Geophysical Journal International

Geophys. J. Int. (2013) **194,** 1701–1726
Advance Access publication 2013 June 19

doi: 10.1093/gji/ggt180

Bayesian inversion for finite fault earthquake source models I—theory and algorithm

S. E. Minson, M. Simons and J. L. Beck

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

Accepted 2013 May 2. Received 2013 May 1; in original form 2012 October 15

SUMMARY

The estimation of finite fault earthquake source models is an inherently underdetermined problem: there is no unique solution to the inverse problem of determining the rupture history at depth as a function of time and space when our data are limited to observations at the Earth's surface. Bayesian methods allow us to determine the set of all plausible source model parameters that are consistent with the observations, our *a priori* assumptions about the physics of the earthquake source and wave propagation, and models for the observation errors and the errors due to the limitations in our forward model. Because our inversion approach does not require inverting any matrices other than covariance matrices, we can restrict our ensemble of solutions to only those models that are physically defensible while avoiding the need to restrict our class of models based on considerations of numerical invertibility. We only use prior information that is consistent with the physics of the problem rather than some artefice (such as smoothing) needed to produce a unique optimal model estimate. Bayesian inference can also be used to estimate model-dependent and internally consistent effective errors due to shortcomings in the forward model or data interpretation, such as poor Green's functions or extraneous signals recorded by our instruments. Until recently, Bayesian techniques have been of limited utility for earthquake source inversions because they are computationally intractable for problems with as many free parameters as typically used in kinematic finite fault models. Our algorithm, called cascading adaptive transitional metropolis in parallel (CATMIP), allows sampling of high-dimensional problems in a parallel computing framework. CATMIP combines the Metropolis algorithm with elements of simulated annealing and genetic algorithms to dynamically optimize the algorithm's efficiency as it runs. The algorithm is a generic Bayesian Markov Chain Monte Carlo sampler; it works independently of the model design, *a priori* constraints and data under consideration, and so can be used for a wide variety of scientific problems. We compare CATMIP's efficiency relative to several existing sampling algorithms and then present synthetic performance tests of finite fault earthquake rupture models computed using CATMIP.

Key words: Inverse theory; Probability distributions; Computational seismology.

1 INTRODUCTION

To study the physics of earthquakes, we need observations of earthquake ruptures, but the earthquake rupture process can only be inferred from measurements taken at the surface of the Earth. Using limited surface observations to constrain a possibly complex and heterogeneous source process is a fundamentally ill-posed inverse problem. Conventionally, regularization is used to transform such inverse problems into a well-conditioned optimization problem for a single source model. Typical regularization schemes include Laplacian smoothing, minimizing the length of the solution (which is equivalent to moment minimization for finite fault earthquake models), positivity constraints, and sparsity constraints (e.g. Du *et al.* 1992; Arnadottir & Segall 1994; Ji *et al.* 2002; Evans & Meade 2012). Some of these constraints, such as positivity, can be defended based on the physical processes being modelled. However, other forms of regularization are often employed to make the inversion numerically stable or to prevent overfitting the data due to limitations of the forward model. The choice of which form and strength of regularization to use is often arbitrary. Yet even a slight change in inversion design can lead to different solutions, thereby limiting our ability to distill the physics of the rupture process from a given source model. When different inversion methodologies yield very different rupture models for the same earthquake (e.g. Fig. 1), it is not obvious what conclusions, if any, can be drawn about the source process.

© The Authors 2013. Published by Oxford University Press on behalf of The Royal Astronomical Society.

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

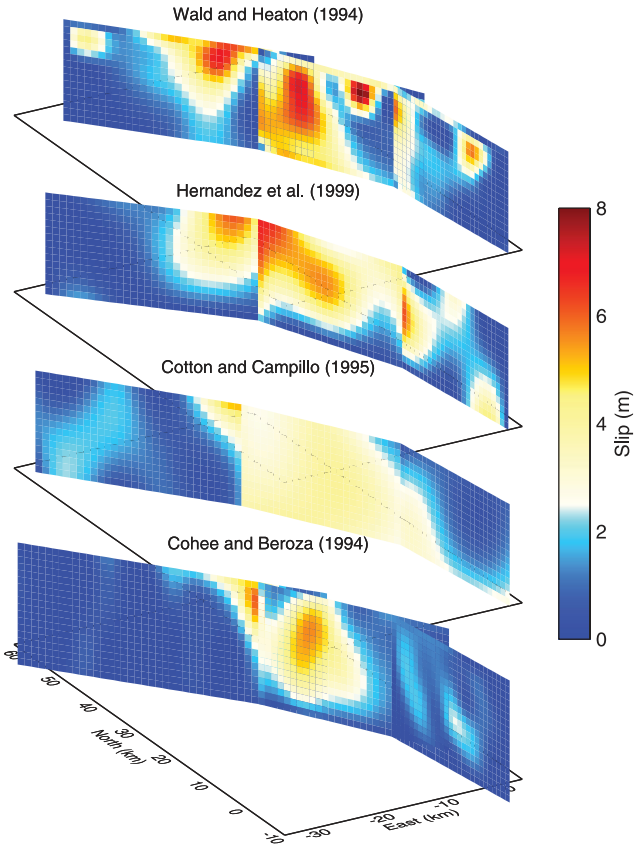

Figure 1. An illustration of the variability in earthquake source models. Small differences in inversion techniques and data can lead to large differences in inferred earthquake slip models. We show here four published slip models for the 1992 M_w 7.3 Landers, California earthquake available through the ETH Zurich (Swiss Federal Institute of Technology, Zurich) Finite-Source Rupture Model Database (Cohee & Beroza 1994b; Wald & Heaton 1994; Cotton & Campillo 1995; Hernandez *et al.* 1999).

In a Bayesian inversion framework, the regularization of the inverse problem is accomplished by the choice of prior distribution and different choices can be assessed by their posterior probability based on the data (e.g. Tarantola 2005; Beck 2010). Further, the choice of the prior can be based purely on knowledge about the physics of the problem that is not encapsulated in the forward model rather than any requirement to produce a unique source model (as required in inversions using regularized least squares). Finally, since Bayesian methods return the ensemble of all models that are consistent with the data and chosen prior information, the question of how to choose one single solution to a problem that does not have a unique solution becomes moot.

Regularized optimization returns only the part of the null space required to satisfy the imposed smoothness requirement. Bayesian sampling (which uses probabilistic methods to determine the family of all possible models that are consistent with the data and our *a priori* constraints), will theoretically produce models from everywhere in the parameter space, with density proportional to the posterior probability content: that is, this sampling naturally produces more models in regions which fit the data better (and so are deemed more plausible) and fewer in regions with lower probability. We can then analyse these models however we want. For example, we can plot histograms and 2-D or 3-D projections of the posterior samples to image the topology of the complete solution space including the locations and sizes of its minima.

The term 'Bayesian' was not coined until Fisher (1921) (Fienberg 2006), but Bayesian techniques have been used in many scientific fields for centuries under such names as inverse probability and subjective probability. Thomas Bayes' only paper on the topic was published posthumously (Bayes 1763). Pierre-Simon Laplace derived many important fundamental probabilistic inference results starting in 1774 and culminating in his treatise on probability (Laplace 1812). He was perhaps the first to use these techniques in a scientific context when he employed Bayesian inference to derive a posterior probability distribution on the mass of a moon of Saturn. Bayesian inference has been used to study geophysical problems at least since the work of Sir Harold Jeffreys (e.g. Jeffreys 1931, 1939), and there has been a recent resurgence in interest by geophysicists (e.g. Mosegaard & Tarantola 1995; Malinverno 2002; Sambridge & Mosegaard 2002; Tarantola 2005).

An ideal goal for inversion of earthquake rupture models is to use the physics of the rupture process as our only constraint so that we can determine what is and what is not constrained by the data and the assumed physics. However, a full Bayesian solution to an inverse problem using only prior constraints based on the physics of the process being modelled can be very computationally expensive, especially for high-dimensional problems like the seismic rupture models we are studying. For seismic source inversions, there has been an effort to develop a computationally tractable proxy for the Bayesian posterior probability density function (PDF; e.g. Monelli & Mai 2008), as well as studies using Bayesian analysis to calculate the solution to the traditional inverse problem with non-physical regularization (e.g. Fukuda & Johnson 2008). But using non-physical prior constraints makes it difficult to interpret the inversion results.

In contrast, we have developed a full Bayesian approach that uses a new Markov Chain Monte Carlo (MCMC) sampling technique to produce finite fault earthquake models that is based on the well-known Metropolis algorithm (Metropolis *et al.* 1953). Because of the increase in sampling efficiency and massively parallel computing, we are able to solve modelling problems that would up to now have been computationally intractable. We note that the sampling technique is in principle completely independent from the data and model under consideration, and thus has the potential to be applied to a wide variety of parameter estimation problems.

We begin by providing a brief background on the theory of Bayesian inversion. This background is followed by a description of our new MCMC sampling algorithm, cascading adaptive transitional metropolis in parallel (CATMIP), including performance tests relative to existing algorithms. We then derive a physics-based, minimally constrained finite fault earthquake rupture parametrization suitable for Bayesian techniques and present a series of performance tests of this model using CATMIP sampling. (Application to real observations is reserved for a following paper: Minson *et al.* Bayesian inversion for finite fault earthquake source models II–the 2011 great Tohoku-oki, Japan earthquake, in preparation. We will refer to this as Paper II.) Finally, we present various potentially useful methods to explore ensembles of earthquake source models.

2 BAYESIAN APPROACH TO INVERSION

Broadly, Bayesian methods for inverse modelling use probability models to quantify our state of knowledge by explicitly treating the uncertainties related to the observation process and the uncertainties due to missing information or errors in the model design, and from this we can ascribe an *a posteriori* plausibility to each model in a

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

set of proposed models (Jaynes 2003; Beck 2010). This posterior probability distribution describing the plausibility of each member of the ensemble of models is the 'solution' to our inverse problem. We can derive the Bayesian solution to a generic inverse problem [the posterior PDF, $p(\boldsymbol{\theta}|\mathbf{D})$] using Bayes' Theorem,

$$p(\boldsymbol{\theta}|\mathbf{D}) \propto p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ is a k-dimensional vector of model parameters whose set of values specify a set of possible models and $p(\boldsymbol{\theta})$ is the prior PDF that defines the relative plausibility of each possible value of $\boldsymbol{\theta}$ *a priori* (i.e. without reference to the data). For example, if we were fitting a straight line to some data, $\boldsymbol{\theta}$ would be a two-element vector containing the slope and intercept that specify a possible line, and $p(\boldsymbol{\theta})$ would be a measure of the relative plausibility assigned to a specific line that is given by the slope and intercept in $\boldsymbol{\theta}$. The data likelihood, $p(\mathbf{D}|\boldsymbol{\theta})$, is a PDF describing the probability of having observed our data, $\mathbf{D}$, given a value for $\boldsymbol{\theta}$.

The Bayesian approach is completely generic. The data, the model and the form of the probability distributions are not restricted. It is this generality that allows for greater specificity. There are no simplifying assumptions required in formulating the model. The model can be linear or non-linear. Prior information that exists about the physics of the problem but is not specific enough to build into the forward model can be incorporated in a prior probability distribution.

Confusingly, there are also optimization techniques that are sometimes described as Bayesian because the choice of regularization in the optimization scheme is based on some rule derived from Bayes' Theorem. But, at best, these are only partial Bayesian analyses. [Traditional regularized optimization can be viewed as a partial Bayesian analysis; for details see Appendix A and Menke (2012).] However, when we discuss Bayesian analysis, we refer to methods that characterize the complete posterior distribution, $p(\boldsymbol{\theta}|\mathbf{D})$. We now consider each component of Bayes' Theorem (eq. 1) in turn.

2.1 Observed data: D

The observed data represent a superset of possible data sets. For earthquake source modelling, these data sets could be seismic, GPS, InSAR, tsunami data, etc.,

$$\begin{aligned}\mathbf{D} &= \{\mathbf{D}_{\text{seismic}}, \mathbf{D}_{\text{geodetic}}, \mathbf{D}_{\text{tsunami}}, \ldots\} \\ &= \left\{\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{N_{\text{ds}}}\right\},\end{aligned} \tag{2}$$

where each $\mathbf{D_i}$ is a vector of data points (or observations) comprising each of N_{ds} data sets.

These data are sets of numbers obtained from measurement and so are known at the time of the inversion analysis. On the other hand, our model-based predictions of these measurements contain uncertainty from two sources: the uncertainty about the errors in the prediction of the observed quantities based on a geophysical model, plus the uncertainty about the errors in the measurements based on a model of the observation process. (The latter is often referred to as 'data uncertainty' but our perspective is that, in an inversion analysis, the data is certain and it is our corresponding predictions that are uncertain.) In the next subsection, these two sources of uncertainty are quantified by a stochastic forward model for the *predictions*, $\mathbf{d_i}$ (a random variable), corresponding to the *actual measurements*, $\mathbf{D_i}$, for the ith data set.

2.2 Stochastic forward model, p(d|θ), and likelihood function, p(D|θ)

Consider a generic data set, $\mathbf{D}$, and corresponding prediction of these measurements, $\mathbf{d}$, where vectors $\mathbf{D}$ and $\mathbf{d}$ both have N_{dp} elements. Given a deterministic forward model design, $\mathbf{G}(\boldsymbol{\theta})$, the stochastic forward model, $p(\mathbf{d}|\boldsymbol{\theta})$, that we use to express the uncertainty in the predicted measurements is based on,

$$\mathbf{d} = \mathbf{G}(\boldsymbol{\theta}) + \mathbf{e} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{e}$ represents the uncertain measurement errors (the difference between the predicted measurements and the true values of the observed physical quantities) and $\boldsymbol{\epsilon}$ represents the uncertain model prediction errors (the difference between the true observed quantities and the predictions of the deterministic forward model).

A common choice of the probability models for the measurement errors, $\mathbf{e}$, and the model prediction errors, $\boldsymbol{\epsilon}$, is to use independent Gaussian PDFs. [This choice can be justified by using the principle of maximum entropy to select the probability models for $\mathbf{e}$ and $\boldsymbol{\epsilon}$ in eq. (3). See, for examples, Jaynes (2003) and Beck (2010).] In this case, the sum $(\mathbf{e} + \boldsymbol{\epsilon})$ in eq. (3) is Gaussian, so the stochastic forward model is given by,

$$\begin{aligned}p(\mathbf{d}|\boldsymbol{\theta}) &= \mathcal{N}\left(\mathbf{d}|\mathbf{G}(\mathbf{m}) + \boldsymbol{\mu}, \mathbf{C}_\chi\right) \\ &= \frac{1}{(2\pi)^{N_{\text{dp}}/2}|\mathbf{C}_\chi|^{\frac{1}{2}}} e^{-\frac{1}{2}[\mathbf{d}-\mathbf{G}(\mathbf{m})-\boldsymbol{\mu}]^T \cdot \mathbf{C}_\chi{}^{-1} \cdot [\mathbf{d}-\mathbf{G}(\mathbf{m})-\boldsymbol{\mu}]},\end{aligned} \tag{4}$$

where $\mathbf{C}_\chi$ and $\boldsymbol{\mu}$ are the covariance matrix and mean of the sum $(\mathbf{e} + \boldsymbol{\epsilon})$, respectively. Thus, $\boldsymbol{\mu}$ represents a possible bias in our predictions. Further, because $\mathbf{e}$ and $\boldsymbol{\epsilon}$ are modelled as probabilistically independent,

$$\mathbf{C}_\chi = \mathbf{C}_\mathbf{d} + \mathbf{C}_\mathbf{p}, \tag{5}$$

where $\mathbf{C}_\mathbf{d}$ and $\mathbf{C}_\mathbf{p}$ are the covariance matrices for $\mathbf{e}$ and $\boldsymbol{\epsilon}$, respectively. Note that for additional generality we have written the deterministic forward model in eq. (4) as $\mathbf{G}(\mathbf{m})$ instead of $\mathbf{G}(\boldsymbol{\theta})$. In many applications, $\boldsymbol{\theta} = \mathbf{m}$. However, we may also want to use the data, $\mathbf{D}$, to learn about the parameters in the probability models for $\mathbf{e}$ and $\boldsymbol{\epsilon}$. Then we would have $\boldsymbol{\theta} = (\mathbf{m}, \boldsymbol{\mu}, \mathbf{C}_\chi)$.

The likelihood function, $p(\mathbf{D}|\boldsymbol{\theta})$, gives the probability of the observed data according to the model given by $\boldsymbol{\theta}$, and represents the forward model's goodness of fit to the data, $\mathbf{D}$. If a model gives low probability to the observations, then it is highly unlikely that this model accurately describes the source of those observations. The likelihood function $p(\mathbf{D}|\boldsymbol{\theta})$ is not a probability model of the actual measured data, $\mathbf{D}$, which is known. Instead, the likelihood function gives the probability of observing the actual data, $\mathbf{D}$, given by setting $\mathbf{d} = \mathbf{D}$ in the stochastic forward model $p(\mathbf{d}|\boldsymbol{\theta})$. For example, if $\boldsymbol{\theta} = (\mathbf{m}, \boldsymbol{\mu}, \mathbf{C}_\chi)$, then the posterior is $p(\mathbf{m}, \boldsymbol{\mu}, \mathbf{C}_\chi|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{m}, \boldsymbol{\mu}, \mathbf{C}_\chi)p(\mathbf{m}, \boldsymbol{\mu}, \mathbf{C}_\chi)$ where the likelihood function, $p(\mathbf{D}|\mathbf{m}, \boldsymbol{\mu}, \mathbf{C}_\chi)$, is given by eq. (4) with $\mathbf{d} = \mathbf{D}$, and $p(\mathbf{m}, \boldsymbol{\mu}, \mathbf{C}_\chi)$ represents the chosen prior PDF on the forward model parameters and on the mean and covariance of the errors. Results for inversions will be presented in Section 5 for this case as well as when $\boldsymbol{\mu}$ and $\mathbf{C}_\chi$ are fixed *a priori* (so that $\boldsymbol{\theta} = \mathbf{m}$).

In many inverse problems, the difference $\mathbf{D} - \mathbf{G}(\mathbf{m})$ is viewed as simply measurement errors. But, in many cases, the measurement errors may be dwarfed by the model prediction errors, $\boldsymbol{\epsilon}$, produced by differences between the model and the real world. For finite fault earthquake source processes, possible error sources of this type include having the wrong source geometry, a poorly located hypocentre, an incorrect elastic structure and simply parametrizing the earthquake source evolution in a way that is not amenable to

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

representing the 'true' source process. Further, some measurement errors may be better considered as model prediction errors. For example, an atmospheric region of high water content may lead to a spurious deformation signal in a radar interferogram or a large truck driving by a seismometer may create detectable ground motion. These are not errors in the accuracy of the measurements. The change in phase between two radar scenes is no more or less difficult to estimate on a rainy day than on a dry one, nor is the sensitivity of a seismometer affected by its proximity to a highway.

In theory, we could create an earthquake source model which included not only spatially and temporally varying slip but also parameters to describe the variation in the propagation velocity of radar signals and ground motions due to near-seismometer activity. Similarly, parameters for the Earth structure and hypocentre could be simultaneously modelled as part of the process of sampling possible earthquake ruptures. But it is not usually tractable to directly and simultaneously model the earthquake, Earth structure and all possible error sources. However, it is possible to combine all model prediction errors into one probabilistic representation by casting the unknown prediction errors produced by any model as the total uncertainties due to all sources of errors between the observations and the predictions of the deterministic forward model except for the measurement errors produced by the sensors (which can be independently modelled; e.g. Beck & Katafygiotis 1998; Tarantola 2005; Beck 2010). We have adopted this strategy in eqs (4) and (5) to construct the stochastic forward model where the covariance matrix $\mathbf{C_d}$ is pre-determined (and usually taken as diagonal) by a separate study of the measurement process, whereas $\mathbf{C_p}$, or some defining parameters for it, is included in the parameter vector $\boldsymbol{\theta}$ and updated in the posterior distribution. We defer discussion of the form of model prediction covariance matrix, $\mathbf{C_p}$, that we use in our finite fault model to Section 4.

As with traditional optimization approaches, if we underestimate our errors, we will overfit the data and produce posterior distributions that are too tightly peaked. (In traditional optimization, this problem is overtly or sometimes unknowingly dealt with through regularization.) Similarly, if we overestimate our errors, we will underfit the data and our posterior distributions will be too broad.

The stochastic forward model corresponding to the N_{ds} data sets in eq. (2) is a joint distribution,

$$\mathrm{p}(\mathbf{d}|\boldsymbol{\theta}) = \mathrm{p}(\mathbf{d_1}, \ldots, \mathbf{d}_{\mathrm{N_{ds}}}|\boldsymbol{\theta}). \tag{6}$$

We treat the predictions, $\mathbf{d_1}, \ldots, \mathbf{d}_{\mathrm{N_{ds}}}$, as probabilistically independent, so,

$$\begin{aligned}\mathrm{p}(\mathbf{d}|\boldsymbol{\theta}) &= \mathrm{p}(\mathbf{d_1}|\boldsymbol{\theta})\mathrm{p}(\mathbf{d_2}|\boldsymbol{\theta})\cdots\mathrm{p}(\mathbf{d}_{\mathrm{N_{ds}}}|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N_{\mathrm{ds}}} \mathrm{p}(\mathbf{d_i}|\boldsymbol{\theta}),\end{aligned} \tag{7}$$

where each $\mathrm{p}(\mathbf{d_i}|\boldsymbol{\theta})$ is chosen as in eqs (4) and (5).

