Geophysical Journal International

Geophys. J. Int. (2024) **239,** 1236–1256
Advance Access publication 2024 September 11
GJI Seismology

https://doi.org/10.1093/gji/ggae334

Variational prior replacement in Bayesian inference and inversion

Xuebin Zhao and Andrew Curtis

School of GeoSciences, University of Edinburgh, Edinburgh EH8 9XP, United Kingdom. E-mail: xuebin.zhao@ed.ac.uk

Accepted 2024 September 10. Received 2024 September 9; in original form 2024 June 6

SUMMARY

Many scientific investigations require that the values of a set of model parameters are estimated using recorded data. In Bayesian inference, information from both observed data and prior knowledge is combined to update model parameters probabilistically by calculating the posterior probability distribution function. Prior information is often described by a prior probability distribution. Situations arise in which we wish to change prior information during the course of a scientific project. However, estimating the solution to any single Bayesian inference problem is often computationally costly, as it typically requires many model samples to be drawn, and the data set that would have been recorded if each sample was true must be simulated. Recalculating the Bayesian inference solution every time prior information changes can therefore be extremely expensive. We develop a mathematical formulation that allows the prior information that is embedded within a solution, to be changed using variational methods, without recalculating the original Bayesian inference. In this method, existing prior information is removed from a previously obtained posterior distribution and is replaced by new prior information. We therefore call the methodology *variational prior replacement* (VPR). We demonstrate VPR using a 2-D seismic full waveform inversion example, in which VPR provides similar posterior solutions to those obtained by solving independent inference problems using different prior distributions. The former can be completed within minutes on a laptop computer, whereas the latter requires days of computations using high-performance computing resources. We demonstrate the value of the method by comparing the posterior solutions obtained using three different types of prior information: uniform, smoothing and geological prior distributions.

Key words: Bayesian inference; Inverse theory; Probability distributions; Seismic tomography; Waveform inversion.

1 INTRODUCTION

In a wide variety of scientific and engineering applications, researchers seek to estimate unknown (latent) parameters using observed data by solving an inverse or inference problem. By approximating the physical system, it is usually possible to calculate a forward function which estimates the synthetic data that would have been observed if any particular set of model parameter values were true, and this function commonly has unique values. However, direct inversion of this function is difficult if not impossible, due to uncertainties in the finite number of measurable data, and to the non-linearity of the forward function (Boyd & Vandenberghe 2004). Typically in practise, solutions to such inverse problems are non-unique (Mosegaard & Tarantola 1995; Mosegaard & Sambridge 2002; Tarantola 2005; Valentine & Sambridge 2023).

Bayesian inference solves fully non-linear, non-unique inverse problems under a probabilistic framework by seeking to define the family of all plausible solutions within model parameter space. The solution to Bayesian inference is described by the so-called posterior probability distribution function (*pdf* – either a probability density function for continuous variables, or a set of probabilities for discrete variables), obtained by updating prior information about model parameters with new information from observed data. It provides a statistical description of how consistent is each solution with both the data and prior information, and allows uncertainties in the inverse problem solution to be estimated (Tarantola 2005; Arnold & Curtis 2018).

Global search or sampling based methods are often used to solve Bayesian inference problems. Samples of parameter values with non-zero posterior probability values are retained, sometimes in proportion to their probabilities, to build an ensemble of model solutions. These solutions are used to estimate statistical properties of the posterior distribution that characterize the solution uncertainty. Monte Carlo is one of the most frequently used

© The Author(s) 2024. Published by Oxford University Press on behalf of The Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

sampling methods, including Metropolis-Hastings Markov chain Monte Carlo (MH-McMC–Metropolis *et al.* 1953; Hastings 1970; Press 1968; Mosegaard & Tarantola 1995; Sambridge & Mosegaard 2002), trans-dimensional Monte Carlo (Green 1995; Malinverno 2002; Sambridge *et al.* 2006; Bodin & Sambridge 2009; Galetti *et al.* 2017), gradient-based Monte Carlo methods (Welling & Teh 2011; Girolami & Calderhead 2011; Fichtner *et al.* 2019; Gebraad *et al.* 2020; Zhao & Sen 2021; Biswas & Sen 2022; de Lima *et al.* 2023a; Berti *et al.* 2023) and informed-proposal Monte Carlo (Khoshkholgh *et al.* 2021, 2022). Global search methods that do not use Markov chains have been developed to solve Bayesian problems, either using optimization to search for the most probable solution such as in simulated annealing (Kirkpatrick *et al.* 1983; Sen & Stoffa 2013; Zhao *et al.* 2022b) and genetic algorithms (Stoffa & Sen 1991; Sambridge & Drijkoningen 1992), or algorithms that characterize the posterior distribution such as the neighbourhood algorithm (Sambridge 1999a, b), prior sampling (Meier *et al.* 2007a, b; Käufl *et al.* 2016; Mosser *et al.* 2020; Bloem *et al.* 2024), exact sampling (Propp & Wilson 1996; Walker & Curtis 2014a) and direct estimation of posterior pdfs without sampling (Nawaz & Curtis 2016). However, all such sampling methods present deficiencies when applied to complex or high dimensional inference problems such as slow convergence (Atchadé & Rosenthal 2005; Andrieu & Thoms 2008), which implies that a large number of model samples and their corresponding forward simulations are required.

Variational inference provides an alternative to random sampling methods to solve Bayesian problems. Variational methods select one optimal approximation to the true posterior pdf from a predefined family of known and computationally tractable probability distributions (referred to as the variational family). This is accomplished by minimizing the difference between the posterior and variational pdfs (Bishop 2006; Blei *et al.* 2017; Zhang *et al.* 2021), thus solving Bayesian problems using optimization rather than stochastic sampling. This approach can often be computationally efficient, ease the detection of convergence, and scale well to high dimensional inference problems with large data sets, while still producing a valid probability distribution.

In geophysics, variational inference was first applied to estimate subsurface geological facies and petrophysical parameters using seismic data (Nawaz & Curtis 2018, 2019; Nawaz *et al.* 2020), where a mean-field approximation (which ignores correlations between parameters) is used to simplify the mathematical formulation of the variational problem (Bishop 2006; Kucukelbir *et al.* 2017). Since then, more advanced variational methods have been developed for different geophysical problems, such as travel time tomography (Zhang & Curtis 2020a; Zhao *et al.* 2021; Levy *et al.* 2022), seismic migration (Siahkoohi *et al.* 2021, 2023), seismic amplitude inversion (Zidan *et al.* 2022), earthquake hypocentre inversion (Smith *et al.* 2022), slip distribution inversion (Sun *et al.* 2023), full waveform inversion (Zhang & Curtis 2021b; Bates *et al.* 2022; Wang *et al.* 2023; Lomas *et al.* 2023; Izzatullah *et al.* 2024; Zhao & Curtis 2024b; Yin *et al.* 2024a) and experimental design (Strutz & Curtis 2024).

Most studies mentioned above, whether using random sampling or variational methods, focus on performing Bayesian inference efficiently and accurately given a specific set of observed data and fixed prior knowledge. Over recent years, researchers made use of neural networks and other machine learning architectures to implement efficient Bayesian inference in which the posterior pdf can be obtained rapidly for any newly observed data set (Devilee *et al.* 1999; Meier *et al.* 2007a, b; Shahraeeni & Curtis 2011; Shahraeeni *et al.* 2012; de Wit *et al.* 2013; Käufl *et al.* 2014, 2016; Earp & Curtis 2020; Earp *et al.* 2020; Zhang & Curtis 2021a; Mosher *et al.* 2021; Wang *et al.* 2022; Hansen & Finlay 2022; Alyaev & Elsheikh 2022; Grana *et al.* 2022; Guan *et al.* 2024; Sun & Williamson 2024). However, almost no publications consider situations where we want to change (update) the prior information used in a previously performed inference process, or where we have multiple plausible prior hypotheses to be tested for the same observed data. In such cases one might have to perform the inference repeatedly with different prior distributions; this would become extremely expensive in many applications, even though the data used in each individual case do not change.

Walker & Curtis (2014b) introduced a method called prior replacement, which allows prior information to be changed rapidly in Bayesian inference without repeating the full inference procedure for each individual prior pdf (on occasion below, we may refer to this simply as the prior). This is achieved by dividing the obtained posterior distribution by the current prior pdf in an attempt to remove the effect of the latter, and then multiplying the result by a new prior pdf to inject (update) the results with new prior information. The method was demonstrated to be effective for varying prior information when estimating rock physics parameters (clay volume and sandstone matrix porosity) using seismic impedance data. Walker & Curtis (2014b) used (semi-)analytic methods to perform prior replacement, which requires the calculation of integrals of probability distributions over the entire parameter space, making it difficult to calculate for high dimensional problems. Moreover, the analytic calculation is only applicable under stringent conditions: first, the existing posterior distribution is represented by a mixture of Gaussian distributions. And second, the old and new prior distributions should be uniform, Gaussian, or (possibly) other probability distributions whose integral over the parameter space and whose multiplication and division by Gaussian distributions are analytically tractable, such that the replacement of the new posterior distribution can be calculated analytically.

In this paper, we develop a prior replacement methodology under the framework of variational inference, hence the name *variational prior replacement* (VPR). VPR addresses (relaxes) the issues mentioned above, making it applicable for high dimensional and complicated Bayesian inference problems. We test and demonstrate the method on a full waveform inversion (FWI) problem. Until the last few years there was no published fully non-linear Bayesian solution to any FWI problem that approached a practical scale, due to the huge computational cost, and associated theoretical and algorithmic challenges (Gebraad *et al.* 2020; Zhang & Curtis 2020b). While to-date studies have extended the method to three dimensional cases (Zhang *et al.* 2023; Zhao & Curtis 2024c), the computational issues to make this approach mainstream remain. Advances that deliver even approximate results using greatly reduced computation are therefore significant.

The rest of this paper is organized as follows. In Section 2, we review the Bayesian inference and prior replacement concept developed in Walker & Curtis (2014b). To perform prior replacement more efficiently, we introduce variational inference and derive the variational prior replacement (VPR) framework. In Section 3, we demonstrate the method using a seismic full waveform inversion example, in which we compare the results obtained using VPR with those found using independent Bayesian inference for each prior pdf. We demonstrate the effectiveness of the method by testing three different prior distributions using the same observed data. Finally, we provide a brief discussion and draw conclusions from this study.

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

2 METHODOLOGY

2.1 Bayesian inference

In Bayesian inference, inverse problems are solved under a probabilistic framework by calculating the so-called *posterior* probability distribution function (pdf) of model vector $\mathbf{m}$ given observed data $\mathbf{d}_{obs}$ using Bayes' rule:

$$p(\mathbf{m}|\mathbf{d}_{obs}) = \frac{p(\mathbf{d}_{obs}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d}_{obs})} \tag{1}$$

where $p(\mathbf{m})$ is the *prior* pdf that describes available information about model parameter $\mathbf{m}$ before inference process, and $p(\mathbf{d}_{obs}|\mathbf{m})$ is the *likelihood* function, which calculates the probability of observing data $\mathbf{d}_{obs}$ given any model value $\mathbf{m}$. The likelihood is used to describe how well $\mathbf{d}_{obs}$ matches synthetic data generated by a particular model $\mathbf{m}$. Term

$$p(\mathbf{d}_{obs}) = \int_{\mathbf{m}} p(\mathbf{d}_{obs}|\mathbf{m})p(\mathbf{m})d\mathbf{m} \tag{2}$$

is a normalization constant called the *evidence*. It ensures that the right hand side of eq. (1) is a valid probability distribution. Bayesian inference combines information from both data and prior knowledge in a probabilistic manner, and the resulting posterior distribution describes all possible model solutions that fit the data and the prior.

2.2 Prior replacement

Consider a situation where we have different sets of prior information (e.g., different hypotheses, or differing beliefs held by different people) about model parameter $\mathbf{m}$, defined by prior probability distributions $p_1(\mathbf{m})$, $p_2(\mathbf{m})$, and so on, and we wish to evaluate the implications of these various priors by calculating the corresponding posterior distributions. Such an array of prior distributions might originate from the views of different groups of experts, or might invoke different assumptions about the structures and properties that might pertain to the model, perhaps representing a range of different hypotheses to be tested and discriminated (e.g., Bloem *et al.* 2024). A straightforward strategy is to apply Bayes' rule to each prior distribution, and solve independent (prior specific) Bayesian inverse problem whenever prior information changes. However, such a prior specific approach is not practically feasible, since it is already expensive to perform a single inference process, especially for high dimensional problems with large data sets such as typically occurs in seismic tomography and FWI problems.

Alternatively, suppose we have obtained a posterior distribution for data $\mathbf{d}_{obs}$ based on one specific type of prior information. When additional prior information becomes available or when different prior hypotheses exist and need to be discriminated, we might remove the effect of the existing prior information from the current posterior distribution, then inject different prior information. Thus we would obtain the desired posterior pdf without explicitly applying Bayes' rule a second time. Below we denote prior and posterior pdfs that have been considered or calculated previously as *old* pdfs and those obtained by updating the prior distribution in this way as *new* ones; the words *old* and *new* in this context do not refer to situations where we update information because additional data have been collected, but rather to the order in which different prior distributions are combined with information in a fixed data set.

Walker & Curtis (2014b) mathematically formulated the above idea as follows: the new posterior distribution $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ given the new prior distribution $p_{new}(\mathbf{m})$ can be calculated by

$$\begin{aligned} p_{new}(\mathbf{m}|\mathbf{d}_{obs}) &= \frac{p(\mathbf{d}_{obs}|\mathbf{m})p_{new}(\mathbf{m})}{p_{new}(\mathbf{d}_{obs})} \\ &= \frac{p(\mathbf{d}_{obs}|\mathbf{m})p_{old}(\mathbf{m})}{p_{old}(\mathbf{d}_{obs})} \frac{p_{new}(\mathbf{m})}{p_{old}(\mathbf{m})} \frac{p_{old}(\mathbf{d}_{obs})}{p_{new}(\mathbf{d}_{obs})} \end{aligned} \tag{3}$$

The first line is simply Bayes' rule. In both new and old distributions, we assume that the observed data are the same. In the second line, $p_{old}(\mathbf{d}_{obs})$ and $p_{new}(\mathbf{d}_{obs})$ are two constants which are independent of model vector $\mathbf{m}$ according to eq. (2), so we define $k = p_{old}(\mathbf{d}_{obs})/p_{new}(\mathbf{d}_{obs})$ for later convenience. Term $p_{new}(\mathbf{m})/p_{old}(\mathbf{m})$ essentially takes the role of changing (replacing) the old prior by the new prior pdf in Bayesian inference, and states how we inject new prior information. Denote

$$p_{old}(\mathbf{m}|\mathbf{d}_{obs}) = \frac{p(\mathbf{d}_{obs}|\mathbf{m})p_{old}(\mathbf{m})}{p_{old}(\mathbf{d}_{obs})} \tag{4}$$

as the old posterior distribution given the old prior $p_{old}(\mathbf{m})$. Then eq. (3) becomes

$$p_{new}(\mathbf{m}|\mathbf{d}_{obs}) = k\ p_{old}(\mathbf{m}|\mathbf{d}_{obs})\ \frac{p_{new}(\mathbf{m})}{p_{old}(\mathbf{m})} \tag{5}$$

Equation 5 has a form that allows us to evaluate the new posterior distribution from the old one by updating (replacing) prior information after Bayesian inference, assuming that we know both $p_{old}(\mathbf{m})$ and $p_{new}(\mathbf{m})$, and that we can evaluate the normalization constant k. Using this formulation, there is no need to perform new likelihood evaluations, which is normally the most computationally expensive step in solving an inverse problem, no matter how many different prior distributions we wish to inject. However, prior replacement is valid only under one condition: the new prior must have zero (or in practise, very small) probability where the old prior has zero probability values to avoid a numerically unstable situation of dividing by zero (Walker & Curtis 2014b). Intuitively, the support of the new prior pdf must be a subset of that of the old one.

Eqs. (4) and (5) define the two main operations involved in prior replacement. The former is the solution to a typical Bayesian problem in which the (old) posterior pdf is evaluated given the observed data and existing prior information. The latter can be viewed as a quasi-Bayesian problem in the sense that Bayes rule applied to the new prior distribution is implicit within the formula, and the new posterior pdf is obtained by combining information from three probability distributions–of similar form to Bayes' rule, but without the need to re-calculate the likelihood function from scratch. Calculation of these expressions can nevertheless be computationally expensive, so in the following two sections we introduce efficient methods for each of these operations, respectively.

