Geophysical Journal International

Geophys. J. Int. (2013) **195,** 408–422
Advance Access publication 2013 June 26

doi: 10.1093/gji/ggt220

Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks

Ralph W. L. de Wit, Andrew P. Valentine and Jeannot Trampert

Department of Earth Sciences, Utrecht University, Budapestlaan 4, 3584 *CD, Utrecht, the Netherlands. E-mail: r.w.l.dewit@uu.nl*

Accepted 2013 May 29. Received 2013 May 15; in original form 2013 February 14

SUMMARY

How do body-wave traveltimes constrain the Earth's radial (1-D) seismic structure? Existing 1-D seismological models underpin 3-D seismic tomography and earthquake location algorithms. It is therefore crucial to assess the quality of such 1-D models, yet quantifying uncertainties in seismological models is challenging and thus often ignored. Ideally, quality assessment should be an integral part of the inverse method. Our aim in this study is twofold: (i) we show how to solve a general Bayesian non-linear inverse problem and quantify model uncertainties, and (ii) we investigate the constraint on spherically symmetric *P*-wave velocity (V_P) structure provided by body-wave traveltimes from the EHB bulletin (phases *Pn*, *P*, *PP* and *PKP*). Our approach is based on artificial neural networks, which are very common in pattern recognition problems and can be used to approximate an arbitrary function. We use a Mixture Density Network to obtain 1-D marginal posterior probability density functions (pdfs), which provide a quantitative description of our knowledge on the individual Earth parameters. No linearization or model damping is required, which allows us to infer a model which is constrained purely by the data.

We present 1-D marginal posterior pdfs for the 22 V_P parameters and seven discontinuity depths in our model. *P*-wave velocities in the inner core, outer core and lower mantle are resolved well, with standard deviations of ∼0.2 to 1 per cent with respect to the mean of the posterior pdfs. The maximum likelihoods of V_P are in general similar to the corresponding *ak135* values, which lie within one or two standard deviations from the posterior means, thus providing an independent validation of *ak135* in this part of the radial model. Conversely, the data contain little or no information on *P*-wave velocity in the D″ layer, the upper mantle and the homogeneous crustal layers. Further, the data do not constrain the depth of the discontinuities in our model. Using additional phases available in the ISC bulletin, such as *PcP*, *PKKP* and the converted phases *SP* and *ScP*, may enhance the resolvability of these parameters. Finally, we show how the method can be extended to obtain a posterior pdf for a multidimensional model space. This enables us to investigate correlations between model parameters.

Key words: Neural networks, fuzzy logic; Inverse theory; Probability distributions; Body waves; Seismic tomography; Statistical seismology.

GJI Seismology

1 INTRODUCTION

Since the start of the 20th century, the illumination of the Earth's interior by seismic waves has enabled seismologists to infer its seismic velocity and density structure. Current 3-D tomographic models show structural variations in great detail (see e.g. Nolet 2008; Rawlinson *et al.* 2010, for an overview). Such tomographic inversions are often built upon radial (1-D) earth models. The quality of a 3-D tomographic model is thus intrinsically linked to the robustness of a 1-D model; therefore, it is crucial to assess the quality of the latter. Further, seismological models are frequently used to determine earthquake locations. The spherically symmetric *ak135* model (Kennett *et al.* 1995), for instance, is used in the location algorithm of the International Seismological Centre (ISC). However, any imperfections in the earth model will map into the source location estimate (e.g. Valentine & Trampert 2012b). Clearly, an accurate estimation of seismic source parameters requires a precise knowledge of the underlying earth model and its uncertainties. However, determining the quality of earth models is non-trivial.

In many geophysical inverse problems, a single 'optimal' solution is obtained via a linearized approach (e.g. Parker 1994; Tarantola 2005). In reality, the dependence of the data on the model often is non-linear. Further, not all model parameters are equally resolved by the data, and seismological inversions usually suffer from a strong model non-uniqueness (e.g. de Wit *et al.* 2012). Therefore, it is important to understand the uncertainties and resolution associated

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

with the one 'optimal' model. Neglecting these is likely to lead to flaws in any interpretation of the final model. Various approaches are available to assess model quality. For instance, Kennett *et al.* (1995) used a non-linear search procedure to assess the robustness of the spherically symmetric *ak135* model in the lower mantle and core, but the velocity bounds for the search procedure itself were based on the final model and relatively narrow, that is, within 0.5 per cent from *ak135*. de Wit *et al.* (2012) showed how to explore the model null-space to investigate the non-uniqueness of a 3-D tomographic model. Alternatively, a resolution analysis is often employed to investigate the robustness of the inferred Earth structure, using for instance the linear framework provided by Backus & Gilbert (1968, 1970). For example, the resolution of 1-D density structure, as determined from normal mode data, was investigated using both linear (Masters & Gubbins 2003) and non-linear techniques (Kennett 1998). However, it would be better to take the non-linearity and model non-uniqueness into account in our inversion framework, rather than treating them *ex post facto*.

A more general approach, which allows us to solve a non-linear inverse problem and to quantify uncertainties, involves the description of our knowledge about earth parameters by probability distributions (e.g. Tarantola & Valette 1982; Tarantola 2005). Following Bayes' theorem (Bayes 1763), our posterior knowledge of the model is our prior knowledge updated by the observed data, using a physical theory that relates the model to the data. The aim of Bayesian inference is to obtain the posterior probability distribution $\sigma(\mathbf{m}|\mathbf{d})$ of the model $\mathbf{m}$, conditioned on the observed data $\mathbf{d}$. A common approach is to directly sample the posterior model probability density using Monte Carlo techniques (e.g. Mosegaard & Tarantola 1995; Sambridge 1999a,b; Resovsky & Trampert 2003; Käufl *et al.* 2013). Beghein *et al.* (2006) constructed probability density functions (pdfs) to assess whether radial anisotropy in existing 1-D mantle models is robust. While such sampling methods are powerful for solving non-linear inverse problems, they quickly deteriorate as the dimension of the model space increases, a phenomenon which Bellman (1961) termed the curse of dimensionality. In practice, this currently limits the use of sampling methods to inverse problems which involve at most a few tens of model parameters.

As an alternative to Monte Carlo techniques, we use artificial neural networks to solve the Bayesian inverse problem. Neural networks can be viewed as non-linear filters and are very common in pattern recognition problems. They can approximate an arbitrary non-linear relation between two parameter spaces, inferring the mapping from a set of training data. As such, neural networks can be very useful in situations where the forward relation is known, but the inverse mapping is unknown or difficult to establish by more conventional analytical or numerical methods. This situation applies to many geophysical inverse problems. In addition, neural networks can interpolate between available model samples, as opposed to conventional Monte Carlo methods, which only sample the model space discretely. This helps to address the dimensionality issue mentioned above. Common references on artificial neural networks include Bishop (1995) and MacKay (2003).

Neural networks are widely applied in many different research areas, such as finance, medicine and engineering. Applications include bankruptcy risk predictions (e.g. Odom & Sharda 2002), breast cancer detection (e.g. Baker *et al.* 1995), face recognition (e.g. Rowley 1998), landslide susceptibility estimation (e.g. Lee *et al.* 1998) and traffic flow forecasting (e.g. Jiang & Adeli 2005). Extensive reviews of geophysical applications of neural networks are given by van der Baan & Jutten (2000) and Poulton (2002). Recent examples include Devilee *et al.* (1999) and Meier *et al.* (2007a, 2009), who invert surface wave phase velocities for (local) Earth structure, and Shahraeeni & Curtis (2011) and Shahraeeni *et al.* (2012), who infer petrophysical parameters from seismic velocity data on the reservoir scale. Valentine & Woodhouse (2010) use neural networks for the quality assessment of seismic waveforms, while Valentine & Trampert (2012a) investigate neural network-based dimensionality reduction of seismograms and its potential applications.

Here we perform a Bayesian inversion of *P*-wave traveltime curves for the Earth's spherically symmetric *P*-wave velocity (V_P) structure. We use traveltimes from the EHB bulletin (Engdahl *et al.* 1998) for the *Pn*, *P*, *PP* and *PKP* phases. The inverse problem is non-linear, as the ray paths of the seismic phases depend on the underlying velocity structure of the Earth. Our focus is twofold. First, we demonstrate how to solve a Bayesian non-linear inverse problem and assess model uncertainties using neural networks. Second, we quantify the constraint on radial V_P structure which is provided by the traveltime data for these major seismic phases. To solve our non-linear inverse problem, we use a particular class of neural networks, known as a Mixture Density Network (MDN, Bishop 1995). An MDN outputs a parametric distribution, which approximates the continuous posterior model probability density. This distribution reflects our updated state of knowledge on the earth model parameters.

First, we briefly outline the Bayesian inversion framework, followed by an introduction to artificial neural networks. Second, we describe the model parametrization and the traveltime data used for this study. Last, we invert the traveltime data using neural networks and show the 1-D marginal pdfs for *P*-wave velocity parameters and discontinuity depths. We emphasize that our focus lies on the constraints on individual model parameters; we do not present a new 1-D earth model.

2 METHODOLOGY

2.1 The inverse problem

In the Bayesian formalism, all information is described by probability distributions that represent degrees of belief for each parameter. Following Tarantola & Valette (1982), the posterior state of knowledge is given by the conjunction of our prior knowledge and the information contained in the data, expressed by the likelihood. The solution to the general inverse problem can then be given by the conditional posterior probability distribution (Tarantola 2005)

$$\sigma(\mathbf{m}|\mathbf{d}) = k\rho(\mathbf{m})L(\mathbf{m}|\mathbf{d}), \tag{1}$$

where k is a normalizing constant, $\rho(\mathbf{m})$ is the prior distribution for the l-dimensional model $\mathbf{m}$ and $L(\mathbf{m}|\mathbf{d})$ is the likelihood function, which reflects how well a model explains the data. Both the posterior pdf and the likelihood function are conditioned on the observed data $\mathbf{d}$.

Instead of the pdf of the full model $\mathbf{m}$, it is often desirable to study the marginal probability distribution for a subset of the model parameters, that is,

$$\begin{aligned}\sigma(\mathbf{m}'|\mathbf{d}) &= k\rho(\mathbf{m}')L(\mathbf{m}'|\mathbf{d}) \\ &= \int \sigma(\mathbf{m}|\mathbf{d})\mathrm{d}m_{c+1}\mathrm{d}m_{c+2}\ldots\mathrm{d}m_l,\end{aligned} \tag{2}$$

where $\mathbf{m}'$ is a c-dimensional model vector (with $c \leq l$). The marginal posterior pdf represents the final state of knowledge of $\mathbf{m}'$, given

the variations in the remaining l–c model parameters. Usually, $c =$ 1 or 2, in which case the marginal probability distribution in eq. (2) represents 1-D or 2-D marginal posterior pdfs, respectively. The former reflect our knowledge of a single model parameter, while the latter are useful to investigate the correlation between any two parameters.

2.2 Neural networks

An artificial neural network is essentially a mathematical model of an arbitrary mapping between two parameter spaces. By changing the free parameters of the mathematical model, during a so-called training process, the mapping can be modified to represent the desired relation. Network training is driven by presenting the network with examples of corresponding input–output pairs. The fundamental idea is to represent the potentially complicated mapping as a combination of many simpler univariate activation functions. Common choices for the activation functions are the logistic and hyperbolic tangent functions. We use the latter in this work, because symmetric sigmoids, such as the hyperbolic tangent, often display better convergence properties during network training (e.g. LeCun *et al.* 1998). It is the non-linear nature of such functions that helps neural networks to approximate non-linear relations.

