Hydrogeological model selection among complex spatial priors

C. Brunetti[1], M. Bianchi[2], G. Pirot[1], N. Linde[1]

[1]Applied and Environmental Geophysics Group, Institute of Earth Sciences, University of Lausanne, 1015 Lausanne, Switzerland

[2]British Geological Survey, Environmental Science Centre, Nottingham, UK

Key Points:

- Full Bayesian method for model selection among geologically-realistic conceptual models
- Thermodynamic integration and stepping-stone sampling provide consistent ranking of conceptual models
- Method applied to solute concentration data collected during a tracer test at the MADE site

Corresponding author: Carlotta Brunetti, Carlotta.Brunetti@unil.ch

Abstract

Hydrogeological field studies rely often on a single conceptual representation of the subsurface. This is problematic since the impact of a poorly chosen conceptual model on predictions might be significantly larger than the one caused by parameter uncertainty. Furthermore, conceptual models often need to incorporate geological concepts and patterns in order to provide meaningful uncertainty quantification and predictions. Consequently, several geologically-realistic conceptual models should ideally be considered and evaluated in terms of their relative merits. Here, we propose a full Bayesian methodology based on Markov chain Monte Carlo (MCMC) to enable model selection among 2D conceptual models that are sampled using training images and concepts from multiple-point statistics (MPS). More precisely, power posteriors for the different conceptual subsurface models are sampled using sequential geostatistical resampling and Graph Cuts. To demonstrate the methodology, we compare and rank five alternative conceptual geological models that have been proposed in the literature to describe aquifer heterogeneity at the MAcroDispersion Experiment (MADE) site in Mississippi, USA. We consider a small-scale tracer test (MADE-5) for which the spatial distribution of hydraulic conductivity impacts multilevel solute concentration data observed along a 2D transect. The thermodynamic integration and the stepping-stone sampling methods were used to compute the evidence and associated Bayes factors using the computed power posteriors. We find that both methods are compatible with MPS-based inversions and provide a consistent ranking of the competing conceptual models considered.

1 Introduction

The geological structure of the subsurface is a key controlling factor on groundwater flow and solute transport in aquifers [*Maliva*, 2016; *Renard and Allard*, 2013; *Zheng and Gorelick*, 2003] and, therefore, it needs to be properly represented and accounted for in modelling studies. The needs for quantitative and reliable subsurface modelling and management [*Refsgaard and Henriksen*, 2004; *Scheidt et al.*, 2018] are driving hydrogeologists to consider conceptual models with increasing geological realism and complexity (e.g., see reviews by *Linde et al.* [2015a]; *Hu and Chugunova* [2008]). Traditionally, (hydro)geological subsurface heterogeneity has often been described in terms of mean values and covariances of the relevant physical properties (e.g., through the widely used multi-Gaussian models). However, such conceptualisations may be too simplistic in certain subsurface systems and, therefore, insufficient to accurately reproduce and predict flow and transport processes [*Gómez-Hernández and Wen*, 1998; *Zinn and Harvey*, 2003; *Journel and Zhang*, 2006; *Kerrou et al.*, 2008]. Multiple-point statistics (MPS) [*Guardiano and Srivastava*, 1993; *Strebelle*, 2002; *Hu and Chugunova*, 2008; *Mariethoz and Caers*, 2014] offers a means to effectively reproduce complex geological structures such as curvilinear features. By using a training image, MPS enables geostatistical simulations that honour point data and the higher-order spatial statistics that are captured in the training image. The training image is a conceptual representation summarising prior geological understanding about the system under study. It can be constructed from sketches drawn by hand, digitalised outcrops or generated by, for example, process-imitating, structure-imitating, or descriptive simulation methods [*Koltermann and Gorelick*, 1996; *De Marsily et al.*, 2005].

In many real world applications, generally because of the sparsity of direct observations, several alternative conceptualisations of subsurface heterogeneity (e.g., describing the spatial distribution of hydraulic conductivity) might be plausible and proposed by one or several experts. Unfortunately, uncertainty pertaining to the choice of the conceptual model is often ignored in modelling studies, even if it might be a dominant source of uncertainty [*Bond et al.*, 2007; *Rojas et al.*, 2008; *Refsgaard et al.*, 2012; *Lark et al.*, 2014; *Scheidt et al.*, 2018; *Randle et al.*, 2018]. Indeed, geostatistical model realisations generated from one training image might lead to a vastly different range of predictions than those generated from another training image, as shown, for example, by *Pirot et al.* [2015]. Conceptual uncertainty should, therefore, be integrated in modelling and inversion studies. Ideally, this should be achieved by using formal methods to test and rank alternative conceptual geological models based on available hydrogeological and geophysical data [*Linde*, 2014; *Linde et al.*, 2015a; *Schöniger et al.*, 2014; *Dettmer et al.*, 2010]. Bayesian model selection [*Jeffreys*, 1935, 1939; *Kass and Raftery*, 1995] offers a quantitative approach to perform such comparisons by computing the so called evidence (i.e., the denominator in Bayes' theorem) which allows to identify the conceptual model, in a chosen set, that is the most supported by the data. However, a complication arises when performing Bayesian model selection with complex spatial priors that are represented by training images. Most MPS-based inversions are non-parametric which implies that they rely on samples being drawn proportionally to the prior distribution, while it is generally not possible within a MPS framework to evaluate the prior probability of a given model proposal.

Hence, MPS-based inversions cannot build on many state-of-the-art concepts to enhance the performance of the MCMC (e.g., *Laloy and Vrugt* [2012]) and associated approaches for calculating the evidence [*Volpi et al.*, 2017; *Brunetti et al.*, 2017]. Similarly, it is not possible within a MPS-framework to calculate approximate evidence estimates using the Laplace-Metropolis method [*Lewis and Raftery*, 1997].

It is only recently that MPS-based inversions have been proposed (see review by *Linde et al.* [2015a]). Markov chain Monte Carlo (MCMC) inversions with MPS (e.g., *Mariethoz et al.* [2010a]; *Hansen et al.* [2012]) generally rely on model proposals obtained by sequential geostatistical resampling of the prior (Gibbs sampling) that are used within the extended Metropolis algorithm to accept model proposals based on the likelihood ratio [*Mosegaard and Tarantola*, 1995]. Sequential geostatistical resampling generates model proposals of the spatially-distributed parameters of interest by conditional resimulations of a random fraction of the current field proportionally to the prior as defined by the training image. There exist several MPS methods to sample complex spatial priors with sequential Gibbs sampling. Examples include the versatile direct sampling method [*Mariethoz et al.*, 2010b] or the recent Graph Cuts approach [*Zahner et al.*, 2016; *Li et al.*, 2016] that enables speed-ups by one to two orders of magnitude. Since high-dimensional MCMC inversions necessitate many evaluations of model proposals by forward modelling, it is essential that the geostatistical model proposal process is fast compared to the forward simulation time while ensuring model realisations of high quality that honour geological patterns in the training image. Various advances have been made to enhance MPS-based inversions both in a non-parametric MCMC framework (e.g., parallel tempering by *Laloy et al.* [2016]) and in a parametric framework using, for example, spatial generative adversarial neural networks [*Laloy et al.*, 2018]. Also, ensemble-based exploration schemes have been explored [*Jäggli et al.*, 2017].

State-of-the-art evidence estimators that are compatible with non-parametric spatial priors include thermodynamic integration [*Gelman and Meng*, 1998; *Friel and Pettitt*, 2008a], stepping-stone [*Xie et al.*, 2011] and nested sampling [*Skilling*, 2004, 2006]. The thermodynamic integration method takes the name from its original application, which was to compute the difference in a thermodynamic property (usually free energy) of a system at two given states. Thermodynamic integration and the stepping-stone method sample from a sequence of so-called power posterior distributions that connect the prior to the posterior distribution. The nested sampling method is based on a constrained local sampling procedure in which the prior distribution is sampled under the constraint of a lower bound on the log-likelihood function that increases with time. Thermodynamic integration and nested sampling transform the evidence, that is, a multi-dimensional integral over the parameter space, into a one-dimensional integral over unit range in the log-likelihood space. The stepping-stone sampling estimator approximates the evidence by importance sampling using the power posteriors as importance distributions. To the best of our knowledge, thermodynamic integration and stepping-stone sampling have never been used to estimate the evidence of subsurface models built with MPS in the context of Bayesian model selection, while this is the case for nested sampling [*Elsheikh et al.*, 2015]. Recent studies in hydrology suggest that nested sampling is less accurate and stable than thermodynamic integration [*Liu et al.*, 2016; *Zeng et al.*, 2018] and that it is strongly dependent on the efficiency of the constrained local sampling procedure. Unfortunately, MPS-based inversions cannot benefit from recent improvements in constrained local sampling approaches as they require parametric (analytical) forms of the prior [*Schöniger et al.*, 2014; *Liu et al.*, 2016; *Zeng et al.*, 2018; *Cao et al.*, 2018]. Even if thermodynamic integration and stepping-stone sampling are computationally expensive, they are easily parallelised such that the computational time is equivalent to the time needed to run a single MCMC chain. Moreover, these two methods are easy to implement and flexible in the sense that any suitable MCMC method can, provided minimal changes, be used to explore the power posterior distributions. The classical brute force Monte Carlo (MC) method [*Hammersley and Handscomb*, 1964] can also be used to estimate the evidence when considering non-parametric spatial priors. However, *Brunetti et al.* [2017] show that MC often requires a prohibitive computational time to obtain reliable evidence estimates even for very simple subsurface conceptualizations (e.g., layered models) when considering as few as seven unknowns. This limits its application to realistic high-dimensional MPS-based conceptual models.

One way to circumvent the challenges of non-parametric priors in Bayesian model selection is to reduce the model parameter space, for example, by cluster-based polynomial chaos expansion [*Bazargan and Christie*, 2017] or by truncated discrete cosine transform combined with summary metrics from training images [*Lochbühler et al.*, 2015]. Bayesian inference and model selection is then applied on the reduced dimension space whose prior distribution is parametric (e.g., multivariate Gaussian distribution). The main drawback of such approaches is that truncation may smoothen sharp interfaces found in the training images.

In this study, we propose the first full Bayesian method that enables Bayesian model selection among geologically-realistic conceptual subsurface models. To do so, we combine sequential geostatistical resampling based on Graph Cuts, the extended Metropolis acceptance criterion and evidence estimation by power posteriors using either thermodynamic integration or stepping-stone sampling. The advantages and the drawbacks of this new methodology are assessed using a challenging application. In this study, we compare and rank five alternative conceptual geological models that have been proposed in the literature to characterise the spatial heterogeneity of the aquifer at the Macrodispersion Experiment (MADE) site in Mississippi, USA [*Zheng et al.*, 2011]. Among this set of five conceptual models of hydraulic conductivity spatial distribution, we aim to identify the one that is in the best agreement with multilevel concentration data acquired during a small-scale dipole tracer test (MADE-5) [*Bianchi et al.*, 2011a]. The case-study at the MADE site is used to demonstrate the ability of our Bayesian model selection method to deal with widely different conceptual hydrogeological models. We stress that the 2D modeling framework used herein limits our ability to generalize the findings to actual 3D field conditions. Extensions to 3D is methodologically straightforward, but computationally very challenging.

2 Theory

2.1 Bayesian inference and model selection

Bayesian inference approaches express the posterior pdf, $p(\theta|\widetilde{\mathbf{Y}})$, of a set of unknown model parameters, $\theta = \{\theta_1, \ldots, \theta_d\}$, given n measurements, $\widetilde{\mathbf{Y}} = \{\widetilde{y}_1, \ldots, \widetilde{y}_n\}$, via Bayes' theorem

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

The prior pdf, $p(\theta)$, quantifies all the information that is available about the model parameters before considering the observed data. Typically, $p(\theta)$ is represented by multivariate analytical functions (e.g., Gaussian, uniform, exponential) describing marginal distributions of each parameter and their spatial correlation. With the advent of MPS methods, higher-order spatial statistics of θ can be incorporated in inversions by means of training images. In this case, the description of prior knowledge is typically non-parametric and sequential geostatistical resampling techniques are used to sample $p(\theta)$. The likelihood function, $p(\widetilde{\mathbf{Y}}|\theta)$, summarises in a single scalar value the probability that the observed data has been generated by a proposed set of model parameters. We consider a Gaussian likelihood characterised by uncorrelated and normally distributed measurement errors with constant standard deviation, $\sigma_{\widetilde{\mathbf{Y}}}$,

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

As the residuals between the observed data, $\widetilde{y}_h$, and the simulated forward responses, $\mathcal{F}_h(\theta)$, tends toward 0, the likelihood increases and, in particular, $p(\widetilde{\mathbf{Y}}|\theta) \rightarrow \left(\sqrt{2\pi\sigma_{\widetilde{\mathbf{Y}}}^2}\right)^{-n}$. The denominator in Bayes' theorem is the evidence (or marginal likelihood), $p(\widetilde{\mathbf{Y}})$, and it is the cornerstone quantity in most Bayesian model selection problems. It should be noted, however, that the explicit computation of the evidence can be avoided by using reversible jump (trans-dimensional) MCMC methods [*Green*, 1995]. The conceptual model with the highest evidence [*Jeffreys*, 1935, 1939] is the one that is the most supported by the data. A noteworthy feature of the evidence is that it implicitly accounts for the trade-off between goodness of fit and model complexity [*Gull*, 1988; *Jeffreys*, 1939; *Jefferys and Berger*, 1992; *MacKay*, 1992]. More precisely, the evidence quantifies how likely it is that a given conceptual model, $\eta \in \mathbb{N}$, with model parameters, θ, and prior distribution, $p(\theta|\eta)$, has generated the data $\widetilde{\mathbf{Y}}$,

$$p(\widetilde{\mathbf{Y}}|\eta) = \int p(\widetilde{\mathbf{Y}}|\theta, \eta)p(\theta|\eta)d\theta. \tag{3}$$

The evidence is used to calculate Bayes factors [*Kass and Raftery*, 1995], that is, evidence ratios of one conceptual model with respect to an other. For instance, the Bayes factor of η_1 with respect to η_2, or $B_{(\eta_1,\eta_2)}$, is defined as

$$B_{(\eta_1,\eta_2)} = \frac{p(\widetilde{\mathbf{Y}}|\eta_1)}{p(\widetilde{\mathbf{Y}}|\eta_2)}. \tag{4}$$

Conceptual models with large Bayes factors are preferred statistically and the conceptual model with the largest evidence is the one that best honours the data on average over its prior. However, the evidence computation is analytically intractable for most problems of interest and the multi-dimensional integral in Eq. 3 must be approximated by numerical means. In this work, the different conceptual models represent alternative spatial representations of hydraulic conductivity in the subsurface.

