Geophysical Journal International

Geophys. J. Int. (2014) **196,** 1676–1693
Advance Access publication 2013 December 27

doi: 10.1093/gji/ggt473

A framework for fast probabilistic centroid-moment-tensor determination—inversion of regional static displacement measurements

Paul Käufl,[1] Andrew P. Valentine,[1] Thomas B. O'Toole[2] and Jeannot Trampert[1]

[1]*Department of Earth Sciences, Universiteit Utrecht, Budapestlaan* 4*, NL-*3584 *CD Utrecht, The Netherlands. E-mail: p.j.kaufl@uu.nl*
[2]*Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX*1 3*AN, UK*

Accepted 2013 November 19. Received 2013 October 4; in original form 2013 July 8

SUMMARY

The determination of earthquake source parameters is an important task in seismology. For many applications, it is also valuable to understand the uncertainties associated with these determinations, and this is particularly true in the context of earthquake early warning (EEW) and hazard mitigation. In this paper, we develop a framework for probabilistic moment tensor point source inversions in near real time. Our methodology allows us to find an approximation to $p(m|d)$, the conditional probability of source models (m) given observations (d). This is obtained by smoothly interpolating a set of random prior samples, using Mixture Density Networks (MDNs)—a class of neural networks which output the parameters of a Gaussian mixture model. By combining multiple networks as 'committees', we are able to obtain a significant improvement in performance over that of a single MDN. Once a committee has been constructed, new observations can be inverted within milliseconds on a standard desktop computer. The method is therefore well suited for use in situations such as EEW, where inversions must be performed routinely and rapidly for a fixed station geometry. To demonstrate the method, we invert regional static GPS displacement data for the 2010 M_W 7.2 El Mayor Cucapah earthquake in Baja California to obtain estimates of magnitude, centroid location and depth and focal mechanism. We investigate the extent to which we can constrain moment tensor point sources with static displacement observations under realistic conditions. Our inversion results agree well with published point source solutions for this event, once the uncertainty bounds of each are taken into account.

Key words: Neural networks, fuzzy logic; Inverse theory; Probabilistic forecasting; Probability distributions; Earthquake source observations; Early warning.

1 INTRODUCTION

Studying earthquake sources is one of the fundamental tasks of seismology. Knowledge of the source is required for many applications, including the quantification and mitigation of seismic hazard, seismic tomography and the enforcement of the nuclear-test-ban treaty. It is important to determine not only the parameters describing a physical source model, but also their associated uncertainties. In particular, the latter are valuable for earthquake early warning (EEW) systems, where quickly determined source parameters may be used to alert residents of imminent ground shaking and to assess where to direct resources in the aftermath of an earthquake. Realistic uncertainty bounds on source parameters are also required for other applications. However, most existing methods only provide point estimates; any uncertainty estimation has to be carried out retrospectively.

In this study, we will develop a method capable of inverting a wide variety of data for point source parameters within a Bayesian framework. Our method is based on using a neural network to approximate posterior distributions of source parameters, allowing uncertainties and trade-offs to be represented. The method also allows results to be obtained rapidly—within a fraction of a second—once observations are available, making it particularly useful for EEW. We demonstrate our method using static displacement data observed by a continuous GPS network. However, the method is not limited to a specific type of data and can in general be used for joint inversions of different observables, such as strong motion data and displacement waveforms.

We describe earthquake sources using a moment tensor point source description, which has proven to be sufficient for many applications. Moment tensor solutions are routinely calculated and collected in comprehensive catalogues, for example, the Global Centroid Moment Tensor Project (GCMT, www.globalcmt.org). Apart from the underlying centroid-moment-tensor (CMT) algorithm (Dziewonski *et al.* 1981; Ekström *et al.* 2012), which is very robust and widely applied, there are a variety of other

methods that have been used to invert long-period body and surface waves for global moment tensor solutions, for example, Kanamori (1993); Kanamori & Rivera (2008); Duputel *et al.* (2011).

Within this framework, a source is described by six independent moment tensor components. The wave equation is linear in these, but the source inversion problem becomes non-linear if the centroid location and origin time are also to be determined, in which case solutions are commonly obtained by iterating with a linearized inversion algorithm. Furthermore, the problem can be ill-posed due to poor station coverage and noisy data, leading to non-unique and uncertain solutions. This arises when multiple models, perhaps lying in disjunct subsets of the model space, give rise to a good data fit.

Owing to their linearized nature, estimating realistic uncertainty bounds within the framework of classical CMT-type inversions is challenging. Typically, analysis is restricted to calculations of 'standard errors' due to noise within the inversion, and a more thorough error analysis is often lacking. Recent studies (Duputel *et al.* 2012; Valentine & Trampert 2012b) suggest that the reported errors typically underestimate the uncertainties significantly.

This motivates the use of a Bayesian statistical framework, giving rise to posterior distributions rather than just point estimates of model parameters (Tarantola 2005). Our methodology is based on parametrizing posterior probability distributions as sums of Gaussians. We use neural networks that output the coefficients of these Gaussian kernels to approximate the inverse mapping. Bishop (1995) coined the term Mixture Density Network (MDN) for this particular type of neural network.

Most Bayesian geophysical inversions to date have been based on Monte Carlo methods, which directly generate samples of the posterior distribution. For every new observation, the sampling procedure must be repeated, which is often expensive and time-consuming. In our neural network framework, in contrast, the sampling stage is separated from the inversion stage. Once a network has been trained using a set of previously generated samples, it can be presented with new observations, and rapidly outputs the corresponding posterior distribution of model parameters.

Neural networks have successfully been applied to a great variety of classification and regression problems. Geophysical examples include seismic reflection data inversion (Röth & Tarantola 1994), probabilistic inversion of surface wave velocities for Eurasian crustal thickness (Devilee *et al.* 1999), automated data selection and quality assessment (Valentine & Woodhouse 2010) and dimensionality reduction of seismograms (Valentine & Trampert 2012a). MDNs have been used to invert surface wave data for global Moho depth (Meier *et al.* 2007a,b), water content in the transition zone (Meier *et al.* 2009) and more recently for the inversion of *P*- and *S*-wave velocity for petrophysical parameters (Shahraeeni & Curtis 2011). Shahraeeni *et al.* (2012) extend this approach to 3-D seismic data and de Wit *et al.* (2013) use MDNs to infer the Earth's 1-D seismic velocity structure from body wave traveltimes.

In this paper, we intend to recover centroid location, event magnitude and the moment tensor from coseismic static displacements as observed by GPS sensors (Blewitt 2007). From our perspective, static offset data provide a simple and manageable data set for testing and developing this method. However, due to the increasing availability of GPS stations in seismically active regions, there is also growing interest in the possibilities that this type of data provides for seismology.

Conventionally, displacement time-series are retrieved from velocity or acceleration seismograms by integration. However, this is only possible in a limited frequency range, due to ground tilts and rotations that result in distortions and baseline shifts and the accumulation of other observational errors. In particular, it is difficult to recover the static offset (the displacement remaining once all ground shaking has ceased). Furthermore, due to limitations in the dynamic range of the broad-band instruments, velocity seismograms recorded close to the source often suffer from clipping. Neither of these problems affect GPS sensors, which directly measure the ground displacement (e.g. Larson *et al.* 2003; Bock *et al.* 2011). GPS data may therefore complement traditional seismic observations and provide additional information on earthquake magnitude and fault mechanism in close proximity to the hypocentre or for very large events (Wang *et al.* 2013). These properties are particularly valuable in the context of EEW systems (Crowell *et al.* 2009; Melgar *et al.* 2013).

Displacement time-series from GPS sensors have previously been used for rapid point source inversions. A recent example is given by Crowell *et al.* (2009), who estimate the earthquake hypocentre and magnitude using a grid search combined with an empirical scaling relation in near real time. High-rate GPS waveforms are inverted by O'Toole *et al.* (2012) using an adapted CMT inversion algorithm, and Zheng *et al.* (2012) invert 5 Hz GPS displacement records for the focal mechanism using a grid search approach. Allen & Ziv (2011) use an early estimate of the static offset to rapidly determine the earthquake magnitude given prior information on the hypocentre and fault geometry. Moreover, the static offset has recently been used for fast moment tensor determination by Melgar *et al.* (2012) and O'Toole *et al.* (2013).

We begin by describing our MDN-based Bayesian inversion framework and introduce a suitable parametrization of the space of source models. We then validate our method by means of a synthetic experiment and subsequently apply it to data recorded in southern California after the 2010 M_W 7.2 El Mayor Cucapah event. Finally, we demonstrate how uncertainties in earth model can be incorporated into our analysis, and illustrate how our approach may be used to explore trade-offs between model parameters.

2 PROBABILISTIC CENTROID-MOMENT-TENSOR INVERSION USING MDNS

For any event, we wish to obtain a probability density function (pdf) describing our state of knowledge of each source parameter. We approximate these posterior probability densities using a neural network, which outputs the parameters of a Gaussian mixture model (GMM). This approximation is based on a set of examples of the mapping between observable data and source parameters, obtained by forward modelling. This is not too different from other advanced probabilistic methods, such as the Neighbourhood Algorithm (Sambridge 1999a), where the interpolation is done in the appraisal stage (Sambridge 1999b) by importance sampling a piecewise constant approximation of the posterior probability. Here, we use a neural network—a general function approximator—to retrieve a smooth interpolation. The advantage of our approach is that inversions are fast and can be performed repeatedly for new observations without significant additional computational effort. The evaluation of this neural network approximation only involves repeated matrix multiplications and evaluation of a simple non-linear function, and can be performed within milliseconds on a standard desktop computer.

2.1 The inverse problem

Solving the inverse problem—the problem of finding source models that fit an observed datum—is equivalent to finding the conditional probability density (Tarantola 2005)

$$p(\mathbf{m}|\mathbf{d}) = kp(\mathbf{m})p(\mathbf{d}|\mathbf{m}), \tag{1}$$

where $\mathbf{m} \in \mathbb{M}$ is a source model, $\mathbf{d} \in \mathbb{D}$ a vector of observable data points, $p(\mathbf{m})$ is an unconditional or prior probability density on the model space and k a normalization constant.

We will find approximations of the conditional distribution (1) based on samples $\{\mathbf{m}_i, \mathbf{d}_i\}$ obtained by forward modelling. Under the assumption of Gaussian measurement errors, we have

$$p(\mathbf{d}|\mathbf{m}) \propto \exp\left\{-\frac{1}{2}[\mathbf{d}-\mathbf{g}(\mathbf{m})]^{\mathrm{T}}\mathbf{C}_d^{-1}[\mathbf{d}-\mathbf{g}(\mathbf{m})]\right\}, \tag{2}$$

where $\mathbf{g}(\mathbf{m})$ denotes the deterministic forward relation between model $\mathbf{m}$ and data $\mathbf{d}$, and $\mathbf{C}_d$ is a covariance matrix. Having generated a sample $\mathbf{m}_i$ from the prior $p(\mathbf{m})$, which is straightforward to do for a suitable choice of distribution, we subsequently compute the corresponding 'noisy' datum

$$\mathbf{d}_i = \mathbf{g}(\mathbf{m_i}) + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon}$ is a simulated measurement noise vector drawn from a Gaussian distribution with zero mean and covariance $\mathbf{C}_d$.

In order to present and interpret the posterior, we use the concept of marginalization (e.g. MacKay 2003), that is, integrating out all model parameters except those of interest. We therefore define the 1-D marginal probability density

$$p(m_i|\mathbf{d}) = \int p(\mathbf{m}|\mathbf{d}) \prod_{k\neq i} \mathrm{d}m_k \tag{4}$$

and the 2-D marginal density

$$p(m_i, m_j|\mathbf{d}) = \int p(\mathbf{m}|\mathbf{d}) \prod_{k\neq i,j} \mathrm{d}m_k. \tag{5}$$

Note that by using the definition of conditional probability (e.g. MacKay 2003), we can express probability densities defined on an l-dimensional space in terms of marginal and conditional probability densities over an $(l-1)$-dimensional space, since

$$p(m_1, ..., m_l|\mathbf{d}) = p(m_1, ..., m_{k-1}, m_{k+1}, ..., m_l|m_k, \mathbf{d})p(m_k|\mathbf{d}). \tag{6}$$

By recursively applying eq. (6), we can reconstruct distributions over arbitrary dimensional subspaces of $\mathbb{M}$ from 1-D marginal and conditional distributions.

2.1.1 Gaussian mixture models

We model 1-D probability densities, such as in eqs (4) and (6), as mixtures of Gaussians with input-dependent parameters. We define

$$p(m_k|\mathbf{d}) \simeq \sum_{i=1}^{M} \alpha_i(\mathbf{d})\phi_i(m_k|\mathbf{d}), \tag{7}$$

where M is the number of kernels, $\alpha_i(\mathbf{d})$ are input-dependent mixture coefficients that sum to unity and

$$\phi_i(m_k|\mathbf{d}) = \frac{1}{\sqrt{2\pi}\sigma_i(\mathbf{d})} \exp\left\{-\frac{[m_k-\mu_i(\mathbf{d})]^2}{2\sigma_i(\mathbf{d})^2}\right\} \tag{8}$$

are Gaussian kernels with input-dependent mean $\mu_i(\mathbf{d})$ and standard deviation $\sigma_i(\mathbf{d})$.

