Geophysical Journal International

Geophys. J. Int. (2014) **196,** 357–374
Advance Access publication 2013 October 22

doi: 10.1093/gji/ggt342

A Parallel Tempering algorithm for probabilistic sampling and multimodal optimization

Malcolm Sambridge

Research School of Earth Sciences, Australian National University, Canberra, ACT 0200*, Australia. E-mail: malcolm.sambridge@anu.edu.au*

Accepted 2013 August 28. Received 2013 August 7; in original form 2013 June 11

SUMMARY

Non-linear inverse problems in the geosciences often involve probabilistic sampling of multimodal density functions or global optimization and sometimes both. Efficient algorithmic tools for carrying out sampling or optimization in challenging cases are of major interest. Here results are presented of some numerical experiments with a technique, known as Parallel Tempering, which originated in the field of computational statistics but is finding increasing numbers of applications in fields ranging from Chemical Physics to Astronomy. To date, experience in use of Parallel Tempering within earth sciences problems is very limited. In this paper, we describe Parallel Tempering and compare it to related methods of Simulated Annealing and Simulated Tempering for optimization and sampling, respectively. A key feature of Parallel Tempering is that it satisfies the detailed balance condition required for convergence of Markov chain Monte Carlo (McMC) algorithms while improving the efficiency of probabilistic sampling. Numerical results are presented on use of Parallel Tempering for trans-dimensional inversion of synthetic seismic receiver functions and also the simultaneous fitting of multiple receiver functions using global optimization. These suggest that its use can significantly accelerate sampling algorithms and improve exploration of parameter space in optimization. Parallel Tempering is a meta-algorithm which may be used together with many existing McMC sampling and direct search optimization techniques. It's generality and demonstrated performance suggests that there is significant potential for applications to both sampling and optimization problems in the geosciences.

Key words: Numerical solutions; Inverse theory.

1 INTRODUCTION

Two classes of approach to inversion are common in the earth sciences. The first is to seek a single set of unknowns via optimization of a data misfit function, often combined with some regularization term (Parker 1994). The second is via probabilistic sampling of an *a posteriori* probability density function (PDF) within a Bayesian framework (Tarantola 2005). There are numerous examples of both approaches in the literature, and in some cases combinations (Aster *et al.* 2012; Sen & Stoffa 2013). Over the past 20 yr, geoscientists have attempted an ever expanding range of data inference problems in terms of both size and complexity. For situations where the data–model relationship is highly non-linear, the corresponding optimization or sampling problem becomes difficult because of the multimodality of the data misfit or log-likelihood term. In these circumstances, gradient-based optimization algorithms can become ineffective due to entrapment in secondary minima and likewise probabilistic sampling methods can become inefficient in converging to regions of parameter space where *a posteriori* probability density is high. Accordingly, there is an ongoing need to extend the range of problems that can be addressed through more efficient and robust inversion algorithms. Here, we define efficiency as the time taken for an algorithm to converge to a solution, and robust as the likelihood that the obtained solution is acceptable.

In this paper, we discuss a class of approach known as Parallel Tempering (PT; Geyer 1991; Falcioni & Deem 1999), which has gained considerable attention in the field of computational statistics over the last decade but to date appears to have been overlooked by earth scientists. PT was devised as a technique for probabilistic sampling of multimodal density functions, but as we shall argue here also has applications to global optimization. We introduce the rationale for PT by briefly outlining two other related techniques, Simulated Annealing (SA) and Simulated Tempering (ST), the first of which will be familiar to geophysicists through numerous applications to optimization problems over more than two decades. Several numerical examples are presented applying PT to multimodal sampling and optimization problems. Comparisons with some

existing approaches are presented both in theoretical context and by way of numerical examples. We conclude that for inverse sampling, PT appears to provide a significant acceleration of convergence, and is also able to solve complex multimodal optimization problems. The paper is concluded with a discussion of related approaches and future directions.

2 TEMPERING FOR OPTIMIZATION AND SAMPLING

The class of geophysical inverse problem considered here is one in which the likelihood function measuring discrepancies between data and model predictions is non-quadratic and most likely multimodal. This situation arises in non-linear inverse problems where the data/model relationship is sufficiently complex that gradient-based optimization methods are of limited use. An example is the well known work of Ammon *et al.* (1990) where seismic receiver functions are used to constrain 1-D shear wave velocity profiles as a function of depth. For the class of problems considered in this paper, we will assume that gradient methods are of little use, either due to their need for an ambitiously optimistic starting model, or simply because no adequately accurate or efficient approach is available for gradient calculations. If our objective is optimization of some combination of likelihood and regularization terms, then we seek the global minimum of a function $\phi(\mathbf{m})$ with respect to parameters $\mathbf{m}$ representing an earth model. Global optimization is commonly used in geosciences for many problems, examples include survey design (Curtis & Wood 2004), Gibbs' free energy minimization (Bina 1998) and phase detection in seismology (Garcia *et al.* 2006), and all such problems can be cast in a similar form (see Sen & Stoffa 2013, for a discussion of many such applications). A common approach is to optimize $\phi(\mathbf{m})$ by instead sampling a PDF $\pi(\mathbf{m}, T)$, where

$$\pi(\mathbf{m}, T) = \exp^{-\phi(\mathbf{m})/T}. \quad (1)$$

For $T > 0$ the minimum of $\phi(\mathbf{m})$ corresponds to the maximum of $\pi(\mathbf{m})$, hence an optimization problem has been converted to a statistical sampling problem where samples drawn from $\pi(\mathbf{m}|T)$ will be attracted to the peak in the distribution and hence the global minimum of $\phi(\mathbf{m})$.

Tempering will be familiar to many readers from the work of Rothman (1985, 1986) who used it to perform optimization of residual statics parameters in reflection seismology. In brief, it is the process of introducing the variable T, or temperature whose role is to rescale the optimization function, all quantities being dimensionless. The role of T is best illustrated through a simple example. Fig. 1 shows a plot of $\pi(\mathbf{m}, T)$ for a least-squares data misfit function representing the difference between observed and predicted receiver functions (see Section 3.2). Note that $\pi(\mathbf{m}, 1)$ is a multipeaked target PDF for sampling. As temperatures are raised, the shape of $\pi(\mathbf{m}, T)$ becomes relatively flat with all maxima less pronounced. Since probabilistic sampling algorithms, by definition, seek to draw samples of $\mathbf{m}$ proportional to the relative heights of the target PDF, then their performance is affected by a change in temperature. At higher temperatures the flatter PDF means that it is much easier to escape from local maxima as each is less prominent in the landscape. At the extremes as $T \to \infty$, $\pi(\mathbf{m}, T)$ tends to a uniform PDF and as $T \to 0$, $\pi(\mathbf{m}, T)$ tends to a delta function located at the global maximum in π (minimum in $\phi(\mathbf{m})$). For an optimization problem then, tempering is the process of embedding the objective function $\phi(\mathbf{m})$ into a higher dimensional sampling space $(\mathbf{m}, T)$,

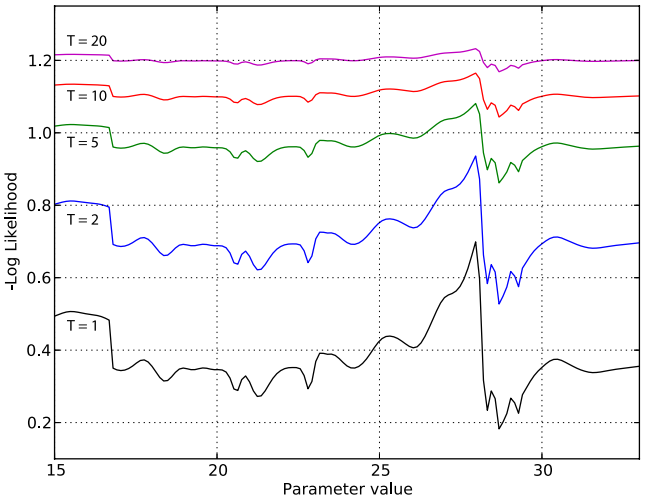

Figure 1. An example of a tempered likelihood function according to eq. (1) corresponding to inversion of a seismic receiver function. Details of the problem setup are given in Section 3.2. Here higher values represent increased likelihood or lower data misfit. The temperatures are shown on the left-hand side. Increasing temperature flattens the profile and reduces the chance that a Metropolis random sampler will get trapped by secondary peaks. To aid visualization, curves are offset vertically by 0.1, 0.15, 0.2 and 0.25 units, respectively, in cases for $T > 1$.

where the prominence of local and global minima are controlled by temperature.

In inversion settings where probabilistic sampling is the objective (for summaries see Mosegaard & Tarantola 1995; Mosegaard & Sambridge 2002; Tarantola 2005), stochastic algorithms are used to directly draw samples from Bayesian *a posteriori* PDFs given by the well known Bayes' rule

$$p(\mathbf{m}|\mathbf{d}) = k^{-1} p(\mathbf{d}|\mathbf{m}) p(\mathbf{m}), \quad (2)$$

where, as usual, $\mathbf{d}$ represents the data, $p(\mathbf{d}|\mathbf{m})$ the likelihood term and $p(\mathbf{m})$ the *a priori* PDF (or prior) on the unknowns $\mathbf{m}$ (Tarantola & Valette 1982a,b), and k is a normalizing constant. In this context, tempering may be used in a number of ways. We can write

$$\pi(\mathbf{m}, T) = [p(\mathbf{d}|\mathbf{m}) p(\mathbf{m})]^{1/T}, \quad (3)$$

and hence simply raise the unnormalized target PDF to a power of inverse temperature, as in (1). For notational convenience we ignore the dependence of π on the data $\mathbf{d}$. As before, temperature acts as a convenient way to embed the actual PDF whose sampling is desired, that is $\pi(\mathbf{m}, 1)$, into a larger augmented space of $[\mathbf{m}, T]$ over which the PDF $\pi(\mathbf{m}, T)$ is defined and whose sampling is in a sense easier due to its propensity towards flatness for $T > 1$. If $\pi(\mathbf{m}, T)$ can be sampled for the augmented set of variables, $[\mathbf{m}, T]$, then, as we shall see, samples from the conditional distribution $\pi(\mathbf{m}|T = 1)$ are usually available as a subset. In contrast to the optimization case here, we restrict the range of temperatures to $T \geq 1$ to avoid needless sampling of distributions even more peaked than the target $p(\mathbf{m}|\mathbf{d})$. Details of sampling algorithms follow in the next section, but intuitively one can recognize tempering as a way to improve the efficiency of probabilistic samplers, by allowing them to escape from local maxima and move more freely about the space at higher temperatures.

Alternate versions of tempering are possible in Bayesian computations. For example, (3) could be replaced with

$$\pi(\mathbf{m}, T) = k^{-1} p(\mathbf{d}|\mathbf{m})^{1/T} p(\mathbf{m}). \quad (4)$$

The reader will recognize that as $T \to \infty$ the tempered distribution $\pi(\mathbf{m}, T)$ becomes the *a priori* distribution, $p(\mathbf{m})$, which may not be uniform. At the other extreme $T = 1$, the tempered distribution becomes the *a posteriori* PDF, $p(\mathbf{m}|\mathbf{d})$. Here, the aim would be to draw samples from the conditional distribution at $T = 1$ using the augmented PDF, which may be an advantage in cases where there is a convenient algorithm available to draw from the particular *a priori* distribution, or indeed the prior itself is implicitly defined by such an algorithm. For an example of the latter see Mosegaard & Tarantola (1995). This form of tempering has been used to good effect by Minson *et al.* (2013) for finite fault inversion of a seismic source (see also Beck & Au 2002; Ching & Chen 2007).

Tempering in a sampling context should not be confused with hierarchical inversion schemes (such as Bodin *et al.* (2012b)) where a parameter, representing data noise variance, for example, σ^2, is also introduced into the exponent of the likelihood function, in a similar manner to T, and then sampled over. They differ because, in the hierarchical case, the data noise parameter also appears in the normalizing constant, that is, k in (4). The competing roles of σ in both the likelihood exponent and normalizing constant means that noise parameters can be constrained by the data (see Appendix B). This is in contrast to the temperature parameter in (4), which only appears in the exponent and always flattens the likelihood as it increases.

We see then that both optimization (1) and Bayesian sampling problems (3) can be treated similarly, that is, as a sampling problem of a tempered distribution over an augmented space $[\mathbf{m}, T]$. In the next section, we briefly outline a few sampling algorithms that can be used to solve problems of this kind and point out some of their limitations. This leads to the concept of PT which is the main focus of the paper.

2.1 Simulated Annealing

SA is an optimization method introduced by Kirkpatrick *et al.* (1983). It was recognized early on that SA could find practical solutions to difficult combinatorial optimization problems which generated much interest in its use. The idea was first applied in the geosciences by Rothman (1985, 1986) to fit seismic reflection waveforms in residual statics. For a detailed account of SA and its variants, as well as a summary of numerous applications that have appeared see Sen & Stoffa (2013). For our purposes a generic description will suffice and many of these details will be familiar to readers.

SA makes use of a Metropolis algorithm (Metropolis & Ulam 1949), also called a Markov chain Monte Carlo (McMC) random walker, to draw samples $\mathbf{m}$ from the conditional distribution $\pi(\mathbf{m}|T)$ given by (1) for a fixed temperature, T. A temperature ladder $T_i (i = 1, \ldots, n)$ is constructed either in advance, or dynamically, and the algorithm progressively samples from the conditional distributions $\pi(\mathbf{m}|T_i)$ as temperature is adjusted from high to low values as in Fig. 2(a). The initial temperature, T_n, is chosen sufficiently high so that the tempered distribution $\pi(\mathbf{m}|T_n)$ is relatively flat (as in Fig. 1) with the effect that changes to the model, or 'state space' $\mathbf{m}$, which both decrease and increase the objective function $\phi(\mathbf{m})$ are allowed. At each fixed temperature, the Markov chain is required to be in 'equilibrium' meaning that it is sampling from the corresponding conditional distribution $\pi(\mathbf{m}|T_n)$ without bias. As the temperature is decreased, to the next level, T_{n-1}, the algorithm switches to sampling a new conditional distribution $\pi(\mathbf{m}|T_{n-1})$ and since $T_{n-1} < T_n$ the new distribution is more peaked than the last resulting in increased preference for downhill steps in $\phi(\mathbf{m})$.

