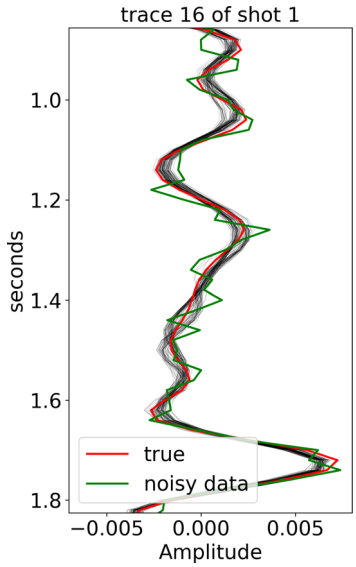

Figure 19. Zooming into one of the traces from Fig. 18, with the noisy data (green) and true model response (red) overlain.

5.1 Examination of data fit and residual statistics

An important check after any inversion is an examination of the data fit and residuals. With real data, correlated residuals are indicative of theory error, an incorrect likelihood function, coherent noise, or some combination of the above. These cannot be always be avoided, but residuals can tell us how much an inversion can be trusted—for example, in Ray *et al.* (2016) it was expected that the residuals would be correlated (due to processing/acquisition artefacts) but Gaussian, and indeed they were. For the synthetic examples in this paper, we added uncorrelated Gaussian random noise and expect that our residuals should therefore also be uncorrelated and Gaussian. For our reflection experiment, we selected 100 random models from the posterior PDF and forward calculated the shot gathers. We have plotted all 100 modelled responses at select traces as shown in Fig. 18. Zooming into one of the traces as shown in Fig. 19, we can see how the added Gaussian noise has allowed for a spread of allowable model responses and hence contributed to uncertainty in inverted velocity.

We can examine the data fit for 100 random posterior models for both shots, as shown in Fig. 20. On the left-hand side is the mean seismic response calculated by summing the posterior model responses for each shot. On the right, is the noisy synthetic data. One can see that the mean of the model responses is remarkably similar to the observed data. All major events and their amplitude versus offset (AVO) characteristics, multiples and refractions have for the most part been well reproduced. The normalized inversion residuals for all time samples, for both shots, for the same 100 random models from the posterior ensemble are shown in Fig. 21. This is further proof that the sampling / inversion is working as intended. We had assumed a Gaussian likelihood, and the sampled models have not overfit the data, producing residuals which when normalized by their standard deviation, approximate an analytic standard normal PDF. We can also compare the mean of the model responses with the true, noiseless synthetic data as shown in Fig. 22.

6 CONCLUSIONS

We have demonstrated with two synthetic examples, the feasibility of carrying out a fully nonlinear, 2-D Bayesian inversion with

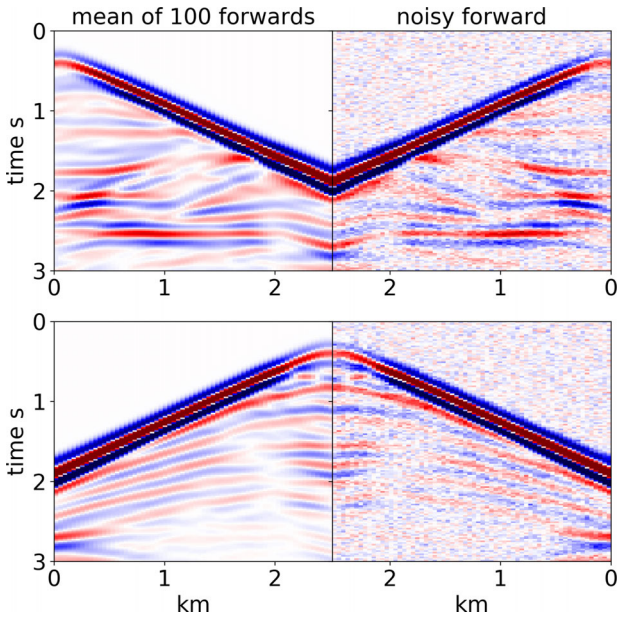

Figure 20. The so called 'butterfly plots' to examine data match *a posteriori*. Left: mean seismic response calculated by summing the posterior model responses for each shot, for one hundred randomly selected posterior velocity models. Right: noisy shot gathers. The AVO characteristics as well as kinematics for both shot gathers are well matched.

