Geophysical Journal International

Geophys. J. Int. (2021) **226,** 302–326
Advance Access publication 2021 March 26
GJI General Geophysical Methods

doi: 10.1093/gji/ggab114

Bayesian inversion using nested trans-dimensional Gaussian processes

Anandaroop Ray

Geoscience Australia, Symonston, Australian Capital Territory, GPO Box 378 *Canberra ACT* 2601*, Australia. E-mail: a2ray@ucsd.edu*

Accepted 2021 March 22. Received 2021 March 18; in original form 2020 September 25

SUMMARY

To understand earth processes, geoscientists infer subsurface earth properties such as electromagnetic resistivity or seismic velocity from surface observations of electromagnetic or seismic data. These properties are used to populate an earth model vector, and the spatial variation of properties across this vector sheds light on the underlying earth structure or physical phenomenon of interest, from groundwater aquifers to plate tectonics. However, to infer these properties the spatial characteristics of these properties need to be known in advance. Typically, assumptions are made about the length scales of earth properties, which are encoded *a priori* in a Bayesian probabilistic setting. In an optimization setting, appeals are made to promote model simplicity together with constraints which keep models close to a preferred model. All of these approaches are valid, though they can lead to unintended features in the resulting inferred geophysical models owing to inappropriate prior assumptions, constraints or even the nature of the solution basis functions. In this work it will be shown that in order to make accurate inferences about earth properties, inferences can first be made about the underlying length scales of these properties in a very general solution basis. From a mathematical point of view, these spatial characteristics of earth properties can be conveniently thought of as 'properties' of the earth properties. Thus, the same machinery used to infer earth properties can be used to infer their length scales. This can be thought of as an 'infer to infer' paradigm analogous to the 'learning to learn' paradigm which is now commonplace in the machine learning literature. However, it must be noted that (geophysical) inference is not the same as (machine) learning, though there are many common elements which allow for cross-pollination of useful ideas from one field to the other, as is shown here. A non-stationary trans-dimensional Gaussian Process (TDGP) is used to parametrize earth properties, and a multichannel stationary TDGP is used to parametrize the length scales associated with the earth property in question. Using non-stationary kernels, that is kernels with spatially variable length scales, models with sharp discontinuities can be represented within this framework. As GPs are multidimensional interpolators, the same theory *and* computer code can be used to solve geophysical problems in 1-D, 2-D and 3-D. This is demonstrated through a combination of 1-D and 2-D non-linear regression examples and a controlled source electromagnetic field example. The key difference between this and previous work using TDGP is generalized nested inference and the marginalization of prior length scales for better posterior subsurface property characterization.

Key words: Inverse theory; Probability distributions.

1 INTRODUCTION

An aspect of Bayesian inference which is simultaneously exalted and much debated (e.g. Scales & Sneider 1997; Backus 1988, respectively) is the the fact that posterior inference through prior specification is necessarily transparent, yet subjective. In other words, posterior inference is dependent on prior notions about the problem at hand. However, with a suitable mathematical treatment, deterministic and Bayesian approaches can be reconciled as is elegantly shown by multiple authors such as Calvetti & Somersalo (2018) and Malinverno & Parker (2006).

Bayes' theorem bridges posterior and prior knowledge about the earth through the acquired geophysical data (Tarantola & Valette 1982). Crucially, this specification of prior knowledge and its parametrization is often overlooked in both Bayesian as well as optimization contexts. Using an informative Bayesian prior or optimization regularization operator which appropriately reflects the earth's spatial character is key

to making meaningful inferences about the earth's subsurface structure (e.g. Valentine & Sambridge 2020a). With choices of solution basis based on the physics of the problem, well-designed inversion and inference algorithms can infer the level of detail with which we can resolve the earth (e.g. Hawkins & Sambridge 2015; Muir & Tsai 2020). Critical for distinguishing the appropriate level of detail provided by two distinctly parametrized families of posterior models which fit the data, is the ability to compute and compare their *marginal likelihood* (Kass & Raftery 1995) or *evidence* for each parametrization. However, this is a difficult task, given that this quantity requires sampling the model likelihood according to the prior (as opposed to the posterior) distribution. One method to avoid doing this yet compare different levels of complexity within a model basis, is to use reversible jump Markov chain Monte Carlo (RJ-McMC, Green 1995; Green & Hastie 2009). This technique was popularized for geophysical inference as the trans-dimensional or Trans-D method (see Malinverno & Leaney 2000; Malinverno 2002; Sambridge *et al.* 2006). Another novel technique for performing model selection is cross-validation using the methods proposed by Vehtari *et al.* (2017). A recent geophysical example using cross-validation can be found in Muir & Tkalčić (2020).

Gaussian processes constitute a highly researched interpolation, regression and inference method in the machine learning (ML) literature, with well understood qualities of spatial variability (see Rasmussen & Williams 2006). As the mathematics of GP interpolation are spatial dimension-agnostic, Ray & Myer (2019) proposed using Trans-D methods in conjunction with GPs. This approach has three advantages. First, the same theory *and* computer code can be used for 1-D, 2-D and 3-D geophysical inference, as opposed to piecewise constant functions in a 1-D earth and Voronoi cells for 1-D, 2-D or 3-D earth models. Secondly, piecewise constant and Voronoi tessellations are effectively able to capture sharp discontinuities, but not smooth transitions in subsurface earth properties—though an ensemble of Voronoi partitions can certainly capture uncertainty in the location of a partition (e.g. Bodin & Sambridge 2009; Ray *et al.* 2014). Using TDGP, smooth changes with a specified length scale can be used to interpolate models to the shape of subsurface property variations, as demanded by surface geophysical data. Finally, the number and position of the GP nuclei can be determined in a Trans-D fashion using the existing mathematical machinery of RJ-McMC. This leads to a theoretically rich (e.g. Roininen *et al.* 2019; Valentine & Sambridge 2020b) yet practical formulation for general Bayesian geophysical inversion.

While using TDGP for 2-D magnetotelluric (MT) inversion (Blatter 2020, Blatter *et al*. in review), it was found that it was necessary to use a warping of the input space (e.g. Sampson & Guttorp 1992; MacKay 1998), thus allowing for a change in the length scale of posterior earth properties (resistivity) with depth. Further, a presumption of smoothness for non-linear problems is particularly problematic as the posterior model may indeed be smooth in areas of low sensitivity, while demanding sharp changes in other parts of the model space. Examples illustrating the challenges in, and utility of, inferring local sharp changes in globally smooth models have been shown in a geophysical context in chapter 7 of Hawkins (2017). Similar themes, but for statistical inference and prediction using GPs were investigated by Paciorek & Schervish (2004). In this work, the formulation of Ray & Myer (2019) has been extended to allow sharp transitions and varying length scales by using non-stationary GP kernels as laid out in Paciorek & Schervish (2004). In particular, a stationary TDGP 'length scale model' is used to parametrize a non-stationary TDGP 'properties model'. This allows for the representation of a much larger set of earth models and makes Trans-D adaptation to subsurface structure powerfully general. Remarkably, though another layer of complexity is added to the parametrization (a second TDGP model)—Bayesian parsimony naturally ensures that the ensuing posterior inference is not overly complicated (MacKay 2003, chapter 28). Purely statistical regression that operates in a similar, sparse fashion is described in Snelson & Ghahramani (2005). In one sense, this is analogous to overcomplete basis representations for enhanced coding efficiency and signal characterization as demonstrated by Lewicki & Sejnowski (2000). The idea of using GPs to recursively parametrize GPs has received much recent attention in the machine learning and statistics literature (e.g. Lindgren *et al.* 2011; Dunlop *et al.* 2018; Roininen *et al.* 2019; Emzir *et al.* 2020), but this work is among the first implementations in the earth sciences which will surely follow. The line of research pursued in this work is complementary to other promising parametrizations using neural networks in earth science (e.g. Laloy *et al.* 2017). To summarize, the difference between using TDGP in purely stationary mode versus using TDGP as proposed in this work is captured in Fig. 1. The bottom row shows the mean reconstruction using McMC via TDGP in its purely stationary (left-hand side) and non-stationary (right-hand side) forms as is detailed in later sections. A commonly used metric to define image reconstruction quality is PSNR (peak signal-to-noise ratio), defined in Appendix C. Higher values in dB are indicative of better image restoration, and indicates that the non-stationary reconstruction is superior. The Trans-D machinery used in this example can also be used for geophysical inverse problems to construct 1-D, 2-D or 3-D property fields such as conductivity or velocity within the earth.

2 THEORY

Details of the purely stationary mode, fixed length scale TDGP method including an introduction to GPs are given fully in Ray & Myer (2019). In this section instead, a brief description of GPs is given. Particular attention is paid to their mathematical formalism, especially the extension to the non-stationary aspects of the new TDGP sampling formulation.

2.1 Gaussian processes

As described in Rasmussen & Williams (2006), a GP is a stochastic process that is completely determined by its mean and covariance. Gaussian processes are a method of non-parametric regression that do not require a fixed discretization, providing both a prediction and uncertainty around the prediction (see Fairbrother *et al.* 2018,for a modern implementation). GPs have been successfully used in many fields

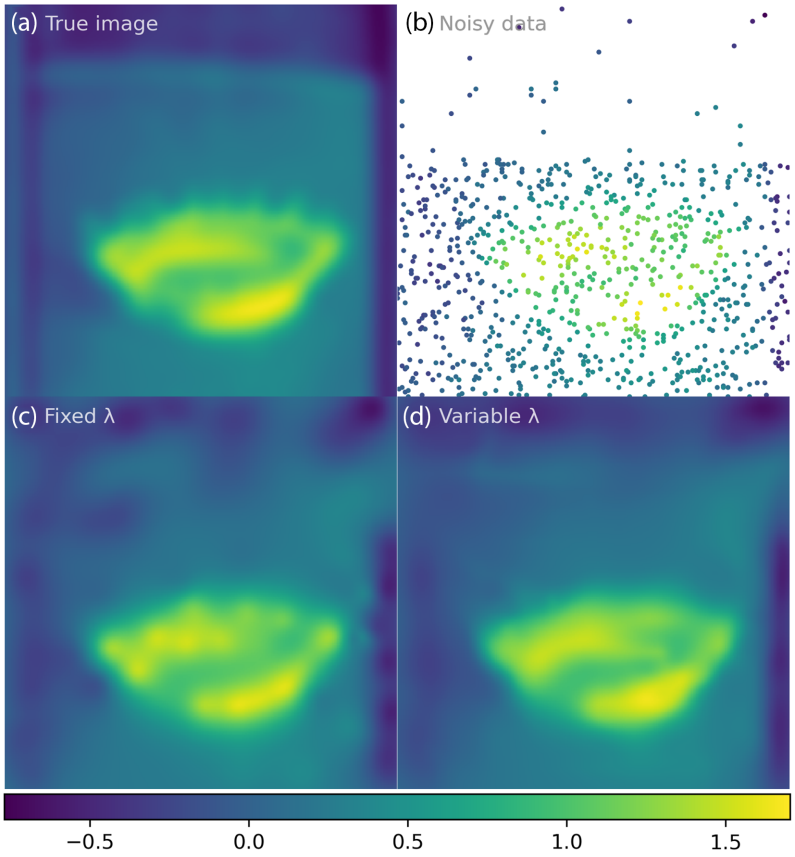

Figure 1. A 2-D non-linear regression problem. (a) The true image to be recovered with a parsimonious representation. (b) The provided data were 851 out of 65 536 noisy pixels. As detailed in later sections, on average *only* 70–80 GP nuclei in total were required to reconstruct the images (c, d) in the bottom row. Clearly, the mean reconstruction with a kernel of variable length scale λ in (d) provides a better approximation to the true image. However, the fixed length scale mean reconstruction in (c) captures the main features of the true image as well. Both reconstructions fit the provided 851 data to within noise. Using PSNR (peak signal-to-noise-ratio) as an objective measure of the reconstruction, the variable length reconstruction (d) with PSNR = 23.69 dB outperforms the stationary length scale reconstruction (c) with PSNR = 21.70 dB.

from spatial statistics (Cressie 1992), statistics (Williams & Rasmussen 1996), robotics (Ko & Fox 2009), weather prediction (Chen *et al.* 2014), reinforcement learning (Deisenroth *et al.* 2015), automated image analysis (Luthi *et al.* 2018) to classification (Galy-Fajou *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). Until recently GPs have largely been used in the geosciences within reservoir modelling or mining as an interpolation tool and their potential as a mathematical framework for geophysical inversion has only just been investigated (e.g. Valentine & Sambridge 2020b). Apart from the fact that GP theory is well understood, GPs have a close connection with neural networks and other modern ML methods (e.g. Damianou & Lawrence 2013; Jiaxuan *et al.* 2017). Neal (1996) proved that Bayesian regression models based on a one layer neural network converge to a GP in the limit of an infinite number of neurons. Given the recent progress in the field of ML, further investigations exploiting the synergies between inference in geophysical problems and ML are eagerly awaited.

2.2 Formalism for a stationary GP

Since GPs are defined by a mean and a covariance, a single GP necessarily cannot express multimodality. However, the ensemble of means of Gaussian processes can indeed represent multimodality. This is where Ray & Myer (2019) depart from the realm of conventional geostatistics or traditional GP regression and allow for non-linear uncertainty estimation using a Trans-D approach. To elaborate, a single GP mean $\boldsymbol{\mu}_* \in \mathbb{R}^{n_{\text{test}}}$ in the forward modelling domain is represented by the following equation (Murphy 2012):

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

where $\mathbf{m} \in \mathbb{R}^{n_{\text{train}}}$ contains the n_{train} Trans-D property values (e.g. resistivity) defining an earth model in n_D spatial dimensions. These points are located at $\mathbf{x}_i \in \mathbb{R}^{n_D}$, $i = 1, \ldots, n_{\text{train}}$. To be explicit, each $\mathbf{x}_i$ contains the spatial co-ordinates of the points in the Trans-D property vector $\mathbf{m}$. The n_{test} elements of the mean $\boldsymbol{\mu}_*$ are similarly located at n_{test} vectors $\mathbf{x}_{*i} \in \mathbb{R}^{n_D}$, $i = 1, \ldots, n_{\text{test}}$. To define the matrices $\mathbf{K}_m \in \mathbb{R}^{n_{\text{train}} \times n_{\text{train}}}$ and $\mathbf{K}_* \in \mathbb{R}^{n_{\text{test}} \times n_{\text{train}}}$ a correlation function or kernel R and a quadratic distance Q_{ij} are defined as follows. R can take various forms, popular

choices being the squared Euclidean and Matern kernels (Rasmussen & Williams 2006):

$$R(\xi) = \begin{cases} \exp(-\xi^2/2) & \text{if kernel is Squared Euclidean,} \\ (1+\sqrt{5}\xi+5\xi^2/3)\exp(-\sqrt{5}\xi) & \text{if kernel is Matern 5/2,} \\ (1+\sqrt{3}\xi)\exp(-\sqrt{3}\xi), & \text{if kernel is Matern 3/2.} \end{cases} \tag{2}$$

In order of smoothness, squared Euclidean kernels have the slowest spatial decay, and Matern 3/2 the most rapid. Following Paciorek & Schervish (2004), the quadratic form Q_{ij} is defined as shown underneath:

$$Q_{ij}(\mathbf{y}_i, \mathbf{y}_j) = \left[\mathbf{y}_i - \mathbf{y}_j\right]^t \mathbf{C}_{\text{avg}}^{-1} \left[\mathbf{y}_i - \mathbf{y}_j\right], \text{ where } \mathbf{y} \in \mathbb{R}^{n_D}, \tag{3}$$

and $\mathbf{C}_{\text{avg}}$ is an $n_D \times n_D$ symmetric positive definite matrix with the squared length scales for a stationary GP along the diagonal. For a stationary GP, a correlation function can then be written as follows:

$$k(\mathbf{y}_i, \mathbf{y}_j) = R(\sqrt{Q_{ij}}), \tag{4}$$

where $k(\mathbf{y}_i, \mathbf{y}_j)$ populate the elements of matrices $\mathbf{K}_m$ and $\mathbf{K}_*$ in the following manner. By first calculating (4) pairwise among the vectors $\mathbf{x}_i \in \mathbb{R}^{n_D}$, $i = 1, \ldots, n_{\text{train}}$ to obtain a matrix $\mathbf{K} \in \mathbb{R}^{n_{\text{train}} \times n_{\text{train}}}$ and adding a 'nugget' or diagonal damping term $\boldsymbol{\sigma}_m^2$ (see Pyrcz & Deutsch 2014,for details) the following is arrived at:

$$\mathbf{K}_m = \mathbf{K} + \boldsymbol{\sigma}_m^2. \tag{5}$$

In a similar fashion, by using $\mathbf{x}_{*i} \in \mathbb{R}^{n_D}$, $i = 1, \ldots, n_{\text{test}}$ and $\mathbf{x}_j \in \mathbb{R}^{n_D}, j = 1, \ldots, n_{\text{train}}$ pairwise in (4) the matrix $\mathbf{K}_* \in \mathbb{R}^{n_{\text{test}} \times n_{\text{train}}}$ is obtained. At this point, all that is necessary to compute a GP mean, given a set of training points and a stationary length scale (in each dimension) has been detailed. As shown in Ray & Myer (2019), a Trans-D prior over $\mathbf{x}_i$ and $\mathbf{m}$ is required to compute a $\boldsymbol{\mu}_*$ which can be used in purely stationary mode, both for regression or feeding into a forward modelling physics engine. Given natural Bayesian parsimony (Malinverno 2002; MacKay 2003), usually the required $n_{\text{train}} < < n_{\text{test}}$ for solving the problem at hand.