To apply Bayes Theorem (eq. 1), we substitute the observed data, $\mathbf{D}$, for $\mathbf{d}$ in eq. (7). This is done by setting $\mathbf{d_i} = \mathbf{D_i}$ in each stochastic forward model. Then the posterior PDF is,

$$\begin{aligned}\mathrm{p}(\boldsymbol{\theta}|\mathbf{D}) &\propto \mathrm{p}(\mathbf{D_1}|\boldsymbol{\theta})\mathrm{p}(\mathbf{D_2}|\boldsymbol{\theta})\cdots\mathrm{p}(\mathbf{D}_{\mathrm{N_{ds}}}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}) \\ &= \mathrm{p}(\boldsymbol{\theta})\cdot\prod_{i=1}^{N_{\mathrm{ds}}} \mathrm{p}(\mathbf{D_i}|\boldsymbol{\theta}),\end{aligned} \tag{8}$$

where $\boldsymbol{\theta}$ includes all parameters needed to define the suite of stochastic forward models for all $\mathbf{d_i}$, comprised of the parameters for all deterministic forward models as well as the parameters defining the probability models for the model prediction errors for each $\mathbf{d_i}$. Thus, we have the freedom to fuse as many data sets together as we want.

3 CASCADING ADAPTIVE TRANSITIONAL METROPOLIS IN PARALLEL: CATMIP

3.1 CATMIP introduction

CATMIP is a parallel MCMC algorithm that efficiently samples high-dimensional spaces. It is based on the Transitional Markov Chain Monte Carlo (TMCMC) algorithm of Ching & Chen (2007) which combines transitioning (akin to simulated annealing or tempering) and resampling with the MCMC simulation of the Metropolis algorithm (Metropolis *et al.* 1953). The Metropolis algorithm uses a random walk to explore the model space and probabilistically chooses whether or not to take a proposed step based on the probability associated with the candidate model. Intrinsically, the basic Metropolis algorithm is not parallelizable because it uses a single Markov chain. Further, the efficiency of the Metropolis algorithm depends strongly on the probability distribution used to produce the random walk steps, and it has difficultly sampling multiply peaked PDFs efficiently.

CATMIP and TMCMC belong to a class of samplers which use transitioning, an approach which shares several characteristics with simulated annealing optimization (Kirkpatrick *et al.* 1983; Cerny 1985; Rothman 1985). In the Bayesian literature, it is more common to find the term 'tempering' than 'annealing', although a distinction is often made in which annealing is used to describe algorithms which go from an initial 'hot' solution to a final 'cold' solution while tempering algorithms allow both cooling and heating of the solution. The use of tempering in Bayesian sampling dates back at least to Marinari & Parisi (1992). CATMIP and TMCMC use the annealing strategy. The initial 'hot' state of the solution is the prior PDF which is broader than the final 'cold' posterior PDF. Because we start with a broad distribution and then slowly 'cool' it to the compact posterior distribution, it is easier for the sampler to find all of the peaks of the posterior distribution. More importantly, the particular variant of annealing we use, called transitioning (see eq. 9), ensures that our population of samples are almost always at equilibrium with the PDF we are trying to simulate, which makes the sampling very efficient.

Both CATMIP and TMCMC employ resampling (Fig. 2), a process in which less probable samples from the previous transitioning step are replaced with more probable models. Resampling allows samples trapped in regions of low probability to be transplanted to become seeds for new Markov chains in regions of higher probability. As we will see later, this combination of transitioning and resampling allows the CATMIP and TMCMC algorithms to outperform the Metropolis algorithm at sampling a PDF with multiple peaks. Although the Markov chains do not mix with each other, information from all Markov chains is combined to calculate a model covariance used to define the proposal PDF from which the candidate samples at each transitioning step are drawn. This adaptive updating of the proposal PDF tunes the algorithm for maximum efficiency.

The main difference between CATMIP and TMCMC is in how the Metropolis algorithm is employed. In TMCMC, more probable models are assembled into longer Markov chains. In CATMIP, more probable models spawn more Markov chains, leading to a concentration of multiple chains in regions of high probability. In this

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

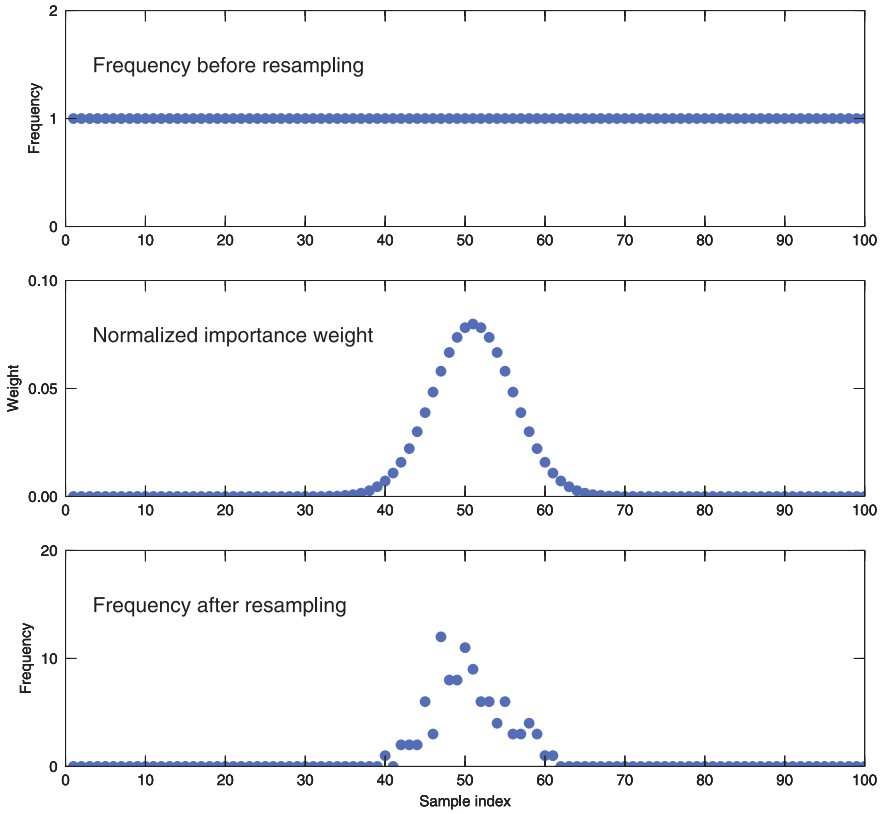

Figure 2. Schematic illustration of resampling process. This example begins with a set of 100 unique samples, and thus each sample has a frequency of one (top). Normalized importance weights are then calculated for each sample from eq. (11) (middle). Finally, we make 100 random draws from the original set of samples where the chance of selecting each sample is given by the normalized importance weights. The frequency with which each sample was drawn in this trial is plotted (bottom). Note that there are still 100 samples although they are no longer unique as some samples have been duplicated while others have been eliminated. The resampling process thus allows a set of samples to more closely approximate a target PDF (the normalized importance weights give the ratio of the target probability for each sample to that sample's probability in the PDF from which it was drawn) without the computational effort of having to generate new samples. However, the gain in computational efficiency is at the loss of the number of unique samples of the target PDF.

respect, CATMIP is more similar to the Neighbourhood algorithm (Sambridge 1999) which explores the parameter space by concentrating random walk sampling in regions which produce better (more plausible) models.

3.2 CATMIP algorithm

Following Beck & Au (2002) and Ching & Chen (2007), we sample from a series of 'transitional' intermediate PDFs that are controlled by a tempering (or annealing) parameter, β,

$$f(\boldsymbol{\theta}|\mathbf{D}, \beta_m) \propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{D}|\boldsymbol{\theta})^{\beta_m}$$

$$m = 0, 1, \ldots, M$$

$$0 = \beta_0 < \beta_1 < \beta_2 < \cdots < \beta_M = 1. \qquad (9)$$

Since $f(\boldsymbol{\theta}|\mathbf{D}, \beta_0 = 0) = \mathrm{p}(\boldsymbol{\theta})$, it can be simulated directly by drawing samples from the prior distribution. We then sample from a succession of PDFs, each of which is approximately equal to the PDF we have just sampled, and which therefore is much easier to simulate than it would be to directly sample from the final posterior PDF without the information from the preceding cooling steps. Finally, we sample $f(\boldsymbol{\theta}|\mathbf{D}, \beta_M = 1) \propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{D}|\boldsymbol{\theta}) \propto \mathrm{p}(\boldsymbol{\theta}|\mathbf{D})$, and thus we have sampled the posterior PDF.

For each transitional stage, CATMIP requires three steps. First, a new value for β is chosen. Secondly, the samples from the previous transitional stage are resampled in proportion to their relative likelihoods as given by the next intermediate PDF. Third, each output of the resampling process is used as the seed for a separate instance of the Metropolis algorithm.

The total number of cooling steps, M, is not a parameter of the CATMIP algorithm. Instead, M is simply the number of cooling steps needed to reach $\beta = 1$ where each successive value of β is calculated adaptively so that the difference between $f(\boldsymbol{\theta}|\mathbf{D}, \beta_m)$ and $f(\boldsymbol{\theta}|\mathbf{D}, \beta_{m+1})$ is small, ensuring that the next transitional PDF to be simulated is fairly well approximated by the current set of samples. This approach makes sampling $f(\boldsymbol{\theta}|\mathbf{D}, \beta_{m+1})$ very efficient.

By choosing β dynamically, the algorithm transitions optimally. If the data are relatively uninformative, then β will converge to 1 quickly because increasing β has little effect on the intermediate PDFs. The more informative the data, the more dissimilar the posterior will be from the prior, and the more transitional PDFs are required for efficiency. The most efficient cooling rate is obtained by choosing β_{m+1} so that the equivalent number of independent samples after the resampling step, called the *effective sample size*, is approximately $\frac{N}{2}$ where N is the total number of samples (for definition, see Beck & Zuev 2013). Beck & Zuev (2013) show that this optimal cooling rate can be obtained by choosing β_{m+1} such that $c_v[w(\boldsymbol{\theta}_{m,i})] = 1$, where $\boldsymbol{\theta}_{m,i}$ is the ith model at cooling step m, and c_v denotes the coefficient of variation defined as the ratio of the standard deviation to the mean of $\{w(\boldsymbol{\theta}_{m,i}) : i = 1, \ldots, N\}$. $\{w(\boldsymbol{\theta}_{m,i}) : i = 1, \ldots, N\}$ is a set of N importance (or plausibility) weights each of which is the ratio of the ith sample's probability at

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

β_{m+1} to its probability at β_m,

$$w(\boldsymbol{\theta}_{m,i}) = \frac{\mathrm{p}(\boldsymbol{\theta}_{m,i})\mathrm{p}(\mathbf{D}|\boldsymbol{\theta}_{m,i})^{\beta_{m+1}}}{\mathrm{p}(\boldsymbol{\theta}_{m,i})\mathrm{p}(\mathbf{D}|\boldsymbol{\theta}_{m,i})^{\beta_m}} = \mathrm{p}(\mathbf{D}|\boldsymbol{\theta}_{m,i})^{\beta_{m+1}-\beta_m}. \quad (10)$$

After an updated value for β is calculated, the current population of models are resampled according to their probabilities at β_{m+1} so that the likelihood of choosing any model is proportional to its updated probability. To correct for the change in probability distribution between β_m and β_{m+1}, each sample $\boldsymbol{\theta}_{m,i}$ must be assigned the weight $w(\boldsymbol{\theta}_{m,i})$ in eq. (10). During resampling, the chance of drawing each model $\boldsymbol{\theta}_{m,i}$ is proportional to $w(\boldsymbol{\theta}_{m,i})$. Resampling has the benefits of both redistributing the density of samples so that they better approximate $f(\boldsymbol{\theta}|\mathbf{D}, \beta_{m+1})$ as well as creating a genetic algorithm-like behaviour in which more unlikely models are removed from the population in favour of increasing the number of more likely models.

Each resampled model is used as the initial seed for an instance of the Metropolis algorithm (see Appendix B). The Metropolis algorithm uses a random walk through model space to produce models whose density are proportional to the target PDF. The candidate samples in the random walk are drawn from a proposal PDF from which random numbers can be generated (typically a Gaussian PDF centred on the current sample in the Markov chain). The efficiency of the Metropolis algorithm at generating samples of the target PDF is controlled by how similar the proposal PDF is to the target transitional PDF. Thus, it is important that the proposal PDF mimic the target PDF as closely as possible.

We use a multivariate Gaussian PDF as our proposal PDF. To make our Metropolis sampling as efficient as possible, we set the covariance equal to the sample model covariance, $\mathbf{C_m}$, calculated from the current samples and scaled according to the acceptance rate. By using the sample model covariance, $\mathbf{C_m}$, in our proposal PDF, the random walk will automatically take larger steps in directions in which the target PDF is broad and thus has large variance (or covariance if there are trade-offs between model parameters), and take smaller steps in directions where the target PDF is highly peaked. We also dynamically rescale $\mathbf{C_m}$ so that the random walk sampler takes larger steps and explores model space more efficiently when the acceptance rate of candidate samples is high, but shrinks the step size when too few acceptable models are found.

To construct the sample model covariance, $\mathbf{C_m}$, we first note that we have a set of samples from the wrong probability distribution: our samples are from $f(\boldsymbol{\theta}|\mathbf{D}, \beta_m)$ while we wish to calculate the covariance of $f(\boldsymbol{\theta}|\mathbf{D}, \beta_{m+1})$. We need to correct for this by weighting each sample by the ratio of probabilities in eq. (10) and renormalizing (see e.g. Gelman *et al.* 2004). The weights of the N samples at the $(m+1)^{th}$ intermediate level are given by,

$$p_{m,i} = \frac{w(\boldsymbol{\theta}_{m,i})}{\sum_{j=1}^{N} w(\boldsymbol{\theta}_{m,j})}, \quad (11)$$

where the importance weights, $w(\boldsymbol{\theta}_{m,i})$, are defined in eq. (10). The sample mean for the $(m+1)$th level is then,

$$\bar{\boldsymbol{\theta}} = \sum_{i=1}^{N} p_{m,i}\boldsymbol{\theta}_{m,i}, \quad (12)$$

and the sample covariance matrix is,

$$\mathbf{C_m} = \sum_{i=1}^{N} (\boldsymbol{\theta}_{m,i} - \bar{\boldsymbol{\theta}}_m)(\boldsymbol{\theta}_{m,i} - \bar{\boldsymbol{\theta}}_m)^T p_{m,i}. \quad (13)$$

The proposal density for the Metropolis sampler (see Appendix B) in CATMIP is $q(y|x) = \mathcal{N}(x, \boldsymbol{\Sigma}_\mathbf{m})$ with $\boldsymbol{\Sigma}_\mathbf{m} = c_m^2\mathbf{C_m}$ and $c_m = a + bR$ where R is the observed acceptance rate of the Metropolis sampling and a and b are selected constants (Muto, personal communication, 2008). In this way, we rescale our proposal density by the acceptance rate of our sampler. When the acceptance rate is higher, we increase the size of our random walk steps, allowing greater exploration of the model space. When our acceptance rate decreases, we take smaller steps to increase the chance that a candidate model will be accepted. a and b are not expected to have a major effect on the efficiency of sampling, but merely act to tweak the rescaling of the proposal PDF, and so we consider an investigation of possible optimal values for a and b to be beyond the scope of this paper. Instead, for the performance tests and earthquake modelling presented here, we will arbitrarily adopt $a = \frac{1}{9}$ and $b = \frac{8}{9}$ (Muto, personal communication, 2008; these values were also used to produce the results in Muto & Beck 2008).

Adopting the current best approximation to the model covariance for the Metropolis proposal density has the advantage that sampling automatically adapts to both trade-offs between model parameters and variations in model parameter resolution. For example, if we have a two-variable problem in which the first model parameter is well resolved and the second is poorly resolved, then the first parameter will have a small posterior sample variance, whereas the second will have a large sample variance. The sampler will then update the current set of samples by taking random walk steps that make small changes to the value of the first parameter and large changes to the second parameter, efficiently exploring the range of possible values for both. As a second example, consider a two-variable parameter vector in which the values of the two model parameters trade-off with each other, resulting in a large sample model covariance. This information will be passed to the sampler through the proposal PDF covariance, and the random walk will take large steps in the direction of strongest correlation and small steps in the perpendicular direction, optimizing the efficiency of the sampler and obviating the need to carefully choose the model design to avoid trade-offs.

The steps of the basic CATMIP algorithm without cascading and a schematic illustration of these steps are given in Table 1 and Fig. 3, respectively. (We will introduce cascading in Section 3.3.) The behaviour of the CATMIP algorithm while sampling a biased mixture of 2-D Gaussians is shown in Fig. 4. In this example of a bi-modal target PDF, the proposal PDF is initially broad and oriented across the two regions of high probability, allowing the random walkers to efficiently visit both high-probability areas. In later cooling steps, the proposal PDF becomes more highly peaked, and the random walk chains tend to 'freeze' into peaks as $\beta \to 1$, allowing the accumulation of samples within each peak of the target PDF. So, though our proposal density is not necessarily optimally efficient for multiply peaked model spaces, our adaptive proposal PDF based on $\mathbf{C_m}$ can still improve the exploration of model parameters in multiply peaked model spaces.

The number of samples (as specified by both the number of Markov chains, N, and the length of each Markov chain, N_{steps}) needed to adequately represent the posterior distribution is mainly governed by the number of model parameters in $\boldsymbol{\theta}$, with more and longer chains needed to simulate higher-dimensional models. The 'Curse of Dimensionality' (Bellman 1957) requires that the number of samples, N, be large enough to represent the target PDF. The length of each Markov chain must be at least long enough to exceed the 'burn-in' period and, in practice, must be significantly longer to allow sufficient exploration of the parameter space. Appropriate

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

Table 1. CATMIP algorithm.

(1)	Set $m = 0$. Generate N samples $\{\boldsymbol{\theta}_0\} = \{\boldsymbol{\theta}_{0,1} \ldots \boldsymbol{\theta}_{0,N}\}$ from the prior PDF $f_0 = \mathrm{p}(\boldsymbol{\theta})$.
(2)	Set $m = m + 1$. Choose β_m such that the $c_v[w]$ equals some target value, where w is the vector of N weights in eq. (10) for all N samples $\{\boldsymbol{\theta}_{m-1}\}$.
(3)	Calculate $\boldsymbol{\Sigma}_\mathbf{m} = c_m^2 \mathbf{C}_\mathbf{m}$ using eq. (13).
(4)	Draw N samples from $\{\boldsymbol{\theta}_{m-1}\} = \{\boldsymbol{\theta}_{m-1,1}, ..., \boldsymbol{\theta}_{m-1,N}\}$ with probability distribution $\{p_{m-1}\}$ from eq. (11). This set of N resampled models is $\{\hat{\boldsymbol{\theta}}_{m-1}\}$.
(5)	Use each resampled model in $\{\hat{\boldsymbol{\theta}}_{m-1}\}$ as the seed for generating N_{steps} models from the Metropolis algorithm using a Gaussian proposal density with covariance $\boldsymbol{\Sigma}_\mathbf{m}$.
(6)	Collect $\{\boldsymbol{\theta}_m\}$, the set of samples comprised of the final model from each of the N Markov chains. Thus, the total number of samples, N, is unchanged.
(7)	Repeat steps 2–6 until sampling at $\beta_M = 1$ is completed.

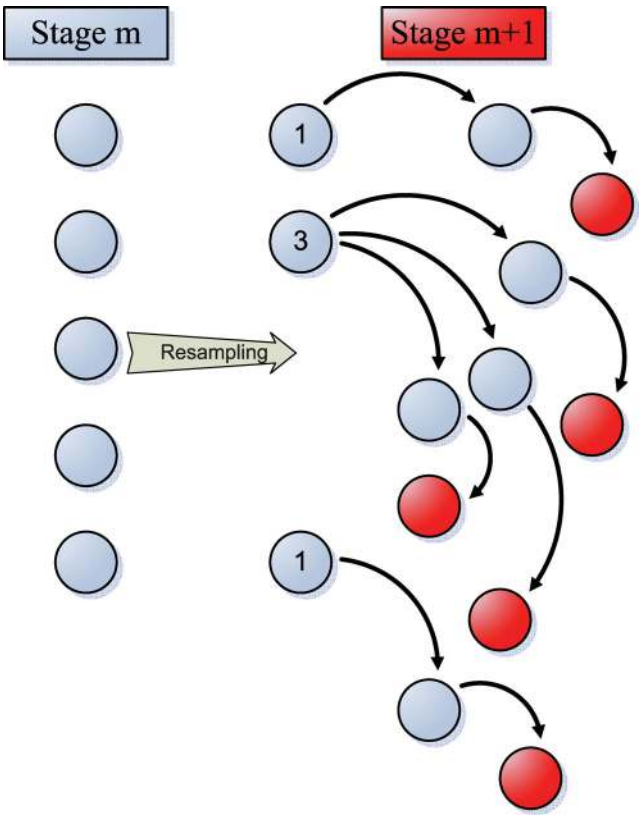

Figure 3. CATMIP algorithm schematic. This cartoon illustrates one complete cooling stage of the CATMIP algorithm. The five samples from β_m are resampled and then an instance of the Metropolis algorithm is run for each of the resulting samples. Numbers indicate the frequency of each model after resampling. The five red samples comprise the posterior distribution for β_{m+1}. The algorithm is plotted with very short Markov chains and a 100 per cent acceptance rate for simplicity. In applications, the Markov chains would be much longer and the acceptance rate much lower.

values for N and N_{steps} can be determined through preliminary performance tests in which synthetic data from known source models are inverted for a variety of values of N and N_{steps} to determine the minimum number of samples needed to recover the source model with sufficient precision. The results of a series of such trials for a number of fault parametrizations will be presented in Figs 11 and 12 in Section 5.

