Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Original Article

Bayesian model selection in hydrogeophysics: Application to conceptual subsurface models of the South Oyster Bacterial Transport Site, Virginia, USA

Carlotta Brunetti[a,*], Niklas Linde[a], Jasper A. Vrugt[b,c]

[a] Applied and Environmental Geophysics Group, Institute of Earth Sciences, University of Lausanne, Lausanne, 1015, Switzerland
[b] Department of Civil and Environmental Engineering, University of California Irvine, Irvine, CA, 92697-2175, USA
[c] Department of Earth Systems Science, University of California Irvine, Irvine, CA, 92697-2175, USA

ARTICLE INFO

Article history:
Received 22 July 2016
Revised 6 February 2017
Accepted 6 February 2017
Available online 7 February 2017

Keywords:
Bayesian model selection
Evidence
Laplace-Metropolis
Importance sampling
Conceptual model
Ground-penetrating radar

ABSTRACT

Geophysical data can help to discriminate among multiple competing subsurface hypotheses (conceptual models). Here, we explore the merits of Bayesian model selection in hydrogeophysics using crosshole ground-penetrating radar data from the South Oyster Bacterial Transport Site in Virginia, USA. Implementation of Bayesian model selection requires computation of the marginal likelihood of the measured data, or evidence, for each conceptual model being used. In this paper, we compare three different evidence estimators, including (1) the brute force Monte Carlo method, (2) the Laplace-Metropolis method, and (3) the numerical integration method proposed by Volpi et al. (2016). The three types of subsurface models that we consider differ in their treatment of the porosity distribution and use (a) horizontal layering with fixed layer thicknesses, (b) vertical layering with fixed layer thicknesses and (c) a multi-Gaussian field. Our results demonstrate that all three estimators provide equivalent results in low parameter dimensions, yet in higher dimensions the brute force Monte Carlo method is inefficient. The isotropic multi-Gaussian model is most supported by the travel time data with Bayes factors that are larger than 10^{100} compared to conceptual models that assume horizontal or vertical layering of the porosity field.

1. Introduction

Geophysical methods are used widely in near-surface applications, because of their innate ability to infer, with high resolution, the properties and spatial structure of the subsurface. Geophysical data, for instance, warrant a detailed characterization of the hydrologic properties of the vadose zone and aquifers (Binley et al., 2010; 2015; Hubbard and Linde, 2011; Hubbard and Rubin, 2005). Most published studies in the hydrogeophysical literature rely on a single conceptual representation of the subsurface, without recourse to explicit treatment of the actual uncertainty associated with the choice of a single conceptual model (Linde, 2014; Linde et al., 2015). Geophysics-based model selection has received relatively limited attention, which is somewhat surprising, as geophysical data contain a wealth of information about the structure of the subsurface. In contrast to current practice, we should not rely only on a single conceptualization and parameterization of the subsurface, but instead determine as well the proper spatial arrangement of variables of interest such as porosity and moisture content. One approach of doing this is to implement model selection, and use the geophysical data to provide guidance about which representation of the subsurface is most supported by the available data among a set of competing conceptual models (Linde, 2014). Such an approach will not only enhance the fidelity of our subsurface investigations, but will also further promulgate and disseminate the importance of geophysical data in hydrologic and environmental studies. By providing knowledge about suitable geostatistical descriptions of the subsurface, model selection might also help in closing the gap in scale between plot-based geophysical investigations and the much larger spatial domains relevant to water resources management, contaminant transport and risk assessment. In this way, geophysics is used to define an appropriate geostatistical model that can later be used to produce unconditional geostatistical realizations at larger scales.

Many different approaches have been suggested in the statistical literature to help select the ”best” model among a group of competing hypotheses. This includes frequentist and Bayesian solutions. The application of such approaches to geophysical studies

* Corresponding author.
E-mail addresses: Carlotta.Brunetti@unil.ch (C. Brunetti), Niklas.Linde@unil.ch (N. Linde), jasper@uci.edu (J.A. Vrugt).

http://dx.doi.org/10.1016/j.advwatres.2017.02.006
0309-1708/© 2017 Elsevier Ltd. All rights reserved.

has its own special challenges. For instance, a parameter-rich, but geologically-unrealistic model may fit the data equally well or perhaps even better than a more parsimonious model with more appropriate conceptualization of the subsurface (Rosenkrantz, 1977). What is more, the decision about which model is favoured, is also heavily influenced by the choice of the models' prior parameter distribution, even for geophysical data comprised of many different measurements. With the use of an inappropriate prior the model can be made to fit the data arbitrarily poorly, changing fundamentally our opinion about which model should be favoured, a phenomenon known as the Jeffreys–Lindley paradox (Jeffreys, 1939; Lindley, 1957).

To describe accurately this trade-off between model complexity and goodness of fit, we here use Bayesian model selection, and investigate in detail the denominator in Bayes theorem. This normalizing constant, referred to as the evidence, marginal likelihood or integrated likelihood, conveys all information necessary to determine which of the competing subsurface models (given their prior parameter distributions) is most supported by the geophysical data. The conceptual model with the largest evidence over the prior model space is the one that is most supported by the experimental data. The foundation of Bayesian model selection originates from Jeffreys (Jeffreys, 1935; 1939) and builds on the principles of Occam's razor, that is, parsimony is favoured over complexity. In other words, if two models exhibit a (nearly) equivalent fit to the data, the model with the least number of "free" parameters is preferred statistically (Gull, 1988; Jefferys and O. Berger, 1992; Jeffreys, 1939; MacKay, 1992). Statisticians prefer the use of so-called Bayes factors (Kass and Raftery, 1995) to quantify the odds of each model with respect to every other competing model. This Bayes factor of two models A and B, is equivalent to the ratio of the evidences of both models. The larger the value of this ratio, the stronger the support for hypothesis A. In cases when the evidence values are of similar magnitude (e.g., within the same one or two orders of magnitude), then it is recommended to use Bayesian model averaging to combine predictions from different conceptual models and, thus, obtain a more appropriate description of posterior parameter uncertainty (Hoeting et al., 1999).

Another distinct advantage of Bayesian model selection is that model comparison is relative to the existing conceptual models at hand, and consequently, the "true" model does not have to be part of the ensemble considered for hypothesis testing. To paraphrase Box and Draper (1987): *All our conceptual models are wrong, but some are useful. It is the task of Bayesian model selection to determine which of the considered conceptual models is the most useful.* Of course, the answer to which model is most useful depends critically on the purpose and intended goal of model application. Within the realm of model selection we can, however, answer the question of which model is most supported by the available data. Yet, this task is not particularly easy for subsurface models, as the integral of the posterior parameter distribution is, in general, high-dimensional and without analytic solution. This probably explains why Bayesian model selection is seldom used in hydrogeophysics and near-surface geophysics. Instead, we have to resort to numerical methods to approximate the value of the evidence for each competing conceptual model. Gelfand and Dey (1994) suggest that the integral of the posterior distribution can be estimated via numerical integration using, for instance, Monte Carlo methods (Hammersley and Handscomb, 1964), asymptotic solutions (e.g., Bayesian information criterion, BIC) (Schwarz et al., 1978) or Laplace's method (De Bruijn, 1970). In the field of geophysics, BIC (Dettmer et al., 2009), annealed importance sampling (Dettmer et al., 2010) and the deviance information criterion, DIC, (Spiegelhalter et al., 2002; Steininger et al., 2014) have been used for calculation of the evidence.

In a separate line of research, transdimensional (or reversible jump) Markov chain Monte Carlo (MCMC) methods (Green, 1995) are receiving a surge of attention to determine the optimal complexity (number of parameters) in geophysical modeling investigations (e.g., Bodin and Sambridge (2009); Bodin et al. (2012); Sambridge et al. (2006); Steininger et al. (2014)). In reversible jump MCMC, the number of model parameters is treated as an unknown and parsimony is preferred as this method incorporates directly the evidence in its calculations which makes it extremely efficient for model selection. Notwithstanding this progress made, transdimensional MCMC is poorly adaptable to situations with multiple different conceptual models that each use a different geologic description (structure) of the target of interest (Chib and Jeliazkov, 2001). Moreover, this method performs relative ranking of the considered conceptual models, which implies that the whole inversion procedure must be re-run if additional candidate models are to be considered at a later stage.

In the field of hydrology, metrics such as Akaike's information criterion (AIC) (Akaike, 1973), BIC, and Kashyap's information criterion (KIC) (Kashyap, 1982) are used widely to select the most adequate model (Li and Tsai, 2009; Marshall et al., 2005; Tsai and Li, 2008; Ye et al., 2010). A recent study by Schöniger et al. (2014) elucidates that AIC and BIC do a rather poor job in ranking hydrologic models. The authors of this study therefore concluded that AIC and BIC are a relatively poor proxy of the evidence. The same study found that the brute force Monte Carlo method provides the most accurate and bias-free estimates of the evidence. Yet, this method is not particularly adequate in high dimensions and for peaky posteriors. What is more, the brute force Monte Carlo method is known to be affected by the so-called curse of dimensionality which degenerates the evidence estimates and make them unusable in high dimensions (Lewis and Raftery, 1997). In cases where reliable brute force Monte Carlo integration is infeasible, Schöniger et al. (2014) promote the use of KIC for model selection, evaluated at the maximum a-posteriori (MAP) density parameter values of the posterior distribution. Note that the KIC is a simple transform of evidence estimates obtained by the Laplace-Metropolis method (Lewis and Raftery, 1997).

The purpose of this study is twofold. In the first place, we investigate to what extent evidence estimates and Bayes factors derived for moderately high parameter dimensionalities (i.e., up to 105 unknowns) can be used to perform Bayesian model selection in synthetic and real-world case studies. For this purpose, we compare evidence estimates computed by (1) the brute force Monte Carlo method (Hammersley and Handscomb, 1964), (2) the Laplace-Metropolis method (Lewis and Raftery, 1997) and (3) the Gaussian mixture importance sampling (GMIS) estimator of Volpi et al. (2016). This latter method approximates the evidence by importance sampling from a Gaussian mixture model fitted to a large sample of posterior solutions generated with the DREAM$_{(ZS)}$ algorithm (Laloy and Vrugt, 2012; Vrugt, 2016; Vrugt et al., 2008). Then, we present an application of Bayesian model selection to subsurface modeling using geophysical data from the South Oyster Bacterial Transport Site in Virginia (USA) (Chen et al., 2001; 2004; Hubbard et al., 2001; Linde et al., 2008; Linde and Vrugt, 2013). These data consist of travel time observations made by crosshole ground-penetrating radar (GPR), and exhibit small measurement errors typical of most near-surface geophysical sensing methods.

2. Theory and methods

2.1. Bayesian inference with MCMC

Given n measurements, $\widetilde{\mathbf{Y}} = \{\widetilde{y}_1, \ldots, \widetilde{y}_n\}$, and a d-dimensional vector of model parameters, $\theta = \{\theta_1, \ldots, \theta_d\}$, it is possible to back out the posterior probability density function (pdf) of the parameters, $p(\theta|\widetilde{\mathbf{Y}})$, via Bayes theorem

$$p(\theta|\widetilde{\mathbf{Y}}) = \frac{p(\theta)p(\widetilde{\mathbf{Y}}|\theta)}{p(\widetilde{\mathbf{Y}})}, \tag{1}$$

where, $p(\theta)$ signifies the prior pdf, $L(\theta|\widetilde{\mathbf{Y}}) \equiv p(\widetilde{\mathbf{Y}}|\theta)$, denotes the likelihood function, and $p(\widetilde{\mathbf{Y}})$ is equivalent to the marginal likelihood, or evidence. The larger the likelihood the better the model, $\mathcal{F}(\theta)$, explains the observed data, $\widetilde{\mathbf{Y}}$. Bayesian model selection can be carried out for any type of likelihood function. However, in this work, we conveniently assume that the error residuals, $E(\theta) = \{e_1(\theta), \ldots, e_n(\theta)\}$, are normally distributed with constant variance and negligible covariance. These three assumptions lead to the following definition of the likelihood function:

$$L(\theta|\widetilde{\mathbf{Y}}, \sigma_{\widetilde{\mathbf{Y}}}) = \left(\sqrt{2\pi\sigma^2_{\widetilde{\mathbf{Y}}}}\right)^{-n} \exp\left[-\frac{1}{2}\sum_{h=1}^{n}\left(\frac{\mathcal{F}_h(\theta) - \widetilde{y}_h}{\sigma_{\widetilde{\mathbf{Y}}}}\right)^2\right], \tag{2}$$

where $\sigma_{\widetilde{\mathbf{Y}}}$ denotes the standard deviation of the measurement data error. This entity can be fixed a-priori by the user if deemed appropriate, or alternatively, the measurement data error can be treated as nuisance variable and the value of $\sigma_{\widetilde{\mathbf{Y}}}$ is inferred jointly with the d-vector of model parameters, θ. The Gaussian likelihood function of Eq. (2) has found widespread application and use in the field of geophysics, nevertheless it is important to stress that the error residuals hardly ever satisfy the rather restrictive assumptions of normality, constant variance, and lack of serial correlation. The Gaussian likelihood in Eq. (2) is sufficient, though, to illustrate the power and usefulness of Bayesian model selection.