SA is illustrated in Fig. 2(a) which shows eight chains at different temperatures. We refer to the model updates of $\mathbf{m}$ as being 'within-chain', corresponding to horizontal steps in Fig. 2(a) (see the appendix for a discussion). Changes of temperature are referred to as being an update 'between-chains' (vertical steps). In practice, there may be an ensemble of independent walkers all acting in parallel, but the key feature to emphasize here is that SA involves the sampling of a series of conditional distributions, $\pi(\mathbf{m}|T_i), (i = 1 \ldots, T_n)$ transitioned in a deterministic fashion, which decrease T 'slowly enough', to drive the algorithm towards a global minimum in $\phi(\mathbf{m})$. In practice, the performance of the algorithm is always dependent on the choice of cooling schedule, that is, how and when transitions occur between temperature levels. Experience shows that the optimal cooling schedule will usually be problem dependent and hence tuning is often required for each application.

2.2 Simulated Tempering

Introduced independently by Marinari & Parisi (1992) and Geyer & Thompson (1995), ST was devised not as an optimization tool *per se* but rather for probabilistic sampling of PDFs in the form of (3). The situation is similar to SA in that a temperature ladder is required and a Metropolis algorithm draws samples of $\mathbf{m}$ within the chain according to the conditional distribution $\pi(\mathbf{m}|T_i)$ in the usual way. In ST, however, two new elements appear: (1) the temperature level may either increase or decrease, and (2) the decision to change level becomes stochastic where proposals are made at random and accepted or rejected according to the same Metropolis–Hastings rule used in a standard McMC random walker. This situation is depicted in Fig. 2(b). At points along the Markov chain, a jump is proposed between temperatures T_i and T_j and accepted with probability $\alpha(i, j)$. The Metropolis–Hastings rule for determining the acceptance probability in this situation has been determined by Geyer & Thompson (1995) and can be written as

$$\alpha(i, j) = 1 \wedge \frac{\tilde{p}(\mathbf{m}|\mathbf{d})^{1/T_j} c(T_i) q(i|j)}{\tilde{p}(\mathbf{m}|\mathbf{d})^{1/T_i} c(T_j) q(j|i)}, \tag{5}$$

where

$$\tilde{p}(\mathbf{m}|\mathbf{d}) = p(\mathbf{d}|\mathbf{m}) p(\mathbf{m}) \tag{6}$$

is the unnormalized *a posteriori* PDF; $c(T_i)$ and $c(T_j)$ are the normalizing constants

$$c(T_i) = \int \tilde{p}(\mathbf{m}|\mathbf{d})^{1/T_i} \, \mathrm{d}\mathbf{m}, \tag{7}$$

$$c(T_j) = \int \tilde{p}(\mathbf{m}|\mathbf{d})^{1/T_j} \, \mathrm{d}\mathbf{m}, \tag{8}$$

and $q(j|i)$ is the probability of proposing a move from temperature level i to j, $q(j|i) = q(i|j) = 1/2$, for $j = i \pm 1$ with $q(2|1) = q(n - 1|n) = 1$, and n is the number of temperature levels. In (5) the function $a \wedge b$ represents the minimum of a and b. The stochastic accept–reject process using (5) is necessary to kept the chain in equilibrium, a property known as 'detailed balance'. After many such steps, the ST walker will spend time at all temperatures and produce samples $(\mathbf{m}, T_i)$ whose density converges to the set of conditional distributions $\pi(\mathbf{m}|T_i), (i = 1, \ldots, n)$.

A comparison of ST and SA is illuminating when one recalls that the whole idea of tempering in an optimization context was to achieve sampling of $\pi(\mathbf{m}|T)$ as T decreases, thereby encouraging samples near the global maximum. It is evident that ST becomes

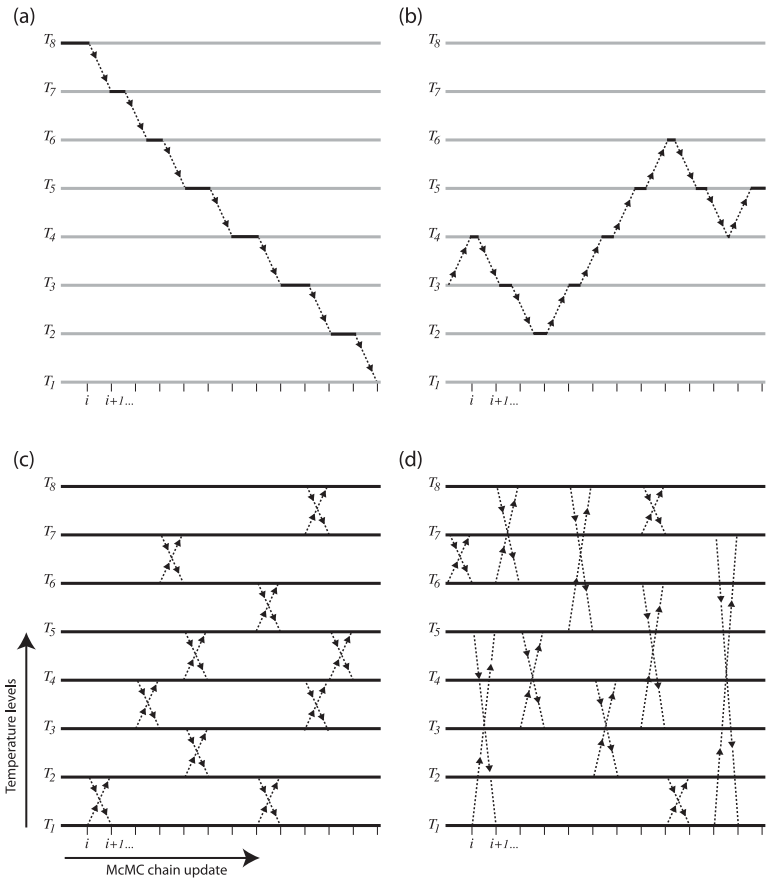

Figure 2. Schematic illustration of various tempering-based sampling algorithms. Each horizontal line represents a Markov chain random walker at a different temperature, $T_1 < T_2 < \ldots < T_8$. A standard McMC update of variables is represented by a horizontal step along the black line, whereas a change in temperature is represented by a vertical jump between levels (dashed line). Panel (a) represents a Simulated Annealing algorithm where the temperature of a single walker is steadily reduced, often according to a prescribed function; (b) corresponds to Simulated Tempering where the temperature of a single walker can increase or decrease in fixed increments; (c) shows a standard Parallel Tempering algorithm applied to eight walkers, with randomized swaps between neighbouring temperatures; (d) shows the Parallel Tempering variant used in this study with randomized swaps between any pair of temperature levels.

identical to SA if $j = i - 1$ and $\alpha = 1$, but of course, in general, this will not be the case. SA, therefore, violates the condition of detailed balance in transitioning between a finite set of temperature levels, while ST maintains detailed balance (Earl & Deem 2005). In SA the Markov chain must be in balance at each temperature and the rule of thumb is that transitions between two temperature levels must be done 'slowly' enough to avoid entrapment in local minima. The Metropolis–Hastings condition (5) provides a quantification of this statement since 'slow' translates to taking small enough increments in temperature so that $\alpha \approx 1$. In contrast, ST guarantees detailed balance for finite steps in temperature by satisfying (5).

ST would appear to have attractive properties in both optimization and sampling problems. However, looking again at (5) we note that the ratios of the normalizing constants $c(T_i)/c(T_j)$, $(j = i \pm 1; i = 1, \ldots, n)$ are required for evaluation of $\alpha(i, j)$. Marinari & Parisi (1992) suggest that these can be determined in advance by some experimentation. From (7) we see that each of them is an integral of the *a posteriori* distribution, raised to a power, over the entire model space. In Bayesian problems, this is comparable to calculation of the quantity known as the evidence, accurate determination of which is often a major computational challenge (see Sambridge *et al.* 2006). Hence, while ST might be attractive in maintaining detailed balance it is difficult to implement in cases where normalizing constants must be determined numerically.

2.3 Parallel Tempering

The preceding discussion of SA and ST helps set the scene for a description of PT, which was initially devised by Geyer (1991) with the more modern version attributed to Falcioni & Deem (1999). In this case, a temperature ladder is again employed and an ensemble of walkers are distributed across all levels of the ladder. PT is, therefore, naturally an ensemble-based approach. Appendix C contains a pseudo-code representation of a basic PT algorithm which illustrates the main idea. For PT the within-chain steps, updating model parameters $\mathbf{m}$, are unchanged from ST and SA, and so again the standard McMC Metropolis algorithm is used to sample the respective conditional distribution $\pi(\mathbf{m}|T_i)$. The difference between ST and PT lies in the nature of the between-chain steps, that is, transitions from one conditional distribution to another. Again between-chain steps are proposed either randomly, or simply alternately to within-chain steps. However, rather than a single walker moving either up or down a level in the ladder as in ST, in PT the between-chain step consists of a swap of models at two neighbouring temperature levels, a process referred to as an 'exchange swap'. We write

$$(\mathbf{m}_i, T_i), (\mathbf{m}_j, T_j) \rightarrow (\mathbf{m}_i, T_j), (\mathbf{m}_j, T_i), \tag{9}$$

where $\mathbf{m}_i$ and $\mathbf{m}_j$ are the model parameter vectors in chains i and j immediately before the proposed swap. Fig. 2(c) illustrates the situation. Pairs of neighbouring chains are seen to always swap

together. As shown in Appendix A, the probability, $\alpha(i, j)$, that an exchange swap between models $\mathbf{m}_i$ and $\mathbf{m}_j$ at temperature levels T_i and T_j, respectively, should be accepted is

$$\alpha(i, j) = 1 \wedge \left\{ \frac{\tilde{p}(\mathbf{m}_i|\mathbf{d})^{1/T_j} c(T_i) q(i|j)}{\tilde{p}(\mathbf{m}_i|\mathbf{d})^{1/T_i} c(T_j) q(j|i)} \times \frac{\tilde{p}(\mathbf{m}_j|\mathbf{d})^{1/T_i} c(T_j) q(j|i)}{\tilde{p}(\mathbf{m}_j|\mathbf{d})^{1/T_j} c(T_i) q(i|j)} \right\}. \quad (10)$$

Cancelling the like terms, the detailed balance condition for the swap becomes

$$\alpha(i, j) = 1 \wedge \left[\frac{\tilde{p}(\mathbf{m}_j|\mathbf{d})}{\tilde{p}(\mathbf{m}_i|\mathbf{d})} \right]^{1/T_i} \left[\frac{\tilde{p}(\mathbf{m}_i|\mathbf{d})}{\tilde{p}(\mathbf{m}_j|\mathbf{d})} \right]^{1/T_j}, \quad (11)$$

and hence the troublesome normalizing constants cancel out, leaving $\alpha(i, j)$ dependent on only the values of the unnormalized *a posteriori* distribution in the two chains at the time of the swap. We see then that PT provides a way to sample the combined distributions $\pi(\mathbf{m}|T_i)$, $(i = 1, \ldots, n)$ maintaining equilibrium but without the need to calculate normalizing constants. In the author's view this is it most appealing feature.

For an optimization problem (1), the corresponding expression is derived in Appendix A and becomes

$$\alpha(i, j) = 1 \wedge \exp\{(1/T_i - 1/T_j)(\phi(\mathbf{m}_i) - \phi(\mathbf{m}_j)\}, \quad (12)$$

and again this is a simple term to evaluate requiring only known quantities. It is seen then that the use of the term 'Parallel' in the name does not refer to a need for parallelized computer architecture but rather in the nature of the algorithm itself, and the way it communicates information between otherwise independent Markov chains using the exchange swap process. In practice, it makes a lot of sense to implement PT on paralleled hardware for which it is ideally suited.

In a Bayesian sampling problem our interest is only in the distribution at $T = 1$; however, by augmenting the space in this way the ensemble of models collected at $T = 1$ will have been cycled through all other temperature levels each of which allows considerably more exploration of the parameter space. This should, in principle, improve 'mixing' of the Markov chain and hence efficiency of convergence. In an optimization framework this translates to an improved ability to escape from entrapment in local minima. Examples of applications appear below.

A clear difference between PT and SA in an optimization context, is that the latter by definition starts at high temperatures and hence more exploratory parameter search and moves to lower temperatures and more localized search. In contrast, with PT effort is spread across all temperatures at all times and hence while low temperature chains are exploring locally, higher temperature chains are exploring more globally and communication between all changes continues at a constant rate throughout. One might argue that PT would be less efficient for some problems since effort (i.e. objective function evaluations) is continually expended in the more exploratory higher temperature sampling, regardless of the value of the objective function. This may well be true in some cases, however, all direct search optimization algorithms are a trade-off between efficiency, that is, how quickly one gets to an acceptable answer, and robustness, that is, the likelihood of not getting an answer at all (due to entrapment in secondary minima). In SA the search is initially expansive and finally concentrated which will be efficient provided the rate of transition between the two states is appropriate for the particular problem. The harder the optimization problem, the more carefully the cooling schedule of temperatures needs to be chosen, otherwise entrapment in local minima will result. In PT these issues are avoided because the balance of effort between higher temperature exploratory search and lower temperature localized search is kept constant throughout. Thereby, at least, providing the potential that sampling can always climb out of deep wells in the misfit landscape, because some chains are never stuck there in the first place. An example of this aspect appears in the next section.

The first applications of search algorithms involving exchange swaps were to protein folding, then under the title of replica exchange molecular dynamics (Swendsen & Wang 1986; Sugita & Okamoto 1999; Habeck *et al.* 2005). In recent years, use of the PT algorithm has become widespread across a number of fields including chemical physics (Falcioni & Deem 1999), Bayesian statistics (Brooks *et al.* 2011) and gravitational wave astronomy (Cornish 2012). More recently, the first applications appeared in ocean acoustics, (Dettmer & Dosso 2012, 2013; Dosso *et al.* 2012). We are not aware of any applications in solid earth geophysics. Further discussions of tempering algorithms can be found in Li *et al.* (2004), Earl & Deem (2005) and Geyer (2011). In the next section, we illustrate PT through some numerical examples. Issues addressed include the number and distribution of temperatures in the ladder and its affect on performance.

