Geophysical Journal International

Geophys. J. Int. (2019) **217,** 1706–1726
Advance Access publication 2019 February 28
GJI Marine geosciences and applied geophysics

doi: 10.1093/gji/ggz111

Bayesian geophysical inversion with trans-dimensional Gaussian process machine learning

Anandaroop Ray [1] and David Myer [2]

[1] *Geoscience Australia, Symonston, ACT 2609, Australia E-mail: a2ray@ucsd.edu*
[2] *BlueGreen Geophysics, Encinitas, California 92024, USA*

Accepted 2019 February 27. Received 2019 February 22; in original form 2018 December 7

SUMMARY

A key aspect of geophysical inversion is the ability to model the earth with a low dimensional representation. There exist various approaches to solve the inverse problem. However, most methods do not automatically adapt inverse model complexity or the number of active model parameters as dictated by data noise and sparse receiver coverage, do not quantify inverse model uncertainty or do not work equally well for 1-D, 2-D or 3-D earth models. Low-frequency electromagnetic (EM) inversion, for example, can require for 3-D problems upwards of 10^6 cells to forward model. Only a small fraction of these cells are effectively resolvable and there are significant trade-offs between them. To address such problems and get around these limitations we present a novel approach to earth model parametrization by using a Gaussian Processes (GP) machine learning (ML) technique, coupled with a parsimonious Bayesian trans-dimensional (trans-D) Markov chain Monte Carlo sampling scheme. One aspect that sets our approach apart from recent spatial dimension agnostic algorithms in the trans-D or ML literature is the ability to specify inversion property priors directly, as opposed to doing so in a transform domain of the property. We develop the theory, describe the effects of specifying different geological priors and apply the trans-D-GP method to a 1-D controlled source EM and 2-D nonlinear regression problem, using actual field data from the Northwest Australian Shelf for the former. The key advantages in using our method are the simplicity of prior specification, parsimonious low dimensional representations and ease of representing large-scale models in 1-D, 2-D or even 3-D with the same parametrization and computer code.

Key words: Inverse theory; Probability distributions; Electrical properties.

1 INTRODUCTION

Geophysical electromagnetic (EM) inversion with linearized, gradient-based methods are efficient, well understood and have been extensively used (e.g. Constable *et al.* 1987; MacGregor & Sinha 2000; Newman & Alumbaugh 2000; de Groot-Hedlin & Constable 2004; Abubakar *et al.* 2008; Key 2009; Mittet & Gabrielsen 2013; Sasaki 2013; Myer *et al.* 2015). However, to stabilize matrix inversions required to converge to a solution, keep the solution close to a preferred model and enforce smoothness in the solution, some form of regularization must be imposed for the inverted solution to be meaningful. Most regularization schemes can be interpreted in a Bayesian framework in which regularization is looked at as a means of incorporating additional information to arrive at a desired solution (see Calvetti & Somersalo 2018 for a detailed discussion). This requires that we interpret the solution model as a random variable instead of the solution possessing one single value.

In a Bayesian framework, given observed data with a description of the data noise, we aim to find the distribution of data-compatible solution values, through a forward model and prior knowledge (or belief) about the solution. This distribution, known as the *posterior* distribution encapsulates our state of knowledge (and hence uncertainty) about the solution space (in our case, the earth's subsurface conductivity). A particularly engaging discussion around the legitimate use of priors in this context can be found in Scales & Sneider (1997). Bayes' theorem bridges posterior and prior knowledge through the acquired EM data. This specification of prior knowledge (e.g. Hansen & Minsley 2017) and its parametrization is often overlooked in Bayesian inversions of geophysical data (see Pasquale & Linde 2017 for a discussion). Designing an informative prior that accurately reflects the earth's spatial character given the resolution we expect our data to possess is key to drawing meaningful inferences about subsurface geology. While this may sound like a chicken-and-egg situation, a Bayesian perspective lays bare the fact that we must make choices in designing an inversion scheme, whether it be regularized or otherwise implemented. With choices based on the physics of the problem, as we will discuss in this work, well-designed Bayesian algorithms can infer the resolution with which

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

© The Author(s) 2019. Published by Oxford University Press on behalf of The Royal Astronomical Society.

we can 'see' into the earth. Recent geophysical work highlighting the importance of choosing *a priori* appropriate basis functions in a Bayesian framework can be found in Hawkins & Sambridge (2015); Pasquale & Linde (2017) and Ray *et al.* (2017). In the field of hydrogeophysics, Cordua *et al.* (2012); Lochbühler *et al.* (2015) and Laloy *et al.* (2017) have used Training Image (TI) based priors for this purpose. Training Images, through multiple-point-statistics (Strebelle 2002) provide a realistic specification of subsurface hydrology. However, sampling the posterior distribution either through a Markov chain Monte Carlo (McMC) scheme or optimizing with gradient descent is difficult as consecutively generated TIs in a naive implementation usually have disparate properties. To overcome this issue, Laloy *et al.* (2018) have successfully used Generative Adversarial Networks (GANs; Goodfellow *et al.* 2014), a recent machine learning (ML) technique to train a low-dimensional latent space which can effectively mimic a high-dimensional solution space. In their case, the high-dimensional target to mimic is the earth's spatial hydrological variation as represented by TIs. The reduced dimension latent space is then used as the solution basis in which conventional McMC inversion is carried out, thus guaranteeing that posterior solution models are geologically realistic. This approach is promising, as it has two sought-after qualities in any geophysical inversion—geological realism and low dimension—which are often at odds with each other. The former quality is necessary for the solution models to be interpretable, while the latter is required for the solution to be stable and trustworthy. This is a geological restatement of the variance-bias trade-off (e.g. Lever *et al.* 2016)—simple earth models tend to be biased about earth structure but have low variance, while complicated earth models tend to behave in the opposite manner. However, it is unclear how to choose the low, fixed dimension of the latent space and there are many hyperparameters (nuisance variables) to be tuned in the training of the GAN. The training of GANs in a stable manner is not a trivial process (Laloy *et al.* 2017), though it may well be possible using techniques such as Bayesian Optimization (Shahriari *et al.* 2016) and worthwhile for many near-surface geophysical problems. However, for intermediate-to-deep earth geophysics, there is not enough prior knowledge to know what a TI should look like. Further, given that low-frequency information (acquired surface geophysical data) cannot with great fidelity reproduce high wavenumber information at depth (earth properties), it may not be useful to specify high wavenumber (highly detailed) priors. This is because the posterior model ensemble will marginalize to produce little detail at depth – we can think of this as being akin to a high standard deviation of geophysical properties with depth.

In such cases, as with low-frequency EM inversion, for prior representation we propose to use Gaussian Processes (GPs), which have well understood qualities of spatial variability—see Rasmussen & Williams (2006) for a thorough review. By using GPs in conjunction with the trans-dimensional (trans-D) McMC method (Green 1995; Malinverno 2002; Bodin & Sambridge 2009) for model solution dimension reduction, we can effectively model high dimension, but keep the number of inverted parameters small (i.e. achieve parsimony, Malinverno 2002). This small model dimension is what makes McMC tractable or gradient inversion stable (Laloy *et al.* 2017). As is usual with trans-D methods, the posterior model ensemble will quantify the nonlinear spatial resolution of the data.

2 THEORY

2.1 Gaussian processes

A Gaussian process is a stochastic process that is completely determined by its mean and covariance. As we will show, it is defined by priors and posteriors over *functions*. Broadly speaking, GPs are a method of non-parametric regression that do not require a fixed discretization, providing both a prediction and uncertainty around the prediction. GPs have been successfully used in many fields including spatial statistics (Cressie 1992), statistics (Williams & Rasmussen 1996), robotics (Ko & Fox 2009), weather prediction (Chen *et al.* 2014), reinforcement learning (Deisenroth *et al.* 2015) and automated image analysis (Luthi *et al.* 2018). In the ML literature, they have been extensively used to model 'black box' functions and even optimize them (e.g. Snoek *et al.* 2012). In the geosciences, they have been known by the name 'kriging' (Krige 1952; Pyrcz & Deutsch 2014) and are closely related to radial basis functions (Broomhead & Lowe 1988).

To gain insight into the workings of GPs, we follow the Bayesian exposition of Williams & Rasmussen (1996) through an example shown in Fig. 1. First, we specify *prior* notions of spatial smoothness through a covariance, typically defined by a similarity kernel which ensures that spatially close locations have similar values. Training observations are then regarded as realizations from an updated, *posterior* multivariate Gaussian. Test outputs at all unobserved points are then simply conditional realizations from the posterior Gaussian. The mathematics behind this methodology, referring to this example, is explained in detail in the remainder of this section.

In mathematical form, following the textbook of Murphy (2012), we write this as follows:

$$\begin{bmatrix} \mathbf{m} \\ \mathbf{m}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_m & \mathbf{K}_* \\ \mathbf{K}_*^t & \mathbf{K}_{**} \end{bmatrix}\right), \tag{1}$$

where the vector of values $\mathbf{m} \in \mathbb{R}^{n_{\text{train}}}$ has been observed at spatial locations $\mathbf{x} \in \mathbb{R}^{n_{\text{train}} \times n_d}$. n_{d} is the number of spatial dimensions under consideration. $\mathbf{m}_* \in \mathbb{R}^{n_{\text{test}}}$ is a vector specifying predicted values at all desired spatial locations $\mathbf{x}_* \in \mathbb{R}^{n_{\text{test}} \times n_d}$. To define the covariance matrix $\begin{bmatrix} \mathbf{K}_m & \mathbf{K}_* \\ \mathbf{K}_*^t & \mathbf{K}_{**} \end{bmatrix}$ in (1), we first define the following correlation function:

$$K(\mathbf{y}, \mathbf{y}') = \exp\left(-\frac{1}{2}[\mathbf{y}-\mathbf{y}']^t \mathbf{C}_\lambda{}^{-1}[\mathbf{y}-\mathbf{y}']\right), \text{ where } \mathbf{y} \in \mathbb{R}^{n_d}. \tag{2}$$

$\mathbf{y}$ and $\mathbf{y}'$ are any two points in n_{d} spatial dimensions. The square of the correlation length scale in each spatial dimension is specified along the diagonal of a symmetric positive definite matrix $\mathbf{C}_\lambda \in \mathbb{R}^{n_d \times n_d}$ and spatial anisotropy (rotation) by the off-diagonal entries. Geologically speaking, this matrix encapsulates our prior knowledge of the length scales and predominant strike directions to be represented by the GP. In the example shown in Fig. 1, with $n_{\text{d}} = 1$ (one spatial dimension), C_λ is a scalar. A similarity length scale λ is set equal to 0.1 spatial units *a priori*, with $C_\lambda^{-1} = \frac{1}{\lambda^2} = \frac{1}{0.1^2}$. With n_{train} observed training points located at $\mathbf{x}$, we can define a matrix $\mathbf{K} \in \mathbb{R}^{n_{\text{train}} \times n_{\text{train}}}$ using (2) for all pairwise distances between training points. We then define $\mathbf{K}_m \in \mathbb{R}^{n_{\text{train}} \times n_{\text{train}}}$ through the addition of an additive noise term such that

$$\mathbf{K}_m = \mathbf{K} + \boldsymbol{\sigma}_m^2, \tag{3}$$

where $\boldsymbol{\sigma}_m^2$ is a diagonal covariance matrix of the observed additive noise in $\mathbf{m}$ at the locations $\mathbf{x}$. For the example in Fig. 1, $\boldsymbol{\sigma}_m^2$ was set diagonal with the *a priori* constant value 0.0025 across the diagonal. If we would like to predict $\mathbf{m}_*$ at n_{test} locations $\mathbf{x}_*$ then

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

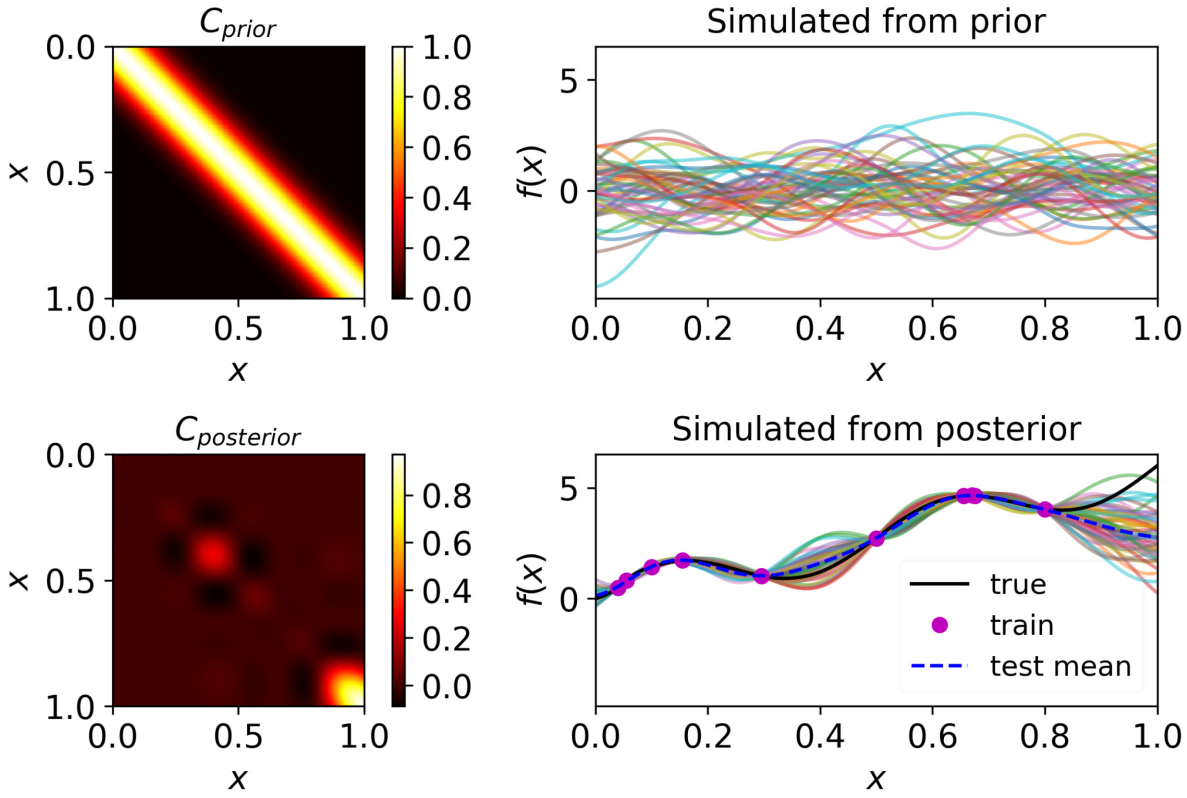

Figure 1. Top left-hand panel: a prior covariance C_{prior} defined by a smoothly decaying stationary similarity kernel such that values differing by up to 0.1 units in x have high correlation. Top right-hand panel: 50 random realizations from a Gaussian with zero mean and covariance C_{prior}. Bottom left-hand panel: a GP posterior covariance formed by modifying the prior covariance by bringing in 10 training observations. Bottom right-hand panel: The 10 training observations are shown with magenta circles. The unknown true function being approximated is shown in black. The inferred posterior mean is shown in dashed blue. Also shown are 50 random test realizations from a Gaussian with mean set to the posterior mean and covariance $C_{\text{posterior}}$. Henceforth we will refer to the mean of all posterior test realizations simply as the GP mean. We should note that regions in x with fewer training points have high posterior variance and vice versa.

$\mathbf{K}_* \in \mathbb{R}^{n_{\text{train}} \times n_{\text{test}}}$ is defined using (2) and pairs of $(\mathbf{x}, \mathbf{x}_*)$. Finally, $\mathbf{K}_{**}$ is an $n_{\text{test}} \times n_{\text{test}}$ matrix defined using (2) for all pairwise distances between testing points. For the example in Fig. 1, $\mathbf{K}_{**}$ provides the prior covariance matrix (top left), and one way to think of this is that $\mathbf{K}_{**}$ has no input from training data and is purely derived from prior knowledge.

