Geophysical Journal International

Geophys. J. Int. (2020) **223,** 1741–1757
Advance Access publication 2020 August 08
GJI Seismology

doi: 10.1093/gji/ggaa328

Probabilistic neural network tomography across Grane field (North Sea) from surface wave dispersion data

S. Earp,[1] A. Curtis,[1,2] X. Zhang[1] and F. Hansteen[3]

[1] *University of Edinburgh, School of Geosciences, Edinburgh, UK. E-mail: stephanie.earp@ed.ac.uk*
[2] *Institute of Geophysics, ETH Zurich, Switzerland*
[3] *Equinor ASA, Bergen, Norway*

Accepted 2020 July 5. Received 2020 June 22; in original form 2019 August 19

SUMMARY

Surface wave tomography uses measured dispersion properties of surface waves to infer the spatial distribution of subsurface properties such as shear wave velocities. These properties can be estimated vertically below any geographical location at which surface wave dispersion data are available. As the inversion is significantly non-linear, Monte Carlo methods are often used to invert dispersion curves for shear wave velocity profiles with depth to give a probabilistic solution. Such methods provide uncertainty information but are computationally expensive. Neural network (NN) based inversion provides a more efficient way to obtain probabilistic solutions when those solutions are required beneath many geographical locations. Unlike Monte Carlo methods, once a network has been trained it can be applied rapidly to perform any number of inversions. We train a class of NNs called mixture density networks (MDNs), to invert dispersion curves for shear wave velocity models and their non-linearized uncertainty. MDNs are able to produce fully probabilistic solutions in the form of weighted sums of multivariate analytic kernels such as Gaussians, and we show that including data uncertainties as additional inputs to the MDN gives substantially more reliable velocity estimates when data contains significant noise. The networks were applied to data from the Grane field in the Norwegian North sea to produce shear wave velocity maps at several depth levels. Post-training we obtained probabilistic velocity profiles with depth beneath 26 772 locations to produce a 3-D velocity model in 21 s on a standard desktop computer. This method is therefore ideally suited for rapid, repeated 3-D subsurface imaging and monitoring.

Key words: Inverse theory; Neural networks, fuzzy logic; Surface waves.

1 INTRODUCTION

Seismic surface waves travel around the surface of the Earth but are sensitive to heterogeneity in elastic properties within the subsurface. Different frequencies of surface waves travel at different speeds since they depend mainly on the shear wave velocity structure at different depths. Surface wave tomography uses this property (called dispersion) to infer the spatial distribution of subsurface shear velocities over global scales (Woodhouse & Dziewonski 1984; Trampert & Woodhouse 1995; Shapiro & Ritzwoller 2002; Zhou *et al.* 2006; Meier *et al.* 2007a,b), regional scales (Montagner & Jobert 1988; Curtis & Woodhouse 1997; Curtis *et al.* 1998; Ritzwoller & Levshin 1998; Devilee *et al.* 1999; Villasenor *et al.* 2001; Simons *et al.* 2002) and reservoir scales (Bussat & Kugler 2011; de Ridder & Dellinger 2011; Mordret *et al.* 2014).

Surface wave tomography is often performed using a two-step inversion scheme (Trampert & Woodhouse 1995; Ritzwoller *et al.* 2002). In step 1, traveltimes of surface waves between pairs of known locations are measured at various fixed periods, then used to create geographical phase or group velocity maps at each period using 2-D tomography. In step 2, the dispersion properties (speed of the waves at different periods—often referred to as a dispersion curve) at each point on the 2-D map are then inverted to estimate a 1-D shear wave velocity profile with depth below that point. The 1-D velocity profiles beneath many geographical locations can then be placed side-by-side and interpolated to create a 3-D model of the subsurface.

Both of the two-step surface wave inverse problems are non-linear. They can be solved approximately by partially linearized (Bodin & Sambridge 2009), or fully non-linear (Rawlinson *et al.* 2014; Galetti *et al.* 2015, 2016) Monte Carlo methods. These types of approaches provide relatively robust estimates of the range of possible shear wave velocity structures with depth that are consistent with the measured surface wave speeds (often referred to as the solution *uncertainty*) by using the Markov chain Monte Carlo (McMC) algorithm to perform the inversions in a Bayesian framework. However, all existing sampling based methods, including the direct (one-step) 3-D Monte Carlo tomography method of

Zhang *et al.* (2018), are extremely demanding computationally. If large data sets are to be inverted rapidly while maintaining our ability to assess post-inversion uncertainties without making linearizing approximations to the Physics, different methods are needed to speed up fully non-linear inversions.

We take an alternative approach and use neural networks (NNs) to perform non-linear inversion of the phase velocities of Rayleigh-type Scholte surface waves (we refer to these simply as Rayleigh waves) for subsurface shear wave velocity over length scales ∼1-10 km. NNs approximate a non-linear mapping between two parameter spaces. The mapping is inferred from a set of examples of inputs and corresponding outputs of the true mapping (these examples are called *training* data). Using certain types of NN-based methods, uncertainties in the mapping can be output by the network. NNs are therefore useful for problems where the forward mapping is well known or simple to calculate (in order to construct many training data) but the inverse mapping is complex or costly to determine directly. In such cases training data can be generated by applying the forward mapping to many sets of model parameter values, after which the NN can be trained to map in the inverse direction, taking the measurable data as input and outputting model parameter estimates and their uncertainties.

Once trained, NNs can be applied to calculate the mapping for any input parameters extremely efficiently. For this reason NNs have become increasingly popular for solving geophysical problems in recent years. Applications include well-log analysis (Aristodemou *et al.* 2005; Maiti *et al.* 2007), first arrival picking (Murat & Rudman 1992; McCormack *et al.* 1993), fault detection (Araya-Polo *et al.* 2017; Huang *et al.* 2017) and velocity analysis (Roth & Tarantola 1994; Calderón-Macías *et al.* 2000; Araya-Polo *et al.* 2018). However all of these methods provide only deterministic estimates of the inverse problem solution (in most cases, the mean model estimate). NNs can also be used in a Bayesian sense to give fully probabilistic solutions. They were first used in Geophysics to estimate Bayesian uncertainties by Devilee *et al.* (1999) who inverted surface wave phase and group velocities for large-scale subsurface velocity structure and interface depths. They inverted regional dispersion curves for discretized probability distributions of crustal thickness across Eurasia using histogram and median networks, and analysed the trade-off between crustal thickness and velocity structure. Meier *et al.* (2007a,b) improved this method by using mixture density networks (MDNs) to give continuous probabilistic estimates of global crustal thickness and crustal velocity structure. A MDN is a type of network that maps an input vector to a probability density function (pdf) rather than to a single set of output values (Bishop 1995). Since the work of Meier *et al.* (2007a,b), MDNs have been used to perform petrophysical inversion of surface wave data for global water content and temperature in the mantle transition zone (Meier *et al.* 2009), inversion of industrial seismic data sets for subsurface porosity and clay content (Shahraeeni & Curtis 2011; Shahraeeni *et al.* 2012), inference of the Earth's 1-D global average structure using body-wave traveltimes (De Wit *et al.* 2013) and for earthquake detection and source parameter estimation (Käufl *et al.* 2014, 2015).

To produce a 3-D shear wave velocity versus depth model on any scale using the two-step method, the inverse problem for structure with depth must be solved at many geographical locations (usually many thousands) over the area of interest. McMC inversion methods are computationally expensive and it is generally impractical to apply them in cases where parameters or data sets are large, where computational efficiency and processing time are usually limiting constraints due to the need to forward model many samples (of the order of thousands or millions) at each location. On the other hand, once trained, NNs and MDNs can often solve such inverse problems in seconds with no additional sampling. In addition, in cases where we wish to monitor changes in the subsurface, the same network can be applied rapidly to repeated data measurements, enabling the possibility of near-real time monitoring provided that the inputs to the networks can be produced rapidly from the raw measured data (Cao et al., 2020). Our aim herein is to investigate whether this is possible in practice.

In what follows we first introduce NNs and MDNs and how they can be used to invert Rayleigh wave phase velocities for models of 1-D shear wave velocity with depth. We discuss the effect of data uncertainty and how to incorporate this within a NN, then apply trained networks to data from the Grane field in the Norwegian North Sea to create 2-D shear wave velocity maps of specific depth intervals. We compare the results from the MDN to non-linearized McMC methods, and thus prove that MDN surface wave inversion methods are both efficient and robust at the scale of reservoirs.

2 METHOD

2.1 Grane data

Grane is an offshore oil field in the Norwegian North Sea. A permanent reservoir monitoring (PRM) system was installed in 2014 over approximately 50 km^2 of the Grane seabed (Thompson *et al.* 2015). Ambient seismic noise is recorded continuously at the field using four-component sensors—three-component geophones (Vertical, North and East) and a hydrophone. The data used in this study was preprocessed according to the protocol of Zhang *et al.* (2020), summarized as follows. Data from the vertical and hydrophone components were selected over a 6.5-hr interval. The data were bandpass-filtered between 0.35 and 1.5Hz and data from every pair of stations are cross-correlated using overlapping half-hour recording sections, then correlations are stacked over the full 6.5-hr interval. Cross-correlations of hydrophone and vertical component noise mainly contain information about Rayleigh-type waves. Phase velocities were automatically picked for the cross-correlation of each station-pair. Seventeen phase velocity maps and their corresponding standard deviation (uncertainty) maps were produced using eikonal tomography for periods between 0.6 and 2.2 s at 0.1 s intervals over a grid of cells with cell size 50 m × 50 m grid. Fig. 1(a) shows four examples of the phase velocity maps at periods 0.7, 1.0, 1.3 and 1.9 s and their corresponding uncertainties.

Zhang *et al.* (2020) perform 1-D, 2-D and 3-D McMC tomography over the Grane field to produce maps of the shear velocity structure with depth. However, McMC solutions are relatively slow to compute as they require ∼10^6 3-D or ∼10^9 1-D forward modelling simulations to obtain robust results, and for 4-D applications the set of simulations needs to be performed for every repeat survey. A more efficient method to carry out 3-D tomography is desirable.

Continuous seismic monitoring is a relatively new field, enabled by PRM systems and high bandwidth transfer and compute resources. To fully utilize the large volumes of data collected by PRM systems efficient algorithms are important. Whilst data from ambient noise surface wave interferometry does not generally give sufficient spatial resolution at typical depths of industrial operations it is useful for overburden monitoring. Effects such as fracturing,

