Year : 2018

Bayesian model selection in hydrogeophysics and hydrogeology

Brunetti Carlotta

UNIL | Université de Lausanne

Faculté des géosciences et de l'environnement
Institut des sciences de la Terre

Bayesian model selection in hydrogeophysics and hydrogeology

Thèse de doctorat

Présentée à la
Faculté des géosciences et de l'environnement
Institut des sciences de la Terre
de l'Université de Lausanne
par

Carlotta Brunetti

Diplôme (M.Sc.) en Physique du Système Terre
Université de Bologna

Jury
Prof. Dr. Niklas Linde, directeur de thèse
Prof. Dr. James Irving, expert interne
Prof. Dr. Wolfgang Nowak, expert externe
Prof. Dr. Christian Kull, président du jury

Lausanne, 2018

UNIL | Université de Lausanne
Faculté des géosciences et de l'environnement
bâtiment Géopolis bureau 4631

IMPRIMATUR

Vu le rapport présenté par le jury d'examen, composé de

Président de la séance publique :	M. le Professeur Christian Kull
Président du colloque :	M. le Professeur Christian Kull
Directeur de la thèse :	M. le Professeur Niklas Linde
Expert interne :	M. le Docteur James Irving
Expert externe :	M. le Professeur Wolfgang Nowak

Le Doyen de la Faculté des géosciences et de l'environnement autorise l'impression de la thèse de

Madame Carlotta BRUNETTI

Titulaire d'un
Master in Physics of Earth System
de l'Université de Bologne

intitulée

Bayesian model selection in hydrogeophysics and hydrogeology

Lausanne, le 22 février 2019

Pour le Doyen de la Faculté des géosciences et de l'environnement

Professeur Christian Kull

Acknowledgements

It is an honour to thank all the people who have contributed to the successful completion of my PhD thesis and who have supported me all along the four years of this unique experience. My special thanks go to:

- my supervisor **Niklas Linde** for teaching me how excellent and ethical scientific research is done, for always being available, trustworthy and rigorous, for always helping me whenever there was a need, for his sharp mind and vast knowledge and for making my PhD experience so productive and stimulating. I could not have imagined having a better supervisor and mentor for my doctoral studies
- the committee members **Wolfgang Nowak** and **James Irving** for the fruitful and inspiring discussion during my private defense as well as the president of the jury **Christian Kull** for his availability and kindness
- the **Swiss National Science Foundation** [grant number 200021_155924] for funding this work
- the scientists involved in peer-reviewing the individual chapters of this thesis such as the Editor **Paolo D'Odorico**, the Associated Editor **Harrie-Jan Hendricks-Franssen**, **Anneli Guthke**, **Jan Dettmer** and all the anonymous reviewers for their very constructive and insightful comments
- my coauthors **Marco Bianchi**, **Jasper A. Vrugt** and **Guillaume Pirot** for their enriching collaboration
- **Eric Laloy** for making available the code used for building the multi-Gaussian models, **John Peterson** and **Susan Hubbard** for providing the crosshole GPR data from the South Oyster Bacterial Transport site and **Jasper A. Vrugt** for making DREAM$_{\text{(ZS)}}$ available and for providing an initial version of the code used for Gaussian mixture importance sampling
- the system administrator **Philippe Logean** and my colleague **Jürg Hunziker** for helping me with the clusters of the university
- my faithful PhD partner **Giulia de Pasquale** for sharing this four years adventure and the field camps under (almost always) stormy and rainy weather with me
- my colleagues **Corinna Koepke**, **Emily B. Voytek** and **Laureline Josset** for sharing the office and the good and hard times of the PhD with me
- my **parents** and my **grandmother** for their unconditional love and endless support. In particular, to my **mother** for helping me with her pragmatism and strength that only a woman can have and to my **father** for helping me with his thoughtful mind and sharp wisdom that only a real yogi can achieve
- my **boyfriend** for his contagious optimism and for bringing me always in amazing places in the Swiss Alps during the weekends where I could refresh my mind all along my PhD
- the **lac Léman** for the energizing walks and runs along its shores

