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, 1015 Lausanne, Switzerland, Carlotta.Brunetti@unil.ch, Niklas.Linde@unil.ch*

^b*Department of Civil and Environmental Engineering, University of California Irvine, Irvine, CA 92697-2175, USA, jasper@uci.edu*

^c*Department of Earth Systems Science, University of California Irvine, Irvine, CA 92697-2175, USA*

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) [1]. 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

*Corresponding author at: University of Lausanne, Géopolis - bureau 3123, 1015 Lausanne, Switzerland

Email address: Carlotta.Brunetti@unil.ch (Carlotta Brunetti)

conceptual models that assume horizontal or vertical layering of the porosity field.

Keywords: Bayesian model selection, evidence, Laplace-Metropolis, importance sampling, conceptual model, ground-penetrating radar

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 [2, 3, 4, 5]. 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 [6, 7]. 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 [6]. 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

26 be used to produce unconditional geostatistical realizations at larger scales.

27 Many different approaches have been suggested in the statistical literature
28 to help select the "best" model among a group of competing hypotheses. This
29 includes frequentist and Bayesian solutions. The application of such approaches
30 to geophysical studies has its own special challenges. For instance, a parameter-
31 rich, but geologically-unrealistic model may fit the data equally well or perhaps
32 even better than a more parsimonious model with more appropriate conceptu-
33 alization of the subsurface [8]. What is more, the decision about which model is
34 favoured, is also heavily influenced by the choice of the models' prior parameter
35 distribution, even for geophysical data comprised of many different measure-
36 ments. With the use of an inappropriate prior the model can be made to fit
37 the data arbitrarily poorly, changing fundamentally our opinion about which
38 model should be favoured, a phenomenon known as the Jeffreys-Lindley para-
39 dox [9, 10].

40 To describe accurately this trade-off between model complexity and good-
41 ness of fit, we here use Bayesian model selection, and investigate in detail the
42 denominator in Bayes theorem. This normalizing constant, referred to as the
43 evidence, marginal likelihood or integrated likelihood, conveys all information
44 necessary to determine which of the competing subsurface models (given their
45 prior parameter distributions) is most supported by the geophysical data. The
46 conceptual model with the largest evidence over the prior model space is the one
47 that is most supported by the experimental data. The foundation of Bayesian
48 model selection originates from Jeffreys [11, 9] and builds on the principles of
49 Occam's razor, that is, parsimony is favoured over complexity. In other words,
50 if two models exhibit a (nearly) equivalent fit to the data, the model with the
51 least number of "free" parameters is preferred statistically [12, 9, 13, 14]. Statis-
52 ticians prefer the use of so-called Bayes factors [15] to quantify the odds of each
53 model with respect to every other competing model. This Bayes factor of two
54 models A and B, is equivalent to the ratio of the evidences of both models. The
55 larger the value of this ratio, the stronger the support for hypothesis A. In cases
56 when the evidence values are of similar magnitude (e.g., within the same one

57 or two orders of magnitude), then it is recommended to use Bayesian model
58 averaging to combine predictions from different conceptual models and, thus,
59 obtain a more appropriate description of posterior parameter uncertainty [16].

60 Another distinct advantage of Bayesian model selection is that model com-
61 parison is relative to the existing conceptual models at hand, and consequently,
62 the "true" model does not have to be part of the ensemble considered for hy-
63 pothesis testing. To paraphrase Box and Draper [17]: *All our conceptual models*
64 *are wrong, but some are useful. It is the task of Bayesian model selection to de-*
65 *termine which of the considered conceptual models is the most useful.* Of course,
66 the answer to which model is most useful depends critically on the purpose and
67 intended goal of model application. Within the realm of model selection we can,
68 however, answer the question of which model is most supported by the avail-
69 able data. Yet, this task is not particularly easy for subsurface models, as the
70 integral of the posterior parameter distribution is, in general, high-dimensional
71 and without analytic solution. This probably explains why Bayesian model
72 selection is seldom used in hydrogeophysics and near-surface geophysics. In-
73 stead, we have to resort to numerical methods to approximate the value of the
74 evidence for each competing conceptual model. Gelfand and Dey [18] suggest
75 that the integral of the posterior distribution can be estimated via numerical
76 integration using, for instance, Monte Carlo methods [19], asymptotic solutions
77 (e.g., Bayesian information criterion, BIC) [20] or Laplace's method [21]. In
78 the field of geophysics, BIC [22], annealed importance sampling [23] and the
79 deviance information criterion, DIC, [24, 25] have been used for calculation of
80 the evidence.

81 In a separate line of research, transdimensional (or reversible jump) Markov
82 chain Monte Carlo (MCMC) methods [26] are receiving a surge of attention to
83 determine the optimal complexity (number of parameters) in geophysical model-
84 ing investigations (e.g., [27, 28, 29, 24]). In reversible jump MCMC, the number
85 of model parameters is treated as an unknown and parsimony is preferred as
86 this method incorporates directly the evidence in its calculations which makes
87 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 [30]. 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) [31], BIC, and Kashyap’s information criterion (KIC) [32] are used widely to select the most adequate model [33, 34, 35, 36]. A recent study by Schöniger et al. [37] 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 [38]. In cases where reliable brute force Monte Carlo integration is infeasible, Schöniger et al. [37] 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 [38].

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 [19], (2) the Laplace-Metropolis method [38] and (3) the Gaussian mixture importance sampling (GMIS) estimator of Volpi et al. [1]. 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

119 DREAM_(ZS) algorithm [39, 40, 41]. Then, we present an application of Bayesian
 120 model selection to subsurface modeling using geophysical data from the South
 121 Oyster Bacterial Transport Site in Virginia (USA) [42, 43, 44, 45, 46]. These
 122 data consist of travel time observations made by crosshole ground-penetrating
 123 radar (GPR), and exhibit small measurement errors typical of most near-surface
 124 geophysical sensing methods.

125 2. Theory and Methods

126 2.1. Bayesian inference with MCMC

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

$$130 \quad p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) = \frac{p(\boldsymbol{\theta})p(\tilde{\mathbf{Y}}|\boldsymbol{\theta})}{p(\tilde{\mathbf{Y}})}, \quad (1)$$

131 where, $p(\boldsymbol{\theta})$ signifies the prior pdf, $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \equiv p(\tilde{\mathbf{Y}}|\boldsymbol{\theta})$, denotes the likelihood
 132 function, and $p(\tilde{\mathbf{Y}})$ is equivalent to the marginal likelihood, or evidence. The
 133 larger the likelihood the better the model, $\mathcal{F}(\boldsymbol{\theta})$, explains the observed data,
 134 $\tilde{\mathbf{Y}}$. Bayesian model selection can be carried out for any type of likelihood
 135 function. However, in this work, we conveniently assume that the error residuals,
 136 $E(\boldsymbol{\theta}) = \{e_1(\boldsymbol{\theta}), \dots, e_n(\boldsymbol{\theta})\}$, are normally distributed with constant variance and
 137 negligible covariance. These three assumptions lead to the following definition
 138 of the likelihood function:

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

139 where $\sigma_{\tilde{\mathbf{Y}}}$ denotes the standard deviation of the measurement data error. This
 140 entity can be fixed a-priori by the user if deemed appropriate, or alternatively,
 141 the measurement data error can be treated as nuisance variable and the value of
 142 $\sigma_{\tilde{\mathbf{Y}}}$ is inferred jointly with the d -vector of model parameters, $\boldsymbol{\theta}$. The Gaussian
 143 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, $p(\boldsymbol{\theta})$, quantifies our knowledge about the expected distribution of the model parameters before considering the observed data. The evidence, $p(\tilde{\mathbf{Y}})$, acts as a normalization constant of the posterior distribution, and for fixed model parameterizations, is therefore often ignored in Bayesian inference. The posterior pdf, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, 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 [47, 48, 49, 39].