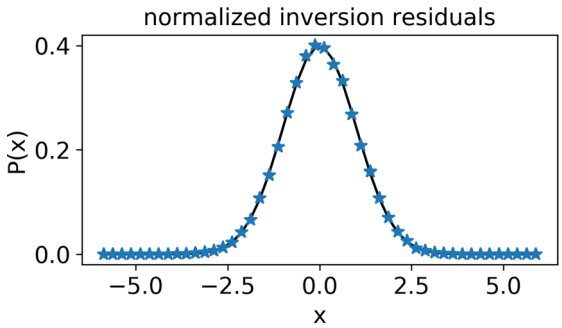

Figure 21. The stars represent the PDF of residuals from each of the hundred models used to construct Fig. 20, normalized by their standard deviations. Comparison is made with a standard normal PDF (solid line), showing that the algorithm has worked as intended.

adaptive model complexity in a tree based framework. There are numerous advantages to doing this, chief among them being an easy to use parametrization which works equally well across 1-D, 2-D and 3-D earth models. Using the tree based parametrization, we easily obtain acceptance rates for birth and death as high as 25 per cent, ensuring good mixing of the McMC, which is very difficult with a Voronoi cell parametrization (Hawkins & Sambridge 2015). Specifying prior coefficient bounds as we have done here, restricts prior models to being within only a certain range of feasible models, while not being an overly restrictive constraint. The use of Parallel Tempering enables us to escape local misfit minima, a major hindrance for reflection based FWI. Finally, the DWT provides an easy means of switching to the model basis most appropriate for solving the current problem. Of course, there is an inherent subjectivity in the use of Bayesian priors (Backus 1988) and different basis functions (Hawkins & Sambridge 2015). However, for practical purposes, almost all geophysical inversion via optimization takes advantage of sensible constraints (e.g. Esser *et al.* 2016). Bayesian inversion methods as demonstrated here and in the recent work of Pasquale

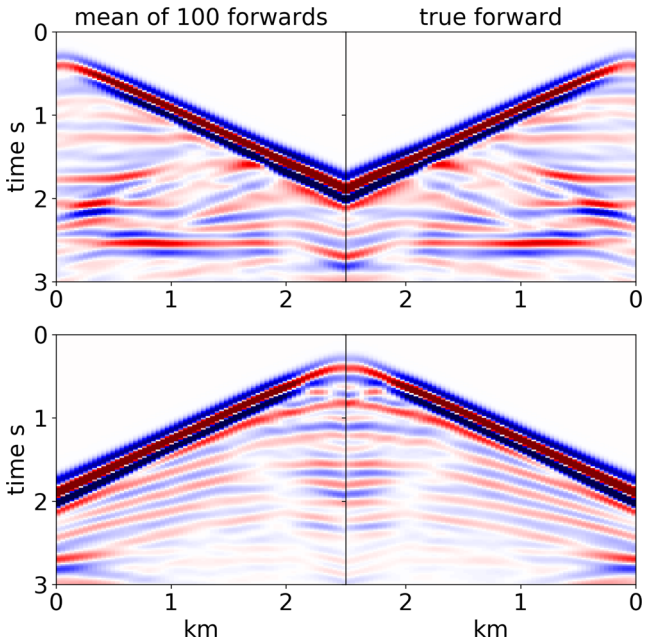

Figure 22. A comparison between the mean forward response for one hundred models with the noiseless synthetic data. Despite a fairly good match, we still have the associated velocity uncertainty shown in Figs 14–16, because of data noise, receiver geometry and physics of the experiment.

& Linde (2017) are naturally able to incorporate multiple structural constraints as prior information. While it is undoubtedly true that a Bayesian appraisal is more time consuming than optimization, fast methods to speed up sampling by an order of magnitude are being researched actively in both the geophysics (Sen & Biswas 2017) and particularly the statistics communities (e.g. Neal 2011; Hoffman & Gelman 2014), coupled with increasingly easy availability of parallel computing from commercial vendors. In this context, our analysis can be extended to higher frequencies and more shots. The fact that a Bayesian inversion of geophysical data provides an uncertainty analysis is invaluable, as it can be a risk mitigation factor for many decisions informed by geophysical data.

ACKNOWLEDGEMENTS

We would like to thank Chevron Energy Technology Company for providing permission to publish this work. We would also like to thank Rhys Hawkins for providing valuable advice on setting up tree based RJ-McMC while maintaining detailed balance. Malcolm Sambridge and Jan Dettmer provided further insight into the workings of sparse parametrizations.

All calculations were carried out using the Julia language (Bezanson *et al.* 2012, 2015), available under the MIT license. We must mention that an invaluable resource while dealing with integer sequences is the Online Encyclopedia of Integer Sequences at https://oeis.org. We would like to thank Alberto Malinverno, Jan Dettmer and an anonymous reviewer for their constructive comments.