Where a model parameter is periodic and defined on the domain $[0, 2\pi)$ rather than on $\mathbb{R}$, we replace (8) with the wrapped normal kernel

$$\tilde{\phi}_i(\theta|\mathbf{d}) = \sum_{n=-P}^{P} \phi_i(\theta + n2\pi|\mathbf{d}). \tag{9}$$

Note that (9) is normalized on $[0, 2\pi)$ for $P = \infty$, since

$$\int_0^{2\pi} \tilde{\phi}(\theta|\mathbf{d})\,\mathrm{d}\theta = \int_{-\infty}^{\infty} \phi(\chi|\mathbf{d})\,\mathrm{d}\chi = 1, \tag{10}$$

with $\chi = \theta + n2\pi$ (Bishop & Legleye 1994). In practice we have to limit P, while ensuring that it remains large enough for the integral in (10) to contain all areas where the underlying Gaussian functions $\phi_i(\chi|\mathbf{d})$ do not vanish. Since the standard deviation of the Gaussian kernels is typically much smaller than a few periods, and the means $\mu_i(\mathbf{d})$ can be assumed to be not further apart than one period from the range $[0, 2\pi)$, it is not necessary to sum over many cycles and $P = 7$ is found to be sufficient in our case.

It has been shown that, given a sufficient number of kernels, any probability density can be approximated to arbitrary accuracy with a GMM (McLachlan & Basford 1988). In particular, GMMs are able to capture multimodal distributions. Fig. 1 shows examples

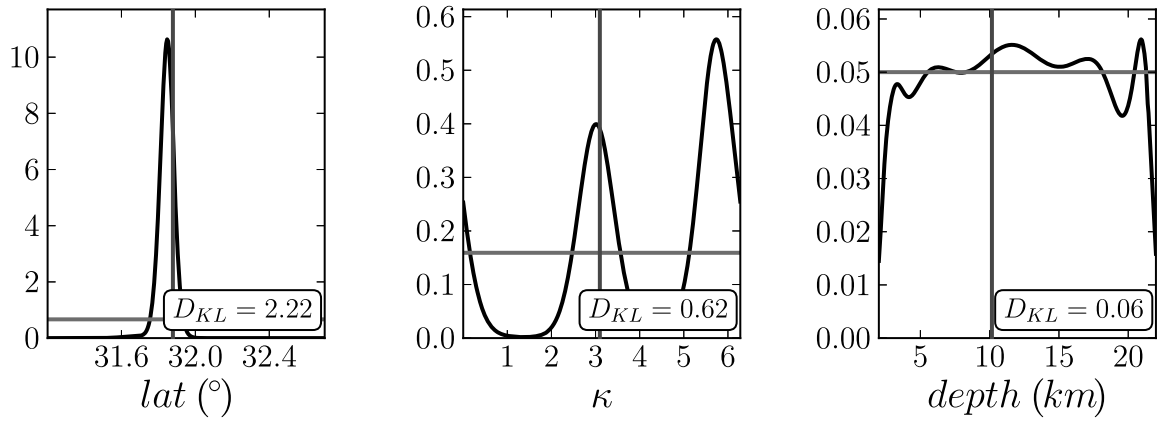

Figure 1. Examples of 1-D posterior marginal distributions (black line). Left-hand panel: Unimodal distribution, the target value (vertical bar) is close to the mode of the distribution. Middle-panel: Bimodal distribution. Although the target value coincides with a region of high probability, the mode of the distribution is comparatively distant. Right-hand panel: Approximately uniform posterior. The mode of the distribution becomes irrelevant. The values for the information gain D_{KL} are determined according to eq (14). Note that the information gain is low when the posterior resembles the prior (green line) closely.

of GMMs approximating a unimodal, a bimodal and a uniform distribution. Note that for (7) to be a meaningful approximation of the conditional distribution (1), it is necessary that the set of samples $\{\mathbf{m}_i, \mathbf{d}_i\}$ is sufficiently dense that it captures all possible variations of the underlying relation. We believe that this requirement can be met in our case, since we are working in a relatively low-dimensional model space.

2.1.2 Mixture Density Networks

In order to find the marginal posterior probability (7), we need to find functional relations $\alpha_i(\mathbf{d})$, $\mu_i(\mathbf{d})$ and $\sigma_i(\mathbf{d})$. We will approximate these mappings by a feed-forward neural network as shown in Fig. 2. A fully connected feed-forward neural network is a general function approximator, able to capture a wide class of functional mappings to arbitrary accuracy (Hornik *et al.* 1989). Similar to Kolmogorov's superposition theorem (Kolmogorov 1957), a neural network models a non-linear function by applying non-linear basis functions to linear combinations of input variables, which are in turn combined linearly to form the input to a subsequent layer. Such a model can be visualized graphically and we use a two-layer structure as depicted in Fig. 2. A detailed description and analytical expressions for the network outputs is given in Appendix Section A1. Neural networks with outputs that are regarded as the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ of a GMM, are termed MDNs and have been introduced by Bishop (1995).

The network model $f(\mathbf{d}; \mathbf{w})$ of a function is controlled by a set of free parameters, the 'network weights' $\mathbf{w}$, and we can assign a probability $p(\mathbf{w}|\mathcal{D})$ to a specific set of weights given a set of samples $\mathcal{D} = \{\mathbf{d}_n, \mathbf{m}_n\}$, consisting of input–output pairs, which are—in our case—synthetic displacement vectors $\mathbf{d}_n \in \mathbb{D}$ and the corresponding source models $\mathbf{m}_n \in \mathbb{M}$. We can therefore express the posterior distribution (4) explicitly as a marginal over the distribution of network weights,

$$p(m_k|\mathbf{d}) = \int p(m_k|\mathbf{d}, \mathbf{w}) p(\mathbf{w}|\mathcal{D})\, \mathrm{d}\mathbf{w}. \tag{11}$$

In practical implementations, it is hard to evaluate the integral in (11), since it would involve sampling from the posterior weight distribution $p(\mathbf{w}|\mathcal{D})$, which is not known explicitly. Instead a common workaround replaces the integral over the weight space with a single set of weights $\mathbf{w}^*$ for which the probability $p(\mathbf{w}^*|\mathcal{D})$ is maximized (Bishop 1995). This assumes that $p(\mathbf{w}|\mathcal{D})$ is sufficiently narrow and centred around $\mathbf{w}^*$.

The set of optimal parameters $\mathbf{w}^*$ is found by maximizing the likelihood of a 'training set' $\mathcal{D}_{\mathrm{tr}} = \{\mathbf{d}_n, \mathbf{m}_n\}$, or equivalently minimizing the error function

$$E[\mathcal{D}_{\mathrm{tr}}] = -\sum_n \ln p\left[(m_k)_n|\mathbf{d}_n, \mathbf{w}\right], \tag{12}$$

where the sum runs over all examples in $\mathcal{D}_{\mathrm{tr}}$. The minimization of (12) is done using the limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) method (Nocedal 1980). As a quasi-Newton optimizer it makes use of second-order information about the error surface but replaces the full approximation of the Hessian with a sparse representation, which makes it suitable for problems with a large number of parameters. The required derivatives of (12) are efficiently calculated using error backpropagation (Rumelhart *et al.* 1986).

Since the method is iterative, a starting point $\mathbf{w}^0$ in weight space has to be chosen. Following Bishop (1995), we randomly initialize the network weights such that the output of the untrained network follows the unconditional distribution of the target data $p(m_k|\mathcal{D}_{tr})$. See Appendix Section A2 for a detailed description. Once a sufficient number of training iterations have been performed, we obtain an optimal set of weights $\mathbf{w}^*$ and can express (11) as

$$p(m_k|\mathbf{d}) \simeq p(m_k|\mathbf{d}, \mathbf{w}^*). \tag{13}$$

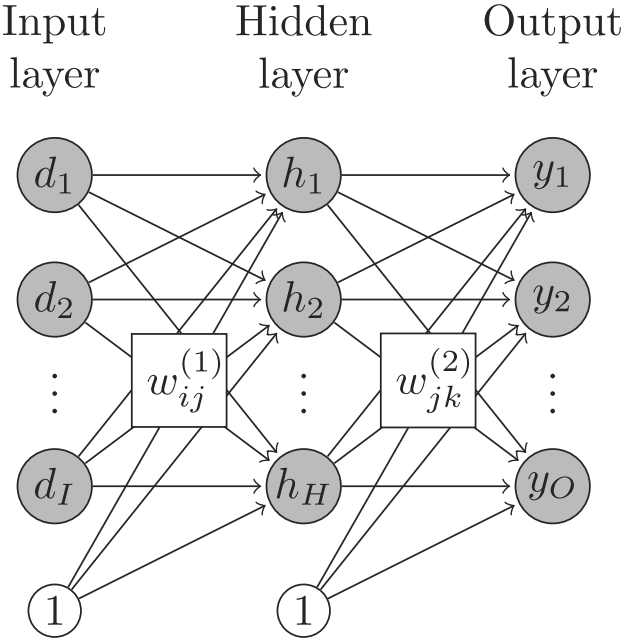

Figure 2. The two-layer feed-forward neural network used to model the functional relation between data vectors $\mathbf{d}$ and GMM parameters α_i, μ_i and σ_i. The filled circles represent computational units, which apply a linear or sigmoidal 'activation function' to their according inputs. The input to any unit is given by a weighted sum of the outputs of the previous layer and a 'bias unit', which is fixed to one, with weights $\mathbf{w}$ (interconnecting lines). The units of the hidden layer h_j are referred to as 'hidden units' in the main text. For details and analytical expressions, see Appendix Section A1.

2.1.3 Assessing the network performance

Once a trained network is available we can assess its performance using a third independent test data set $\mathcal{D}_{\mathrm{test}}$. We can evaluate the test set error $E[\mathcal{D}_{\mathrm{test}}]$—that is, the negative log-likelihood of the test set. The test set error can be used to compare different networks' performance in predicting m_k for the examples in the data set. However, a relatively large test set error does not necessarily indicate that the network is performing poorly: it can also arise if the data are insensitive to the parameter m_k. In this case, we cannot learn much upon seeing the training set and it is impossible to find networks that give a low error.

We can quantify how much has been learned about a certain parameter by measuring the difference between prior and posterior distributions. This information gain can be measured using the Kullback–Leibler divergence (e.g. MacKay 2003)

$$D_{KL} = \int \ln\left(\frac{p(m_k|\mathbf{d})}{p(m_k)}\right) p(m_k|\mathbf{d})\, \mathrm{d}m_k, \tag{14}$$

which is a dimensionless number given in logarithmic information units (nats). A few examples for distribution pairs and their corresponding information gain are given in Fig. 1. Note that in the

case that nothing has been learned upon seeing the data vector $\mathbf{d}$, $p(m_k|\mathbf{d}) = p(m_k)$ and $D_{KL} = 0$.

2.2 Regularization and complexity of the network model

The generalization performance of a trained network—that is, how well it performs on previously unseen data—is influenced by a number of factors. In particular these include the network architecture, the number of hidden units and the size and sampling distribution of the training set. In addition, the fact that we replaced the full predictive distribution (11) with a term accounting for only a single set of weights, can limit the generalization capabilities, since only a limited part of the weight space has been explored by the training algorithm. It is possible that other networks exist that can explain the training data equally well. In order to mitigate these effects we 'regularize' during the network training process. This is discussed in Appendix Section A3. Furthermore by combining multiple networks—each of them trained using a different starting point in weight space—into 'network committees', we can achieve a significant improvement in generalization performance. A detailed description and illustrative results can be found in Appendix Section A4.

2.3 Parametrization of point sources

Seismic point sources are fully determined by a point in space and time and a second-order tensor—the 'moment tensor'. Any seismic moment tensor can be written as a symmetric 3-by-3 matrix with six independent components (e.g. Madariaga 2007).

It might appear straightforward to use the six moment tensor components as independent parameters directly. Doing so, however, has several disadvantages. The moment tensor encodes three distinct types of information: the total amount of energy released; the radiation pattern and the orientation of the source. Using the moment tensor components as parameters does not allow for an easy distinction between the three. This is particularly problematic in probabilistic inversions, since uncertainties on parameters directly relate to distances in parameter space. In order to find interpretable probability distributions it is important to choose a parametrization in which distances are physically meaningful and where the amount of duplication—that is, the existence of two parameter configurations that correspond to the same physical source—is minimized. Furthermore, correlations among multiple parameters should reflect physical relations, rather than interdependencies caused by the particular parametrization.

Therefore, a description that clearly separates magnitude, radiation pattern and source orientation has advantages. We thus follow Tape & Tape (2012), who give a geometric approach to efficiently parametrize all possible moment tensors. A similar exposition is also found in Chapman & Leaney (2012). We will briefly review the parametrization, but refer the reader to their publications for further details.

A moment tensor is fully determined by a set of three eigenvalues $\lambda_i \in \mathbb{R}$ and an orthonormal basis $\mathbf{U}$, determining the orientation of the source. If the trace of the moment tensor vanishes, the net moment is zero and the source is called 'deviatoric' (e.g. Shearer 1999). We restrict ourselves to deviatoric sources, since this representation is chosen by most established catalogues for earthquakes, but it is straightforward to extend our method to admit all possible point sources.

Table 1. The eight parameters used to describe all deviatoric point sources.