Contents

List of Figures

List of Tables

Abbreviations

ABC	Approximate Bayesian Computation
AIC	Akaike's Information Criterion
AR	Acceptance Rate
BFMC	Brute-Force Monte Carlo
BIC	Bayesian Information Criterion
BTC	Breakthrough Curve
DIC	Deviance Information Criterion
EM	Expectation-Maximization
GMIS	Gaussian Mixture Importance Sampling
GPR	Ground Penetrating Radar
HM	Harmonic Mean
KIC	Kashyap's Information Criterion
LM	Laplace-Metropolis
MADE	MacroDispersion Experiment
MAP	Maximum A-Posteriori
MCMC	Markov chain Monte Carlo
MLS	Multi-Level Sampler
MPS	Multiple-Point Statistics
PC	POLYCHORD
pdf	posterior density function
RMSE	Root Mean Square Error
SS	Stepping-Stone sampling
TH	Thermodynamic Integration

Résumé

Les eaux souterraines sont une ressource fondamentale. Avec la croissance démographique, le changement d'utilisation du sol, les activités économiques, l'urbanisation et le changement climatique, une gestion sûre et durable des ressources en eau souterraine devient de plus en plus cruciale. Cela doit reposer sur une caractérisation précise de l'hétérogénéité des propriétés hydrogéologiques du sous-sol, tâche qui représente toutefois un défi. Premièrement, le sous-sol est caché et la collecte locale de données renseignant sur les propriétés hydrogéologiques est difficile ou trop coûteuse. Deuxièmement, les méthodes géophysiques peuvent permettre une acquisition efficace de telles mesures, elles nécessitent néanmoins la définition des relations pétrophysiques qui sont souvent incertaines et mal connues. Troisièmement, la structure géologique des systèmes hébergeant les eaux souterraines est complexe et la définition d'un modèle conceptuel correspondant n'est pas unique. Cela conduit à l'une des sources d'incertitude majeure (et souvent ignorée) dans les études de modélisation, appelée *incertitude conceptuelle*. La sélection bayésienne de modèles, reposant sur le calcul de l'évidence et sur les facteurs de Bayes, fournit une approche quantitative permettant de comparer et de classer des modèles conceptuels alternatifs et, par conséquent, de prendre en compte l'incertitude conceptuelle. Dans cette thèse, nous étudierons l'utilisation de la sélection bayésienne de modèles en hydrogéophysique et en hydrogéologie en répondant aux questions de recherche suivantes : (1) Les données géophysiques sont-elles appropriées pour guider la sélection de modèles en hydrogéologie ? (2) L'incertitude pétrophysique et sa structure spatiale peuvent-elles être déduites dans des études hydrogéophysiques et quel impact ont-elles sur l'inversion bayésienne et la sélection de modèles ? (3) Comment pouvons-nous réaliser la sélection de modèles lorsque nous ciblons des modèles conceptuels aux structures géologiques réalistes représentés par des images d'entraînement ? Ces objectifs seront traités en utilisant une approche bayésienne complète basée sur les algorithmes de Monte Carlo par chaînes de Markov. Les objectifs de la recherche seront ensuite explorés via des études de cas synthétiques et réels, dans le but de caractériser spatialement les champs de porosité ou de conductivité hydraulique dans les aquifères. Dans notre première étude de sélection bayésienne de modèles en hydrogéophysique, nous concluons que les méthodes géophysiques peuvent être utiles pour choisir la représentation hydrogéologique du sous-sol qui est la plus étayée par les données disponibles, parmi un ensemble de modèles conceptuels concurrents. Nous proposons une méthode pour prendre en compte et déduire l'incertitude pétrophysique et sa corrélation spatiale. Nous constatons que cette approche conduit à une diminution du biais et à une quantification plus réaliste de l'incertitude et du classement des modèles conceptuels. De plus, nous proposons et appliquons avec succès une nouvelle méthodologie pour effectuer la sélection bayésienne de modèles parmi des modèles conceptuels géologiquement réalistes.

