Geophysical Journal International

Geophys. J. Int. (2021) **227,** 1879–1892
Advance Access publication 2021 August 09
GJI Marine Geosciences and Applied Geophysics

https://doi.org/10.1093/gji/ggab315

Probabilistic inversion of seafloor compliance for oceanic crustal shear velocity structure using mixture density neural networks

S. G. Mosher,[1] Z. Eilon,[2] H. Janiszewski[3] and P. Audet[1]

[1] *Department of Earth and Environmental Sciences, University of Ottawa, K1N 6N5, Canada. E-mail: stephenmosher@protonmail.com*
[2] *Department of Earth Science, University of California, Santa Barbara,* 93106*, CA, USA*
[3] *Department Earth Sciences, SOEST University of Hawai'i, Mānoa, Honolulu HI,* 96822*, USA*

Accepted 2021 August 4. Received 2021 July 29; in original form 2020 November 23

SUMMARY

Measurements of various physical properties of oceanic sediment and crustal structures provide insight into a number of geological and geophysical processes. In particular, knowledge of the shear wave velocity (V_S) structure of marine sediments and oceanic crust has wide ranging implications from geotechnical engineering projects to seismic mantle tomography studies. In this study, we propose a novel approach to nonlinearly invert compliance signals recorded by colocated ocean-bottom seismometers and high-sample-rate pressure gauges for shallow oceanic shear wave velocity structure. The inversion method is based on a type of machine learning neural network known as a mixture density neural network (MDN). We demonstrate the effectiveness of the MDN method on synthetic models with a fixed deployment depth of 2015 m and show that among 30 000 test models, the inverted shear wave velocity profiles achieve an average error of 0.025 km s^{-1}. We then apply the method to observed data recorded by a broad-band ocean-bottom station in the Lau basin, for which a V_S profile was estimated using Monte Carlo sampling methods. Using the mixture density network approach, we validate the method by showing that our V_S profile is in excellent agreement with the previous result. Finally, we argue that the mixture density network approach to compliance inversion is advantageous over other compliance inversion methods because it is faster and allows for standardized measurements.

Key words: Composition and structure of the oceanic crust; Inverse Theory; Neural Networks and fuzzy logic; Probability distributions.

1 INTRODUCTION

Accurate measurements of the physical properties (i.e. seismic velocity, density, temperature, thickness, porosity, conductivity, etc.) of oceanic sediment and crustal structures are critical for understanding a variety of geological and geophysical processes. Measurements of sediment thicknesses and their seismic velocities can constrain marine sedimentation rates, which improve our understanding of palaeoclimate and tectonic uplift (Agius *et al.* 2018). Near-surface seismic parameters have provided insight into hotspot volcanism (Doran & Laske 2019), hydrothermal fluid circulation and crustal formation (Crawford *et al.* 1991), and the study of oceanic gravity waves (Yamamoto & Torii 1986). These parameters also inform geotechnical engineering projects (Yamamoto & Torii 1986) and hazard assessment (Ruan *et al.* 2014). Furthermore, accurate profiles of shallow oceanic crustal structure have direct bearing on the measurement and interpretation of gravity anomaly residuals (Herceg *et al.* 2015) and for resolving tradeoffs between crustal structure and deeper anomalies in mantle tomographic models (Marone & Romanowicz 2007), even at long periods (Montagner & Jobert 1988).

The shear wave velocity structure (V_S) of oceanic sediments is of particular interest in marine seismology since, due to their typically low V_S values (Hamilton 1971), sediments strongly affect shear wave traveltime measurements recorded by ocean-bottom seismometers (OBSs). Additionally, the elastic properties of sediments, on which V_S depends, are also responsible for large site effects that bias seismic amplitudes recorded by OBSs. Therefore, V_S is also required for seismic studies of oceanic structures that rely upon seismic amplitude information (Ruan *et al.* 2014).

Both active and passive methods exist to measure seismic velocities within oceanic structures. However, since active methods involve the excitation of seismic energy through the use of explosives or air gun shots (e.g. Sauter *et al.* 1986; Davy *et al.* 2020), and because subsurface shear wave energy is difficult to excite acoustically (Sauter *et al.* 1986; Whitmarsh & Miles 1991), direct

measurements of V_S via active methods are difficult. A passive technique to measure 1-D shear modulus profiles within oceanic structures that exploits the phenomenon of seafloor compliance was first pioneered by Yamamoto & Torii (1986). Seafloor compliance is the phenomenon whereby long-period ocean infragravity waves propagating along the ocean's surface induce a deformation of the seafloor. Longuet-Higgins (1950) was the first to provide a physical mechanism for this process, which involves nonlinear interactions between wind-generated waves (Ardhuin *et al.* 2014). Such infragravity waves typically propagate in the open ocean with wavelengths up to tens of kilometres, periods of several minutes, and with wave height displacements ranging from millimetres to centimetres (Aucan & Ardhuin 2013). Yamamoto & Torii (1986) measured 1-D, layered shear modulus profiles using a linearized inversion of compliance signals recorded by shallow-water OBSs. Since their original work, the most significant developments to the method have been; (1) its adaptation to deep sites (deployment depths greater than 1000 m) and for V_S rather than shear modulus (Crawford *et al.* 1991); (2) its extension to the 2-D case involving laterally varying, layered structures (Crawford *et al.* 1998); and (3) its adaptation for the case when seafloor deformation is induced by Rayleigh waves rather than ocean infragravity waves, which allows the method to be used in a different frequency band (Ruan *et al.* 2014; Bell *et al.* 2015). In this study, we modify the approach taken by Crawford *et al.* (1991) and demonstrate the possibility of nonlinearly inverting compliance signals for V_S within oceanic sediment and crustal structures through the use of mixture density neural networks (MDNs). The motivation for pursuing MDN inversion over other inverse methods is that MDN inversion is often faster than both linear and nonlinear methods (Earp *et al.* 2020, see e.g.), and allows for easily standardized measurements between researchers. The latter point comes from the fact that even the most sophisticated of trained neural networks are entirely specifiable by only thousands or millions of numbers, which amount to mere kilobytes of data. Thus, neural networks are easily shareable, and because they operate in a deterministic manner, they produce repeatable measurements.

2 THEORY

2.1 The forward problem

The compliance $\xi(\omega)$ of a uniform half-space as a function of angular frequency ω due to the forcing caused by ocean infragravity waves was first derived by Sorrells & Goforth (1973) and is given by

$$\xi(\omega) = -\frac{1}{k(\omega)}\left(\frac{V_P^2}{2\rho V_S^2(V_P^2 - V_S^2)}\right), \tag{1}$$

where $k(\omega)$ is the wavenumber of the ocean infragravity waves, V_P is the P-wave velocity, V_S is the shear wave velocity and ρ is the density of the medium. Following Crawford *et al.* (1991), we prefer to work with normalized compliance $\eta(\omega)$ in which the filtering effect of ocean infragravity waves is removed

$$\eta(\omega) = k(\omega)\xi(\omega) = -\frac{V_P^2}{2\rho V_S^2(V_P^2 - V_S^2)}. \tag{2}$$

The normalized compliance of a 1-D layered Earth model described by $V_P(z)$, $V_S(z)$ and $\rho(z)$ can be forward computed numerically by applying the matrix-propagator method (Aki & Richards 2002) to eq. (2).

2.2 Compliance measurement

Given a normalized compliance signal measured by an OBS, we wish to predict the most probable range of structures (V_P, V_S and ρ) that correspond to that signal. Normalized compliance signals measured by OBSs are computed using the complex vertical displacement $Z(\omega)$ and pressure $P(\omega)$ spectra of these instruments as (Crawford 2004)

$$\eta(\omega) = k(\omega)\gamma_{PZ}(\omega)\sqrt{\frac{|Z(\omega)|}{|P(\omega)|}}, \tag{3}$$

where the coherence between the pressure and vertical displacement spectra $\gamma_{PZ}(\omega)$ is given by

$$\gamma_{PZ}(\omega) = \sqrt{\frac{|C_{PZ}(\omega)|^2}{C_{PP}(\omega)C_{ZZ}(\omega)}}, \tag{4}$$

and $C_{XY}(\omega)$ indicates the cross-spectral density between signals X and Y or the auto-spectral density when $X = Y$.

2.3 Frequency considerations

The compliance response of the seafloor to ocean infragravity wave forcing is sensitive to different depth ranges at different frequencies. This is analogous to how surface waves at different frequencies have different sensitivities to structures at different depths. However, unlike surface wave dispersion measurements, compliance measurements are limited by the coherence observed between the pressure and vertical channels. Furthermore, not all ocean infragravity waves are physically capable of generating pressure fluctuations on the seafloor. The amplitude P_B of the pressure signal generated on the seafloor at depth H due to an infragravity wave with displacement height ζ, wavenumber k and wavelength λ is given by (Webb *et al.* 1991)

$$P_B = \frac{\rho_w g\zeta}{\cosh(kH)} = \frac{\rho_w g\zeta}{\cosh(2\pi H/\lambda)}, \tag{5}$$

where ρ_w is the water density and g is the gravitational acceleration. Critically, P_B depends on the water depth H, and the only infragravity waves that will produce measurable seafloor deformation are those with wavelengths greater than or equal to the water depth (Fig. 1). Moreover, using the dispersion relation for ocean infragravity waves (Apel 1987)

$$\omega^2 = gk \cdot \tanh(kH) \tag{6}$$

and the requirement that $\lambda \geq H$ in order to produce compliance effects, it can be shown that the maximum frequency f_c at which infragravity waves will produce measurable compliance signals is

$$f_c \approx \sqrt{\frac{g}{2\pi H}}. \tag{7}$$

Thus, the frequency domain over which compliance signals can be inverted is fundamentally limited by the station deployment depth H and the pressure-vertical coherence $\gamma_{PZ}(\omega)$. While there is no theoretical lower frequency bound f_l at which compliance effects can be observed, and while ocean infragravity waves with periods as large as 1000 s are possible (Aucan & Ardhuin 2013), practical limits are set by instrument sensitivities (Doran & Laske 2019). We further speculate that the physical mechanism generating ocean surface waves at periods beyond 1000 s and wavelengths greater than 10 km changes so that the preceding analysis is likely no longer relevant. Examples of normalized compliance and pressure-vertical