2.3 Variational inference

The Bayesian posterior distribution in eq. (4) can be estimated using either random sampling or variational inference methods. Markov chain Monte Carlo (McMC) is a typical sampling method that generates an ensemble of samples distributed according to the posterior distribution as the number of samples tends to infinity (Mosegaard & Tarantola 1995). However, McMC can be expensive in practise since the required number of samples increases exponentially with the dimensionality of model vector $\mathbf{m}$ – a concept referred to as the curse of dimensionality (Curtis & Lomax 2001).

Variational inference is an alternative to McMC that can be more efficient in certain situations. In variational inference, we define a

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

family of probability distributions $\mathcal{Q}(\mathbf{m}) = \{q(\mathbf{m})\}$ with fixed (predefined) complexity, within which we select a member $q^*(\mathbf{m})$ that best approximates the unknown posterior distribution. Therefore, variational inference solves Bayesian problems using optimization rather than random sampling. The optimal distribution can be found by minimizing the discrepancy between the variational and posterior distributions.

The Kullback-Leibler (KL) divergence (Kullback & Leibler 1951) is often used to measure the distance between two distributions

$$\mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})] = \mathbb{E}_{q(\mathbf{m})}[\log q(\mathbf{m}) - \log p(\mathbf{m}|\mathbf{d}_{obs})] \tag{6}$$

where the expectation is taken with respect to the variational distribution $q(\mathbf{m})$. The KL divergence is non-negative and equals zero only when the two distributions are identical. Substituting Bayes' rule (eq. 1) into eq. (6), we have

$$\log p(\mathbf{d}_{obs}) = \mathbb{E}_{q(\mathbf{m})}[\log p(\mathbf{m}, \mathbf{d}_{obs})] - \mathbb{E}_{q(\mathbf{m})}[\log q(\mathbf{m})] + \mathrm{KL}[q(\mathbf{m}) ||p(\mathbf{m}|\mathbf{d}_{obs})] \tag{7}$$

Since $\log p(\mathbf{d}_{obs})$ is a constant and independent of $q(\mathbf{m})$, minimizing $\mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})]$ is equivalent to maximizing the first two terms on the right hand side of eq. (7). In addition, since $\mathrm{KL}[q||p] \geq 0$, these two terms together act as a lower bound on the logarithmic evidence, and they are usually defined as the *evidence lower bound* (ELBO) of $\log p(\mathbf{d}_{obs})$:

$$\mathrm{ELBO}[q(\mathbf{m})] = \mathbb{E}_{q(\mathbf{m})}[\log p(\mathbf{m}, \mathbf{d}_{obs}) - \log q(\mathbf{m})] \tag{8}$$

Evaluating $\mathrm{ELBO}[q(\mathbf{m})]$ is easier than evaluating $\mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})]$ since it does not explicitly require the evidence term $p(\mathbf{d}_{obs})$ to be calculated, which is often computationally intractable. The variational problem is therefore often solved by maximizing the $\mathrm{ELBO}[q(\mathbf{m})]$. The optimization result is a probability distribution $q^*(\mathbf{m})$ with

$$q^*(\mathbf{m}) = \underset{q \in \mathcal{Q}}{\operatorname{argmax}}\, \mathrm{ELBO}[q(\mathbf{m})] \tag{9}$$

which serves as the best approximation to $p(\mathbf{m}|\mathbf{d}_{obs})$ within $\mathcal{Q}(\mathbf{m})$.

In variational inference, there is a trade-off when choosing the variational family: it needs to be sufficiently expressive to provide an accurate approximation to the (potentially complex) posterior distribution, yet simple enough for efficient optimization. Different choices of the family often result in different variational distributions, and also in different algorithms.

2.4 Variational prior replacement

Solving the second (quasi) Bayesian problem in eq. (5) requires $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ to be evaluated. On the right hand side of eq. (5), while in most cases both old and new prior probability values can be calculated efficiently, the main challenge lies in computing $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$ for a new model $\mathbf{m}$ without invoking eq. (4) (Bayes' rule) which involves forward simulation of data corresponding to $\mathbf{m}$ in order to evaluate the likelihood. Otherwise prior replacement would reduce to solving multiple independent Bayesian inverse problems with different prior distributions.

We use variational inference to solve the first Bayesian problem described in eq. (4), providing a known and parametrized probability distribution (called the variational distribution according to the previous section) $q_{old}(\mathbf{m})$ that approximates the old posterior pdf:

$$q_{old}(\mathbf{m}) \approx\ p_{old}(\mathbf{m}|\mathbf{d}_{obs}) = \frac{p(\mathbf{d}_{obs}|\mathbf{m})p_{old}(\mathbf{m})}{p_{old}(\mathbf{d}_{obs})} \tag{10}$$

This distribution is found by solving a variational optimization problem described in eq. (9). Once we obtain $q_{old}(\mathbf{m})$, we can use it to replace $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$ in eq. (5). Since $q_{old}(\mathbf{m})$ is only an approximation (rather than exactly equal) to $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$, we end up with an approximate expression for $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$:

$$p_{new}(\mathbf{m}|\mathbf{d}_{obs}) = k\ p_{old}(\mathbf{m}|\mathbf{d}_{obs})\ \frac{p_{new}(\mathbf{m})}{p_{old}(\mathbf{m})} \approx k\ q_{old}(\mathbf{m})\ \frac{p_{new}(\mathbf{m})}{p_{old}(\mathbf{m})} \tag{11}$$

This means that to estimate $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ we do not need to evaluate $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$ and so avoid the calculation of likelihood function. Since forward simulation is the most computationally expensive component in an inverse problem, evaluating $q_{old}(\mathbf{m})$ would normally be far cheaper than evaluating $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$. Eq. (11) indicates that we can estimate $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ from $q_{old}(\mathbf{m})$, up to a normalization constant which can be absorbed into k.

Note that not all variational inference methods can provide a variational distribution ($q_{old}(\mathbf{m})$ in this case) whose probability value can be evaluated easily. For example, Stein variational gradient descent (SVGD–Liu & Wang 2016) and its stochastic version (sSVGD–Gallego & Insua 2018) iteratively update a set of samples (also called particles) such that they become distributed according to an approximation to the posterior distribution. The output is the optimized set of particles, which are used to estimate statistical properties of the posterior distribution. However, evaluating the probability value $q_{old}(\mathbf{m})$ for a particular $\mathbf{m}$ is not at all straightforward with these methods.

An alternative suite of variational methods approximates the posterior distribution as a known structure with given complexity (for example a Gaussian distribution), and can thus be expressed by a given parametric (often analytic) representation. Variational inference finds the optimal values of hyperparameters that control the parametric expression, thus defining a variational distribution that best approximates the true posterior pdf. Since we obtain a parametric (closed form) expression for the variational distribution, we can easily evaluate its probability value for any model $\mathbf{m}$. We refer to this kind of variational method as *parametric variational inference* (Sjölund 2023). Examples of typical parametric variational inference methods include automatic differentiation variational inference (ADVI–Kucukelbir *et al.* 2017), normalizing flows (Rezende & Mohamed 2015), boosting variational inference (BVI–Guo *et al.* 2016; Miller *et al.* 2017), and physically structured variational inference (PSVI–Zhao & Curtis 2024b).

If $q_{old}(\mathbf{m})$ is constructed using a parametric variational inference method then its probability value can be calculated easily, and eq. (11) can in principle be evaluated using any probabilistic inference method since the probability value $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ can be approximated efficiently. However, even though $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ can be evaluated efficiently, the curse of dimensionality may nevertheless make the problem expensive, if not impossible, to solve using Monte Carlo sampling methods. We therefore introduce a second variational distribution $q_{new}(\mathbf{m})$ to approximate the new posterior distribution $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ given the new prior information $p_{new}(\mathbf{m})$. This new variational distribution can be obtained by minimizing the KL-divergence between $q_{new}(\mathbf{m})$ and $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$:

$$\begin{aligned}&\mathrm{KL}[q_{new}(\mathbf{m})||p_{new}(\mathbf{m}|\mathbf{d}_{obs})] = \mathbb{E}_{q_{new}(\mathbf{m})}[\log q_{new}(\mathbf{m}) - \log p_{new}(\mathbf{m}|\mathbf{d}_{obs})],\\ &\approx \mathbb{E}_{q_{new}(\mathbf{m})}[\log q_{new}(\mathbf{m}) - \log q_{old}(\mathbf{m}) - \log p_{new}(\mathbf{m}) + \log p_{old}(\mathbf{m})] - \log k\end{aligned} \tag{12}$$

where the second line is obtained by substituting eq. (11) into the first line. Note that $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ is the exact distribution from eq. (11), which is then approximated by using the same approximation as in eq. (11) by introducing the approximate Bayesian solution

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

$q_{old}(\mathbf{m})$. The last term $\log k$ is a constant and can safely be ignored when minimizing $\mathrm{KL}[q_{new}(\mathbf{m})||p_{new}(\mathbf{m}|\mathbf{d}_{obs})]$. This optimization problem can be solved in exactly the same way as conventional variational problem, and the result satisfies $q_{new}(\mathbf{m}) \approx p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ obtained by solving

$$q^*_{new}(\mathbf{m}) = \underset{q\in Q}{\operatorname{argmin}}\, \mathrm{KL}[q_{new}(\mathbf{m})||p_{new}(\mathbf{m}|\mathbf{d}_{obs})] \tag{13}$$

Using the framework of variational inference, the two main operations in the original prior replacement problem described in eqs. (4) and (5) are converted into two variational problems to estimate $q_{old}(\mathbf{m})$ and $q_{new}(\mathbf{m})$, containing two approximate steps. We therefore call this new methodology *variational prior replacement* (VPR).

Note that eq. (10) illustrates that the variational solution to the old Bayesian problem is an approximation, and VPR makes an additional approximation. Even if $q_{old}(\mathbf{m})$ equals $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$ or if we somehow find an exact and analytic solution for $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$, we still need to introduce $q_{new}(\mathbf{m})$ to approximate the true posterior pdf $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ by minimizing the KL divergence $\mathrm{KL}[q_{new}(\mathbf{m})||p_{new}(\mathbf{m}|\mathbf{d}_{obs})]$ in eq. (12) since direct calculation of $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ using eq. (11) requires the normalization constant k to be evaluated which is intractable in high dimensional inverse problems. While the first operation (eq. 10) which estimates $q_{old}(\mathbf{m})$ must be performed using a parametric variational method so that its probability value can be evaluated in eqs. (12) and (13), the second problem can be solved using any variational method.

The most expensive step in the VPR algorithm is to solve the first variational problem, because this requires the likelihood term (forward function) to be calculated. Provided that the support of all other prior pdf's are subsets of this support, then this step needs to be performed only once, after which we can replace prior information rapidly whenever it changes. This makes the proposed method attractive in real problems, especially when multiple different prior distributions are possible for a single observed data set (Earp & Curtis 2020; Bloem *et al.* 2024).

2.5 Physically Structured Variational Inference (PSVI)

In this paper, we use physically structured variational inference (PSVI–Zhao & Curtis 2024b) to solve the two variational problems to calculate both $q_{old}(\mathbf{m})$ and $q_{new}(\mathbf{m})$. PSVI is an efficient parametric variational inference method that defines a Gaussian variational family with a physics-based correlation structure. When the model parameters to be estimated have physical constrains (for example, seismic velocity should be a positive number and earthquake source location should be below the Earth's surface), a bijective function (an invertible transform) is usually applied to the Gaussian random variables to ensure that the transformed model parameters satisfy their physical constrains. For example, the following *logit* functions

$$\begin{aligned} m_i &= f(\theta_i) = a_i + \frac{b_i - a_i}{1+\exp(-\theta_i)} \\ \theta_i &= f^{-1}(m_i) = \log(m_i - a_i) - \log(b_i - m_i) \end{aligned} \tag{14}$$

are often used to convert a Gaussian distributed variable θ_i defined in an unconstrained space (from minus to plus infinity) into the physical model parameter m_i to be estimated which is bounded by the lower and upper bounds a_i and b_i, respectively. The transformed probability distribution can be calculated through the change of variable formula

$$\log p(\mathbf{m}) = \log p(\Theta) - \log|\det(\partial_\Theta f(\Theta))| \tag{15}$$

where $p(\Theta)$ is the Gaussian variational distribution in the unbounded space. Term $|\det(\cdot)|$ calculates the absolute value of the determinant of the Jacobian matrix $\partial_\Theta f(\Theta)$, which accounts for the volume change corresponding to this transform (Kucukelbir *et al.* 2017).

A Gaussian variational distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. To ensure that $\boldsymbol{\Sigma}$ always remains positive semi-definite, we re-parametrize it using a Cholesky factorization $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{\mathbf{T}}$, where $\mathbf{L}$ is a lower triangular matrix. A full covariance matrix can be constructed to include correlation information between pairs of model parameters. However, this incurs huge memory requirements and computational costs (for an n dimensional problem, $\mathbf{L}$ requires $n(n+1)/2$ real-valued entries). Alternatively, a mean-field (factorized) Gaussian variational approximation may be used for high dimensional problems, which defines a diagonal covariance matrix, thus ignoring all correlations between model parameters. These two options are respectively referred to as full rank ADVI and mean-field ADVI in Kucukelbir *et al.* (2017). Unfortunately for the full waveform inversion (FWI) problems considered in this paper, mean-field ADVI normally underestimates uncertainties of the posterior distribution whereas full rank ADVI is intractable due to the dimensionality of models (Zhang *et al.* 2023; Zhao & Curtis 2024b).

PSVI embodies a method with intermediate cost that lies between mean-field ADVI and full rank ADVI, by modelling only the most important (dominant) correlation information in model vector $\mathbf{m}$, guided by physical properties (prior knowledge) of imaging problems. Specifically, in spatial inverse (imaging) problems, model correlations are shown to be strong mainly between pairs of locations that are in spatial proximity to each other, and the magnitude of correlations decreases rapidly as the distance between two locations increases (Gebraad *et al.* 2020; Zhang & Curtis 2021a; Biswas & Sen 2022). This suggests that it might be sufficient to model correlations only between parameters that define properties which are spatially close (e.g., for FWI, parameters of cells that lie within a dominant wavelength of one another), and ignore correlations between those that are further apart.

Since off-diagonal elements of the lower triangular matrix $\mathbf{L}$ dominantly represent correlations between parameter pairs, we impose the following sparse structure on $\mathbf{L}$

$$\mathbf{L} = \begin{bmatrix} l_{0,1} & & & & & & \\ l_{1,1} & l_{0,2} & & & & & \\ 0 & l_{1,2} & l_{0,3} & & & & \\ \dots & 0 & l_{1,3} & \dots & & & \\ l_{i,1} & \dots & 0 & \dots & l_{0,n-2} & & \\ 0 & \dots & \dots & \dots & l_{1,n-2} & l_{0,n-1} & \\ \dots & 0 & l_{i,n-i} & \dots & 0 & l_{1,n-1} & l_{0,n} \end{bmatrix} \tag{16}$$

For each element, the first subscript i indicates a block of off-diagonal elements that are i rows below the main diagonal (i.e., at an offset of i from the main diagonal), and the second subscript j indicates that $l_{i,j}$ is the jth element of that off-diagonal block. In eq. (16), sparsely distributed off-diagonal elements in red are used to capture main correlations between parameter pairs that are assumed to be important (in this case, are spatially close), and their values are optimized during variational inference. All other off-diagonal elements of $\mathbf{L}$ are set to be zero by assuming independence of the corresponding model parameter pairs. Note that we only impose a sparse structure on $\mathbf{L}$ rather than setting constraints on the values of the non-zero off-diagonal elements in red: those values are updated freely during the variational optimization (Zhao & Curtis 2024b).

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

In PSVI, we can impose any desired correlation structure on $\mathbf{L}$ by setting only the corresponding off-diagonal blocks as unknown hyperparameters and optimizing them. The total number of parameters to define $\mathbf{L}$ can thus be greatly reduced compared to that in full rank ADVI. The covariance matrix $\mathbf{\Sigma}$ obtained in this way then also represents a sparse correlation structure with specific non-zero off-diagonal blocks (similar to the red elements in eq. (16) but located below and above the main diagonal elements). Since a *priori* we expect that most of the important correlations are included in PSVI, the obtained variational distribution would capture parameter correlations of interest. Thus, inference results are significantly improved compared to those from mean-field ADVI (Zhao & Curtis 2024b).

Variational parameters $\boldsymbol{\mu}$ and $\mathbf{L}$ are updated by maximizing the EBLO in eq. (8) or equivalently minimizing the KL divergence in eqs. (6) and (12) using gradient based optimization methods, and their gradients are calculated using automatic differentiation libraries (Abadi *et al.* 2016; Paszke *et al.* 2019). The expectation terms are estimated by Monte Carlo integration with a relatively small number of samples drawn from the variational distribution, because the optimization is performed over many iterations so that statistically the parameters will converge towards the correct solution (Kucukelbir *et al.* 2017).