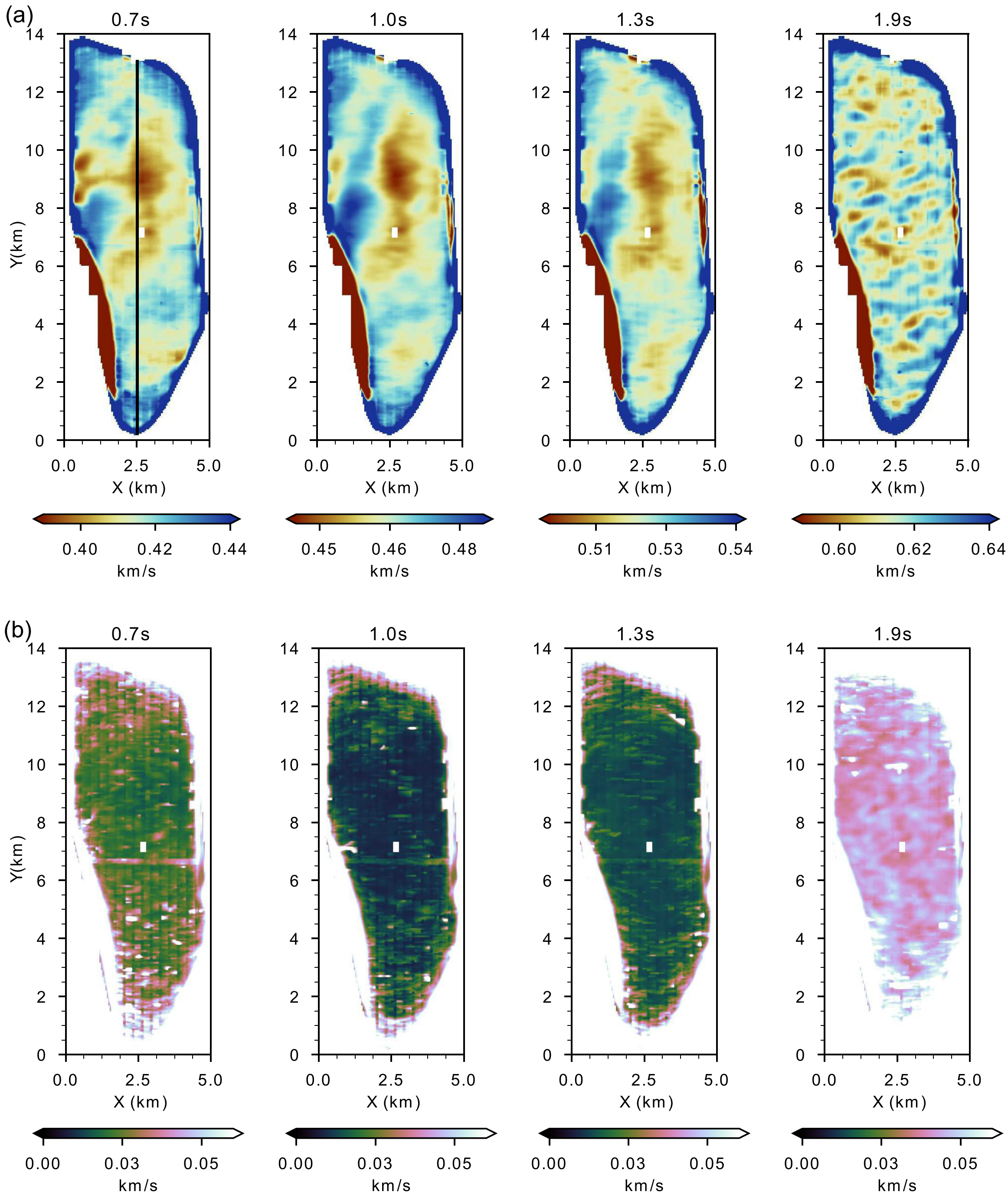

Figure 1. (a) A selection of four phase velocity maps used to compute discretized dispersion curves. Periods shown are 0.7, 1.0, 1.3 and 1.9 s. (b) Maps of estimated standard deviation of uncertainties in the phase velocity at each location. Velocities and uncertainty colour scales are saturated at either end to prevent domination of outliers, and to highlight structure across the field. The vertical black line in the top-left-hand plot shows the location of a cross section shown in other figures.

fault-reactivation and fluid migration or leakage happen on relatively short time scales, and it is important to detect these as early as possible. There is also a potential application at CO_2 storage sites for containment monitoring (Stork et al., 2018). On a longer timescale, months to years, geomechanical effects due to pressure depletion and reservoir compaction can be observed by changes to shallow *S*-wave and *P*-wave velocities. Although a rapid detection system is not needed as urgently in that case, frequent repeat measurements could help improve the signal-to-noise ratio and may enable a better understanding of how these effects evolve over time.

2.2 Bayesian Inference

We wish to solve the surface wave inversion problem in a probabilistic framework to find the Bayesian posterior distribution of subsurface velocity structure parameters **m** that fit the given data **d**,

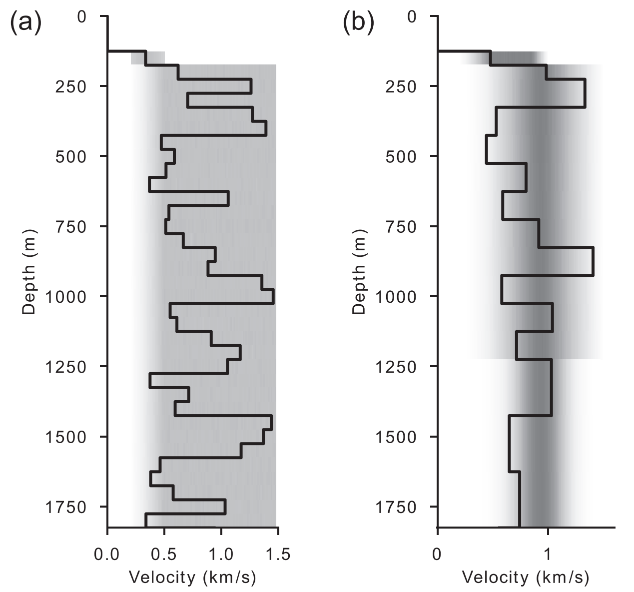

Figure 2. (a) Initial distribution of velocity structures created with a piecewise-constant discretization over depth. (b) Distribution of velocity structures created after averaging structures in (a) over larger depth intervals. Grey-scale shows the probability density distribution, darker colours represent higher density of velocity structures, and the black line is an example of a randomly selected velocity structure in each panel which also illustrates the depth intervals used in cases (a) and (b).

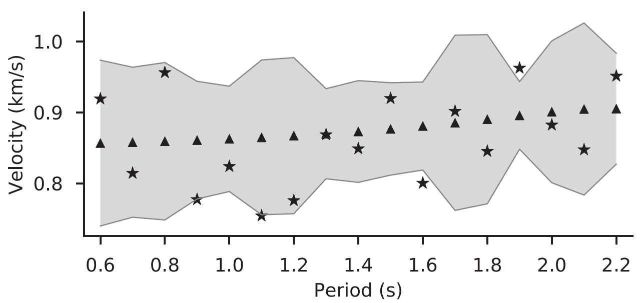

Figure 3. Graph showing a synthetic dispersion curve **d** (triangles) compared to a dispersion curve with added noise $\tilde{\mathbf{d}}$ (stars). The grey shaded area is the standard deviation **u** from eq. (10).

written as $p(\mathbf{m}|\mathbf{d})$. This is defined as (e.g. Tarantola 2005):

$$p(\mathbf{m} \mid \mathbf{d}) = k\, p(\mathbf{d} \mid \mathbf{m})\, p(\mathbf{m}), \tag{1}$$

where $p(\mathbf{m})$ represents the prior probability density on the velocity parameter space which describes information about **m** known prior to using data **d**, $p(\mathbf{d}|\mathbf{m})$ is known as the likelihood and represents the conditional probability of measuring data **d** given the velocity parameters **m**, and k is a normalization constant. In multidimensional problems where the dimensionality of **m** is greater than 1, we often wish to infer the posterior inversion information about a single parameter with index i and hence must calculate the marginal posterior distribution $p(m^i|\mathbf{d})$. This is obtained by integrating over all parameters m^j that are not of interest:

$$p(m^i \mid \mathbf{d}) = \int_{\forall m^j \neq m^i} p(\mathbf{m} \mid \mathbf{d})\, dm^j \tag{2}$$

In this study, we focus on estimating marginal distributions $p(m^i|\mathbf{d})$.

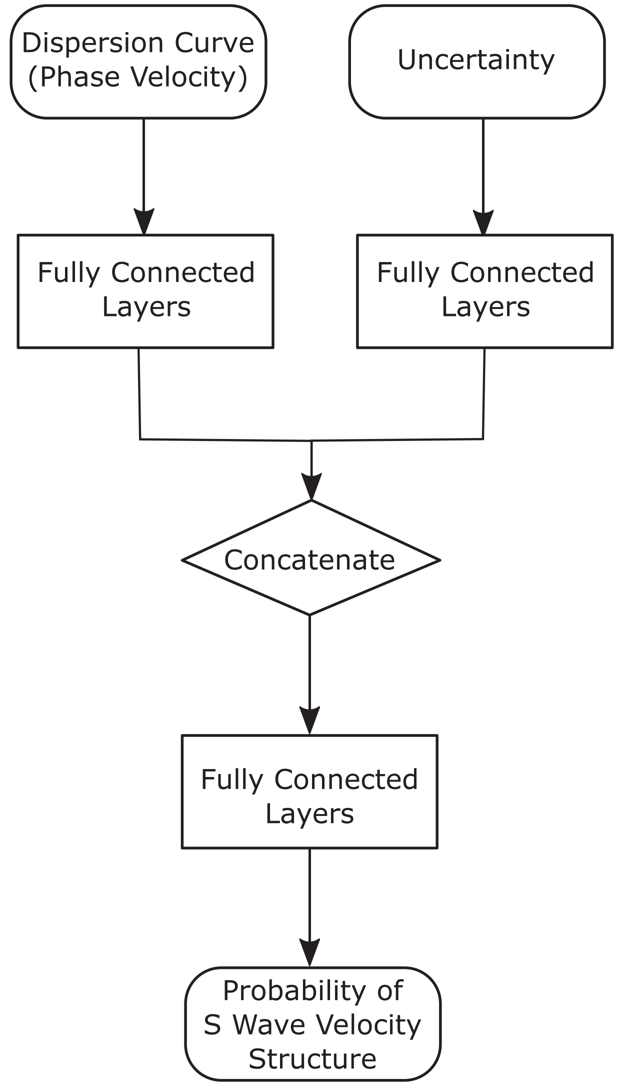

Figure 4. Diagram of network used to include standard deviation estimates in the input vector, used for the Variable-Noise-MDN. Rounded edged boxes represent inputs/outputs of network. Squared edges boxes represent one or more fully connected layers within the model where the internal model weights are optimized during training. The structure of these is described in Table A2. The diamond box represents the concatenation of layers: this step involves no new weights and simply concatenates the outputs of the previous layers. The arrows represent the direction of flow of data through the network.

2.3 MDNs

A NN can be *trained* to represent the arbitrary non-linear mapping between the spaces of input data **d** and output parameters **m** by presenting the network with a set of N training pairs $T = \{(\mathbf{d}_i, \mathbf{m}_i) : i = 1, \ldots, N\}$ and minimizing a cost function that measures the difference between the NN output and the defined output, often called the 'error'. For example if the set of training velocity structures $\mathbf{m}_i$ are distributed according to the prior pdf, then a network trained to output $\mathbf{m}_i$ given input $\mathbf{d}_i$ by minimizing the sum-of-squared errors across set T will output an approximation to the mean of the Bayesian posterior distribution $p(\mathbf{m} \mid \mathbf{d})$ when presented with data **d** (Bishop 1995). By contrast, in this paper we use a class of NNs called MDNs. These provide a framework for modelling complete probability distributions. They are trained on