Altogether, CATMIP contains a number of different features which increase the efficiency with which it can sample even complex and high-dimensional PDFs. Use of multiple Markov chains allow the sampler to explore a wider range of the parameter space. Since we only keep the final sample from each Markov chain, these samples are much less correlated than if a single Markov chain was used. The sampling efficiency of the Markov chains is optimized by using a proposal PDF based on the current best estimate of the target PDF being simulated, and the proposal PDF is rescaled according to the sampler's rejection rate. Transitioning acts to make the sampling process easier by not only ensuring that our current population of samples is always nearly in equilibrium with the current target PDF but also by allowing the sampler to begin working over a broader support region of this target PDF. Finally, through resampling, individual samples can be 'teleported' directly from regions of low probability to regions of high probability without having to first random walk to the new location.

3.3 Cascading

To handle even larger parameter spaces, we wrap the basic sampling algorithm in an approach we call cascading, in which we analyse a subset of the data and model parameters and then use the resulting posterior PDF as the basis of the prior PDF for the full inverse problem. Consider a case in which we have two data sets, $\mathbf{D}_1$ and $\mathbf{D}_2$, each of which may contain multiple types of data, and that we can divide the parameters in $\boldsymbol{\theta}$ into two groups so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ are taken as independent *a priori*. Further, suppose that the data likelihood for $\mathbf{D}_1$ depends only on model parameters $\boldsymbol{\theta}_1$ while the data likelihood for $\mathbf{D}_2$ depends on both $\boldsymbol{\theta}_1$ and additional model parameters $\boldsymbol{\theta}_2$, and that given $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, our predictions for $\mathbf{D}_1$ and $\mathbf{D}_2$ are independent. We can write our posterior distribution as

$$\begin{aligned}\mathrm{p}(\boldsymbol{\theta}|\mathbf{D}) &\propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{D}|\boldsymbol{\theta}) \\ &= \mathrm{p}(\boldsymbol{\theta}_1)\mathrm{p}(\boldsymbol{\theta}_2)\mathrm{p}(\mathbf{D}_1|\boldsymbol{\theta}_1)\mathrm{p}(\mathbf{D}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= [\mathrm{p}(\boldsymbol{\theta}_1)\mathrm{p}(\mathbf{D}_1|\boldsymbol{\theta}_1)]\mathrm{p}(\boldsymbol{\theta}_2)\mathrm{p}(\mathbf{D}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &\propto \mathrm{p}(\boldsymbol{\theta}_1|\mathbf{D}_1)\mathrm{p}(\boldsymbol{\theta}_2)\mathrm{p}(\mathbf{D}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \end{aligned} \quad (14)$$

Thus, we can first update $\boldsymbol{\theta}_1$ using $\mathbf{D}_1$, and then update $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ using $\mathbf{D}_2$ with the joint prior PDF $\mathrm{p}(\boldsymbol{\theta}_1|\mathbf{D}_1)\mathrm{p}(\boldsymbol{\theta}_2)$.

We can incorporate eq. (14) into our transitioning scheme by rewriting our transitional distribution in eq. (9) as,

$$f(\boldsymbol{\theta}|\mathbf{D}, \beta_m, \gamma_n) \propto \mathrm{p}(\boldsymbol{\theta}_1)\mathrm{p}(\boldsymbol{\theta}_2)\mathrm{p}(\mathbf{D}_1|\boldsymbol{\theta}_1)^{\beta_m}\mathrm{p}(\mathbf{D}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{\gamma_n}. \quad (15)$$

To sample this joint distribution, use the algorithm in Table 1 twice to sample the following distributions:

$$\begin{aligned} &1.\ f(\boldsymbol{\theta}_1|\mathbf{D}_1, \beta_m) \propto \mathrm{p}(\boldsymbol{\theta}_1)\mathrm{p}(\mathbf{D}_1|\boldsymbol{\theta}_1)^{\beta_m} \\ &\qquad 0 \le \beta_m \le 1 \\ &2.\quad f(\boldsymbol{\theta}|\mathbf{D}, \gamma_n) \propto \mathrm{p}(\boldsymbol{\theta}_1)\mathrm{p}(\mathbf{D}_1|\boldsymbol{\theta}_1)\mathrm{p}(\boldsymbol{\theta}_2)\mathrm{p}(\mathbf{D}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{\gamma_n} \\ &\qquad 0 \le \gamma_n \le 1. \end{aligned} \quad (16)$$

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

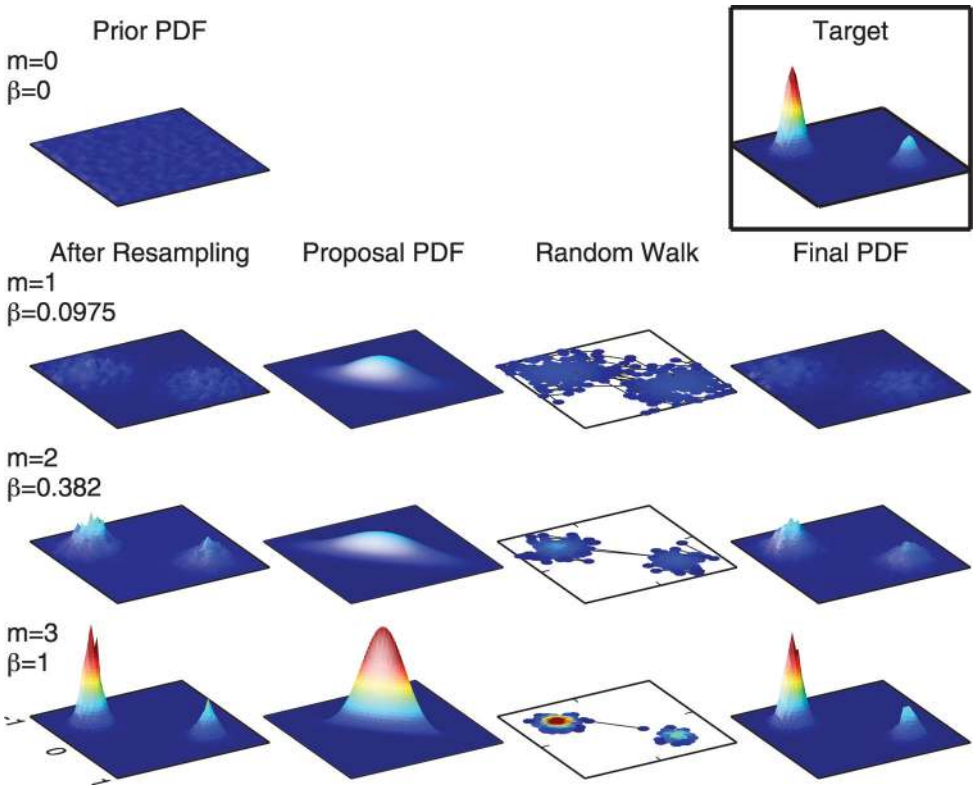

Figure 4. CATMIP algorithm example. The algorithm begins by directly sampling the prior distribution (row 1). A new value for β is calculated and the distribution is resampled (column 1). The covariance of samples and acceptance rate is used to design a proposal PDF (column 2) for use in the Metropolis algorithm (column 3). The final sample from each of 10 000 Markov chains comprise the new PDF (column 4). In this example, the prior distribution is uniform and the target distribution is the sum of two Gaussians one of which has a factor of three greater amplitude than the other. Both Gaussians in the target PDF have variance $\Sigma_{ii} = 0.01$ and covariance $\Sigma_{ij} = 0$. The target PDF is plotted in the top right corner for reference.

Cascading allows us to sample the full joint posterior PDF created from a variety of different deterministic models, different error models and/or different data sets by first sampling a smaller and potentially much lower-dimensional inverse problem and then leveraging the information from that distribution to more efficiently simulate the solution to the full inverse problem. Our final posterior PDF is exactly the joint posterior PDF for the joint inverse problem, and thus cascading is not an approximation. Cascading is simply a substantially more efficient way to sample the posterior PDFs for large inverse problems.

3.4 CATMIP versus TMCMC versus Metropolis

To evaluate the efficiency of CATMIP, we compare CATMIP to TMCMC and the Metropolis algorithm. Loosely based on Example 2(VIII) in Ching & Chen (2007), we used all three samplers to simulate a target distribution which is a biased mixture of 10-D Gaussians:

$$0.1\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}_{10}) + 0.9\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2\mathbf{I}_{10})$$

$$\boldsymbol{\mu}_1 = \left[\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right]$$

$$\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$$

$$\sigma = 0.1$$

$$\mathbf{I}_{10} = 10 - \text{by} - 10 \text{ identity matrix.} \qquad (17)$$

The prior PDF for this test is the uniform distribution $\mathcal{U}(-2, 2)$.

Both CATMIP and TMCMC are run with a target c_v of 1. All three samplers were set up so that they drew approximately 400 000 samples and thus had equal computational expenses. Note that we cannot prescribe the total number of samples precisely because both CATMIP and TMCMC choose the cooling (or transitioning) schedule dynamically and thus the final number of model evaluations is not known in advance of running the algorithm. The Metropolis algorithm was run for 400 000 samples with the random walk initiating at the origin. TMCMC was run with 20 000 samples; it took 20 cooling steps to complete ($M = 19$ in eq. 9), for a total of 400 000 samples over the lifetime of the algorithm. CATMIP was run with $N = 2200$ Markov chains where each Markov chain required 15 forward model evaluations, or 33 000 samples per cooling step; it completed in 13 cooling steps ($M = 12$), for a total of 398 200 samples over the lifetime of the algorithm.

The marginal distributions for one dimension of the target distribution are shown in Fig. 5. (These are calculated by taking the ith component of all samples $\boldsymbol{\theta}$.) The Metropolis algorithm has difficulty with the multimodal distribution, and becomes trapped in one peak of the target PDF. (Of course, given enough samples, the Metropolis random walk would eventually find the other peak.) TMCMC and CATMIP image both peaks of the distribution, but TMCMC fails to reproduce the relative amplitudes of the two peaks, demonstrating the importance of the fact that CATMIP includes more MCMC exploration of the parameter space than TMCMC. Overall, CATMIP can better reproduce the target distribution with less computational expense.

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

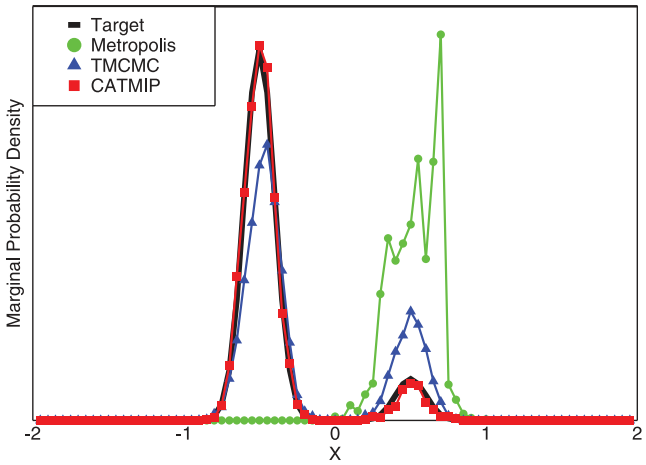

Figure 5. Comparison of CATMIP, TMCMC and Metropolis algorithms. The marginal distributions for one of the 10 spatial axes is shown.

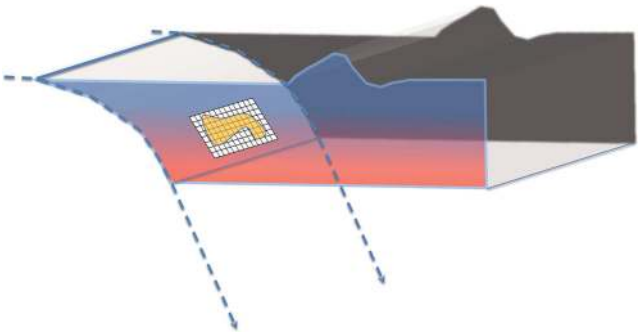

Figure 6. Cartoon showing typical parametrization of fault slip. The fault surface is discretized into a set of patches. The sampling process then finds the distribution of the average slip in that patch.

4 A BAYESIAN FINITE FAULT PARAMETRIZATION

As the name suggests, a finite fault earthquake source model consists of the history of an earthquake source (its spatial and temporal evolution) over a fault surface of finite extent (Fig. 6). Thus, we must determine all faulting parameters at many points in space and time. For Bayesian sampling, we need a model with a tractable number of free parameters, a computationally fast forward model and a prior distribution. There is not much flexibility when choosing the spatial complexity of the fault model: it is mostly determined by the spatial resolution of the available data and the frequency content of the kinematic data being used. So our only option is to describe each rupture source in space with as few parameters as possible. In an attempt to balance computational and sampling cost with reasonable flexibility in the source model, we use four faulting parameters per source: slip in two directions, rupture velocity and source duration. This model, thus, requires that each point only ruptures once with a prescribed functional form for its temporal evolution.

4.1 Kinematic source model

The displacements due to a kinematic seismic source in an elastic medium can be represented by the sum of a series of discrete sources,

$$\hat{d}_i(\boldsymbol{\zeta}, t) = \sum_{j=1}^{2} \sum_{k=1}^{n_s} U_j^k \tilde{g}_{i,j}^k(\boldsymbol{\zeta}, t - t_0^k | T_r^k), \tag{18}$$

where k is an index over all n_s source locations, U_j^k is the final slip in the jth direction at the kth source, and $\tilde{g}_{i,j}^k(\boldsymbol{\zeta}, t - t_0^k | T_r^k) = \int_0^{\min(t-t_0^k, T_r^k)} s(\tau | T_r^k) \tilde{G}_{i,j}^k(\boldsymbol{\zeta}, t - \tau) \mathrm{d}\tau$ given that $\tilde{G}$ is the Green's function that maps a unit dislocation in the jth direction at the kth source location to the ith direction at receiver location $\boldsymbol{\zeta}$, $s(\tau | T_r^k)$ is a source-time function with slip duration T_r^k at the kth source location and t_0^k is the time of rupture initiation at the kth source location. The index j would normally run 1 to 3 but, since we are interested in shear dislocations and not tensile faults, we will only consider the two components of slip that lie parallel to the fault plane. Also, we choose to parametrize our source-time function, s, as a triangle.

There are four free parameters in eq. (18) for each point source: two components of slip, slip duration and initial rupture time. Rather than directly modelling t_0 (for which there is no intuitive *a priori* knowledge), we instead solve for a rupture velocity, V_r, at each source location. To determine how much to time-shift each source in our model, we then map our hypocentre location and rupture velocity field into initial rupture times at each patch. This mapping can be done quickly and efficiently using the Fast Sweeping Algorithm (Zhao 2005), a level-set method which uses a Godunov upwind difference scheme (Rouy & Tourin 1992) for solving the eikonal equation. This V_r-based parametrization ensures that the resulting rupture propagation is causal.

Eq. (18) convolves the source-time function, $s(\tau | T_r^k)$, and the point source Green's function, $\tilde{G}$, to evaluate the modified Green's function, $\tilde{g}$. We pre-compute a set of modified Green's functions for a wide variety of values for T_r, and, at each forward model evaluation, use a stored version of $\tilde{g}$. This approach significantly increases efficiency since convolving the source-time function is one of the costliest parts of evaluating the kinematic forward model.

Our kinematic model has one triangular source-time function per patch and freely varying rupture velocity. This model design is almost the opposite approach to Kikuchi & Kanamori (1982) who used a complex source-time function with a fixed rupture velocity. Cohee & Beroza (1994a) concluded, somewhat unsurprisingly, that the former approach does a better job of recovering the rupture velocity of the source at the expense of doing a worse job at estimating rise time; but they also found that source models which can only rupture once do a better job at estimating the seismic moment.

The data likelihood comes from our kinematic stochastic forward model (as described in Section 2.2),

$$\mathrm{p}(\mathbf{D_k} | \boldsymbol{\theta}) = \mathrm{p}(\mathbf{D_k} | \boldsymbol{\theta_s}, \boldsymbol{\theta_k}) = \frac{1}{(2\pi)^{\frac{N_k}{2}} |\mathbf{C}_\chi^k|^{\frac{1}{2}}} e^{-\frac{1}{2}[\mathbf{D_k} - \mathbf{G_k}(\boldsymbol{\theta}) - \boldsymbol{\mu_k}]^T \mathbf{C}_\chi^{k\,-1} [\mathbf{D_k} - \mathbf{G_k}(\boldsymbol{\theta}) - \boldsymbol{\mu_k}]}, \tag{19}$$

where $\mathbf{D_k}$ is an N_k-dimensional vector of the observed kinematic time-series such as seismograms or high-rate GPS, $\mathbf{G_k}(\boldsymbol{\theta}) = \hat{\mathbf{d}}_\mathbf{k}$ is the corresponding output vector of the kinematic forward model (eq. 18) and $\boldsymbol{\theta}$ is a vector of model parameters. $\boldsymbol{\mu_k}$ is the combined bias of our observation and prediction errors, and may be taken to be $\mathbf{0}$. Following the cascading approach (Section 3.3), we separate $\boldsymbol{\theta}$ into $\boldsymbol{\theta_s}$, the set of parameters sufficient to define the static source model, and $\boldsymbol{\theta_k}$, a vector of kinematic fault parameters. $\boldsymbol{\theta_s}$ is identified with the vector of $2 \times n_s$ slip parameters, U_j^k, in eq. (18). $\boldsymbol{\theta_k}$ contains T_r and V_r for each source and, optionally, the location

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

of the hypocentre on the fault surface. $\mathbf{C}_\chi^k$ is a covariance matrix which models the uncertainty from the measurement errors and from model prediction errors as introduced in Section 2. ($\mathbf{C}_\chi^k$ is used instead of $\mathbf{C}_\chi$ to emphasize that this is the covariance matrix for the kinematic model.)

4.2 Static source model

In the static case, the time evolution of the seismic source drops out and eq. (18) simplifies to,

$$\hat{d}_i(\boldsymbol{\zeta}) = \sum_j \sum_{k=1}^{n_s} U_j^k \cdot \tilde{g}_{i,j}^k(\boldsymbol{\zeta}, \infty). \tag{20}$$

We can rewrite the above in matrix notation for a discrete set of observation locations,

$$\hat{\mathbf{d}}_\mathbf{s} = \mathbf{G}_\mathbf{s} \cdot \boldsymbol{\theta}_\mathbf{s},$$

where $\boldsymbol{\theta}_\mathbf{s}$ is the same as in eq. (19), $\mathbf{G}_\mathbf{s}$ is a matrix of Green's functions, and $\hat{\mathbf{d}}_\mathbf{s}$ is a vector of data predictions.

Introducing a covariance matrix, $\mathbf{C}_\chi^s$, and a bias, $\boldsymbol{\mu}_\mathbf{s}$ (which again may be taken to be $\mathbf{0}$), as in eq. (19), the associated data likelihood for static data, $\mathbf{D}_s$, a given N_s-dimensional vector, is,

$$\mathrm{p}(\mathbf{D}_\mathbf{s}|\boldsymbol{\theta}_\mathbf{s}) = \frac{1}{(2\pi)^{\frac{N_s}{2}} |\mathbf{C}_\chi^s|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{D}_\mathbf{s}-\mathbf{G}_\mathbf{s}\cdot\boldsymbol{\theta}_\mathbf{s}-\boldsymbol{\mu}_\mathbf{s})^T \mathbf{C}_\chi^{s\,-1}(\mathbf{D}_\mathbf{s}-\mathbf{G}_\mathbf{s}\cdot\boldsymbol{\theta}_\mathbf{s}-\boldsymbol{\mu}_\mathbf{s})}. \tag{21}$$

Note that $\boldsymbol{\theta}_\mathbf{s}$ is a subvector of $\boldsymbol{\theta}$ in eq. (19). In fact, $\mathbf{D}_\mathbf{s}$ could be viewed as a subset of $\mathbf{D}_\mathbf{k}$ since it contains just the static part of the measured kinematic time history.