2.2.1 The Multilayer Perceptron (MLP)

Fig. 1 shows a two-layer feed-forward MLP. This particular type of neural network consists of two layers of free parameters (weights), which are represented by lines in the figure. The weight $w_{ij}^{(1)}$ in the first layer connects the input unit x_i with the hidden neuron h_j, while the second layer weight $w_{jk}^{(2)}$ connects the hidden unit h_j to the output neuron z_k. In addition, the biases of the first ($b_j^{(1)}$) and second layer ($b_k^{(2)}$) provide a constant offset as input to the neurons in a subsequent layer. For the activation functions we use here, the bias controls the threshold at which the output of a neuron changes sign. Information flows only in the forward direction from the input to the output neurons (feed-forward). The network output $\mathbf{z}(\mathbf{x}; \mathbf{w})$ is an explicit function of both the input $\mathbf{x}$ and network parameters $\mathbf{w}$. The K units in the MLP output layer are given by

$$z_k = g\left[\sum_j^J w_{jk}^{(2)} h_j + b_k^{(2)}\right], \tag{3}$$

where $g(\cdot)$ represents the activation function for the output neurons, $w_{jk}^{(2)}$ and $b_k^{(2)}$ are the second layer weights and biases, respectively, and h_j are the outputs of the J hidden neurons

$$h_j = f\left[\sum_i^I w_{ij}^{(1)} x_i + b_j^{(1)}\right]. \tag{4}$$

Here, $f(\cdot)$ is the activation function for the hidden units, $w_{ij}^{(1)}$ and $b_j^{(1)}$ are the first layer weights and biases, respectively, and x_i represents the values of the I input units. For the hidden neurons, we choose hyperbolic tangent functions, $f(a) = \tanh(a)$, while for the output units we use a linear activation function, $g(a) = a$. These are common choices, and such an MLP can learn an arbitrary continuous mapping from a finite data set, provided the number of hidden units is sufficient (Cybenko 1989; Hornik *et al.* 1989). Commonly, a trial-and-error procedure is adopted to determine the appropriate number of hidden neurons.

Learning corresponds to the minimization of a cost function with respect to the network weights. The cost function measures the difference between the network output and the desired output, the target vector. The necessary derivatives are given by the backpropagation algorithm, as introduced by Werbos (1974) and Rumelhart *et al.* (1986). The network is trained on a synthetic data set $D = \{\mathbf{x}_n, \mathbf{t}_n\}$, where $n = 1, \ldots, N$ labels the statistically independent patterns in the data set. Every pattern consists of a pair of input and target vectors $\mathbf{x}$ and $\mathbf{t}$, respectively. Once successfully trained, the network can be applied to unseen input to produce an estimate of the unknown output.

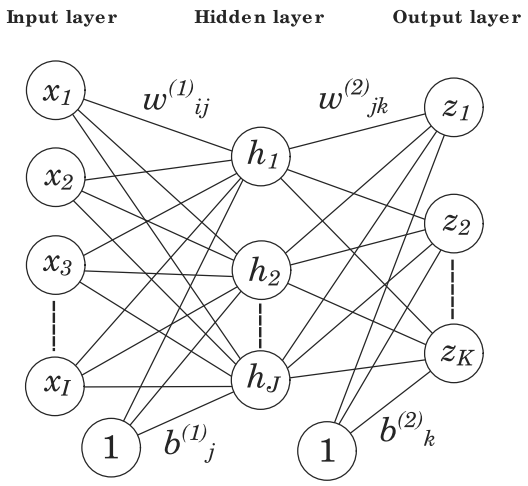

Figure 1. A two-layer feed-forward Multilayer Perceptron (MLP). The lines represent the two layers of free parameters in the network, represented by the weights $w_{ij}^{(1)}$ and $w_{jk}^{(2)}$ and biases $b_j^{(1)}$ and $b_k^{(2)}$. The I input neurons x_i feed into the J hidden neurons h_j, which form the input to the K output units z_k. An additional input of value 1 feeds into the hidden and output layer, which is associated with the biases. Information flows only from the input to the output neurons (feed-forward).

2.2.2 The Mixture Density Network

Bishop (1995) shows that an MLP, as shown in Fig. 1, outputs the mean of the conditional probability distribution $p(\mathbf{t}|\mathbf{x})$ of the target $\mathbf{t}$, conditioned on the input $\mathbf{x}$. This will give meaningless results if the underlying function, which relates input and target, is multivalued; therefore, it is desirable to obtain the full conditional distribution of the target (e.g. Bishop 1995; Meier *et al.* 2007a). We thus employ an MDN, which can model an arbitrary probability distribution, in the same fashion that an MLP can approximate an arbitrary function (McLachlan & Basford 1988).

In our study, the network input $\mathbf{x}$ corresponds to the body-wave traveltime curves $\mathbf{d}$ and the target $\mathbf{t}$ is given by the model parameters of interest $\mathbf{m}'$, a subspace of the radial P-wave velocity earth model $\mathbf{m}$. The precise composition of $\mathbf{m}$, $\mathbf{m}'$ and $\mathbf{d}$ will be discussed below. An MDN gives a continuous approximation to the corresponding marginal posterior pdf $\sigma(\mathbf{m}'|\mathbf{d})$ (eq. 2) as a linear sum of Gaussian kernels

$$\sigma(\mathbf{m}'|\mathbf{d}; \mathbf{w}) \approx \sum_{j=1}^{M} \alpha_j(\mathbf{d}; \mathbf{w}) \phi_j(\mathbf{m}'|\mathbf{d}; \mathbf{w}), \tag{5}$$

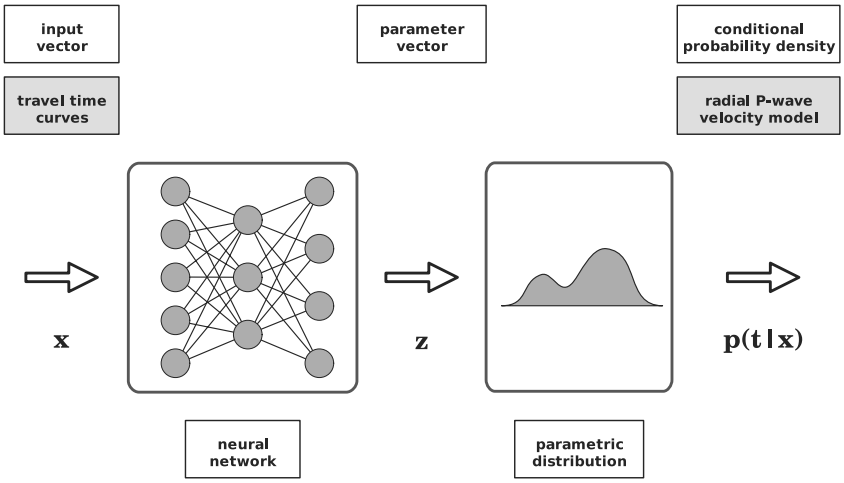

Figure 2. A Mixture Density Network (MDN), as introduced by Bishop (1995). The output of an MDN approximates a parametric distribution $p(\mathbf{t}|\mathbf{x})$ for the target $\mathbf{t}$, conditioned on the input $\mathbf{x}$. The parameters describing this distribution are given by the output $\mathbf{z}$ of a neural network, such as the MLP shown in Fig. 1. In this study, the input consists of traveltime data $\mathbf{d}$, while the target represents the parameters of interest $\mathbf{m}'$, which form a subspace of the l-dimensional radial P-wave velocity earth model $\mathbf{m}$ (eqs 1 and 2).

where α_j gives the relative importance of the jth kernel and the M Gaussian kernels ϕ_j are defined as

$$\phi_j(\mathbf{m}'|\mathbf{d};\mathbf{w}) = \frac{1}{(2\pi)^{c/2}[\sigma_j(\mathbf{d};\mathbf{w})]^c} \exp\left\{-\frac{\|\mathbf{m}' - \boldsymbol{\mu}_j(\mathbf{d};\mathbf{w})\|^2}{2[\sigma_j(\mathbf{d};\mathbf{w})]^2}\right\}, \quad (6)$$

where c is the dimensionality of the target vector $\mathbf{m}'$. The Gaussian mixture model is fully described by the mean vectors $\boldsymbol{\mu}_j(\mathbf{d};\mathbf{w})$ of size c, the variances $\sigma_j^2(\mathbf{d};\mathbf{w})$ and the mixing coefficients $\alpha_j(\mathbf{d};\mathbf{w})$. Each spherical Gaussian kernel ϕ_j is described by a single variance $\sigma_j^2(\mathbf{d};\mathbf{w})$, regardless of the dimensionality c of the target vector $\mathbf{m}'$. Note the explicit dependence of the network output $\sigma(\mathbf{m}'|\mathbf{d},\mathbf{w})$ on the network weights $\mathbf{w}$. We emphasize that an MDN models the posterior pdf in eq. (2) directly. Thus, the likelihood function $L(\mathbf{m}'|\mathbf{d})$ is not explicitly evaluated in the neural network, nor is the prior distribution $\rho(\mathbf{m}')$. The posterior distribution $\sigma(\mathbf{m}'|\mathbf{d})$ is evaluated implicitly through network training.

Fig. 2 illustrates an MDN, as introduced by Bishop (1995). The parameters describing the mixture model form the output $\mathbf{z}(\mathbf{d};\mathbf{w})$ of a conventional MLP, as shown in Fig. 1. For M spherical Gaussian kernels, the MLP will have $(c+2)\cdot M$ output parameters. Alternatively, more complex Gaussian kernels, such as Gaussians with full covariance matrices (Williams 1996), could be used. This is computationally more demanding, however, and we find that spherical kernels are flexible enough to model the probability densities of interest here. See Bishop (1995) for a detailed description of the MDN.

Once the parametric form of the probability distribution has been defined (eqs 5 and 6), the associated parameters can be found by training an MLP. Training corresponds to finding the weight values that maximize the likelihood for the desired pdf $\sigma(\mathbf{m}'|\mathbf{d};\mathbf{w})$. Since maximizing the likelihood is equivalent to minimizing the negative logarithm of the likelihood, the error function for the MDN is defined as (Bishop 1995)

$$E = -\sum_{n=1}^{N} \ln\left\{\sum_{j=1}^{M} \alpha_j(\mathbf{d}_n;\mathbf{w})\phi_j(\mathbf{m}_n'|\mathbf{d}_n;\mathbf{w})\right\}, \quad (7)$$

where the outer summation runs over the N patterns in the synthetic data set $D = \{\mathbf{d}_n, \mathbf{m}_n'\}$. Analytical expressions for the derivatives of E with respect to the adjustable network parameters are given by Bishop (1995) and allow an optimization algorithm to be implemented.

2.2.3 *Network training*

MDN training corresponds to the minimization of the cost function in (eq. 7). Commonly, gradient-based optimization algorithms are used for this task. We use the Scaled Conjugate Gradient (SCG) algorithm (Møller 1993), which avoids the expensive line-search procedure of the conjugate gradient algorithm. Conjugate gradient methods acquire second order information about the error surface and are therefore more efficient than simpler gradient descent methods.