3 APPLICATION TO FULL WAVEFORM INVERSION

Seismic full waveform inversion (FWI) estimates subsurface physical properties, such as seismic velocities and density, using seismic waveform data (Tarantola 1984; Fichtner *et al.* 2009; Virieux & Operto 2009). FWI is a highly nonlinear and non-unique inverse problem, and thus deterministic methods often fail to find a truly representative Earth model that generates the observed data, and to estimate reliable uncertainties in the inversion results. In recent years, researchers have started to use various Bayesian inference methods to solve probabilistic FWI problems, including Monte Carlo sampling methods (Ray *et al.* 2016, 2018; Visser *et al.* 2019; Gebraad *et al.* 2020; Guo *et al.* 2020; Kotsi *et al.* 2020; Zhao & Sen 2021; Khoshkholgh *et al.* 2022; Biswas & Sen 2022; Fu & Innanen 2022; de Lima *et al.* 2023a, b; Berti *et al.* 2023) and variational inference (Zhang & Curtis 2021b; Bates *et al.* 2022; Wang *et al.* 2023; Lomas *et al.* 2023; Izzatullah *et al.* 2024; Zhao & Curtis 2024b; Yin *et al.* 2024b). Bayesian FWI often requires huge computational resources since (1) the dimensionality (number of unknown parameters) of an FWI problem is usually high (Curtis & Lomax 2001), and (2) the forward and adjoint simulations are expensive (Wang *et al.* 2019; Zhao *et al.* 2020). Therefore, reducing computational overhead (while still obtaining reasonably accurate inversion results) is a top priority in Bayesian FWI, especially if we have multiple prior distributions to consider.

We apply variational prior replacement (VPR) to a 2D Bayesian acoustic FWI problem to explore its effectiveness. Fig. 1(a) displays the true velocity model used in the following tests, which is obtained by truncating and downsampling the original Marmousi model (Martin *et al.* 2006) to a grid size of 110 × 250 cells, with each cell measuring 20 m in both horizontal and vertical directions. Top 10 rows of the grid is fixed at their true velocity values during inversion. We place 12 sources (red stars in Fig. 1a) on the surface with a spacing of 400 m, and 250 receivers (white line in Fig. 1a) on the seabed (200 m depth) with a horizontal interval of 20 m. The waveform data are 4 s long with a sample interval of 2 ms, which are generated by solving a 2D acoustic wave equation using a time-domain finite difference method. We further add Gaussian random noise with zero mean and a standard deviation value of 0.1 (∼ 1% of the average value of the maximum amplitude of each seismic trace) to the obtained waveform data, which is treated as the observed data set in this example. The source function is a Ricker wavelet with a dominant frequency of 10 Hz.

We define a Gaussian likelihood function to represent data uncertainties

$$p(\mathbf{d}_{obs}|\mathbf{m}) \propto \exp\left[-\frac{(\mathbf{d}_{syn}-\mathbf{d}_{obs})^T\Sigma_{\mathbf{d}}^{-1}(\mathbf{d}_{syn}-\mathbf{d}_{obs})}{2}\right] \tag{17}$$

In this example, the covariance matrix of data noise $\Sigma_{\mathbf{d}}$ is assumed to be a diagonal matrix (uncorrelated data noise), and all diagonal elements are set to be 0.1 to represent the level of noise added to the synthetic waveform data. That is to say, we assume the data covariance matrix $\Sigma_{\mathbf{d}}$ is known. The same finite difference forward modelling method is used to calculate synthetic data $\mathbf{d}_{syn}$, and data-model gradients are computed using the adjoint-state method (Plessix 2006). During forward and adjoint simulations, we fix velocity values in the water layer at their true values. Prior distributions used in this study are discussed below.

3.1 Prior information

We consider three different types of prior information in the FWI problem. We first define a uniform prior distribution $p_1(\mathbf{m})$ for the velocity values at each grid cell with lower and upper bounds at each depth displayed in Fig. 1(b), similar to that used in Zhang & Curtis (2021b). This is a non-informative (weak) and thus broad prior distribution with no correlations between neighbouring cells. It has the advantage that any type of velocity contrast between neighbouring cells would be consistent with this prior pdf (all model samples have the same prior probability density) as long as they lie within the prior bounds, and hence can in principle be discriminated only by comparing their consistency with the observed waveform data.

The second prior distribution is a spatially smoothed version of the uniform distribution, obtained by applying a second-order finite difference (smoothing) operator $\mathbf{S}$:

$$\mathbf{S} = \begin{bmatrix} & & \cdots & & & & \\ \cdots & 1 & -2 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & -2 & 1 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & -2 & 1 & \cdots \\ & & & \cdots & & & \end{bmatrix} \tag{18}$$

to model parameter $\mathbf{m}$. Define a Gaussian distribution for $\mathbf{Sm}$

$$p(\mathbf{Sm}) = k_1 \exp\left[-\frac{1}{2}(\mathbf{Sm})^T\Sigma_{\mathbf{Sm}}^{-1}(\mathbf{Sm})\right] \tag{19}$$

where k_1 is a normalization constant, and $\Sigma_{\mathbf{Sm}}$ is a diagonal matrix with its diagonal elements controlling the strength of the spatial smoothness (larger values correspond to weaker spatial smoothness). In this paper, the diagonal elements of $\Sigma_{\mathbf{Sm}}$ are set to 500. This can be interpreted as applying a Tikhonov (regularization) matrix $\mathbf{S}$ to $\mathbf{m}$ (Golub *et al.* 1999; Aghamiry *et al.* 2018). Then the smoothed prior distribution $p_2(\mathbf{m})$ can be written as

$$p_2(\mathbf{m}) = \frac{p(\mathbf{Sm})p_1(\mathbf{m})}{k_2} \tag{20}$$

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

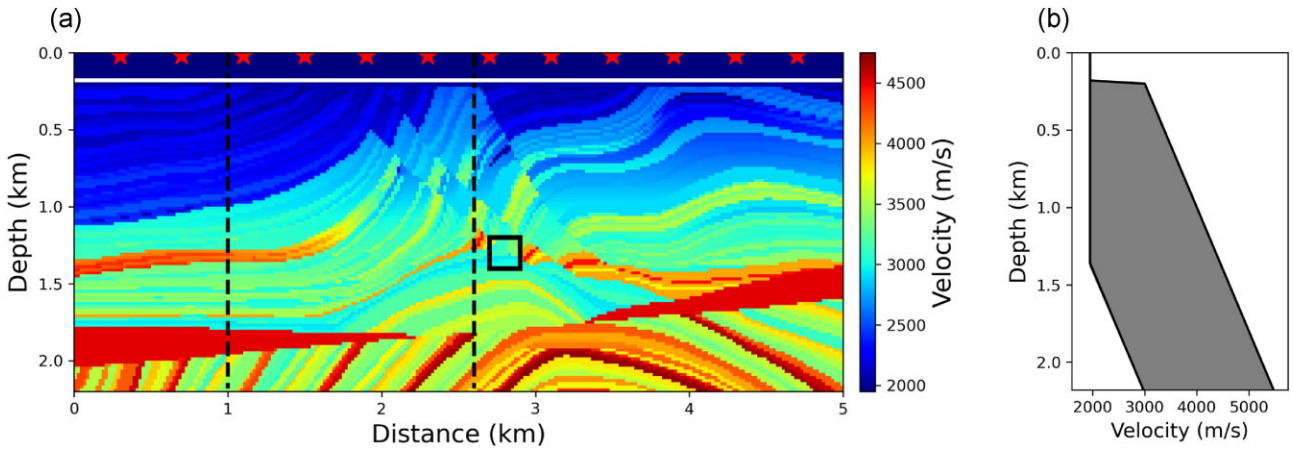

Figure 1. (a) Truncated and down-sampled P wave velocity of the Marmousi model used in this paper. Source locations are indicated by red stars and the receiver line is marked by a white line. Dashed black lines display the locations of two vertical profiles used to compare the posterior marginal probability distributions in Figs 6 and 10 in the main text. (b) Upper and lower bounds for the uniform prior distribution at different depths.

where $p_1(\mathbf{m})$ is the uniform distribution defined above, and k_2 is another normalization constant which can be absorbed into the evidence term in Bayes' rule so we do not need to calculate its value.

This prior distribution $p_2(\mathbf{m})$ embodies strong prior information in which model samples with smaller velocity contrasts between spatially neighbouring cells have higher probability values. Therefore, velocities in neighbouring cells should be positively correlated. This information may or may not be advantageous depending on the true (geological) prior information about the form of the velocity structure being estimated. Compared to the uniform prior distribution, large velocity contrasts between neighbouring cells are almost excluded by this prior information as they have relatively low prior probability values. This effectively reduces the hypervolume of parameter space spanned by significantly non-zero values of the posterior pdf. In other words, it provides more information than $p_1(\mathbf{m})$. More detailed comparisons of these two prior distributions and their effects on the posterior pdfs can be found in Earp & Curtis (2020).

The third prior distribution $p_3(\mathbf{m})$ is a Gaussian distribution with real geology-informed inter-parameter correlation information. The mean and standard deviation vectors of the Gaussian prior distribution are set to be those of the uniform prior distribution $p_1(\mathbf{m})$ for consistency. Considering the dimensionality of this FWI problem (100×250), it is difficult to build a full prior covariance matrix to describe detailed geological prior information. We therefore build a prior covariance matrix that incorporates only a local correlation structure estimated from real geology.

To achieve this, we first select a set of realistic geological images. Fig. 2 displays one such image (John 2012), and while the images are all at much smaller scale than the Marmousi model was designed to represent, we assume scale invariance of geological correlations (only for the purposes of this test). From each of the pictures, we randomly sample 1000 subimages using a window with 20×20 pixels, which together represent a local correlation structure between parameters. Fig. 3(a) shows the calculated full correlation matrix with a size of 400×400, each element denoting correlation information of one parameter pair within the 20×20 window. To analyse a more detailed structure of this correlation matrix, Figs 3(b) and (c) present its first 60×60 elements and 20×20 elements, respectively. Note that we reshape the 20×20 (2D) images into 1D vectors in a row-major order (i.e., for each training image the first 20 elements of the 1D vector comprise the first row

Figure 2. A picture of real geological structures with a scale of metres (John 2012) used to calculate a local correlation matrix and define the geological prior distribution $p_3(\mathbf{m})$.

of the 2D image, the second 20 elements comprise the second row, and so on). Therefore, off-diagonal blocks observed in Figs 3(a) and (b) represent correlation information in the vertical direction; only one such obvious block is visible, meaning that vertical correlations exist predominantly between vertically adjacent cells and decay rapidly with greater inter-cell distance. On the other hand, off-diagonals directly below and above the main diagonal elements (displayed in Fig. 3c) denote horizontal correlations: three or four strong off-diagonal elements have correlation values larger than 0.7, implying that strong horizontal correlations exist within approximately 4 neighbouring cells. Such horizontally smoother and vertically rougher correlation features are also clear in Fig. 2.

We use this correlation matrix to construct a full correlation matrix $\mathbf{R}$ that describes correlations in model vector $\mathbf{m}$ (with a dimensionality of 100×250 in this example) by considering correlations between pairs of parameters that are located only inside a 20×20 window. We set all other elements to zero for reasons discussed in Section 2.5 and in Zhao & Curtis (2024b). The covariance matrix of this Gaussian prior distribution $\mathbf{\Sigma}_{p_3}$ can then be calculated by

$$\mathbf{\Sigma}_{p_3} = \mathbf{D}_{std}\, \mathbf{R}\, \mathbf{D}_{std} \tag{21}$$

where $\mathbf{D}_{std}$ is a diagonal matrix with diagonal elements being the standard deviations of $p_3(\mathbf{m})$, and $\mathbf{R}$ is the correlation matrix obtained above. Finally, the Gaussian prior distribution $p_3(\mathbf{m})$ can be

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

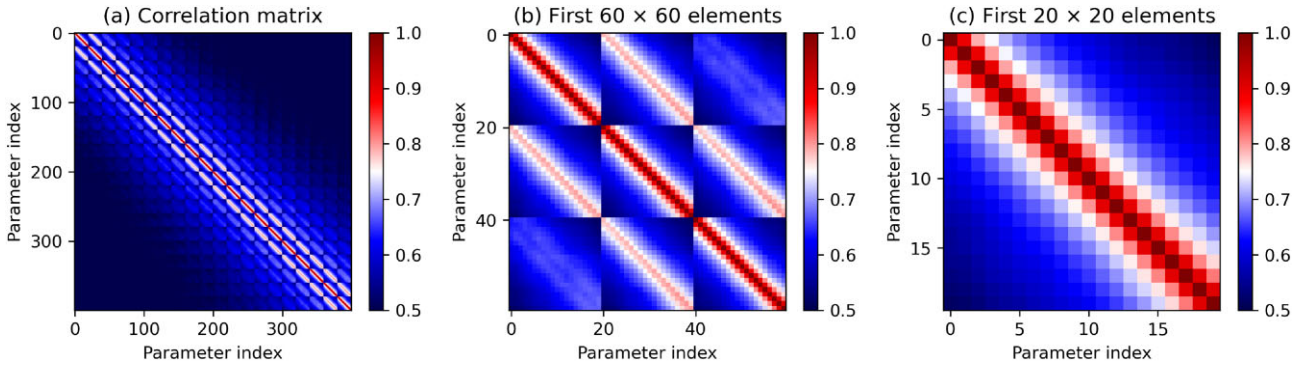

Figure 3. (a) Prior correlation matrix calculated from a set of 20 × 20 sized training images sampled from pictures of real geology such as Fig. 2. (b) and (c) show magnifications of the first 60 × 60 and 20 × 20 elements in (a), respectively. Due to parametrization of the training images, distinct off-diagonal blocks below and above the main diagonal block in (a) and (b) represent vertical correlations, and off-diagonals directly below and above the main diagonal elements in (c) denote horizontal correlations.

defined as

$$p_3(\mathbf{m}) = k_3 \exp\left[-\frac{1}{2}\left(\mathbf{m}-\boldsymbol{\mu}_{p_3}\right)^T \boldsymbol{\Sigma}_{p_3}^{-1}\left(\mathbf{m}-\boldsymbol{\mu}_{p_3}\right)\right] \tag{22}$$

where $\boldsymbol{\mu}_{p_3}$ is the mean vector of this Gaussian distribution. Similarly to eq. (20), k_3 is a normalization constant whose value is not required in VPR. Below, $p_3(\mathbf{m})$ is referred to as the geological prior distribution since it captures spatial correlation information from real geological structures.

Fig. 4 displays one random sample drawn from each of the three prior distributions. Since no spatial correlation is considered in the uniform prior distribution, we observe large velocity contrasts between neighbouring cells in Fig. 4(a). The smoothed and geological prior distributions impose spatially correlated information, thus the prior samples presented in Figs 4(b) and (c) are spatially smoother. In addition, local velocity structures in Fig. 4(c) show rectangular patterns with larger sizes in the horizontal direction and smaller sizes in the vertical direction due to horizontal smoothness and vertical roughness as represented in the geological prior pdf and illustrated in Figs 2 and 3. This pattern is not observed in Fig. 4(b) since in $p_2(\mathbf{m})$ we impose the same magnitude of smoothness in both horizontal and vertical directions (this was not a requirement, but in eq. (19) we used equal horizontal and vertical smoothing to contrast with the geological prior pdf). Fig. 4 thus proves that the three prior distributions encapsulate significantly different prior information that we may wish to inject into FWI inversion results.

3.2 Verifying variational prior replacement

In the first test, we verify that VPR produces correct results by comparing them to those obtained using a conventional approach where an independent Bayesian inversion is performed for each prior distribution (referred to herein as *prior specific inversion*). For this test, we consider the uniform prior distribution $p_1(\mathbf{m})$ and the smoothed prior $p_2(\mathbf{m})$. For the prior specific case, PSVI is used to solve these two FWI problems with their respective priors. We update variational parameters (mean vector $\boldsymbol{\mu}$ and lower triangular matrix $\mathbf{L}$ mentioned in Section 2.5) for 5,000 iterations. During each iteration, 2 random samples are used to approximate the ELBO[$q(\mathbf{m})$] (eq. 8) using Monte Carlo integration; such a low number of samples has been shown to be reasonable in a stochastic sense in previous studies, because of the large number of iterations (Kucukelbir *et al.* 2017). For VPR, the uniform distribution $p_1(\mathbf{m})$ is treated as the old prior distribution, which is then removed from the old posterior distribution (the posterior pdf calculated using $p_1(\mathbf{m})$ by prior specific inversion) and replaced by the smoothed (new) prior distribution $p_2(\mathbf{m})$. This is achieved by solving the variational problem described in eq. (13), through minimization of the KL divergence expressed in eq. (12). Similarly to the prior specific case, PSVI is used to solve this problem, where variational parameters are updated for 5,000 iterations with 10 samples per iteration used to estimate the expectation term in eq. (12). Note that it might be impossible to replace a smoothed (old) prior by a Uniform (new) prior distribution using VPR since in this case the support of the old prior may only be a subset of that of the new prior distribution (or it might be effectively so due to sampling and numerical approximations). This might make $p_{new}(\mathbf{m})/p_{old}(\mathbf{m})$ numerically unstable because for some parameters $\mathbf{m}$ the value $p_{old}(\mathbf{m})$ could be small, poorly determined or even zero.