Mots clefs : Sélection bayésienne de modèles, hydrogéophysique, évidence, incertitude pétrophysique, incertitude conceptuelle, méthode de Monte Carlo par chaînes de Markov, image d'entraînement

Abstract

Groundwater is a fundamental source of drinking water. With population growth, land use changes, economic activities, urbanisation and climate change, a safe and sustainable management of groundwater resources is becoming more and more critical. This needs to rely on an accurate characterisation of the hydrogeological heterogeneity in the subsurface, which is a challenging task. First, the subsurface is hidden from sight and collecting local hydrogeological measurements is difficult or too expensive. Second, geophysical methods can effectively support such measurements but, at the same time, they require the definition of petrophysical relationships that are often uncertain and poorly known. Third, the spatial geological structure of groundwater systems is complex and the definition of the corresponding conceptual model is non-unique. This leads to one of the main (and often ignored) sources of uncertainty in modelling studies, namely *conceptual uncertainty*. Bayesian model selection relying on evidence computation and Bayes factors provides a quantitative approach for comparing and ranking alternative conceptual models and, therefore, accounting for conceptual uncertainty. In this thesis, we will investigate the use of Bayesian model selection in hydrogeophysics and hydrogeology by answering the following research questions: (1) Are geophysical data suitable for guiding model selection in hydrogeology? (2) Can petrophysical uncertainty and its spatial structure be inferred in hydrogeophysical studies and how do they impact Bayesian inversion and model selection? (3) How can we achieve model selection when targeting geologically-realistic hydrogeological conceptual models represented by training images? These objectives will be addressed using a full Bayesian approach based on Markov chain Monte Carlo algorithms. The research goals will be then explored in light of synthetic and field-based case studies with the purpose of characterising spatially-distributed porosity or hydraulic conductivity fields in aquifers. From the first comparative study of Bayesian model selection in hydrogeophysics ever, we conclude that geophysical methods can be valuable in providing guidance about which hydrogeological representation of the subsurface is the most supported by the available data among a set of competing conceptual models. We then propose a method to account for and infer the spatially-correlated uncertainty of petrophysical relationships. We find that this approach leads to less bias, more realistic uncertainty quantification and less overconfident model selection. Moreover, we propose and successfully apply a new methodology for performing Bayesian model selection among geologically-realistic conceptual models represented by training images.

Key words: Bayesian model selection, hydrogeophysics, evidence, petrophysical uncertainty, conceptual uncertainty, Markov chain Monte Carlo, training image

Chapter 1

Introduction

Water is essential for human life and nature. Population growth, land use changes, economic activities, urbanisation and climate change contribute to both a decreasing water supply and an increasing water demand (IPCC report by Jiménez Cisneros et al. (2014)). These combining factors are expected to lead to an estimated 40% global water supply shortage by 2030 as reported by the European Commission (2012). Over 95% of the freshwater on the planet, excluding glaciers and ice caps, is found underground and it is a fundamental source of drinking water. Groundwater systems are prone to over-pumping and contamination from agricultural and industrial activities and they are vulnerable to extreme events such as droughts and floods. Over the past 30 years, European water policy has been focused on water resources protection (e.g., quality and sufficient quantity of water). A safe and sustainable management of groundwater resources, as well as reliable assessment of water policies, are becoming more and more critical and they should, in fact, rely on quantitative subsurface modelling studies (Scheidt et al., 2018) that are able to simulate past and present conditions and predict future responses of aquifers to natural and anthropogenic stresses (Maliva, 2016).