2.2. Evidence and Bayes factor

Bayesian hypothesis testing uses Bayes factors [15] 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

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

which is simply equivalent to the ratio of the evidences, $p(\tilde{\mathbf{Y}}|\eta_1)$ and $p(\tilde{\mathbf{Y}}|\eta_2)$, 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 $B_{(\eta_1, \eta_2)}$.

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

$$175 \quad p(\tilde{\mathbf{Y}}|\eta_l) = \int L(\boldsymbol{\theta}_l, \eta_l|\tilde{\mathbf{Y}})p(\boldsymbol{\theta}_l|\eta_l)d\boldsymbol{\theta}_l \quad l = 1, 2. \quad (4)$$

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

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

196 The Bayes factor is a sufficient statistic for hypothesis testing, yet renders
 197 necessary the definition of "formal" guidelines on how to interpret its value
 198 before we can proceed with model selection. Table 1 articulates an interpretation
 199 of the Bayes factor as advocated by Kass and Raftery [15]. This interpretation
 200 differentiates four (increasing) levels of support for proposition η_1 relative to
 201 η_2 . In general, the evidence in favor of η_1 increases with the value of its Bayes

202 factor. Thus, the larger the value of $B_{(\eta_1, \eta_2)}$, the more the data $\tilde{\mathbf{Y}}$ supports the
 203 hypothesis η_1 relative to η_2 , and the easier it becomes to reject this alternative
 204 hypothesis. It is suggested that the Bayes factor must be larger than 3 (or
 205 smaller than 1/3) to discriminate positively among two competing hypotheses.

Table 1: Interpretation of Kass and Raftery [15] 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

206 Unfortunately, the integral in Eq. (4) cannot be derived by analytic means
 207 nor by analytic approximation, and we therefore resort to numerical methods to
 208 calculate the evidence of each conceptual model. In the next section, we review
 209 briefly three different methods for estimating the evidence, including the brute
 210 force Monte Carlo method (BFMC), the Laplace-Metropolis (LM) method and
 211 the Gaussian mixture importance sampling (GMIS) approach recently developed
 212 by Volpi et al. [1].

213 2.2.1. Brute force Monte Carlo method

214 The BFMC method [19] approximates the evidence in Eq. (4) as an average
 215 of the likelihoods of N different samples drawn randomly from the (multivariate)
 216 prior distribution [15]

$$p_{\text{BFMC}}(\tilde{\mathbf{Y}}) \approx \frac{1}{N} \sum_{i=1}^N L(\boldsymbol{\theta}_i | \tilde{\mathbf{Y}}). \quad (5)$$

217 The validity of this estimator is ensured by the law of large numbers, and the
 218 standard deviation of the evidence can be monitored using the central limit
 219 theorem [50]. Many published studies have shown that this estimator works well
 220 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 [38] builds on the assumption that the posterior parameter distribution is characterized adequately with a (multi)normal distribution

$$p_{\text{LM}}(\tilde{\mathbf{Y}}) \approx (2\pi)^{d/2} |\mathbf{H}(\boldsymbol{\theta}^*)|^{1/2} p(\boldsymbol{\theta}^*) L(\boldsymbol{\theta}^* | \tilde{\mathbf{Y}}), \quad (6)$$

where $\boldsymbol{\theta}^*$ denotes the mean of this distribution, and $|\mathbf{H}(\boldsymbol{\theta}^*)|^{1/2}$ signifies the determinant of the Hessian matrix at $\boldsymbol{\theta}^*$. The two terms $(2\pi)^{d/2}$ and $p(\boldsymbol{\theta}^*) L(\boldsymbol{\theta}^* | \tilde{\mathbf{Y}})$ 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 $\boldsymbol{\theta}^*$. This derivation appears in [38], and interested readers are referred to this publication for further details. The mean of the multinormal distribution, $\boldsymbol{\theta}^*$, is assumed equivalent to the MAP solution of the posterior parameter distribution, and the Hessian matrix, $\mathbf{H}(\boldsymbol{\theta}^*)$, is computed from the J posterior samples, $\boldsymbol{\theta}_j$, as follows [51]

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

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

The KIC [32]

$$\text{KIC}_{\boldsymbol{\theta}^*} = -2 \log(p_{\text{LM}}(\tilde{\mathbf{Y}})) \quad (8)$$

is closely related to the LM approach, with $\boldsymbol{\theta}^*$ 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. [1]. 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 [39]. 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 [52, 53]. Let us assume that MCMC simulation with DREAM has produced J realizations, $\Theta = \{\theta_1, \dots, \theta_J\}$, of the d -variate posterior parameter distribution under hypothesis, η_1 . We approximate these samples' probability density function, $p(\theta|\tilde{\mathbf{Y}})$, with a mixture distribution

$$q(\theta, K) = \sum_{k=1}^K \alpha_k f_k(\theta; \mu_k, \Sigma_k), \quad (9)$$

of $K > 0$ multivariate normal densities, $f_k(\cdot | \mu_k, \Sigma_k)$ in $\mathbb{R}^d$, where α_k , μ_k and Σ_k signify the scalar weight, the d -dimensional mean vector, and the $d \times d$ -covariance matrix of the k th 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 Σ_k 's must be symmetric, $\Sigma_k(\theta_i, \theta_j) = \Sigma_k(\theta_j, \theta_i)$, and positive semi-definite.

The Expectation-Maximization (EM) algorithm [52, 53] is used to determine the values of the d_{mix} -variables of the mixture distribution, $\Phi = \{\alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_K\}$, where each $\beta_k = \{\mu_k, \Sigma_k\}$ characterizes the mean and covariance matrix of a different normal density of the mixture, and $k = \{1, \dots, K\}$. This algorithm

272 maximizes the log-likelihood, $\log\{L(\Phi|\Theta, K)\}$, of the mixture density

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

273 by alternating between an expectation (E) step and a maximization (M) step,
 274 until convergence of the values of Φ is achieved for a given number of compo-
 275 nents, K . The optimum complexity of the mixture distribution is determined
 276 via minimization of the Bayesian information criterion, or BIC

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

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

$$\hat{K} = \arg \min_{K \in \mathbb{N}_+} \text{BIC}(K), \quad (12)$$

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

282 The optimal mixture distribution now serves as importance density, $q(\theta, \hat{K})$,
 283 in GMIS to estimate the marginal likelihood, $p_{\text{GMIS}}(\tilde{\mathbf{Y}})$. To this end, we draw at
 284 random from the importance distribution, $Q(\theta, \hat{K})$, a total of N different sam-
 285 ples, $\{\theta_1^{\text{imp}}, \dots, \theta_N^{\text{imp}}\}$. We then evaluate each of these N parameter vectors in
 286 our hypothesis (conceptual model), and calculate their unnormalized posterior
 287 densities, $p(\theta_r^{\text{imp}})L(\theta_r^{\text{imp}}|\tilde{\mathbf{Y}})$, where $r = \{1, \dots, N\}$. The evidence, $p_{\text{GMIS}}(\tilde{\mathbf{Y}})$,
 288 is now computed by GMIS as a weighted mean of the ratios of the samples'
 289 unnormalized posterior densities and corresponding importance densities [54]