Gradient-based optimization methods typically operate iteratively and thus require a user-defined starting point for the network weights. The starting point is crucial to ensure that the network training converges to an appropriate solution. For the hyperbolic tangent function, the summed input should be of order unity. If not, the activation functions become saturated, that is, their first derivative tends to zero. Consequently, the error surface will become almost flat, so that training ceases to be useful. To aid the initialization of the training process, it is common practice to pre-process the input and target vectors (Appendix A).

The initial network weights are drawn from a Gaussian distribution of zero mean. The variance of this distribution is inversely proportional to the number of input units I for the first layer weights $w_{ij}^{(1)}$ and the number of hidden units J for the second layer weights $w_{jk}^{(2)}$ (Bishop 1995). Further, the network parameters are initialized such that $\sigma(\mathbf{m}'|\mathbf{d},\mathbf{w}) \approx \rho(\mathbf{m}')$, that is, the initial posterior pdf resembles the prior pdf. This requires setting some initial values for the biases of the output layer in the MLP ($b_k^{(2)}$ in Fig. 1). Following (Bishop 1995; Nabney 2002), such an initialization leads to faster convergence and reduces the risk of the optimization method getting stuck in a poor local minimum.

Regardless of the initialization, every network run will be sensitive to the specific initial network parameters. It is therefore common practice to train multiple networks with different random weight initializations, all other settings being equal. The optimal weight vector

$\mathbf{w}^*$, which minimizes the cost function (eq. 7), is used to estimate the marginal posterior pdf $\sigma(\mathbf{m}'|\mathbf{d}, \mathbf{w}^*)$ through eq. (5).

2.2.4 Generalization and regularization

The goal of network training is to approximate the relationship between two parameter spaces. Such a mapping is in general believed to be found when the optimal network produces accurate results for previously unseen data, that is, data that was not used for network training. In that case, the network is said to display a good generalization behaviour. This can be verified by using a data set that is independent of the training data.

To simulate the presence of measurement uncertainties in the real data, noise is added to the synthetic data. This discourages the network from fitting the details of the training data set, referred to as overfitting. Instead, it enhances the generalization behaviour by encouraging the network to map the underlying relation between input and output. Bishop (1995) shows that such noise addition is similar to using a regularization constraint (such as simple weight decay), thereby forcing the network to find a smooth mapping, that is, a mapping that is insensitive to variations in the data on the order of the noise level. Meier *et al.* (2007a) demonstrate this concept in the context of a Bayesian inversion of surface wave data.

In addition, we employ early stopping, which is a common procedure to improve generalization. The network is trained using the conventional training set, but training is halted when the error (eq. 7) for an independent validation set starts to increase. Since the validation set is used to determine the optimal set of network weights, a third (test) set is used to verify the accuracy of the network on unseen data.

3 MODEL PARAMETRIZATION

We adopt a piecewise continuous representation for the *P*-wave velocity model, as was used for the *prem* (Dziewonski & Anderson 1981) and *iasp91* (Kennett & Engdahl 1991) models. The piecewise continuous functions can be used to evaluate the model at any depth exactly. We parametrize the depths of seven discontinuities in the V_P profile: the inner core boundary (ICB) and core–mantle boundary (CMB), the top of the D″ layer, the discontinuities around 660, 410 and 210 km depth and the Moho. We define the lower mantle (LM) as the region between the top of the D″ layer and the 660 km discontinuity, while the transition zone (TZ) spans the region between the 660 and 410 km discontinuities.

Between the discontinuities, we parametrize the *P*-wave velocity structure at L depths (knots). Subsequently, we construct the V_P profile $f(z)$, that is, the 1-D velocity structure between the discontinuities, by interpolating between the L knots using a set of L natural cubic spline functions

$$f(z) = \sum_{i=1}^{L} a_i \psi_i(z). \tag{8}$$

Each spline function $\psi_i(z)$ is a function of the depth z and equals 1 at one knot, while being 0 at the remaining $L - 1$ knots (Fig. 3). The coefficients a_i represent the V_P values at each knot. This yields a piecewise continuous representation of the model.

The transition zone and the region between the 410 and 210 km discontinuities are parametrized using eq. (8) with $L = 2$. This results in linear velocity gradients with depth in these layers. We separate the region between the Moho and 210 km discontinuity in two linear ($L = 2$) segments, that is, 210–120 km and 120–Moho km, as the linear velocity gradient in these two segments

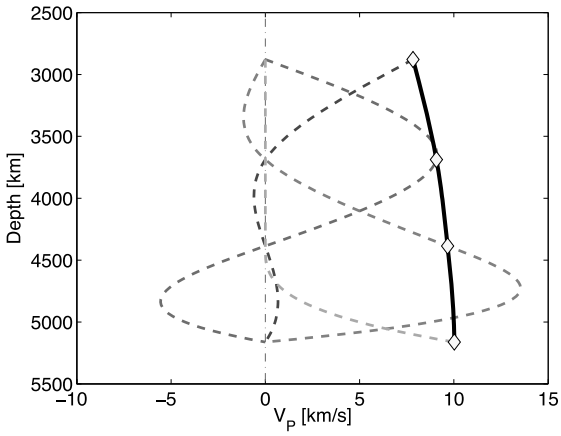

Figure 3. An example of the natural cubic splines used to construct the 1-D earth models, following eq. (8). The coefficients a_i represent V_P values in the outer core (yellow diamonds), drawn from the prior model distribution $\rho(\mathbf{m})$. The black line shows the resulting V_P profile $f(z)$ in the outer core, which is constructed by summing the products $a_i\psi_i(z)$ for the four splines (dashed).

differs significantly in existing 1-D models such as *ak135*. Thus, the velocity profile is continuous at 120 km depth, but its first derivative is not. The depth of the transition at 120 km is not varied. The lower mantle and outer core are parametrized by 4 knots, and the inner core by 3, which results in non-linear gradients with depth (Fig. 3). The crust is parametrized by two homogeneous layers. No sediment or water layers are present. We thus have 29 parameters in our model $\mathbf{m}$: 22 V_P parameters (the coefficients a_i in eq. 8) and 7 discontinuity depths.

Tables 1 and 2 define the prior model distribution $\rho(\mathbf{m})$ (eq. 1). Discontinuity depths are independently drawn from uniform priors, as are the V_P values directly below the discontinuities (Table 1). The prior distributions are centred on the corresponding values of the *ak135* model. We choose conservative priors by allowing for a large variation in the independent model parameters, that is, ±3 per cent with respect to *ak135* for V_P in the core and lower mantle and ±5 per cent in the upper mantle. We emphasize that by choosing such broad prior distributions, we ensure that the results of our probabilistic inversion are not driven by the actual values in the *ak135* model.

The V_P values at the other $L - 1$ knots in each region are calculated using the new value at the first knot (m_d^1 in Table 1) and the gradient of the *ak135* model. Subsequently, these values are perturbed, with the amount of perturbation drawn from a uniform prior (Table 2). This introduces a correlation between the V_P parameters in each region. In general, radial *P*-wave velocity increases with depth, that is, the velocity gradient is mostly positive. By using the gradient in *ak135*, we aim to exclude physically implausible models and restrict the size of our model space.

We generate 99 862 synthetic models, which are drawn from the prior model distribution $\rho(\mathbf{m})$. Ten synthetic models in the training set are shown in Fig. 4, along with the *ak135* model. Note that locally negative velocity gradients can still exist in the models.

4 TRAVELTIME DATA

4.1 EHB data

The EHB bulletin (Engdahl *et al.* 1998) contains millions of routinely determined traveltime measurements, which have been

Table 1. Prior information on independent model parameters. Prior distributions are uniform over the specified ranges. The ranges for the *P*-wave velocity parameters are given as percentile perturbations from *ak135* (Kennett *et al.* 1995), except for V_P in the two crustal layers. V_P parameters are indicated by m_d^1 and represent the knots located directly below a discontinuity d. Note that the tops of the transition zone (TZ) and the lower mantle (LM) are formed by the 410 and 660 km discontinuities, d_{410} and d_{660}, respectively. The interpolation style for the V_P profile in every region, following eq. (8), is given in the last column.

Discontinuity	Parameter	Range (km)	
Inner–outer core (ICB)	d_{ICB}	5133.5–5173.5	
Core–mantle (CMB)	d_{CMB}	2871.5–2911.5	
D″ layer (top)	$d_{D''}$	2720–2760	
660 discontinuity	d_{660}	630–690	
410 discontinuity	d_{410}	380–440	
210 discontinuity	d_{210}	190–230	
Moho	d_{Moho}	25–75	
Region	**Parameter**	**Range (± per cent)**	**Interpolation style**
Inner core (IC)	m_{IC}^1	3	3 cubic splines
Outer core (OC)	m_{OC}^1	3	4 cubic splines
D″ layer	$m_{D''}^1$	3	Linear
Lower mantle (LM)	m_{LM}^1	3	4 cubic splines
Transition zone (TZ)	m_{TZ}^1	5	Linear
410–210	m_{210}^1	5	Linear
210–Moho	m_{Moho}^1	5	
210–120			Linear
120–Moho			Linear
Region	**Parameter**	**Range (km s^{-1})**	
Lower crust (LC)	m_{LC}	6.4–7.4	
Upper crust (UC)	m_{UC}	5.6–6.3	

Table 2. Prior information on dependent model parameters. Prior distributions are uniform over the specified ranges, which are given as percentile perturbations from the updated model value (see text). The indices m_d^i represent the correlated model parameters in every region d, with higher indices i corresponding to deeper V_P knots in the parametrization. The corresponding independent parameters m_d^1 are listed in Table 1.

Region	Parameter	Range (± per cent)
Inner core (IC)	$m_{\mathrm{IC}}^{2,3}$	1
Outer core (OC)	$m_{\mathrm{OC}}^{2,3,4}$	1
D″	$m_{D''}^2$	1
Lower mantle (LM)	$m_{\mathrm{LM}}^{2,3,4}$	1
Transition zone (TZ)	m_{TZ}^2	2
410–210	m_{210}^2	2
210–Moho	$m_{\mathrm{Moho}}^{2,3}$	2

corrected for source mislocation. We select traveltime data for the *Pn*, *P*, *PP*, *PKPab*, *PKPbc* and *PKPdf* phases for the years 2001–2008 (Fig. 5). For simplicity, we exclude *PnPn*, the upgoing phases *pP*, *pwP* and *sP* and the crustal phases *Pb* and *Pg*.

Several corrections are provided with the EHB bulletin. We correct the raw EHB data for the Earth's ellipticity and station elevation. We do not apply the regionally smoothed station corrections, which perform regional averaging (5 × 5° patches) to smooth out effects of lateral heterogeneities in the upper mantle. In our setup, the imprint

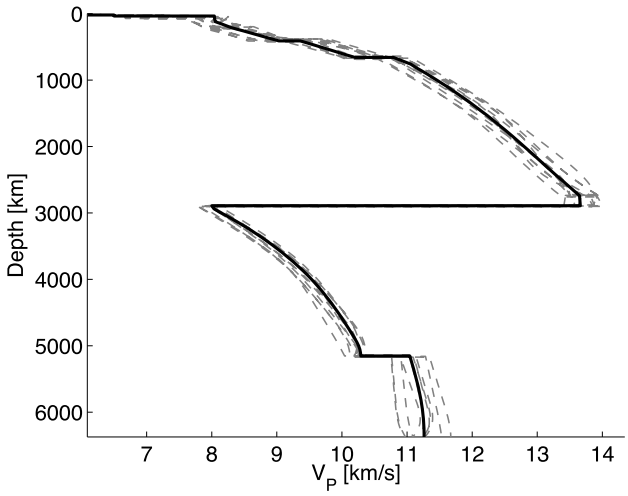

Figure 4. Ten model realizations (red), drawn from the prior distribution $\rho(\mathbf{m})$ and *ak135* (black) for V_P.

