Geophysical Journal International

Geophys. J. Int. (2015) **203,** 972–1000 doi: 10.1093/gji/ggv326
GJI Seismology

Geophysical imaging using trans-dimensional trees

Rhys Hawkins and Malcolm Sambridge

Research School of Earth Sciences, Australian National University, Canberra ACT 2601*, Australia. E-mail: Rhys.Hawkins@anu.edu.au*

Accepted 2015 July 31. Received 2015 July 30; in original form 2015 May 6

SUMMARY

In geophysical inversion, inferences of Earth's properties from sparse data involve a trade-off between model complexity and the spatial resolving power. A recent Markov chain Monte Carlo (McMC) technique formalized by Green, the so-called trans-dimensional samplers, allows us to sample between these trade-offs and to parsimoniously arbitrate between the varying complexity of candidate models. Here we present a novel framework using trans-dimensional sampling over tree structures. This new class of McMC sampler can be applied to 1-D, 2-D and 3-D Cartesian and spherical geometries. In addition, the basis functions used by the algorithm are flexible and can include more advanced parametrizations such as wavelets, both in Cartesian and Spherical geometries, to permit Bayesian multiscale analysis. This new framework offers greater flexibility, performance and efficiency for geophysical imaging problems than previous sampling algorithms. Thereby increasing the range of applications and in particular allowing extension to trans-dimensional imaging in 3-D. Examples are presented of its application to 2-D seismic and 3-D teleseismic tomography including estimation of uncertainty.

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

1 INTRODUCTION

Seismic tomography is the inference of the spatial distribution of properties of the Earth's interior using recorded seismograms. The history of seismic tomography dates back nearly 40 yr to the work of Aki (1977). Rapid increases in computing power coupled with greater availability of data has provided fertile ground for more advanced and effective inversion methods. For recent review articles see Rawlinson & Sambridge (2003) and Rawlinson *et al.* (2014).

In general, inversion methods fall into one of two categories: those that produce a single optimal model given some parametrization and data fit criteria, for example, Thurber (1983), and those that produce an ensemble of models (Mosegaard & Tarantola 1995). In the first category, model estimation methods make use of optimization techniques to minimize a combination of error norms. In the second approach, an ensemble of trial models are generated, often using probabilistic sampling methods such as Markov chain Monte Carlo (McMC; Gamerman & Lopes 2006; Brooks *et al.* 2011). From the ensemble, statistical inferences can be made and representative single models extracted such as the ensemble mean, mode, or some transform of the model parameters. Mosegaard & Tarantola (1995) provides a comprehensive overview of this type of inversion.

In optimization for a single regularized solution to an inverse problem, a number of parameters need to be tuned. Often these involve choices to be made, for example, deciding on an underlying maximal grid resolution as well as damping and smoothing parameters to prevent over fitting of the data. These choices are subjective and while there are criteria for choosing the smoothing and damping parameters such as the L-curve test (Hansen 1992), these are not without their problems (Hanke 1996; Vogel 1996). Additionally, smoothing and damping operations are often imposed through a globally tuned parameter whose choice is a compromise across the whole domain. This is particularly problematic in regional or global seismic tomography where there is often highly uneven coverage of a region of interest by seismic waves. In addition, uncertainty estimates based on a regularized single realization of the inverse problem, necessarily reflect the form of damping and smoothing imposed and can be overly optimistic.

Recently, computing power has advanced sufficiently to allow Monte Carlo sampling approaches to be applied to seismic tomography problems. With the introduction of Birth/Death Monte Carlo (Geyer & Moller 1994) and the more general Reversible Jump McMC (Green 1995; Denison *et al.* 2002), these sampling schemes are able to make trans-dimensional steps between different model parametrizations. This allows both the characteristic, for example, basis function, and the number of unknowns to vary from step to step of the inversion algorithm.

This is a flexible approach that results in a 'parsimonious' solution to problems in that the model complexity is driven by the data without the use of parameters that require tuning. These trans-dimensional samplers, introduced to the geophysics community by Malinverno (2002), have been successfully applied to a number of geophysical inverse problems (Sambridge *et al.* 2006;

Hopcroft *et al.* 2007; Bodin & Sambridge 2009; Piana Agostinetti & Malinverno 2010; Minsley 2011; Dettmer *et al.* 2012; Iaffaldano *et al.* 2014; Piana Agostinetti *et al.* 2015).

A common class of parametrization used in the trans-dimensional solution of spatial problems are Voronoi cells built from a set of nuclei (Okabe *et al.* 1992). By specifying the location of the nuclei as well as the value (or values) of Earth properties within each cell, a mobile Voronoi model can be used to represent Earth properties spatially in 2-D (Bodin *et al.* 2012). In the first 3-D application we are aware of, Piana Agostinetti *et al.* (2015) have recently extended the Voronoi cell approach to local earthquake tomography problems. These Voronoi cell parametrizations are grid free and locally adapt to regions of increased heterogeneity tempered by the resolving power of the data. Although the application of the trans-dimensional Voronoi cell method is now well established for seismic imaging, there are a number of short comings that hinder its application as the number of data and complexity of the Earth model increases.

In ray-based seismic tomography, numerical integration along ray paths requires the evaluation of the model at hundreds to thousands of spatial points per observation. For each point along the ray, we need to determine from the Voronoi cell parametrization the Earth properties, for example, seismic wave speed, and this involves determining in which cell the point resides. A naive algorithm would simply find the nearest Voronoi nuclei by computing the distance to every nuclei of the model and this results in an $\mathcal{O}(n)$ operation, where n is the number of Voronoi cells (Sambridge & Gudmundsson 1998). In 2-D problems, we can use a Delaunay triangulation to speed up the cell look up operation to an $\mathcal{O}(\log n)$ operation. Even with fast algorithms for incrementally maintaining the Delaunay triangulation (Lawson 1977), the accounting cost of maintaining the triangulation can be prohibitive for large problems.

A second feature of the Voronoi cell approach is that they do not lend themselves well to representing a continuous field. In a Voronoi cell parametrization, the Earth properties within each cell are often represented with constant values, although in principle, any order polynomial is possible. This means that each Earth model consists of an irregular polygonal mesh with discontinuities, both in the function and in its derivatives, at the interfaces between cells. Typically, any single Earth model in the ensemble is rather crude and implausible and it is only by averaging over many such crude representations that it is possible to generate a continuous field. This means that the Voronoi cell approach must utilize multiple independent Markov chains or very large numbers of samples in a single chain in order to produce a continuous field through averaging.

Use of Voronoi cells in 3-D imaging has two additional complications. The first is that there is no analogue of fast 2-D incremental Delaunay calculation algorithms and so Voronoi cells must be determined from 'scratch' each time the mesh is updated, further adding to the computational burden. The second is that the shape of Voronoi cells in 3-D is particularly sensitive to the choice of spatial scaling between lateral and radial directions. For example, Voronoi cells built around nuclei at depth can easily protrude to the surface.

In this paper, we introduce a new class of parametrization for trans-dimensional imaging problems which overcomes the limitations of Voronoi cells while providing a general efficient framework for dealing with 1-D, 2-D and 3-D problems in Cartesian or spherical geometries. The new framework allows a great deal of flexibility in terms of the choice of basis functions, including multiscale parametrizations such as wavelets and subdivision surfaces. We show that with our new algorithm, we are able to make larger scale 3-D tomographic problems practical using trans-dimensional sampling for velocity model estimation with uncertainties.

2 REPRESENTATION OF GEOPHYSICAL IMAGES WITH TREES

Before introducing the trans-dimensional framework for sampling over trees, we show how the concept of trees can be used to represent a tomographic Earth model. There are many examples of using hierarchical or multiresolution analyses of images in 2-D, for example, the Laplacian pyramid (Burt & Adelson 1983) and the wavelet transform (Mallat 1989). Broadly speaking, within each of these schemes an image is subsampled to obtain a coarser but more compact representation. Error terms are computed representing the difference between the subsampled and true image so that with a combination of the subsampled image and error terms, we can accurately reconstruct the original image. This process can be repeated recursively on each subsampled image until the result is a single pixel, representing the mean of the image, and a hierarchical set of error terms for each resolution scale. It is a property of continuous tone images that individual pixels are often highly correlated with their neighbours, and as a result, many of the error terms are near zero. For this reason, such multiresolution image analysis techniques have been used for image compression, for example, the JPEG 2000 image compression standard (Unser & Blu 2003).

This hierarchy of a single mean value of an image through successive levels of perturbative terms can naturally be represented by a tree structure. Fig. 1(a) shows how a quaternary tree in which each node has 4 child branches, spans from the single pixel representation of an average value of a field, through successive levels of local perturbation terms. In this example, each node of the tree contains a single parameter value. At the root of the tree, the highest level in Fig. 1(a), this value represents the mean of the image and all other descendant nodes represent a local deviation from that mean. In this way, each tree level creates an image with a corresponding spatial resolution and each child node adds detail by perturbing the previous level at a finer spatial resolution. From this multiresolution tree representation, we can construct arbitrary 2-D images which can be used to represent, for example, the seismic wave speed or slowness of a region of the Earth. This same principle, of spanning the subdivision of grid with a tree, applies equally to 1-D, 2-D, 3-D Cartesian geometries and equally to non-Cartesian geometries such as spherical geometry (Samet 2006).

It is important to point out here that the tree needn't completely span the underlying 2-D grid as shown in the Fig. 1(a). An incomplete tree is shown in Fig. 1(b). The 2-D image from this tree is constructed in the same way as the full tree with the parameter values of 0 at the missing nodes of the tree. This has the potential to compress the model space, or the number of model parameters, by locally adapting to structure or data coverage.

The use of adaptive mesh refinement has been used previously in geophysical inversion, for example, Sambridge & Faletic (2003), where a criterion based on the maximum spatial gradients in seismic velocity perturbation was used to iteratively subdivide a tetrahedral grid during the inversion. A similar approach was presented by Plattner *et al.* (2012) for electrical resistivity tomography where a multiscale wavelet parametrization was adaptively refined through optimization.

Rather than a fixed criterion, we propose to sample over such subdivision refinement choices to obtain posterior information on where our data requires finer scale features. By itself, recasting geophysical inverse problems within a tree structure offers little advantage, but as we shall see it is highly suited to coupling with a trans-dimensional algorithm within a fully Bayesian framework.

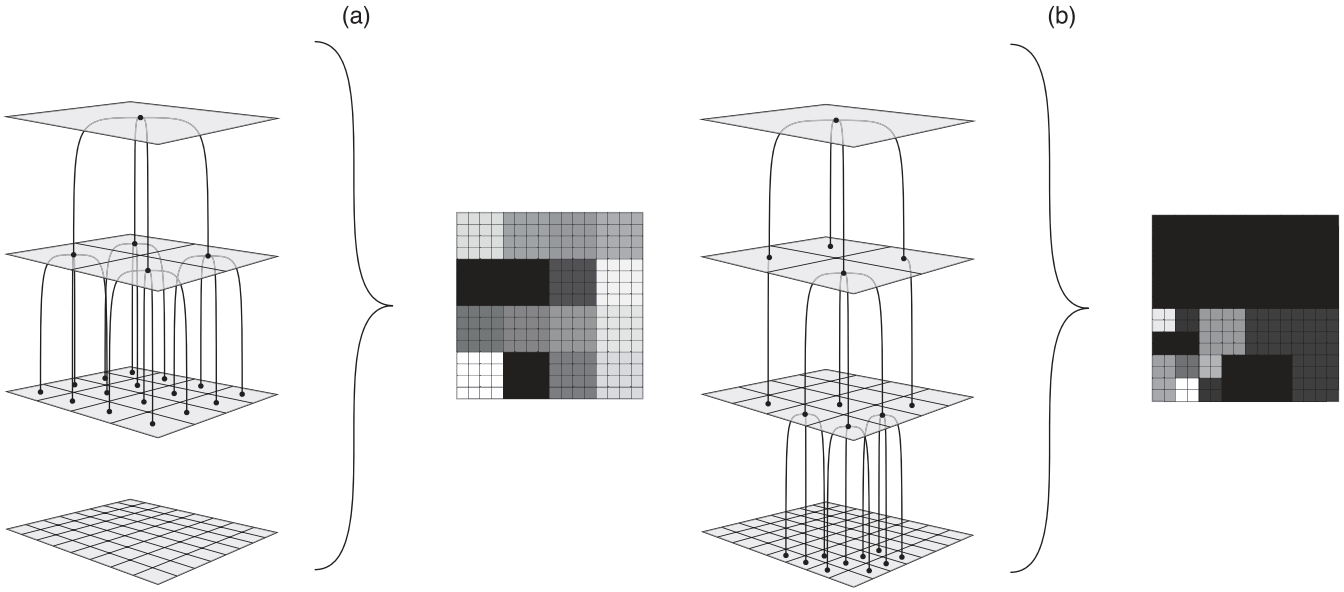

Figure 1. The Laplacian Pyramid subdivision showing how a quaternary tree can span from the coarsest resolution to the finest error terms. At the top level, we have a single pixel representation of a 2-D domain, the root node of the tree, which is subdivided into four subpixels at the next level and so on. In (a), we show the complete tree structure to the third level. In (b), we show how an incomplete quaternary tree can still be used to parametrize a 2-D Earth model and how this can locally adapt to regions of localized heterogeneity. In both (a) and (b), the two models have the same number of parameters but represent very different structure.

3 A GENERAL BAYESIAN TRANS-DIMENSIONAL FRAMEWORK FOR TREES

In a Bayesian approach to inference, the solution we obtain is a numerical estimate of the *a posteriori* probability distribution or posterior (see Gelman *et al.* 2004 for a general overview and Mosegaard & Tarantola 1995, Sambridge & Mosegaard 2002 for an overview of Bayesian inference in a geophysical context). This is the probability density of the model space given the observed data, or written mathematically, $p(\theta|\mathbf{d})$, where θ is our vector of model parameters and $\mathbf{d}$ our vector of observations. In all but the simplest of problems, this probability density function is approximated numerically using McMC techniques and Bayes theorem (Bayes 1763), that is,

$$p(\theta|\mathbf{d}) = \frac{p(\theta)p(\mathbf{d}|\theta)}{p(\mathbf{d})}. \tag{1}$$

This states that the posterior probability density, $p(\theta|\mathbf{d})$, is equal to the prior probability distribution, $p(\theta)$, times the likelihood $p(\mathbf{d}|\theta)$, which we will abbreviate to $\mathcal{L}(\theta)$, normalized by the evidence, $p(\mathbf{d})$. An McMC sampling approach can be applied to the numerator of the right-hand side of eq. (1) to obtain an estimate of the posterior probability distribution up to the normalizing constant of the evidence, which is often difficult to compute explicitly (Sambridge *et al.* 2006).

An McMC sampler requires the specification of the prior probability distribution, which represents *a priori* information we may have on the distribution, or plausible range, of the model parameters, and the likelihood which is a probabilistic measure of the fit of the model to the data. An McMC sampler operates by starting from some model at step i of θ_i, then creating a new proposed model θ_i' using a proposal in the form of a reversible probability density function $q(\theta_i'|\theta_i)$. The new model is accepted, that is, $\theta_{i+1} = \theta_i'$ or rejected, that is, $\theta_{i+1} = \theta_i$, based on an acceptance probability, commonly the Metropolis–Hastings acceptance criterion (Metropolis *et al.* 1953; Hastings 1970)

$$\alpha(\theta', \theta) = \min\left\{1, \frac{p(\theta')}{p(\theta)}\frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta)}\frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}\right\}. \tag{2}$$

This can be read as the prior ratio times the likelihood ratio times the proposal ratio. We are not limited to a single proposal probability density function at every step, it is perfectly feasible to select randomly from a set of proposal distributions. The Metropolis–Hastings criteria satisfies the mathematical condition known as 'detailed balance' (Gamerman & Lopes 2006) which allows the Markov chain to converge and correctly sample the target posterior distribution.

It is common practice to remove some initial number of steps from the final ensemble which are believed to be pre-converged or 'burn-in' samples. In most cases, the fact that we only obtain the posterior probability distribution up to a normalizing constant is not a problem as relative inferences are generally sufficient.

An extension to McMC samplers is the Birth/Death scheme of Geyer & Moller (1994), generalized to the trans-dimensional framework developed by Green (1995). In trans-dimensional samplers, a proposal distribution is allowed to change the parametrization of the model and dimension, that is, the size of the vector θ of model parameters. A key benefit of allowing the sampling to jump between dimensions is that the data dictates the model complexity resulting in a parsimonious result (Malinverno 2002; Sambridge *et al.* 2006). Additionally, we can obtain a posterior probability distribution on the number of model parameters required by the data given the noise rather than fixing this *a priori*.

The generalization of the Metropolis–Hastings acceptance criteria to support trans-dimensional steps is

$$\alpha(\theta', \theta) = \min\left\{1, \frac{p(\theta')}{p(\theta)}\frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta)}\frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)} \mid \mathcal{J} \mid\right\}, \tag{3}$$

where the additional term from eq. (2), $|\mathcal{J}|$, is determinant of the Jacobian that maintains detailed balance through variable

transformations resulting from trans-dimensional steps. Expression (3) may also be used if the dimension is unchanged, but the proposal involves a step from one class of parametrization to another.

The complexity of the models generated from trans-dimensional samplers is dependent on the level of noise applied, that is, in general, the lower the noise, the higher the complexity. For this reason, in the case where the noise on the data us unknown, it is advantageous to use a hierarchical Bayesian step which allows noise parameters to be inverted for as part of the sampling of model parameters as shown by Bodin *et al.* (2012).

A birth/death trans-dimensional sampler will consist of three classes of proposal, a birth proposal where the model vector θ will increase in size, a death proposal where some model parameters are removed, and a value proposal where the model vector remains the same size, but one or more values will be changed (i.e. the normal class of proposal in fixed dimension McMC samplers).

Our aim here is to apply the trans-dimensional framework to the McMC sampling of tree structures that we can use to represent geophysical models of the Earths internal structure. In this framework, a birth proposal will consist of adding one or more new nodes to the tree, a death proposal will consist of removing one or more nodes from the tree, and a value proposal will perturb one or more values located within the existing tree. To our knowledge no general treatment of trans-dimensional sampling over tree structures has previously been presented. The only work we are aware of is Denison *et al.* (1998) which is limited to binary classification trees. Here we apply the trans-dimensional formalism of Green (1995) to general trees with known structure.

We define a 'general' tree as one in which the maximum number of child nodes, of any node, is fixed. With this restriction a prior can be computed. We have yet to encounter a situation where this restriction limits the application of this new framework. In general, the structure of the tree will be restricted by the geometry of the physical application. For example, in the 2-D image example earlier, each pixel is subdivided into four subpixels and this is the upper limit on the number of child nodes. For a 3-D volume, each voxel will subdivide into eight subvoxels which gives an upper limit on the number of child nodes of 8.

In the following subsections, we describe the components of the acceptance criteria and introduce the full general expressions for each type of model perturbation.

3.1 The model

In the earlier 2-D example, the tree structure 'template' consists of the complete spanning quaternary tree and two possible tree models conforming to this template are shown in Fig. 1. A simpler example of such a tree model in a binary tree template appears in Fig. 2 where the template is shown in outline and an example tree model, consisting of active nodes and value(s) at each node, is shown in solid shading.

Each active node in the tree model has one or more associated values, so given a number of nodes, k, our model space vector would be

$$\theta = \langle \mathcal{T}_k, \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle , \tag{4}$$

where $\mathcal{T}_k$ represents the arrangement of the k nodes within the template tree structure and $\mathbf{v}$ is the vector of parameters at each node (which may be a single parameter). If we have a unique counter for each tree node, we can represent $\mathcal{T}_k$ as a set of indices, that is, $\mathcal{T}_k = \langle t_1, \ldots, t_k \rangle$.