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

290 This concludes our description of the GMIS estimator. We refer interested
 291 readers to [1] for a more detailed treatment and explanation of the theory,
 292 concepts, and main building blocks of GMIS. This paper also documents a
 293 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 DREAM_(ZS) algorithm [41, 39, 40]. 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 [46, 55, 56, 57, 58] 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, $\sigma_{\mathbf{Y}}$ in Eq. (2) is jointly estimated with the model parameters (e.g., [57]). 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 200000 and 500000 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 $\hat{R}$ -diagnostic [59] 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 [60] 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 unusual 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 [61], but under the assumption of a known geostatistical model. The method developed by Laloy et al. [61] 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 [61]. 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 [62] 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

$$s = \sqrt{\phi^m c^{-2} [\varepsilon_w + (\phi^{-m} - 1) \varepsilon_s]}, \quad (14)$$

where $\varepsilon_w = 81$ (-) 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

353 2D travel time solver (*time 2d*) of Podvin and Lecomte [63] to compute first-
 354 arrival travel times from slowness fields obtained by applying the petrophysical
 355 relationship of Eq. (14) to each porosity field.

356 3. Illustrative toy example

357 To benchmark the different evidence estimators of Section 2.2, we first con-
 358 sider an illustrative example involving a simple crosshole GPR experiment. A
 359 total of 10 transmitter and receiver antennas are placed at multiple different
 360 depths (uniform intervals) in boreholes located in the left and right side of the
 361 domain, respectively (see Fig. 1a). This results in a total of 100 different
 362 transmitter-receiver antenna pairs. The spatial domain that necessitates poros-
 363 ity characterization covers an area of $7.2 \text{ m} \times 7.2 \text{ m}$. To warrant accurate model
 364 simulations, a spatial discretization of $0.04 \times 0.04 \text{ m}$ is considered. We contam-
 365 inate the $n = 100$ first-arrival travel time data with Gaussian white noise using
 366 a measurement error, $\sigma_{\tilde{Y}} = 2 \text{ ns}$. This comparatively high error level was cho-
 367 sen to facilitate comparison with the BFMC method, which is known to work
 368 better in the presence of large measurement errors. This leads to a likelihood
 369 function that is less peaked, and, consequently, a posterior distribution that is
 370 more dispersed as it will distribute more evenly the probability mass over the
 371 parameter space. The "true" porosity field of the subsurface is made up of four
 372 different layers of equal thickness with porosity values of 0.3, 0.45, 0.35 and 0.4,
 373 in the downward direction, respectively (see Fig. 1a). We varied the number
 374 of horizontal layers of constant thickness from $d = 1$ to $d = 16$, and assume
 375 a uniform prior distribution for the porosity, ϕ , of each respective layer using
 376 upper and lower bound values of 0.25 and 0.50, respectively. The petrophysical
 377 parameters of Eq. (14) are assumed fixed using values of $m = 1.5$ and $\varepsilon_s = 5$,
 378 respectively.

379 Figure 1b-e presents the posterior mean porosity field derived from the
 380 DREAM_(ZS) algorithm for four different model conceptualizations. The two
 381 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.

Figure 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.

391 Now we calculate the marginal likelihood of each hypothesis using the BFMC,
 392 LM, and GMIS estimators. The results of this analysis are presented in Fig.
 393 2 using at the left hand-side a plot of the mean evidence computed by each
 394 method against model complexity, and at the right-hand-side a graph of the as-
 395 sociated uncertainty of each estimator. We consider subsurface models with up
 396 to $d = 16$ horizontal porosity layers of equal thickness. To simplify graphical in-
 397 terpretation of the results, we plot $\log_{10}$ transformed values of the evidence, and
 398 refer to this entity as $\mathcal{P}(\tilde{\mathbf{Y}})$. Colour coding is used to differentiate between the
 399 results of the three different methods. The results highlight several important
 400 findings. In the first place, the evidence estimates confirm that the model with
 401 four different porosity layers, that is $d = 4$, is most supported by the available
 402 data (Fig. 2a). This finding is not surprising as this model uses the exact same
 403 layering of the porosity field as used in the synthetic GPR experiment that was
 404 used to create the "measured" travel time data. Secondly, the BFMC (black),
 405 the LM (blue) and the GMIS (red) estimators are in excellent agreement and
 406 provide nearly identical values of the evidence for conceptual models with just
 407 a few parameters (horizontal layers)(Fig. 2a). Thirdly, the BFMC starts to
 408 deviate from the LM and GMIS methods at seven model dimensions and sub-
 409 stantial differences appear for models with more than nine layers (Fig. 2a). This
 410 behavior is explained by the fact that the BFMC estimates did not converge for
 411 model dimensions higher than six. The convergence analysis was performed by
 412 a bootstrap analysis with 1000 bootstrap estimates (results not shown herein).
 413 In the fourth place, notice in Fig. 2b that the LM and GMIS estimators exhibit
 414 a negligible uncertainty compared to the range of evidence values considered
 415 and that the upper and lower bound values of the evidence derived from both
 416 methods appear rather similar. Evidence estimates derived from the BFMC
 417 method, on the contrary, exhibit a much larger uncertainty due to the fact that
 418 the BFMC does not reach convergence for model dimensions higher than six.
 419 This provides further support for the claim that, in our implementation and
 420 algorithmic settings, the BFMC method is inefficient when applied to models
 421 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.

Figure 2: Mean values of the evidence in log₁₀ space, $\mathcal{P}(\tilde{\mathbf{Y}})$ (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.

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

435 The results in Fig. 3 provide further evidence for our earlier conclusions.
 436 Indeed, the three methods provide rather similar evidence values (Fig. 3a) for
 437 the simpler subsurface models (i.e., up to $d = 6$ different porosity layers). For
 438 larger model complexities the LM and the GMIS estimators differ a bit from each
 439 other - but this difference is very small in comparison to their mean estimates.
 440 It is now evident that the difference in the evidence estimates derived from LM
 441 and GMIS increases with model complexity. Note that the maximum deviation
 442 between both methods is on the order of 0.7 unit in $\mathcal{P}(\tilde{\mathbf{Y}})$ space, which, with
 443 mean estimates on the order of one-hundred (see Fig. 2a), equates to a difference
 444 smaller than 1%. However, it is important to stress here that there is no reason
 445 to expect that the two methods provide equivalent results since they are based
 446 on very different assumptions (details in Sections 2.2.2 and 2.2.3). Results from
 447 Fig. 3 also confirm that the evidence values derived from the BFMC method
 448 start to deviate from the other two methods for model dimensions higher than six
 449 since the method does not reach convergence for those models (Fig. 3b-c). These
 450 differences grow as large as 6-7% in $\mathcal{P}(\tilde{\mathbf{Y}})$ space for the most complex subsurface
 451 models with $d = 14$ and $d = 16$ porosity layers. It is worth noting that we are
 452 primarily interested in an accurate model ranking, while the accuracy of the
 453 evidence estimates themselves are of secondary importance. In light of this, we
 454 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.