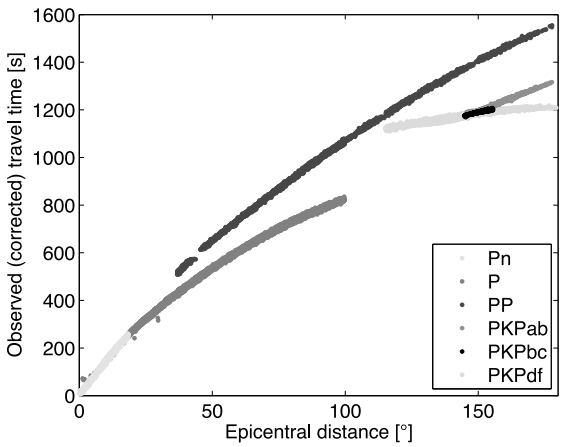

Figure 5. Travel time measurements in the EHB bulletin for 2001–2008. Event depths are restricted to lie between 14 and 16 km.

of 3-D structure on the traveltime measurements is treated as noise; therefore, such a correction is unnecessary here.

We select the traveltimes for which the EHB estimated source depth lies between 14 and 16 km. Note that this range of depths is chosen to approximate a fixed source around 15 km depth and not to represent the uncertainty in EHB source depth estimates; we will discuss our data noise model below in Section 4.3. Each event in the EHB bulletin is given a three-letter label that characterizes the quality of hypocentral determination. We exclude the EHB measurements for which the source depth is fixed to a standard depth by Engdahl, denoted by FEQ in the EHB data, and solutions for which the uncertainty in source depth is expected to be >15 km (LEQ, XEQ). The remaining data correspond to 1100 events that were registered at 5268 different stations (Fig. 6). Both sources and receivers are globally distributed, that is, within the typical limitations of seismological data coverage.

4.2 Synthetic data

Neural network training and validation requires a data set containing many examples of input–output pairs. For this purpose, we calculate synthetic first-arrival traveltime curves for 99 862 synthetic models

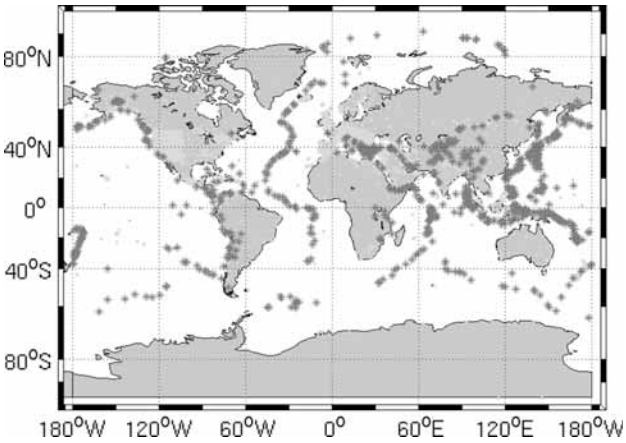

Figure 6. Locations of 1100 sources (red asterisks) and 5268 stations (green dots) in the EHB data for 2001–2008. Event depths are restricted to lie between 14 and 16 km.

using the TauP package (Crotwell *et al.* 1999). The synthetic models are parametrized as described in the previous section. All 29 model parameters are allowed to vary in each model realization. The source depth is fixed to 15 km for all synthetic data. Thus, source depth is a fixed parameter in our setup and any uncertainties in EHB depth estimates are regarded as a source of data noise, as will be discussed below in Section 4.3. The three PKP branches are calculated separately. Travel times are computed for a phase-specific range of epicentral distances at one degree intervals (Table 3). For these phases, the distance ranges are comparable to those used by Kennett *et al.* (1995), who used traveltime data collected by the ISC. We include *Pn*, which refracts along the Moho, to provide a better constraint on the structure of the uppermost mantle.

If for a given distance no arrival is computed by TauP, the traveltime is set to zero. By doing so, the number of elements in every traveltime branch is constant. This is a requirement of the network architecture, which only permits input vectors of constant dimension. The zeros represent gaps in the traveltime curves, which commonly result from low-velocity zones, and associated negative gradients, in the earth model. This information is therefore available to the neural network. Note that for the epicentral distance ranges used (Table 3), no gaps occur in the globally distributed EHB data (Fig. 5). This could indicate that there are no global low-velocity zones in the parts of the Earth sampled by the data, or that some of the EHB traveltimes do not represent a direct geometric arrival (or a combination of both).

4.3 Data uncertainties

Uncertainties exist in both the epicentral distance, through the source location estimate, and the traveltime measurements. We add noise to the synthetic data to simulate these two types of uncertainty.

For every synthetic traveltime measurement, we draw a perturbation to the epicentral distance from a uniform distribution $\mathcal{U}(-\epsilon_{\text{dist}}, +\epsilon_{\text{dist}})$ with $\epsilon_{\text{dist}} = 0.1^\circ (\sim 10\,\text{km})$. The value of ϵ_{dist} is similar to the average test-event mislocation reported by Engdahl *et al.* (1998). The corresponding traveltime is updated by applying the local gradient in the traveltime curve to this perturbation.

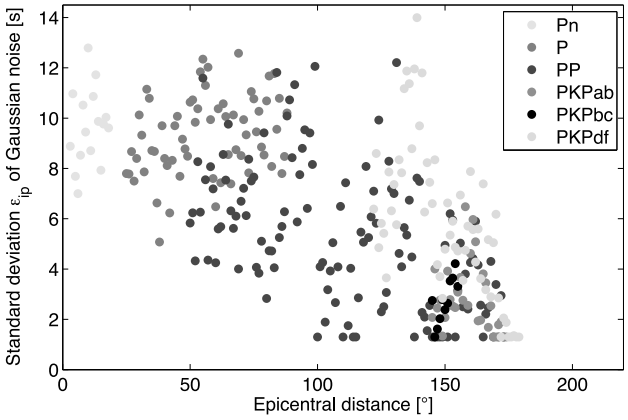

Figure 7. The half-width of the spread in the observed traveltimes ϵ_{ip} is used as the standard deviation of the Gaussian noise model, which differs for every phase p and epicentral distance i. Different colours denote different phases.

Table 4. Phase-specific measurement error ϵ_{ISC}, as documented in ISC (2008), which serves as a minimum for the standard deviation ϵ_{ip} (Fig. 7).

Phase	Pn	P	PP	PKPab	PKPbc	PKPdf
$\epsilon_{\text{ISC}}[s]$	0.8	0.8	1.3	1.3	1.3	1.3

Second, the synthetic data have to be corrupted to reflect noise in the traveltime data. The scatter in the observed traveltime data (Fig. 5) originates from lateral heterogeneities in the Earth, measurement errors, phase misidentifications and uncertainties in the estimated source depth. For a given phase p and epicentral distance i, we estimate the noise as the spread in the EHB traveltimes. The half-width of this spread ϵ_{ip} (Fig. 7) is used to define a Gaussian noise distribution $\mathcal{N}(0, \epsilon_{ip}^2)$, that is, with zero mean and standard deviation ϵ_{ip}. A random sample from this noise distribution is added to every synthetic datum. The scatter in the data may be small if few data are available, which would result in an unrealistically low noise estimate. Therefore, we impose a minimum phase-specific noise level (Table 4), which is based on measurement error estimates documented in a recent ISC report (ISC 2008).

4.4 Data processing

The input **d** to the neural network is a concatenation of the traveltime curves for the different phases. Since the curves are rather smooth, a large (linear) correlation exists between the traveltime at different epicentral distances. Therefore, we sample the traveltime curves at intervals of 2°. This reduces the size of the input vector and thus the number of network parameters, thereby making network training faster. In light of the strong correlations, we assume that this downsampling does not result in a significant loss of information on our earth model. The resulting input vector consists of 152 traveltimes, which is a concatenation of the data for the Pn

Table 3. Epicentral distance range for the seismic phases. The ranges used by Kennett *et al.* (1995) are added as a reference.

Distance range (°)	Pn	P	PP	PKPab	PKPbc	PKPdf
This study	3:18	25:88	50:173	145:174	145:155	122:179
Kennett *et al.* (1995)	–	25:99	53:180	156:178	151:153	118:180

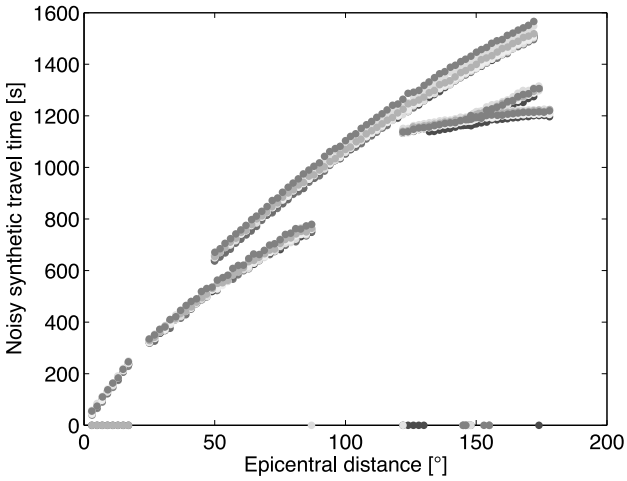

Figure 8. Examples of noisy synthetic traveltime data that form the input to the network. The synthetic data are sampled at distance intervals of 2° for the epicentral distance ranges specified in Table 3. Note the zeros in the traveltime curves, which indicate that TauP did not compute an arrival at the corresponding distance.

(8 traveltime measurements), P (32), PP (62), PKPab (15), PKPbc (6) and PKPdf (29) phases.

For every synthetic model, the input pattern is constructed by sampling the synthetic data at 2° intervals and subsequently adding a random noise sample (Fig. 8). The input vectors for the observed data are constructed in a slightly different fashion. For each distance sample d_{ip}, with phase p and epicentral distance i, we extract all observations from the EHB data for which the epicentral distance lies within the range $d_{ip} \pm \epsilon_{\text{dist}}$, where $\epsilon_{\text{dist}} = 0.1°$ as before. Consequently, multiple EHB observations are available at each distance point. We draw one random sample from these multiple possibilities. This yields the EHB traveltime curves that serve as input to the trained networks (Fig. 9). The differences between these curves are regarded as noise (see Section 4.3). The variations in the real data vectors are significantly smaller than the variations in synthetic training patterns (*cf.* Fig. 8). Both observed and synthetic data contain noise, but the variations in the synthetic data are larger due to the differences in the underlying synthetic earth models.