Finally, it must be mentioned that multiple channels of output (e.g. resistivity anisotropy in addition to resistivity) can be independently handled by the same correlation function and the GP mean formulation in (1) by simply appending columns to $\mathbf{m}$ (Rasmussen & Williams 2006). Consequently, $\boldsymbol{\mu}_*$ will have as many columns as there are columns of input in $\mathbf{m}$.

2.3 Extension of formalism for non-stationarity

The advantage of using the approach of Paciorek & Schervish (2004) is that the methodology described in Section 2.2 and eqs (1) through (5) can be used to compute the GP mean with a variable length scale, with only two modifications. In eq. (3), $\mathbf{C}_{\text{avg}}$ needs to be computed as the mean of the length scale covariances $\mathbf{C}_i$ and $\mathbf{C}_j$ at spatial locations $\mathbf{y}_i$ and $\mathbf{y}_j$:

$$\mathbf{C}_{\text{avg}} = \frac{\mathbf{C}_i + \mathbf{C}_j}{2}. \tag{6}$$

Finally, eq. (4) is modified as follows:

$$k(\mathbf{y}_i, \mathbf{y}_j) = |\mathbf{C}_i|^{\frac{1}{4}} |\mathbf{C}_j|^{\frac{1}{4}} |\mathbf{C}_{\text{avg}}|^{-\frac{1}{2}} R(\sqrt{Q_{ij}}), \tag{7}$$

which falls back to the stationary form if $\mathbf{C}_i = \mathbf{C}_j$. For non-stationary mode TDGP, the departure is made from purely stationary mode TDGP at this juncture.

2.4 Bayesian trans-D inversion

Trans-D McMC is well suited for sampling earth models $\boldsymbol{\theta}_{\boldsymbol{m}}$ of variable (parameter) dimension k. Trans-D inversion (Malinverno 2002; Sambridge *et al.* 2006) is based on birth/death Monte Carlo (Geyer & Møller 1994) and the more general RJ-McMC method (Green 1995). However, as discussed in detail in Section 2.2 of Ray & Myer (2019) most implementations have required the use of different parametrizations, for 1-D, 2-D and 3-D models (e.g. Bodin & Sambridge 2009; Agostinetti & Malinverno 2010; Minsley 2011; Brodie & Sambridge 2012; Ley-Cooper 2016; Burdick & Lekić 2017; Zhang *et al.* 2018; Galetti & Curtis 2018; Hawkins *et al.* 2019, to list but a few). The applications have ranged from electromagnetic sounding to seismic imaging from shallow sediment to mantle-deep length scales. The work of Hawkins & Sambridge (2015) with Trans-D wavelet trees can be effectively utilized for computationally intensive geophysical posterior inference and is spatial dimension agnostic (e.g. Dettmer *et al.* 2016; Hawkins *et al.* 2017; Ray *et al.* 2018). However, wavelet tree formulations require prior specification in the wavelet transform domain, as opposed to the familiar space domain. It is somewhat unclear what appropriate prior bounds are, or how to specify them in the wavelet domain. Further, all the above listed references provide geophysical model realizations which are either intrinsically smooth or sharp, depending on the underlying model representation basis. In chapter 7 of Hawkins (2017) some promising new Trans-D formulations are developed which can provide both smoothness and sharp changes in the same 1-D model.

The stationary TDGP formulation introduced in Ray & Myer (2019) is flexible enough to approximate discontinuities reasonably, incorporate prior knowledge of length scales within the earth, while using the same theory and code for 1-D, 2-D and 3-D inverse problems or

regression. The computational cost in TDGP is dominated by the dimension of the GP mean vector $\boldsymbol{\mu}_*$, not its spatial dimension (timing and computational aspects are discussed in Appendix A4). However, as shown in Fig. 1 there are advantages in allowing the inference process to infer the length scales of the quantity to be inferred, and discontinuities are better approximated with variable length scales as will be shown shortly. Keeping these facts in mind, in concert with with the theoretical developments around the use of GPs for Bayesian inversion (see Roininen *et al.* 2019; Valentine & Sambridge 2020a,b), the development of non-stationary TDGP is motivated and is discussed underneath.

2.5 Bayes' theorem

For observed data $\mathbf{d}$ and TDGP earth models $\boldsymbol{\theta_m}$ it can be written:

$$p(\boldsymbol{\theta_m}|\mathbf{d}) \propto p(\mathbf{d}|\boldsymbol{\theta_m})p(\boldsymbol{\theta_m}), \tag{8}$$

Reading from right to left, $p(\boldsymbol{\theta_m})$ is the prior probability of $\boldsymbol{\theta_m}$, which is known independent of the observations $\mathbf{d}$. The prior importance of $\boldsymbol{\theta_m}$ is re-assessed by carrying out a geophysical experiment which shows how likely it is that $\boldsymbol{\theta_m}$ fits the observations. The probability of fit is provided by the likelihood function $p(\mathbf{d}|\boldsymbol{\theta_m})$. The result of re-weighting or updating the prior importance of $\boldsymbol{\theta_m}$ by the likelihood of $\boldsymbol{\theta_m}$ provides the *posterior* probability of observing the model $\boldsymbol{\theta_m}$. The posterior probability is represented by the term $p(\boldsymbol{\theta_m}|\mathbf{d})$.

2.5.1 Likelihood function

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

$$\mathcal{L}(\boldsymbol{\theta_m}) = p(\mathbf{d}|\boldsymbol{\theta_m}) = \frac{1}{\sqrt{|2\pi\mathbf{C}_d|}} \exp\left(-\frac{1}{2}\Big[\mathbf{f}(\boldsymbol{\theta_m}) - \mathbf{d}\Big]^t \mathbf{C}_d^{-1}\Big[\mathbf{f}(\boldsymbol{\theta_m}) - \mathbf{d}\Big]\right), \tag{9}$$

where $[\mathbf{f}(\boldsymbol{\theta_m}) - \mathbf{d}]$ is the residual vector of misfit between the forward model calculation and the data for the model $\boldsymbol{\theta_m}$. The covariance matrix of data errors is given by $\mathbf{C}_d$. A Gaussian likelihood is generally encouraged by the ubiquitous application of stacking to geophysical data for noise attenuation. Stacking implies Central Limiting for the resulting noise estimates on the mean data, and the implication of Gaussianity. However, care must be taken to remove outlying data to justify this assumption. The frequency domain Gaussian likelihood for complex data is detailed in Appendix B. Non-Gaussian likelihoods can also be accommodated within a Bayesian framework.

2.5.2 Prior specification

If $\boldsymbol{\theta}$ represents a k parameter stationary *or* non-stationary Trans-D model, then a k parameter prior model probability can be written as

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

where $\mathbf{m}_k$ is a vector of GP 'training' property values (e.g. resistivity) *or* length scales of this property. Index k can now be identified with n_{train} for a stationary *or* non-stationary TDGP model. $\mathbf{x}_k$ contains k vectors in $\mathbb{R}^{n_D \times k}$ that specifies the locations of $\mathbf{m}_k$. n_D is the number of spatial dimensions of the model (e.g. $n_D = 2$ for 2-D). Using the chain rule of probabilities, the following expansion is arrived at:

$$p(\mathbf{m}_k, \mathbf{x}_k, k) = p(\mathbf{m}_k|\mathbf{x}_k, k)p(\mathbf{x}_k|k)p(k). \tag{11}$$

If it is assumed that each of k training values (for property *or* length scale) can be independently and uniformly sampled within a range δ, and that they can be arranged in any of $k!$ ways uniformly within a length, area or volume given by $\prod_{i=1}^{n_D} \Delta x_i$, the above equation can be rewritten as:

$$p(\mathbf{m}_k, \mathbf{x}_k, k) = \frac{1}{\delta^k} \frac{k!}{(\prod_{i=1}^{n_D} \Delta x_i)^k} p(k). \tag{12}$$

Common choices for $p(k)$, the prior probability on the number of GP nuclei 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.

At this juncture, it is useful to point out again that the prior formulation in eqs (10) through (12) are valid for *both* stationary as well as non-stationary TDGP models, and they will in general have different prior bounds on their Trans-D count k. For clarity of exposition, it helps to separate $\boldsymbol{\theta}$ into stationary and non-stationary components $\boldsymbol{\theta}_{\text{s}}$ and $\boldsymbol{\theta}_{\text{ns}}$ as follows:

$$\boldsymbol{\theta_m} = [\boldsymbol{\theta}_{\text{s}}, \boldsymbol{\theta}_{\text{ns}}], \tag{13}$$

where $\boldsymbol{\theta_m}$ represents the full stationary *and* non-stationary McMC model parametrizations (Fig. 2). The stationary TDGP model $\boldsymbol{\theta}_{\text{s}}$ when plugged into (1) with a fixed *a priori* length scale $\boldsymbol{\lambda}_{\text{s}} \in \mathbb{R}^{n_D}$ and fixed *a priori* scalar nugget σ_{s}, provides an earth property (e.g. velocity or resistivity) length scale $\boldsymbol{\mu}_{*\text{s}} \in \mathbb{R}^{n_{\text{test}}}$ which varies in n_D spatial dimensions. This length scale informs $\mathbf{C}_i, i = 1, \ldots, n_{\text{test}}$ for the calculations of the non-stationary kernel given by (7). The calculation of these non-stationary kernels is performed using the locations of the property values (e.g. velocity or resistivity) in the non-stationary TDGP model $\boldsymbol{\theta}_{\text{ns}}$. Together with the property values within $\boldsymbol{\theta}_{\text{ns}}$ and a separate fixed *a priori*

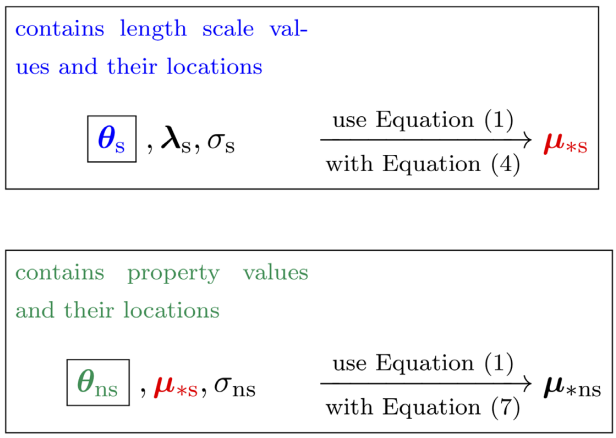

Figure 2. The relationship between $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_{ns}$, respectively the stationary and non-stationary McMC models in non-stationary mode TDGP is illustrated here. The stationary TDGP McMC process is shown in the top box and the non-stationary TDGP McMC process is enclosed in the bottom box. $\boldsymbol{\lambda}_s$ is the fixed, stationary length scale underlying $\boldsymbol{\mu}_{*s}$, the GP mean representing the length scale for the geophysical property. σ_s is a fixed, scalar additive noise (akin to damping) in the stationary TDGP McMC process. σ_{ns} plays a similar role for the non-stationary TDGP McMC process. The two McMC models $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_{ns}$ are *a priori* independent of each other, though the final non stationary GP mean $\boldsymbol{\mu}_{*ns}$ representing geophysical property depends on the state of both $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_{ns}$. It is $\boldsymbol{\mu}_{*ns}$ which is fed into a forward modelling engine or used for regression. Both $\boldsymbol{\mu}_{*ns}$ and $\boldsymbol{\mu}_{*s}$ have the same unchanging vector dimension n_{test}.

scalar nugget σ_{ns}, the calculation of a geophysical properties vector $\boldsymbol{\mu}_{*ns}$ is again performed through the use of (1). $\boldsymbol{\mu}_{*ns} \in \mathbb{R}^{n_{test}}$ also varies in n_D spatial dimensions. Further, $\boldsymbol{\mu}_{*ns}$ is dependent on $\boldsymbol{\mu}_{*s}$. Changing the state of $\boldsymbol{\theta}_s$ changes $\boldsymbol{\mu}_{*s}$ and therefore $\boldsymbol{\mu}_{*ns}$. However, changing the state of $\boldsymbol{\theta}_s$ does not change the state of $\boldsymbol{\theta}_{ns}$ and vice versa. The two McMC parametrizations $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_{ns}$ are therefore *a priori* independent. This insight allows for the use of two independent Trans-D processes to be used within the same Markov chain (e.g. Dettmer & Dosso 2013; Blatter *et al.* 2019), with their own birth, death, position change and update values moves which are commonly used in Trans-D methods (e.g. Bodin & Sambridge 2009; Ray & Myer 2019). For purely stationary mode TDGP (Ray & Myer 2019), only the top box in Fig. 2 was used, and $\boldsymbol{\theta}_s$ was used to parametrize geophysical properties. It must be mentioned that $\boldsymbol{\mu}_*$ can take on values outside defined prior bounds on $\mathbf{m}_k$ as the mean of Gaussian realizations will have support outside this bounded interval (see Ray & Myer 2019, for implications of this support through examples of prior sampling).

It must be highlighted here that all prior specifications in (12) are made in the familiar domains of geophysical property or its spatial length scales as a function of 1-D, 2-D or 3-D Euclidean space, irrespective of the Euclidean spatial dimension n_D of the earth model. Simple transformations of this Euclidean space such as a geometric progression can also be used without modifying the theory above (e.g. Blatter 2020, Blatter *et al*. 2020, in review). To ensure that no negative length scales are computed in $\boldsymbol{\mu}_{*s}$, prior specification is carried out in the $\log_{10}$ domain of the length scale within $\boldsymbol{\theta}_s$. This is similar to specifying $\boldsymbol{\theta}_{ns}$ in the $\log_{10}$ domain of resistivity for an electromagnetic inverse problem. All hierarchical representations end at some level, and it is ended in this implementation at the underlying stationary TDGP model $\boldsymbol{\theta}_s, \boldsymbol{\lambda}_s, \sigma_s$. In theory more hierarchical layers can be added *ad infinitum* without an obvious benefit from this further complexity (see Duvenaud *et al.* 2014, for a discussion).

Finally, if multiple channels are required (e.g. property anisotropy or as many length scales as there are spatial dimensions) then δ in (12) is modified to be the product of the uniform ranges of the multiple quantities (e.g. see eq. B14 of Ray *et al.* 2016).