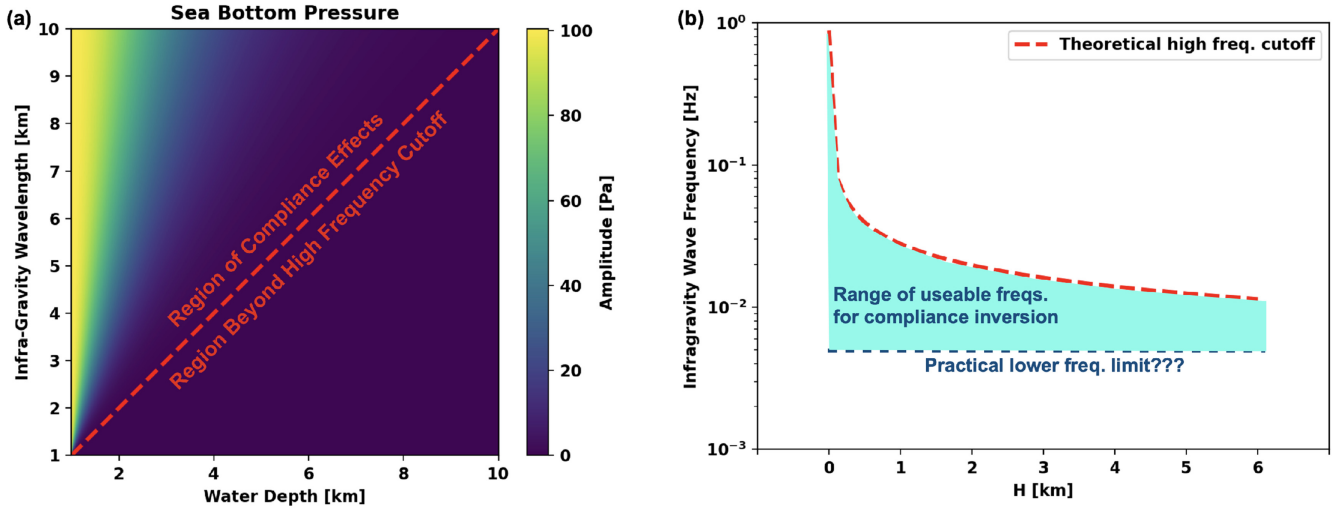

Figure 1. (a) Sea bottom pressure due to ocean infragravity waves as a function of water depth and wavelength, assuming a wave height of 1 cm. The region of observable compliance effects is clearly demarcated. (b) Usable frequency band for compliance inversion as a function of OBS deployment depth H.

coherence functions computed for OBS A02W of the Eastern Lau Spreading Center Seismic Experiment are shown in Fig. 2. The functions shown in Fig. 2 were computed from an ensemble of 75 d of hour-long noise recordings, using the ATaCR package (Janiszewski *et al.* 2019; Audet & Janiszewski 2020).

3 NEURAL NETWORK INVERSION

3.1 Standard networks

The adoption of machine learning techniques as tools for solving problems in seismology, particularly neural networks, has had a significant impact on the field (for a recent review of machine learning in the geosciences see Bergen *et al.* 2019). Originally, neural networks were intended as models of interconnected neurons in biological brains (Bishop 1995; Valentine & Woodhouse 2010). A single neuron, the fundamental unit of a neural network, operates by applying some known function $f(\cdot)$ (referred to as an activation function) to a number of inputs to produce a single output. Several such neurons are thus connected in a given architecture to form a network (Fig. 3). The simplest class of neural networks, known as multilayer perceptrons (MLPs), are defined as feed-forward networks which have sigmoidal or threshold activation functions (Bishop 1995). Feed-forward networks refer to networks in which the information processing flows unilaterally (i.e. this class of networks excludes feedback elements such as in recurrent neural networks). Popular activation functions originally included the standardized logistic function $L(x) = 1/(1 + e^{-x})$ or the hyperbolic tangent $\tanh(x)$, and were biologically motivated. However, the rectified linear unit (ReLU) activation function, defined as $max(0, x)$ has become the predominant activation function used in most neural networks due to its superior performance over a wider range of problem types (Glorot *et al.* 2011; Ramachandran *et al.* 2017).

An example of a simple, three-layer MLP is shown in Fig. 3. The first layer in such a network is called the input layer, the final layer is called the output layer and any layers between the input and output layers are referred to as hidden layers. The particular MLP in Fig. 3 takes two features as input as a vector $\mathbf{X} = (x_1, x_2)^T$. The activation of neuron i in the hidden layer, a_i, is computed by applying the activation function $f(\cdot)$ to the linear combination of weights Θ_{ij} and inputs from the previous layer ($\mathbf{X}$ in this case). The final output (equivalently the activation in the final layer) is similarly computed from the hidden layer immediately before it. Training a neural network refers to the process of randomly initializing network weights and then pushing inputs with known outputs through the network. Once the output for a given input is computed by the network, a suitably chosen objective function is used to compute the *loss* (the misfit) between the true value of the output and the result computed by the network. In a process known as backpropagation, the network then uses a gradient-descent style algorithm to make adjustments to its weights in order to reduce the error. The training process is then run for a large number of iterations until the network is sufficiently trained.

The utility of neural networks comes from the fact that they can be used to learn arbitrarily complex functions from R^m to R^n given a set of examples consisting of inputs with known outputs (Lapedes & Farber 1988; Valentine & Woodhouse 2010). Recent applications of neural networks in seismology include signal versus noise discrimination (Meier *et al.* 2019), automatic arrival-time picking (Zhu & Beroza 2019), and automatic detection and location of seismic events (Mosher & Audet 2020).

3.2 Mixture density networks

MDNs were first devised by Bishop (1994). Whereas standard neural networks learn to map a vector from R^m to R^n, MDNs learn to map a vector from R^m to an n-dimensional conditional probability distribution. Moreover, the probability distribution learned by the MDN is not restricted to be Gaussian, rather, MDNs learn arbitrary probability distributions by parametrizing them as Gaussian mixture models (GMMs). Mathematically, a multidimensional conditional probability distribution parametrized as a GMM can be expressed as (Bishop 2006)

$$P(\mathbf{t}|\mathbf{X}) = \sum_{k=1}^{K} \Pi_k(\mathbf{X}) N(\mathbf{t}|\mu_k(\mathbf{X}), \sigma_k(\mathbf{X})), \tag{8}$$

where the probability $P(\cdot)$ of observing target vector $\mathbf{t}$, given input vector $\mathbf{X}$, is given by the sum of K n-dimensional parametric Gaussian PDFs $N(\mathbf{t}|\mu, \sigma)$, each with their own mean $\mu_k(\mathbf{X})$, standard deviation $\sigma_k(\mathbf{X})$ and mixture weight $\Pi_k(\mathbf{X})$. Note that the

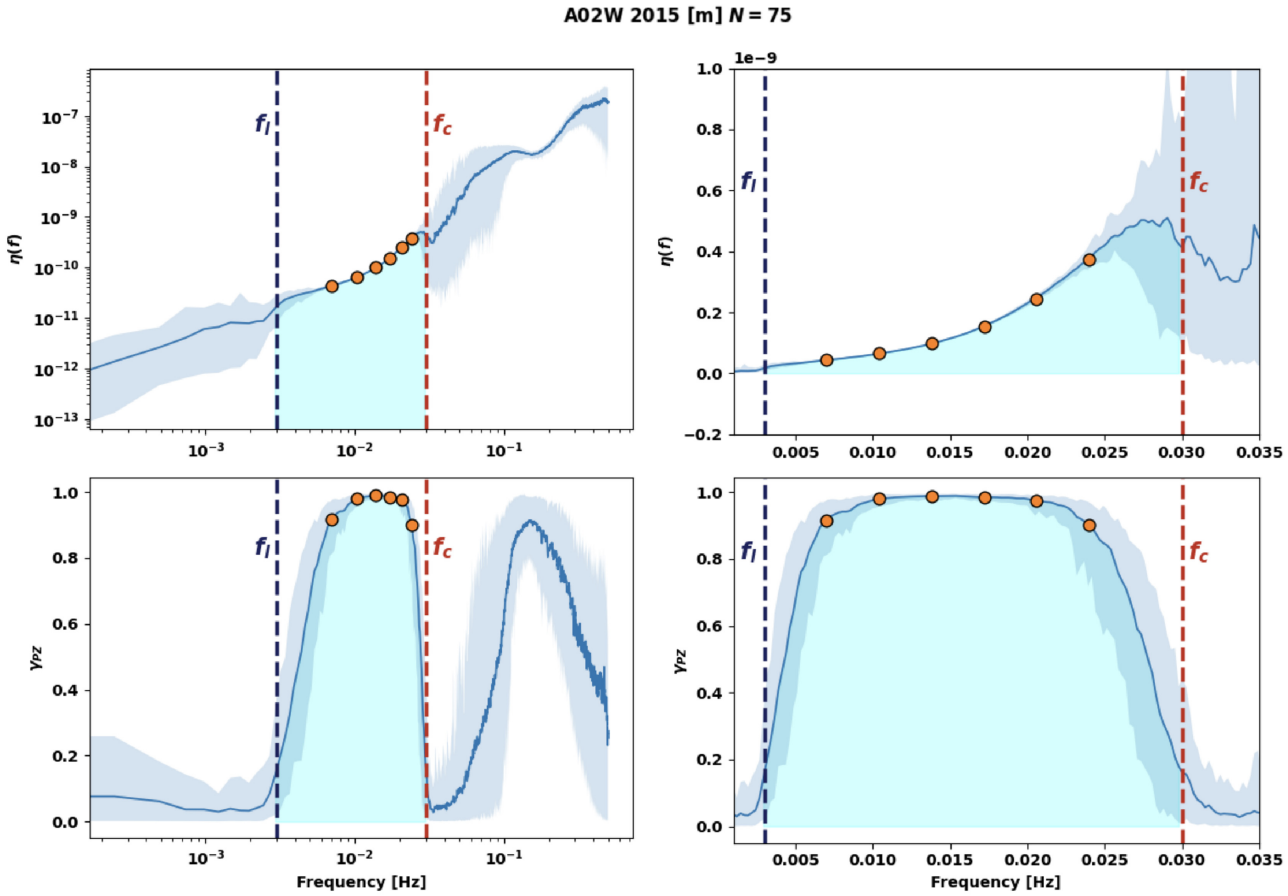

Figure 2. Normalized compliance (top) and pressure-vertical coherence (bottom) measured for OBS A02W of the Eastern Lau Spreading Center Seismic Experiment, deployed at a depth of 2015 m, and computed from an ensemble of 75 d of hour-long noise recordings. The blue shaded regions denote 95 per cent confidence intervals. The cyan shaded regions denote the compliance frequency band. The theoretical high-frequency cut-off f_c and empirical low-frequency cut-off f_l are denoted by the red and blue vertical dashed lines, respectively. Signals on the left have been plotted in log–frequency space. Signals on the right have been centred on the compliance frequency band and plotted in linear-frequency space. The orange circles denote the compliance and coherence values used in the inversion.