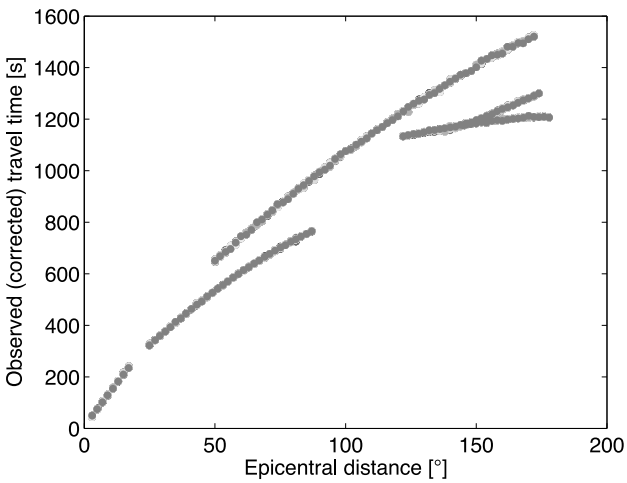

Figure 9. Ten input patterns, which are constructed from the EHB data (Fig. 5). Note that the variation in these real data vectors is significantly smaller than the variations between synthetic training patterns (Fig. 8).

Note that upon applying a trained network to one input pattern for the EHB data, only 152 measurements are 'inverted'. For a given distance and phase, however, all available observations should contain the same information on the radial earth model, given the measurement noise defined above. When repeated with a different EHB input pattern, constructed as described here, the inversion should yield similar results.

5 RESULTS

We present inversion results for all 22 *P*-wave velocity and seven discontinuity depth parameters. For each model parameter m_i, we investigate the constraint that is provided by the traveltime data and quantify the associated uncertainty. We therefore train MDNs on 1-D target vectors $\mathbf{m}' = m_i$ (eq. 2). Since all model parameters are allowed to vary in the synthetic models, the output of each MDN forms a 1-D marginal posterior pdf $\sigma(m_i|\mathbf{d})$. This is equivalent to marginalizing the full joint posterior pdf $\sigma(\mathbf{m}|\mathbf{d})$ over all model parameters other than m_i via the integration in eq. (2). $\sigma(m_i|\mathbf{d})$ reflects our knowledge of m_i given the variations in the other 28 model parameters.

5.1 Network configuration

For all results presented in this study, we train MDNs with 40 hidden units and a Gaussian mixture consisting of 15 Gaussian kernels ($M = 15$). We verified that the precise number of hidden units is not of paramount importance to final network performance. The same applies to the number of Gaussian kernels. During training, the mixing coefficient α_j can be set close to zero for redundant kernels, or kernels can be combined by having a similar mean and variance (Bishop 1995).

For a 1-D target ($c = 1$), the MDN has $(c + 2) \cdot M = 45$ output parameters: the means μ_j, the variances σ_j^2 and the mixing coefficients α_j (eqs 5 and 6). In combination with the 152-D input pattern (Figs 8 and 9), the corresponding MLP has 7725 free parameters (the weights $w_{ij}^{(1)}$ and $w_{jk}^{(2)}$ and biases $b_j^{(1)}$ and $b_k^{(2)}$ in Fig. 1). We use 80 per cent of the 99 862 patterns in the synthetic data set for training, 15 per cent for the validation set, which is used to evaluate the early stopping criterion, and the remaining 5 per cent for the test set.

Theoretically, there is no limit to the size of a neural network. However, a larger network consists of more free parameters and thus takes longer to train. More importantly, more network parameters require a larger training set to succesfully train the network. Therefore, computational facilities restrict the network size. In general, the number of free parameters should not exceed the number of training patterns (e.g. Bishop 1995; Duda *et al.* 2001). In this study, we ensure that the training set is approximately a factor of 10 larger than the number of network parameters. The main computational requirement thus lies in the generation of synthetic training patterns, that is, repetitively solving the forward problem. For the ∼100 000 patterns, this took ∼100 hr on a standard desktop computer. Once the training set is available, network training is relatively fast: the training time for a single network is on the order of tens of minutes in this study.

Due to the random initialization of the network weights, the optimization algorithm can become stuck in local minima of the error surface. To verify that network training converges properly, we train 30 independent networks. For each of these networks, training commences at a different point in weight space due to the random

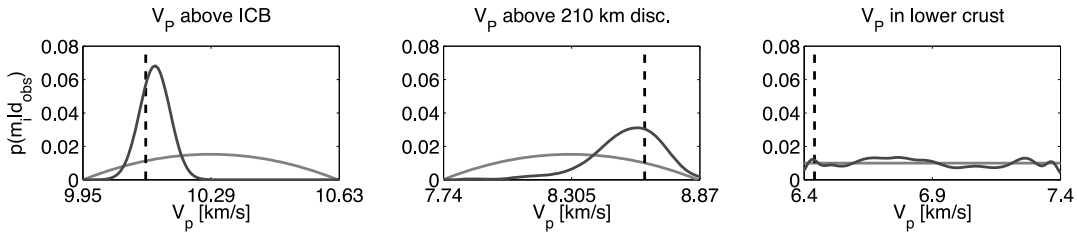

Figure 10. 1-D marginal posterior pdf (blue line), prior pdf (red) and true model value (black, dashed) for one pattern in the test data set for V_P (left-hand frame) directly above the ICB, (middle frame) directly above the 210 km discontinuity and (right-hand frame) in the lower crust.

initialization. To reduce the chance of overfitting, all synthetic patterns are divided randomly over the training, validation and test sets before the training of each independent network commences. We find that results for the 30 independent networks are similar and choose the network that produces the lowest pattern-averaged error for the test set.

5.2 Network evaluation

For each V_P parameter, we apply the optimal MDN to the ~5000 patterns in the synthetic test set, which are not used for network training. Network performance can be evaluated by comparing the resulting ~5000 1-D marginal pdfs with the true synthetic model value. As an example, we show 1-D marginal posterior pdfs for one test pattern for three model parameters: V_P directly above the ICB, directly above the 210 km discontinuity and in the lower crust (Fig. 10). For the *P*-wave velocity near the ICB (left-hand frame) and 210 km discontinuity (middle frame), most probability mass in the posterior distribution lies close to the target value (black line). We conclude that network performance is accurate for this particular input pattern. It is clear from the marginal pdf of V_P in the lower crust (right-hand frame) that the traveltime data do not constrain this part of the model. Consequently, the MDN output represents the uniform prior distribution for this parameter. The difference between the MDN output and the true uniform prior pdf is due to the fixed shape of the finite number of Gaussian kernels.

It is difficult to quantitatively evaluate network performance from such marginal distributions, however. A more pragmatic validation method is to analyse the correlation between the target value and the mean of the dominant Gaussian kernel in the MDN output (Bishop 1995). We take the kernel associated with the largest mixing coefficient α_j (eq. 5) to be dominant. Although this simple measure ignores the information provided in the full posterior pdf, such an analysis provides a practical way of evaluating the trained networks.

For all 22 V_P parameters, Fig. 11 shows the mean μ_j of the dominant Gaussian kernel versus the true target value for all patterns in the test data set. Every row in the figure represents a region between two discontinuities, with the depth of the V_P parameter (knot) decreasing from left to right. The corresponding correlation coefficient R is given above every frame ($R = 0$ indicates no correlation, whereas $R = 1$ represents a perfect correlation). Network performance on these unseen input patterns is good ($R \geq 0.87$) for the *P*-wave velocity in the inner and outer core and lower mantle (first, second and fourth row, respectively). For V_P in the D″ layer (third row), the upper mantle (fifth and sixth row) and crust (bottom row), correlations are in general low or absent ($R \approx 0$).

Besides network evaluation, the performance on synthetic input is a good indicator of the constraint that the data provide on the model. For the *P*-wave velocity directly above the ICB, for instance, the 1-D marginal posterior pdf is unimodal and narrow relative to the width of the prior pdf (left-hand frame, Fig. 10). This parameter is constrained well by the data, as indicated by $R = 0.94$ in Fig. 11 (second row, first column). Conversely, for V_P in the lower crust the mean of the 'dominant' kernel does not relate to the true value ($R = 0.05$, Fig. 11, seventh row, first column). The traveltime data do not constrain this part of the model, as is apparent from the corresponding marginal pdf in Fig. 10 (right).

One can expect similar results, for both resolved and unresolved model parameters, when applying the trained networks to the observed traveltime data. As the data provide very little or no constraint on the seven discontinuity depth parameters, we do not show the corresponding performance on the test set here and restrict ourselves to the application to the EHB data for these parameters.

5.3 Application to EHB data

Fig. 12 shows 1-D marginal pdfs for *P*-wave velocities for the ten EHB input patterns in Fig. 9. Recall that these ten input patterns are random realizations from the available EHB data set, as described in Section 4.4. The differences between these input patterns are regarded as noise. The network should be insensitive to such variations, since we have used a similar noise level during network training. Consequently, network output should be approximately the same for these different input vectors.

The data constrain V_P in the outer core (OC) and lower mantle (LM) best (second and fourth row, respectively). PKPdf, the only seismic phase in our data set that is sensitive to the inner core (IC), constrains V_P in this region (first row). The most notable feature is the proximity of the posterior maxima to the *ak135* values, which are indicated by the green dashed lines. However, in addition to a most likely model value, we obtain uncertainties in the *P*-wave velocity estimate. Recall that our posterior pdfs are based on our conservative prior pdfs and are therefore taken to be independent of the actual values in the *ak135* model.

For each V_P parameter in the inner core, outer core and lower mantle, we extract statistics from the ten posterior distributions in Fig. 12. Since the ten distributions are similar for these parameters, we calculate the mathematical expectation $\langle V_P \rangle$ and standard deviation σ_{V_P} of the 10 pdfs combined (Table 5). When expressed as a percentage of the mean $\langle V_P \rangle$, the standard deviation of the posterior pdfs σ_{V_P} is smaller than 1 per cent for every model parameter. The corresponding *ak135* values are given as a reference and lie within one standard deviation from $\langle V_P \rangle$ for the parameters in Table 5, except for $m^{2,3}_{\mathrm{LM}}$ in the lower mantle, for which *ak135* lies within two standard deviations. It is difficult to compare the uncertainties found here, represented by the standard deviations σ_{V_P}, as uncertainty estimates are scarce in the literature. Kennett *et al.* (1995) used a non-linear search procedure to evaluate a range of models around *ak135* for various data misfit measures. They used velocity