3 NUMERICAL EXAMPLES

3.1 A bi-modal toy problem

As a first illustration of PT, we draw samples from a bi-modal PDF given by

$$\pi(x) = 2^{-x} + 2^{-(100-x)}, \quad (13)$$

which is shown in Fig. 3(a). This has two peaks of value $\pi(x \approx 1$ at $x = 0$ and 100 separated by an extremely low probability region in between where $\pi(x) \approx 10^{-15}$. This function was used by Atchadé *et al.* (2011) to study optimal temperature ladders in PT. It is a simple way of illustrating the 'mixing problem', where McMC random walkers at one peak have considerable difficulty in moving to the other. In a Bayesian setting, the test function presents a challenge because we need to sample all significant probability peaks for meaningful results, while in optimization we want to explore all significant peaks to ensure an optimal solution, that is, higher values of $\pi(x)$. In this example, the x-axis is discretized into 101 points, $x_i = i$ $(i = 0, \ldots, 100)$. The within-chain steps of the Metropolis–Hastings walker consist of a randomly chosen proposal to either increase or decrease the walker position, x_i, by one unit. This is represented by a proposal distribution $q(x_j|x_i)$, which is the probability of the chain moving from position x_i to x_j and given by

$$q(x_j|x_i) = \begin{cases} 1/2 & : \quad j = i \pm 1, \\ 1 & : \quad i = 0, j = 1 \text{ or } i = 100, j = 99, \\ 0 & : \quad \text{otherwise.} \end{cases} \quad (14)$$

According to the Metropolis–Hastings rule, this proposal is accepted with probability

$$\alpha_{ij} = \frac{\pi(x_j) q(x_i|x_j)}{\pi(x_i) q(x_j|x_i)}. \quad (15)$$

Fig. 3(b) shows the position of the (non-tempered) McMC random walker produced in this way. The walker starts in the left-hand peak at $x = 0$, and, despite numerous attempts, after 10 000 steps is unable to move far from its initial position. This is because the extreme low probability region between the two peaks forms a barrier to

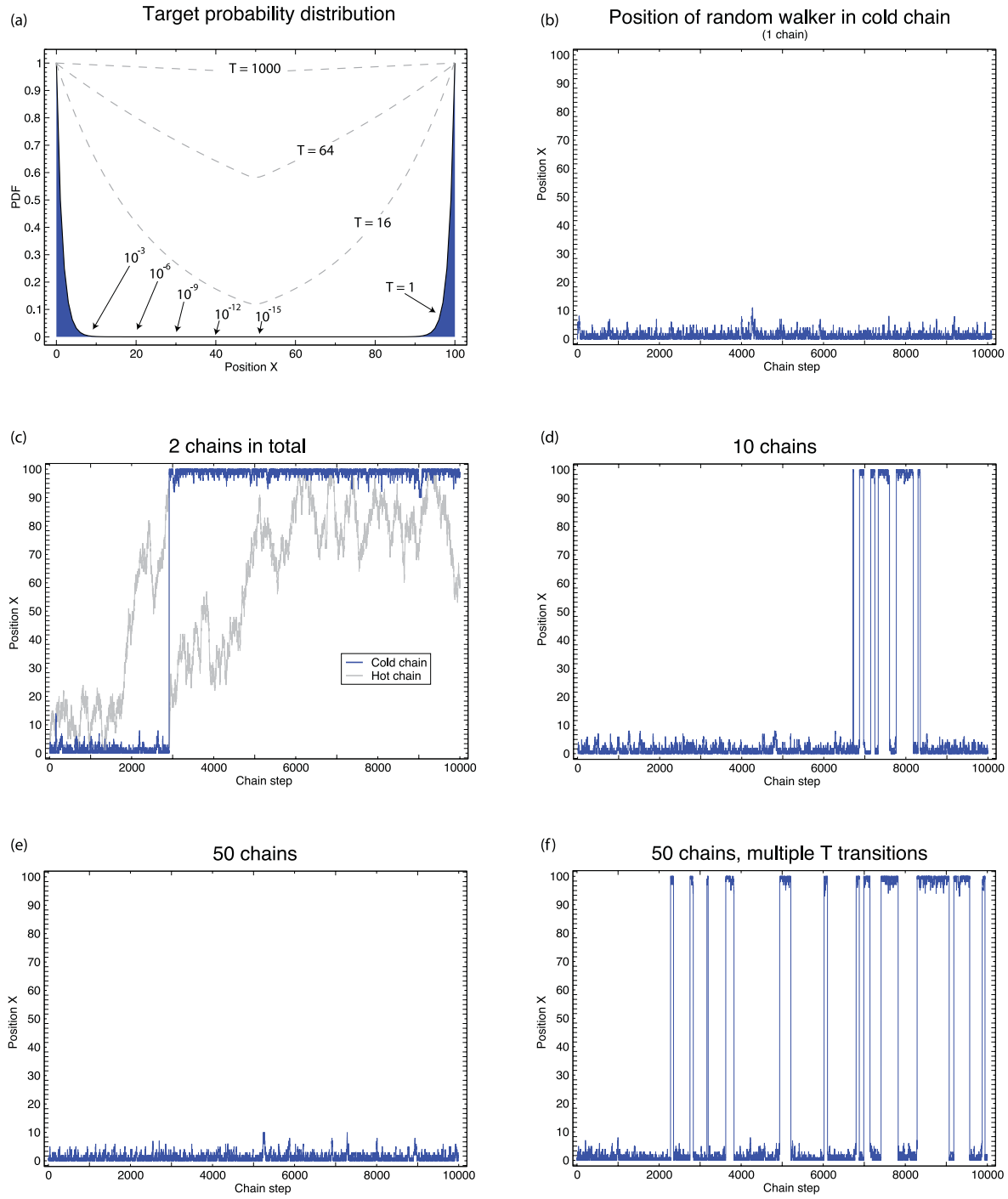

Figure 3. (a) Target probability density function for the bi-modal test problem corresponding to (13). Random walkers have difficulty passing through the channel of low probability separating the two peaks. Dashed lines indicate tempered distributions for $T = 16$, 64 and 1000. (b) Position, x, of non-tempered random McMC walker (y-axis) as a function of chain step (x-axis). The random walker starts in the left-hand peak at $x = 0$ but is unable to move far away even after 10 000 samples. (c) Positions of two random walkers at $T = 1$ (cold chain) and $T = 1000$ (hot chain) with exchange swaps allowed between them. The cold chain is now able to traverse the low probability region and reach the second high probability peak at $x = 100$. For reference, the grey line shows the hot chain ($T = 1000$) which moves more freely around the space. Panels (d) and (e) show position of the cold chain with 10 and 50 temperature levels, respectively, between $1 \leq T \leq 1000$, allowing exchange swaps between neighbouring temperatures (see Fig. 2c). For $n = 10$, the cold chain transitions multiple times between the peaks, but these disappear for $n = 50$. (f) A repeat of case (e) at the same 50 temperature levels only with exchange swaps permitted between all chains, as shown in Fig. 2(d). Traversals are observed and at an increased rate over the 10 temperature case.

transition. Fig. 3(c) shows the situation with two tempered chains. The cold chain at $T = 1$ 'sees' the same $\pi(x)$ given by (13) as in the previous case, but now is able to perform exchange swaps with the hot chain at $T = 1000$ which wanders about virtually uniformly across the domain (see grey curve in Fig. 3a). Here, within-chain steps perturbing x and exchange swaps between temperature levels are proposed alternately. After about 3000 chain steps a successful transition occurs from the peak at $x = 0$ to the one at $x = 100$

because of an accepted exchange swap between the two chains. This simple example demonstrates the power of the exchange swap process in enabling the cold chain to pass across regions of extreme low probability in parameter space.

Figs 3(d) and (e) show the situation with a ladder of $n = 10$ and 50 temperature levels, respectively. Temperatures are distributed according to the formula $T_i = 10^{3(i-1)/(n-1)}$, $(i = 1, \ldots, n)$ as proposed by Atchadé *et al.* (2011). For the 10 temperature level case, the cold chain transitions multiple times between the two peaks in the first 10 000 steps suggesting that smaller temperature jumps and more levels improves the ability of the cold chain to move about the space compared to the $n = 1$ case. However, with 50 temperature levels the transitions between peaks is absent even though there are five times as many chains present.

One might expect that an increased numbers of chains and temperatures would mean better sampling of the temperature variable and hence better performance. However, this is not the case because an increase in the number of temperature levels means a decrease in the gap between temperatures. While the probability of a successful exchange swap is increased as the temperature gap decreases, at the same time the cold chain needs to undergo many more successful swaps in order to reach the hottest chain. It turns out that these two effects do not simply trade-off with one another. In fact, it is a generally observed feature of PT that the ability of the cold chain to move around the parameter space will initially be improved as the number of temperature levels is increased, because of higher acceptance probability per swap, but ultimately inhibited, because of the increased number of successful swaps needed (Earl & Deem 2005).

This effect is seen in our example. Fig. 4 shows a plot of the average distance moved by the cold chain (a proxy for algorithm efficiency used by Atchadé *et al.* 2011) as a function of the number of temperature intervals. As can be seen, the cold chain initially increases in efficiency with n and then decreases again. Atchadé *et al.* (2011) cite this as motivation for the need to tune the size of temperature gaps to allow the cold chain to efficiently move about the space. Indeed, there has been considerable focus on the tuning of temperature ladders in both PT and ST (Geyer & Thompson 1995; Kofke 2002, 2004; Pedrescu *et al.* 2004; Kone & Kofke 2005).

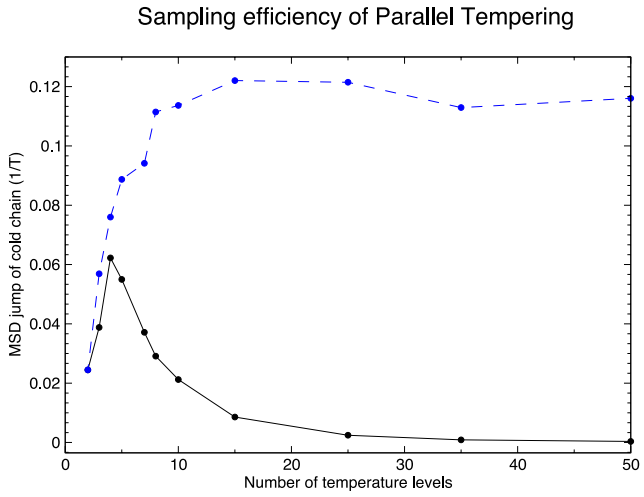

Figure 4. Sampling efficiency of a Parallel Tempering algorithm for the bi-modal problem in Fig. 3(a) as a function of the number of temperature levels. Here efficiency is represented by the size of the mean square jump distance of the cold chain in units of inverse temperature (see Atchadé *et al.* 2011, for a discussion). The solid curve corresponds to the standard case where transitions between temperatures are restricted to neighbouring temperature levels, and the dashed curve where transitions are allowed between all temperature levels. Increased efficiency is achieved by allowing transitions between all temperature levels, and, in this case, finding an appropriate number of temperature levels is relatively straightforward.

In our experiments we found that the situation can be remedied with a simple adjustment to the standard PT algorithm. Rather than restricting exchange swaps to neighbouring temperature levels, we relax this condition and allow exchange swaps to be proposed between any pair of levels randomly. The situation is depicted in Fig. 2(d). This simple change increases the number of possible swap pairs from $N - 1$ to $\frac{1}{2}N(N-1)$ but has no effect on the computational effort because the total number of swaps are unchanged. Nevertheless, the result is a significant improvement in performance. Fig. 3(f) shows a repeat of the 50 temperature level example in Fig. 3(e) allowing proposed swaps between any pair of temperature levels. The result is dramatic with multiple and regular transitions of the cold chain between the two peaks in the PDF. Fig. 4 shows a plot of the average distance moved by the cold chain (in units of inverse temperature) with number of temperature levels for the two cases of restricted (solid) and unrestricted (dashed) temperature jumps. As is clearly seen, the efficiency of the unrestricted case increases with n in a simple fashion and avoids the need for the careful tuning apparent in the restricted case. In all of our experiments we choose n after some experimentation in this way. Since most implementations of PT are likely to be on parallelized hardware, there is little additional cost in increasing the number of temperature levels and hence tuning can be straightforward. Dettmer & Dosso (2012) also implement PT by allowing transitions between multiple temperature levels and found it to be beneficial in geoacoustic inverse problems.

3.2 Parallel Tempering for probabilistic sampling

To illustrate the effectiveness of PT on a more challenging problem, we applied it to the trans-dimensional inversion of seismic receiver functions (Bodin *et al.* 2012b). In a trans-dimensional inversion, the number of unknowns is also unknown and an ensemble of solutions can be obtained by sampling with a Markov chain over a variable dimension parameter space (see Sambridge *et al.* 2006; Hopcroft *et al.* 2007; Gallagher *et al.* 2009, 2011; Dettmer & Dosso 2012; Bodin *et al.* 2012a, for some recent examples). One aspect of this problem is the need for efficient sampling, particularly in the model dimension parameter. The problem setup is illustrated in Fig. 5(a) which is identical to that described by Bodin *et al.* (2012b).

In brief, the parametrization consists of a 1-D shear wave velocity profile as a function of depth made from N_L layers, each of which has a velocity parameter and an interface depth as unknown, $(V_{s,i}, c_i)$, $(i = 1, \ldots, N_L)$, where N_L is a variable. As shown in Fig. 5(a), each interface depth is defined as the midpoint of consecutive nodes at c_i. The likelihood function takes the familiar Gaussian form

$$p(\mathbf{d}|\mathbf{m}) = \frac{1}{\sqrt{(2\pi)^N |C_d|}} \times \exp\left\{-1/2(\mathbf{d} - \mathbf{d}_p(\mathbf{m}))^T C_d^{-1}(\mathbf{d} - \mathbf{d}_p(\mathbf{m}))\right\}, \quad (16)$$

where $\mathbf{d}$ is the synthetic receiver function, $\mathbf{m}$ the model, $\mathbf{d}_p(\mathbf{m})$ the predicted receiver function from the model, C_d is the noise covariance matrix and N is the number of data values (= number of samples in each receiver function × number of receiver functions).