REFERENCES

Aravkin, A.Y., Van Leeuwen, T., Burke, J.V. & Herrmann, F.J., 2011. A nonlinear sparsity promoting formulation and algorithm for full waveform inversion, *EAGE Expanded Abstracts,* pp. 23–26, doi:10.3997/2214-4609.20149053.

Backus, G.E., 1988. Bayesian inference in geomagnetism, *Geophys. J. Int.,* **92**(1), 125–142.

Bezanson, J., Karpinski, S., Shah, V.B. & Edelman, A., 2012. Julia: a fast dynamic language for technical computing, arXiv:1209.5145, pp. 1–27.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V.B., 2015. Julia: a fresh approach to numerical computing, *SIAM Ref.,* **59**(1), 1–37.

Biondi, B. & Almonin, A., 2014. Simultaneous inversion of full data bandwidth by tomographic full waveform inversion (TFWI), *Geophysics,* **79**(3), WA129–WA140.

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *Geophys. J. Int.,* **178**(3), 1411–1436.

Bodin, T., Sambridge, M. & Gallagher, K., 2009. A self-parametrizing partition model approach to tomographic inverse problems, *Inverse Probl.,* **25**(5), 55009.

Bodin, T., Sambridge, M., Rawlinson, N. & Arroucau, P., 2012a. Transdimensional tomography with unknown data noise, *Geophys. J. Int.,* **189**(3), 1536–1556.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K. & Rawlinson, N., 2012b. Transdimensional inversion of receiver functions and surface wave dispersion, *J. geophys. Res.,* **117**(B2), doi:10.1029/2011JB008560.

Brown, G.G. & Shubert, B.O., 1984. On random binary trees, *Math. Oper. Res.,* **9**(1), 43–65.

Brunetti, C., Linde, N. & Vrugt, J.A., 2017. Bayesian model selection in hydrogeophysics: application to conceptual subsurface models of the South Oyster Bacterial Transport Site, Virginia, USA, *Adv. Water Res.,* **102,** 127–141.

Burdick, S. & Lekić, V., 2017. Velocity variations and uncertainty from transdimensional*P*-wave tomography of North America, *Geophys. J. Int.*, **209**(2), 1337–1351.

Choi, Y. & Alkhalifah, T., 2016. Waveform inversion with exponential damping using a deconvolution-based objective function, *SEG Technical Program Expanded Abstracts,* pp. 1467–1471, doi:10.1190/segam2016-13818075.1.

Cohen, A., Daubechies, I. & Feauveau, J.-C., 1992. Biorthogonal bases of compactly supported wavelets, *Commun. Pure appl. Math.,* **45**(5), 485–560.

Constable, S.C., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics,* **52**(3), 289–300.

Daubechies, I., 1992. *Ten Lectures on Wavelets,* Society for Industrial and Applied Mathematics.

Dettmer, J. & Dosso, S.E., 2012. Trans-dimensional matched-field geoacoustic inversion with hierarchical error models and interacting Markov chains, *J. acoust. Soc. Am.,* **132**(4), 2239–2250.

Dettmer, J. & Dosso, S.E., 2013. Probabilistic two-dimensional water-column and seabed inversion with self-adapting parameterizations., *J. acoust. Soc. Am.,* **133**(5), 2612–2623.

Dettmer, J., Dosso, S.E. & Osler, J.C., 2010. Bayesian evidence computation for model selection in non-linear geoacoustic inference problems, *J. acoust. Soc. Am.,* **128**(6), 3406–3415.

Dettmer, J., Benavente, R., Cummins, P.R. & Sambridge, M., 2014. Trans-dimensional finite-fault inversion, *Geophys. J. Int.,* **199**(2), 735–751.

Dettmer, J., Dosso, S.E., Bodin, T. & Stipˇ, J., 2015. Direct-seismogram inversion for receiver-side structure with uncertain source–time functions, *Geophys. J. Int.,* **203,** 1373–1387.

Dettmer, J., Hawkins, R., Cummins, P.R., Hossen, J., Sambridge, M., Hino, R. & Inazu, D., 2016. Tsunami source uncertainty estimation: The 2011 Japan tsunami, *J. geophys. Res.,* **121**(6), 4483–4505.

Esser, E., Guasch, L., Leeuwen, T.V., Aravkin, A.Y. & Herrmann, F.J., 2016. Total-variation regularization strategies in full-waveform inversion, arxiv:1608.06159, pp. 1–46.

Fang, Z., Herrmann, F.J. & Silva, C.D., 2014. Fast uncertainty quantification for 2D full-waveform inversion with randomized source subsampling, in *76th EAGE Conference and Exhibition,* Extended abstract, doi:10.3997/2214-4609.20140715.