The prior pdf, $p(\theta)$, quantifies our knowledge about the expected distribution of the model parameters before considering the observed data. The evidence, $p(\widetilde{\mathbf{Y}})$, acts as a normalization constant of the posterior distribution, and for fixed model parameterizations, is therefore often ignored in Bayesian inference. The posterior pdf, $p(\theta|\widetilde{\mathbf{Y}})$, for a given conceptual model, quantifies the probability density of a vector with parameter values given the initial knowledge embedded in the prior distribution and the information provided by the measurement data via the likelihood. In the absence of closed-form analytic solutions of the posterior distribution, MCMC methods are often used to approximate this distribution using sampling (Hastings, 1970; Metropolis et al., 1953; Robert and Casella, 2013; Vrugt, 2016).

2.2. Evidence and Bayes factor

Bayesian hypothesis testing uses Bayes factors (Kass and Raftery, 1995) to determine which conceptual model is most supported by the available data, and prior distribution. These Bayes factors quantify the odds of two competing models. For the time being, let us assume that we have two competing hypotheses, η_1 and η_2, that differ in their spatial description of the main variable of interest, say porosity. The first hypothesis (model) could assume horizontal layering of the porosity field, whereas the second model adopts a multi-Gaussian description of the spatial configuration of the porosity values. Now the Bayes factor (”odds”) of η_1 with respect to the alternative hypothesis, η_2, or $B_{(\eta_1,\eta_2)}$, can be calculated using

$$B_{(\eta_1,\eta_2)} = \frac{p(\widetilde{\mathbf{Y}}|\eta_1)}{p(\widetilde{\mathbf{Y}}|\eta_2)}, \tag{3}$$

which is simply equivalent to the ratio of the evidences, $p(\widetilde{\mathbf{Y}}|\eta_1)$ and $p(\widetilde{\mathbf{Y}}|\eta_2)$, of the two conceptual models. It then logically follows that the Bayes factor of model two, or the alternative hypothesis η_2, is equal to the reciprocal of $B_{(\eta_1,\eta_2)}$.

The evidence (scalar) of a given conceptual model, η_l, is defined as the (multidimensional) integral of the likelihood function over the prior distribution

$$p(\widetilde{\mathbf{Y}}|\eta_l) = \int L(\theta_l, \eta_l|\widetilde{\mathbf{Y}})p(\theta_l|\eta_l)d\theta_l \qquad l = 1, 2. \tag{4}$$

In practice, it is often not necessary to integrate over the entire prior distribution, as large portions of this space are made up of areas with a negligible posterior density whose contributions to the integral of Eq. (4) are negligibly small. Instead, we can restrict our attention to those areas of the parameter space occupied by the posterior distribution.

Table 1
Interpretation of Kass and Raftery (1995) for the Bayes factor of two conceptual models η_1 and η_2.

$2\log B_{(\eta_1,\eta_2)}$	$B_{(\eta_1,\eta_2)}$	Evidence against η_2
0 to 2	1 to 3	barely worth mentioning
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
$>$ 10	$>$ 150	very strong

It should be evident from the above that models with large Bayes factors are preferred statistically. Indeed, the subsurface conceptual model with largest value of its evidence is most supported by the geophysical data, $\widetilde{\mathbf{Y}}$. In practice, however the computed Bayes factors might not differ substantially from unity and each other to warrant selection of a single model. Bayes factors differ most from each other if relatively simple models are used with widely different characterizations of the subsurface as their flexibility is insufficient to compensate for epistemic errors due to improper system representation and conceptualization. This inability introduces relatively large differences in the models' quality of fit, and consequently their Bayes factors, which simplifies model selection. Highly parameterized models on the contrary, have a much improved ability to correct for system misrepresentation, thereby making it more difficult to judge which hypothesis is preferred statistically. Nevertheless, poor conceptual models should exhibit relatively low Bayes factors in response to their relatively low likelihoods.

The Bayes factor is a sufficient statistic for hypothesis testing, yet renders necessary the definition of ”formal” guidelines on how to interpret its value before we can proceed with model selection. Table 1 articulates an interpretation of the Bayes factor as advocated by Kass and Raftery (1995). This interpretation differentiates four (increasing) levels of support for proposition η_1 relative to η_2. In general, the evidence in favor of η_1 increases with the value of its Bayes factor. Thus, the larger the value of $B_{(\eta_1,\eta_2)}$, the more the data $\widetilde{\mathbf{Y}}$ supports the hypothesis η_1 relative to η_2, and the easier it becomes to reject this alternative hypothesis. It is suggested that the Bayes factor must be larger than 3 (or smaller than 1/3) to discriminate positively among two competing hypotheses.

Unfortunately, the integral in Eq. (4) cannot be derived by analytic means nor by analytic approximation, and we therefore resort to numerical methods to calculate the evidence of each conceptual model. In the next section, we review briefly three different methods for estimating the evidence, including the brute force Monte Carlo method (BFMC), the Laplace-Metropolis (LM) method and the Gaussian mixture importance sampling (GMIS) approach recently developed by Volpi et al. (2016).

2.2.1. Brute force Monte Carlo method

The BFMC method (Hammersley and Handscomb, 1964) approximates the evidence in Eq. (4) as an average of the likelihoods of N different samples drawn randomly from the (multivariate) prior distribution (Kass and Raftery, 1995)

$$p_{\text{BFMC}}(\widetilde{\mathbf{Y}}) \approx \frac{1}{N}\sum_{i=1}^{N} L(\theta_i|\widetilde{\mathbf{Y}}). \tag{5}$$

The validity of this estimator is ensured by the law of large numbers, and the standard deviation of the evidence can be monitored using the central limit theorem (James, 1980). Many published studies have shown that this estimator works well for rather parsimonious models with relatively few fitting parameters. Indeed, for such models it is not that difficult to sample exhaustively the prior parameter distribution, and to evaluate the likelihood function for each of these points. Unfortunately, the computational requirements of this BFMC method become rather impractical for parameter-rich models as many millions or even billions of model evaluations are required to characterize adequately the likelihood surface.

2.2.2. Laplace-Metropolis method

The LM method (Lewis and Raftery, 1997) builds on the assumption that the posterior parameter distribution is characterized adequately with a (multi)normal distribution

$$p_{\mathrm{LM}}(\widetilde{\mathbf{Y}}) \approx (2\pi)^{d/2}|\mathbf{H}(\theta^*)|^{1/2}p(\theta^*)L(\theta^*|\widetilde{\mathbf{Y}}), \tag{6}$$

where θ^* denotes the mean of this distribution, and $|\mathbf{H}(\theta^*)|^{1/2}$ signifies the determinant of the Hessian matrix at θ^*. The two terms $(2\pi)^{d/2}$ and $p(\theta^*)L(\theta^*|\widetilde{\mathbf{Y}})$ scale the density of the normal distribution so as to consider explicitly the effect of parameter dimensionality, and quality of fit, on the evidence, respectively. This estimator is derived from an asymptotic approximation of the evidence and uses a quadratic Taylor series expansion around θ^*. This derivation appears in Lewis and Raftery (1997), and interested readers are referred to this publication for further details. The mean of the multinormal distribution, θ^*, is assumed equivalent to the MAP solution of the posterior parameter distribution, and the Hessian matrix, $\mathbf{H}(\theta^*)$, is computed from the J posterior samples, θ_j, as follows (Rousseeuw and Van Zomeren, 1990)

$$\mathbf{H}(\theta^*) = \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j - \theta^*)^T(\theta_j - \theta^*). \tag{7}$$

For a large enough sample, the Hessian matrix converges to the posterior covariance matrix.

The KIC (Kashyap, 1982)

$$\mathrm{KIC}_{\theta^*} = -2\log(p_{\mathrm{LM}}(\widetilde{\mathbf{Y}})) \tag{8}$$

is closely related to the LM approach, with θ^* assumed equivalent to the MAP solution.

2.2.3. Gaussian mixture importance sampling

As third and last method we consider the GMIS evidence estimator developed recently by Volpi et al. (2016). This method uses multidimensional numerical integration of the posterior parameter distribution via bridge sampling (a generalization of importance sampling) of a mixture distribution fitted to samples of the target derived from MCMC simulation with the DREAM algorithm (Vrugt, 2016). This approach has elements in common with the BFMC method, yet draws samples directly from the posterior distribution, rather than the prior distribution (as in BFMC) to approximate the evidence. One would therefore expect a much higher sampling efficiency of the GMIS method. The use of a Gaussian mixture distribution allows GMIS to approximate as closely and consistently as possible the actual posterior target distribution. Indeed, this distribution can be multimodal, truncated, and "quasi-skewed" - properties that can be emulated with a mixture model if a sufficient number of normal components is used. The Expectation-Maximization (EM) algorithm is used to construct the Gaussian mixture distribution (Dempster et al., 1977; Hoogerheide et al., 2012). Let us assume that MCMC simulation with DREAM has produced J realizations, $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_J\}$, of the d-variate posterior parameter distribution under hypothesis, η_1. We approximate these samples' probability density function, $p(\theta|\widetilde{\mathbf{Y}})$, with a mixture distribution

$$q(\theta, K) = \sum_{k=1}^{K}\alpha_k f_k(\theta; \mu_k, \mathbf{\Sigma}_k), \tag{9}$$

of $K > 0$ multivariate normal densities, $f_k(\cdot|\mu_k, \mathbf{\Sigma}_k)$ in $\mathbb{R}^d$, where α_k, μ_k and $\mathbf{\Sigma}_k$ signify the scalar weight, the d-dimensional mean vector, and the $d \times d$-covariance matrix of the kth Gaussian component. The weights, or mixing probabilities, must lie on the unit Simplex, Δ^K, that is, $\alpha_k \geq 0$ and $\sum_{k=1}^{K}\alpha_k = 1$, and the $\mathbf{\Sigma}_k$'s must be symmetric, $\mathbf{\Sigma}_k(\theta_i, \theta_j) = \mathbf{\Sigma}_k(\theta_j, \theta_i)$, and positive semi-definite.

The Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Hoogerheide et al., 2012) is used to determine the values of the d_{mix}-variables of the mixture distribution, $\mathbf{\Phi} = \{\alpha_1, \ldots, \alpha_K, \beta_1, \ldots, \beta_K\}$, where each $\beta_k = \{\mu_k, \mathbf{\Sigma}_k\}$ characterizes the mean and covariance matrix of a different normal density of the mixture, and $k = \{1, \ldots, K\}$. This algorithm maximizes the log-likelihood, $\log\{L(\mathbf{\Phi}|\mathbf{\Theta}, K)\}$, of the mixture density

$$\log\{L(\mathbf{\Phi}|\mathbf{\Theta}, K)\} = \sum_{j=1}^{J}\log\left\{\sum_{k=1}^{K}\alpha_k f_k(\theta_j; \mu_k, \mathbf{\Sigma}_k)\right\}, \tag{10}$$

by alternating between an expectation (E) step and a maximization (M) step, until convergence of the values of $\mathbf{\Phi}$ is achieved for a given number of components, K. The optimum complexity of the mixture distribution is determined via minimization of the Bayesian information criterion, or BIC

$$\mathrm{BIC}(K) = -2\log\{L(\mathbf{\Phi}|\mathbf{\Theta}, K)\} + d_{\mathrm{mix}}(K)\log(J). \tag{11}$$

This criterion strikes a balance between quality of fit (first-term) and the complexity of the mixture distribution (second term). In practice, we use different values for K and then select the "optimal" mixture distribution by minimizing the value of the BIC criterion, or

$$\widehat{K} = \arg\min_{K\in\mathbb{N}_+}\mathrm{BIC}(K), \tag{12}$$

where $\mathbb{N}_+$ is the collection of strictly positive integer values.