The data noise covariance matrix represents the standard deviation of the noise and correlation parameters in time. The noise

standard deviation parameter is also treated as an unknown, σ and we write

$$C_d = \sigma^2 \tilde{C}_d, \tag{17}$$

where $\tilde{C}_d$ is the correlation matrix of the synthetic data. In this synthetic problem, the true model is the 6 layer shear wave velocity profile shown in Fig. 5(b) representing a total of 12 model parameters plus the data noise standard deviation. The receiver function with noise added, Fig. 5(c) acts as the data for our experiments and was calculated with the algorithm of Shibutani *et al.* (1996) with true noise value of $\sigma = 0.01$. The matrix $\tilde{C}_d$ represents the temporal correlation of the noise and is non-diagonal. Here, we fix it at the true value, meaning that only the size of the noise, σ, is sampled over and not the correlation parameters. The *a priori* PDF is uniform in all variables between fixed bounds which are: $2.0\,\mathrm{km\,s^{-1}} \leq V_{s,i} \leq 5.0\,\mathrm{km\,s^{-1}}$, $0\,\mathrm{km} \leq c_i \leq 60$ km, $10^{-3}\,\mathrm{s} \leq \sigma \leq 10^{-1}$ s. This setup is identical to earlier work. See eqs (4) and (5) of Bodin *et al.* (2012b) for further details.

To illustrate the multimodal character of the data misfit function, we plot the $-$log-likelihood about a randomly chosen point in parameter space. Fig. 6 shows four panels each of which is produced by changing a single model parameter while keeping all other parameters fixed. Figs 6(a) and (b) show the effect on the data misfit, $-\log(p(\mathbf{d}|\mathbf{m})$, by varying the velocity, V_i, of an upper crustal and mid-crustal layer, respectively. Figs 6(c) and (d) show how the corresponding node depth parameters, c_i, influence the data misfit. The multimodal character of this conditional likelihood is clearly evident. In the numerical experiments that follow each Markov chain is initiated at a random point derived from a uniform prior and so the landscapes seen in Fig. 6 are likely to be typical of that experienced during early stages of the McMC chains. From a sampling perspective, such likelihood functions are reminiscent of the toy problem in Fig. 3(a), with peaks replaced by troughs, and represent a challenge in drawing unbiased samples. From an optimization perspective, the multimodal character would largely preclude the use of 'downhill' gradient-based algorithms due to a likely entrapment in secondary minima.

The Bayesian inference problem is to sample the multimodal *a posteriori* PDF, $p(\mathbf{m}|\mathbf{d})$, for this problem setup with a variable number of layers. The birth–death McMC implementation of Bodin *et al.* (2012b) is used to do this and constitutes the within-chain sampler. In this case, perturbations of the model consist in equal proportion of a layer birth, layer death, as well as change of the velocity parameter within a layer using a Gaussian proposal distribution. The mean of the proposal distribution is equal to the velocity parameter value immediately prior to the perturbation, and the standard deviations are tuned *a priori* as proposed by Rosenthal (2000).

In all the examples presented here, we choose to only temper the likelihood function, $p(\mathbf{d}|\mathbf{m})$, rather than the full *a posteriori* PDF, $p(\mathbf{m}|\mathbf{d})$. This is because explicit evaluation of the prior PDF was not required in the sampling algorithm of Bodin *et al.* (2012b), which constitutes the within-chain sampler. In trans-dimensional inversion, the prior PDF may significantly influence the number of unknowns in the model and so without tempering we expect a more limited influence on the ability of the Markov chain to jump dimensions. An explicitly defined prior for this 1-D spatial problem is possible (see Hopcroft *et al.* 2007; Steininger *et al.* 2013), and with this tempering could be applied. With no tempering of the prior, the acceptance term (11) becomes

$$\alpha(i,j) = 1 \wedge \left[\frac{\tilde{p}(\mathbf{d}|\mathbf{m}_j)}{\tilde{p}(\mathbf{d}|\mathbf{m}_i)}\right]^{1/T_i} \left[\frac{\tilde{p}(\mathbf{d}|\mathbf{m}_i)}{\tilde{p}(\mathbf{d}|\mathbf{m}_j)}\right]^{1/T_j}. \tag{18}$$

Our primary interest is to examine the effect of introducing exchange swaps into the ensemble of random walkers that would otherwise be independent. To do this we performed an experiment with 380 tempered McMC chains with 25 per cent at $T = 1$ and the remainder with temperatures generated randomly according to a log-uniform distribution in the range $1 \leq T \leq 1000$. The upper limit in T was chosen somewhat arbitrarily but is sufficiently large to ensure that the corresponding walker is uniformly random. All

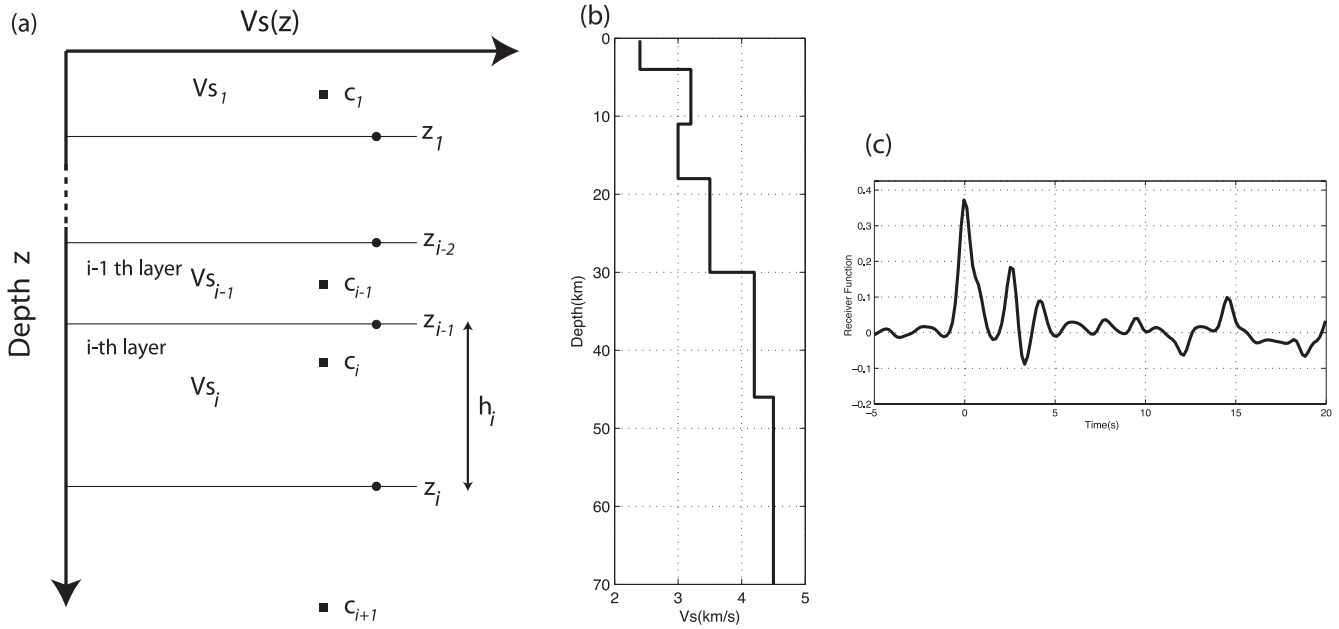

Figure 5. (a) Parametrization of the trans-dimensional 1-D velocity profile. Layers are represented by nodes at position c_i spaced equidistant between interfaces at depths $z_i = \frac{1}{2}(c_{i+1} + c_i)$. The ith layer has velocity parameter $V_{s,i}$; (b) shows the 1-D velocity profile used to calculate the synthetic receiver function in the right panel. (c) Noise is added using the approach of Shibutani *et al.* (1996) and this becomes the synthetic data for the numerical example.

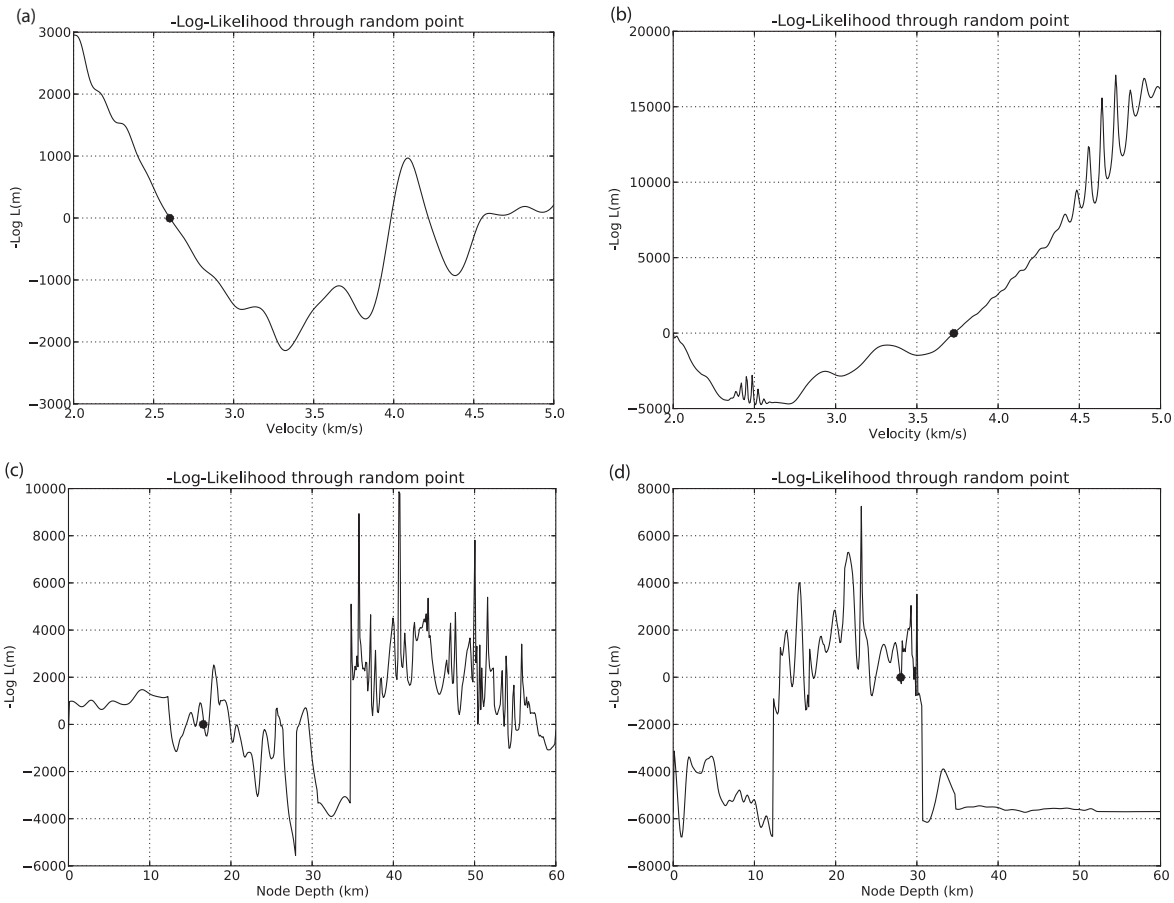

Figure 6. Profiles of changes in the negative log-likelihood (or data misfit) through a randomly chosen point in parameter space for the receiver function problem described in Fig. 5. Panels (a) and (b) show the effect on the misfit when a single velocity parameter in the model is varied and all other parameters are kept fixed. The randomly chosen reference model has values of 2.60 km s^{-1} and 3.73 km s^{-1}, respectively. Panels (c) and (d) are the same but for two interface depth parameters with the random model at 16.68 km and 27.96 km, respectively. The location of the reference model is indicated by a solid dot. The character of the likelihood surface varies dramatically from multimodal, but relatively smoothly varying, as in (a), to combinations of smooth and high-frequency oscillations, (c) and (d).

chains were initiated randomly using the prior. In the first experiment, exchange swaps are only allowed after 10^5 McMC steps have been completed in order to examine the effect of PT.

Fig. 7 shows results of some average properties of the cold chains at $T = 1$. In the upper panel, the average $-$log-likelihood (or receiver function misfit) is plotted as a function of chain step, while the middle panel shows the average number of layers in the trans-dimensional Markov chains and the lower panel the standard deviation of the receiver function noise. Fig. 7 shows that in all three cases the chains begin to pass through the initial 'burn-in' phase, typified by a reduction in average data misfit. For the first 10^5 chain steps no exchange swaps are allowed and all chains work independently. In this stage then standard McMC is being used. PT begins at 10^5 steps and a significant influence on the Markov chain is observed as the chains start to communicate through exchange swaps. It is seen that both the $-$log-likelihood and the average data noise parameter decrease virtually instantaneously. Changes in the average number of layers are more subtle, but in all three cases the variance of the mean also becomes larger, suggesting more mixing within the chains. This result suggests that the convergence of the McMC process is significantly accelerated by PT.

Results of a second experiment are shown in Fig. 8. In this case, the average $-$log-likelihood is compared between two separate runs. The solid black curve shows a PT run from Fig. 7, only now with exchange swaps present from the start and the temperature range $1 \leq T \leq 50$. The grey curve is the average of 380 independent McMC chains all at $T = 1$. We emphasize here that the total amount of work in each ensemble (represented by the number of chain steps x number of chains) is identical between these two, however, PT shows markedly improved convergence. An estimate of the gradient of McMC curve in Fig. 8 after 4×10^5 steps suggests that approximately 10 times the number of chain steps would be required to reduce the average $-$log-likelihood to the value of the PT at 4×10^5 steps. In terms of this measure, convergence of PT is at least 10 times faster than the non-tempered McMC.