Fichtner, A. & Trampert, J., 2011. Resolution analysis in full waveform inversion, *Geophys. J. Int.,* **187**(3), 1604–1624.

Fu, L. & Symes, W.W., 2017. An adaptive multiscale algorithm for efficient extended waveform inversion, *Geophysics,* **82**(3), R183–R197.

Fukuda, J. & Johnson, K.M., 2010. Mixed linear-non-linear inversion of crustal deformation data: Bayesian inference of model, weighting and regularization parameters, *Geophys. J. Int.,* **181**(3), 1441–1458.

Galetti, E., Curtis, A., Meles, G.A. & Baptie, B., 2015. Uncertainty loops in travel-time tomography from nonlinear wave physics, *Phys. Rev. Lett.,* **114**(14), 1–5.

Geman, S. & Geman, D., 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Trans. Pattern Anal. Mach. Intell.,* **6**(6), 721–741.

Geyer, C.J., 1991. Markov chain Monte Carlo maximum likelihood, in *Proceedings of the 23rd Symposium on the Interface*, New York, p. 156, American Statistical Association.

Geyer, C.J. & Møller, J., 1994. Simulation procedures and likelihood inference for spatial point processes, *Scand. J. Stat. Theory Appl.,* **21,** 359–373.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika,* **82**(4), 711–732.

Haar, A., 1910. Zur Theorie der orthogonalen Funktionensysteme, *Math. Ann.,* **69**(3), 331–371.

Hastie, D. & Green, P., 2012. Model choice using reversible jump Markov chain Monte Carlo, *Stat. Neerlandica,* **66**(3), 309–338.

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

Hawkins, R., Brodie, R.C. & Sambridge, M., 2017. Trans-dimensional Bayesian inversion of airborne electromagnetic data for 2D conductivity profiles, *Explor. Geophys.,* doi.org/10.1071/EG16139.

Hoffman, M.D. & Gelman, A., 2014. The No-{U}-Turn Sampler: Adaptively Setting Path Lengths in {H}amiltonian {M}onte {C}arlo, *J. Mach. Learn. Res.,* **15**(April), 1593–1623.

Jaynes, E.T., 2003. Probability theory: the logic of science, *Math. Intelligencer,* **27**(2), 83–83.

Jeffreys, H., 1939. *Theory of Probability,* Oxford University Press.

Kass, R.E. & Raftery, A.E., 1995. Bayes Factor, *J. Am. Stat. Assoc.,* **90,** 773–795.

Lever, J., Krzywinski, M. & Altman, N., 2016. Points of Significance: Model selection and overfitting, *Nat. Methods,* **13**(9), 703–704.

Lin, Y., Abubakar, A. & Habashy, T.M., 2012. Seismic full-waveform inversion using truncated wavelet representations, in *SEG Las Vegas 2012 Annual Meeting,* pp. 1–6.

MacKay, D.J.C., 2003. *Information Theory, Inference and Learning Algorithms,* Cambridge Univ. Press.

Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.,* **151**(3), 675–688.

Malinverno, A. & Briggs, V.A., 2004. Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes, *Geophysics,* **69**(4), 1005–1016.

Malinverno, A. & Leaney, S., 2000. A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data, in *SEG Annual Meeting,* no. 3, pp. 2393–2396.

Malinverno, A. & Leaney, W.S., 2005. Monte-Carlo Bayesian look-ahead inversion of walkaway vertical seismic profiles, *Geophys. Prospect.,* **53**(5), 689–703.

Mallat, S.G., 1989. A Theory for Multiresolution Signal Decomposition, *IEEE Trans. Pattern Anal. Mach. Intell.,* **11**(7), 674–693.

Mazzotti, A., Bienati, N., Stucchi, E., Tognarelli, A., Aleardi, M. & Sajeva, A., 2016. Two-grid genetic algorithm full-waveform inversion, *Leading Edge,* **35**(12), 1068–1075.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E. & Virieux, J., 2016. Increasing the robustness and applicability of full-waveform inversion: an optimal transport distance strategy, *Leading Edge,* **35**(12), 1060–1067.

Minsley, B.J., 2011. A trans-dimensional Bayesian Markov chain Monte Carlo algorithm for model assessment using frequency-domain electromagnetic data, *Geophys. J. Int.,* **187**(1), 252–272.

Morton, G.M., 1966. A computer Oriented Geodetic Data Base; and a New Technique in File Sequencing, Tech. rep., IBM Ltd., Ottawa.

Neal, R.M., 2011. MCMC using Hamiltonian dynamics, in *Handbook of Markov Chain Monte Carlo,* pp. 113–162, eds Brooks, S., Gelman, A., Jones, G.L. & Meng, X.-L., Chapman & Hall/CRC.