Parameter	Description
γ	Deviation from a pure DC, for which $\gamma = 0$
κ	Strike
σ	Rake
h	cos(dip)
ρ	$\sqrt{2}M_0$
lat	Centroid latitude
lon	Centroid longitude
depth	Centroid depth

The eigenvalues control radiation pattern and magnitude. The latter is given by $\rho = ||\mathbf{\Lambda}|| = \sqrt{2}M_0$, with the eigenvalue triplet $\mathbf{\Lambda}$ and the scalar seismic moment M_0. We can thus find a magnitude-independent radiation pattern

$$\hat{\mathbf{\Lambda}} = \frac{\mathbf{\Lambda}}{\rho}, \tag{15}$$

where the two remaining degrees of freedom, governing $\hat{\mathbf{\Lambda}}$, can be parametrized using spherical coordinates on the unit sphere. Tape & Tape (2012) show that only a subset of the sphere—the fundamental lune, corresponding to one of the six eigenvalue permutations—is required to produce all possible radiation patterns. The fundamental lune is given by (Tape & Tape 2012, eq. 17)

$$\mathbb{L} = \left\{\mathbf{\Lambda} \in \mathbb{R}^3 : \lambda_1 \geq \lambda_2 \geq \lambda_3, ||\mathbf{\Lambda}|| = 1\right\}. \tag{16}$$

In the following. we use the set of coordinates (γ, β) to parametrize the fundamental lune, where $-\pi/6 \leq \gamma \leq \pi/6$ and $0 \leq \beta \leq \pi$. For deviatoric sources, we fix $\beta = \pi/2$ and γ becomes a measure for the extent to which a moment tensor departs from a pure double-couple (DC), for which $\gamma = 0$. The parameter γ is related to the more commonly known parameter $\epsilon = \lambda_2/\max(|\lambda_1|, |\lambda_3|)$ (e.g. Giardini 1984) via the relation $\tan\gamma = \sqrt{3}\epsilon/(2 - |\epsilon|)$ (Tape & Tape 2012, section 8).

The space of source orientations can be parametrized by three angles (κ, σ, θ) with ranges $0 \leq \kappa \leq 2\pi$, $-\pi/2 \leq \sigma \leq \pi/2$ and $0 \leq \theta \leq \pi/2$, which correspond to the strike, slip and dip angle, respectively, for DC sources. For a given beachball pattern $\mathbf{\Lambda}$, a uniform distribution of orientations is retrieved if (κ, σ, h), with $h = \cos(\theta)$, are uniformly distributed (Tape & Tape 2012).

We thus work in the domain $\mathbb{T}^{\mathrm{DEV}}$ of deviatoric moment tensors given by

$$\mathbb{T}^{\mathrm{DEV}} = \left\{(\gamma, \kappa, \sigma, h, \rho) : -\frac{\pi}{6} \leq \gamma \leq \frac{\pi}{6},\ 0 \leq \kappa \leq 2\pi,\right.$$
$$\left.-\frac{\pi}{2} \leq \sigma \leq \frac{\pi}{2},\ 0 \leq h \leq 1,\ \rho > 0\right\}. \tag{17}$$

The eight-dimensional model space is the joint space of deviatoric moment tensors and hypocentral locations

$$\mathbb{M} = \mathbb{T}^{\mathrm{DEV}} \times \{(\mathrm{lat}, \mathrm{lon}, \mathrm{depth})\}. \tag{18}$$

Note that there is no temporal coordinate, since this cannot be constrained using static data only. Table 1 summarizes the model parameters and their relations to more commonly used source parameters.

2.4 Synthetic static displacements

In order to train networks, we need to generate samples according to (3). For any $\mathbf{m} \in \mathbb{M}$, we can calculate a set of three-component

synthetic static displacements $\mathbf{g}(\mathbf{m}) \in \mathbb{D} = \mathbb{R}^{3N_r}$, where N_r is the number of receivers. We solve the forward problem using a method developed and implemented by O'Toole & Woodhouse (2011), who use an adapted Thomson–Haskell propagator matrix method to solve the elastic wave equation in a plane layered, isotropic medium. In particular, the method is exact and stable in the zero-frequency limit and is thus well suited to calculate static displacements. While there is no need for the calculation of the dynamic wavefield in this demonstration, the flexibility of this forward solver allows us to easily extend our approach to joint inversions of waveform and static data at a later stage. Note that the implicit flat-Earth approximation is valid for shallow events and epicentral distances up to $\sim 20^\circ$ (Okada 1985).

2.5 Data pre-processing

Pre-processing of input and target vectors can significantly increase the convergence speed of the training stage (Bishop 1995) and ideally the input vectors follow a standard normal distribution. We find that we achieve suitable input data distributions by transforming the three-component static displacements $\mathbf{u}$, expressed in Cartesian *N–E–Z*-coordinates, as follows:

$$\tilde{u}_1 = \log ||\mathbf{u}||, \tag{19}$$

$$\tilde{u}_2 = \arccos\left(\frac{u_N}{||\mathbf{u}||}\right), \tag{20}$$

$$\tilde{u}_3 = \operatorname{arctan2}\left(u_E, u_Z\right), \tag{21}$$

where arctan2 denotes the version of the arctan function that takes the phase into account. A $3N_r$-dimensional network input vector $\mathbf{d}$ is subsequently formed by concatenating the $\tilde{u}_i$ at all N_r receivers. The training set $\mathcal{D}_{tr}$ is standardized to have variance one and mean zero (Appendix B). While this increases convergence speed, it also leads to an equal weighting of the input vector components. Note that noise is added to the synthetics before performing the transformation. New data vectors to be presented to the network have to be transformed accordingly.

Target variables m_k are rescaled to $[-1, 1]$, or to $[0, 2\pi)$ for periodic variables. However, in what follows we do not explicitly distinguish between the rescaled parameters $\tilde{m}_k$ and the untransformed parameters m_k, in order to keep the notation simple.

3 DEMONSTRATION

To demonstrate this approach and to study the resolvability of point source parameters using coseismic displacement data we perform two experiments. First, we work in an idealized scenario using only synthetic data and complete azimuthal coverage of receivers. We then extend the setup to a more realistic station distribution and invert observed data for a 2010 M_W 7.2 event recorded by the California Real-Time GPS network (CRTN). We compare our results to several previously published point source inversions and catalogue solutions.

3.1 The 2010 M_W 7.2 El Mayor Cucapah event

An M_W 7.2 event, which occurred on 2010 April 4 in northern Baja California, was recorded by 105 GPS receivers of the CRTN network. The tectonic structure in the source region is comparatively complex and the rupturing process involves a combination of normal and right-lateral faulting (Hauksson *et al.* 2010; Wei *et al.* 2011; Oskin *et al.* 2012). Moment tensor inversions for this event using a similar data set have been performed using a variety of deterministic approaches (Melgar *et al.* 2012; O'Toole *et al.* 2013) and can serve as a benchmark for our results.

We use a set of observed post-earthquake coseismic displacements for the inversion, which have been determined from post-processed site positions before and after the event (Nikolaidis 2002, see also http://sopac.ucsd.edu/processing/refinedModelDoc.html). We simulate wave propagation in the same layered crustal model as used by Melgar *et al.* (2012) and O'Toole *et al.* (2013). The 1-D model is based on the California Community Velocity Model version 4 (Kohler *et al.* 2003), which has been averaged in a box with corners (117°W, 32°N) and (115°W, 34°N) and manually tuned to account for the fact that most of the relevant stations sit on soft, sedimentary material (Melgar, personal communication, 2013).

3.2 Prior constraints

The Bayesian approach requires us to choose prior probability distributions on the parameters and a suitable noise model for the observed data. While prior information on the rupturing process and the faults involved is certainly available for historic events, we aim to investigate the extent to which we can constrain source parameters from the static offset data alone. Furthermore, in an EEW context we want to be able to monitor multiple fault zones and include the possibility of earthquakes happening on previously unknown faults. We therefore choose prior distributions that incorporate all possible deviatoric sources. Within our parametrization (Section 2.3) we obtain a distribution of moment tensors that is approximately uniform, by drawing the orientation- and type-governing parameters (γ, κ, σ, h) from uniform distributions. We do not draw ρ directly, but instead consider distributions on M_W (Hanks & Kanamori 1979), which are related by $\rho = \sqrt{2} \cdot 10^{1.5 \cdot (M_W + 10.7)}$. This aids network training, since ρ itself can vary by multiple orders of magnitude. M_W is subsequently drawn from a uniform distribution in the range from 6.5 to 8.0. The epicentral location has been restricted to a 1.5°-by-2° box centred around the location of the GCMT solution; depth ranges from 2 to 22 km. Fig. 3 shows a set of samples drawn from the prior.

In an EEW setup, the location prior would correspond to the monitored volume of possible source locations, which could be a region encompassing a set of known faults. Note that predictions for events that occur outside the prior range are likely to be meaningless, since they would require the trained networks to extrapolate to previously unseen regions; we have not addressed this in our performance tests in this case. The potentially time-consuming generation of this set of prior samples, forming the training set for the neural networks, can be performed before the EEW system becomes operational.

Throughout this study, we assume that observational noise is Gaussian and uncorrelated. We assign a station-independent noise level of 1 mm in horizontal and 10 mm in vertical direction, which is slightly more conservative than the standard error estimates provided with the data set. Note that if a real-time algorithm (such as Allen & Ziv 2011) were used to retrieve an early estimate of the final static displacement using an initial portion of the displacement waveform, we would typically have to assume a noise level that is up to one order of magnitude higher than that used here. Table 2 summarizes the *a priori* bounds on all eight parameters and the noise standard deviations.

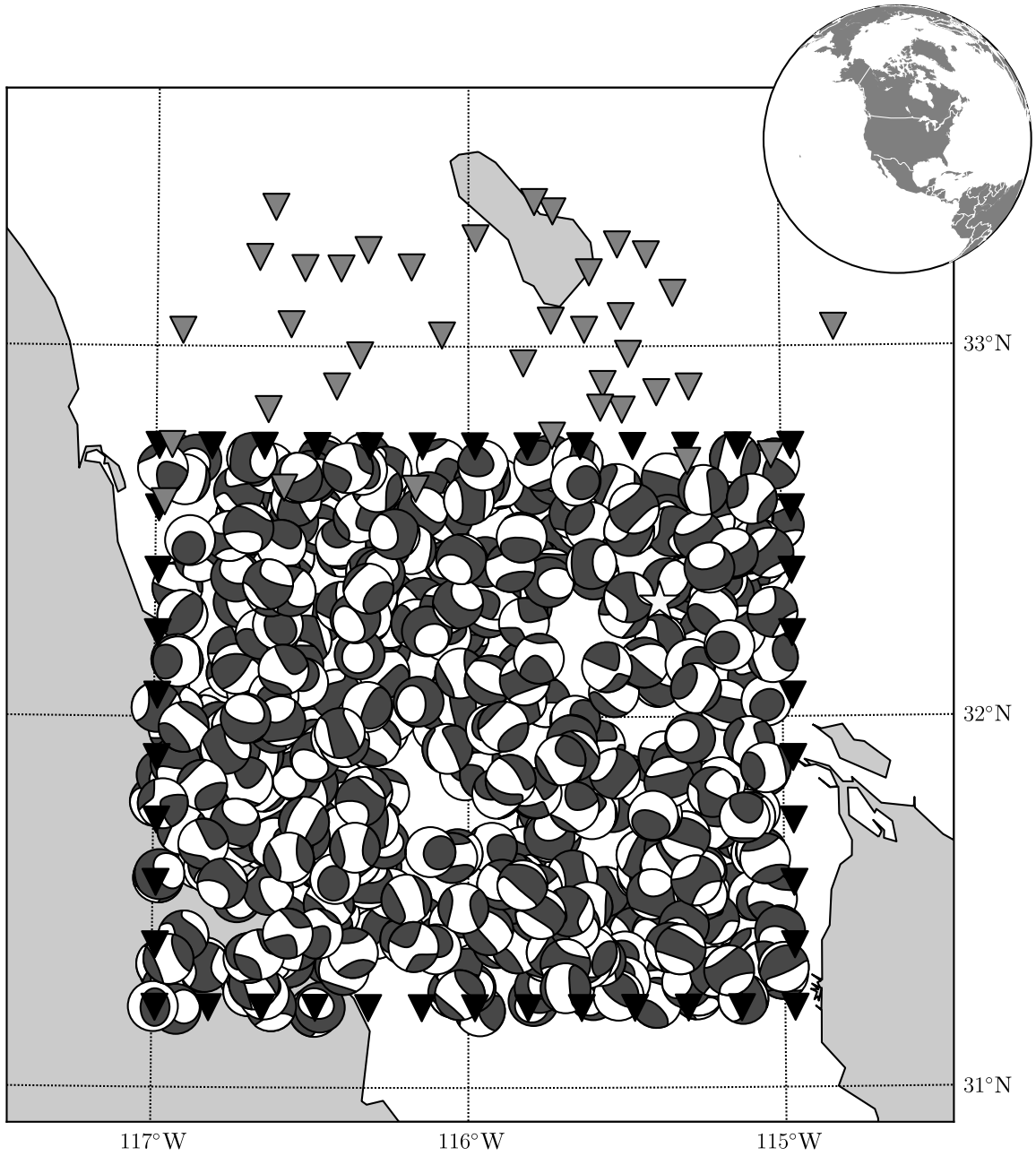

Figure 3. Shown are 500 prior samples covering the study region. The black and red triangles indicate the positions of the 42 virtual receivers used in the synthetic experiment and the 37 CRTN stations, respectively. The reported position of a recorded 2010 M_W 7.2 event is denoted by the yellow star.

3.3 A synthetic experiment

In order to test our methodology and analyse the resolution power of static displacement data we first perform a synthetic test, in which we use an idealized distribution of stations at the surface (black triangles in Fig. 3).

3.3.1 Network training and validation

Having chosen prior distributions and a noise model, we generate a total of 100 000 examples $\{\mathbf{d}_i, \mathbf{m}_i\}$, of which 80 000 form the training set $\mathcal{D}_{\text{tr}}$, 16 000 serve as validation set $\mathcal{D}_{\text{val}}$ and the remaining 4000 examples are used as test set $\mathcal{D}_{\text{test}}$. We subsequently train network committees consisting of $C = 50$ members on marginal distributions $p(m_k|\mathbf{d})$ for each of the eight model parameters. The number of Gaussian mixture components per member has been set to $M = 6$, leading to an overall number of $C \cdot M = 300$ components per committee. The number of hidden units for the committee members is randomly chosen in the range [20, 50] ([40, 60] for latitude and longitude) based on the considerations given in Appendix A4 (see also Fig. A1).