2.5.3 Posterior sampling

The process of finding the posterior probability $p(\boldsymbol{\theta}_m|\mathbf{d})$ for various models $\boldsymbol{\theta}_m$ admissible by the prior is repeated until an ensemble of models representative of the probability density function or PDF $p(\boldsymbol{\theta}_m|\mathbf{d})$ is obtained. For the Trans-D method sampling this is done 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 $\boldsymbol{\theta}_m$ to $\boldsymbol{\theta}'_m$ in the McMC chain:

$$\alpha(\mathbf{m}'|\mathbf{m}) = \min\left[1, \frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta})}\left\{\frac{p(\mathbf{d}|\boldsymbol{\theta}')}{p(\mathbf{d}|\boldsymbol{\theta})}\right\}^{1/T}\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}|\mathbf{J}|\right]. \tag{14}$$

For either of the stationary or non-stationary McMC processes, the McMC model $\boldsymbol{\theta}$ is perturbed to $\boldsymbol{\theta}'$ via a proposal PDF $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$. 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 $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$. 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. The 'birth-death' algorithm has also been used in this work. As the mechanics of the McMC birth/death proposals are the same as in Bodin & Sambridge (2009), they

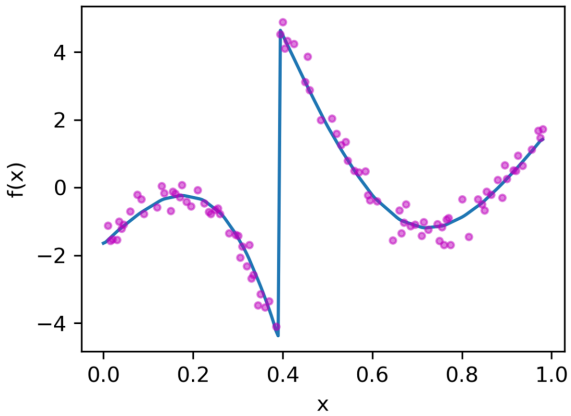

Figure 3. A digitization of the Jump1D test function (Plagemann *et al.* 2008) which originally appeared in Paciorek & Schervish (2004). The function displays different amounts of curvature on opposite sides of a sharp jump near $x = 0.4$. Half of the points in x were randomly selected as training data locations, and Normal noise with a standard deviation of 0.275 was added to the digitized values at these points (magenta dots). These values were presented to TDGP to try and recover the shape of the Jump1D function.

satisfy the conditions for unity Jacobian determinant as laid out in detail in Dosso *et al.* (2014). 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). Sampling using interacting chains is often referred to as a parallel tempering (PT) algorithm (for a thorough review see Earl & Deem 2005).

Since the stationary and non-stationary TDGP McMC models are *a priori* independent of each other but use the same prior formulation, they are sampled using the same, standard Trans-D McMC moves as defined in Ray & Myer (2019). The moves are laid out in the Appendix A for completeness. The only difference with the addition of the non-stationary process is that each stationary TDGP McMC step is followed by a non-stationary TDGP McMC step. This procedure is the same as in Dettmer & Dosso (2013) and Blatter *et al.* (2019), where independent Trans-D models have been used to parametrize different spatial locations within an earth properties model, but both Trans-D processes see the same data when evaluating likelihoods. Approximations required for the efficient computation of kernels are also laid out in Appendix A4. Detailed expressions for the acceptance probabilities α for all McMC moves are given in Appendix A. Pseudo-code for non-stationary mode TDGP McMC sampling with interacting chains is provided in Algorithm 1.

3 APPLICATIONS OF TDGP AND COMPARISON OF STATIONARY AND NON-STATIONARY INFERENCE

A 1-D non-linear regression example, followed by details of the methodology behind the 2-D non-linear regression example summarized in Fig. 1 will now be discussed. The machinery behind these examples is statistically the same as is required for geophysical inverse problems. To demonstrate the algorithm in a concrete geophysical setting—a 1-D controlled source electromagnetic (CSEM) example from the Scarborough Gas Field (Myer *et al.* 2010, North West Australian shelf) is discussed. The 2-D non-linear image regression and the real data examples were introduced in Ray & Myer (2019), but are revisited here with the non-stationary TDGP formulation. Also different from Ray & Myer (2019), in this work a Matern 3/2 kernel has been used for all examples to facilitate rapid spatial decay in all GP representations.

3.1 1-D non-linear regression

This example (Fig. 3) is motivated by the Jump1D test function (Plagemann *et al.* 2008) which originally appeared in Paciorek & Schervish (2004). The objective in this regression example is to represent the true function *parsimoniously* with uncertainty in the locations not sampled while fitting the data points to within the noise. Geochemical examples of this kind of application can be found in Gallagher *et al.* (2011). The results of posterior inference using purely stationary mode TDGP are shown in Fig. 4, switching the axis labels to a geophysical context. From this figure, it is evident that a visually satisfactory fit to the true function is achieved. A fixed correlation length $\lambda_s = 0.1$ units was used, and prior ranges for the number of GP nuclei was set to [2,30]. The prior range for the GP nucleus properties was the extremal range of the data values to regress, and the prior ranges of the GP nucleus positions was set to [0,1]. A nugget value $\sigma_s = 0.05$ was also fixed through trial and error. Lower nugget values tend to favour close adherence to the GP nuclei, and sharper changes can be represented. However this can lead to instabilities in the Cholesky decomposition of $\mathbf{K}_m$ for $\mathbf{K}_m^{-1}\mathbf{m}$ in (5). Values of the GP nuclei were realized within the maximum and minimum ranges of the data. Sampling statistics can be seen in Fig. 5, indicating healthy McMC chain mixing, and that only 13 nodes on average were required to represent 197 function locations, a compression of nearly 15 times. Also evident is the fact that the data are not quite fit to within the noise—primarily because of the jump, as will be discussed underneath.

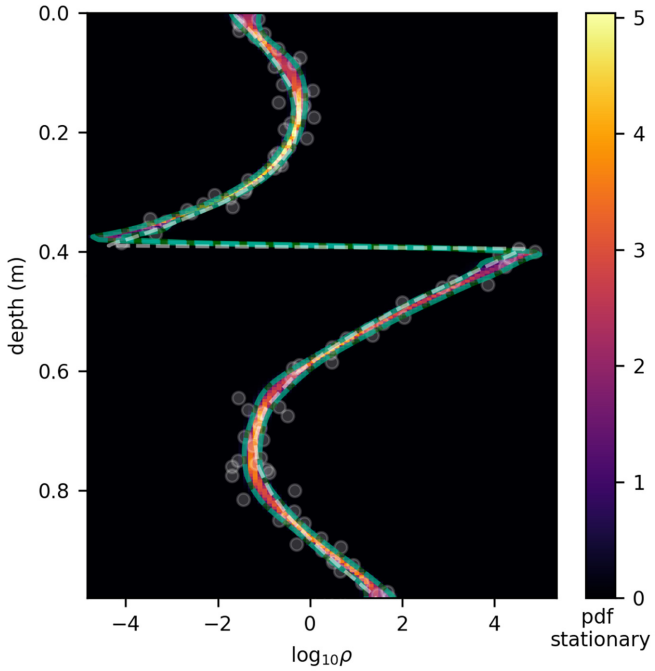

Figure 4. The Jump1D function recovered by purely stationary mode TDGP McMC sampling. The image has been rotated by 90° and the vertical axis is represented as depth and the horizontal axis as the $\log_{10}$ of resistivity ρ, as is conventional for a 1-D geophysical electromagnetic experiment. Green dashed lines represent the 90 per cent credible interval (CI) and hotter colours are more probable. The true function is plotted with a dashed white line. Noisy data are shown as translucent white circles. Stationary TDGP recovers the true function reasonably, including remarkably, the jump.

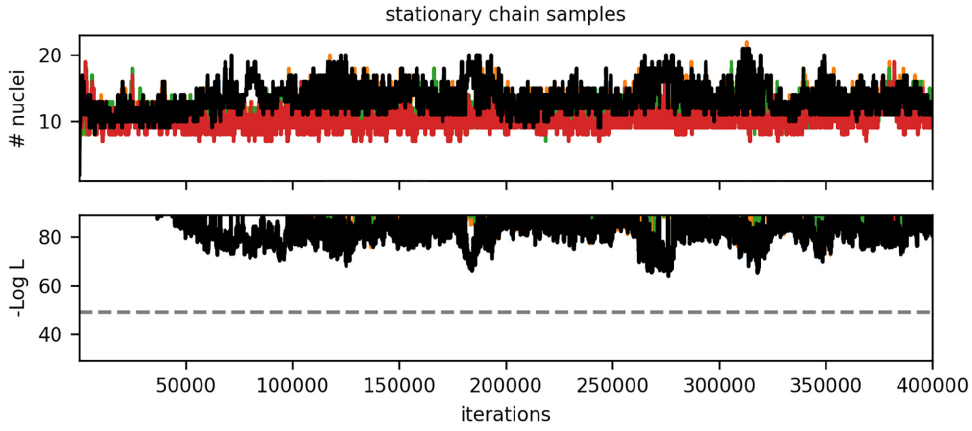

Figure 5. Sampling statistics for the stationary inference in Fig. 4. Each McMC chain at a particular temperature is denoted by a particular colour. The $T = 1$ chain used for inference is shown in black. The first quarter of the samples were discarded during burnin. From the first row it is clear that no more than 21 GP nuclei or nodes are ever required to fit the data and represent the 197 Jump1D function locations shown in Fig. 4. The second row shows the negative log likelihood or $\chi^2/2$ value. The dashed line shows the expected $\chi^2/2$ level and it becomes apparent that though the stationary TDGP algorithm does quite well, it does not quite fit the data to within the noise.

Posterior marginal distributions from the non-stationary mode TDGP inference are shown in Fig. 6, with accompanying sampling statistics shown in Fig. 7. As evidenced by the $\chi^2/2$, the data are now fit to within the noise. There are now two sets of sampling statistics, one each for the stationary McMC model $\boldsymbol{\theta}_{\rm s}$ parametrizing the length scales $\boldsymbol{\mu}_{*\rm s}$, and the other for the McMC model $\boldsymbol{\theta}_{\rm ns}$ parametrizing the properties $\boldsymbol{\mu}_{*\rm ns}$. The compression is now near 7 as on average, 10 non stationary GP nuclei together with 18 stationary GP nuclei (28 in total), are required to represent the 197 true function locations. However, the PSNR is also better for the non-stationary mode TDGP (Table 1), in addition to fitting data within the noise. The length scales McMC posterior models are also parsimonious and required to compute the properties GP mean, but the length scales can be considered as an abstract hierarchical layer the exact form of which is not the goal of the inference exercise. This will become clearer with the 2-D example that follows. As expected from the theory, the stationary McMC chain is able to parametrize the jump in the non-stationary properties as seen in Fig. 6 near 0.4 units of depth. The prior ranges for the properties non-stationary McMC model $\boldsymbol{\theta}_{\rm ns}$ were set to be the same as for the stationary mode TDGP, and for the length scales McMC model $\boldsymbol{\theta}_{\rm s}$, the prior ranges in the $\log_{10}$ length scale were set to be $[-1.2, -0.8]$, with the same mean (central) value as that of the stationary mode TDGP

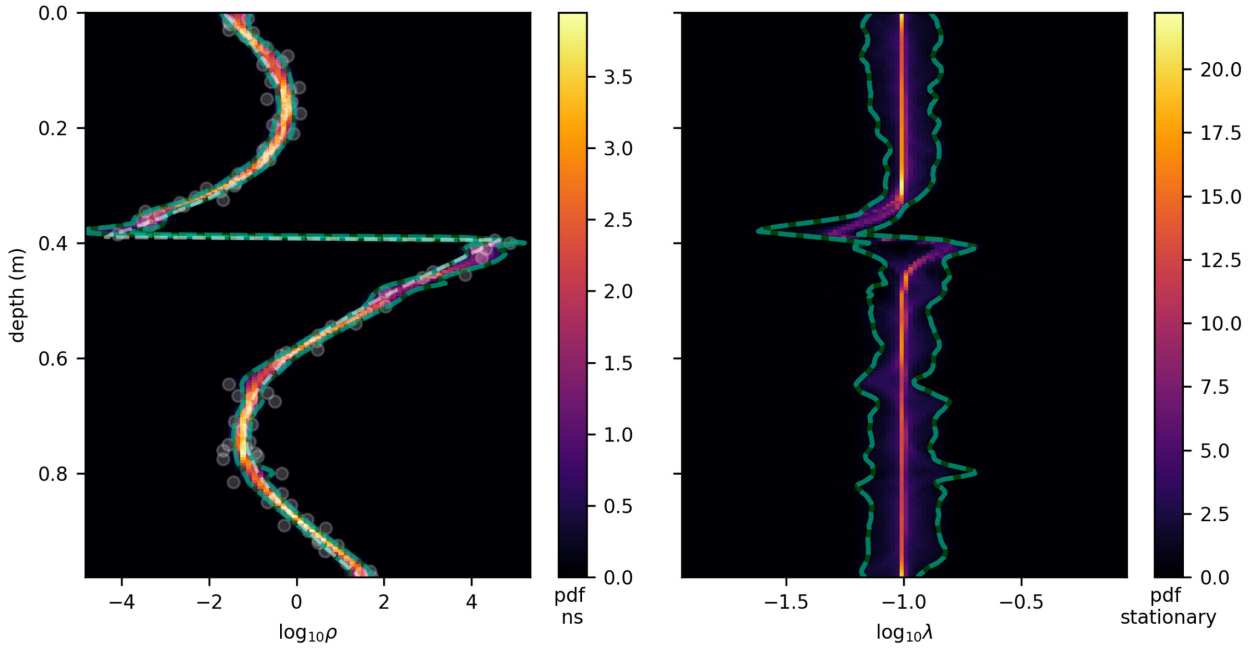

Figure 6. The Jump1D function recovered by non-stationary TDGP McMC sampling. As before, green dashed lines represent the 90 per cent CI and hotter colours are more probable. On the left, are shown the marginal posterior values corresponding to $\boldsymbol{\mu}_{*\mathrm{ns}}$, computed from the properties McMC model $\boldsymbol{\theta}_{\mathrm{ns}}$ and the length scales vector $\boldsymbol{\mu}_{*\mathrm{s}}$. The true function is plotted with a dashed white line. Noisy data are shown as translucent white circles. On the right are the marginal posterior values corresponding to the length scales $\boldsymbol{\mu}_{*\mathrm{s}}$ represented as $\log_{10}\lambda$ since they represent non-negative length scales, computed from McMC model $\boldsymbol{\theta}_{\mathrm{s}}$. Non-stationary TDGP recovers the true function, including the jump, to within data noise as the sampling statistics in Fig. 7 show. Of particular interest is the self-parametrization reflected in the length scales posterior on the right, which show a sharp decrease reflecting a jump in the properties posterior on the left, near depth = 0.4 m.

length scale of 10^{-1}. Both nuggets $\sigma_{m_{\mathrm{s}}}$ and $\sigma_{m_{\mathrm{ns}}}$ were set to 0.05, the same as in the purely stationary case after some experimentation on damping required in the ensuing respective GP mean. This example provides confidence that the algorithm works as it has been designed—a stationary TDGP can indeed be used to infer the length scales required to represent properties through a non-stationary TDGP model. This is further evidenced by zooming into the region near the jump as shown in Fig. 8. A PSNR comparison is presented in Table 1.

3.2 2-D Non-linear regression

The motivating 256 × 256 dimensional 2-D example shown in Fig. 1 is now discussed in detail. The purely stationary mode TDGP version of the same problem was discussed in Ray & Myer (2019) with a Squared Euclidean kernel. Similar geoscientific examples of this kind of problem have been investigated using Trans-D methods by Hopcroft *et al.* (2009); Bodin *et al.* (2012) and Hawkins *et al.* (2019) for borehole temperature inversion, Moho surface reconstruction and sea level rise. Depending on the specifics of the problem, interfaces were used for one spatial dimension and Voronoi cells for two. However, as in Ray & Myer (2019), the exact same theory (discussed in Section 2) and code framework holds for TDGP, no matter the number of spatial dimensions. The true function is a low-passed 256 × 256 pixel image of the standard milk drop test image 'splash' available from the SIPI database at the University of Southern California (http://sipi.usc.edu/database/). Eight hundred fifty-one of the original 65 536 pixels were randomly sampled, with a deliberate bias resulting in the upper part being sparsely sampled to investigate the algorithm's posterior adaptation to irregular, non-stationary data coverage. Random Gaussian noise with standard deviation equal to 5 per cent of the max value was also added to the 851 data points. As in the 1-D example, the objective is to find 2-D representations, within data noise, of the true image and associated uncertainty at all 65 536 locations. Again, kriging methods could be used to solve this problem with all 851 points as GP nodes, but *parsimonious* representations of this image are sought. It is this parsimony which enables tractable McMC sampling for geophysical applications over a spatially vast part of the earth, forward modeling the physics for which would require many pixels. The image is scaled to extend to 2560 units by 2560 units as may be expected for a seismic region of investigation in metres.

Purely stationary mode TDGP and non-stationary mode TDGP were run for 500 000 McMC iterations. The mean image reconstructions from both are shown in (c) and (d) of Fig. 1. Clearly (see Table 2), the non-stationary TDGP McMC does better by providing variable length scales with which to parametrize the pixel values.