2.2 Evidence estimation by power posteriors

Thermodynamic integration, also called path sampling [*Gelman and Meng*, 1998], and stepping-stone sampling [*Xie et al.*, 2011] are two methods to estimate the evidence (Eq. 3) numerically. The key idea behind both methods is to sample from a sequence of so-called power posterior distributions, $p_\beta(\theta|\widetilde{\mathbf{Y}})$, in order to create a path in the probability density space that connects the prior to the posterior distribution [*Friel and Pettitt*, 2008a]. The power posterior distribution is proportional to the prior pdf multiplied by the likelihood function raised to the power of $\beta \in [0, 1]$:

$$p_\beta(\theta|\widetilde{\mathbf{Y}}) \propto p(\theta)p(\widetilde{\mathbf{Y}}|\theta)^\beta. \tag{5}$$

Decreasing β has the effect of flattening the likelihood function. For $\beta = 1$, the posterior distribution is sampled, $p_1(\theta|\widetilde{\mathbf{Y}}) \propto p(\theta)p(\widetilde{\mathbf{Y}}|\theta)$; for $\beta = 0$, the prior distribution is sampled, $p_0(\theta|\widetilde{\mathbf{Y}}) \propto p(\theta)$. In thermodynamic integration and stepping-stone sampling, the priors are assumed to be proper and a sequence of β-values needs to be defined (see Section 2.2.3). For each β value, one (or more) MCMC runs are used to draw N samples from the corresponding power posterior distribution and the corresponding likelihood values are recorded. The Markov chains for the different β-values can be run independently in parallel or sequentially from $\beta = 0$ to $\beta = 1$ (serial MCMC) as described in *Friel and Pettitt* [2008a]. Thermodynamic integration and stepping-stone sampling have several attractive characteristics: (1) the total computing time is equivalent to a normal MCMC inversion provided that all MCMC runs are carried out in parallel, (2) they can be applied for any MCMC inversion method with only minimal intervention (it is only necessary to add the exponent β to the likelihood function) and (3) the only information needed is the series of likelihoods obtained from MCMC simulations with different β-values. Once the power posterior distributions have been sampled, the thermodynamic integration and stepping-stone sampling methods use the recorded likelihood values in two different ways to estimate the evidence (Sections 2.2.1-2.2.2).

2.2.1 Thermodynamic integration

Thermodynamic integration reduces the multi-dimensional integral of Eq. 3 into a one-dimensional integral of the expectation of the log-likelihood, $\log p(\widetilde{\mathbf{Y}}|\theta, \eta)$, as:

$$\log p(\widetilde{\mathbf{Y}}|\eta) = \int_0^1 \mathrm{E}_{\theta|\widetilde{\mathbf{Y}},\beta}\left[\log p(\widetilde{\mathbf{Y}}|\theta, \eta)\right] d\beta. \tag{6}$$

For the derivation of Eq. 6, we refer to *Friel and Pettitt* [2008a] and *Lartillot and Philippe* [2006]. The integral in Eq. 6 is estimated by a quadrature approximation over a discrete set of β-values, $0{=}\beta_1 < \cdots < \beta_j < \cdots < \beta_J{=}1$. To simplify the notation, we define the expectations of the log-likelihood functions as $\ell_j \equiv \mathrm{E}_{\theta|\widetilde{\mathbf{Y}},\beta_j}\left[\log p(\widetilde{\mathbf{Y}}|\theta,\eta)\right]$ and their corresponding variances as $\sigma_j^2 \equiv \mathrm{V}_{\theta|\widetilde{\mathbf{Y}},\beta_j}\left[\log p(\widetilde{\mathbf{Y}}|\theta,\eta)\right]$. In this work, we use the corrected composite trapezoidal rule:

$$\log p(\widetilde{\mathbf{Y}}|\eta) \approx \sum_{j=2}^{J} \frac{(\beta_j - \beta_{j-1})}{2}(\ell_j + \ell_{j-1}) - \sum_{j=2}^{J} \frac{(\beta_j - \beta_{j-1})^2}{12}(\sigma_j^2 - \sigma_{j-1}^2), \tag{7}$$

which provides more accurate estimates compared with the classical composite trapezoidal rule (first term in Eq. 7) as it also considers the second-order correction term (second term in Eq. 7). This corrected composite trapezoidal rule was originally employed by *Friel et al.* [2014] and later used by other authors including *Oates et al.* [2016] and *Grzegorczyk et al.* [2017].

The accuracy of the resulting evidence estimates depends on how the β-values are discretised, the number of β-values used, J, (details provided in Section 2.2.3), the number, N, and the degree of correlation of the power posterior samples obtained by MCMC. The uncertainties associated with the evidence estimation by thermodynamic integration are often summarised by two error types: the sampling error, e_s, and the discretisation error, e_d [*Lartillot and Philippe*, 2006; *Calderhead and Girolami*, 2009]. The sampling error is related to the standard errors of the MCMC posterior expectations of the log-likelihoods obtained for each β_j. To avoid underestimation of these errors, the autocorrelation in the MCMC samples should be accounted for in order to calculate the effective sample size, N_{eff}, (i.e., number of independent samples within each MCMC chain) as suggested by *Kass et al.* [1998]. The effective sample size is defined as:

$$N_{\text{eff},j} = \frac{N_j}{1 + 2\sum_{z=1}^{\infty} \rho_j(z)}, \tag{8}$$

where $\rho_j(z)$ is the autocorrelation at lag z. Applying the rules for uncertainty propagation to the first leading term in Eq. 7 and assuming the errors of ℓ_j to be independent of those associated to ℓ_{j-1}, the sampling error is:

$$\sigma_s^2 = \sum_{j=2}^{J} \frac{(\beta_j - \beta_{j-1})^2}{4}\left(\frac{\sigma_j^2}{N_{\text{eff},j}} + \frac{\sigma_{j-1}^2}{N_{\text{eff},j-1}}\right). \tag{9}$$

Discretisation errors arise as the continuous integral of Eq. 6 is estimated using a finite number of evaluation points (Eq. 7). Following *Lartillot and Philippe* [2006], *Baele et al.* [2013] and *Friel et al.* [2014], we define e_d as the worst-case discretisation error that arises from the approximation of Eq. 6 with a rectangular rule. Hence, e_d is half the difference of the areas between the upper and lower step functions and it can be interpreted as the variance of the trapezoidal rule:

$$\sigma_d^2 = \sum_{j=2}^{J} \frac{(\beta_j - \beta_{j-1})^2}{4} (\ell_j - \ell_{j-1})^2. \tag{10}$$

As a consequence, the variance on the evidence estimates can be summarised as $\widehat{\mathrm{Var}} \log p(\widetilde{\mathbf{Y}}|\eta) = \sigma_d^2 + \sigma_s^2$.

2.2.2 Stepping-stone sampling

Stepping-stone sampling [*Xie et al.*, 2011] computes the evidence by combining power posteriors with importance sampling. The key underlying idea is to write the evidence as the ratio, r, of the normalising factors in Bayes' theorem for β=1 (posterior sampling) and β=0 (prior sampling):

$$r = \frac{p(\widetilde{\mathbf{Y}}|\eta, \beta = 1)}{p(\widetilde{\mathbf{Y}}|\eta, \beta = 0)}. \tag{11}$$

Since the prior integrates to one, the evidence is equivalent to r as $p(\widetilde{\mathbf{Y}}|\eta, \beta = 0)$ equals 1. The ratio can be expressed as a product of J ratios, r_j:

$$r = \prod_{j=2}^{J} r_{j-1} = \prod_{j=2}^{J} \frac{p(\widetilde{\mathbf{Y}}|\eta, \beta_j)}{p(\widetilde{\mathbf{Y}}|\eta, \beta_{j-1})}. \tag{12}$$

Then, importance sampling is applied to the numerator and denominator of Eq. 12 using the power posterior $p_{\beta_{j-1}}(\theta|\widetilde{\mathbf{Y}})$ as the importance distribution:

$$r_{j-1} = \frac{1}{N} \sum_{i=1}^{N} p(\widetilde{\mathbf{Y}}|\theta_{j-1,i})^{\beta_j - \beta_{j-1}} \tag{13}$$

and, finally, the log-evidence is computed as:

$$\log p(\widetilde{\mathbf{Y}}|\eta) = \sum_{j=2}^{J} \log r_{j-1} = \sum_{j=2}^{J} \log \left\{ \frac{1}{N} \sum_{i=1}^{N} \exp \left[(\beta_j - \beta_{j-1}) \cdot \log p(\widetilde{\mathbf{Y}}|\theta_{j-1,i}) \right] \right\}. \tag{14}$$

In contrast to thermodynamic integration, the evidence estimated by stepping-stone sampling does not suffer from discretisation errors. The sampling error can be evaluated as:

$$\widehat{\mathrm{Var}} \log p(\widetilde{\mathbf{Y}}|\eta) = \sum_{j=2}^{J} \frac{1}{N_{\mathrm{eff},j-1} \cdot N} \sum_{i=1}^{N} \left(\frac{p(\widetilde{\mathbf{Y}}|\theta_{j-1,i})^{\beta_j - \beta_{j-1}}}{r_{j-1}} - 1 \right)^2. \tag{15}$$

The derivation of Eq. 14 and 15 appears in *Xie et al.* [2011]; *Fan et al.* [2011], and interested readers are referred to this publication for further details. The only difference in our Eq. 15 is that we consider the effective sample size as defined in Eq. 8. Note that Eq. 13 is only valid for the specific choice of $p_{\beta_{j-1}}(\theta|\widetilde{\mathbf{Y}})$ as the importance distribution.

2.2.3 Discretisation scheme for β-values

For small increases of β close to 0, l_j increases dramatically and the corresponding power posteriors quickly turn from being similar to the prior to being similar to the posterior distribution (e.g, *Friel et al.* [2014]; *Oates et al.* [2016]; *Liu et al.* [2016]). As a consequence, the accuracy of the evidence estimates increases when placing most of the β-values close to 0 (e.g., *Friel and Pettitt* [2008b]; *Liu et al.* [2016]; *Grzegorczyk et al.* [2017]). This is especially true for the thermodynamic integration method that estimates the evidence as the area below the curve of the expectation of the log-likelihood, l_j, as a function of β_j (Eq. 6). Starting from an initial set of sampling points, *Liu et al.* [2016] use an empirical method that places additional β-values based on a qualitative search for locations where l_j changes strongly in order to target additional β-values to use. However, this method is subjective and it increases the computing time when using parallel computations as the β-values are not defined at the outset. *Friel and Pettitt* [2008a] are the first to employ a discretisation scheme of β-values that follows a power law spacing as:

$$\beta_j = \left(\frac{j-1}{J-1}\right)^c \quad \text{with} \quad j = 1, 2 \ldots, J. \tag{16}$$

Calderhead and Girolami [2009] demonstrate that this scheme significantly improve the accuracy of the evidence estimates with respect to the uniform spacing used by *Lartillot and Philippe* [2006].

3 Method

3.1 General framework

It is common to sample the unnormalised posterior pdf of Eq. 1 with MCMC simulations. This is here achieved by combining the extended Metropolis acceptance criterion [*Mosegaard and Tarantola*, 1995] with a sequential geostatistical resampling technique (e.g., Graph Cuts) that provides conditional model proposals at each iteration featuring similar geological patterns as those found in the corresponding training image. For each proposed model, $\theta_{\mathbf{prop}}$, we calculate the forward response and compare it with the observed data and, according to the extended Metropolis algorithm, accept $\theta_{\mathbf{prop}}$ with probability:

$$\alpha = min\left\{1, \frac{p(\widetilde{\mathbf{Y}}|\theta_{\mathbf{prop}})}{p(\widetilde{\mathbf{Y}}|\theta_{\mathbf{cur}})}\right\}. \tag{17}$$

To sample the power posteriors, we simply modify the extended Metropolis acceptance criteria by raising the likelihoods in Eq. 17 with the corresponding β_k-values. We report below the overall algorithm (Algorithm 1), in which we combine model proposals based on MPS with the extended Metropolis acceptance criteria followed by evidence estimation using power posteriors.

3.2 Graph Cuts model proposals

In this work, to sample spatially correlated parameters, we rely on model proposals based on the Graph Cuts algorithm introduced by *Zahner et al.* [2016] with some of the improvements proposed by *Pirot et al.* [2017a,b]. The main steps in the Graph Cuts algorithm are depicted in Figure 1. Basically, a section of the same size as the model domain, $\theta_{\mathbf{new}}$ (Figure 1b), is randomly drawn from the training image and the absolute difference between $\theta_{\mathbf{new}}$ and the current model realisation, $\theta_{\mathbf{cur}}$ (Figure 1a), is computed and raised to the power of the cost power, δ_{cp}, [*Pirot et al.*, 2017b] to obtain the cost image, $\boldsymbol{\delta}$ = $|\theta_{\mathbf{cur}}\text{-}\theta_{\mathbf{new}}|^{\delta_{cp}}$ (Figure 1d). Two distinct regions of high cost, similar size and containing at least p pixels are randomly selected (Figure 1e). To choose these terminals, *Pirot et al.* [2017a] introduce the cutting threshold, $\delta_{th} \in [0, 100]$, defined as a percentile of max($\boldsymbol{\delta}$), which limits the possible terminals to those regions where $\boldsymbol{\delta} > \delta_{th} \cdot max(\boldsymbol{\delta})$. A patch is defined as the region enclosed by a minimum cost line separating the two terminals using the min-cut/max-flow algorithm by *Boykov and Kolmogorov* [2004] (Figure 1f) and the new model proposal, $\theta_{\mathbf{prop}}$ (Figure 1c), is built by cutting the patch from $\theta_{\mathbf{new}}$ and replacing the corresponding area in $\theta_{\mathbf{cur}}$.