The advantage of this formalism expressed through Gaussians is that the posterior conditional GP in (1), again following Murphy (2012), can be written as

$$p(\mathbf{m}_*|\mathbf{x}_*, \mathbf{x}, \mathbf{m}) = \mathcal{N}(\mathbf{m}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*), \tag{4}$$

where analytical formulae allow us to write out this normal probability for the random variable $\mathbf{m}_*$ with mean $\boldsymbol{\mu}_*$ and covariance $\boldsymbol{\Sigma}_*$ as follows:

$$\boldsymbol{\mu}_* = \mathbf{K}_*^t \mathbf{K}_m^{-1} \mathbf{m}, \tag{5}$$

and

$$\boldsymbol{\Sigma}_* = \mathbf{K}_{**} - \mathbf{K}_*^t \mathbf{K}_m^{-1} \mathbf{K}_*. \tag{6}$$

A GP fitting exercise as shown in Fig. 1 is a classic example of Bayesian 'probability updating.' We start with a prior notion (top row of Fig. 1) and update it as training data comes in. This update is according to a posterior Gaussian described entirely by mean (5) and covariance (6) as shown in the bottom row of Fig. 1. We can explicitly see the prior covariance $\mathbf{K}_{**}$ being updated to the posterior by the subtraction of the term $\mathbf{K}_*^t \mathbf{K}_m^{-1} \mathbf{K}_*$ in (6). The theory presented here has been for a zero mean GP without loss of generality. For instance, in the example shown in Fig. 1, the mean of posterior test realizations (or the GP mean) $\boldsymbol{\mu}_*$ was calculated by de-meaning $\mathbf{m} \in \mathbb{R}^{10}$ and then using (5), finally adding to this quantity the mean of the 10 training data points provided in $\mathbf{m}$. The 50 test posterior realizations in the bottom right-hand panel of Fig. 1 were obtained by randomly sampling from a Gaussian with the aforementioned GP mean and covariance provided by (6).

In the work presented here, we do not use the GP for inferring earth properties directly, which is how we depart from the realm of geostatistics—where the posterior Gaussian as represented by the mean and variance given by eqs (5) and (6) would completely statistically represent the earth. Instead, we use the GP mean (5) as a sparse yet smooth representation of earth model properties, ideal for low-frequency geophysical inverse problems such as EM inversion. To illustrate how we use the GP mean as a statistical interpolator, we turn to another simple example shown in Fig. 2. Using 10–15 training points $\mathbf{m}$, randomly selected in $\mathbf{x}$, we are able to approximate a smooth, unknown (to the GP) function $\mathbf{m}_*$ everywhere at the test locations $\mathbf{x}_*$. This is done by using eq. (5) and the 1-D version of (2) with a similarity length scale λ set equal to 0.05 spatial units *a priori*, with $C_\lambda^{-1} = \frac{1}{\lambda^2} = \frac{1}{0.05^2}$. The matrix $\boldsymbol{\sigma}_m^2$ was set to be diagonal with the *a priori* constant value 0.0025 across the diagonal. Now if we were to rotate the figure by 90°, we could imagine the true function to be a profile of log-resistivity in the earth, to be resolved through a parsimonious inversion scheme inverting surface EM data.

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

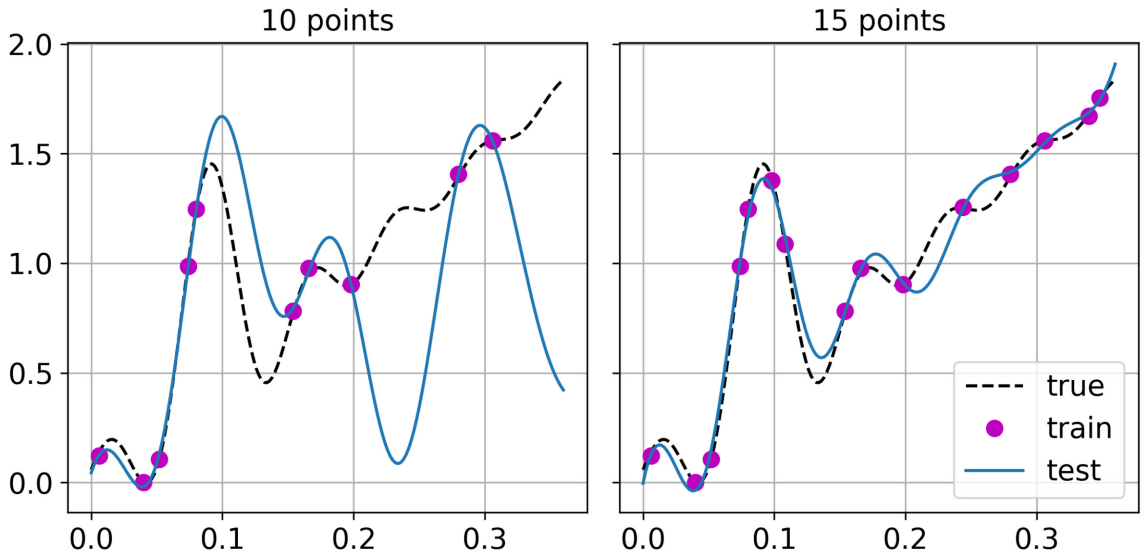

Figure 2. Using 10 and then 15 randomly selected training points to approximate an unknown function with a GP. The true function is plotted in dashed black, and the inferred GP mean is plotted with a blue solid line. Note how the GP approximation improves with the addition of more points. In this work, we propose placing GP training points in a solution model space (e.g. earth conductivity) proportional to a Bayesian inversion posterior probability. The placement and number of training points will be guided by both prior information and the model likelihood (i.e. geophysical data misfit) through a 'birth-death' trans-D scheme (Geyer & Møller 1994).

In order to use a GP as presented in this section, we need in addition to the training data, prior knowledge of $\boldsymbol{\sigma}_m^2$ and length scales in $C_{\boldsymbol{\lambda}}$. In Section 3.1 we detail the effects of choosing different values for these prior parameters. Generally speaking for regression problems (as opposed to geophysical or geological problems), if prior knowledge is not readily available, one method to obtain it is through cross-validation (see Friedman *et al.* 2001 for details). Other methods to obtain these hyperparameters are through hierarchical sampling (e.g. Gelman *et al.* 1995) or by making maximum likelihood estimates of these parameters (e.g. Plagemann *et al.* 2008), which we will detail in Section 5.

With trans-D-GP we can add and subtract training points via trans-D 'birth' and 'death' McMC steps (Geyer & Møller 1994; Green 1995; Sambridge *et al.* 2006) to represent an earth property model, say conductivity, using a GP. The key advantage of this approach lies in the fact that with the same formalism presented in this section we can represent a 1-D, 2-D or 3-D earth—instead of layers for a 1-D earth, Voronoi cells for a 2-D earth or other parametrizations in 3-D. It follows that the same inversion code and GP parametrization can be used without the computational geometry overhead required to go from 1-D to 2-D or 3-D earth models. The most costly parts of earth model construction are the matrix operations in (5). However, unlike in geostatistics, we are not drilling the earth to get more training samples of conductivity to improve the GP—we intend to use EM data, which at best only smoothly resolve the earth's subsurface. We can use a parsimonious 'training' set $\mathbf{m}$, which can be updated via a trans-D Bayesian inversion scheme that samples according to the data misfit. With Bayesian parsimony we will show that low-frequency inversion, even in 2-D, requires model representation with $\mathbf{m} \in \mathbb{R}^k$, where k need not exceed 100. This makes the calculation of (5) feasible through Cholesky decomposition for the inversion of $\mathbf{K}_m$. Further, only small parts of $\mathbf{K}_m$ and $\mathbf{K}_*$ in (5) need to be updated at every trans-D McMC step, thus obviating the need to reconstruct a large matrix with successive iterations. Using a single thread on a 2.8 GHz laptop processor, for a 2-D earth model with $k = 100$ representing 201×201 cells, a mean time of 0.074 s was required to update a model from $k = 100$ to $k = 101$. This is a typical sampling step without the forward call in a trans-D McMC iteration. Further decreases in computation time can be brought about by storing only half of $\mathbf{K}_m$ and updating the Cholesky decomposition instead of doing a new decomposition every time $\mathbf{K}_m$ changes, multithreading and/or using hierarchical off diagonal low-rank (HODLR) methods (Ambikasaran *et al.* 2016).

Although we have only investigated smooth kernel functions in the GP, it is possible to use the Matérn family of kernels to define $\mathbf{K}$ in (2) and model sharp discontinuities. We can also use a non-stationary kernel as described in Section 5. Further details on various types of kernel functions can be found in chapter 4 of Rasmussen & Williams (2006). Finally, representation of multiple earth properties at the same spatial location, say, for example, conductivity anisotropy in terms of a horizontal and vertical conductivity, can be modelled by a GP through the use of covariance between these properties as detailed in various geostatistical approaches (e.g. Cressie 1992).

2.2 Bayesian trans-D inversion

For the purpose of probabilistic inversion, trans-D McMC is well suited for sampling earth models $\mathbf{m}$ of variable dimension k. Trans-D inversion (Sambridge *et al.* 2006) is based on birth/death Monte Carlo (Geyer & Møller 1994) and the more general Reversible Jump McMC method (Green 1995). Previously, for a 1-D earth model, researchers have sampled over a variable number of layers (Malinverno & Leaney 2000; Minsley 2011; Bodin *et al.* 2012b; Dettmer *et al.* 2015; Ray *et al.* 2016; Blatter *et al.* 2018; Gao & Lekić 2018). For 2-D models, Voronoi representations with different numbers of cells have been widely used (e.g. Bodin & Sambridge 2009; Dettmer *et al.* 2014; Ray *et al.* 2014; Galetti *et al.* 2015; Saygin *et al.* 2016; Galetti & Curtis 2018). In effect, the trans-D algorithm via Bayes' theorem performs the task of model selection with regard to the complexity of the model (i.e. number of dimensions k). The fact

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

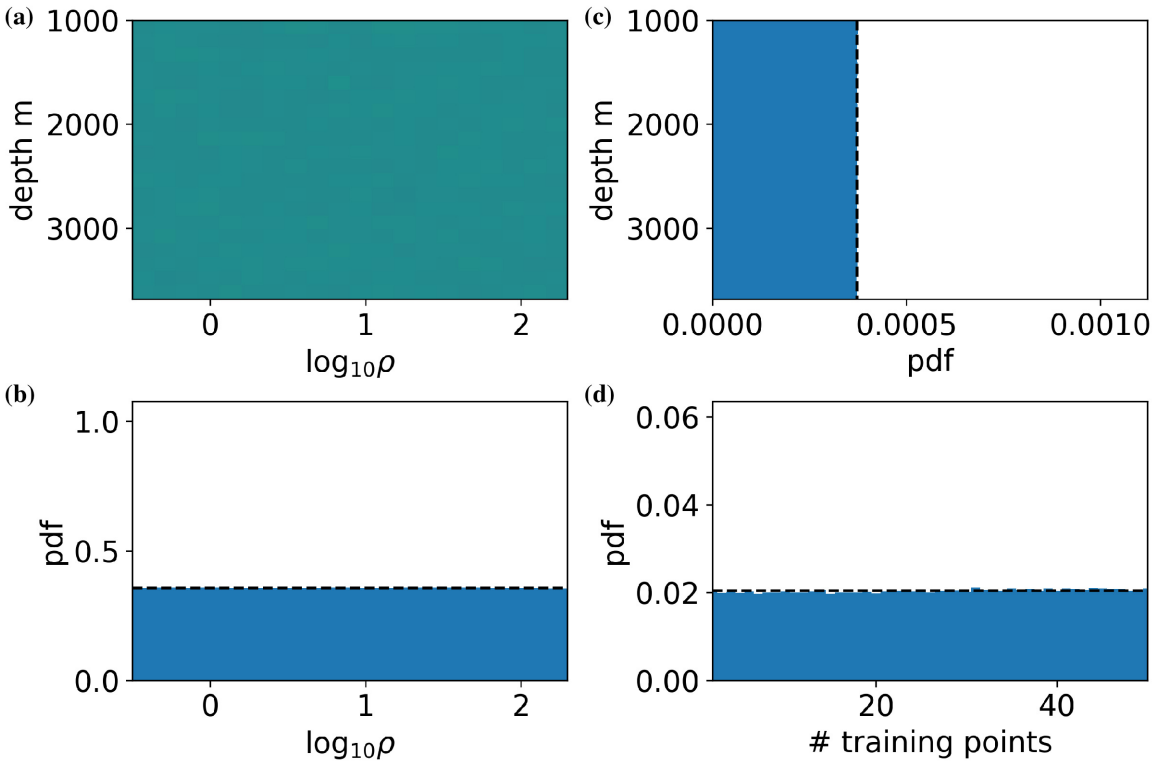

Figure 3. Sampling the prior for $\mathbf{m}_k, \mathbf{x}_k$ and related marginal distributions. The dashed line represents a uniform prior. The sampled marginal PDFs are approximately uniform as well, indicating that our prior specifications have been honoured by the implemented trans-D McMC sampler.

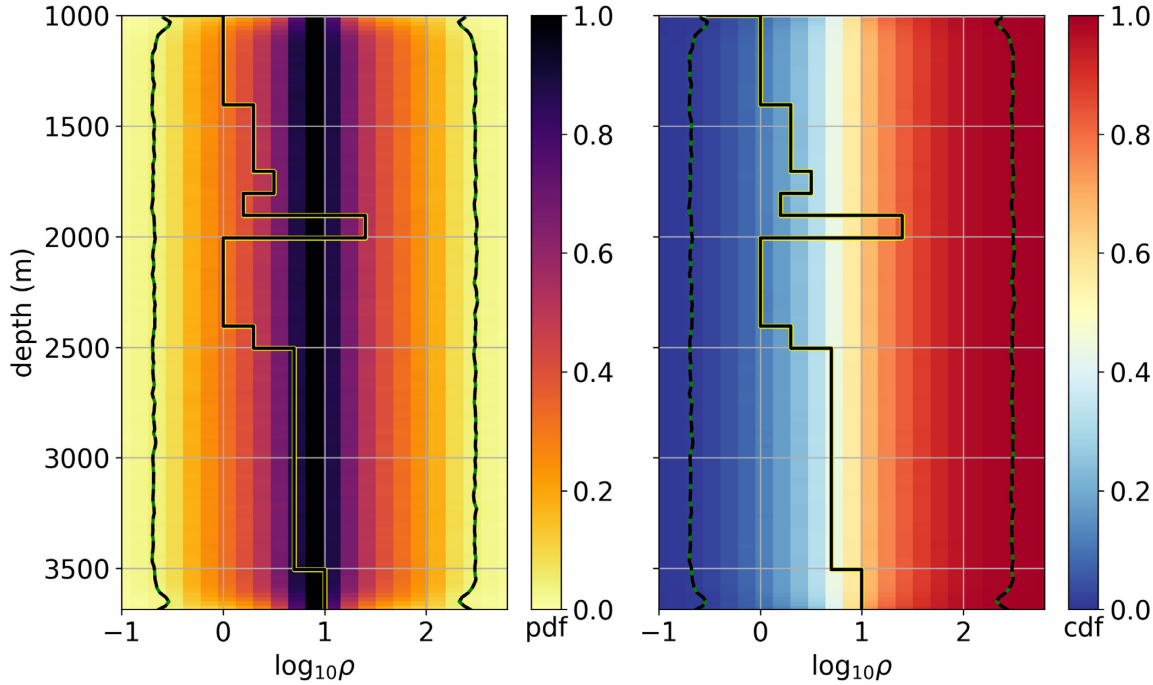

Figure 4. Left-hand panel: resulting marginal PDFs on the resistivity, that is, $\boldsymbol{\mu}_*$ after sampling prior $\mathbf{m}$ in Fig. 3. Right-hand panel: accompanying CDFs of resistivity at every depth, with each colour representing a quantile. The stair-step resistivity model to be inverted with this prior specification is shown in black. The 98 % credible interval is indicated by the dashed lines, with only 2 % of sampled resistivities outside this zone.

that models are neither overfit nor underfit is based on the idea of Bayesian parsimony, introduced to geoscience by Malinverno & Leaney (2000) and Malinverno (2002). An 'Occam factor' that penalizes overly complicated models is built into the framework of Bayes' theorem when formulated appropriately (MacKay 2003). Galetti & Curtis (2018) point out that this is not as straightforward as was previously assumed for trans-D and this issue is discussed further in Appendix A. Theoretically speaking, the Bayesian model