Figure 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.

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

469 **4. Field example**

470 *4.1. Field site and available data*

471 We now focus our attention on the South Oyster Bacterial Transport Site in
 472 Virginia, USA, and use geophysical data measured at this experimental site to
 473 determine which model of the subsurface is preferred statistically. The geological
 474 characteristics of the South Oyster Bacterial Transport Site are described in [44].
 475 GPR travel time data were measured at the S14-M13 borehole transect using a
 476 PulseEKKO 100 system with a 100-MHz nominal-frequency antenna. A domain
 477 of 7.2×7.2 m was measured with a total of 57 sources and 57 receivers, leading to
 478 a data set of 3248 observations of first-arrival travel times (one value is missing).
 479 We assume the measurement errors of the travel time to be uncorrelated and
 480 normally distributed with constant standard deviation, $\sigma_{\tilde{\mathbf{Y}}}$. A relatively fine
 481 spatial discretization consisting of square cells with length 0.04 m was used in
 482 our forward simulations with the non-linear 2D travel time solver (*time 2d*) of
 483 Podvin and Lecomte [63] to compute the first-arrival travel times for the 7.2×7.2
 484 m domain of interest. The models used in this study differ in their conceptual
 485 representation of the subsurface, and use horizontal and vertical layering of the

486 porosity. The numbers of porosity layers (equal thickness) is varied between
 487 1 to 60, thereby providing a large array of competing hypotheses. Table 2
 488 lists the parameters of both spatial porosity configurations which are subject
 489 to inference with the DREAM_(ZS) algorithm. This includes, the porosity, ϕ , of
 490 each individual layer, and the values of m , ε_s and $\sigma_{\tilde{\mathbf{Y}}}$. We list their symbol,
 491 unit, range, type of prior distribution, and respective number of unknowns.

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
$\sigma_{\tilde{\mathbf{Y}}}$	ns	0.3-2	Log-uniform	1

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

492 The use of horizontal and vertical layering of the porosity is perhaps con-
 493 venient computationally, but might not describe properly the subsurface of an
 494 actual field site. Indeed, the subsurface might exhibit much more complex poros-
 495 ity structure. We therefore augment the ensemble of hypotheses with a model
 496 that assumes a multi-Gaussian porosity field. This field is generated over a regu-
 497 lar 2D grid of size 180×180 with geostatistical properties and spatial structure
 498 described with the Matérn variogram. Fortunately, the values of the integral
 499 scales in the x and z -direction, I_x and I_z , respectively, anisotropy angle, φ , and
 500 smoothness parameter, ν , of this variogram have been reported in the literature
 501 for the South Oyster Bacterial Transport Site [42, 44]. Their values are listed in
 502 the second column of Table 3, and assume horizontal anisotropy of the porosity
 503 field, that is $\varphi = 90^\circ$. This model is referred to as MGha. For completeness,
 504 we also consider herein a multi-Gaussian model with vertical anisotropy, $\varphi = 0^\circ$
 505 (third column), coined MGva, and include an isotropic description of the poros-
 506 ity (fourth column), hereafter referred to as MGis, and enforced by setting I_x
 507 and I_z equal to the geometric mean of the integral scales of the first two multi-
 508 Gaussian models. We fix the value of $\nu = 0.5$ in the Matérn variogram, as we
 509 expect an exponential variogram model. Interested readers are referred to [61]
 510 for a more detailed description of the Matérn variogram.

Table 3: Integral scales in x - and z -direction, I_x and I_z , 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

511 We now focus our attention to the "unknown" parameters in each model
 512 (see Table 4), which are subject to inference using the observed travel time

513 data. In our MCMC inversions we infer jointly the petrophysical parameters,
514 ε_s and m of Eq. (14), mean porosity, $\bar{\phi}$, and its associated variance, v , the
515 measurement data error, $\sigma_{\tilde{\mathbf{Y}}}$, of the travel time data, and 100 dimensionality
516 reduction variables, **DR** (details in Section 2.4).

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

517 4.2. Results

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

Figure 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.

528 We now move on to our inversion results and present in Fig. 5 for each model
529 of the ensemble (different rows), four different draws of the posterior distribu-
530 tion (first four columns), the posterior mean porosity field (fifth column) and the
531 associated standard deviation (last column) derived from the DREAM_(ZS) al-
532 gorithm. The order of the presentation matches exactly Fig. 4, that is, the first
533 three rows presents the results of the multi-Gaussian models with (a) isotropy,

534 (b) horizontal anisotropy, and (c) vertical anisotropy of the porosity values, and
 535 the bottom two rows illustrate the results of the models with (d) horizontal and
 536 (e) vertical layering. The different conceptual models provide quite different
 537 characterizations of the porosity field. Some commonalities can be observed,
 538 though. For instance, the isotropic multi-Gaussian model, the multi-Gaussian
 539 model with horizontal anisotropy and the horizontally layered model (Fig. 5a-
 540 b-d) all depict the presence of a low-porosity zone just below the surface and
 541 at a depth of 4-5 m. They also demonstrate high-porosity zones at depths of
 542 2 m and 6 m, and at 3 m below the ground surface a small high-porosity area
 543 is also visible, although this is not so evident for the isotropic multi-Gaussian
 544 model. The porosity fields parametrized by these three conceptual models are
 545 estimated with relatively low uncertainties (i.e., maximum of posterior standard
 546 deviations equals to or less than ± 0.01), especially, in the case of the horizontal
 547 layering. Also, the conceptual subsurface model with vertically oriented poros-
 548 ity structures (i.e., the vertically layered model and the multi-Gaussian model
 549 with vertical anisotropy) exhibit more variation in their porosity values (first
 550 four columns in Fig. 5c-e) and characterized by larger uncertainties (last column
 551 in Fig. 5c-e) than the other models.

552 Note that the posterior mean porosity field of the multi-Gaussian model
 553 with horizontal anisotropy (fifth column in Fig.5b) is in good agreement with
 554 the velocity field obtained by Linde et al. [45] and Linde and Vrugt [46] for the
 555 exact same data set.

Figure 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.

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.

561 The prior distribution is separately indicated in each plot with the red line.
 562 Note, to simplify graphical notation, the density of all the distributions was
 563 scaled to be between 0 and 1. This figure highlights several interesting findings.
 564 In the first place, notice that the three parameters appear to be well defined in
 565 each of the five conceptual models with posterior distributions that occupy only
 566 a small portion of their respective prior distributions. This is particularly true
 567 for the marginal distribution of $\sigma_{\tilde{\mathbf{Y}}}$, the measurement error of the travel time
 568 data. Secondly, notice that the use of a vertically layered porosity (Fig. 6e)
 569 results in truncated histograms of the parameters m and ε_s and a large inferred
 570 data error, $\sigma_{\tilde{\mathbf{Y}}} > 1.5$ ns. These are possible signs of model malfunctioning, a
 571 claim that we will investigate next by looking in detail at the evidence estimates
 572 of each model, but supported thus far by the much larger posterior values of
 573 $\sigma_{\tilde{\mathbf{Y}}}$ for the vertically layered model than the other four competing subsurface
 574 models. Thirdly, notice that the histograms of the petrophysical parameters m
 575 and ε_s differ quite substantially between the conceptual models. These param-
 576 eters probably compensate in different ways for imperfections in each model's
 577 porosity structure. The histograms of the nuisance parameter $\sigma_{\tilde{\mathbf{Y}}}$ appear almost
 578 similar with the exception of the model with vertically layered porosity values.
 579 Altogether, the lowest value of the measurement data error, $\sigma_{\tilde{\mathbf{Y}}} = 0.457$ ns, is
 580 found for the isotropic multi-Gaussian model (Fig. 6a), which should suggest
 581 that this model most closely matches the observed travel time data.