Fig. 9 contains Bayesian *a posteriori* PDFs derived from both McMC and PT ensembles. Figs 9(a) and (b) show information on the shear wave velocity as a shaded image of stacked marginals. The left panel in each set shows marginals of the *a posteriori* PDF of shear wave velocity profiles, V_s, aligned in depth with warmer colours representing higher relative probability. The marginal image appears to be better resolved with PT sampling than non-tempered McMC which is considerably more blurred. The middle panels show the peak of each marginal represented as a single velocity depth model. In each case the thicker piecewise curve is the true one. Again the PT result is closer to the truth than that of non-tempered McMC although the difference is less pronounced. Since the interface depth position is variable, then in the right-hand panel,

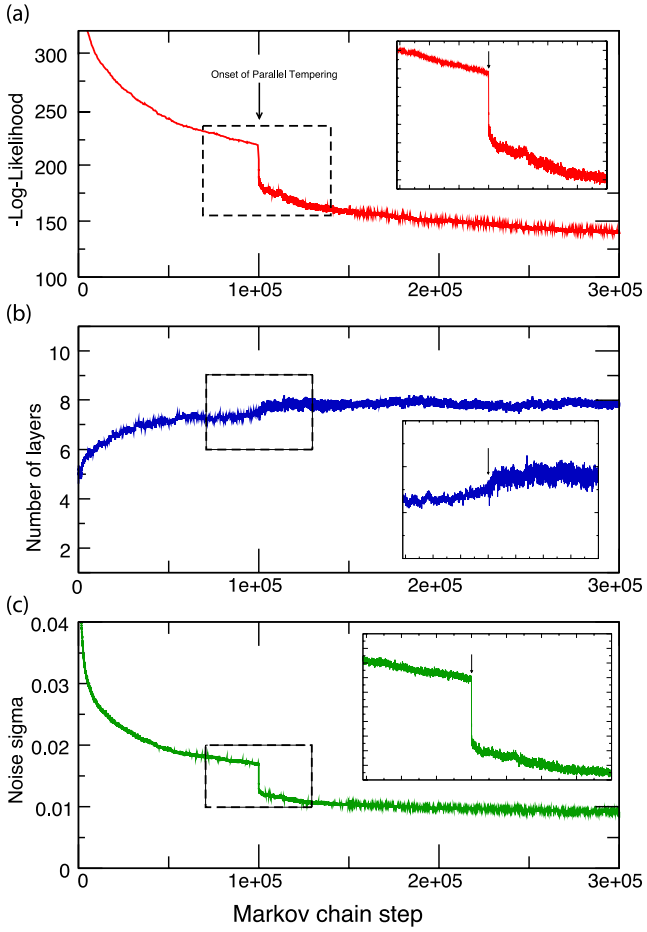

Figure 7. Comparison of average convergence properties for non-tempered versus tempered Markov chains. Results show averages of the 95 chains at temperature, $T = 1$. Panel (a) is the negative log-likelihood; (b) is the number of layers in the earth model and (c) is the estimated standard deviation of the noise. In all cases, the horizontal axis is the number of steps of the Markov chain. Exchange swaps between chains are only allowed after 1×10^5 steps have elapsed. The insets show more detail in the region of the transition between non-tempered McMC and Parallel Tempering. The effect of allowing exchange swaps between chains clearly produces a dramatic acceleration of convergence which is reflected in the behaviour of all three parameters.

the *a posterior* PDF profile for the interface depths is shown. Again the PT appears better converged with PDF peaks at the true depths more pronounced than in the McMC case. Fig. 9(c) shows the *a posteriori* marginal PDF of the number of layers and the standard deviation of the data noise, both of which are peaked about the true solutions of $N_L = 6$ and $\sigma = 0.01$, respectively. For reference Fig. 9(d) shows the fit of the original receiver function dashed and that calculated from the model within the ensemble with maximum *a posteriori* PDF. The closeness of fit indicates that the chain has converged. All of the properties of the ensemble displayed suggest that the use of PT has accelerated convergence of the Markov chains.

As noted above, PT is naturally a multichain technique, whereas McMC could be run as a single chain. An alternate comparison with equal numbers of likelihood evaluations would be between the PT results as in Fig. 8 and a single McMC chain run for 380 times as long. By that stage one might well expect the McMC chain to have also converged, in which case one could argue that McMC is able to do equally as well as PT. This would be true and in fact we have not done that experiment because it is impractical to

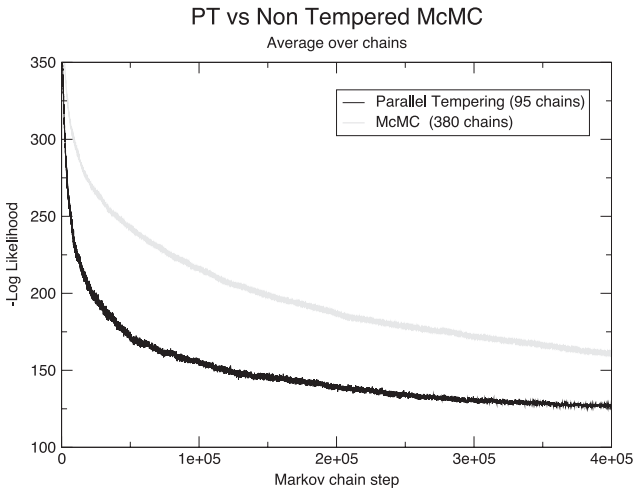

Figure 8. Negative log-likelihood as a function of chain step for two sampling algorithms. The grey curve is the average of 380 non-tempered McMC chains (all at $T = 1$) and black is an average of 95 tempered chains ($T = 1$) when exchange swaps are allowed between 380 chains with temperatures spanning the range $1.0 \leq T \leq 50.0$. The Parallel Tempering converges at least 10 times faster in this case. See text for details.

do so. Calculations here are feasible because they are performed on parallel computer architectures which ideally suits a multichain framework, because each chain can be handled simultaneously by a separate processor. A single McMC chain run for 380 times as many steps could not take advantage of hardware parallelism and hence would take 380 times as long on a single processor which is a significant disadvantage. Furthermore, experience shows that multiple independent McMC chains are usually more robust than single chains in that the latter can experience exponentially long wait times sampling deep secondary maxima in the PDF.

Fig. 10 shows results of a third experiment with trans-dimensional sampling where the McMC and PT Markov chains are initiated at the same model, which is the maximum likelihood (ML) model from the second experiment. Here the number of layers of the shear wave model is displayed for a single random walker. Separate panels show how the dimension parameter changes for the non-tempered McMC walker (upper) and PT (lower) over a window of the cold Markov chain. By initiating the two chains from the same point we can examine the effect of dimension mixing. Ideally, one would prefer higher mixing rates which means the chain successfully transitions to as many different values of N_L, which are consistent with the data and prior PDF. As can be seen the frequency of dimension changes is much higher for PT than McMC. For this window the PT chain changes dimension 90 times over the range $8 \leq N_L \leq 14$, whereas the non-tempered McMC changes dimension 10 times between $12 \leq N_L \leq 13$. As noted above, these results are achieved without tempering the prior PDF which suggests that the broader dimension sampling is driven in response to the data. This result demonstrates the superior dimension mixing ability of PT, which is consistent with the results of Dettmer & Dosso (2012) who observed a similar effect.

Fig. 11 shows a matrix representing the rates of successful transitions between temperature levels for the PT algorithm in the second experiment. Here the temperature range of 1–50 is divided into 16 regularly spaced bins on a log-uniform axis. The intensity of shade in each pixel represents the success rate of exchange swaps between chains whose temperatures are in the respective bins. Fig. 11 appears to show a healthy degree of mixing across all

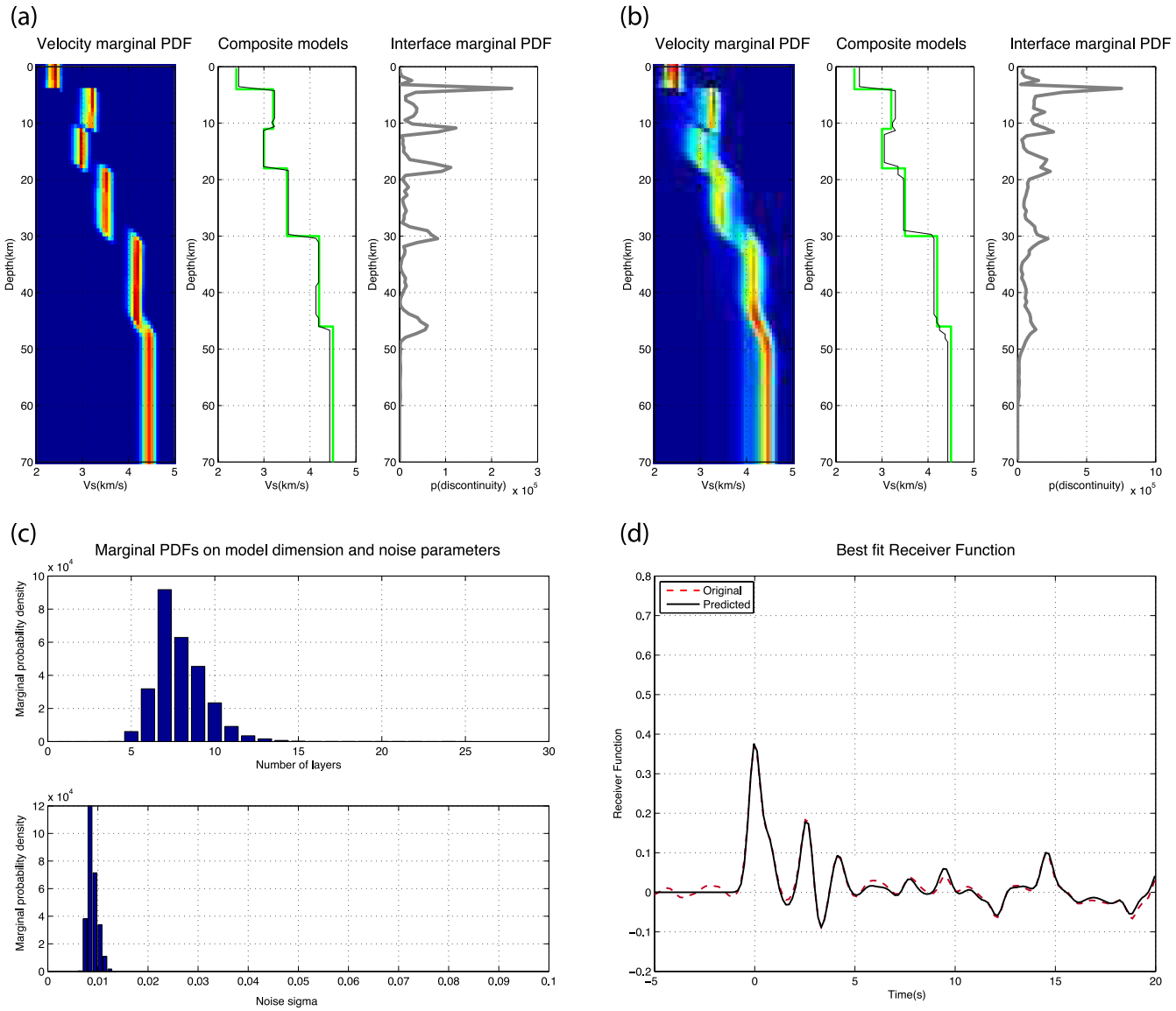

Figure 9. Results of the two sampling algorithms shown in Fig. 8(a) shows the case for the Parallel Tempering algorithm and (b) for the non-tempered McMC sampler. Both after 300 000 McMC steps, following 100 000 burn-in steps. Figures are similar in format to that of Bodin *et al.* (2012b). Displayed are marginal PDFs of the shear wave velocity (left panels); peaks of those marginal PDFs compared to the true solution (middle panels) and PDFs of the interface depths (right panels). The PT solution is better resolved indicating more efficient sampling of the parameter space. Parts (c) and (d) show additional results for the PT case. Panels show the *a posteriori* marginal for the numbers of layers with true value of 6, the noise parameter marginal centred on true value 0.01 s and the fit of the maximum likelihood model (solid) compared to the original noisy receiver function (dashed).

temperature levels. High values in the off-diagonal positions also indicate that many successful exchange swaps have occurred between non-neighbouring temperature levels, which provides some justification for allowing this to happen in the first place. As the chain progresses the models in each chain are likely to move across all temperature levels and thereby more rapidly explore the parameter space.

3.3 Multimodal optimization

The final numerical example is a demonstration of PT applied to global optimization. Here the setup is described in Fig. 12, where the data consist of 18 receiver functions recorded at points along a 2-D profile, as shown in Fig. 12(b), and we seek to recover the best fitting laterally varying 2-D velocity model shown in Fig. 12(a). Here the parametrization of the model consists of six 1-D piecewise constant control models. As before, these are parametrized by a combination of depth nodes and shear wave velocities pairs, $(V_{s,i}, c_i)$, with one per layer. At locations between each pair of control profiles the shear wave velocity is constructed via linear interpolation, while beyond the first and last control model no lateral gradients are assumed. An approximation is introduced to the governing physics so that the receiver function at any point along the profile is only dependent on the 1-D shear wave model immediately beneath the receiver location. This has the effect of allowing standard, and rapid, 1-D synthetics to be used.

The 18 receiver locations are indicated by the dots along the surface of the profile in Fig. 12(a). Overall, there are 36 structural parameters representing velocities and interface depths, which are again controlled by nodal parameters in each layer, as in Fig. 5. The influence of the control models on the receiver locations is

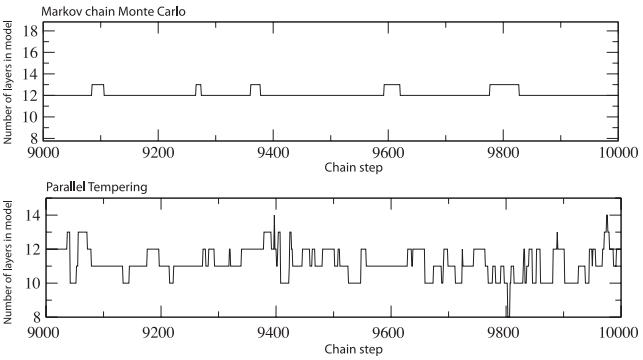

Figure 10. Comparison of dimension mixing ability of a non-tempered McMC chain (upper) and Parallel Tempering(lower) for trans-dimensional receiver function sampling. The y-axis in both cases is the number of layers in a McMC run (upper panel) and PT run (lower panel), while the x-axis is the chain step. The non-tempered chain makes only 10 changes in the number of layers, whereas the tempered chain makes about an order of magnitude more transitions in this window. In this test both chains are initiated at the same best data fit model found by the Parallel Tempering algorithm solution shown in Fig. 9.

indicated by a series of dotted lines in the upper panel of Fig. 12(a). As can be seen, receiver functions depend on at most two control models, while each control model influences between four and five receiver functions. From left to right the control models form a 2-D profile containing several dipping layers in shear wave velocity. The combined effect is a multimodal optimization problem for structural and data noise parameters.