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

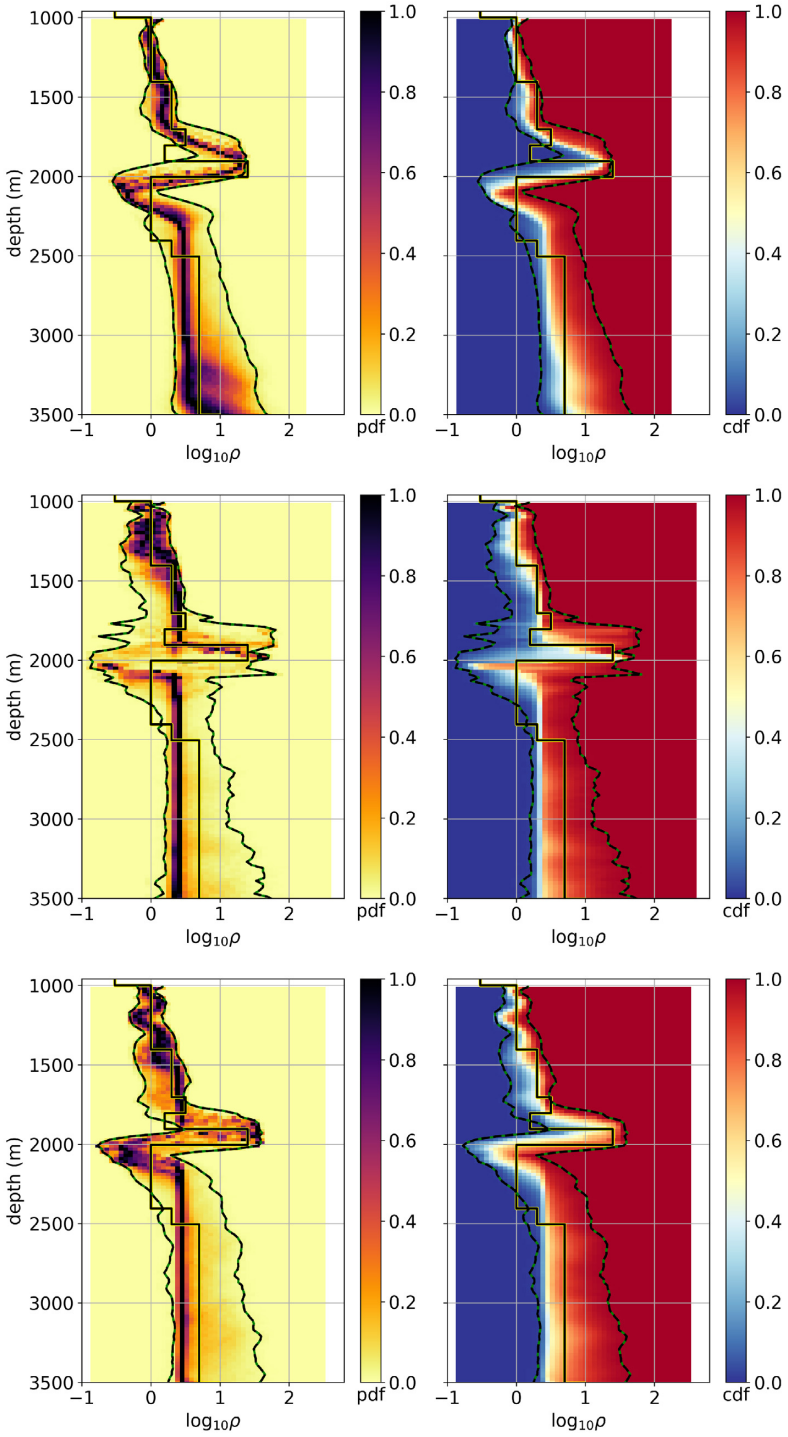

Figure 5. From top to bottom: three different correlations lengths $\lambda = 100$, 25 and 50 m chosen *a priori* and the resulting inversion posteriors. Note how assuming a longer correlation length provides a more optimistic picture of uncertainty—simpler models have less variance.

selection principles demonstrated for 1-D and 2-D earth models are equally applicable for 3-D inversion. However, Hawkins & Sambridge (2015) point out that computationally efficient parametrizations for trans-D problems in 2-D or 3-D (e.g. Piana Agostinetti *et al.* 2015; Burdick & Lekić 2017; Belhadj *et al.* 2018; Zhang *et al.* 2018) are not easy to construct (though it is certainly possible as the aforementioned 3-D applications show), or the specification of prior knowledge about geometric structure is difficult. The recent work of Hawkins & Sambridge (2015) has to some extent successfully overcome this issue. They demonstrate that any basis function set that is representable by a tree structure can be used as a valid model representation for trans-D inversion. As a consequence, tree-based trans-D is agnostic to the spatial dimensionality of the earth model, be it 1-D, 2-D or 3-D. This is a promising research route that allows us to tackle difficult high-dimensional problems

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

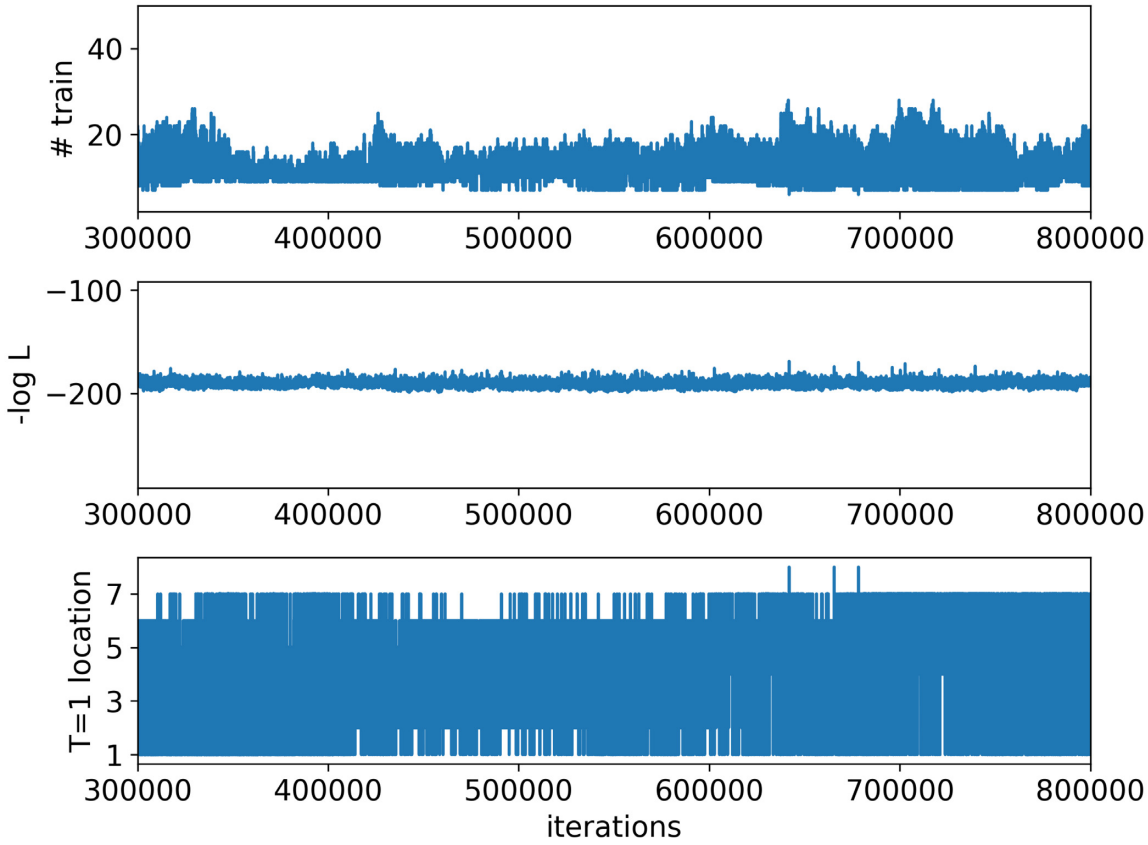

Figure 6. Stationary convergence statistics for $\lambda = 50$ m. Top: the number of training points k as a function of McMC sample number. Middle: the negative log likelihood, or the argument of the exponential in (8) together with a constant. This is a proxy for the χ^2 misfit when the data error is considered unknown. Bottom: Location of the unbiased $T = 1$ chain for carrying out posterior inference. Since we are using parallel tempering and exchanging temperatures between chains to facilitate navigation of 'peaky' likelihoods, this plot shows a healthy exchange of temperatures indicative of good 'chain mixing' and effective sampling.

such as probabilistic seismic full waveform inversion (see Ray *et al.* 2017), without making impractically limiting constraining assumptions about the posterior distribution. However, tree-based trans-D requires the specification of priors in a wavelet or other transform domain (e.g. Mallat 1989). This is not intuitive, requires some experimentation and abrupt cutoffs in the wavelet transform domain can lead to edge effects in the space domain. Calculating the prior probability of a dimension k requires a particularly clever 'memorized' computation (similar to using a lookup table) with the use of Big Integers to avoid integer overflow while counting arrangements of trees. Further, arbitrary earth model aspect ratios require the juxtaposition of more than one tree. While these difficulties are clearly not insurmountable (Dettmer *et al.* 2016; Hawkins *et al.* 2017), the trans-D-GP method avoids them altogether. In particular, prior specification with a different length scale in each dimension can be made in the familiar space domain as described in the following section.

2.3 Bayes' theorem

For observed data $\mathbf{d}$ and earth models $\mathbf{m}$ we can write:

$$p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m})p(\mathbf{m}). \tag{7}$$

Reading from right to left, $p(\mathbf{m})$ is the prior probability of $\mathbf{m}$, which we know independent of the observations $\mathbf{d}$. We re-assess our prior notion of $\mathbf{m}$ by carrying out an EM experiment that shows us how likely it is that $\mathbf{m}$ fits the observations. This weight is given by the likelihood function $p(\mathbf{d}|\mathbf{m})$. The result of re-weighting or updating our prior notion by the likelihood provides the *posterior* probability of observing the model $\mathbf{m}$. The posterior probability is represented by the term $p(\mathbf{m}|\mathbf{d})$.

The likelihood function $p(\mathbf{d}|\mathbf{m})$ for Gaussian data noise can be written as:

$$\mathcal{L}(\mathbf{m}) = p(\mathbf{d}|\mathbf{m}) = \frac{1}{\sqrt{|2\pi\mathbf{C_d}|}} \times \exp\left(-\frac{1}{2}\Big[\mathbf{f}(\mathbf{m}) - \mathbf{d}\Big]^{\mathrm{t}}\mathbf{C_d^{-1}}\Big[\mathbf{f}(\mathbf{m}) - \mathbf{d}\Big]\right), \tag{8}$$

where $[\mathbf{f}(\mathbf{m}) - \mathbf{d}]$ is the vector of misfit between the forward model calculation and the data for the model $\mathbf{m}$. The covariance matrix of data errors is given by $\mathbf{C_d}$. To be clear, $\mathbf{f}(\mathbf{m})$ represents the forward calculation for a k parameter trans-D model as represented by a GP mean. A k parameter prior model probability can be written as

$$p(\mathbf{m}) = p(\mathbf{m}_k, \mathbf{x}_k, k), \tag{9}$$

where $\mathbf{m}_k$ is a vector of GP 'training' resistivities. $\mathbf{x}_k$ is a vector in $\mathbb{R}^{k \times n_d}$ that specifies the locations of $\mathbf{m}_k$. n_d is the number of spatial dimensions of the model (e.g. $n_\mathrm{d} = 2$ for 2-D). Using the chain rule of probabilities, we can write:

$$p(\mathbf{m}_k, \mathbf{x}_k, k) = p(\mathbf{m}_k|\mathbf{x}_k, k)p(\mathbf{x}_k|k)p(k). \tag{10}$$

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

If we assume that each of k training resistivities can be independently and uniformly sampled within a log resistivity range $\Delta\rho$, and that we can arrange them in any of $k!$ ways uniformly within a length, area or volume given by $\prod_{i=1}^{n_d}\Delta x_i$, we can rewrite the above equation as

$$p(\mathbf{m}_k, \mathbf{x}_k, k) = \frac{1}{\Delta\rho^k}\frac{k!}{(\prod_{i=1}^{n_d}\Delta x_i)^k}p(k). \tag{11}$$

Common choices for $p(k)$, the prior probability on the number of interfaces are uniform $p(k) = \frac{1}{k_{\max}-k_{\min}+1}$ as we have used in our work here, or the Jeffreys (1939) prior where $p(k) = \frac{1}{k}$. The Jeffrey's prior is particularly useful in cases when the observed geophysical data are not informative. We highlight here that all prior specifications in (11) are done in the familiar domains of log-resistivity ρ and space $\mathbf{x}$, irrespective of the spatial dimension n_d of the earth model. We have tacitly omitted explicit mention of the length scales in $\mathbf{C}_\boldsymbol{\lambda}$, but they need to be specified *a priori* as can be seen through eqs (9), (5) and (2). We will describe the effects of selection of $\boldsymbol{\lambda}$ in detail in the next section and later in the text.

We repeat the process of finding the posterior probability $p(\mathbf{m}|\mathbf{d})$ for various models $\mathbf{m}$ admissible by our prior notions until we obtain an ensemble of models representative of the probability density function or PDF $p(\mathbf{m}|\mathbf{d})$. For the trans-D method we do this sampling using the Metropolis–Hastings–Green McMC algorithm (Metropolis *et al.* 1953; Hastings 1970; Green 1995; Hastie & Green 2012). Sampling proportional to the posterior probability is carried out by using the following acceptance probability to move from model $\mathbf{m}$ to $\mathbf{m}'$ in the McMC chain:

$$\alpha(\mathbf{m}'|\mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})}\left[\frac{p(\mathbf{d}|\mathbf{m}')}{p(\mathbf{d}|\mathbf{m})}\right]^{1/T}\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}|\mathbf{J}|\right]. \tag{12}$$

$\mathbf{m}$ is perturbed to $\mathbf{m}'$ via a proposal PDF $q(\mathbf{m}'|\mathbf{m})$. The Jacobian determinant term $|\mathbf{J}|$ is not to be confused with the model Jacobian needed for gradient-based inversions (e.g. Constable *et al.* 1987), but is a matrix that incorporates changes in model dimension when moving from $\mathbf{m}$ to $\mathbf{m}'$. There are various implementations of trans-D McMC, and in all the examples cited so far, a 'birth-death' scheme (Geyer & Møller 1994) has been used. As shown in Bodin & Sambridge (2009), Dettmer *et al.* (2010) and Sen & Biswas (2017) for most 'birth-death' trans-D McMC schemes, $|\mathbf{J}|$ is unity. We have adopted the 'birth-death' algorithm in this work. T is a tempering parameter used to anneal hard-to-sample likelihoods, with $T = 1$ used for unbiased sampling in a sequence of interacting Markov chains (see Dettmer & Dosso 2012 for details). Detailed expressions for the acceptance probabilities α are given in Appendix A. Pseudo-code for trans-D McMC sampling with interacting chains is provided in Algorithm 1.

3 CSEM INVERSION

The marine controlled source EM (CSEM) method is an active source sounding technique. It has been in use for over three decades for the detection of geology with high resistivity contrasts (Young & Cox 1981; Chave & Cox 1982). Conductive media such as sea-water or brine filled sediments have a characteristic EM scale length (skin depth) $\delta = \sqrt{\frac{2\rho}{\mu\omega}}$ that is dependent on both the medium resistivity ρ and the frequency of propagation ω, where μ is the permeability of the medium. Owing to the fact that δ is smaller in conductive (low ρ) media, marine geophysical EM methods operate in the lower frequency quasi-static regime with physics that is more diffusive than wave like (Loseth *et al.* 2006). To first order, it is this diffusive decay that can characterize the conductivity of a given medium. For hydrocarbon bearing geology, it is the high resistivity of the hydrocarbon accumulation with respect to its surroundings that produces a detectable EM signature. This signature is different from what would have been observed in the absence of hydrocarbons (e.g. Constable 2006, 2010). However, reliable inferences from CSEM can only be made by means of an inversion, and a Bayesian inversion is ideal to quantify the uncertainty inherent in the inversion process (e.g. Hou *et al.* 2006; Chen *et al.* 2007; Gunning *et al.* 2010; Buland & Kolbjornsen 2012). The aforementioned references, while Bayesian, used a fixed number of dimensions k dictated by the user and not by the likelihood. Trans-D Bayesian methods have been used to invert CSEM data with both 1-D and 2-D parametrizations (e.g. Ray & Key 2012; Ray *et al.* 2014). Though the theory was similar in both cases, the implementation of the trans-D method required parametrization with layers/interfaces for a 1-D earth and Voronoi cells for 2-D.