We manually tune three algorithmic parameters to obtain model proposals that preserve the patterns found in the training image: the minimum number, p, of pixels in each of the two terminals, the cutting threshold, δ_{th}, and the cost power, δ_{cp}. We have set the cost power to 1 or 2 depending on the type of conceptual model considered. The main reason for using graph-cut proposals in this work is its computational speed relatively to other MPS algorithms (see comparisons by *Zahner et al.* [2016]). However, slower pixel-based geostatistical resimulation strategies that implement sequential Gibbs sampling, such as, those presented by *Mariethoz et al.* [2010b] or *Hansen et al.* [2012] could also be used.

3.3 Field site and available data

The MADE site is characterised by an unconsolidated shallow alluvial aquifer composed by a mixture of gravel, sand, and finer sediments. The high heterogeneity at the

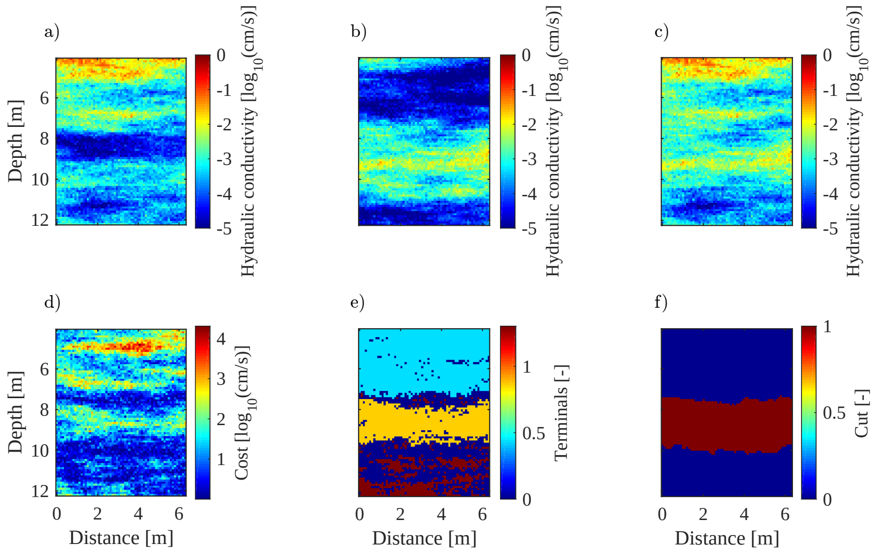


Figure 1. Illustration of how model proposals are obtained using the Graph Cuts algorithm. (a) Current model realisation, $\theta_{\mathbf{cur}}$, (b) section drawn randomly from the training image, $\theta_{\mathbf{new}}$, and (c) the resulting model proposal, $\theta_{\mathbf{prop}}$. This model proposal is obtained as follows: (d) the cost image, $\boldsymbol{\delta}$, is defined as the absolute difference raised to the cost power, δ_{cp}, that is $\boldsymbol{\delta} = |\theta_{\mathbf{cur}}\text{-}\theta_{\mathbf{new}}|^{\delta_{cp}}$, (e) two disconnected regions of high differences (light blue and orange areas) of similar size are randomly selected and (f) the cut of minimum cost that separates the two regions is calculated and the resulting dark red region is cut from (b) $\theta_{\mathbf{new}}$ and pasted into (a) $\theta_{\mathbf{cur}}$ to create (c) $\theta_{\mathbf{prop}}$.

MADE site got the attention of the hydrogeological community in the mid-1980s and numerous studies have been carried out since then (see *Zheng et al.* [2011] for a review). Previous interpretations of two large-scale tracer tests suggest that the structure is consistent with a network of highly permeable sediments embedded in a less permeable matrix [*Harvey and Gorelick*, 2000; *Feehley et al.*, 2000; *Bianchi and Zheng*, 2016]. The case-study considered herein focuses on determining the most appropriate conceptual model of hydraulic conductivity in a reduced set given the multilevel solute concentration data collected during the MADE-5 tracer experiment [*Bianchi et al.*, 2011a]. The test was performed in an array of four aligned boreholes with a maximum separation of 6 metres. The concentration data used in this work was collected in the two inner multi-level sampler (MLS) wells between the outer injection and abstraction wells, which were screened over the entire aquifer thickness. Before tracer injection, a steady-state dipole flow field was established by injecting clean water. Then, a known volume of bromide solution was injected along the entire vertical profile of the aquifer for 366 min followed by continuous injection of clean water for 32 days. The flow rates at both the injection and extraction wells were kept practically constant during all the steps of the test. Bromide concentrations in the MLS wells were recorded at 19 different times and at seven depth levels (sampling ports) in each of the two MLS wells resulting in 266 concentration measurements. Full technical details about the experiment can be found in *Bianchi et al.* [2011a]. Given the particular design of the borehole array, groundwater flow and bromide tracer transport could be simulated only along the 2D transect intercepting the four wells (the forward model used is described in Appendix A). This was necessary to reduce the computational demands in this application of the proposed Bayesian model selection method. In practice, the 2D model assumes that the concentrations measured at the inner MLS wells are mainly the result of transport along straight flow paths between the injection and the abstraction wells. To enable such 2D modeling, we performed a simple 3D-to-2D transformation of the data as described in Appendix A.

3.3.1 Conceptual models at the MADE site and corresponding training images

We consider five training images that may represent spatially distributed hydraulic conductivity fields at the MADE site (Figure 2). The multi-Gaussian training image in Figure 2a was created as a 2D unconditional realisation obtained with the Sequential Gaussian SIMulation (SGSIM) algorithm of the Stanford Geostatistical Modeling Software (SGeMS) [*Remy et al.*, 2009]. The corresponding variogram parameters (Table 1) were calculated by *Bianchi et al.* [2011a] from the analysis of more than 1000 hydraulic conductivity values estimated by means of borehole flowmeter tests [*Rehfeldt et al.*, 1992]. According to *Bianchi et al.* [2011a], the mean and variance in $\log_{10}$(cm/s) is set equal to -2.37 and 1.95, respectively.

The training images in Figure 2b-d were generated following *Linde et al.* [2015b]. The highly conductive and connected channels in an homogeneous matrix (Figure 2b) is built from the original training image of *Strebelle* [2002] modified according to the channel properties proposed by *Ronayne et al.* [2010] for the MADE site. The channel hydraulic conductivity is equal to -0.54 in $\log_{10}$(cm/s), the channel thickness is 0.2 m and the channel fraction is 3.25 %. The training image in Figure 2c is based on hydrogeological facies and their hydraulic conductivity values correspond to those of an outcrop located near the MADE site [*Rehfeldt et al.*, 1992] and reported in Table 2.

The training image in Figure 2d is chosen solely on the knowledge that the aquifer at the MADE site is constituted by alluvial deposits [*Boggs et al.*, 1992]. *Linde et al.* [2015b] and *Lochbühler et al.* [2014] used the training image of Figure 2d as derived from a detailed mapping study at the Herten site in Germany [*Bayer et al.*, 2011; *Comunian et al.*, 2011] featuring representative alluvial deposit structures and adapted it to the hydrogeological facies observed at the MADE site (Table 2).

The training image of Figure 2e is built based on five hydrogeological facies identified from lithological borehole data at the MADE site [*Bianchi and Zheng*, 2016] and reported in Table 3. This training image is a stochastic unconditional realisation that was generated following *Bianchi and Zheng* [2016].

Training images should be stationary and approach ergodicity [*Caers and Zhang*, 2004]. This implies that the type of patterns found should not change over the domain covered by the training image (stationarity). Moreover, the size of the training image should be sufficiently large (at least the double) compared to the largest pattern to enable adequate simulations (ergodicity). Small training images lead to large ergodic fluctuations that deteriorates pattern reproduction [*Renard et al.*, 2005]. Note that the smallest training image considered herein (Figure 2b) is four times wider than the size of the model domain in the horizontal direction.

In this work, we compare the five conceptual models of hydraulic conductivity that, in the following, we refer to as (1) *multi-Gaussian* as built from the training image in Figure 2a; (2) *hybrid* that consists of the highly conductive channels of Figure 2b overlaid on the multi-Gaussian background of Figure 2a; (3) *outcrop-based* built from the training image in Figure 2c; (4) *analog-based* built from the training image in Figure 2d; (5) *lithofacies-based* built from the training image in Figure 2e. This selection of conceptual models allows us to compare very different parameterisations of the spatial heterogeneity at the MADE site. Note that a full assessment of all conceptual models that has been published for the MADE site is outside the scope of this study. Since computational limitations prohibit full 3D simulations, we acknowledge that our findings in terms of the suitability of different conceptual models at the MADE site should be treated with some caution. Instead, the focus is on a new versatile methodology that enables comparison of widely different conceptual models.

3.4 Evidence estimation in practice

We discretise the power coefficients β using the commonly used power law of Eq. 16 [*Grzegorczyk et al.*, 2017; *Höhna et al.*, 2017; *Baele and Lemey*, 2013; *Xie et al.*, 2011; *Calderhead and Girolami*, 2009; *Friel and Pettitt*, 2008a]. According to these studies, the parameter c should be set equal to 3 or 5 and J as large as possible with the common choice of $20 \leq J \leq 100$. In this study, we chose $c = 5$ and $J = 40$. For each β value, we run one MCMC chain of 10^5 iterations. These choices are dictated by computational constraints. The most challenging power posterior to sample is for β=1, for which we run 3 chains to better explore the posterior distribution. Consequently, we run 42 MCMC chains for each conceptual model. Given that the log-likelihoods obtained from the MCMC simulations are the basis for evidence estimations by power posteriors, we define the burn-in period (i.e., number of MCMC iterations required before reaching the target distribution) by considering the evolution of the log-likelihoods. To assess when the log-likelihood values start to oscillate around a constant value, we apply the Geweke method [*Geweke*, 1992] on the log-likelihoods of each chain. This diagnostic compares the mean computed on the last half of the considered chain length against the one derived from a smaller interval in the beginning of the chain (in our case, 20% of the chain length). At first, the Geweke's method is applied to the whole chain (no burn-in), and if its statistics is outside the 95% confidence interval of the standard normal distribution, we apply it again after discarding the first 1%, 2%, ...,95% of the total chain length. The burn-in is determined in this way for β=1, as this is the most challenging case for which burn-in takes the longest time to achieve. The evidence estimates are computed using the thermodynamic integration method based on both the corrected trapezoidal rule (Eq. 7), as well as with the stepping-stone sampling method (Eq. 14). In order to correctly estimate the uncertainty of the evidence estimates, the effective sample size (Eq. 8) in each chain needs to be assessed. When evaluating Eq. 8, we truncate the sum in the denominator at the lag at which $\rho_j(z)$ is within 95% confidence interval of the normal distribution with standard deviation equal to the standard error of the sample autocorrelation. The evidence estimates are updated continuously after burn-in to visualise their evolution with the number of MCMC iterations. The uncertainty associated with the evidence estimates are summarised by standard errors, $\mathrm{SE} = \sqrt{\widehat{\mathrm{Var}} \log p(\widetilde{\mathbf{Y}}|\eta)}$ with corresponding 95% confidence intervals. The variances $\widehat{\mathrm{Var}} \log p(\widetilde{\mathbf{Y}}|\eta)$ are computed using Eqs. 9-10 for the thermodynamic integration and using Eq. 15 for the stepping-stone sampling method.

4 Results for the MADE-5 case study

4.1 Bayesian inference

For each of the conceptual models considered, we first show prior MPS-realisations (i.e., $\beta = 0$) of hydraulic conductivity fields that are generated with the Graph Cuts method (Figure 3). Each set of prior realisations shows considerable spatial variability and is in broad agreement with the original training image (Figure 2). This is valid for both continuous (Figure 3b), categorical (Figures 3c-e) and hybrid conceptual models (Figure 3a).

The posterior distributions (i.e., $\beta = 1$) are obtained by assuming that the standard deviation of the measurement errors, $\sigma_{\widetilde{\mathbf{Y}}}$ [mg/L], follows a log-uniform prior distribution in the range [1,10] mg/L (last column of Table 4). The lowest mean of the inferred $\sigma_{\widetilde{\mathbf{Y}}}$ is obtained for the hybrid conceptual model (5.8 mg/L) suggesting that this model enables the best match with the data. The highest $\sigma_{\widetilde{\mathbf{Y}}}$ is found for the outcrop-based model (9.4 mg/L). The acceptance rates are lower (second column in Table 4) than the ideal range between 15% and 40% proposed by *Gelman et al.* [1996], which suggests a slow convergence of the Markov chains. The burn-in time for each chain is obtained by the Geweke method (Table 4) as described in Section 3.4.

The different conceptual models provide quite different posterior distributions of the hydraulic conductivity field (Figure 4), even if certain commonalities are observed. For instance, all the posterior models have a high-conductive zone at a depth of 7 m that extends to a depth of 8 m on the right hand-side of the model domain. These features are visible in both the posterior mean and the maximum a-posteriori fields (first and second column of Figure 4). The analog- and outcrop-based conceptual models exhibit more variability in the inferred hydraulic conductivity values (Figures 4c and 4e) with respect to the others and the lithofacies-based conceptual model is characterised by the smallest posterior standard deviations (Figure 4d). The Gelman-Rubin statistic [*Gelman and Rubin*, 1992] is commonly used to assess if the MCMC chains has adequately sampled the posterior distribution, which is generally considered to be the case if this statistic is below 1.2. We see in the last column of Figure 4 that this is not the case for all pixel values, especially in the high-conductivity region, and that a larger number of iterations is required for a full convergence. However, we note that the evidence estimates are valid as long as the MCMC chains reach burn-in, while enhanced sampling decreases the estimation error.