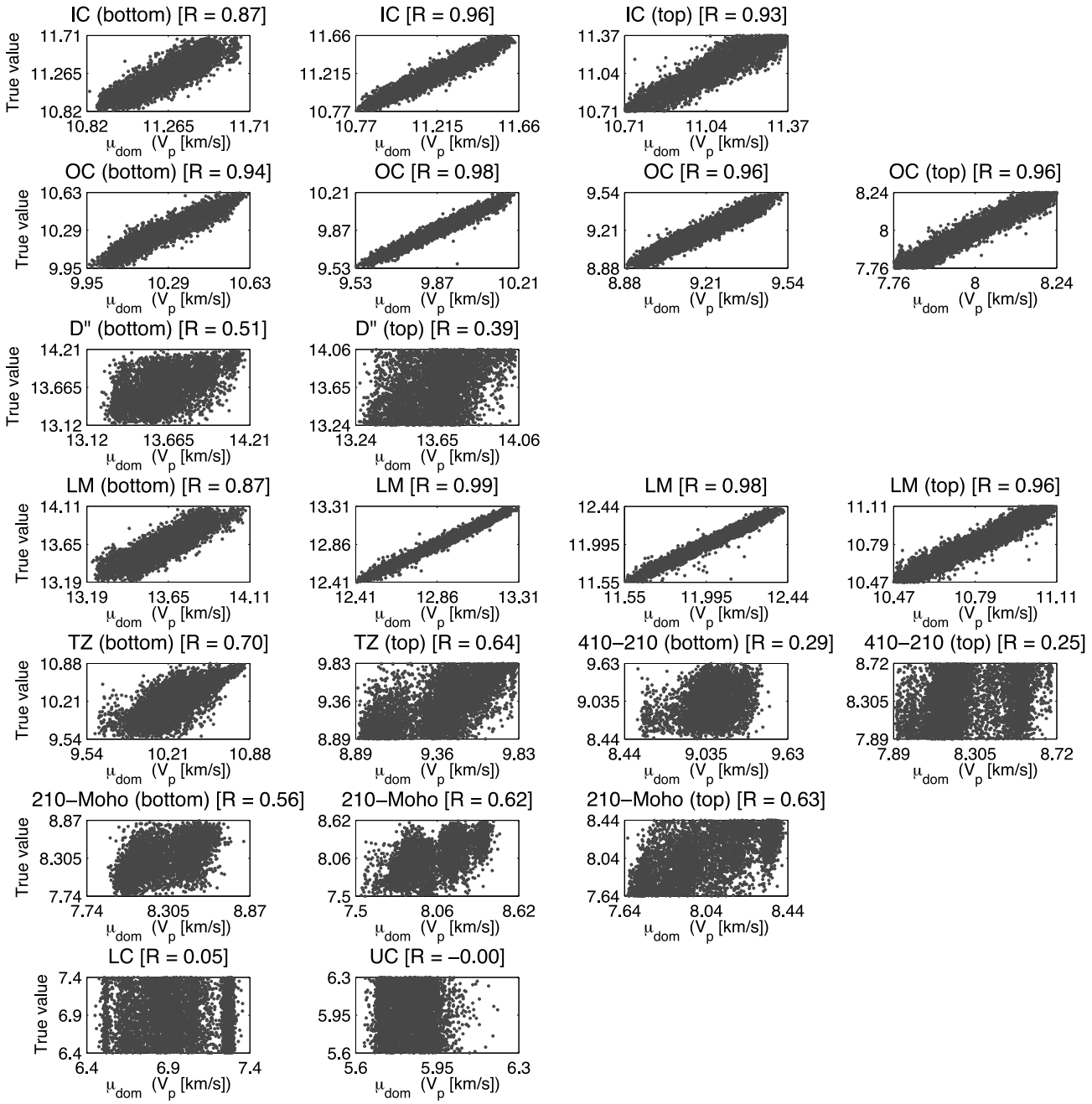

Figure 11. Mean μ_j (eq. 6) of the dominant Gaussian kernel (maximum α_j in eq. 5), labelled μ_{dom} in the figure, versus the true synthetic value for all patterns in the test data set for the 22 V_P parameters (Tables 1 and 2). The regions between discontinuities are displayed on different rows. For every region, depth decreases from left to right in the figure. The corresponding correlation coefficient R is given above each frame.

bounds of $\pm 0.02\,\mathrm{km\,s^{-1}}$ for the lower mantle and $\pm 0.04\,\mathrm{km\,s^{-1}}$ for the lowermost mantle and core. The σ_{V_P} values we find here are on the order of these velocity bounds or a factor of 2–3 larger (Table 5). We emphasize that our uncertainty estimates are not representative of the uncertainties in *ak135*, given the differences in the data selection and the model parametrization. However, the constraint on individual model parameters, as investigated here, may be indicative of the uncertainties in *ak135* and similar models.

The 1-D marginals for V_P between the Moho and the 210 km discontinuity (Fig. 12, sixth row) indicate that the traveltime data contain some information on this region. We find that this is mainly due to the addition of the Pn phase, which refracts along the Moho. The data indicate a (very) weak preference for velocities slightly higher than in *ak135*. We should point out, however, that the posterior distributions include all but the very low P-wave velocities. Thus, for these model parameters the data cannot falsify any of the assumptions on V_P contained in our model prior.

The data contain no or at most a very limited signal on the parameters of the D″ layer (third row, Fig. 12), the transition zone (TZ, fifth row), the region between the 410 and 210 km discontinuities (410–210, fifth row) and the two homogeneous crustal layers

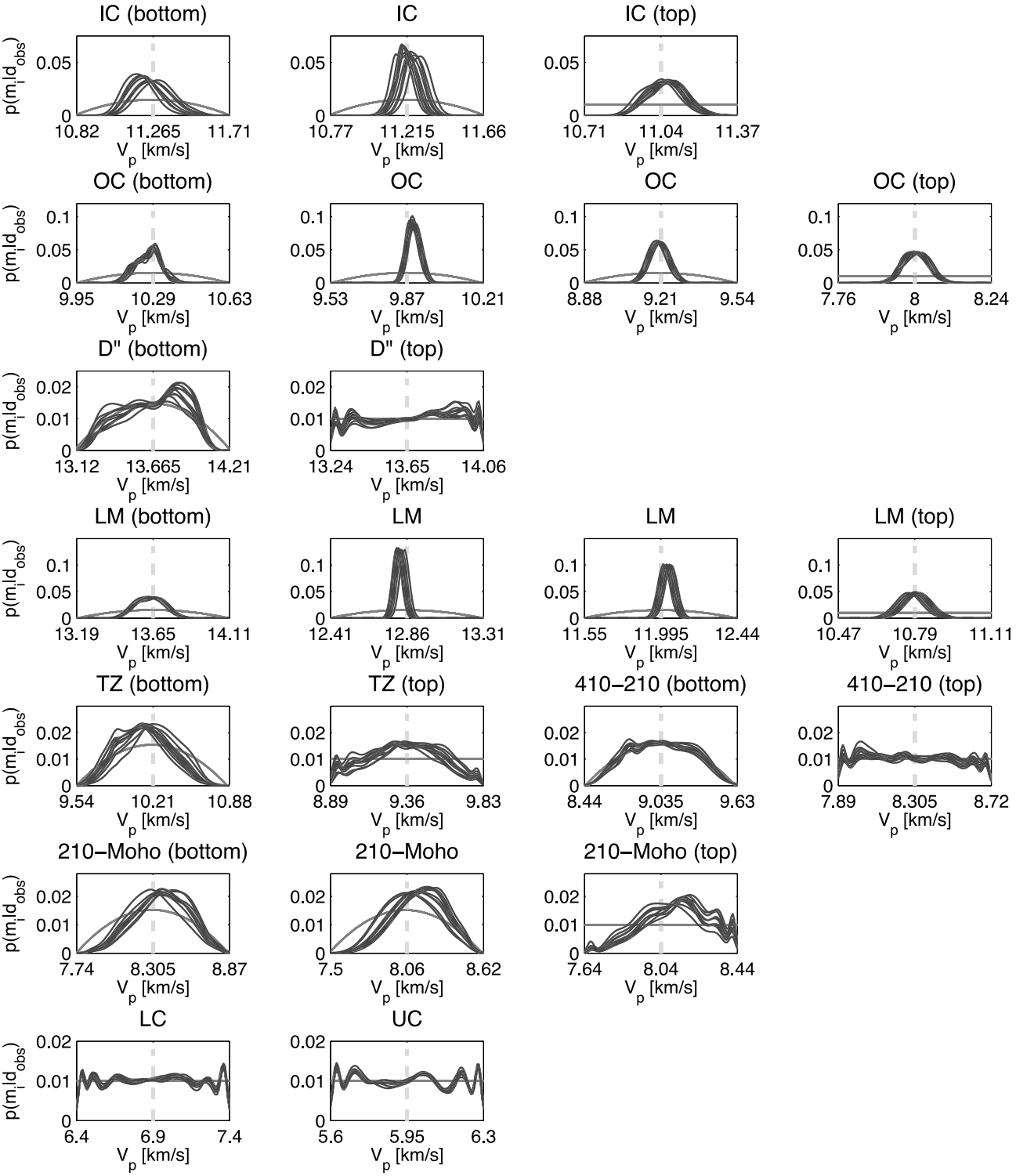

Figure 12. 1-D marginal posterior pdf (blue line), prior pdf (red) for the 22 V_P parameters (Tables 1 and 2). The same trained networks as used for Fig. 10 are applied to ten different EHB input patterns (Fig. 9). The regions between discontinuities are displayed on different rows. For every region, depth decreases from left to right in the figure. Note that the range of the vertical axis, that is, normalized probability, differs between rows.

(seventh row). The poor constraint on upper mantle structure is not surprising, given the teleseismic epicentral distances for which *P* (above 25°) and *PP* (above 50°) traveltimes are used (Table 3).

The traveltime data we invert here contain practically no information on the depth of the discontinuities (Fig. 13). For the upper mantle, this may be explained by the near-vertical incidence under which the teleseismic rays travel through the discontinuities in this region. The inclusion of phases that reflect of discontinuities, for example, *PcP*, which reflects of the CMB, could improve the constraint on discontinuity depths.

Table 5. Mean $\langle V_P \rangle$ and standard deviation σ_{V_P} of the ten 1-D marginal posterior pdfs for the *P*-wave velocity parameters (Fig. 12) in the inner core (IC), outer core (OC) and lower mantle (LM). All values are in (km s^{-1}), except for the fourth column, which shows the standard deviation σ_{V_P} as a percentage of the mean $\langle V_P \rangle$. The corresponding value in *ak135* is given for comparison (V_P^{ak135}). Recall that the depth of the knots m^i decreases with decreasing index number *i* (see Tables 1 and 2).

Parameter	$\langle V_P \rangle$	σ_{V_P}	σ_{V_P} (per cent)	V_P^{ak135}
m^3_{IC}	11.248	0.107	0.952	11.265
m^2_{IC}	11.229	0.066	0.589	11.215
m^1_{IC}	11.063	0.081	0.736	11.040
m^4_{OC}	10.266	0.056	0.545	10.290
m^3_{OC}	9.899	0.030	0.304	9.870
m^2_{OC}	9.201	0.043	0.466	9.210
m^1_{OC}	8.004	0.041	0.506	8.000
m^4_{LM}	13.638	0.086	0.629	13.650
m^3_{LM}	12.819	0.031	0.239	12.860
m^2_{LM}	12.034	0.039	0.320	11.995
m^1_{LM}	10.793	0.057	0.528	10.790

6 DISCUSSION

We trained neural networks to invert *P*-wave traveltime data for the radial *P*-wave velocity structure of the Earth. We obtained a continuous probabilistic description of the individual model parameters from the conjunction of our prior knowledge with the information contained in the data. The 1-D marginal posterior pdfs enable us to assess the uncertainty in the model parameters, which reflects the non-uniqueness of the non-linear inverse problem.

A visual comparison of the prior and posterior pdfs enables us to assess how well a model parameter is resolved by the data. Alternatively, one can quantify the constraint on a model parameter by comparing the information content of the prior and posterior pdfs (Tarantola & Valette 1982), as was done by for instance Meier *et al.* (2007b). Quantifying the information content, or gain, is useful when quantitatively comparing the resolving power of various data types or when it is not possible to show the posterior pdfs for all model parameters. We do not include such a measure in this study, as we explicitly show the prior and posterior distributions for all 29 model parameters (Figs 12 and 13).