Once the trained network committees have been generated, we present the test set to each committee. For every test set example we retrieve a posterior pdf, which we can compare to the true target value for this example. Quantifying this difference can be difficult, depending on the complexity of the pdf. A simple measure is the

Table 2. Prior parameter ranges and noise amplitudes. Distributions on the source parameters are uniform with hard bounds as indicated. The noise is assumed to be uncorrelated normal with standard deviations as given.

Parameter	Prior range
γ	$-\frac{\pi}{6}$ to $\frac{\pi}{6}$
κ	0–2π
σ	$-\frac{\pi}{2}$ to $\frac{\pi}{2}$
h	0–1
M_{W}	6.5–8.0
lat	31.2°N–32.7°N
lon	115°W–117°W
depth	2–22 km
Noise level	$\sigma_x = \sigma_y = 1$ mm; $\sigma_z = 10$ mm

distance of the mode from the target value. This distance can be visualized as in Fig. 4. If the pdf shows a pronounced maximum and is unimodal for all test set examples, ideally we would see a straight diagonal line, indicating that the predicted modes align well with the target values. However, this measure does not take the full pdf into account. This is particularly significant if the pdf is multimodal. The target value might still lie in a range to which a high probability is assigned despite the mode being comparatively distant from the target value. The measure is rendered completely meaningless in the limiting case of a uniform posterior distribution. In this case, the mode is insignificant, since the same probability is assigned to any parameter value. Illustrative examples can be seen in Fig. 1.

In order to more easily interpret Fig. 4, we take a second measure into account—the information gain. This measures the distance of the posterior from the prior distribution using eq. (14). As discussed in Section 2.1.3, if a posterior distribution resembles the prior closely, the information gain will be close to zero, indicating that we did not learn much about this parameter, and the distance of the mode from the target value is less relevant. This information gain is colour-coded in Fig. 4. The figure also reveals an artefact due to the fact that we approximate marginal distributions by means of GMMs. For marginal distributions that closely resemble a uniform distribution, the mixtures of Gaussians tend to overshoot at the edges of the prior range, since Gaussian kernels that are placed close to a hard bound of the model space are forced to fall to zero sharply, leading to a very narrow kernel (*cf.* right-hand panel of Fig. 1). This effect causes the vertical features in the plots for γ, h and depth at the boundaries of the model space.

We provide histograms of the information gain over the whole test set in Fig. 5. These give some insight into the average resolving power of the data. A large average information gain indicates that a particular parameter is well resolved. Thus, we see that the average information gain is comparatively high for M_{W}, latitude and longitude with values of around ∼2 nats, indicating that these parameters are well resolved. We observe intermediate average values for strike κ and rake σ and low values, corresponding to a very poor resolution, for γ, cosine of dip (h) and depth. Finally, Fig. 6 shows 1-D posterior marginals for a test set example. In addition to the posterior distribution (thick black line), the predictions of the individual committee members are shown in light grey. The known target values for this example are highlighted by vertical lines.

3.3.2 *Conclusion*

We find that, given the noise model, prior bounds and with an idealized station distribution, we are able to determine epicentral location and magnitude well. In general, γ, which indicates how far the source deviates from a pure DC, the cosine of the dip (h) and the source depth seem poorly constrained. For strike (κ) and rake (σ), we find an intermediate resolution.

The strike angle κ is constrained up to an intrinsic non-uniqueness, which is due to a duplication of moment tensors, that occurs in $\kappa\sigma h$-space. Depending on rake and dip, a rotation of the strike by 180° can result in an unchanged moment tensor for a fixed radiation pattern. Given the limited resolution in h and σ, we observe bimodal distributions in most cases for κ, as apparent from Fig. 4. In addition, this issue is addressed in Fig. 7. The two panels show different approximations of the joint distribution $p(\kappa, h|\mathbf{d})$, decomposed according to eq. (6). Note that high probability is assigned to the areas in κ–h space, that correspond to the target value (big diamond) and the theoretical duplication point (small diamond) given by $(\kappa + \pi, \sin\theta)$.

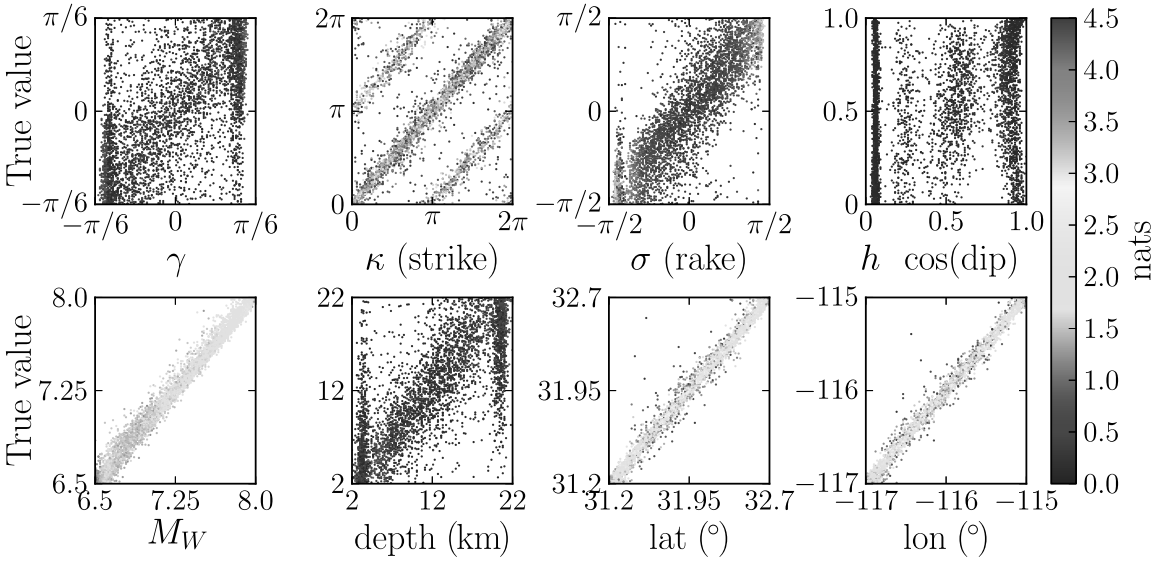

Figure 4. Position of the posterior mode plotted against the true target value for the 4000 test set examples. The colour indicates the information gain for every example according to eq. (14).

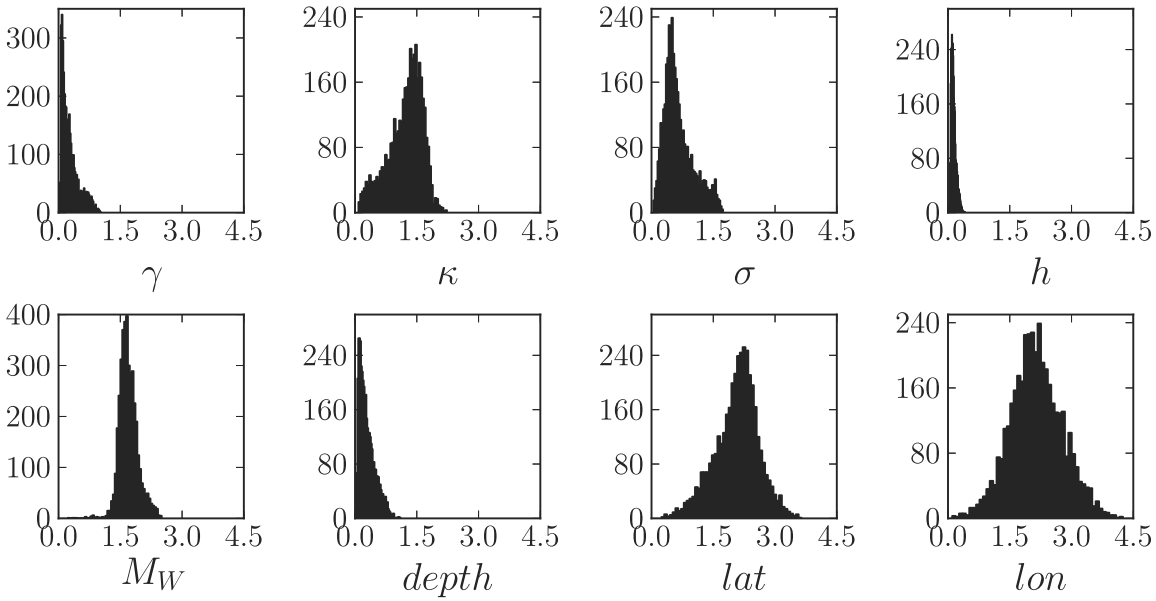

Figure 5. Information gain distributions measured in 'nats' for the 4000 test set examples (see eq. 14). The information gain is a measure for how much has been learnt about a parameter after the data vector has been presented. A higher information gain indicates that the posterior distribution is narrower than the uniform prior. See also Fig. 1 for explanatory examples.

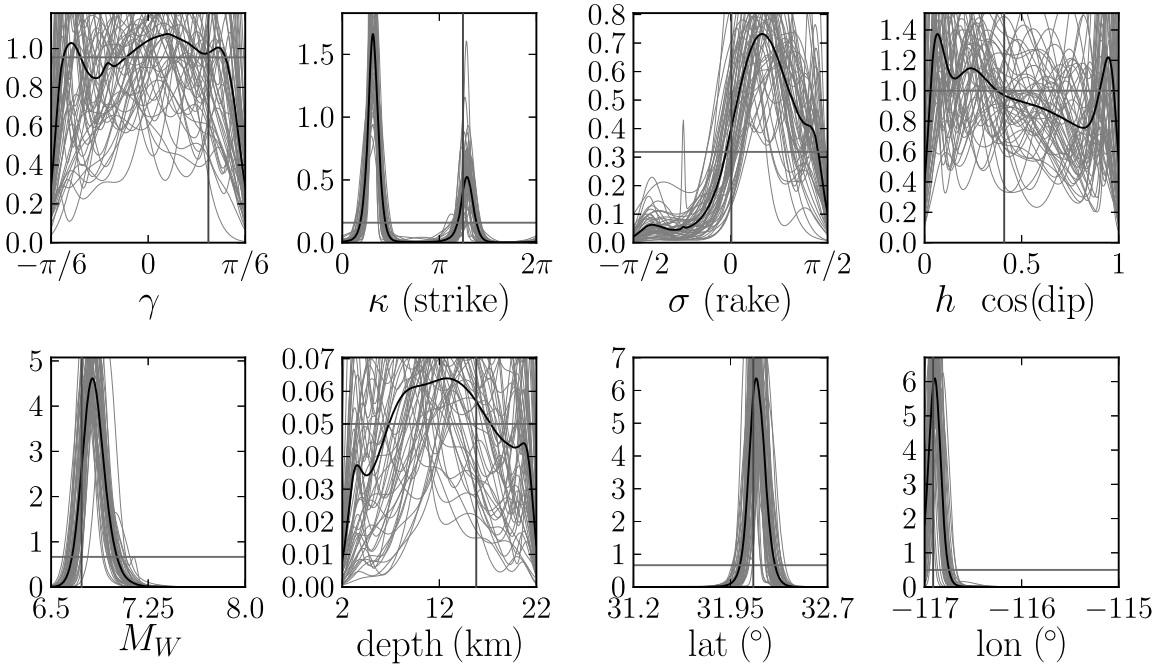

Figure 6. Posterior marginal distributions for a test set example. Shown in grey are the predictions of the individual committee members. The position of the known target value is indicated by the blue vertical line, the prior distribution is shown in green.

Finally, we note that posterior uncertainties in general seem to vary strongly across the test set, indicating a rather complex interdependence among multiple parameters and possible trade-offs that are hidden in a condensed 1-D representation of the posterior.

3.4 Inversion of the El Mayor Cucapah data set

In the previous section, we used an idealized station distribution to investigate the resolvability of source parameters using static displacement data. We will now move to a more realistic situation and use 37 stations of the CRTN network (red triangles in Fig. 3), which recorded the 2010 El Mayor Cucapah event. The prior parameter distributions and the noise model are the same as in the previous experiment.

Figs 8 and 9 again show performance characteristics for the 4000 test set examples. A comparison with Figs 4 and 5 reveals that most parameters are now less well resolved. In particular, the average information gain for the location parameters is significantly lower, as expected due to the uneven station coverage.