4.3 Choice of prior distribution

The Bayesian formulation of the inverse problem requires that we specify a prior distribution to use with our model. The one source parameter of which we usually have good *a priori* information is the seismic moment tensor (from teleseismic data). Although the seismic moment tensor does not tell us anything about the distribution of slip, it tells us something about the average slip direction (rake) and the total amount of slip summed over all fault patches. To turn these observations into a prior distribution of slip, we use a rotated coordinate system with one axis, $U_\parallel$, aligned with the teleseismic rake direction and the other axis, $U_\perp$, perpendicular to the rake direction (Fig. 7). [For efficiency, we rotate our Green's functions into ($U_\perp$, $U_\parallel$) coordinates thus eliminating the need to transform the forward model for each likelihood evaluation.] The prior on $U_\perp$ is a Gaussian with zero mean and standard deviation, σ, that is chosen to be smaller than the anticipated magnitude of the parallel component of slip (since we assume that $U_\parallel$ is aligned with the dominant slip direction). Thus, we allow variation in the rake but assume that the most probable variation is no variation. For the prior on $U_\parallel$, we use a one-sided semi-infinite uniform prior probability: we allow any positive value for $U_\parallel$, but forbid large amounts of back-slip. (This choice of prior is equivalent to a uniform distribution between a small negative amount of slip and some large positive value of slip, $u_{\max}$, which the sampler will never reach.) We allow for slightly negative $U_\parallel$ to avoid undersampling models with small slips due to the hard bound on minimum $U_\parallel$. Thus our prior distribution on the static slip model is,

$$\begin{aligned}\mathrm{p}(\boldsymbol{\theta}_\mathbf{s}) &= \mathrm{p}(\mathbf{U}_\perp)\mathrm{p}(\mathbf{U}_\parallel) \\ &= \mathcal{N}\left(0, \sigma^2 \mathbf{I}_{\mathbf{n_s}}\right) \mathcal{U}(u_{\min}, u_{\max})^{n_s},\end{aligned} \tag{22}$$

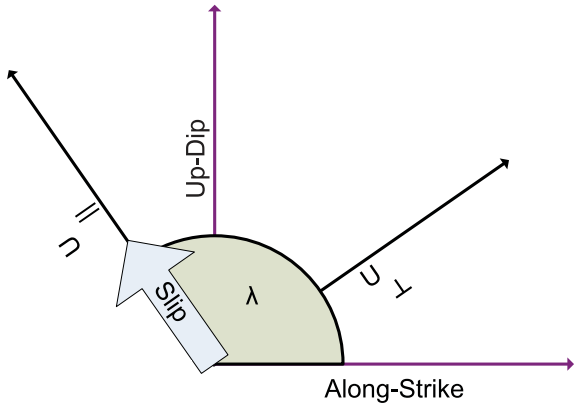

Figure 7. Slip coordinate system. The Bayesian finite fault parametrization uses components of slip, $U_\perp$ and $U_\parallel$, which lie in the fault plane but are orthogonal to each other. $U_\parallel$ is aligned with the direction of hangingwall motion given by the rake angle, λ, which is chosen from the catalogue moment tensor solution for the earthquake.

where $\mathbf{I}_{\mathbf{n_s}}$ denotes the n_s-by-n_s identity matrix, $\mathcal{U}(u_{\min}, u_{\max})^{n_s}$ is the uniform PDF in n_s dimensions, $\mathbf{U}_\perp$ and $\mathbf{U}_\parallel$ are vectors in $\mathbb{R}^{n_s}$ and n_s represents the number of discrete source locations in our forward model (or, equivalently, the number of patches on our tessellated fault plane), as in eq. (18).

For the transitioning described in Section 3.2, we must first simulate $f(\boldsymbol{\theta}|\mathbf{D}, 0) = \mathrm{p}(\boldsymbol{\theta})$. Thus, we must draw samples from the prior slip distributions. Since our prior on $U_\perp$ is a zero-mean Gaussian, the average net perpendicular slip of each slip model is zero. So the moment of each random slip model is controlled by $U_\parallel$. If initially we drew samples from the uniform distribution, $\mathcal{U}(u_{\min}, u_{\max})$, our initial population of models would include slip models whose corresponding moments spanned many orders of magnitude. So for efficiency, we begin at $m = 0$, $\beta = 0$, with a population of samples which have plausible moment magnitudes. We accomplish this by drawing our initial samples of slip in the $U_\parallel$ direction from the Dirichlet distribution which is applicable to problems where a finite number of states must sum to a particular value (see e.g. Gelman *et al.* 2004). A n_s-dimensional sample of the Dirichlet distribution produces a set of n_s positive random numbers which sum to a specified value, allowing us to pre-determine the total moment for each random slip model in $f(\boldsymbol{\theta}|\mathbf{D}, 0)$ (Fig. 8). We take a Gaussian uncertainty on the moment magnitude, M_w, for the earthquake. The standard deviation of the Gaussian is set to 0.5 magnitude units based on our intuition about the typical errors associated with magnitudes reported in seismic catalogues. For each prior sample, we draw a magnitude from a Gaussian distribution and then generate random slips on each patch using the Dirichlet distribution so that the total slip equals the proposed magnitude (Fig. 9). The use of samples from the Dirichlet distribution is a shortcut to ensure that the initial pool of models have enough probable models. Otherwise, almost all of the initial models would be useless for future cooling steps since even a slight inclusion of the data likelihood would reject the models with slips corresponding to moments that were wrong by an order of magnitude or more. It should be emphasized that this entire process of drawing samples from the Dirichlet distribution scaled by a reasonable array of seismic moments is simply our method of generating our seed models for initializing CATMIP. Neither the Gaussian distribution on M_w nor the Dirichlet PDF are used as priors on our slip model and thus have no effect on which candidate models are accepted or rejected for the lifetime of the CATMIP algorithm. Our use of the Dirichlet distribution for

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

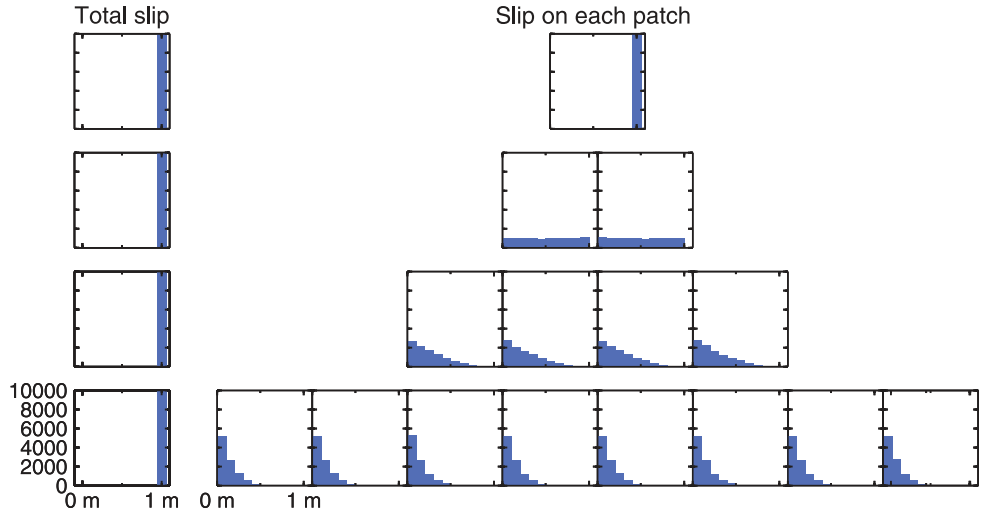

Figure 8. Dirichlet distribution. One sample of a k-dimensional Dirichlet distribution produces k random numbers which sum to one. Here, we plot the results of 10 000 draws from the Dirichlet distribution for one, two, four and eight patch slip models (right). The distribution becomes more highly peaked near zero as the number of patches increases. This behaviour can be intuited from the fact that in order for any one patch to have slip approaching one, the slips on all other patches must approach zero. Since each patch has an equal probability of observing a given slip value, for each time that a particular patch has a value near one, it approaches nearly zero $k - 1$ times. Note that the sum of the slip on the patches equals 1 for each of the 10 000 draws (left).

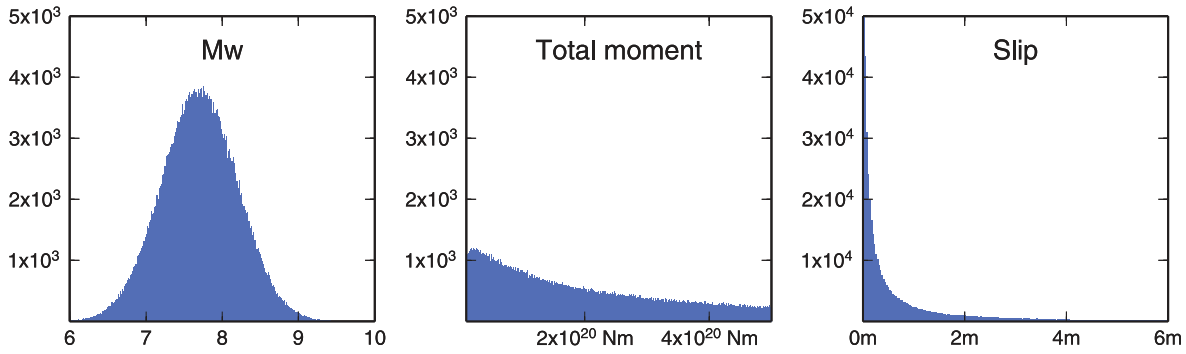

Figure 9. Methodology for generating initial slip distributions. We describe the magnitude of the earthquake by a Gaussian distribution centred on the catalogue magnitude for the event (left-hand panel). This yields a log-normal distribution for scalar seismic moment (centre panel). We then use the Dirichlet distribution to translate the distribution of total moment for the earthquake into a distribution of slip on each patch, and we assume that this slip is aligned with the rake direction of the earthquake as shown for a selected fault patch (right-hand panel).

initialization purposes merely increases the efficiency of our sampling by ensuring that our random walk seeds come from models which we cannot eliminate *a priori* due to unreasonably low or high seismic moments.

We use uniform priors on T_r and V_r. The bounds on T_r are based on reasonable slip durations. V_r is allowed to vary between 0 and the P-wave velocity of our elastic Earth structure (if we want to include all physically allowable rupture velocities) or the S-wave velocity (if we want to forbid super-shear rupture). We choose the prior PDF on the hypocentre position, $\mathbf{H_0}$, to be Gaussian with its mean ($\boldsymbol{\mu}_{\mathbf{H_0}}$) centred on a location from an earthquake catalogue or independent study, and its covariance ($\boldsymbol{\Sigma}_{\mathbf{H_0}}$) based on the formal error associated with that hypocentre location or our intuition about typical hypocentre location errors. (For the example in Section 5, we will assume that our hypocentre location has a standard deviation of 10 km.) We define the hypocentre location by two coordinates: the distance of the hypocentre along-strike and the distance of the hypocentre downdip. Thus, $\mathbf{H_0}$ is a two-element vector, and $\mathbf{T_r}$ and $\mathbf{V_r}$ are each n_s-element vectors where n_s is the number of discrete seismic sources. We can then write our prior on the kinematic model as,

$$\mathrm{p}(\mathbf{m}) = \mathrm{p}(\boldsymbol{\theta}_\mathbf{s}, \boldsymbol{\theta}_\mathbf{k}) = \begin{cases} \mathrm{p}(\boldsymbol{\theta}_\mathbf{s})\mathcal{U}(\mathrm{T_{r_{min}}}, \mathrm{T_{r_{max}}})^{\mathrm{n_s}}\,\mathcal{U}(\mathrm{V_{r_{min}}}, \mathrm{V_{r_{max}}})^{\mathrm{n_s}} \\ \qquad \text{for fixed hypocentre} \\ \mathrm{p}(\boldsymbol{\theta}_\mathbf{s})\mathcal{U}(T_{r_{\min}}, T_{r_{\max}})^{n_s}\,\mathcal{U}(V_{r_{\min}}, V_{r_{\max}})^{n_s}\,\mathcal{N}(\boldsymbol{\mu}_{\mathbf{H_0}}, \boldsymbol{\Sigma}_{\mathbf{H_0}}) \\ \qquad \text{for uncertain hypocentre.} \end{cases} \tag{23}$$

4.4 Implementation of cascading

For many earthquakes we have both kinematic and static data, allowing us to take advantage of the cascading technique. As the static data depend only on the static slip distribution, we can use the posterior static slip distribution as a prior distribution for a full kinematic model which also includes rupture velocity, slip duration, and possibly hypocentre location. Given our static parameters ($\boldsymbol{\theta}_\mathbf{s}$) and our kinematic-only parameters ($\boldsymbol{\theta}_\mathbf{k}$), the posterior PDF of the

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

full model ($\boldsymbol{\theta} = [\boldsymbol{\theta}_\mathbf{s}, \boldsymbol{\theta}_\mathbf{k}]$) can be written as (see eq. 14)

$$\begin{aligned} \mathrm{p}(\boldsymbol{\theta}|\mathbf{D}) &\propto \mathrm{p}(\boldsymbol{\theta}_\mathbf{s})\mathrm{p}(\boldsymbol{\theta}_\mathbf{k})\mathrm{p}(\mathbf{D}_\mathbf{s}|\boldsymbol{\theta}_\mathbf{s})\mathrm{p}(\mathbf{D}_\mathbf{k}|\boldsymbol{\theta}_\mathbf{k}, \boldsymbol{\theta}_\mathbf{s}) \\ &\propto \mathrm{p}(\boldsymbol{\theta}_\mathbf{s}|\mathbf{D}_\mathbf{s})\mathrm{p}(\boldsymbol{\theta}_\mathbf{k})\mathrm{p}(\mathbf{D}_\mathbf{k}|\boldsymbol{\theta}_\mathbf{k}, \boldsymbol{\theta}_\mathbf{s}), \end{aligned} \tag{24}$$

where $\boldsymbol{\theta}_\mathbf{s} = (\mathbf{U}_\perp, \mathbf{U}_\parallel)$ and $\boldsymbol{\theta}_\mathbf{k} = (\mathbf{T}_\mathbf{r}, \mathbf{V}_\mathbf{r})$, for fixed hypocentre, or $\boldsymbol{\theta}_\mathbf{k} = (\mathbf{T}_\mathbf{r}, \mathbf{V}_\mathbf{r}, \mathbf{H}_\mathbf{0})$, for uncertain hypocentre.

When fusing static and kinematic earthquake data into updating a single source model, we solve the static problem first. Then, through cascading, we use the posterior distribution from modelling the static data to make the prior distribution for the full kinematic problem. This approach significantly decreases the computational cost in multiple ways. First, we can treat the model as two problems, one of which has half as many free parameters as the full problem without cascading. Secondly, we can explore the parameter space of the slip distribution using only the static forward model, which is significantly faster to compute than the full kinematic forward model. (The kinematic forward model is about an order of magnitude computationally slower than the static forward model.) Third, once we have sampled the slip distribution based on the static data, we have explored a large part of the joint kinematic parameter space, making it much easier to sample for the kinematic model than if we started without any knowledge of the slip distribution. Finally, cascading has the additional value that it allows us to assess the progressive impact of adding additional data sets.

4.5 Model prediction error

As discussed in Section 2, the deficiencies in our Earth structure model, source parametrization and data interpretation can far exceed the lack of accuracy in our observations. To determine the effects of these additional error sources on our solution and thus update our *a posteriori* uncertainties, we include the uncertainty in the model prediction error. For our particular problem, shortcomings in our Earth structure are probably the single largest source of error and these errors increase with the size of the source displacement. Consider a set of Green's functions which have a 10 per cent error so that if a given source displacement should produce 10 units of surface displacement these Green's functions only predict 9 units of displacement, leaving a residual between the observations and the model's predictions of 1 unit. Then an earthquake that was an order of magnitude larger would produce 10 times as much source displacement and 100 units of surface displacement, but our Green's functions yield only 90 units of slip leaving a residual misfit of 10. The size of the residual is not constant but instead grows proportionally with the size of the input. We could parametrize our model prediction error variance as a percentage of $\mathbf{G}(\mathbf{m})$, but this would bias our solution since models which overpredicted the observations would be more plausible than models that underpredict because the larger output models would be accompanied by larger prediction errors. Instead, we use our observations as a proxy for $\mathbf{G}(\mathbf{m})$ and adopt a Gaussian distribution for the predicted data (eq. 4) with covariance matrix, $\mathbf{C}_\chi$, that has the form,

$$\begin{aligned} \mathbf{C}_\chi &= \mathbf{C}_\mathbf{d} + \mathbf{C}_\mathbf{p} \\ &= \mathbf{C}_\mathbf{d} + \alpha^2 \mathrm{diag}(D_1^2, ..., D_{N_{\mathrm{dp}}}^2), \end{aligned} \tag{25}$$

where α represents the fractional error of our forward model. Note that, since we can specify a different $\mathbf{C}_\chi$ for each data set, we can solve for a different fractional error, α, for each data set. Thus we need not assume that the Green's functions for tsunami data contain the same errors as the Green's functions for GPS data, and we can let the inversion determine the errors associated with the prediction errors for each data set (and thus, in a sense, the appropriate relative weights for each data set).

The parameter α must be non-negative. Tarantola (2005) argues that positive scale quantities, termed Jeffreys parameters after Jeffreys (1939), should be replaced with the log of that parameter to both acknowledge that the inverse of the parameter could be used in its place and to preserve the scale invariance of both the original and inverse quantities. Thus, we sample for $\ln\alpha$ instead of α, and adopt a Gaussian prior on $\ln\alpha$. As with all prior PDFs, our choice for the prior PDF on $\ln\alpha$ must be based on our *a priori* intuition about the quantity in question, in this case plausible Green's functions errors due to the fact that the elastic structure is poorly known. For example, say that a researcher felt that it is likely that there is a 5 per cent error in a set of Green's functions but that it is unlikely that these errors exceed 20 per cent so that there is 95 per cent probability that the error is 20 per cent or less. Then the prior on $\ln\alpha$ could be set to $\ln\alpha = \mathcal{N}(\ln 0.05, \sigma)$ where σ is chosen so that the cumulative density function (CDF) at $\ln\alpha = \ln 0.2$ is 0.95. (However, in the examples presented in Section 5, we will use very broad priors on $\ln\alpha$ to demonstrate the ability of the sampler to recover the error in the Green's functions with poor prior information.)

Our parametrization of the model prediction error has the advantage of providing error estimates that are insensitive to (by automatically scaling with) the magnitude of the earthquake, but it lacks any sense of temporal dependence for waveforms (for contrast, see Yagi & Fukahata 2011), and thus may be less optimal for seismic data, tsunami records and the like. To better capture the physics of the real problem, we could develop a more complex model for the prediction error uncertainty than our current model of it. For example, the predicted data from the Green's functions for each source–receiver pair should not be modelled as independent: the predicted surface displacements slowly vary as a function of distance and azimuth from the source and so should be modelled as correlated. Similarly, any error in the elastic structure should produce a slowly spatially varying set of model prediction errors. Thus the model prediction errors at neighbouring station locations should not be modelled as independent but should instead have non-zero covariance. However, choosing an appropriate form for this covariance matrix is beyond the scope of this paper.

Except in the case of very large amounts of data and model parameters, evaluating the data likelihood is quite fast if $\mathbf{C}_\chi$ is fixed *a priori* and $\mathbf{C}_\chi{}^{-1}$ is pre-computed. If the model prediction error variance is included as a free parameter, then $\mathbf{C}_\mathbf{p}$ in eq. (5) changes with every sample of the target PDF, and $\mathbf{C}_\chi{}^{-1}$ must be recalculated for every evaluation of the forward model. In this scenario, it is generally necessary to assume that $\mathbf{C}_\mathbf{d}$ and $\mathbf{C}_\mathbf{p}$, and thus $\mathbf{C}_\chi$, are diagonal, simply to make calculating $\mathbf{C}_\chi{}^{-1}$ computationally efficient. Thus, due to computational expense, we may be forced to choose between either updating a diagonal model prediction variance which ignores the error covariances or instead using the full error covariances while holding $\mathbf{C}_\chi{}^{-1}$ partially or totally fixed.

5 PERFORMANCE TESTS

We present a series of tests of Bayesian finite fault models using synthetic data. These tests demonstrate the performance of our Bayesian sampling scheme and provide intuition about Bayesian inversion of finite fault earthquake models. We apply our methodology to real earthquake data in Paper II.

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

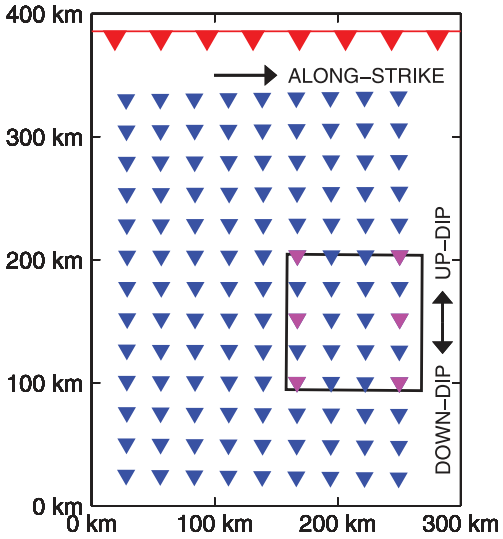

Figure 10. Source–receiver geometry for synthetic finite fault models. Synthetic three-component surface displacements were calculated for each of the locations denoted by blue triangles. Pink triangles represent colocated static and kinematic observations for the synthetic kinematic models. The surface projection of the fault plane is shown with a thick black line. The fault dips towards the bottom of the figure at an angle of 18°. The depth to the top of the fault is 40 km.