The optimal mixture distribution now serves as importance density, $q(\theta, \widehat{K})$, in GMIS to estimate the marginal likelihood, $p_{\mathrm{GMIS}}(\widetilde{\mathbf{Y}})$. To this end, we draw at random from the importance distribution, $Q(\theta, \widehat{K})$, a total of N different samples, $\{\theta_1^{\mathrm{imp}}, \ldots, \theta_N^{\mathrm{imp}}\}$. We then evaluate each of these N parameter vectors in our hypothesis (conceptual model), and calculate their unnormalized posterior densities, $p(\theta_r^{\mathrm{imp}})L(\theta_r^{\mathrm{imp}}|\widetilde{\mathbf{Y}})$, where $r = \{1, \ldots, N\}$. The evidence, $p_{\mathrm{GMIS}}(\widetilde{\mathbf{Y}})$, is now computed by GMIS as a weighted mean of the ratios of the samples' unnormalized posterior densities and corresponding importance densities (Perrakis et al., 2014)

$$p_{\mathrm{GMIS}}(\widetilde{\mathbf{Y}}) \approx \frac{1}{N}\sum_{r=1}^{N}\frac{p(\theta_r^{\mathrm{imp}})L(\theta_r^{\mathrm{imp}}|\widetilde{\mathbf{Y}})}{q(\theta_r^{\mathrm{imp}})}. \tag{13}$$

This concludes our description of the GMIS estimator. We refer interested readers to Volpi et al. (2016) for a more detailed treatment and explanation of the theory, concepts, and main building blocks of GMIS. This paper also documents a diverse set of case studies (up to $d = 100$) which evaluate and benchmark the performance of GMIS against other commonly used evidence estimation methods.

2.3. Evidence estimation in practice

The posterior distribution and the MAP solution that is used by the LM (Section 2.2.2) and GMIS (Section 2.2.3) methods are derived from MCMC simulation using the $\mathrm{DREAM}_{(ZS)}$ algorithm (Laloy and Vrugt, 2012; Vrugt, 2016; Vrugt et al., 2008). This multi-chain method creates proposals on the fly from an historical archive of past states using a mix of parallel direction and snooker updates. We refer the reader to Linde and Vrugt (2013); Lochbühler et al. (2014); 2015); Rosas-Carbajal et al. (2013); 2015) for various geophysical case-studies in which this algorithm is used. For the actual field application, we use a hierarchical Bayesian formulation, in which the data error, $\sigma_{\widetilde{\mathbf{Y}}}$ in Eq. (2) is jointly estimated with the model parameters (e.g., Rosas-Carbajal et al. (2013)). For numerical reasons we work with a log-likelihood formulation of Eq. (2). A total of four chains were deemed sufficient for 25 parameters, five chains were used for model dimensions between 26 and 64, and eight chains for models with more than 65 parameters. The number of generations varied between 200,000 and 500,000 depending on the dimensionality of the target distribution. The scaling factor, β_0 of the jump rate was tuned to achieve an adequate acceptance rate and the univariate $\widehat{R}$-diagnostic (Gelman and Rubin, 1992) was used to judge when convergence had been achieved of the DREAM$_{(ZS)}$ algorithm to a limiting distribution.

We report the evidence estimates of the BFMC method using three different sample sizes involving $N = 10^5$, $N = 10^6$ and $N = 10^7$ samples in Eq. (5). In GMIS, we use a total of $N = 10^5$ importance samples (Eq. (13)). We repeat each of these two numerical experiments ten times, and summarize the mean evidence and associated range in the results section. Lastly, in the case of the LM method, we report the evidence computed as the mean of the estimates on the different Markov chains (Van Haasteren, 2013) together with the range.

2.4. Conceptual subsurface models

To demonstrate the usefulness of model selection in a hydrogeophysical setting, we consider two common parameterizations for the porosity structure, (a) horizontal layering with fixed thickness of each layer, hereafter referred to as Lh, and (b) a multi-Gaussian model, coined MG. In addition to these, we also consider vertical layering of the porosity, using fixed layer thicknesses, abbreviated Lv. This parameterization is rather unusal and uncommon, but serves herein to illustrate that a poor conceptual model exhibits low odds. We also compare and juxtapose much finer discretizations of the two layered models and considered three different variants of the multi-Gaussian model involving horizontal anisotropy (MGha), vertical anisotropy (MGva) and isotropy (MGis). The multi-Gaussian model we use herein is adopted from Laloy et al. (2015), but under the assumption of a known geostatistical model. The method developed by Laloy et al. (2015) generates a zero-mean stationary multi-Gaussian field through the circulant embedding method (CEM) of the covariance matrix together with a dimensionality reduction which is useful when dealing with finely discretized fields. The dimensionality is reduced by generating two low-dimensional vectors of standard normal random numbers (i.e., in our case, each vector has 50 dimensionality reduction (**DR**) variables) which are subsequently resampled to the original dimension through a one-dimensional Fast Fourier Transform interpolation (Laloy et al. (2015)). This method decreases substantially model dimensionality, and, as a consequence, lowers significantly the computational cost of MCMC simulation to sample the target distribution.

2.4.1. Petrophysics and forward modelling

The case-studies considered herein focus on porosity estimation using first-arrival travel time data from crosshole GPR. We use the petrophysical relationship by Pride (1994) to link the geophysical properties (i.e., radar slowness, s) to the hydrologic properties of primary interest (i.e., porosity, ϕ) in a water saturated media

$$s = \sqrt{\phi^m c^{-2}[\varepsilon_{\mathrm{w}} + (\phi^{-m} - 1)\varepsilon_{\mathrm{s}}]}, \tag{14}$$

where $\varepsilon_{\mathrm{w}} = 81$ (-) denotes the relative permittivity of water, $c = 3 \cdot 10^8$ (m/s) is the speed of light in a vacuum, ε_{s} (-) signifies the relative permittivity of the mineral grains and m is a unitless cementation index. We use the non-linear 2D travel time solver (*time 2d*) of Podvin and Lecomte (1991) to compute first-arrival travel times from slowness fields obtained by applying the petrophysical relationship of Eq. (14) to each porosity field.

3. Illustrative toy example

To benchmark the different evidence estimators of Section 2.2, we first consider an illustrative example involving a simple crosshole GPR experiment. A total of 10 transmitter and receiver antennas are placed at multiple different depths (uniform intervals) in boreholes located in the left and right side of the domain, respectively (see Fig. 1a). This results in a total of 100 different transmitter-receiver antenna pairs. The spatial domain that necessitates porosity characterization covers an area of 7.2 m × 7.2 m. To warrant accurate model simulations, a spatial discretization of 0.04 × 0.04 m is considered. We contaminate the $n = 100$ first-arrival travel time data with Gaussian white noise using a measurement error, $\sigma_{\widetilde{\mathbf{Y}}} = 2$ ns. This comparatively high error level was chosen to facilitate comparison with the BFMC method, which is known to work better in the presence of large measurement errors. This leads to a likelihood function that is less peaked, and, consequently, a posterior distribution that is more dispersed as it will distribute more evenly the probability mass over the parameter space. The ”true” porosity field of the subsurface is made up of four different layers of equal thickness with porosity values of 0.3, 0.45, 0.35 and 0.4, in the downward direction, respectively (see Fig. 1a). We varied the number of horizontal layers of constant thickness from $d = 1$ to $d = 16$, and assume a uniform prior distribution for the porosity, ϕ, of each respective layer using upper and lower bound values of 0.25 and 0.50, respectively. The petrophysical parameters of Eq. (14) are assumed fixed using values of $m = 1.5$ and $\varepsilon_{\mathrm{s}} = 5$, respectively.

Fig. 1b–e presents the posterior mean porosity field derived from the DREAM$_{(ZS)}$ algorithm for four different model conceptualizations. The two layer model (Fig. 1b) is an overly simplistic representation of the true porosity field which is, by construction, perfectly described by the conceptual model with four layers shown in Fig. 1c. The posterior mean porosity field of the six layers model presented in Fig. 1d exhibits a relatively poor agreement with the reference porosity field. Finally, the porosity values for the eight layer model (Fig. 1e) correspond rather closely with their counterparts of the reference field (Fig. 1a). The bottom panel, in Fig. 1f–i, display the posterior standard deviation of the porosity estimates for the different layers of our four model conceptualizations. As expected, the uncertainty of the porosity estimates increases with the number of layers that are used in our subsurface characterizations.

Now we calculate the marginal likelihood of each hypothesis using the BFMC, LM, and GMIS estimators. The results of this analysis are presented in Fig. 2 using at the left hand-side a plot of the mean evidence computed by each method against model complexity, and at the right-hand-side a graph of the associated uncertainty of each estimator. We consider subsurface models with up to $d = 16$ horizontal porosity layers of equal thickness. To simplify graphical interpretation of the results, we plot $\log_{10}$ transformed values of the evidence, and refer to this entity as $\mathcal{P}(\widetilde{\mathbf{Y}})$. Colour coding is used to differentiate between the results of the three different methods. The results highlight several important findings. In the first place, the evidence estimates confirm that the model with four different porosity layers, that is $d = 4$, is most supported by the available data (Fig. 2a). This finding is not surprising as this model uses the exact same layering of the porosity field as used in

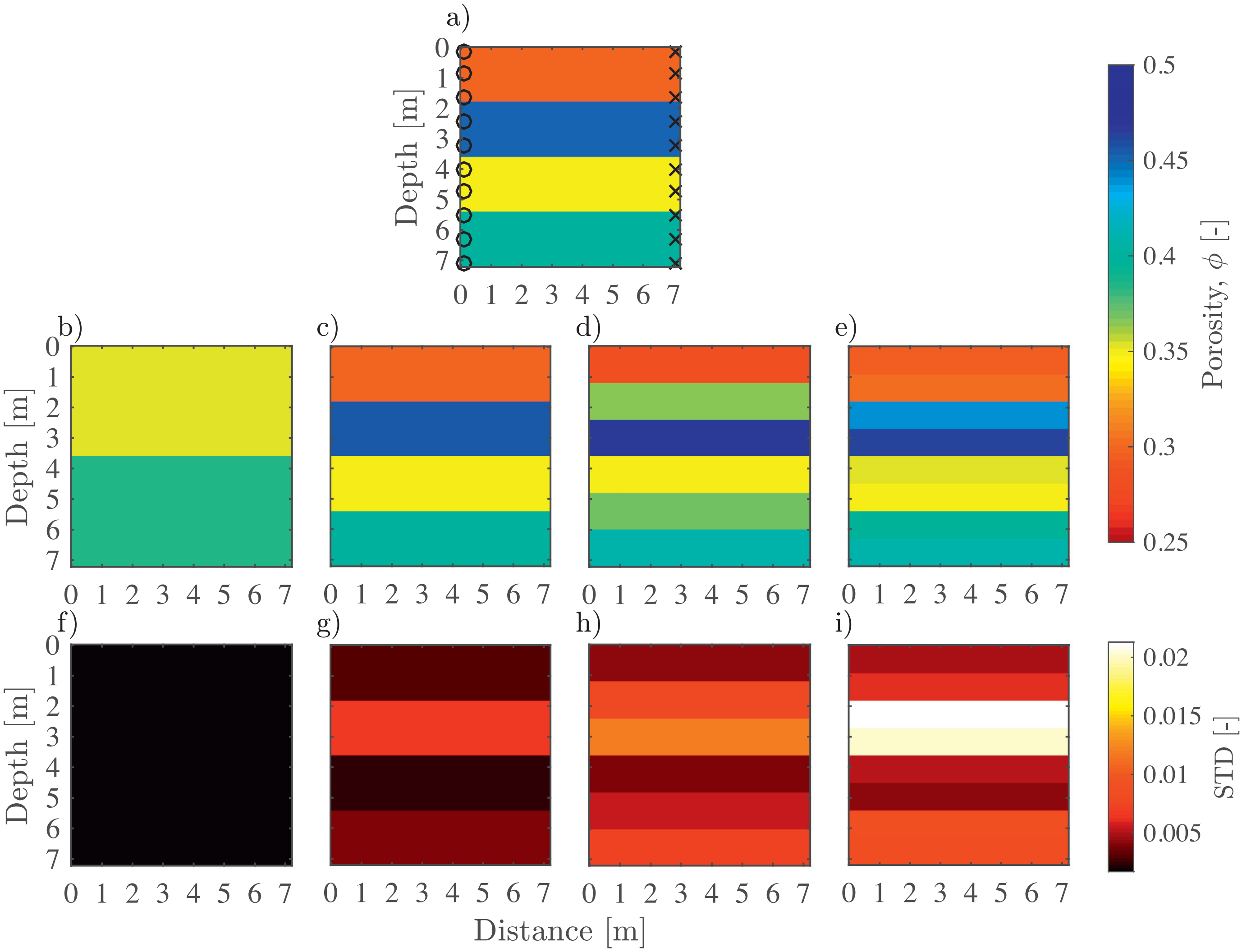

Fig. 1. a) The "true" subsurface porosity model used in our synthetic crosshole-GPR experiment. The different measurement depths of the transmitter antenna (black crosses) and receiver antenna (black circles) are separately indicated. Mean porosity fields of the posterior distribution derived from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using four different conceptualizations of the subsurface involving (b) two, (c) four, (d) six, and (e) eight horizontal layers. The corresponding posterior standard deviations of the porosity estimates for the four different conceptualizations of the subsurface are shown in (f), (g), (h) and (i), respectively.

the synthetic GPR experiment that was used to create the "measured" travel time data. Secondly, the BFMC (black), the LM (blue) and the GMIS (red) estimators are in excellent agreement and provide nearly identical values of the evidence for conceptual models with just a few parameters (horizontal layers)(Fig. 2a). Thirdly, the BFMC starts to deviate from the LM and GMIS methods at seven model dimensions and substantial differences appear for models with more than nine layers (Fig. 2a). This behavior is explained by the fact that the BFMC estimates did not converge for model dimensions higher than six. The convergence analysis was performed by a bootstrap analysis with 1000 bootstrap estimates (results not shown herein). In the fourth place, notice in Fig. 2b that the LM and GMIS estimators exhibit a negligible uncertainty compared to the range of evidence values considered and that the upper and lower bound values of the evidence derived from both methods appear rather similar. Evidence estimates derived from the BFMC method, on the contrary, exhibit a much larger uncertainty due to the fact that the BFMC does not reach convergence for model dimensions higher than six. This provides further support for the claim that, in our implementation and algorithmic settings, the BFMC method is inefficient when applied to models of high dimensionality since large numbers of samples (implying prohibitively large CPU-costs) are needed to properly characterize the likelihood surface and obtain reliable results.