Here, the dimension is fixed and we minimize the likelihood function only. Another simplification from the earlier sampling problem is that we solve for the noise parameter σ by setting its value to the ML estimate throughout, that is, we find σ in (17) which maximizes

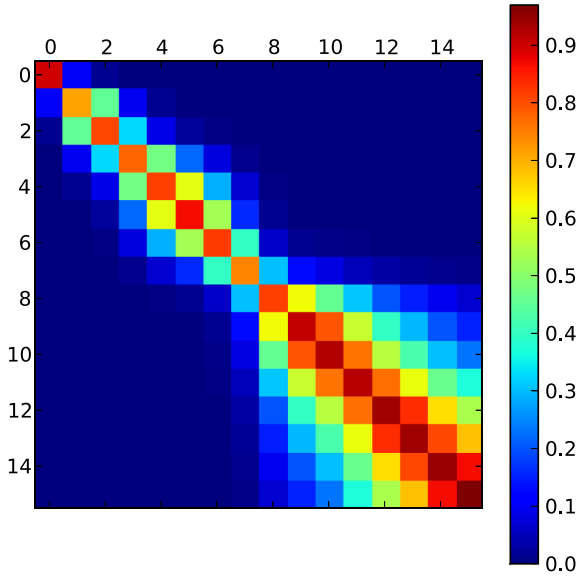

Figure 11. Probability transition matrix between 16 temperature bins for all 380 chains of the Parallel Tempering algorithm in the receiver function example. Temperature bins span $T = 1$–50 and are log-uniformly distributed. Warmer colours indicate higher rates of successful transitions between chains in corresponding bins. Successful jumps are seen to occur between chains at similar temperatures and off-diagonal elements increase in size with temperature indicating significant, and desirable, mixing of the chains across multiple temperature levels.

(16). It can be shown (see Appendix B) that the ML estimate of σ is

$$\sigma(\mathbf{m}) = \left[\frac{1}{N}\mathbf{r}^T\tilde{C}_d^{-1}\mathbf{r}\right]^{1/2}, \tag{19}$$

where the residual vector, $\mathbf{r} = \mathbf{d} - \mathbf{d}_p(\mathbf{m})$ and data vector $\mathbf{d}$ is a concatenation of the 18 receiver functions, Fig. 12(b), and $\mathbf{d}_p(\mathbf{m})$ are the corresponding predictions from the model $\mathbf{m}$. Substitution of (19) into (16) and dropping additive constants gives a modified $-$log-likelihood (data misfit) expression which is independent of the noise parameter σ

$$-\log p(\mathbf{d}|\mathbf{m}) = \frac{N}{2}\log\left(\mathbf{r}^T\tilde{C}_d^{-1}\mathbf{r}\right). \tag{20}$$

For details see Appendix B. This can be a useful substitution for many inverse problems where data variances are poorly known. For our test problem, it reduces the dimension of the parameter space while automatically solving for the noise parameter during the optimization. Previously this approach has been used by Dosso & Wilmut (2006) to good effect for non-linear inverse problems in ocean acoustics (see also Dosso *et al.* 2012).

Parameters of the six control models are sought which minimize misfit between the 18 observed and predicted receiver functions simultaneously, by optimization of (20). Results from our PT algorithm are displayed in Figs 13 and 14. Here, as before, we make use of 380 McMC chains with a temperature ladder spanning $1.0 \leq T \leq 50$, and 95 chains fixed at $T = 1$. Each chain is initiated at a random velocity model calculated from uniform random variables in the range $0\,\mathrm{km} \leq c_i \leq 60$ km for each depth node and $2.0\,\mathrm{km\,s^{-1}} \leq V_{s,i} \leq 5.0\,\mathrm{km\,s^{-1}}$ for each velocity parameter. These are quite wide bounds and hence initial models in each chain are typically very poor fits to the data.

Within-chain steps were performed with the same McMC algorithm used in previous experiments. Exchange swaps were proposed uniform randomly between all pairs of temperature levels. Fig. 13 shows three curves of the data misfit as a function of chain step. The solid 'staircase' curve is the lowest value of the $-$log-likelihood according to expression (20) as a function of chain step; the thick black line is the average value over the cold chains and the grey curve is the average over chains with temperatures in the range $1.5 \leq T \leq 2.5$. Note that due to the form of (20) values of $-$log-likelihood can become negative when data residuals, $\mathbf{r}$, are small. Fig. 13 shows that while the optimum model is largely converged after 25 per cent of the steps, the ensemble of models at $T = 1$ show some statistical fluctuations and continue to reduce throughout. As expected, models at higher temperature have notably slower convergence due to the tempered Markov chain being more explorative.

Fig. 14(a) shows best fit solutions found for the six control models (solid curves) together with the true solutions (dashed), while Fig. 14(b) contains a comparison of the corresponding predicted and original receiver functions. All synthetic data is fit very well in this case and the recovered models are close to true values. Given that the true model is unlikely to be at the global minimum of the objective function (20) due to added noise, the fit is sufficiently good to conclude that the algorithm has found the global solution. Overall, this example shows that PT is able to optimize complicated multimodal functions.

4 DISCUSSION

The results of experiments presented here indicate that exchange swaps between tempered McMC chains is an effective mechanism

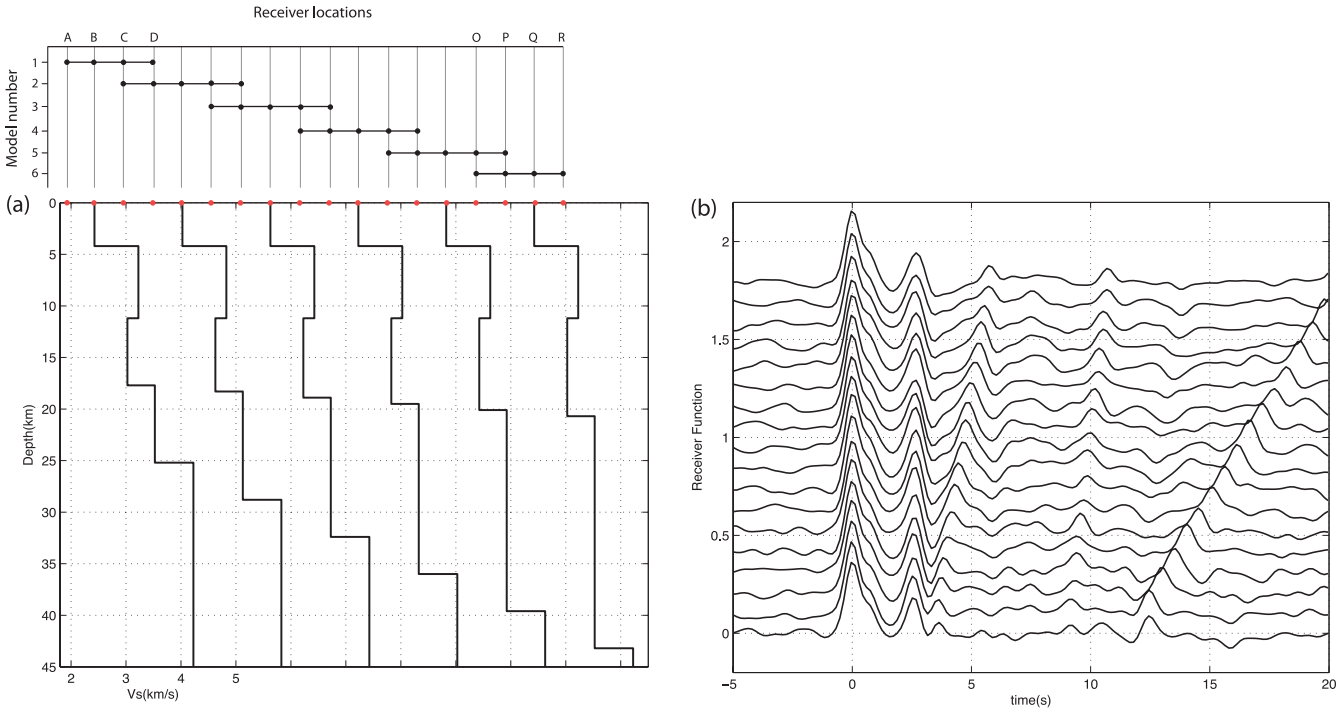

Figure 12. (a) 2-D profile of the laterally varying wave speed model used in the optimization test. The six 1-D control models are shown as a function of position. Velocity values and interface positions vary linearly along the profile between control models. The velocity axis applies to the first model only. Depth range is not to the full extent of the model. Locations of 18 receivers are represented by dots at zero depth. (b) Synthetic receiver functions with added noise calculated at the 18 receiver locations. Each receiver function is calculated from the 1-D model beneath its location, which is determined through linear interpolation of neighbouring control models. The interdependence of earth models and receiver functions is indicated by horizontal bars in the panel above part (a).

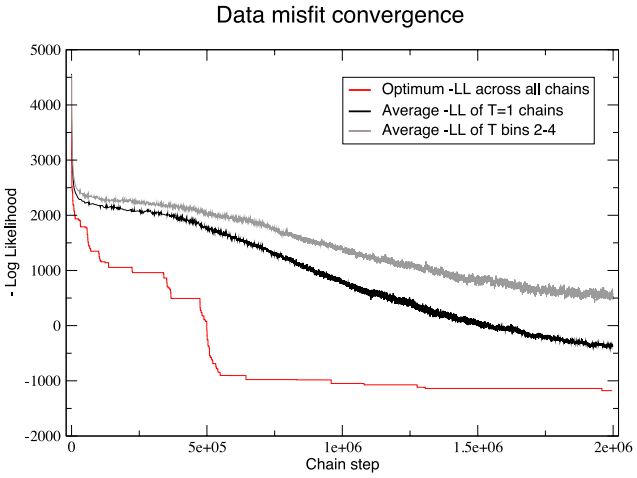

Figure 13. Convergence of negative log-likelihood as a function of chain step for the PT algorithm. The solid (staircase) curve is the lowest data misfit across all chains. The black line is the average data misfit across the 95 chains at the lowest temperature ($T = 1$). The grey line is the average data misfit across chains in temperature bins 2–4, corresponding to temperatures $1.5 \leq T \leq 2.5$. The optimum model converges by 5×10^5 steps, whereas, as expected, the convergence rates of the ensembles are ordered inversely with temperature because higher temperatures are more explorative of the parameter space.

for increasing efficiency in sampling algorithms, as well as a novel approach to global optimization. Several numerical examples are presented to illustrate the central idea. These are intended to be illustrative. In particular, it is not argued that PT should necessarily replace alternate approaches for the inversion of receiver functions. Convergence can often be achieved for this problem with existing trans-dimensional McMC samplers, as proposed by Bodin *et al.* (2012b), provided they are run for long enough. The intention here is to demonstrate the potential of PT in optimization and sampling of multimodal functions and thereby encourage further applications to geoscience problems.

An interesting feature of PT is that it is in essence a 'meta'-algorithm, in that it incorporates a McMC sampler at its centre, but is not dependent on any details of that algorithm. This is clearly demonstrated in the pseudo-code representation of the algorithm shown in Appendix C. PT can be applied to fixed or trans-dimensional within-chain sampling, use any form of model perturbation or indeed any form of model parametrization. The exchange-swapping process merely requires multiple chains to exist that are swapped in pairs using the corresponding Metropolis–Hastings rule (11). This has considerable advantages in software construction because libraries can be written that make no assumption about the nature of the parameter space. Real, integer or combinatorial unknowns can be treated equally well. Furthermore, it is always possible to apply an exchange-swapping process to any existing McMC sampler, and hence investment in refining such algorithms for a particular problem is not lost. The generality of PT is an appealing feature.

While we argue that PT may be used to accelerate convergence of McMC algorithms, many other techniques exist which have the same goal. An area of much current focus is the choice of proposal distribution used for the within-chain sampling. This decision has been shown to have considerable effect on convergence rates of McMC algorithms (Dosso & Wilmut 2006). In our experiments, there is a single structural parameter per layer and so a 1-D proposal Gaussian distribution is used to perturb layers independently.

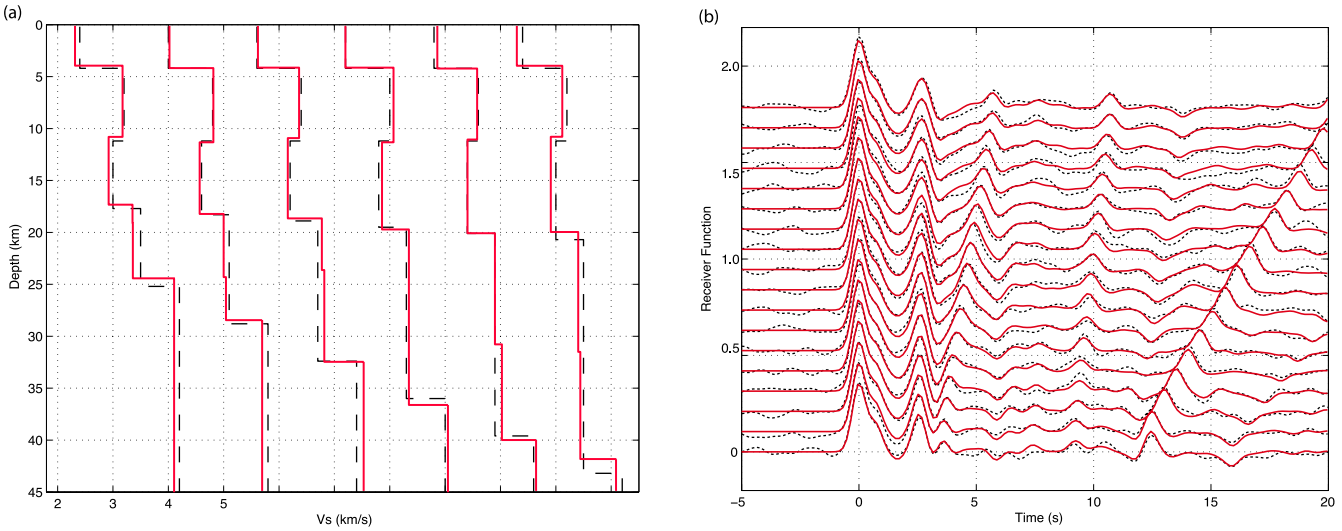

Figure 14. (a) Maximum likelihood solution for the six 1-D control models found by the PT algorithm. Recovered models are solid and are an excellent fit to the original profiles (dashed). (b) Predicted (solid) and original (dashed) receiver functions, showing all major features fit to a high level. All other details are as in Fig. 12.