An aquifer is a geological unit that can store useable amounts of water. The characterisation of the (hydro)geological heterogeneity of an aquifer is fundamental because, at small scales, it is a key controlling factor in flow and transport of contaminants and, at larger scales, it influences the rate, position and magnitude of recharge and discharge areas (Maliva, 2016). The properties, structure and processes taking place in an aquifer are difficult to characterise and not fully understood. Indeed, the subsurface is hidden from sight and collecting local measurements of aquifer hydraulic and transport parameters (e.g., porosity, hydraulic conductivity) is challenging or too expensive. Conventional methods in the field of hydrogeology to gather such measurements consist, for instance, by drilling boreholes for collecting soil samples, logging the penetrated geological formations and/or installing fluid sampling instruments to be used, for example, in pumping and tracer tests (Maliva, 2016). These techniques are invasive, costly and provide only point measurements. Geophysical methods (e.g., ground penetrating radar (GPR), direct current resistivity, electromagnetics, seismics) can effectively support the conventional hydrogeological techniques by collecting complementary information over more extended areas and at a lower cost. This combined approach of using hydrogeological and geophysical data forms since the 1990s the discipline of hydrogeophysics (Rubin and Hubbard, 2005; Vereecken et al., 2006). Hydrogeophysics explores the potential of using geophysical methods to infer, with high resolution, hydrologic parameters and processes and the spatial structure of the subsurface relevant for hydrological investigations. The combination of geophysical and hydrogeological investigations is valuable to improve the mapping and the understanding of subsurface systems.

However, the underlying geology and corresponding hydrogeological parameters will never be exactly known, the processes taking place in subsurface systems will never be fully understood and data are noisy and sparse. Quantifying and acknowledging the degree of "ignorance" is fundamental to reliably manage groundwater systems and to effectively support decision-making (Scheidt et al., 2018).

1.1 Motivation: the Bayesian approach

Studies of subsurface systems are often formulated and solved as inverse problems (Tarantola, 2005). Solving an inverse problem consists in using observed data (measurements collected in the field) and prior information to infer the parameters of interest that describe the system under study and their uncertainties. The scientific approach to tackle such problems involves three steps.

(i) *Parameterization*: the natural system is conceptualised and parametrised when defining a conceptual model (e.g., simplified representation of the geological structure of an aquifer) and assigning the model parameters values that characterise the system (e.g., hydraulic conductivity values of each lithofacies).

(ii) *Forward modelling*: simulation of the response of a given conceptual model and values of the model parameters using a forward model. A forward model is often a numerical solver implementing physical laws. For instance, if the data are solute concentrations measured during a tracer experiment, the forward model may consist of a set of equations that are solved on a discretised mesh to simulate flow and transport of a solute in a porous medium.

(iii) *Inverse modelling*: the observed data are compared with the simulated ones and model parameter values are updated in order to infer the actual values of the model parameters.

The accuracy of any inference about underlying parameters of interest is affected by several sources of uncertainty. We can distinguish between the uncertainty in the data, the forward model, the petrophysical relationship and the choice of the conceptual model and its assumptions (i.e., values and parameters associated to it). Indeed, the data are noisy and has a limited spatial coverage. The forward model is a simplified physical description of the system and capture only the main processes of the system that are relevant to the study at hand. Geophysical methods are directly sensitive to physical properties of the subsurface and petrophysical (also called rock physics) relationships need to be defined to link these properties to the hydrogeological parameters and state variables of interest (Binley et al., 2010). One challenge in hydrogeophysics is that the petrophysical relationships in shallow subsurfaces are often non-stationary, non-unique and poorly understood (Rubin and Hubbard, 2005; Linde et al., 2006b), thereby, affecting the predictive power of the inferred parameters of interest. The identification and conceptualisation of the geological structure of groundwater systems are challenging due to their high heterogeneity and spatial variability.

The definition of a conceptual model for representing a subsurface system is non-unique (Backus and Gilbert, 1967) and it is one of the main sources of uncertainty in modelling studies (Refsgaard et al., 2006; Bond et al., 2007; Rojas et al., 2008; Pirot et al., 2015; Scheidt et al., 2018); it is often referred to as *conceptual uncertainty*. The predominant practice in most hydrogeophysical and hydrogeological studies is to estimate the parameters of interest under the assumption of one single (often rather simple) representation of the subsurface and to ignore the uncertainty associated with this choice. Testing alternative conceptual models should be promoted to account for conceptual uncertainty (Refsgaard and Henriksen, 2004; Linde, 2014; Linde et al., 2015b; Nilsson et al., 2006; Refsgaard et al., 2012).