We now investigate in more detail the discrepancies between the results of the three estimators, and plot in Fig. 3 the differences between the logarithmic values of the marginal likelihoods, $\mathcal{P}(\tilde{\mathbf{Y}})$, computed by the methods for the competing models used in this study. The solid black line depicts the difference in the mean evidence estimates derived by comparing each pairs of methods, and the grey shaded region quantifies the range associated with the differences in evidence estimates (i.e., the upper and lower boundaries of the grey shaded region are, respectively, the maximum and minimum difference in evidence estimate computed by each pairs of methods). Note, we use $N = 10^7$ in the BFMC method and report results for subsurface models with number of horizontal porosity layers (equal thickness) that ranges from $d = 1$ to $d = 16$.

The results in Fig. 3 provide further evidence for our earlier conclusions. Indeed, the three methods provide rather similar evidence values (Fig. 3a) for the simpler subsurface models (i.e., up to $d = 6$ different porosity layers). For larger model complexities the LM and the GMIS estimators differ a bit from each other - but this difference is very small in comparison to their mean estimates. It is now evident that the difference in the evidence estimates

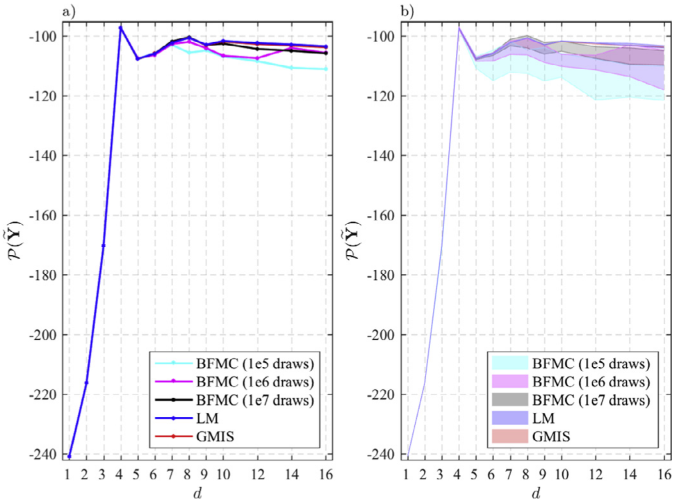

Fig. 2. Mean values of the evidence in $\log_{10}$ space, $\mathcal{P}(\widetilde{\mathbf{Y}})$ (a: left graph), and their associated uncertainty (b: right graph) derived from the BFMC, LM, and GMIS estimators for each model complexity, d used herein. Color coding is used to differentiate among the different methods. The evidence estimates of the LM and GMIS estimators are in excellent agreement and their uncertainty is negligibly small. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

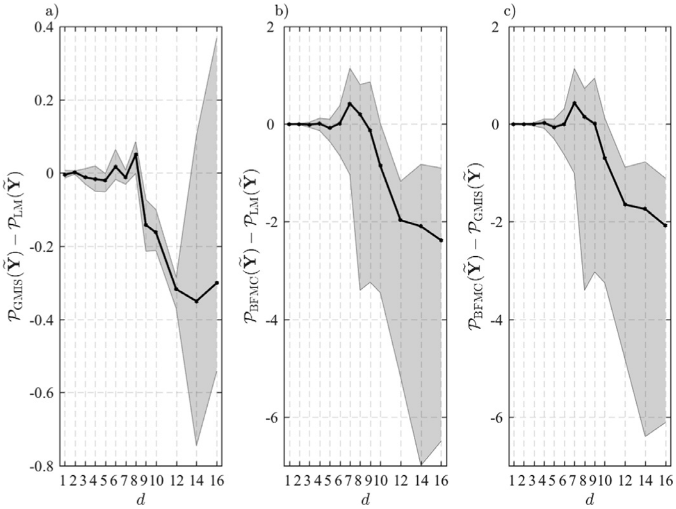

Fig. 3. Difference in the evidence estimates derived from different pairs of two methods as function of model complexity, (a) GMIS and LM, (b) BFMC and LM, and (c) BFMC and GMIS. The solid black line in each graph portrays the difference in the mean evidence estimates, and the grey shaded region quantifies the range associated with the difference in the mean evidence estimates of each method. Note, we use $\log_{10}$ transformed value of the evidence estimates.

Table 2
Parameters of the conceptual subsurface models with horizontal and vertical porosity layering. The last three columns summarize the range, prior distribution, and number, of each parameter, respectively as used in our MCMC inversion with the DREAM$_{(ZS)}$ algorithm. The variable n_{layer} defines the number of layers that is used in each conceptual model.

Parameter	Units	Prior range	Prior	***n*° parameters**
ϕ	–	0.25-0.5	Uniform	n_{layer}*
m	–	1.3-1.7	Uniform	1
ε_s	–	2–6	Uniform	1
$\sigma_{\widetilde{\mathbf{Y}}}$	ns	0.3–2	Log-uniform	1

* $1 \leq n_{\text{layer}} \leq 60$

derived from LM and GMIS increases with model complexity. Note that the maximum deviation between both methods is on the order of 0.7 unit in $\mathcal{P}(\widetilde{\mathbf{Y}})$ space, which, with mean estimates on the order of one-hundred (see Fig. 2a), equates to a difference smaller than 1%. However, it is important to stress here that there is no reason to expect that the two methods provide equivalent results since they are based on very different assumptions (details in Sections 2.2.2 and 2.2.3). Results from Fig. 3 also confirm that the evidence values derived from the BFMC method start to deviate from the other two methods for model dimensions higher than six since the method does not reach convergence for those models (Fig. 3b–c). These differences grow as large as 6–7% in $\mathcal{P}(\widetilde{\mathbf{Y}})$ space for the most complex subsurface models with $d = 14$ and $d = 16$ porosity layers. It is worth noting that we are primarily interested in an accurate model ranking, while the accuracy of the evidence estimates themselves are of secondary importance. In light of this, we find that the differences in the evidence estimates obtained by the three different estimators do not have an impact on which models are ranked first and second in the presented synthetic example.

This illustrative toy example shows that results from the three methods successfully agree on which model is most supported by the available data. The LM and GMIS methods provide similar values of the evidence, with associated uncertainty that appears rather small. The evidence estimates derived from the BFMC method, on the contrary, are trustworthy only for the most parsimonious subsurface conceptualizations (models) consisting only of a few porosity layers. Beyond this complexity, the 10 million BFMC samples used herein are insufficient to declare convergence and obtain reliable evidence estimates. Of course, we could further increase BFMC's sample size, yet this would increase further its already prohibitive computational time. Based on these findings, we discard the BFMC method and carry forward to the next case study the LM and GMIS estimators that are relatively CPU-efficient.

4. Field example

4.1. Field site and available data

We now focus our attention on the South Oyster Bacterial Transport Site in Virginia, USA, and use geophysical data measured at this experimental site to determine which model of the subsurface is preferred statistically. The geological characteristics of the South Oyster Bacterial Transport Site are described in Hubbard et al. (2001). GPR travel time data were measured at the S14-M13 borehole transect using a PulseEKKO 100 system with a 100-MHz nominal-frequency antenna. A domain of 7.2 × 7.2 m was measured with a total of 57 sources and 57 receivers, leading to a data set of 3248 observations of first-arrival travel times (one value is missing). We assume the measurement errors of the travel time to be uncorrelated and normally distributed with constant standard deviation, $\sigma_{\widetilde{\mathbf{Y}}}$. A relatively fine spatial discretization consisting of square cells with length 0.04 m was used in our forward simulations with the non-linear 2D travel time solver (*time 2d*) of Podvin and Lecomte (1991) to compute the first-arrival travel times for the 7.2 × 7.2 m domain of interest. The models used in this study differ in their conceptual representation of the subsurface, and use horizontal and vertical layering of the porosity. The numbers of porosity layers (equal thickness) is varied between 1 to 60, thereby providing a large array of competing hypotheses. Table 2 lists the parameters of both spatial porosity configurations which are subject to inference with the DREAM$_{(ZS)}$ algorithm. This includes, the porosity, ϕ, of each individual layer, and the values of m, ε_s and $\sigma_{\widetilde{\mathbf{Y}}}$. We list their symbol, unit, range, type of prior distribution, and respective number of unknowns.

The use of horizontal and vertical layering of the porosity is perhaps convenient computationally, but might not describe properly the subsurface of an actual field site. Indeed, the subsurface

Table 3
Integral scales in x- and z-direction, I_x and I_y, respectively, anisotropy angle, φ, and smoothness parameter, ν for the multi-Gaussian model with horizontal anisotropy (second column, MHha), vertical anisotropy (third column, MGva), and isotropy (last column, MGis).

	MGha	MGva	MGis
I_x	1.5 m	1.5 m	$\sqrt{1.5 \cdot 0.2}$ m
I_z	0.2 m	0.2 m	$\sqrt{1.5 \cdot 0.2}$ m
φ	90°	0°	90°
ν	0.5	0.5	0.5

Table 4
Parameters of multi-Gaussian models (first column) and their respective units (second column), range (third column), prior distribution (fourth column), and number (last column).

Parameter	Units	Prior range	Prior	***n*° parameters**
DR	–	–	Normal	100
$\bar{\phi}$	–	0.3 – 0.4	Uniform	1
v	–	$10^{-4} - 2.5 \cdot 10^{-3}$	Log-uniform	1
m	–	1.3 – 1.7	Uniform	1
ε_s	–	2 – 6	Uniform	1
$\sigma_{\tilde{\mathbf{Y}}}$	ns	0.3 – 2	Log-uniform	1

might exhibit much more complex porosity structure. We therefore augment the ensemble of hypotheses with a model that assumes a multi-Gaussian porosity field. This field is generated over a regular 2D grid of size 180 × 180 with geostatistical properties and spatial structure described with the Matérn variogram. Fortunately, the values of the integral scales in the x and z-direction, I_x and I_z, respectively, anisotropy angle, φ, and smoothness parameter, ν, of this variogram have been reported in the literature for the South Oyster Bacterial Transport Site (Chen et al., 2001; Hubbard et al., 2001). Their values are listed in the second column of Table 3, and assume horizontal anisotropy of the porosity field, that is $\varphi = 90°$. This model is referred to as MGha. For completeness, we also consider herein a multi-Gaussian model with vertical anisotropy, $\varphi = 0°$ (third column), coined MGva, and include an isotropic description of the porosity (fourth column), hereafter referred to as MGis, and enforced by setting I_x and I_z equal to the geometric mean of the integral scales of the first two multi-Gaussian models. We fix the value of $\nu = 0.5$ in the Matérn variogram, as we expect an exponential variogram model. Interested readers are referred to Laloy et al. (2015) for a more detailed description of the Matérn variogram.

We now focus our attention to the "unknown" parameters in each model (see Table 4), which are subject to inference using the observed travel time data. In our MCMC inversions we infer jointly the petrophysical parameters, ε_s and m of Eq. (14), mean porosity, $\bar{\phi}$, and its associated variance, v, the measurement data error, $\sigma_{\tilde{\mathbf{Y}}}$, of the travel time data, and 100 dimensionality reduction variables, **DR** (details in Section 2.4).