Figs 5(a) and (b) display prior specific inversion (PSI) and VPR results obtained using the smoothed prior distribution. In each column, a random posterior sample, the mean velocity map, standard deviation and the relative error of the posterior distribution are presented from top to bottom row. The relative error defined to be the difference between the mean and true velocity models (Fig. 1a) divided by the standard deviation at each point, reflecting the relative deviation between the true and inverted mean models. The most important feature is that the first-order posterior statistics displayed in Figs 5(a) and (b) are almost identical; Fig. 5(b) was produced from the results obtained using the uniform prior (displayed in Fig. 9a). This supports the statement that VPR is able to replace the old prior information and inject (update) the new prior information into the inversion results, without solving a Bayesian inverse problem again from scratch.

We do observe some discrepancies between these two sets of results. For example, a vertically oriented low velocity structure is present in Fig. 5(a) inside dashed black and red boxes, which is not present in Fig. 5(b). Although this feature is not observed in the true velocity model in Fig. 1(a), there is no strong evidence to discriminate which result is better.

For other local regions such as those below and to the right of the boxes in Fig. 5, the mean velocity model from VPR (Fig. 5b) appears to be closer to the old posterior distribution using the uniform prior pdf (displayed in Fig. 9a) than it does to the PSI result using the smoothed prior (Fig. 5a). This might be because this

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

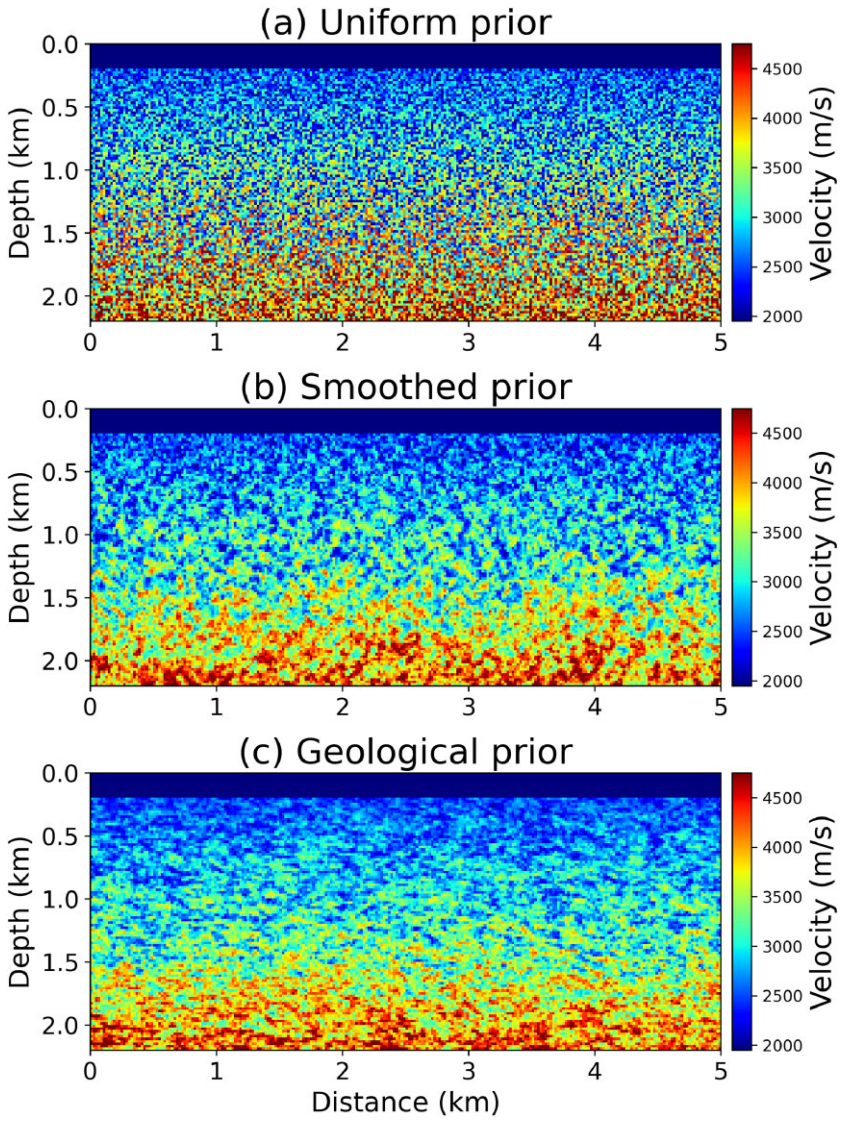

Figure 4. (a) – (c) One random prior sample drawn from the (a) uniform, (b) smoothed and (c) geological prior distributions defined in the main text, respectively.

implementation of PSI (which is an optimization problem) converged to a local minimum, or that it has not fully converged (full convergence might require a larger number of forward evaluations which is very expensive). Alternatively, it is also possible that the VPR procedure (also an optimization process) has not fully converged and thus might indicate an incomplete prior replacement in this test. On the other hand, for some other statistics such as the characteristic of spatial variations in each posterior sample, standard deviations, or posterior marginal pdfs displayed in Fig. 6, VPR results are clearly closer to the PSI results using the smoothed prior than to those using the uniform prior, supporting the statement that VPR produces reasonable statistical accuracy. We also note that it is reasonable that there remain some discrepancies, caused by the fact that in VPR we introduce a variational distribution $q_{new}(\mathbf{m})$ to approximate the new posterior distribution $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ as expressed in eq. (13), rather than calculating the actual $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ as in Walker & Curtis (2014b).

Fig. 6 compares the posterior marginal pdfs of the above two results along two vertical velocity profiles at horizontal locations of 1 km (top row) and 2.6 km (bottom row). Their locations are displayed by dashed black lines in Fig. 1(a). Red and black lines represent the true and mean velocity values, respectively. Despite some small discrepancies here and there in Figs 6(a) and (b), they provide similar posterior marginal pdfs. Interestingly, as displayed in the top row, the two methods (PSI and VPR) find very similar yet *incorrect* posterior solutions given the same data and prior information (especially below 1.3 km depth where true velocity values are excluded by the high probability region of the posterior pdfs). This is because that the PSVI algorithm may have converged around an incorrect solution in this region caused by cycle skipping, which often occurs in FWI problems. We also compare correlation information of the two posterior pdfs in Fig. 7, which displays the posterior correlation matrices for velocity values in a 2D window with a size of 10 × 10 cells inside the black box in Fig. 1(a). The top row shows the full correlation matrices (with a size of 100 × 100), and the bottom row shows the first 30 × 30 elements. Highly consistent posterior correlation values are obtained, which again proves the effectiveness of VPR.

To further test the performance of VPR, in Figs 8(a) and (b) we compare one observed shot gather with data simulated by one

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

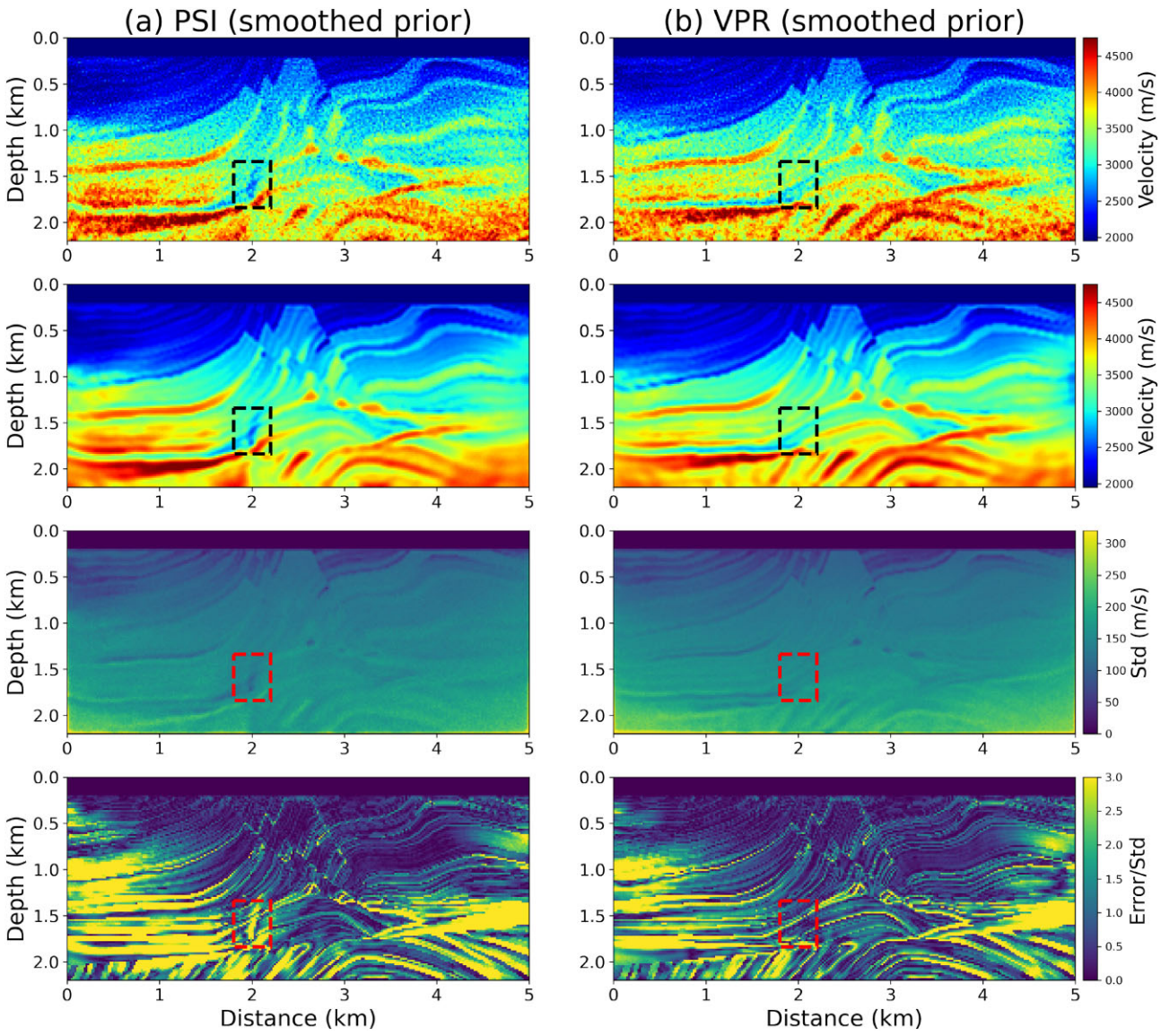

Figure 5. (a) Prior specific inversion (PSI) and (b) variational prior replacement (VPR) results obtained using the smoothed prior distribution $p_2(\mathbf{m})$. The latter is obtained by removing the uniform prior information $p_1(\mathbf{m})$ from the old posterior pdf (displayed in Fig. 9a), and imposing the smoothed prior $p_2(\mathbf{m})$ using VPR. From top to bottom row, they are: a random posterior sample, mean velocity map, standard deviation and relative error of the obtained posterior distribution, respectively. The relative error is the absolute error between the mean and true models divided by the corresponding standard deviation at each point. Red and black dashed boxes highlight differences between (a) and (b).

randomly chosen posterior sample obtained from PSI and VPR results, respectively. In both figures, the simulated data are highly consistent with the observed (noisy) data, which demonstrates that VPR can provide models that produce accurate waveform data that fit the observed data to within data uncertainties.

In Appendix A, we present a second example to test the accuracy of VPR using a Gaussian prior distribution for $p_1(\mathbf{m})$, which again shows that VPR and PSI provide highly consistent inversion results and posterior statistics. In conclusion, since variational prior replacement and prior specific inversion provide almost identical posterior random samples, first-order statistics (posterior mean, standard deviation and marginal pdfs) and second-order statistics (correlation matrices), we assert that the proposed method is effective and accurate for varying prior information in Bayesian inference without performing repeated independent inversions.

3.3 FWI using different priors

In this section we analyse the effect of the three different priors defined previously and compare the corresponding inversion results. We use PSVI to perform a single variational Bayesian FWI using the uniform distribution $p_1(\mathbf{m})$. $p_1(\mathbf{m})$ is then replaced by both the smoothed prior $p_2(\mathbf{m})$ and the geological prior $p_3(\mathbf{m})$ using VPR. Note that in VPR the old prior distribution should be broader than the new prior to avoid numerical instability issues such as occur when attempting to divide by zero (Walker & Curtis 2014b). However, the Gaussian geological prior distribution is defined in the space of real numbers, which spans a broader parameter space than the uniform distribution. Therefore, we truncate $p_3(\mathbf{m})$ within the lower and upper bounds of the uniform prior distribution, and renormalize $p_3(\mathbf{m})$ by another normalization constant (which does not need to be evaluated, similar to those in eqs. 20 and 22).

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

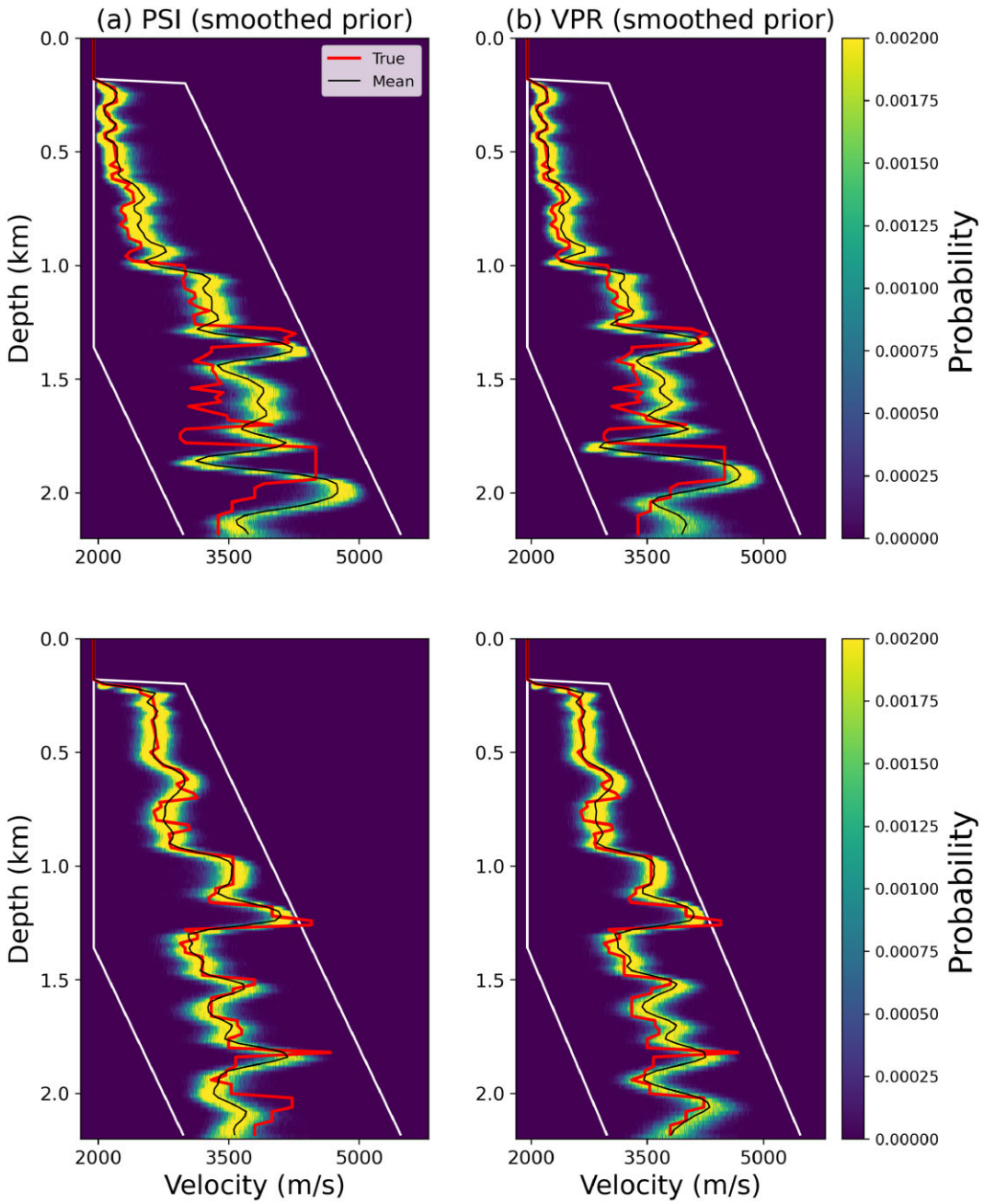

Figure 6. Posterior marginal distributions coloured from dark blue (zero probability) to yellow (maximum value of marginal pdf's in each plot), along two vertical velocity profiles at horizontal locations of 1 km (top row) and 2.6 km (bottom row) whose locations are marked by black dashed lines in Fig. 1(a). (a) and (b) Posterior marginal pdfs from prior specific inversion (PSI) and variational prior replacement (VPR), obtained using the smoothed prior distribution.