Figure 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.

582 We now turn our attention to the evidence of each model. Fig. 7 presents
583 the results of this analysis using a $\log_{10}$ transformation of the evidence values.
584 The left graph (Fig. 7a) displays the results for the three multi-Gaussian mod-
585 els with isotropy (circle), horizontal 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 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 $\mathcal{P}(\tilde{\mathbf{Y}})$ at $d = 40$ for the vertically layered model. This spike is observed in the empirical $\mathcal{P}(\tilde{\mathbf{Y}})$ 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 $\mathcal{P}(\tilde{\mathbf{Y}})$ 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 high-

light 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 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 -1025 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.

Figure 7: Mean values of the evidence in $\log_{10}$ space, $\mathcal{P}(\tilde{\mathbf{Y}})$, 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.

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 the ranking of the models using as metric the posterior values of the measurement data error, $\sigma_{\tilde{\mathbf{Y}}}$. 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, $\sigma_{\tilde{\mathbf{Y}}} = 0.457$ ns. Note, that the LM and GMIS methods disagree in their assessment of the second best model. The more approximate LM method

639 achieves the second highest support for the horizontally layered model with 37
640 layers ($d = 40$), whereas GMIS favours instead the multi-Gaussian model with
641 horizontal anisotropy.

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		
$\mathcal{P}_{\text{LM}}(\tilde{\mathbf{Y}})$	$\mathcal{P}_{\text{GMIS}}(\tilde{\mathbf{Y}})$	$\sigma_{\tilde{\mathbf{Y}}} [\text{ns}]$
MGis	MGis	MGis
L40	MGha	MGva
L39	L40	MGha
L43	L41	L43
L41	L43	L41

642 We now calculate the Bayes factor ("odds") for the best model (isotropic
643 multi-Gaussian) of the ensemble in relationship to each conceptual model. The
644 "odds" of the isotropic multi-Gaussian model are on the order of 10^{118} and 10^{151}
645 relative to the second best model of the ensemble according to the LM and GMIS
646 estimators (Table 5; Fig. 8). Figure 8a depicts twice the natural logarithm of
647 the Bayes factors with respect to the multi-Gaussian model with horizontal
648 anisotropy (square symbol), and vertical anisotropy (triangle symbol), and Fig.
649 8b-c displays the same entity with respect to the horizontally and vertically
650 layered models, respectively. Colour coding is used to differentiate between the
651 LM (blue) and GMIS (red) evidence estimators. It is evident that the isotropic
652 multi-Gaussian model receives most support by the data - the values listed
653 on the y -axis in each plot are all larger than 600, which according to Table
654 1 suggests that there is very strong evidence against each of these alternative
655 hypotheses.

Figure 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.

656 The results presented thus clearly favour the use of an isotropic multi-
657 Gaussian model for the porosity structure of the subsurface at the South Oyster
658 Bacterial Transport Site. This conclusion is at odds with findings presented in
659 the literature [42, 44] using geostatistical analysis of the porosity structure. The
660 results of these studies support the use of a multi-Gaussian model with horizon-
661 tal anisotropy.

662 4.3. A synthetic experiment

663 To shed some more light on the selection of the isotropic multi-Gaussian
664 model, we proceed with a synthetic experiment. We use the exact same domain
665 (7.2×7.2 m) and setup as in our real-world study (Section 4.1), and simulate

666 first-arrival travel times for a multi-offset GPR experiment with 57 transmitter
 667 and 57 receiver antennas using as reference porosity a multi-Gaussian field with
 668 horizontal anisotropy. This "true" porosity field is constructed without the
 669 use of dimensional reduction using values of the integral scales and smoothness
 670 parameter listed in Table 3. The mean of this porosity field is, $\bar{\phi} = 0.39$ and
 671 the variance is, $v = 2 \cdot 10^{-4}$. The $57 \times 57 = 3249$ simulated travel times are
 672 corrupted with Gaussian white noise using $\sigma_{\tilde{\mathbf{y}}} = 0.5$ ns, and these distorted
 673 values are now used for numerical inversion using the DREAM_(ZS) algorithm.

674 Table 6 presents the evidence estimates of the LM (first row) and GMIS
 675 (bottom row) methods using as competing hypotheses multi-Gaussian models
 676 with horizontal anisotropy (second column), isotropy (third column) and vertical
 677 anisotropy (right column). The numerical setup of these three conceptual mod-
 678 els follows exactly Tables 3 and 4. The results of Table 6 demonstrate that both
 679 evidence estimators provide a similar ranking of the three subsurface models.
 680 As is to be expected, the most support is found for the multi-Gaussian model
 681 with horizontal anisotropy (second column). This is followed by the isotropic
 682 multi-Gaussian model (third column) and the multi-Gaussian model with ver-
 683 tical anisotropy (last column). This latter model, though, receives rather low
 684 evidence values. These results illustrate that both evidence estimators correctly
 685 identify the "best" model of the ensemble. We thus feel confident with the main
 686 conclusions of our real-world experiment, that the porosity field of the subsurface
 687 at the South Oyster Bacterial Transport Site is best described with an isotropic
 688 multi-Gaussian model. This conclusion is different from Chen et al. [42] and
 689 Hubbard et al. [44] whose results favoured the use of a multi-Gaussian model
 690 with horizontal anisotropy. These works considered the geophysical tomogram
 691 as data within a geostatistical analysis. Possible reasons for this discrepancy is
 692 that previous studies relied on forward modeling with straight ray paths and
 693 geophysical tomograms with inversions that did not consider an explicit under-
 694 lying geostatistical model.

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
$\mathcal{P}_{\text{LM}}(\tilde{\mathbf{Y}})$	-1325.39	-1413.53	-1562.47
$\mathcal{P}_{\text{GMIS}}(\tilde{\mathbf{Y}})$	-1293.94	-1371.91	-1516.72

5. Discussion

The transdimensional (or reversible jump) MCMC algorithm [26] 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 [19], LM [21] and GMIS [1] 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. [37, 64, 65] 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. [37]). 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

717 evidence estimates are similar for simpler subsurface conceptual models but
 718 that the difference between them increases with model complexity. Indeed, we
 719 do not expect to obtain equivalent results since the two methods are built on
 720 different assumptions (see details in Sections 2.2.2 and 2.2.3). For instance, the
 721 LM method is built on the assumption that a Gaussian model can properly
 722 describe the posterior distribution. This is different for GMIS (or BFMC for
 723 that matter) that is based on importance sampling within the prior parameter
 724 bounds. It is clear then that the more the posterior distributions are far from
 725 being Gaussian, the more the LM and GMIS methods will provide different
 726 estimates.