Pasquale, G.D. & Linde, N., 2017. On structure-based priors in Bayesian geophysical inversion, *Geophys. J. Int.,* **208,** 1342–1358.

Piana Agostinetti, N., Giacomuzzi, G. & Malinverno, a., 2015. Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.,* **201**(3), 1598–1617.

Plattner, A., Maurer, H.R., Vorloeper, J. & Blome, M., 2012. 3-D electrical resistivity tomography using adaptive wavelet parameter grids, *Geophys. J. Int.,* **189**(1), 317–330.

Ray, A., Alumbaugh, D.L., Hoversten, G.M. & Key, K., 2013. Robust and accelerated Bayesian inversion of marine controlled-source electromagnetic data using parallel tempering, *Geophysics,* **78**(6), E271–E280.

Ray, A., Key, K., Bodin, T., Myer, D. & Constable, S., 2014. Bayesian inversion of marine CSEM data from the Scarborough gas field using a transdimensional 2-D parametrization, *Geophys. J. Int.,* **199,** 1847–1860.

Ray, A., Sekar, A., Hoversten, G.M. & Albertin, U., 2016. Frequency domain full waveform elastic inversion of marine seismic data from the Alba field using a Bayesian trans-dimensional algorithm, *Geophys. J. Int.,* **205**(2), 915–937.

Routh, P.S., Behura, J. & Tanis, M., 2016. Introduction to this special section: Full-waveform inversion Part I, *Leading Edge,* **35**(12), 1024–1024.

Sambridge, M., 1999. Geophysical inversion with a neighborhood algorithm - II. Appraising the ensemble, *Geophys, J. Int.,* **138,** 727–746.

Sambridge, M., 2013. A Parallel Tempering algorithm for probabilistic sampling and multimodal optimization, *Geophys. J. Int.,* **196**(1), 357–374.

Sambridge, M. & Faletič, R., 2003. Adaptive whole Earth tomography, *Geochem. Geophys. Geosyst.,* **4**(3), 1–20.

Sambridge, M., Gallagher, K., Jackson, A. & Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.,* **167**(2), 528–542.

Saygin, E. *et al.*, 2016. Imaging architecture of the Jakarta Basin, Indonesia with transdimensional inversion of seismic noise, *Geophys. J. Int.,* **204**(2), 918–931.

Scales, J.A. & Gouveia, W.P., 1998. Bayesian seismic waveform inversion: parameter estimation and uncertainty, **103**(B2), 2759–2779.

Sen, M.K. & Biswas, R., 2017. Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm, *Geophysics,* **82**(3), R119–R134.

Stuart, G.K., Yang, W., Minkoff, S.E. & Pereira, F., 2016. A two-stage Markov chain Monte Carlo method for velocity estimation and uncertainty quantification, in *SEG Technical Program Expanded Abstracts 2016,* pp. 3682–3687.

Swendsen, R.H. & Wang, J.S., 1987. Nonuniversal Critical Dynamics in Monte Carlo Simulations, *Phys. Rev. Lett.,* **58**(2), 86–88.

Tanis, M. & Behura, J., 2017. Introduction to this special section: full-waveform inversion Part II, *Leading Edge,* **36**(1), 58–58.

Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, **49**(8), 1259–1266.

Taubman, D.S. & Marcellin, M.W., 2002. JPEG2000: standard for interactive imaging, *Proc. IEEE,* **90**(8), 1336–1357.

Tikhonov, A.N., 1963. Solution of Incorrectly Formulated Problems and the Regularization Method, *Sov. Math. Dokl.,* **5,** 1035–1038.

Van Leeuwen, T. & Herrmann, F.J., 2015. A penalty method for PDE-constrained optimization in inverse problems, *Inverse Probl.,* **32,** 015007, doi:10.1088/0266-5611/32/1/015007.

Versteeg, R., 1994. The Marmousi experience: Velocity model determination on a synthetic complex data set, *Leading Edge,* **13**(9), 927–936.

Vigh, D., Jiao, K., Cheng, X., Sun, D. & Lewis, W., 2016. Earth-model building from shallow to deep with full-waveform inversion, *Leading Edge,* **35**(12), 1025–1030.

Virieux, J. & Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics, *Geophysics,* **74**(6), WCC1–WCC26.

Warner, M. & Guasch, L., 2016. Adaptive waveform inversion: Theory, *Geophysics,* **81**(6), R429–R445.

Wheelock, B.D. & Parker, R.L., 2013. Exploring L1 model space in search of conductivity bounds for the MT problem, in *AGU Fall Meeting Abstracts*.