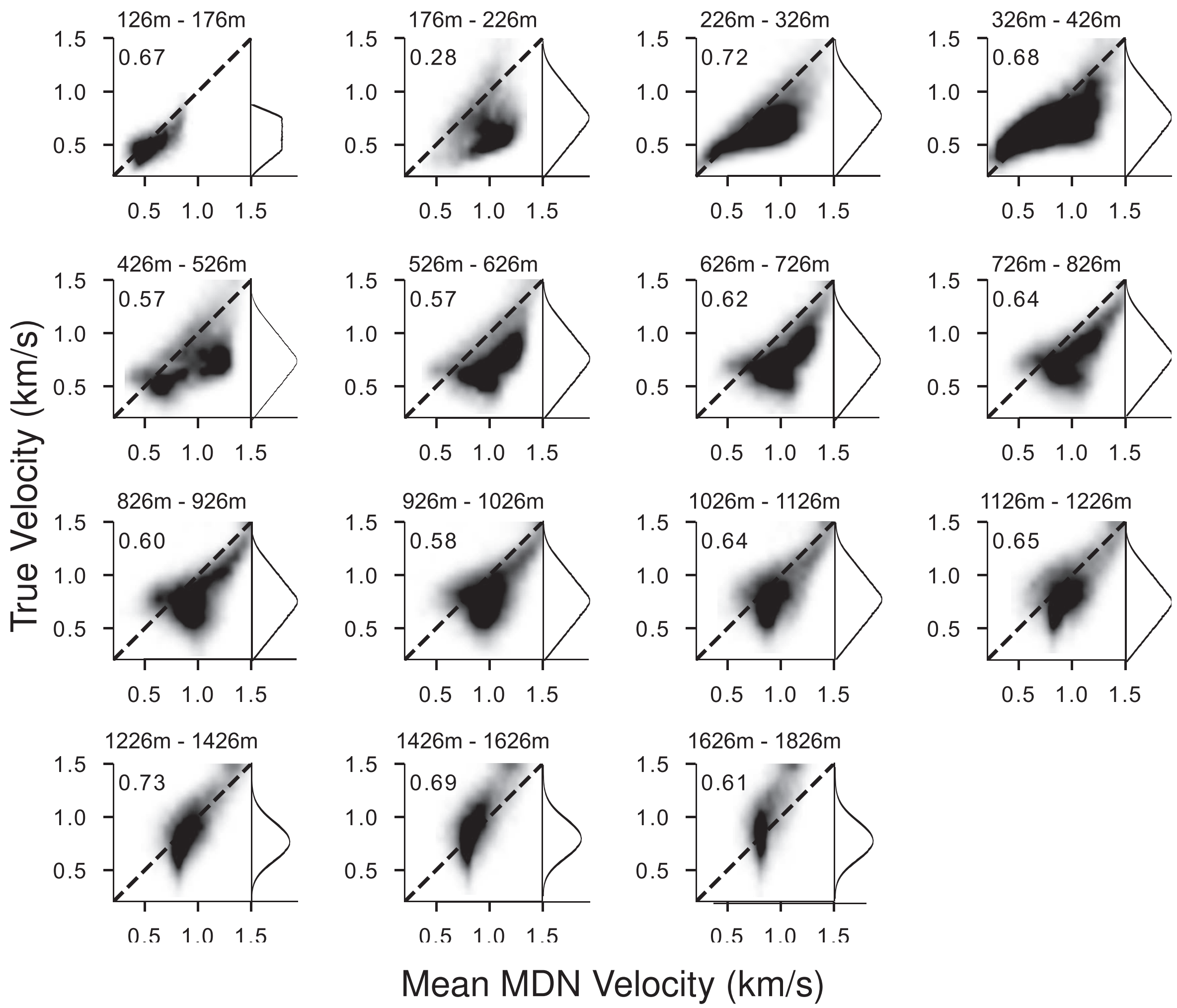

Figure 5. Mean of the posterior marginal pdfs from Fixed-Noise-MDN inversions, versus the true value of velocity for each velocity structure in the set of smooth models. Each graph represents a different depth interval as indicated above the graph. The black solid line on the right-hand side of each graph shows the prior distribution on the training set. The corresponding Pearson correlation coefficient R is given in the top left-hand corner of each graph.

the same set T of data–velocity structure pairs, but instead of providing the mean estimate of the velocity structure, they provide an estimate of the Bayesian posterior probability distribution $p(\mathbf{m} \mid \mathbf{d})$. The estimate is parametrised by a mixture (sum) of Gaussian kernels

$$p(\mathbf{m} \mid \mathbf{d}) = \sum_{k=1}^{M} \alpha_k \Theta_k(\mathbf{m} \mid \mathbf{d}), \tag{3}$$

where α_k are amplitude parameters that attach relative importance to each Gaussian kernel, M is the number of Gaussians in the mixture and Θ_k are Gaussian density functions given by

$$\Theta_k(\mathbf{m} \mid \mathbf{d}) = \frac{1}{(2\pi)^{c/2}\sigma_k^c(\mathbf{d})} \exp\left\{-\frac{(m_k - \mu_k(\mathbf{d}))^2}{2\sigma_k^2(\mathbf{d})}\right\} \tag{4}$$

where c is the dimensionality of $\mathbf{m}$.

The set of mixture parameters α_k, means μ_k and standard deviations σ_k fully define the set of Gaussian kernels and hence the output of the MDN. Training an MDN thus requires that we create a way to predict appropriate values for these parameters given any input data. For this task we use a standard feed-forward NN which contains a set number of layers and nodes. At each layer the inputs of each node are weighted and summed before being passed through a function that induces non-linearity in the mapping. This provides an output value that can become the input for all units in the following layer. These weights are adjusted during training to provide the optimum mapping. The number of mixtures M dictates the complexity of the final probability distribution, and the number of network outputs is given by $(c + 2) \times M$ compared with the output of a NN such as a multilayer perception or convolutional network that have c outputs. The number of kernels that should be used depends on the complexity of the problem; however, as long as more kernels than necessary are used, an accurate number is not required as the network can reduce the amplitude of any mixture parameter to near zero for redundant kernels (Bishop 1995). For a full description of MDN and NN structures see Bishop (1995), or in a geophysical context see Meier *et al.* (2007a) or Shahraeeni & Curtis (2011).

During training the internal weights of the network are adjusted to maximize the likelihood of the desired pdf given the training data. The cost function minimized is the negative log likelihood function (Bishop 1995)

$$E_{MDN} = -\sum_{n=1}^{N} ln\left\{\sum_{k=1}^{M} \alpha_k(d_n)\Theta_k(m_n \mid d_n)\right\}. \tag{5}$$

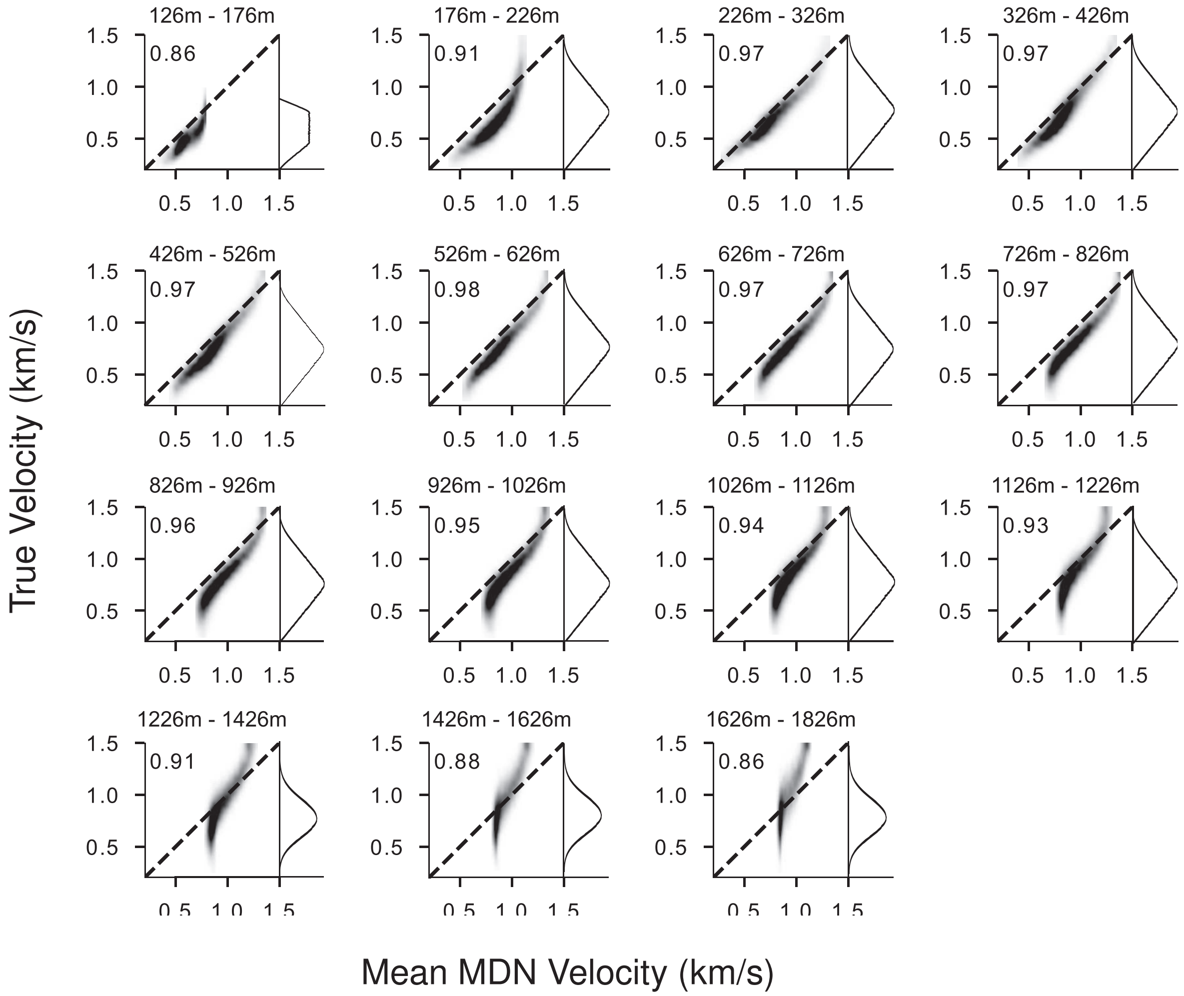

Figure 6. Mean of the posterior marginal pdfs from Variable-Noise-MDN inversions, versus the true value of velocity for each velocity structure in the set of smooth models. Each graph represents a different depth interval as indicated above the graph. The black solid line on the right-hand side of each graph shows the prior distribution on the training set. The corresponding Pearson correlation coefficient R is given in the top left-hand corner of each graph.

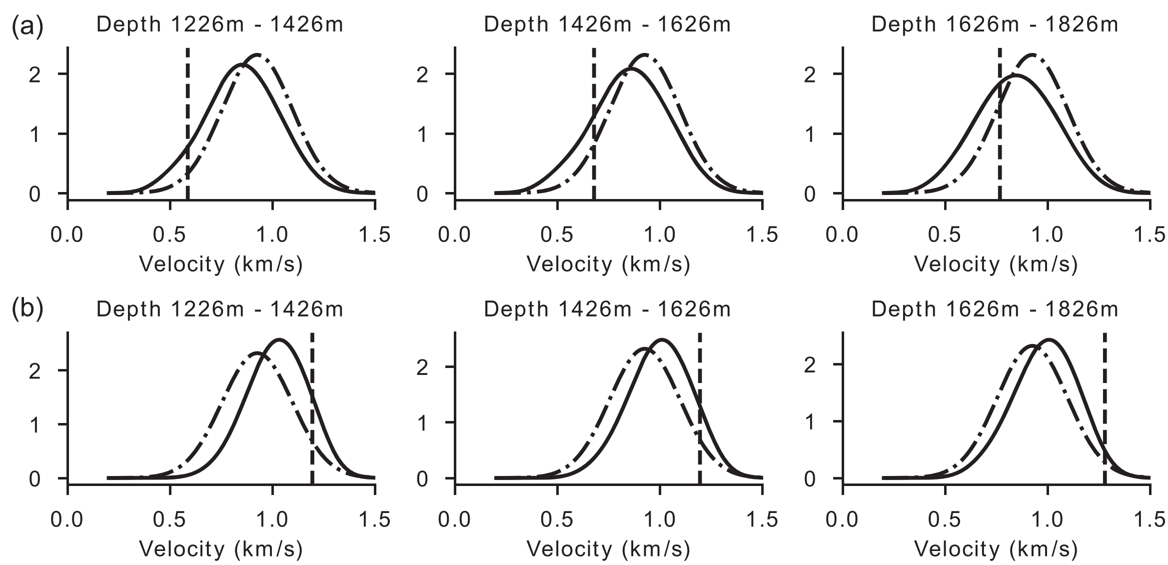

Figure 7. Individual probability density functions (pdfs) for depths below 1226 m for two synthetic velocity structures in (a) and (b), respectively. The solid line is the marginal posterior pdf from the MDN, the vertical dashed line is the true velocity value, and the dot–dashed line is the prior pdf.