In the following sections, we demonstrate how our new (spatial) dimension agnostic method, parametrized with GPs (trans-D-GP), can be used to efficiently perform Bayesian trans-D inversion. Since CSEM forward computation is computationally expensive and all computations in this paper were carried out on a 4-core / 8-thread processor @2.8 GHz, CSEM inversions were done in 1-D, and a nonlinear regression inverse problem is carried out in 2-D.

3.1 Synthetic examples

3.1.1 Prior sampling

Before we turn to real CSEM data in the next sections, we demonstrate the effectiveness of our method on synthetic CSEM data modelled on our knowledge of the Scarborough gas field (see Driscoll & Karner 1998; Myer *et al.* 2013 for details). First and foremost, before using any sampling based Bayesian method with any data we advise a thorough examination of the sampled prior in the absence of informative data. By setting the likelihood function to a constant in (7)—the equivalent of assigning all sampled models the same misfit—we are only sampling the prior (11). The sampled prior should reflect what we have specified in the theory. When implementing a new algorithm, small errors in the theoretical derivation or the actual computer code can lead to serious biases in the prior specification, which will in turn affect the posterior sampling (see Piana Agostinetti & Malinverno 2010 for a discussion). Prior sampling is shown in Fig. 3. In (a) we plot the joint prior PDF of $\mathbf{m}_k$, $\mathbf{x}_k$. As theoretically specified by our choice of prior, it is near-uniform. This is clear also from (b) where depth has been marginalized out of the joint pdf and we see the prior pdf on $\mathbf{m}_k$ which are the training values for resistivity. In (c) the resistivity has been marginalized out and we see the prior PDF on $\mathbf{x}_k$, which in this 1-D case is the prior PDF of placing training points in z. Finally in (d) we show the marginal pdf on k, the number of training points sampled, also close to uniform as was specified in the prior. This shows that the sampling of the prior works according to the theory specified.

We would now like to examine the resulting high-dimensional models $\boldsymbol{\mu}_*(\mathbf{m}_k)$, as shown in Fig. 4, to see if the GP parametrization samples appropriately the region of high-dimensional earth model space we are interested in.

On the left-hand panel of Fig. 4 are shown marginal PDFs of resistivity with depth, which tend to be a little jittery, and on the right are shown the corresponding cumulative density functions (CDFs) which of course are smoother. At every depth, in accordance

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

with the definition of a PDF the sharpest change in the CDF marks the area of highest probability in the PDF. All sampled 'training' points $\mathbf{m}_k$ are de-meaned before we apply (5), after which the mean is added back in. This results in the marginalized prior $\boldsymbol{\mu}_*$ being centred at the middle of the interval $[-0.5, 2.3]$. There are small 'edge effects' at both ends of the z co-ordinate we could address by extending the sampling domain through 'padding' (see Turner 2011 for more sophisticated methods) but have elected not to do so as they are relatively minor. In both panels, the 98 % Bayesian credible interval (CI) is the zone at every depth in-between the dashed lines, with 2 % of resistivities outside of it. The synthetic resistivity model we are going to invert for (in black) is well within this interval. It must be mentioned that PDFs at all depths are normalized to one to show the location of maximum probability. We had chosen a λ of 50 m, implying that resistivities are highly correlated within 50 m of each other. Equally important, is the fact that we have specified a σ_{m} value or 'noise' in the observed 'training' points at 0.2. This corresponds to 7 % of the interval as wiggle room within which to fit the training points. The lower σ_{m} is, the more the GP tries to fit the training points exactly, which causes $\boldsymbol{\mu}_*$ to oscillate more. This causes the CI to get wider and allow extreme, δ-function-like resistivities. Some experimentation of this sort is necessary to set these hyperparameters for the algorithm. Since no forward computations are required, this is not a time consuming exercise.

3.1.2 Posterior sampling

Using the prior probabilities specified in the above subsection, we performed a set of synthetic experiments based on the geology of the Scarborough gas field in the North West Australian shelf. The target layer, as shown in the previous experiments for sampling the prior, is at 1900–2000 m depth with a resistivity of 25 ohm-m ($\log_{10}$ of 1.4). However, a confounding layer corresponding to the Gearle siltstone formation with a moderately high resistivity of 3.16 ohm-m ($\log_{10}$ of 0.5) at 1700–1800 m depth is also included in the model. Previous studies have shown that at typical CSEM noise levels, it is not possible to invert both of these closely spaced resistive bodies in the same earth model (Myer *et al.* 2012). However, the bulk resistivity amounting to a sizable hydrocarbon saturation or its absence near 2000 m depth can indeed be inverted as shown by Myer *et al.* (2015) and Ray *et al.* (2014). In the synthetic study, frequencies at 0.25, 0.5, 0.5, 1.75 and 3.25 Hz were used for the inversion, the same as in the actual field studies. Gaussian noise proportional to 5 % of the amplitude was added at every receiver, independently to the real and imaginary radial inline electric fields, with a source normalized noisefloor of $10^{-14} V/(Am^2)$. The transmitter was placed at 975 m depth, with 32 seafloor receivers placed at 1000 m depth. The receivers were spaced at intervals of 177 m between inline radial offsets of 500 and 6000 m.

As we have mentioned in Section 2.3, implicit in the way we have defined the prior probability (11) is a correlation length λ. It is a fact of most inversion methods, whether deterministic, Bayesian or ML based, that we must include constraints or reliable prior knowledge. In a hierarchical Bayesian framework, we can place *hyper-priors* on the priors themselves, and sample over a range of priors to understand sensitivity to different prior specifications. We could do the same here, except that it will require reconstructing the low dimensional matrix $\mathbf{K}_m$ and the large dimensional matrix $\mathbf{K}_*$ in (5). Though this is not a concern for a 1-D earth, matrix construction for a large 2-D or 3-D model can be time consuming. Further, allowing local length scales to be inferred from the data is the direction we would like to proceed in as discussed later in Section 5. In this section, we instead study the impact of choosing correlation lengths of 100, 25 and then 50 m as can be seen in Fig. 5. As predicted in all cases, the data were unable to resolve the Gearle layer, though high resistivity corresponding to the reservoir was found by the marginalized posterior resistivity distribution peaking near 2000 m depth. The marginalized CDFs provide, in our opinion, the clearest indications of a resistive anomaly at depth.

The convergence plots for the inversion with λ equal to 50 m are shown in Fig. 6. The algorithm was run for 800 000 iterations, the last 500 000 of which were used to infer the posterior and ensure that samples are not trapped in low-probability regions. The number of training points is shown in the first row indicating that the algorithm never used any less than five training points k nor any more than 35, with a mean of 15. The prior limits were set for k between 2 and 50. The middle row shows the negative log-likelihood, a proxy for the χ^2 misfit in (8). We did not consider the data noise to be known, and found a maximum likelihood estimate of the noise per frequency, as detailed by eq. (B10) in Appendix B. This is a common signal processing approach borrowed from the field of geoacoustics (e.g. Mecklenbrauker & Gerstoft 2000; Dosso & Wilmut 2012). We used parallel tempering (PT; Swendsen & Wang 1987; Geyer 1991; Dettmer & Dosso 2012; Ray *et al.* 2013a; Sambridge 2013; Bottero *et al.* 2016) with eight interacting Markov chains to accelerate convergence. PT ensures that the likelihood is thoroughly explored via a sequence of concurrently running McMC chains with gradually annealed likelihoods. By this we simply mean that the temperatures T are logarithmically spaced. In optimization parlance, this provides a good means of escaping local misfit minima. However, instead of exchanging models between adjacent chains as is traditionally done (e.g. Earl & Deem 2005), we allowed any chain to exchange information with any other chain, which allows for more efficient sampling (Sambridge 2013). Further, the exchange of temperatures, especially in a parallel computing environment is equivalent to the exchange of models, but more efficient as it cuts down the parallel communication overhead. We show the exchange of temperatures in the third row of Fig. 6 from which we can infer healthy exchange of information between different McMC chains. Rapid changes in the number of training points with sample number are also indicative of efficient sampling. Since the birth/death acceptance rates within a single chain can be on average as low as 3%, a well known difficulty of using trans-D methods, we circumvent this issue with PT exchanges-between-chains. Average acceptance rates for change of spatial location x (here the z co-ordinate) were 53%, and for change in property m, here conductivity, were 21%. Posterior inference was carried out as usual with the McMC chain that is not annealed, with $T = 1$. Eight McMC chains with logarithmically spaced temperatures between 1 and 2.5 were run in parallel using PT for a total runtime of 21 hr.

The data fit for 100 randomly chosen posterior models for the $\lambda = 50$ m inversion is shown in Fig. 7. Note how the assumption of Gaussian noise has been qualitatively met, given that there are no large outliers. This indicates that our maximum likelihood method for estimating data noise within our McMC scheme is working as expected. The convergence statistics for the longer and shorter λ were similar, with all target chains able to sample traditionally calculated root mean square (RMS) data errors ≈ 1. However one crucially different aspect in the three cases is that the shorter $\lambda = 25$ m inversion sampled on average a higher number of points k, while the longer $\lambda = 100$ m inversion sampled on average a lower number of points k than $\lambda = 50$ m. The effect of λ can also be seen in the widths of the respective credible intervals in Fig. 5. With the

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

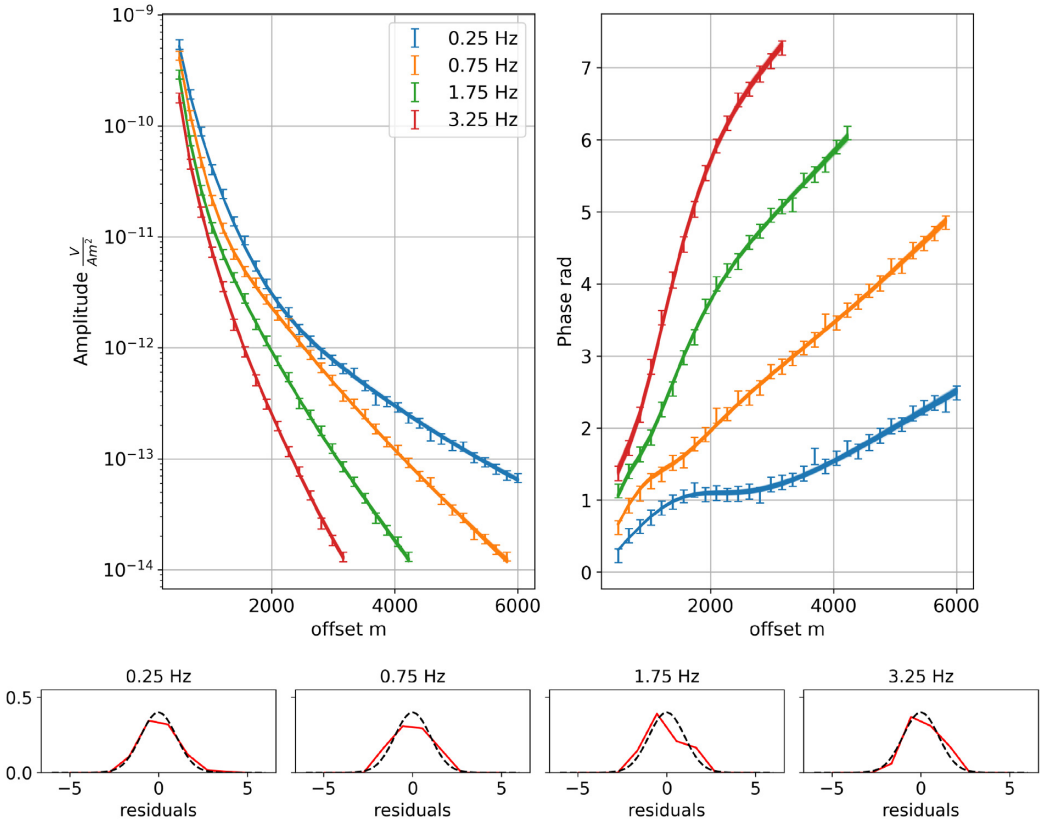

Figure 7. Top: Data fit for $\lambda = 50$ m and 100 randomly selected posterior models. The error bars correspond to 2σ Gaussian noise at 5 per cent relative to $|E_r|$. Bottom: the data noise was an unknown in the inversion and estimated through a maximum likelihood procedure at every McMC step. The histogram of standardized posterior inversion residuals at every frequency (in red) are shown together with an analytic Gaussian (dashed black).

highest λ (top row) the anomalous resistivity distribution at reservoir level is quite smooth but it appears that posterior distributions of resistivity are narrow. With the lowest λ (middle row) we get more 'resolution' of the reservoir anomaly and can separately infer the reservoir top and bottom. However, the posterior distribution of resistivity is quite broad. Based on this study, we chose to go with $\lambda = 50$ m (bottom row of Fig. 5) for the Scarborough real data CSEM inversion.

At this point, we would like to point out that had we used a trans-D parametrization with layers, we would effectively have chosen $\lambda = 0$ at interfaces and $\lambda = \infty$ in-between interfaces. Using the GP-based prior, though we need to fix λ, we at least give ourselves a choice about the correlation length of geology in the earth—which is certainly not zero or infinity. The same argument holds for Voronoi cells and abrupt changes in 2-D. We have within this restriction of a chosen correlation length, allowed the data and Bayesian parsimony to determine the location and number of training points that define an earth model **m**. If we compare with 'classic' layered trans-D CSEM inversion results as shown in fig. 10 of Ray *et al.* (2014), the posterior using trans-D-GP is always smoother, as we should expect—given that the choice of prior parametrization determines the behaviour of inferred posterior models (see Hawkins & Sambridge 2015; Ray *et al.* 2017 for a discussion).

3.2 Scarborough field CSEM inversion

We applied trans-D-GP to data from the Scarborough gas field, which lies inside the Exmouth Plateau in the North West Australian Shelf. The plateau is covered by a number of nearly horizontal layers with resistivity varying between 1 and 10 ohm-m (Myer *et al.* 2012). Five exploration wells have been drilled in the Scarborough gas field and the well data together with 3-D seismic data were used to delineate the approximate extent of the reservoir. The reservoir itself is between 20 and 30 m thick at a depth of ∼2000 m below sea level. The bathymetry, also quite flat is at a depth of ∼950 m. Resistivity at reservoir level is moderate at 25 ohm-m and the reservoir is overlain by several thin 5–10 ohm-m layers. We inverted data from two sites located in the 'off reservoir' and 'on reservoir' parts of a CSEM tow-line. The posterior resistivity with depth for the both sites is shown in Fig. 8. In the off-reservoir part there is evidence of weak 8–10 ohm-m anomalies with accompanying changes in the CDF above 2000 m depth. Contrast this with the on reservoir posteriors indicating high probability of moderately resistive material of 10–25 ohm-m at similar depths. Our results are in line with the previous findings of Ray *et al.* (2014), who showed that the posterior PDFs of resistivity (not just the mode) near 2000 m depth move *en masse* to more resistive values as we tow the transmitter from off-reservoir to on-reservoir sites. Interestingly, the 'jumping' back and forth between conducting to resistive or multimodal nature of posterior resistivities between 1500 and 2500 m depth is also visible in previous studies of the area (see Ray *et al.* 2014). We conjecture that this is a sign of macroscale conductivity anisotropy due to rapidly alternating (in depth) layers of resistive shale (or siltstone) and briny conducting fluid fill.