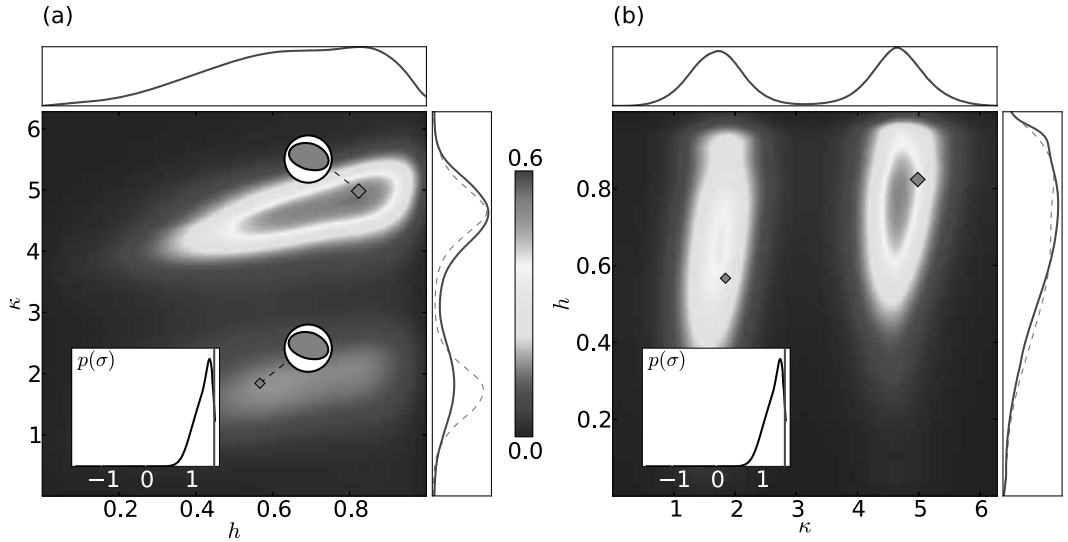

Figure 7. Duplication of moment tensors in the κ–h plane. Both panels show the probability $p(\kappa, h|\mathbf{d})$ for the same example, an event with a value of σ close to $\pi/2$. The large red diamond indicates the position of the target value, the small diamond the position of the duplicate source in the case $\sigma = \pi/2$. The inset shows $p(\sigma|\mathbf{d})$ for this example. The distribution in (a) has been constructed according to eq. (6) by decomposing $p(\kappa, h|\mathbf{d}) = p(\kappa|h, \mathbf{d})p(h|\mathbf{d})$, where $p(h|\mathbf{d})$ is shown above the figure. The right-hand panel (b) uses $p(\kappa, h|\mathbf{d}) = p(h|\kappa, \mathbf{d})p(\kappa|\mathbf{d})$. Note the flipped axes. Two network committees are thus required for each 2-D distribution—one for the conditional and one for the marginal distribution. The marginals above and right of the figure are determined by integrating the 2-D joint distribution. The output of an independent committee trained on $p(\kappa|\mathbf{d})$ (left-hand panel) and $p(h|\mathbf{d})$ (right-hand panel) serves as additional consistency check and is shown as green dashed line. The reason for the two distributions not being identical are different approximation errors in the network committees involved. However, they both show essentially the same features.

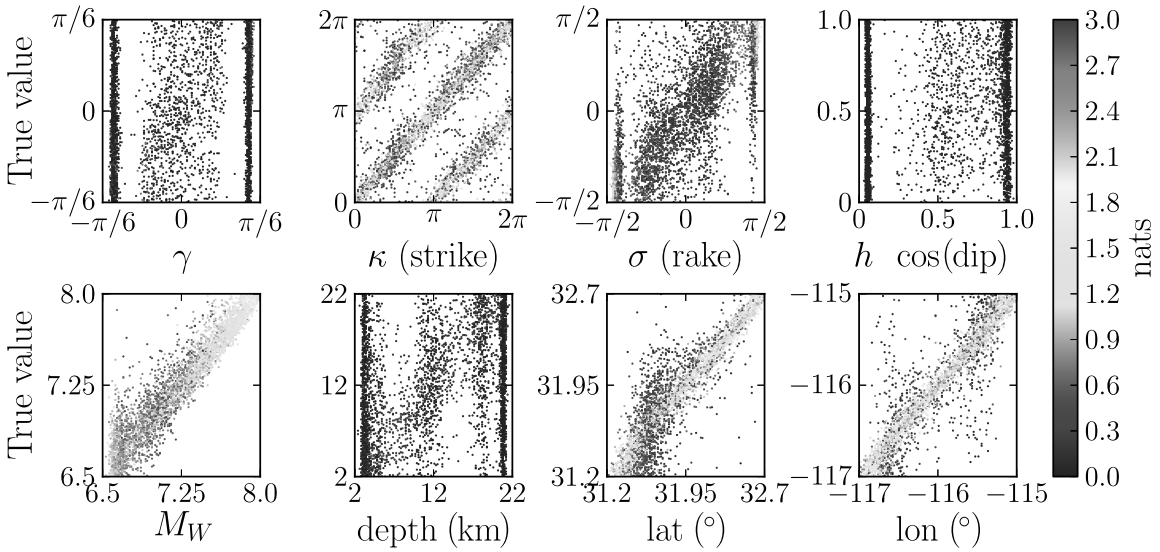

Figure 8. Same as Fig. 4, but for the 37 CRTN stations. Note the change in colour scale.

Fig. 10 shows 1-D marginals—our inversion results—obtained from presenting the network with the observed data. Vertical lines denote the position of other published solutions. The spread of published solutions is quite large, due to differing data sets, types of data, methods and prior information. However, for most parameters all deterministic solutions lie in ranges to which we assign high probability. That is, our solution does not explicitly rule out any of the other solutions for this event, with the exception of longitude from the W phase & CMT catalogue solutions (green and blue solid vertical lines, respectively, in Fig. 10). This could possibly be caused by neglected 3-D structure in our earth models, which is not averaged out due to the uneven azimuthal station coverage. Furthermore, these inversions fix depth to values $\geq$12 km, which our inversion indicates to be unrealistically deep. This possibly leads to a different location estimate due to trade-offs with depth. We further address this discrepancy in the Section 4.

We find that our solution favours values for γ close to $-\pi/6$, indicating a strong non-DC component, although this is not well resolved. This is in agreement with most other solutions. We furthermore find that we can resolve the strike κ, up to the known ambiguity (Fig. 7), rake σ, magnitude M_W and the epicentral location comparatively well, while the cosine of the dip angle (h) and the

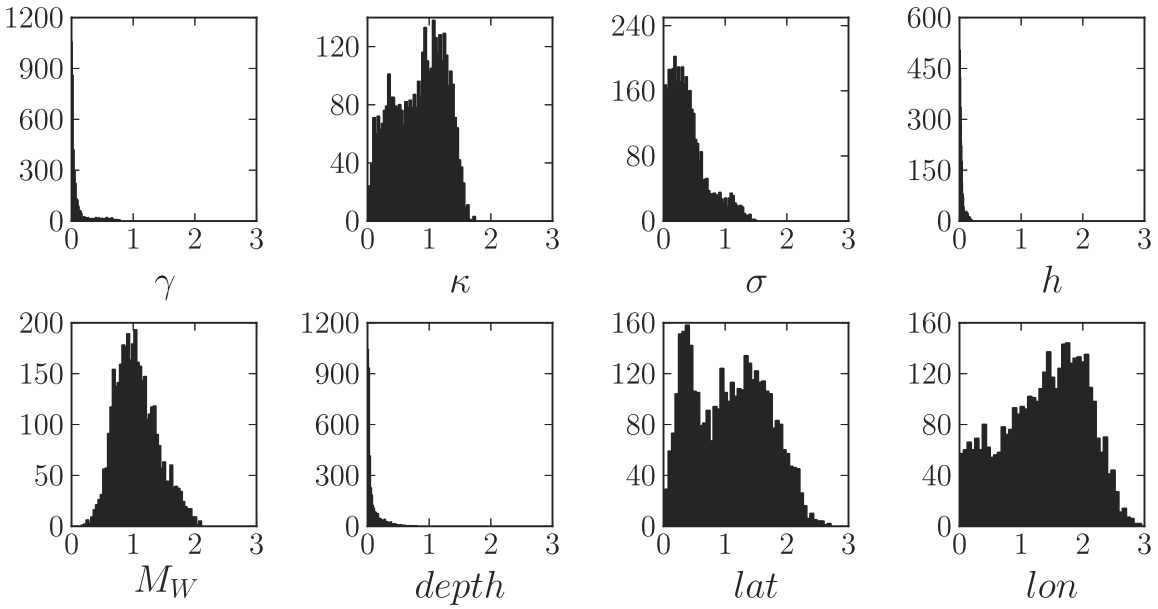

Figure 9. Same as Fig. 5, but for the 37 CRTN stations. As compared to Fig. 5, the information gain is lowered for all parameters, due to the uneven station coverage. Despite that, magnitude, latitude and longitude are still comparatively well resolved. There is a class of test set examples for which the epicentral location cannot be determined (peak at ∼0.5 nats for latitude and longitude). These examples correspond to sources that are further away from the station network, as can also be seen from Fig. 8 (bottom right-hand panels). The signal for those sources is below the noise level for most receivers.

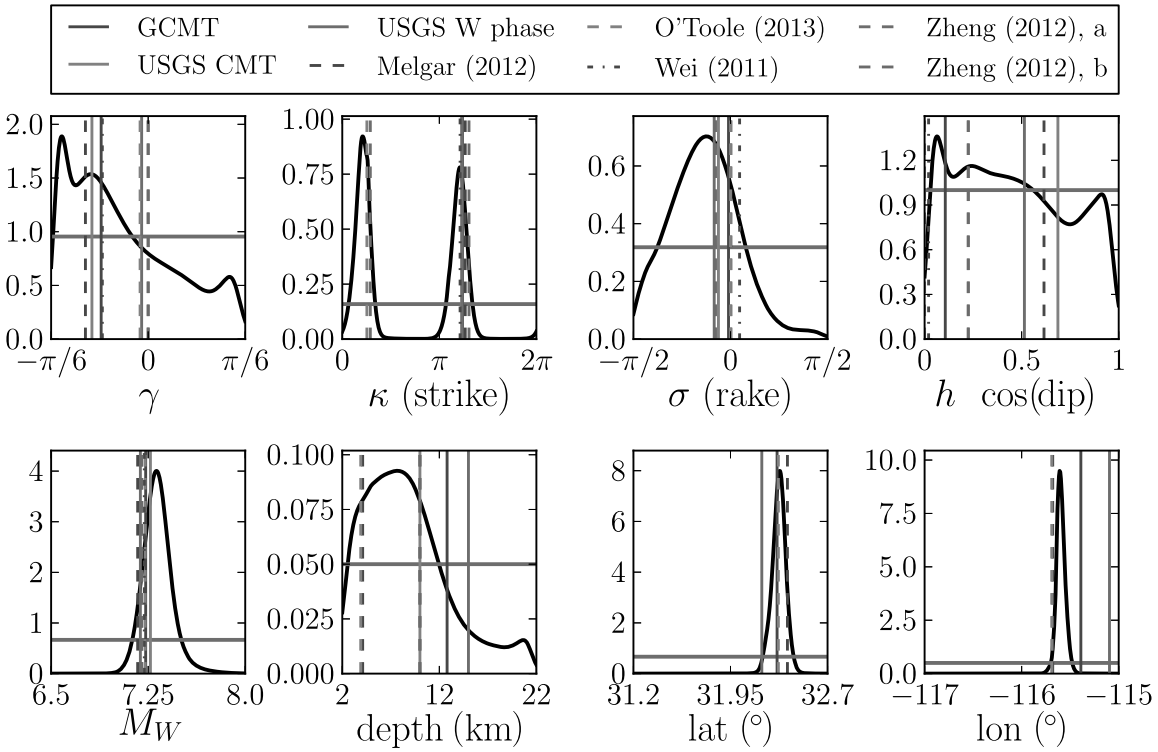

Figure 10. Posterior 1-D marginal distributions (solid black line) for the 2010 7.2 El Mayor Cucapah event. The horizontal green line is the uniform prior distribution and vertical lines correspond to other published point source solutions for this event. Solid lines hereby correspond to teleseismic body and surface waves' data sets, dashed lines to data sets using GPS displacement data and the dash-dotted line corresponds to a joint inversion of GPS and seismic data for the focal mechanism. Zheng *et al.* (2012) *a priori* assumed a double-couple and the two solutions (a and b) only differ by their respective strike angle. Note that they give solution (b) as their preferred choice based on prior information on the fault orientation.

depth are subject to large uncertainties. Note that the large probability densities assigned to the very low ends of the prior range in the case of γ and h are likely to be artefacts of the GMM-approximation (see also Section 3.3.1 and Figs 4 and 8).

4 DISCUSSION

We have shown inversion results for static displacements observed after the 2010 El Mayor Cucapah event. Wei *et al.* (2011) performed a comprehensive analysis of this event and they find slip distributions and a moment tensor description by jointly inverting several seismic and geodetic data sets. Our results are in good agreement with theirs. In particular, they find that the faulting was geometrically complex and involved strike-slip movement combined with normal faulting on several distinct subfaults, which is reflected by a large non-DC component (first panel in Fig. 10). They also find that most of the energy is released at shallow depths above 10 km, which is in agreement with our solution, which likewise prefers very shallow centroid depths (Fig. 10). Note that although some other depth solutions shown in Fig. 10 are less meaningful, since depth has been fixed prior to the inversion in these cases, a very shallow depth is also in accordance with the solutions by Melgar *et al.* (2012) and O'Toole *et al.* (2013), who both used a comparable data set. Zheng *et al.* (2012) determined the focal mechanism for this event from 5 Hz GPS displacement waveforms. They *a priori* assume a pure DC and perform a grid search for strike, dip, rake, moment and depth. They subsequently find two distinct misfit minima for the strike angle, corresponding to the two modes in our posterior distribution. These two solutions are labelled (a) and (b), respectively in Fig. 10.

A slight disagreement appears for the epicentral location. Our solution seems to rule out the GCMT and W-phase inversion results, which both suggest an event location slightly more to the east. This could be caused by a number of reasons: first, uncertainties in location for CMT-type inversions can be on the order of $\sim$50 km in lateral direction (e.g. Valentine & Trampert 2012a), which would reconcile the CMT results with our posterior pdf. Secondly, we used a layered, 1-D earth model despite the very heterogeneous southern Californian crust (Kohler *et al.* 2003), which can—in combination with a one-sided station coverage (red triangles in Fig. 3)—lead to a significant bias in the estimation of the source location. In addition, the centroid location depends on the frequency content of the respective data and we should not necessarily expect centroids obtained at different frequencies to coincide. Finally, we have not yet taken into account any effect that uncertainties in the 1-D model may have upon the location estimate. We now address this issue.