We use neural networks as an alternative to Monte Carlo methods. A succesful comparison of the two types of technique was presented by for instance Meier *et al.* (2007a); Shahraeeni & Curtis (2011). The ∼100 000 samples in our data set can be used to sample the posterior pdf with the straightforward Independent Metropolis–Hastings algorithm. However, we find that the data set is insufficient to produce robust results. To obtain the posterior pdf, we would need many more samples or a more sophisticated approach, such as the Neighbourhood Algorithm (Sambridge 1999a). This illustrates the efficiency of the neural network to interpolate between the limited number of available samples

In general, the deeper parts of the model (inner and outer core, lower mantle) are constrained well by the data and the corresponding posterior pdfs contain V_P values similar to those in *ak135* (Kennett *et al.* 1995). By contrast, the same data provide little information on the upper mantle V_P structure and the depth of discontinuities in the radial V_P profile. Kennett *et al.* (1995) derived the *ak135* model from the traveltime data provided by the ISC. They used the same phases as we use here, except for *Pn*, and in addition used *PcP*, *PKKP*, *P′P′ (PKPPKP)* and the converted phases *ScP* and *SP*. The inclusion of such complimentary phases could have enhanced our knowledge on parts of the V_P model. These phases, however, are not included in the EHB bulletin for 2001–2008 and we decided to restrict our inversion to the phases listed in Table 3.

Ideally, the output for all ten input patterns in Figs 12 and 13 is similar (see Section 4.4). This is the case for the well-resolved parameters in the outer core and lower mantle. Differences between the posterior pdfs are larger for the inner core parameters, as they are for the *P*-wave velocity in the uppermost mantle (210–Moho). The trained networks are thus not completely insensitive to random variations in the input on the order of the noise level. In our view, however, these differences are minor and we argue that similar

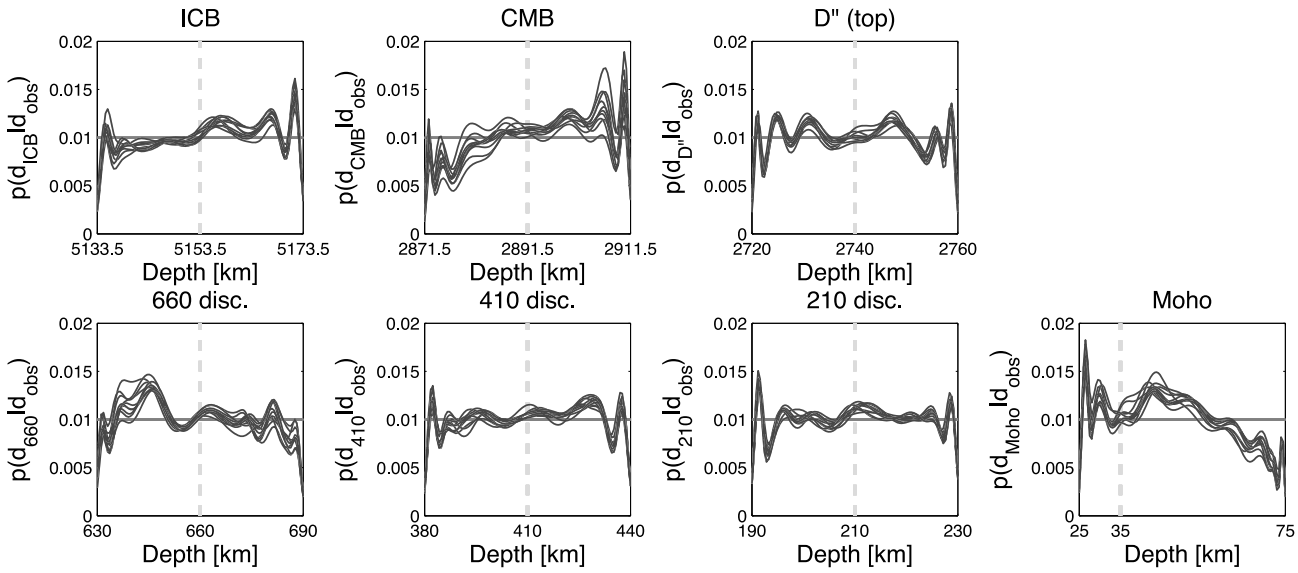

Figure 13. 1-D marginal posterior pdf (blue line), prior pdf (red) for the seven discontinuity depths (Table 1). Trained networks are applied to 10 different EHB input patterns (Fig. 9).

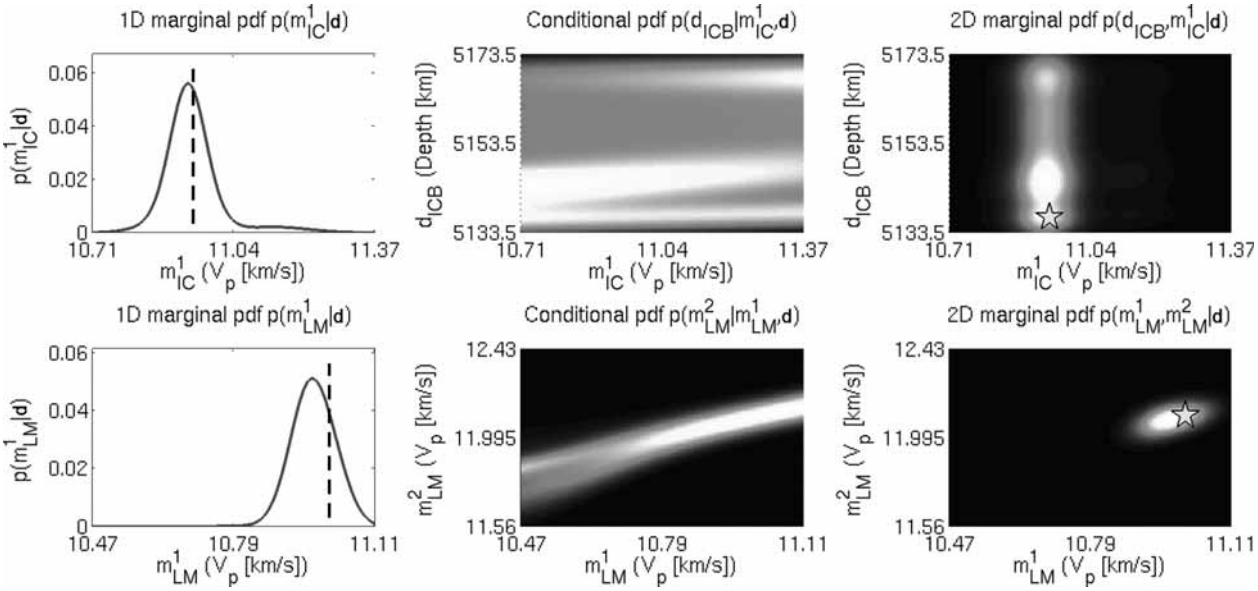

Figure 14. Construction of 2-D marginal posterior pdfs via eq. (9) for (first row) V_P at the top of the inner core (m^1_{IC}) and the ICB depth (d_{ICB}) and (second row) V_P in the lower mantle (m^1_{LM} and m^2_{LM}). See Tables 1 and 2 for the model parametrization. The three panels in each row show (left-hand panel) the 1-D marginal pdf, (middle panel) the conditional pdf and (right-hand panel) the 2-D marginal pdf for one of the patterns in the test set. Lighter colours denote higher probabilities. The corresponding target values are denoted by the black line (left-hand panel) and the cyan star (right-hand panel).

inferences would be drawn from the ten 1-D marginal posterior pdfs for these parameters.

For all parameters, the networks have been trained on the same type of traveltime data as input. Obviously, any information redundancy between the individual units in the input pattern is inefficient from the perspective of network training. This is the case for the inner core parameters, for which we know only PKPdf carries information. It would thus be most efficient if the networks trained for the inner core parameters would only take PKPdf as input. We verified that prediction accuracy is similar, regardless of whether the networks are trained on the full input pattern or only on PKPdf traveltimes. Thus, network training is able to focus on the systematic relation between the PKPdf data and the inner core *P*-wave velocities.

Our final results comprise the 1-D marginal posterior pdfs $\sigma(m_i|\mathbf{d})$ of the model parameters m_i. These represent all available knowledge on the individual model parameters, given the traveltime data, associated measurement errors, our choices regarding the model parametrization and the variations in the other model parameters. Such 1-D distributions do not contain information on any correlations between model parameters. Further, we emphasize that the maximum likelihood values of the individual model parameters, when taken together, do not necessarily represent the maximum likelihood model. Therefore, it is often desirable to analyse the joint posterior pdf, that is, the posterior probability distribution of the full model $\sigma(\mathbf{m}|\mathbf{d})$ (eq. 1). This distribution could be obtained by training a network on the complete 29-D model $\mathbf{m}$. Such a network, however, would contain a lot of free parameters and thus require a large training set. Further, network training may converge slowly or not at all for such a high-dimensional target space.

Alternatively, a joint distribution for an n-dimensional model can be constructed from the product of conditional and marginal pdfs (e.g. Tarantola 2005)

$$\sigma(m_1, \ldots, m_n|\mathbf{d}) = \sigma(m_n|m_1, \ldots, m_{n-1}, \mathbf{d}) \times \sigma(m_1, \ldots, m_{n-1}|\mathbf{d}). \quad (9)$$

For each pdf in the r.h.s. product, a separate network would be trained. Note that it is straightforward to train networks on conditional pdfs such as $\sigma(m_n|m_1, \ldots, m_{n-1}, \mathbf{d})$. This only requires the input pattern to be extended with the model parameters on which the distribution is conditioned. Once successfully approximated, the joint pdf has to be sampled to analyse its properties. This requires the evaluation of the individual pdfs in the decomposition for many random model realizations and performing the necessary multiplication following eq. (9). This can be done efficiently, as for every trained network computing the probability for a specific input datum and model value is very fast, that is, consumes a fraction of a second on a standard desktop computer. By doing so, one can approximate the joint posterior model distribution via eq. (9) and construct a representative ensemble of models. This will aid the interpretation of the information on Earth structure that is contained in seismological data.

As a simple example, we construct 2-D marginal posterior pdfs using eq. (9), that is, for a 2-D $\mathbf{m}'$ (eq. 2), for one of the patterns in the test set. Fig. 14 shows the three distributions in eq. (9) for two different combinations of parameters: V_P at the top of the inner core (m^1_{IC}) and the ICB depth d_{ICB} (first row) and V_P in the lower mantle (m^1_{LM} and m^2_{LM}, second row). The inner core V_P is resolved well, whereas the ICB depth is unresolved, as was apparent from the 1-D marginals in Figs 12 and 13. No correlation between these parameters is observed in the 2-D marginal pdf (Fig. 14). The two shallowest V_P parameters in the lower mantle are resolved well (*cf.* Fig. 12, fourth row). Despite the good constraint provided by the data, a (weak) positive correlation between the two parameters is visible in the corresponding 2-D marginal posterior pdf (Fig. 14).