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

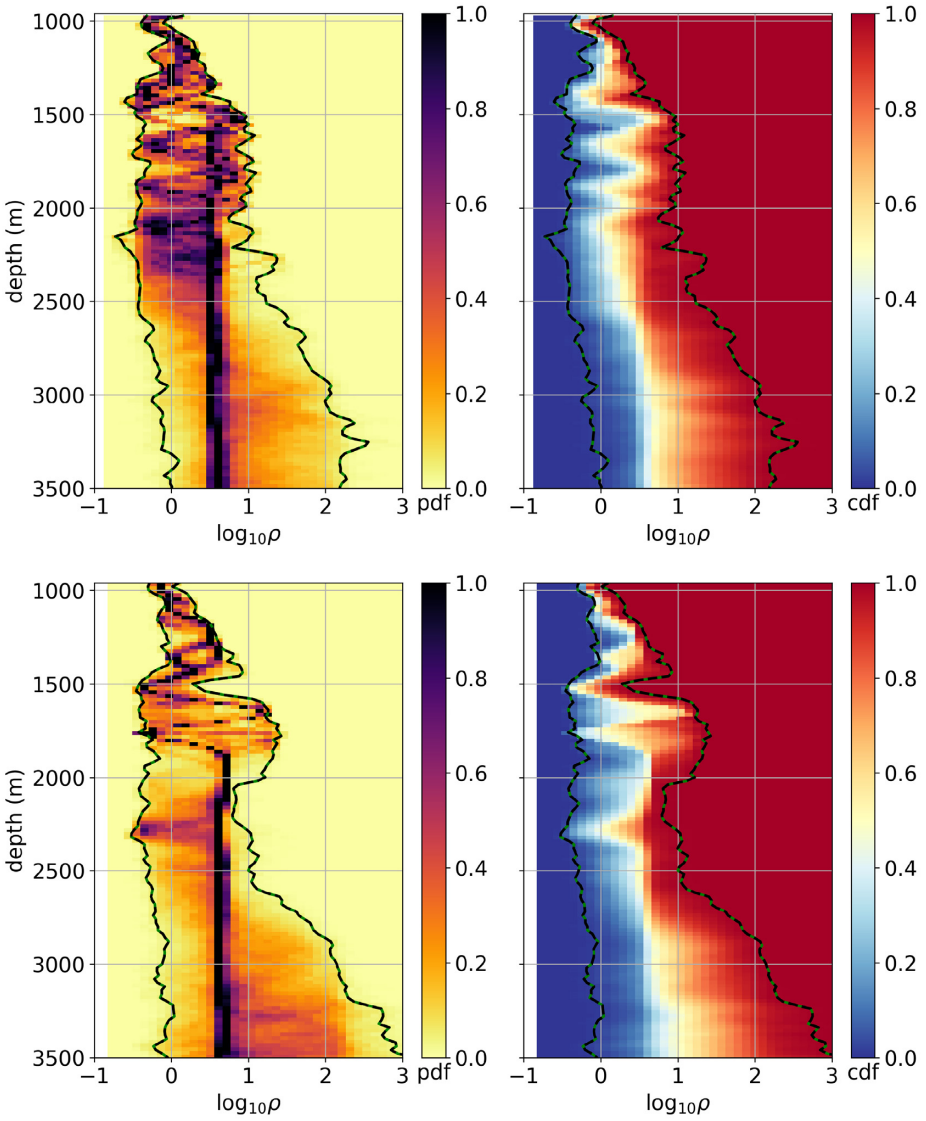

Figure 8. Top: Off reservoir posterior resistivity distributions. Bottom: On reservoir posterior resistivity distributions. Note how the resistivity PDFs near and above 2000 m depth shift in bulk to more resistive for the on reservoir case. This is indicative of the resistive hydrocarbon bearing reservoir as also found by Ray *et al.* (2014).

The algorithm was run for 2 000 000 samples with the last 1 000 000 samples used for posterior inference. Data fits and convergence statistics for both sites can be seen in Figs 9 and 10. Convergence for the off reservoir case as seen in Fig. 9 is perhaps questionable, given that the sampling of the number of training points seems not to be stationary. However, the sampled square misfit or negative log likelihood is indeed stationary, and the values of posterior resistivity with depth do not change appreciably if inference is made using the last 1 000 000 samples or the last 500 000. As noted by Bodin & Sambridge (2009) conventional McMC diagnostics are not useful for trans-D given that the number of parameters vary from step to McMC step. We have followed their approach of instead focusing on near-stationarity in the values of geophysical property (resistivity in our case) at spatial locations across the model. Given the similarity with previous results using an entirely different parametrization (Ray *et al.* 2014), we deem the target chain converged for all practical purposes. While we agree that accounting for correlated data error is necessary for drawing robust inferences, we have at least attempted to ensure that our residuals are Gaussian. We note that correlated data error is a significant source of confusion for posterior inference. It is also the likely cause for the greater number of samples to reach convergence than in the synthetic studies. Though we have not attempted to deal with correlated error here, see Ray *et al.* (2013b) for hierarchical approaches to data covariance matrix estimation. Another approach to reducing inversion artefacts from correlated data error is to use a 2-D parametrization and treat navigation data with a common mid-point approach (e.g. Ray *et al.* 2014). Of course, it would be best to forward model this data with a 2-D earth model using a 3-D source (e.g. Key & Ovall 2011). This was not possible given the computational resources available to us, though it is well within the means of academic research consortia and industry. We have instead tried to deal with inconsistent and correlated data error estimates by using maximum likelihood data estimates and by inverting both the in-tow and out-tow data – the errors in which, at similar offsets, are not correlated. Given that our

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

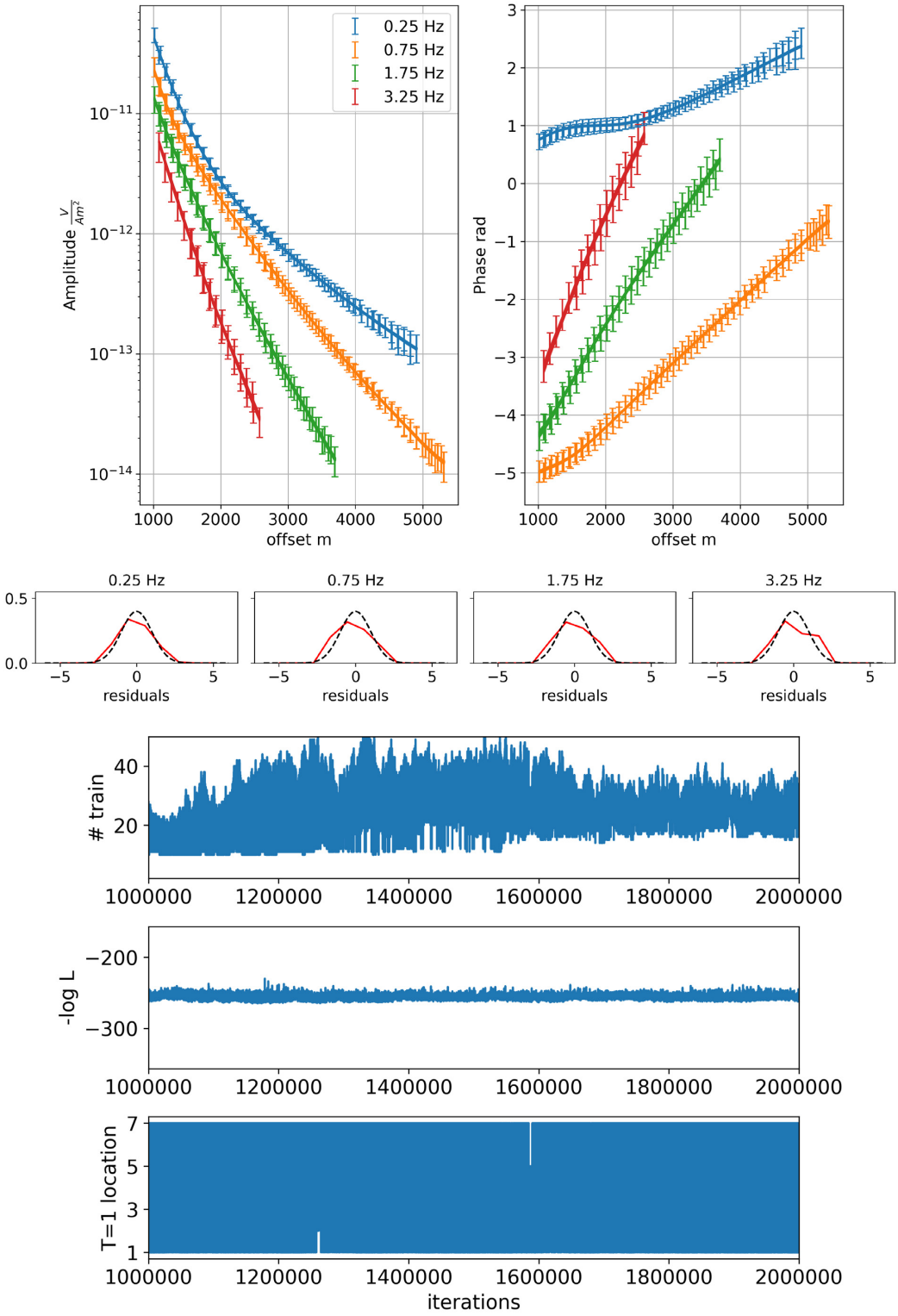

Figure 9. Off reservoir data fit, inversion residuals and convergence statistics. The data noise was considered an unknown in this inversion and maximum likelihood estimates were used in the likelihood function. 2σ error bars are from the error analysis made by Myer *et al.* (2012).

results are in line with previous work, this demonstrates that our trans-D-GP methodology works in a real-world setting.

4 EXTENSION TO 2-D: A NONLINEAR REGRESSION APPLICATION

As an example of extending to higher spatial dimensions, we solve a nonlinear regression problem with 'non-function data.' By this we mean the data to be fit are not the outcome of a single valued function

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

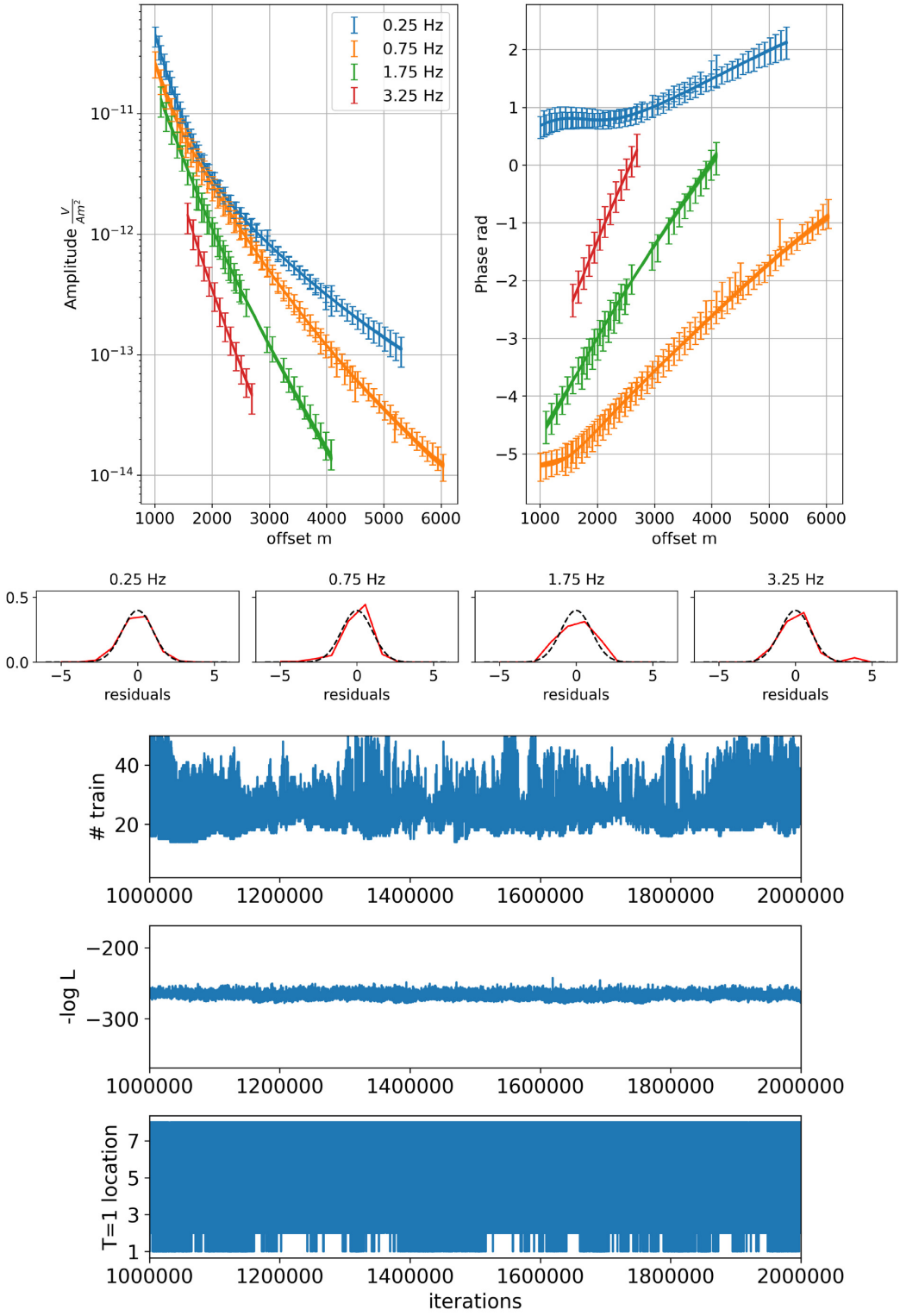

Figure 10. On reservoir data fit, inversion residuals and convergence statistics. The data noise was considered an unknown in this inversion and maximum likelihood estimates were used in the likelihood function. 2σ error bars are from the error analysis made by Myer *et al.* (2012).

as a function of a distance co-ordinate (see Criminisi *et al.* 2011 for further examples). We could also think of this as a 2-D spatial regression problem. Geoscientific applications of this type using trans-D methods have been investigated by Gallagher *et al.* (2011) and Bodin *et al.* (2012a). Depending on the specifics of the problem, they used interfaces for one spatial dimension and Voronoi cells for two. However with a GP, the exact same theory in Section 2.1 holds no matter the number of spatial dimensions. With a different misfit function, using the same code as we did for the CSEM problem we now solve a problem involving a parameter space with two spatial

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

dimensions (Fig. 11). On the left is a low-passed 256 × 256 image of the standard test image 'splash' available from the SIPI database at the University of Southern California (http://sipi.usc.edu/database/). On the right, we sample at random 851 of the original 65 536 pixels, deliberately sampling the upper part sparsely to see how the algorithm adapts to irregular non-stationary data coverage. Random Gaussian noise with standard deviation equal to 5 % of the max value is also added. The objective is to find 2-D representations (and also their uncertainty) that approximate the true image at locations not sampled. Naturally, one could use kriging methods to solve this problem, but we are interested in one further property that a standard kriging methodology cannot ensure. We would like *parsimonious* representations of this image, as we would require for geophysical applications over a spatially vast part of the earth, forward modelling the physics for which we would require many pixels. Further, we demand that the data coverage and noise levels should determine the complexity of the model representation(s) in concert with our prior knowledge. To these ends, we define a likelihood function such that the residual misfit vector is simply the difference in the values on the right of Fig. 11 from the values of sampled GP models $\boldsymbol{\mu}_*(\mathbf{m})$ at those same spatial locations. Similar to the CSEM case, the data noise variance was determined using a maximum likelihood method (see eq. B5 of Sambridge 2013).

The progress of trans-D sampling with $\mathbf{C}_\lambda = \begin{pmatrix} 100^2 & 0 \\ 0 & 100^2 \end{pmatrix}$, that is, λ set to 100 in both spatial dimensions is shown in Fig. 12. Note how the misfit (negative log likelihood) decreases as the model complexity k increases, achieving near-stationarity 20 000 samples onwards. $38 < k < 66$ after achieving stationarity, though the maximum permissible prior value for k is 100. Sample 50 000 is shown in the left of Fig. 13 with the accompanying 55 training points needed to define the model. The mean of samples from 50 000 onwards is shown on the right. Note that this is a nonlinear process and multimodal distributions of parameter values cannot be represented by only a mean and a variance. Another advantage of our trans-D-GP method, unlike a traditional GP with a unimodal Gaussian at every spatial location, is that the full posterior distribution of inverted parameter values can be shown at any spatial location (e.g. Ray *et al.* 2017; Galetti & Curtis 2018). This has already been evidenced by the CSEM examples where posterior distributions of resistivity at certain depths were seen to be multimodal. In Fig. 14 we show how our method adapts to *both* model complexity as well as the manner in which the data have been sampled, a hall-mark of trans-D methods that we have preserved in our algorithm. Where there is less data, there should be high posterior uncertainty, as we can see in the figure to the left. On the right, we can see that where the data are informative, there is a dense nucleation of GP points.