Xue, Z. & Zhu, H., 2015. Full waveform inversion with sparsity constraint in seislet domain, in *SEG Expanded Abstracts,* pp. 1382–1387.

APPENDIX: OUTLINE OF THE TRANS-D TREE ALGORITHM

We start the algorithm with a very simple model, typically a tree with only one root node. We then allow the algorithm to iteratively add active nodes to the tree ('birth'), prune them ('death'), or simply modify the coefficient value at an existing active node ('update'). This is all done as the data may demand via the acceptance probability α in (11). This process is repeated until the McMC chain converges to a stationary chain of samples. Details of convergence monitoring for the trans-D inversion and the parallel tempering algorithm used to escape local likelihood maxima (misfit minima) are detailed in Ray *et al.* (2016).

Following the notation of Hawkins & Sambridge (2015), we need to keep track of the set of active nodes S_v, the set of nodes from which to give birth S_b, and the set of active nodes which have no children ('leaves' of the tree) for death S_d. An example tree model with $k = 2$ active nodes and the active, birth and death sets illustrated is shown in Fig. A1.

A1 Construction of the algorithm

At every step of the McMC, one of three possible moves is randomly chosen with equal probability.

A1.1 Update a node coefficient

A node is selected at random from the sets of nodes S_v, and the coefficient value is perturbed using a Gaussian proposal. Typically, we set the standard deviation of the update to be 5 per cent of the width of the uniform bounds at the particular node's depth. This move does not change the model dimension.

A1.2 Birth

The birth move involves the following steps:
If $k < k_{\max}$,

(1) make a copy $\mathbf{m}'$ of the initial tree model $\mathbf{m}$ (i.e. coefficient values and the three node sets);
(2) randomly select node to activate from birth set S_b of initial model;
(3) remove selected node from birth set S'_b of proposed model;
(4) propose coefficient value v' uniformly from the uniform prior coefficient range for the selected node's depth level;
(5) add selected node to active set S'_v of proposed model;
(6) add selected node to death set S'_d of proposed model;
(7) unless parent is the root node, remove selected node's parent from the death set S'_d (if the parent is in the death set), as the parent is no longer a leaf node.
(8) find children of the selected node, add them to the birth set of proposed model S'_b (if the children are within the max tree depth restriction).

This move proposes an increase in dimension, $k' = k + 1$.

A1.3 Death

The death move involves the following steps, and is the reverse of the birth step:
If $k > k_{\min}$,

(1) make a copy $\mathbf{m}'$ of the initial tree model $\mathbf{m}$ (i.e. coefficient values and the three node sets);
(2) randomly select a tree node to remove from death set S_d of start model;
(3) assign zero to the selected node coefficient (simply for completeness);
(4) remove the selected node from the death set S'_d of the proposed model;
(5) find and remove the selected node from the active set S'_v of the proposed model;
(6) find and remove children of the selected node from the birth set S'_b (if children are within the depth restriction)
(7) add the selected node to birth set S'_b of the proposed model;
(8) unless the parent of the selected node is the root, add parent node to the death set S'_d if it is now a leaf node.

This move proposes a decrease in dimension, $k' = k - 1$

A2 Acceptance probability α for the different move types

The probability that the McMC chain moves from a model $\mathbf{m}$ to $\mathbf{m}'$ is given by the acceptance probability (11). For tree based trans-D McMC, it takes different forms for each of the three different move types, and the expressions given below are derived in detail by Hawkins & Sambridge (2015).

For the update move, there is no change in dimension, and when proposing from a uniform prior coefficient range as we have done, it is simply the likelihood ratio:

$$\alpha(\mathbf{m} \rightarrow \mathbf{m}') = \min\left[1, \frac{\mathcal{L}(\mathbf{m}')}{\mathcal{L}(\mathbf{m})}\right]. \tag{A1}$$

For the birth move, the acceptance probability is

$$\alpha(\mathbf{m} \rightarrow \mathbf{m}') = \min\left[1, \frac{p(k+1)}{p(k)} \frac{p(T|k+1,h)}{p(T|k,h)} \frac{\mathcal{L}(\mathbf{m}')}{\mathcal{L}(\mathbf{m})} \frac{|S_b|}{|S'_d|}\right], \tag{A2}$$

where $|S_x|$ is the number of elements in set S_x and h is the maximum depth level restriction. For the death move, the acceptance probability is

$$\alpha(\mathbf{m} \rightarrow \mathbf{m}') = \min\left[1, \frac{p(k-1)}{p(k)} \frac{p(T|k-1,h)}{p(T|k,h)} \frac{\mathcal{L}(\mathbf{m}')}{\mathcal{L}(\mathbf{m})} \frac{|S_d|}{|S'_b|}\right]. \tag{A3}$$