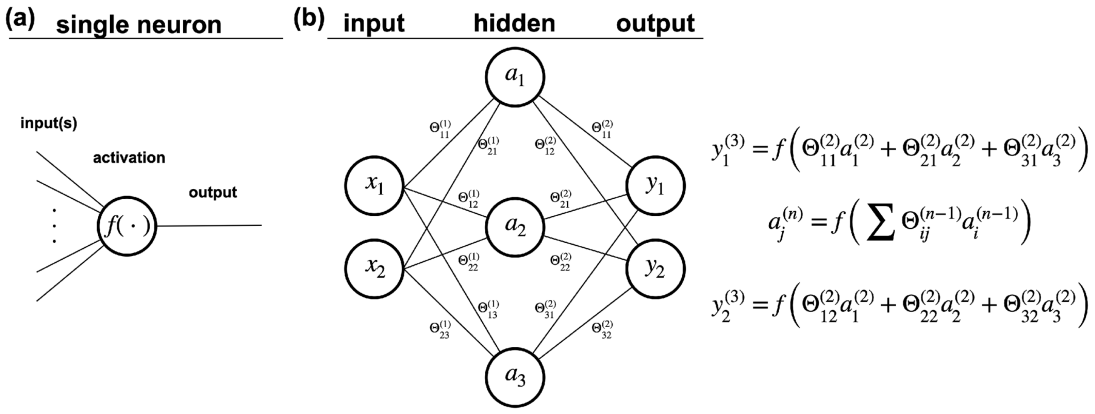

Figure 3. (a) A graphical representation of a single neuron, the fundamental unit of a neural network. The neuron takes an input vector and applies a known activation function $f(\cdot)$ to generate a single output. (b) A graphical representation of a simple three-layer MLP. The input to this network is a 2-D vector $\mathbf{X} = (x_1, x_2)^T$. The output of the network is the 2-D vector $\mathbf{Y} = (y_1, y_2)^T$. The jth activation in the nth layer, $a_j^{(n)}$, is computed by applying the activation function $f(\cdot)$ to the linear combination of weights and activations from the previous layer ($\Theta_{ij}^{(n-1)}$ and $a_i^{(n-1)}$, respectively).

mixture weights must satisfy the following constraints

$$\sum_{k=1}^{K} \Pi_k(\mathbf{X}) = 1 \qquad 0 \leq \Pi_k(\mathbf{X}) \leq 1. \tag{9}$$

An MDN can be understood as an MLP augmented with a final layer whose outputs represent the parameters of an n-dimensional parametric GMM (Fig. 4). In general, if $\mathbf{t}$ is D-dimensional then the final layer of an MDN will contain $3KD - K(D-1)$ units consisting of the means, standard deviations, and weights of each mixture component. To ensure that the units in the final MDN layer correctly represent the components of the GMM, the following operations are applied to the final outputs of the MLP, which we refer to as the vector $\mathbf{Z} = (z_1, z_2, ..., z_i)^T$ (Bishop 1995, 2006)

$$\Pi_k(\mathbf{X}) = \frac{\exp(Z_k^{\Pi})}{\sum_{l=1}^{K} \exp(Z_l^{\Pi})} \tag{10}$$

and

$$\sigma_k(\mathbf{X}) = \exp(Z_k^{\sigma}). \tag{11}$$

In eq. (10), the softmax operation is applied to the components of $\mathbf{Z}$ intended to represent the GMM weights (denoted as $\mathbf{Z}^{\Pi}$) to ensure the weights satisfy the constraints in eq. (9). In eq. (11), the exponential is applied to the components of $\mathbf{Z}$ intended to represent the GMM standard deviations (denoted as $\mathbf{Z}^{\sigma}$) so as to ensure that $\sigma_k^2(\mathbf{X}) \geq 0$. Finally, the components of $\mathbf{Z}$ intended to represent the GMM means (denoted as $\mathbf{Z}^{\mu}$) can be represented directly by the final MLP network activations, thus

$$\mu_k(\mathbf{X}) = Z_k^{\mu}. \tag{12}$$

Typical training protocols for MLPs include the minimization of loss quantified by least-squares or cross-entropy objective functions. However, Bishop (1994) showed that MLPs trained with either of these protocols approximate conditional averages of target data. For the class of problems in which the probability distribution of the target variable is either Gaussian or unimodal, networks trained with such protocols will give meaningful results, however, if the distribution of the target variable is either non-Gaussian or multimodal then these training protocols are unsuitable (Meier *et al.* 2007). In the context of MDNs, training is instead accomplished by minimizing the following negative log-likelihood function (assuming N independent pieces of data)

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \Pi_k(\mathbf{X}_n, \mathbf{w}) N(\mathbf{t}_n | \mu_n(\mathbf{X}_n, \mathbf{w}), \times \sigma_k(\mathbf{X}_n, \mathbf{w})) \right\}, \tag{13}$$

where the vector $\mathbf{w}$ consists of the network parameters, and the explicit dependence of Π_k, μ_k and σ_k on $\mathbf{w}$ has been included (Bishop 2006). This objective function, along with the GMM approach, allows MDNs to obtain a more complete description of the target data (Bishop 1994). In the next section, we illustrate these points with a toy problem.

3.3 A toy problem

Consider the problem of predicting output variable x_2 from input variable x_1 for the data set generated by

$$x_2 = \pm\sqrt{x_1 - 1} + 3 + \epsilon, \tag{14}$$

where ϵ indicates the addition of a small amount of Gaussian noise (Fig. 5a). The essential features of this problem are that it is (1) nonlinear and (2) multimodal. In other words, given any x_1 we expect two distinct ranges of possibilities for $P(x_2|x_1)$. The result established by Bishop (1994) is that a typical MLP trained with a least-squares or cross-entropy minimization objective will approximate the conditional average of the target data (x_2 in this case). This result is demonstrated in Fig. 5(a), in which predictions of a simple MLP trained on the data set generated by eq. (14) uniformly approximate the conditional average of the target variable x_2. Thus, the conditional average of x_2 is an incomplete description and the inappropriateness of attempting to train an MLP in such a manner for this style of problem is evident. By contrast, since we know that this data set is bimodal, we should be able to train a simple MDN to learn $P(x_2|x_1)$ using $K = 2$ GMM components. Once $P(x_2|x_1)$ is learned, we can then interrogate it for a complete description of the relationship between x_1 and x_2. Samples taken from the probability distribution learned by an MDN trained on the data generated by eq. (14) are also shown in Fig. 5(a). The MDN adequately learns the relationship between x_1 and x_2 and the components of the GMM ($\mu_k(x_1)$, $\Pi_k(x_1)$, and $\sigma_k(x_1)$) learned by the MDN are shown in Figs 5(a) and (b). With $K = 2$ components in this example, the two mixture components parametrize the upper and lower branches, respectively, of the parabola. In general, the optimal number of mixture components required for a given problem is not known *a priori*. However, MDNs are parsimonious in that they typically assign zero weight to unnecessary mixture coefficients $\Pi_k(\mathbf{X})$ and use as few mixture components as needed (Bishop 1995). K is rarely required to be larger than 15 for most problems (Meier *et al.* 2007; de Wit *et al.* 2013; Earp *et al.* 2020).

3.4 Prior applications of MDNs in seismology

The motivation for implementing MDNs over MLPs or other types of neural networks is in their superior handling of nonlinear inverse problems. Indeed MDNs have been successfully used in this capacity for a number of studies in seismology: Meier *et al.* (2007) pioneered this technique to invert surface wave data for a model of global crustal thickness; Shahraeeni *et al.* (2012) predicted porosity, clay content and water saturation of reservoirs from V_P and V_S; de Wit *et al.* (2013) trained an MDN to invert P-wave traveltime curves for Earth's spherically symmetric V_P structure; de Wit *et al.* (2014) obtained 1-D velocity and density profiles by inverting degree-zero spheroidal mode splitting function measurements with an ensemble of MDNs; Küufl *et al.* (2016) inverted coseismic displacement observations for point source parameter estimates; and, in a study similar to ours, Earp *et al.* (2020) used an ensemble of MDNs to invert surface wave dispersion data for shear wave velocity models and their nonlinearized uncertainty beneath the Grane field in the Norwegian North Sea.

4 APPLICATION TO SYNTHETIC DATA

4.1 Validating the method with a synthetic recovery test

To test whether an MDN can invert compliance signals for shallow V_S we must first select a station deployment depth. Recall that the station deployment depth controls the theoretical high-frequency limit f_c at which compliance effects will be measurable. Therefore, when modelling compliance signals for an OBS with an arbitrary

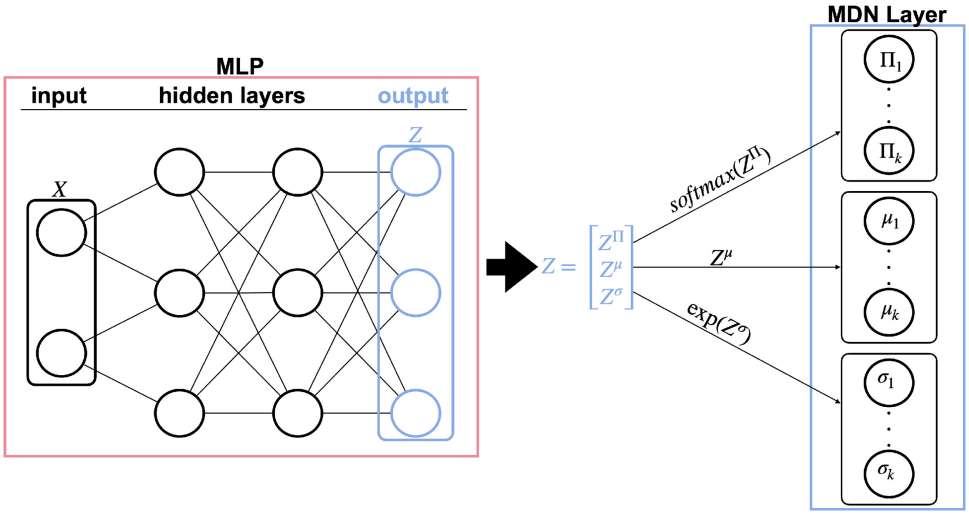

Figure 4. A graphical representation of an MDN. The entire MDN consists of an MLP augmented with a final MDN layer. The MDN layer applies the transformations indicated to the output **Z** of the MLP. The softmax operator is applied to the components of **Z** intended to represent GMM mixing coefficients (Π_k). The exponential operator is applied to the components of **Z** intended to represent the GMM standard deviations σ_k. No operation is applied to the components of **Z** intended to represent the GMM means μ_k.