Figs 9(a)–(c) display the obtained inversion results. Each figure includes a random posterior sample, the mean velocity, standard deviation and the relative error maps of the posterior pdf from top to bottom row. Note that Figs 5(b) and 9(b) represent the same results obtained using VPR. In a previous study (Zhao & Curtis 2024b), we compared the inversion results obtained without using prior replacement displayed in Fig. 9(a) with two entirely independent variational methods using exactly the same uniform prior distribution and observed data and obtained highly consistent results, proving that we obtain approximately correct posterior uncertainty statistics for this specific prior. In this study we focus on the inversion results obtained using different priors.

The posterior random sample in Fig. 9(a) shows significant velocity 'speckle' - strong, short wavelength contrasts - since no correlation is introduced by the uniform prior distribution. The relatively non-informative prior information results in significantly higher uncertainties at greater depths, increasing up to around 800 m/s. The two posterior samples in Figs 9(b) and (c) are smoother since extra (smooth) prior information is injected into the two inversion results which precludes sharp velocity variations between neighbouring cells. The smoothed prior pdf imposes spatial smoothness explicitly, and the geological prior pdf injects similar information implicitly, as illustrated by the positive correlations displayed in Fig. 3. Therefore, the true velocity

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

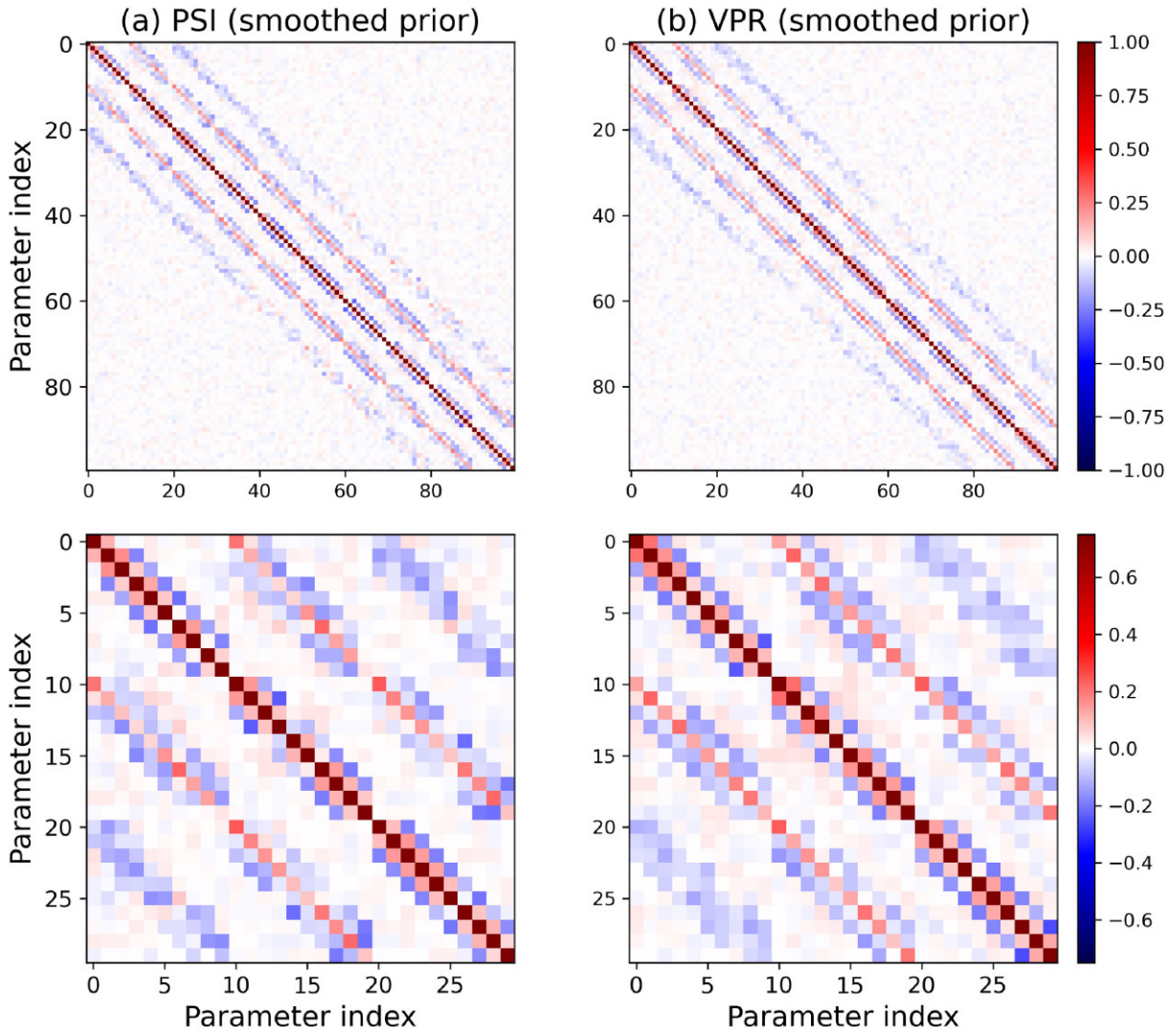

Figure 7. Posterior correlation matrices from (a) PSI and (b) VPR results for velocity values in a 10 × 10 window inside the black box in Fig. 1(a). Top row shows the full 100 × 100 sized posterior correlation matrices and bottom row shows the first 30 × 30 elements for better comparison.

structures are better resolved since they are indeed laterally fairly smooth.

The three mean velocity maps are quite smooth and similar to each other, generally resembling the true velocity map. The methods fail to recover some thin layers in the deeper part of the model due to the limited frequency band of the waveform data (10 Hz dominant frequency). Also, the mean can sometimes have a very low, even zero, probability density, especially in the case of the uniform prior distribution where such a smoothed mean model might be precluded by the data: this is because purely observed waveform data would inject negative correlations between neighbouring cells as displayed below in Fig. 11(a) and in Gebraad *et al.* (2020); Zhang *et al.* (2023); Zhao & Curtis (2024b). Therefore, smoothed velocity structures (such as the mean model) which imply positive correlations between adjacent cells, might have low probability values.

Standard deviation values displayed in Figs 9(b) and (c) are smaller than those in Fig. 9(a) (note that different colorbars are used in the former figures). This makes sense because by imposing prior information that velocity structures should be relatively smooth, we have removed the possibility of including large velocity contrasts between laterally proximal cells. In fact, the introduction of prior information, if it correctly reflects the true state of nature, should lead to models that better reflect the true state of nature overall (other than in pathalogical cases). In our case, the geological prior information introduced is derived from pictures that represent real geology and is therefore reasonably reflective of the true model, so in this case at least, the introduction of prior information should improve the result. Nevertheless, in some cases the uncertainty reduction displayed in Figs 9(b) and (c) might not be a good outcome. Normally there is less information at greater depths from waveform data recorded at the surface. The decreased uncertainties at depth occur because information from data and prior knowledge are combined. This only leads to more accurate models if the prior information is also accurate. In the case of our smoothed prior, smoothing is applied equally in vertical and horizontal directions which does not correctly reflect the structure of the true model. In that case we therefore would not expect that results are more accurate after applying this prior information (but they may be, for example if the relative errors in vertical smoothing are more than compensated by increased accuracy in horizontal smoothing). In

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

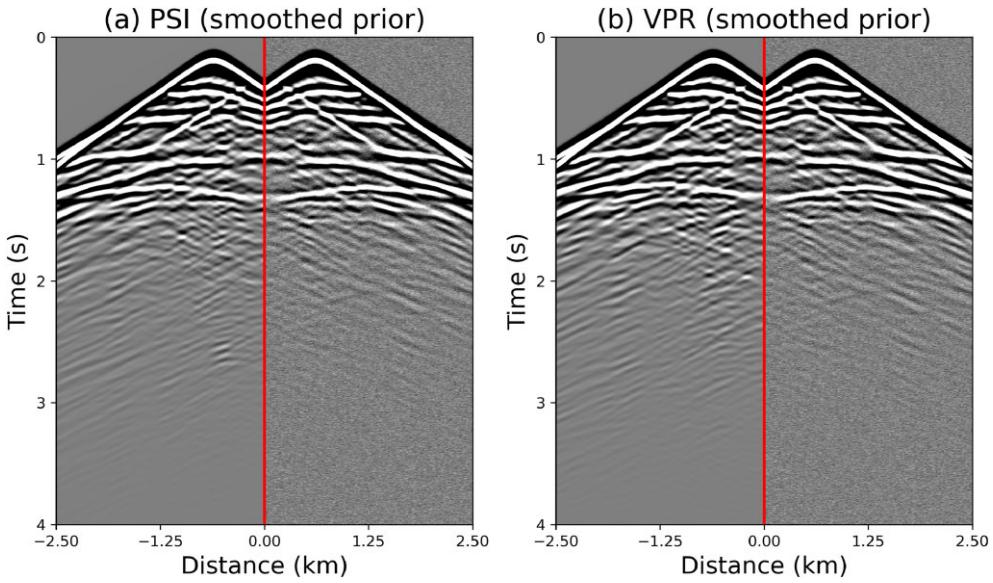

Figure 8. The "butterfly plot" of data comparison for one common shot gather. (a) The data predicted by a random posterior sample from PSI (left hand side of red line) and the observed waveform data (right hand side of red line). (b) The data predicted by a posterior sample from VPR (left) and the observed waveform data (right). In both figures, the simulated data are highly consistent with the observed (noisy) data.

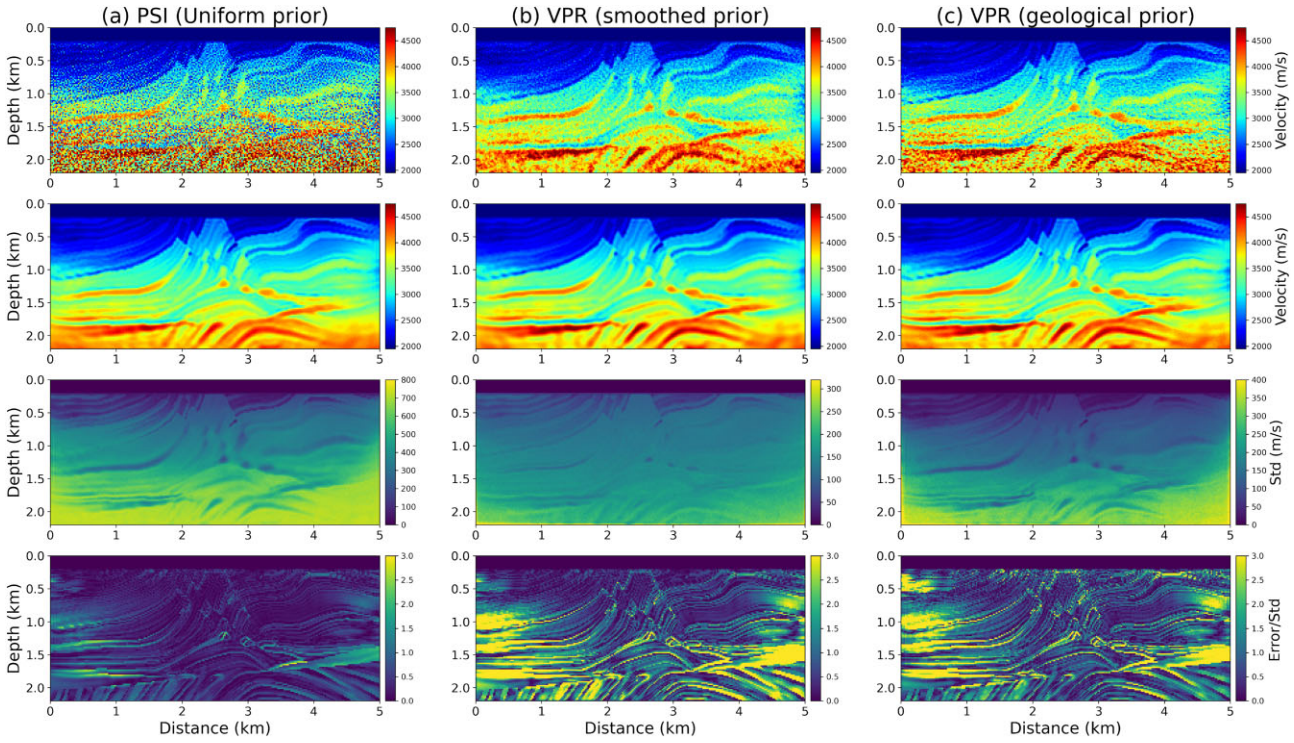

Figure 9. (a) Prior specific inversion (PSI) results obtained using the uniform prior distribution $p_1(\mathbf{m})$. (b) and (c) Variational prior replacement (VPR) results obtained by replacing the uniform prior distribution $p_1(\mathbf{m})$ by the smoothed prior $p_2(\mathbf{m})$ and the geological prior $p_3(\mathbf{m})$, respectively. In each column, a random posterior sample, mean velocity, standard deviation and relative error maps are displayed from top to bottom row, respectively.

addition, the posterior uncertainties in Fig. 9(c) are higher than those in Fig. 9(b), possibly due to the different magnitude of smoothness applied to these two results (one controlled by the predefined parameter $\Sigma_{\mathbf{Sm}}$ in eq. (19) and the other by correlations calculated from the training images such as Fig. 2).

In addition to the magnitude of the standard deviation values, strong prior information also suppresses overall changes in the uncertainty structures: obvious spatial (vertical) variations are observed in the standard deviation map in Fig. 9(a). For example, uncertainties increase at depth since data sensitivity decreases at

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

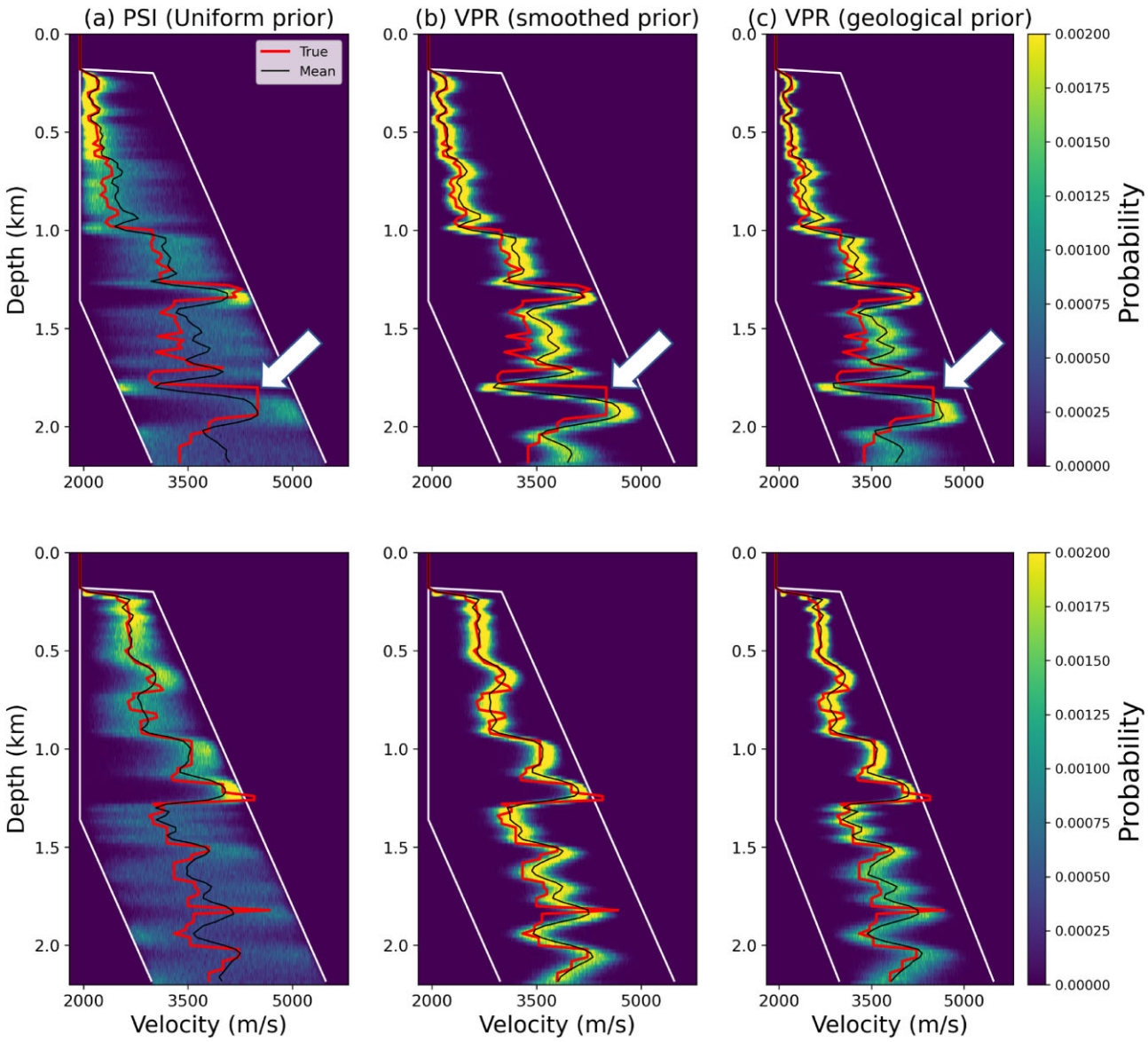

Figure 10. Posterior marginal distributions along two vertical velocity profiles at locations of 1 km (top row) and 2.6 km (bottom row) marked by two black dashed lines in Fig. 1(a). Columns (a) – (c) correspond to the same inversion results as those displayed in Figs 9(a)–(c).

depth and lower uncertainties are observed around some high velocity layers. However, those features are less significant in Fig. 9(c) and almost invisible in Fig. 9(b). This is because the geological prior information imposes weaker vertical smoothness, as illustrated in Figs 3 and 4(c), whereas the smoothed prior distribution imposes the same magnitude of smoothness in both vertical and horizontal directions. As a result, we observe larger relative errors in Figs 9(b) and (c), especially at layer boundaries where higher uncertainties *should* be expected (Galetti *et al.* 2015).