727 In our application to the South Oyster Bacterial Transport Site (Section 4),
 728 we found that the isotropic multi-Gaussian model has the highest evidence (Fig.
 729 7a). The corresponding Bayes factors (Eq. (3)), computed with respect to each
 730 tested conceptual models, are all larger than 10^{100} . This result is in agreement
 731 with the findings by Schöniger et al. [37]: one decisive winning conceptual
 732 model is often obtained when using large data sets and small data errors. We
 733 also considered the field example described in Section 4.1, but using less data
 734 (i.e., $n = 224$ instead of $n = 3248$) and we found (results not shown) that: (1) the
 735 isotropic multi-Gaussian model is still the winner, (2) all the evidence estimates
 736 are much larger (e.g., in the case of the isotropic multi-Gaussian model, the
 737 evidence increases from about 10^{-1000} to 10^{-100}) and that (3) the Bayes factors
 738 are much smaller (e.g., when comparing the multi-Gaussian model with vertical
 739 anisotropy and the one with isotropy, the Bayes factor decreases approximately
 740 from 10^{190} to 10^{10}). Hence, even if we can still identify one clear winning
 741 conceptual model, the magnitudes of the Bayes factors have been drastically
 742 decreased.

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

747 Based on previous geostatistical analysis at the South Oyster Bacterial Trans-

port Site [42, 44] 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 [7], for example, multiple-point statistics (MPS) [66].
- 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) [67, 68, 69, 70] and lithological tomography [71]. 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 [72]. 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 [73] and the Galilean Monte Carlo algorithm [74]. 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. [1] provide overall consistent relative evidence estimates and with rather small errors in both the

807 synthetic cases where simple and low-dimensional (Section 3) and more complex
 808 and high-dimensional conceptual models (Section 4.3) were considered. Applica-
 809 tion of the Laplace-Metropolis and the Gaussian mixture importance sampling
 810 estimator to conceptual subsurface models of the South Oyster Bacterial Trans-
 811 port Site in Virginia, USA, revealed that the isotropic multi-Gaussian model
 812 was most supported by the available GPR travel time data. This model had
 813 the largest evidence and its Bayes factors were all larger than 10^{100} relative to
 814 all other plausible conceptualizations of the subsurface. Finally, the model with
 815 the least geological realism (i.e., vertically layered model) has extremely low
 816 evidence values for all of its discretizations (i.e., more than 10^{1500} times smaller
 817 than the evidences computed for the horizontally layered or multi-Gaussian
 818 models). Future research will focus on including the statistical nature of petro-
 819 physical relationships, model errors, and more realistic conceptual models of the
 820 subsurface.

821 **Acknowledgements**

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

829 References

- 830 [1] E. Volpi, G. Schoups, G. Firmani, J. A. Vrugt, Bayesian model selection
831 using MCMC simulation and bridge sampling, (submitted to Water Resour
832 Res) (2016).
- 833 [2] A. Binley, G. Cassiani, R. Deiana, Hydrogeophysics: opportunities and
834 challenges, *B Geofis Teor Appl* 51 (4) (2010) 267–284.
- 835 [3] A. Binley, S. S. Hubbard, J. A. Huisman, A. Revil, D. A. Robinson,
836 K. Singha, L. D. Slater, The emergence of hydrogeophysics for improved
837 understanding of subsurface processes over multiple scales, *Water Resour*
838 *Res* 51 (6) (2015) 3837–3866. doi:10.1002/2015WR017016.
- 839 [4] S. Hubbard, N. Linde, Hydrogeophysics, in: P. Wilderer (Ed.), *Trea-*
840 *tise on Water Science*, Elsevier, 2011, pp. 401–434. doi:10.1016/
841 B978-0-444-53199-5.00043-9.
- 842 [5] S. S. Hubbard, Y. Rubin, Introduction to Hydrogeophysics, in: *Hydrogeo-*
843 *physics*, Springer, 2005, pp. 3–21. doi:10.1007/1-4020-3102-5_1.
- 844 [6] N. Linde, Falsification and corroboration of conceptual hydrological models
845 using geophysical data, *Wiley Interdiscip Rev Water* 1 (2) (2014) 151–171.
846 doi:10.1002/wat2.1011.
- 847 [7] N. Linde, P. Renard, T. Mukerji, J. Caers, Geological realism in hydrogeo-
848 logical and geophysical inverse modeling: A review, *Adv Water Resour* 86
849 (2015) 86–101. doi:10.1016/j.advwatres.2015.09.019.
- 850 [8] R. D. Rosenkrantz, Inference, method and decision: Towards a
851 Bayesian philosophy of science, Vol. 115, Springer, 1977. doi:10.1007/
852 978-94-010-1237-9.
- 853 [9] H. Jeffreys, *Theory of Probability*, 3rd Edition, Oxford University Press,
854 1939.

- 855 [10] D. V. Lindley, A statistical paradox, *Biometrika* 44 (1-2) (1957) 187–192.
856 doi:10.1093/biomet/44.1-2.187.
- 857 [11] H. Jeffreys, Some Tests of Significance, Treated by the Theory of Prob-
858 ability, *Math Proc Cambridge* 31 (2) (1935) 203–222. doi:10.1017/
859 S030500410001330X.
- 860 [12] S. F. Gull, Bayesian inductive inference and maximum entropy, in:
861 Maximum-entropy and Bayesian methods in Science and Engineering, Vol.
862 31-32, Springer, 1988, pp. 53–74. doi:10.1007/978-94-009-3049-0_4.
- 863 [13] W. H. Jefferys, J. O. Berger, Ockham’s Razor and Bayesian Analysis, *Am*
864 *Sci* 80 (1) (1992) 64–72.
- 865 [14] D. J. MacKay, Bayesian interpolation, *Neural Comput* 4 (3) (1992) 415–
866 447. doi:10.1162/neco.1992.4.3.415.
- 867 [15] R. E. Kass, A. E. Raftery, Bayes factors, *J Am Stat Assoc* 90 (430) (1995)
868 773–795. doi:10.1080/01621459.1995.10476572.
- 869 [16] J. A. Hoeting, D. Madigan, A. E. Raftery, C. T. Volinsky, Bayesian
870 model averaging: a tutorial, *Stat Sci* (1999) 382–401doi:10.1214/ss/
871 1009212519.
- 872 [17] G. E. Box, N. R. Draper, Empirical model-building and response surfaces,
873 Vol. 424, Wiley New York, 1987.
- 874 [18] A. E. Gelfand, D. K. Dey, Bayesian model choice: asymptotics and exact
875 calculations, *J Roy Stat Soc B Met* (1994) 501–514.
- 876 [19] J. M. Hammersley, D. C. Handscomb, Monte Carlo methods, Vol. 1,
877 Springer Netherlands, 1964. doi:10.1007/978-94-009-5819-7.
- 878 [20] G. Schwarz, et al., Estimating the dimension of a model, *Ann Stat* 6 (2)
879 (1978) 461–464. doi:10.1214/aos/1176344136.