5.1 Synthetic static models with abundant data

We first consider the case of a shallowly dipping thrust fault (Fig. 6) located beneath dense geodetic observations (Fig. 10). We use synthetic three-component displacements at 117 surface locations to constrain static finite fault slip models using various parametrizations. The fault is a single plane embedded in a 1-D layered elastic structure. Although the model is perfect and the data noise-free, for the purposes of computing the inverse, the data variance (diagonal elements of $\mathbf{C_d}$) is taken to be 0.1 cm^2 for all observations. Our prior PDFs for each component of slip on each patch are $U_\perp = \mathcal{N}(0\text{m}, 3\text{m})$ and $U_\parallel = \mathcal{U}(-1\text{m}, 10\text{m})$.

Since the slip distribution consists of two components of motion on each fault patch, the total number of free parameters in each synthetic model is twice the number of fault patches. When the fault surface is discretized into a few large patches, the inverse problem is essentially overdetermined, and the resulting posterior PDF is tightly peaked with a mean that perfectly matches the synthetic source model (Fig. 11). As the number of patches (and thus the number of free parameters) increases, the number of samples required to reproduce the synthetic slip distribution increases. At some point, the quality of the mean solution begins to decline, not because of undersampling, but because the patch size has become so small that the data can no longer resolve the model given this source–receiver geometry and assumed error structure. (In the optimization approach, one would say that the inverse problem has become underdetermined.) When model resolution is lost, the displacements on neighbouring patches begin to trade-off with each other. (We note that underparametrization can conversely create small posterior variances that may lead to over confidence in the model.) The mean of all of these possible models results in a smoother slip distribution than the synthetic source model. Our data only resolve local averages and not the slip on each patch individually. So the posterior PDF for each patch is no longer highly peaked. There are many possible models that are consistent with the data and, in real applications, we should consider all of these models.

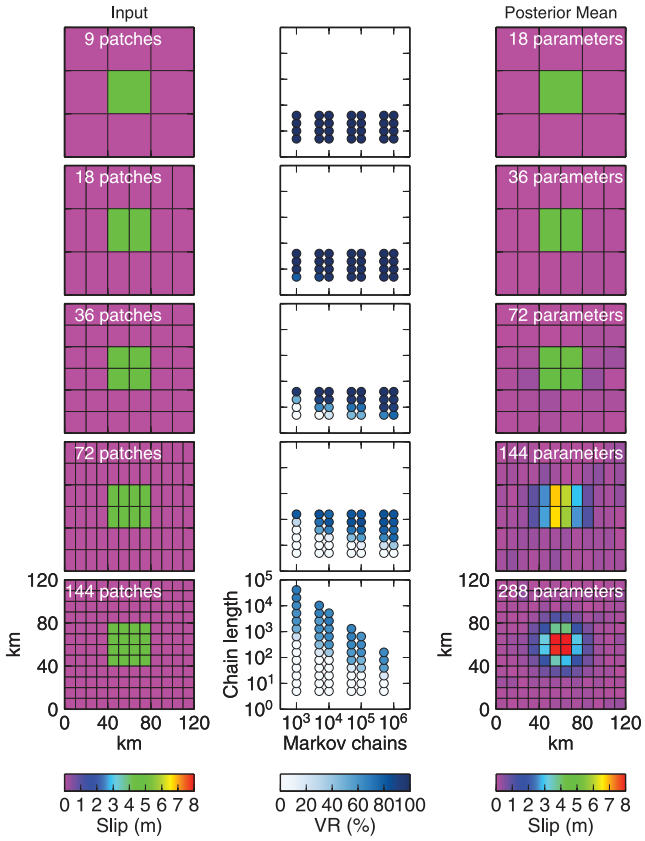

Figure 11. Sampling a synthetic static finite fault model. The left-hand column shows the input, that is the slip distributions used to create the synthetic surface displacements for each test. The quality of the output of CATMIP sampling is shown in the middle column as evaluated by the variance reduction between the input and the mean of the output slip distributions. The number of Markov chains and their lengths correspond to the parameters N and N_{steps} in the CATMIP algorithm, respectively. The mean of the posterior samples for the CATMIP run with the largest number of samples is shown in the right-hand column. The source–receiver geometry for this test is mapped in Fig. 10.

The computational cost of each CATMIP cooling step scales with the number of samples, which is the product of the number of Markov chains with the length of each chain. We see in Fig. 11 that the quality of the posterior distribution is approximately equal for equal products of chain length and number of chains. There is some improvement in the output for sampling runs with longer chain lengths (larger N_{steps}) and smaller number of chains (smaller N), but some of this effect may simply be a manifestation of the need to have sufficiently long random walks to exceed the 'burn-in' period of each Markov chain (Fig. 12). Regardless, N must always be large enough so that the number of samples can fully define the posterior PDF and, to take maximum advantage of CATMIP's embarassingly parallel architecture, N should be a large multiple of the number of worker CPUs.

5.2 Synthetic static models with prediction error estimation

To test the effects of errors in our elastic structure on slip modelling, we generated synthetic observations using the same layered elastic space as in Section 5.1 and then added Gaussian noise to those surface displacements. For this test, we used the correct formal

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

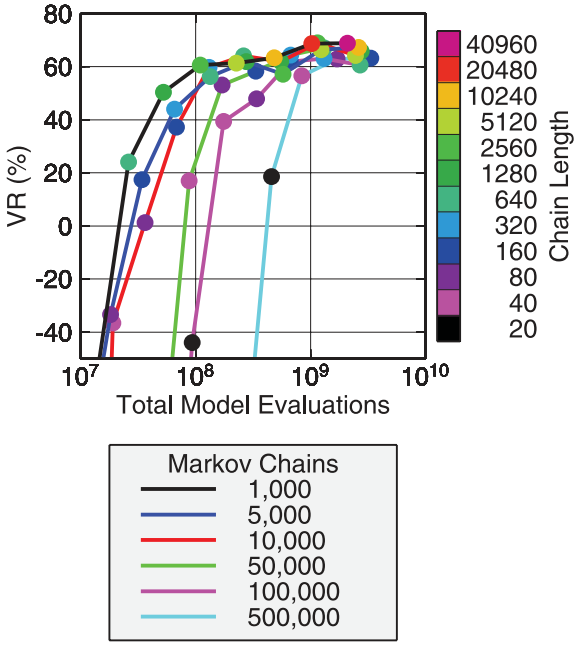

Figure 12. Computational cost for different CATMIP runs using a variety of input parameters. Results using the 144 patch synthetic static slip model in Fig. 11 are shown. The quality of the posterior mean is measured by the variance reduction with respect to the input slip model. The computational cost of CATMIP consists almost entirely of the number of samples drawn (or, equivalently, number of forward models evaluated) over the lifetime of the algorithm. The x-axis is this computational cost. CATMIP has two input parameters: the number of Markov chains and the length of each chain. Each line in the plot represents a different number of Markov chains, and the colour of the symbols indicates the length of those Markov chains. The total number of model evaluations is a function of not only the number of Markov chains and their lengths, but also the number of cooling steps, which is chosen dynamically by CATMIP during the run.

observational error variance, σ_d^2, and source geometry but introduced a prediction error into our forward model by using Green's functions for a homogeneous elastic half-space. The difference between our synthetic observations and the predicted displacements for the true slip distribution propagated through the corrupted forward model are shown in Figs 13 and 14. It is clear from these figures that the formal observation errors significantly underestimate the true error in our model.

We used two modelling schemes for the Bayesian inversion of the slip. The first mimics traditional inversion techniques which ignore the prediction error and only model the formal observational errors. This approach is equivalent to $\alpha = 0$ in eq. (25), so the covariance matrix in the likelihood function is $\mathbf{C}_\chi = \mathbf{C_d} = \sigma_d^2 \mathbf{I}_{N_{dp}}$. The second scheme models the prediction error as outlined in Section 4.5. The results from these two techniques are compared in Fig. 15. The inclusion of model prediction error does not significantly improve the accuracy of the mean of the posterior PDF. Once observational noise has corrupted the data or prediction errors bias our forward model, the model which best fits the observations is likely no longer the true source model, at least for overparametrized inverse problems. Thus, the mean or peak of our posterior PDF will not match the true source model. There is no way to recover the model resolution that has been lost through observational noise and a poor forward model, and estimating the prediction error does nothing to help this. Instead, estimating the model prediction error increases the effective errors in our forward model and thus our posterior

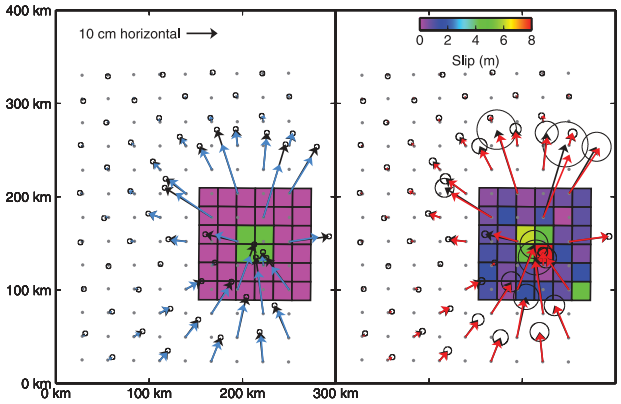

Figure 13. Prediction errors for synthetic static finite fault model. Left-hand panel: Synthetic surface displacements were calculated using the source model plotted in colour. The sense of fault motion is thrust and the fault geometry is the same as in Fig. 10. Vectors are horizontal displacements at selected observation points. The black vectors show our synthetic observations which were generated using a 1-D layered elastic structure and Gaussian zero-mean observation noise with a variance of 0.1 cm^2. 95 per cent confidence ellipses for the formal observational error are plotted as black circles. The blue vectors are the predicted surface displacements from propagating the synthetic source model through an imperfect elastic structure, specifically a homogeneous elastic half-space. The difference between these two sets of vectors is the total prediction error and is plotted in Fig. 14. Right-hand panel: The background slip model is the mean of the posterior PDF for the inversion including model prediction error (i.e. the same slip model that is plotted in red in Fig. 15), and the red vectors are the predicted surface displacements due to this slip model. The observations (black vectors) are unchanged, but the uncertainties on the observations are now the 95 per cent confidence ellipses based on the covariance matrix of the total prediction error: $\mathbf{C}_\chi = \mathbf{C_d} + \tilde{\alpha}^2 \mathrm{diag}(D_1^2, \ldots, D_{N_{dp}}^2)$, where $\mathbf{C_d} = 0.1\ \mathrm{cm}^2\mathbf{I}$ and $\tilde{\alpha} = 0.1250$ is the mean of the posterior distribution on α.

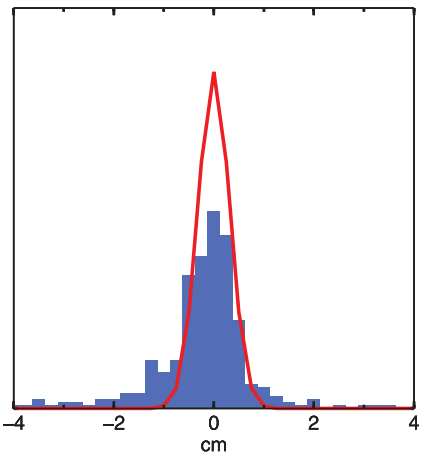

Figure 14. Comparison of histogram of total errors (data errors plus prediction errors; blue histogram) for the model in Fig. 13 with the formal observation error distribution (red line). The formal observational error is Gaussian with mean zero and variance $\sigma_d^2 = 0.1\ \mathrm{cm}^2$ for the true slip with correct Green's functions. The actual prediction error for the true slip with incorrect Green's functions is biased (i.e. the mean is not zero) and the total error histogram is significantly broader than the formal observational error. The mean and standard deviation of the total error for the true slip model are −0.2743 and 1.1451 cm, respectively.

PDF, helping to ensure that the true model lies within the posterior 95 per cent confidence ellipses 95 per cent of the time. In contrast, ignoring the prediction errors gives too much weight to fitting the data.

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

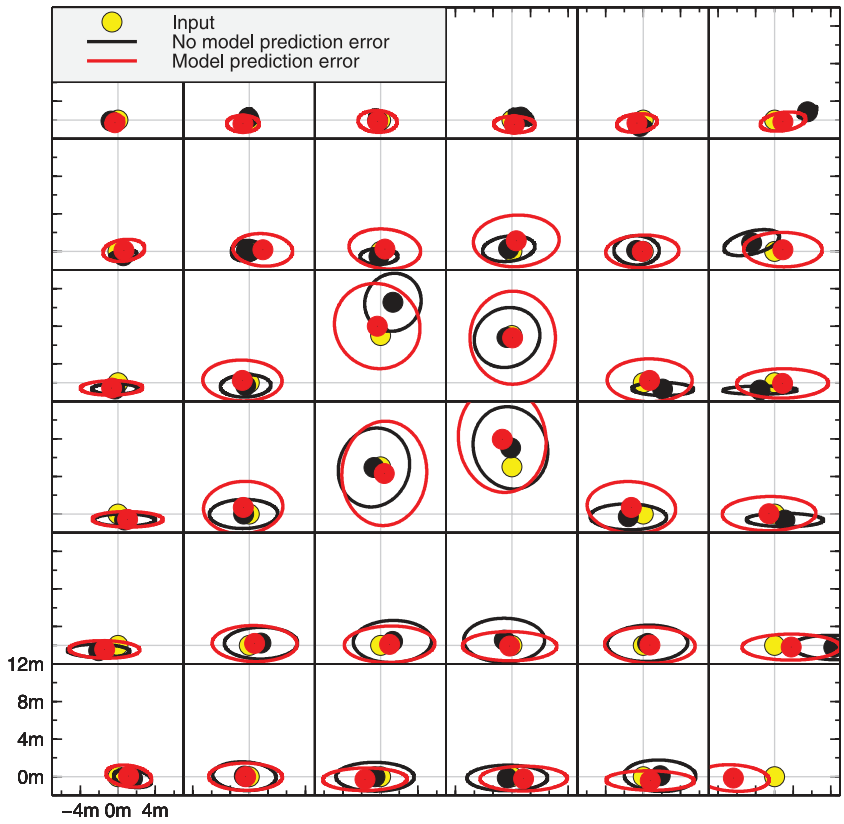

Figure 15. Effects of including model prediction errors for the 36-patch model in Fig. 11. The *x*-axis and *y*-axis represent the strike-slip and dip-slip components of displacement on each fault patch, respectively. Yellow circles indicate input slip used to generate the synthetic observations. The small circles and ellipses are the mean and 95 per cent confidence ellipses for posterior PDFs generated with (red) and without (black) including the model prediction error in the inversion.

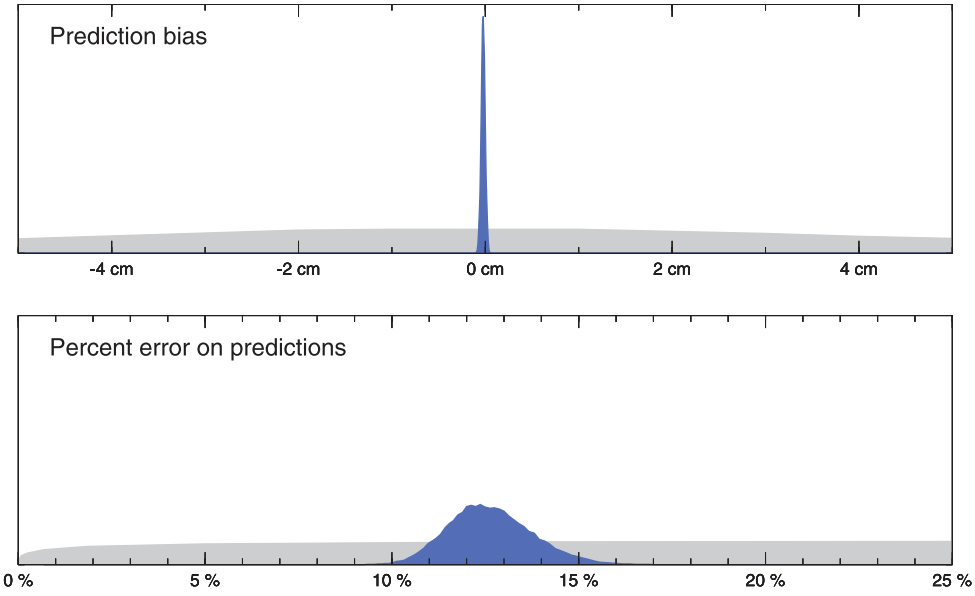

Figure 16. Posterior model prediction error for synthetic static finite fault model. The PDFs for the prior (grey) and posterior (blue) model prediction error distributions are shown. In this example, both μ_s and α (as defined in eqs (21) and (25)) were estimated as part of the sampling process. The prior on μ_s is a Gaussian, $\mathcal{N}(0\,\text{cm}, 5\,\text{cm})$, whereas α has a log-normal prior such that $\ln\alpha \sim \mathcal{N}(0, 5)$.

In Fig. 15, only the horizontal component of slip on the bottom right fault patch is definitely outside of our 95 per cent confidence bounds when the prediction error effect is included. Of course, for a 95 per cent confidence estimate, the true answer will lie outside of the calculated bounds 5 per cent of the time. Our fault model has thirty-six fault patches and two components of slip on each patch. Thus, misestimating one out of seventy-two inputs is only 1.4 per cent of our parameters. So, if anything, we are overestimating the size of the error in our model. In contrast, at least five components of slip are outside of the 95 per cent confidence bounds

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

for the run without model prediction error estimation, which is about 7 per cent of our model parameters and thus an under-estimate of the size of the error in our model. Explicitly including the model prediction error as part of the inversion process, allows us to produce posterior PDFs of the model prediction error (Fig. 16) and to update the estimated total uncertainties in the inverse problem (Fig. 13).

We also note that the quality of the entire model prediction error estimation process depends on the quality of the model used to parametrize the model prediction error. Here, we have adopted a design that attempts to approximate any errors in our elastic structure with a factor that scales with the amplitude of the data. We chose this design because a first-order effect of errors in the Green's functions is that, for a given Earth structure and fault model, the size of the model prediction error scales with the amplitude of slip (and the amplitude of the observations can be used as a proxy for the amplitude of the predicted data). This may not be an ideal approximation for model prediction errors due to errors in the elastic structure, especially since changes to the elastic structure are expected to produce highly correlated changes in the associated Green's functions; and, furthermore, there may be other significant sources of model prediction error other than those originating from the elastic structure model.

5.3 Synthetic kinematic models

We present a series of synthetic kinematic finite fault earthquake models using the fault geometry and distribution of geodetic data in Fig. 10 combined with synthetic seismic records from six receivers. All of the synthetic data are perfect and free of any noise, but in the inversion we assume observational errors for the static and kinematic data of $\mathbf{C}_\chi^s = 0.1\text{cm}^2\mathbf{I}_{\mathbf{N_s}}$ and $\mathbf{C}_\chi^k = 1\text{cm}^2\mathbf{I}_{\mathbf{N_k}}$, respectively. The prior PDFs on slip duration and rupture velocity are $T_r = \mathcal{U}(0\text{s}, 10\text{s})$ and $V_r = \mathcal{U}(0\,\text{km}\,\text{s}^{-1},\ 5\,\text{km}\,\text{s}^{-1})$, respectively. The assumed uncertainty for the hypocentre location is 10 km.

We use the cascading technique; so for each joint static-kinematic model, we first produce samples of the posterior static slip distribution using the static geodetic offsets as our observations, $\mathbf{D_s}$. The samples of the posterior static slip distribution are then used as samples of the prior slip distribution for the joint kinematic model. We present results for two fault parametrizations: one fault discretized into nine patches (Fig. 17) and one discretized into 36 patches (Fig. 18). Comparisons between the synthetic model and the mean of the posterior models are shown in Figs 17–20. The inversion does a good job of recovering the input source model and all of the posterior distributions are tightly peaked. The quality of the solution for the kinematic rupture parameters (slip duration and rupture velocity) seems slightly poorer than that for the slip parameters, although this might be due to the fact that the posterior PDFs on the slip distribution are constrained by synthetic GPS offsets in addition to synthetic seismograms. Also, it should be noted that the posterior PDFs on the slip duration are somewhat coarse because the slip durations are only evaluated in discrete increments of 1 s, the sampling rate of the kinematic data used in the inversion. The hypocentre location is well recovered (Fig. 19), especially considering the fact that our slip patches are 20 km wide and are populated with point sources spaced 5 km apart.