4.2. Results

We first display in Fig. 4 five realizations of the prior porosity field (columns) for each of the conceptual models (different rows) used in this case study. This includes the three multi-Gaussian models with (a) isotropy, (b) horizontal anisotropy, and (c) vertical anisotropy, and more simplistic models that assume (d) horizontal and (e) vertical layering of the porosity values. It is evident that these five model types provide very different descriptions of the porosity field of the subsurface at the experimental site. The multi-Gaussian models exhibit most spatial diversity with realizations that differ substantially in their mean porosity and associated variance. The porosity values of the layered models change abruptly from one depth to the next.

We now move on to our inversion results and present in Fig. 5 for each model of the ensemble (different rows), four different draws of the posterior distribution (first four columns), the posterior mean porosity field (fifth column) and the associated standard deviation (last column) derived from the DREAM$_{(ZS)}$ algorithm. The order of the presentation matches exactly Fig. 4, that is, the first three rows presents the results of the multi-Gaussian models with (a) isotropy, (b) horizontal anisotropy, and (c) vertical anisotropy of the porosity values, and the bottom two rows illustrate the results of the models with (d) horizontal and (e) vertical layering. The different conceptual models provide quite different characterizations of the porosity field. Some commonalities can be observed, though. For instance, the isotropic multi-Gaussian model, the multi-Gaussian model with horizontal anisotropy and the horizontally layered model (Fig. 5a-b-d) all depict the presence of a low-porosity zone just below the surface and at a depth of 4–5 m. They also demonstrate high-porosity zones at depths of 2 m and 6 m, and at 3 m below the ground surface a small high-porosity area is also visible, although this is not so evident for the isotropic multi-Gaussian model. The porosity fields parametrized by these three conceptual models are estimated with relatively low uncertainties (i.e., maximum of posterior standard deviations equals to or less than ± 0.01), especially, in the case of the horizontal layering. Also, the conceptual subsurface model with vertically oriented porosity structures (i.e., the vertically layered model and the multi-Gaussian model with vertical anisotropy) exhibit more variation in their porosity values (first four columns in Fig. 5c–e) and characterized by larger uncertainties (last column in Fig. 5c–e) than the other models.

Note that the posterior mean porosity field of the multi-Gaussian model with horizontal anisotropy (fifth column in Fig. 5b) is in good agreement with the velocity field obtained by Linde et al. (2008) and Linde and Vrugt (2013) for the exact same data set.

To provide more insights into the posterior parameter distributions of each model, Fig. 6 plots histograms of the marginal distributions of the cementation index, m (first column), the relative permittivity of the mineral grains, ε_s (second column), and the inferred data error, $\sigma_{\tilde{\mathbf{Y}}}$ (third column) for the multi-Gaussian (top three rows) and layered (bottom two rows) subsurface models. The prior distribution is separately indicated in each plot with the red line. Note, to simplify graphical notation, the density of all the distributions was scaled to be between 0 and 1. This figure highlights several interesting findings. In the first place, notice that the three parameters appear to be well defined in each of the five conceptual models with posterior distributions that occupy only a small portion of their respective prior distributions. This is particularly true for the marginal distribution of $\sigma_{\tilde{\mathbf{Y}}}$, the measurement error of the travel time data. Secondly, notice that the use of a vertically layered porosity (Fig. 6e) results in truncated histograms of the parameters m and ε_s and a large inferred data error, $\sigma_{\tilde{\mathbf{Y}}} > 1.5$ ns. These are possible signs of model malfunctioning, a claim that we will investigate next by looking in detail at the evidence estimates of each model, but supported thus far by the much larger posterior values of $\sigma_{\tilde{\mathbf{Y}}}$ for the vertically layered model than the other four competing subsurface models. Thirdly, notice that the histograms of the petrophysical parameters m and ε_s differ quite substantially between the conceptual models. These parameters probably compensate in different ways for imperfections in each model's porosity structure. The histograms of the nuisance parameter $\sigma_{\tilde{\mathbf{Y}}}$ appear almost similar with the exception of the model with vertically layered porosity values. Altogether, the lowest value of the measurement data error, $\sigma_{\tilde{\mathbf{Y}}} = 0.457$ ns, is found for the isotropic

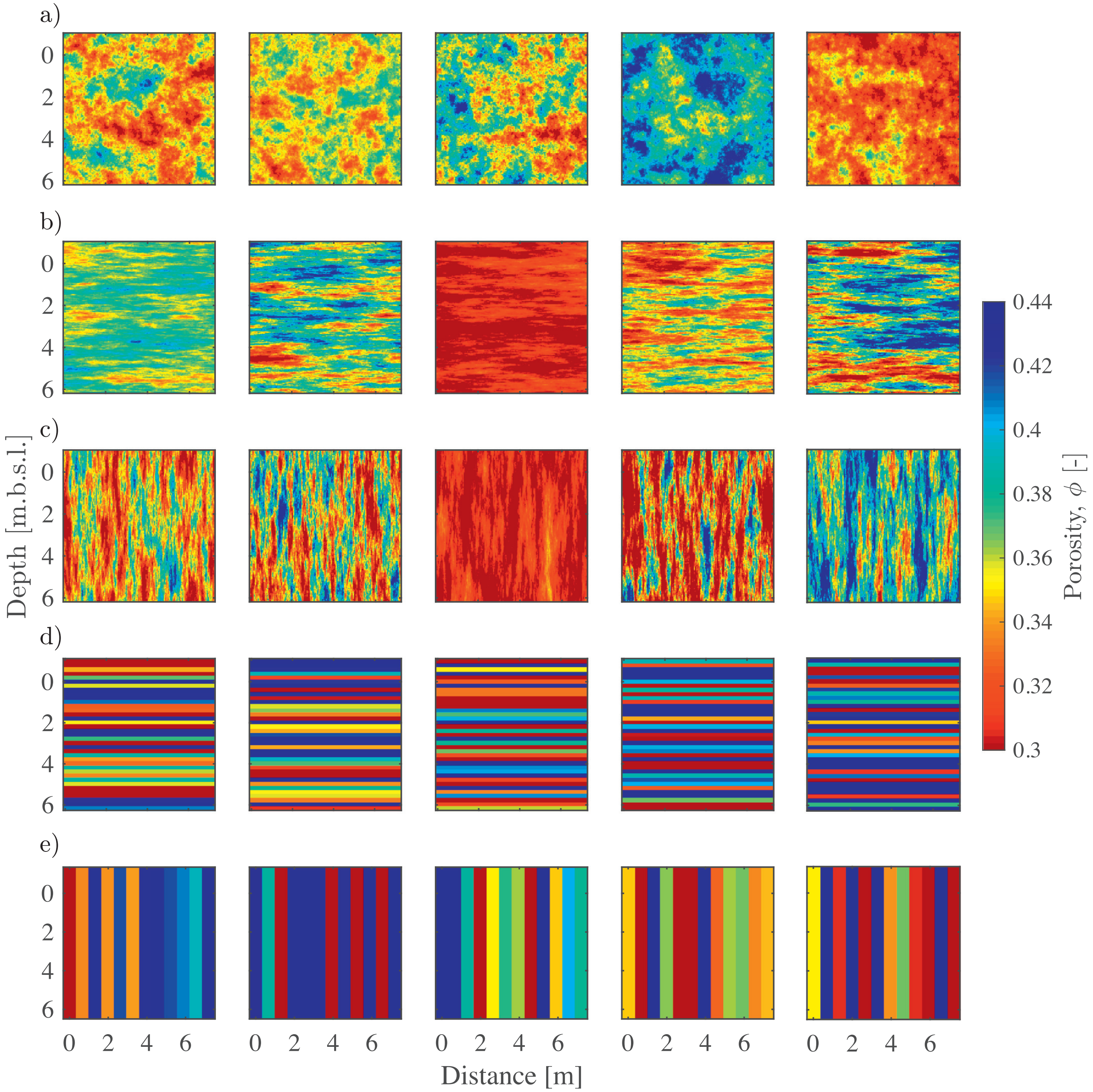

Fig. 4. Realizations drawn randomly from the prior distribution for the (a) isotropic multi-Gaussian model, (b) multi-Gaussian model with horizontal anisotropy, (c) multi-Gaussian model with vertical anisotropy, (d) horizontally layered model with 37 layers of equal thickness, and (e) vertically layered model with 12 layers of equal thickness.

multi-Gaussian model (Fig. 6a), which should suggest that this model most closely matches the observed travel time data.

We now turn our attention to the evidence of each model. Fig. 7 presents the results of this analysis using a $\log_{10}$ transformation of the evidence values. The left graph (Fig. 7a) displays the results for the three multi-Gaussian models with isotropy (circle), horiziontal anisotropy (square) and vertical anisotropy (triangles), respectively, using a single complexity involving $d = 105$ parameters. The graph in the middle (Fig. 7b) and on the right (Fig. 7c) depict the results for the conceptual models with horizontal and vertical layering, respectively, using between 1 to 60 different porosity layers. Colour coding is used in all the three plots to differentiate between the LM (blue) and GMIS (red) estimators. The vertical bars in Fig. 7a and shaded regions in Fig. 7b–c depict the uncertainty of the evidence estimates derived from the different trials with the LM and GMIS methods.

The most important conclusions are as follows. In the first place, the evidence estimates derived from both methods appear similar for model complexities with less than 30 (unknown) parameters. Beyond this, the difference between the marginal likelihoods derived from both methods grows up to 2% in $\log_{10}$ space for $d = 105$. Secondly, the evidence estimates derived from the different trials are quite similar, particularly for the GMIS method. Thirdly, the use of a larger number of layers in the two layered

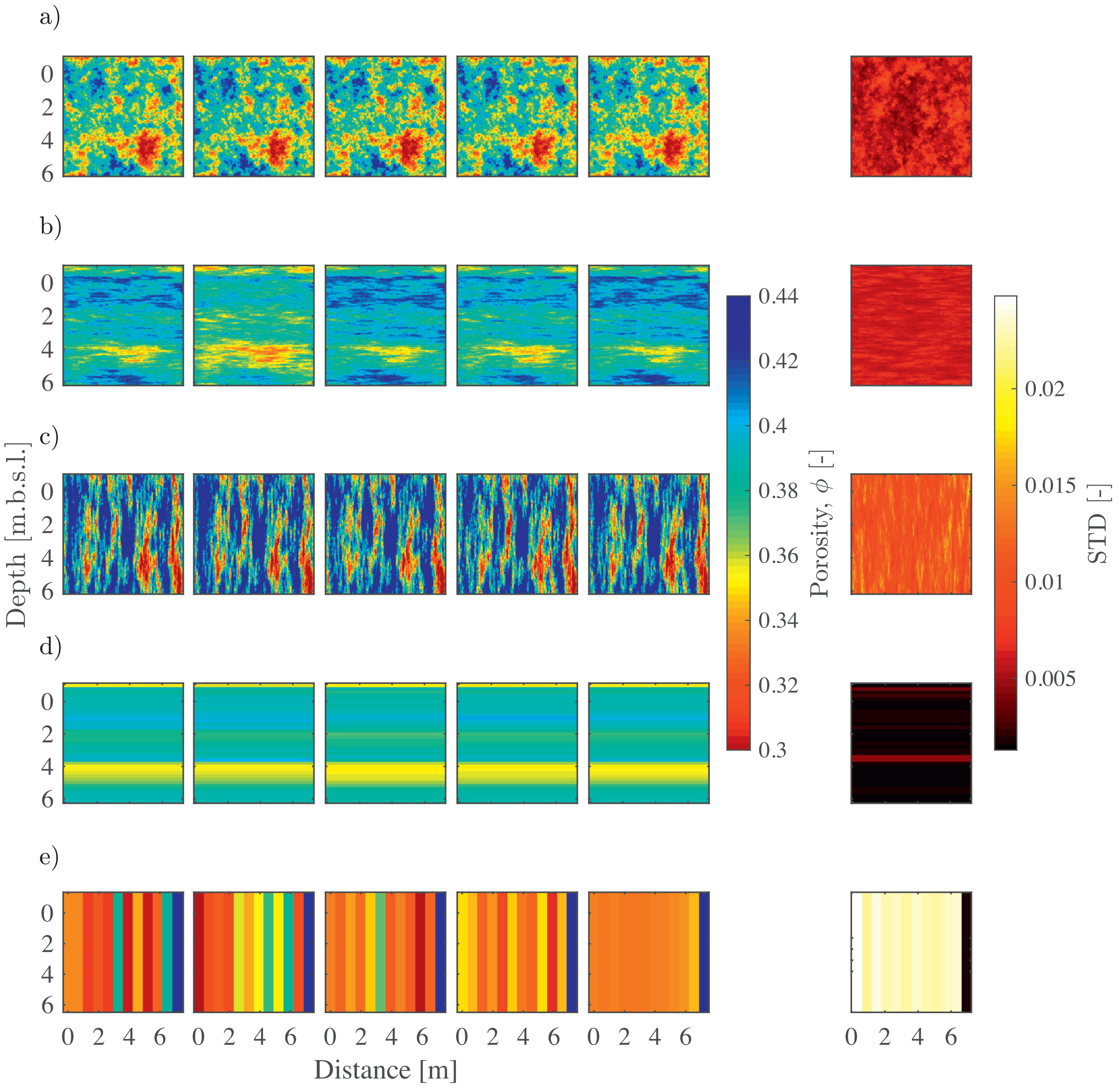

Fig. 5. Four realizations drawn randomly from the posterior distribution (first four columns), the posterior mean porosity field (fifth column) and the standard deviations of the posterior porosity estimates (last column) for the (a) isotropic multi-Gaussian model, (b) multi-Gaussian model with horizontal anisotropy, (c) multi-Gaussian model with vertical anisotropy, (d) horizontally layered model with 37 layers of equal thickness, and (e) vertically layered model with 12 layers of equal thickness.