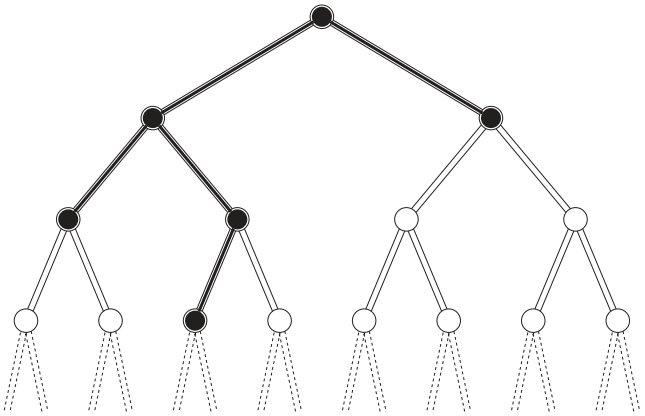

Figure 2. The first four levels of a binary tree template shown as outline with an individual tree model highlighted with solid lines.

3.2 The prior

Given the parametrization in eq. (4), we can write the prior on the model in general terms as a product of conditional probability distribution functions (PDFs),

$$p(\theta) = \prod_{i=1}^{k} p(\mathbf{v}_i \mid \mathcal{T}_k, k) p(\mathcal{T}_k|k) p(k). \tag{5}$$

Stated simply, the prior is a combination of the probability on the number of nodes in the tree, the probability of the arrangement of the tree within its template and the parameter values at each of the nodes. This prior specification reasonably assumes that each term is independent which results in a separable prior PDF.

3.2.1 *Prior on the number of nodes*

The prior for the number of nodes is a choice that will be dependent on how the model is mapped from the tree structure. Here, we leave the prior as a general expression, $p(k)$, but highlight two common choices. First, a uniform prior

$$p(k) = \frac{1}{k_{\max} - k_{\min} + 1}, \tag{6}$$

where $k_{\max}$ and $k_{\min}$ (usually 1) are chosen as the upper and lower bounds on the number of nodes. An alternative is to use a Jeffreys' prior (Jeffreys 1939; Jaynes 2003), that is,

$$p(k) \propto \begin{cases} \frac{1}{k} & k > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{7}$$

This prior is improper because the limit of the integral of $p(k)$ is unbounded as k goes to infinity. Nevertheless, a useful feature is that there is no imposed restriction on the number of nodes unlike in the uniform prior case (see page 238 of Jeffreys 1939). In experiments to be described here, the posterior PDF of k with either prior is similar, which shows that it is primarily the data which constrains the dimension of the model.

3.2.2 *Prior on homogeneous unrestricted trees*

The prior on the arrangement of the nodes within the tree template, $p(\mathcal{T}_k|k)$, is the most complicated component of this algorithm and is derived here for our general tree parametrization. The prior we have used is a uniform prior on the structure of the tree, by this we mean that given a number of nodes, k, any arrangement of the nodes into a

valid tree within its template has equal probability to any other. This is the least informative prior on a tree structure and also the most tractable to compute for the acceptance criteria. The consequences of this prior are that a model that has an even distribution of detail across the region is equally as likely as a model that has localized fine detail. This is illustrated in Fig. 1 where both models shown have the same number of active tree nodes. In this prior, both of these models are equally likely.

This reduces computing the prior on the structure of the tree into a problem of computing the number of valid tree arrangements possible given a tree structure template and the number of active nodes, that is,

$$p(\mathcal{T}_k \mid k) = \frac{1}{\mathcal{N}_k}, \tag{8}$$

where $\mathcal{N}_k$ is the number of valid trees with k nodes. To evaluate $\mathcal{N}_k$, we first consider unrestricted homogeneous trees, which we define as those where each node has the same upper limit on the number of child nodes. Binary and quaternary trees fall into this class. By unrestricted we mean that there are no other constraints on the structure of the tree such as a maximum height and therefore that the tree can grow infinitely large. For this class of trees, there are analytical expressions for computing the number of arrangements, $\mathcal{N}_k$. For binary trees, it is known that the number of arrangements follows the sequence of Catalan numbers (Catalan 1844; Hilton & Pedersen 1991; Knuth 2004), that is,

$$\mathcal{N}_k = \frac{1}{k+1}\binom{2k}{k}, \tag{9}$$

where $\binom{2k}{k}$ is the standard binomial coefficient. This result has been generalized by Aval (2008) to trees with a maximum number of n children

$$\mathcal{N}_k = \frac{1}{(n-1)k+1}\binom{nk}{k}. \tag{10}$$

When $n = 2$, this reduces to eq. (9). This expression allows closed form expressions for the prior for homogeneous unrestricted trees. However, this only represents a small subclass of possible trees and we need to extend this further.

3.2.3 *Restricted and heterogeneous trees and their priors*

The first restriction on a tree template is an upper limit on height. As seen in the earlier 2-D example, the height of the tree corresponds to the level of subdivision of the region. As such, a restriction on the height of the tree imposes a strict upper limit on the minimum resolution scale of the model. In addition, it also constrains the computational complexity of the problem as we no longer need to deal with arbitrarily large trees.

A second variant to be considered is a heterogeneous tree which contains nodes with varying upper limits on the number of child nodes. In later examples, where we use wavelet parametrizations in 2-D, we will make use of heterogeneous trees where the root of the tree has three possible child nodes, and all subsequent nodes have four possible offspring. Analytic expressions for the number of arrangements of a tree given the number of nodes are only known for trees where each node has the same maximum number of possible child nodes. For both the restricted height and heterogeneous trees, we need to calculate number of arrangements given k.

The Catalan sequence for binary trees can be derived from a recurrence relationship using generating functions (see Eq. 2.5.10 of Wilf 1990). The general solution to both these problems is to compute the number of arrangements from a recurrence relationship. Starting from the recurrence relationship for binary trees,

$$\mathcal{N}_k = \begin{cases} 1 & k \leq 0 \\ \sum_{i=0}^{k-1} \mathcal{N}_i \mathcal{N}_{k-i-1} & \text{otherwise} \end{cases}, \tag{11}$$

where k is the number of tree nodes, we recognize that we have a simple integer partitioning problem (Stanley 1997) and the modification of eq. (11) from a binary tree to a ternary tree requires the addition of a third partitioning of the k nodes among the three child branches. To include restrictions on the height of the tree, we simply add the relevant terminating conditions, for example, rewriting the equation

$$\mathcal{N}_{k,h} = \begin{cases} 0 & h = 0 \\ 1 & k \leq 0 \\ \sum_{i=0}^{k-1} \mathcal{N}_{i,h-1} \mathcal{N}_{k-i-1,h-1} & \text{otherwise} \end{cases}. \tag{12}$$

Further details of the recurrence relationships and our algorithm for computing them in an efficient fashion is outlined in Appendix A. From here on, we assume that $\mathcal{N}_{k,h}$ is known from a recurrence relationship like eq. (12) and that the prior on the structure of the tree can be calculated as the inverse of the number of arrangements of trees given a number of nodes, that is,

$$p(\mathcal{T}_k \mid k, h) = \frac{1}{\mathcal{N}_{k,h}}, \tag{13}$$

where h is a maximum height restriction.

3.2.4 *Prior on each parameter value*

The prior on the Earth model parameters at each node of the tree will depend the particular basis functions used. Again this prior is often a choice and we briefly mention some alternatives. The simplest prior is a uniform prior which constrains the parameter values to be within an upper and lower bound. It has been shown that the distribution of wavelet coefficients for continuous images follows a generalized Gaussian distribution (Antonini *et al.* 1990, 1992) suggesting that a generalized Gaussian distribution may be a suitable prior for wavelet based parametrizations. For Bayesian approaches to wavelet based Compressive Sensing, 'spike and slab' priors have been used (Ishwaran & Rao 2005; He & Carin 2009). Any of these choices are possible and we leave the prior on the Earth model parameters at each active tree node unspecified and simply write $p(\mathbf{v}_i|\mathcal{T}_k, k)$. In the case of $\mathbf{v}_i$ being of dimension m, this becomes

$$p(\mathbf{v}_i|\mathcal{T}_k, k) = \prod_{j=1}^{m} p(v_{i,j}|\mathcal{T}_k, k), \tag{14}$$

where $p(v_{i,j}|\mathcal{T}_k, k)$ is the prior on the jth component of the ith tree node.

3.2.5 *Prior ratios*

For each class of proposal, that is, birth, death and change value, we can write down the prior ratios. For a simple change in the jth component of the parameter value in the ith tree node, the structure of the tree does not alter and the prior ratio is

$$\frac{p(\theta')}{p(\theta)} = \frac{p(v'_{i,j}|\mathcal{T}_k, k)}{p(v_{i,j}|\mathcal{T}_k, k)}. \tag{15}$$

For uniform priors, $p(v'_{i,j}|\mathcal{T}_k, k) = p(v_{i,j}|\mathcal{T}_k, k)$ the prior ratio is unity.

For a birth proposal, the structure of the tree changes due to the addition of a new node and the prior of the values cancels except for those of the new node, hence the prior ratio is

$$\frac{p(\theta')}{p(\theta)} = \frac{p(k+1)p(\mathcal{T}_{k+1})p(\mathbf{v}_i|\mathcal{T}_k, k)}{p(k)p(\mathcal{T}_k)}. \tag{16}$$

If the prior on k, the number of nodes, is uniform then $\frac{p(k+1)}{p(k)}$ will cancel. Analytical expressions for the prior ratio on the structure of the tree are generally not available except for some simple unrestricted trees of which we give examples in following sections.

For the death proposal, the prior ratio is

$$\frac{p(\theta')}{p(\theta)} = \frac{p(k-1)p(\mathcal{T}_{k-1})}{p(k)p(\mathcal{T}_k)p(\mathbf{v}_i|\mathcal{T}_k, k)}. \tag{17}$$

3.3 The likelihood

It is assumed that the model vector can be mapped into the same data space as our vector of observations, **d**, so that a standard misfit can be computed as

$$\Phi(\theta) = (\mathbf{G}(\theta) - \mathbf{d})^T C_e^{-1} (\mathbf{G}(\theta) - \mathbf{d}), \tag{18}$$

where **G** is the operator that represents the predictions of data observations from a model and C_e is our data error covariance matrix which assumes errors follow a Gaussian distribution. We can then use the standard normal error distribution for computing the likelihood

$$p(\mathbf{d} \mid \theta) = \frac{1}{\sqrt{(2\pi)^n |C_e|}} \exp\left\{-\frac{\Phi(\theta)}{2}\right\}, \tag{19}$$

where n is the number of observations. The operator **G** can take many forms, in Fig. 1 we showed two examples of how a quaternary tree can be mapped into a 2-D image which could be compared to measured data. In later examples, we similarly show how trees with the node values representing wavelet coefficients can be mapped into 2-D and 3-D images.

3.4 The proposals

For the proposal distribution, $q(\theta'|\theta)$, we have three different classes of proposal: birth, death and change parameter value. Throughout these explanations we use the prime superscript to represent proposed quantities, for example, θ' is a proposed model generated from the current model, θ, via proposal distribution $q(\theta'|\theta)$.

To aid the explanation of the operation of these proposal classes, we introduce three sets of nodes within a general tree structure. The first set is simply the set of all currently active tree nodes which we label $\mathcal{S}_v$. Note that $\mathcal{S}_v$ is always non-empty because it will always have at least the root of the tree as an element. The second set, which we label $\mathcal{S}_d$, is the set of all nodes in the tree that have no active child nodes. It is from this set that we choose nodes to remove from the tree during the death proposal of the algorithm. The third set, $\mathcal{S}_b$, is the set of empty nodes in the tree structure that are direct children of the nodes in set $\mathcal{S}_d$. This set represents possible locations for adding new tree nodes during the birth proposal of the algorithm. It should be noted that the set $\mathcal{S}_d$ is a subset of $\mathcal{S}_v$, whereas the set $\mathcal{S}_b$ is disjoint of the other 2 sets.

An example showing each set for a binary tree can be seen in Fig. 3 with the nodes of each set shaded with a different colour.

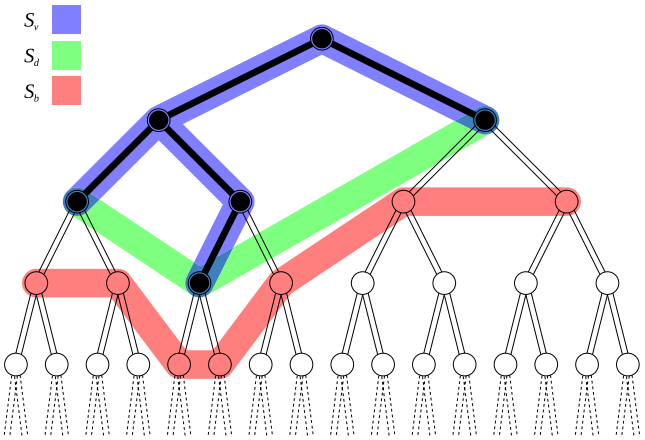

Figure 3. The first five levels of a binary tree template are shown in outline with a representative individual tree model drawn with solid lines. The nodes shaded in blue correspond to nodes in the current tree model and are members of the set $\mathcal{S}_v$, or the set of nodes that can be perturbed during a change value proposal. The nodes shaded in green are members of the set $\mathcal{S}_d$ and represent nodes that can be removed by the next death proposal. Conversely, the nodes shaded in red are members of the set $\mathcal{S}_b$ that contains inactive nodes that could be added to the tree model by the next birth proposal. Although we have shown only a binary tree here, these sets can apply equally to any tree structure.

3.4.1 *Value proposals*

The first and simplest proposal is the change value proposal. This perturbation updates the value of an existing node of the tree. If we take the general case of selecting the jth embedded parameter at the ith node in the tree, then the forward proposal probability density becomes

$$q(\theta' \mid \theta) = q(\Delta v_{i,j} \mid i, j) q(j \mid i) q(i \mid \mathcal{S}_v). \tag{20}$$

The last term of the above equation represents the probability of choosing the ith node given $\mathcal{S}_v$. Generally, the choice of which node to perturb for a value proposal will be a uniform one and so we have

$$q(i \mid \mathcal{S}_v) = \frac{1}{|\mathcal{S}_v|}, \tag{21}$$

where $|\mathcal{S}_v|$ is the number of elements in $\mathcal{S}_v$.

The second term represents the probability of selecting the jth component of the vector of value(s) at the ith tree node. For cases where there is only one value at each node this term disappears. The first term is the actual perturbation of the Earth model parameter value itself. A common approach to perturbing values in McMC samplers is to draw from a symmetric distribution centred about the current value with a pre-defined width which is tuned to achieve a desired acceptance rate. A common choice is the Normal distribution and in this case the proposal probability will be

$$q(\Delta v_{i,j} \mid i, j) = \frac{1}{\sqrt{2\pi}\sigma_{i,j}} \exp\left\{-\frac{\Delta v_{i,j}^2}{2\sigma_{i,j}^2}\right\}, \tag{22}$$

where $\sigma_{i,j}$ is the standard deviation of the normal distribution for the perturbation of the parameter. Using a proposal of this form, rather than sampling from the prior, can cause proposed values to be outside prior bounds, in which case the proposal is rejected.

The standard deviation may be the same for all tree nodes or set separately to achieve good acceptance rates. It is also straight forward to use adaptive schemes such as the Single Component

Adaptive Monte Carlo approach of Haario *et al.* (2005) and the adaptive approach of Atchade & Rosenthal (2005).

Regardless of how the standard deviation or width is set, in all cases the new value is generated from a symmetrically distributed random variable. This results in the reverse proposal probability density equal to that of the forward, so the proposal ratio for changing values is unity

$$\frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)} = 1. \tag{23}$$

3.4.2 Birth proposals

The birth proposal probability density may be written

$$q(\theta' \mid \theta) = q(\mathbf{v}_i \mid i) q(i \mid \mathcal{S}_b). \tag{24}$$

Similarly to the change value proposal, the last term represents the probability of choosing where to place the new node. Unlike in the case of the change value proposal, in some cases there is merit in preferentially choosing to birth nodes closer to the root of the tree. We have performed experiments with using weighted proposal densities of the form

$$q(i \mid \mathcal{S}_b) = \begin{cases} \frac{\mathcal{D}(i)^\alpha}{\sum_{j \in \mathcal{S}_b} \mathcal{D}(j)^\alpha} & |\mathcal{S}_b| > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{25}$$

where $\mathcal{D}(i)$ is the depth or height of node i and α is the weighting factor. Negative values of α preferentially select lower height nodes and, conversely, positive values preferentially select higher height nodes, whereas a 0 value results in a uniform choice of the birth node. In our experiments with a weighted proposal, we obtained on average poorer results than simply using a uniform proposal to select the position of the new node, so we prefer a simpler uniform proposal

$$q(i \mid \mathcal{S}_b) = \begin{cases} \frac{1}{|\mathcal{S}_b|} & |\mathcal{S}_b| > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{26}$$

The case for the condition when $|\mathcal{S}_b| = 0$ can only occur when there is some restriction on the tree structure template on the total number of nodes in the tree. An example of this would be a tree with a maximum height.

The first term of the proposal probability density in eq. (24) reflects how the new parameters are chosen for the new tree node. The simplest method of performing this is to sample the new values from the prior, that is,

$$q(\mathbf{v}_i \mid i) = p(\mathbf{v}_i | \mathcal{T}_k, k). \tag{27}$$

Although this is an 'unfocused' proposal, birthing from the prior has been shown to result in good mixing by Dosso *et al.* (2014). It also simplifies the calculation of the acceptance terms as the prior probability density in the proposal cancels with the prior ratio in the full acceptance expression.

The probability density for the reverse step can be written

$$q(\theta \mid \theta') = q(i \mid \mathcal{S}'_d). \tag{28}$$

This states that the reverse proposal is simply the probability of selecting the newly added node i from the set $\mathcal{S}'_d$. $\mathcal{S}'_d$ is the set of nodes that may be deleted after the proposed birth.

With uniform selection from the two sets involved and sampling from the prior for the new values, we obtain a general expression for the proposal ratio

$$\frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)} = \frac{|\mathcal{S}_b|}{|\mathcal{S}'_d| p(\mathbf{v}_i | \mathcal{T}_k, k)}. \tag{29}$$

3.4.3 Death proposals

The proposal probability distribution is essentially the reverse of the birth proposal, so again, for a uniform selection of the node to remove, and sampling from the prior on the reverse step, we can write the proposal ratio for a death step as

$$\frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)} = \frac{|\mathcal{S}_d| p(\mathbf{v}_i | \mathcal{T}_k, k)}{|\mathcal{S}'_b|}, \tag{30}$$

where the set $\mathcal{S}'_b$ represents the set of available points to add nodes to the tree after the selected node is removed.

3.4.4 Jacobian

The last component of the acceptance criteria is the Jacobian. For the change value proposal, the dimension of the model, θ, is constant. Since we only perturb one value at a time using a simple function of a random variable, the Jacobian will always be equal to 1 in this case.

For a birth proposal, our model space vector can be written as

$$\theta = \langle (t_1, \mathbf{v}_1), \ldots, (t_k, \mathbf{v}_k) \rangle, \tag{31}$$

where we use unique indices $t_1 \ldots t_k$ to define the currently active nodes of the tree and hence the model vector becomes a set of tuples consisting of the node index and the vector of values associated with that node. We can then write the transform, which must be bijective, for a birth step as

$$\begin{array}{c} \langle (t_1, \mathbf{v}_1), \ldots, (t_k, \mathbf{v}_k), (u, \mathbf{w}) \rangle \\ \Longleftrightarrow \\ \langle (t'_1, \mathbf{v}_1)', \ldots, (t'_k, \mathbf{v}'_k), (t_{k+1}, \mathbf{v}_{k+1})) \rangle \end{array}, \tag{32}$$

where u is a random variable used to choose the unique index of the location of the new node in the tree and $\mathbf{w}$ is the vector of random variables used to generate the values for the new node. To build the Jacobian we construct a matrix of partial derivatives of the functions used to map values from one model space to the other. For existing nodes in a birth step no change is required and

$$t_i, \mathbf{v}_i = t'_i, \mathbf{v}'_i \quad \forall i \in 1 \ldots k. \tag{33}$$

Therefore, the partial derivatives for these will be 1. The proposals as described in previous sections for the choice of the location of the new node will always mean that $t_{k+1} = u$ and likewise this will result in a partial derivative of 1.

In the case where we sample from the prior for the values in the new node, we will similarly have $\mathbf{v}_{k+1} = \mathbf{w}$ which will result in an identity matrix for the Jacobian and therefore 1 for the determinant. This is the scenario that we have generally used but we would like to highlight a potential extension that results in some modification to the Jacobian.