4.1 Uncertainties in the crustal model

Since we lack realistic information on uncertainties in the 1-D earth model we have not taken them into account in our inversion of the El Mayor Cucapah data set. However, it is straightforward to include modelling uncertainties—if available—into the inversion procedure. In the following we investigate the effect different amounts of earth model variation have upon the 1-D posterior marginals.

We draw horizontally layered crustal models $\mathbf{M}_i = (v_p^{(1)}, v_s^{(1)}, \rho^{(1)}, d^{(1)}, \ldots, v_p^{(4)}, v_s^{(4)}, \rho^{(4)}, d^{(4)}, v_p^{(5)}, v_s^{(5)}, \rho^{(5)})_i$, with parameters *P*-wave speed v_p, *S*-wave speed v_s, density ρ and layer thickness d, respectively, for a total of four layers above a half-space from the prior distribution

$$p(M_k) = \frac{1}{\sqrt{2\pi}\sigma_M} \exp\left\{ -\frac{(M_k - \overline{M}_k)^2}{2\sigma_{M_k}^2} \right\}. \quad (22)$$

In this expression, $\overline{\mathbf{M}}$ is the unperturbed crustal model used throughout Section 3 and the standard deviations σ_{M_k} are set to $0.05 \cdot \overline{M}_k$, $0.1 \cdot \overline{M}_k$ and $0.2 \cdot \overline{M}_k$, respectively. A few models drawn from this prior are shown in Fig. 11. Note that since the perturbations are drawn in an uncorrelated fashion the v_p/v_s ratios are also allowed to change across the training set.

We find that variations below 5 per cent do not give rise to changes in the synthetic data above the noise level, indicating that in general static displacements show little sensitivity to the 1-D crustal model structure. Fig. 12 shows the influence of the model variation on the 1-D posterior marginal distributions for different amounts of model perturbation. It reveals that the parameters governing the radiation pattern and orientation are less sensitive to model variations than the parameters governing location and magnitude, which show a clear broadening with increasing variations. Interestingly, both the latitude and magnitude distributions also show a slight shift. This could indicate that we have reached a regime in which the neural network interpolation becomes poor and the number of training samples is not sufficient to constrain the mapping well, since a large portion of the training set features very unrealistic earth models. This is a particular consequence of a poor prior distribution in a relatively high-dimensional space. Since this effect is only observed for very strong model perturbations we do not investigate this any further, but note that the results might be misleading in this particular case. The longitude estimate seems not to be affected at all, pointing to a very robust result and suggesting that the observed deviations from catalogue solutions might indeed stem from large uncertainties in the CMT inversions or from neglected 3-D structure or station distribution effects. We note, furthermore, that we do in fact expect some disagreement between the location estimates of the different solutions due to the different nature of the data sets on which they are based, and their respective frequency content.

We conclude that our estimates do not significantly change unless we impose unrealistically strong model variations of 20 per cent, suggesting that static displacement measurements can provide robust information on source parameters. Effects due to neglected 3-D structure obviously may still strongly affect the results. In order to monitor a specific seismically active region for an EEW application, ideally we would generate a training set using 3-D wave-propagation in a heterogeneous local crustal earth model. Due to the high computational demands for creating the training set, this was not feasible for this demonstration.

4.2 Trade-offs

If 1-D marginals are broad, this does not necessarily mean that the information on the corresponding parameter is limited. Relatively broad marginals can also result from dependencies between multiple parameters. We can use 2-D marginal distributions as a tool to discover these trade-offs. Fig. 13 shows the seven possible combinations of h (cosine of dip) with other parameters. The 2-D marginals reveal a slight linear trade-off with strike κ and a more complex dependency of magnitude M_W on h in a certain range. We investigate possible trade-offs of magnitude with source location in Fig. 14. A linear trade-off between magnitude and latitude is revealed, which is to be expected due to the one-sided station distribution. We can

either favour a smaller source located closer to the station network, or a larger source, further away. This trade-off might explain why our 1-D M_{W} posterior seems to suggest slightly larger magnitudes than catalogue solutions (*cf.* Fig. 10). An increase in resolution for latitude would therefore improve estimates of M_{W} as well.

4.3 Limitations of the method

The trained networks smoothly interpolate a set of training examples. This requires that the training examples cover the full range of observations that are likely to occur. Naturally the quality of the approximation strongly depends on the number of training examples, the smoothness of the inverse mapping and the intrinsic dimensionality of the problem. More samples are likely to be needed if the complexity of the mapping increases or the number of model parameters grows. We are, however, able to test the quality of the network approximation by means of a synthetic test data set. If the predictions for the test data set become unreasonable, this is an indication that a larger training set is needed. Furthermore, care must be taken to ensure that no observations are presented to the network which correspond to an event outside the prior range, since the network would then be forced to extrapolate. In practice, this might be achieved by, for example, applying an amplitude threshold and presenting only those observations whose amplitudes lie within the accepted range.

A further issue is related to the nature of the source description. Clearly, a moment tensor point source is not a sufficient physical model to fully explain the near-field deformation pattern of a big earthquake such as the El Mayor Cucapah event. In fact, Oskin *et al.* (2012) show that the total length of the highly complex multi-fault rupture was 120 km, leading to a complicated post-earthquake deformation pattern at the surface. Despite that, our results seem compatible with far-field point source solutions and we therefore believe that they can still provide a useful characterization of the source.

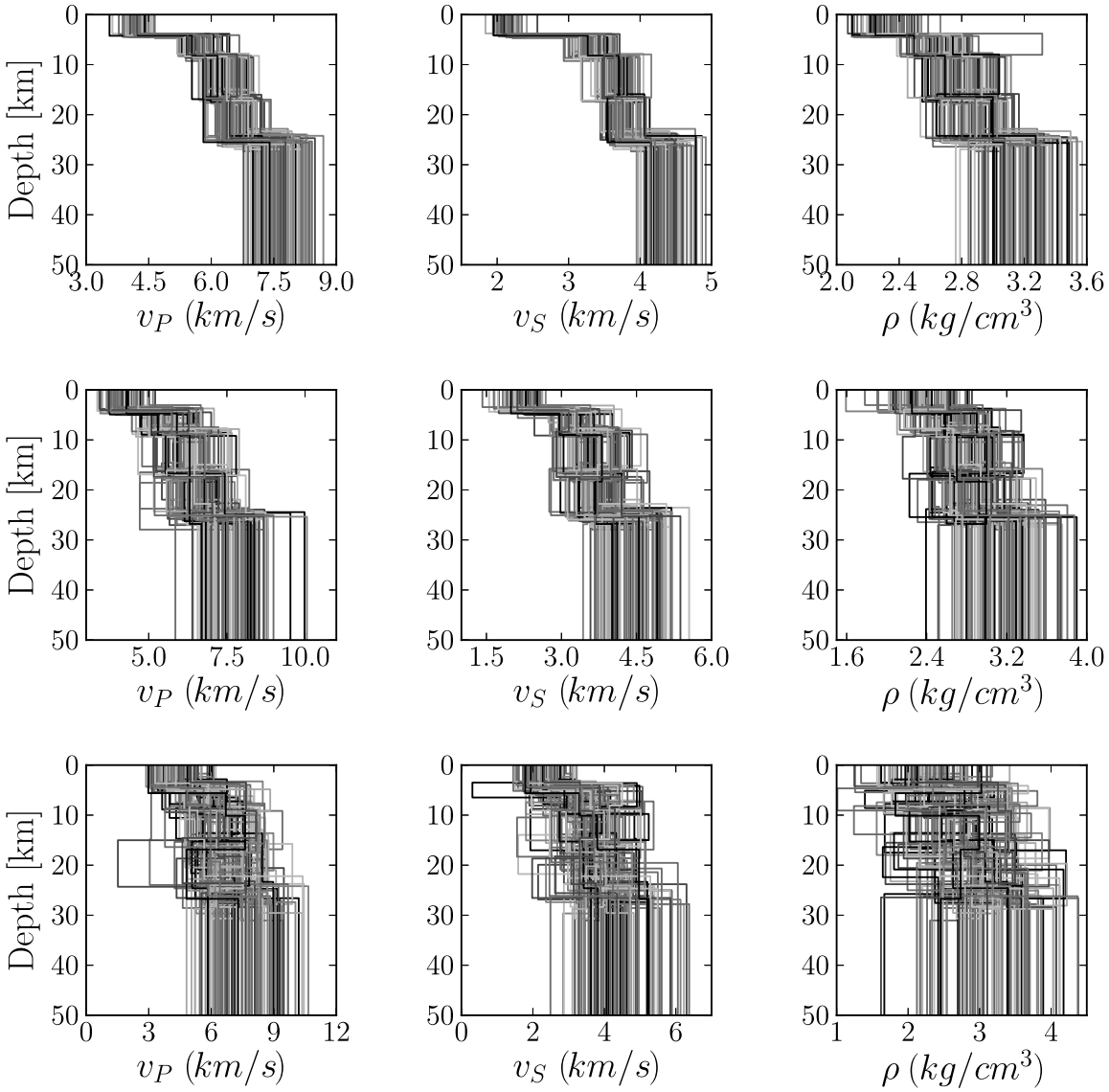

Figure 11. Each row shows 100 samples drawn from the prior earth model distribution, according to eq. (22). With standard deviations 5 per cent (top panels), 10 per cent (middle panels) and 20 per cent (bottom panels), respectively.

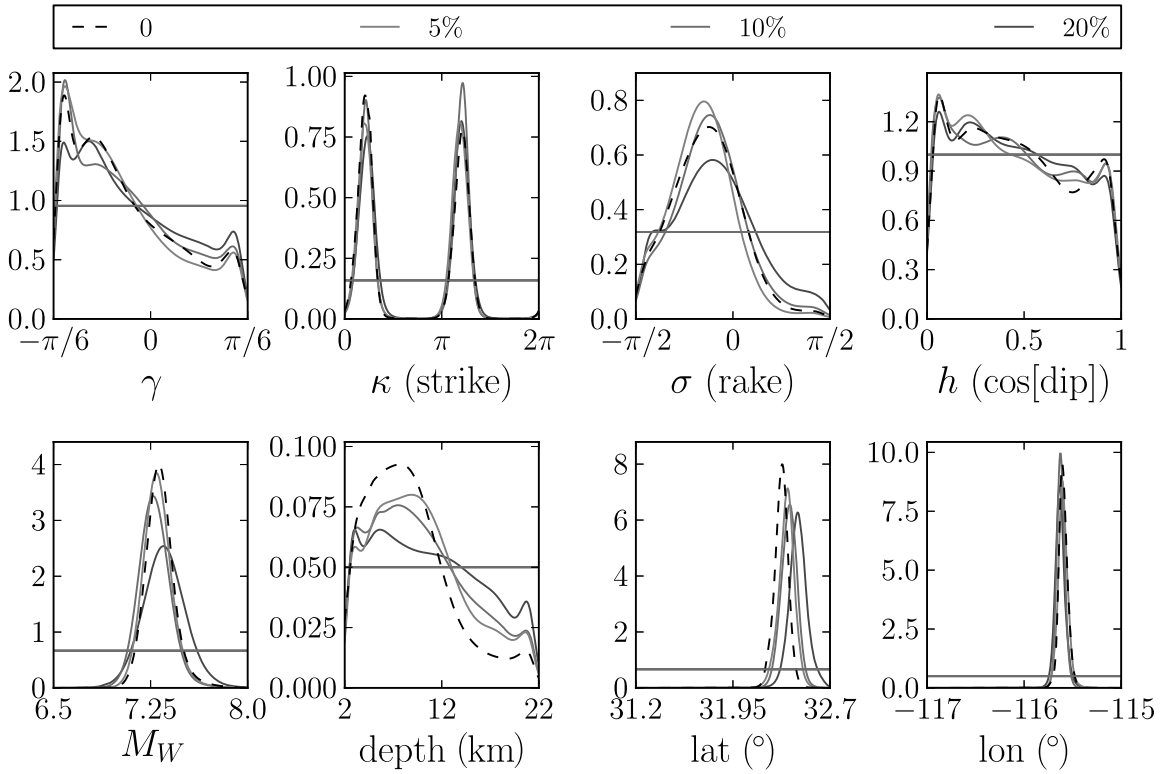

Figure 12. Influence of perturbations of the crustal model on the 1-D posterior marginal distributions for the observed datum. The four distributions correspond to the unperturbed case (dashed black, same as the solid line in Fig. 10), 5 per cent (red), 10 per cent (green) and 20 per cent (blue) variation.

5 CONCLUSIONS

We have presented a new neural network–based methodology for fast Bayesian point source inversion. We have tested and applied this method using static displacement data observed by dense GPS networks, such as the CRTN network in southern California. Our results reveal that static displacement data contain robust information on epicentral location and magnitude. Furthermore, we have shown that the observable is relatively insensitive to the 1-D crustal earth model and our parameter estimates do not change fundamentally even for very large model perturbations on the order of 20 per cent in thickness, v_p, v_s and ρ.

We inverted observed data for the 2010 M_W 7.2 El Mayor Cucapah event, confirming previously published moment tensor point source solutions for that earthquake. In particular, we found that our posterior parameter uncertainties encompass most other solutions. A comparison with CMT catalogue solutions reveals that, although derived from local data, our moment tensor solution is comparable with solutions inferred from teleseismic body waves. A slight, albeit not unexpected, disagreement appears for the centroid location, due to the biased azimuthal coverage, the differing earth models and the different frequency content of the respective data.