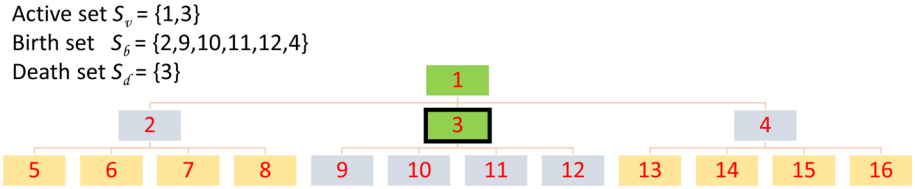

Figure A1. The active (shaded green), birth (shaded blue) and death sets (boxed black) used to propose trans-D moves for an example tree model with $k = 2$ active nodes and a maximum of three allowed levels.

If the prior probability on the number of nodes is uniform then

$$\frac{p(k+1)}{p(k)} = \frac{p(k-1)}{p(k)} = 1. \tag{A4}$$

However, if a Jeffrey's prior has been used as we have done in this work, then

$$\frac{p(k+1)}{p(k)} = \frac{k}{k+1}, \tag{A5}$$

and

$$\frac{p(k-1)}{p(k)} = \frac{k}{k-1}. \tag{A6}$$

The last remaining piece, with pseudo-code to compute $p(T|k)$ or to be more specific, $p(T|k, h)$ for a tree depth restriction h, is provided in the next subsection. If a proposed model is accepted with probability α, it is stored as the next sample. If the proposal is rejected, then the previous model in the McMC chain is retained as the next sample.

A3 Obtaining $p(T|k)$ by counting tree arrangements

The most difficult part, conceptually, of this algorithm is the counting of the number of possible arrangements of a tree given the number of active nodes k in (10), required to calculate α for birth and death proposals in (A2) and (A3).

For a binary tree, if there are n nodes, then for node i, say we can have C_{i-1} arrangements of the nodes preceding it. This leaves C_{n-i} arrangements possible for the remaining nodes. Since the arrangements are independent, the total number of arrangements for node i is $C_{i-1} \cdot C_{n-i}$. But since there are n nodes we have to sum over all i and so the total number of arrangements for n nodes is

$$C_n = \begin{cases} \sum_{i=1}^{n} C_{i-1} C_{n-i}, & \text{if} \quad n \geq 1 \\ 1, & \text{if} \quad n = 0. \end{cases} \tag{A7}$$

For $n = 1$, we set $C_0 = 1$ as there is exactly one way to make a tree with only 1 node. This defines the Catalan number sequence via a recurrence relation, with a base case defining $C_0 = 1$. One can use this logic to construct the number of arrangements of higher order and more general trees as well (Hawkins & Sambridge 2015). (A7) can easily be solved via recursion, but on closer examination we see that to obtain C_3 we need to compute C_2 and C_1. But if we have already computed C_2, we can store this value and re-use it without another recursive call. This is known as *memoization*, a technique extensively used in dynamic programming. This becomes very useful when there are many recursive calls made, as in the case of a pure quaternary tree, where the number of arrangements Y_n can be written thus

$$Y_n = \begin{cases} \sum_{i=1}^{n} Y_{i-1} \sum_{j=1}^{n-i+1} Y_{j-1} \sum_{k=1}^{n-i-j+2} Y_{k-1} \\ \qquad \times Y_{n-i-j-k+2}, & \text{if} \quad n \geq 1 \\ 1, & \text{if} \quad n = 0. \end{cases} \tag{A8}$$

Further, for (A8), in addition to memoizing Y_n we can memoize each of the partial sums over j and k, as the partial sums are functions of the sum upper limit. The modified quaternary tree required for the Cartesian DWT has one root node and three children (Figs 4 and A1), each of these three children follow pure quaternary tree structures. We can write the number of arrangements thus:

$$T_n = \sum_{i=1}^{n} Y_{i-1} \sum_{j=1}^{n-i+1} Y_{j-1} Y_{n-i-j+1}, \tag{A9}$$

taking advantage of the fact that we can again memoize partial sums. Finally, we can treat restricted tree depths with another index representing the depth level restriction. For the case of binary trees, a restriction to a depth h is given by modifying (A7) according to eq. (2.6) of Brown & Shubert (1984),

$$C_{n,h+1} = \sum_{i=1}^{n} C_{i-1,h} C_{n-i,h}, \tag{A10}$$

with

$$C_{n,h} = \begin{cases} 1, & \text{if} \quad n = 0 \quad \text{and} \quad h = 0, \\ 0, & \text{if} \quad n > 0 \quad \text{and} \quad h = 0, \\ 1, & \text{if} \quad n = 0 \quad \text{and} \quad h \geq 0. \end{cases} \tag{A11}$$

We can apply exactly the same restricted binary tree arrangement logic (A10) to the modified restricted quaternary tree arrangement count (A9). All we need to do is modify the numbers of arrangements at any level h by simply making the calculation depend on the previous level $h - 1$.