Fig. 10 displays the posterior marginal pdfs of the three inversion results at the same two locations as in Fig. 6. Similarly to the standard deviation maps, the marginal pdfs in Figs 10(b) and (c) are narrower than those in Fig. 10(a) due to the additional prior information injected. Fig. 10(c) presents larger vertical variations than Fig. 10(b). As marked by the three white arrows in the first row, the (old) posterior pdf using the uniform prior distribution fails to find the true solution, thus neither do the two VPR results. This is reasonable considering the variational prior replacement methodology: the method can only *replace* prior information imposed previously into the inversion results, but it cannot *correct (improve)* the old estimate of the posterior distribution in cases where the old estimate is poor. Nevertheless, in places where the old posterior distribution includes the true model solution, VPR injects new prior information properly as displayed in the bottom row.

To analyse posterior correlation information in both horizontal and vertical directions, we calculate the correlation matrices for velocity values selected from 10 horizontally contiguous and 10 vertically contiguous grid cells (marked by top and left boundaries of the black box in Fig. 1a). The corresponding results are displayed in the top and bottom rows of Fig. 11. Similarly to Figs 9 and 10, each column (Figs 11 a–c) represents the calculated correlation matrix from one specific prior pdf. Since no correlation information is

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

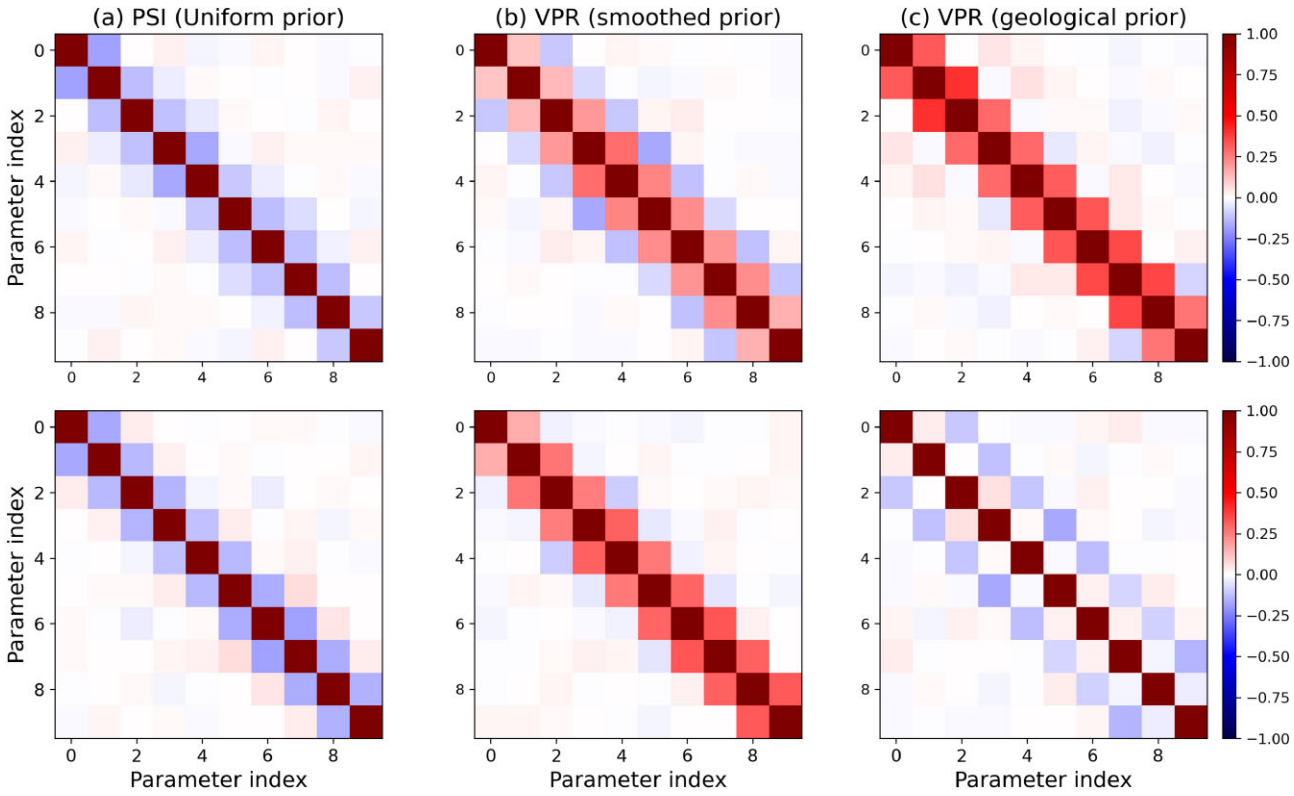

Figure 11. Posterior correlations for velocity values in 10 horizontally (top row) and vertically (bottom row) contiguous cells along the top and left of the black box in Fig. 1(a), resulting in correlation matrices with sizes of 10 × 10. From left to right, (a) – (c) correspond to the same inversion results as those displayed in Figs 9 and 10.

introduced by the uniform prior distribution, the posterior correlations displayed in Fig. 11(a) are purely determined by observed data. Negative correlations are observed between neighbouring cells and positive correlations are vaguely presented between every second neighbouring cells. Since FWI is a highly underdetermined inverse problem (the effective number of independent data points is significantly smaller than the number of unknown model parameters), velocity values oscillate between adjacent cells to achieve a better data fit, especially for grid cells within one wavelength. Similar posterior correlation patterns using a uniform prior distribution were observed in previous studies (Gebraad *et al.* 2020; Zhang *et al.* 2023; Zhao & Curtis 2024b). This would also happen in linearized (deterministic) FWI if no regularization term (prior information) was added.

As discussed previously, the smoothed prior distribution imposes the same magnitude of smoothness in both directions to prevent sharp velocity changes (i.e., positive correlations between adjacent cells). The posterior correlations presented in Fig. 11(b) are thus a result of the combination of correlation information from both the prior pdf (positive correlations) and observed data (negative correlations). Positive correlation values between neighbouring cells in Fig. 11(b) indicate that correlation information from the smoothed prior is stronger than that from the waveform data. The geological prior pdf injects different levels of smoothness in horizontal and vertical directions, as illustrated in Fig. 3. In the horizontal direction, the magnitude of the smoothness injected by the prior pdf is presumably stronger than the magnitude of anti-correlation injected by the data, resulting in positive correlations between adjacent cells (top row in Fig. 11c). On the other hand, we observe almost zero correlations between vertically neighbouring cells (bottom row in Fig. 11c), implying that the vertical correlation information injected from the prior and data have similar strength, thus being cancelled out completely.

3.4 Computational cost

In Table 1 we summarize the detailed computational cost for the four different inversion tests performed in the previous two sections. The first two rows in Table 1 present the computational cost for the prior specific inversion (PSI) method using the uniform and smoothed prior distributions. In each test, we update the variational parameters for 5,000 iterations with 2 samples per iteration, resulting in a total number of 10,000 forward and adjoint (FWI) simulations. This process is performed using 36 CPU cores with a wall clock time of approximately 2 days. The other two results are obtained using variational prior replacement (VPR) with smoothed and geological prior distributions, based on the inversion results from the uniform prior pdf. To solve these two VPR problems, we use 5,000 iterations and 10 samples per iteration to minimize the KL divergence in eq. (12). This does not require any FWI simulations to calculate the likelihood (data misfit) value for any sample. We need only to evaluate the probability value of the old posterior pdf represented by $q_{old}(\mathbf{m})$. To achieve this, in VPR we construct $q_{old}(\mathbf{m})$ using a parametric variational inference method (in this case we use PSVI introduced in Section 2.5). Note that the computational cost for evaluating the probability value $q_{old}(\mathbf{m})$ is almost zero compared to the forward and gradient simulations in FWI. Therefore, the two VPR results can be obtained within 5 minutes using 1 CPU core, which can be performed efficiently even on a laptop. The costs for both PSI and VPR depend on subjective assessments of

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

Table 1. A comparison of computational cost of the different tests performed in this study. PSI and VPR stand for *prior specific inversion* and *variational prior replacement*, respectively.

Prior pdf	Method	Number of iterations	Samples per iteration	Number of FWI simulations	CPU cores	Elapsed time
Uniform	PSI	5000	2	10,000	36	2 days
Smoothed	PSI	5000	2	10,000	36	2 days
Smoothed	VPR	5000	10	0	1	5 minutes
Geological	VPR	5000	10	0	1	5 minutes

the point of convergence, so the absolute computational time listed in Table 1 might not be entirely accurate. Nevertheless, it is still obvious that VPR is significantly cheaper than PSI since no further FWI simulation is involved once we have obtained the old posterior distribution. In Section 3.2, we showed that PSI and VPR provide almost identical results. This makes the proposed method attractive when multiple different priors are available or need to be tested using the same observed data, as presented in this paper.

4 DISCUSSION

We demonstrated that variational prior replacement (VPR) can change prior information efficiently post Bayesian inference. The updated posterior distribution is found by solving a variational problem, in which a variational distribution $q_{new}(\mathbf{m})$ is introduced and optimized iteratively to approximate $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$, as expressed in eq. (13). Therefore, we do not expect VPR and prior specific inversion (PSI) to provide exactly the same results, especially if PSI itself is performed using variational inference as in this study which also results in an approximation. This is part of the reason why we observe some small discrepancies between the results obtained using VPR and PSI displayed in Figs 5 – 7.

A similar effect was observed in the original prior replacement paper (Walker & Curtis 2014b) which used mixture density networks (MND–Bishop 1994) to estimate the old posterior pdf, and used (semi-)analytic methods to calculate the new posterior pdf using eq. (5). In that case, the obtained pdf was still not the actual posterior distribution $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$ given observed data and new prior information, since the old posterior distribution $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$ used in eq. (5) remained an approximation represented by the MND. This explains why the results obtained using prior replacement and direct Monte Carlo sampling (i.e., prior specific inversion) displayed in Figures 2 and 3 in Walker & Curtis (2014b) are not identical. Nevertheless, most of the posterior statistics in this current paper are nearly identical between PSI and VPR, implying that VPR is effective at updating prior information.

In addition, (semi-)analytic calculation of eq. (5) requires the evaluation of the normalization constant k by integrating $p_{new}(\mathbf{m})$, $p_{old}(\mathbf{m})$ and $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$ over the entire parameter space analytically, which might be intractable for high dimensional inference problems, and indeed only under certain circumstances can this be done (Walker & Curtis 2014b). In the proposed VPR framework, we introduce a second variational distribution $q_{new}(\mathbf{m})$, which is found by minimizing the KL-divergence in eq. (12); the specific value of k then need not be estimated explicitly, so VPR can be implemented in a more straightforward manner. On the other hand, this implies that VPR is itself an approximate method which uses $q_{new}(\mathbf{m})$ to approximate $p_{new}(\mathbf{m}|\mathbf{d}_{obs})$.

One essential requirement for prior replacement developed here and in Walker & Curtis (2014b) is that the probability value for the old posterior pdf must be able to be evaluated cheaply (otherwise, there is no reason to use VPR instead of prior specific inversion). In this study, we use physically structured variational inference for this purpose (Zhao & Curtis 2024b), which constructs a transformed Gaussian distribution with a specific correlation structure to approximate the old posterior distribution $p_{old}(\mathbf{m}|\mathbf{d}_{obs})$ so that the probability value $q_{old}(\mathbf{m})$ can be calculated efficiently. Other well-established parametric variational inference methods, such as normalizing flows (Rezende & Mohamed 2015; Dinh *et al.* 2015; Kobyzev *et al.* 2020; Papamakarios *et al.* 2021; Zhao *et al.* 2022a; Siahkoohi *et al.* 2020; Levy *et al.* 2022), automatic differentiation variational inference (Kucukelbir *et al.* 2017; Zhang & Curtis 2020a; Bates *et al.* 2022; Sun *et al.* 2023) and boosting variational inference (Guo *et al.* 2016; Miller *et al.* 2017; Locatello *et al.* 2018; Zhao & Curtis 2024a), can also be used to construct $q_{old}(\mathbf{m})$. The choice of method should be based on the specific problem at hand, since the No Free Lunch theorem states that no method is better than any other when averaged across all problems (Wolpert & Macready 1997).

Note that the prior replacement step (the second step described by eqs. 5 and 11) does not necessarily need to be solved using parametric variational methods or even variational inference. Various Monte Carlo sampling methods can also be used for this purpose as long as the dimensionality of the inverse problem is not too high (Curtis & Lomax 2001).

Walker & Curtis (2014b) used mixture density networks (MDN) to approximate the old posterior distribution, and it has been shown to be difficult to capture posterior correlations between different parameters using this method (Zhang & Curtis 2021a; Bloem *et al.* 2024). Nevertheless, as shown in numerous studies (Devilee *et al.* 1999; Meier *et al.* 2007a, b; Shahraeeni & Curtis 2011; Shahraeeni *et al.* 2012; Käufl *et al.* 2014, 2016; Earp & Curtis 2020; Earp *et al.* 2020; Cao *et al.* 2020; Lubo-Robles *et al.* 2021; Hansen & Finlay 2022; Bloem *et al.* 2024), an advantage of using MDN is that they can determine the posterior pdf corresponding to any data set extremely rapidly once the networks have been trained. In other words, varying observed data in Bayesian inference can be accomplished using MDN with almost no additional cost. Prior to the work of Walker & Curtis (2014b) and this current work, if we wished to change prior information we would have to re-train the MDN which typically requires millions of training samples and the calculation of their forward function values. A possible extension of the current work might combine VPR and MDN, in which MDN is used to calculate the old posterior distribution (using a set of prior samples obtained from the old prior pdf) and VPR is used to evaluate any potential new posterior distribution when prior information changes. Under this framework, both prior information and observed data can be replaced efficiently with one single training of an MDN (using the old prior). This opens the possibility that advanced real-time monitoring of subsurface changes and the corresponding uncertainties can be implemented efficiently for some problems.