models does not necessarily increase the statistical support for this model. Indeed, the value of the evidence is maximized when using 37 horizontal porosity layers or 15 vertical porosity layers. Beyond this number of porosity layers, the evidence values deteriorate slowly but with the exception of a sudden increase in $\mathcal{P}(\widetilde{\mathbf{Y}})$ at $d = 40$ for the vertically layered model. This spike is observed in the empirical $\mathcal{P}(\widetilde{\mathbf{Y}})$ functions of both evidence estimators (LM and GMIS), inspiring confidence in their results. Notice that the GMIS estimator produces a secondary peak at $d = 63$ (sixty layers), which causes the LM and GMIS methods to diverge in the rightmost part of their $\mathcal{P}(\widetilde{\mathbf{Y}})$ curves. Since it is not particularly clear which of the two estimators is at fault, we further test this case with GMIS by using 10^6 instead of 10^5 posterior realizations to construct the $d = 63$-variate importance distribution. The results (not shown herein) confirm the presence of the peak at $d = 63$ which suggests that the secondary peak is real. Fortunately, this does not affect at all model ranking as the evidence values of the vertically layered porosity model are many orders of magnitude smaller than their counterparts of the multi-Gaussian models. These results illustrate the importance of hypothesis testing and highlight the need for (statistical) methods that help us to determine, in an efficient and robust manner, an appropriate model complexity. In fact, the marginalization approach that is used to determine the model evidence can be viewed as a formalization

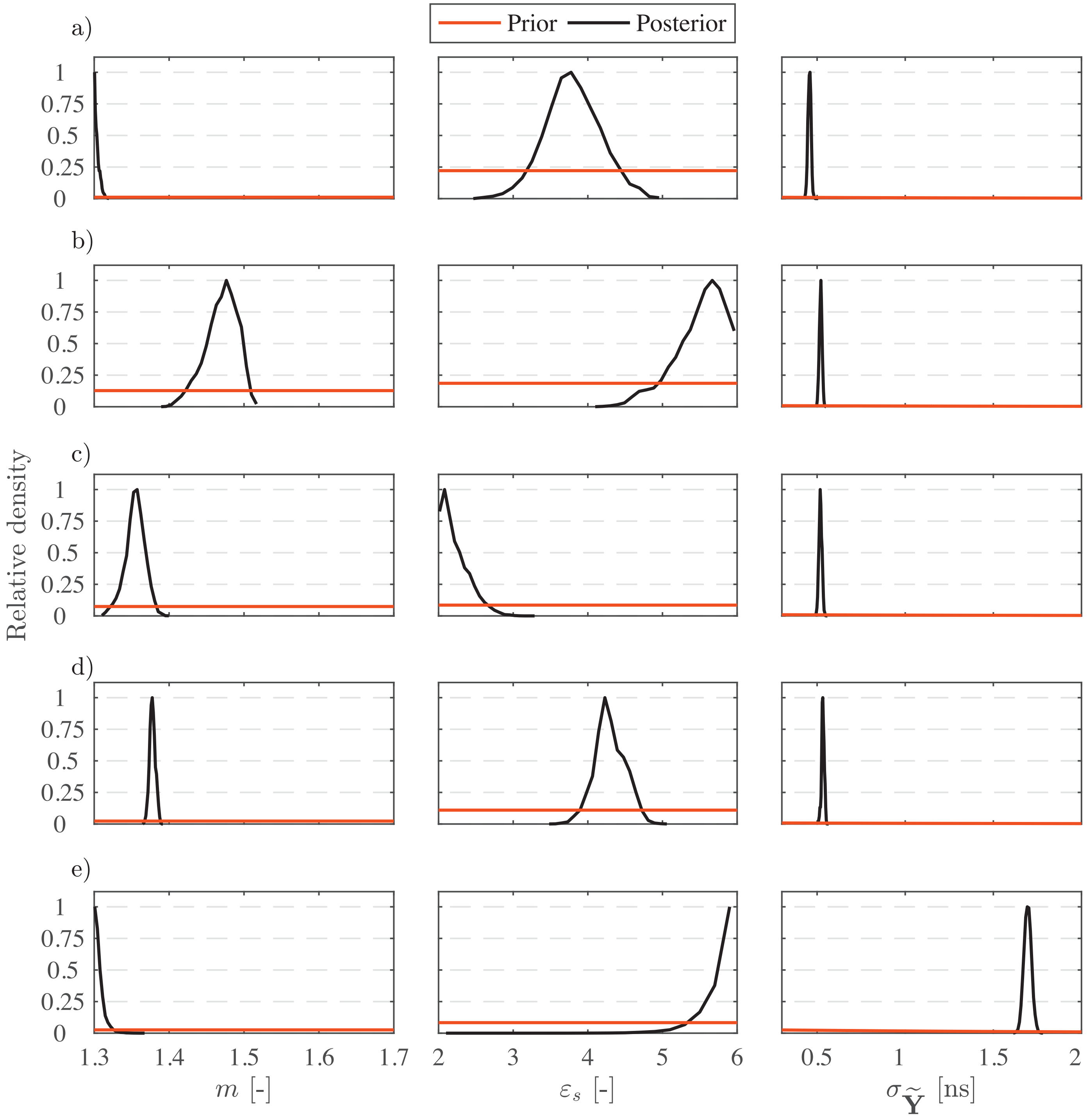

Fig. 6. Marginal posterior distributions of the inferred cementation index, m (first column), the relative permittivity of the mineral grains ε_s (second column), and the inferred data error, $\sigma_{\tilde{\mathbf{Y}}}$ (third column) for the multi-Gaussian models with (a) isotropy, (b) horizontal anisotropy, and (c) vertical anisotropy of the porosity values, and the two models with (d) horizontal and (e) vertical layering. The prior distribution is indicated separately in each plot using the red lines. The densities in each plot are normalized so that they all share the units of the y-axis on the left. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of Occam's razor and leads to a subsurface characterization that is not too simple nor too complex. Furthermore, and perhaps most important from the perspective of the present paper, the isotropic multi-Gaussian model receives the largest evidence values. This is true for both methods. Note, also that the vertically layered model exhibits very low evidence values. Indeed, the best vertically layered model has an evidence in $\log_{10}$ units of about −2757, much lower than the values of approximately −2757, and −1178 for the multi-Gaussian and horizontally layered models, respectively. This latter result confirms our earlier conclusion that the vertically layered model is deficient and inadequate.

Table 5 shows the five top-ranking conceptual models based on their evidence estimates derived from the LM (first column), and GMIS (second column) methods. The conceptual model that is most supported by the experimental data appears on top of the list (first row). For completeness, we also present in the third column

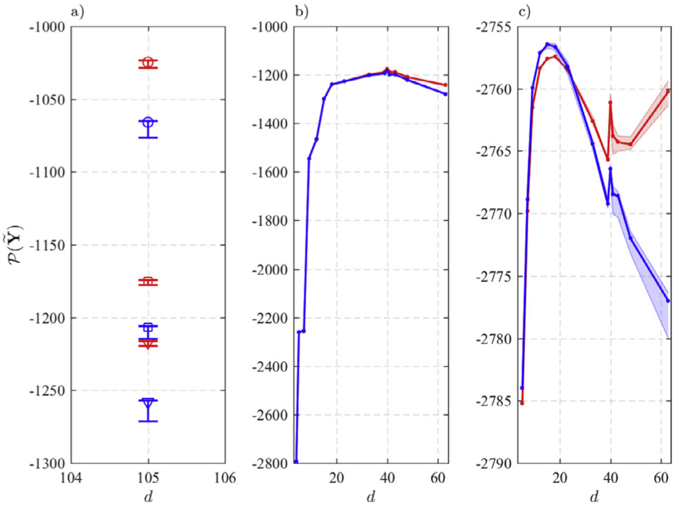

Fig. 7. Mean values of the evidence in $\log_{10}$ space, $\mathcal{P}(\widetilde{\mathbf{Y}})$, derived from the LM (blue) and GMIS (red) methods for (a) the multi-Gaussian models with isotropy (circles), horizontal anisotropy (squares), and vertical anisotropy (triangles), and the two models with (b) horizontal, and (c) vertical layering of the porosity. The error bars in (a) and the shaded areas in (b) and (c) summarize the ranges of the evidence estimates as derived from the different independent trials with both methods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

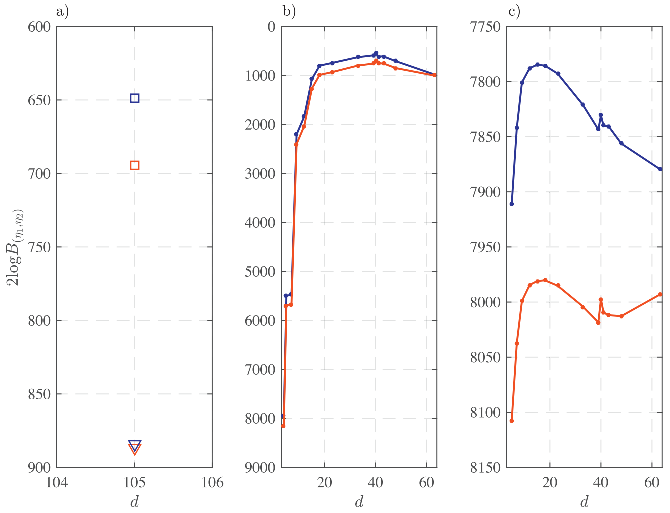

Fig. 8. Twice the natural logarithm of the Bayes factors of the best model (isotropic multi-Gaussian) of the ensemble with respect to the (a) multi-Gaussian model with horizontal anisotropy (squares) and vertical anisotropy (triangles), and the two conceptual models with (b) horizontal and (c) vertical layering of the porosity. Results are shown for the LM (blue) and the GMIS (red) methods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Ranking of the different conceptual models for the South Oyster Bacterial Transport Site based on evidence estimates derived from the LM (first column) and GMIS (second column) methods. The third column ranks the models based on their respective values of the measurement data error inferred from MCMC simulation using the DREAM$_{(ZS)}$ algorithm.

Ranking of conceptual models		
$\mathcal{P}_{\mathrm{LM}}(\widetilde{\mathbf{Y}})$	$\mathcal{P}_{\mathrm{GMIS}}(\widetilde{\mathbf{Y}})$	$\sigma_{\widetilde{\mathbf{Y}}}$ [ns]
MGis	MGis	MGis
L40	MGha	MGva
L39	L40	MGha
L43	L41	L43
L41	L43	L41

the ranking of the models using as metric the posterior values of the measurement data error, $\sigma_{\widetilde{\mathbf{Y}}}$. All three rankings demonstrate conclusively that the isotropic multi-Gaussian model is preferred. This model receives the highest evidence with both estimators and lowest value of the measurement data error, $\sigma_{\widetilde{\mathbf{Y}}} = 0.457$ ns. Note, that the LM and GMIS methods disagree in their assessment of the second best model. The more approximate LM method achieves the second highest support for the horizontally layered model with 37 layers ($d = 40$), whereas GMIS favours instead the multi-Gaussian model with horizontal anisotropy.

We now calculate the Bayes factor (”odds”) for the best model (isotropic multi-Gaussian) of the ensemble in relationship to each conceptual model. The ”odds” of the isotropic multi-Gaussian model are on the order of 10^{118} and 10^{151} relative to the second best model of the ensemble according to the LM and GMIS estimators (Table 5; Fig. 8). Fig. 8a depicts twice the natural logarithm of the Bayes factors with respect to the multi-Gaussian model with horizontal anisotropy (square symbol), and vertical anisotropy (triangle symbol), and Fig. 8b–c displays the same entity with respect to the horizontally and vertically layered models, respectively. Colour coding is used to differentiate between the LM (blue) and GMIS (red) evidence estimators. It is evident that the isotropic multi-Gaussian model receives most support by the data - the values listed on the y-axis in each plot are all larger than 600, which according to Table 1 suggests that there is very strong evidence against each of these alternative hypotheses.