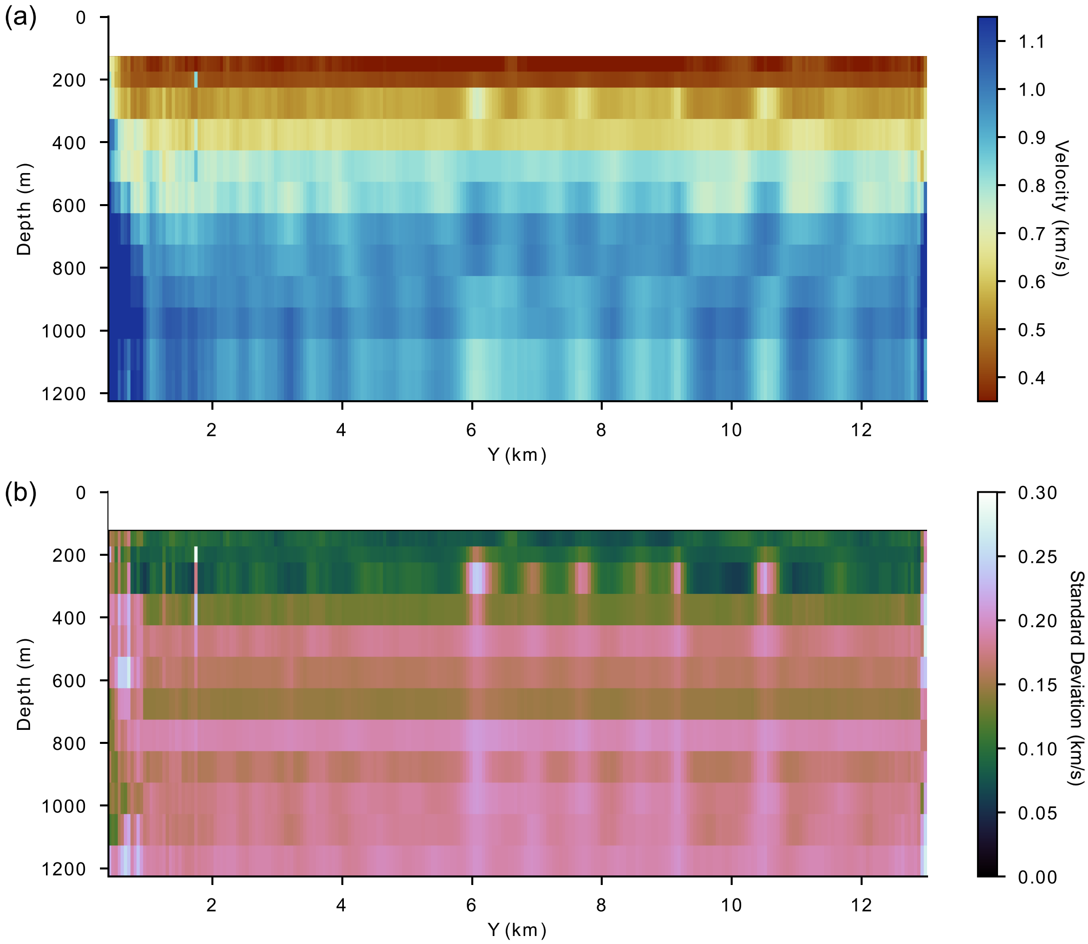

Figure 8. (a) Mean shear velocity cross-section, and (b) corresponding posterior standard deviation cross-sections from the Fixed-Noise-MDN inversion. The top white layer represents the water layer where shear velocity is zero.

We train multiple networks with different network structures, and each network's parameters are randomly initialized before training begins. The networks are then combined using a weighted average of network outputs in order to improve generalization and prediction accuracy (Dietterich 2000). The weights for each network are determined by the cost function E_{MDN} evaluated over the so-called test set—a portion of the data set removed before training and used to test the network once training has completed. An approximation to the posterior probability distribution of a set of velocity parameters $\mathbf{m}$ given data $\mathbf{d}$ is thus given by

$$p(\mathbf{m} \mid \mathbf{d}) \simeq \sum_{k=1}^{L} \sum_{j=1}^{c} \frac{E_k \alpha_{kj}}{\sum_{l=1}^{L} E_l} \Theta_{kj}(\mathbf{m} \mid \mathbf{d}), \tag{6}$$

where

$$E_k = -\exp(E_{\mathrm{MDN},k}) \tag{7}$$

and where L is the number of networks in the ensemble, $E_{\mathrm{MDN},\,k}$ is the cost function value of the kth network, α_{kj} is the jth weighting parameter of the kth network and Θ_{kj} is the jth Gaussian kernel of the kth network. Once the networks have been trained, the outputs can be used to estimate the posterior probability distributions using eqs (4) and (6). This gives a more complete description of the family of velocity structures that are consistent with the data than does the output of a multilayer perception with only a deterministic output.

2.4 Creating a training set

The velocity structures $\mathbf{m}$ are parametrized as follows: each 1-D structure has a water layer of the true depth 126 m, followed by constant velocity layers every 25 m to a depth of 100 m below the water layer, then 50-m-thick layers down to 2000 m below the water layer, beneath which there is a homogeneous half-space. For each velocity structure the *S*-wave velocity of the top solid layer was selected randomly from the Uniform probability distribution $v_{\mathrm{top}} \sim U(0.2\,\mathrm{km\,s^{-1}}, 0.5\,\mathrm{km\,s^{-1}})$ to represent unconsolidated near-surface sediments. For a fundamental mode surface wave to be observed, the top solid layer must have the lowest velocity (Galetti *et al.* 2016), therefore the following layers were randomly selected from distribution $U(v_{\mathrm{top}}, 1.5\,\mathrm{km\,s^{-1}})$. We generated 1000 000 velocity structures and Fig. 2(a) shows the resulting distribution of velocities along with an example velocity structure.

The forward problem is solved for each of the generated velocity structures using the DISPER80 subroutines by Saito (1988) to obtain corresponding fundamental mode Rayleigh wave dispersion curves. The phase velocities were calculated for periods 0.6–2.2 s at 0.1 s intervals in order to match the range available from ambient

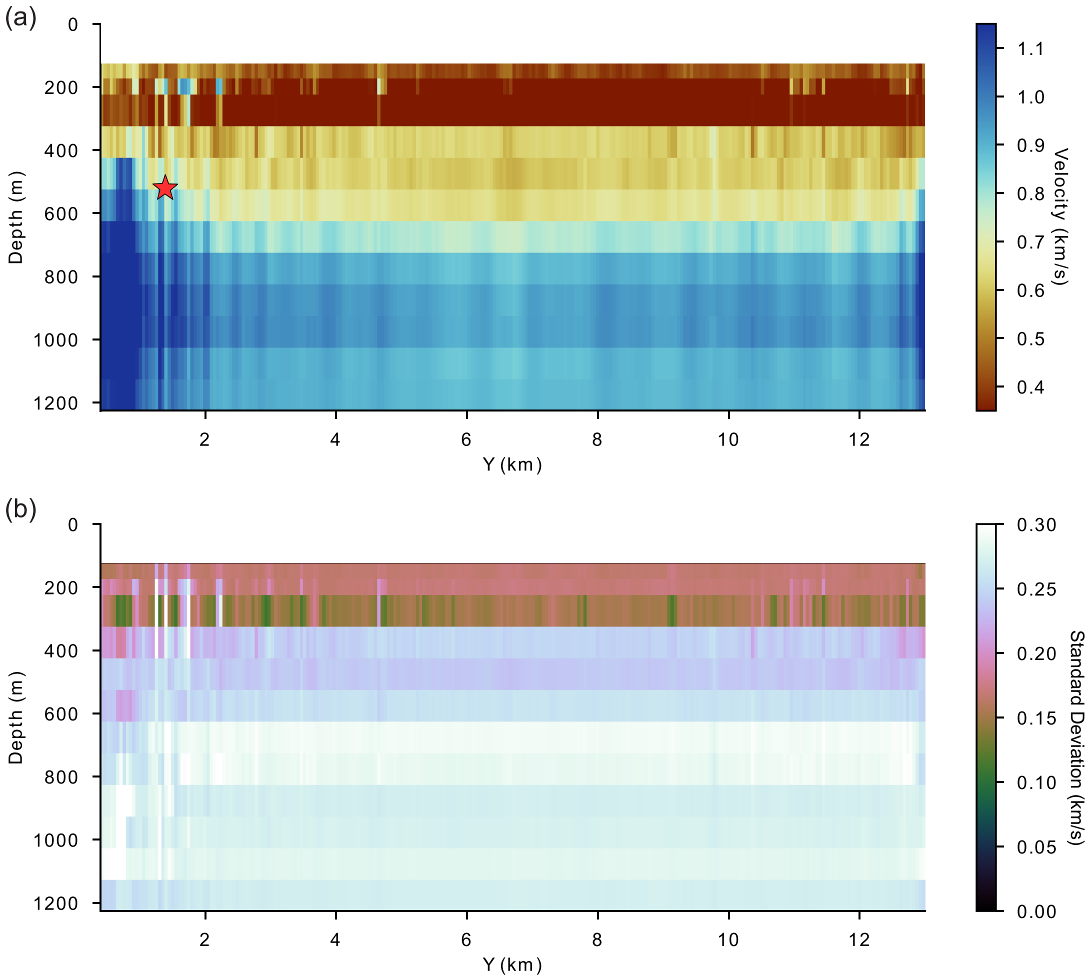

Figure 9. (a) Mean shear velocity cross-section, and (b) corresponding posterior standard deviation cross-sections from the Variable-Noise-MDN inversion. The top white layer represents the water layer where shear velocity is zero. Red star shows location of the joint pdf shown in Fig. 15.

noise recorded at Grane. The DISPER80 forward modeller needs *P*- and *S*-wave velocity and density for each layer in depth in order to calculate the phase velocities at each of any set of discrete periods. From our *S*-wave velocity structures calculated previously we computed a corresponding *P*-wave velocity v_p and the density ρ for each velocity layer based on typical values for sedimentary rocks using (Castagna *et al.* 1985; Brocher 2005)

$$v_p = 1.16v_s + 1.36 \tag{8}$$

$$\rho = 1.74v_p^{0.25}. \tag{9}$$

Rather than attempt to invert surface wave speeds at 17 periods for shear velocities in 40 depth layers, before training the velocity model is averaged over seventeen larger fixed-depth intervals (Fig. 2b). This averaging results in the prior distribution tending towards a Gaussian pdf. We then train networks to invert for the velocity in each of these larger depth intervals.

2.5 Uncertainties

In past work, uncertainty information about the data is only included by adding random Gaussian noise to the training data set (Devilee *et al.* 1999; Meier *et al.* 2007b; Shahraeeni & Curtis 2011; De Wit *et al.* 2013). Adding noise acts to regularize the network, helps to generalize when the network is inverting new data, and accounts for the data uncertainty in the Bayesian solution. However the disadvantage of such an approach is that noise added is assumed to be at a fixed level for all inversions, this is rarely the case in the real Earth. This means that when presenting the network with new data, updated uncertainty information for those particular data is not included in the inversion; indeed, that network would invert the data assuming that the incorrect data uncertainties still pertain.

By contrast, here the data uncertainty is included as an additional set of inputs to the network, so that the noise level can vary in amplitude between different data. This makes sense because uncertainty is in fact additional pertinent information for each inversion. To train the MDN the clean synthetically modelled data set is augmented with varying levels of noised data. For each data point in the dispersion curve in the original synthetic data set, a random percentage of noise ϵ is selected between the bounds outlined in Table 1 for six different Uniform distributions of ϵ. The noise is then added to the data according to

$$u_j = \epsilon \times d_j \tag{10}$$

$$\tilde{d}_j = \mathcal{N}(0, 1)u_j + d_j, \tag{11}$$

where u_j is the standard deviation value of the noised data $\tilde{d}_j$ and $\mathcal{N}(0, 1)$ is a Standard Normal distribution with mean 0 and standard