Alternate proposal distributions include multidimensional Gaussians based on linearized *a posteriori* model covariances (Malinverno 2002; Dosso & Wilmut 2006; Minsley 2011) and those which dynamically detect the local scale and shape of the target distribution during the random walk. Examples of which are delayed rejection (Tierney & Mira 1999; Mira 2001; Haario *et al.* 2006), 'multiple try' or 'snooker' moves which utilize the entire history of ensemble to create model perturbations (Laloy & Vrugt 2012), and schemes suited to cases where the target PDF may be approximated as a mixture of Gaussians (Craiu *et al.* 2009; Bai *et al.* 2011). The focus of this paper, however, is on the effect of introducing exchange swaps into the ensemble and not the nature of the proposal distribution itself. As noted above, PT is independent of the choice of within-chain sampler and hence these adaptive proposal distributions could be used together with a tempering framework. For example, it would make sense to make the within-chain proposal distributions a function of temperature of the chain, something not done in our experiments, so that higher temperature chains tend to propose larger steps in model space while lower temperature chains propose smaller steps.

In optimization, PT is widely applicable. For example, it could be applied to any problem where, say SA is used, thereby taking advantage of the same central Markov chain approach but with parallel exchange swaps present and governed by the acceptance rule (12). More generally, one could imagine applying exchange swaps to any existing numerical algorithm (not just an McMC sampler), which sampled, or optimized, tempered probability distributions in the form of (3) or (1), respectively. In this way, PT could be generalized for use with other direct search optimization methods, however, this aspect does not appear to have been explored to date.

The retention of 'detailed balance' in PT, which keeps the swapping process in equilibrium, is in principle an advantage over SA. However, this does not mean that PT will always outperform SA or ST in any particular application. There are many examples of successful applications of SA to optimization problems in the geosciences, where the lack of equilibrium in the Markov chain as temperatures are reduced is not seriously detrimental to performance. However, in general, too rapid a cooling is known to cause disequilibrium in SA, with entrapment in secondary minima the likely result (Aarts & Korst 1989). PT offers an alternate in these cases.

While we have presented five separate numerical examples showing applications of PT in various situations, these are clearly not exhaustive. For example, we have not provided numerical examples comparing PT with alternate approaches for parameter search and optimization, which is beyond the scope of this paper. In addition to SA and ST, another search algorithm which has found several applications in the geosciences is the neighbourhood algorithm (NA) of Sambridge (1999). In this case a few comments are possible. Specifically, NA, like PT, is an ensemble-based parameter search technique, but one which is restricted to real-valued parameter spaces of fixed dimension. NA has the property of being driven only by a ranking of models in parameter space according to an objective function. Since a change of temperature as shown in (1) does not change the rank of models in an ensemble, then NA is unaffected by tempering. A second observation is that NA has largely found success in problems where the numbers of unknowns is relatively small, say less than 50. PT, on the other hand, may be applied to any problem where SA can be used, and there are applications across the sciences where these run to the hundreds to thousands of unknowns (Sen & Stoffa 2013). The NA and tempering do not appear to be readily combined and one might expect the latter to find application in a broader class of problems.

The central aim of this paper has been to show how the efficiency of randomized sampling is improved through tempering of probability distributions, (3) and (1). Taking a broader view, we can recognize that tempering is in essence just one way of replacing a difficult multimodal sampling, or optimization problem, with a series of less difficult versions. In the author's view, this might be a principal with general applicability for non-linear inverse problems. For example, if it is possible to replace a single difficult problem with a family of related versions of the problem with decreasing complexity, then considering them in unison may pay benefits. For example, in parameter search and sampling, solutions to simplified cases can provide useful starting points for more complex problems. The key feature of PT is that all such problems are tackled at once with information continually exchanging between random pairs.

5 CONCLUSION

A discussion of algorithms to sample-tempered PDFs is presented. Certain multimodal optimization and probabilistic sampling problems common in geophysical treatments of inverse problems may be addressed by sampling an augmented model space consisting of the original model parameters and an additional temperature variable. The technique known as PT is described and illustrated through several numerical examples. A key element of this approach is the use of exchange swaps between pairs of Markov chains each sampling a tempered version of the target probability distribution. These allow sampling of the augmented parameter space while retaining detailed balance and hence convergence to stationary distributions (and global minima). A practical solution is proposed to the question of how to define a temperature ladder upon which PT may be performed, which requires some adjustments to the standard version of the algorithm, to allow transitions between arbitrary, rather than adjacent temperature levels. Results of numerical tests suggest that inclusion of exchange swaps provides significant benefits in terms of acceleration of convergence of Markov chain samplers. Since the tempering framework is independent of the choice of McMC algorithm used to sample the parameter space, it may be combined with the most appropriate sampling algorithm for any given problem. The results here provide encouragement for future applications of PT more broadly within the geosciences, which to date have been almost entirely absent.

ACKNOWLEDGEMENTS

I am grateful to Thomas Bodin for the trans-dimensional McMC sampler code used in the receiver function examples and Jan Dettmer who made useful suggestions in respect of the numerical examples. Thomas Bodin, Jan Dettmer and Kerry Gallagher provided useful feedback on an earlier draft of this manuscript. Constructive reviews of an earlier draft were received from Burke Minsley and Juan Carlos Afonso. Aspects of the research reported here were supported under Australian Research Council Discovery grant scheme (project number DP110102098). Calculations were performed on the Terrawulf III cluster, a computational facility supported through the AuScope Australian Geophysical Observing System (AGOS). Software development implementing the algorithms described here was possible with support from the AuScope Inversion Laboratory. Software for Parallel Tempering is available upon request.

REFERENCES

Aarts, E. & Korst, J., 1989. *Simulated Annealing and Boltzmann Machines,* John Wiley & Sons.

Ammon, C.J., Randall, G.E. & Zandt, G., 1990. On the nonuniqueness of receiver function inversions, *J. geophys. Res.,* **95,** 15 303–15 318.

Aster, R., Borchers, B. & Thurber, C.H., 2012. *Parameter Estimation and Inverse Problems,* 2nd edn, Elsevier Academic Press.

Atchadé, Y.F., Roberts, G.O. & Rosenthal, J.S., 2011. Towards optimal scaling of metropolis-coupled Markov chain Monte Carlo, *Stat. Comput.,* **21**(4), 555–568.

Bai, Y., Craiu, R.V. & Di Narrzo, F., 2011. Divide and conquer: a mixture-based approach to regional adaptation for MCMC, *J. Comp. Graph. Stat.,* **20**(1), 63–79.

Beck, J.L. & Au, S.-K., 2002. Bayesian updating of structural models and reliability using Markov chain Monte Carlo simulation, *J. Eng. Mech.,* **128,** 380–391.

Bina, C.R., 1998. Free energy minimization by simulated annealing with applications to lithospheric slabs and mantle plumes, *Pure appl. Geophys.,* **151,** 605–618.

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

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K. & Rawlinson, N., 2012b. Transdimensional inversion of receiver functions and surface wave dispersion, *J. geophys. Res.,* **117,** B02301, doi:10.1029/2011JB008560.

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

Ching, J. & Chen, Y.-C., 2007. Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging, *J. Eng. Mech.,* **133**(7), 816–832.

Cornish, N.J., 2012. Gravitational wave astronomy: needle in a haystack, *Phil. Trans. R. Soc. Lond., A,* **371**(1984), doi: 10.1098/rsta.2011.0540.

Craiu, R.V., Rosenthal, J. & Yang, C., 2009. Learn from thy neighbor: parallel-chain and regional adaptive McMC, *J. Am. Stat. Assoc.,* **104**(488), 1454–1466.

Curtis, A. & Wood, R., 2004. Optimal elicitation of probabilistic information from experts, in *Geological Prior Information,* Publication 239, pp. 1–14, eds Curtis, A. & Wood,, R., The Geological Society of London.

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,** 2239–2250.

Dettmer, J. & Dosso, S.E., 2013. Probabilistic two dimensional joint water-column and seabed inversion, *J. acoust. Soc. Am.,* **133,** 2612–2623.

Dosso, S.E. & Wilmut, M.J., 2006. Data uncertainty estimation in Matched-Field geoacoustic inversion, *IEEE J. Ocean. Eng.,* **31,** 470–479.

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

Earl, D.J. & Deem, M.W., 2005. Parallel tempering: theory, applications, and new perspectives, *Phys. Chem. Chem. Phys.,* **7,** 3910–3916.

Falcioni, M. & Deem, M.W., 1999. A biased Monte Carlo scheme for zeolite structure solution, *J. Chem. Phys.,* **110,** 1754–1766.

Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M. & Stephenson, J., 2009. Markov chain Monte Carlo (McMC) sampling methods to determine optimal models, model resolution and model choice for earth science problems, *J. Mar. Petrol. Geol.,* **26,** 525–535.

Gallagher, K., Bodin, T., Sambridge, M., Weiss, D., Kylander, M. & Large, D., 2011. Inference of abrupt changes in noisy geochemical records using transdimensional change point models, *Earth planet. Sci. Lett.,* **311,** 182–194.

Garcia, R.H., Tkalčić, H. & Chevrot, S., 2006. A new global PKP data set to study earth's core and deep mantle, *Phys. Earth planet. Inter.,* **159,** 15–31.

Geyer, C.J., 1991. Markov Chain Monte Carlo maximum likelihood, in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* American Statistical Association, New York, pp. 156–163.

Geyer, C.J., 2011. Importance sampling, simulated tempering and umbrella sampling, in *Handbook of Markov Chain Monte Carlo,* pp. 295–311, eds Brooks, S., Gelman, A., Jones, G.L. & Meng, X., Chapman & Hall/CRC.

Geyer, C.J. & Thompson, E.A., 1995. Annealing Markov chain Monte Carlo with applications to ancestral inference, *J. Am. Stat. Assoc.,* **90,** 909–920.

Haario, H., Laine, M., Mira, A. & Saksman, E., 2006. DRAM: efficient adaptive MCMC, *Stat. Comput.,* **16**(4), 339–354.

Habeck, M., Nilges, M. & Rieping, W., 2005. Replica-exchange Monte Carlo scheme for Bayesian data analysis, *Phys. Rev. Lett.,* **94,** 018105–1–018105–4.

Hopcroft, P., Gallagher, K. & Pain, 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.

Kirkpatrick, S., Gelatt, C.D. Jr & Vecchi, M.P., 1983. Optimization by simulated annealing, *Science,* **220,** 671–680.

Kofke, D.A., 2002. On the acceptance probability of replica-exchange Monte Carlo trials, *J. Chem. Phys.,* **117,** 6911–6914.
Kofke, D.A., 2004. Comment on the incomplete beta function law for parallel tempering sampling of classical canonical systems [J. Chem. Phys. 120, 4119 (2004)], *J. Chem. Phys.,* **121,** 1167, doi:10.1063/1.1758211.
Kone, A. & Kofke, D.A., 2005. Selection of temperature intervals for parallel-tempering simulations, *J. Chem. Phys.,* **122,** 206101–1–206101–2.
Laloy, E. & Vrugt, J.A., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM and high performance computing, *Water Resour. Res.,* **48,** W01526, doi: 10.1029/2011WR010608.
Li, Y., Protopopescu, V.A. & Gorin, A., 2004. Accelerated simulated tempering, *Phys. Lett. A,* **328**(45), 274–283.
Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.,* **151,** 675–688.
Marinari, E. & Parisi, G., 1992. Simulated tempering: a new Monte Carlo scheme, *Europhys. Lett.,* **19**(6), 451–458.
Metropolis, N. & Ulam, S.M., 1949. The Monte Carlo method, *J. Am. Stat. Assoc.,* **44,** 335–341.
Minsley, B., 2011. A trans-dimensional Bayesian Markov chain Monte Carlo algorithm for model assessment using frequency-domain electromagnetic data, *Geophys. J. Int.,* **187,** 252–272.
Minson, S.E., Simons, M. & Beck, J.L., 2013. Bayesian inversion for finite fault earthquake source models: I – theory and algorithm, *Geophys. J. Int.,* **194**, 1701–1726.
Mira, A., 2001. On Metropolis–Hastings algorithms with delayed rejection, *Metron,* **59**(3–4), 231–241.
Mosegaard, K. & Sambridge, M., 2002. Monte Carlo analysis of inverse problems, *Inverse Probl.,* **18,** R29–R54.
Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.,* **100,** 12 431–12 447.
Parker, R.L., 1994. *Geophysical Inverse Theory,* Princeton University Press.
Pedrescu, C., Pedrescu, M. & Ciobanu, C.V., 2004. The incomplete beta function law for parallel tempering sampling of classical canonical systems, *J. Chem. Phys,* **120,** 4119–4128.
Press, W.H., Tuekolsky, S.A., Vetterling, W.T. & Flannery, B.P., 1992. *Numerical Recipes in FORTRAN,* Cambridge University Press.
Rosenthal, J., 2000. Parallel computing and Monte Carlo algorithmas, *Far East J. Theor. Stat.,* **4**(2), 207–236.
Rothman, D.H., 1985. Nonlinear inversion statistical mechanics and residual statics corrections, *Geophysics,* **50,** 2784–2796.
Rothman, D.H., 1986. Automatic estimation of large residual statics corrections, *Geophysics,* **51,** 332–346.
Sambridge, M., 1999. Geophysical inversion with a neighbourhood algorithm–I. Searching a parameter space, *Geophys. J. Int.,* **138,** 479–494.
Sambridge, M., Gallagher, K., Jackson, A. & Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.,* **167,** 528–542.
Sen, M.K. & Stoffa, P.L., 2013. *Global Optimization Methods in Geophysical Inversion,* 2nd edn, Cambridge University Press.
Shibutani, T., Sambridge, M. & Kennett, B.L.N., 1996. Genetic algorithm inversion for receiver functions with application to crust and uppermost mantle structure beneath eastern australia, *Geophys. Res. Lett.,* **23,** 1829–1832.
Steininger, G.A.M.W., Dettmer, J., Holland, C.W. & Dosso, S.E., 2013. Trans-dimensional joint inversion of seabed scattering and reflection data, *J. acoust. Soc. Am.,* **133,** 1347–1357.
Sugita, Y. & Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding, *Chem. Phys. Lett.,* **314**(12), 141–151.
Swendsen, R.H. & Wang, J.S., 1986. Replica Monte Carlo simulation of spin glasses, *Phys. Rev. Lett.,* **57,** 2607–2609.
Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* SIAM.
Tarantola, A. & Valette, B., 1982a. Inverse problems = quest for information, *J. Geophys.,* **50,** 159–170.
Tarantola, A. & Valette, B., 1982b. Generalized nonlinear inverse problems solved using the least squares criterion, *Rev. Geophys. Space Phys.,* **20,** 219–232.
Tierney, L. & Mira, A., 1999. Some adaptive Monte Carlo methods for Bayesian inference, *Stat. Med.,* **18**(1718), 2507–2515.