In Figure 5, we show some of the simulated and observed breakthrough curves. We have chosen the ones at a depth of 7 m in the monitoring wells MLS-1 (Figure 5a) and MLS-2 (Figure 5b) because they correspond to a region of high conductivity (high concentrations) and the ones at a depth of 11 m that correspond to low concentrations in MLS-1 (Figure 5c) and MLS-2 (Figure 5d). Note that the range of measured concentration values spans two orders of magnitude (Figure 5). In general, the outcrop-based conceptual model is the worst in reproducing the observed breakthrough curves while the hybrid model is the best performing one; this is particularly clear in Figure 5d. Corresponding plots at all measurement locations are found in the Supporting Information. The Pearson correlation coefficients between the simulated posterior mean concentrations and the observed ones are 0.96 for the hybrid model, 0.94 for the multi-Gaussian and analog-based models, 0.91 for the lithofacies- and outcrop-based models.

4.2 Bayesian model selection

In this section, we present the estimated evidence values for each conceptual model considered. Overall, the evidence values obtained using stepping-stone sampling and thermodynamic integration based on the corrected trapezoidal rule are in good agreement with each other considering their 95% confidence intervals (Figure 6). Moreover, except for some fluctuations at the early stage after burn-in, the evidence estimates evolve only slowly as a function of the number of MCMC iterations after burn-in (Figure 6). We find that stepping-stone sampling provides evidence values that are always lower than the ones estimated with the thermodynamic integration. This behaviour is somewhat surprising as the stepping-stone sampling technique is not based on a discretisation, while this is the case for thermodynamic integration leading to an expected underestimation of the evidence. The uncertainty associated with the stepping-stone evidence estimator decreases at a sustained pace when increasing the number of MCMC iterations and it is lower than the one associated with thermodynamic integration (Figure 6 and Table 5). Thermodynamic integration is more affected by discretisation errors, an error source that is independent of the number of MCMC iterations, than by sampling errors (Figure 8). For this reason, the width of the confidence intervals obtained by thermodynamic integration does not reduce significantly with increasing numbers of MCMC iterations (Figure 6).

Both evidence estimators lead to the same ranking of the conceptual models with the hybrid conceptual model having the largest evidence and the outcrop-based conceptual model having the lowest one (Table 5). The multi-Gaussian and the analog-based conceptual models have very similar evidence estimates and they are the second-best performing conceptual models (Table 5).

For each conceptual model, the means of the log-likelihood functions, ℓ, increase with increasing β as we move from sampling the prior distribution ($\beta = 0$) to sampling the posterior distribution ($\beta = 1$) (Figure 7). From $\beta = 0$ to $\beta = 0.1$, the ℓ-estimates span three orders of magnitude. At very small values of β (i.e., $< 10^{-6}$), the outcrop-based conceptual model (green line in Figure 7) has mean log-likelihoods that are almost one order of magnitude higher than the other models. With increasing β, the outcrop-based model shows a much less steep increase of ℓ and at $\beta = 10^{-3}$, they start to be lower than the log-likelihood means of the other models. At higher power posteriors ($\beta > 0.1$), the ℓ-estimates for the hybrid conceptual model are the highest (red line in Figure 7), which explains why the highest evidence value is found for the hybrid conceptual model. We also note that the mean log-likelihood is not increasing continuously when β is close to one, which we attribute to random fluctuations of the MCMC chains (Figure 7).

The percentage ratio of independent MCMC samples after burn-in is never above 10% and it decreases to values as low as 0.01% for $\beta = 1$ (Figure 8). This is a consequence of the slow mixing of the MCMC chains and it explains the increase of the sampling errors with increasing β for both thermodynamic integration (Figure 8c) and stepping-stone sampling (Figure 8d). The sampling errors of the stepping-stone sampling method are always at least two orders of magnitude higher than the ones related to the thermodynamic method, but this method is devoid of discretisation errors, which constitutes the dominant error type for thermodynamic integration. As mentioned before, using a power law to distribute β-values (Eq. 16) ensures that the discretisation errors for small β are relatively small (i.e., between 10^{-10} and 10^{-3}, Figure 8b). The pronounced fluctuations of the discretisation errors especially for $\beta > 0.1$ (Figure 8b) are related to the fact that the mean of the log-likelihoods does not increase monotonically for high β-values.

We now compute the Bayes factors for the best conceptual model (hybrid) with respect to each of the other competing conceptual models. In particular, we follow the guideline proposed by *Kass and Raftery* [1995] and we present twice the natural logarithm of the Bayes factors (Figures 9a-b). The Bayes factors of the hybrid conceptual model are on the order of 10^{15} and 10^{16} relative to the second best models (multi-Gaussian and analog-based) and 10^{58} relative to the worst model (outcrop-based) for both thermodynamic integration and stepping-stone sampling. Note that the measure of twice the natural logarithms of the Bayes factors are all larger than 50 (Figures 9a-b). According to the interpretation of *Kass and Raftery* [1995], we can safely claim that the hybrid model shows very strong evidence of being superior to the other considered conceptual models. The Bayes factors computed with the stepping-stone sampling method have smaller uncertainties (Figure 9b) than the ones based on thermodynamic integration (Figure 9a). We note that the relative rankings of the competing models obtained with the thermodynamic integration and the stepping-stone sampling methods are consistent and stable as long as the MCMC chains has reached burn-in. Practically, this suggests that we can perform and obtain reliable Bayesian model selection results at less computational cost and, again, that formal convergence of the MCMC chains are not necessary.

5 Discussion

We have proposed a new methodology targeted at Bayesian model selection among geologically-realistic conceptual models that are represented by training images. For MCMC inversions, we use sequential geostatistical resampling through Graph Cuts that is two orders of magnitude faster than the forward simulation time (i.e., 0.08 versus 8.35 sec). In addition to being fast, the model realisations based on Graph Cuts are of high quality and honour the geological patterns in the training images. This is true for the five different types of conceptual models considered (Figures 3-4). Moreover, the Graph Cuts algorithm can include point conditioning [*Li et al.*, 2016] even if this is not considered herein. In our 2D analysis, we find that the hybrid conceptual model allows for the best fit of the observed breakthrough curves (Figure 5). The inclusion of highly conductive channels in a multi-Gaussian background enables enhanced simulations of the maximal concentrations and it is in general agreement with the expected subsurface structure at the MADE site (i.e., highly permeable network of sediments embedded in a less permeable matrix [*Harvey and Gorelick*, 2000; *Zheng and Gorelick*, 2003; *Liu et al.*, 2010; *Ronayne et al.*, 2010; *Bianchi et al.*, 2011a,b]). We find that the outcrop model has not enough degrees of freedom to properly fit the solute concentration data (Figure 5). Furthermore, we expect that an improved data fit would have been possible if we additionally would have inferred certain model parameter values (e.g., hydraulic conductivity for each facies and for the geostatistical parameters of the multi-Gaussian field).

In the light of the MADE-5 solute concentration data considered, the best fitting model (hybrid) is also the one that has the highest evidence, while the outcrop-based conceptual model has a Bayes factor of 10^{-58} with respect to the hybrid one, the lowest evidence and the lowest data fit (Table 4, Figure 6, Table 5). *Linde et al.* [2015b] rank different conceptual models (only the analog- and outcrop-based models are exactly the same as in the present work) of the region between the MLS-1 and MLS-2 wells using the maximum likelihood estimate based on geophysical data (cross-hole ground-penetrating radar data). In agreement with our results, *Linde et al.* [2015b] find that the analog-based conceptual model explains the data much better than the outcrop-based conceptual model and that the latter is the worst performing one in the considered set.

Our results suggest that when comparing complex conceptual models represented by training images in data-rich environments, it may sometimes be possible to simply rank the performance of the competing conceptual models based on the inferred standard deviation of the measurement errors, $\sigma_{\widetilde{\mathbf{Y}}}$ (Table 4), or the maximum likelihood estimate. Similar results for more traditional spatial priors were also found in other studies [*Schöniger et al.*, 2014; *Brunetti et al.*, 2017]. However, note that maximum likelihood-based model ranking will sometimes fail miserably as Bayesian model selection considers the trade-off between parsimony and goodness of fit. For instance, we expect that if we would have considered an uncorrelated hydraulic conductivity field, it would have produced the best fitting model but not the highest evidence. Moreover, it is also clear from these results that simply sampling the prior ($\beta = 0$) and then ranking the competing conceptual models based on the mean of the sampled likelihoods may provide misleading results. Indeed, the outcrop-based model has mean likelihoods of the prior model that are almost one order of magnitude higher than the ones of the other models (Figure 7) and, therefore, such a ranking would suggest that the outcrop-based conceptual model is the best one in describing the data while it is actually the worst one.

We find that stepping-stone sampling almost always provides slightly lower evidence estimates than thermodynamic integration (Table 5). This is in disagreement with the theory and with results by *Xie et al.* [2011] and *Friel et al.* [2014]. We attribute these unexpected results to the facts that (1) the MCMC chains for β close to 1 do not reach full convergence and the stepping-stone sampling is sensitive to poor convergence [*Friel et al.*, 2014] and (2) most of the contribution to the total evidence estimate comes from the terms of Eq. 7 computed for $\beta > 0.1$, a region where the mean log-likelihood does not increase monotonically due to random fluctuations of the MCMC chains (Figure 7). We also highlight that the comparison between the uncertainty estimates of the evidence values provided by thermodynamic integration and stepping-stone sampling (Figure 6) is not completely fair since the discretisation errors affecting thermodynamic integration are based on a worst-case scenario that arises from the approximation of Eq. 6 with a rectangular rule.

We stress again that our main intent is to present and demonstrate the proposed methodology targeted at Bayesian model selection among geologically-realistic conceptual models. Computational constrains made it infeasible to perform model selection in 3D. Instead, given the particular design of the tracer experiment (i.e., array of four aligned boreholes), we used a 2D flow and transport model and the data were corrected using a 3D-to-2D transformation that account for differences in flowpaths for a homogeneous subsurface (Appendix A). Since 3D heterogeneity is important at the MADE site, our 2D model ranking should only be considered approximate.

Future work should better account for model errors caused by the 3D-to-2D flow and transport approximation described in Appendix A. This would enhance the ability to make more definite statements about aquifer heterogeneity at the MADE site. How to properly account and represent model errors is a challenging task especially in problems involving many data, high-dimensional parameter spaces and non-linear forward models (e.g., *Linde et al.* [2017]). Another interesting topic that could be explored is to apply parallel tempering and use the resulting chains for computing the evidence with thermodynamic integration or stepping-stone sampling [*Vlugt and Smit*, 2001; *Bailer-Jones*, 2015; *Earl and Deem*, 2005]. Parallel tempering allows swapping between chains and, thereby, improving sampling efficiency. This may contribute to more robust results, faster convergence and, thereby, increase the number of effective samples (Figure 8a).

6 Conclusions

Inversions with geologically-realistic priors can be performed using training images and model proposals that honour their multiple-point statistics. Unfortunately, such inversions cannot rely on many state-of-the-art inversion methods and associated approaches for calculating the evidence needed when performing Bayesian model selection. In this work, we introduce a new full Bayesian methodology to enable Bayesian model selection among complex geological priors. To demonstrate this methodology, we have evaluated its performance in the context of determining, in a reduced set, the conceptual model that best explains the concentration data for the case study considered (MADE-5). Our methodology is applicable to both continuous and categorical conceptual models (e.g., a geologic facies image) and it could be used at other sites, scales and for different data types. Thermodynamic integration and stepping-stone sampling methods are used for evidence computation using a series of power posteriors obtained from MPS-based inversions. They provide a consistent ranking of the competing conceptual models regardless of the number of MCMC iterations after burn-in. This suggests that one can perform and obtain reliable Bayesian model selection results with MCMC chains that have only achieved limited sampling after burn-in. Both thermodynamic integration and stepping stone sampling are suitable evidence estimators. However, we recommend the stepping-stone sampling method because it is not affected by discretisation errors and its uncertainty (sampling errors) is significantly decreased with increasing numbers of MCMC iterations. This is not the case for the thermodynamic integration because it is affected by discretisation errors that dominate over the sampling errors. From the power posteriors derived from the test case, we find that (1) ranking the conceptual models based on prior sampling only ($\beta = 0$) favours the conceptual model with the lowest evidence and (2) model ranking based on the maximum posterior likelihood estimates ($\beta = 1$) provides, for this specific example, the same results as the formal Bayesian model selection methods considered herein. For improved sampling, we suggest that future work should investigate the use of parallel tempering results for evidence computations. Moreover, a full 3D analysis or a more formal treatment of model errors due to the considered 3D-to-2D approximation would enhance the confidence in statements about the suitability of alternative conceptual models at highly heterogeneous field sites.

A: Forward model: from 3D to 2D

The forward model used by *Bianchi et al.* [2011a] to simulate the bromide concentrations during the MADE-5 experiment is a 3D block-centred finite-difference model based on MODFLOW (3D flow simulator) [*Harbaugh*, 2005] and MT3DMS (3D transport simulator) [*Zheng*, 2010]. We initially consider a fine spatial discretisation of 0.1 m in the area around the wells (Figure A.1a-b). However, running such a 3D model is computationally prohibitive for evidence computations (i.e., 15 minutes of computing time to get one forward response and we need 10^5 forward evaluations for each MCMC chain and power posterior considered). To reduce the computing time, we perform a simple 3D to 2D correction of the data followed by 2D flow and transport simulations using the finite-volume algorithm MaFloT [*Künze and Lunati*, 2012]. Moreover, we restrict the simulations to the best fitting cross section (red segment in Figures A.1a-b) between the positions of the injection, extraction and the two MLS wells, which results in an area of 6.3 m × 8.1 m (Figure A.1c). For the transport equation, we set Dirichlet boundary conditions with the normalised concentration to the given fluxes on the left side of the model domain (Figure A.1c) corresponding to the injection well location. For the pressure equation, we set Dirichlet boundary conditions at the west and east sides (i.e., pressure difference), and Neumann boundary conditions at the north and south sides of the model domain (Figure A.1c).