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

Papamakarios & Murray (2016) introduced a similar approach compared to VPR for likelihood-free inference using MDN. Traditionally one would use a large number (millions) of prior samples and their forward function values to train an MDN. If we are only interested in a posterior pdf for one specific data set, such a strategy is inefficient since most of the prior samples and their forward evaluations would result in near-zero probabilities in that specific posterior pdf. Therefore, Papamakarios & Murray (2016) defined a proposal prior distribution $\tilde{p}(\mathbf{m})$, from which samples are drawn to train an MDN. The proposal prior distribution is updated iteratively to generate samples that are highly informative while training an MDN for a specific observation (for example, if $\tilde{p}(\mathbf{m})$ could be set equal to the true posterior pdf then all samples would be useful). After the training process, the proposal prior is replaced by the original prior pdf to obtain the true posterior pdf–as described in eqs. (5) and (13). However, in that case the corresponding posterior pdf needs to be normalized (Walker & Curtis 2014b), and for cases in which the proposal prior is narrower than the original prior distribution replacing it would lead to a numerical issue of dividing by zero; both issues are likely to be especially prevalent in high dimensional problems and are mitigated in our methodology.

Based on a Markov chain Monte Carlo (McMC) framework, Mosegaard & Tarantola (1995) introduced an approach to draw samples from one probability distribution when one only has samples from another distribution. They achieved this by resampling using a simple Metropolis accept-reject criterion based on the ratio of two probability values $p_{new}(\mathbf{m})/p_{old}(\mathbf{m})$, potentially avoiding further likelihood function evaluations. In this sense, that method can be interpreted as a Monte Carlo version of prior replacement that provides a sampling-based solution, compared to the VPR approach which is based on variational inference. Both approaches of Mosegaard & Tarantola (1995) and in this paper have a similar pre-requisite: that the support of the old prior distribution $p_{old}(\mathbf{m})$ should include that of the new prior so that numerical instability issues are avoided when performing the division $p_{new}(\mathbf{m})/p_{old}(\mathbf{m})$.

In our numerical examples, geological prior information is represented by a Gaussian distribution with a local correlation structure estimated from images of real geology. A direct generalization is to use a mixture of Gaussian distributions to model the geological prior distribution. In addition, normalizing flows (Dinh *et al.* 2015, 2017; Papamakarios *et al.* 2017; Kingma & Dhariwal 2018) are often used as a deep generative model in the machine learning community, which construct a complex probability distribution by passing a simple and analytically known probability distribution (such as a uniform or a standard normal distribution) through a series of invertible and differentiable transforms. After training a normalizing flows model, the prior probability value of any model sample can be evaluated and used in the VPR framework. Future work might explore the use of these methods to build a more sophisticated prior distribution. On the other hand, geological prior information might be simulated through geological processing models (Tetzlaff *et al.* 1989; Paola 2000; Burgess *et al.* 2001; Hill *et al.* 2009; Tetzlaff 2023), which can then be parametrized by advanced neural network models and used during Bayesian inference (Laloy *et al.* 2018; Mosser *et al.* 2020; Levy *et al.* 2022; Scheiter *et al.* 2022; Hillier *et al.* 2023; Liu *et al.* 2024; Bloem *et al.* 2024; Bloem & Curtis 2024). These approaches could lead to more accurate and realistic representations of geological structures and their associated uncertainties.

The main purpose of VPR is to update (replace) prior information efficiently in Bayesian inference. As demonstrated herein, new inversion results can be obtained with no further forward simulation. On the other hand, this also indicates that we cannot obtain new information about data misfit values purely from VPR results - indeed, that is the point of the method. In future, VPR might be used to compare different prior hypotheses (e.g., including different magnitudes of smoothness for a smoothed prior distribution): say we have obtained approximate posterior pdfs for a set of different prior assumptions by applying VPR. For the different inversion results, we could perform a small number of additional forward simulations using posterior samples drawn from VPR results, based on which the different prior hypotheses can be compared, and one or two close-to-optimal options could be selected (one example is presented in Zhao & Curtis 2024c). Although this procedure does require additional forward simulations, it still provides a far more efficient approach to test different prior options compared to carrying out a sequence of independent inversions as is typically done using linearized inversion. In addition, for cases when VPR becomes less accurate (e.g., when the dimensionality and complexity of the inverse problem increase such as in 3D FWI problems), one might use a relatively lower cost forward function with different data types to refine (fine tune) the outcomes obtained from VPR, by invoking Bayes' rule again (Zhao & Curtis 2024c).

5 CONCLUSIONS

We develop a variational prior replacement (VPR) methodology designed to efficiently update prior information in Bayesian inference solutions. This approach involves replacing the existing prior information in a posterior distribution obtained from a previous inference process with a new prior distribution. The new posterior distribution is then found using variational inference. VPR eliminates the need to re-solve Bayesian inverse problems from scratch each time prior information changes. The results from a 2D full waveform inversion example support the effectiveness of VPR for varying prior information, in which VPR provides consistent statistics of the posterior probability distribution compared to those obtained using the conventional prior specific inversion scheme. This similarity holds for individual posterior samples, first- and second-order statistics, as well as simulated waveform data. The key advantage of VPR lies in its computational efficiency: achieving the same results in a matter of minutes compared to two days required by the conventional approach. Additionally, we show that VPR can be used to investigate the impact of different prior distributions on Bayesian inference results. This methodology has significant potential for applications in more computationally demanding inverse problems, such as 3D Bayesian FWI, especially when multiple priors are available or need to be tested and discriminated for the same data set.

ACKNOWLEDGMENTS

We thank the Edinburgh Imaging Project (EIP - https://blogs.ed.ac.uk/imaging/) sponsors (BP and TotalEnergies) for supporting this research. For the purpose of open access, we have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

DATA AVAILABILITY

The code underlying this article will be shared on reasonable request to the corresponding author.

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

REFERENCES

Abadi, M. *et al.*, 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems, preprint (arXiv.1603.04467).

Aghamiry, H., Gholami, A. & Operto, S., 2018. Hybrid Tikhonov+ total-variation regularization for imaging large-contrast media by full-waveform inversion, *Paper presented at the 2018 SEG International Exposition and Annual Meeting,* 14–19 October 2018, Anaheim, CA, USA, Paper Number: SEG-2018-2996968, SEG.

Alyaev, S. & Elsheikh, A.H., 2022. Direct multi-modal inversion of geophysical logs using deep learning, *Earth Space Sci.,* **9**(9), e2021EA002186.

Andrieu, C. & Thoms, J., 2008. A tutorial on adaptive MCMC, *Stat. Comput.,* **18,** 343–373.

Arnold, R. & Curtis, A., 2018. Interrogation theory, *Geophys. J. Int.,* **214**(3), 1830–1846.

Atchadé, Y.F. & Rosenthal, J.S., 2005. On adaptive Markov chain Monte Carlo algorithms, *Bernoulli,* **11**(5), 815–828.

Bates, O., Guasch, L., Strong, G., Robins, T.C., Calderon-Agudo, O., Cueto, C., Cudeiro, J. & Tang, M., 2022. A probabilistic approach to tomography and adjoint state methods, with an application to full waveform inversion in medical ultrasound, *Inverse Problem,* **38**(4), doi:10.1088/1361-6420/ac55ee.

Berti, S., Aleardi, M. & Stucchi, E., 2023. A computationally efficient Bayesian approach to full-waveform inversion, *Geophys. Prospect.,* **72**(2), 580–603.

Bishop, C.M., 1994. *Mixture Density Networks,* Aston Univ.

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

Biswas, R. & Sen, M.K., 2022. Transdimensional 2d full-waveform inversion and uncertainty estimation, preprint (arXiv.2201.09334).

Blei, D.M., Kucukelbir, A. & McAuliffe, J.D., 2017. Variational inference: a review for statisticians, *J. Am. Stat. Assoc.,* **112**(518), 859–877.

Bloem, H. & Curtis, A., 2024. Bayesian geochemical correlation and tomography, *Sci. Rep.,* **14**(1), 9266.

Bloem, H., Curtis, A. & Tetzlaff, D., 2024. Introducing conceptual geological information into Bayesian tomographic imaging, *Basin Res.,* **36**(1), e12811.

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

Boyd, S. & Vandenberghe, L., 2004. *Convex Optimization,* Cambridge Univ. Press.

Burgess, P., Wright, V. & Emery, D., 2001. Numerical forward modelling of peritidal carbonate parasequence development: implications for outcrop interpretation, *Basin Res.,* **13**(1), 1–16.

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

Curtis, A. & Lomax, A., 2001. Prior information, sampling distributions, and the curse of dimensionality, *Geophysics,* **66**(2), 372–378.

de Lima, P.D.S., Corso, G., Ferreira, M.S. & de Araújo, J.M., 2023a. Acoustic full waveform inversion with Hamiltonian Monte Carlo method, *Phys. A: Stat. Mech. Appl.,* **617,** doi:10.1016/j.physa.2023.128618.

de Lima, P.D.S., Ferreira, M.S., Corso, G. & de Araújo, J.M., 2023b. Bayesian time-lapse full waveform inversion using Hamiltonian Monte Carlo, *Geophys. Prospect.,* online ahead of print, doi: 10.1111/1365-2478.13604.

de Wit, R.W., Valentine, A.P. & Trampert, J., 2013. Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks, *Geophys. J. Int.,* **195**(1), 408–422.

Devilee, R., Curtis, A. & Roy-Chowdhury, K., 1999. An efficient, probabilistic neural network approach to solving inverse problems: inverting surface wave velocities for Eurasian crustal thickness, *J. geophys. Res.,* **104**(B12), 28 841–28 857.

Dinh, L., Krueger, D. & Bengio, Y., 2015. NICE: Non-linear independent components estimation, preprint (arXiv.1410.8516).

Dinh, L., Sohl-Dickstein, J. & Bengio, S., 2017. Density estimation using real NVP, preprint (arXiv.1605.08803).

Earp, S. & Curtis, A., 2020. Probabilistic neural network-based 2D travel-time tomography, *Neur. Comput. Appl.,* **32**(22), 17 077–17 095.

Earp, S., Curtis, A., Zhang, X. & Hansteen, F., 2020. Probabilistic neural network tomography across Grane field (North Sea) from surface wave dispersion data, *Geophys. J. Int.,* **223**(3), 1741–1757.

Fichtner, A., Kennett, B.L., Igel, H. & Bunge, H.-P., 2009. Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods, *Geophys. J. Int.,* **179**(3), 1703–1725.

Fichtner, A., Zunino, A. & Gebraad, L., 2019. Hamiltonian Monte Carlo solution of tomographic inverse problems, *Geophys. J. Int.,* **216**(2), 1344–1363.

Fu, X. & Innanen, K.A., 2022. A time-domain multisource Bayesian Markov chain Monte Carlo formulation of time-lapse seismic waveform inversion, *Geophysics,* **87**(4), R349–R361.

Galetti, E., Curtis, A., Baptie, B., Jenkins, D. & Nicolson, H., 2017. Transdimensional Love-wave tomography of the British Isles and shear-velocity structure of the East Irish Sea Basin from ambient-noise interferometry, *Geophys. J. Int.,* **208**(1), 36–58.

Galetti, E., Curtis, A., Meles, G.A. & Baptie, B., 2015. Uncertainty loops in travel-time tomography from nonlinear wave physics, *Phys. Rev. Lett.,* **114**(14), doi:10.1103/PhysRevLett.114.148501.

Gallego, V. & Insua, D.R., 2018. Stochastic gradient MCMC with repulsive forces, Vol. **1050,** 30, preprint (arXiv.1812.00071).

Gebraad, L., Boehm, C. & Fichtner, A., 2020. Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo, *J. geophys. Res.,* **125**(3), e2019JB018428.

Girolami, M. & Calderhead, B., 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods, *J. R. Stat. Soc., B: Stat. Methodol.,* **73**(2), 123–214.

Golub, G.H., Hansen, P.C. & O'Leary, D.P., 1999. Tikhonov regularization and total least squares, *SIAM J. Matrix Anal. Appl.,* **21**(1), 185–194.

Grana, D., Azevedo, L., De Figueiredo, L., Connolly, P. & Mukerji, T., 2022. Probabilistic inversion of seismic data for reservoir petrophysical characterization: Review and examples, *Geophysics,* **87**(5), M199–M216.

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

Guan, X., Wang, X., Wu, H., Yang, Z & Yu, P., 2024. Efficient Bayesian inference using physics-informed invertible neural networks for inverse problems, *Mach. learn.: sci. technol.,* **5**(3), 035026, doi: 10.1088/2632-2153/ad5f74.

Guo, F., Wang, X., Fan, K., Broderick, T. & Dunson, D.B., 2016. Boosting variational inference, *Adv. Neur. Inf. Proc. Syst.,* preprint (arXiv:1611.05559).

Guo, P., Visser, G. & Saygin, E., 2020. Bayesian trans-dimensional full waveform inversion: synthetic and field data application, *Geophys. J. Int.,* **222**(1), 610–627.

Hansen, T.M. & Finlay, C.C., 2022. Use of machine learning to estimate statistics of the posterior distribution in probabilistic inverse problems - an application to airborne EM data, *J. geophys. Res.,* **127**(11), e2022JB024703.

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

Hill, J., Tetzlaff, D., Curtis, A. & Wood, R., 2009. Modeling shallow marine carbonate depositional systems, *Comput. Geosci.,* **35**(9), 1862–1874.

Hillier, M., Wellmann, F., de Kemp, E.A., Brodaric, B., Schetselaar, E. & Bédard, K., 2023. Geoinr 1.0: an implicit neural network approach to three-dimensional geological modelling, *Geosci. Model Dev.,* **16**(23), 6987–7012.

Izzatullah, M., Alali, A., Ravasi, M. & Alkhalifah, T., 2024. Physics reliable frugal uncertainty analysis for full waveform inversion, *Geophys. Prospect.,* **72**(7), 2718–2738.

John, J.S., 2012. Folded gyprock (Castile formation, Upper Permian; State line outcrop, southern Eddy county, New Mexico, USA) 7, https://www.flickr.com/photos/jsjgeology/8280544003.

Käufl, P.,, & Trampert, J., 2016. Solving probabilistic inverse problems rapidly with prior samples, *Geophys. J. Int.,* **205**(3), 1710–1728.

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

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

Khoshkholgh, S., Zunino, A. & Mosegaard, K., 2021. Informed proposal Monte Carlo, *Geophys. J. Int.,* **226**(2), 1239–1248.

Khoshkholgh, S., Zunino, A. & Mosegaard, K., 2022. Full-waveform inversion by informed-proposal Monte Carlo, *Geophys. J. Int.,* **230**(3), 1824–1833.

Kingma, D.P. & Dhariwal, P., 2018. Glow: generative flow with invertible 1x1 convolutions, in *Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018),* pp. 10 215–10 224, eds Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R., NIPS.

Kirkpatrick, S., Gelatt, C.D., Jr & Vecchi, M.P., 1983. Optimization by simulated annealing, *Science,* **220**(4598), 671–680.

Kobyzev, I., Prince, S.J. & Brubaker, M.A., 2020. Normalizing flows: An introduction and review of current methods, *IEEE T. Pattern Anal. Mach. Intell.,* **43**(11), 3964–3979.

Kotsi, M., Malcolm, A. & Ely, G., 2020. Uncertainty quantification in time-lapse seismic imaging: a full-waveform approach, *Geophys. J. Int.,* **222**(2), 1245–1263.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D.M., 2017. Automatic differentiation variational inference, *J. Mach. Learn. Res.,* **18**(1), 430–474.

Kullback, S. & Leibler, R.A., 1951. On information and sufficiency, *Ann. Math. Stat.,* **22**(1), 79–86.

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

Levy, S., Laloy, E. & Linde, N., 2022. Variational Bayesian inference with complex geostatistical priors using inverse autoregressive flows, *Comput. Geosci.,* **171,** doi:10.1016/j.cageo.2022.105263.

Liu, M., Grana, D. & Mukerji, T., 2024. Geostatistical inversion for subsurface characterization using Stein variational gradient descent with autoencoder neural network: an application to geologic carbon sequestration, *J. geophys. Res. ,* **129**(7), doi:10.1029/2024JB029073.

Liu, Q. & Wang, D., 2016. Stein variational gradient descent: a general purpose Bayesian inference algorithm, in *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016),* Barcelona, Spain, 2378–2386, NIPS.

Locatello, F., Dresdner, G., Khanna, R., Valera, I. & Rätsch, G., 2018. Boosting black box variational inference, *Adv. Neur. Inf. Proc. Syst.,* **31**.

Lomas, A., Luo, S., Irakarama, M., Johnston, R., Vyas, M. & Shen, X., 2023. 3D probabilistic full waveform inversion: application to Gulf of Mexico field data, in *Proceedings of the 84th EAGE Annual Conference & Exhibition,* Vol. 2023, European Association of Geoscientists & Engineers, pp. 1–5.

Lubo-Robles, D., Ha, T., Lakshmivarahan, S., Marfurt, K.J. & Pranter, M.J., 2021. Exhaustive probabilistic neural network for attribute selection and supervised seismic facies classification, *Interpretation,* **9**(2), T421–T441.