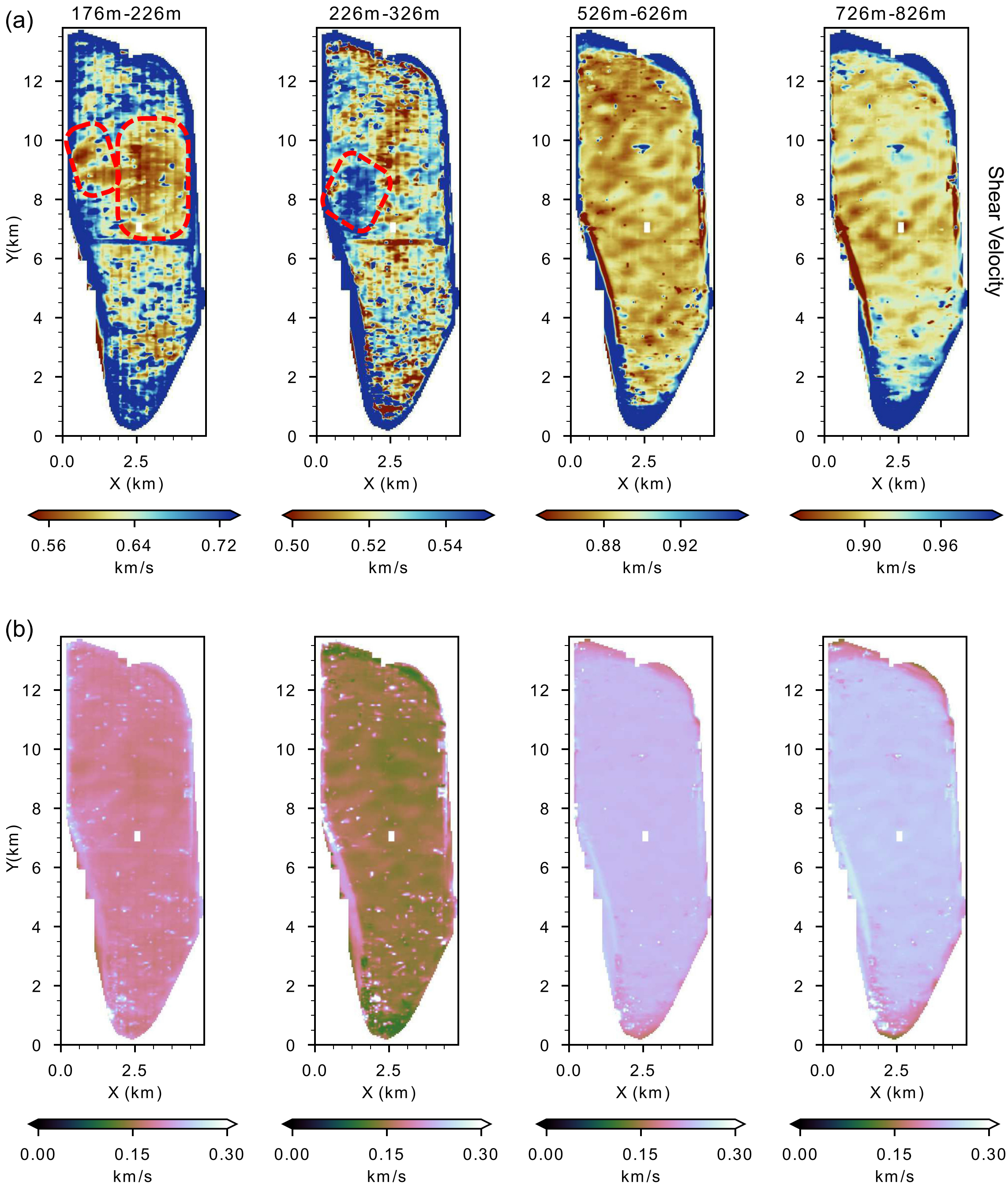

Figure 10. Fixed depth maps of (a) the mean and (b) the standard deviation of the shear velocity from Variable-Noise-MDN inversion of fundamental mode Rayleigh dispersion at depth slices 176–226 m, 226–326 m, 526–626 m and 726–826 m. The dotted red boxes are locations where the structures in the shear velocity maps are consistent with the phase velocity maps in Fig. 1(a).

deviation 1. An example of noised data and the randomly chosen noise level is shown in Fig. 3. We thus generate an augmented training set of data-velocity structure pairs $T_{uncer} = \{([\tilde{\mathbf{d}}_j, \mathbf{u}_j], \mathbf{m}_j) : j = 1, ..., N\}$, where our data consists of the noised dispersion curves $\tilde{\mathbf{d}}_j$ and their associated uncertainties $\mathbf{u}_j$. The final data sets T and T_{uncer} are then shuffled and split into a training set (90 per cent of training pairs) that is used to train the network for the optimum mapping, a validation set (5 per cent) used during training to check the network is not overfitting the training examples (see below), and a test set (5 per cent) which is used post training to assess the network performance on previously unseen data. This final assessment provides weights E_i for the network ensemble in eq. (6). Early stopping is used to prevent overfitting of the network to the data: this is where the cost function is periodically checked on the validation set during training. When the cost function stops decreasing it is assumed that the network is already fit to the training data but is no longer improving its generalization to new data. Training is then stopped.

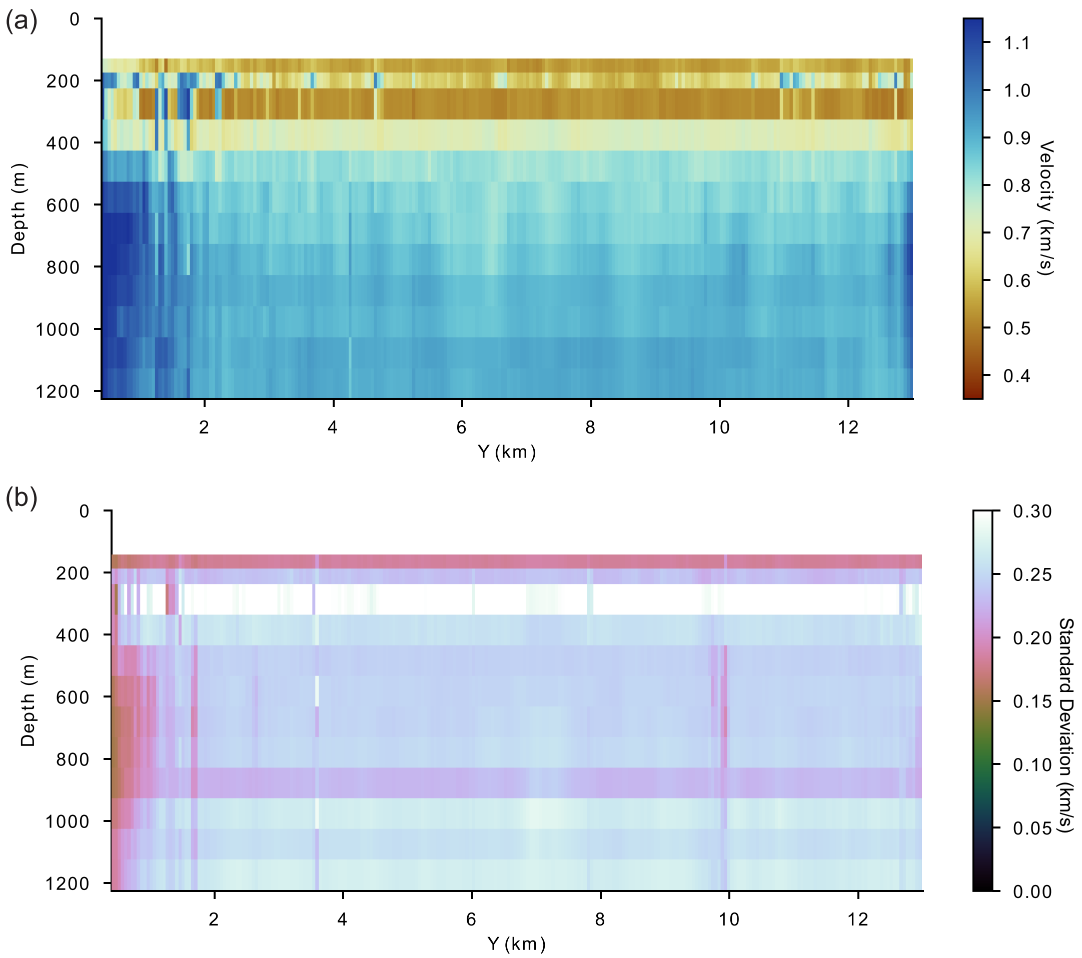

Figure 11. (a) Mean shear velocity cross-section, and (b) corresponding posterior standard deviation cross-sections from the Variable-Noise-MDN inversion of fundamental and first higher mode Rayleigh dispersion. The top white layer represents the water layer where shear velocity is zero.

3 RESULTS

3.1 Network design

Networks are trained for two different data sets: first a training set T in which data were perturbed by 10 per cent Gaussian noise is used to train what we refer to herein as a Fixed-Noise-MDN. This MDN does not include standard deviation in its input vector. A second training set T_{uncer} includes a variable standard deviation vector as described in the previous section, which is used to train what we refer to as a Variable-Noise-MDN. To include both the dispersion curve and their uncertainties in the latter networks two inputs are included as shown in Fig. 4. The dispersion curve is passed through two layers, whilst the standard deviation is input separately and passed through one layer. The outputs of these two layers is then concatenated before being passed through two more layers to output the parameter vector that defines the probability distribution of the shear wave velocity structure. For the Fixed-Noise-MDN a single input is used (dispersion curve only) and there is no need for concatenation.

Separate MDNs are trained for each depth interval in the velocity structure defined in Fig. 2(b). For each interval approximately 40 networks are trained from which we select for the ensemble the 10 networks with the lowest cost value across the validation set. The weights and biases are randomly initialized using the Glorot uniform initializer (Glorot & Bengio 2010) for each training run, and we use different sizes of layers in the different networks to create diversity. The different layer sizes were determined using a form of Bayesian optimization using the Python library hyperopt (Bergstra *et al.* 2015). In Appendix A, we describe the network configurations trained. The networks each use a Gaussian mixture with 15 kernels, so by using an ensemble of 10 networks a total of 150 kernels potentially contribute to each posterior distribution. However, we found that normally only 3 or 4 kernels with different means and standard deviations were assigned significantly non-zero amplitudes by each individual network.

3.2 Network evaluation

A set of 100 000 synthetic velocity structures to which no network has previously been exposed were then created. These simulate relatively smooth velocity structures by not allowing the velocity to vary more than 400 m s^{-1} between neighbouring depth intervals. Corresponding data are created using the DISPER80 forward modeller, to which 10 per cent Gaussian noise was added. For each depth interval in the velocity structure we apply the MDN ensemble to each of the 100 000 synthetic data and calculate the mean of each

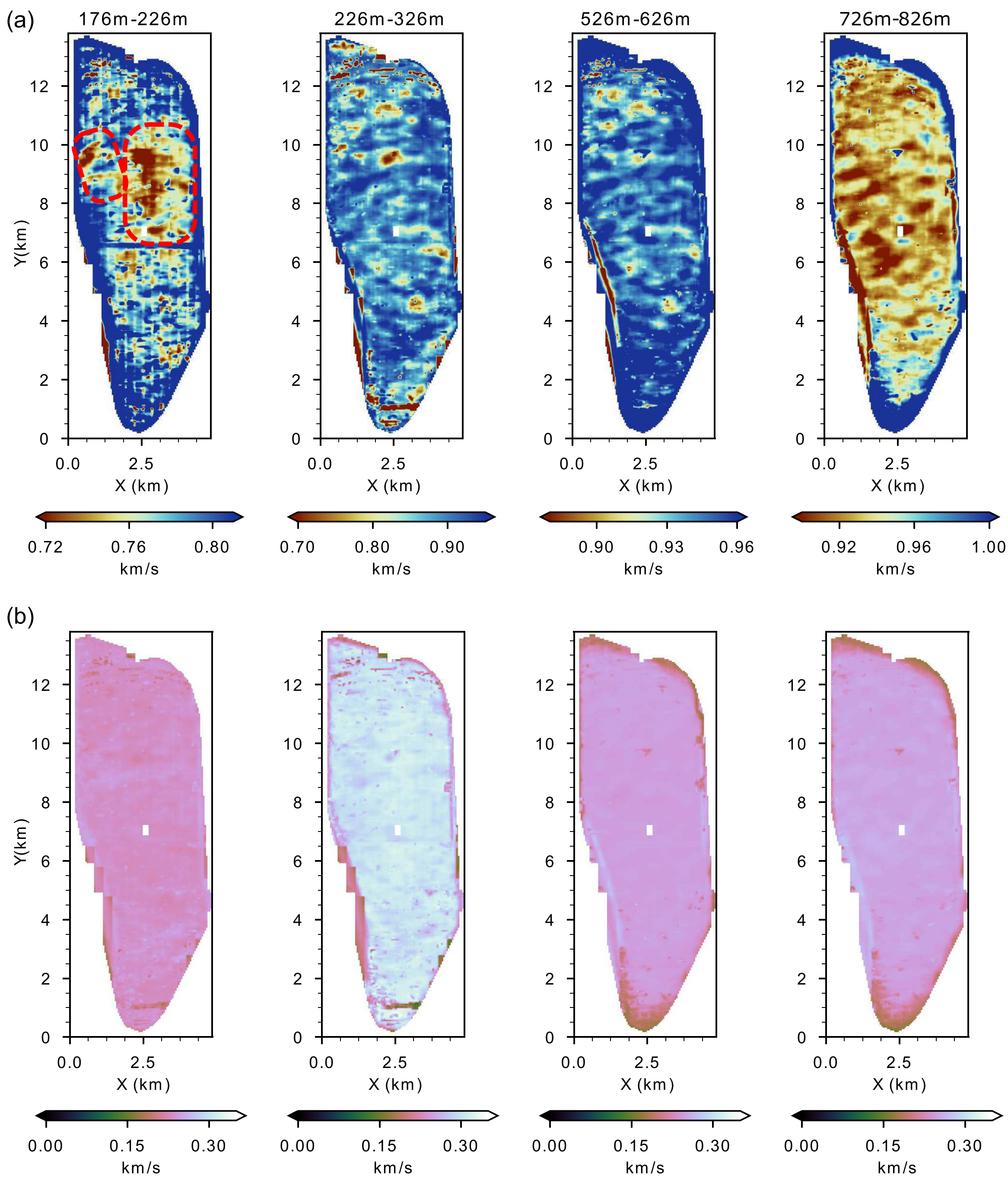

Figure 12. Fixed depth maps of (a) the mean and (b) the standard deviation of the shear velocity from Variable-Noise-MDN inversion of fundamental and first higher mode Rayleigh dispersion at depth slices 176m-226m, 226m-326m, 526m-626m, 726m-826m.