The application of the tree structure suggests a multiresolution hierarchy and as such we expect there to be some relationship between either the parent node and the newly added child node or a newly added child node and its siblings. For example, we may

expect that the values at the child will be less than that of the parent so we may wish to choose random values scaled by those of the parent. Alternatively, we may expect the mean of the child nodes to be near zero, and so if there are existing child nodes then we scale and offset the new values according to the values of the siblings. In either case, we find that the mapping takes the form

$$\mathbf{v}_{k+1} = f(\mathbf{w}, \mathbf{v}_j), \tag{34}$$

where f is some function of both the random variables and one or more of the existing values of other tree nodes (e.g. the parent or other sibling nodes). This will result in off-diagonal values in the Jacobian matrix. Some choices of the function, f, may also result in non-unity values along the diagonal of the Jacobian and care must be taken to correctly compute the Jacobian scaling term.

3.4.5 *The general acceptance criteria*

We are now in a position to write down the general acceptance criteria for a trans-dimensional sampler on tree structures by combining the expressions from the previous sections. For a value proposal, the acceptance criterion is

$$\alpha(\theta' \mid \theta) = \min\left\{1, \frac{p(v'_{i,j}|\mathcal{T}_k, k)}{p(v_{i,j}|\mathcal{T}_k, k)} \frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta)}\right\}. \tag{35}$$

When using a uniform prior, $p(v'_{i,j}|\mathcal{T}_k, k) = p(v_{i,j}|\mathcal{T}_k, k)$, and the above expression reduces to the likelihood ratio.

For a birth step, with the values of the new node sampled from the prior, the acceptance criterion is

$$\alpha(\theta' \mid \theta) = \min\left\{1, \frac{p(k+1)p(\mathcal{T}_{k+1})}{p(k)p(\mathcal{T}_k)} \frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta)} \frac{|\mathcal{S}_b|}{|\mathcal{S}'_d|}\right\}. \tag{36}$$

And likewise for a death step, the general acceptance criterion is

$$\alpha(\theta' \mid \theta) = \min\left\{1, \frac{p(k-1)p(\mathcal{T}_{k-1})}{p(k)p(\mathcal{T}_k)} \frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta)} \frac{|\mathcal{S}_d|}{|\mathcal{S}'_b|}\right\}. \tag{37}$$

When using a uniform prior on the number of nodes we have $p(k) = p(k+1) = p(k-1)$ and these terms cancel from the birth and death acceptance criteria.

These are conceptually simple criteria for sampling over general tree structures, however, a practical difficulty is in efficiently computing the tree structure prior ratios $\frac{p(\mathcal{T}_{k+1})}{p(\mathcal{T}_k)}$ and $\frac{p(\mathcal{T}_{k-1})}{p(\mathcal{T}_k)}$ for which we describe a fast algorithm in Appendix A.

It is generally acknowledged that the construction of acceptance criteria for trans-dimensional samplers is non-trivial. A small error in these criteria can easily result in a sampler that superficially appears to be working but will nonetheless bias the results. In Appendix B, we show some of the tests performed to validate the correctness of the new framework.

3.4.6 *A simple synthetic regression test*

To give a simple example of the application of this general framework, we implemented a simulated 1-D regression problem. This uses a binary tree template and a box car basis function of varying width and location at each node of the tree. This is the 1-D equivalent of the parametrization shown in Fig. 1. Given a boxcar

$$B(x)_{i,j} = \begin{cases} 1 & 2^{-i}j \le x < 2^{-i}(j+1) \\ 0 & \text{otherwise} \end{cases}, \tag{38}$$

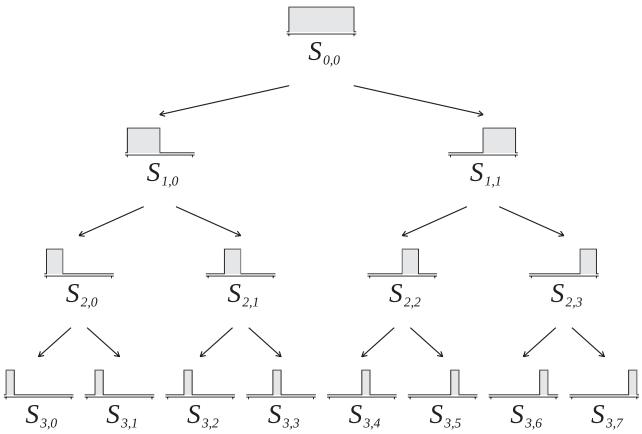

Figure 4. In a binary tree template, we can associate a boxcar basis function with each tree node. In the figure above, we show the boxcar basis functions graphically embedded in a binary tree structure. Along each row or at each height of the tree, the basis functions are orthogonal to each other. Conversely, from any parent node, the two child node basis functions are bisecting subdividers of the parents basis function. By storing scaling terms at each node of the tree, $S_{i,j}$, we can construct a 1-D function from the tree expressed as the sum of scaled versions of the basis functions using eq. (39).

where i represents its width and j its offset, we can construct a binary tree template containing coefficients, $S_{i,j}$, at each node. The 1-D regression function to be estimated is then constructed from

$$g(x) = \sum_{i=0}^{i_{\max}} \sum_{j=0}^{2^i-1} S_{i,j} B(x)_{i,j}. \tag{39}$$

The i coordinate maps to the height in the tree and j runs horizontally starting at 0 for each row. This is shown graphically in Fig. 4.

To verify that we can recover information from noisy data, we used this binary tree template with boxcar basis functions to invert data samples from a synthetic step function with added Gaussian noise. The true model is shown in Fig. 5(a) together with the data samples which are irregularly sampled to create areas of sparse coverage.

A single Markov chain was run with 1 million steps with the first 500 000 samples discarded. We set the probabilities of the birth, death and change value proposals as $p(\text{birth}) = p(\text{death}) = 0.25$ and $p(\text{changevalue}) = 0.5$. The choice of these probabilities is arbitrary except that $p(\text{birth})$ must equal $p(\text{death})$ and that they sum to one. In principle, these could be tuned for better performance in larger more complex problems, but for this simple problem this is unnecessary. The prior on the coefficients at each node was set to uniform between ± 1, and the change value proposals were normally distributed with standard deviation of 0.1. The initial model was set to have one node (the root of the tree) with its initial value sampled from the prior.

We show the mean result (solid line) in Fig. 5(b) compared to the true model (dotted line). The recovery is accurate and additionally we have not over fit the data and introduced spurious artefacts, even in regions of poor data coverage. The variance obtained from the posterior is also low which is expected in this case as our parametrization can perfectly represent the true model.

Fig. 5(c) shows the posterior histogram on the number of tree nodes used to represent the data. The modal number of tree nodes is six which matches the true model. It is interesting to note that over the course of the sampling, the entire prior range of the number of tree nodes has been sampled, as evidenced by the small number of

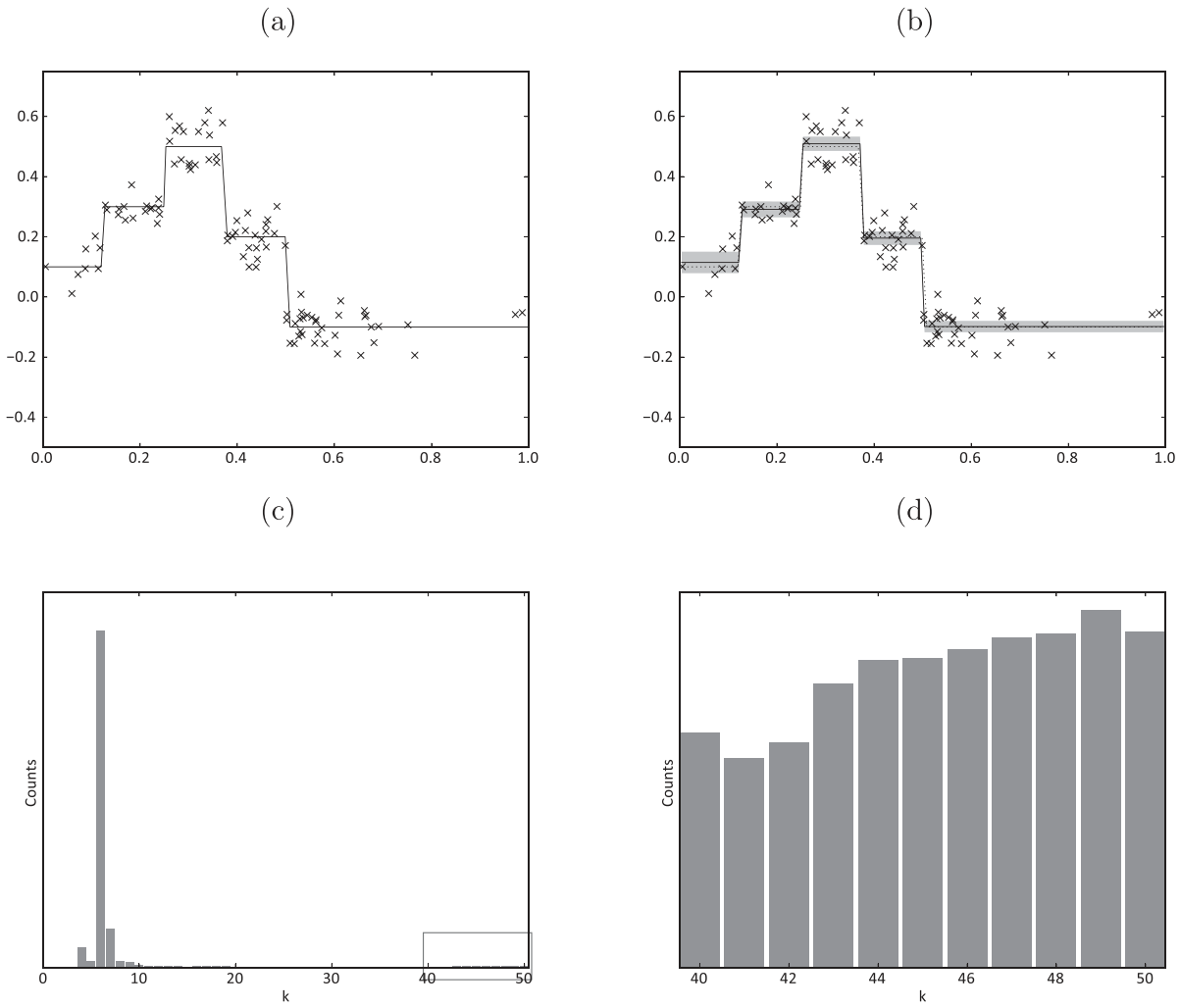

Figure 5. Here we show a 1-D regression experiment using unrestricted binary trees with boxcar basis functions. The synthetic data are shown in (a) which consists of sparsely located data points shown with crosses and the underlying true model is shown with a solid line. The recovered model is shown in (b) with a solid line compared to the true model represented with a dotted line. The shaded region represents ± 3 times the point estimate of the standard deviation from the ensemble models. In (c), we show the posterior probability density (PPD) of k, the number of nodes of the tree which has a modal value at 6. In (d), we show the PPD of k zoomed in at the higher values of k highlighted with the box in (c) to show that the posterior has sampled across the entire range of the prior.

counts at 50 nodes, shown enlarged in Fig. 5(d), even though the Markov chain is initiated at $k = 1$ nodes.

This simple tests lend confidence that the algorithm and acceptance criteria of our general framework are correct.

4 APPLICATION TO 2-D AMBIENT NOISE TOMOGRAPHY

Ambient Noise Tomography is a technique of obtaining near surface structure information from correlation of noise measurements between spatially distributed receiver stations, introduced to the seismological field by Shapiro & Campillo (2004) (see also review articles by Larose *et al.* 2006; Snieder & Larose 2013).

Trans-dimensional traveltime tomography using a Voronoi cell parametrization was introduced by Bodin & Sambridge (2009) and has been successfully used for inversion of ambient noise measurements for group velocity structure in several regional studies, for example, Young *et al.* (2011), Pilia *et al.* (2015), and Saygin *et al.* (2015). In the following sections, we show how this problem can be solved with our new trans-dimensional tree algorithm using wavelets as basis functions.

4.1 A tree-structured wavelet parametrization

Wavelet analysis may be used to decompose bounded signals in both time and frequency at multiple scales. This is in contrast to Fourier analysis which decomposes signals by frequency only (for an introduction to wavelets see Daubechies 1992 and Mallat 1999). The fast discrete wavelet transform (DWT), following the multiresolution wavelet transform of Mallat (1989), has been used in a variety of image based problems, notably image compression. Wavelet bases have been previously used in several studies for resolving seismic tomography at various scales, for example, Chiao & Kuo (2001), Simons *et al.* (2011), Chevrot *et al.* (2012), Charlety *et al.* (2013) and Fang & Zhang (2014).

The DWT in Cartesian domains has a natural multiscale hierarchy that can be traversed with a tree structure. In image compression, this has been utilized by Shapiro (1993) and Said & Pearlman (1996). In

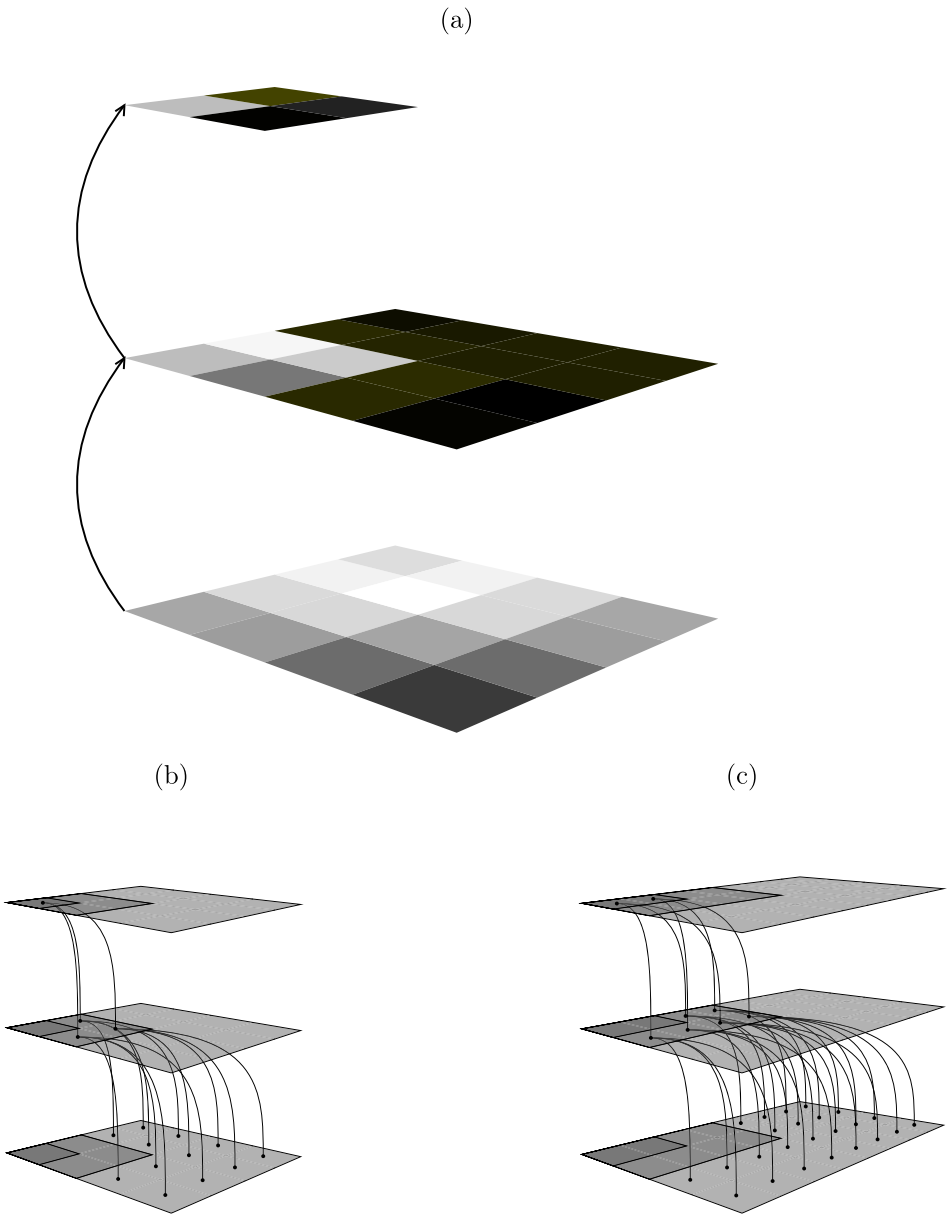

Figure 6. In (a), we show a simple 4 × 4 image in the lowest panel and two successive wavelet transforms of this image in the panels above. The first forward wavelet transform results in 2 × 2 lower resolution approximation of the input image and a set of wavelet coefficients (shown in darker shade). The next step performs the forward wavelet transform on the 2 × 2 image to obtain a 1 pixel approximation and 3 wavelet coefficients. With this 1 pixel approximation and the 3 plus 12 wavelet coefficients, we can recover the original 4 × 4 pixel image using the inverse wavelet transform. In (b), we show the tree structure that spans the 1 pixel approximation and wavelet coefficients of a 4 × 4 square image. Each level of decomposition is shaded a progressively lighter shade of grey and note how each branching of the tree coincides with the next wavelet decomposition level. In (c), we show how a variation of the tree structure can equally apply to rectangular regions by beginning from two top level coefficients.

Compressive Sensing the same tree structure has been used for 1-D signal recovery by La & Do (2005) and 2-D image reconstruction by He *et al.* (2010).

In Fig. 6(a), we show the progressive decomposition of a small 4 × 4 pixel image (bottom) by a wavelet transform. As can be seen, at each step the image is reduced by half in each dimension. The wavelet based tree structure of this wavelet decomposition is illustrated in Fig. 6(b). The progressively shaded regions indicate each level of wavelet decomposition with the darkest top left corner representing the scaling coefficient of the wavelet decomposition at the coarsest level which also corresponds to the root of our tree.

The tree has three children from the root node, and four children from every other node with the exception of the last nodes representing the finest level of detail which have no children. This is the case for a region in which the width and height are equal. For rectangular regions, a tree can be constructed by treating the initial scaling coefficients of a wavelet decomposition of a rectangular region as a 2-D subdivision grid. An example is seen in Fig. 6(c). In the following examples we use square images for simplicity, however, the only limitation when working with wavelets and this framework is that each image dimension must be a power of two.

At the root of the tree, the parameter value represents the scaling coefficient from a wavelet decomposition of the tomographic image.