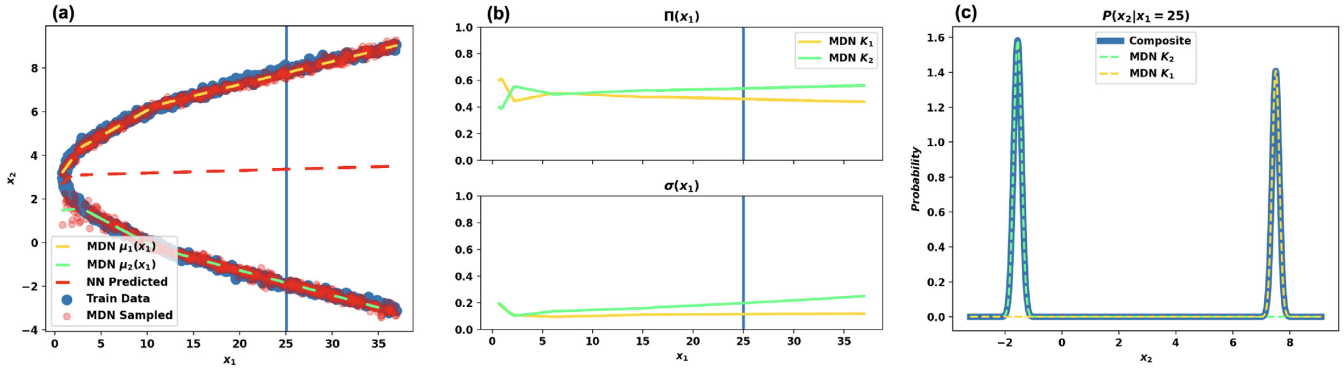

Figure 5. (a) An example of noisy, multimodal, quadratic data (the training data), as well as the predictions of a standard neural network, and samples taken from a GMM learned by an MDN, each trained on this data. With $K = 2$ mixture components, the MDN learns to parametrize the ith branch of the parabola using the ith mean $\mu_i(x_1)$ of the GMM. (b) The remaining parameters of the GMM learned by the MDN, namely $\Pi_i(x_1)$ and $\sigma_i(x_1)$. (c) The MDN estimate of the conditional probability $P(x_2|x_1 = 25)$ for the data in (a). The contributions from each mixture component are superimposed on the conditional probability distribution.

deployment depth, eq. (7) may be used to determine f_c, and the low-frequency limit f_l may be set according to a particular instrument response. However, because we intend to later invert the compliance signal measured by A02W (Fig. 2), we choose to model synthetic data using the particulars of A02W. Thus, we set the station deployment depth in the synthetic case to be 2015 m, and we determine the range of inversion frequencies available to use by considering the pressure-vertical coherence spectrum of A02W (Fig. 2). Doing so, we choose to invert compliance values measured at six frequencies equally spaced between 0.007 and 0.024 Hz. Having determined our inversion frequencies we then compute theoretical sensitivity kernels for these six frequencies in order to get a sense of what structures the inversion will be able to resolve (Fig. 6). Because the theoretical sensitivity kernels inform us that compliance effects for a station deployed at a depth of 2015 m will be most sensitive to earth structure shallower than 2 km, we therefore limit our structural models to a depth of 2 km and include a half-space layer below. Due to this shallow structure focus, we thus choose to parametrize our structural models with the intent of predominantly characterizing oceanic sediment structure.

In many of the previous studies involving MDNs in seismology, structural parametrizations were accommodated using cubic splines, which allow for continuous representations of Earth structures with depth across discontinuities (Meier *et al.* 2007; de Wit *et al.* 2013, 2014). However, in this study we prefer a model parametrization based on cubic Bernstein polynomials. A depth-dependent model parameter u on the interval [0, $z_{\max}$] can be represented using Bernstein polynomials according to

$$u(\tilde{z}) = \sum_{j=0}^{J} B_j \beta_j(\tilde{z}, J), \tag{15}$$

where $\tilde{z} = z/z_{\max}$ is the normalized depth ($z_{\max}$ is the maximum depth of the polynomial representation), J is the polynomial order of the representation and the B_j terms are the coefficients of the Bernstein basis functions $\beta_j(\tilde{z}, J)$, which are given by

$$\beta_j(\tilde{z}, J) = \binom{J}{j}(1-\tilde{z})^{(J-j)}z^j. \tag{16}$$

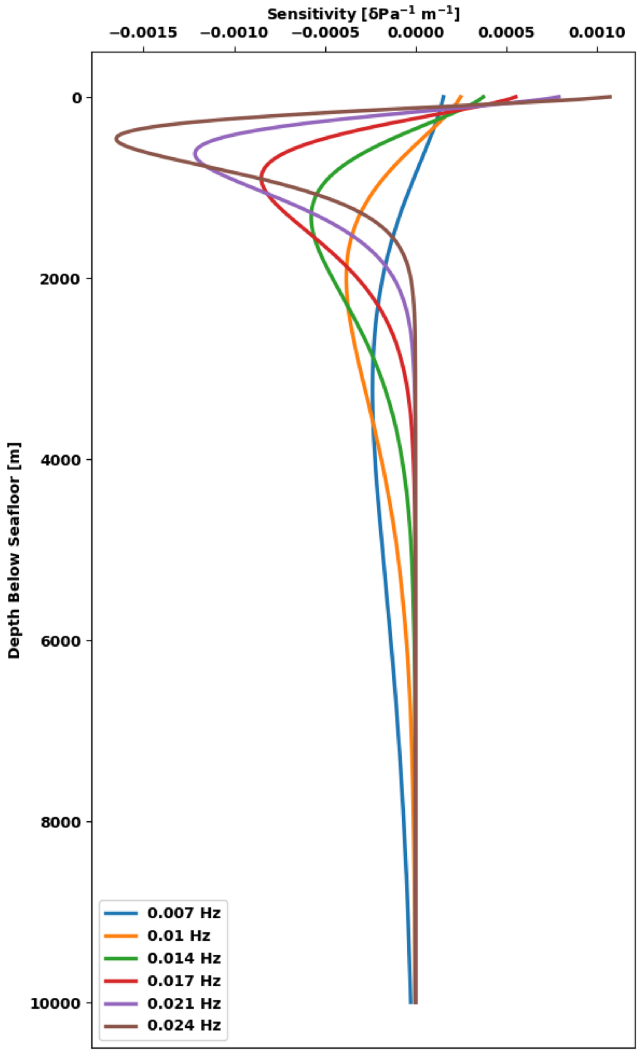

Figure 6. Theoretical compliance sensitivity kernels computed for the six inversion frequencies used in this study for a station deployed at a water depth of 2015 m.

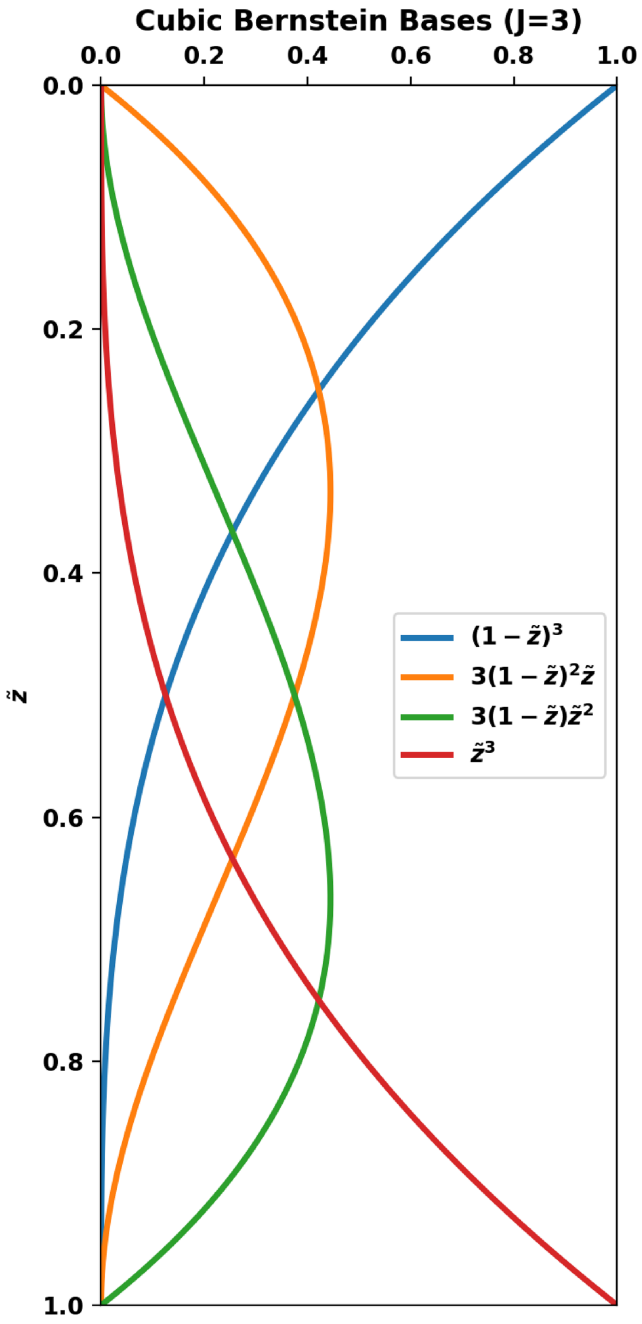

Figure 7. Third-order ($J = 3$) Bernstein polynomial basis functions.

The use of cubic Bernstein polynomials (Fig. 7) allows for a low number of model parameters (4) and allows for an efficient parametrization of continuous V_S profiles (which are expected in oceanic sediments) without the need to prescribe a layered structure. Furthermore, this representation has optimal stability for a given polynomial order (Farouki & Rajan 1987; Farouki 2012). An additional attractive feature of Bernstein polynomials is that, since they sum to unity at all depths, the width of the prior bounds on the coefficients is equivalent to the width of the prior bounds on the geophysical parameter being represented (i.e. $V_S(z)$, Gosselin *et al.* 2017). It should be noted that this property of Bernstein polynomials was the primary reason for selecting this polynomial basis over another. Therefore, we generate V_S profiles by randomly selecting cubic Bernstein coefficients from the uniform distribution $U(0.1, 3)$ and we also enforce V_S profiles generated in this manner to be strictly monotonically increasing with depth. Furthermore, we assume a constant V_P profile of 6.0 km s^{-1} and a constant density profile of 2.0 g cm^{-3} for our structural models. Fixing V_P and ρ helps aid the inversion of V_S, since we do not include these parameters as additional input to the MDN. Although more sophisticated fixed parametrizations could be used for V_P and ρ, compliance is much less sensitive to V_P and ρ at high V_P/V_S (i.e. $V_P/V_S > 2$) (Zha & Webb 2016). Because V_S is limited to values from 0.1 to 3 km s^{-1} in our models, our V_P/V_S is always >2, and hence, assuming constant values for these quantities is not detrimental.