The Bayesian approach is a powerful tool to solve inverse problems and to account for different sources of uncertainty that affect natural system investigations. The main advantages and limitations of the Bayesian approach are listed below.

Advantages
Uncertainty quantification. The Bayesian approach provides a general and flexible probabilistic framework for solving inverse problems (Bayesian inference, Section 1.3.1) that fully quantifies uncertainty. The Bayesian approach allows to simultaneously account for different sources of uncertainty such as those mentioned above.

Clarity. As opposed to a "black box", the Bayesian approach requires that each source of uncertainty is explicitly described and that each step in the inverse problem solution is clearly stated.

Easy interpretation. The uncertainty is quantified in terms of a probability distribution that represents the degree of belief about an unknown parameter of interest. This Bayesian interpretation of probability is more straightforward than the classical (frequentist) point of view (Gelman et al., 2013). Indeed, the uncertainty in classical statistics is expressed in terms of confidence intervals that quantify the probability of a certain parameter value in terms of the fraction of times that value occurred after an infinitely repeated number of inferences.

Conceptual uncertainty. The Bayesian approach enables inclusion of conceptual uncertainty. Bayesian model selection (Section 1.3) addresses this type of uncertainty using results from Bayesian inference and the computation of Bayes Factors to answer questions as: Which model among a set of competing conceptual models is most supported by the available data? How well does it perform relative to the other conceptual models in the set?.

Definition of a prior. The Bayesian approach requires an explicit description of prior knowledge. This stimulates scientists to think, discuss and delve deeper into this aspect, thereby, contributing to a better understanding of the system at hand. For instance, in subsurface modelling, significant effort has been dedicated in recent years to build more geologically-realistic priors using, for example, multiple-point statistics (MPS). Moreover, describing properly the prior knowledge about a system is a process that may involve different experts and stakeholders, thereby, possibly increasing the confidence in the decision-making process.

Limitations
Definition of a prior. There is ongoing research and discussions on how to properly describe the prior knowledge and how the choice of the prior impacts the inference and model selec-

tion results. The definition of a prior is problem specific, subjective, non-trivial and general guidelines on how to choose it does not exist when dealing with spatial hydrogeological property fields. In subsurface system studies, this issue becomes acute when no or only little prior information is available and the scientist is often pushed to choose possibly inadequate prior descriptions, such as, uniform and multi-Gaussian distributions (Scheidt et al., 2018; de Pasquale and Linde, 2017).

Computationally intensive. The Bayesian approach is known to be computationally demanding, especially when using large datasets, complex models and many parameters that are to be estimated. The need for integrating competences and increasing collaboration between computational and statistical science in this domain has been identified by the International Society for Bayesian Analysis (Jordan, 2011).

Note that the definition of a prior in Bayesian approaches is controversial and it can been seen both as an advantage and a limitation.

1.2 Conceptual models

A conceptual model is a simplified representation of a real system that is built in order to achieve an improved understanding of that reality and to meet the goals of the modelling study at hand. Conceptual models are built based on the prior knowledge that is available about the system using equations, assumptions, governing relationships and spatial parametrisation in order to make interpretations about the system. *"The conceptual model in other words constitutes the scientific hypothesis or theory that we assume for our particular modelling study"* (Refsgaard and Henriksen, 2004).

In the field of subsurface systems and in this thesis, a conceptual model is a geological interpretation of the subsurface through the definition of (i) the spatial discretisation and parameterisation of the (hydro)geological heterogeneity, (ii) the prior probability density functions (pdf) that represent all the possible values that each model parameter can take. The model parameters can be assumed to be known or unknown (inferred from the data). If they are unknown, the most probable values and the associated uncertainties given the data are derived by Bayesian inference (Section 1.3.1). In the field of hydrogeophysics, a conceptual model can also include the definition of a petrophysical model and the prior pdf of their parameters if they need to be inferred. Appropriate conceptual models of (hydro)geological heterogeneity in the subsurface are crucial for a reliable and accurate groundwater modelling (Maliva, 2016).