- 880 [21] N. G. De Bruijn, Asymptotic methods in analysis, Vol. 4, Dover Publica-
881 tions, 1970.
- 882 [22] J. Dettmer, S. E. Dosso, C. W. Holland, Model selection and Bayesian
883 inference for high-resolution seabed reflection inversion, *J Acoust Soc Am*
884 125 (2) (2009) 706–716. doi:10.1121/1.3056553.
- 885 [23] J. Dettmer, S. E. Dosso, J. C. Osler, Bayesian evidence computation for
886 model selection in non-linear geoaoustic inference problems, *J Acoust Soc*
887 *Am* 128 (6) (2010) 3406–3415. doi:10.1121/1.3506345.
- 888 [24] G. Steininger, S. E. Dosso, C. W. Holland, J. Dettmer, Estimating seabed
889 scattering mechanisms via Bayesian model selection, *J Acoust Soc Am*
890 136 (4) (2014) 1552–1562. doi:10.1121/1.4892752.
- 891 [25] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. Van Der Linde, Bayesian
892 measures of model complexity and fit, *J Roy Stat Soc B* 64 (4) (2002)
893 583–639.
- 894 [26] P. J. Green, Reversible jump Markov chain Monte Carlo computation and
895 Bayesian model determination, *Biometrika* 82 (4) (1995) 711–732. doi:
896 10.1093/biomet/82.4.711.
- 897 [27] T. Bodin, M. Sambridge, Seismic tomography with the reversible jump
898 algorithm, *Geophys J Int* 178 (3) (2009) 1411–1436. doi:10.1111/j.
899 1365–246X.2009.04226.x.
- 900 [28] T. Bodin, M. Sambridge, H. Tkalčić, P. Arroucau, K. Gallagher, N. Rawl-
901 inson, Transdimensional inversion of receiver functions and surface wave
902 dispersion, *J Geophys Res-Sol Ea* 117 (B2) (2012) 1–24. doi:10.1029/
903 2011JB008560.
- 904 [29] M. Sambridge, K. Gallagher, A. Jackson, P. Rickwood, Trans-dimensional
905 inverse problems, model comparison and the evidence, *Geophys J Int*
906 167 (2) (2006) 528–542. doi:10.1111/j.1365–246X.2006.03155.x.

- 907 [30] S. Chib, I. Jeliazkov, Marginal likelihood from the Metropolis-Hastings
908 output, *J Am Stat Assoc* 96 (453) (2001) 270–281. doi:10.1198/
909 016214501750332848.
- 910 [31] H. Akaike, Information theory and an extension of the maximum likelihood
911 principle, in: B. N. Petrov, F. Csaki (Eds.), 2nd International Symposium
912 on Information Theory, 1973, pp. 267–281.
- 913 [32] R. L. Kashyap, Optimal choice of AR and MA parts in autoregressive
914 moving average models, *IEEE T Pattern Anal PAMI-4* (2) (1982) 99–104.
915 doi:10.1109/TPAMI.1982.4767213.
- 916 [33] X. Li, F. T.-C. Tsai, Bayesian model averaging for groundwater head pre-
917 diction and uncertainty analysis using multimodel and multimethod, *Water*
918 *Resour Res* 45 (9) (2009) 1–14. doi:10.1029/2008WR007488.
- 919 [34] L. Marshall, D. Nott, A. Sharma, Hydrological model selection: A
920 Bayesian alternative, *Water Resour Res* 41 (10) (2005) 1–11. doi:10.
921 1029/2004WR003719.
- 922 [35] F. T.-C. Tsai, X. Li, Inverse groundwater modeling for hydraulic conduc-
923 tivity estimation using Bayesian model averaging and variance window,
924 *Water Resour Res* 44 (9) (2008) 1–15. doi:10.1029/2007WR006576.
- 925 [36] M. Ye, K. F. Pohlmann, J. B. Chapman, G. M. Pohll, D. M. Reeves,
926 A model-averaging method for assessing groundwater conceptual model
927 uncertainty, *Ground Water* 48 (5) (2010) 716–728. doi:10.1111/j.
928 1745-6584.2009.00633.x.
- 929 [37] A. Schöniger, T. Wöhling, L. Samaniego, W. Nowak, Model selection on
930 solid ground: Rigorous comparison of nine ways to evaluate Bayesian
931 model evidence, *Water Resour Res* 50 (12) (2014) 9484–9513. doi:
932 10.1002/2014WR016062.

- 933 [38] S. M. Lewis, A. E. Raftery, Estimating Bayes factors via posterior simu-
 934 lation with the Laplace-Metropolis estimator, *J Am Stat Assoc* 92 (438)
 935 (1997) 648–655. doi:10.1080/01621459.1997.10474016.
- 936 [39] J. A. Vrugt, Markov chain Monte Carlo simulation using the DREAM soft-
 937 ware package: Theory, concepts, and MATLAB implementation, *Environ*
 938 *Modell Softw* 75 (2016) 273–316. doi:10.1016/j.envsoft.2015.08.013.
- 939 [40] J. A. Vrugt, C. J. Ter Braak, M. P. Clark, J. M. Hyman, B. A. Robinson,
 940 Treatment of input uncertainty in hydrologic modeling: Doing hydrology
 941 backward with Markov chain Monte Carlo simulation, *Water Resour Res*
 942 44 (12). doi:10.1029/2007WR006720.
- 943 [41] E. Laloy, J. A. Vrugt, High-dimensional posterior exploration of hydrologic
 944 models using multiple-try DREAM_{ZS} and high-performance computing,
 945 *Water Resour Res* 48 (1) (2012) 1–18. doi:10.1029/2011WR010608.
- 946 [42] J. Chen, S. Hubbard, Y. Rubin, Estimating the hydraulic conductivity at
 947 the South Oyster Site from geophysical tomographic data using Bayesian
 948 techniques based on the normal linear regression model, *Water Resour Res*
 949 37 (6) (2001) 1603–1613. doi:10.1029/2000WR900392.
- 950 [43] J. Chen, S. Hubbard, Y. Rubin, C. Murray, E. Roden, E. Majer, Geochemi-
 951 cal characterization using geophysical data and Markov Chain Monte Carlo
 952 methods: A case study at the South Oyster bacterial transport site in Vir-
 953 ginia, *Water Resour Res* 40 (12) (2004) 1–14. doi:10.1029/2003WR002883.
- 954 [44] S. S. Hubbard, J. Chen, J. Peterson, E. L. Majer, K. H. Williams, D. J.
 955 Swift, B. Mailloux, Y. Rubin, Hydrogeological characterization of the South
 956 Oyster Bacterial Transport Site using geophysical data, *Water Resour Res*
 957 37 (10) (2001) 2431–2456. doi:10.1029/2001WR000279.
- 958 [45] N. Linde, A. Tryggvason, J. E. Peterson, S. S. Hubbard, Joint inversion
 959 of crosshole radar and seismic traveltimes acquired at the South Oyster