Although we construct V_S profiles by sampling cubic Bernstein coefficients from uniform distributions, it is important to note that uniform sampling of Bernstein coefficients on a fixed interval does not correspond to uniform sampling of V_S. That being said, the distribution of V_S profiles used in this study adequately covers the range of real compliance signals observed. An example of the distribution of 100 000 V_S profiles generated in the described manner is shown in Fig. 8 along with a single structural model (V_S only).

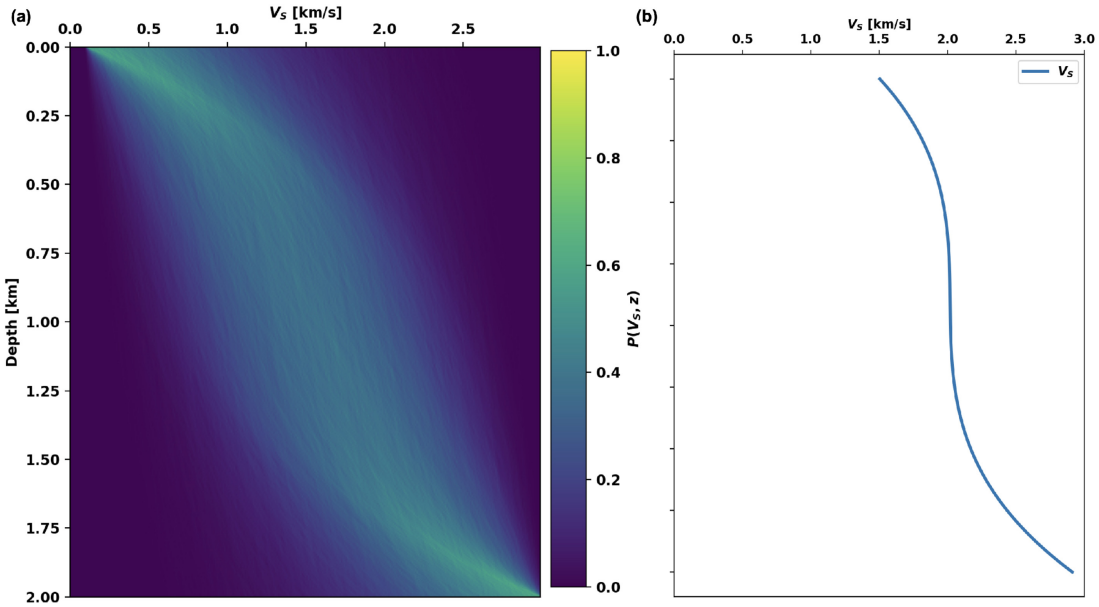

Figure 8. (a). An example of the prior distribution of V_S models used in this study generated from 100 000 models. (b) A particular V_S profile selected randomly from the prior distribution in (a).

While a higher order polynomial basis could be used to parametrize V_S profiles in principal, we argue that third-order (cubic) polynomials are sufficient for parametrizing shallow V_S for two reasons. First, quadratic V_S relations have been determined for marine sediment structures such as those in the Cascadia subduction zone (Ruan *et al.* 2014; Bell *et al.* 2015). Second, it is possible to compute the effective number of model parameters M_{eff} required to invert compliance signals at six frequencies using the theoretical sensitivity kernels in Fig. 6. Assuming the sensitivity kernels in Fig. 6 as the data kernel $\mathbf{G}$ for a minimum-length inverse problem, we compute M_{eff} by calculating the trace of the model resolution matrix given by $\mathbf{R} = \mathbf{G}^{-\mathbf{g}}\mathbf{G}$, where $\mathbf{G}^{-\mathbf{g}}$ is the generalized inverse. Doing so, we find that as long as we limit our inversion structure to a maximum depth of 2 km, then $M_{\text{eff}} = 3.99$, otherwise M_{eff} increases if greater structural depths are considered.

In preparation for training an MDN, we generate 100 000 random structural models to use as a training set and 30 000 models to use as a testing set following the above process. We then use eq. (2) and the 1-D matrix-propagator method (Crawford *et al.* 1991; Aki & Richards 2002) to forward model the normalized compliance signals that each of the models would produce at the six inversion frequencies. The forward computation is carried out by discretizing the training and testing models such that they have one layer per metre of structure, for a total of 2000 layers. Once the signals have been forward computed, we add random noise to each compliance value using the 95 per cent signal statistics computed for A02W at the appropriate frequencies (Fig. 2). Ensemble averaged compliance signals such as that computed for A02W typically have low uncertainty and are well constrained at frequencies where γ_{PZ} is large. Thus, the inversion method will perform best for such frequencies.

In other compliance inversion studies, the role of the pressure-vertical coherence has been to simply inform practitioners of the frequency domain over which compliance effects can be observed. Therefore, the multiplication by γ_{PZ} (which is essentially a weighting function) in eq. (3) is not performed when calculating $\eta(\omega)$ in practice. Instead, the inversion of $\eta(\omega)$ is usually restricted to frequencies where $\gamma_{PZ} \geq 0.8$ (e.g. Zha & Webb 2016; Doran & Laske 2019). The reason for this is that if significant noise sources persist on the horizontal OBS components, then even large values of γ_{PZ} may underestimate compliance signals by 40 per cent (Crawford & Singh 2008). However, Zha & Webb (2016) demonstrate that compliance estimates for frequencies where $\gamma_{PZ} \geq 0.8$ are not significantly biased in this manner. In essence, such approaches end up assuming a pressure-vertical coherence of unity. Rather than follow the threshold approach and assume $\gamma_{PZ} = 1$, however, we prefer to include the coherence multiplication when forward modelling synthetic compliance signals. In this way, we more accurately reproduce the distribution of compliance signals likely to be measured by a given station. In the general synthetic case, this means assuming a coherence function for a hypothetical station and deployment depth. In the current case, we assume the coherence function computed for A02W (Fig. 2). The potential impact of such an assumption on the generalizability of the method will be discussed further below.

Because normalized compliance values are typically quite small (values on the order of 10^{-9} Pa^{-1} or lower), rather than work with raw signals, we instead take the $\log_{10}$ of our compliance signals and then standardize them before they are fed to the MDN by applying Z-score normalization. This type of feature scaling ensures that the compliance values with respect to each frequency in the training and testing sets have zero mean and unit variance. Note that this type of feature scaling assumes zero covariance between the compliance values at each frequency being used, which is unrealistic. While in principle the MDN framework can be modified to account for fully covariant input parameters (Bishop 2006), MDNs trained under the assumption of a diagonal covariance matrix have often been found to adequately model target probability distributions even when the input parameters are covariant (e.g. Meier *et al.* 2007; de Wit *et al.* 2013). Moreover, we found that this type of feature scaling provided the best network performance. We refer to these transformed signals to be fed to the MDN as $\hat{\eta}(\omega)$.

We train an MDN to estimate $P(\mathbf{B}|\hat{\eta}(\omega))$ using five hidden layers, each with 42 units, and allowing for $K = 6$ GMM components, where $\mathbf{B}$ is a vector of cubic Bernstein polynomial coefficients (i.e. $\mathbf{B} = (B_0, B_1, B_2, B_3)^T$). Since any given $\mathbf{B}$ in our framework is a

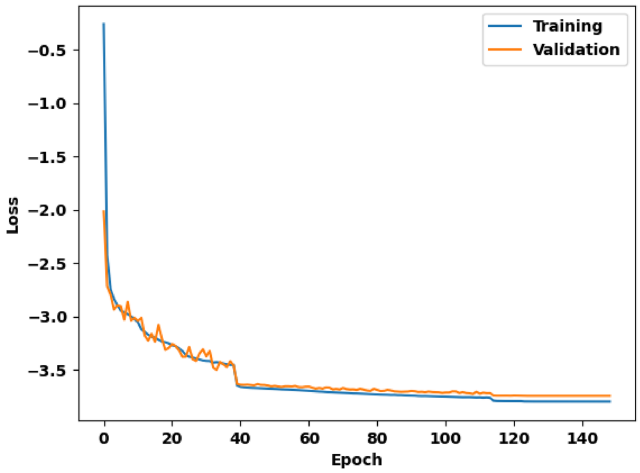

Figure 9. The training curve of the final, trained MDN used in this study. The similar performance of the MDN against the training and validation sets demonstrates that the network is not likely overfitting.

four-dimensional vector, the final MDN layer contains 42 units, and the entire MDN consists of 9840 tuneable parameters (including bias units). When training neural networks it is best to keep the number of network parameters as low as possible, in order to avoid overfitting (Valentine & Woodhouse 2010). A general rule of thumb is to ensure that there are at least 10 times as many training examples as there are network parameters. In our case this rule was the main principle for deciding upon the network architecture, since the number of hidden layers, as well as the number of units in the hidden layers are not critically important in these types of problems (Meier *et al.* 2007; de Wit *et al.* 2013). We confirmed this ourselves by testing several different network architectures, noting that no single architecture significantly outperformed another. To train the MDN, we use mini-batch gradient descent with batch sizes of 32 samples and train the network for a maximum of 1000 epochs. Rather than train the network for the maximum possible number of epochs, however, we instead employ early stopping, whereupon training ceases if the performance of the network, as measured against a portion of the training set held out for validation, does not improve after eight epochs. Moreover, we use such a validation set to ensure that the trained MDN will not be prone to overfitting (Fig. 9). Network training takes approximately 20 min on a 2.8 GHz quad-core i7 powered laptop.

Recall that MDNs do not learn a function from one vector space to another, but rather, learn to associate a probability distribution to a vector. In our case, we train an MDN to approximate $P(\mathbf{B}|\hat{\eta}(\omega))$, that is, the conditional probability of a set of cubic Bernstein coefficients given an observed compliance signal. Thus, we must decide on a sampling strategy in order to determine *the* inverted V_S profile estimated from $P(\mathbf{B}|\hat{\eta}(\omega))$. We determine the final inversion result by taking 1000 samples of Bernstein coefficients from $P(\mathbf{B}|\hat{\eta}(\omega))$. We then use the sampled coefficients to construct 1000 V_S profiles and take the mean of these samples as the final inversion result for a particular compliance signal. The process is illustrated in Fig. 10, in which we show a single inversion result for one of our test models. This sampling approach also allows us to compute the 95 per cent confidence intervals on our V_S estimates.