5 INTRODUCING LOCAL LENGTH SCALES

As mentioned in Section 3.1.2, we would like to put forward the idea of allowing length scales $\boldsymbol{\lambda}$ which can vary spatially. This 'stationarity' of the length scale is not a requirement, as has been proved by various workers such as Gibbs (1997), Higdon (1998) and Paciorek (2003). Following the approach of Paciorek (2003) we redefine the GP kernel $K(\mathbf{y}, \mathbf{y}')$ (2) (see eqns 4.33 and 4.34 of Rasmussen & Williams (2006)):

$$K_{ns}(\mathbf{y}, \mathbf{y}') = 2^{n_d/2}|\Sigma|^{1/4}|\Sigma'|^{1/4}|\Sigma + \Sigma'|^{-1/2} \times \exp\left(-\frac{1}{2}[\mathbf{y}-\mathbf{y}']'\mathbf{C}_{\lambda_{ns}}^{-1}[\mathbf{y}-\mathbf{y}']\right), \quad (13)$$

where

$$\mathbf{C}_{\lambda_{ns}} = \frac{\Sigma + \Sigma'}{2}, \quad (14)$$

and Σ, Σ' are the local length scale covariances at spatial locations $\mathbf{y}$ and $\mathbf{y}'$.

Using this approach we were able to find a GP mean with minimal trial-and-error, that can approximate a 1-D function with two abrupt changes (Fig. 15). Similar to Plagemann *et al.* (2008), in the bottom row we show the smoothly varying length scale over the abruptly varying function. This variation in λ enabled us to model the true function with less oscillation than the GP mean with a fixed λ. Though we have not used this technique in our algorithm, it can in principle be used to make $\boldsymbol{\lambda}$ a spatially varying model parameter. It would require us to use another GP to model the non-stationary length scales $\boldsymbol{\lambda}$ everywhere (e.g. Plagemann *et al.* 2008), given that $\boldsymbol{\lambda}$ is defined sparsely at a few spatial locations. This will add twice as much computational overhead for modelling a GP, but for large-scale models where this time is negligible compared to the forward modelling time, coupled with the fact that $\mathbf{K}_m$ and the length scales for the continuous property $\boldsymbol{\lambda}$ need only be defined by fewer than 100 points, this is not such a hindrance as it may at first seem. Given that the trans-D algorithm is able to place training points with appropriate earth model property values (e.g. conductivity) at appropriate spatial locations, this idea should extend hierarchically to appropriately inferring the unknown length scales over the unknown earth property values. The idea belongs to the 'learning to learn' paradigm in ML (e.g. Andrychowicz *et al.* 2016; Chen *et al.* 2017). To demystify this line of thought we might say that in order to learn the earth's properties, we must also learn its length scales.

6 CONCLUSIONS

We have developed a new methodology that incorporates the well-known Gaussian Process ML technique into a parsimonious trans-D framework, demonstrating its use in 1-D, 2-D and field applications. We have shown that ML techniques can be easily incorporated into a Bayesian geophysical inversion framework through the specification of prior information (e.g. Laloy *et al.* 2017). While our method does adapt earth model complexity according to the data noise and receiver coverage, it is not truly multilength scale (e.g. Hawkins & Sambridge 2015, whose tree method is multiscale). However in Section 5, we have shown with examples that it is theoretically possible and perhaps even desirable to incorporate multiple length scales into the technique. The key advantage in using our method is the simplicity of prior specification and ease of representing large-scale models in 1-D, 2-D or even 3-D. Further, the inclusion of 'fixed' prior values in the earth model may be achieved by keeping part of $\mathbf{K}_*$, $\mathbf{K}_m$ and $\mathbf{m}$ fixed, discounting these elements from the trans-D count k. This is possible as GPs are based on conditional realizations of Gaussians, while this is not easily done in dimension-reduced latent-space methods (see Laloy *et al.* 2018 for workarounds). We contend that for low-frequency geophysical inversion, trans-D-GP is simple enough to implement from scratch without the use of an ML library and provides the scalability for inverting large 2-D or 3-D earth models with a small number of effective parameters. Last but not least, it provides uncertainty estimates on inverted earth properties.

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

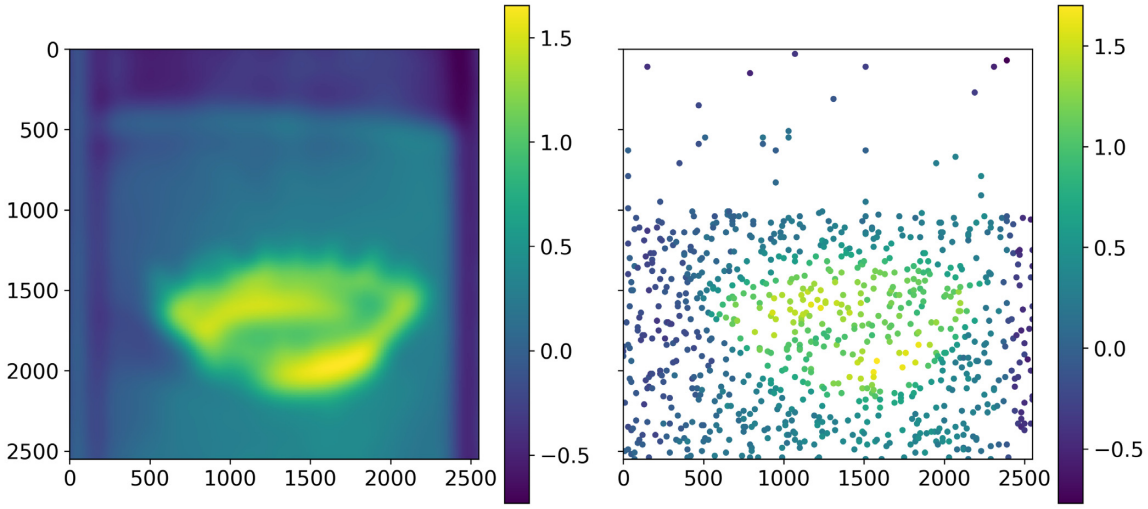

Figure 11. Left-hand panel: the 'splash' test image downsampled to 256×256 and lowpassed with a 7×7 Gaussian kernel. Right-hand panel: unequally and irregularly sampled data from the left image with added Gaussian noise.

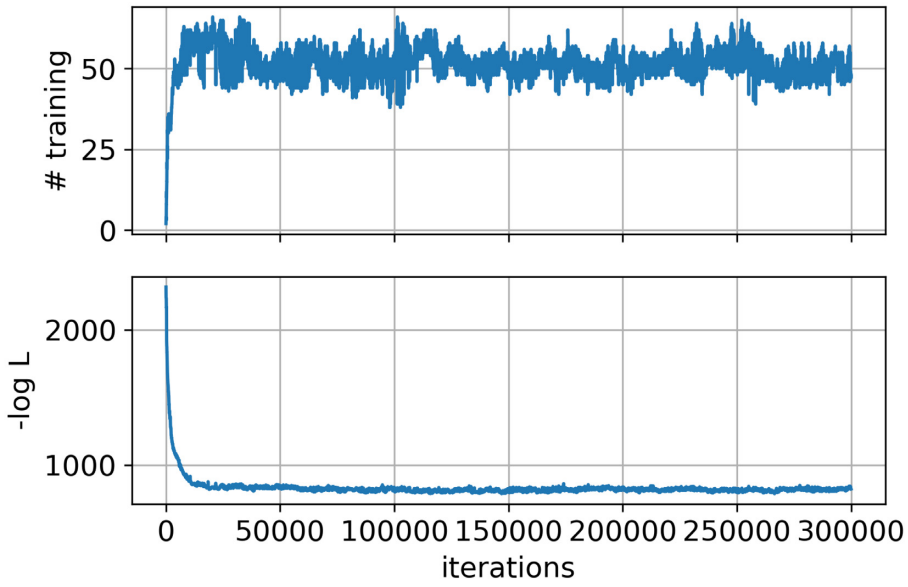

Figure 12. Top: number of training points k required in the reconstruction of the 'splash' test image. Bottom: misfit reduction by the trans-D McMC chain. Note how k never rises above 66 though the permissible maximum is 100.

ACKNOWLEDGEMENTS

All calculations were carried out using the Julia language (Bezanson *et al.* 2017, 2015, 2012), available under the MIT license. The authors thank BHP Billiton Petroleum and the Seafloor Electromagnetic Methods Consortium at the Scripps Institution of Oceanography for making the Scarborough data publicly available. Eric Laloy and Niklas Linde provided many useful insights into the use of GANs and TIs in a Bayesian context. AA read an early draft manuscript and greatly improved its readability. IB provided uninterrupted computation on their machine for which we are grateful. We would like to thank Jan Dettmer and Andrew Curtis for detailed and constructive reviews which have greatly improved the exposition of our methodology.

REFERENCES

Abubakar, A., Habashy, T.M., Druskin, V.L., Knizhnerman, L. & Alumbaugh, D., 2008. 2.5D forward and inverse modeling for interpreting low-frequency electromagnetic measurements, *Geophysics,* **73**(4), F165–F177.

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D.W. & O'Neil, M., 2016. Fast direct methods for Gaussian processes, *IEEE Trans. Pattern Anal. Mach. Intell.,* **38**(2), 252–265.

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M.W., Pfau, D., Schaul, T. & De Freitas, N., 2016. Learning to learn by gradient descent by gradient descent, in *Advances in Neural Information Processing Systems,* pp. 3981–3989, eds., Jordan M.I., LeCun, Y. & Solla, S. A., MIT Press.

Belhadj, J., Romary, T., Gesret, A., Noble, M. & Figliuzzi, B., 2018. New parameterizations for Bayesian seismic tomography, *Inverse Problems,* **34**(6), doi:10.1088/1361-6420/aabce7.

Bezanson, J., Karpinski, S., Shah, V.B. & Edelman, A., 2012. Julia: a fast dynamic language for technical computing, arXiv:1209.5145, pp. 1–27.

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

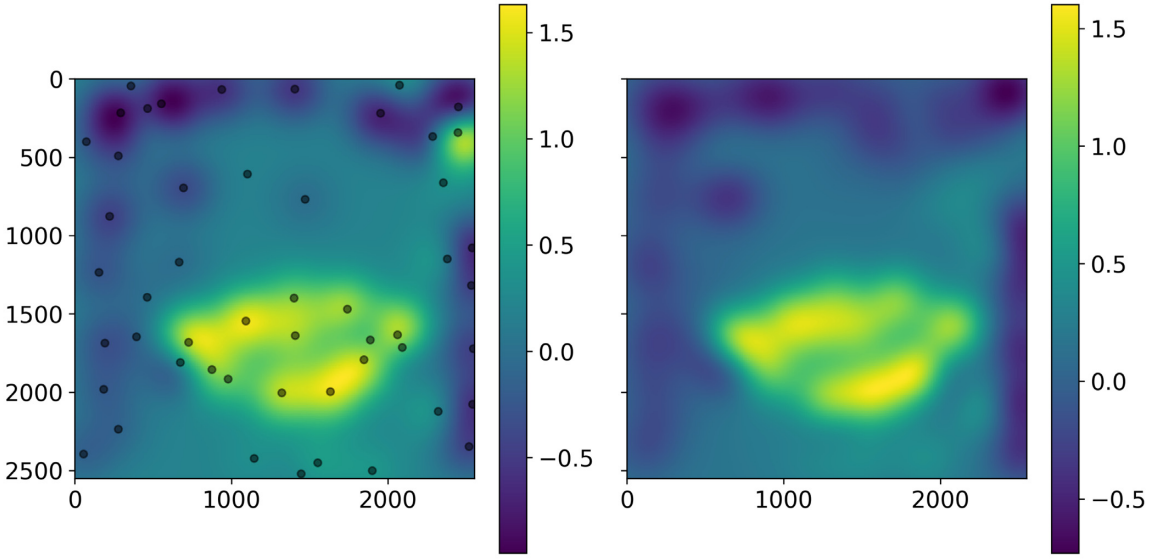

Figure 13. Left-hand panel: $\boldsymbol{\mu}_*(\mathbf{m})$ at sample 50 000 and the 55 training points needed to define it. Right-hand panel: the mean of $\boldsymbol{\mu}_*(\mathbf{m})$ samples 50 000 onwards.

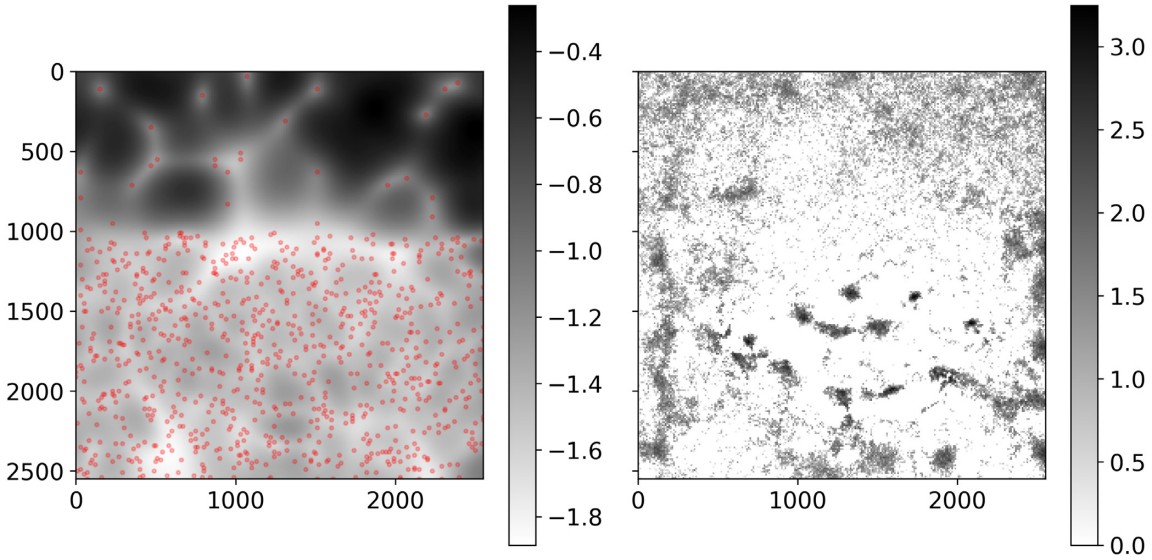

Figure 14. Left-hand panel: $\log_{10}$ of the standard deviation of posterior samples. Locations of the noisy data are overlain in red. Right-hand panel: $\log_{10}$ of the hit count of posterior training samples in the space domain. On the left we can see high standard deviation (darker shades) when the data coverage is poor, as we should reasonably expect. On the right, we see that sampled points for the GP parametrization are densely nucleated near resolvable features in the model, and loosely clustered when there is poor data coverage or features are not resolvable. This indicates that the algorithm adapts to complexity in the model as well as the density of the observations.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V.B., 2015. Julia: a fresh approach to numerical computing, *SIAM Rev.,* **59**(1), 1–37.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V.B., 2017. Julia: a fresh approach to numerical computing, *SIAM Rev.,* **59**(1), 65–98.

Blatter, D., Key, K., Ray, A., Foley, N., Tulaczyk, S. & Auken, E., 2018. Trans-dimensional Bayesian inversion of airborne transient EM data from Taylor Glacier, Antarctica, *Geophys. J. Int.,* **214,** 1919–1936.

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

Bodin, T., Salmon, M., Kennett, B.L.N. & Sambridge, M., 2012a. Probabilistic surface reconstruction from multiple data sets: an example for the Australian Moho, *J. geophys. Res.,* **117**(B10), B10307, doi: 10.1029/2012JB009547.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K. & Rawlinson, N., 2012b. Transdimensional inversion of receiver functions and surface wave dispersion, *J. geophys. Res.,* **117**(B2), doi:10.1029/2011JB008560.

Bottero, A., Gesret, A., Romary, T., Noble, M. & Maisons, C., 2016. Stochastic seismic tomography by interacting Markov chains, *Geophys. J. Int.,* **207**(1), 374–392.

Broomhead, D.S. & Lowe, D., 1988. Radial basis functions, multi-variable functional interpolation and adaptive networks, *Tech. rep.,* Royal Signals and Radar Establishment Malvern, United Kingdom.

Buland, A. & Kolbjornsen, O., 2012. Bayesian inversion of CSEM and magnetotelluric data, *Geophysics,* **77**(1), E33–E42.