Formal approaches to account for model errors in MCMC inversions exist (e.g., *Cui et al.* [2011]), but they are outside the scope of the present contribution. In the following, we introduce a simple error model that allows us to correct for the leading effects of the 3D to 2D transformation. These modelling errors stem primarily from the 2D linear approximation of the 3D radial distribution of the hydraulic heads, which results in a time shift in the breakthrough curves at the MLS wells. To estimate the correction factors, we consider a uniform hydraulic conductivity model with the geometric mean hydraulic conductivity at the MADE site (i.e., $4.3 \cdot 10^{-5}$ m/s [*Rehfeldt et al.*, 1992]). For this model, we perform 3D and 2D simulations of the MADE-5 experiment with MODFLOW/MT3DMS and MaFloT, respectively. As expected, the 3D simulated hydraulic heads between the injection and extraction wells does not change linearly as for the 2D simulation (Figure A.2). We tune the injection rate in the MODFLOW simulations to achieve simulated hydraulic heads that are as close as possible to the measured ones. We then perform MaFloT simulations using the MODFLOW simulated hydraulic heads at the injection and extraction wells as boundary conditions and we compute correction factors at the MLS wells. These multiplicative correction factors are those that maximise the correlation between the concentrations simulated with MT3DMS and MaFloT. The mean correction factors over the seven sampling ports in each of the two MLS wells are 1.09 and 1.92. Once the correction factors have been applied, the earlier time shifts (Figures A.2b-c) are removed (Figures A.2d-e). These correction factors are used in all subsequent simulations. Note that no attempt is made to correct for tracer movement due to 3D heterogeneity; the correction is a simple geometrical correction to account for the transformation of a uniform 3D to 2D flow field. We acknowledge that this is a crude approximation, but we deem it sufficient for the purposes of the present paper.

Acknowledgments

This work was supported by the Swiss National Science Foundation under grant number 200021_155924. Niklas Linde thanks Arnaud Doucet for initially suggesting the use of thermodynamic integration. Marco Bianchi publishes with the permission of the Executive Director of the British Geological Survey. The training images are available at https://doi.org/10.5281/zenodo.2545587 and the concentration data of the MADE-5 tracer experiment will be soon available at https://www.bgs.ac.uk/services/NGDC/.

References

Baele, G., and P. Lemey (2013), Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency, *Bioinformatics*, *29*(16), 1970–1979, doi:10.1093/bioinformatics/btt340.

Baele, G., P. Lemey, and S. Vansteelandt (2013), Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution, *BMC bioinformatics*, *14*(1), 85, doi:10.1186/1471-2105-14-85.

Bailer-Jones, C. A. (2015), A general method for Bayesian time series modelling, *Tech. rep.*, Max Planck Institute for Astronomy, Heidelberg.

Bayer, P., P. Huggenberger, P. Renard, and A. Comunian (2011), Three-dimensional high resolution fluvio-glacial aquifer analog: Part 1: Field study, *Journal of Hydrology*, *405*(1-2), 1–9, doi:10.1016/j.jhydrol.2011.03.038.

Bazargan, H., and M. Christie (2017), Bayesian model selection for complex geological structures using polynomial chaos proxy, *Computational Geosciences*, *21*(3), 533–551, doi:10.1007/s10596-017-9629-0.

Bianchi, M., and C. Zheng (2016), A lithofacies approach for modeling non-Fickian solute transport in a heterogeneous alluvial aquifer, *Water Resources Research*, *52*(1), 552–565, doi:10.1002/2015WR018186.

Bianchi, M., C. Zheng, G. R. Tick, and S. M. Gorelick (2011a), Investigation of small-scale preferential flow with a forced-gradient tracer test, *Groundwater*, *49*(4), 503–514, doi:10.1111/j.1745-6584.2010.00746.x.

Bianchi, M., C. Zheng, C. Wilson, G. R. Tick, G. Liu, and S. M. Gorelick (2011b), Spatial connectivity in a highly heterogeneous aquifer: From cores to preferential flow paths, *Water Resources Research*, *47*(5), doi:10.1029/2009WR008966.

Boggs, J. M., S. C. Young, L. M. Beard, L. W. Gelhar, K. R. Rehfeldt, and E. E. Adams (1992), Field study of dispersion in a heterogeneous aquifer: 1. Overview and site description, *Water Resources Research*, *28*(12), 3281–3291, doi:10.1029/92WR01756.

Bond, C. E., A. D. Gibbs, Z. K. Shipton, and S. Jones (2007), What do you think this is? "Conceptual uncertainty" in geoscience interpretation, *GSA today*, *17*(11), 4, doi:10.1130/GSAT01711A.1.

Boykov, Y., and V. Kolmogorov (2004), An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, *IEEE transactions on pattern analysis and machine intelligence*, *26*(9), 1124–1137, doi:10.1109/TPAMI.2004.60.

Brunetti, C., N. Linde, and J. A. Vrugt (2017), Bayesian model selection in hydrogeophysics: Application to conceptual subsurface models of the South Oyster Bacterial Transport Site, Virginia, USA, *Advances in Water Resources*, *102*, 127–141, doi:10.1016/j.advwatres.2017.02.006.

Caers, J., and T. Zhang (2004), Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models, in *Integration of Outcrop and Modern Analogs in Reservoir Modeling*, edited by G. M. Grammer, P. M. Harris, and G. P. Eberli, chap. 18, American Association of Petroleum Geologists, doi:10.1306/M80924C18.

Calderhead, B., and M. Girolami (2009), Estimating Bayes factors via thermodynamic integration and population MCMC, *Computational Statistics & Data Analysis*, *53*(12), 4028–4045, doi:10.1016/j.csda.2009.07.025.

Cao, T., X. Zeng, J. Wu, D. Wang, Y. Sun, X. Zhu, J. Lin, and Y. Long (2018), Integrating MT-DREAMzs and nested sampling algorithms to estimate marginal likelihood and comparison with several other methods, *Journal of Hydrology*, *563*, 750–765, doi:10.1016/j.jhydrol.2018.06.055.

Comunian, A., P. Renard, J. Straubhaar, and P. Bayer (2011), Three-dimensional high resolution fluvio-glacial aquifer analog: Part 2: Geostatistical modeling, *Journal of hydrology*, *405*(1-2), 10–23, doi:10.1016/j.jhydrol.2011.03.037.

Cui, T., C. Fox, and M. O'sullivan (2011), Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm, *Water Resources Research*, *47*(10), doi:10.1029/2010WR010352.

De Marsily, G., F. Delay, J. Gonçalvès, P. Renard, V. Teles, and S. Violette (2005), Dealing with spatial heterogeneity, *Hydrogeology Journal*, *13*(1), 161–183, doi: 10.1007/s10040-004-0432-3.

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

Earl, D. J., and M. W. Deem (2005), Parallel tempering: Theory, applications, and new perspectives, *Physical Chemistry Chemical Physics*, *7*(23), 3910–3916, doi: 10.1039/B509983H.

Elsheikh, A. H., V. Demyanov, R. Tavakoli, M. A. Christie, and M. F. Wheeler (2015), Calibration of channelized subsurface flow models using nested sampling and soft probabilities, *Advances in Water Resources*, *75*, 14–30, doi:10.1016/j.advwatres.2014.10.006.

Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis (2011), Choosing among partition models in Bayesian phylogenetics, *Molecular biology and evolution*, *28*(1), 523–532, doi:10.1093/molbev/msq224.

Feehley, C. E., C. Zheng, and F. J. Molz (2000), A dual-domain mass transfer approach for modeling solute transport in heterogeneous aquifers: Application to the Macrodispersion Experiment (MADE) site, *Water Resources Research*, *36*(9), 2501–2515, doi: 10.1029/2000WR900148.

Friel, N., and A. N. Pettitt (2008a), Marginal likelihood estimation via power posteriors, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(3), 589–607, doi:10.1111/j.1467-9868.2007.00650.x.

Friel, N., and A. N. Pettitt (2008b), Marginal likelihood estimation via power posteriors, *Journal of the Royal Statistical Society. Series B*, *70*(3), 589–607, doi:10.1111/j.1467-9868.2007.00650.x.

Friel, N., M. Hurn, and J. Wyse (2014), Improving power posterior estimation of statistical evidence, *Statistics and Computing*, *24*(5), 709–723, doi:10.1007/s11222-013-9397-1.

Gelman, A., and X.-L. Meng (1998), Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, *Statistical Science*, pp. 163–185, doi: 10.1214/ss/1028905934.

Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Statistical Science*, *7*(4), 457–472, doi:10.1214/ss/1177011136.

Gelman, A., G. O. Roberts, and W. R. Gilks (1996), Efficient Metropolis jumping rules, *Bayesian statistics*, *5*, 599–608.

Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments, *Bayesian statistics*, *4*, 641–649.

Gómez-Hernández, J. J., and X.-H. Wen (1998), To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology, *Advances in Water Resources*, *21*(1), 47–61, doi: 10.1016/S0309-1708(96)00031-0.

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

Grzegorczyk, M., A. Aderhold, and D. Husmeier (2017), Targeting Bayes factors with direct-path non-equilibrium thermodynamic integration, *Computational Statistics*, *32*(2), 717–761, doi:10.1007/s00180-017-0721-7.

Guardiano, F. B., and R. M. Srivastava (1993), Multivariate geostatistics: beyond bivariate moments, in *Geostatistics Tróia '92*, edited by A. Soares, pp. 133–144, Springer, doi: 10.1007/978-94-011-1739-5_12.

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

Hammersley, J. M., and D. C. Handscomb (1964), *Monte Carlo methods*, vol. 1, VIII, 178 pp., Springer Netherlands, doi:10.1007/978-94-009-5819-7.

Hansen, T. M., K. S. Cordua, and K. Mosegaard (2012), Inverse problems with non-trivial priors: Efficient solution through sequential Gibbs sampling, *Computational Geosciences*, *16*(3), 593–611, doi:10.1007/s10596-011-9271-1.

Harbaugh, A. W. (2005), *MODFLOW-2005, The US Geological Survey modular ground-water model: the ground-water flow process*, US Department of the Interior, US Geological Survey Reston.

Harvey, C., and S. M. Gorelick (2000), Rate-limited mass transfer or macrodispersion: Which dominates plume evolution at the Macrodispersion Experiment (MADE) site?, *Water Resources Research*, *36*(3), 637–650, doi:10.1029/1999WR900247.

Höhna, S., M. L. Landis, and J. P. Huelsenbeck (2017), Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics, *bioRxiv*, doi: doi.org/10.1101/104422.

Hu, L., and T. Chugunova (2008), Multiple-point geostatistics for modeling subsurface heterogeneity: A comprehensive review, *Water Resources Research*, *44*(11), doi: 10.1029/2008WR006993.

Jäggli, C., J. Straubhaar, and P. Renard (2017), Posterior population expansion for solving inverse problems, *Water Resources Research*, *53*(4), 2902–2916, doi: 10.1002/2016WR019550.

Jefferys, W. H., and J. Berger (1992), Ockham's Razor and Bayesian Analysis, *American Scientist*, *80*(1), 64–72.

Jeffreys, H. (1935), Some Tests of Significance, Treated by the Theory of Probability, *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(2), 203–222, doi: 10.1017/S030500410001330X.

Jeffreys, H. (1939), *Theory of Probability*, third ed., Oxford University Press.

Journel, A., and T. Zhang (2006), The necessity of a multiple-point prior model, *Mathematical Geology*, *38*(5), 591–610, doi:10.1007/s11004-006-9031-2.

Kass, R. E., and A. E. Raftery (1995), Bayes factors, *Journal of the American Statistical Association*, *90*(430), 773–795, doi:10.1080/01621459.1995.10476572.

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998), Markov chain Monte Carlo in practice: a roundtable discussion, *The American Statistician*, *52*(2), 93–100, doi: 10.1080/00031305.1998.10480547.

Kerrou, J., P. Renard, H.-J. H. Franssen, and I. Lunati (2008), Issues in characterizing heterogeneity and connectivity in non-multigaussian media, *Advances in Water Resources*, *31*(1), 147–159, doi:10.1016/j.advwatres.2007.07.002.

Koltermann, C. E., and S. M. Gorelick (1996), Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches, *Water Resources Research*, *32*(9), 2617–2658, doi:doi.org/10.1029/96WR00025.

Künze, R., and I. Lunati (2012), An adaptive multiscale method for density-driven instabilities, *Journal of Computational Physics*, *231*(17), 5557–5570, doi: 10.1016/j.jcp.2012.02.025.

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

Laloy, E., N. Linde, D. Jacques, and G. Mariethoz (2016), Merging parallel tempering with sequential geostatistical resampling for improved posterior exploration of high-dimensional subsurface categorical fields, *Advances in Water Resources*, *90*, 57–69, doi: 10.1016/j.advwatres.2016.02.008.

Laloy, E., R. Hérault, D. Jacques, and N. Linde (2018), Training-image based geostatistical inversion using a spatial generative adversarial neural network, *Water Resources Research*, *54*(1), 381–406, doi:10.1002/2017WR022148.

Lark, R., S. Thorpe, H. Kessler, and S. Mathers (2014), Interpretative modelling of a geological cross section from boreholes: sources of uncertainty and their quantification, *Solid Earth*, *5*(2), 1189–1203, doi:10.5194/se-5-1189-2014.

Lartillot, N., and H. Philippe (2006), Computing Bayes factors using thermodynamic integration, *Systematic Biology*, *55*(2), 195–207, doi:10.1080/10635150500433722.

Lewis, S. M., and A. E. Raftery (1997), Estimating Bayes factors via posterior simulation with the Laplace-??Metropolis estimator, *J Am Stat Assoc*, *92*(438), 648–655, doi: 10.1080/01621459.1997.10474016.

Li, X., G. Mariethoz, D. Lu, and N. Linde (2016), Patch-based iterative conditional geostatistical simulation using graph cuts, *Water Resources Research*, *52*(8), 6297–6320, doi:10.1002/2015WR018378.

Linde, N. (2014), Falsification and corroboration of conceptual hydrological models using geophysical data, *Wiley Interdisciplinary Reviews: Water*, *1*(2), 151–171, doi: 10.1002/wat2.1011.

Linde, N., P. Renard, T. Mukerji, and J. Caers (2015a), Geological realism in hydrogeological and geophysical inverse modeling: A review, *Advances of Water Resources*, *86*, 86–101, doi:10.1016/j.advwatres.2015.09.019.