The progress of purely stationary mode Trans-D sampling with $\mathbf{C}_{\mathrm{avg}} = \begin{pmatrix} 141^2 & 0 \\ 0 & 141^2 \end{pmatrix}$, that is λ_{s} set to 141 in both spatial dimensions is shown in Fig. 9(a). The data are fit to within noise without difficulty as the negative log likelihood shows. $65 < k < 100$ after achieving stationarity, though the maximum permissible prior value for k is 100. As in the 1-D regression example, prior bounds for the properties are

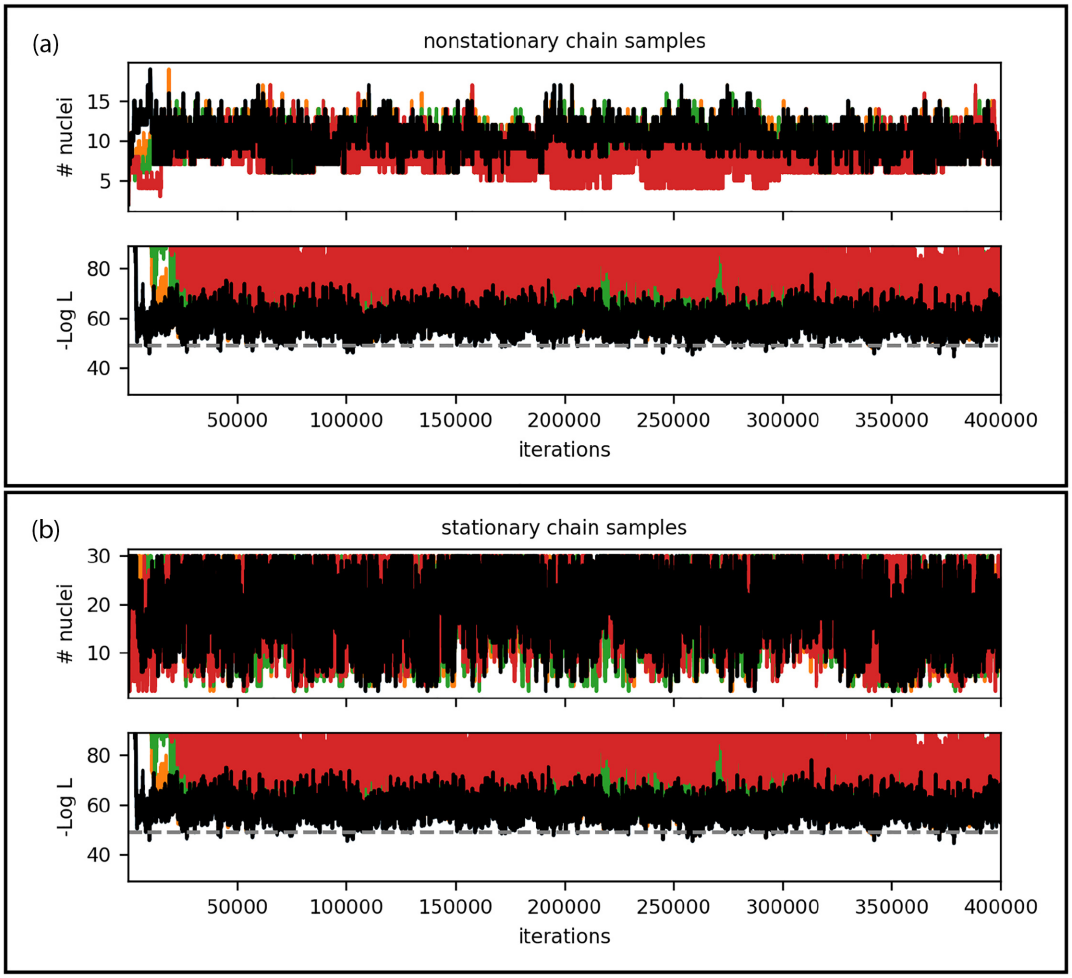

Figure 7. Sampling statistics for the non-stationary TDGP inference in Fig. 6. As before, McMC chains at each temperature are indicated by a different colour. The $T = 1$ chain used for inference is shown in black. The first quarter of the samples were discarded during burnin. a) Statistics for the non-stationary properties McMC model $\boldsymbol{\theta}_{\rm ns}$ corresponding to the left-hand column in Fig. 6. From the first row it is clear that no more than 17 GP nuclei representing the properties are ever required to fit the data. The second row shows the negative log likelihood, and the dashed line shows the expected $\chi^2/2$ value. It is evident that the non-stationary TDGP algorithm is able to fit the data to within the noise. (b) A similar set of plots as in (a) except, now for the stationary length scales chain with McMC model $\boldsymbol{\theta}_{\rm s}$ corresponding to the right-hand column in Fig. 6. From the first row it is clear that no more than 30 GP nuclei representing the underlying length scales are ever required to fit the data.

Table 1. Mean reconstruction PSNR comparison for the Jump1D function.

Type of length scale	PSNR dB
Stationary (Fig. 4)	17.78
Non-stationary (Fig. 6)	18.45

set to the extremal values of the noisy 851 data. In Fig. 9(b) it is apparent from the posterior CI that uncertainties are greater where data coverage is sparse (see Fig. 1), a distinct characteristic of appropriately formulated Bayesian algorithms.

Sampling statistics from non-stationary mode TDGP sampling are shown in Fig. 10. On average, ∼25 non-stationary GP nuclei representing property (pixel value) and ∼75 stationary GP nuclei representing ($\log_{10}$) length scales are required to produce the mean image shown in Fig. 1(d). This is a compression of 65 536:100 or 655 times. Posterior inference from the second half of samples after burnin can be seen in Figs 11 and 12. The maximum number of nodes for both the stationary and non-stationary McMC models was 100. The length scales were set to be in the prior $\log_{10}$ uniform range [1, 3.301] corresponding to linear bounds of 10 and 2000 distance units in both x and y. What is initially surprising, is that on using not one but two *a priori* independent TDGP McMC models—a stationary model $\boldsymbol{\theta}_{\rm s}$ for the length scales and a non-stationary model $\boldsymbol{\theta}_{\rm ns}$ for the properties (Fig. 2 illustrates how they are used in relation to one another), the non-stationary mode TDGP inference is superior to the purely stationary variant. This is evidenced in Figs 1, 9(b), 11 and 12 and quantified in Table 2. The reasons as to why this is so are speculated on in the conclusions section.

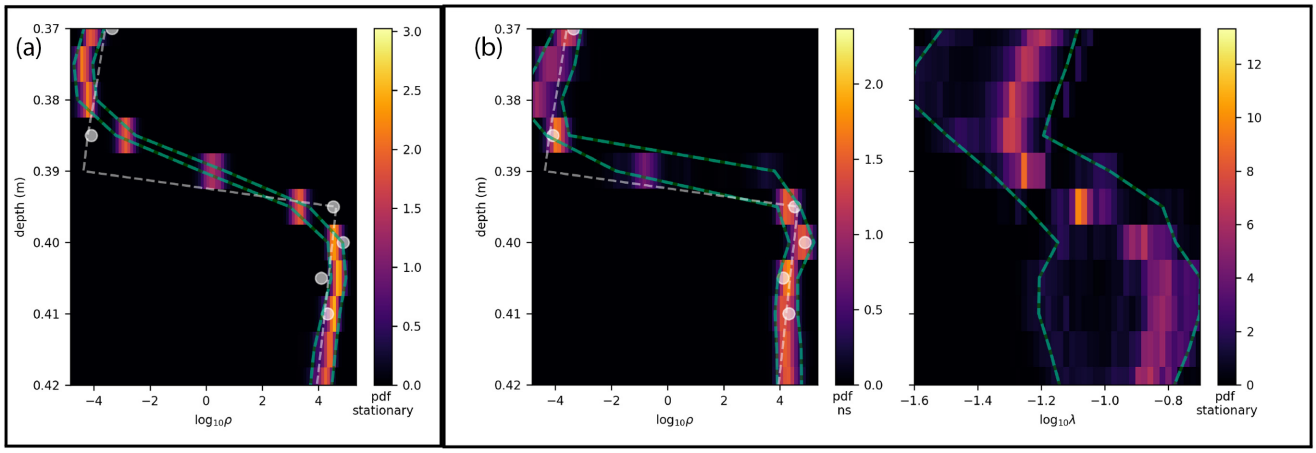

Figure 8. Zooming into the (a) purely stationary mode and (b) non-stationary mode TDGP posterior inferences near the jump. The true function is shown with a dashed white line, as before. The non-stationary mode TDGP represents the true function more faithfully. Further, where there is a paucity of data (white circles), the non-stationary mode TDGP reports higher posterior uncertainty (wider 90 per cent CI indicated by green dashed lines), non-oscillatory high probability regions and higher adaptability to the data.

Table 2. Mean reconstruction PSNR comparison for the milk-drop image.

Type of length scale	PSNR dB
Stationary (Fig. 1c)	21.70
Non-stationary (Fig. 1d)	23.69

3.3 CSEM inversion

The marine CSEM method is an active source sounding technique, in use for nearly four 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 electromagnetic scale length (skin depth) $\delta_{\rm EM} = \sqrt{\frac{2\rho}{\mu\omega}}$ 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 $\delta_{\rm EM}$ 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 which can characterize the conductivity of a given medium. The high resistivity of target geology with respect to its surroundings produces a detectable EM signature in the data, which needs to be inverted to a resistivity model before interpretation. Resistive targets can range from offshore freshwater aquifers (Blatter *et al.* 2019; Gustafson *et al.* 2019) to hydrocarbon accumulations (e.g. Constable 2006). However, owing to the presence of noise and the numerous trade-offs possible in the inversion of CSEM data, Bayesian inversion is ideal to quantify the associated resistivity model uncertainty (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; Gehrmann *et al.* 2015; Blatter *et al.* 2019). Purely stationary mode TDGP was used by Ray & Myer (2019) to invert data from the Exmouth plateau in the Northwest Australian Shelf. The flat stratigraphy and bathymetry lend themselves well to inversion with 1-D physics, which has been used here. This concluding example with the Exmouth data (Constable *et al.* 2019; Myer *et al.* 2015, 2010; Myer 2012) compares purely stationary mode TDGP and non-stationary mode TDGP, both modes using GP input warping (Sampson & Guttorp 1992; Blatter 2020; Blatter *et al*. 2020, in review). Instead of parametrizing depth $\mathbf{x}$ in eq. (1) in linear or logarithmic increments of Euclidean space, the sum of a geometric progression in terms of a thickness δz and an expansion fraction f has been used. The depth variable z representing $\mathbf{x}$ is now written as

$$z = \delta z \frac{(1 - f^n)}{(1 - f)}. \tag{15}$$

The insight through the use of this simple transform is that correlation lengths can now be specified in terms of a linear index n, instead of through depth. This allows for the input spatial dimension to be warped (e.g. Sampson & Guttorp 1992), resulting in longer correlation lengths with increasing depth. Further, this is accomplished without changing the GP math in Section 2 for *both* purely stationary mode and non-stationary mode TDGP. For purposes of prior specification, this length n need not be an integer value. A stationary correlation length of 1 implies that adjacent depths defined through (15) are well correlated. In other words, one TDGP nucleus will have an influence region of one thickness unit above and below itself, and this influence dies off rapidly with distance. However, as thickness increases by the fraction f with depth, larger spatial regions are correlated in the vicinity of TDGP nuclei as depth increases. Of course, with non-stationary TDGP,

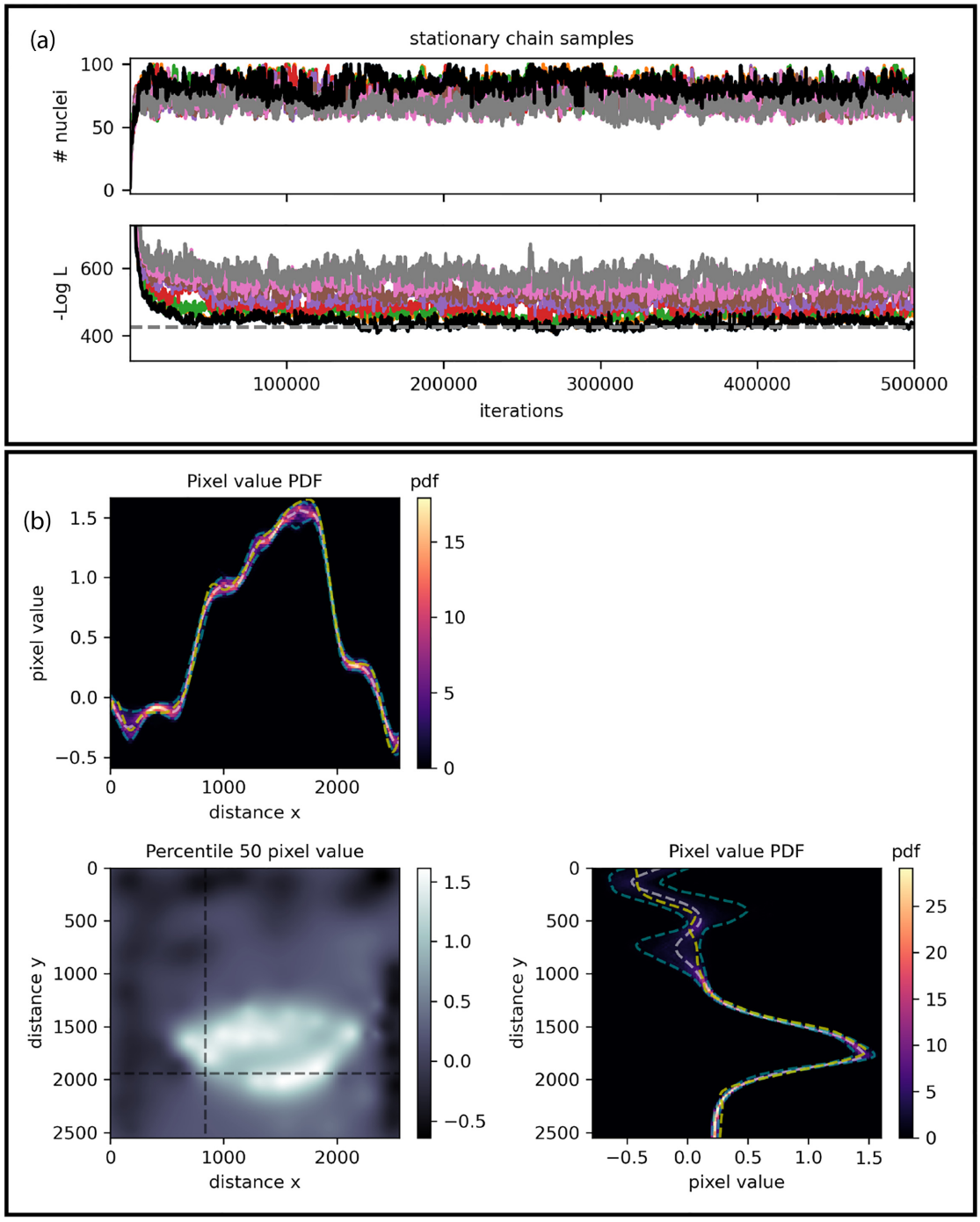

Figure 9. (a) Sampling statistics for the purely stationary mode TDGP inference for the 2-D image regression. As before, McMC chains in PT at each temperature are shown in a different colour and the $T = 1$ chain used for inference is shown in black. The first half of the samples were discarded during burnin. From the first row it is clear that on average 83 GP nuclei were able to represent the 256 × 256 image. The second row shows the negative log likelihood, and the dashed line is the expected $\chi^2/2$ value, indicating the data have been fit to within noise. (b) This figure shows posterior marginal distributions of the pixel values through row 195 (top left) and column 85 (bottom right) and the posterior median image in the bottom left. The profile locations are shown with dashed black lines. In the marginal posterior distributions through the corresponding row and column, hotter colours are more probable. The 90 per cent CI has been shown between green dashed lines. The median posterior values are shown with a dashed white line, and the true pixel values are shown by a dashed yellow line. For the row going through the high data density section, the CI is quite narrow. For the column which passes through low data density between 0 and 1000 units in y, the CI is quite wide as expected. Again, the purely stationary mode TDGP is able to adapt to both the data density as well as its underlying features.