4.2 MDN performance assessment

To quantitatively assess the overall performance of the MDN against the test set, we compute the following error metrics. First, we compute the L_2-norm between mean predicted Bernstein coefficients and the true Bernstein coefficients for each test model. Doing so, we can assess the overall coefficient error of the MDN across all test models, as well as the error associated with each individual Bernstein coefficient. Second, we compute the absolute difference between the true and inverted V_S profiles across all test models. As with the coefficient errors, the errors associated with the velocity profiles can be assessed as a function of depth or depth-averaged. In Fig. 11, we show these metrics computed for each of the 30 000 test models we created. Looking at Fig. 11, we see that the MDN trained on the synthetic structures has a depth-averaged absolute velocity error of 0.025 km s^{-1} and an average coefficient error of 0.2. Moreover, the MDN achieves its lowest errors when predicting the Bernstein coefficients B_0 and B_3, whereas the larger spreads for coefficients B_1 and B_2 are likely indicative of model parameter trade-offs. Fig. 11 also shows that, as a function of depth, the inversion results obtained from the MDN have the largest errors at depths between 0–0.25 and 1.25–1.5 km, which correlate spatially with areas of low and decreasing compliance sensitivity (Fig. 6). Nevertheless, the metrics in Fig. 11 demonstrate the effectiveness of using MDNs to invert compliance signals for V_S in oceanic sediments and the shallow crust.

5 RESULTS AND DISCUSSION

5.1 A02W

In this section, we apply the MDN technique to invert the compliance signal recorded by OBS station A02W of the Eastern Lau Spreading Center Seismic Experiment. This station offers a convenient benchmark result since it was recently analysed in detail by Zha & Webb (2016) who used a Markov Chain Monte Carlo (MCMC) method to nonlinearly invert compliance signals for V_S. As described previously, rather than use the deployment depth of A02W (2015 m) to determine the highest frequency available to us for compliance inversion, and decide the lower limit by considering the instrument response, we do so by considering the pressure-vertical coherence of the station (Fig. 2). Furthermore, it should be noted that we invert compliance using a slightly different frequency band than Zha & Webb (2016), namely 0.007–0.024 Hz, rather than 0.006–0.02 Hz. Additionally, while we parametrize our synthetic models assuming the same density that Zha & Webb (2016) use for sediment layers in their inversion (ρ=2.0 g cm^{-3}), Zha & Webb (2016) parametrize V_P using a relation obtained by Dunn *et al.* (2013), whereas we assume a constant V_P profile of 6.0 km s^{-1}. Because our model parametrization remains unchanged from the synthetic scenario, we are able to invert the real compliance signal measured by A02W using the network trained previously on the synthetic data. The inversion result is shown in Fig. 12 as well as the fit between the predicted compliance signal (i.e. that computed from the MDN inverted V_S profile) and the measured signal. Fig. 12 demonstrates an excellent agreement with the inversion result obtained by Zha & Webb (2016) for A02W (insofar as the results share the same structural domain), and validates the method. Furthermore, the predicted compliance signal agrees quite well with the measured signal, except for the highest frequency value. This high-frequency discrepancy is likely reflected in the absolute depth-error histogram of Fig. 11(c). Otherwise, all predicted compliance

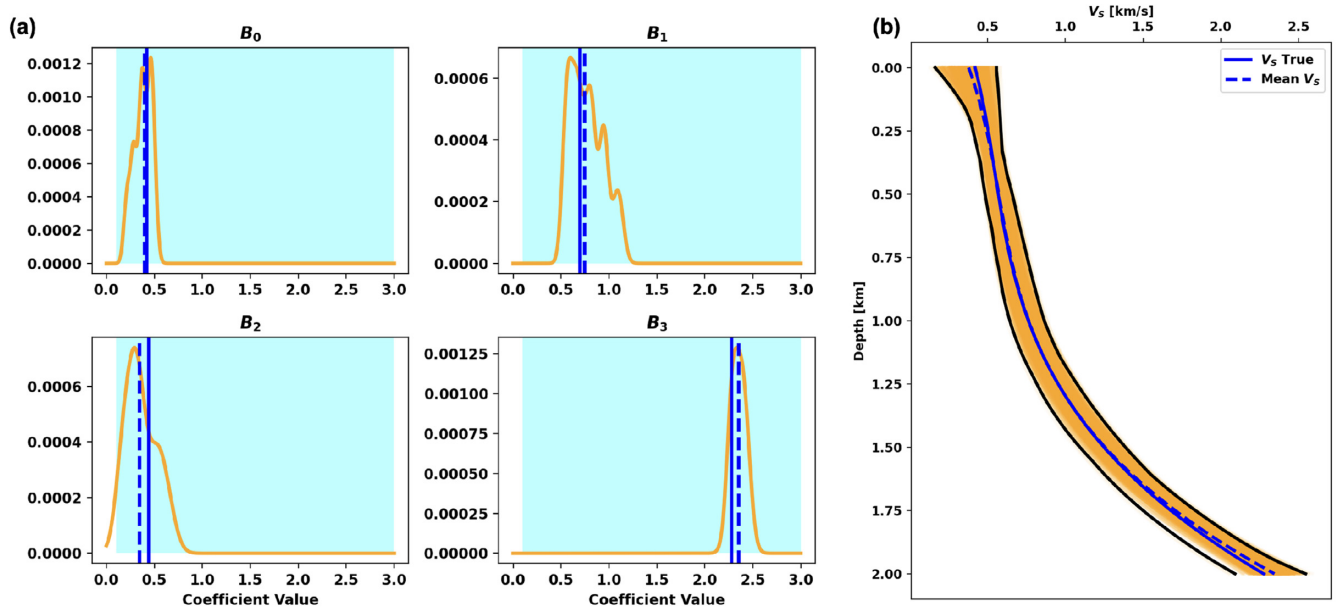

Figure 10. A single synthetic inversion example. (a) The posterior distributions of each inverted cubic Bernstein polynomial coefficient (orange). Also shown are the prior distributions of Bernstein coefficients (the shaded cyan boxes indicating $U(0.1, 3.0)$). True Bernstein coefficients are indicated by solid vertical blue lines. The recovered coefficients are indicated by dashed vertical blue lines (the means of 1000 samples taken from $P(\mathbf{B}|\hat{\eta}(\omega))$. (b) The V_S profiles generated from the 1000 sets of Bernstein coefficients, sampled from the GMM learned by the MDN (orange). The true V_S profile is shown in solid blue. The mean profile (our estimate of V_S) is shown in dashed blue. The 95 per cent confidence bounds are plotted in black.

values are within the 95 per cent confidence levels of the measured signal.

5.2 Limitations

The MDN approach to inverting compliance signals recorded by OBSs for V_S depends principally on the following aspects; the station deployment depth, the measured or assumed pressure-vertical coherence function, the type of structural parametrization, and the noise assumptions implicit in the training process. Each of these aspects have important implications on the ability of the method to generalize or its limitations, which we in turn discuss here.

Because seafloor compliance is a depth-dependent signal, rather than training a single MDN to invert compliance signals recorded by any OBS, one must train different MDNs to invert compliance signals recorded by OBSs deployed at different depths. Although it would likely be unnecessary to train separate MDNs for stations whose deployment depths differ only on the order of metres, an important test to perform in future analyses will be to determine at what point a difference in deployment depth between stations necessitates the training of a separate MDN. For example, consider that it is found that MDNs trained for a given depth can also satisfactorily invert signals recorded by stations whose deployment depths differ by up to 50 m. Then, the physically possible deployment depths on Earth could be binned into 50 m intervals, and a suite of MDNs could be trained to invert compliance signals for any depth. In this way, the method could be generalized to various depths. Additionally, it may not be necessary to train different MDNs for stations deployed in large ocean depths (such as 3 km and beyond, e.g.) since infragravity wave dispersion eventually saturates with increasing deployment depth (as can be seen from equation 6 where $\tanh(x) \approx 1$ when $x \geq 3$). This means that beyond a certain point, compliance becomes depth-independent.

In addition to deployment depth however, real compliance signals observed by OBSs depend on the pressure-vertical coherence at a given station. For the purposes of generalizing the method, this is a more subtle issue than the deployment depth. This becomes apparent when one recognizes the distinction between the compliance signal generated by an infragravity wave, which is always present, versus the ability to measure it. Therefore, one must consider how noise artefacts (e.g. time-variable tilt noise) may affect the ability to robustly calculate pressure-vertical coherence and compliance functions. Currently, it is unclear if such noise artefacts vary systematically with features such as instrument design or deployment environment, in addition to depth. We speculate that by conducting a statistical analysis of coherence functions observed by instruments deployed at various depths globally, it may be possible at the very least to determine empirical depth dependent coherence distributions which could be used to train general MDNs for various depths. Indeed, this will be the subject of future manuscripts. For example, when generating training samples by forward modelling compliance signals from structural models, rather than multiplying signals by a fixed coherence, as we have done, one could multiply by a randomly selected coherence function chosen from a suitable distribution of coherence functions for the selected depth. If these coherence values are recorded, then, during the training process, they can be fed to the MDN as prior information in conjunction with the forward modelled compliance signals. Using such an approach the method could be made station independent and thus, entirely depth dependent. Similarly, in the implementation presented here we have also effectively assumed the noise statistics for a single station when creating training signals. Again, it may be possible to generalize compliance noise in a station independent way when creating training data for the network. For now, however, the MDNs we have trained in this study are depth and station specific.