Linde, N., T. Lochbühler, M. Dogan, and R. L. Van Dam (2015b), Tomogram-based comparison of geostatistical models: Application to the Macrodispersion Experiment (MADE) site, *Journal of Hydrology*, *531*, 543–556, doi:10.1016/j.jhydrol.2015.10.073.

Linde, N., D. Ginsbourger, J. Irving, F. Nobile, and A. Doucet (2017), On Uncertainty Quantification in Hydrogeology and Hydrogeophysics, *Advances in Water Resources*, *110*, 166–181, doi:10.1016/j.advwatres.2017.10.014.

Liu, G., C. Zheng, G. R. Tick, J. J. Butler, and S. M. Gorelick (2010), Relative importance of dispersion and rate-limited mass transfer in highly heterogeneous porous media: Analysis of a new tracer test at the Macrodispersion Experiment (MADE) site, *Water Resources Research*, *46*(3), doi:10.1029/2009WR008430.

Liu, P., A. S. Elshall, M. Ye, P. Beerli, X. Zeng, D. Lu, and Y. Tao (2016), Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods, *Water Resources Ressearch*, *52*(2), 734–758, doi: 10.1002/2014WR016718.

Lochbühler, T., G. Pirot, J. Straubhaar, and N. Linde (2014), Conditioning of multiple-point statistics facies simulations to tomographic images, *Mathematical Geosciences*, *46*(5), 625–645, doi:10.1007/s11004-013-9484-z.

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

MacKay, D. J. (1992), Bayesian interpolation, *Neural Computation*, *4*(3), 415–447, doi: 10.1162/neco.1992.4.3.415.

Maliva, R. G. (2016), *Aquifer Characterization Techniques*, Springer, doi:10.1007/978-3-319-32137-0.

Mariethoz, G., and J. Caers (2014), *Multiple-point Geostatistics: Stochastic Modeling with Training Images*, John Wiley & Sons, doi:10.1002/9781118662953.

Mariethoz, G., P. Renard, and J. Caers (2010a), Bayesian inverse problem and optimization with iterative spatial resampling, *Water Resources Research*, *46*(11), doi: doi.org/10.1029/2010WR009274.

Mariethoz, G., P. Renard, and J. Straubhaar (2010b), The direct sampling method to perform multiple-point geostatistical simulations, *Water Resources Research*, *46*(11), doi: 10.1029/2008WR007621.

Mosegaard, K., and A. Tarantola (1995), Monte Carlo sampling of solutions to inverse problems, *Journal of Geophysical Research: Solid Earth*, *100*(B7), 12,431–12,447, doi: 10.1029/94JB03097.

Oates, C. J., T. Papamarkou, and M. Girolami (2016), The controlled thermodynamic integral for Bayesian model evidence evaluation, *Journal of the American Statistical Association*, *111*(514), 634–645, doi:10.1080/01621459.2015.1021006.

Pirot, G., P. Renard, E. Huber, J. Straubhaar, and P. Huggenberger (2015), Influence of conceptual model uncertainty on contaminant transport forecasting in braided river aquifers, *Journal of Hydrology*, *531*, 124–141, doi:10.1016/j.jhydrol.2015.07.036.

Pirot, G., N. Linde, G. Mariethoz, and J. H. Bradford (2017a), Probabilistic inversion with graph cuts: Application to the Boise Hydrogeophysical Research Site, *Water Resources*

Research, *53*(2), 1231–1250, doi:doi.org/10.1002/2016WR019347.

Pirot, G., M. Cardiff, G. Mariethoz, J. Bradford, and N. Linde (2017b), Towards 3D Probabilistic Inversion with Graphcuts, in *23rd European Meeting of Environmental and Engineering Geophysics*.

Randle, C. H., C. E. Bond, R. M. Lark, and A. A. Monaghan (2018), Can uncertainty in geological cross-section interpretations be quantified and predicted?, *Geosphere*, doi: 10.1130/GES01510.1.

Refsgaard, J. C., and H. J. Henriksen (2004), Modelling guidelines—-terminology and guiding principles, *Advances in Water Resources*, *27*(1), 71–82, doi: 10.1016/j.advwatres.2003.08.006.

Refsgaard, J. C., S. Christensen, T. O. Sonnenborg, D. Seifert, A. L. Højberg, and L. Troldborg (2012), Review of strategies for handling geological uncertainty in groundwater flow and transport modeling, *Advances in Water Resources*, *36*, 36–50, doi: 10.1016/j.advwatres.2011.04.006.

Rehfeldt, K. R., J. M. Boggs, and L. W. Gelhar (1992), Field study of dispersion in a heterogeneous aquifer: 3. Geostatistical analysis of hydraulic conductivity, *Water Resources Research*, *28*(12), 3309–3324, doi:10.1029/92WR01758.

Remy, N., A. Boucher, and J. Wu (2009), *Applied geostatistics with SGeMS: a user's guide*, Cambridge University Press.

Renard, P., and D. Allard (2013), Connectivity metrics for subsurface flow and transport, *Advances in Water Resources*, *51*, 168–196, doi:10.1016/j.advwatres.2011.12.001.

Renard, P., H. Demougeot-Renard, and R. Froidevaux (2005), *Geostatistics for environmental applications*, Springer, doi:10.1007/s11004-018-9733-2.

Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resources Research*, *44*(12), doi:10.1029/2008WR006908.

Ronayne, M. J., S. M. Gorelick, and C. Zheng (2010), Geological modeling of submeter scale heterogeneity and its influence on tracer transport in a fluvial aquifer, *Water Resources Research*, *46*(10), doi:10.1029/2010WR009348.

Scheidt, C., L. Li, and J. Caers (2018), *Quantifying Uncertainty in Subsurface Systems*, vol. 236, John Wiley & Sons, doi:10.1002/9781119325888.

Schöniger, A., T. Wöhling, L. Samaniego, and W. Nowak (2014), Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, *Water*

Resources Research, *50*(12), 9484–9513, doi:10.1002/2014WR016062.

Skilling, J. (2004), Nested sampling, in *AIP Conference Proceedings*, vol. 735, pp. 395–405, AIP, doi:10.1063/1.1835238.

Skilling, J. (2006), Nested sampling for general Bayesian computation, *Bayesian analysis*, *1*(4), 833–859, doi:10.1214/06-BA127.

Strebelle, S. (2002), Conditional simulation of complex geological structures using multiple-point statistics, *Mathematical Geology*, *34*(1), 1–21, doi: 10.1023/A:1014009426274.

Vlugt, T. J., and B. Smit (2001), On the efficient sampling of pathways in the transition path ensemble, *PhysChemComm*, *4*(2), 11–17, doi:10.1039/B009865P.

Volpi, E., G. Schoups, G. Firmani, and J. Vrugt (2017), Sworn testimony of the model evidence: Gaussian Mixture Importance (GAME) sampling, *Water Resources Research*, doi:10.1002/2016WR020167.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen (2011), Improving marginal likelihood estimation for Bayesian phylogenetic model selection, *Systematic Biology*, *60*(2), 150–160, doi:10.1093/sysbio/syq085.

Zahner, T., T. Lochbühler, G. Mariethoz, and N. Linde (2016), Image synthesis with graph cuts: a fast model proposal mechanism in probabilistic inversion, *Geophysical Journal International*, *204*(2), 1179–1190, doi:10.1093/gji/ggv517.

Zeng, X., M. Ye, J. Wu, D. Wang, and X. Zhu (2018), Improved nested sampling and surrogate-enabled comparison with other marginal likelihood estimators, *Water Resources Research*, *54*(2), 797–826, doi:10.1002/2017WR020782.

Zheng, C. (2010), MT3DMS v5. 3 supplemental user's guide, *Department of Geological Sciences, University of Alabama, Tuscaloosa, Alabama*.

Zheng, C., and S. M. Gorelick (2003), Analysis of solute transport in flow fields influenced by preferential flowpaths at the decimeter scale, *Groundwater*, *41*(2), 142–155, doi:10.1111/j.1745-6584.2003.tb02578.x.

Zheng, C., M. Bianchi, and S. M. Gorelick (2011), Lessons learned from 25 years of research at the MADE site, *Groundwater*, *49*(5), 649–662, doi:10.1111/j.1745-6584.2010.00753.x.

Zinn, B., and C. F. Harvey (2003), When good statistical models of aquifer heterogeneity go bad: A comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields, *Water Resources Research*, *39*(3), doi:

10.1029/2001WR001146.

Algorithm 1: MCMC inversion workflow based on MPS and the extended Metropolis algorithm to enable evidence estimation using power posteriors.

Input: T, maximum number of MCMC iterations; J, number of power coefficients β distributed according to Eq. 16; a training image

Output: Λ_j, matrices containing power posteriors and log-likelihoods; $\log p(\widetilde{\mathbf{Y}}|\eta)$, evidence

Set $t = 1$;

Draw $\theta_{\mathbf{1}}$ from the training image;

Solve the forward problem;

Compute likelihood (e.g., Eq. 2);

for $j = 1,..., J$ **do**

    for $t = 2,..., T$ **do**

        Set $\theta_{\mathbf{cur}} = \theta_{\boldsymbol{t-1}}$;

        Draw $\theta_{\mathbf{prop}}$ based on MPS (e.g., using Graph Cuts proposals);

        Solve the forward problem;

        Compute likelihood (e.g., Eq. 2);

        Accept $\theta_{\mathbf{prop}}$ with probability, $\alpha = min\left\{1, \frac{p(\widetilde{\mathbf{Y}}|\theta_{\mathbf{prop}})^{\beta_j}}{p(\widetilde{\mathbf{Y}}|\theta_{\mathbf{cur}})^{\beta_j}}\right\}$;

        if $\theta_{\mathbf{prop}}$ *accepted* **then**

            Set $\theta_{\boldsymbol{t}}$=$\theta_{\mathbf{prop}}$;

        else

            Set $\theta_{\boldsymbol{t}}$=$\theta_{\mathbf{cur}}$;

        end

        Store $\theta_{\mathbf{t}}$ and the corresponding log-likelihood in matrix Λ_j;

        Set t=t+1;

    end

end

Compute $\log p(\widetilde{\mathbf{Y}}|\eta)$ (Eqs. 7 and 14) and corresponding variances (Eqs. 9-10 and 15) using the information stored in Λ_j after the removal of the burn-in period.

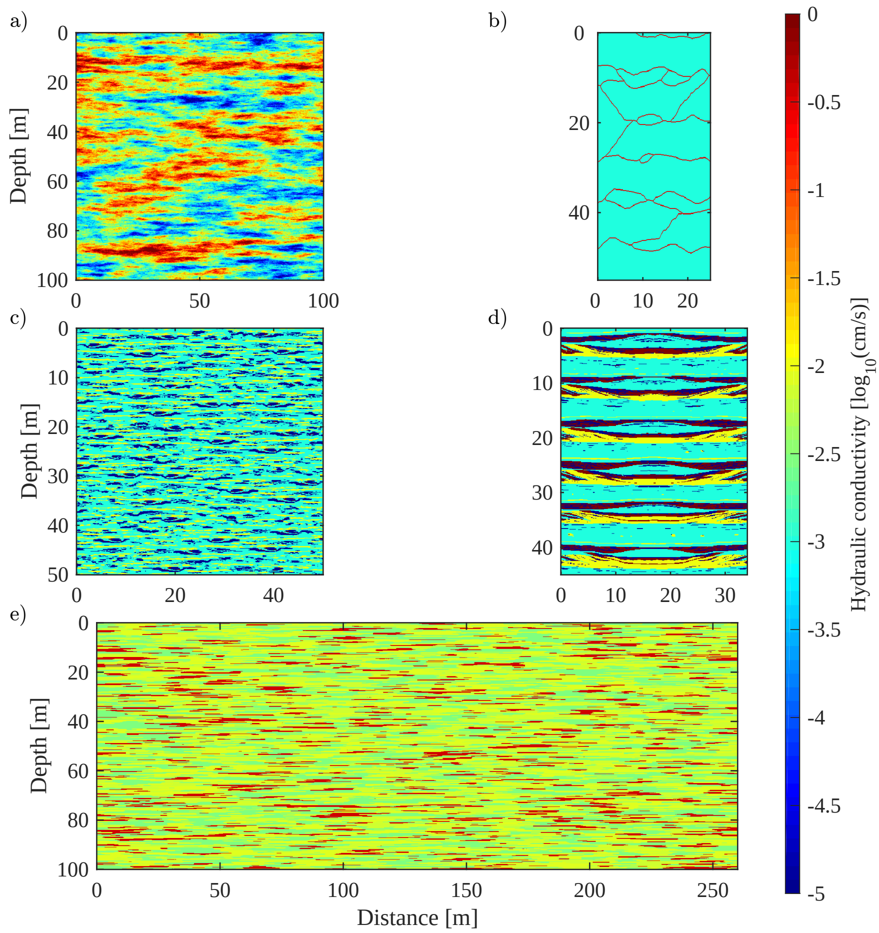


Figure 2. Training images used in the MPS-based inversion to represent spatial hydraulic conductivity of the MADE site: (a) multi-Gaussian field [*Bianchi et al.*, 2011a], (b) highly conductive channels in an homogeneous matrix [*Strebelle*, 2002; *Ronayne et al.*, 2010; *Linde et al.*, 2015b], (c) model based on a mapping study of a MADE outcrop [*Rehfeldt et al.*, 1992; *Linde et al.*, 2015b], (d) model based on a mapping study at the Herten site in Germany [*Bayer et al.*, 2011; *Comunian et al.*, 2011; *Linde et al.*, 2015b] featuring representative alluvial deposit structures and (e) model based on lithological borehole data collected at the MADE site [*Bianchi and Zheng*, 2016].

Table 1. Geostatistical parameters of the multi-Gaussian training image (Figure 2a) proposed by *Bianchi et al.* [2011a] for the MADE site. The actual variogram model was a linear combination of a spherical and an exponential model.

	Variogram model	
Variogram parameters	Spherical	Exponential
Maximum range [m]	76	21
Minimum range [m]	4.6	5
Nugget	0.2	-
Sill	1.75	3.0

Table 2. Hydrogeological facies and their hydraulic conductivity values [*Rehfeldt et al.*, 1992] observed at the MADE site outcrop and used for the training images in Figure 2c-d.