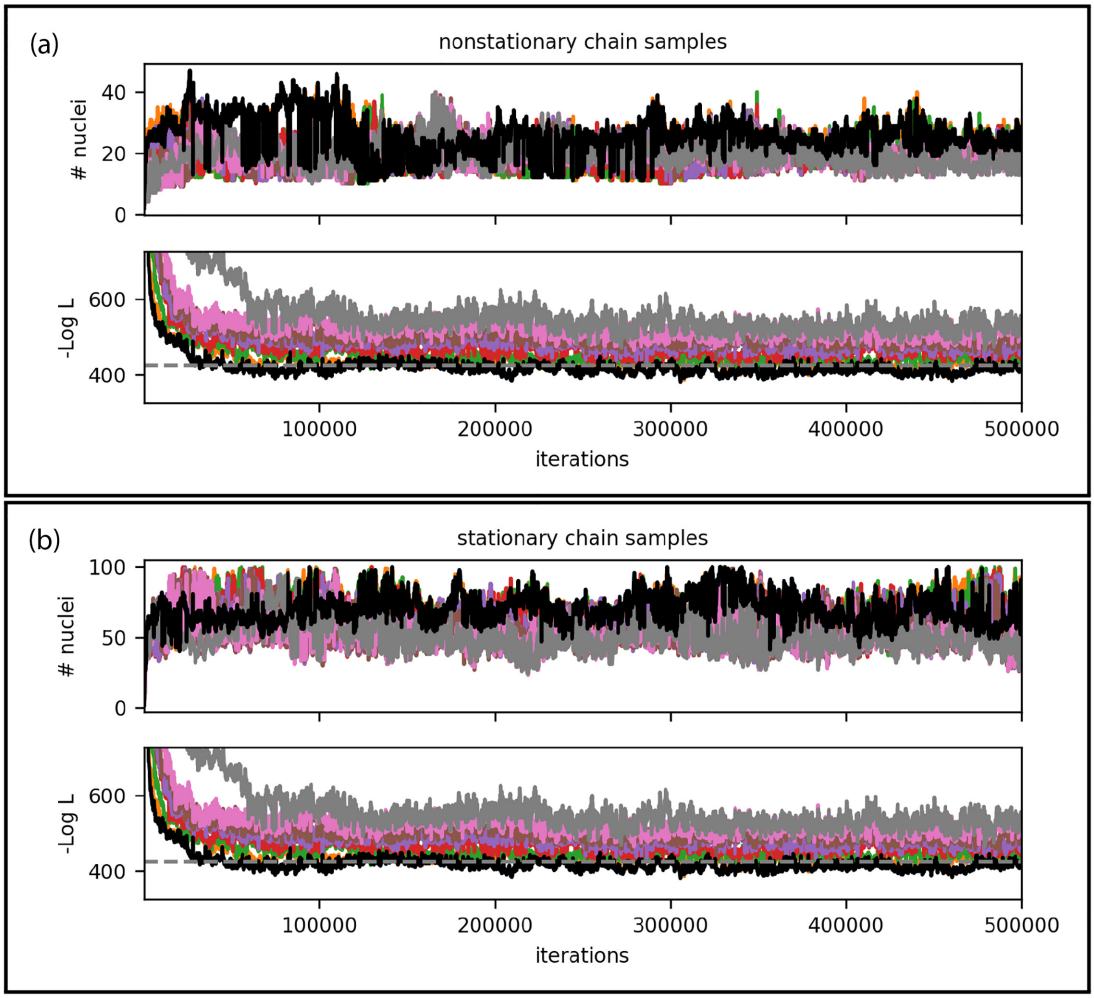

Figure 10. Sampling statistics for non-stationary TDGP inference for the 2-D non-linear regression problem. As before McMC chains in PT are shown with a different colour per temperature. The $T = 1$ chain used for inference is shown in black. The first half of the samples were discarded during burnin. (a) Statistics for the non-stationary properties McMC model $\boldsymbol{\theta}_{\rm ns}$. From the first row it is clear that no more than 39 GP nuclei representing the properties are ever required to fit the data post burnin. The second row shows the negative log likelihood, and the dashed line shows the expected $\chi^2/2$ value. The non-stationary mode TDGP algorithm is able to fit the data to within the noise. (b) A similar set of plots as in (a) except, now for the stationary length scales chain with McMC model $\boldsymbol{\theta}_{\rm s}$. From the first row it is clear that no more than 100 GP nuclei representing the underlying length scales are ever required to fit the data.

as has been described throughout this work, ranges of this length scale can be sampled to allow for sharp changes or smoothness where the geophysical data demand.

Using a maximum likelihood scaling factor on the assigned data error (e.g. Ray & Myer 2019; Sambridge 2013; Dosso & Wilmut 2012; Mecklenbrauker & Gerstoft 2000), both purely stationary mode TDGP and non-stationary TDGP can fit the on-reservoir Scarborough gas field data (Myer 2012) to within the assigned error bars as shown in Fig. 13. The stationary TDGP used a correlation length in index units of 2, corresponding to ∼40 m length at 2000 m depth (see Fig. 13), near the known resistive accumulation which is approximately 50–100 m thick with interbedded shales and siltstones. Prior limits for the properties McMC model in both the purely stationary mode and non-stationary mode TDGP inversions were set to $[-0.5, 2.3]$ in the $\log_{10}$ resistivity domain, or 0.5–200 ohm-m linear. The purely stationary mode TDGP had a uniform prior in the number of nuclei in the range [2, 40]. The non-stationary mode TDGP had a uniform prior over the number of resistivities nuclei in the range [2, 20] and a uniform prior in the range [2, 40] over the number of length scales nuclei. In addition, the non-stationary mode TDGP inversion McMC model had the the length scale set to lie in the prior $\log_{10}$ uniform range [0, 1.5] corresponding to bounds of 1 and 31.6 depth index units. The left-hand panel in Fig. 13 illustrates how depth is represented non-linearly with a linear index.

It should be noted that the models from non-stationary mode TDGP are smoother than their purely stationary TDGP counterparts, except at certain depth locations of note. These CSEM data are from the on-reservoir part of CSEM tow Line 2 (see Myer 2012). All across the tow line, there exists between ∼1.4 and 2.0 km depth the resistive Gearle siltstones formation. This regional resistor makes inferring the resistive gas reservoir which is present only in the on-reservoir part at ∼2 km depth quite difficult. Further, the reservoir is a laminated structure with low bulk resistivity (10–25 ohm-m, 1–1.39 in $\log_{10}$), which makes it hard to infer amidst the background as is discussed in detail by Myer *et al.* (2012). For deterministic inversion, the resistor location had to be fixed, with deliberate smoothness penalty relaxations applied between

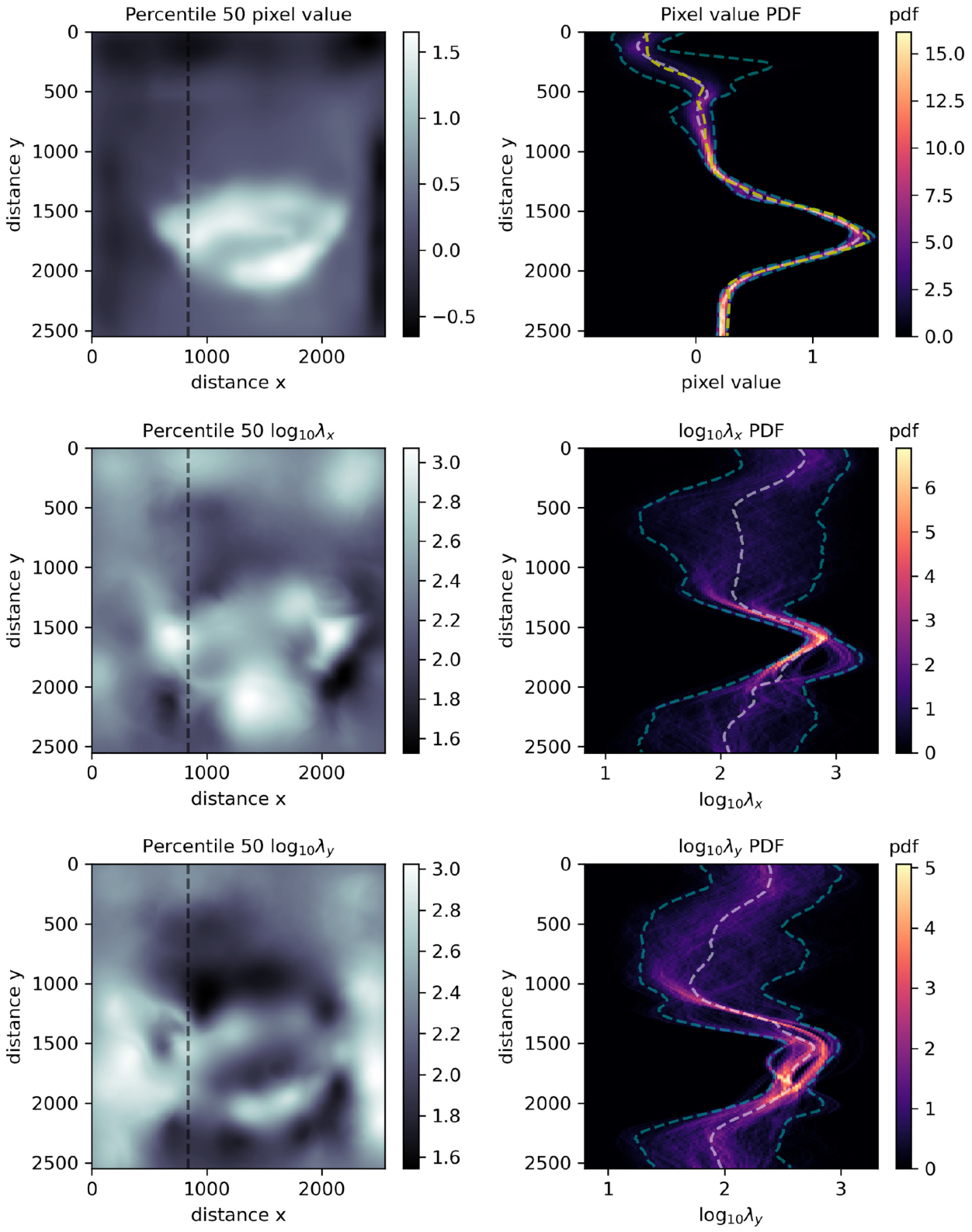

Figure 11. Posterior median (left-hand column) and profile marginal distributions (right-hand column) for the non-stationary mode TDGP in the 2-D regression problem. A profile through image column 85 is shown in the left-hand column with a dashed black line. In the right-hand column, the 90 per cent CI has been shown with green dashed lines. The median posterior value is shown with a dashed white line, and the true pixel values are shown by a dashed yellow line. Hotter histogram colours are more probable as usual. As this profile goes through variable data density, the pixel value CI is quite narrow for high data density and wide between 0 and 1000 units in y where data are sparse. The pixel values are parametrized by $\boldsymbol{\theta}_{\mathrm{ns}}$ and represented by the non-stationary GP mean $\boldsymbol{\mu}_{*\mathrm{ns}}$. There are two length scales λ_x and λ_y which are parametrized by a multichannel GP through $\boldsymbol{\theta}_{\mathrm{s}}$ and represented by two stationary GP means within each column of $\boldsymbol{\mu}_{*\mathrm{s}}$, one each for the x and y directions. The median values of λ_y and its CI tend to dip in the vicinity of rapid changes in pixel value in y.

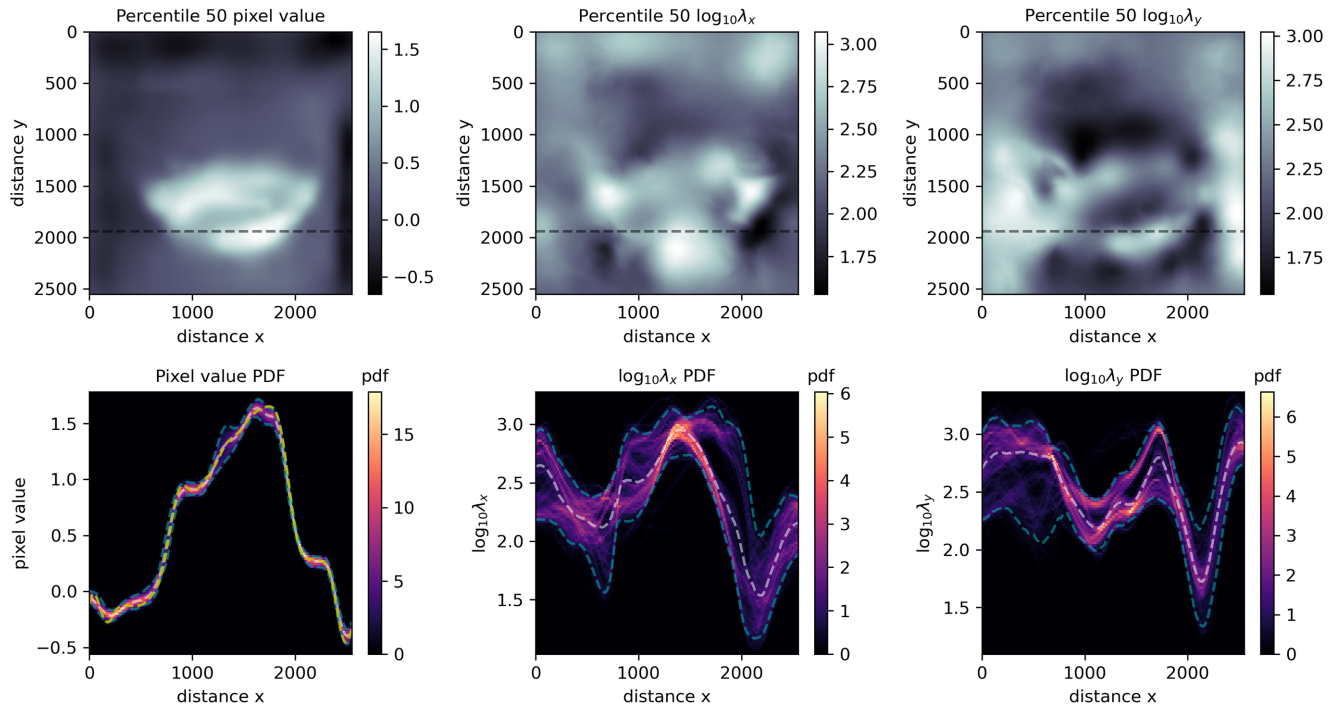

Figure 12. Posterior median (top row) and profile marginal distributions (bottom row) for the non-stationary mode TDGP in the 2-D regression problem. A profile through row 195 is shown in the top row with a dashed black line. In the bottom row, the 90 per cent CI has been shown with green dashed lines. The median posterior value is shown with a dashed white line, and the true pixel values are shown by a dashed yellow line. As this row goes through a high data density, the CI is quite narrow for the pixel values. The pixel values are parametrized by $\boldsymbol{\theta}_{\rm ns}$ and represented by the non-stationary GP mean $\boldsymbol{\mu}_{*\rm ns}$. There are two length scales λ_x and λ_y which are parametrized by a multichannel GP through $\boldsymbol{\theta}_{\rm s}$ and represented by two stationary GP means within each column of $\boldsymbol{\mu}_{*\rm s}$, one each for the x and y directions. The median values of λ_x and its CI tend to dip in the vicinity of changes in pixel value in x. Note how the right edge of the image has consistently high length scales in λ_y both in the median inference and the marginal histograms, as is expected given the constant sliver of dark pixels in the y direction at that location.

1900 and 2000 m. This allowed for sharp resistivity changes commensurate with a gas accumulation to be inverted. However, as can be seen from both Figs 13(a) and (b) there are high resistivities greater than 10 ohm-m at ∼2000 m depth. However the anomalous nature of the feature appears to stand out more in the non-stationary TDGP posterior models in Fig. 13(b). This will be discussed further when looking at posterior marginal distributions of resistivity.

Both the purely stationary mode and non-stationary mode TDGP McMC were run for 4 000 000 iterations and the first half of the samples were discarded during burnin. Sampling statistics are shown in Fig. 14. The originally supplied data errors were diagonal and produced very small residuals (see Ray *et al.* 2013b). As a consequence, similar to Ray & Myer (2019) these errors have been scaled by a maximum likelihood approach (see Appendix B). By inverting both in-tow and out-tow data which are not correlated at similar source–receiver offsets, the issue of correlated error has been somewhat circumvented. Hierarchical attempts to deal with correlated CSEM errors for Scarborough gas field are detailed in Ray *et al.* (2013b) but are not the focus of this work.