posterior marginal pdf $\bar{\mu}_{\text{post}}$

$$\bar{\mu}_{\text{post}} = \sum_{i=1}^{M} \alpha_i \mu_i . \tag{12}$$

The correlation between the mean of the posterior and the true target value for each data vector can be used to evaluate the performance of the networks when presented with new data. This evaluation does not use all of the information contained in each posterior pdf, but does provide a practical way to begin to evaluate network performance. Fig. 5 shows the means of the posterior pdf of the fundamental mode Rayleigh wave Fixed-Noise-MDN inversions versus the true velocity values across all of the synthetic smooth velocity models, for each depth interval. The corresponding Pearson correlation coefficient, R, is shown in the top-left-hand corner of the plot. The plots show a clear tendency for the mean of the network to overestimate the true velocity value. When the same inversions are performed using the Variable-Noise-MDNs (Fig. 6) the correlation between the mean MDN velocities and the true velocities improves at every depth level. The additional information

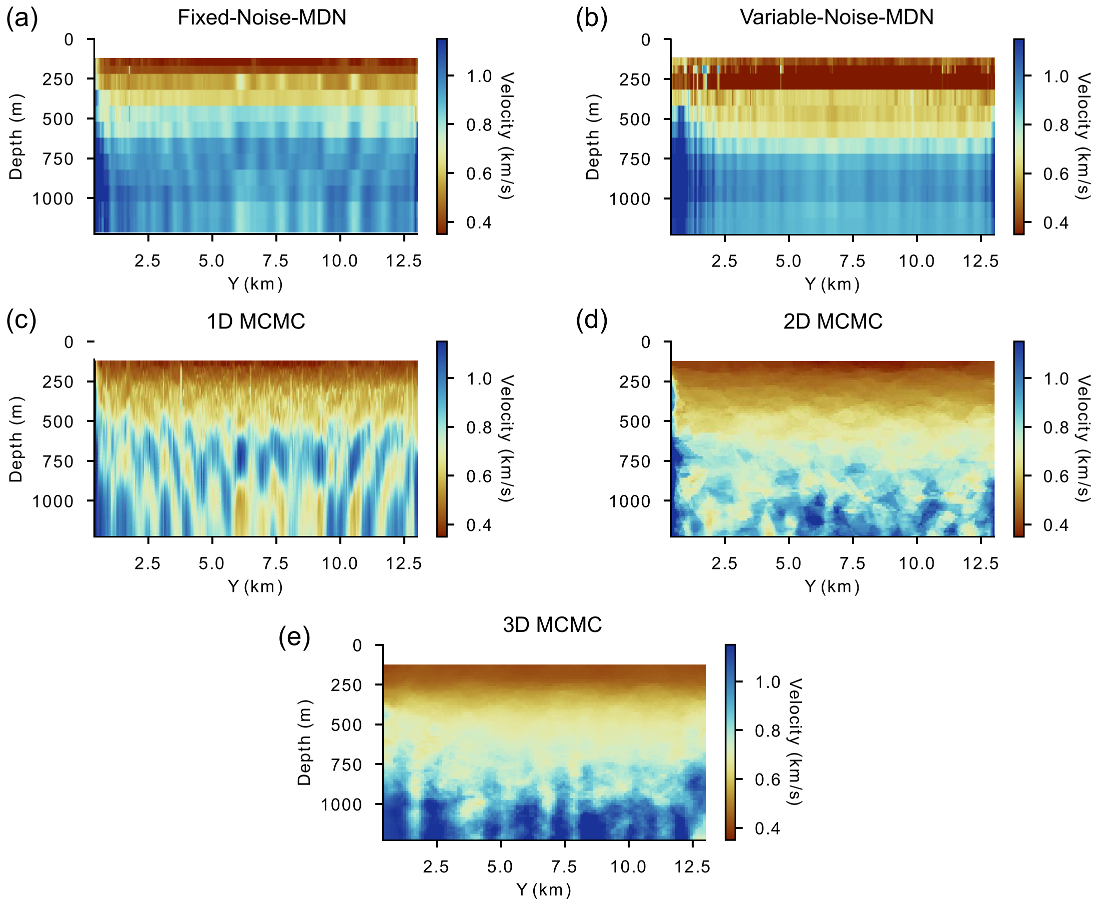

Figure 13. Mean shear velocity along the cross-section in Fig. 1(a) from (a) MDN inversions using a training set with added Gaussian noise of fixed standard deviation, (b) MDN inversions using estimated data uncertainties as added input data, (c) independent 1-D Monte Carlo inversions, (d) a single 2-D Monte Carlo inversion and (e) a 3-D Monte Carlo inversion, where results in (c), (d) and (e) are from Zhang *et al.* (2020). The top white layer represents the water layer, where the shear velocity is zero.

provided to the network that describes uncertainties in the data results in a significantly more accurate estimate of the velocity structure.

A full example inversion of noised synthetic data corresponding to two velocity structures using Fixed-Noise-MDNs is shown in the Supplementary Material as Figure S1. We observe significant biases in the estimated shear velocity structure such that velocities at several depths lie outside of the main range of uncertainty. Figure S2 shows the results when we train the network to take the data uncertainties as explicit inputs using Variable-Noise-MDNs: the biases are entirely removed. This demonstrates that it is extremely important to train networks to make explicit use of the additional information contained in data uncertainty estimates.

The plots in Fig. 6 allow us to compare how the networks perform at different depth levels. The performance of the networks decrease with depth, and at the deeper levels (1626–1826 m) the mean of the Variable-Noise-MDN tends towards the mean of the prior. Fig. 7 shows an example of the marginal posterior pdf for two synthetic velocity structures at depths below 1226 m. In both plots the true velocity structure is far away from the mean of the prior distribution and the predicted marginal posterior distribution remains close to the prior: this shows that at these depths the networks are unable to add any information to the prior pdf given the data presented to the network. This could be due to the depth sensitivity of the periods of Rayleigh waves selected for the inversion or the training set is not suitable at these depths. For this reason the following results are only shown down to a depth of 1226 m.

3.3 Field data

The final trained MDNs are applied to invert Rayleigh wave phase velocities from the Grane field in the Norwegian North Sea. Dispersion curves were extracted at each grid point producing 26 772 dispersion curves to be inverted for 1-D depth–velocity structures. The standard deviations shown in Fig. 1(b) were extracted at each point and used as the standard deviation vector input to the Variable-Noise-MDNs (Fig. 4). Figs 8 and 9 show the mean and associated standard deviations (representing uncertainty) of the posterior pdf estimated at the location of the black line in Fig. 1(a) from Fixed-Noise-MDNs and Variable-Noise-MDNs, respectively. The standard deviation, σ_{post}, of the weighted mixture of Gaussians is computed by:

$$\sigma_{\text{post}} = \sqrt{\sum_{i=1}^{M} \alpha_i \left(\sigma_i^2 + \mu_i^2\right) - \left(\sum_{i=1}^{M} \alpha_i \mu_i\right)^2} \tag{13}$$

Both plots of the mean show a reasonably similar structure: a near-surface low velocity layer down to 300 m, then an increased velocity down to 600 m, with yet higher velocities below this. However, the

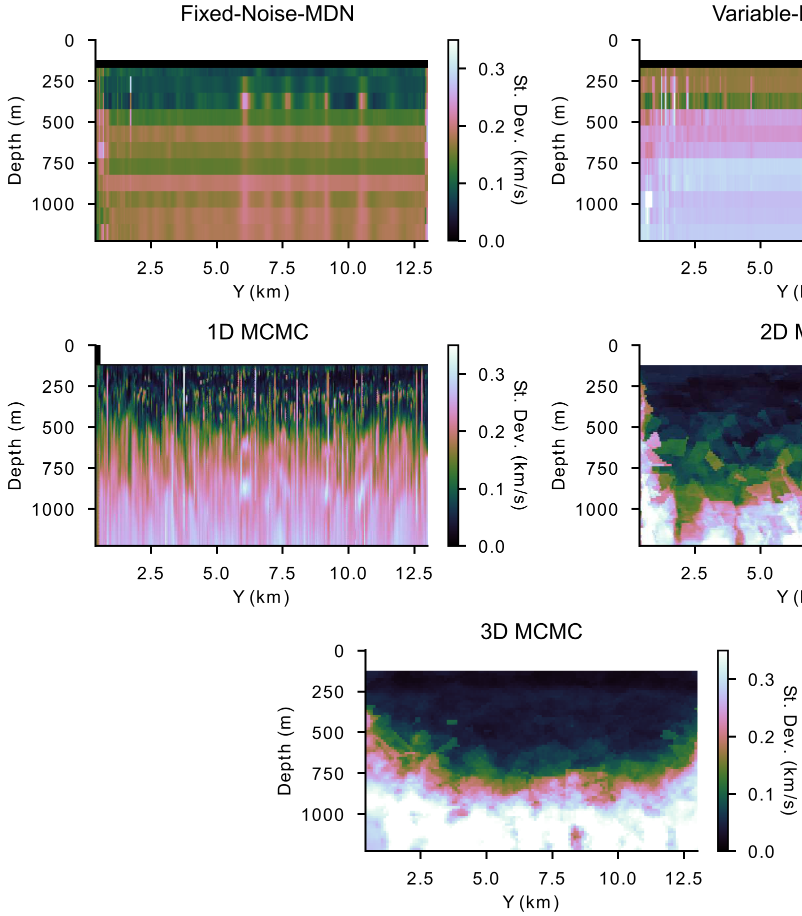

Figure 14. Standard deviation of shear velocity along the cross-section in Fig. 1(a) from (a) MDN inversions using a training set with added Gaussian noise of fixed standard deviation, (b) MDN inversions using estimated data uncertainties as added input data, (c) independent 1-D Monte Carlo inversions, (d) a single 2-D Monte Carlo inversion, and (e) a 3-D Monte Carlo inversion, where results in (c), (d) and (e) are from Zhang *et al.* (2020). The top white layer represents the water layer, where the shear velocity is zero.

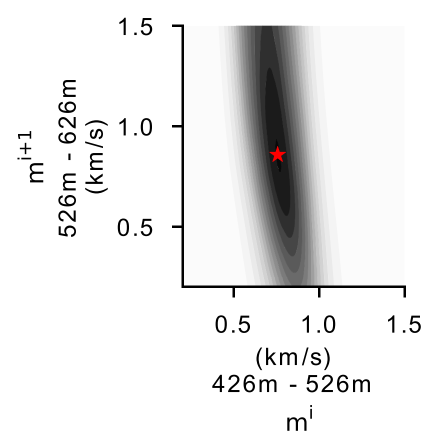

Figure 15. Joint pdf comparing the velocity trade-off between two adjacent layers m^i and m^{i+1} at depths given in the axis labels. The red star represents the mean velocity shown in Fig. 9(a).

layers are more distinct in the inversion using the Variable-Noise-MDNs. Fig. 8(a) from the Fixed-Noise-MDN shows a higher variability in the velocity below 600 m than does the mean in Fig. 9(a), and the velocity highs in Fig. 8(a) coincide with higher uncertainties in Fig. 8(b). When networks are trained including the full standard deviation information (Fig. 9) these velocity highs disappear so that the mean velocity and uncertainties are laterally smoother across the section. We therefore now focus on the Variable-Noise-MDN results.