In Fig. 20, we compare the synthetic faulting parameters with the 1σ Bayesian credibility intervals from the posterior distribution. At 1σ confidence, we would expect about 68 per cent of the input model parameters to lie in the confidence bounds, or about 99 of the 146

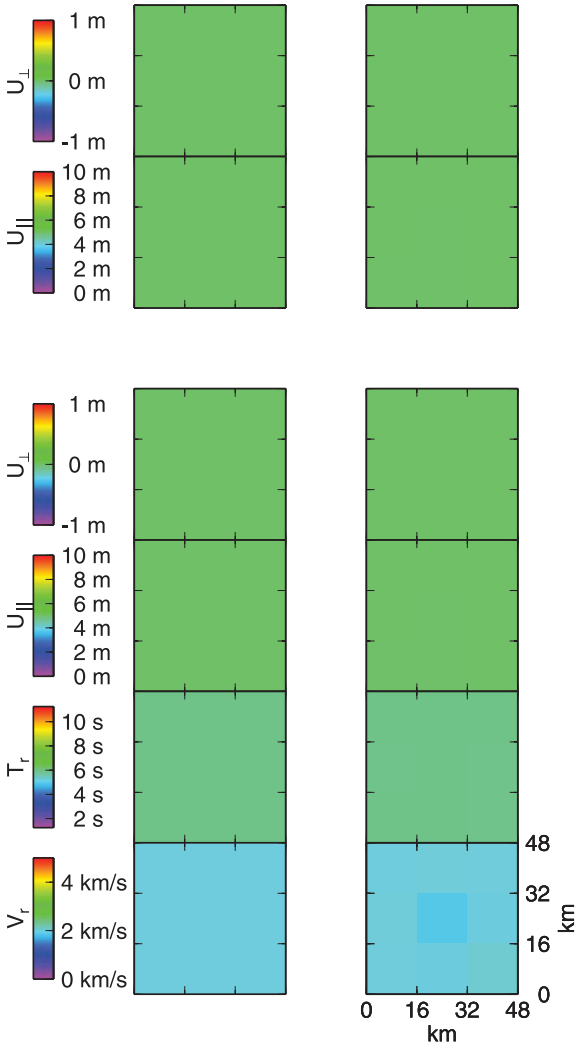

Figure 17. Sampling a synthetic kinematic finite fault model with nine fault patches. The left-hand column shows the faulting parameters used to create the synthetic surface displacements for each performance test. The right-hand column is the mean of the posterior samples for a CATMIP run with 500 000 Markov chains. For the initial static-only run, each chain is 10 steps long. The joint kinematic-static run used Markov chains 100 steps long. (Top panel) Static model. (Bottom panel) Kinematic model.

model parameters. For this sampling run, 114 parameters lie within their credibility interval. The kinematic forward model has four parameters per fault patch compared to two parameters per patch for the static forward model. Thus, the quality of the posterior after sampling could be compared to the 18-patch (36-parameter) and 72-patch (144-parameter) static solutions. However, these are not entirely fair comparisons. First, the static models were calculated by sampling the complete 36-parameter and 144-parameter spaces directly, whereas the kinematic solutions were produced through cascading from lower-dimension static models. Further, the suggested comparisons are not fair in terms of comparing the computational cost of the models given the considerably higher computational cost of the kinematic forward model.

The evolution of the faulting parameters for one patch during cascading is shown in Fig. 21. Because the kinematic and static data sets were generated from the same fault slip, the mean of the posterior PDF on slip is nearly the same for the static and kinematic

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

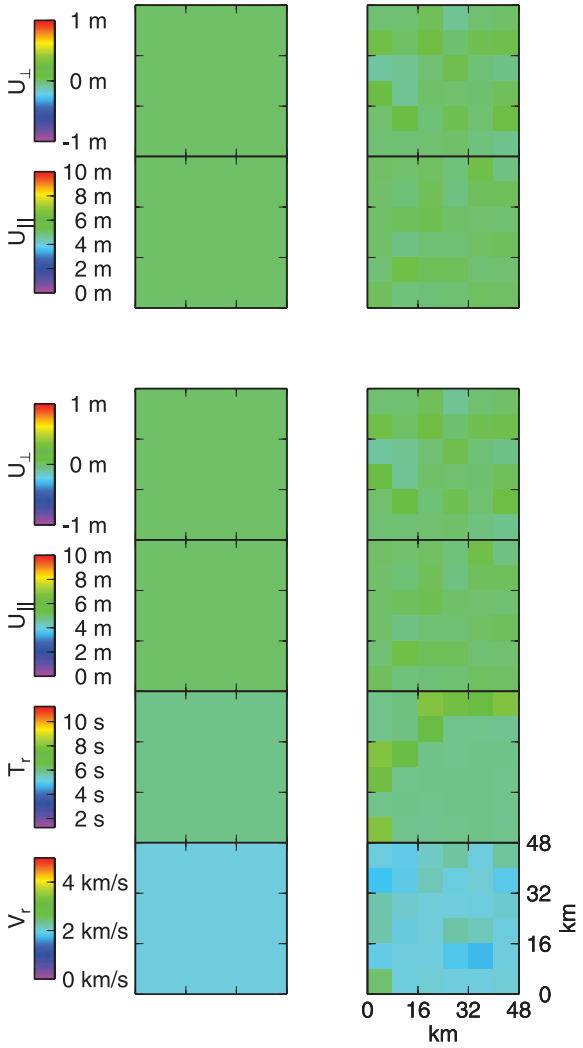

Figure 18. Sampling a synthetic kinematic finite fault model with 36 fault patches using a similar setup as in Fig. 17. The posterior distribution was simulated by a CATMIP run of 1000 Markov chains with each chain 40 960 steps long.

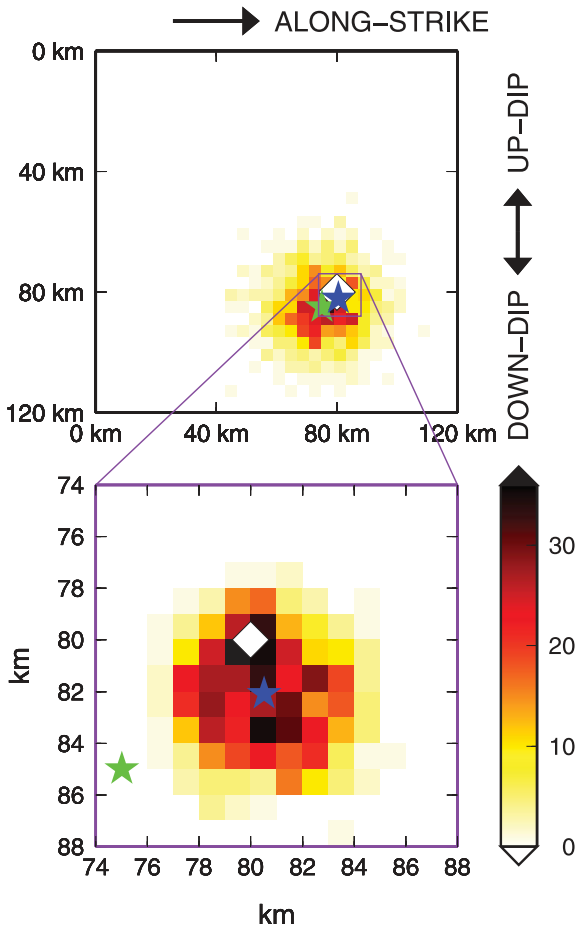

Figure 19. Histograms of prior and posterior PDFs for hypocentre location for the run in Fig. 18. (Top) 2-D histogram of prior PDF. The prior PDF is a Gaussian centred at the location denoted by the green star with a standard deviation of 10 km in both the along-strike and downdip directions. The mean of the posterior PDF is shown with the blue star. The true hypocentre location used to generate the synthetic data is 80 km along-strike and 80 km downdip, and is marked with a white diamond. Background colour is number of samples of the prior PDF in each 4-by-4 km region. (Bottom) 2-D histogram of posterior PDF. Background colour is number of samples of the posterior PDF in each 1-by-1 km region.

solutions. However, the uncertainty on the slip model decreases after the addition of the kinematic data. The reason behind this can be intuited by thinking about the familiar inverse optimization problem. In least-squares problems, the uncertainty on model parameters decreases with increasing numbers of observations. The additional information in the kinematic data and the sheer quantity of data points in the kinematic time-series act together to greatly increase the ostensible number of observations, making the posterior uncertainties small. This behaviour is further evidence that additional effort is required to quantify the correlation in our data. If these correlations are properly accounted for, the effective number of data points may be much smaller and the posterior uncertainties larger.

The patch-to-patch correlations are shown in Figs 22 and 23. Slip tends to be highly anticorrelated with its neighbours, showing that there are trade-offs between slip on adjoining patches and indicating that spatial resolution is poor. Slip duration and rupture velocity show less spatial correlation, although it should be noted that we are looking at a very narrow type of correlation (the Gaussian covariance between one parameter on one patch and that same parameter on other patches) and does not rule out the possibility that other correlations may be present in the posterior PDF. The posterior distributions for the faulting parameters are highly peaked in Fig. 21, indicating that they are well resolved with this network geometry and these data. However, this apparent high level of success is mostly due to the unrealistically perfect synthetic data and source model combined with a lack of prediction error.

6 EXPLORING THE ENSEMBLE OF POSTERIOR SAMPLES

An advantage of the sampling approach to modelling is that instead of producing one optimal model, our sampling yields an arbitrarily large ensemble of all plausible models based on the fit to the data and the *a priori* information relevant to the model. However, analysing this ensemble is non-trivial; simply visualizing a high-dimensional PDF is difficult. We can look at individual models, such as the mean of the posterior samples, the median of the

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

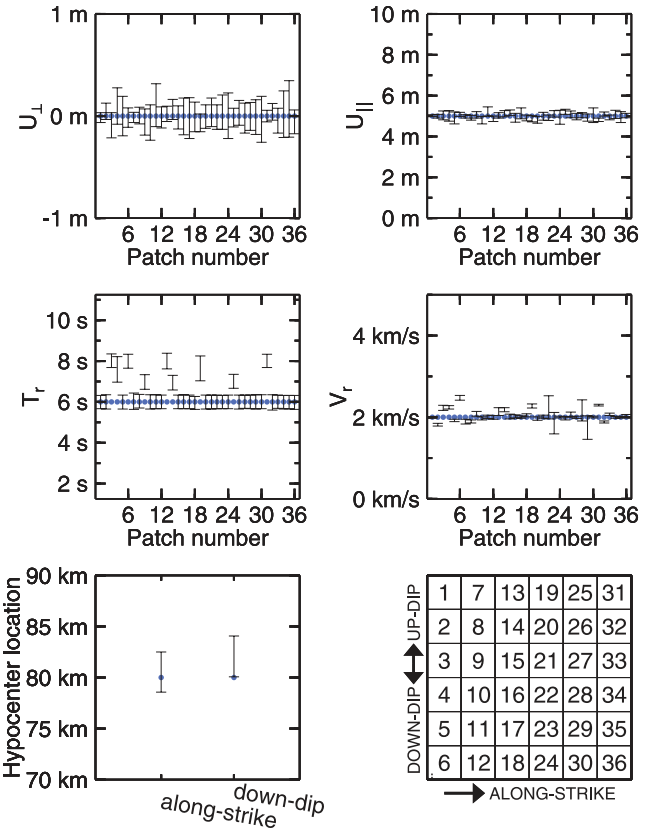

Figure 20. Bayesian credibility intervals for the run in Fig. 18. The values of the synthetic source model are shown with blue circles. Error bounds represent range of 68 per cent ($\pm 1\sigma$) credibility intervals. The two parameters defining the hypocentre location are the distance along-strike and distance downdip, respectively. A key to the location of each patch on the fault plane is provided in the bottom-right-hand panel.

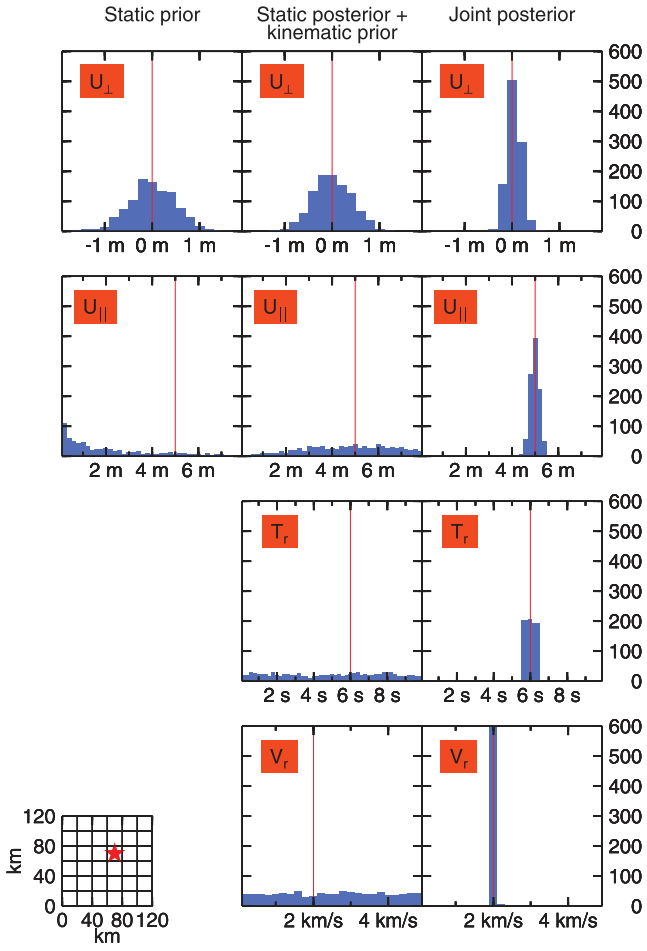

Figure 21. Evolution of faulting parameters for the patch indicated with a star from the 36-parameter kinematic model. Red lines indicate the values of the synthetic source model. Blue histograms represent the distribution of samples of each PDF. The left-hand column is the prior distribution for the static solution. The middle column is the prior distribution for the joint kinematic solution that uses the posterior static slip distribution as an *a priori* constraint. The right-hand column is the final joint posterior kinematic solution. The rows from top to bottom are $U_\perp$, $U_{||}$, T_r and V_r.

posterior samples, the maximum *a posteriori* (MAP) which is the model that maximizes the posterior probability, or the maximum likelihood estimate (MLE) which is the model that maximizes the data likelihood. However, these models can be potentially unrepresentative of the PDFs from which they are derived. The mean and median may be very misleading for long-tailed asymmetric PDFs, and the MAP and MLE can differ significantly depending on how informative the prior distribution is. Even for simple PDFs for which the mean, median, MAP and/or MLE represent well-behaved statistical quantities, any one model may contain features which are not well constrained and thus should not necessarily be considered 'real'.

The mean of the posterior samples was plotted for the examples in Fig. 11 in Section 5.1. In that section, we discussed how the mean model solution changes as the size of the fault patches decreases and the trade-off between the model parameters increases due to a loss of model resolution. However, there is very little information in Fig. 11 itself to illustrate how this process occurs or even that it is happening. In this section, we explore more advanced methods of analysing the posterior PDF using the examples from Fig. 11.

To understand which features of the solution are well-constrained and which are uncertain, the posterior standard deviation or variance may be computed for each model parameter. However, there are two main limitations to this analysis. First, the standard deviation or variance of each model parameter is only a good metric for describing the posterior PDF if the PDF is approximately Gaussian. Unless you are able to look at the posterior distribution itself, which is quite difficult in high dimensions, you have no way of knowing what the variance of the PDF implies. Secondly, the posterior PDFs must be analysed in the context of their respective prior PDFs because the posterior PDF only has meaning given a specific prior PDF. Consider the following. If the data are completely uninformative, then $p(\mathbf{D}|\boldsymbol{\theta}) = \text{constant}\ \forall\ \boldsymbol{\theta}$ and $p(\boldsymbol{\theta}|\mathbf{D}) \propto p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})$. Thus if the data are nearly uninformative, then the posterior PDF will nearly be the same as the prior PDF. The posterior distribution on a model parameter can be understood to be well-constrained by the observed data not if its posterior PDF is highly peaked (and thus has small variance) but rather if its posterior PDF is substantially different from its prior PDF.

The posterior PDF on each parameter can be plotted by computing histograms of the posterior samples, that is, the samples from the final cooling step (Figs 24 and 25). Most of the posterior PDFs are Gaussian-like, but the posterior PDFs on $U_{||}$ are truncated on patches with zero slip because our prior on $U_{||}$ is a uniform distribution which does not allow back-slip in excess of 1 m. Note that the posterior PDFs in the centre four squares in Fig. 25 (corresponding

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

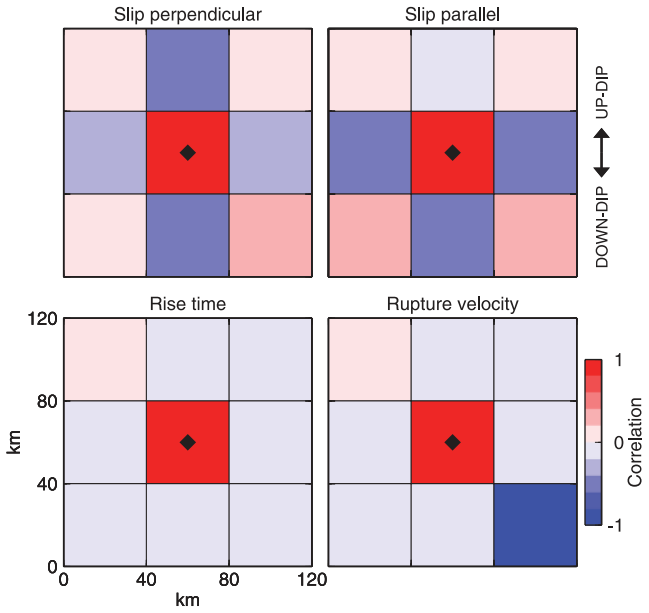

Figure 22. Posterior correlation between the patch marked with a diamond and the other patches for each faulting parameter. These correlations are calculated for the nine-patch kinematic model.

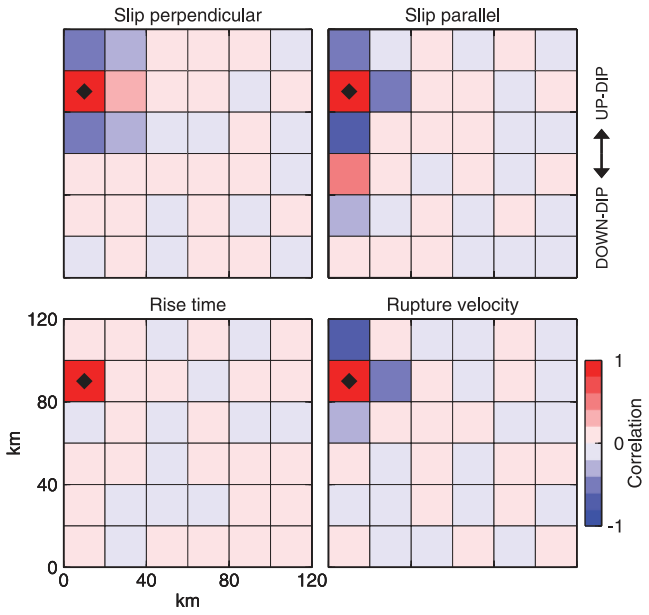

Figure 23. Posterior correlation between the patch marked with a diamond and the other patches for each faulting parameter. These correlations are calculated for the 36-patch kinematic model.

to patches with high slip) are relatively broad but very different from the prior PDF. This indicates that the data are informative and substantially change the posterior distribution relative to the prior assumptions. However, the broadness of the posterior PDFs indicate that there must be significant uncertainty on the value of slip on each patch likely due to trade-offs between slip on neighbouring patches. This is also revealed by the posterior variances in Fig. 26.

To understand how the possible values for each model parameter trade-off with each other, the complete posterior model covariance or correlation must be considered (Fig. 27). In general, the PDF on any model parameter may be asymmetric or multiply peaked,

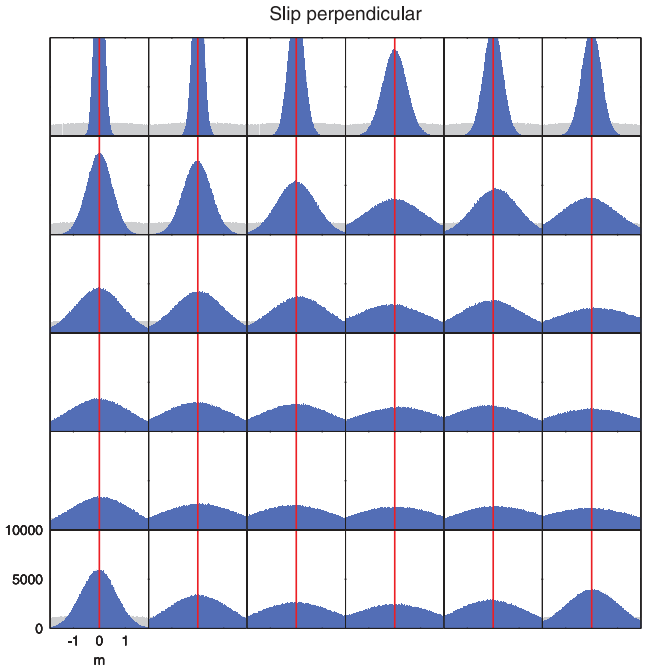

Figure 24. Histograms of the rake-perpendicular component of slip on each patch for the prior (grey) and posterior (blue) PDFs from the 72-parameter model in Fig. 11. Red lines mark the slip values from the actual source model.