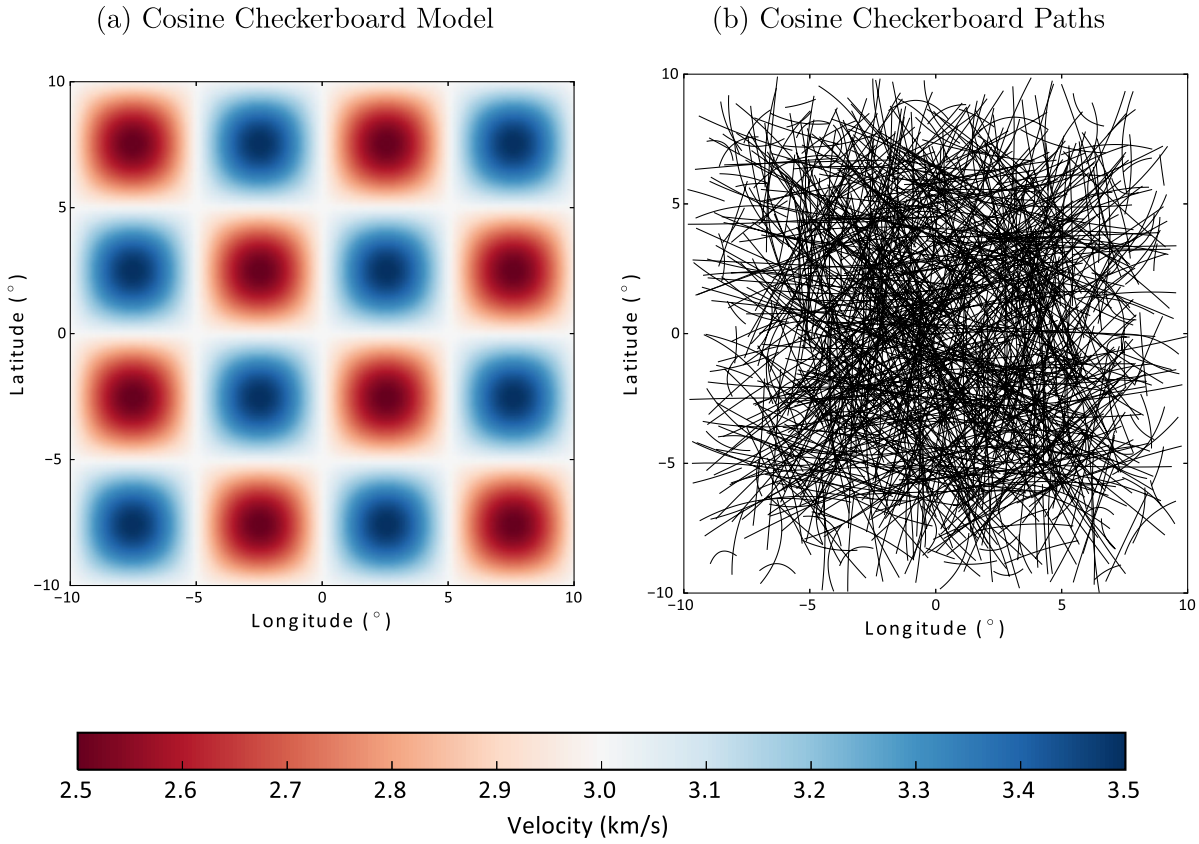

Figure 7. The synthetic models used in our tests is a smooth (cosine) checkerboard with seismic velocities between 2.5 and 3.5 km s^{-1}. We generate 1000 random ray paths through the region from which we integrate traveltimes to obtain our synthetic observations to which we add Gaussian noise.

The parameter values of the remaining tree nodes represent the hierarchy of wavelet coefficients. In contrast to the earlier 2-D image example, where values at the tree nodes are directly summed into an output image, we reconstruct an image from these coefficients by using the inverse wavelet transform (Mallat 1999).

4.2 The synthetic model and test procedure

To demonstrate the new trans-dimensional tree algorithm, we compare it to the Voronoi parametrization in some synthetic checkerboard tests with 1000 fixed ray paths. The ray paths remain fixed during the sampling to allow a direct comparison between the various parametrizations, however, there is no impediment in the new method that prevents either ray path updates at every step for a fully non-linear inversion (Galetti *et al.* 2015) or periodic updates for an iterative non-linear scheme (Bodin & Sambridge 2009).

The true model and the ray coverage are shown in Fig. 7. The region of the test is set to a square bounded at ±10 degrees longitude and latitude. The model is a smooth (cosine) checkerboard, we also show results for a discontinuous (boxcar) checkerboard in Appendix C. The observed traveltimes are computed by integrating along each path and Gaussian noise is added with a standard deviation of 5 s which corresponds to approximately a 2.5 per cent error on the mean traveltime.

For the wavelet parametrization, we repeat the exercise with three different wavelet bases. These are the Haar wavelet (Haar 1910), the Daubechies 6-tap wavelet (Daubechies 1988) and the Cohen-Daubechies-Feauveau 9/7 wavelet (Cohen *et al.* 1992, see table 6.2) as used in the JPEG-2000 image compression standard (Usevitch 2001). The choice of these wavelet bases is designed to give a representative selection of available wavelets with varying degrees of smoothness.

We have endeavoured to perform the tests under comparable conditions. To that end, 64 independent Markov chains are used in each case with 10 million steps. At an interval of 1 million steps, we restart each chain by randomly choosing a new starting model from current population with probability proportional to the mean likelihood of each chain. This approach, detailed by Dettmer *et al.* (2011), accelerates convergence to sampling the high-probability region of the posterior PDF and prevents individual chains from becoming stuck in local modes.

For the Voronoi case, all chains are started with a single cell corresponding to a tree with a single root node. We use ‘birth from the prior’ for both the Voronoi and Wavelet parametrization in the birth and death steps. For change value proposals, we use fixed Gaussian perturbations of the cell values/wavelet coefficients where we have reasonably tuned these to obtain acceptance rates of approximately 20 to 40 per cent.

The prior on the number of parameters, $p(k)$, is set to be uniform between 1 and 5000 parameters (eq. 6). In the Voronoi parametrization, we set a uniform prior on the wave speed between 2.0 and 4.0 km s^{-1} which encompasses the true range of 2.5 and 3.5 km s^{-1}.

For the Wavelet parametrization, the prior specification is complicated by the fact that the range of values of the coefficients can vary by several orders of magnitude, that is, from the coarsest to finest

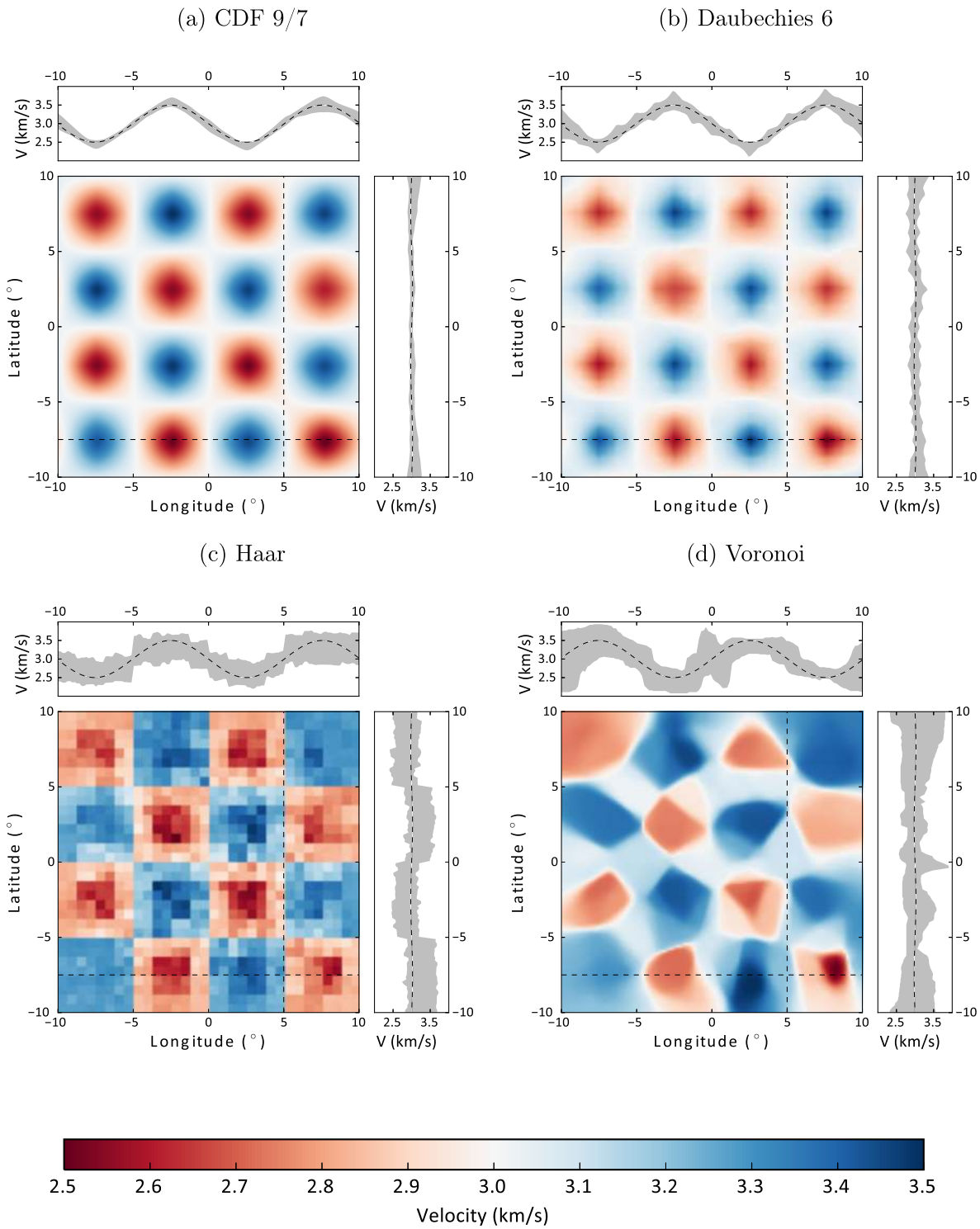

Figure 8. The mean of the ensembles obtained for the four different parametrizations. In each plot, we also show the uncertainties along longitudinal and latitudinal transects indicated by the dashed lines. These show the 95 per cent credible interval as a grey shaded region with the true model overplotted with black and white dashes.

resolution. This means that it is sensible to set a different uniform prior for each level of wavelet decomposition with the prior bounds determined by examining likely velocity variations. This approach suffices for these simulation tests, but a more advanced scheme such as that of Lochbühler *et al.* (2015), would also be possible.

4.3 Ensemble mean and credible intervals

We now present the results of these simulations to compare the two approaches. First, the mean of the ensembles across all chains is shown in Fig. 8. Subjectively, the CDF 9/7 and Daubechies 6 wavelets have recovered the smooth model better. The Haar wavelet

has performed poorly while the Voronoi parametrization reasonably recovered the broad pattern of the model but introduced polygonal artefacts.

In addition to the mean of the ensemble, we can extract point wise 95 per cent credible intervals. In Fig. 8, we have also plotted the 95 per cent credible interval along transects indicated by the dashed line through the ensemble means. The credible interval is shown as a shaded grey range and the true model is shown with a dashed line. The two transects are chosen in this example so that the longitudinal transect samples along peaks and troughs while the latitudinal transect samples along a constant velocity. From Figs 8(a) and (b), we can see that the CDF 9/7 and Daubechies 6 wavelets have low uncertainties, a characteristic of model parametrizations suited to the underlying data. Contrasting this we can see that the magnitude of the uncertainties for the Haar wavelet and the Voronoi cell are significantly higher. These results highlight the point that the choice of parametrization is important both to the recovery and, more importantly, to the uncertainties recovered.

4.4 Number of model parameters

The number of parameters (coefficients in the trans-dimensional tree based wavelet parametrization and cells in the Voronoi parametrization) gives a simple measure of model complexity. Direct comparison between the two parametrizations is a little difficult because in the Voronoi parametrization each cell has three parameters, the cell value and its (x, y) coordinates. For the trans-dimensional tree wavelet parametrization, the most reasonable approach is to assume the model is written as in eq. (31) where each parameter has a coefficient value and a unique tree node identifier as variables. This would mean that we should multiply the number of Voronoi cells by 3 and the number of wavelet coefficients by 2 to obtain a fair comparison. In the Fig. 9, we show the histograms of the raw number of cells/wavelet coefficients.

For the wavelet parametrizations, the number of coefficients is higher than that of the Voronoi cell parametrization, particularly for the Haar wavelet parametrization. This may suggest that the wavelet parametrizations is over-parametrized, however, as shown in Section 4.7 this is not necessarily the case.

4.5 Computational time

We recorded the compute time for the last 1 million steps for each independent chain and averaged these to obtain an estimate of the relative computational cost of each of the parametrizations. The computed times are shown in Table 1.

Since the Voronoi parametrization is grid free, comparing the cost of integrating traveltimes along ray paths will depend on the sampling rate along the ray paths. To ensure equivalency, as much as possible, of the two methods in terms of forward model accuracy, the ray paths were sampled at approximately the upper limit of grid resolution used by the wavelet parametrization. As a 128×128 grid was used in a 20×20 degree region, this sampling spacing was approximately 0.16 degrees.

In the tree-based wavelet parametrization, the forward model cost is dominated by the inverse wavelet transform (Mallat 1999). As a general rule, a smoother wavelet will require more computational effort in the transform. For the Daubechies 6 wavelet, we used the standard DWT whereas both the Haar and CDF 9/7 used the Fast Lifted Wavelet transform (Sweldens 1996; Daubechies & Sweldens 1998). This explains the relatively poor performance of the Daubechies 6 parametrization. It is possible to use a lifted transform version of the Daubechies 6 wavelet in which case the expected time for this transform would lie between that of the Haar and the CDF 9/7 transform. However, the number of active coefficients does factor into the computational time as evidenced by the fact that the Haar computational time is greater than that of the CDF 9/7 transform (this is reversed in other examples presented in Appendix C, where more coefficients are needed by the CDF 9/7 parametrization).

Taking the median of the tree-based wavelet parametrization compute times, we can see that for these examples, the Voronoi parametrization is roughly an order of magnitude slower. These synthetic tests have relatively few coefficients. As the complexity of the models increase, the Voronoi parametrization scales in computational effort as $\mathcal{O}(\log n)$, with n the number of cells, in the best case. In contrast, the dominant cost in the forward model of the trans-dimensional tree wavelet parametrization, the inverse wavelet transform, is independent of the number of coefficients, suggesting that the wavelet method will scale better to more complex and larger scale tomographic problems.

4.6 Convergence

Monitoring convergence is notoriously difficult in McMC. In the trans-dimensional case, measures such as the Gelman-Rubin statistic (Gelman & Rubin 1992) are not applicable. In this work, we simply assume that the independent Markov chains have converged when the negative log likelihood as reached an equilibrium value consistent with the data and errors. This is sufficient for the simulated problems we introduce here but robust convergence metrics for trans-dimensional sampling is an area of further research.

The evolution of the negative log likelihood of each Markov chain is plotted in Fig. 10 for the first million steps. We can see that the trans-dimensional tree wavelet parametrization has lower variability in the log likelihood across the chains and in some cases convergence has been achieved after a relatively small number of steps.

One reason for this is that, in general, the acceptance rates for a birth or death proposal is higher in the trans-dimensional tree wavelet parametrization than for the Voronoi parametrization. In rough figures, the acceptance rates were approximately 10 per cent for the tree based wavelet method and around 5 per cent for the Voronoi method. Hence a birth proposal is approximately twice as likely to be accepted in the tree based wavelet parametrization than the Voronoi. It is a common criticism of trans-dimensional samplers that the acceptance rates for the birth/death proposal are generally quite low and therefore the convergence is hindered due to lack of mixing between model spaces. It is this higher acceptance rate for birth/death proposal that results in the faster convergence of the trans-dimensional tree wavelet parametrization.

This higher acceptance rate is a result of the tree structure coupled with a multiscale basis. To explain why this is the case we can see that in the Voronoi case, the order of the births of its cells does not matter. Contrast this with a trans-dimensional tree model where the ordering of the birth does matter as a parent node must be birthed before its child nodes. In a multiscale parametrization such as wavelets, this means that coefficients that represent broad scale features will be birthed first, and often well constrained, before finer scale features. It also means that from any particular model, any birth will be at a scale length that is appropriate to refining the model rather than wasted on large-scale feature changes.

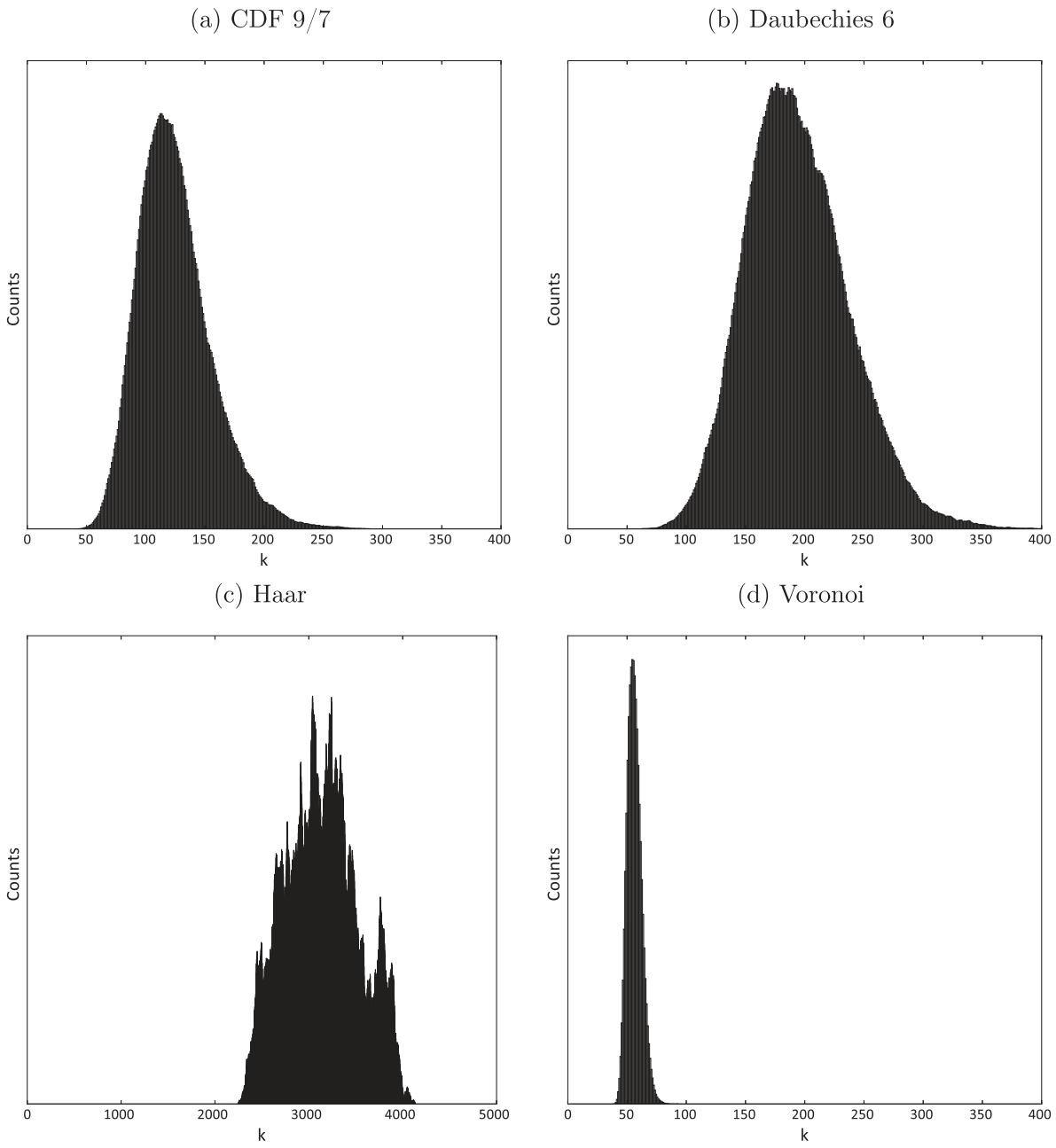

Figure 9. The estimated posterior probability distribution on the number of nodes/cells for the different parametrizations from the cosine checkerboard test.

Table 1. Mean computational time per 1 million steps for cosine checkerboard model.

Parametrization	Time (s)	Relative time
Haar	2452.8	1.4
CDF 9/7	1760	1.0
Daubechies 6	4735.7	2.7
Voronoi	30684.8	17.4

In the case of the two smooth wavelet bases, the rapid convergence and small spread of the negative log likelihood values suggests that these tree based wavelet parametrizations have efficiently explored the parameter space. This implies that large numbers of independent chains, as is needed in the Voronoi based approach, may be less important with the wavelet parametrization, given an appropriate choice of basis. In Fig. 11, we plot a comparison of the mean and MAP models of all chains combined compared to a single chain. The single chain that was chosen was the chain with the largest minimum likelihood, notionally the worst performing chain. As can be seen in the figure, even the 'worst' chain is barely distinguishable from the overall mean.