From the posterior marginal distributions of resistivity (Fig. 15), a comparison of the the purely stationary mode and non-stationary mode TDGP results can be made. Both show large changes in resistivity at a depth close to 2000 m, where there is a known, moderate resistivity gas reservoir. It appears that the reservoir is more prominently characterized using the adaptive parametrization within non-stationary mode TDGP, while it is primarily the base of the reservoir which is imaged with the purely stationary mode TDGP. It is known from stratigraphy of the region, summarized in Myer *et al.* (2012), that below the moderately resistive Gearle siltstones at ∼1800 m, there are fine, alternating sand/shale/silt sequences above and within the laminated hydrocarbon bearing reservoir at ∼2000 m. This leads to a gradual change in the bulk resistivity at the reservoir top, and in comparison, a sharper change at the bottom. As detailed in Myer *et al.* (2012) and Ray *et al.* (2014), the Gearle formation by itself is not resolvable given its low bulk resistivity increase compared to the background. The confounding effect of the Gearle resistor and the gradual change in bulk resistivity underneath makes inference of a clear reservoir top difficult. It is surmised that the chosen fixed correlation length with purely stationary mode TDGP is only able to resolve the larger resistivity contrast at the reservoir bottom. The non-stationary mode TDGP however, it is conjectured, can adapt length scales to be longer within the main reservoir, and shorter at the top and bottom. There is clearly a set of modal length scales which display this behaviour in the right panel of Fig. 15(b). However, it must be stressed there is no escaping non-uniqueness in the inverted resistivity, as other non-reservoir lithologies may also be interpreted with the purely non-stationary mode TDGP. As with all geological interpretation exercises, ancillary data such as nearby well logs and seismic imaging should be used to make a robust interpretation regarding the presence of significant hydrocarbon accumulation, which is not the objective in this work. However, interpreting the probabilities from an adaptively parametrized inversion, open up possibilities which are

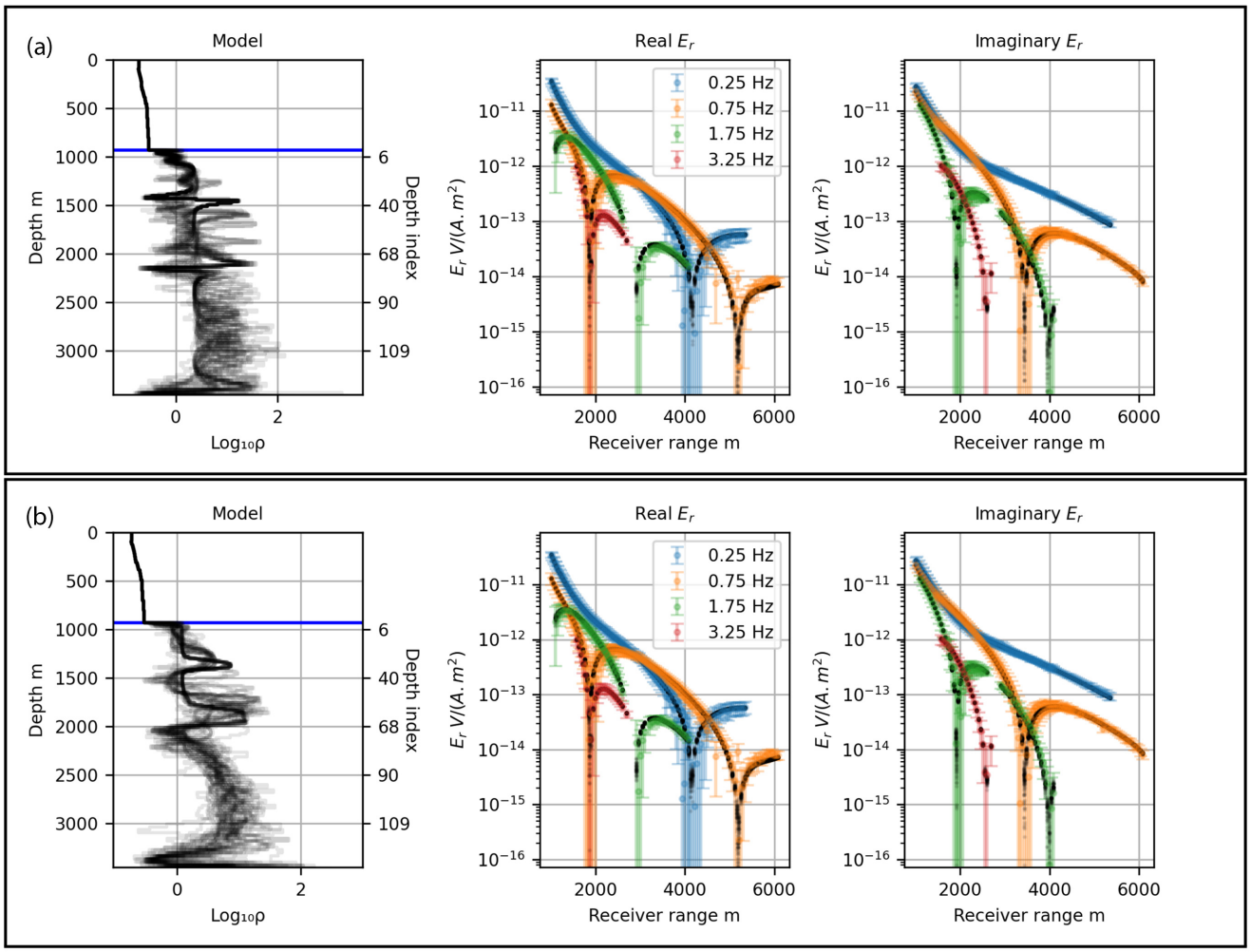

Figure 13. 50 randomly selected posterior models and their CSEM responses from inverting the on reservoir Scarborough field CSEM data using (a) purely stationary mode TDGP (b) non-stationary mode TDGP. Note how the depth indices introduced in (15) start from the sea floor (blue horizontal line in left-hand panel) and encompass larger depth intervals with increasing depth. The black dots in the middle and right-hand panels indicate the real and imaginary responses from the 30 models. Using a maximum likelihood scaling factor per frequency for the data error, each frequency is fit to within the originally supplied error bars.

not brought to light either in a deterministic interpretation workflow or if the inversion length scale is fixed. This is especially true if the regularization or fixed length scale used is not reflective of the interplay between depositional geology and geophysical sensitivity at depth.

4 DISCUSSION AND CONCLUSIONS

All inference, inversion or learning requires appropriate context. In Bayesian terms, this can be presented as making a choice in-between two extremes: Complete uncertainty about the solution sought after and therefore trust in only the data, and, complete certainty about the solution thus rendering information in the data worthless. This is where context, or a balance between prior probability (uncertainty about the solution) and the likelihood (uncertainty about the data), plays an important role in providing a useful solution. In a Bayesian framework, this balance is organically achieved through the specification of prior probabilities and likelihoods. In this work, we have attempted to demonstrate that this process can be hierarchically structured, such that we can infer the characteristics of the prior via the same inference machinery used to sample the posterior. This is where the approach presented here differs from hierarchical methods such as those found in Malinverno & Briggs (2004). This nested manner of thinking is not new, and in the ML literature can be found in the design of ML algorithms which are designed by these ML algorithms themselves (e.g. Andrychowicz *et al.* 2016). Why this line of thinking is particularly advantageous, is because it allows the algorithms to exploit structure in the high dimensional posterior space, in a generalizable way that often outperforms manual prior specification for the problem (e.g. Roininen *et al.* 2019).

When a function representing a property (e.g. regression data, image pixel values, or electromagnetic conductivity) displays spatial changes, spatial derivatives of this function will reflect these changes. For example, a sharp change in the Jump1D function is characterized by small underlying length scales near the jump (Fig. 6). The magnitude of the spatial derivative of the Jump1D function also has a high

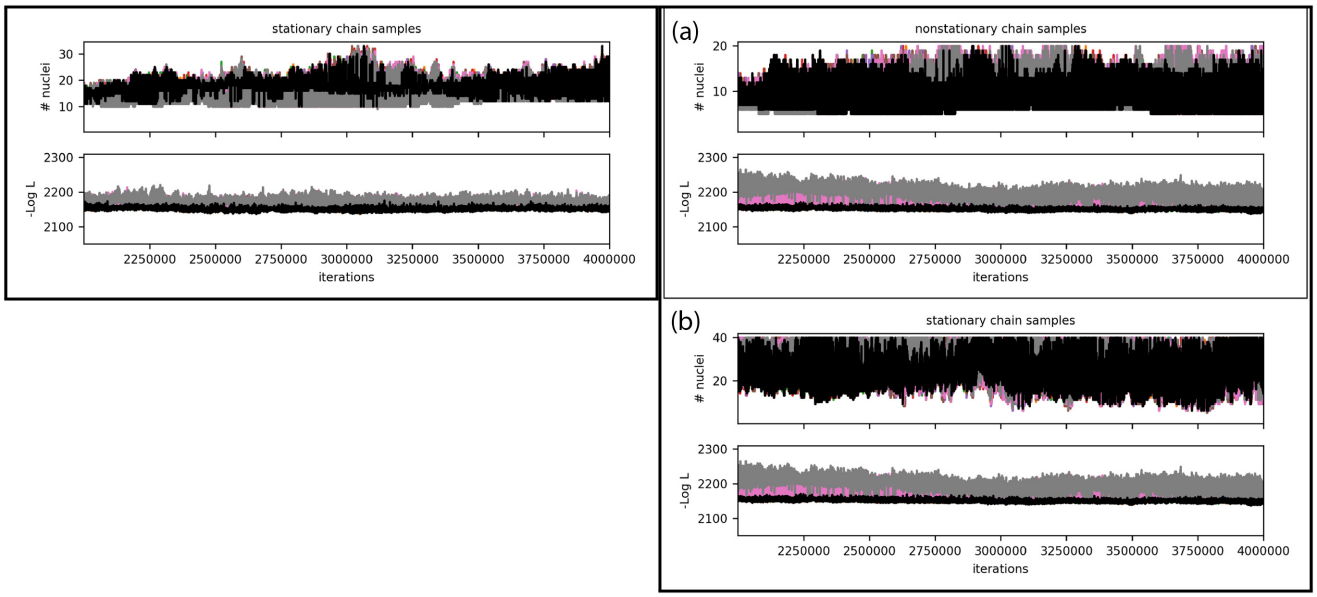

Figure 14. Sampling statistics for the Exmouth plateau CSEM inversion, using purely stationary mode TDGP (left-hand column) and non-stationary mode TDGP (right-hand column). As before McMC chains in PT are shown in a different colour per temperature. The $T = 1$ chain used for inference is shown in black. In the first row (left-hand column), 10–30 nuclei representing subsurface resistivity are required to fit the data to a stable negative log-likelihood (second row). (a) Statistics for the non-stationary mode TDGP properties (resistivity) McMC model $\boldsymbol{\theta}_{\text{ns}}$. From the first row it is clear that no more than 20 GP nuclei representing the subsurface resistivity are ever required to fit the data. The second row shows the negative log likelihood, which is similar to that attained by the purely stationary mode TDGP (left). (b) A similar set of plots as in (a) except, now for the stationary length scales chain with McMC model $\boldsymbol{\theta}_{\text{s}}$. From the first row it is clear that no more than 40 GP nuclei representing the underlying length scales are ever required to fit the CSEM data.

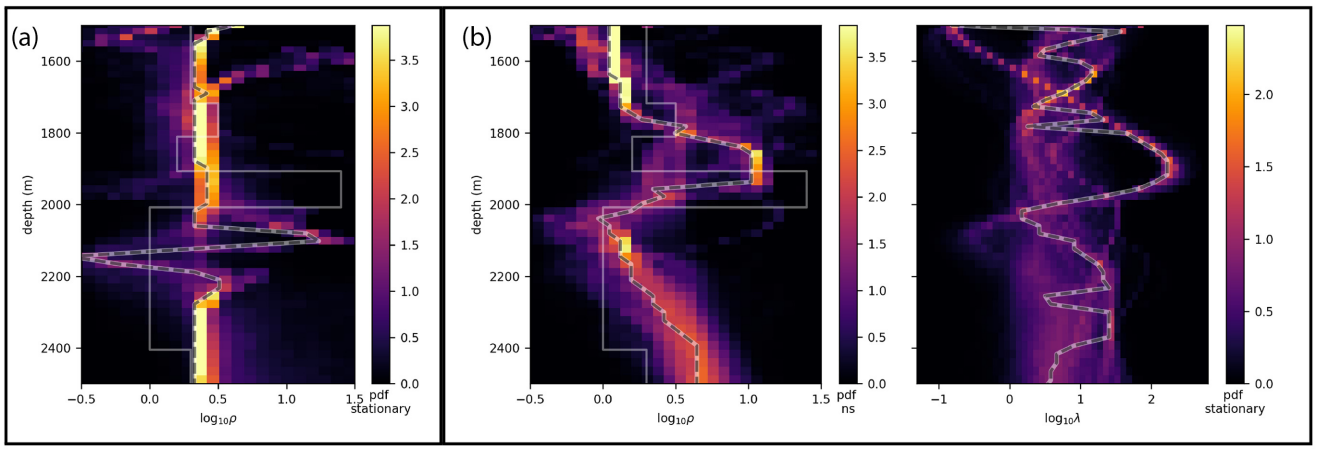

Figure 15. Zoomed in, posterior marginal resistivity distributions of on-reservoir Scarborough CSEM data for the (a) purely stationary mode TDGP and (b) non-stationary mode TDGP with associated length scales. The location of the marginal posterior maximum has been plotted with a faded dashed black-and-white line. Hotter colours are more probable as before. The conceptual, simplified resistivity model is shown as a faded white line. While the purely stationary mode TDGP shows that the posterior resistivity PDF below 2000 m depth alternates from resistive back to conducting, the resistive reservoir near 2000 m depth appears more prominent with the non-stationary TDGP formulation. As detailed in the text, the difference in behaviour between the two posteriors can be attributed to the laminated stratigraphy of the reservoir, GP length scales (fixed vs sampled), associated resistivity contrasts and diffusive CSEM sensitivity.

value at the jump. If the underlying TDGP length scales are analogous to the inverse of the spatial derivative, parallels with the Taylor series expansion for a function can be drawn. To be explicit, when using two levels of abstraction (e.g. non-stationary mode TDGP), the analogy is with a first-order Taylor approximation. The constant value is given by the properties GP and the gradient value is provided by the (inverse of the) length scales GP. It should be stressed that this observation is made purely by way of analogy. However, the connection between Gaussian processes, deep neural networks and their Taylor expansions in a Bayesian context is given a rigorous mathematical treatment by Lee *et al.* (2019).

As pointed out by Fuglstad *et al.* (2015), the use of non-stationary length scales is not always necessary. A significant hurdle in the implementation of the Bayesian framework proposed through non-stationary mode TDGP, is that two Trans-D processes need to be run to generate the posterior. For the same number of McMC samples of the geophysical property of interest, there needs to be an equal number of McMC samples of the length scales. This requires the use of exactly twice the amount of computation, for the same number of purely

stationary mode TDGP samples. However, the length scales could perhaps be sampled less often than the properties, which will lead to significant computational savings. Approximations required for the efficient computation of stationary and non-stationary kernels are detailed in Appendix A4.

It is hoped through this work that purely stationary mode TDGP and the newly introduced nested, non-stationary mode TDGP Bayesian inference have been shown to be useful, generalizable and adaptable methods for geophysical inference. The same theory and code can be applied across a wide range of geophysical problems in 1-D, 2-D and potentially 3-D. This becomes apparent, when considering that the theory and code from Ray & Myer (2019) and the extensions proposed here have allowed the algorithm to parametrize even itself. Subjective choices are unavoidable in Bayesian inference and in reality most other forms of inference, since infinite quantities of informative data are not available. If these choices can be made in a systematic manner as demonstrated in this work, it will make the ensuing inferences and the real world decisions which stem from them, more robust.

ACKNOWLEDGEMENTS

All calculations were carried out using the Julia language (Bezanson *et al.* 2017, 2015, 2012), available under the MIT license. Richard Taylor, Yusen Ley-Cooper and Ross C. Brodie are thanked for early reviews which increased the clarity of the manuscript. David Myer is thanked for providing the Scarborough CSEM data used in this work, and Stephen Constable is thanked for making the data publicly available (https://marineemlab.ucsd.edu/Projects/Scarborough/Data.html). Jan Dettmer, Andrea Licciardi and an anonymous reviewer are thanked for their constructive reviews. Marina Costelloe, Richard Blewett and David Robinson are thanked for actively encouraging the author's mathematical forays into inverse theory and inference. The research presented in this paper was made possible by the Exploring for the Future (EFTF) program (https://www.ga.gov.au/eftf). This paper is published with the permission of the CEO, Geoscience Australia.

REFERENCES

Agostinetti, N.P. & Malinverno, A., 2010. Receiver function inversion by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.,* **181**(2), 858–872.

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.

Backus, G.E., 1988. Bayesian inference in geomagnetism, *Geophys. J. Int.,* **92**(1), 125–142.