In this thesis, the subsurface heterogeneity is spatially discretised on regular grids. The choice of the grid cell size is driven, for example, by the scale at which we are interested to investigate the system heterogeneity, the purpose of the study and computational limitations. The spatial parameterisation of the subsurface heterogeneity is here carried out using a zonation approach, a variogram-based (or two-point) geostatistical approach and a MPS approach. In the zonation approach, the subsurface is subdivided into zones within which the value of the parameter of interest is assumed constant because its variation is negligibly small compared to the variations between different zones. The zonation approach is adequate for representing sharp discontinuities in the subsurface geological structure. However, poorly defined locations of the boundaries may lead to biased model parameter estimation (Vanrolleghem, 2010; Linde et al., 2006b). A very simple example of a conceptual model parametrised with the zonation approach is a horizontally layered model (Figure 1.1a).

The variogram-based geostatistical approach describes the spatial distribution of the parameters of interest as a random field with a correlation structure defined by means and covariances that convey information on the variance and the integral scales of spatial parameter correlation in different directions (i.e., anisotropy). A classical example of conceptual models built from two-point geostatistics are the multi-Gaussian fields in Figure 1.1b-d. This type of conceptualisation is widely used. However, it is well recognised that they may be simplistic and inadequate to capture the complexity of the subsurface geological structure and, thus, to properly reproduce and predict flow and transport processes in subsurface systems (Gómez-Hernández and Wen, 1998; Journel and Zhang, 2006; Kerrou et al., 2008).

Training images offer a means to conceptualise the prior geological knowledge of the system under study and MPS allows to effectively reproduce the complex geological patterns (e.g., curvilinear features) found in the training image (Guardiano and Srivastava, 1993; Strebelle, 2002; Hu and Chugunova, 2008; Mariethoz and Caers, 2014). The first simulation algorithm based on training images and MPS is SNESIM (Strebelle, 2002) that is limited to categorical fields. Efficient and computationally fast simulation algorithms that are able to sample from both categorical and continuous images are, for example, the direct sampling (Mariethoz et al., 2010b) and the recent graph cuts (Zahner et al., 2016) methods. Examples of conceptual models built using graph cuts are shown in Figure 1.1e-h. The prior geological understanding is informed by expert knowledge, outcrops and geophysical and borehole data. These informations are then used to create a training image from sketches drawn by hand, digitalised outcrops, process-imitating, structure-imitating or descriptive simulation methods (Koltermann and Gorelick, 1996; De Marsily et al., 2005).

Conceptual models built with zonation or variogram-based geostatistics imply explicit formulas for the prior pdfs (e.g., parametric priors such as Gaussian or exponential functions) of the model parameters. On the other hand, the conceptual models obtained from training images circumvent the definition of parametric priors by using a pseudo-random process (e.g., sequential geostatistical resampling) that produces samples according to the prior distribution (Mosegaard and Tarantola, 1995). From this prospective, the prior pdf is represented by a series of simulation steps rather than an explicit formula (Ruggeri et al., 2015) and a specific acceptance criterion in the Markov chain Monte Carlo (MCMC) algorithm is needed in these cases (Section 1.5). For a criticism of this type of approach, the reader is referred to Emery and Lantuéjoul (2014), in which it is suggested that the training image must be of infinite extent to enable a complete uncertainty quantification.