One of the primary reasons for employing multiple chains in the Voronoi parametrization is to improve robustness of the chain by averaging. In the Voronoi cell case, this is the only way to obtain a more plausible result for ambient noise tomography. The results of these experiments have shown that with a trans-dimensional tree based method, and an appropriate choice of wavelet basis function, multichain averaging may be unnecessary. Hence with the new approach it suffices to employ a smaller number of Markov chains, although with more complex and non-linear problems,

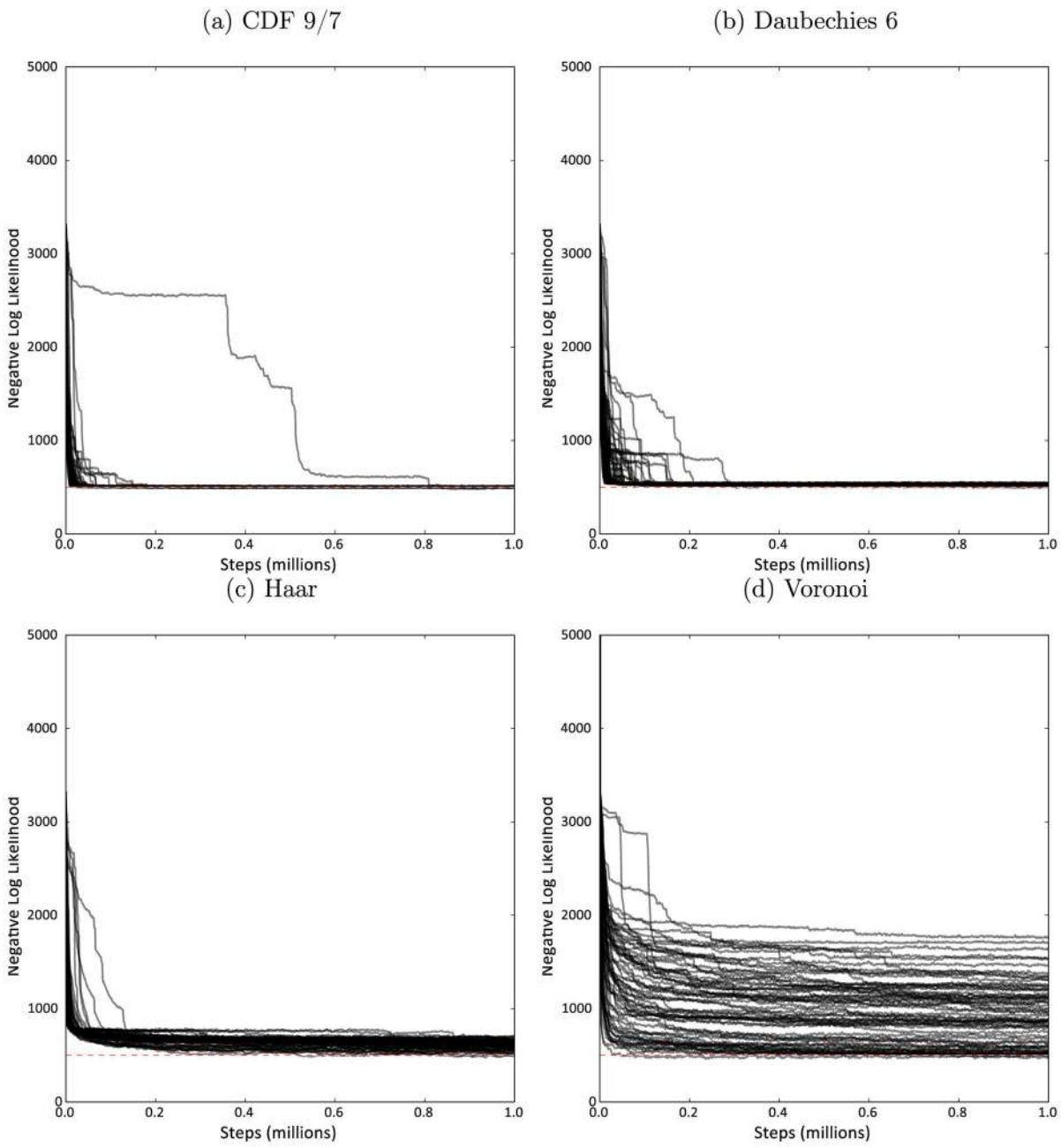

Figure 10. For each of the parametrizations compared, we plot the history of the negative log-likelihood for each of the 64 chains for the first 1 million steps during the recovery tests of the cosine checkerboard model.

parallel interacting chain approaches such as Parallel Tempering (Earl & Deem 2005; Dettmer & Dosso 2012; Dosso *et al.* 2012; Sambridge 2014) may be necessary to adequately overcome local modes and multimodalities.

4.7 Model comparisons

With the new trans-dimensional tree approach, we have flexibility in the choice of basis function. With this flexibility comes the problem of determining the best basis to use for a given problem. To compare the results of different parametrizations, in synthetic tests we can use some error norm from the 'true' model such as the mean squared error. One issue with this approach is it does not take into account model complexity and therefore may prefer over fitted models. A second issue is that in real inversions, we will not have the 'true' model with which to compare.

We therefore require a flexible model comparison criterion. A direct comparison between the trans-dimensional tree wavelet approach and the Voronoi method using the Bayesian Information Criteria (BIC; Schwarz 1978) is difficult due to the already alluded to issue of fairly estimating the number of parameters in the tree based wavelet parametrization. Here we propose the use of the Deviance Information Criteria (DIC; Spiegelhalter *et al.* 2002) which has previously been applied in trans-dimensional model comparison by Steininger *et al.* (2014). We use a variation of the DIC proposed in Gelman *et al.* (2004, chapter 12), where the DIC is computed as

$$\mathrm{DIC} = \overline{D(\theta)} + \mathrm{var}(D(\theta)), \tag{40}$$

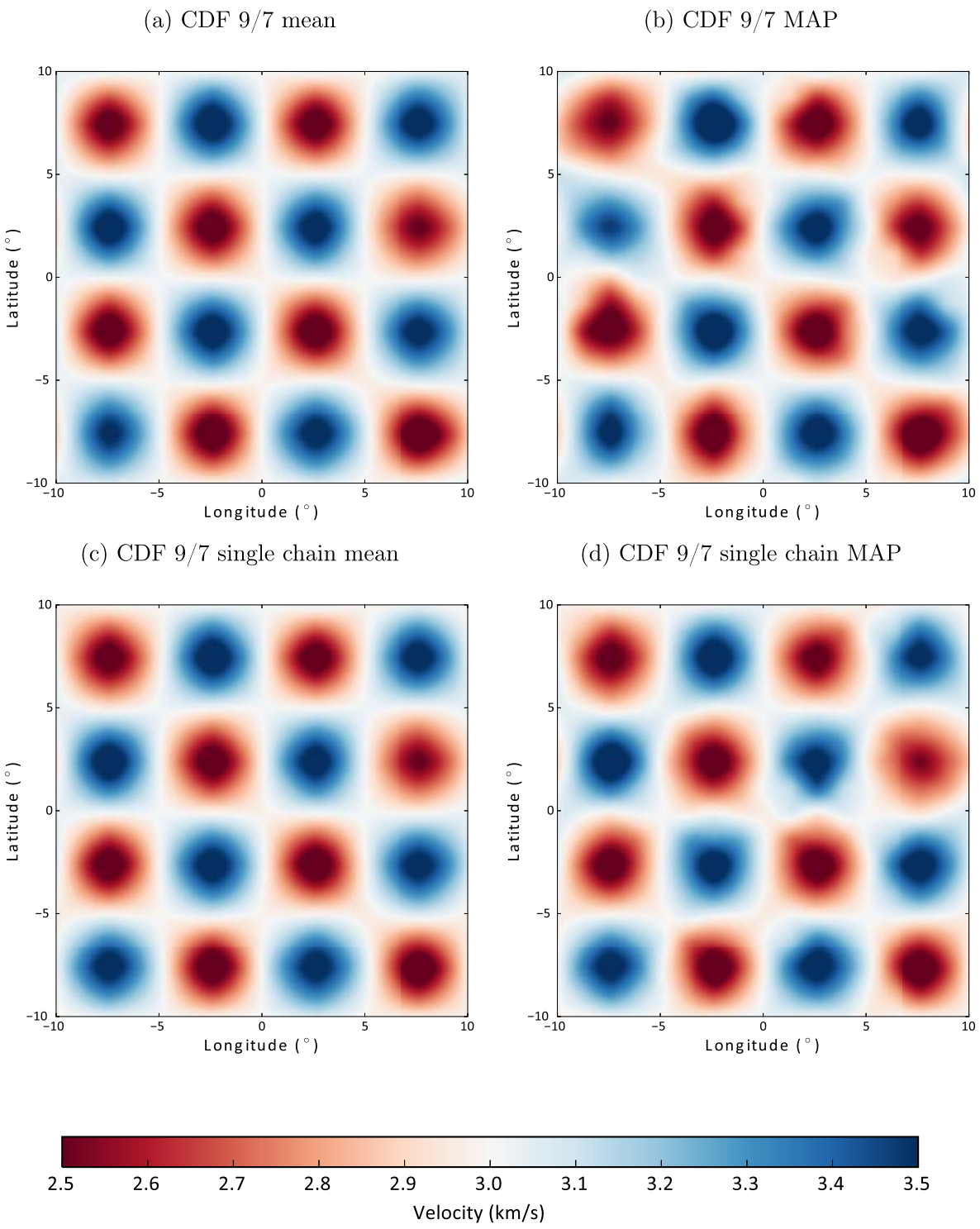

Figure 11. Here we show that for the CDF 9/7 parametrization recovering the cosine checkerboard model, that even the 'worst' performing Markov chain of the 64 parallel chains obtains results comparable to the overall ensemble solution. In (a) and (c), we compare the mean of the ensemble of the 64 chains to the mean of the single 'worst' performing chain respectively. In (b) and (d), we show the over all best Bayesian maximum *a posteriori* (MAP) and the MAP model of the 'worst' performing chain to show that they contain many similar features.

where the overbar represents the mean, and $D(\theta)$, the deviance, is given by

$$D(\theta) = -2\log\mathcal{L}(\theta) + 2\log f(\mathbf{d}). \tag{41}$$

$f(\mathbf{d})$ is a normalizing function of the data which cancels out in model comparison applications and can be ignored when computing the DIC. We prefer this formulation because in trans-dimensional sampling, point estimates can be over parametrized and from experience, using the variance results in a more stable calculation.

The first term in eq. (40) rewards a low mean negative log likelihood which penalizes too simplistic an ensemble of models. The second term penalizes model complexity since more unknowns tend

Table 2. The DIC of the various parametrizations from the cosine checkerboard recovery test.

Parametrization	$\overline{D(\theta)}$	var($D(\theta)$)	DIC
(i) *All chains*			
CDF 9/7	9280.2	82.7	9321.6
Daubechies 6	9294.7	256.4	9422.9
Haar	9232.2	481.4	9472.9
Voronoi	9207.9	540.5	9478.2
(ii) *Best chain*			
CDF 9/7	9274.4	149.8	9349.3
Daubechies 6	9282.9	194.2	9380.0
Haar	9212.1	528.4	9476.3
Voronoi	9191.2	566.1	9474.2
(iii) *Steps 750 000 to 1 000 000*			
CDF 9/7	9283.0	252.7	9409.3
Daubechies 6	9336.2	571.3	9621.8
Haar	9509.7	10817.3	14918.4
Voronoi	10193.4	467600.2	243993.5

to result in ensembles with larger likelihood variance (Spiegelhalter *et al.* 2002; Gelman *et al.* 2004). A model is said to be a better fit to the data if it has a lower DIC value. We show the results of the DIC in Table 2. We have computed the DIC across all Markov chains (i), with just a single best chain (ii), and across all chains early in the simulation (iii) (for steps 750 000 to 1 000 000).

The DIC results confirm earlier subjective visual comparisons of the mean of the ensemble (Fig. 8) to the true input models (Fig. 7) where the CDF 9/7 and Daubechies wavelet parametrizations had recovered the true model more accurately. Here we should note that the mean deviance of the Voronoi parametrization is less than that of the CDF 9/7 parametrization implying a better fit to the data. This is an example where using the misfit alone for model comparisons is insufficient. Previously we showed in Fig. 9 that the number of parameters in the wavelet parametrization was higher, suggesting over-fitting. However, the DIC shows low variance of the deviance in the wavelet case suggesting a smaller number of effective parameters.

The results of computing the DIC across all chains and a single chain are similar. We also computed the DIC during the last quarter of the first 1 million steps representing the tail end of the burn-in period. In these results, we have a great deal more variance, particularly for the Voronoi parametrization, and these results clearly show the more rapid convergence of the trans-dimensional tree approach in this problem.

With the new trans-dimensional tree wavelet method we now have the ability to choose from a variety of bases. Although we can use prior knowledge of the expected heterogeneity of the tomography to guide the choice of basis, this choice will necessarily be based on incomplete knowledge. A potential solution is to run a sweep of inversions with different basis functions and then compute the DIC (or similar criteria) of the obtained ensembles.

An alternative, which is beyond the scope of this work, is to select the wavelet bases in a hierarchical fashion itself using a trans-dimensional sampler. In this way, the choice of basis could be driven by the data.

5 3-D TELESEISMIC TOMOGRAPHY

For a more substantive test of the new trans-dimensional tree framework, we apply it to the teleseismic inversion of body waves to recover 3-D lithospheric structure. We use the inversion result and ray paths published in Rawlinson *et al.* (2011) of a large-scale regional area centred around Victoria, Australia.

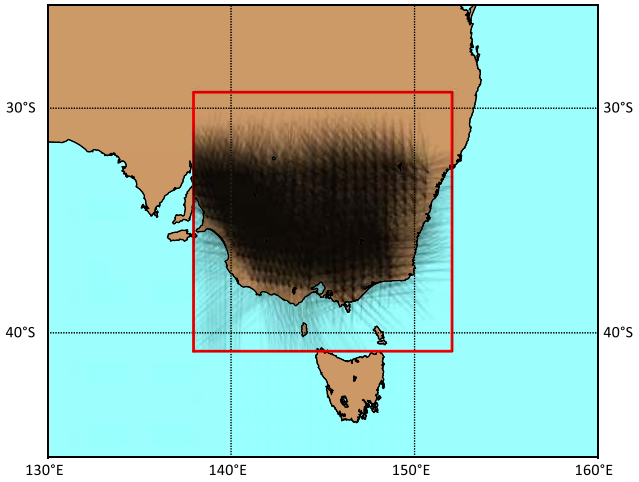

Figure 12. The teleseismic paths clipped to the 3-D region (red rectangle) used in the inversion are sourced from the published study of Rawlinson *et al.* (2011). There are 19 897 body wave ray paths in our region of interest located in southeastern Australia.

To construct simulated data for our inversion, we apply a Gaussian filter on the model obtained by Rawlinson *et al.* (2011) to remove streak artefacts and use this as our 'true' model. We then embed this model as a deviation from the AK135 Earth reference model (Kennett *et al.* 1995) in the region of interest, shown in Fig. 12, and reintegrate the 19 897 of the original 19 922 ray paths through this model to obtain true traveltimes (some paths were removed as they were outside our region of interest). Gaussian noise is then added with a standard deviation of 0.5 s which corresponds to an approximately 1 per cent error on the average traveltime through the region. As in the earlier 2-D experiments, we are linearizing the problems by using fixed ray paths.

The parametrization we use for the inversion of this region mostly follows that of the ambient noise tomography example shown earlier. We set a grid that is 128 longitude cells × 128 latitude cells × 32 radial cells to represent the region, this equates to nearly cubical voxels of approximately 10 km size. In this problem, we have a 3-D rectangular region which requires a tree starting with a 4 × 4 subdivision grid laterally, that is, 16 children from the root of the tree, which then progresses to the standard 3-D wavelet tree consisting of 7 children from these subdivision nodes and 8 children thereafter (recall that in the 2-D case this was 3 children from the root node and 4 thereafter).

In this large-scale simulation study, we use the CDF 9/7 wavelet basis. This results in a maximum of 524 288 wavelet coefficients, to sample all of which would be computationally prohibitive. However, with the trans-dimensional tree approach, we impose a hierarchy of scale over these coefficients from coarse to fine and this results in the trans-dimensional sampler selecting a far smaller (of the order of 500) number of coefficients that are required to support the resolvable model.

Nonetheless, to start from a single node of a tree as in the 2-D case would likely take a long time to burn in. To accelerate this process, we use a simple stochastic optimization scheme to generate an initial model. At each iteration, this scheme generates a large number of trial birth proposals in parallel and selects the proposed birth with the highest likelihood. To prevent this optimization method from

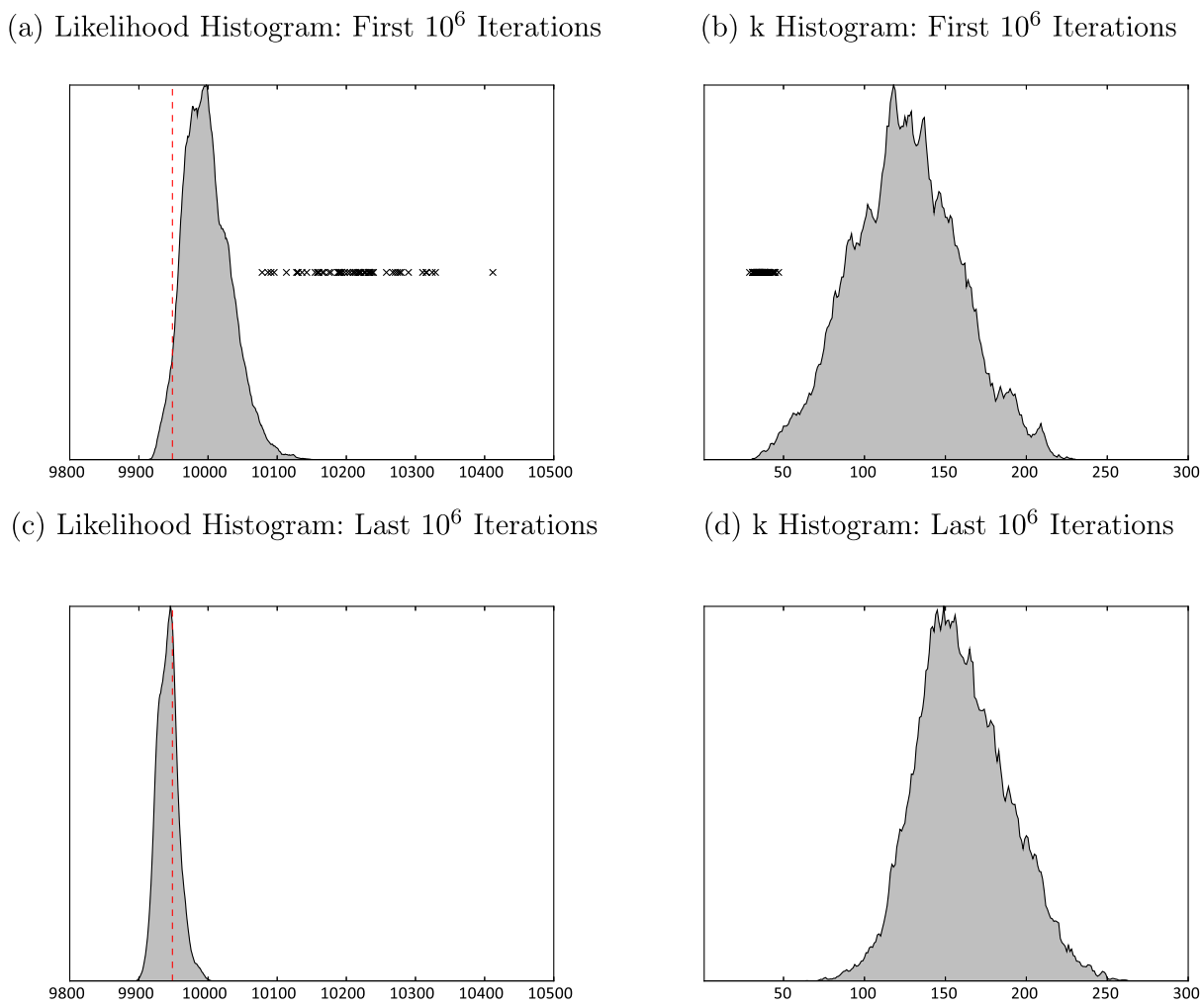

Figure 13. The histogram of likelihood and k, the number of wavelet coefficients, of all Markov chains for the first 10^6 iterations with the height restriction in place are shown in (a) and (b), respectively. Overplotted with crosses are the spread of likelihoods and k generated via the optimization scheme for the initial models. The histograms for the last 10^6 iterations are shown in (e) and (f). In the likelihood plots, the vertical red dashed line represents the theoretical χ^2 limit of the data. These plots illustrate the convergence of the likelihood and the number of coefficients through the three phases we used during the inversion.

generating over fit initial models, we halt when the BIC (Schwarz 1978) fails to decrease.