Facies	$\log_{10}$ K [cm/s]
Open framework gravel	$-6.83 \cdot 10^{-4}$
Sand	-2.00
Undifferentiated sandy gravel	-3.00
Sandy, clayey gravel	-5.00

Table 3. Hydrogeological facies and their hydraulic conductivity values based on lithological data from the MADE site [*Bianchi and Zheng*, 2016] and used for the training image in Figure 2e.

Facies	$\log_{10}$ K [cm/s]
Highly conductive gravel	-0.45
Sand and gravel	-2.05
Gravel with sand	-2.11
Well-sorted sand	-2.18
Sand gravel and fines	-2.53

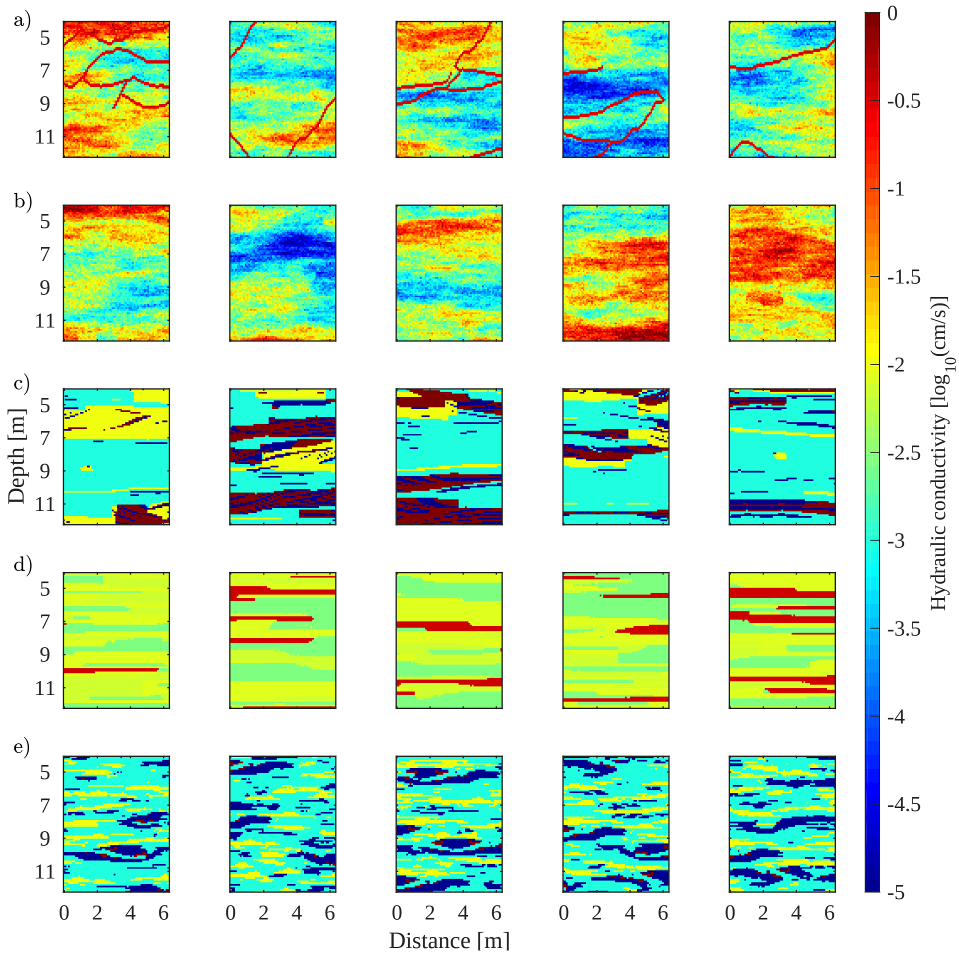


Figure 3. Five prior realisations of hydraulic conductivity fields generated from the training images of Figure 2 with the Graph Cuts algorithm for the (a) hybrid, (b) multi-Gaussian, (c) analog-based, (d) lithofacies-based and (e) outcrop-based conceptual model of the MADE site.

Table 4. Summary of MCMC results using the MADE-5 tracer data for three MCMC chains of 10^5 steps for each conceptual model with $\beta = 1$. First column, conceptual model considered; second column, average acceptance rate (AR); third to fifth column, burn-in percentage based on the Geweke method for each of the three chains (when no value is displayed, the chain failed to reach burn-in); last two columns, means and standard deviations of the standard deviation of the measurement errors inferred with MCMC.

		Burn-in [%]			$\sigma_{\widetilde{\mathbf{Y}}}$ [mg/L]	
Conceptual model	AR [%]	Chain 1	Chain 2	Chain 3	Mean	Std
Hybrid	0.6	-	58	87	5.81	0.27
Multi-Gaussian	8.0	48	45	62	7.14	0.33
Analog	4.1	-	64	84	7.22	0.34
Lithofacies	1.2	55	38	74	8.92	0.60
Outcrop	5.5	76	97	-	9.36	0.35

Table 5. Estimates of the natural logarithm of the evidence, $\log p(\widetilde{\mathbf{Y}}|\eta)$, with corresponding standard errors, SE, for each conceptual model (first column) based on the stepping-stone sampling method (second and third column) and on the thermodynamic integration method with the corrected trapezoidal rule (last two columns).

	Stepping-stone sampling		Thermodynamic integration	
Conceptual model	$\log p(\widetilde{\mathbf{Y}}|\eta)$ [-]	SE [-]	$\log p(\widetilde{\mathbf{Y}}|\eta)$ [-]	SE [-]
Hybrid	-903.99	1.17	-902.68	4.02
Multi-Gaussian	-939.43	0.64	-939.15	0.93
Analog	-941.48	0.87	-941.40	1.30
Lithofacies	-1009.01	1.18	-1008.76	3.90
Outcrop	-1037.58	1.11	-1036.45	1.47

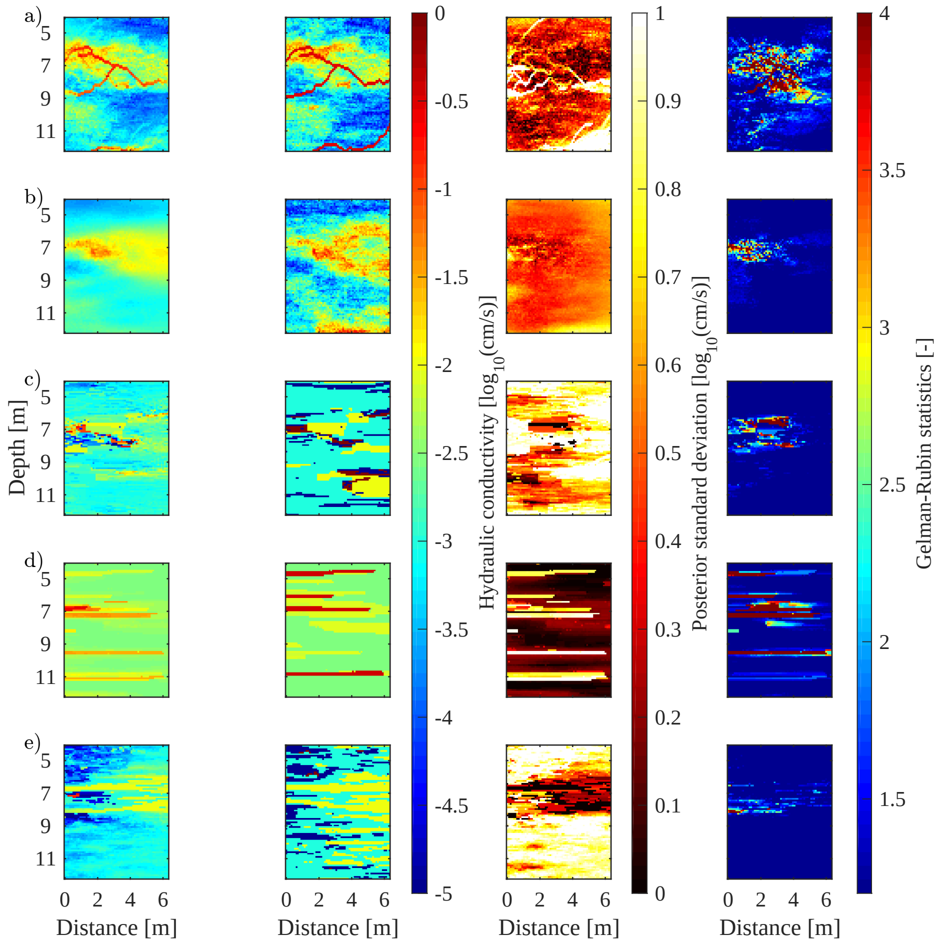


Figure 4. Mean (first column), maximum a-posteriori (second column), and standard deviation (third column) of the posterior hydraulic conductivity realisations for the (a) hybrid, (b) multi-Gaussian, (c) analog-based, (d) lithofacies-based and (e) outcrop-based conceptual model at the MADE site. In the last column, the Gelman-Rubin statistic for each pixel is reported. Dark-blue regions represent values equal or less than 1.2 and indicate that convergence has been reached for those pixels.

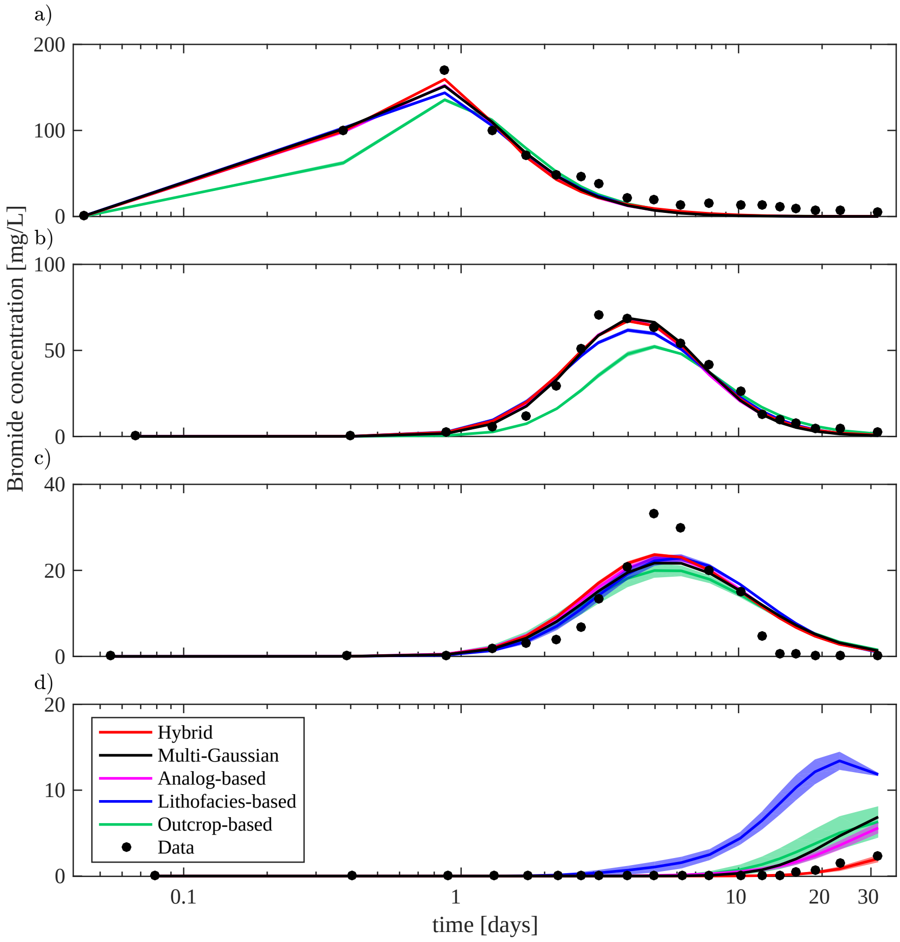


Figure 5. Simulated (solid lines) and measured (black dots) bromide breakthrough curves from the MADE-5 experiment in the two monitoring wells MLS-1 and MLS-2 at a depth of 7 m (a-b) and 11 m (c-d), respectively. The simulated breakthrough curves are summarised by the mean of the posterior realisations (solid lines) and their 95% confidence intervals (shaded areas).