APPENDIX A: METROPOLIS–HASTINGS SAMPLERS

A1 Within-chain sampling

A Markov chain Monte Carlo algorithm is a method for drawing random samples of a multidimensional parameter vector, $\mathbf{x}$, from an arbitrary normalized target probability density distribution, $\pi(\mathbf{x})$. An up to date discussion of McMC samplers as well as much of the recent research in the area can be found in Brooks *et al.* (2011). Here, we provide a brief summary of the main elements. An McMC algorithm starts at an initial point, $\mathbf{x}_0$ and generates a new dependent vector $\mathbf{x}_1$ using a Metropolis–Hastings (M-H) sampler.

The M-H sampler consists of two steps: the first is to generate a proposed new random vector $\mathbf{x}'$ and the second is to decide whether to accept or reject it. If accepted the chain moves to $\mathbf{x}'$, if rejected it stays at its original position $\mathbf{x}$. In the first step, the random vector $\mathbf{x}'$ is drawn from the distribution $q(\mathbf{x}'|\mathbf{x})$. Here, it is assumed that some method is available to do this. An example is a multidimensional Gaussian distribution centred on $\mathbf{x}$ for which convenient algorithms exist (see Press *et al.* 1992):

$$q(\mathbf{x}'|\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left\{-1/2(\mathbf{x}' - \mathbf{x})^T \Sigma^{-1}(\mathbf{x}' - \mathbf{x})\right\}. \qquad \text{(A1)}$$

Having drawn the new vector $\mathbf{x}'$, this is accepted with probability $\alpha(\mathbf{x}'|\mathbf{x})$. To ensure convergence to the target PDF, $\pi(\mathbf{x})$, the M-H rule is used to determine the acceptance probability

$$\alpha(\mathbf{x}'|\mathbf{x}) = 1 \wedge \left\{ \frac{\pi(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{\pi(\mathbf{x})q(\mathbf{x}'|\mathbf{x})} |J| \right\}, \qquad \text{(A2)}$$

where $|J|$ is the determinant of the Jacobian between the space in which $\mathbf{x}$ and $\mathbf{x}'$ lies. Typically, these are the same space and so $|J| = 1$. The distribution $q(\mathbf{x}|\mathbf{x}')$ is the proposal probability for the reverse step from $\mathbf{x}'$ to $\mathbf{x}$, and since often this is symmetrical as in (A1) the proposal ratio cancels. Markov chains that involve transitions between dimension can be dealt with in the same way, only here the Jacobian may not always be unity (see Hopcroft *et al.* 2007; Bodin *et al.* 2012a,b, for examples).

After many steps of the M-H sampler, the Markov chain formed in this way has the history $\mathbf{x}_i$ $(i = 1, \ldots)$, and provided (A2) is satisfied, the distribution will converge to the target $\pi(\mathbf{x})$. In practice, one has to collect a subset of the samples in the chain by ignoring the initial vectors generated in the 'burn-in' phase, because these are dependent on the initial model, and also 'thin the chain' by keeping only every nth sample, thereby reducing correlation between vectors. The within-chain sampling used in all algorithms described in this paper follow this structure.

A2 Acceptance rule for an exchange swap

In parallel tempering exchange swaps occur between a pair of levels in a temperature ladder, and, in this case, the vector $\mathbf{x}$ describes the joint system of model vectors at all levels of the ladder, $\mathbf{x} = [\mathbf{m}_i]$, $(i = 1, \ldots, n)$, where $\mathbf{m}_i$ is the state of the model vector in chain i.

The target PDF, $\pi(\mathbf{x})$, then becomes the joint distribution over all chains

$$\pi(\mathbf{x}) = \prod_{i=1}^{n} \frac{\tilde{p}(\mathbf{m}_i|\mathbf{d})^{1/T_i}}{c(T_i)}, \tag{A3}$$

where $c(T_i)$ are normalizing constants

$$c(T_i) = \int \tilde{p}(\mathbf{m}|\mathbf{d})^{1/T_i} d\mathbf{m}. \tag{A4}$$

In an exchange swap, the model vectors of only two chains alter, all others are kept constant. For an exchange between models $\mathbf{m}_i$ at temperature T_i and $\mathbf{m}_j$ at T_j, the state of the system prior to swap is $\mathbf{x} = [\mathbf{m}_i, \mathbf{m}_j]$ and after is $\mathbf{x}' = [\mathbf{m}_j, \mathbf{m}_i]$. For simplicity of notation, we drop variables at all other temperature levels as they are unchanged by the swap. In an exchange swap, the proposal probability is symmetric ($q(\mathbf{x}'|\mathbf{x}) = q(\mathbf{x}|\mathbf{x}')$) and Jacobian equal to unity, and so combining this information and substituting (A3) into (A2), the M-H rule for calculating the acceptance probability for this type of transition becomes

$$\alpha(i, j) = 1 \wedge \left\{ \frac{\tilde{p}(\mathbf{m}_i|\mathbf{d})^{1/T_j} c(T_i)}{\tilde{p}(\mathbf{m}_i|\mathbf{d})^{1/T_i} c(T_j)} \times \frac{\tilde{p}(\mathbf{m}_j|\mathbf{d})^{1/T_i} c(T_j)}{\tilde{p}(\mathbf{m}_j|\mathbf{d})^{1/T_j} c(T_i)} \right\}, \tag{A5}$$

which is eq. (10) of the main text and simplifies to

$$\alpha(i, j) = 1 \wedge \left[\frac{\tilde{p}(\mathbf{m}_j|\mathbf{d})}{\tilde{p}(\mathbf{m}_i|\mathbf{d})} \right]^{1/T_i} \left[\frac{\tilde{p}(\mathbf{m}_i|\mathbf{d})}{\tilde{p}(\mathbf{m}_j|\mathbf{d})} \right]^{1/T_j}, \tag{A6}$$

which is eq. (11) of the main text.

A3 Acceptance rule for optimization

We aim to show that for an optimization problem that the general M-H rule (A2) is equivalent to (12). In this case, the model vector $\mathbf{x}$ is the same as for sampling above and so the corresponding normalized target PDF becomes

$$\pi(\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{-\phi(\mathbf{m}_i)/T_i}}{c(T_i)}, \tag{A7}$$

where now the normalizing constants are

$$c(T_i) = \int e^{-\phi(\mathbf{m}_i)/T_i} d\mathbf{m}_i. \tag{A8}$$

As before, exchange swaps occur between models $\mathbf{m}_i$ at temperature T_i and $\mathbf{m}_j$ at T_j, and the state of the joint system moves from $\mathbf{x} = [\mathbf{m}_i, \mathbf{m}_j]$ to $\mathbf{x}' = [\mathbf{m}_j, \mathbf{m}_i]$. Substituting (A7) into (A2) gives

$$\alpha(i, j) = 1 \wedge \left\{ \frac{e^{-\phi(\mathbf{m}_i)/T_j}}{e^{-\phi(\mathbf{m}_i)/T_i}} \times \frac{e^{-\phi(\mathbf{m}_j)/T_i}}{e^{-\phi(\mathbf{m}_j)/T_j}} \times \frac{c(T_i)c(T_j)}{c(T_j)c(T_i)} \right\}, \tag{A9}$$

which reduces to

$$\alpha(i, j) = 1 \wedge \exp\left\{(1/T_i - 1/T_j)(\phi(\mathbf{m}_i) - \phi(\mathbf{m}_j)\right\}, \tag{A10}$$

and this is eq. (12) of the main text.

APPENDIX B: MAXIMUM LIKELIHOOD ESTIMATION OF NOISE PARAMETERS

By substituting (17) into the general likelihood expression (16), we get an expression for the likelihood which depends on the unknown variance, σ^2, and the assumed known data correlation matrix, $\tilde{C}_d$,

$$p(\mathbf{d}|\mathbf{m}) = \frac{1}{\sqrt{(2\pi)^N \sigma^{2N} |\tilde{C}_d|}} \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} \right\}, \tag{B1}$$

where the residual vector $\mathbf{r} = \mathbf{d} - \mathbf{d}_p(\mathbf{m})$. In situations where the data noise σ is not known, it can be solved for together with the model $\mathbf{m}$, either in a Bayesian framework (as in Bodin *et al.* 2012a), or using a maximum likelihood approach as shown here. In particular, we follow Dosso *et al.* (2012) and find the value of σ which maximizes (B1). To simplify algebra, we first take logs of (B1) which gives

$$-\log p(\mathbf{d}|\mathbf{m}) = N \log \sigma + \frac{1}{2\sigma^2} \mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} + \frac{1}{2} \log[(2\pi)^N |\tilde{C}_d|]. \tag{B2}$$

An optimal value for σ is found by differentiating (B2) with respect to σ and setting to zero

$$\frac{\partial}{\partial \sigma}[-\log p(\mathbf{d}|\mathbf{m})] = \frac{N}{\sigma} - \frac{1}{\sigma^3} \mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} = 0, \tag{B3}$$

$$\Rightarrow \sigma^2 = \frac{1}{N} \mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r}, \tag{B4}$$

$$\Rightarrow \sigma = \left[\frac{1}{N} \mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} \right]^{1/2}, \quad (\sigma > 0), \tag{B5}$$

which is eq. (19) of the main text. To find the modified likelihood expression, we substitute this expression for σ into (B2) and get

$$-\log p(\mathbf{d}|\mathbf{m}) = N \log \left[\frac{1}{N} \mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} \right]^{1/2} + \frac{N}{2} + \frac{1}{2} \log[(2\pi)^N |\tilde{C}_d|], \tag{B6}$$

$$\Rightarrow -\log p(\mathbf{d}|\mathbf{m}) = \frac{N}{2} \log\left(\mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} \right) + \left\{ \frac{N}{2}(1 - \log N) + \frac{1}{2} \log[(2\pi)^N |\tilde{C}_d|] \right\}. \tag{B7}$$

The term in curly brackets does not depend on the residual vector $\mathbf{r}$ and so we write

$$-\log p(\mathbf{d}|\mathbf{m}) = \frac{N}{2} \log\left(\mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} \right) + \text{Const}, \tag{B8}$$

which gives

$$p(\mathbf{d}|\mathbf{m}) \propto \exp\left\{ -\frac{N}{2} \log\left(\mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} \right) \right\}, \tag{B9}$$

which is eq. (20) of the main text. Note that (B9) can also be written as

$$p(\mathbf{d}|\mathbf{m}) \propto \left(\mathbf{r}^T \tilde{C}_d^{-1} \mathbf{r} \right)^{-N/2}. \tag{B10}$$

APPENDIX C: PARALLEL TEMPERING PSEUDO-CODE

The pseudo-code below shows the basic structure of a Parallel Tempering algorithm. Upon initialization, the n-vector $\mathbf{T}$ contains the preset temperatures of the n chains, m is the number of within-chain steps and n_b is the number of within-chain burn-in steps executed before results are collected. The user-supplied routine '*AdvanceChain*' performs within-chain McMC sampling by updating the model in the ith temperature level over the j time step and returns the updated value of the target PDF, $\pi_{i,j}$. This routine also stores any results along the chain in a form suitable for the parameter space. The function $U(a, b)$ represents a random draw from a uniform PDF between a and b. The pseudo-code shows that the PT routine is independent of both the details of the McMC sampler and also the dimension and nature of the parameter space.

Algorithm 1 Exchange swapping McMC chains in a tempered space

procedure PT($\mathbf{T}, n, m, n_b$)
  for $j \leftarrow 1, m$ **do** ▷ Loop over time step of Markov chains
    for $i \leftarrow 1, n$ **do** ▷ Loop over temperature ladder
      $\pi_{i,j} \leftarrow$ **AdvanceChain**(i, j, T_i, n_b) ▷ Advance ith chain over the jth step
    end for
    for $i \leftarrow 1, n$ **do** ▷ Swap random pairs of chains using eq. (11)
      $p \leftarrow U(1, n)$ ▷ Select uniform random chain
      $q \leftarrow U(1, n)$, $q \neq p$ ▷ Select chain partner for swap
      $r_1 \leftarrow [\pi_{q,j}/\pi_{p,j}]^{1/T_p}$
      $r_2 \leftarrow [\pi_{p,j}/\pi_{q,j}]^{1/T_q}$
      $\alpha \leftarrow min[1, r_1 r_2]$
      if $\alpha < U(0, 1)$ **then**
        $T_p \leftarrow T_q$ ▷ Swap chain temperatures
        $T_q \leftarrow T_p$
      end if
    end for
  end for
end procedure