In the tree-based parametrization, we can also restrict the height that the tree is allowed to sample. Given our grid is approximately 10 km on edge, we can equate each depth our tree of seven levels with an approximate length scale: that is, level 0 represents the overall mean of the model velocity variations, level 1 represents scale lengths of 320 km, level 2 scale lengths of 160 km and so on down to level 7 which corresponds approximately to a 10 km scale length. As an additional restriction, we set an initial height restriction at level 5 (levels 6 and 7 unavailable) so the optimization scheme only generated models with scale length features down to approximately 40 km. The height restriction is an optional feature of the trans-dimensional tree method that may be used to improve convergence in higher dimension and problems of greater complexity.

We generated 60 independent models using the optimization scheme and from these starting models ran 60 Markov chains for 2 million steps. For the first 1 million steps, the height restriction remained in place but was removed for the last million. In Fig. 13, we show the spread of the negative log likelihood and number of coefficients generated from the optimization, the histograms of these during the first million steps with the height restriction, and the last million steps once this restriction is removed. In the negative log likelihood plots, we also show with a red dashed line the theoretical χ^2 limit given the number of data and the level of independent Gaussian noise.

The benefit of of the height restriction is that it allows broader scale features to converge before sampling of fine scale features commences. Fig. 13 shows how the negative log likelihood decreases from the initial models, which are clearly too simple. In the first million steps, the algorithm resolves only medium scale features due to the height restriction. In the last million steps, the chains converge to the target posterior and the likelihood distribution is tightly focused on the theoretical χ^2 limit of the data. Similarly for the posterior on k, the number of coefficients, starts from a relatively small number in the optimization phase and converges to a higher number in the final 1 million steps.

The time taken for this simulation is approximately 15 hr in total with 1 hr required for the optimization phase and 7 hr for each of the 1 million steps (Intel Xeon CPU E5-2620 at 2.10 GHz). This

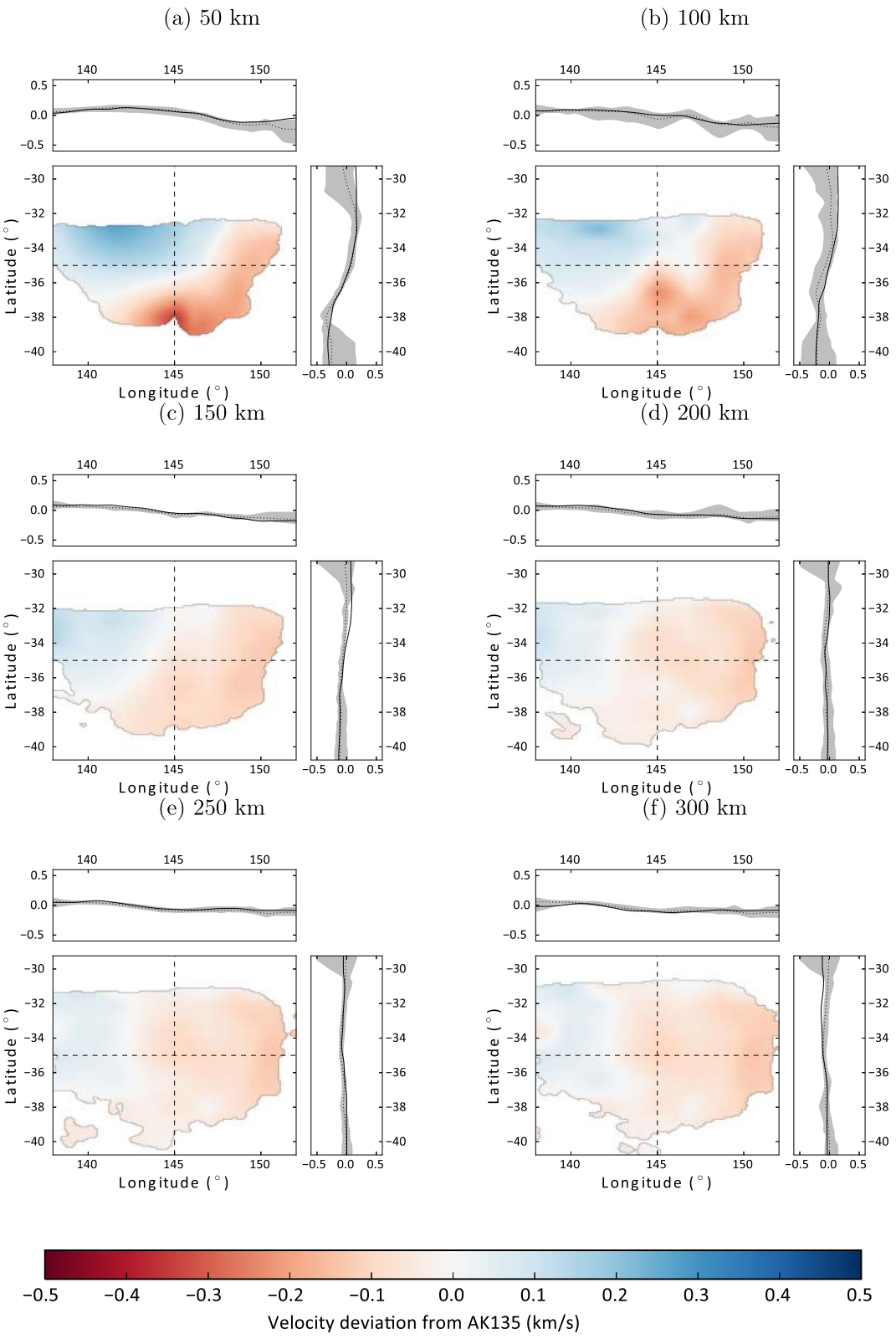

Figure 14. The ensemble mean model is shown with slices at six different depths with regions of no ray coverage masked out. In each plot, we also show uncertainties along transects indicated with the dashed lines. In the uncertainty plots, the grey region represents the 95 per cent credible interval, the ensemble mean along the transect is shown with a dotted line, and the true model with a solid line. Generally, the true model falls close to the ensemble mean and is within the uncertainty bounds indicating good recovery in this simulation.

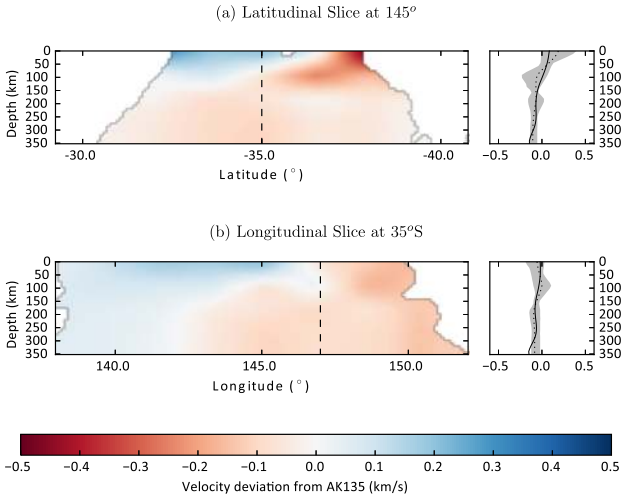

Figure 15. The ensemble mean model is shown with a slice along lines of constant longitude in (a) and latitude in (b). In each plot, we also show uncertainties along transects indicated with the dash lines. In the uncertainty plots, the grey region represents the 95 per cent credible interval, the ensemble mean along the transect is shown with a dotted line, and the true model with a solid line.

equates to approximately 25 ms per iteration. We have not performed a comparable inversion with the Voronoi parametrization. The only work we are aware of that has attempted 3-D trans-dimensional tomographic inversion is that of Piana Agostinetti *et al.* (2015) who report running times of approximately one month, however their inversion included hypocentre relocations and ray-path updates which adds significant computational complexity so a direct comparison is not meaningful. We expect that, as with the 2-D case, we would obtain approximately an order of magnitude decrease in computational time for a single chain in the trans-dimensional tree approach when compared to the Voronoi parametrization for the same scale of problem.

In Fig. 14, we present the ensemble mean results of the volume with lateral slices at varying depth. Similar to the 2-D results earlier, for each depth we show the uncertainty along lateral transects indicated by the dashed line. In the uncertainty plots, the shaded grey region shows the 95 per cent credible range, the solid line the true input model, and the dotted line the ensemble mean along the transect. We can see that we have in general achieved good recovery of the true input model, but there are some cases where the true model does not entirely reside in uncertainty bounds.

Similarly, in Fig. 15, we show two slices of the ensemble mean volume longitudinally and latitudinally to show the recovery as a function of depth. Again the recovery is quite good and the algorithm has not introduced any noticeable streaking artefacts due to the highly anisotropic ray distribution. Of minor concern is that a feature of both of the plots is a subtle underestimation of the velocity perturbation at the deepest part of the model. This is most likely a result of the poor resolvability of features at depth inherent in this teleseismic data set.

These results show that the tree based wavelet parametrization can be used for large-scale 3-D geophysical tomography problems. Further work on the parallelization or domain decomposition of the wavelet transform, coupled with parallel evaluation of the likelihood (i.e. the integration along the ray paths to obtain the model predicted traveltimes), would likely improve performance further.

6 CONCLUSIONS AND FUTURE WORK

In this study, we have presented a new trans-dimensional framework for solving general image based geophysical inverse problems which is both efficient and flexible. This new approach is efficient because of three main factors: the first is that we map our models back to regular grids which enables efficient forward model processing. Second, we can take advantage of existing fast algorithms such as the Fast Lifted Wavelet Transform for building the Earth models from the trans-dimensional tree representation. Lastly, the tree-based approach is inherently multiscale and therefore constructs models in a top down, coarse to fine scale, fashion.

Our trans-dimensional framework is flexible because it allows a wide variety of basis functions to be used for representing Earth models, while performing all sampling operations on a common tree structure. We have shown examples of simple boxcar basis functions and wavelet bases, however, more advanced bases can be used such as higher order orthogonal polynomials, curvelets (Candes & Donoho 1999), and wavelets on the sphere (Schröder & Sweldens 1995; Leistedt *et al.* 2013). Some early work on Earth Mantle inversions has been performed using a combination of spherical wavelets laterally and Cartesian wavelets radially (Hawkins & Sambridge 2014).

With this flexibility in parametrization, a choice of basis function represents *a priori* information that is at best weakly informed but generally a subjective decision. To evaluate the efficacy of these choices of bases objectively there are two approaches. First, we can perform multiple inversions with different basis functions and perform a model comparison with an objective measure such as the DIC. Second, and the subject of future work, we hope to be able to add a 'best basis' trans-dimensional selection process to the framework that will allow the data to drive this choice.

In all experiments here, we have used uninformative uniform priors, rudimentary model perturbations and limited interactions between Markov chains. As stated earlier, setting the prior on a hierarchy of coefficients for parametrizations such as wavelets is a difficult problem and worthy of further study. More sophisticated model perturbation schemes are possible, for example, the trans-dimensional tree approach lends itself well to adaptive proposal schemes, multiple model parameter updates and parallel interacting Markov chain techniques such as Parallel Tempering (Earl & Deem 2005; Dettmer & Dosso 2012; Dosso *et al.* 2012; Sambridge 2014).

In our 3-D teleseismic inversion example, we used a simple optimization scheme to seed the trans-dimensional sampling. It is possible to use more advanced optimization schemes, such as l_1 norm based sparsity maximizing schemes, previously used by Simons *et al.* (2011), Charlety *et al.* (2013), and Fang & Zhang (2014) for wavelet based parametrizations in seismic tomography problems.

From our preliminary results, the trans-dimensional tree approach appears to show promise in the probabilistic solution of large-scale geophysical inverse problems including robust uncertainty estimates.

ACKNOWLEDGEMENTS

We would like to thank Thomas Bodin for the original version of the Voronoi tomography code, Nicholas Rawlinson for making his 3-D model and rays available for our 3-D teleseismic tests, Jan Dettmer, Hrvoje Tkalčić and Stephen Roberts for constructive feedback on this work.

We would like to thank Niklas Linde and Alberto Malinverno for their constructive comments that improved the manuscript.

Aspects of this research were supported under Australian Research Council Discovery grant scheme, project DP110102098. Some calculations were performed on the Terrawulf cluster, a computational facility supported through the AuScope Australian Geophysical Observing System (AGOS) and the National Collaborative Research Infrastructure Strategy (NCRIS), both Australian Federal Governments programmes.

REFERENCES

Aki, K., 1977. Determination of the three-dimensional seismic structure of the lithosphere, *J. geophys. Res.,* **82**(2), 277–296.

Antonini, M., Barlaud, M., Mathieu, P. & Daubechies, I., 1990. Image coding using vector quantization in the wavelet transform domain, in *International Conference on Acoustics, Speech, and Signal Processing,* IEEE.

Antonini, M., Barlaud, M., Mathieu, P. & Daubechies, I., 1992. Image coding using wavelet transform, *IEEE Trans. Image Process.,* **1**(2), 205–220.

Atchade, Y.F. & Rosenthal, J.S., 2005. On adaptive Markov chain Monte Carlo algorithms, *Bernoulli,* **11**(5), 815–828.

Aval, J.C., 2008. Multivariate Fuss-Catalan numbers, *Discrete Math.,* **308**(20), 4660–4669.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances, *Phil. Trans. R. Soc.,* **53,** 370–418.

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

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

Brooks, S., Gelman, A., Jones, G.L. & Meng, X. (eds), 2011. *Handbook of Markov Chain Monte Carlo,* Chapman and Hall/CRC.

Burt, P.J. & Adelson, E.H., 1983. The Laplacian pyramid as a compact image code, *IEEE Trans. Commun.,* **31**(4), 532–540.

Candes, E.J. & Donoho, D.L., 1999. Curvelets: a surprisingly effective nonadaptive representations for objects with edges, in *Curves and Surface Fitting : Saint-Malo 1999,* Vanderbilt University Press.

Catalan, E., 1844. Note extraite d'une lettre adressée à l'éditeur par Mr. E. Catalan, répétiteur à l'école polytechnique de paris, *Journal für die reine und angewandte Mathematik,* **27,** 192–192.

Charlety, A., Voronin, S., Nolet, G., Loris, I., Simons, F.J., Sigloch, K. & Daubechines, I.C., 2013. Global seismic tomography with sparsity constraints: comparison with smoothing and damping regularization, *J. geophys. Res.,* **118,** 4887–4899.

Chevrot, S., Martin, R. & Komatitsch, D., 2012. Optimized discrete wavelet transforms in the cubed sphere with the lifting scheme - implications for global finite-frequency tomography, *Geophys. J. Int.,* **191,** 1391–1402.

Chiao, L. & Kuo, B., 2001. Multiscale seismic tomography, *Geophys. J. Int.,* **145,** 517–527.

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

Daubechies, I., 1988. Orthonormal bases of compactly supported wavelets, *Commun. Pure appl. Math.,* **41,** 909–996.

Daubechies, I., 1992. *Ten Lectures on Wavelets,* SIAM.

Daubechies, I. & Sweldens, W., 1998. Factoring wavelet transforms into lifting steps, *J. Fourier Anal. Appl.,* **4**(3), 247–269.

Denison, D.G.T., Mallick, B.K. & Smith, A.F.M., 1998. A Bayesian CART algorithm, *Biometrika,* **85**(2), 363–377.

Denison, D.G.T., Holmes, C.C., Mallick, B.K. & Smith, A.F.M., 2002. *Bayesian Methods for Non-linear Classification and Regression,* John Wiley and Sons.

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. & Holland, C.W., 2011. Sequential trans-dimensional Monte Carlo for range-dependent geoacoustic inversion, *J. acoust. Soc. Am.,* **129**(4), 1794–1806.

Dettmer, J., Molnar, S., Steininger, G., Dosso, S.E. & Cassidy, J.F., 2012. Trans-dimensional inversion of microtremor array dispersion data with hierarchical autoregressive error models, *Geophys. J. Int.,* **188,** 719–734.

Dosso, S.E., Holland, C.W. & Sambridge, M., 2012. Parallel tempering in strongly nonlinear geoacoustic inversion, *J. acoust. Soc. Am.,* **132**(5), 3030–3040.

Dosso, S.E., Dettmer, J., Steininger, G. & Holland, C.W., 2014. Efficient trans-dimensional bayesian inversion for geoacoustic profile estimation, *Inverse Problems,* **30**(11), 114018, doi:10.1088/0266-5611/30/11/114018.

Earl, D.J. & Deem, M.W., 2005. Parallel tempering: Theory, applications, and new perspectives, *Physical Chemistry Chemical Physics,* **7**(23), 3910–3916.

Fang, H. & Zhang, H., 2014. Wavelet-based double-difference seismic tomography with sparsity regularization, *Geophys. J. Int.,* **199,** 944–955.

Galetti, E., Curtis, A., Meles, G.A. & Baptie, B., 2015. Uncertainty loops in travel-time tomography from nonlinear wave physics, *Phys. Rev. Lett.,* **114,** 148501, doi:10.1103/PhysRevLett.114.148501.

Gamerman, D. & Lopes, H.F., 2006. *Markov Chain Monte Carlo,* 2nd edn, Chapman and Hall/CRC.

Gelman, A. & Rubin, D.B., 1992. *Bayesian Statistics,* chap. A single series from the Gibbs sampler provides a false sense of security, pp . 625–631, Oxford University Press.

Gelman, A., Carlin, J.B., Hal, S. & Rubin, D.B., 2004. *Bayesian Data Analysis,* 2nd edn, CRC Press.

Geyer, C.J. & Moller, J., 1994. Simulation procedures and likelihood inference for spatial point processes, *Scandinavian Journal of Statistics,* **21,** 359–373.

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

Haar, A., 1910. Zur theorie der orthogonalen funktionensysteme, *Mathematische Annalen,* **69**(3), 331–371.

Haario, H., Saksman, E. & Tamminen, J., 2005. Componentwise adaptation for high dimensional McMC, *Comput. Stat.,* **20,** 265–273.

Hanke, M., 1996. Limitations of the L-curve method in ill-posed problems, *BIT Numerical Mathematics,* **36**(2), 287–301.

Hansen, P.C., 1992. Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Rev.,* **34**(4), 561–580.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications, *Biometrika,* **57**(1), 97–109.

Hawkins, R. & Sambridge, M., 2014. A multi-scale framework for trans-dimensional tomography, in *2014 Fall Meeting,* AGU, Abstract S53A-4492.

He, L. & Carin, L., 2009. Exploiting structure in wavelet-based Bayesian compressive sensing, *IEEE Trans. Signal Process.,* **57**(9), 3488–3497.

He, L., Chen, H. & Carin, L., 2010. Tree-structured compressive sensing with variational Bayesian analysis, *IEEE Signal Process. Lett.,* **17**(3), 233–236.

Hilton, P. & Pedersen, J., 1991. Catalan numbers, their generalization, and their uses, *The Mathematical Intelligencer,* **13**(2), 64–75.

Hopcroft, P.O., Gallagher, K. & Pain, C.C., 2007. Inference of past climate from borehole temperature data using Bayesian reversible jump Markov chain Monte Carlo, *Geophys. J. Int.,* **171**(3), 1430–1439.

Iaffaldano, G., Hawkins, R. & Sambridge, M., 2014. Bayesian noise-reduction in Arabia/Somalia and Nubia/Arabia finite rotations since ∼20 Ma: Implications for Nubia/Somalia relative motion, *Geochem. Geophys. Geosyst.,* **15**(4), doi:10.1002/2013GC005089.

Ishwaran, H. & Rao, J.S., 2005. Spike and slab variable selection: frequentist and Bayesian strategies, *The Annals of Statistics,* **33**(2), 730–773.

Jaynes, E.T., 2003. *Probability Theory: The Logic of Science,* Cambridge Univ. Press.

Jeffreys, H., 1939. *Theory of Probability,* 3rd edn, Clarendon Press.