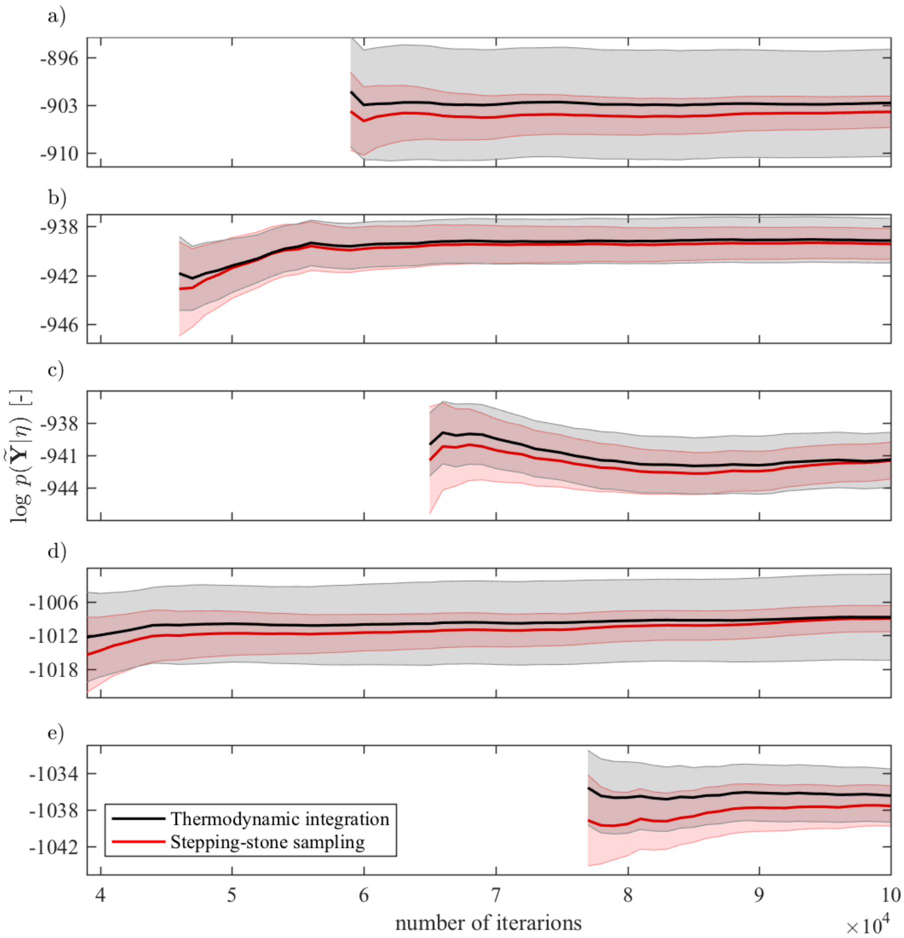
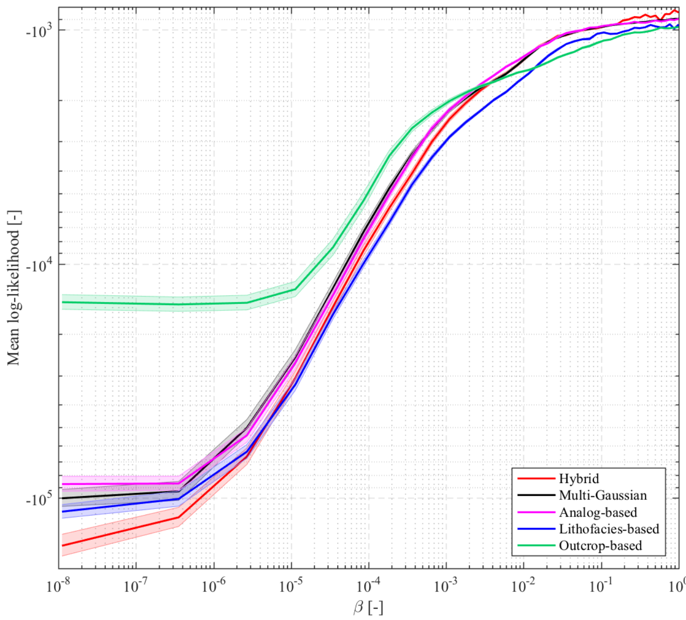
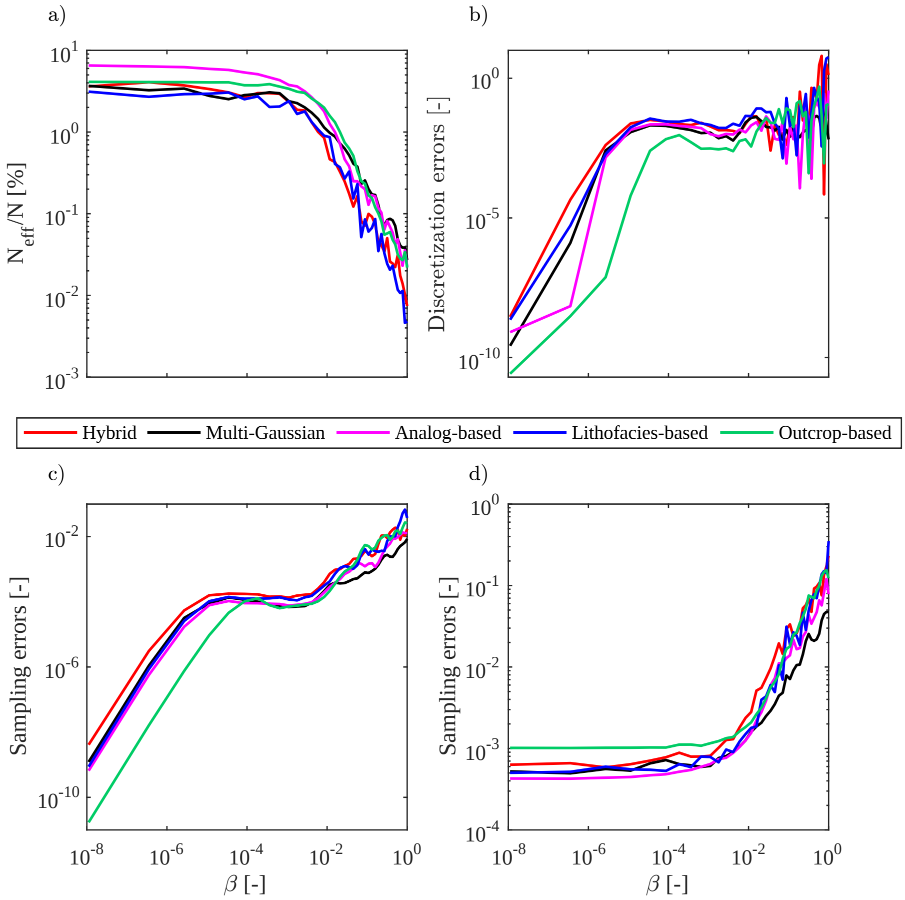


Figure 6. Natural logarithm of the evidence estimates, $\log p(\widetilde{\mathbf{Y}}|\eta)$, as a function of the number of MCMC iterations. Evidences are computed with the stepping-stone sampling method (red line) and the thermodynamic integration method based on the corrected trapezoidal rule (black line) for the (a) hybrid, (b) multi-Gaussian, (c) analog-based, (d) lithofacies-based and (e) outcrop-based model at the MADE site. The evidence computation starts after a different number of MCMC iterations because each of the conceptual models has a specific burn-in period. The shaded areas represent the 95% confidence interval of the evidence estimates (pink for stepping-stone sampling and grey for thermodynamic integration).



570 **Figure 7.** Mean (line) of the natural logarithm of the likelihood functions, ℓ, computed for each β value and
571 the 95% confidence interval of the ℓ-estimates (shaded areas). Note that the *x*- and *y*-axes are in $\log_{10}$ scale.



584 **Figure 8.** (a) Percentage ratio between the effective and the total number of MCMC samples, (b) discretisation errors in the thermodynamic integration method (square root of Eq. 10), (c) sampling errors in the thermodynamic integration method (square root of Eq. 9) and (d) sampling errors in the stepping-stone sampling method (square root of Eq. 15) as a function of β-values. Note that all the x- and y-axes are in $\log_{10}$ scale.

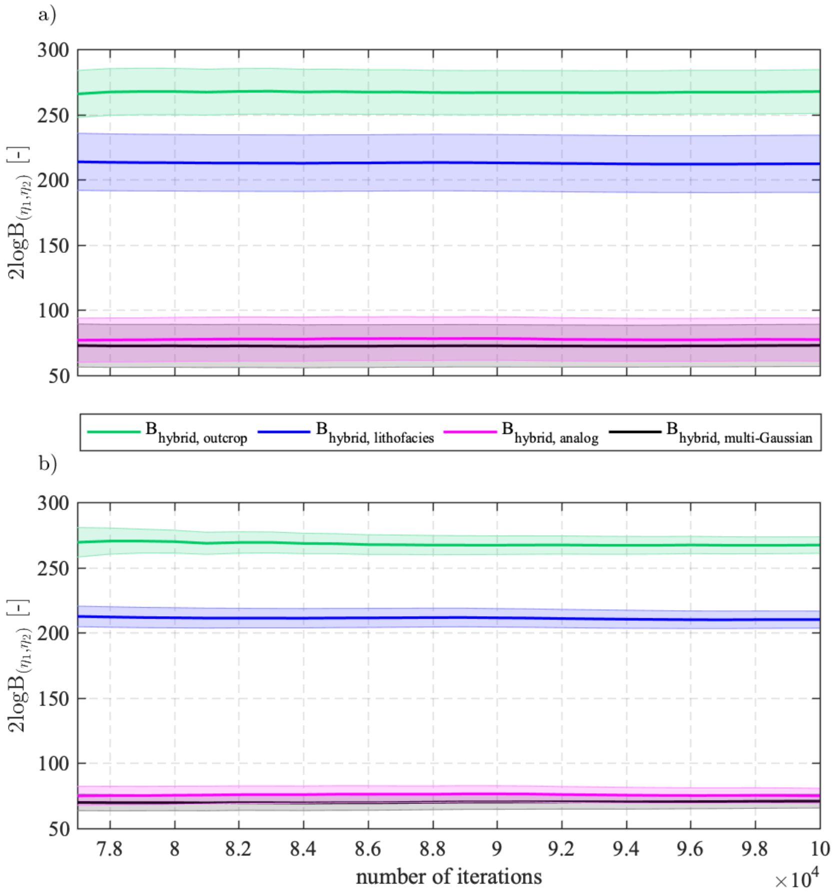
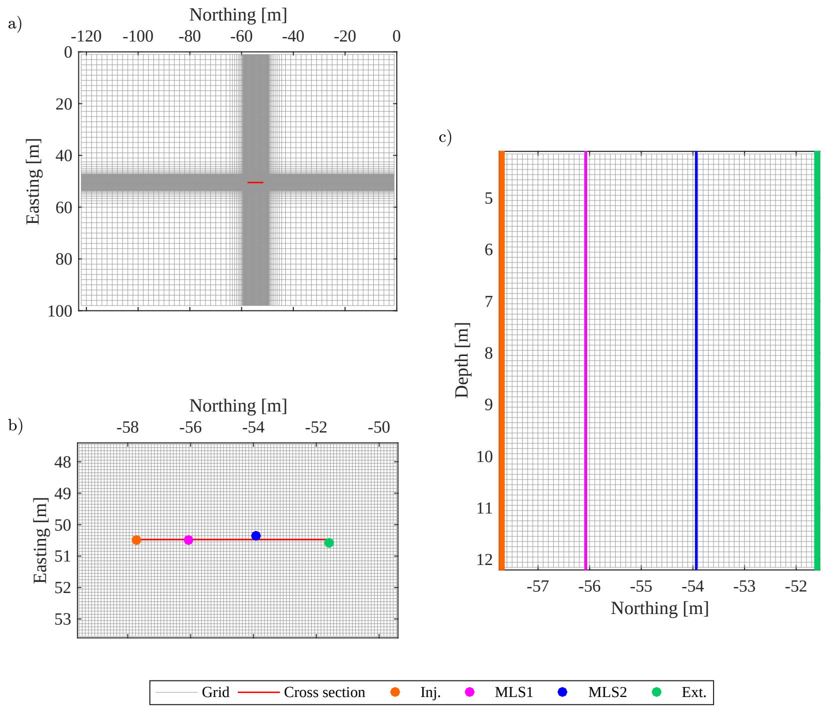


Figure 9. Twice the natural logarithm of the Bayes factors of the "best model" (hybrid) with respect to the outcrop-based (green line), lithofacies-based (blue line), analog-based (magenta line) and multi-Gaussian (black line) conceptual model at the MADE site. Results are shown for (a) the thermodynamic integration method based on the corrected trapezoidal rule and for the (b) stepping-stone sampling method. The shaded areas represent the 95% confidence interval of the $2\log B_{\eta_1,\eta_2}$ measures.



739 **Figure A.1.** (a) Aerial view of the 3D grid used for simulations with MODFLOW/MT3DMS; (b) zoom
740 in the tracer test area, in which the grid size was refined to 0.1 m; (c) cross section used for simulations with
741 MaFloT. The width of the lines in (c) is representative of the diameter of the four wells.

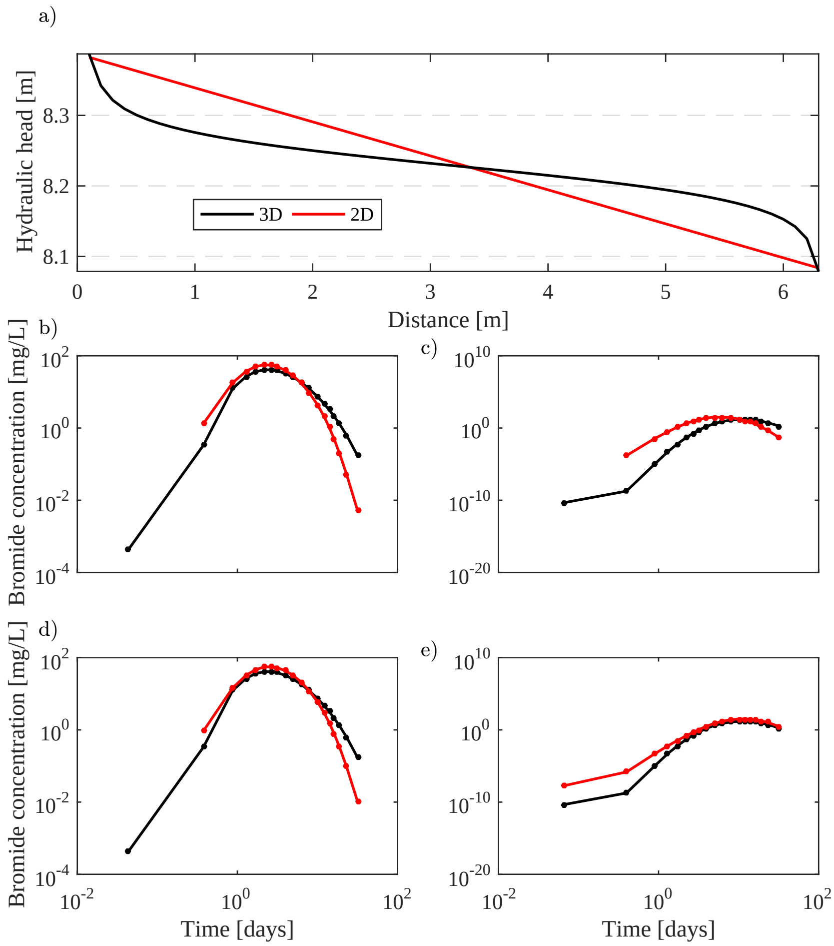


Figure A.2. (a) Hydraulic head profiles between the injection and extraction wells arising from 2D and 3D flow simulations in a uniform hydraulic conductivity field. Simulated breakthrough curves at 7 m depth in (b) MLS-1 and (c) MLS-2 without corrections. The shifts in the 2D simulations are removed when (d-e) applying the correction factors.