7 CONCLUDING REMARKS

We used artifical neural networks to solve a non-linear Bayesian inverse problem. Neural networks are flexible and can be used to

approximate an arbitrary function. No linearization or model damping is required, which allows for an optimal use of the information on the model that is contained in the data. We used an MDN to acquire a continuous probabilistic description of each model parameter. Each 1-D marginal posterior pdf represents our knowledge of the parameter and provides the necessary quantification of uncertainties, which plays a crucial role in any interpretation of seismological models.

We investigated the information on the Earth's radial *P*-wave velocity structure that is available in the EHB traveltime data for the *Pn*, *P*, *PP*, *PKPab*, *PKPbc* and *PKPdf* phases. Our results comprise 1-D marginal posterior probabiliy distributions for the 22 V_P parameters and seven discontinuity depths in our model. These 1-D marginal pdfs enable us to assess the uncertainty in the individual model parameters. We have shown how the method can be extended to obtain a posterior pdf for a multidimensional model space. This enables us to investigate potential correlations between model parameters.

The *P*-wave velocities in the inner core, outer core and lower mantle are resolved well, that is, standard deviations of ∼0.2 to 1 per cent with respect to the means of the 1-D marginal posterior pdfs. The maximum likelihoods of V_P are in general similar to the corresponding *ak135* values, which lie within one or two standard deviations from the means of the posterior pdfs (Table 5). This provides an independent validation of this part of the *ak135* model, which is often used in 3-D seismic tomography and earthquake location algorithms. Conversely, the data contain little or no information on *P*-wave velocity in the D″ layer, the upper mantle and the homogeneous crustal layers. For the upper mantle, this is not surprising, given that the traveltime data used here are of a teleseismic nature, that is, $>25^\circ$ epicentral distance. Using additional phases available in the ISC bulletin, such as *PcP*, *PKKP* and the converted phases *SP* and *ScP*, may enhance the resolvability of our model parameters. However, the major phases we used here give a good indication of how much information on the radial V_P structure is contained in typical body-wave traveltime data. We included *Pn*, which led to a weak constraint on the V_P structure in the uppermost mantle. The data do not constrain the depth of the discontinuities in our model. Again, this is common knowledge, as teleseismic rays tend to travel perpendicular to discontinuities and thus provide a poor sampling of these structures. Reflected phases, such as *PcP*, which reflects of the CMB, are known to contain much more information on discontinuities.

Seismograms contain more information on the seismic source and the Earth's structure. We aim to apply Mixture Density Networks to Bayesian seismic waveform inversion in the future. However, for such an application the dimensionality of the data is much larger than for the traveltime inversion performed in this study, which presents additional challenges that must be overcome.

ACKNOWLEDGEMENTS

We thank Malcolm Sambridge and an anonymous referee for constructive reviews. We appreciate helpful discussions with Paul Käufl and Hanneke Paulssen. Ralph de Wit and Andrew Valentine are funded by the Netherlands Organization for Scientific Research (NWO) under the grant ALW Top-subsidy 854.10.002. Computational resources for this work were provided by the Netherlands Research Center for Integrated Solid Earth Science (ISES 3.2.5 High End Scientific Computation Resources).

REFERENCES

Backus, G.E. & Gilbert, F., 1968. The resolvingpower of gross Earth data, *Geophys. J. R. astr. Soc.,* **16,** 169–205.

Backus, G.E. & Gilbert, F., 1970. Uniqueness in the inversion of inaccurate gross Earth data, *Phil. Trans. R. Soc. Lond.,* **266,** 123–192.

Baker, J.A., Kornguth, P.J., Lo, J.Y., Williford, M.E. & Floyd, C.E., Jr, 1995. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon, *Radiology,* **196,** 817–822.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M.A. and F.R.S., *Phil. Trans. R. Soc. Lond.,* **53,** 370–418.

Beghein, C., Trampert, J. & van Heijst, H.J., 2006. Radial anisotropy in seismic reference models of the mantle, *J. geophys. Res.,* **111,** B02303, doi:10.1029/2005JB003728.

Bellman, R.E., 1961. *Adaptive Control Processes,* Princeton Univ. Press.

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition,* Oxford Univ. Press.

Crotwell, H.P., Owens, T.J. & Ritsema, J., 1999. The taup toolkit: flexible seismic travel-time and ray-path utilities, *Seismol. Res. Lett.,* **70,** 154–160.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function, *Math. Contl., Signal. Syst.,* **2,** 304–314.

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,** 28 841–28 857.

de Wit, R.W.L., Trampert, J. & van der Hilst, R.D., 2012. Toward quantifying uncertainty in travel time tomography using the null-space shuttle, *J. geophys. Res.,* **117,** B03301, doi:10.1029/2011JB008754.

Duda, R.O., Hart, P.E. & Stork, D.G., 2001. *Pattern Classification,* Wiley, NY, USA.

Dziewonski, A.M. & Anderson, D.L., 1981. Preliminary reference Earth model, *Phys. Earth planet. Inter.,* **25,** 297–356.

Engdahl, E.R., van der Hilst, R.D. & Buland, R., 1998. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination, *Bull. seism. Soc. Am.,* **88,** 722–743.

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

, 2008. Summary of the Bulletin of the International Seismological Centre. ftp://colossus.iris.washington.edu/pub/docs/BulletinSummary_draft.pdf.

Jiang, X. & Adeli, H. ISC, 2005. Dynamic wavelet neural network model for traffic flow forecasting, *J. Transp. Eng.,* **131,** 771–779.

Käufl, P.J., Fichtner, A. & Igel, H., 2013. Probabilistic full waveform inversion based on tectonics regionalisation—development and application to the Australian upper mantle, *Geophys. J. Int.,* doi:10.1093/gji/ggs131.

Kennett, B., 1998. On the density distribution within the Earth, *Geophys. J. Int.,* **132,** 374–382.

Kennett, B. & Engdahl, E.R., 1991. Travel times for global earthquake location and phase association, *Geophys. J. Int.,* **105,** 429–465.

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

LeCun, Y., Bottou, L., Orr, G. & Muller, K., 1998. Efficient BackProp, in *Neural Networks: Tricks of the Trade,* eds Orr, G. & Muller, K., Springer.

Lee, S., Ruy, J.-H., Won, J.-S. & Park, H.-J., 1998. Determination and application of the weights for landslide susceptibility mapping using an artificial neural network, *Eng. Geol.,* **71,** 289–302.

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

Masters, G. & Gubbins, D., 2003. On the resolution of density within the Earth, *Phys. Earth planet. Inter.,* **140,** 159–167.

McLachlan, G.J. & Basford, K.E., 1988. *Mixture Models: Inference and Applications to Clustering,* Marcel Dekker.

Meier, U., Curtis, A. & Trampert, J., 2007a. Global crustal thickness from neural network inversion of surface wave data, *Geophys. J. Int.,* **169,** 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,** L16304, doi:10.1029/2007GL030989.

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. Res. Lett.,* **282,** 91–101.

Møller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning, *Neural Networks,* **6,** 525–533.

Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.,* **100,** 12 431–12 447.

Nabney, I.T., 2002. *Netlab: Algorithms for Pattern Recognition, Advances for Pattern Recognition,* Springer Verlag.

Nolet, G., 2008. *A Breviary of Seismic Tomography: Imaging the Interior of the Earth and Sun,* Cambridge Univ. Press, 344 pp.

Odom, M. & Sharda, R., 2002. A neural network model for bankruptcy prediction, in *Proceedings of the International Joint Conference on Neural Networks,* San Diego, CA, Vol. 2, pp. 163–168, IEEE Press.

Parker, R.L., 1994. *Geophysical Inverse Theory,* Princeton Univ. Press.

Poulton, M.M., 2002. Neural networks as an intelligence amplification tool: a review of applications, *Geophysics,* **67,** 979–993.

Rawlinson, N., Pozgay, S. & Fishwick, S., 2010. Seismic tomography: a window into deep Earth, *Phys. Earth planet. Inter.,* **178,** 101–135.

Resovsky, J. & Trampert, J., 2003. Using probabilistic seismic tomography to test mantle velocity-density relationships, *Earth Planet. Res. Lett.,* **215,** 121–134.

Rowley, H.A., 1998. Neural network-based face detection, *IEEE Trans. Pattern Anal. Mach. Intell.,* **20,** 23–38.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J., 1986. Learning internal representations by error propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* Vol. 1: Foundations, pp. 318–362, MIT Press.

Sambridge, M., 1999a. Geophysical inversion with a neighbourhood algorithm—I. Searching a parameter space, *Geophys. J. Int.,* **138,** 479–494.

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

Shahraeeni, M.S. & Curtis, A., 2011. Fast probabilistic nonlinear petrophysical inversion, *Geophysics,* **76,** 45–58.

Shahraeeni, M.S., Curtis, A. & Chao, G., 2012. Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data, *Geophysics,* **77,** 1–19.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* SIAM.

Tarantola, A. & Valette, B., 1982. Inverse Problems = Quest for Information, *J. Geophys.,* **50,** 159–170.

Valentine, A.P. & Trampert, J., 2012a. Data-space reduction, quality assessment and searching of seismograms: autoencoder networks for waveform data, *Geophys. J. Int.,* **189,** 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,** 1001–1012.

van der Baan, M. & Jutten, C., 2000. Neural networks in geophysical applications, *Geophysics,* **65,** 1032–1047.

Werbos, P.J., 1974. Beyond regression: new tools for prediction and analysis in the behavioral sciences, *PhD thesis,* Harvard Univ., Cambridge, MA, USA.

Williams, P.M., 1996. Using neural networks to model conditional multivariate densities, *Neural Comput.,* **8,** 843–854.

APPENDIX A: DATA PRE-PROCESSING

We process the input data **x** to have zero mean and unit variance for each input neuron x_i, which is commonly referred to as standardizing (Bishop 1995):

$$\bar{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_i^n \tag{A1}$$

$$\mathrm{Var}(x_i) = \frac{1}{N-1}\sum_{n=1}^{N}\left(x_i^n - \bar{x}_i\right)^2, \tag{A2}$$

where $i = 1, \ldots, I$ denotes the input units (Fig. 1) and $n = 1, \ldots, N$ labels the patterns in the training data set. We can then apply a linear transformation so that we obtain a set of rescaled variables given by

$$\tilde{x}_i^n = \frac{x_i^n - \bar{x}_i}{[\mathrm{Var}(x_i)]^{1/2}}. \tag{A3}$$

Note that as long as the same linear transformation is applied to every pattern in the data set, the information content of the data set is not altered.

In addition to pre-processing the input, we find that it is beneficial to pre-process the target data and thus perform a similar operation as in eq. (A3) to the target data. Obviously, this linear transformation turns the targets into dimensionless numbers. Therefore, once the network is trained we reverse the linear transformation in eq. (A3) and apply it to the Gaussian kernel means $\boldsymbol{\mu}_j$ and standard deviations σ_j in the MDN output (eq. 6). By doing so, these parameters are given in the true physical dimensions of the earth model parameters. We do not correct the standard deviations σ_j for the translation in eq. (A3), however, as the variance of a probability distribution is invariant under translations. The mixing coefficients α_j are not transformed, as they are dimensionless and sum to one due to the application of the softmax function (Bishop 1995).

Note that the validation and test sets are pre-processed following eq. (A3) by using the mean x_i and the variance $\mathrm{Var}(x_i)$ calculated for the training data set (eqs A1 and A2). Thus the same transformation is applied to the three different synthetic data sets. The same is true for the observed data.