This method also relies upon accurate knowledge of the frequency response of differential pressure gauges (DPGs), colocated

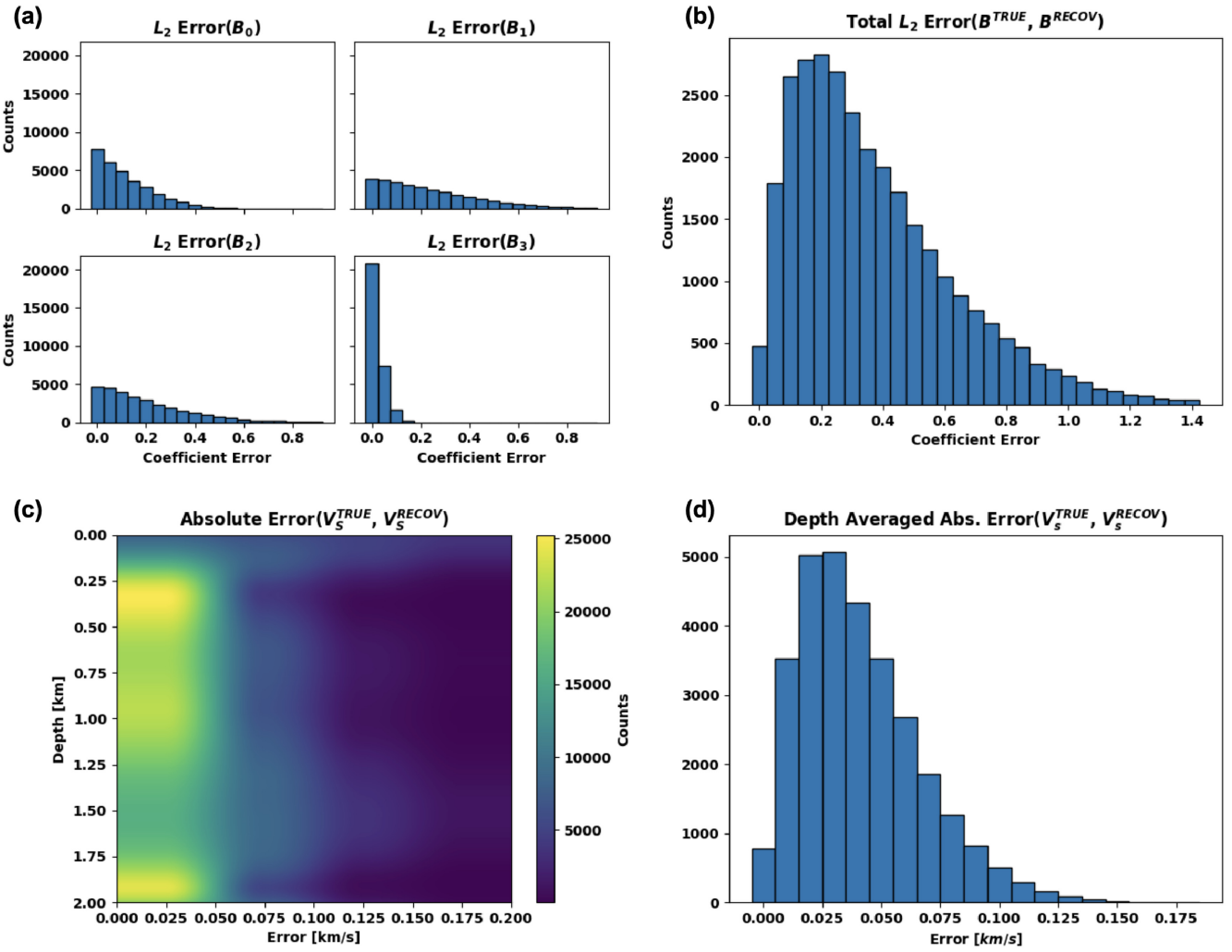

Figure 11. Compiled error metrics computed for each of the 30 000 synthetic test models. (a) Histograms of the L_2 error between each true and recovered Bernstein coefficient individually. (b) Histogram of the overall L_2 error between true and recovered Bernstein coefficients. (c) The 2-D histogram of the absolute error between true and recovered V_S profiles as a function of error and depth. (d) The depth-averaged absolute error between true and recovered V_S profiles.

with OBSs, from which we estimate compliance. Ideally, these instruments are calibrated before deployment. However, since it is difficult to calibrate DPGs in the lab due to their output varying with ambient temperature and pressure, sometimes this information is not available (Zha & Webb 2016). In this case, a constant offset is added to the DPG response values. To address this, some OBS instrument groups have developed calibration DPGs that empirically determine the response prefactor *in situ*. We therefore encourage practitioners to carefully check response metadata and evaluate compliance curves for intra-deployment consistency.

Finally, in our implementation of this method, we used a simple parametrization scheme and assumed constant density profiles and V_P profiles. In general, the trained MDN performance will depend on the parametrization scheme used, and more sophisticated schemes may enable greater utility of the method, albeit at the expense of more complex setup. In a supplement to this study we indeed verify that the choice of parametrization for V_P does not have a significant impact on our results. Additionally, the focus of this study has been on inverting for V_S in oceanic sediments, which we have implicitly assumed to be smooth. For instance, we have chosen to parametrize V_S using a smooth polynomial basis. While V_S profiles in sediments are expected to behave smoothly, due to compaction and pressure effects this is another potential limitation of the method in cases where this assumption does not hold. All that said, this study is primarily a proof of concept; adapting this method for use in general settings would be relatively straightforward.

5.3 Advantages

Despite the limitations above, many of which may be addressed in principle, the main advantage of MDN inversion over other inverse methods is that MDN inversion is a nonlinear method capable of directly estimating Bayesian posterior probability distributions. In fact, the MDN inversion procedure is a rigorously formulated Bayesian inference procedure and is equivalent to MCMC methods (Sambridge & Mosegaard 2002) given sufficient training data (Küufl *et al.* 2016). However, unlike other nonlinear inverse techniques such as MCMC methods, which construct posterior probability distributions by sampling from them, MDNs can be understood to estimate posteriors through prior sampling (Küufl *et al.* 2016). In essence, MDNs evaluate all prior samples without reference to any particular data observations and, for this reason, MDN inversion is orders of magnitude faster than both MCMC approaches and

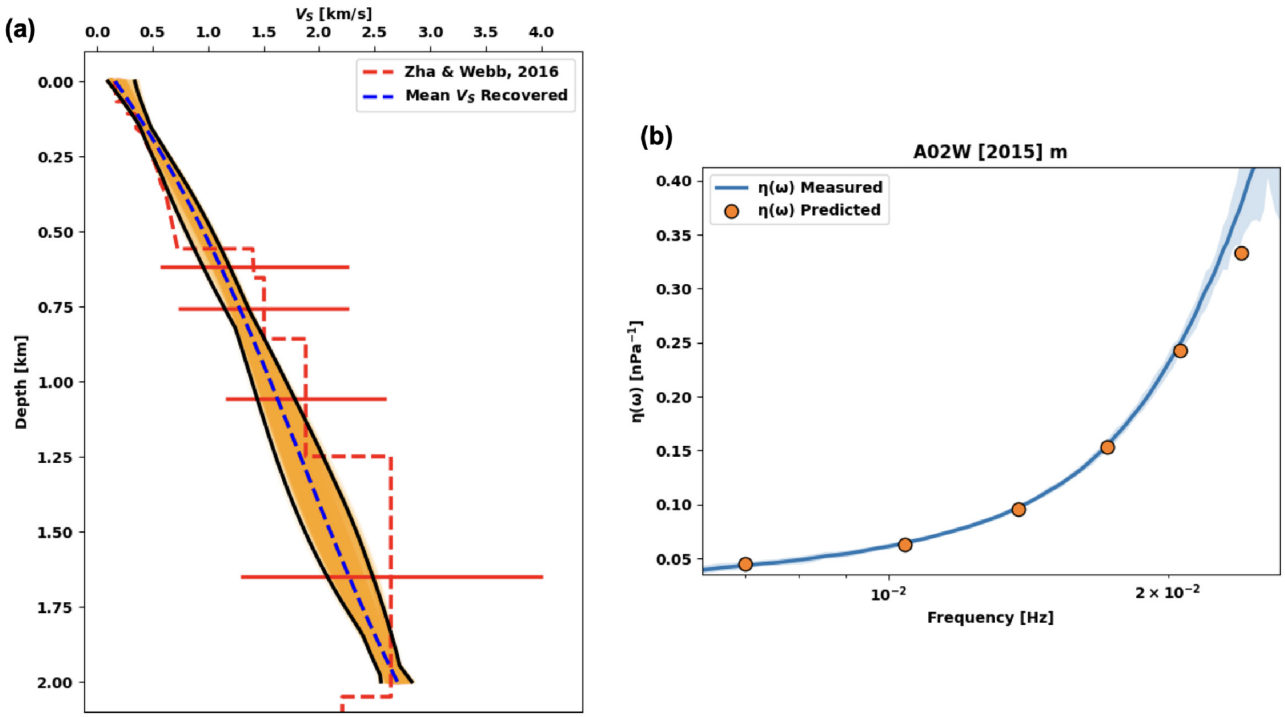

Figure 12. Inversion of the normalized compliance signal measured by A02W. (a) The V_S profiles generated from the 1000 sets of Bernstein coefficients, sampled from the GMM learned by the MDN (orange). The mean profile (our estimate of V_S) is shown in dashed blue. The 95 per cent confidence bounds are plotted in black. The inversion result obtained by Zha and Webb (2016) is plotted in dashed red along with their corresponding confidence bounds (solid red bars). (b) A comparison of the normalized compliance signal predicted by our result in (a), shown as orange circles, against the signal measured by A02W. The 95 per cent confidence bounds for the measured signal are shown in shaded blue. All values of the predicted signal are within the confidence bounds of the measured signal, apart from the highest frequency value.

linearized inverse methods (Earp *et al.* 2020). That being said, the computationally intensive aspect of MDNs lies within their training, which can take up to several hours depending on the exact problem. However, once trained, MDN inversion is on the order of seconds, thus they are especially suited to situations in which the same inverse problem needs to be solved repeatedly at different points in space or time (Küufl *et al.* 2016). Finally, MDNs can be standardized and are easily shared. For example, in the case of compliance inversion pursued here, the MDN used to invert data from A02W occupies 185 kB on disk and, if used by anyone else, will produce repeatable measurements.

6 CONCLUSIONS

In this study, we demonstrate the effectiveness of a novel approach to nonlinearly invert compliance signals recorded by colocated OBSs and high-sample-rate pressure gauges for V_S in shallow oceanic crustal structures within a probabilistic framework. Rather than use a nonlinear inverse technique such as MCMC, which estimates posterior probability distributions through prior sampling, we use a machine learning technique known as an MDN to learn conditional probability distributions between compliance signals and structural V_S models via prior sampling (Meier *et al.* 2007; de Wit *et al.* 2013; Küufl *et al.* 2016). Using this approach, we were able to train a network to adequately invert for oceanic V_S profiles in several thousand synthetic models. Among all 30 000 synthetic inversion tests, the average velocity error was 0.025 km s^{-1}. We then applied the trained MDN to invert compliance data recorded by OBS A02W of the Eastern Lau Spreading Center Seismic Experiment, for which a V_S profile was recently estimated by Zha & Webb (2016), using an MCMC approach. The resulting V_S profile obtained using the MDN in this study is in excellent agreement with the result of Zha & Webb (2016).

In the context of V_S inversion from compliance data, the contrast between nonlinear sampling techniques (e.g. MCMC) and MDN is that, within the prior sampling framework, the process of computing realizations of data is separate from the actual inversion. Therefore, while the act of training an MDN can be computationally demanding, once a network is trained, the actual inversion process itself is computationally extremely advantageous. Furthermore, we note that the use of MDNs allows for repeatable measurements and therefore helps to standardize geophysical inversions. Finally, while the MDN approach to compliance inversion pursued in this work led to networks that are not able to generalize to other stations, we discussed improvements to the method which would allow it to do so.