Kennett, B.L.N., Engdahl, E.R. & Buland, R., 1995. Constraints on seismic velocities in the earth from travel times, *Geophys. J. Int.,* **122,** 108–124.

Knuth, D.E., 2004. *The Art of Computer Programming,* Vol. 4, Addison-Wesley.

La, C. & Do, M.N., 2005. Signal reconstruction using sparse tree representations, *Proceedings of SPIE,* **5914,** 59140W.
Larose, E. *et al.*, 2006. Correlation of random wavefields: An interdisciplinary review, *Geophysics,* **71**(4), S111–S121.
Lawson, C.L., 1977. Software for C^1 Surface Interpolation, in *Mathematical Software III,* pp. 161–194, ed. Rice, J.R. Academic Press.
Leistedt, B., McEwen, J.D., Vandergheynst, P. & Wiaux, Y., 2013. S2LET: a code to perform fast wavelet analysis on the sphere, *Astron. Astrophys.,* **558**(A128).
Lochbühler, T., Vrugt, J.A., Sadegh, M. & Linde, N., 2015. Summary statistics from training images as prior information in probabilistic inversion, *Geophys. J. Int.,* **201**(1), 157–171.
Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.,* **151,** 675–688.
Mallat, S., 1989. Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$, *Transactions of the American Mathematical Society,* **315**(1), 69–87.
Mallat, S., 1999. *A Wavelet Tour of Signal Processing,* 2nd edn, Academic.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E., 1953. Equation of state calculations by fast computing machines, *J. Chem. Phys.,* **21**(6), 1986–1992.
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,** 252–272.
Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.,* **100**(B7), 12 431–12 447.
Okabe, A., Boots, B. & Sugihara, K., 1992. *Spatial Tesselations: Concepts and Applications of Voronoi Diagrams,* John Wiley & Sons.
Piana Agostinetti, N. & Malinverno, A., 2010. Receiver function inversion by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.,* **181**(2), 858–872.
Piana Agostinetti, N., Giacomuzzi, G. & Malinverno, A., 2015. Local 3D earthquake tomography by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.,* **201**(3), 1598–1617.
Pilia, S., Rawlinson, N., Direen, N.G., Reading, A.M., Cayley, R., Pryer, L., Arroucau, P. & Duffett, M., 2015. Linking mainland Australia and Tasmania using ambient seismic noise tomography: implications for the tectonic evolution of the east Gondwana margin, *Gondwana Research,* **28,** 1212–1227.
Plattner, A., Maurer, H.R., Vorleoper, J. & Blome, M., 2012. 3-d electrical resistivity tomography using adaptive wavelet parameter grids, *Geophys. J. Int.,* **189,** 317–330.
Rawlinson, N. & Sambridge, M., 2003. Seismic traveltime tomography of the crust and lithosphere, *Adv. Geophys.,* **46,** 81–198.
Rawlinson, N., Kennett, B. L.N., Vanacore, E., Glen, R.A. & Fishwick, S., 2011. The structure of the upper mantle beneath the Delamerian and Lachlan orogens from simultaneous inversion of multiple teleseismic datasets, *Gondwana Research,* **19,** 788–799.
Rawlinson, N., Fichtner, A., Sambridge, M. & Young, M., 2014. Seismic tomography and the assessment of uncertainty, *Adv. Geophys.,* **55,** 1–76.
Said, A. & Pearlman, W.A., 1996. A new fast and efficient image codec based on set partitioning in hierarchical trees, *IEEE Trans. Circuits Syst. Video Technol.,* **6**(3), 243–250.
Sambridge, M., 2014. A parallel tempering algorithm for probabilistic sampling and multimodal optimization, *Geophys. J. Int.,* **192,** 357–374.
Sambridge, M. & Faletic, R., 2003. Adaptive whole Earth tomography, *Geochem. Geophys. Geosyst.,* **4**(3), 1022–1042.
Sambridge, M. & Gudmundsson, O., 1998. Tomographic systems of equations with irregular cells, *J. geophys. Res.,* **103**(B1), 773–781.
Sambridge, M. & Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems, *Rev. Geophys.,* **40**(3), 1–29.
Sambridge, M., Gallagher, K., Jackson, A. & Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.,* **167,** 528–542.
Samet, H., 2006. *Foundations of Multidimensional and Metric data structures,* Morgan Kaufmann.
Saygin, E., Cummins, P., Cipta, A., Hawkins, R., Pandhu, R., Murjaya, J., Masturyono Irsyam, M., Widiyantoro, S. & Kennett, B.L.N., 2015. Imaging architecture of the Jakarta basin, Indonesia with trans-dimensional Bayesian seismic noise tomography, *Geophys. J. Int.,* in press.
Schröder, P. & Sweldens, W., 1995. Spherical wavelets: efficiently representing functions on the sphere, in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques,* ACM.
Schwarz, G.E., 1978. Estimating the dimension of a model, *Ann. Statist.,* **6**(2), 461–464.
Shapiro, J.M., 1993. Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. Signal Process.,* **41**(12), 3445–3462.
Shapiro, N.M. & Campillo, M., 2004. Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise, *Geophys. Res. Lett.,* **31,** L07614, doi:10.1029/2004GL019491.
Simons, F.J. *et al.*, 2011. Solving or resolving global tomographic models with spherical wavelets, and the scale and sparsity of seismic heterogeneity, *Geophys. J. Int.,* **187**(2), 969–988.
Snieder, R. & Larose, E., 2013. Extracting Earth's elastic wave response from noise measurements, *Annu. Rev. Earth planet. Sci.,* **41,** 183–206.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A., 2002. Bayesian measures of model complexity and fit, *J. R. Astron. Soc.: Series B,* **64,** 583–639.
Stanley, R.P., 1997. *Enumerative Combinatorics,* Vol. 1, Cambridge Univ. Press.
Steininger, G., Dosso, S.E., Holland, C.W. & Dettmer, J., 2014. Estimating seabed scattering mechanisms via Bayesian model selection, *J. acoust. Soc. Am.,* **136**(4), 1552–1562.
Sweldens, W., 1996. The lifting scheme: a custom-design construction of biorthogonal wavelets, *Appl. Comput. Harmon. Anal.,* **3**(15), 186–200.
Thurber, C.H., 1983. Earthquake locations and three-dimensional crustal structure in the Coyote lake area, central California, *J. geophys. Res.,* **88**(B10), 8226–8236.
Unser, M. & Blu, T., 2003. Mathematical properties of the JPEG2000 wavelet filters, *IEEE Trans. Image Process.,* **12**(9), 1080–1090.
Usevitch, B.E., 2001. A tutorial on modern lossy wavelet image compression: foundations of JPEG2000, *IEEE Signal Process. Mag.,* **18,** 22–35.
Vogel, C.R., 1996. Non-convergence of the L-curve regularization parameter selection method, *Inverse Problems,* **12,** 535–547.
Wilf, H.W., 1990. *Generating Functionology,* Academic Press.
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.

APPENDIX A: COUNTING ARRANGEMENTS OF GENERAL TREES

In this appendix, we give an overview of the method to computing the number of possible arrangements of a tree within a template structure given a number of nodes, k. Recall that the recurrence relationship for computing the number of arrangements in a binary tree, from eq. (11), is

$$\mathcal{N}_k = \begin{cases} 1 & k \leq 0 \\ \sum_{i=0}^{k-1} \mathcal{N}_i \mathcal{N}_{k-i-1} & \text{otherwise} \end{cases}. \tag{A1}$$

We can extend this to a ternary tree, or a tree in which every node has three possible children, as follows:

$$\mathcal{N}_k = \begin{cases} 1 & k \leq 0 \\ \sum_{i=0}^{k-1} \mathcal{N}_i \left[\sum_{j=0}^{k-1-i} \mathcal{N}_j \mathcal{N}_{k-i-j-1} \right] & \text{otherwise} \end{cases}. \tag{A2}$$

In generalizing this further, it should be recognized that this is essentially a restricted integer partitioning problem (Stanley 1997),

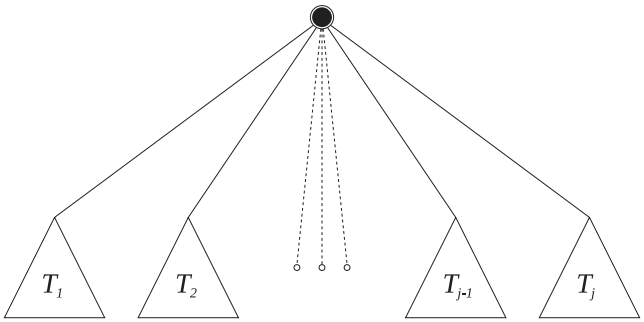

Figure A1. An abstract tree node with j subtrees.

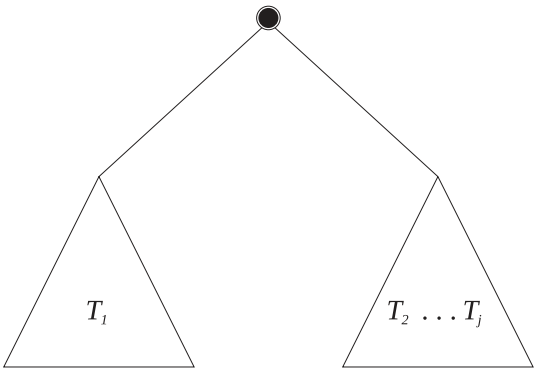

Figure A2. Rearrangement of the subtrees into a binary tree structure by amalgamation $j-1$ right most subtrees.

or stated simply as how many ways can an integer number of nodes be distributed among some arbitrary number of subtrees. In Fig. A1, we show a general tree node with j possible subtrees labeled $T_1 \ldots T_j$. It should be noted that each of these subtrees may have a different structure, that is, a different limit on the number of child nodes at the next level, to each other and to the parent tree. From this generalization, we can construct any tree structure.

By grouping the subtrees appropriately, we recognize that any number of subtrees can be reformulated into an expression of the same form as the binary tree case by treating subtree T_1 as itself and subtrees $T_2 \ldots T_k$ as an amalgamated collection of subtrees. This is shown graphically in Fig. A2.

Alternatively, when j, the number of subtrees, is even, we can split the subtrees evenly into two amalgamated collection of subtrees as shown in Fig. A3.

In either of these cases where we amalgamate multiple subtrees into two super-sub-trees, if we label these subtrees $\mathcal{A}$ and $\mathcal{B}$, we can rewrite the recurrence relationship as

$$\mathcal{N}_k = \begin{cases} 1 & k \leq 0 \\ \sum_{i=0}^{k-1} \mathcal{A}_i \mathcal{B}_{k-i} & \text{otherwise} \end{cases}. \tag{A3}$$

Note that there is a small difference between this equation and Eq. (A1) in that the number of nodes partitioned to the right branch is $k-i$ rather than $k-i-1$, that is, we have $\mathcal{B}_{k-i}$ instead of $\mathcal{N}_{k-i-1}$. The reason for this is that we effectively split the tree in two and compute the left- and right-hand sides which results in the root of the tree needing to be counted twice.

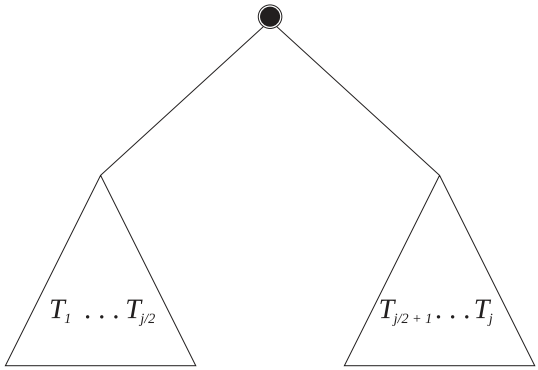

Figure A3. Rearrangement of an even number subtrees into a binary tree structure by an even amalgamation of the subtrees.

For trees or subtrees with some restriction, for example, a restriction on the height, this can be enforced by adding an extra restriction in the recurrence relationship such that

$$\mathcal{N}_k = \begin{cases} 0 & k > k_{\max} \\ 1 & k \leq 0 \text{ or } k = k_{\max} \\ \sum_{i=0}^{k-1} \mathcal{A}_i \mathcal{B}_{k-i} & \text{otherwise} \end{cases}, \tag{A4}$$

where $k_{\max}$ represents the maximum number of nodes of the current subtree. This can be computed recursively using

$$k_{\max} = 1 + k_{\max}(\mathcal{A}) + k_{\max}(\mathcal{B}). \tag{A5}$$

In all our work thus far, our $k_{\max}$ is specified as a height restriction on the tree so that for some subtrees, that is, those with a fixed number of child nodes, we can use an analytical expression to compute the maximum number of nodes. For other trees and subtrees, these are generally constructed piece wise from generic trees and it is therefore easy and efficient to compute the maximum number of nodes recursively.

We now describe the general algorithm for computing the number of arrangements of trees. The first point is that the algorithm incrementally computes the number of arrangements for a given k rather than for all values of k. Second, we memoize the result for previously computed k in the each subtree and the full tree. The memoize operation is a method of reusing previously computed results, so when we memoize some computation, the first time it is actually computed and every other time it is simply a look-up operation in a stored list of results. For recurrence relationship computations, this is vital to speed up the computation as the same partial results are frequently required. The algorithm for this is shown in Algorithm .

To give an appreciation of the need to use such an algorithm for computing the number of arrangements we have timed the algorithm for computing the number of arrangements for k equal 1 to 100 for the trees used in the 2-D wavelet parametrization in Section 4. For a naive algorithm, this takes approximately 148 min to compute and with this algorithm we can compute the same range of numbers in approximately 6 ms, over a million times faster.

APPENDIX B: VALIDATION

B1 Sampling the prior

A key test of the correctness of a set of acceptance criteria for a trans-dimensional sampler is that the criteria do not bias the posterior on k, which in our case represents the number of nodes. The simplest

Algorithm 1. Algorithm for computing the number of tree arrangements.

```
𝒟 = ∅
function MemoizeArrangements(Tree 𝒯, Integer k)
    if k = 0 or k = k_max(𝒯) then
        return 1
    end if
    if k < 0 or k > k_max(𝒯) then
        return 0
    end if
    if (𝒯, k) ∉ 𝒟 then
        j ← NumSubTrees(𝒯)
        if j = 1 then
            𝒜 ← Subtrees(𝒯, 1, 1)
            𝒟(𝒯, k) ← MemoizeArrangements(𝒜, k)
        else if j mod 2 = 1 then
            𝒜 ← Subtrees(𝒯, 1, 1)
            ℬ ← Subtrees(𝒯, 2, j)
            𝒟(𝒯, k) ← ComputeSubtrees(𝒜, ℬ, k)
        else
            𝒜 ← Subtrees(𝒯, 1, j/2)
            ℬ ← Subtrees(𝒯, j/2 + 1, j)
            𝒟(𝒯, k) ← ComputeSubtrees(𝒜, ℬ, k)
        end if
    end if
    return 𝒟(𝒯, k)
end function
function ComputeSubtrees(Tree 𝒜, Tree ℬ, Integer k)
    sum ← 0
    for i = 0 . . . k do
        a ← MemoizeArragements(𝒜, i)
        b ← MemoizeArragements(ℬ, k − i − 1)
        sum ← sum + a × b
    end for
    return sum
end function
```

way to demonstrate this is to allow the algorithm to run for a large number of steps with the likelihood kept at a constant value (hence the likelihood ratio is unity) and ensure that the posterior on the number of nodes matches the known prior.

In order to test the general algorithm, we first implemented the acceptance criteria for simple homogeneous trees, that is, binary, ternary, quaternary trees etc. For these trees, we can use the result of Aval (2008) to write down analytical expressions for the number of arrangements of the trees for a given number of nodes which then results in closed form solutions for the acceptance criteria which we produce here for the birth and death proposals

$$\alpha(\theta', \theta)_{\text{birth}} = \min\left\{1, \frac{\left[\prod_{j=2}^{n}(k(n-1)+j)\right](k+1)}{\prod_{j=1}^{n}(nk+j)} \frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta)} \frac{|\mathcal{S}_b|}{|\mathcal{S}'_d|}\right\}, \tag{B1}$$

$$\alpha(\theta', \theta)_{\text{death}} = \min\left\{1, \frac{n\prod_{j=1}^{n-1}(nk-j)}{\prod_{j=1}^{n-1}((n-1)k-j+2)} \frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta)} \frac{|\mathcal{S}_d|}{|\mathcal{S}'_b|}\right\}. \tag{B2}$$

Where n represents the number of child nodes for each tree node, that is, $n = 2$ corresponds to a binary tree, $n = 3$ corresponds to a ternary tree, etc. We performed a test of 1 million Markov steps with a uniform prior on the number of nodes between 1 and a variable

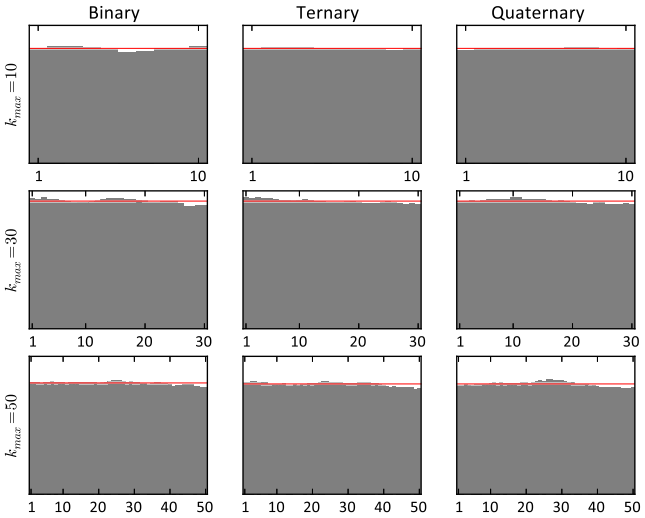

Figure B1. The sampled prior of k, the number of active tree nodes, is plotted as a grey histogram for a variety of uniform prior widths with three different classes of trees (binary, ternary and quaternary). In each plot, the solid red line represents the input prior showing there is good agreement between the prior and sampled histogram. This gives confidence that the algorithm maintains detailed balance and therefore will correctly sample the true PPD.

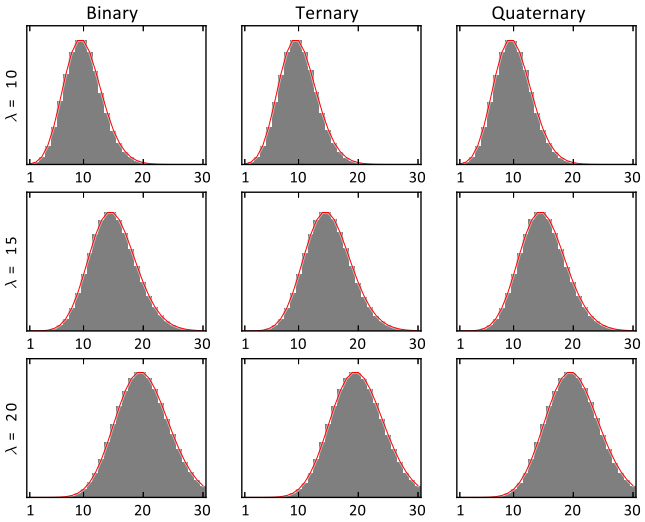

Figure B2. The sampled prior obtained when using a truncated Poisson prior is shown with a grey histogram. In each of these tests, the maximum k is fixed at 30 and the λ parameter of the Poission prior is varied with different classes of trees (binary, ternary and quaternary). The prior is shown with a solid red line and agrees well with the sampled histogram.

$k_{\max}$ and for three different values of n. The results of this test are shown in Fig. B1 with expected histogram shown with a red solid line. In all cases the McMC results approximately match with the uniform prior.