Unlike most deterministic methods, we are also able to invert for event depth, although this is subject to large uncertainties. We

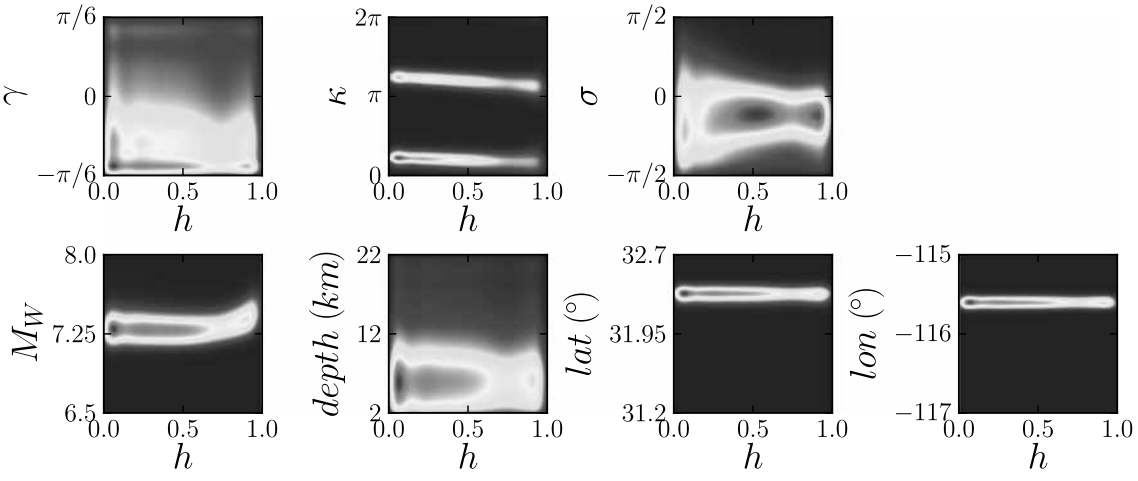

Figure 13. Shown are 2-D posterior distributions of h with all other parameters for the observed El Mayor Cucapah data set. A trade-off between h and M_W for large values of h (second row, first panel) and a slight dependency of κ with h (first row, second panel) is revealed.

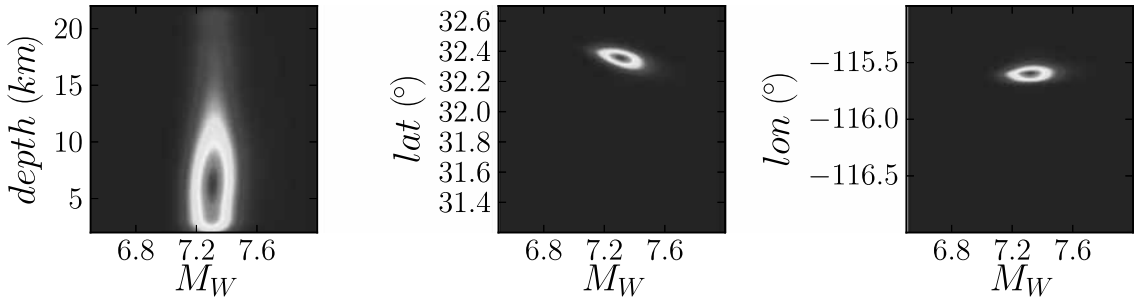

Figure 14. Shown are 2-D posterior distributions of M_W and depth, M_W and latitude and M_W and longitude, respectively, for the observed El Mayor Cucapah data set. A trade-off between latitude and M_W is revealed.

find that our solution clearly prefers a shallow source, which is in good agreement with studies that performed comprehensive finite source modelling. Our result additionally reveals a large non-DC component, pointing to a complex subsurface fault geometry.

The potential of our method lies in real-time applications, such as EEW, since it is able to rapidly invert new observations on a routine basis without additional significant computational effort. The method stores information obtained from a set of samples representative of the relation between model parameters and observable data—which may have been expensive to produce—in a compact neural network structure able to interpolate smoothly between the samples. This committee of networks can be prepared beforehand and could be kept in memory on a standard desktop computer in an EEW centre. A new observation, made by the same set of receivers for which the committees of networks have been trained, can then be inverted within milliseconds and yields approximations of 1-D and 2-D posterior marginal probability distributions on model parameters.

Due to the flexible treatment of input data, the method can readily be extended to incorporate other types of data such as real-time GPS waveforms or accelerograms from strong-motion sensors—either individually or as a joint determination. In particular, waveform data would allow us to invert for the temporal characteristics of an earthquake as well. This potential remains to be fully explored.

ACKNOWLEDGEMENTS

We thank Ralph de Wit and Diego Melgar for fruitful discussions and Carl Tape and an anonymous reviewer for their constructive and careful feedback that helped improving the manuscript. The GPS displacement time-series and coseismic offsets for the El Mayor Cucapah event have been made available by the Scripps Orbit and Permanent Array Center (SOPAC) and can be downloaded from http://geoapp03.ucsd.edu/gridsphere/gridsphere?cid=El+Mayor+Cucapah. The following software libraries have been used for this work: PyBrain (Schaul *et al.* 2010), ALGLIB (Bochkanov & Bystritsky 2012) and ObsPy (Beyreuther *et al.* 2010).

REFERENCES

Allen, R.M. & Ziv, A., 2011. Application of real-time GPS to earthquake early warning, *Geophys. Res. Lett.,* **38**(16), 1–7.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y. & Wassermann, J., 2010. ObsPy: a Python toolbox for seismology, *Seism. Res. Lett.,* **81**(3), 530–533.

Bishop, C. & Legleye, C., 1994. Estimating conditional probability densities for periodic variables, in *Advances in Neural Information Processing System 7,* pp. 641–648, MIT Press.

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition,* Vol. 92, Oxford University Press.

Blewitt, G., 2007. GPS and space-based geodetic methods, in *Treatise on Geophysics,* Elsevier.

Bochkanov, S. & Bystritsky, V., 2012. ALGLIB. Available at: www.alglib.net, last accessed 1 November 2012.

Bock, Y., Melgar, D. & Crowell, B., 2011. Real-time strong-motion broadband displacements from collocated GPS and accelerometers, *Bull. seism. Soc. Am.,* **101**(6), 2904–2925.

Carney, M., Cunningham, P., Dowling, J. & Lee, C., 2005. Predicting probability distributions for surf height using an ensemble of mixture density networks, in *Proceedings of the 22nd International Conference on Machine learning - ICML '05,* pp. 113–120.

Chapman, C.H. & Leaney, W.S., 2012. A new moment-tensor decomposition for seismic events in anisotropic media, *Geophys. J. Int.,* **188**(1), 343–370.

Cornford, D., Nabney, I.T. & Bishop, C.M., 1999. Neural network-based wind vector retrieval from satellite scatterometer data, *Neural Comput. Appl.,* **8**(3), 206–217.

Crowell, B.W., Bock, Y. & Squibb, M.B., 2009. Demonstration of earthquake early warning using total displacement waveforms from real-time GPS networks, *Seism. Res. Lett.,* **80**(5), 772–782.

Devilee, R.J.R., Curtis, A. & Roy-Chowdhury, K., 1999. An efficient, probabilistic neural network approach to solving inverse problems: inverting surface wave velocities for Eurasian crustal thickness, *J. geophys. Res.,* **104**(B12), 28 841–28 857.

Duputel, Z., Rivera, L., Kanamori, H., Hayes, G.P., Hirshorn, B. & Weinstein, S., 2011. Real-time W phase inversion during the 2011 off the Pacific coast of Tohoku earthquake, *Earth Planets Space,* **63**(7), 535–539.

Duputel, Z., Rivera, L., Fukahata, Y. & Kanamori, H., 2012. Uncertainty estimations for seismic source inversions, *Geophys. J. Int.,* **190**(2), 1243–1256.

Dziewonski, A.M., Chou, T.-A. & Woodhouse, J.H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. geophys. Res.,* **86**(B4), 2825–2852.

Ekström, G., Nettles, M. & Dziewonski, A.M., 2012. The global CMT project 2004–2010: centroid-moment tensors for 13,017 earthquakes, *Phys. Earth planet. Inter.,* **200–201,** 1–9.

Giardini, D., 1984. Systematic analysis of deep seismicity: 200 centroid-moment tensor solutions for earthquakes between 1977 and 1980, *Geophys. J. Int.,* **77**(3), 883–914.

Hanks, T. & Kanamori, H., 1979. A moment magnitude scale, *J. geophys. Res.: Solid Earth,* **84**(B5), 2348–2350.

Hauksson, E., Stock, J., Hutton, K., Yang, W., Vidal-Villegas, J.A. & Kanamori, H., 2010. The 2010 Mw 7.2 El Mayor-Cucapah earthquake sequence, Baja California, Mexico and southernmost California, USA: active seismotectonics along the Mexican Pacific margin, *Pure appl. Geophys.,* **168**(8–9), 1255–1277.

Hornik, K., Stinchcombe, M. & White, H., 1989. Multilayer feedforward networks are universal approximators, *Neural Netw.,* **2**(5), 359–366.

Kanamori, H., 1993. W phase, *Geophys. Res. Lett.,* **20**(16), 1691–1694.

Kanamori, H. & Rivera, L., 2008. Source inversion of W phase: speeding up seismic tsunami warning, *Geophys. J. Int.,* **175**(1), 222–238.

Kohler, M., Magistrale, H. & Clayton, R., 2003. Mantle heterogeneities and the SCEC reference three-dimensional seismic velocity model version 3, *Bull. seism. Soc. Am.,* **93**(2), 757–774.
Kolmogorov, A.K., 1957. On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition, *Dokl. Akad. Nauk SSSR,* **114,** 369–373.
Larson, K.M., Bodin, P. & Gomberg, J., 2003. Using 1-Hz GPS data to measure deformations caused by the Denali fault earthquake, *Science,* **300**(5624), 1421–1424.
MacKay, D., 2003. *Information Theory, Inference and Learning Algorithms,* Cambridge Univ. Press.
Madariaga, R., 2007. Seismic source theory, in *Treatise on Geophysics,* Vol. 4, pp. 59–82, Elsevier.
McLachlan, G.J. & Basford, K.E., 1988. *Mixture Models: Inference and Applications to Clustering,* Statistics: Textbooks and Monographs 84, Marcel Dekker.
Meier, U., Curtis, A. & Trampert, J., 2007a. Global crustal thickness from neural network inversion of surface wave data, *Geophys. J. Int.,* **169**(2), 706–722.
Meier, U., Curtis, A. & Trampert, J., 2007b. Fully nonlinear inversion of fundamental mode surface waves for a global crustal model, *Geophys. Res. Lett.,* **34**(16), 1–6.
Meier, U., Trampert, J. & Curtis, A., 2009. Global variations of temperature and water content in the mantle transition zone from higher mode surface waves, *Earth planet. Sci. Lett.,* **282**(1–4), 91–101.
Melgar, D., Bock, Y. & Crowell, B.W., 2012. Real-time centroid moment tensor determination for large earthquakes from local and regional displacement records, *Geophys. J. Int.,* **188**(2), 703–718.
Melgar, D., Crowell, B.W., Bock, Y. & Haase, J.S., 2013. Rapid modeling of the 2011 Mw 9.0 Tohoku-oki earthquake with seismogeodesy, *Geophys. Res. Lett.,* **40**(12), 2963–2968.
Nikolaidis, R., 2002. Observation of geodetic and seismic deformation with the Global Positioning System, *PhD thesis,* University of California, San Diego.
Nocedal, J., 1980. Updating quasi-Newton matrices with limited storage, *Math. Comput.,* **35**(151), 773–782.
Okada, Y., 1985. Surface deformation due to shear and tensile faults in a half-space, *Bull. seism. Soc. Am.,* **75**(4), 1135–1154.
Oskin, M.E. *et al.*, 2012. Near-field deformation from the El Mayor-Cucapah earthquake revealed by differential LIDAR, *Science,* **335**(6069), 702–705.
O'Toole, T.B. & Woodhouse, J.H., 2011. Numerically stable computation of complete synthetic seismograms including the static displacement in plane layered media, *Geophys. J. Int.,* **187**(3), 1516–1536.
O'Toole, T.B., Valentine, A.P. & Woodhouse, J.H., 2012. Centroid-moment tensor inversions using high-rate GPS waveforms, *Geophys. J. Int.,* **191**(1), 257–270.
O'Toole, T.B., Valentine, A.P. & Woodhouse, J., 2013. Earthquake source parameters from GPS-measured static displacements with potential for real-time application, *Geophys. Res. Lett.,* **40,** 60–65.
Röth, G. & Tarantola, A., 1994. Neural networks and inversion of seismic data, *J. geophys. Res.,* **99**(B4), 6753–6768.
Rumelhart, D., Hinton, G. & Williams, R., 1986. Learning representations by back-propagating errors, *Nature,* **323,** 533–536.
Sambridge, M., 1999a. Geophysical inversion with a neighbourhood algorithm—I. Searching a parameter space, *Geophys. J. Int.,* **138**(2), 479–494.
Sambridge, M., 1999b. Geophysical inversion with a neighbourhood algorithm—II. Appraising the ensemble, *Geophys. J. Int.,* **138**(3), 727–746.
Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., Rückstieß, T. & Schmidhuber, J., 2010. PyBrain, *J. Mach. Learn. Res.,* **11,** 743–746.
Shahraeeni, M.S. & Curtis, A., 2011. Fast probabilistic nonlinear petrophysical inversion, *Geophysics,* **76**(2), E45–E58.
Shahraeeni, M.S., Curtis, A. & Chao, G., 2012. Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data, *Geophysics,* **77**(3), O1–O19.
Shearer, P.M., 1999. *Introduction to Seismology,* Cambridge Univ. Press.
Tape, W. & Tape, C., 2012. A geometric setting for moment tensors, *Geophys. J. Int.,* **190**(1), 476–498.
Tarantola, A., 2005. *Inverse Problem Theory,* 4, SIAM.
Valentine, A.P. & Trampert, J., 2012a. Data space reduction, quality assessment and searching of seismograms: autoencoder networks for waveform data, *Geophys. J. Int.,* **189**(2), 1183–1202.
Valentine, A.P. & Trampert, J., 2012b. Assessing the uncertainties on seismic source parameters: towards realistic error estimates for centroid-moment-tensor determinations, *Phys. Earth planet. Inter.,* **210–211,** 36–49.
Valentine, A.P. & Woodhouse, J.H., 2010. Approaches to automated data selection for global seismic tomography, *Geophys. J. Int.,* **182**(2), 1001–1012.
Wang, R., Parolai, S., Ge, M., Jin, M., Walter, T.R. & Zschau, J., 2013. The 2011 Mw 9.0 Tohoku earthquake: comparison of GPS and strong-motion data, *Bull. seism. Soc. Am.,* **103**(2B), 1336–1347.
Wei, S. *et al.*, 2011. Superficial simplicity of the 2010 El Mayor–Cucapah earthquake of Baja California in Mexico, *Nature Geosci.,* **4**(9), 615–618.
de Wit, R.W., Valentine, A.P. & Trampert, J., 2013. Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks, *Geophys. J. Int.,* **195**(1), 408–422.
Zheng, Y., Li, J., Xie, Z. & Ritzwoller, M.H., 2012. 5Hz GPS seismology of the El Mayor-Cucapah earthquake: estimating the earthquake focal mechanism, *Geophys. J. Int.,* **190**(3), 1723–1732.