- 960 Bacterial Transport Site, *Geophysics* 73 (4) (2008) G29–G37. doi:10.
961 1190/1.2937467.
- 962 [46] N. Linde, J. A. Vrugt, Distributed soil moisture from crosshole ground-
963 penetrating radar travel times using stochastic inversion, *Vadose Zone J*
964 12 (1) (2013) 1–21. doi:10.2136/vzj2012.0101.
- 965 [47] W. K. Hastings, Monte Carlo sampling methods using Markov chains and
966 their applications, *Biometrika* 57 (1) (1970) 97–109. doi:10.1093/biomet/
967 57.1.97.
- 968 [48] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller,
969 Equation of state calculations by fast computing machines, *J Chem Phys*
970 21 (6) (1953) 1087–1092. doi:10.1063/1.1699114.
- 971 [49] C. Robert, G. Casella, Monte Carlo statistical methods, Springer Science
972 & Business Media, 2013. doi:10.1007/978-1-4757-4145-2.
- 973 [50] F. James, Monte Carlo theory and practice, *Rep Prog Phys* 43 (9) (1980)
974 1145–1189. doi:10.1088/0034-4885/43/9/002.
- 975 [51] P. J. Rousseeuw, B. C. Van Zomeren, Unmasking multivariate outliers and
976 leverage points, *J Am Stat Assoc* 85 (411) (1990) 633–639. doi:10.1080/
977 01621459.1990.10474920.
- 978 [52] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from
979 incomplete data via the EM algorithm, *J Roy Stat Soc B Met* 39 (1) (1977)
980 1–38.
- 981 [53] L. Hoogerheide, A. Opschoor, H. K. Van Dijk, A class of adaptive im-
982 portance sampling weighted EM algorithms for efficient and robust pos-
983 terior and predictive simulation, *J Econometrics* 171 (2) (2012) 101–120.
984 doi:10.2139/ssrn.2027967.
- 985 [54] K. Perrakis, I. Ntzoufras, E. G. Tsionas, On the use of marginal posteriors
986 in marginal likelihood estimation via importance sampling, *Comput Stat*
987 *Data An* 77 (2014) 54–69. doi:10.1016/j.csda.2014.03.004.

- 988 [55] T. Lochbühler, S. J. Breen, R. L. Detwiler, J. A. Vrugt, N. Linde, Probabilistic electrical resistivity tomography of a CO₂ sequestration analog, *J Appl Geophys* 107 (2014) 80–92. doi:10.1016/j.jappgeo.2014.05.013.
- 989
- 990
- 991 [56] T. Lochbühler, J. A. Vrugt, M. Sadegh, N. Linde, Summary statistics from training images as prior information in probabilistic inversion, *Geophys J Int* 201 (1) (2015) 157–171. doi:10.1093/gji/ggv008.
- 992
- 993
- 994 [57] M. Rosas-Carbajal, N. Linde, T. Kalscheuer, J. A. Vrugt, Two-dimensional probabilistic inversion of plane-wave electromagnetic data: methodology, model constraints and joint inversion with electrical resistivity data, *Geophys J Int* 196 (3) (2013) 1508–1524. doi:10.1093/gji/ggt482.
- 995
- 996
- 997
- 998 [58] M. Rosas-Carbajal, N. Linde, J. Peacock, F. Zyserman, T. Kalscheuer, S. Thiel, Probabilistic 3-D time-lapse inversion of magnetotelluric data: application to an enhanced geothermal system, *Geophys J Int* 203 (3) (2015) 1946–1960. doi:10.1093/gji/ggv406.
- 999
- 1000
- 1001
- 1002 [59] A. Gelman, D. B. Rubin, Inference from iterative simulation using multiple sequences, *Stat Sci* 7 (4) (1992) 457–472. doi:10.1214/ss/1177011136.
- 1003
- 1004 [60] R. Van Haasteren, *Gravitational Wave Detection and Data Analysis for Pulsar Timing Arrays*, Springer Science & Business Media, 2013, Ch. Marginal likelihood calculation with MCMC methods, pp. 99–120. doi:10.1007/978-3-642-39599-4.
- 1005
- 1006
- 1007
- 1008 [61] E. Laloy, N. Linde, D. Jacques, J. A. Vrugt, Probabilistic inference of multi-Gaussian fields from indirect hydrological data using circulant embedding and dimensionality reduction, *Water Resour Res* 51 (6) (2015) 4224–4243. doi:10.1002/2014WR016395.
- 1009
- 1010
- 1011
- 1012 [62] S. Pride, Governing equations for the coupled electromagnetics and acoustics of porous media, *Phys Rev B* 50 (21) (1994) 15678–15696. doi:10.1103/PhysRevB.50.15678.
- 1013
- 1014

- 1015 [63] P. Podvin, I. Lecomte, Finite difference computation of traveltimes in very
1016 contrasted velocity models: a massively parallel approach and its associated
1017 tools, *Geophys J Int* 105 (1) (1991) 271–284. doi:10.1111/j.1365-246X.
1018 1991.tb03461.x.
- 1019 [64] A. Schöniger, W. A. Illman, T. Wöhling, W. Nowak, Finding the right
1020 balance between groundwater model complexity and experimental effort
1021 via Bayesian model selection, *J Hydrol* 531 (2015) 96–110. doi:10.1016/
1022 j.jhydrol.2015.07.047.
- 1023 [65] A. Schöniger, T. Wöhling, W. Nowak, A statistical concept to assess the
1024 uncertainty in Bayesian model weights and its impact on model ranking,
1025 *Water Resour Res* 51 (9) (2015) 7524–7546. doi:10.1002/2015WR016918.
- 1026 [66] S. Strebelle, Conditional simulation of complex geological structures using
1027 multiple-point statistics, *Math Geol* 34 (1) (2002) 1–21. doi:10.1023/A:
1028 1014009426274.
- 1029 [67] M. A. Beaumont, W. Zhang, D. J. Balding, Approximate Bayesian compu-
1030 tation in population genetics, *Genetics* 162 (4) (2002) 2025–2035.
- 1031 [68] P. Marjoram, J. Molitor, V. Plagnol, S. Tavaré, Markov chain Monte Carlo
1032 without likelihoods, *P Natl Acad Sci USA* 100 (26) (2003) 15324–15328.
1033 doi:10.1073/pnas.0306899100.
- 1034 [69] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, M. W. Feldman, Popu-
1035 lation growth of human Y chromosomes: a study of Y chromosome mi-
1036 crosatellites, *Mol Biol Evol* 16 (12) (1999) 1791–1798. doi:10.1093/
1037 oxfordjournals.molbev.a026091.
- 1038 [70] S. Tavaré, D. J. Balding, R. C. Griffiths, P. Donnelly, Inferring coalescence
1039 times from DNA sequence data, *Genetics* 145 (2) (1997) 505–518.
- 1040 [71] M. Bosch, Lithologic tomography: From plural geophysical data to lithol-
1041 ogy estimation, *J Geophys Res-Sol Ea* 104 (B1) (1999) 749–766. doi:
1042 10.1029/1998JB900014.

- 1043 [72] M. Moorkamp, P. G. Lelivre, N. Linde, A. Khan, Integrated Imaging of
1044 the Earth: Theory and Applications, Vol. 218, John Wiley & Sons, 2016.
1045 doi:10.1002/9781118929063.
- 1046 [73] W. Handley, M. Hobson, A. Lasenby, POLYCHORD: nested sampling for
1047 cosmology, Mon Not R Astron Soc 450 (1) (2015) L61–L65. doi:10.1093/
1048 mnrasl/slv047.
- 1049 [74] J. Skilling, Bayesian computation in big spaces: Nested sampling and
1050 Galilean Monte Carlo, AIP Conf Proc 1443 (1) (2012) 145–156. doi:
1051 10.1063/1.3703630.