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.

Bezanson, J., Karpinski, S., Shah, V.B. & Edelman, A., 2012. *Julia: A Fast Dynamic Language for Technical Computing,* 1–27.

Blatter, D., 2020, Constraining fluid properties in the mantle and crust using Bayesian inversion of electromagnetic data, *PhD thesis*, Columbia University.

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.

Blatter, D., Key, K., Ray, A., Gustafson, C. & Evans, R., 2019. Bayesian joint inversion of controlled source electromagnetic and magnetotelluric data to image freshwater aquifer offshore New Jersey, *Geophys. J. Int.,* **218**(3), 1822–1837

Bodin, T., Salmon, M., Kennett, B.L.N. & Sambridge, M., 2012. 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., 2009. Seismic tomography with the reversible jump algorithm, *Geophys. J. Int.,* **178**(3), 1411–1436.

Brodie, R.C. & Sambridge, M., 2012. Transdimensional Monte Carlo inversion of AEM Data, in *Proceedings of the 22nd International Geophysical Conference and Exhibition,* no. 1, Brisbane, Australia.

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.

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, *Wiley Interdiscip. Rev. Comput. Stat.,* **10**(3), 1–19.

Carlsson, K. *et al.*, 2019. NearestNeighbors.jl.

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.

Constable, S., Orange, A. & Myer, D., 2019. Marine controlled-source electromagnetic of the Scarborough gas field Part 3: multicomponent 2D magnetotelluric/controlled-source electromagnetic inversions, *Geophysics,* **84**(6), B387–B401.

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.

Cotter, S.L., Roberts, G.O., Stuart, A.M. & White, D., 2013. MCMC methods for functions: modifying old algorithms to make them faster, *Stat. Sci.,* **28**(3), 424–446.

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

Damianou, A.C. & Lawrence, N.D., 2013. Deep Gaussian processes, in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)* 2013, Scottsdale, AZ, USA, Vol. 31 of JMLR: W&CP 31.

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., 2013. Probabilistic two-dimensional water-column and seabed inversion with self-adapting parameterizations, *J. acoust. Soc. Am.,* **133**(5), 2612–23.

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

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.

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.

Dosso, S.E., Dettmer, J., Steininger, G. & Holland, C.W., 2014. Efficient trans-dimensional Bayesian inversion for geoacoustic profile estimation, *Inverse Problems,* **30**(11), 114018.

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.

Dunlop, M.M., Girolami, M.A., Stuart, A.M. & Teckentrup, A.L., 2018. How deep are deep Gaussian processes? *J. Mach. Learn. Res.,* **19,** 1–46.

Duvenaud, D., Rippel, O., Adams, R.P. & Ghahramani, Z., 2014. Avoiding pathologies in very deep networks, *J. . Learn. Res.,* **33,** 202–210.

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

Emzir, M., Lasanen, S., Purisha, Z., Roininen, L. & Särkkä, S., 2020. Non-stationary multi-layered Gaussian priors for Bayesian inversion, (arXiv:2006.15634) .

Fairbrother, J., Nemeth, C., Rischard, M., Brea, J. & Pinder, T., 2018. GaussianProcesses.jl: a nonparametric bayes package for the Julia Language, (arXiv:1812.09064).

Fisher, R.A. & Yates, F., 1938. *Statistical Tables: For Biological, Agricultural and Medical Research,* Oliver and Boyd.

Fuglstad, G.A., Simpson, D., Lindgren, F. & Rue, H., 2015. Does non-stationary spatial data always require non-stationary random fields?, *Spatial Stat.,* **14**(1), 505–531.

Galetti, E. & Curtis, A., 2018. Transdimensional Electrical Resistivity Tomography, *J. geophys. Res.,* **123**(8), 6347–6377.

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.

Galy-Fajou, T., Wenzel, F., Donner, C. & Opper, M., 2018. Scalable multi-class Gaussian process classification via data augmentation, in *Proceedings of the Symposium on Advances in Approximate Bayesian Inference at NeurIPS 2018*.

Gehrmann, R.A.S., Dettmer, J., Schwalenberg, K., Engels, M., Dosso, S.E. & Özmaral, A., 2015. Trans-dimensional Bayesian inversion of controlled-source electromagnetic data in the German North Sea, *Geophys. Prospect.,* **63**(6), 1314–1333.

Geyer, C.J., 1991. Markov chain Monte Carlo maximum likelihood, in *Proceedings of the 23rd Symposium on the Interface, New York,* pp. 156, American Statistical Association.

Geyer, C.J. & Møller, J., 1994. Simulation procedures and likelihood inference for spatial point processes, *Scand. J. Stat.,* **21**(4), 359–373.

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

Green, P.J. & Hastie, D.I., 2009. Reversible jump MCMC, *Genetics,* **155**(3), 1391–1403.

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.

Gustafson, C., Key, K. & Evans, R.L., 2019. Aquifer systems extending far offshore on the U.S. Atlantic margin, *Scient. Rep.,* **9**(1), 1–10.

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., 2017, Advances in trans-dimensional geophysical inference, *PhD thesis*, Australian National University.

Hawkins, R., Bodin, T., Sambridge, M., Choblet, G. & Husson, L., 2019. Trans-dimensional surface reconstruction with different classes of parameterization, *Geochem. Geophys. Geosyst.,* **20**(1), 505–529.

Hawkins, R., Brodie, R.C. & Sambridge, M., 2017. Trans-dimensional Bayesian inversion of airborne electromagnetic data for 2D conductivity profiles, *Explor. Geophys,* **49**(2), 134–147.

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

Hopcroft, P.O., Gallagher, K. & Pain, C.C., 2009. A Bayesian partition modelling approach to resolve spatial variability in climate records from borehole temperature inversion, *Geophys. J. Int.,* **178**(2), 651–666.

Horé, A. & Ziou, D., 2010. Image quality metrics: PSNR vs. SSIM, in *Proceedings of the 2010 20th International Conference on Pattern Recognition,* pp. 2366–2369.

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 Univ. Press.

Jiaxuan, Y., Xiaocheng, L., Low, M., Lobell, D. & Ermon, S., 2017. Deep Gaussian process for crop yield prediction based on remote sensing data, in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,* pp. 4559–4565.

Kass, R.E. & Raftery, A.E., 1995. Bayes factor, *J. Am. Stat. Assoc.,* **90,** 430, 773–795.

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, *Autonom. Rob.,* **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. S. Afr.,* **52**(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, *Water Resour. Res.,* **54**(1), 381–406.

Lee, J., Xiao, L., Schoenholz, S.S., Bahri, Y., Novak, R., Sohl-Dickstein, J. & Pennington, J., 2019. Wide neural networks of any depth evolve as linear models under gradient descent, in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019),* Vancouver, Canada.

Lewicki, M.S. & Sejnowski, T.J., 2000. Learning overcomplete representations, *Neural Comput.,* **12**(2), 337–365.

Ley-Cooper, A.Y., 2016. Dealing with uncertainty in AEM models (and learning to live with it), *ASEG Extend. Abstr.,* **2016**(1), 1–6.

Lindgren, F., Rue, H. & Lindström, J., 2011. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach, *J. R. Stat. Soc., B,* **73**(4), 423–498.

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.

MacKay, D., 1998. Introduction to {G}aussian Processes, in *Book Neural Networks and Machine Learning,* pp. 84–92, Springer-Verlag.

MacKay, D.J.C., 2003. *Information Theory, Inference and Learning Algorithms,* Cambridge Univ. 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. & Briggs, V.A., 2004. Expanded uncertainty quantification in inverse problems: hierarchical Bayes and empirical Bayes, *Geophysics,* **69**(4), 1005–1016.

Malinverno, A. & Leaney, S., 2000. A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data, in *SEG Annual Meeting,* (3), 2393–2396.

Malinverno, A. & Parker, R.L., 2006. Two ways to quantify uncertainty in geophysical inverse problems, *Geophysics,* **71**(3), W15–W27.
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.
Muir, J.B. & Tkalčić, H., 2020. Probabilistic lowermost mantle P-wave tomography from hierarchical Hamiltonian Monte Carlo and model parametrization cross-validation, *Geophys. J. Int.,* **223**(3), 1630–1643.
Muir, J.B. & Tsai, V.C., 2020. Geometric and level set tomography using ensemble Kalman inversion, *Geophys. J. Int.,* **220**(2), 967–980.
Murphy, K.P., 2012. *Machine Learning: A Probabilistic Perspective,* MIT Press.
Myer, D., 2012, Electromagnetic exploration of the Exmouth and Vøring rifted margins, *PhD thesis*, University of California, San Diego, CA, USA.
Myer, D., Constable, S. & Key, K., 2010. A marine EM survey of the Scarborough gas field, Northwest Shelf of Australia, *First Break,* **28,** 77–82.
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., Key, K. & Constable, S., 2015. Marine CSEM of the Scarborough gas field, Part 2: 2D inversion, *Geophysics,* **80**(3), E187–E196.
Nadipally, M., 2019. Optimization of methods for image-texture segmentation using ant colony optimization, in *Intelligent Data Analysis for Biomedical Applications,* Chapter 2, Intelligent Data-Centric Systems, pp. 21–47, eds, Hemanth, D.J., Gupta, D. & Emilia Balas, V., Academic Press.
Neal, R.M., 1996, *Bayesian Learning for Neural Networks,* Vol. **118,** Springer-Verlag New York.
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. & Meng, X.-L., Chapman and Hall/CRC.
Paciorek, C.J. & Schervish, M.J., 2004. Nonstationary covariance functions for Gaussian process regression, in *Advances in Neural Information Processing Systems*, MIT Press.
Pasquale, G.D. & Linde, N., 2016. On structure-based priors in Bayesian geophysical inversion, *Geophys. J. Int.,* **208** (3), 1342–1358.
Plagemann, C., Kersting, K. & Burgard, W., 2008. 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), 204–219.
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., 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., Kaplan, S., Washbourne, J. & Albertin, U., 2018. Low frequency full waveform seismic inversion within a tree based Bayesian framework, *Geophys. J. Int.,* **212**(1), 522–542.
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., 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,* Vol. 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**(3), 1847–1860.
Ray, A. & Myer, D., 2019. Bayesian geophysical inversion with trans-dimensional Gaussian Process machine learning, *Geophys. J. Int.,* **217,** 1706–1726.
Ray, A., Sekar, A., Hoversten, G. & 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.
Roininen, L., Girolami, M., Lasanen, S. & Markkanen, M., 2019. Hyperpriors for Matérn fields with applications in Bayesian inversion, *Inverse Problems Imag.,* **13**(1), 1–29.
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.
Sampson, P.D. & Guttorp, P., 1992. Nonparametric estimation of nonstationary spatial covariance structure, *J. Am. Stat. Assoc.,* **87**(417), 108–119.
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.
Snelson, E. & Ghahramani, Z., 2005. Sparse Gaussian processes using pseudo-inputs, in *Advances in Neural Information Processing Systems,* Vol. 18, pp. 1257–1264, MIT Press.
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, MIT Press.
Swendsen, R.H. & Wang, J.S., 1987. Nonuniversal critical dynamics in Monte Carlo simulations, *Phys. Rev. Lett.,* **58**(2), 86–88.
Tarantola, A. & Valette, B., 1982. Inverse problems= quest for information, *J. Geophys.,* **50,** 159–170.
Valentine, A.P. & Sambridge, M., 2020a. Gaussian process models-II. Lessons for discrete inversion, *Geophys. J. Int.,* **220**(3), 1648–1656.
Valentine, A.P. & Sambridge, M., 2020b. Gaussian process models-I. A framework for probabilistic continuous inverse theory, *Geophys. J. Int.,* **220**(3), 1632–1647.
Vehtari, A., Gelman, A. & Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Stat. Comput.,* **27**(5), 1413–1432.
Wallin, J. & Vadlamani, S., 2018. Infinite dimensional adaptive MCMC for Gaussian processes, *arXiv,* **1**(1), 1–23.
Williams, C.K.I. & Rasmussen, C.E., 1996. Gaussian processes for regression, in *Advances in Neural Information Processing Systems,* Vol. 8, pp. 514–520, eds Touretzky, D.S., M.C. Mozer & M.E. Hasselmo, MIT Press.
Yang, Z. & Rodriguez, C.E., 2013. Searching for efficient Markov chain Monte Carlo proposal kernels, *Proc. Natl. Acad. Sci.,* **110**(48), 19 307–19 312.
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-D Monte Carlo surface wave tomography, *Geophys. J. Int.,* **215**(3), 1644–1658.

APPENDIX A: MCMC MOVES AND THEIR ACCEPTANCE PROBABILITY

The 'birth-death' McMC method (pseudocode provided in Algorithm 1) has been used in this work. At each step, the length k of elements within the model vector $\boldsymbol{\theta}_m$ either increase by 1 (birth of a GP training point), decrease by 1 (death of a GP training point), or remain 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 (12) is

$$\left[\frac{p(\boldsymbol{\theta}_m')}{p(\boldsymbol{\theta}_m)}\right]_{\text{birth}} = \frac{1}{\delta}\frac{k+1}{\prod_{i=1}^{n_D}\Delta x_i}\frac{p(k+1)}{p(k)}, \tag{A1}$$

where the last fraction is unity for a uniform prior on k. For a birth move, a GP training location is proposed in the region $\prod_{i=1}^{n_D}\Delta x_i$ uniformly at random, and is assigned a value uniformly in $\Delta\rho$. Hence the proposal $q(\boldsymbol{\theta}_m'|\boldsymbol{\theta}_m)$ can be written as

$$\left[q(\boldsymbol{\theta}_m'|\boldsymbol{\theta}_m)\right]_{\text{birth}} = \frac{1}{\prod_{i=1}^{n_D}\Delta x_i}\frac{1}{\delta}, \tag{A2}$$

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

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

Thus the birth proposal ratio is

$$\left[\frac{q(\boldsymbol{\theta}_m|\boldsymbol{\theta}_m')}{q(\boldsymbol{\theta}_m'|\boldsymbol{\theta}_m)}\right]_{\text{birth}} = \frac{\delta\prod_{i=1}^{n_D}\Delta x_i}{k+1}. \tag{A4}$$

Thus, from (12), (A1) and (A4)

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

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 (12) is

$$\left[\frac{p(\boldsymbol{\theta}_m')}{p(\boldsymbol{\theta}_m)}\right]_{\text{death}} = \frac{\delta\prod_{i=1}^{n_D}\Delta x_i}{k}\frac{p(k-1)}{p(k)}, \tag{A6}$$

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

$$\left[q(\boldsymbol{\theta}_m'|\boldsymbol{\theta}_m)\right]_{\text{death}} = \frac{1}{k}. \tag{A7}$$

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 δ, or

$$\left[q(\boldsymbol{\theta}_m|\boldsymbol{\theta}_m')\right]_{\text{death}} = \frac{1}{\prod_{i=1}^{n_D}\Delta x_i}\frac{1}{\delta}. \tag{A8}$$

Thus the death proposal ratio is

$$\left[\frac{q(\boldsymbol{\theta}_m|\boldsymbol{\theta}_m')}{q(\boldsymbol{\theta}_m'|\boldsymbol{\theta}_m)}\right]_{\text{birth}} = \frac{k}{\delta\prod_{i=1}^{n_D}\Delta x_i}. \tag{A9}$$

Thus, from (12), (A6) and (A9)

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

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 2016, for details on reflection). The acceptance probability (14) is then simply the ratio of model likelihoods:

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

Other proposals such as through a Crank–Nicolson step may be better suited to fixed dimension moves in high dimensions (Cotter *et al.* 2013; Wallin & Vadlamani 2018). It should be noted that for a uniform prior over k, then in all cases, whether birth, death or fixed k,

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

A uniform prior over k has been used in this work.