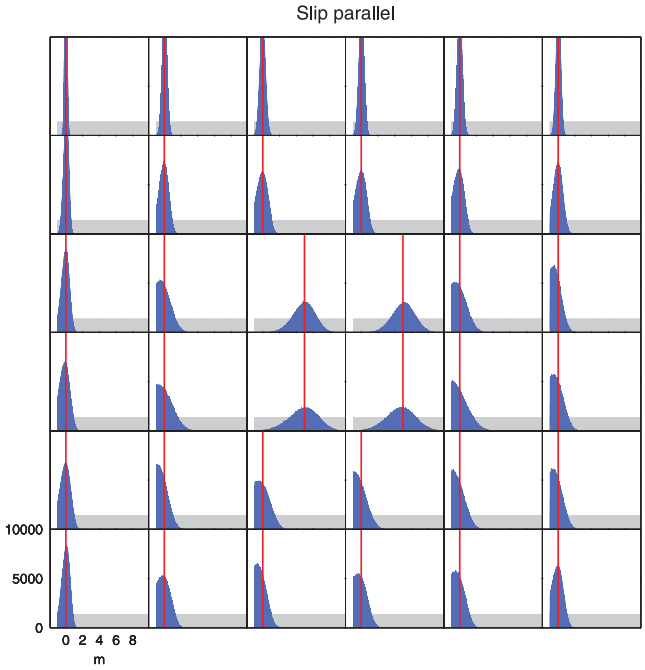

Figure 25. Same as Fig. 24 for the rake-parallel component of slip.

and the way that any two or more model parameters trade-off with each other may not be well described by a Gaussian covariance matrix. (This is equivalent to saying that the joint probability distribution for any two or more model parameters is not a multivariate Gaussian distribution.) Even if individual PDFs are sufficiently Gaussian that the scalar covariance between two parameters is sufficient to characterize the relationship between them, then, for a

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

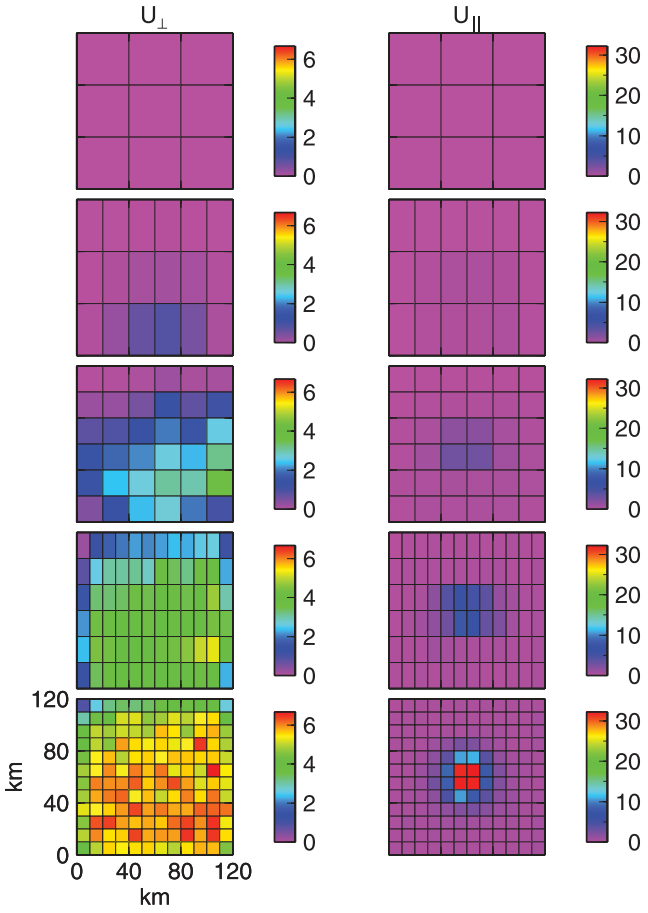

Figure 26. Posterior variances (in m^2) for each component of slip for all models in Fig. 11. The variances are representative of the width of the posterior PDFs in Figs 24 and 25. Note that the $U_{\|}$ variances go to zero for the patches with zero slip due to the positivity constraint eliminating trade-offs between fault patches.

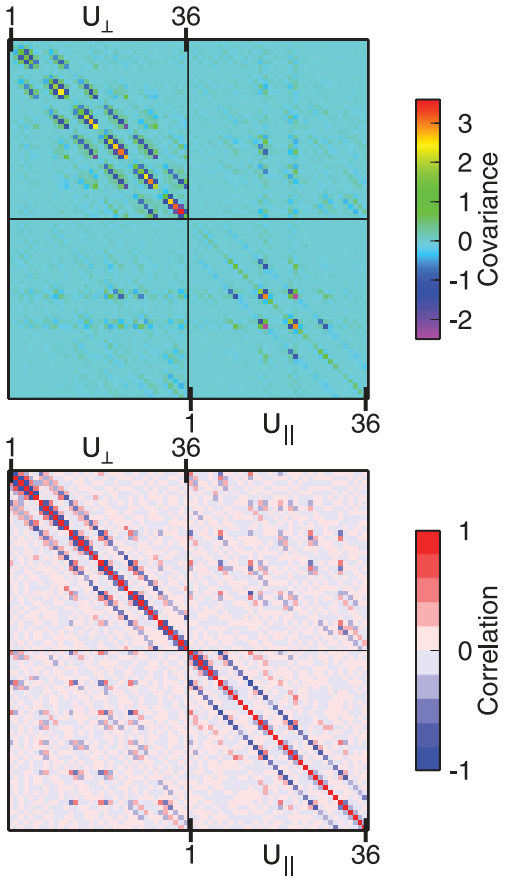

Figure 27. Covariance and correlation coefficients (top and bottom panels, respectively) for the 36-patch/72-parameter model in Fig. 11. The cells of the matrices represent the slips on the individual patches in order according to the numbers in Fig. 20. The rake-perpendicular components of slip on each of the 36 patches are plotted before any of the rake-parallel components of slip are shown. Covariance has units of m^2, whereas correlation is dimensionless.

problem with k model parameters, each model $\boldsymbol{\theta}$ is a vector of k parameters and the model covariance is a k-by-k symmetric matrix. It is very difficult to make visual sense out of a high-dimensional covariance matrix. Consider Fig. 27 which plots the covariance and correlation coefficient matrices for a static finite fault slip model with just 36 fault patches. We solve for two components of slip on each patch, resulting in 72 free parameters. It is possible to discern that individual $U_{\perp}$ slips have greater trade-off with other $U_{\perp}$ slips (upper-left quadrant of the matrix) and the $U_{\|}$ components of slip have greater trade-offs with other $U_{\|}$ components of slip (lower-right quadrant) than any $U_{\perp}$ slip has with the $U_{\|}$ component on the same fault patch (lower-left and upper-right quadrants). But is nearly impossible to intuit any spatial relationships between the correlations when the fault geometry is unwound into matrix form.

To explore the spatial covariances in the posterior PDF, individual pairs of covariances can be plotted on the fault plane as in Figs 22 and 23. But each plot of this type can only show the correlations with all other parameters for one model parameter on one fault patch, requiring hundreds or thousands of such figures to explore the solution for real-world-sized finite fault models. Furthermore, since such figures only plot the parameter-to-parameter correlation between one parameter and each of the others, this method may not capture the full complex ways in which the selected parameter may interact with the other model parameters.

In addition to considering the relationships between individuals pairs of parameters, we should also consider how groups of parameters behave. Such analyses could include exploring the spatial resolution of the slip model by calculating how many adjacent fault patches would need to be averaged together to achieve a minimum acceptable level of resolution. We could also consider the distribution of spatial roughness of slip derived from our distribution of slip models by calculating the norm of the Laplacian of each slip model. Finally, we could explore how each model parameter is correlated to all other model parameters (something akin to the model's total covariance) or, equivalently, evaluate to what extent each parameter is independent of the others. The natural method for this analysis is to compute the mutual information between each model parameter and the rest of the joint posterior PDF, $I(\theta_i; \boldsymbol{\theta}_{-i})$. (Background on mutual information and relative entropy is given in Appendix C.)

We calculate the mutual information between one model parameter, θ_i, and all other parameters by comparing the entropy of that parameter, $h(\theta_i)$, with the entropy of all other parameters, $h(\boldsymbol{\theta}_{-i})$ (see Appendix C). The entropies $h(\theta_i)$ and $h(\boldsymbol{\theta}_{-i})$ as defined in eq. (C4) are plotted in Fig. 28. From this figure we can draw the following conclusions. First, the entropy of any particular slip parameter, θ_i,

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

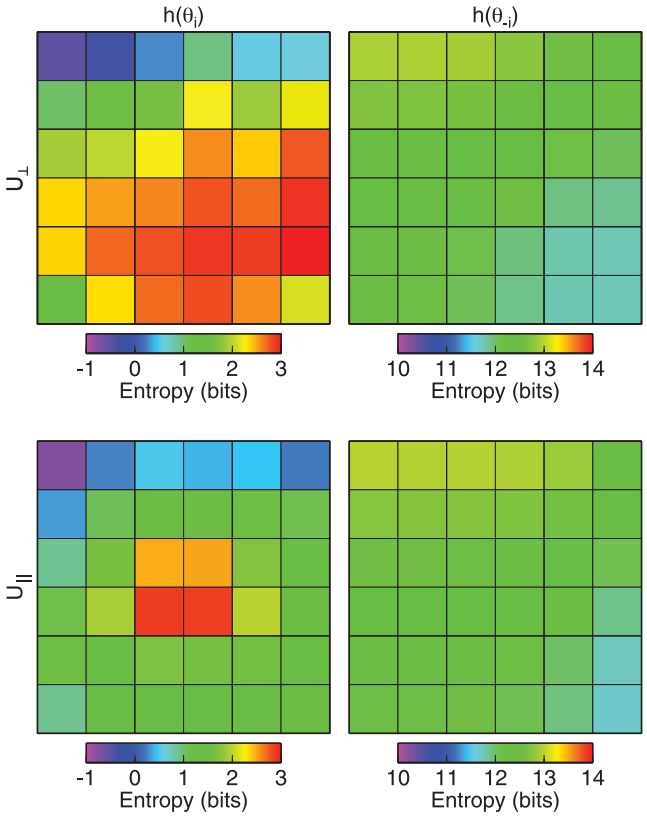

Figure 28. Entropy (in bits) of posterior PDFs on each model parameter for the model in Fig. 27. $h(\theta_i)$ denotes the entropy of a model parameter, $h(\boldsymbol{\theta}_{-\mathbf{i}})$ denotes the entropy for the joint PDF of all model parameters other than θ_i. All colour scales have the same dynamic range. The entropy of the multivariate Gaussian approximation to the posterior PDF $h(\boldsymbol{\theta}|\mathbf{D}) = h(\theta_i, \boldsymbol{\theta}_{-\mathbf{i}}) = 9.965$ bits.

is small compared to the entropy of all of the other variables, $h(\boldsymbol{\theta}_{-\mathbf{i}})$. Secondly, the variation in differential entropy $h(\boldsymbol{\theta}_{-\mathbf{i}})$ as a function of which slip parameter is excluded is small compared to the variability in entropy of the individual slip parameters, $h(\theta_i)$. (Note that $h(\theta_i, \boldsymbol{\theta}_{-\mathbf{i}})$, the entropy of the full posterior PDF, is a constant.)

Finally, the mutual information between each parameter θ_i and all the others is plotted in Fig. 29. We see based on Fig. 28 that the dynamic range in the mutual information for any fault model is due to the contribution of $h(\theta_i)$ while the average value of the mutual information is controlled by $h(\boldsymbol{\theta}_{-\mathbf{i}})$. The mutual information increases as the number of fault patches increase and their size decreases. This growth is especially evident in the mutual information of $U_{\|}$ (Fig. 29). Increasing mutual information means that the individual parameters are less independent of each other and more highly correlated (or anticorrelated). At last, we have arrived at visual proof for why the mean of the posterior distribution in Fig. 11 becomes a blurry version of the synthetic slip model as the size of the fault patches decreases.

Another feature apparent in Fig. 29 is that the mutual information is small for patches with small amounts of slip regardless of the model resolution for that inversion. This is a result of the back-slip constraint we use. If there are two patches whose average slip is large and for which we cannot resolve the slip on each individual patch, there is plenty of 'room' for checkerboard mode uncertainties without violating the positivity constraint. But this is not possible when the average slip is small. Another way of saying this is that, since the variance of the posterior PDF for a patch without slip will be much smaller (due to the back-slip constraint), its covariance with any other patch will also be smaller, and thus so will its mutual information. Thus, low mutual information alone is not enough to identify a particular faulting parameter as being well constrained by the inversion. Once again, the posterior PDF can only be understood in the context of the prior PDF.

7 CONCLUSIONS

We have developed a new framework for Bayesian inversion of finite fault earthquake models that allows imaging of the complete model parameter space for this inherently underdetermined inverse problem without applying any non-physics-based *a priori* constraints on the form of the solution. To make these calculations computationally tractable, we have developed and tested a new sampling algorithm, CATMIP, which is more efficient than comparable existing samplers and can be run in a parallel computing environment. Because of these advances, we can tackle problems with models as large as finite fault parametrizations found in studies using conventional optimization techniques.

It is straightforward to explore any scalar physical quantity or probability derived from the posterior PDF. For example, the full ensemble of model samples can be used to formulate probablistic statements, such as calculating the 95 per cent confidence bounds for a given model parameter. Similarly, the distribution of possible scalar seismic moments for an earthquake can be calculated from the ensemble of all slip models, and the result plotted as a single 1-D PDF.

We have applied our Bayesian methodology to several synthetic finite fault models. For overdetermined inverse problems, the algorithm produces tightly peaked posterior distributions. For underdetermined or poorly resolved models, the posterior PDFs become broad and the slip on neighbouring patches becomes anticorrelated as different patches trade-off with each other. Traditional optimization methods suppress this behaviour through smoothing. This smoothing is not necessary. All possible combinations of trade-offs give plausible models and simply act to broaden the posterior PDF.

In contrast to traditional optimization problems, the methods described here are computationally expensive. However, the reward is a complete characterization of the model. In the process, the Bayesian approach requires a complete evaluation of the trade-offs and covariances of the model parametrization, the observational errors and the model prediction errors, yielding a fuller and richer understanding of the model and data under consideration. In the context of traditional optimization, smoothing acts to reduce understanding nearly as much as it limits model complexity. In the Bayesian context, greater model complexity (through removing model regularization and including the parameters of the error structure in the inversion) creates greater understanding of the strengths, limitations and uncertainties of the inversion process.

ACKNOWLEDGEMENTS

The authors would like to thank Michael Aivazis for helpful discussions. This work is supported by the National Science Foundation through grant number EAR-0941374 and is Caltech Seismological Laboratory contribution 10086.

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

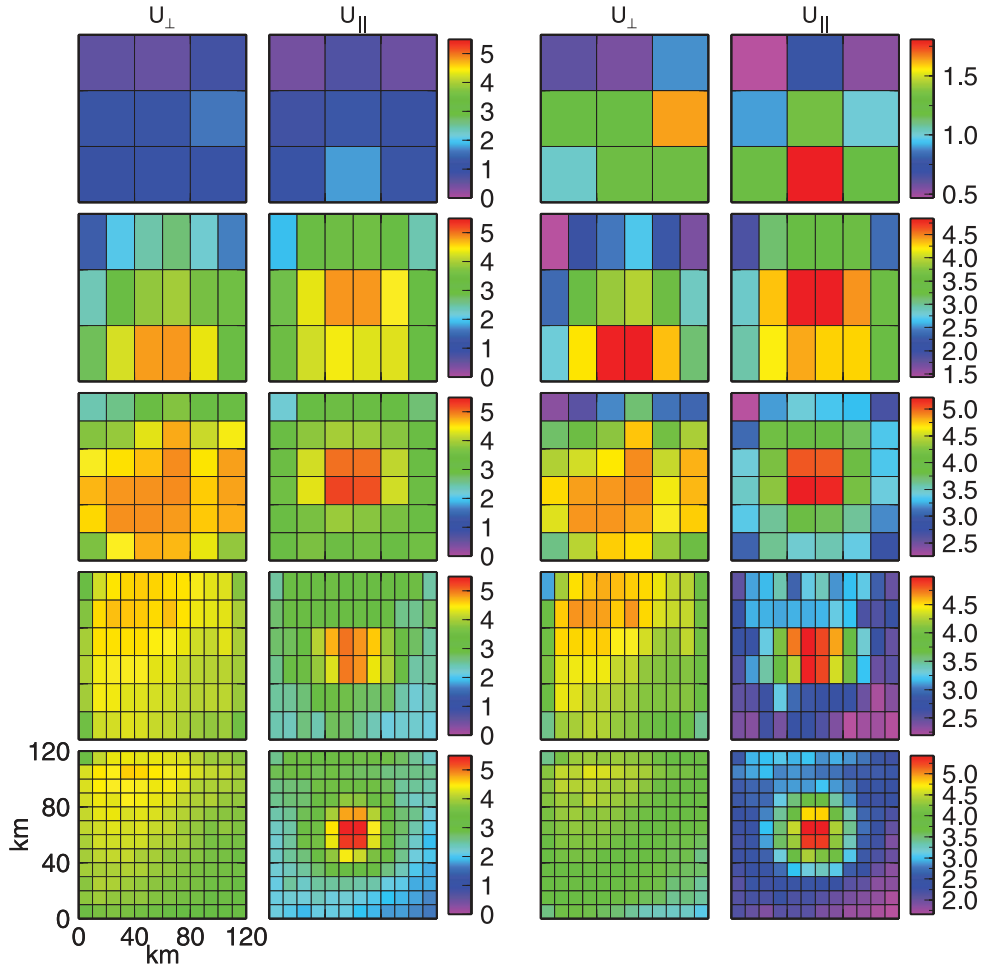

Figure 29. Mutual information of posterior PDFs on each $U_{\perp}$ and $U_{\parallel}$ for each model in Fig. 11. The two left- and two right-hand columns are the same except that the plots in the left-hand columns are shown using the same colour scale as each other, whereas the colour scale in the right-hand columns is rescaled for each fault model. Mutual information is measured in bits.

REFERENCES

Arnadottir, T. & Segall, P., 1994. The 1989 Loma Prieta earthquake imaged from inversion of geodetic data, *J. geophys. Res.*, **99**(B11), 21 835–21 855.

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

Beck, J., 2010. Bayesian system identification based on probability logic, *Struct. Contrl. Health Monit.*, **17,** 825–847.

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

Beck, J. & Katafygiotis, L., 1998. Updating models and their uncertainties. Part I: Bayesian statistical framework, *J. Eng. Mech.*, **124**(4), 455–461.

Beck, J. & Zuev, K., 2013. Asymptotically independent Markov sampling: a new MCMC scheme for Bayesian inference, *Int. J. Uncertain. Quantification,* **3**(5), 445–474.

Bellman, R., 1957. *Dynamic Programming,* Princeton University Press, Princeton, NJ, US.

Cerny, V., 1985. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm, *J. Optim. Theory Appl*. **45**(1), 41–51.

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.

Cohee, B. & Beroza, G., 1994a. A comparison of two methods for earthquake source inversion using strong motion seismograms, *Ann. Geophys.*, **37**(6), 1515–1538.

Cohee, B. & Beroza, G., 1994b. Slip distribution of the 1992 Landers earthquake and its implications for earthquake source mechanics, *Bull. seism. Soc. Am.*, **84**(3), 692–712.

Cotton, F. & Campillo, M., 1995. Frequency domain inversion of strong motions: application to the 1992 Landers earthquake, *J. geophys. Res.*, **100**(B3), 3961–3975.

Cover, T. & Thomas, J., 2006. *Elements of Information Theory*, John Wiley and Sons, Hoboken, NJ, US.

Du, Y., Aydin, A. & Segall, P., 1992. Comparison of various inversion techniques as applied to the determination of a geophysical deformation model for the 1983 Borah Peak earthquake, *Bull. seism. Soc. Am.*, **82**(4), 1840–1866.

Evans, E. & Meade, B., 2012. Geodetic imaging of coseismic slip and postseismic afterslip: sparsity promoting methods applied to the great Tohoku earthquake, *Geophys. Res. Lett.*, **39**(L11314), doi:10.1029/2012GL051990.

Fienberg, S., 2006. When did Bayesian inference become Bayesian?, *Bayesian Anal.*, **1**(1), 1–40.

Fisher, R., 1921. On the 'probable error' of a coefficient of correlation deduced from a small sample, *Metron*, **1,** 3–32.

Fukuda, J. & Johnson, K., 2008. A fully Bayesian inversion for spatial distribution of fault slip with objective smoothing, *Bull. seism. Soc. Am.*, **98**(3), 1128–1146.

Gelman, A., Carlin, J., Stern, H. & Rubin, D., 2004. *Bayesian Data Analysis*, Chapman and Hall, Boca Raton, FL, US.

Hernandez, B., Cotton, F. & Campillo, M., 1999. Contribution of radar interferometry to a two-step inversion of the kinematic process of the 1992 Landers earthquake, *J. geophys. Res.*, **104**(B6), 13 083–13 099.