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

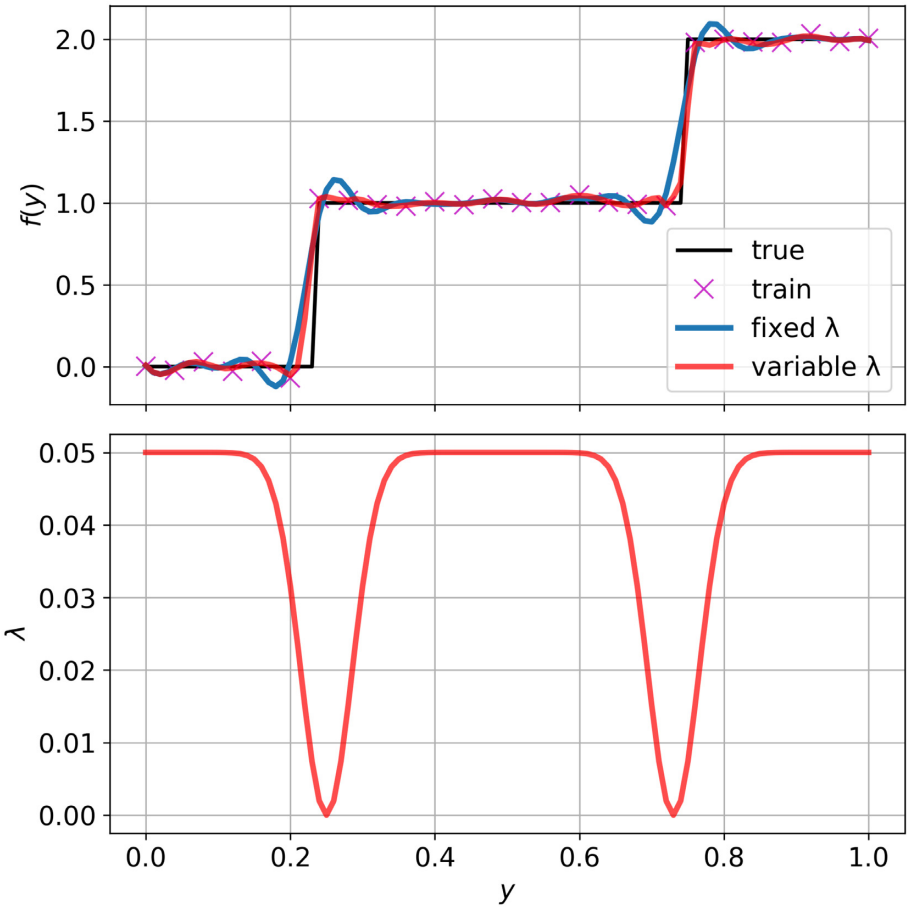

Figure 15. Top: comparison of fixed and variable scale length for modelling a discontinuous function. Bottom: the smoothly varying scale lengths for the variable length scale GP. The fixed length scale GP had λ set to the maximum value in the bottom row. Note how the variable length scale GP oscillates less than the fixed length scale GP.

Burdick, S. & Lekić, V., 2017. Velocity variations and uncertainty from transdimensional *P*-wave tomography of North America, *Geophys. J. Int.,* **209**(2), 1337–1351.

Calvetti, D. & Somersalo, E., 2018. Inverse problems: from regularization to Bayesian inference, *Comput. Stat.,* **10**(3), 1–19.

Chave, A.D. & Cox, C.S., 1982. Controlled electromagnetic sources for measuring electrical conductivity beneath the oceans, 1. Forward problem and model study, *J. geophys. Res.,* **87**(B7), 5327–5338.

Chen, J., Hoversten, G.M., Vasco, D., Rubin, Y. & Hou, Z., 2007. A Bayesian model for gas saturation estimation using marine seismic AVA and CSEM data, *Geophysics,* **72**(2), WA85, doi:10.1190/1.2435082.

Chen, N., Qian, Z., Nabney, I.T. & Meng, X., 2014. Wind power forecasts using gaussian processes and numerical weather prediction, *IEEE Trans. Power Syst.,* **29**(2), 656–665.

Chen, Y., Hoffman, M.W., Colmenarejo, S.G., Denil, M., Lillicrap, T.P., Botvinick, M. & de Freitas, N., 2017. Learning to learn without gradient descent by gradient descent, in *Proc. 34th Int. Conf. Mach. Learn.,* Sydney.

Constable, S., 2010. Ten years of marine CSEM for hydrocarbon exploration, *Geophysics,* **75**(5), 75A67–75A81.

Constable, S.C., 2006. Marine electromagnetic methods - a new tool for offshore exploration, *Leading Edge,* **25,** 438–444.

Constable, S.C., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics,* **52**(3), 289–300.

Cordua, K.S., Hansen, T.M. & Mosegaard, K., 2012. Monte Carlo full-waveform inversion of crosshole GPR data using multiple-point geostatistical a priori information, *Geophysics,* **77**(2), H19–H31.

Cressie, N., 1992. Statistics for spatial data, *Terra Nova,* **4**(5), 613–617.

Criminisi, A., Shotton, J. & Konukoglu, E., 2011. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning, *Tech. rep.,* Microsoft Research.

de Groot-Hedlin, C. & Constable, S., 2004. Inversion of magnetotelluric data for 2D structure with sharp resistivity contrasts, *Geophysics,* **69**(1), 78.

Deisenroth, M.P., Fox, D. & Rasmussen, C.E., 2015. Gaussian processes for data-efficient learning in robotics and control, *IEEE Trans. Pattern Anal. Mach. Intell.,* **37**(2), 408–423.

Dettmer, J. & Dosso, S.E., 2012. Trans-dimensional matched-field geoacoustic inversion with hierarchical error models and interacting Markov chains., *J. acoust. Soc. Am.,* **132**(4), 2239–2250.

Dettmer, J., Dosso, S.E. & Holland, C.W., 2010. Trans-dimensional geoacoustic inversion, *J. acoust. Soc. Am.,* **128**(6), 3393–3405.

Dettmer, J., Molnar, S., Steininger, G., Dosso, S.E. & Cassidy, J.F., 2012. Trans-dimensional inversion of microtremor array dispersion data with hierarchical autoregressive error models, *Geophys. J. Int.,* **188**(2), 719–734.

Dettmer, J., Benavente, R., Cummins, P.R. & Sambridge, M., 2014. Trans-dimensional finite-fault inversion, *Geophys. J. Int.,* **199**(2), 735–751.

Dettmer, J., Dosso, S.E., Bodin, T. & Stipcevic, J., 2015. Direct-seismogram inversion for receiver-side structure with uncertain source-time functions, *Geophys. J. Int.,* **203**(2), 1–4.

Dettmer, J., Hawkins, R., Cummins, P.R., Hossen, J., Sambridge, M., Hino, R. & Inazu, D., 2016. Tsunami source uncertainty estimation: the 2011 Japan tsunami, *J. geophys. Res.,* **121**(6), 4483–4505.

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

Dosso, S.E. & Wilmut, M.J., 2012. Maximum-likelihood and other processors for incoherent and coherent matched-field localization, *J. acoust. Soc. Am.,* **132,** 2273.

Dosso, S.E., Dettmer, J., Steininger, G. & Holland, C.W., 2014. Efficient trans-dimensional Bayesian inversion for geoacoustic profile estimation, *Inverse Problems,* **114018,** doi:10.1088/0266-5611/30/11/114018.

Driscoll, N.W. & Karner, G.D., 1998. Lower crustal extension across the Northern Carnarvon basin, Australia: evidence for an eastward dipping detachment, *J. geophys. Res.,* **103**(B3), 4975–4991.

Earl, D.J. & Deem, M.W., 2005. Parallel tempering: theory, applications, and new perspectives., *Physical Chem. Chem. Phys.,* **7**(23), 3910–3916.

Friedman, J., Hastie, T. & Tibshirani, R., 2001. *The Elements of Statistical Learning,* Vol. **1,** Springer.

Galetti, E. & Curtis, A., 2018. Transdimensional electrical resistivity tomography, *J geophys. Res.,* **123**(8), 6347–6377.

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), 1–5.

Gallagher, K., Bodin, T., Sambridge, M., Weiss, D., Kylander, M. & Large, D., 2011. Inference of abrupt changes in noisy geochemical records using transdimensional changepoint models, *Earth planet. Sci. Lett.,* **311**(1-2), 182–194.

Gao, C. & Lekić, V., 2018. Consequences of parameterization choices in surface wave inversion: insights from transdimensional Bayesian methods, *Geophys. J. Int.,* **215,** 1037–1063.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B., 1995. *Bayesian Data Analysis,* Chapman & Hall.

Geyer, C.J., 1991. Markov chain Monte Carlo maximum likelihood, in *Proc. 23rd Symp. Interface,* p. 156, Am. Stat. Assoc, New York.

Geyer, C.J. & Møller, J., 1994. Simulation procedures and likelihood inference for spatial point processes, *Scand. J. Stat.,* **21,** 359–373.

Gibbs, M., 1997. Bayesian Gaussian processes for regression and classification, *PhD thesis,* University of Cambridge.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y., 2014. Generative adversarial networks, (arXiv:1406.2661).

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika,* **82**(4), 711–732.

Gunning, J., Glinsky, M.E. & Hedditch, J., 2010. Resolution and uncertainty in 1D CSEM inversion: a Bayesian approach and open-source implementation, *Geophysics,* **75**(6), F151–F171.

Hansen, T. & Minsley, B., 2017. Probabilistic inversion of AEM data with an explicit choice of prior information, in *2nd European Airborne Electromagnetics Conference 2017, Held at Near Surface Geoscience Conference and Exhibition 2017,* Malmö, Sweden.

Hastie, D. & Green, P., 2012. Model choice using reversible jump Markov chain Monte Carlo, *Stat. Neerland.,* **66**(3), 309–338.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications, *Biometrika,* **57**(1), 97–109.

Hawkins, R. & Sambridge, M., 2015. Geophysical imaging using trans-dimensional trees, *Geophys. J. Int.,* **203,** 972–1000.

Hawkins, R., Brodie, R.C. & Sambridge, M., 2017. Trans-dimensional Bayesian inversion of airborne electromagnetic data for 2D conductivity profiles, *Explor. geophys.,* **49,** doi: 10.1071/EG16139.

Higdon, D., 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean, *Environ. Ecolog. Stat.,* **5**(2), 173–190.

Hou, Z., Rubin, Y., Hoversten, G.M., Vasco, D. & Chen, J., 2006. Reservoir-parameter identification using minimum relative entropy-based Bayesian inversion of seismic AVA and marine CSEM data, *Geophysics,* **71**(6), O77–O88.

Jeffreys, H., 1939. *Theory of Probability,* Oxford University Press.

Key, K., 2009. 1D inversion of multicomponent, multifrequency marine CSEM data: methodology and synthetic studies for resolving thin resistive layers, *Geophysics,* **74**(2), F9, doi:10.1190/1.3058434.

Key, K. & Ovall, J., 2011. A parallel goal-oriented adaptive finite element method for 2.5-D electromagnetic modelling, *Geophys. J. Int.,* **186**(1), 137–154.

Ko, J. & Fox, D., 2009. GP-BayesFilters: Bayesian filtering using Gaussian process prediction and observation models, *Auton. Robots,* **27**(1), 75–90.

Krige, D.G., 1952. A statistical approach to some basic mine valuation problems on the Witwatersrand, *J. Chem. Metall. Min. Soc. South Afr.,* **52,** No. 6, 201–215.

Laloy, E., Hérault, R., Jacques, D. & Linde, N., 2017. Efficient training-image based geostatistical simulation and inversion using a spatial generative adversarial neural network, arXiv:1708.04975.

Laloy, E., Hérault, R., Jacques, D. & Linde, N., 2018. Training-image based geostatistical inversion using a spatial generative adversarial neural network, *Water Resour. Res.,* **54**(1), 381–406.

Lever, J., Krzywinski, M. & Altman, N., 2016. Points of significance: model selection and overfitting, *Nature Methods,* **13**(9), 703–704.

Lochbühler, T., Vrugt, J.A., Sadegh, M. & Linde, N., 2015. Summary statistics from training images as prior information in probabilistic inversion, *Geophys. J. Int.,* **201**(1), 157–171.

Loseth, L.O., Pedersen, H.M., Ursin, B., Amundsen, L. & Ellingsrud, S., 2006. Low-frequency electromagnetic fields in applied geophysics: waves or diffusion? *Geophysics,* **71**(4), W29–W40.

Luthi, M., Gerig, T., Jud, C. & Vetter, T., 2018. Gaussian process morphable models, *IEEE Trans. Pattern Anal. Mach. Intell.,* **40**(8), 1860–1873.

MacGregor, L. & Sinha, M., 2000. Use of marine controlled-source electromagnetic sounding for sub-basalt exploration, *Geophys. Prospect.,* **48**(6), 1091–1106.

MacKay, D. J.C., 2003. *Information Theory, Inference and Learning Algorithms,* Cambridge University Press.

Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.,* **151**(3), 675–688.

Malinverno, A. & Leaney, S., 2000. A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data, in *SEG Annual Meeting,* no. 3, pp. 2393–2396.

Mallat, S.G., 1989. A theory for multiresolution signal decomposition, *IEEE Trans. Pattern Anal. Mach. Intell.,* **11**(7), 674–693.

Mecklenbrauker, C.F. & Gerstoft, P., 2000. Objective functions for ocean acoustic inversion derived by likelihood methods, *J. Comput. Acoust.,* **8**(2), 259–270.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E., 1953. Equation of state calculations by fast computing machines, *J. Chem. Phys.,* **21**(6), 1087–1092.

Minsley, B.J., 2011. A trans-dimensional Bayesian Markov chain Monte Carlo algorithm for model assessment using frequency-domain electromagnetic data, *Geophys. J. Int.,* **187**(1), 252–272.

Mittet, R. & Gabrielsen, P.T., 2013. Decomposition in upgoing and downgoing fields and inversion of marine CSEM data, *Geophysics,* **78**(1), E1–E17.

Murphy, K.P., 2012. *Machine Learning: A Probabilistic Perspective,* MIT Press.

Myer, D., Constable, S., Key, K., Glinsky, M.E. & Liu, G., 2012. Marine CSEM of the Scarborough gas field, part 1: experimental design and data uncertainty, *Geophysics,* **77**(4), E281–E299.

Myer, D., Constable, S. & Key, K., 2013. Magnetotelluric evidence for layered mafic intrusions beneath the Vøring and Exmouth rifted margins, *Phys. Earth planet. Inter.,* **220,** 1–10.

Myer, D., Key, K. & Constable, S., 2015. Marine CSEM of the Scarborough gas field, part 2: 2D inversion, *Geophysics,* **80**(3), E187–E196.

Neal, R.M., 2011. MCMC using Hamiltonian dynamics, in *Handbook of Markov Chain Monte Carlo,* pp. 113–162, eds Brooks, S., Gelman, A., Jones, G. L. & Meng, X.-L., CRC Press.

Newman, G.A. & Alumbaugh, D.L., 2000. Three-dimensional magnetotelluric inversion using non-linear conjugate gradients, *Geophys. J. Int.,* **140**(2), 410–424.

Paciorek, C.J., 2003. Nonstationary Gaussian Processes for Regression and Spatial Modelling, Ph.D thesis, *Carnegie Mellon University,* **6,** pp. 258

Pasquale, G.D. & Linde, N., 2017. On structure-based priors in Bayesian geophysical inversion, *Geophys. J. Int.,* 208(3), 1342–1358.

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

Piana Agostinetti, N. & Malinverno, A., 2010. Receiver function inversion by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.,* **181,** 858–872.

Piana Agostinetti, N., Giacomuzzi, G. & Malinverno, A., 2015. Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.,* **201**(3), 1598–1617.

Plagemann, C., Kersting, K. & Burgard, W., 2008, pp. 204–219. Nonstationary Gaussian process regression using point estimates of local smoothness, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5212 LNAI, PART 2.

Pyrcz, M.J. & Deutsch, C.V., 2014. *Geostatistical Reservoir Modeling,* Oxford Univ. Press.

Rasmussen, C.E. & Williams, C. K.I., 2006. *Gaussian Processes for Machine Learning,* MIT Press.