We also repeated the test for a case where the prior PDF on k is not uniform, specifically a truncated Poisson prior on the number of nodes of the form

$$p(k) = \frac{\lambda^k}{(e^\lambda - 1)k!}, \tag{B3}$$

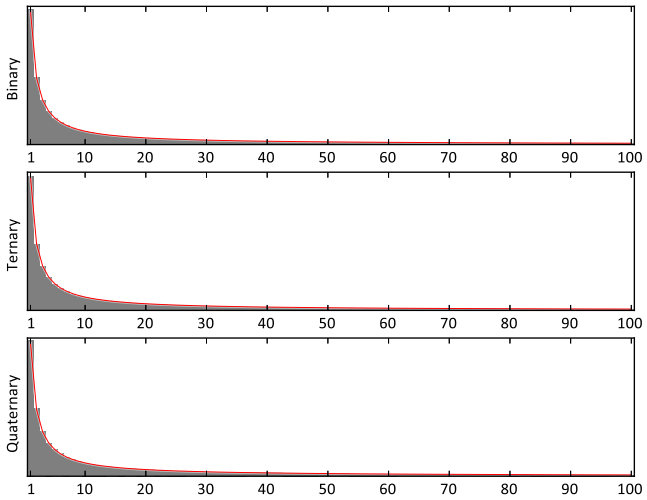

Figure B3. The sampled prior obtained when using a truncated Jeffreys' prior is shown with a grey histogram. The maximum k is fixed at 100 and we show the posterior for three classes of tree, binary, ternary and quaternary. The analytical prior is shown with a solid red line and good agreement is obtained with the sampled histogram.

where λ represents an approximate expected number of nodes in the tree. We show the posterior on the number of tree nodes obtained for varying λ and n in Fig. B2 with the prior over plotted with a solid line. Again the sampled posterior closely matches the analytical prior to within sampling error.

Lastly, we repeated this experiment with a truncated Jeffreys' prior, that is,

$$p(k) = \begin{cases} \frac{c}{k} & 1 \leq k \leq k_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{B4}$$

For some normalizing constant c and an upper limit on k of $k_{\max}$. The posteriors obtained for different n-ary trees with a $k_{\max}$ of 100 are shown in Fig. B3 along with the true distribution plotted with a solid line. In all cases we appear to be correctly sampling the prior on the number of tree nodes.

APPENDIX C: BOXCAR CHECKERBOARD RESULTS

In Section 4.2, we presented the results of a simulated smooth checkerboard 2-D tomography test. Here we repeat the same set of tests with a discontinuous boxcar checkerboard with the true model shown in Fig. C1.

C1 Ensemble mean solutions

We show the ensemble mean solutions for the boxcar checkerboard input model in Fig. C2. The Haar wavelet basis has recovered the input model almost exactly and the Voronoi parametrization has also performed well. Both of the smooth wavelet parametrizations have recovered the underlying model to a lesser degree and have ringing artefacts. This is due to a property of wavelets where the number

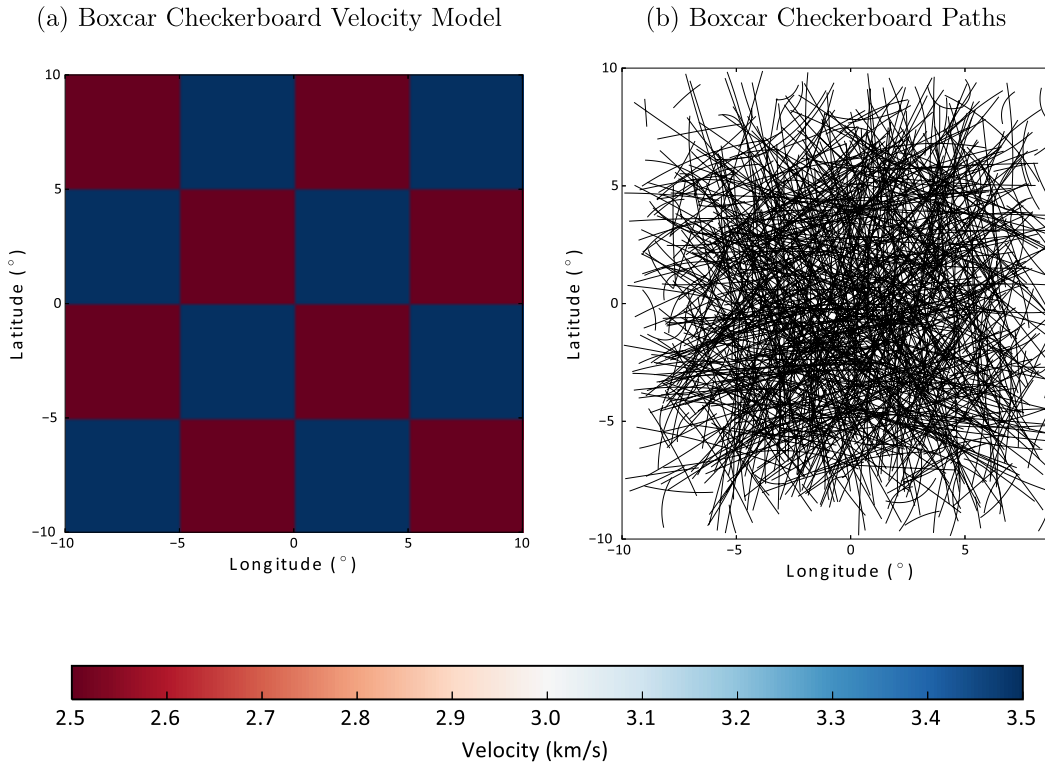

Figure C1. The boxcar synthetic models used in our tests with seismic velocities between 2.5 and 3.5 km s^{-1}. We generate 1000 random ray paths through the region from which we integrate traveltimes to obtain our synthetic observations to which we add Gaussian noise.

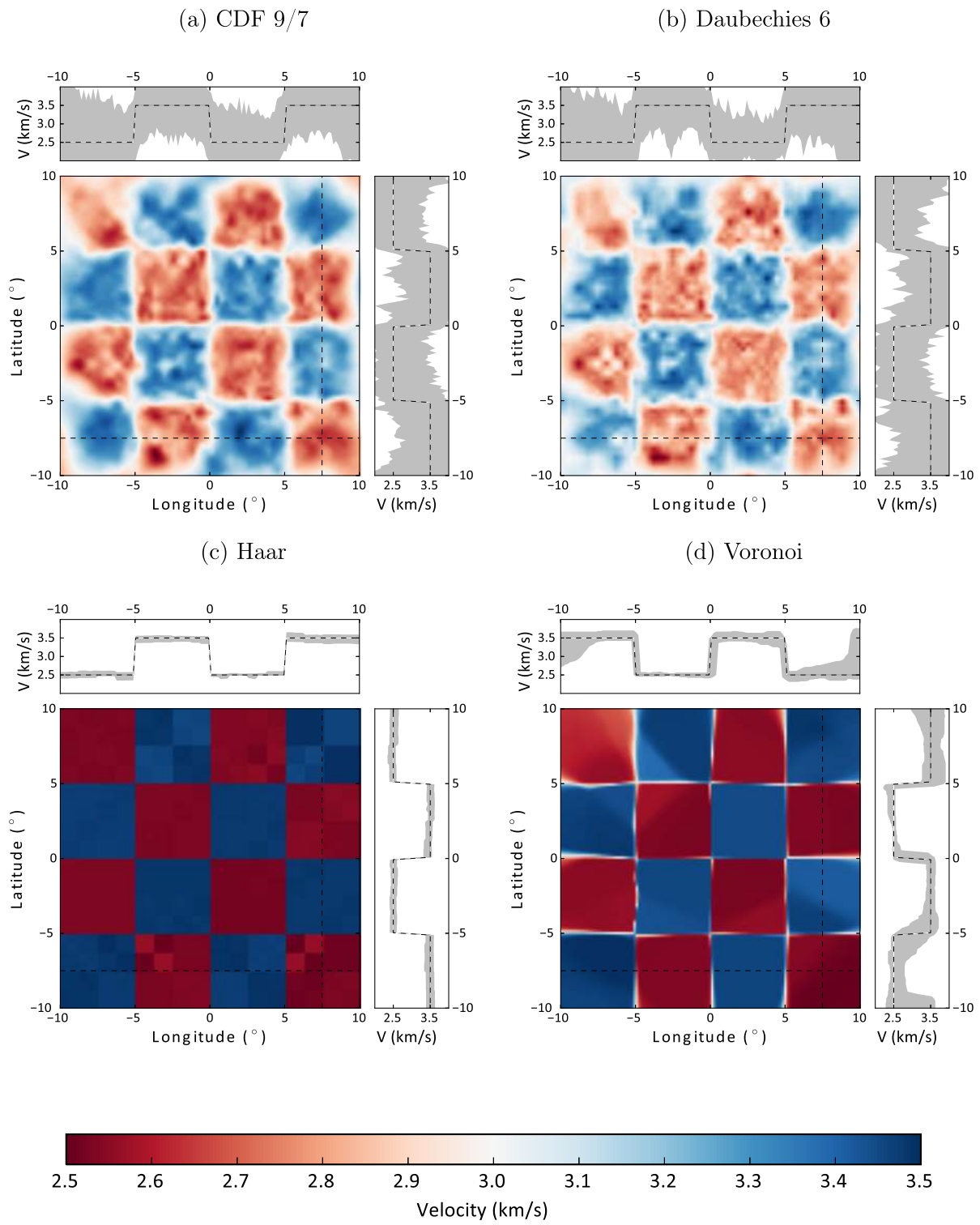

Figure C2. The mean of the ensembles obtained for the four different parametrizations used for the boxcar checkerboard input model.

of coefficients required to represent discontinuities increases as a basis becomes smoother.

C2 Number of model parameters

The histogram on the number of parameters for the boxcar checkerboard tests are shown in Fig. C3. For the CDF 9/7 and Daubechies wavelet inversion of the boxcar checkerboard, we can see that the number of coefficients required to get a poorer representation of the model is substantially larger than the other two methods. This is to be expected as the representation of hard edges with smooth wavelets requires many coefficients. Also this then becomes a more challenging search problem to find these larger number of important coefficients and to sample them sufficiently, resulting in a lengthier convergence time.

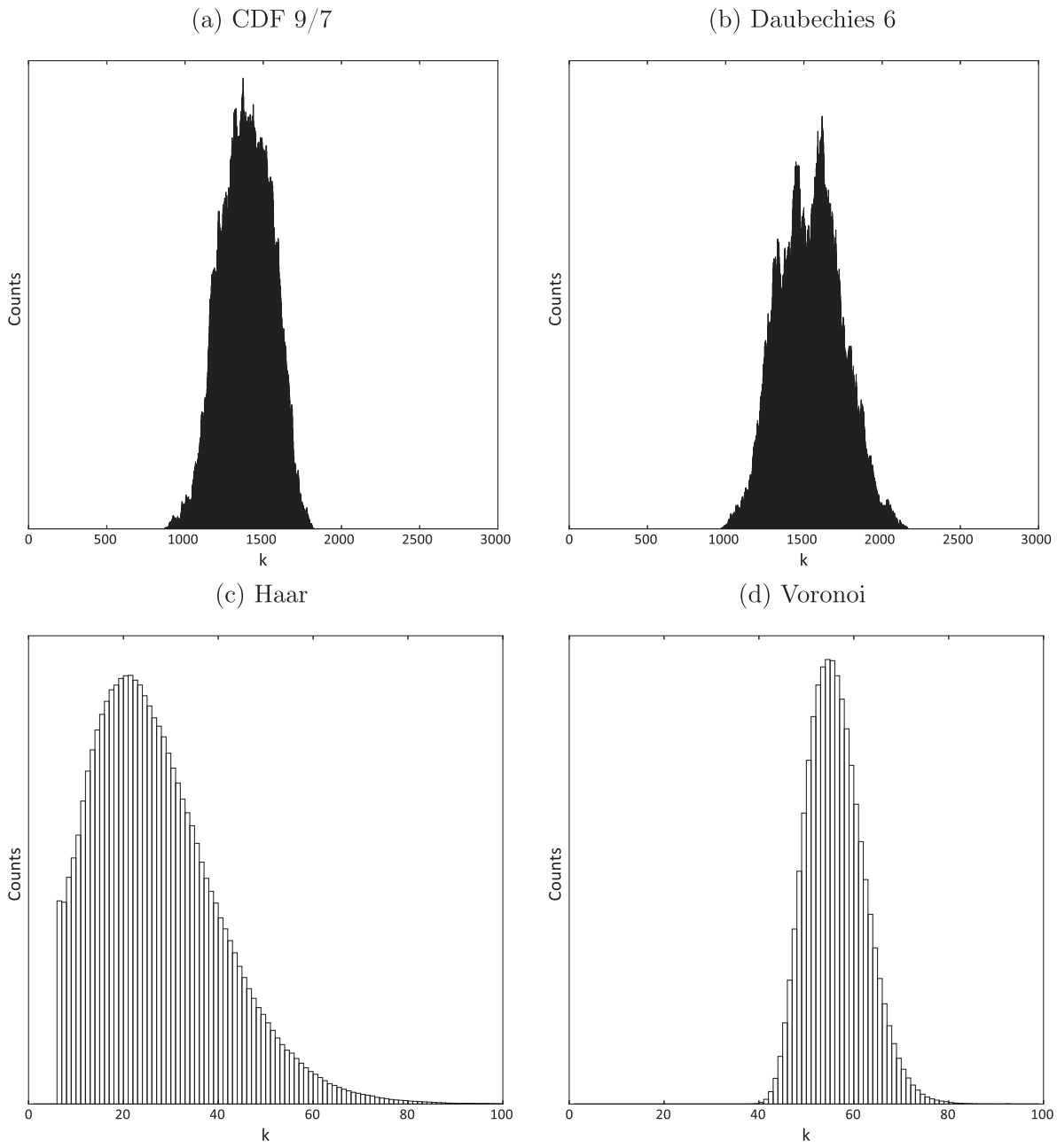

Figure C3. The estimated posterior probability distribution on the number of nodes/cells for the different parametrizations from the boxcar checkerboard test.

Table C1. Mean computational time per 1 million steps for the boxcar checkerboard model.

Parametrization	Time (s)	Relative time
Haar	1742.1	1.0
CDF 9/7	2211.4	1.3
Daubechies 6	5222.6	3.0
Voronoi	19140.1	11.0

C3 Computational time

The computational time for the boxcar checkerboard tests are shown in Table C1. The ordering is the same as for cosine checkerboard simulation with the Haar and CDF 9/7 parametrizations reversed due to the small but not insignificant computational burden resulting from a large number of coefficients.

C4 Convergence

The evolution of the negative log likelihood of each Markov chain for the boxcar checkerboard test is plotted in Fig. C4. The spread in likelihoods of the Voronoi parametrization is noticeably larger than that of the trans-dimensional tree approach, even for wavelet bases that are not a good match for this input model.

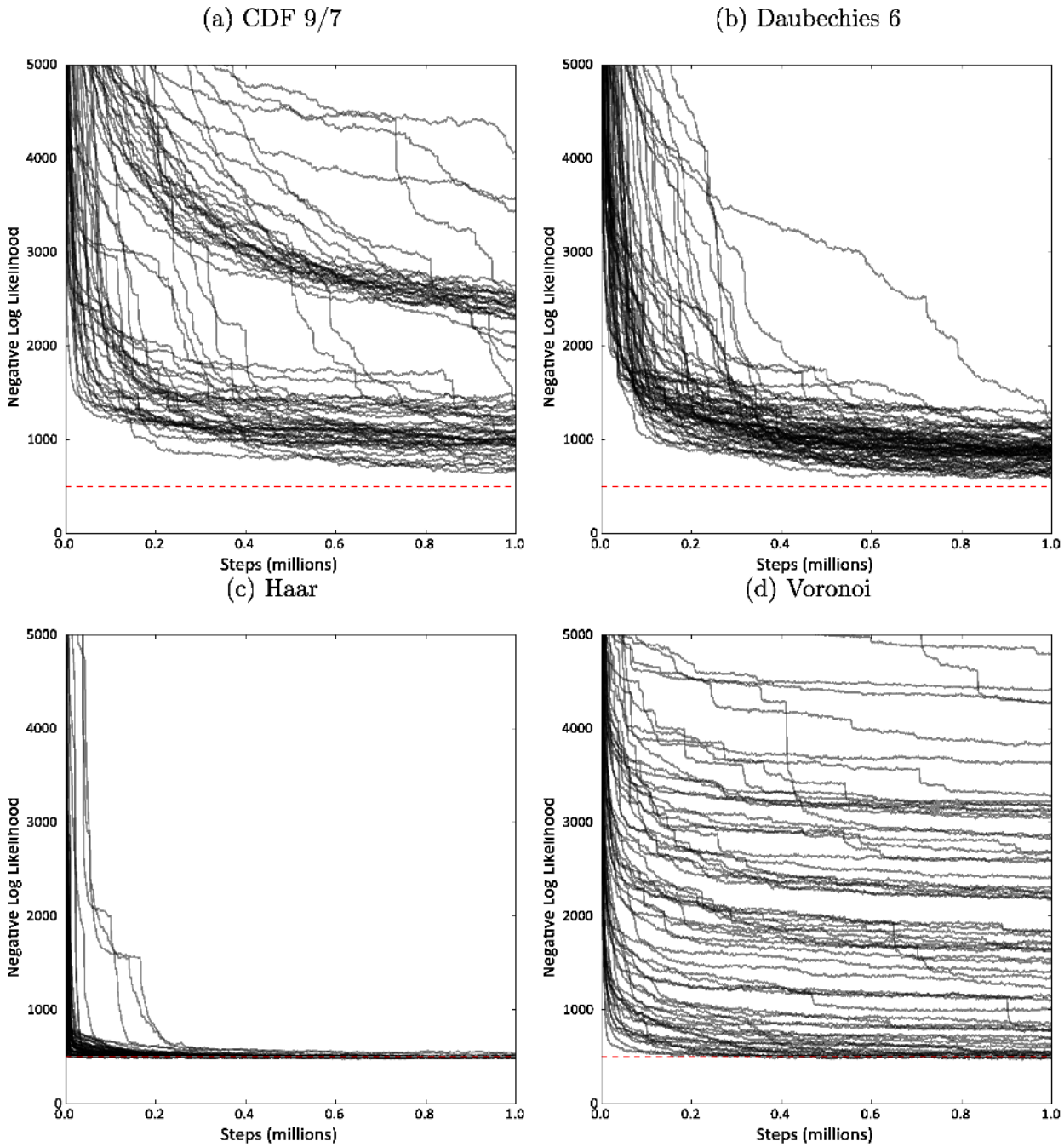

Figure C4. For each of the parametrizations compared, we plot the history of the negative log-likelihood for each of the 64 chains for the first 1 million steps during the recovery tests of the boxcar checkerboard model.

C5 Model comparisons

The DICs for the various parametrizations for the boxcar checkerboard tests are shown in Table C2. The DIC clearly favours the Haar wavelet representation in this case. It is also interesting to note that the DIC for the Haar parametrization is almost exactly the same across all chains after 10 million steps as it is during the steps 750 000 to 1 000 000, implying convergence has been reached very quickly. Again the Voronoi parametrization has the lowest deviance and therefore best overall fit, but is penalized by the variance of the deviance. The other two smooth wavelet parametrization perform more poorly as expected.

Table C2. The DIC of the various parametrizations from the boxcar checkerboard recovery test.

Parametrization	$\overline{D(\theta)}$	var($D(\theta)$)	DIC
(i) *All chains*			
CDF 9/7	9273.9	1262.4	9905.1
Daubechies 6	9292.0	1583.5	10083.8
Haar	9269.9	64.2	9302.0
Voronoi	9269.2	284.5	9411.4
(ii) *Best chain*			
CDF 9/7	9273.5	1050.2	9798.6
Daubechies 6	9253.0	1367.1	9936.6
Haar	9268.5	76.5	9306.7
Voronoi	9257.0	219.8	9366.9
(iii) *Steps 750 000 to 1 000 000*			
CDF 9/7	11727.4	3090233.3	1556844.0
Daubechies 6	10106.9	145477.7	82845.7
Haar	9272.3	332.8	9438.7
Voronoi	12056.0	5641905.0	2833008.5

C6 CONCLUSIONS

In Section 4.2, we observed the wavelet parametrization obtain better results across a series of metrics for a smooth input model. In this appendix, we have repeated this test with a discontinuous input model of the same scale length. In both cases, the trans-dimensional tree approach, with a good choice of wavelet basis, is able to outperform the Voronoi cell based approach.