APPENDIX A: PROBABILITY DENSITY ESTIMATION USING NEURAL NETWORKS

A1 The mixture density network (MDN)

In the following, we give analytical expressions for the outputs of the two-layer feed-forward network architecture depicted in Fig. 2 and relate them to the parameters of the Gaussian mixture model (GMM) described in the main text. An extensive description is given by Bishop (1995).

A feed-forward neural network consists of multiple layers of computational units, which apply a possibly non-linear, scalar activation function to their input, a weighted linear combination of all inbound connections. We use a two-layer structure and refer to the first layer units as 'hidden units' with outputs h_j, while the second layer units yield the network outputs y_k. The hidden unit activations are thus given by

$$h_j = f\left(\sum_{i=1}^{I} w_{ij}^{(1)} d_i + w_{0j}^{(1)}\right) \tag{A1}$$

and the output layer activations by

$$y_k = g\left(\sum_{j=1}^{H} w_{jk}^{(2)} h_j + w_{0k}^{(2)}\right), \tag{A2}$$

with input and hidden layer weight matrices $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, respectively, containing the free parameters of the network. The rows $w_{0j}^{(1)}$ and $w_{0k}^{(2)}$ are referred to as 'biases'. We combine the input and hidden layer weight matrix into one weight matrix $\mathbf{w}$ in the main text for notational brevity. As a non-linearity in the hidden layer we use the hyperbolic tangent $f(a) = \tanh(a)$ whereas in the output layer the identity function $g(a) = a$ suffices. Such a network has been proven to form a general function approximator (Hornik *et al.* 1989; Bishop 1995).

The network outputs y_k are now related to the parameters of a GMM (eq. 7) as follows. For a GMM with M kernels, our network possesses $O = 3M$ output units $y_m^{(\alpha)}$, $y_m^{(\mu)}$ and $y_m^{(\sigma)}$, which are related to the M mixture coefficients α_m, means μ_m and standard deviations σ_m, respectively. We have to impose additional constraints to the otherwise unbounded network outputs y_k to be usable as GMM

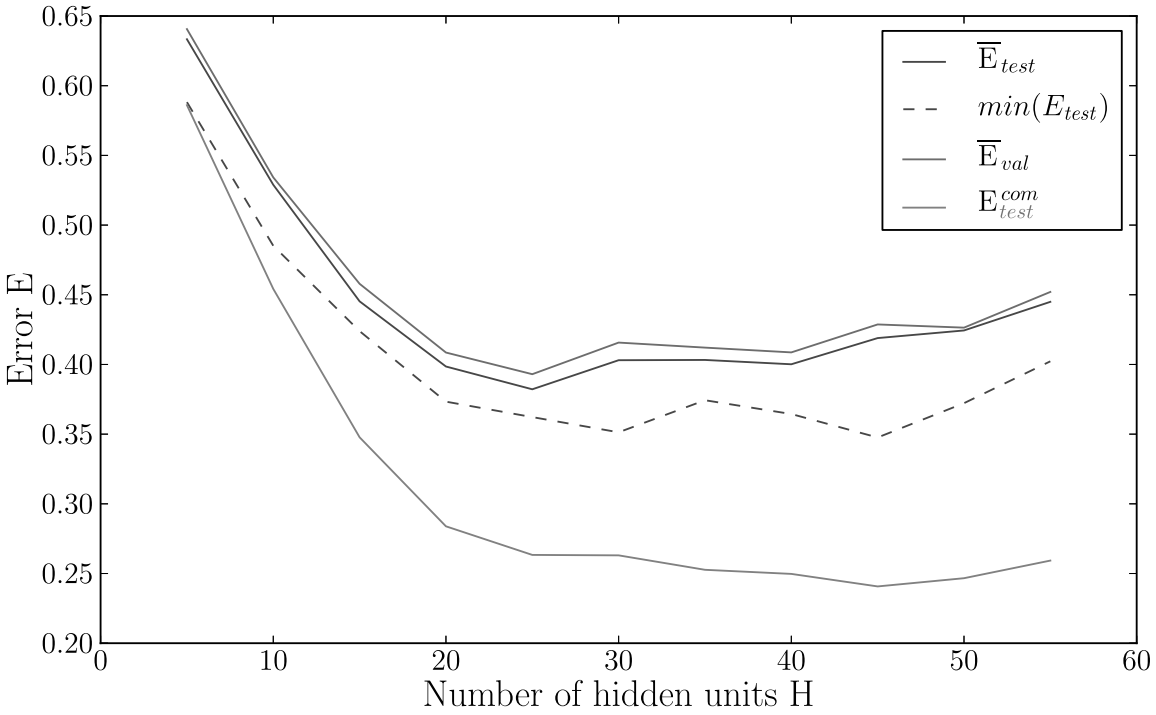

Figure A1. Dependence of the error on the number of hidden units for networks trained on $p(\text{depth}|\mathbf{d})$. The blue and green line show the test and validation set error, respectively, for different choices for the number of hidden units. The error is the average of 10 networks, independently trained and initialized using different random realizations of the weight vector $\mathbf{w}_0$. The dashed blue line is the error of the network that performed best for the respective number of hidden units. Note that the average error shows a minimum at ∼25 hidden units for this particular parameter. The increase of the error for larger networks is due to overfitting. The red line corresponds to committees formed from these 10 independent runs. Note that the committee error stays on approximately the same level for overcomplex members and is significantly lower than the error of the best committee member (dashed blue line).

parameters. The mixture coefficients α_m are required to sum to one and are thus retrieved by applying a 'softmax' function to the raw network outputs

$$\alpha_m = \frac{\exp\left(y_m^{(\alpha)}\right)}{\sum_{m=1}^{M} \exp\left(y_m^{(\alpha)}\right)}. \tag{A3}$$

The standard deviations have to be strictly positive, which is achieved by

$$\sigma_m = \exp\left(y_m^{(\sigma)}\right) \tag{A4}$$

and the output for the means can be used directly

$$\mu_m = y_m^{(\mu)}. \tag{A5}$$

A2 Network initialization

Following Bishop (1995), we draw the initial weight configuration $\mathbf{w}^0$ from a normal distribution

$$\mathbf{w}^0 \sim \mathcal{N}\left(0, \frac{1}{d_h}\right), \tag{A6}$$

where d_h is the number of connections feeding into a unit of layer h (see Fig. 2). The output layer biases $w_{0k}^{(2)}$ are furthermore initialized in such a way that the network initially outputs the prior distribution $p(m_k)$ of the model parameter m_k, independent of the input vector $\mathbf{d}$. Therefore, a GMM with M components is initially fitted to $p(m_k)$ using a k-means clustering algorithm (e.g. MacKay 2003). The resulting means, variances and mixing coefficients are successively used to initialize the biases.

A3 Regularization and complexity of the network model

As in any regression algorithm, we face a problem known as bias–variance trade-off. For a given set of training data, we can find relatively parsimonious network models that can explain the data on average but fail to capture the more detailed structure. Alternatively, we can find more complex models that reproduce the training set well, but do not interpolate smoothly between data points. The optimal solution is thus one that is complex enough to capture the general characteristics of the underlying function, but does not learn the particular realizations of the noise vector $\boldsymbol{\epsilon}$, a phenomenon called 'overfitting'. In order to find an optimal balance, we impose two types of regularization.

First, as in eq. (3), we add random Gaussian noise to the synthetic inputs corresponding to the expected noise in the observed data. Bishop (1995) shows that adding uncorrelated random noise to the network inputs during training is equivalent to a Tikhonov regularization term in the error function (12) with a coefficient proportional to the variance of the random noise component (see also Meier *et al.* 2007a). Secondly, we monitor the error of an independent validation data set $\mathcal{D}_{\text{val}}$ at each iteration during training and keep the set of weights that minimizes $E[\mathcal{D}_{\text{val}}]$. This procedure is referred to as 'early-stopping' (Bishop 1995).

A further parameter affecting the model complexity is the number of hidden units in the network (see Fig. 2 and Appendix A1). Typically, the validation set error shows a minimum or plateau as soon as the model is complex enough to explain the training data. If the model becomes too complex, however, networks start overfitting and the validation set error increases. For an example, see Fig. A1, where the green line shows the validation set error

for networks trained on $p(\text{depth}|\mathbf{d})$ averaged over 10 independent training runs each, for different choices of H. We found a similar behaviour for all model parameters, with a slight preference for more complex models for the networks trained on latitude and longitude.

A4 Committees of networks

We observe that predictions of a certain parameter may depend on the starting point $\mathbf{w}^0$ in weight space of the training algorithm. This is particularly the case if the observed data contains only weak information on that parameter. One solution to this problem lies in limiting the model complexity and imposing stronger regularization constraints. However, the amount of regularization is based on the expected noise in the data. Modifying it would thus implicitly alter our prior observational noise estimate.

Instead we can utilize the variability of the individual networks to improve the overall prediction performance, by combining the different solutions into a committee of networks. By averaging over multiple high-likelihood models we take into account the non-uniqueness of the network model, which we neglected when replacing the integral in (11) with a single set of optimal weights $\mathbf{w}^*$. Bishop (1995) shows that the error of such a committee is bounded above by the average member error. Committees of MDNs have previously been used by, for example, Cornford *et al.* (1999) and Carney *et al.* (2005).

We thus write

$$p(m_k|\mathbf{d}) = \sum_{i=1}^{C} \frac{\omega_i}{\sum_j \omega_j} p(m_k|\mathbf{d}, \mathbf{w}_i^*), \tag{A7}$$

where C is the number of committee members and $\mathbf{w}_i^*$ denotes the set of weights of the ith member. Each member's contribution is weighted by a factor of

$$\omega_i = \exp\left\{-\frac{E[\mathcal{D}_{\text{test}}, \mathbf{w}_i^*]}{N}\right\}, \tag{A8}$$

with $E[\mathcal{D}_{\text{test}}, \mathbf{w}_i^*]$ being the error (eq. 12) of the independent test set $\mathcal{D}_{\text{test}}$ for the ith member and N the number of examples in $\mathcal{D}_{\text{test}}$.

Note that the resulting distribution is still a GMM, now with $C \cdot M$ Gaussian kernels and mixture coefficients given by

$$\beta_{(M\cdot i+m)} = \frac{\omega_i}{\sum_j \omega_j}(\alpha_m)_i, \tag{A9}$$

where $(\alpha_m)_i$ is the mixing coefficient for the mth kernel of the ith committee member. This somewhat empirical approach proves to be reasonable, since the committee prediction leads to significantly lower test set errors than the errors of most of its individual members. See Fig. A1 for networks trained on $p(\text{depth}|\mathbf{d})$. The red line corresponds to the test set error of a committee formed from 10 independent networks, each trained starting from a different random point $\mathbf{w}_0$ in weight space. The test set error varies between the different networks and the average error $\overline{E}_{\text{test}}$ (shown as a solid blue line in Fig. A1) is much higher than the error $E_{\text{test}}^{\text{com}}$ of the committee. Note that in this particular case the committee error is even lower than the error of its best member (dashed blue line in Fig. A1) for this parameter. Committees trained on other parameters behave similarly.

APPENDIX B: DATA SET PRE-PROCESSING

B1 Standardizing

The input vectors of the training set $\{\mathbf{d}_i\}_{tr}$ are standardized according to the following transformation to have zero mean and standard deviation one. This increases the convergence speed of network training significantly.

$$\tilde{d}_k = \frac{d_k - \overline{d_k}}{s_k}, \tag{B1}$$

with the sample mean

$$\overline{d_k} = \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} (d_k)_i \tag{B2}$$

and variance

$$s_k^2 = \frac{1}{N_{\text{tr}} - 1} \sum_{i=1}^{N_{\text{tr}}} \left[(d_k)_i - \overline{d_k}\right]^2. \tag{B3}$$