Fig. 10 shows the mean and standard deviation horizontal depth slices from the Variable-Noise-MDNs. In the near surface maps (126–326 m) the results show similar structures to those in the phase velocity maps in Fig. 1(a) at short periods, for examples within the dotted red boxes in Fig. 10(a). The deeper maps (536–826 m) show structures similar to that of the longer period phase velocity maps, but also a higher standard deviation (Fig. 10b) than shallower layers. As a result, the shear velocity variation in these deeper structures falls within their standard deviation, suggesting that they might not represent true structure.

The method outlined above can easily be extended to joint inversion of fundamental and first higher mode data by adding two additional vectors: the vector of first higher mode phase velocity values generated from the velocity structures in the original training set and a vector of their associated uncertainties. For the method outlined above this would result in 34 additional scalar inputs, two 17 dimensional vectors. Fig. 11 shows the cross-section results and Fig. 12 shows the results from four depths layers, 126–176 m, 226–326 m, 426–526 m, 626–726 m, from Variable-Noise-MDN joint inversion. The same features seen in the shallow layer of Fig. 10(a) are seen in the shallow layer of the joint inversion, highlighted by the red dashed boxes in Fig. 12(a). However, the velocities are on

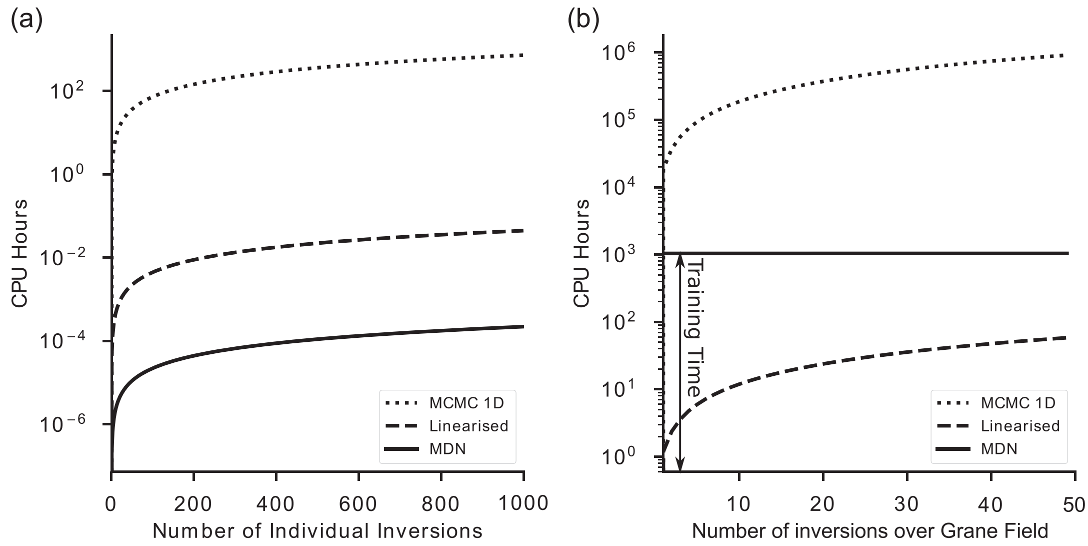

Figure 16. Plots showing the CPU hour time for (a) one inversion per location and (b) the inversion over the entire Grane field using MDNs, linearized inversion and Monte Carlo 1-D (MCMC 1-D) methods. The typical time taken to train the MDN (including the forward modelling runs) is included in (b) but not in (a).

Table 1. Table of percentage range of noise added to data set in each of six different noise scenarios. Uncertainty is added to each datum according to eqs (10) and (11). In the first scenario uncertainties are zero.

	Percentage noise (per cent)	
Noise scenario	Min	Max
1	0	0
2	0	5
3	4	14
4	10	15
5	3	10
6	0	15

average higher than the fundamental mode-only MDN inversions and the standard deviations are larger. In addition, the depth slice at 226–326 m is entirely different to the corresponding slice in Fig. 10(a). Fig. 11(a) shows that the low velocities observed in the top layers of the Variable-Noise-MDN fundamental model inversion (Fig. 9a) no longer exist in the joint inversion with higher modes, showing the latter waves appear to have added additional information to the inversion. However, we are less confident about the quality of the higher mode dispersion measurements than those from the fundamental mode, so we include this result as a demonstration, but in Section 4 we focus mainly on the fundamental mode results.

4 DISCUSSION

We inverted Rayleigh wave phase dispersion curves for subsurface shear wave velocities using MDNs trained with added Gaussian noise at a fixed standard deviation to simulate average data uncertainties, and a second type of MDN with the data uncertainty vector included as an additional input that could include a variable level of noise uncertainty on the data. We showed that to invert noisy data for reliable velocity structures the standard deviation estimates should be included in the network.

A constant number of fixed depth–velocity intervals were used in each MDN inversion, leading to inversions for effective medium (averaged) shear velocities for each fixed depth interval. A trans-dimensional network inversion would have had to include varying depths and number of layers which would significantly increase the dimensionality of the network inversion problem and require a much larger training set and more complex network structure. This in turn would increase training time and the memory needed for training, and would likely make the network outputs less stable and reliable since the posterior would effectively be emulating the inverse function in a higher dimensional space. As discussed in Käufl *et al.* (2016), it is difficult to get meaningful posterior pdfs from MDNs in high dimensional problems. For our intended application (to test our ability to rapidly monitor the overburden of a permanently instrumented field), the inversion for effective medium parameters over fixed depth intervals was sufficient.

4.1 Comparison with Monte Carlo methods

We compare the Noise- and Variable-Noise-MDN inversion results to the McMC results of Zhang *et al.* (2020). Fig. 13 shows the mean shear velocity cross sections of Figs 8(a) and 9(a) along with the same cross sections from 1-D, 2-D and 3-D trans-dimensional McMC. Despite comparing a trans-dimensional result from Monte Carlo methods with fixed-depth layer results from MDNs, all cross sections show a similar, approximately three-layered structure. The 1-D MCMC (Fig. 13c) most represents the networks trained using the Fixed-Noise-MDNs (Fig. 13a) as both contain vertical velocity anomalies in the deeper part of the section. The Variable-Noise-MDN has smoother variations laterally but also has a thicker near surface velocity layer and the second layer extends deeper into the section (to ∼700 m); this is more representative of the 2-D and 3-D MCMC results (Figs 13d and e). This is confirmed by examining the mean-squared difference (MSD) between the mean of each MDN inversion and the Monte Carlo inversions in Table 2: the Fixed-Noise-MDN has a lower MSD compared to the 1-D MCMC inversion and the Variable-Noise-MDN has a lower MSD compared

Table 2. Table showing the mean-squared difference between the Fixed-Noise- and Variable-Noise-MDN inversion cross sections, and the Markov Chain Monte Carlo cross sections of Zhang *et al.* (2020).

	Fixed-Noise-MDN	Variable-Noise-MDN
1-D MCMC	0.0018	0.0025
2-D MCMC	0.0025	0.0026
3-D MCMC	0.0021	0.0019

to the 3-D MCMC inversion. This implies that by adding uncertainties to the MDN training we allow smoothness in the mean estimates which the 3-D MCMC results suggest is reasonable across Grane. The standard deviations of the results are shown in Fig. 14. It can be seen that the MDN inversions have similar levels of uncertainty (standard deviation) as the 1-D MCMC inversion. However the Variable-Noise-MDN has a slightly larger standard deviation than the 1-D MCMC, this is expected as MDNs generally produce conservative estimates of the uncertainty, as discussed in Käufl *et al.* (2016).

4.2 Joint posterior pdfs

The results in the previous section are created from the 1-D marginal posterior pdf $p(m^i|\mathbf{d})$ of the shear velocity in each layer independent of other velocities in each 1-D profile. The correlations between velocities at different depths cannot be derived from such results. To estimate correlations it is necessary to analyse the joint posterior pdf $p(m^i, m^{i+1}|\mathbf{d})$, which can be constructed from the product of the conditional and marginal pdfs.

$$p(m^i, m^{i+1} \mid \mathbf{d}) = p(m^i \mid \mathbf{d}) \times p(m^{i+1} \mid m^i, \mathbf{d}) \tag{14}$$

The marginal pdfs $p(m^i|\mathbf{d})$ are given by the results shown in the previous sections. New networks are trained to estimate the conditional pdfs $p(m^{i+1}|m^i, \mathbf{d})$ by extending the input vector of the data with the velocity to which we want to condition our data: in this example this is the velocity of the layer above the one being estimated.

Fig. 15 shows the results from the location shown by the red star in the Grane cross-section from Fig. 9(a). The plot shows a weak negative correlation, representing the weak trade-off between velocities in subsequent layers. This is likely to be because a relatively coarse parametrization (compare that in Figs 2a and b) was used over depth for the inverse problem. If a finer parametrization was used, such trade-offs would emerge more strongly as demonstrated by Meier *et al.* (2007a). We found the construction of the joint pdfs to be less stable than the construction of the 1-D marginal results shown in this paper. If we reverse the process to construct $p(m^{i+1}, m^i|\mathbf{d})$ the results are not identical. This is probably due to an inadequate training data set, so it is likely that to construct higher dimensional pdfs a larger, more representative training set is required.

4.3 Inversion speed

Post-training, NNs invert new data extremely rapidly: in this study it took approximately 21 CPU seconds to invert all 26 772 locations. The results are compared to Monte Carlo methods which are known to be computationally expensive (Bodin & Sambridge 2009): the MCMC methods used to create the crosslines shown in Fig. 13 took approximately 186 CPU hours for 1-D, 206 CPU hours for 2-D and 4824 CPU hours for 3-D inversions. Despite the higher vertical resolution of results from MCMC methods (since the parametrization over depth varies in those inversions), the compute-time for inversions is between 4 and 6 orders of magnitude larger than for trained MDNs. It should be noted that in this case using NNs makes the inversion extremely efficient because the problem is decomposed into multiple 1-D inversions across the field. For problems where there are only a few data vectors to be inverted this might not be an efficient method once network training time has been taken into account.

A comparison of time per inversion of an individual location for 1-D MDN, 1-D Monte Carlo and a 1-D linearized inversion is shown in Fig. 16(a). Monte Carlo inversion is computationally the most expensive, and MDN inversion is two orders of magnitude faster than even linearized inversion. However, in this comparison we only accounted for the speed of the inversion which is not the full computational expense involved in using NNs. Training a network before the inversion takes significant computational time: in this study we took 1280 CPU hours to train all of the networks used. It should be noted that training the network needs only to be done once, and hence is independent of number of locations to be inverted; therefore inverting more locations renders the MDN inversion method more computationally efficient. Fig. 16(b) compares the CPU hours needed for monitoring-style repeated inversions across the full Grane field as performed in this study, including the time required for training MDNs. The initial cost of training a network before the first inversion is high, but thereafter repeated inversion of new data sets is nearly free. In comparison to 1-D MCMC methods, even accounting for the initial training period, NN methods are more efficient. Therefore, it would be possible to increase the size of the training set and the complexity of the network (which would increase training time) and still be able to produce results more cheaply than 1-D MCMC. The training set and network complexity could be increased even more if a monitoring scenario is considered as the upfront cost of training the network would pay for itself within a few inversions of data since the forward modelling and training only needs to be done once. Increasing the training set size would give a better representation of the data-model relationship and most likely improve the posterior pdf estimation of the MDN.

Linear inversion methods are computationally cheaper that MDN methods: in this case approximately 1000 inversions of the same field would be needed for the NN method to become cheaper. Surface wave tomography is a non-linear inversion problem and despite the linearized inversions involving fewer CPU hours they only provide approximate solutions, in particular for standard deviation estimates, due to their implicit assumption of incorrect (linear) physics. The NN method provides a fully probabilistic, fully non-linear solution that, once a network is trained, can be used to obtain rapid, repeated inversions.