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

Martin, G.S., Wiley, R. & Marfurt, K.J., 2006. Marmousi2: an elastic upgrade for Marmousi, *Leading Edge,* **25**(2), 156–166.

Meier, U., Curtis, A. & Trampert, J., 2007a. Fully nonlinear inversion of fundamental mode surface waves for a global crustal model, *Geophys. Res. Lett.,* **34**(16), doi:10.1029/2007GL030989.

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

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.

Miller, A.C., Foti, N.J. & Adams, R.P., 2017. Variational boosting: iteratively refining posterior approximations, in *International Conference on Machine Learning,* PMLR, pp. 2420–2429.

Mosegaard, K. & Sambridge, M., 2002. Monte Carlo analysis of inverse problems, *Inverse Problem,* **18**(3), R29.

Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.,* **100**(B7), 12 431–12 447.

Mosher, S.G., Eilon, Z., Janiszewski, H. & Audet, P., 2021. Probabilistic inversion of seafloor compliance for oceanic crustal shear velocity structure using mixture density neural networks, *Geophys. J. Int.,* **227**(3), 1879–1892.

Mosser, L., Dubrule, O. & Blunt, M.J., 2020. Stochastic seismic waveform inversion using generative adversarial networks as a geological prior, *Math. Geosci.,* **52**(1), 53–79.

Nawaz, A. & Curtis, A., 2016. Bayesian inversion of seismic attributes for geological facies using a hidden Markov model, *Geophys. J. Int.,* **208**(2), 1184–1200.

Nawaz, A. & Curtis, A., 2018. Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies, *Geophys. J. Int.,* **214**(2), 845–875.

Nawaz, A. & Curtis, A., 2019. Rapid discriminative variational Bayesian inversion of geophysical data for the spatial distribution of geological properties, *J. geophys. Res.,* **124**(6), 5867–5887.

Nawaz, A., Curtis, A., Shahraeeni, M.S. & Gerea, C., 2020. Variational Bayesian inversion of seismic attributes jointly for geological facies and petrophysical rock properties, *Geophysics,* **85**(4), 1–78.

Paola, C., 2000. Quantitative models of sedimentary basin filling, *Sedimentology,* **47,** 121–178.

Papamakarios, G. & Murray, I., 2016. Fast ε-free inference of simulation models with Bayesian conditional density estimation, *Adv. Neur. Inf. Proc. Syst.,* **29**.

Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S. & Lakshminarayanan, B., 2021. Normalizing flows for probabilistic modeling and inference, *J. Mach. Learn. Res.,* **22**(57), 1–64.

Papamakarios, G., Pavlakou, T. & Murray, I., 2017. Masked autoregressive flow for density estimation, in *Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019),* pp. 2338–2347, eds Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R., NIPS.

Paszke, A. *et al.*, 2019. Pytorch: an imperative style, high-performance deep learning library, *Adv. Neur. Inf. Proc. Syst.,* **32**.

Plessix, R.-E., 2006. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications, *Geophys. J. Int.,* **167**(2), 495–503.

Press, F., 1968. Earth models obtained by Monte Carlo inversion, *J. geophys. Res.,* **73**(16), 5223–5234.

Propp, J.G. & Wilson, D.B., 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics, *Rand. Struct. Algorithms,* **9**(1-2), 223–252.

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

Rezende, D.J. & Mohamed, S., 2015. Variational inference with normalizing flows, *International conference on machine learning*. 1530–1538. PMLR. preprint (arXiv:1505.05770).

Sambridge, M. & Drijkoningen, G., 1992. Genetic algorithms in seismic waveform inversion, *Geophys. J. Int.,* **109**(2), 323–342.

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

Sambridge, M., 1999a. Geophysical inversion with a neighbourhood algorithm - I. Searching a parameter space, *Geophys. J. Int.,* **138**(2), 479–494.

Sambridge, M., 1999b. Geophysical inversion with a neighbourhood algorithm - II. Appraising the ensemble, *Geophys. J. Int.,* **138**(3), 727–746.

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.

Scheiter, M., Valentine, A. & Sambridge, M., 2022. Upscaling and downscaling Monte Carlo ensembles with generative models, *Geophys. J. Int.,* **230**(2), 916–931.

Sen, M.K. & Stoffa, P.L., 2013. *Global Optimization Methods in Geophysical Inversion,* Cambridge Univ. Press..

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

Shahraeeni, M.S. & Curtis, A., 2011. Fast probabilistic nonlinear petrophysical inversion, *Geophysics,* **76**(2), E45–E58.
Shahraeeni, M.S., Curtis, A. & Chao, G., 2012. Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data, *Geophysics,* **77**(3), O1–O19.
Siahkoohi, A., Rizzuti, G., Louboutin, M., Witte, P.A. & Herrmann, F.J., 2021. Preconditioned training of normalizing flows for variational inference in inverse problems, preprint (arXiv:2101.03709).
Siahkoohi, A., Rizzuti, G., Orozco, R. & Herrmann, F.J., 2023. Reliable amortized variational inference with physics-based latent distribution correction, *Geophysics,* **88**(3), 1–137.
Siahkoohi, A., Rizzuti, G., Witte, P.A. & Herrmann, F.J., 2020. Faster uncertainty quantification for inverse problems with conditional normalizing flows, preprint (arXiv:2007.07985).
Sjölund, J., 2023. A tutorial on parametric variational inference, preprint (arXiv:2301.01236).
Smith, J.D., Ross, Z.E., Azizzadenesheli, K. & Muir, J.B., 2022. HypoSVI: hypocentre inversion with Stein variational inference and physics informed neural networks, *Geophys. J. Int.,* **228**(1), 698–710.
Stoffa, P.L. & Sen, M.K., 1991. Nonlinear multiparameter optimization using genetic algorithms; inversion of plane-wave seismograms, *Geophysics,* **56**(11), 1794–1810.
Strutz, D. & Curtis, A., 2024. Variational Bayesian experimental design for geophysical applications: seismic source location, amplitude versus offset inversion, and estimating CO2 saturations in a subsurface reservoir, *Geophys. J. Int.,* **236**(3), 1309–1331.
Sun, L., Wang, L., Xu, G. & Wu, Q., 2023. A new method of variational Bayesian slip distribution inversion, *J. Geod.,* **97**(1), 10.
Sun, Y. & Williamson, P., 2024. Invertible neural networks for uncertainty quantification in refraction tomography, *Leading Edge,* **43**(6), 358–366.
Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysics,* **49**(8), 1259–1266.
Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* Vol. **89,** SIAM.
Tetzlaff, D., 2023. Stratigraphic forward modeling software package for research and education, preprint (arXiv:2302.05272).
Tetzlaff, D.M., *et al.*, 1989. *Simulating Clastic Sedimentation,* Vol. **1110,** Springer.
Valentine, A.P. & Sambridge, M., 2023. Emerging directions in geophysical inversion, in *Applications of Data Assimilation and Inverse Problems in the Earth Sciences,* pp. 9–26, eds Ismail-Zadeh, A., Castelli, F., Jones, D. & Sanchez, S., Cambridge Univ. Press.
Virieux, J. & Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics, *Geophysics,* **74**(6), WCC1–WCC26.
Visser, G., Guo, P. & Saygin, E., 2019. Bayesian transdimensional seismic full-waveform inversion with a dipping layer parameterization, *Geophysics,* **84**(6), R845–R858.
Walker, M. & Curtis, A., 2014a. Spatial Bayesian inversion with localized likelihoods: an exact sampling alternative to McMC, *J. geophys. Res.,* **119**(7), 5741–5761.
Walker, M. & Curtis, A., 2014b. Varying prior information in Bayesian inversion, *Inverse Problem,* **30**(6), doi:10.1088/0266-5611/30/6/065002.
Wang, W., McMechan, G.A. & Ma, J., 2023. Re-weighted variational full waveform inversions, *Geophysics,* **88**(4), 1–61.
Wang, Y., Niu, L., Zhao, L., Wang, B., He, Z., Zhang, H., Chen, D. & Geng, J., 2022. Gaussian mixture model deep neural network and its application in porosity prediction of deep carbonate reservoir, *Geophysics,* **87**(2), M59–M72.
Wang, Y., Zhou, H., Zhao, X., Zhang, Q., Zhao, P., Yu, X. & Chen, Y., 2019. Cu Q-RTM: a CUDA-based code package for stable and efficient Q-compensated reverse time migration, *Geophysics,* **84**(1), F1–F15.
Welling, M. & Teh, Y.W., 2011. Bayesian learning via stochastic gradient Langevin dynamics, in *Proceedings of the 28th International Conference on Machine Learning (ICML-11),* Citeseer, pp. 681–688.
Wolpert, D.H. & Macready, W.G., 1997. No free lunch theorems for optimization, *IEEE T. Evol. Comput.,* **1**(1), 67–82.
Yin, Z., Orozco, R. & Herrmann, F.J., 2024a. WISER: multimodal variational inference for full-waveform inversion without dimensionality reduction, preprint (arXiv:2405.10327).
Yin, Z., Orozco, R., Louboutin, M. & Herrmann, F.J., 2024b. WISE: full-waveform variational inference via subsurface extensions, *Geophysics,* **89**(4), 1–31.
Zhang, X. & Curtis, A., 2020a. Seismic tomography using variational inference methods, *J. geophys. Res.,* **125**(4), e2019JB018589.
Zhang, X. & Curtis, A., 2020b. Variational full-waveform inversion, *Geophys. J. Int.,* **222**(1), 406–411.
Zhang, X. & Curtis, A., 2021a. Bayesian geophysical inversion using invertible neural networks, *J. geophys. Res.,* **126**(7), e2021JB022320.
Zhang, X. & Curtis, A., 2021b. Bayesian full-waveform inversion with realistic priors, *Geophysics,* **86**(5), 1–20.
Zhang, X., Lomas, A., Zhou, M., Zheng, Y. & Curtis, A., 2023. 3D Bayesian variational full waveform inversion, *Geophys. J. Int.,* **234**(1), 546–561.
Zhang, X., Nawaz, A., Zhao, X. & Curtis, A., 2021. An introduction to variational inference in geophysical inverse problems, *Adv. Geophys.,* **62,** 73–140.
Zhao, X. & Curtis, A., 2024a. Bayesian inversion, uncertainty analysis and interrogation using boosting variational inference, *J. geophys. Res.,* **129**(1), e2023JB027789.
Zhao, X. & Curtis, A., 2024b. Physically structured variational inference for Bayesian full waveform inversion, ESS Open Archive, 28 May 2024, doi:10.22541/essoar.171691139.96106369/v1.
Zhao, X. & Curtis, A., 2024c. Efficient 3D Bayesian Full Waveform Inversion and Analysis of Prior Hypotheses, Preprint (arXiv:2409.09746) https://arxiv.org/abs/2409.09746.
Zhao, X., Curtis, A. & Zhang, X., 2021. Bayesian seismic tomography using normalizing flows, *Geophys. J. Int.,* **228**(1), 213–239.
Zhao, X., Curtis, A. & Zhang, X., 2022a. Interrogating subsurface structures using probabilistic tomography: an example assessing the volume of Irish Sea basins, *J. geophys. Res.,* **127**(4), e2022JB024098.
Zhao, X., Zhou, H., Chen, H. & Wang, Y., 2020. Domain decomposition for large-scale viscoacoustic wave simulation using localized pseudo-spectral method, *IEEE Trans. Geosci. Remote Sens.,* **59**(3), 2666–2679.
Zhao, Z. & Sen, M.K., 2021. A gradient-based Markov chain Monte Carlo method for full-waveform inversion and uncertainty analysis, *Geophysics,* **86**(1), R15–R30.
Zhao, Z., Sen, M.K., Denel, B., Sun, D. & Williamson, P., 2022b. A hybrid optimization framework for seismic full waveform inversion, *J. geophys. Res.,* **127**(8), e2022JB024483.
Zidan, A., Li, Y. & Cheng, A., 2022. Regularized seismic amplitude inversion via variational inference, *Geophys. Prospect.,* **70**(9), 1507–1527.

APPENDIX A: VERIFYING VPR USING A NORMAL INITIAL PRIOR DISTRIBUTION

In this appendix, we present a second example to test the performance of the proposed variational prior replacement (VPR) methodology. We define a diagonal Gaussian distribution as the old prior pdf, with mean and standard deviation are calculated from the uniform prior distribution used in the main text. Fig. A1(a) shows the inversion results obtained using this normal prior distribution. From top to bottom, each panel represents a random posterior sample, the mean velocity, standard deviation and the relative error maps of the posterior distribution, where the relative error is calculated as the difference between the mean and true velocity models divided by the standard deviation at each point. Given this inversion result, we perform VPR by replacing a smoothed version of the normal distribution (which is defined using exactly the same way as that expressed in eq. (20)) by the original normal prior distribution, and display the corresponding results in Fig. A1(c). Similarly to the main text, in Fig. A1(b) we display prior specific inversion (PSI) results obtained using this same smoothed prior distribution. Since Figs A1(b) and (c) present highly consistent results, we conclude that VPR is accurate and effective in this example.

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

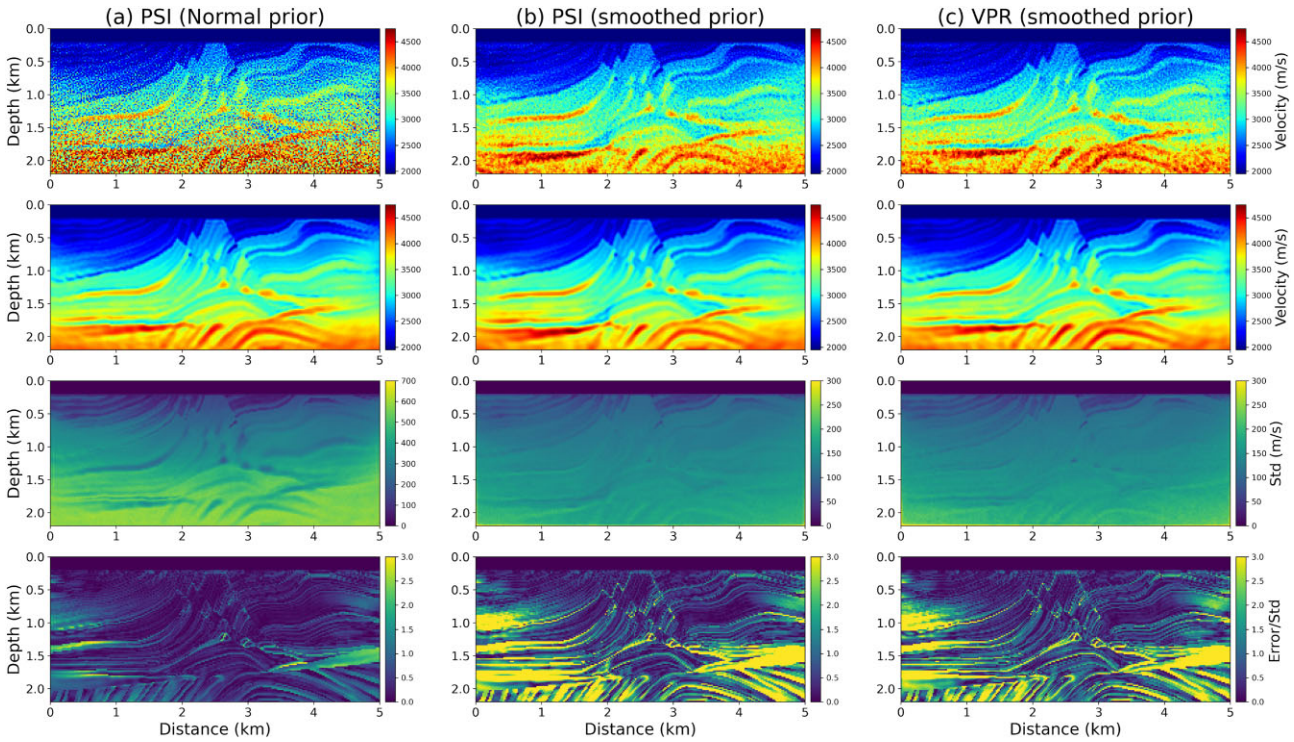

Figure A1. (a) Prior specific inversion (PSI) results obtained using a diagonal Gaussian prior distribution defined in this Appendix. (b) PSI results obtained using a smoothed version of the normal prior distribution. (c) Variational prior replacement (VPR) results obtained by replacing the normal prior distribution by the smoothed prior. In each column, a random posterior sample, mean velocity, standard deviation and relative error maps are displayed from top to bottom row, respectively.

Downloaded from https://academic.oup.com/gji/article/239/2/1236/7755434 by University of Edinburgh user on 25 September 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of The Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.