The results presented thus clearly favour the use of an isotropic multi-Gaussian model for the porosity structure of the subsurface at the South Oyster Bacterial Transport Site. This conclusion is at odds with findings presented in the literature (Chen et al., 2001; Hubbard et al., 2001) using geostatistical analysis of the porosity structure. The results of these studies support the use of a multi-Gaussian model with horizontal anisotropy.

4.3. A synthetic experiment

To shed some more light on the selection of the isotropic multi-Gaussian model, we proceed with a synthetic experiment. We use the exact same domain (7.2 × 7.2 m) and setup as in our real-world study (Section 4.1), and simulate first-arrival travel times for a multi-offset GPR experiment with 57 transmitter and 57 receiver antennas using as reference porosity a multi-Gaussian field with horizontal anisotropy. This ”true” porosity field is constructed without the use of dimensional reduction using values of the integral scales and smoothness parameter listed in Table 3. The mean of this porosity field is, $\overline{\phi} = 0.39$ and the variance is, $\nu = 2 \cdot 10^{-4}$. The $57 \times 57 = 3249$ simulated travel times are corrupted with Gaussian white noise using $\sigma_{\widetilde{\mathbf{Y}}} = 0.5$ ns, and these distorted values are now used for numerical inversion using the DREAM$_{(ZS)}$ algorithm.

Table 6 presents the evidence estimates of the LM (first row) and GMIS (bottom row) methods using as competing hypotheses multi-Gaussian models with horizontal anisotropy (second column), isotropy (third column) and vertical anisotropy (right column). The numerical setup of these three conceptual models follows exactly Tables 3 and 4. The results of Table 6 demonstrate that both evidence estimators provide a similar ranking of the three subsurface models. As is to be expected, the most support is found for the multi-Gaussian model with horizontal anisotropy

Table 6
Synthetic experiment: Evidence estimates derived from the LM and GMIS methods for the multi-Gaussian models with isotropy (MGis), horizontal anisotropy (MGha) and vertical anisotropy (MGva).

	MGha	MGis	MGva
$\mathcal{P}_{\mathrm{LM}}(\widetilde{\mathbf{Y}})$	−1325.39	−1413.53	−1562.47
$\mathcal{P}_{\mathrm{GMIS}}(\widetilde{\mathbf{Y}})$	−1293.94	−1371.91	−1516.72

(second column). This is followed by the isotropic multi-Gaussian model (third column) and the multi-Gaussian model with vertical anisotropy (last column). This latter model, though, receives rather low evidence values. These results illustrate that both evidence estimators correctly identify the ”best” model of the en semble. We thus feel confident with the main conclusions of our real-world experiment, that the porosity field of the subsurface at the South Oyster Bacterial Transport Site is best described with an isotropic multi-Gaussian model. This conclusion is different from Chen et al. (2001) and Hubbard et al. (2001) whose results favoured the use of a multi-Gaussian model with horizontal anisotropy. These works considered the geophysical tomogram as data within a geostatistical analysis. Possible reasons for this discrepancy is that previous studies relied on forward modeling with straight ray paths and geophysical tomograms with inversions that did not consider an explicit underlying geostatistical model.

5. Discussion

The transdimensional (or reversible jump) MCMC algorithm (Green, 1995) is not suitable for comparing conceptual models that are based on completely different model parameterizations (e.g., layered vs. multi-Gaussian). In this study, we investigated to what extent evidence estimates with BFMC (Hammersley and Handscomb, 1964), LM (De Bruijn, 1970) and GMIS (Volpi et al., 2016) can be used to perform Bayesian model selection in the context of synthetic and real-world case studies. This is the first comparative study of evidence estimation in hydrogeophysics and we consider realistically high parameter dimensions (i.e., up to 105), large data sets (several thousands) and small data errors.

The BFMC method is known to provide the most reliable and unbiased evidence estimates in the limit of infinite sample sizes. Schöniger et al. (2015a); (2015b); (2014) found reliable evidence estimates with the BFMC method for different case-studies in hydrology. For our set-up with small errors and high data and model dimensions, we found that reliable evidence estimation with the BFMC method would need prohibitive computation times. If the assumption of a multi-Gaussian posterior density is fulfilled (a reasonable assumption in our test cases), the LM method should provide reliable evidence estimates (see also case-studies by Schöniger et al. (2014)). This is confirmed in our synthetic study in Section 3 by the strong agreement at low model dimensions between BFMC and LM estimates evaluated around the MAP estimate. The comparison of the LM and the more general (but more time-consuming) GMIS method shows that evidence estimates are similar for simpler subsurface conceptual models but that the difference between them increases with model complexity. Indeed, we do not expect to obtain equivalent results since the two methods are built on different assumptions (see details in Sections 2.2.2 and 2.2.3). For instance, the LM method is built on the assumption that a Gaussian model can properly describe the posterior distribution. This is different for GMIS (or BFMC for that matter) that is based on importance sampling within the prior parameter bounds. It is clear then that the more the posterior distributions are far from being Gaussian, the more the LM and GMIS methods will provide different estimates.

In our application to the South Oyster Bacterial Transport Site (Section 4), we found that the isotropic multi-Gaussian model has the highest evidence (Fig. 7a). The corresponding Bayes factors (Eq. (3)), computed with respect to each tested conceptual models, are all larger than 10^{100}. This result is in agreement with the findings by Schöniger et al. (2014): one decisive winning conceptual model is often obtained when using large data sets and small data errors. We also considered the field example described in Section 4.1, but using less data (i.e., $n = 224$ instead of $n = 3248$) and we found (results not shown) that: (1) the isotropic multi-Gaussian model is still the winner, (2) all the evidence estimates are much larger (e.g., in the case of the isotropic multi-Gaussian model, the evidence increases from about 10^{-1000} to 10^{-100}) and that (3) the Bayes factors are much smaller (e.g., when comparing the multi-Gaussian model with vertical anisotropy and the one with isotropy, the Bayes factor decreases approximately from 10^{190} to 10^{10}). Hence, even if we can still identify one clear winning conceptual model, the magnitudes of the Bayes factors have been drastically decreased.

Among the layered models, the GMIS and the LM method both suggest that the conceptual model with 37 layers has the highest evidence (Fig. 7b). Moreover, the model type with the least expected geological realism (i.e., vertically layered model) has, by far, the lowest evidences (Fig. 7c).

Based on previous geostatistical analysis at the South Oyster Bacterial Transport Site (Chen et al., 2001; Hubbard et al., 2001) one would expect that the multi-Gaussian model with horizontal anisotropy would be the one with the highest evidence. To better understand why the isotropic multi-Gaussian model has a higher evidence than the one with horizontal anisotropy, we performed a synthetic example (Section 4.3) in which the true porosity field is described by a multi-Gaussian model with horizontal anisotropy. We found that this conceptual model had the highest evidence, which suggests that the LM and GMIS methods allow us to identify the right conceptual model (Table 6). This suggests that this field-site might display less anisotropy than previously thought or that modeling (e.g., ray-based modeling instead of waveform modeling) and geometrical (e.g., uncertainties in borehole and antenna positions) errors bias the evidence estimates.

Below, we outline three avenues for future research:

- It is necessary to consider conceptual subsurface models with higher geological realism. Multi-Gaussian models are used extensively, but they are poor descriptions of many geological settings. There are many approaches to create more geologically realistic conceptual models (Linde et al., 2015), for example, multiple-point statistics (MPS) (Strebelle, 2002).
- It is essential to account for uncertainty in petrophysical relationships and model errors in order to not overstate the value of geophysical data. This could be accomplished by Approximate Bayesian Computation (ABC) (Beaumont et al., 2002; Marjoram et al., 2003; Pritchard et al., 1999; Tavaré et al., 1997) and lithological tomography (Bosch, 1999). ABC does not require a formal likelihood function and we suspect that this may help to decrease the sensitivity to model errors. Lithological tomography is a formal Bayesian procedure that integrates with the inference process a statistical description of the petrophysical relationships and geological concepts. This approach should spread out more evenly over the parameter space the posterior distribution, thereby decreasing the magnitude and range of the candidate models' Bayes factors, and enhancing the support and evidence for simpler conceptual models. We also highlight that incorporating model errors and petrophysical uncertainty is essential to enable model selection in integrated (joint) earth imaging (Moorkamp et al., 2016). It is also important to better elucidate and understand the relationship between a candidate

model's prior ranges and its evidence estimates. Much work on this topic can be found in the statistical literature (e.g. see Lindley's paradox), but comparatively little work has been done on high-dimensional priors as frequently encountered in subsurface characterization and geophysical inference.

- It would also be fruitful to investigate alternative approaches to evidence computation. In particular, nested sampling algorithms that are suitable to high-dimensional problems, such as the POLYCHORD algorithm (Handley et al., 2015) and the Galilean Monte Carlo algorithm (Skilling, 2012). Initial investigations with POLYCHORD suggest that evidence estimates are consistent with those obtained by LM and GMIS.

6. Conclusions

Hydrogeophysical methods are well suited to guide the critical choice of the most suitable conceptual subsurface hydrological model. Despite its importance, this topic has largely been ignored in the hydrogeophysical literature to date. We have performed a first comparative study of evidence estimation in hydrogeophysical settings. We consider realistically high model dimensions (i.e., about 100 unknowns), large data sets and small data errors that typify hydrogeophysical investigations. In the context of an illustrative synthetic example, we find that the brute force Monte Carlo method provides reliable estimates at low model dimensions but, when applied to higher model dimensions (i.e., in our case, higher than 6), the BFMC method is inefficient since a prohibitively large number of samples (and thus CPU-time) is required to obtain accurate results. This implied that the brute force Monte Carlo method was unsuitable to address our field example from the South Oyster Bacterial Transport Site (Virginia, USA). We find that the Laplace-Metropolis and the recent Gaussian mixture importance sampling estimator by Volpi et al. (2016) provide overall consistent relative evidence estimates and with rather small errors in both the synthetic cases where simple and low-dimensional (Section 3) and more complex and high-dimensional conceptual models (Section 4.3) were considered. Application of the Laplace-Metropolis and the Gaussian mixture importance sampling estimator to conceptual subsurface models of the South Oyster Bacterial Transport Site in Virginia, USA, revealed that the isotropic multi-Gaussian model was most supported by the available GPR travel time data. This model had the largest evidence and its Bayes factors were all larger than 10^{100} relative to all other plausible conceptualizations of the subsurface. Finally, the model with the least geological realism (i.e., vertically layered model) has extremely low evidence values for all of its discretizations (i.e., more than 10^{1500} times smaller than the evidences computed for the horizontally layered or multi-Gaussian models). Future research will focus on including the statistical nature of petrophysical relationships, model errors, and more realistic conceptual models of the subsurface.

Acknowledgements

This work was supported by the Swiss National Science Foundation [grant number 200021_155924]. We thank Eric Laloy for making available the code used for building the multi-Gaussian model. We are grateful to John Peterson and Susan Hubbard, who provided the crosshole GPR data from the South Oyster Bacterial Transport Site. We also thank the Editor Paolo D'Odorico, the Associated Editor Harrie-Jan Hendricks-Franssen, Anneli Guthke and Jan Dettmer for their very constructive and insightful comments.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), 2nd International Symposium on Information Theory, pp. 267–281.

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics 162 (4), 2025–2035.

Binley, A., Cassiani, G., Deiana, R., 2010. Hydrogeophysics: opportunities and challenges. B. Geofis. Teor. Appl. 51 (4), 267–284.

Binley, A., Hubbard, S.S., Huisman, J.A., Revil, A., Robinson, D.A., Singha, K., Slater, L.D., 2015. The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. Water Resour. Res. 51 (6), 3837–3866. http://dx.doi.org/10.1002/2015WR017016.

Bodin, T., Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm. Geophys. J. Int. 178 (3), 1411–1436. http://dx.doi.org/10.1111/j.1365-246X.2009.04226.x.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K., Rawlinson, N., 2012. Transdimensional inversion of receiver functions and surface wave dispersion. J. Geophys. Res.-Solid Earth 117 (B2), 1–24. http://dx.doi.org/10.1029/2011JB008560.

Bosch, M., 1999. Lithologic tomography: from plural geophysical data to lithology estimation. J. Geophys. Res.-Solid Earth 104 (B1), 749–766. http://dx.doi.org/10.1029/1998JB900014.

Box, G.E., Draper, N.R., 1987. Empirical Model-Building and Response Surfaces, 424. Wiley New York.