5 CONCLUSION

We trained MDNs to invert fundamental mode Rayleigh wave dispersion curves for subsurface shear wave velocity using two different methods to represent data uncertainties. The MDNs give a fully probabilistic solution to this non-linear inverse problem giving comparable results to Monte Carlo solutions. We show that inputting data uncertainties explicitly to the network provides a more reliable solution estimate on noisy synthetic data, and a smoother result that is more similar to 3D Monte Carlo inversion results on field data. The same method is used for joint inversion of fundamental and first higher mode data. Once trained, the NN approach gives rapid results that can be repeatedly applied to similar types of data in monitoring scenarios.

ACKNOWLEDGEMENTS

The authors would like to thank the Grane license partners Equinor ASA, Petoro AS, ExxonMobil E&P Norway AS and ConocoPhillips Skandinavia AS for allowing us to publish this work. The views expressed in this paper are the views of the authors and do not necessarily reflect the views of the license partners. The authors thank the Edinburgh Interferometry Project sponsors (Equinor, Schlumberger Cambridge Research and Total) for supporting this research.

REFERENCES

Araya-Polo, M., Dahlke, T., Frogner, C., Zhang, C., Poggio, T. & Hohl, D., 2017. Automated fault detection without seismic processing, *Leading Edge,* **36**(3), 208–214.

Araya-Polo, M., Jennings, J., Adler, A. & Dahlke, T., 2018. Deep-learning tomography, *Leading Edge,* **37**(1), 58–66.

Aristodemou, E., Pain, C., De Oliveira, C., Goddard, T. & Harris, C., 2005. Inversion of nuclear well-logging data using neural networks, *Geophys. Prospect.,* **53**(1), 103–120.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D. & Cox, D.D., 2015. Hyperopt: a python library for model selection and hyperparameter optimization, *Comput. Sci. Discov.,* **8**(1), 014008.

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

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *J. geophys. Int.,* **178**(3), 1411–1436.

Brocher, T.M., 2005. Empirical relations between elastic wave speeds and density in the earth's crust, *Bull. seism. Soc. Am.,* **95**(6), 2081–2092.

Bussat, S. & Kugler, S., 2011. Offshore ambient-noise surface-wave tomography above 0.1 Hz and its applications, *Leading Edge,* **30**(5), 514–524.

Calderón-Macías, C., Sen, M.K. & Stoffa, P.L., 2000. Artificial neural networks for parameter estimation in geophysics, *Geophys. Prospect.,* **48**(1), 21–47.

Cao, R., Earp, S., de Ridder, S., Curtis, A. & Galetti, E., 2020. Near-real time near-surface 3D seismic velocity and uncertainty models by wavefield gradiometry and neural network inversion of ambient seismic noise, *Geophysics,* **85**(1), KS13–KS27, doi.org/10.1190/geo2018-0562.1.

Castagna, J.P., Batzle, M.L. & Eastwood, R.L., 1985. Relationships between compressional-wave and shear-wave velocities in clastic silicate rocks, *Geophysics,* **50**(4), 571–581.

Curtis, A., Trampert, J., Snieder, R. & Dost, B., 1998. Eurasian fundamental mode surface wave phase velocities and their relationship with tectonic structures, *J. geophys. Res.,* **103**(B11), 26 919–26 947.

Curtis, A. & Woodhouse, J.H., 1997. Crust and upper mantle shear velocity structure beneath the tibetan plateau and surrounding regions from interevent surface wave phase velocity inversion, *J. geophys. Res.,* **102**(B6), 11 789–11 813.

de Ridder, S. & Dellinger, J., 2011. Ambient seismic noise eikonal tomography for near-surface imaging at valhall, *Leading Edge,* **30**(5), 506–512.

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.

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, *J. geophys. Int.,* **195**(1), 408–422.

Dietterich, T.G., 2000. Ensemble methods in machine learning, in *Proceedings of the International Workshop On Multiple Classifier Systems,* pp. 1–15, Springer.

Galetti, E., Curtis, A., Baptie, B., Jenkins, D. & Nicolson, H., 2016. Transdimensional love-wave tomography of the British Isles and shear-velocity structure of the east Irish sea basin from ambient-noise interferometry, *J. geophys. Int.,* **208**(1), 36–58.

Galetti, E., Curtis, A., Meles, G.A. & Baptie, B., 2015. Uncertainty loops in travel-time tomography from nonlinear wave physics, *Phys. Rev. Lett.,* **114**(14), 148501.

Glorot, X. & Bengio, Y., 2010. Understanding the difficulty of training deep feed forward neural networks, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics,* pp. 249–256.

Huang, L., Dong, X. & Clee, T.E., 2017. A scalable deep learning platform for identifying geologic features from seismic attributes, *Leading Edge,* **36**(3), 249–256.

Käufl, P., Valentine, A., de Wit, R. & Trampert, J., 2015. Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition, *Bull. seism. Soc. Am.,* **105**(4), 2299–2312.

Käufl, P., Valentine, A.P., de Wit, R.W. & Trampert, J., 2016. Solving probabilistic inverse problems rapidly with prior samples, *J. geophys. Int.,* **205**(3), 1710–1728.

Käufl, P., Valentine, A.P., O'Toole, T.B. & Trampert, J., 2014. A framework for fast probabilistic centroid-moment-tensor determination inversion of regional static displacement measurements, *J. geophys. Int.,* **196**(3), 1676–1693.

Maiti, S., Krishna Tiwari, R. & Kümpel, H.-J., 2007. Neural network modelling and classification of lithofacies using well log data: a case study from ktb borehole site, *J. geophys. Int.,* **169**(2), 733–746.

McCormack, M.D., Zaucha, D.E. & Dushek, D.W., 1993. First-break refraction event picking and seismic data trace editing using neural networks, *Geophysics,* **58**(1), 67–78.

Meier, U., Curtis, A. & Trampert, J., 2007a. Global crustal thickness from neural network inversion of surface wave data, *J. geophys. 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), 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. Sci. Lett.,* **282**(1), 91–101.

Montagner, J.-P. & Jobert, N., 1988. Vectorial tomography. II. Application to the Indian Ocean, *J. geophys. Int.,* **94**(2), 309–344.

Mordret, A., Landès, M., Shapiro, N., Singh, S. & Roux, P., 2014. Ambient noise surface wave tomography to determine the shallow shear velocity structure at Valhall: depth inversion with a neighbourhood algorithm, *J. geophys. Int.,* **198**(3), 1514–1525.

Murat, M.E. & Rudman, A.J., 1992. Automated first arrival picking: a neural network approach, *Geophys. Prospect.,* **40**(6), 587–604.

Rawlinson, N., Fichtner, A., Sambridge, M. & Young, M.K., 2014. Seismic tomography and the assessment of uncertainty, in *Advances in Geophysics,* Vol. **55,** pp. 1–76, Elsevier.

Ritzwoller, M.H. & Levshin, A.L., 1998. Eurasian surface wave tomography: group velocities, *J. geophys. Res.,* **103**(B3), 4839–4878.

Ritzwoller, M.H., Shapiro, N.M., Barmin, M.P. & Levshin, A.L., 2002. Global surface wave diffraction tomography, *J. geophys. Res.,* **107**(B12), ESE–4.

Roth, G. & Tarantola, A., 1994. Neural networks and inversion of seismic data, *J. geophys. Res.,* **99**(B4), 6753–6768.

Saito, M., 1988. *DISPER80 : A Subroutine Package for the Calculation of Seismic Normal Mode Solutions,* Academic Press.

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.

Shapiro, N. & Ritzwoller, M., 2002. Monte-carlo inversion for a global shear-velocity model of the crust and upper mantle, *J. geophys. Int.,* **151**(1), 88–105.

Simons, F.J., Van Der Hilst, R.D., Montagner, J.-P. & Zielhuis, A., 2002. Multimode rayleigh wave inversion for heterogeneity and azimuthal anisotropy of the Australian upper mantle, *J. geophys. Int.,* **151**(3), 738–754.

Stork, A., Allmark, C., Curtis, A., Kendall, M., Huppert, H. & White, D., 2018. Assessing the potential to use repeated ambient

noise seismic tomography to detect CO2 leaks: Application to the Aquistore storage site, *Int. J. Greenhouse Gas Control,* **71,** 20–35, doi.org/10.1016/j.ijggc.2018.02.007.
Tarantola, A., 2005. *Inverse Problem Theory,* SIAM.
Thompson, M., Andersen, M., Elde, R., Roy, S. & Skogland, S., 2015. The startup of permanent reservoir monitoring for Snorre and Grane, in *Proceedings of the 77th EAGE Conference and Exhibition 2015*.
Trampert, J. & Woodhouse, J.H., 1995. Global phase velocity maps of Love and Rayleigh waves between 40 and 150 seconds, *J. geophys. Int.,* **122**(2), 675–690.
Villasenor, A., Ritzwoller, M., Levshin, A., Barmin, M., Engdahl, E., Spakman, W. & Trampert, J., 2001. Shear velocity structure of central Eurasia from inversion of surface wave velocities, *Phys. Earth planet. Inter.,* **123**(2-4), 169–184.
Woodhouse, J.H. & Dziewonski, A.M., 1984. Mapping the upper mantle: three-dimensional modeling of earth structure by inversion of seismic waveforms, *J. geophys. Res.,* **89**(B7), 5953–5986.
Zhang, X., Curtis, A., Galetti, E. & de Ridder, S., 2018. 3-D Monte Carlo surface wave tomography, *J. geophys. Int.,* **215**(3), 1644–1658.
Zhang, X., Hansteen, F., Curtis, A. & de Ridder, S., 2020. 1-D, 2-D, And 3-D Monte Carlo ambient noise tomography using a dense passive seismic array installed on the north sea seabed, *J. geophys. Res.,* **125**(2), e2019JB018552.
Zhou, Y., Nolet, G., Dahlen, F. & Laske, G., 2006. Global upper-mantle structure from finite-frequency surface-wave tomography, *J. geophys. Res.,* **111**(B4), doi:10.1029/2005JB003677.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Supplementary_material.pdf
Figure S1 1-D depth inversion result from Fixed-Noise-MDNs for two synthetic velocity structures with individual probability density functions (pdfs) shown for four depth levels. In the depth inversions dark colours represent areas of higher probability, each row of the posterior integrates to unity, and the black solid line is the true synthetic velocity structure. In the individual pdfs the solid line is the marginal posterior pdf from the MDN, the vertical dashed line is the true velocity structure, and the dot-dash line is the prior pdf.
Figure S2 1-D depth inversion result from Variable-Noise-MDNs for two synthetic velocity structures with individual probability density functions (pdfs) shown for four depth levels. In the depth inversions dark colours represent areas of higher probability, each row of the posterior integrates to unity, and the black solid line is the true synthetic velocity structure. In the individual pdfs the solid line is the marginal posterior pdf from the MDN, the vertical dashed line is the true velocity structure, and the dot–dashed line is the prior pdf.

APPENDIX A: NETWORK CONFIGURATIONS

The terminology used here is standard for NNs and is defined succinctly in Bishop (1995). The networks using Gaussian noise to simulate uncertainty in the data were trained using three fully connected (FC) layers, where each node receives an input from every node in the previous layer. Between each node of the FC layers a rectified linear unit (ReLU) is used. The individual layer sizes and the total number of parameters to be trained in each network is outlined in Table A1.

Table A1. Network configurations of the networks for which Gaussian noise of fixed standard deviation was added to the training set. Each network structure is trained five times with different random initializations of starting parameter values.

FC 1	FC 2	FC 3	Total parameters
200	300	200	133 145
400	200	350	173 545
400	1000	150	565 145
200	1000	350	570 745
400	500	350	398 845
400	1000	200	617 445
300	300	150	147 645
400	1000	350	774 345

The networks that included uncertainties as inputs were trained using two fully connected layers connected to the dispersion curve data and one fully connected layer connected to the standard deviation data, before concatenating the layers together and applying a further two hidden layers of size 250 and 150, respectively (Fig. 4). In between each node of the fully connected layers a ReLU is used. The individual layer sizes and the total number of parameters to be trained in each network is outlined in Table A2.

Table A2. Network configurations of the networks that included standard deviation vectors in the training set. Each network structure is trained five times with different random initializations of starting parameter values.

Dispersion		Standard deviation	Total
FC 1	FC 2	FC 3	parameters
1295	240	500	573 045
1100	900	550	1427 795
960	860	400	1210 635
1000	220	1000	605 915
950	1000	140	1300 315
1100	800	450	1265 895
930	960	100	1221 995
1200	200	600	517 295