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

Jaynes, E., 2003. *Probability Theory: The Logic of Science*, Cambridge University Press, UK.
Jeffreys, H., 1931. *Scientific Inference*, Cambridge University Press, UK.
Jeffreys, H., 1939. *Theory of Probability*, Cambridge University Press, UK.
Ji, C., Wald, D. & Helmberger, D., 2002. Source description of the 1999 Hector Mine, California Earthquake, Part I: wavelet domain inversion theory and resolution analysis, *Bull. seism. Soc. Am.*, **92**(4), 1192–1207.
Kikuchi, M. & Kanamori, H., 1982. Inversion of complex body waves, *Bull. seism. Soc. Am.*, **72**(2), 491–506.
Kirkpatrick, S., Gelatt, C. & Vecchi, M., 1983. Optimization by simulated annealing, *Science*, **220**(4598), 671–680.
Laplace, P., 1812. *Théorie Analytique des probabilités*, Veuve Courcier, Paris.
Malinverno, A., 2002. Parsimonious Bayesian Markov Chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.*, **151**(3), 675–688.
Marinari, E. & Parisi, G., 1992. Simulated tempering: a new Monte Carlo scheme, *Europhys. Lett.*, **19,** 451–458.
Menke, W., 2012. *Geophysical Data Analysis: Discrete Inverse Theory*, Academic Press, Waltham, MA.
Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E., 1953. Equation of state calculations by fast computing machines, *J. Chem. Phys.*, **21**(6), 1087–1092.
Monelli, D. & Mai, P., 2008. Bayesian inference of kinematic earthquake rupture parameters through fitting of strong motion data, *Geophys. J. Int.*, **173**(1), 220–232.
Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.*, **100**(B7), 12–431.
Muto, M. & Beck, J., 2008. Bayesian updating and model class selection for hysteretic structural models using stochastic simulation, *J. Vib. Control*, **14**(1-2), 7–34.
Rothman, D., 1985. Nonlinear inversion, statistical mechanics, and residual statics estimation, *Geophysics*, **50**(12), 2784–2796.
Rouy, S. & Tourin, A., 1992. A viscosity solutions approach to shape-from-shading, *SIAM J. Numer. Anal.*, **29**(3), 867–884.
Sambridge, M., 1999. Geophysical inversion with a neighbourhood algorithm-I. Searching a parameter space, *Geophys. J. Int.*, **138**(2), 479–494.
Sambridge, M. & Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems, *Rev. Geophys.*, **40**(3), 3-1–3-29.
Shannon, C., 1948. A mathematical theory of communication, *Bell Syst. Tech. J.*, **27**(3), 379–423.
Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, PA, US.
Wald, D. & Heaton, T., 1994. Spatial and temporal tistribution of tlip for the 1992 Landers, California, Earthquake, *Bull. seism. Soc. Am.*, **84**(3), 668–691.
Yagi, Y. & Fukahata, Y., 2011. Introduction of uncertainty of Green's function into waveform inversion for seismic source processes, *Geophys. J. Int.*, **186**(2), 711–720.
Zhao, H., 2005. A fast sweeping method for eikonal equations, *Math. Comput.*, **74**(250), 603–628.

APPENDIX A: REGULARIZED LEAST-SQUARES VERSUS FULL BAYESIAN INVERSIONS

If we substitute eq. (4) with $\mathbf{d} = \mathbf{D}$, $\mathbf{C}_\chi = \mathbf{C}$, and $\boldsymbol{\mu} = \mathbf{0}$ into the negative of the logarithm of eq. (1), we get,

$$-\ln \mathrm{p}(\boldsymbol{\theta}|\mathbf{D}) = \frac{1}{2}\ln|\mathbf{C}| + \frac{1}{2}\|\mathbf{D}-\mathbf{G}(\boldsymbol{\theta})\|^2_{\mathbf{C}^{-1}} - \ln \mathrm{p}(\boldsymbol{\theta}) + \text{constant}, \tag{A1}$$

where $\|\cdot\|^2_{\mathbf{C}^{-1}}$ denotes the squared weighted norm implied by the exponent in eq. (4). If the covariance matrix $\mathbf{C}$ is fixed rather than parametrized by uncertain parameters, then minimizing the objective function,

$$J(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{D}-\mathbf{G}(\boldsymbol{\theta})\|^2_{\mathbf{C}^{-1}} - \ln \mathrm{p}(\boldsymbol{\theta}), \tag{A2}$$

is equivalent to maximizing the posterior distribution, $\mathrm{p}(\boldsymbol{\theta}|\mathbf{D})$. In Bayesian statistics, the optimum value, $\hat{\boldsymbol{\theta}}$, is called the maximum *a posteriori* (MAP) estimate of $\boldsymbol{\theta}$; it is simply the most probable value based on the data, $\mathbf{D}$. The regularization in $J(\boldsymbol{\theta})$ is controlled by the prior distribution, $\mathrm{p}(\boldsymbol{\theta})$; for example, if a Gaussian prior $\mathcal{N}(\boldsymbol{\theta}|0, \sigma_0^2\mathbf{I_k})$ is chosen, then,

$$J(\boldsymbol{\theta}) = \|\mathbf{D}-\mathbf{G}(\boldsymbol{\theta})\|^2_{\mathbf{C}^{-1}} + \lambda\,\|\boldsymbol{\theta}\|^2_2\,, \tag{A3}$$

where constants and a common factor of $\frac{1}{2}$ have been dropped since they do not affect the optimal value $\hat{\boldsymbol{\theta}}$. This regularization favours smaller parameter values in the L2-norm sense and the regularization parameter $\lambda = \frac{1}{\sigma_0^2}$ is controlled by the choice of prior variance. Conversely, since regularized least-squares estimation uses an objective function of the same form as eq. (A2), it follows that it is equivalent to Bayesian MAP estimation.

In contrast to the MAP point estimate of $\boldsymbol{\theta}$, a full Bayesian inversion characterizes the whole posterior distribution, $\mathrm{p}(\boldsymbol{\theta}|\mathbf{D})$, and not just a dominant peak. Usually, this cannot be done analytically but it can be accomplished by MCMC sampling that produces samples of $\boldsymbol{\theta}$ that populate the parameter space in a probabilistically appropriate way; that is, the samples are distributed so that the number of them in each region of the parameter space reflects the probability assigned to that region by $\mathrm{p}(\boldsymbol{\theta}|\mathbf{D})$. The MCMC samples can be examined to see if the MAP estimate corresponds to a tightly confined peak of the posterior distribution, or there are multiple such sharp peaks, or one (or a few) broad peaks, implying that there are many models ($\boldsymbol{\theta}$ values) that are almost as probable as the MAP one. Full Bayesian analyses for inversions are the focus of this paper.

APPENDIX B: INTRODUCTION TO THE METROPOLIS ALGORITHM

The Metropolis algorithm uses a Markov chain to simulate draws from an unnormalized target distribution, π, using samples from a chosen probability distribution $q(y|x)$ (termed the 'proposal PDF'), where x is the current sample and y is our proposed new sample. In CATMIP, at the mth cooling step, π is the intermediate PDF in eq. (9), $f(\boldsymbol{\theta}|\mathbf{D}, \beta_m)$. Metropolis *et al.* (1953) uses a proposal PDF of the form $q(y|x) = q'(y-x)$ or, equivalently, $y = x + z$, where $z \sim q'(z)$. (Popularly, the proposal PDF is chosen to be $q' = \mathcal{N}(0, \Sigma)$, so each new candidate sample is the current sample plus some Gaussian perturbation with covariance Σ.) Note that the proposed sample y is produced from a PDF that only depends on a random variable, z, and the current position of the random walker, x. Thus, the Metropolis algorithm describes a random walk through model space that is independent of the history of the walker. This is what makes Metropolis sampling a Markov process and thus an example of MCMC.

Starting with an arbitrary initial sample x_0, the Metropolis algorithm produces N samples of the target distribution by the following:

For $i = 1, 2, \ldots, N$

(i) Draw $z \sim q'$ and compute a candidate sample $y = x_{i-1} + z$.

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

(ii) Generate a sample u from $\mathcal{U}(0, 1)$, the uniform distribution on (0,1).
(iii) Compute $r(x_{i-1}, y) = \min\{\frac{\pi(y)}{\pi(x_{i-1})}, 1\}$.
(iv) If $u \leq r$, $x_i = y$. Otherwise $x_i = x_{i-1}$.

In more conceptual terms, the Metropolis algorithm uses a random walk through model space to produce samples of any PDF we want to simulate. For multidimensional spaces, there are only a few PDFs (e.g. uniform distributions and Gaussian distributions) for which random samples can be directly generated. MCMC sampling is necessary because the PDFs being simulated are arbitrary and not explicitly normalized, and thus samples from the PDF cannot be produced directly. Instead, the Metropolis algorithm draws samples directly from a proposal PDF and then probabilistically chooses whether to keep or eliminate the candidate sample based on the ratio of its probability in the target PDF to the previous sample's probability in the target PDF. Even candidate samples for which this ratio is very low have some chance of being accepted. This ensures that the random walker will not become permanently trapped in a local maximum of the target PDF but instead will eventually escape and visit all parts of the target PDF.

The proposal PDF is critically important for controlling the efficiency of the sampler. The more similar the proposal PDF is to the target PDF, the more efficient is this methodology. In the limit that the proposal PDF equals the target PDF, then we are actually directly sampling from the target PDF.

APPENDIX C: OVERVIEW OF ENTROPY, RELATIVE ENTROPY AND MUTUAL INFORMATION

Information entropy is a measure of the amount of missing information about a variable whose value is uncertain (Shannon 1948). It can be thought of as the average number of yes-or-no questions required to determine the value of the variable (Cover & Thomas 2006). Thus, the typical unit of entropy is bits of information.

For a continuous random variable X with PDF $f(x)$, its *differential entropy* $h(X)$ is defined as,

$$h(X) = -\int f(x) \log f(x)\, \mathrm{d}x. \tag{C1}$$

Note that this definition allows for negative entropy. We can also define the *relative entropy* between two PDFs as,

$$D_{\mathrm{KL}}(f||g) = \int f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x. \tag{C2}$$

The relative entropy is also known as the Kullback–Leibler divergence, Kullback–Leibler distance, Kullback–Leibler information criterion (KLIC) or information gain. D_{KL} is a measure of the differences between two PDFs, although technically it is not a true distance metric because it does not satisfy all of the mathematical requirements to be a distance metric; for example, $D_{\mathrm{KL}}(f||g) \neq D_{\mathrm{KL}}(g||f)$. Alternatively, D_{KL} can be viewed as the inefficiency of assuming that a PDF is g when it is really f (Cover & Thomas 2006). Note that D_{KL} is non-negative and $D_{\mathrm{KL}} = 0$ if and only if $f = g$, that is, the relative entropy between two PDFs is zero if and only if the two PDFs are equal.

If we compute the relative entropy between the joint PDF, $f(x, y)$, of random variables X and Y, and the product of their two marginal PDFs, $f(x)$ and $f(y)$, then we have quantified the difference between their joint PDF and what the joint PDF would be if the two variables were independent. This relative entropy is called the *mutual information* of X and Y. Consider the fact that if the two random variables X and Y are independent (which implies that they are uncorrelated), then their joint distribution is simply given by the product of their marginal PDFs, $f(x, y) = f(x)f(y)$, and the distance between the marginal PDFs and their joint PDF is zero. Thus, the mutual information of X and Y is zero if X and Y are independent. The term mutual information is used because it quantifies the extent to which X and Y contain dependent or redundant information. The concept of mutual information was introduced by Shannon (1948) who applied it to quantifying the capacity of a noisy communication channel.

Following Cover & Thomas (2006), the mutual information, $I(X; Y)$ can be expressed by the equivalent statements,

$$\begin{aligned} I(X;Y) &= D_{\mathrm{KL}}[f(x,y)||f(x)f(y)] \\ &= \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, \mathrm{d}x\, \mathrm{d}y \\ &= h(X) + h(Y) - h(X,Y) \\ &= E_X\left[D_{\mathrm{KL}}\left(f(y|x)||f(y)\right)\right] \\ &= E_Y\left[D_{\mathrm{KL}}\left(f(x|y)||f(x)\right)\right]. \end{aligned} \tag{C3}$$

The last two lines can be read as 'The expectation over x of the relative entropy of y with respect to x' and 'The expectation over y of the relative entropy of x with respect to y.'

It is generally computationally intractable to compute eq. (C3) for high-dimensional PDFs. Thus, we make the simplifying assumption that the posterior PDF can be approximated as a multivariate Gaussian distribution because then the entropies in eq. (C3) can be calculated analytically. This assumption will generally have the effect of overestimating the entropy in our PDFs since there is a well-known result in information theory that a Gaussian has the greatest differential entropy of any probability distribution with a given covariance matrix.

Given a k-dimensional vector of model parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$, our goal is to quantify how independent a particular model parameter, θ_i, is. In other words, we want to calculate the distance between $f(\theta_i|\mathbf{D})f(\theta_1, \theta_2, ...\theta_{i-1}, \theta_{i+1}, ...\theta_k|\mathbf{D})$ and $f(\boldsymbol{\theta}|\mathbf{D})$. For brevity, let us define $\boldsymbol{\theta}_{-\mathbf{i}}$ as all members of $\boldsymbol{\theta}$ except for θ_i, so that we have the following two PDFs,

$$\begin{aligned} f(\theta_i|\mathbf{D}) &= \int f(\boldsymbol{\theta}|\mathbf{D}) \mathrm{d}\boldsymbol{\theta}_{-\mathbf{i}} \\ f(\boldsymbol{\theta}_{-\mathbf{i}}|\mathbf{D}) &= \int f(\boldsymbol{\theta}|\mathbf{D}) d\theta_i = f(\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ...\theta_k|\mathbf{D}), \end{aligned} \tag{C4}$$

which are the marginal distribution for one variable θ_i and the joint distribution for all other variables $\boldsymbol{\theta}_{-\mathbf{i}}$, respectively.

If the posterior PDF is Gaussian, $f(\boldsymbol{\theta}|\mathbf{D}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then,

$$\begin{aligned} f(\theta_i|\mathbf{D}) &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\ f(\boldsymbol{\theta}_{-\mathbf{i}}|\mathbf{D}) &\sim \mathcal{N}(\boldsymbol{\mu}_{-\mathbf{i}}, \boldsymbol{\Sigma}_{-\mathbf{i}}), \end{aligned} \tag{C5}$$

where $\sigma_i^2 = \Sigma_{i,i}$ and $\boldsymbol{\Sigma}_{-\mathbf{i}}$ is $\boldsymbol{\Sigma}$ with the ith row and column deleted. Also μ_i and $\boldsymbol{\mu}_{-\mathbf{i}}$ represent the means of the two Gaussians, respectively, although the mean of a Gaussian is irrelevant to its entropy as will be seen in eq. (C6).

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

From eq. (C1), the entropy for a k-dimensional Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is,

$$h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2}\log_2\left((2\pi e)^k|\boldsymbol{\Sigma}|\right). \tag{C6}$$

Substituting into the third equation of eq. (C3), we find,

$$\begin{aligned} I(\theta_i;\boldsymbol{\theta}_{-\mathbf{i}}) &= h(\theta_i) + h(\boldsymbol{\theta}_{-\mathbf{i}}) - h(\theta_i, \boldsymbol{\theta}_{-\mathbf{i}}) \\ &= +\frac{1}{2}\log_2\left((2\pi e)\sigma_i^2\right) \\ &\quad + \frac{1}{2}\log_2\left((2\pi e)^{k-1}|\boldsymbol{\Sigma}_{-\mathbf{i}}|\right) \\ &\quad - \frac{1}{2}\log_2\left((2\pi e)^k|\boldsymbol{\Sigma}|\right) \\ &= \frac{1}{2}\log_2\frac{\sigma_i^2|\boldsymbol{\Sigma}_{-\mathbf{i}}|}{|\boldsymbol{\Sigma}|}. \end{aligned} \tag{C7}$$

This is the estimate of mutual information that we used for the analyses in Section 6.

APPENDIX D: LISTS OF MATHEMATICAL SYMBOLS AND NOTATION

Table D1. Mathematical symbols.

Symbol	Description
$\mathbf{C}$ (also $\boldsymbol{\Sigma}$)	Covariance matrix.
$\mathbf{C_d}$	Data covariance matrix: a matrix of uncertainties on observations, $\mathbf{D}$.
$\mathbf{C_m}$	Model covariance matrix: a matrix of uncertainties on the model parameters, $\boldsymbol{\theta}$.
$\mathbf{C_p}$	Prediction covariance matrix: a matrix of uncertainties due to errors in forward model, $\mathbf{G}(\boldsymbol{\theta})$.
$\mathbf{C}_{\boldsymbol{\chi}}$	Total covariance matrix of residuals between data and predictions ($\mathbf{C}_{\boldsymbol{\chi}} = \mathbf{C_d} + \mathbf{C_p}$).
$\mathbf{C}^{\boldsymbol{k}}_{\boldsymbol{\chi}}$	$\mathbf{C}_{\boldsymbol{\chi}}$ for kinematic data.
$\mathbf{C}^{\boldsymbol{s}}_{\boldsymbol{\chi}}$	$\mathbf{C}_{\boldsymbol{\chi}}$ for static data.
c_m	A constant used to scale the model covariance matrix: $\boldsymbol{\Sigma}_{\mathbf{m}} = c_m^2\mathbf{C_m}$.
$c_v[\cdot]$	Coefficient of variation, defined as the ratio of the standard deviation to the mean: $c_v = \frac{\sigma}{\mu}$.
$\mathbf{D}$	Observed data (a vector of real numbers).
$\mathbf{d}$	Data predicted by stochastic forward model (an uncertain-valued vector).
$\hat{\mathbf{d}}_{\mathbf{k}}$	Vector of predicted kinematic data produced by deterministic forward model $\mathbf{G}(\boldsymbol{\theta}_{\mathbf{k}})$.
$\hat{\mathbf{d}}_{\mathbf{s}}$	Vector of predicted static data produced by deterministic forward model $\mathbf{G}(\boldsymbol{\theta}_{\mathbf{s}})$.
f (also g, p, q)	Probability density function.
$\mathbf{G}(\cdot)$	A deterministic forward model that accepts a vector of model parameters and returns a vector of predicted observations.
$\mathbf{G_k}(\cdot)$	Deterministic forward model for kinematic data.
$\mathbf{G_s}$	Deterministic forward model for static data comprised of static ($t = \infty$) component of point source Green's function, $\tilde{G}$.
$\tilde{G}$	Point source Green's function.
g (also f, p, q)	Probability density function.
$\tilde{g}$	Green's function, $\tilde{G}$, convolved with a source-time function, s.
$\mathbf{H_0}$	Hypocentre location on the fault plane.
$h(\cdot)$	Differential entropy of $(\cdot)$.
M	Total number of transitional PDFs (cooling stages).
$\mathbf{m}$	Vector of parameter values that specify deterministic forward model $\mathbf{G}(\mathbf{m})$.
$(\cdot)_m$	Index over transitional PDFs (cooling stages).
N	Number of samples of a target PDF. In CATMIP, N is not only the number of samples of the posterior PDF output by the algorithm, it is also equal to the number of Markov chains per cooling step as well as the number of samples output for each transitional PDF $f(\boldsymbol{\theta}\|\mathbf{D}, \beta_m)$.
N_{steps}	Length of a Markov chain (i.e. number of random walk steps).
N_{dp}	Number of data points.
N_{ds}	Number of data sets.
N_k	Number of kinematic data points.
N_s	Number of static data points.
n_s	Number of seismic sources.
$\mathcal{N}$	Gaussian distribution.
p (also f, g, q)	Probability density function.
p	Probability associated with plausibility weight, w.
q (also f, g, p)	Probability density function.
$\mathbb{R}^n$	n-dimensional Euclidian space.
$s(\cdot)$	Source-time function.
T_r	Duration of source-time function, $s(\cdot)$.
t	Time.
$t_0(\cdot)$	First arrival time of the rupture wave front at $(\cdot)$.
U	Displacement on the fault plane.
$U_{\parallel}$	Displacement on the fault plane in the direction aligned with the rake angle

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022

Table D1. (*Continued.*)

Symbol	Description
$U_{\perp}$	Displacement on the fault plane in the direction perpendicular to $U_{\parallel}$.
$\mathcal{U}$	Uniform distribution.
V_r	Rupture velocity.
w	Plausibility weight.
α	A fractional error used to parametrize $\mathbf{C_p}$.
β, γ	'Inverse temperature' for transitioning or annealing.
$\boldsymbol{\zeta}$	Receiver location.
$\boldsymbol{\theta}$	Vector of parameter values which specify full stochastic forward model for data likelihood function $\mathrm{p}(\mathbf{D}\|\boldsymbol{\theta})$. (Often $\boldsymbol{\theta} = \mathbf{m}$.)
$\boldsymbol{\theta}_\mathbf{k}$	Vector of kinematic-only parameter values for the kinematic forward model such that the vector of all parameters for the kinematic model is $\mathbf{m} = (\boldsymbol{\theta}_\mathbf{s}, \boldsymbol{\theta}_\mathbf{k})$.
$\boldsymbol{\theta}_\mathbf{s}$	Vector of parameter values for the static forward model.
μ	Mean.
σ	Standard deviation.
$\boldsymbol{\Sigma}$ (also $\mathbf{C}$)	Covariance matrix.
$\boldsymbol{\Sigma}_\mathbf{m}$	Scaled version of model covariance matrix, $\mathbf{C_m}$.

Table D2. Mathematical notation.

Notation	Description
$\mathrm{p}(x)$	Probability density of continuous variable x.
$\mathrm{p}(x, y)$	Joint probability density function of x and y.
$\mathrm{p}(x\|y)$	Conditional probability density function of x given y.
$D_{\mathrm{KL}}(f\|\|g)$	Relative entropy between PDFs f and g.
$E[X]$	Expected value of X.
$I(X; Y)$	Mutual information of X and Y.
$\bar{(\cdot)}$	Mean of $(\cdot)$.

Downloaded from https://academic.oup.com/gji/article/194/3/1701/645931 by U.S. Department of Justice user on 16 August 2022