Chen, J., Hubbard, S., Rubin, Y., 2001. Estimating the hydraulic conductivity at the South Oyster Site from geophysical tomographic data using Bayesian techniques based on the normal linear regression model. Water Resour. Res. 37 (6), 1603–1613. http://dx.doi.org/10.1029/2000WR900392.

Chen, J., Hubbard, S., Rubin, Y., Murray, C., Roden, E., Majer, E., 2004. Geochemical characterization using geophysical data and Markov Chain Monte Carlo methods: a case study at the South Oyster bacterial transport site in Virginia. Water Resour. Res. 40 (12), 1–14. http://dx.doi.org/10.1029/2003WR002883.

Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis-Hastings output. J. Am. Stat. Assoc. 96 (453), 270–281. http://dx.doi.org/10.1198/016214501750332848.

De Bruijn, N.G., 1970. Asymptotic Methods in Analysis, 4. Dover Publications.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B Met. 39 (1), 1–38.

Dettmer, J., Dosso, S.E., Holland, C.W., 2009. Model selection and Bayesian inference for high-resolution seabed reflection inversion. J. Acoust. Soc. Am. 125 (2), 706–716. http://dx.doi.org/10.1121/1.3056553.

Dettmer, J., Dosso, S.E., Osler, J.C., 2010. Bayesian evidence computation for model selection in non-linear geoacoustic inference problems. J. Acoust. Soc. Am. 128 (6), 3406–3415. http://dx.doi.org/10.1121/1.3506345.

Gelfand, A.E., Dey, D.K., 1994. Bayesian model choice: asymptotics and exact calculations. J. R. Stat. Soc. B Met. 501–514.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7 (4), 457–472. http://dx.doi.org/10.1214/ss/1177011136.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82 (4), 711–732. http://dx.doi.org/10.1093/biomet/82.4.711.

Gull, S.F., 1988. Bayesian inductive inference and maximum entropy. In: Maximum-entropy and Bayesian methods in Science and Engineering, 31–32. Springer, pp. 53–74. http://dx.doi.org/10.1007/978-94-009-3049-0_4.

Hammersley, J.M., Handscomb, D.C., 1964. Monte Carlo Methods, 1. Springer Netherlands http://dx.doi.org/10.1007/978-94-009-5819-7.

Handley, W., Hobson, M., Lasenby, A., 2015. POLYCHORD: nested sampling for cosmology. Mon. Not. R. Astron. Soc. 450 (1), L61–L65. http://dx.doi.org/10.1093/mnrasl/slv047.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57 (1), 97–109. http://dx.doi.org/10.1093/biomet/57.1.97.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 382–401. http://dx.doi.org/10.1214/ss/1009212519.

Hoogerheide, L., Opschoor, A., Van Dijk, H.K., 2012. A class of adaptive importance sampling weighted EM algorithms for efficient and robust posterior and predictive simulation. J. Econometrics 171 (2), 101–120. http://dx.doi.org/10.2139/ssrn.2027967.

Hubbard, S., Linde, N., 2011. Hydrogeophysics. In: Wilderer, P. (Ed.), Treatise on Water Science. Elsevier, pp. 401–434. http://dx.doi.org/10.1016/B978-0-444-53199-5.00043-9.

Hubbard, S.S., Chen, J., Peterson, J., Majer, E.L., Williams, K.H., Swift, D.J., Mailloux, B., Rubin, Y., 2001. Hydrogeological characterization of the south oyster bacterial transport site using geophysical data. Water Resour. Res. 37 (10), 2431–2456. http://dx.doi.org/10.1029/2001WR000279.

Hubbard, S.S., Rubin, Y., 2005. Introduction to Hydrogeophysics. In: Hydrogeophysics. Springer, pp. 3–21. http://dx.doi.org/10.1007/1-4020-3102-5_1.

James, F., 1980. Monte Carlo theory and practice. Rep. Prog. Phys. 43 (9), 1145–1189. http://dx.doi.org/10.1088/0034-4885/43/9/002.

Jefferys, W.H., O. Berger, J., 1992. Ockham's razor and Bayesian analysis. Am. Sci. 80 (1), 64–72.

Jeffreys, H., 1935. Some tests of significance, treated by the theory of probability. Math. Proc. Cambridge 31 (2), 203–222. http://dx.doi.org/10.1017/S030500410001330X.

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

Kashyap, R.L., 1982. Optimal choice of AR and MA parts in autoregressive moving average models. IEEE. Trans. Pattern Anal. PAMI-4 (2), 99–104. http://dx.doi.org/10.1109/TPAMI.1982.4767213.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90 (430), 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572.

Laloy, E., Linde, N., Jacques, D., Vrugt, J.A., 2015. Probabilistic inference of multi-Gaussian fields from indirect hydrological data using circulant embedding and dimensionality reduction. Water Resour. Res. 51 (6), 4224–4243. http://dx.doi.org/10.1002/2014WR016395.

Laloy, E., Vrugt, J.A., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{ZS}$ and high-performance computing. Water Resour. Res. 48 (1), 1–18. http://dx.doi.org/10.1029/2011WR010608.

Lewis, S.M., Raftery, A.E., 1997. Estimating bayes factors via posterior simulation with the laplace-metropolis estimator. J. Am. Stat. Assoc. 92 (438), 648–655. http://dx.doi.org/10.1080/01621459.1997.10474016.

Li, X., Tsai, F.T.-C., 2009. Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod. Water Resour. Res. 45 (9), 1–14. http://dx.doi.org/10.1029/2008WR007488.

Linde, N., 2014. Falsification and corroboration of conceptual hydrological models using geophysical data. Wiley Interdiscip. Rev. Water 1 (2), 151–171. http://dx.doi.org/10.1002/wat2.1011.

Linde, N., Renard, P., Mukerji, T., Caers, J., 2015. Geological realism in hydrogeological and geophysical inverse modeling: a review. Adv. Water Resour. 86, 86–101. http://dx.doi.org/10.1016/j.advwatres.2015.09.019.

Linde, N., Tryggvason, A., Peterson, J.E., Hubbard, S.S., 2008. Joint inversion of crosshole radar and seismic traveltimes acquired at the South Oyster Bacterial Transport Site. Geophysics 73 (4), G29–G37. http://dx.doi.org/10.1190/1.2937467.

Linde, N., Vrugt, J.A., 2013. Distributed soil moisture from crosshole ground-penetrating radar travel times using stochastic inversion. Vadose Zone J. 12 (1), 1–21. http://dx.doi.org/10.2136/vzj2012.0101.

Lindley, D.V., 1957. A statistical paradox. Biometrika 44 (1–2), 187–192. http://dx.doi.org/10.1093/biomet/44.1-2.187.

Lochbühler, T., Breen, S.J., Detwiler, R.L., Vrugt, J.A., Linde, N., 2014. Probabilistic electrical resistivity tomography of a CO_2 sequestration analog. J. Appl. Geophys. 107, 80–92. http://dx.doi.org/10.1016/j.jappgeo.2014.05.013.

Lochbühler, T., Vrugt, J.A., Sadegh, M., Linde, N., 2015. Summary statistics from training images as prior information in probabilistic inversion. Geophys. J. Int. 201 (1), 157–171. http://dx.doi.org/10.1093/gji/ggv008.

MacKay, D.J., 1992. Bayesian interpolation. Neural Comput. 4 (3), 415–447. http://dx.doi.org/10.1162/neco.1992.4.3.415.

Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S., 2003. Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA 100 (26), 15324–15328. http://dx.doi.org/10.1073/pnas.0306899100.

Marshall, L., Nott, D., Sharma, A., 2005. Hydrological model selection: a Bayesian alternative. Water Resour. Res. 41 (10), 1–11. http://dx.doi.org/10.1029/2004WR003719.

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. http://dx.doi.org/10.1063/1.1699114.

Moorkamp, M., Lelièvre, P.G., Linde, N., Khan, A., 2016. Integrated Imaging of the Earth: Theory and Applications, 218. John Wiley & Sons http://dx.doi.org/10.1002/9781118929063.

Perrakis, K., Ntzoufras, I., Tsionas, E.G., 2014. On the use of marginal posteriors in marginal likelihood estimation via importance sampling. Comput. Stat. Data Anal. 77, 54–69. http://dx.doi.org/10.1016/j.csda.2014.03.004.

Podvin, P., Lecomte, I., 1991. Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. Geophys. J. Int. 105 (1), 271–284. http://dx.doi.org/10.1111/j.1365-246X.1991.tb03461.x.

Pride, S., 1994. Governing equations for the coupled electromagnetics and acoustics of porous media. Phys. Rev. B 50 (21), 15678–15696. http://dx.doi.org/10.1103/PhysRevB.50.15678.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W., 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. 16 (12), 1791–1798. http://dx.doi.org/10.1093/oxfordjournals.molbev.a026091.

Robert, C., Casella, G., 2013. Monte Carlo Statistical Methods. Springer Science & Business Media http://dx.doi.org/10.1007/978-1-4757-4145-2.

Rosas-Carbajal, M., Linde, N., Kalscheuer, T., Vrugt, J.A., 2013. Two-dimensional probabilistic inversion of plane-wave electromagnetic data: methodology, model constraints and joint inversion with electrical resistivity data. Geophys. J. Int. 196 (3), 1508–1524. http://dx.doi.org/10.1093/gji/ggt482.

Rosas-Carbajal, M., Linde, N., Peacock, J., Zyserman, F., Kalscheuer, T., Thiel, S., 2015. Probabilistic 3-D time-lapse inversion of magnetotelluric data: application to an enhanced geothermal system. Geophys. J. Int. 203 (3), 1946–1960. http://dx.doi.org/10.1093/gji/ggv406.

Rosenkrantz, R.D., 1977. Inference, Method and Decision: Towards a Bayesian Philosophy of Science, 115. Springer http://dx.doi.org/10.1007/978-94-010-1237-9.

Rousseeuw, P.J., Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc. 85 (411), 633–639. http://dx.doi.org/10.1080/01621459.1990.10474920.

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. http://dx.doi.org/10.1111/j.1365-246X.2006.03155.x.

Schöniger, A., Illman, W.A., Wöhling, T., Nowak, W., 2015. Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. J. Hydrol. 531, 96–110. http://dx.doi.org/10.1016/j.jhydrol.2015.07.047.

Schöniger, A., Wöhling, T., Nowak, W., 2015. A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. Water Resour. Res. 51 (9), 7524–7546. http://dx.doi.org/10.1002/2015WR016918.

Schöniger, A., Wöhling, T., Samaniego, L., Nowak, W., 2014. Model selection on solid ground: rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resour. Res. 50 (12), 9484–9513. http://dx.doi.org/10.1002/2014WR016062.

Schwarz, G., et al., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464. http://dx.doi.org/10.1214/aos/1176344136.

Skilling, J., 2012. Bayesian computation in big spaces: nested sampling and Galilean Monte Carlo. AIP Conf. Proc. 1443 (1), 145–156. http://dx.doi.org/10.1063/1.3703630.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. B. 64 (4), 583–639.

Steininger, G., Dosso, S.E., Holland, C.W., Dettmer, J., 2014. Estimating seabed scattering mechanisms via Bayesian model selection. J. Acoust. Soc. Am. 136 (4), 1552–1562. http://dx.doi.org/10.1121/1.4892752.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. 34 (1), 1–21. http://dx.doi.org/10.1023/A:1014009426274.

Tavaré, S., Balding, D.J., Griffiths, R.C., Donnelly, P., 1997. Inferring coalescence times from DNA sequence data. Genetics 145 (2), 505–518.

Tsai, F.T.-C., Li, X., 2008. Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window. Water Resour. Res. 44 (9), 1–15. http://dx.doi.org/10.1029/2007WR006576.

Van Haasteren, R., 2013. Marginal likelihood calculation with MCMC methods. In: Gravitational Wave Detection and Data Analysis for Pulsar Timing Arrays. Springer Science & Business Media, pp. 99–120. http://dx.doi.org/10.1007/978-3-642-39599-4.

Volpi, E., Schoups, G., Firmani, G., Vrugt, J. A., 2016. Bayesian model selection using MCMC simulation and bridge sampling. (submitted to Water Resour Res).

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. Environ. Modell. Softw. 75, 273–316. http://dx.doi.org/10.1016/j.envsoft.2015.08.013.

Vrugt, J.A., Ter Braak, C.J., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour. Res. 44 (12). http://dx.doi.org/10.1029/2007WR006720.

Ye, M., Pohlmann, K.F., Chapman, J.B., Pohll, G.M., Reeves, D.M., 2010. A model-averaging method for assessing groundwater conceptual model uncertainty. Ground Water 48 (5), 716–728. http://dx.doi.org/10.1111/j.1745-6584.2009.00633.x.