A4 Practical considerations for efficient computation of $\boldsymbol{\mu}_{*\mathbf{s}}$ and $\boldsymbol{\mu}_{*\mathbf{ns}}$

From (A12), it is apparent that likelihood computation at proposed model $\boldsymbol{\theta}'_{\boldsymbol{m}}$ needs to be made to compute the acceptance probability of jumping from $\boldsymbol{\theta}_{\boldsymbol{m}}$. For both the GP mean models $\boldsymbol{\mu}_{*\text{s}}$ and $\boldsymbol{\mu}_{*\text{ns}}$, the size of the kernel matrices $\mathbf{K}_m$ and $\mathbf{K}_*$ in (1) are $n_{\text{train}} \times n_{\text{train}}$ and $n_{\text{test}} \times n_{\text{train}}$, respectively. At a maximum, n_{train} can take the value of max (k), specified in the Trans-D prior. Typically, even for 2-D problems (e.g. Blatter 2020), this has never been greater than 200. However, n_{test} corresponds to the size of the modelling domain vector—and this can be in the range of 10^4–10^5 cells. This number depends on the specifics of the subsurface structure, the earth response of which is to be computed using a forward modelling engine (e.g. Key & Ovall 2011, for 2-D MT responses from resistivity). Since creation of such large matrices at every McMC step is expensive, from the initialization of the algorithm, these need to be pre-allocated to their maximum possible size. Further, calculation of the kernels in (4) and (7) requires the computation of distances in (3) and the evaluation of an exponential in (2). This is quite computationally expensive, even if the kernel matrices are pre-allocated to their maximum possible size. By identifying and updating only those parts of the kernel matrices which change in a given McMC move, the algorithm can proceed efficiently, as detailed underneath. A possible alternative could be to use the stochastic partial differential equation (SPDE) approach (Lindgren *et al.* 2011), but the approach presented underneath appears fit-for-purpose.

A4.1 Changes in the properties model $\boldsymbol{\theta}_{ns}$

This set of changes is the most straightforward. In all cases, $\mathbf{K}_{*m}$ is symmetric and computations only need to be performed on a column and copied to the corresponding row of $\mathbf{K}_{*m}$.

(i) Birth: When birth of a GP nucleus occurs, only the new corresponding column of $\mathbf{K}_*$ and the corresponding new row and column of $\mathbf{K}_{*m}$ need to have kernels re-evaluated before computing (1) for $\boldsymbol{\mu}_{*\text{ns}}$.

(ii) Death: When a GP nucleus is removed, the corresponding column of $\mathbf{K}_*$ and the corresponding row and column of $\mathbf{K}_{*m}$ are simply omitted before computing (1) for $\boldsymbol{\mu}_{*\text{ns}}$.

(iii) Change nucleus position: Only the column of $\mathbf{K}_*$ corresponding to the moved nucleus and the corresponding row and column of $\mathbf{K}_{*m}$ need to have kernels re-evaluated before computing (1) to obtain $\boldsymbol{\mu}_{*\text{ns}}$.

(iv) Property change: As no distances are perturbed, no kernels are recomputed and (1) can be recomputed using the new properties vector $\mathbf{m}_k$ for $\boldsymbol{\mu}_{*\text{ns}}$.

A4.2 Changes in the length scales model $\boldsymbol{\theta}_s$

This set of changes is slightly more involved as it involves making some approximations, but they are conceptually straightforward. As earlier, $\mathbf{K}_{*m}$ is symmetric and computations only need to be performed on a column and copied to the corresponding row of $\mathbf{K}_{*m}$.

(i) Birth: When birth of a GP nucleus occurs, only the new corresponding column of $\mathbf{K}_*$ and the corresponding new row and column of $\mathbf{K}_{*m}$ need to have kernels re-evaluated before computing (1) to obtain $\boldsymbol{\mu}_{*\text{s}}$. The location of the birthed GP nucleus is noted.

(ii) Death: When a GP nucleus is removed, the corresponding column of $\mathbf{K}_*$ and the corresponding row and column of $\mathbf{K}_{*m}$ are simply omitted before computing (1) to obtain $\boldsymbol{\mu}_{*\text{s}}$. The location of the removed GP nucleus location is noted.

(iii) Change nucleus position: Only the column of $\mathbf{K}_*$ corresponding to the moved nucleus and the corresponding row and column of $\mathbf{K}_{*m}$ need to have kernels re-evaluated before computing (1) for $\boldsymbol{\mu}_{*\text{s}}$. The old location of the GP nucleus as well as its new location are noted.

(iv) Property change: As no distances are perturbed, no kernels are recomputed and (1) can be recomputed using (1) and the new properties vector $\mathbf{m}_k$ to obtain $\boldsymbol{\mu}_{*\text{s}}$. The location of the updated GP nucleus is noted.

Table A1. Timing for a birth or death move in successive birth and death for non-stationary mode TDGP as well as stationary mode TDGP, on a single thread of an Intel i7 (2013) 2.7 GHz machine. All McMC models have Trans-D count 150 and the dimension of the associated $\boldsymbol{\mu}_*$ is $n_{\text{test}} = 100 \times 200$. This shows that for large 2-D or 3-D problems, the bulk of the computation time will be taken by the forward modelling engine, not TDGP. For smaller 1-D problems, n_{test} is much smaller, in the order of 100s of cells and computation times are near-linearly reduced.

Type of TDGP McMC model	Time (ms)
Non stationary $\boldsymbol{\theta}_{\text{ns}}$	3.675 ± 0.006
Non stationary $\boldsymbol{\theta}_{\text{s}}$	35.684 ± 0.200
Purely stationary $\boldsymbol{\theta}$	4.136 ± 0.017

Since $\boldsymbol{\theta}_{\text{s}}$ is used to compute $\boldsymbol{\mu}_{*\text{s}}$, which provides the length scales which in turn influence $\boldsymbol{\mu}_{*\text{ns}}$ (see Fig. 2), the non-stationary kernels for computation of $\boldsymbol{\mu}_{*\text{ns}}$ need to be recomputed for acceptance of an McMC move on $\boldsymbol{\theta}_{\text{s}}$. For models with more than a few hundred parameters, recomputing the full non-stationary kernels is prohibitively expensive. However, the fact that the fixed, stationary length scale λ_{s} used to represent changes in the properties length scale $\boldsymbol{\mu}_{*\text{s}}$ is quite small, can be exploited here. This small length scale is necessary for $\boldsymbol{\mu}_{*\text{s}}$ to parametrize sharp changes in $\boldsymbol{\mu}_{*\text{ns}}$ through the non-stationary properties McMC model $\boldsymbol{\theta}_{\text{ns}}$. Keeping in mind the small length scale λ_{s}, only those parts of $\boldsymbol{\mu}_{\text{s}}$ in a small influence region $\delta_l \propto \lambda_{\text{s}}$ around the locations noted in the bulleted list immediately above are important to propagate into the non-stationary kernels. The aforementioned number of changes and locations of note from within the *stationary* model $\boldsymbol{\theta}_{\text{s}}$ are denoted as $\mathbf{x}_{\text{note}_i}, i = 1, \ldots, n_{\text{changes}}$, where $\max(n_{\text{changes}}) = 2$ for the position change move. Keeping in mind that $\mathbf{x}_*$ or the test locations are the same for both $\boldsymbol{\mu}_{*\text{s}}$ and $\boldsymbol{\mu}_{*\text{ns}}$, changes fall into three categories for the *non-stationary kernels* and Trans-D training points contained within $\boldsymbol{\theta}_{*\text{ns}}$:

(i) Changes in rows of $\mathbf{K}_*$ for the forward modelling test locations corresponding to $|\mathbf{x}_* - \mathbf{x}_{\text{note}_i}| < \delta_l, i = 1, \ldots, n_{\text{changes}}$
(ii) Changes in columns of $\mathbf{K}_*$ corresponding to locations of training points within $|\mathbf{x}_{\text{train}} - \mathbf{x}_{\text{note}_i}| < \delta_l, i = 1, \ldots, n_{\text{changes}}$
(iii) Changes in rows and columns of $\mathbf{K}_m$ where $|\mathbf{x}_{\text{train}} - \mathbf{x}_{\text{note}_i}| < \delta_l, i = 1, \ldots, n_{\text{changes}}$

For the examples presented in this work, it is stated without proof that $\delta_l \geq 3.6\lambda_{\text{s}}$ seems to work without making the approximations introduced inaccurate. It must be noted that the McMC model $\boldsymbol{\theta}_{*\text{ns}}$ is not perturbed by these updates to the non-stationary kernels—however, $\boldsymbol{\mu}_{*\text{ns}}$ is changed and this in turn affects the likelihood, which leads to the acceptance or rejection of McMC proposals for $\boldsymbol{\theta}_{*\text{s}}$. Extensive use has been made of *k–d* trees in searching for nearest neighbours [https://github.com/KristofferC/NearestNeighbors.jl, Carlsson *et al.* (2019)] in effecting these approximations or finding the length scale at a particular training point location. Representative timings for the size of n_{test} similar to that used in Blatter (2020) for 2-D MT inversion with stationary mode TDGP are provided in Table A1.

A5 Parallel tempering step

To facilitate the escape of local misfit minima, or equivalently, the navigation of peaky likelihoods, parallel tempering is used to exchange information between McMC chains running in parallel. Temperatures or models are exchanged 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):

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

For a description of why swapping models is effective using (A13) see section 3.2 of Blatter *et al.* (2018). For computational efficiency, temperatures are exchanged during interprocess communication to achieve the exact same effect as swapping models. The entire algorithm is summarized by the pseudocode in Algorithm 1:

All Markov chains were run at log-spaced temperatures between 1 and 2.5. Details of setting a temperature ladder can be found in Dettmer *et al.* (2012) and Ray *et al.* (2013a). The 1-D non-linear regression problem used four temperatures, whereas the image and real data examples used eight temperatures. Larger numbers of temperatures are required to sample rugged likelihoods. Posterior inference is carried out only from models that are at $T = 1$.

APPENDIX B: MAXIMUM LIKELIHOOD DATA ERROR

The model likelihood given in (9), 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, for n_f independent frequencies and n_r receivers at frequency l, the model likelihood

Initialise chains with stationary McMC length scale model $\boldsymbol{\theta}_{\mathrm{s}_j}$ and non-stationary McMC properties model $\boldsymbol{\theta}_{\mathrm{ns}_j}$ 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 $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$
    $\boldsymbol{\theta}_{\mathrm{s}_j}[i] \leftarrow \boldsymbol{\theta}_{\mathrm{s}_j}[i-1]$
    $\boldsymbol{\theta}'_{\mathrm{s}_j} \sim q_{type}(\boldsymbol{\theta}'_{\mathrm{s}_j}|\boldsymbol{\theta}_{\mathrm{s}_j}[i])$
    $u \sim U(0,1)$
    if $u < \alpha_j^{type}(\boldsymbol{\theta}'_{\mathrm{s}_j}|\boldsymbol{\theta}_{\mathrm{s}_j}[i])$ and $p(\boldsymbol{\theta}'_{\mathrm{s}_j}) > 0$ **then**
      $\boldsymbol{\theta}_{\mathrm{s}_j}[i] \leftarrow \boldsymbol{\theta}'_{\mathrm{s}_j}$
    end
    Select $type$ from [**birth, death, fixed**] with probability $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$
    $\boldsymbol{\theta}_{\mathrm{ns}_j}[i] \leftarrow \boldsymbol{\theta}_{\mathrm{ns}_j}[i-1]$
    $\boldsymbol{\theta}'_{\mathrm{ns}_j} \sim q_{type}(\boldsymbol{\theta}'_{\mathrm{ns}_j}|\boldsymbol{\theta}_{\mathrm{ns}_j}[i])$
    $u \sim U(0,1)$
    if $u < \alpha_j^{type}(\boldsymbol{\theta}'_{ns_j}|\boldsymbol{\theta}_{\mathrm{ns}_j}[i])$ and $p(\boldsymbol{\theta}'_{ns_j}) > 0$ **then**
      $\boldsymbol{\theta}_{\mathrm{ns}_j}[i] \leftarrow \boldsymbol{\theta}'_{\mathrm{ns}_j}$
    end
  end
  for $p \leftarrow nTemps$ **to** *2* **do**
    $q \sim U(1, p); q \in \mathbb{I}$
    if $p \neq q$ **then**
      if $u < \alpha^{swap}(p, q)$ **then**
        swap T_p and T_q
      end
    end
  end
end

Algorithm 1: Pseudocode for non-stationary mode McMC with TDGP, and parallel tempering exchanges though a Fisher-Yates shuffle (Fisher & Yates 1938). Forward computation to evaluate $\alpha_{swap}(p, q)$ is not required as likelihoods for models with temperatures p and q have already been computed in the preceding j loop. Inference is carried out from the sorted chain (or chains) with $T = 1$ after an initial 'burn-in' number of samples.

can be written as:

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

where the term in the exponential is $\frac{1}{2}$ the χ^2 misfit as the complex data variance at any receiver in covariance $\mathbf{C}_{d_l}$ is twice that of either the real or imaginary parts. If it assumed that the data error at frequency l is known up to a proportionality constant within the symmetric positive definite matrix $\mathbf{C}_l$, it can be written that

$$\mathbf{C}_{d_l} = \sigma_l^2 \mathbf{C}_l, \tag{B2}$$

where σ_l^2 is an unknown constant scaling at the lth frequency. Eq. (B1) can thus be written as

$$\mathcal{L}(\boldsymbol{\theta_m}) = \prod_{l=1}^{n_f} \frac{1}{(\pi\sigma_l^2)^{n_r}|\mathbf{C}_l|} \exp\left(-\frac{1}{\sigma_l^2}[\mathbf{f}_l(\boldsymbol{\theta_m}) - \mathbf{d}_l]^\dagger \mathbf{C}_l^{-1}[\mathbf{f}_l(\boldsymbol{\theta_m}) - \mathbf{d}_l]\right). \tag{B3}$$

To find the maximum of the likelihood (B3), the negative of the log of the likelihood (i.e. the misfit objective function) is minimized. The approach starts by taking log as follows:

$$-\log \mathcal{L}(\boldsymbol{\theta_m}) = \sum_{l=1}^{n_f} \log(\pi^{n_r}|\mathbf{C}_l|) + 2n_r \log\sigma_l + \left(\frac{1}{\sigma_l^2}[\mathbf{f}_l(\boldsymbol{\theta_m}) - \mathbf{d}_l]^\dagger \mathbf{C}_l^{-1}[\mathbf{f}_l(\boldsymbol{\theta_m}) - \mathbf{d}_l]\right), \tag{B4}$$

writing the data residual $\mathbf{r}$ at frequency l as

$$\mathbf{r}_l = \mathbf{f}_l(\boldsymbol{\theta_m}) - \mathbf{d}_l, \tag{B5}$$

the negative log likelihood is more compactly written as

$$-\log \mathcal{L}(\boldsymbol{\theta_m}) = \sum_{l=1}^{n_f} \log(\pi^{n_r}|\mathbf{C}_l|) + 2n_r \log \sigma_l + \frac{1}{\sigma_l^2}\mathbf{r}_l^\dagger \mathbf{C}_l^{-1}\mathbf{r}_l. \tag{B6}$$

Deriving the negative log likelihood with respect to σ_l and setting equal to zero:

$$\frac{2n_r}{\sigma_l} - \frac{2}{\sigma_l^3}\mathbf{r}_l^\dagger \mathbf{C}_l^{-1}\mathbf{r}_l = 0, \tag{B7}$$

$$\Rightarrow \sigma_l^2 = \frac{1}{n_r}\mathbf{r}_l^\dagger \mathbf{C}_l^{-1}\mathbf{r}_l. \tag{B8}$$

At this point, the similarity of (B8) with eq. (B.5) of Sambridge (2013) should be noted. The latter approach is in the time domain, though this analysis is in frequency. Further, the formulation here differs from Ray & Myer (2019), only in that $\mathbf{C}_l$ is no more a diagonal matrix with the square of the data amplitude at every receiver. By substituting (B8) in (B6) the following is obtained:

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

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

While sampling the posterior models in the McMC chain, the negative log likelihood given by (B10) is used, instead of computing the misfit with unreliable, fixed, data error. The scaled data errors at each frequency are implicitly sampled as a function of the current McMC sample $\boldsymbol{\theta_m}$, thus avoiding the addition of yet another McMC sampling unknown.

APPENDIX C: PEAK SIGNAL-TO-NOISE RATIO

A commonly used metric to measure the quality of an image reconstruction after compression (e.g. Nadipally 2019) is the peak signal-to-noise ratio (PSNR). It is defined in dB (Horé & Ziou 2010) as

$$\text{PSNR} = \log_{10}\left(\frac{R^2}{\frac{1}{n}\mathbf{r}^t\mathbf{r}}\right), \tag{C1}$$

where R is the maximum allowed fluctuation in the pixel values, and the denominator on the right hand side is the mean square error (MSE) across all n pixel values. To be explicit, $\mathbf{r} = \mathbf{f}_{\text{true}} - \mathbf{f}_{\text{reconstructed}}$, where $\mathbf{f} \in \mathbb{R}^n$ represents the vector of true or reconstructed pixel values. Higher values represent better fidelity, and if the reconstruction is perfect, the PSNR is infinite.