DATA STATEMENT

Software tools used in this research include MATLAB open source code written by Wayne Crawford for forward modelling synthetic compliance signals (available at http://www.ipgp.fr/crawford/Homepage/Software.html); the open source ATaCR package, both MATLAB (available at https://github.com/helenjanisz/ATaCR) and Python (available at https://github.com/nfsi-canada/OBStools) versions, written by Helen Janiszewski and Pascal Audet, respectively, for performing various OBS corrections and computing real compliance signals; for the machine learning portion of the project we

used TensorFlow with a custom MDN layer written by Charles Martin and available on GitHub at https://github.com/cpmpercussion/keras-mdn-layer.

ACKNOWLEDGEMENTS

Funding for this research comes from the following funding agencies; the Natural Science and Engineering Research Council of Canada (NSERC) through the NSERC Discovery Grant (RGPIN-2018-03752), NSERC-CGS and NSERC-MSFSS programs; Mitacs, through the Globalink Research Award; the University of Ottawa, through the Student Mobility Scholarship; the National Science Foundation (OCE-1658214); the University of Hawai'i, School of Ocean and Earth Science and Technology (SOEST contribution number 11376.). The facilities of IRIS Data Services, and specifically the IRIS Data Management Center were used for access to waveforms, related metadata, and/or derived products used in this study. IRIS Data Services are funded through the Seismological Facilities for the Advancement of Geoscience and EarthScope (SAGE) Proposal of the National Science Foundation under cooperative agreement EAR-1261681. Finally, we would like to thank the following researchers for their generous time and insight provided while developing this method, Andrew Valentine, Andrew Curtis and Stephanie Earp.

REFERENCES

Agius, M.R., Harmon, N., Rychert, C.A., Tharimena, S. & Kendall, J.-M., 2018. Sediment characterization at the equatorial mid-atlantic ridge from p-to-s teleseismic phase conversions recorded on the pi-lab experiment, *Geophys. Res. Lett.,* **45**(22), 12,244–12,252.

Aki, K. & Richards, P.G., 2002. *Quantitative Seismology,* University Science Books, 2nd edn.

Apel, J., 1987. *Principals of Ocean Physics,* Academic Press.

Ardhuin, F., Rawat, A. & Aucan, J., 2014. A numerical model for free infragravity waves: definition and validation at regional and global scales, *Ocean Model.,* **77,** 20–32.

Aucan, J. & Ardhuin, F., 2013. Infragravity waves in the deep ocean: an upward revision, *Geophys. Res. Lett.,* **40**(13), 3435–3439.

Audet, P. & Janiszewski, H., 2020. OBStools: software for processing broadband ocean- bottom seismic data, doi:10.5281/zenodo.3905412.

Bell, S.W., Ruan, Y. & Forsyth, D.W., 2015. Shear velocity structure of abyssal plain sediments in Cascadia, *Seismol. Res. Lett.,* **86**(5), 1247–1252.

Bergen, K. J., , Johnson, de Hoop, M. V. & Beroza, G.C. 2019. Machine learning for data-driven discovery in solid earth geoscience, *Science,* **363** 6433, doi: 10.1126/science.aau0323

Bishop, C., 1994. Mixture density networks, *NCRG/94/004,*.

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

Bishop, C., 2006. *Pattern Recognition and Machine Learning,* Springer.

Crawford, W.C., 2004. The sensitivity of seafloor compliance measurements to sub-basalt sediments, *J. geophys. Int.,* **157**(3), 1130–1145.

Crawford, W.C. & Singh, S.C., 2008. Sediment shear properties from seafloor compliance measurements: Faroes-shetland basin case study, *Geophys. Prospect.,* **56**(3), 313–325.

Crawford, W.C., Webb, S.C. & Hildebrand, J.A., 1991. Seafloor compliance observed by long-period pressure and displacement measurements, *J. geophys. Res.: Solid Earth,* **96**(B10), 16151–16160.

Crawford, W.C., Webb, S.C. & Hildebrand, J.A., 1998. Estimating shear velocities in the oceanic crust from compliance measurements by two-dimensional finite difference modeling, *J. geophys. Res.: Solid Earth,* **103**(B5), 9895–9916.

Davy, R.G., Collier, J.S., Henstock, T.J. & Consortium, T.V., 2020. Wide-angle seismic imaging of two modes of crustal accretion in mature atlantic ocean crust, *J. geophys. Res.: Solid Earth,* **125**(6), e2019JB019100, doi:10.1029/2019JB019100.

de Wit, R., Küufl, P., Valentine, A. & Trampert, J., 2014. Bayesian inversion of free oscillations for earth's radial (an)elastic structure, *Phys. Earth planet. Inter.,* **237,** 1–17.

de Wit, R.W.L., 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.

Doran, A.K. & Laske, G., 2019. Seismic structure of marine sediments and upper oceanic crust surrounding hawaii, *J. geophys. Res.: Solid Earth,* **124**(2), 2038–2056.

Dunn, R.A., Martinez, F. & Conder, J.A., 2013. Crustal construction and magma chamber properties along the eastern lau spreading center, *Earth planet. Sci. Lett.,* **371-372,** 112–124.

Earp, S., Curtis, A., Zhang, X. & Hansteen, F., 2020. Probabilistic neural network tomography across Grane field (North Sea) from surface wave dispersion data, *J. geophys. Int.,* ggaa328 doi:10.1093/gji/ggaa328.

Farouki, R.T., 2012. The bernstein polynomial basis: a centennial retrospective, *Comput. Aided Geomet. Design,* **29**(6), 379–419.

Farouki, R.T. & Rajan, V.T., 1987. On the numerical condition of polynomials in bernstein form, *Comput. Aided Geomet. Design,* **4**(3), 191–216.

Glorot, X., Bordes, A. & Bengio, Y., 2011. Deep sparse rectifier neural networks, *vol. 15 of Proceedings of Machine Learning Research,* pp. 315–323, JMLR Workshop and Conference Proceedings, Fort Lauderdale, FL, USA.

Gosselin, J.M., Dosso, S.E., Cassidy, J.F., Quijano, J.E., Molnar, S. & Dettmer, J., 2017. A gradient-based model parametrization using Bernstein polynomials in Bayesian inversion of surface wave dispersion, *J. geophys. Int.,* **211**(1), 528–540.

Hamilton, E.L., 1971. Elastic properties of marine sediments, *J. geophys. Res. (1896-1977),* **76**(2), 579–604.

Herceg, M., Artemieva, I. & Thybo, H., 2015. Sensitivity analysis of crustal correction for calculation of lithospheric mantle density from gravity data, *J. geophys. Int.,* **204**(2), 687–696.

Janiszewski, H.A., Gaherty, J.B., Abers, G.A., Gao, H. & Eilon, Z.C., 2019. Amphibious surface-wave phase-velocity measurements of the Cascadia subduction zone, *J. geophys. Int.,* **217**(3), 1929–1948.

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

Lapedes, A. & Farber, R., 1988. *How Neural Nets Work,* pp. 331–346, doi:10.1142/9789814434102_0012.

Longuet-Higgins, M.S., 1950. A theory of the origin of microseisms, *Philos. Trans. R. Soc. Lond. Ser. A, Math. Phys. Sci.,* **243**(857), 1–35.

Marone, F. & Romanowicz, B., 2007. Non-linear crustal corrections in high-resolution regional waveform seismic tomography, *J. geophys. Int.,* **170**(1), 460–467.

Meier, M., *et al.*, 2019. Reliable real-time seismic signal/noise discrimination with machine learning, *J. geophys. Res.: Solid Earth,* **124**(1), 788–800.

Meier, U., Curtis, A. & Trampert, J., 2007. Global crustal thickness from neural network inversion of surface wave data, *J. geophys. Int.,* **169**(2), 706–722.

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

Mosher, S. & Audet, P., 2020. Automatic detection and location of seismic events from time-delay projection mapping and neural network classification, *J. geophys. Res.: Solid Earth,* doi:10.1029/2020JB019426.

Ramachandran, P., Zoph, B. & Le, Q.V., 2017. Searching for activation functions, *CoRR,* http://arxiv.org/abs/1710.05941.

Ruan, Y., Forsyth, D.W. & Bell, S.W., 2014. Marine sediment shear velocity structure from the ratio of displacement to pressure of rayleigh waves at seafloor, *J. geophys. Res.: Solid Earth,* **119**(8), 6357–6371.

Sambridge, M. & Mosegaard, K., 2002. Monte carlo methods in geophysical inverse problems, *Rev. Geophys.,* **40**(3), 3–1-3-29.

Sauter, A.W., Dorman, L.M. & Schreiner, A.E., 1986. *A Study of Sea Floor Structure Using Ocean Bottom Shots and Receivers,* pp. 673–681, Springer US, Boston, MA, doi:10.1007/978-1-4613-2201-6_64.

Shahraeeni, M.S., Curtis, A. & Chao, G., 2012. Fast probabilistic petrophysical mapping of reservoirs from 3d seismic data, *Geophysics,* **77**(3), O1–O19.

Sorrells, G.G. & Goforth, T.T., 1973. Low-frequency earth motion generated by slowly propagating partially organized pressure fields, *Bull. seism. Soc. Am.,* **63**(5), 1583–1601.

Valentine, A.P. & Woodhouse, J.H., 2010. Approaches to automated data selection for global seismic tomography, *J. geophys. Int.,* **182**(2), 1001–1012.

Webb, S.C., Zhang, X. & Crawford, W., 1991. Infragravity waves in the deep ocean, *J. geophys. Res.: Oceans,* **96**(C2), 2723–2736.

Whitmarsh, R.B. & Miles, P.R., 1991, in *Situ Measurements of Shear-Wave Velocity in Ocean Sediments,* pp. 321–328, Springer Netherlands, Dordrecht, doi:10.1007/978-94-011-3568-9_36.

Yamamoto, T. & Torii, T., 1986. Seabed shear modulus profile inversion using surface gravity (water) wave-induced bottom motion, *J. geophys. Int.,* **85**(2), 413–431.

Zha, Y. & Webb, S.C., 2016. Crustal shear velocity structure in the southern lau basin constrained by seafloor compliance, *J. geophys. Res.: Solid Earth,* **121**(5), 3220–3237.

Zhu, W. & Beroza, G.C., 2019. PhaseNet: a deep-neural-network-based seismic arrival-time picking method, *J. geophys. Int.,* **216**(1), 261–273.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Figure S1: A comparison of inversion results obtained from two MDNs each assuming a different parametrization for V_P. The shaded regions denote the confidence intervals of each result.

Figure S2: Forward computed compliance values obtained from the inverted profiles in Fig. S2, compared to the observed signal for station FN04C.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.