We will now provide the pseudo-code to calculate the number of arrangements $\mathcal{N}_{n,h}$, thus providing $p(T|n, h) = \frac{1}{\mathcal{N}_{n,h}}$ for n nodes of a tree with a maximum depth h. A memoization structure needs to be initialized with constituent arrays filled with -1 and the base cases given in A11. This is specified in Algorithm 1. -1 simply indicates that the value is yet to be computed. The number of arrangements of restricted depth quaternary trees is provided by Algorithm 2. Finally, the number of arrangements of modified, restricted height quaternary trees for the 2-D DWT is computed by Algorithm 3.

It is important to note that since the Catalan and related number sequences increase very quickly, all arithmetic operations for the tree arrangement calculations were carried out using arbitrary precision. All the array indexing provided in the following algorithms starts at 1. The pseudo-code assumes that all arrays and structures are handled by reference, or to be more specific, arrays, structures and arrays comprising a structure are *mutable*—they change if modified within a function.

Algorithm 1 Memoization structure

Require: A structure called *memo* with arrays of the following dimensions
$\{C4, C4_2, C4_3\} \in \mathbb{Z}^{n_{\max}+1 \times h_{\max}+1}$
$\{T, T_2\} \in \mathbb{Z}^{n_{\max} \times h_{\max}}$

Ensure: Constituent arrays are initialized to -1, with the exception of $C4$ base cases, following the logic in(A11)
$\{C4, C4_2, C4_3, T, T_2\} \leftarrow -1$
$C4[:, 1] \leftarrow 0$
$C4[1, :] \leftarrow 1$

Algorithm 2 Number of arrangements of height restricted quaternary trees

Require: That the structure *memo* has been initialized
function CATALAN4($n, h, memo$)
$C4 \leftarrow memo.C4$
$C4_2 \leftarrow memo.C4_2$
$C4_3 \leftarrow memo.C4_3$
if $C4[n+1, h+1] == -1$ **then**
$sum1 = 0$
for $i = 1 : n$ **do**
if $C4_2[n-i+1, h+1] == -1$ **then**
$sum2 = 0$
for $j = 1 : (n-i+1)$ **do**
if $C4_3[n-i-j+2, h+1] == -1$ **then**
$sum3 \leftarrow 0$
for $k = 1 : (n-i-j+2)$ **do**
$sum3 \leftarrow sum3 + \text{CATALAN4}(k-1, h-1, memo) \times \text{CATALAN4}(n-i-j-k+2, h-1, memo)$
end for
$C4_3[n-i-j+2, h+1] \leftarrow sum3$
end if
$sum2 \leftarrow sum2 + \text{CATALAN4}(j-1, h-1, memo) \times C4_3[n-i-j+2, h+1]$
end for
$C4_2[n-i+1, h+1] \leftarrow sum2$
end if
$sum1 \leftarrow sum1 + \text{CATALAN4}(i-1, h-1, memo) \times C4_2[n-i+1, h+1]$
end for
$C4[n+1, h+1] \leftarrow sum1$
end if
return $C4[n+1, h+1]$
end function

Algorithm 3 Number of arrangements of modified height restricted quaternary trees for the DWT

Require: That the structure *memo* has been initialized, and that **function** CATALAN4 exists
function IMAGETXTREE($n, h, memo$)
$T \leftarrow memo.T$
$T_2 \leftarrow memo.T_2$
if $T[n, h] == -1$ **then**
$sum1 \leftarrow 0$
for $i = 1 : n$ **do**
if $T_2[n-i+1, h] == -1$ **then**
$sum2 \leftarrow 0$
for $j = 1 : (n-i+1)$ **do**
$sum2 \leftarrow sum2 + \text{CATALAN4}(j-1, h-1, memo) \times \text{CATALAN4}(n-i-j+1, h-1, memo)$
end for
$T_2[n-i+1, h] \leftarrow sum2$
end if
$sum1 \leftarrow sum1 + \text{CATALAN4}(i-1, h-1, memo) \times T_2[n-i+1, h]$
end for
$T[n, h] \leftarrow sum1$
end if
return $T[n, h]$
end function