Ray, A. & Key, K., 2012. Bayesian inversion of marine CSEM data with a trans-dimensional self parametrizing algorithm, *Geophys. J. Int.,* **191**(3), 1135–1151.

Ray, A., Alumbaugh, D.L., Hoversten, G.M. & Key, K., 2013a. Robust and accelerated Bayesian inversion of marine controlled-source electromagnetic data using parallel tempering, *Geophysics,* **78**(6), E271–E280.

Ray, A., Key, K. & Bodin, T., 2013b. Hierarchical Bayesian inversion of marine CSEM data over the Scarborough gas field - A lesson in correlated noise, in *SEG Technical Program Expanded Abstracts,* no. 1, pp. 723–727, Houston.

Ray, A., Key, K., Bodin, T., Myer, D. & Constable, S., 2014. Bayesian inversion of marine CSEM data from the Scarborough gas field using a transdimensional 2-D parametrization, *Geophys. J. Int.,* **199,** 1847–1860.

Ray, A., Sekar, A., Hoversten, G.M. & Albertin, U., 2016. Frequency domain full waveform elastic inversion of marine seismic data from the Alba field using a Bayesian trans-dimensional algorithm, *Geophys. J. Int.,* **205**(2), 915–937.

Ray, A., Kaplan, S., Washbourne, J. & Albertin, U., 2017. Low frequency full waveform seismic inversion within a tree based Bayesian framework, *Geophys. J. Int.,* **212,** 522–542.

Sambridge, M., 2013. A Parallel Tempering algorithm for probabilistic sampling and multimodal optimization, *Geophys. J. Int.,* **196**(1), 357–374.

Sambridge, M., Gallagher, K., Jackson, A. & Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.,* **167**(2), 528–542.

Sasaki, Y., 2013. 3D inversion of marine CSEM and MT data : an approach to shallow-water problem, *Geophysics,* **78**(1), E59–E65.

Saygin, E. *et al.*, 2016. Imaging architecture of the Jakarta Basin, Indonesia with transdimensional inversion of seismic noise, *Geophys. J. Int.,* **204**(2), 918–931.

Scales, J.A. & Sneider, R., 1997. To Bayes or not to Bayes? *Geophysics,* **62**(4), 1045–1046.

Sen, M.K. & Biswas, R., 2017. Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm, *Geophysics,* **82**(3), R119–R134.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P. & Freitas, N.D., 2016. Taking the human out of the loop: a review of Bayesian optimization, *Proc. IEEE,* **104**(1), doi: 10.1109/JPROC.2015.2494218.

Snoek, J., Larochelle, H. & Adams, R.P., 2012. Practical Bayesian optimization of machine learning algorithms, in *Advances in Neural Information Processing Systems,* pp. 2951–2959, Lake Tahoe, Nevada.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics, *Math. Geol.,* **34**(1), 1–21.

Swendsen, R.H. & Wang, J.S., 1987. Nonuniversal critical dynamics in Monte Carlo simulations, *Phys. Rev. Lett.,* **58**(2), 86–88.

Turner, R.D., 2011. Gaussian processes for state space models and change point detection, *PhD thesis,* University of Cambridge.

Williams, C.K.I. & Rasmussen, C.E., 1996. Gaussian processes for regression, in *Advances in Neural Information Processing Systems,* Vol. **8,** eds. Touretzky D. S., Mozer M. C., Hasselmo M. E., MIT Press.

Yang, Z. & Rodriguez, C.E., 2013. Searching for efficient Markov chain Monte Carlo proposal kernels, *Proc. Natl. Acad. Sci.,* **110**(48), 19307–19312.

Young, P.D. & Cox, C.S., 1981. Electromagnetic active source sounding near the East Pacific Rise, *Geophys. Res. Lett.,* **8,** 1043–1046.

Zhang, X., Curtis, A., Galetti, E. & de Ridder, S., 2018. 3-DMonte Carlo surface wave tomography, *Geophys. J. Int.,* **215**(3), 1644–1658.

APPENDIX A: MCMC MOVES AND THEIR ACCEPTANCE PROBABILITY

We have followed the 'birth-death' McMC method (pseudocode provided in Algorithm 1), where in each step, the length k of the model vector $\mathbf{m}$ either increases by 1 (birth of a GP training point), decreases by 1 (death of a GP training point), or remains the same (values of the GP training point or its spatial location are perturbed). It was pointed out by Galetti & Curtis (2018) that Bayesian natural parsimony is not preserved with improperly tuned birth and death steps when using Gaussian proposals. We have obviated the need for such tuning during birth and death steps by simply proposing from the prior as recommended by Dosso *et al.* (2014) and noted in the work of Zhang *et al.* (2018).

A1 Birth step

During a birth move, $k' = k + 1$ and hence the prior ratio from (11) is

$$\left[\frac{p(\mathbf{m}')}{p(\mathbf{m})}\right]_{\text{birth}} = \frac{1}{\Delta\rho}\frac{k+1}{\prod_{i=1}^{n_d}\Delta x_i}\frac{p(k+1)}{p(k)}, \tag{A1}$$

where the last fraction is unity for a uniform prior on k. For a birth move, we propose a GP training location in the region $\prod_{i=1}^{n_d}\Delta x_i$ uniformly at random, and assign it a value uniformly in $\Delta\rho$, hence the proposal $q(\mathbf{m}'|\mathbf{m})$ can be written as

$$\left[q(\mathbf{m}'|\mathbf{m})\right]_{\text{birth}} = \frac{1}{\prod_{i=1}^{n_d}\Delta x_i}\frac{1}{\Delta\rho}, \tag{A2}$$

whereas the reverse proposal in birth involves deletion of a random point out of $k+1$ points and can be written as

$$\left[q(\mathbf{m}|\mathbf{m}')\right]_{\text{birth}} = \frac{1}{k+1}. \tag{A3}$$

Thus the birth proposal ratio is

$$\left[\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}\right]_{\text{birth}} = \frac{\Delta\rho\prod_{i=1}^{n_d}\Delta x_i}{k+1}. \tag{A4}$$

Thus, from (11), (A1) and (A4)

$$\alpha_{\text{birth}}(\mathbf{m}'|\mathbf{m}) = \min\left[1, \left(\frac{\mathcal{L}(\mathbf{m}')}{\mathcal{L}(\mathbf{m})}\right)^{1/T}\frac{p(k+1)}{p(k)}\right], \tag{A5}$$

where the last fraction is unity for a uniform prior on k.

A2 Death step

In the death move, $k' = k - 1$ and hence the prior ratio from (11) is

$$\left[\frac{p(\mathbf{m}')}{p(\mathbf{m})}\right]_{\text{death}} = \frac{\Delta\rho\prod_{i=1}^{n_d}\Delta x_i}{k}\frac{p(k-1)}{p(k)}, \tag{A6}$$

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

where the last fraction is unity for a uniform prior on k. For a death move, we propose to remove one of k existing training locations.

$$\left[q(\mathbf{m}'|\mathbf{m})\right]_{\text{death}} = \frac{1}{k}, \tag{A7}$$

whereas the reverse proposal in death (i.e. the exact opposite of birth) involves addition of a random point uniformly in the region $\prod_{i=1}^{n_d} \Delta x_i$ and assigning it a value uniformly in $\Delta\rho$, or

$$\left[q(\mathbf{m}|\mathbf{m}')\right]_{\text{death}} = \frac{1}{\prod_{i=1}^{n_d} \Delta x_i} \frac{1}{\Delta\rho}. \tag{A8}$$

Thus the death proposal ratio is

$$\left[\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}\right]_{\text{death}} = \frac{k}{\Delta\rho \prod_{i=1}^{n_d} \Delta x_i}. \tag{A9}$$

Thus, from (11), (A6) and (A9)

$$\alpha_{\text{death}}(\mathbf{m}'|\mathbf{m}) = \min\left[1, \left(\frac{\mathcal{L}(\mathbf{m}')}{\mathcal{L}(\mathbf{m})}\right)^{1/T} \frac{p(k-1)}{p(k)}\right], \tag{A10}$$

where the last fraction is unity for a uniform prior on k.

A3 Fixed k step

When k remains the same, the prior model probabilities do not change. One of the existing k training points is chosen at random and the perturbations for either a new position or a new property value (conductivity) are chosen from symmetric Gaussian proposals with reflection to keep parameters within the prior bounds (see Neal 2011; Yang & Rodriguez 2013; Pasquale & Linde 2017 for details on reflection). The acceptance probability (12) is then simply the ratio of model likelihoods:

$$\alpha_{\text{fixed}}(\mathbf{m}'|\mathbf{m}) = \min\left[1, \left(\frac{\mathcal{L}(\mathbf{m}')}{\mathcal{L}(\mathbf{m})}\right)^{1/T}\right]. \tag{A11}$$

Please note that if one uses a uniform prior over k as we have done in this work, then in all cases, whether birth, death or fixed k,

$$\alpha_{\text{unif}\,k}(\mathbf{m}'|\mathbf{m}) = \min\left[1, \left(\frac{\mathcal{L}(\mathbf{m}')}{\mathcal{L}(\mathbf{m})}\right)^{1/T}\right]. \tag{A12}$$

A4 Parallel tempering step

To facilitate the escape of local misfit minima, or equivalently, the navigation of peaky likelihoods, we use parallel tempering to exchange information between McMC chains running in parallel. One can either exchange models or temperatures at the end of each McMC step using the following Metropolis–Hastings criterion (Swendsen & Wang 1987; Geyer 1991; Earl & Deem 2005; Dettmer *et al.* 2012; Ray *et al.* 2013a; Sambridge 2013):

$$\alpha_{swap}(i, j) = \min\left[1, \left(\frac{\mathcal{L}(\mathbf{m}_j)}{\mathcal{L}(\mathbf{m}_i)}\right)^{1/T_i} \left(\frac{\mathcal{L}(\mathbf{m}_i)}{\mathcal{L}(\mathbf{m}_j)}\right)^{1/T_j}\right]. \tag{A13}$$

For a description of why swapping is effective using (A13) see section 3.2 of Blatter *et al.* (2018).

Our entire algorithm is summarized by the pseudocode in Algorithm 1:

initialize chains with models x_j for temperatures T_j where $j = 1, 2, ..., nTemps$
for $i \leftarrow 1$ **to** $nSteps$ **do**
  for $j \leftarrow 1$ **to** $nTemps$ **do**
    Select $type$ from $\{birth, death, fixed\}$ with probability $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$
    $m \leftarrow x_j[i-1]$
    $m' \sim q(m'|m)_{type}$
    $u \sim U(0,1)$
    if $u < \alpha_j(m'|m)_{type}$ and $p(m') > 0$ **then**
      $x_j[i] \leftarrow m'$
    else
      $x_j[i] \leftarrow m$
    end
  end
  for $j \leftarrow 1$ **to** $nTemps$ **do**
    $u \sim U(0,1)$
    $p \sim U(1, nTemps)$, $p \in I$
    $q \sim U(1, nTemps)$, $q \in I$, $p \neq q$
    if $u < \alpha_{swap}(p, q)$ **then**
      swap T_p and T_q
    else
      no swap
    end
  end
end

Algorithm 1: Pseudocode for McMC with trans-D-GP + parallel tempering. Forward computation to evaluate $\alpha_{swap}(p, q)$ is not required as likelihoods for models in chains p and q have already been computed in the preceding i loop. The traditional requirement of allowing only adjacent chains to swap information has been relaxed, as detailed in Sambridge (2013). Finally, we only swap temperatures T and not the models x, as this makes for efficient and minimal exchange of data in a parallel computing environment. Inference is carried out from the chain (or chains) with $T = 1$ after an initial 'burn-in' number of samples.

APPENDIX B: MAXIMUM LIKELIHOOD DATA ERROR

The model likelihood given in (8) is valid when the data (and residuals) are real. For complex data and a circularly symmetric Gaussian variable with equal variance in the real and imaginary parts, we write for n_f frequencies with n_r receivers at frequency l, the model likelihood as

$$\mathcal{L}(\mathbf{m}) = \prod_{l=1}^{n_f} \frac{1}{\pi^{n_r}|\mathbf{C}_{\mathbf{d}l}|} \exp\left(-[\mathbf{f_l}(\mathbf{m}) - \mathbf{d_l}]^\dagger \mathbf{C}_{\mathbf{d}_l}^{-1} [\mathbf{f_l}(\mathbf{m}) - \mathbf{d_l}]\right), \tag{B1}$$

where the term in the exponential is $\frac{1}{2}$ the χ^2 misfit as the complex data variance at any receiver in covariance $\mathbf{C}_{\mathbf{d}l}$ is twice that of either the real or imaginary parts. We assume uncorrelated data error at all offsets and between frequencies, with noise standard deviation proportional to amplitude as follows. The covariance $\mathbf{C}_{\mathbf{d}l}$ at the lth frequency is given by the diagonal matrix

$$\mathbf{C}_{\mathbf{d}l} = \begin{bmatrix} (\sigma_l|d_{l_1}|)^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & (\sigma_l|d_{l_{n_r}}|)^2 \end{bmatrix}, \tag{B2}$$

$$= \sigma_l^2 \mathbf{C}_l, \tag{B3}$$

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019

where σ_l is the constant of proportionality at all receivers to the signal amplitude at the lth frequency. We can thus write (B1) as

$$\mathcal{L}(\mathbf{m}) = \prod_{l=1}^{n_f} \frac{1}{(\pi\sigma_l^2)^{n_r}|\mathbf{C}_l|} \times \exp\left(-\frac{1}{\sigma_l^2}[\mathbf{f_l}(\mathbf{m}) - \mathbf{d_l}]^\dagger \mathbf{C}_l^{-1}[\mathbf{f_l}(\mathbf{m}) - \mathbf{d_l}]\right). \quad \text{(B4)}$$

To find the maximum of the likelihood (B4), we minimize the negative of the log of the likelihood (i.e. the misfit objective function). First we take log as follows:

$$-\log\mathcal{L}(\mathbf{m}) = \sum_{l=1}^{n_f} \log(\pi^{n_r}|\mathbf{C}_l|) + 2n_r \log\sigma_l + \left(\frac{1}{\sigma_l^2}[\mathbf{f_l}(\mathbf{m}) - \mathbf{d_l}]^\dagger \mathbf{C}_l^{-1}[\mathbf{f_l}(\mathbf{m}) - \mathbf{d_l}]\right), \quad \text{(B5)}$$

$$-\log\mathcal{L}(\mathbf{m}) = \sum_{l=1}^{n_f} \log(\pi^{n_r}|\mathbf{C}_l|) + 2n_r \log\sigma_l + \frac{1}{\sigma_l^2}\mathbf{r_l}^\dagger \mathbf{C}_l^{-1}\mathbf{r_l}. \quad \text{(B6)}$$

Next we derive with respect to σ_l and set equal to zero:

$$\frac{2n_r}{\sigma_l} - \frac{2}{\sigma_l^3}\mathbf{r_l}^\dagger \mathbf{C}_l^{-1}\mathbf{r_l} = 0, \quad \text{(B7)}$$

$$\Rightarrow \sigma_l^2 = \frac{1}{n_r}\mathbf{r_l}^\dagger \mathbf{C}_l^{-1}\mathbf{r_l}. \quad \text{(B8)}$$

At this point, we ask that readers note the similarity of (B8) with equation B.5 of Sambridge (2013) who follows a similar approach in the time domain, while we are operating in frequency. Substituting (B8) in (B6) we get

$$-\log\mathcal{L}(\mathbf{m}) = \sum_{l=1}^{n_f} n_r \log\left[\frac{1}{n_r}\mathbf{r_l}^\dagger \mathbf{C}_l^{-1}\mathbf{r_l}\right] + \text{constants not depending on } \mathbf{m}. \quad \text{(B9)}$$

$$-\log\mathcal{L}(\mathbf{m}) = \sum_{l=1}^{n_f} n_r \log\left[\mathbf{r_l}^\dagger \mathbf{C}_l^{-1}\mathbf{r_l}\right] + \text{constants not depending on } \mathbf{m}. \quad \text{(B10)}$$

While sampling the posterior models in the McMC chain, we use the negative log likelihood given by (B10), instead of computing the misfit with unreliable, fixed, data error. Note that using this methodology, the data errors at each frequency are implicitly sampled as a function of the current McMC sample $\mathbf{m}$.

Downloaded from https://academic.oup.com/gji/article-abstract/217/3/1706/5366736 by guest on 01 April 2019