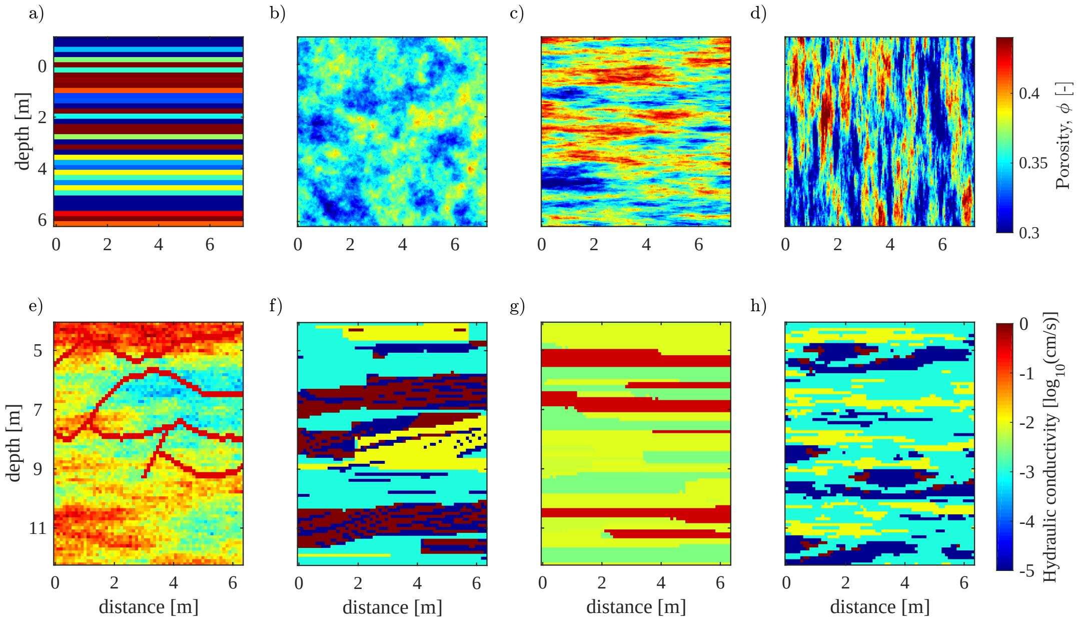


Figure 1.1 – Examples of conceptual models of subsurface (hydro)geological heterogeneity parameterised with the (a) zonation, (b-d) variogram-based and (e-h) multiple-point geostatistics approaches. The different spatial parameterisations consist of (a) horizontal layers, (b) multi-Gaussian with isotropy, (c) multi-Gaussian with horizontal anisotropy, (d) multi-Gaussian with vertical anisotropy and (e) conductive channels overlapped on a multi-Gaussian field (i.e., combination of continuous and categorical fields) and (f-h) categorical fields (i.e., each facies has a specific value of the hydrogeological property).

1.3 Bayesian model selection

Suppose that an aquifer needs to be characterised in order to perform groundwater model predictions. As previously explained (Section 1.1), direct investigations of the subsurface is challenging, the geological structure is heterogeneous and it has high spatial variability. Several experts, such as geologists and hydrogeologists, can be consulted and asked to provide their insights about the expected aquifer structure based on their experience (prior knowledge). In this process, we might come up with several plausible conceptualisations of the aquifer (hydro)geological structure (e.g., layered, multi-Gaussian, outcrop-based) that span a wide range of groundwater predictions. In many real applications, this conceptual uncertainty is a dominant source of uncertainty and ignoring it may imply a drastic underestimation of uncertainty on model predictions (Refsgaard et al., 2006; Bond et al., 2007; Rojas et al., 2008; Pirot et al., 2015; Scheidt et al., 2018). How can we deal with such conceptual uncertainty? Bayesian model selection based on Bayes factors (Jeffreys, 1935, 1939; Kass and Raftery, 1995) provides a quantitative approach for comparing and ranking alternative hypotheses relative to each other (probability of one hypothesis to another) in the light of the observed data. Note that the term hypothesis and conceptual model (as defined in Section 1.2) are here used interchangeably.

This approach of quantifying and improving the state of knowledge about reality by testing and comparing different perceptions of that reality based on the information at hand (data) is fully in line with the concept of falsificationism introduced in Popper's scientific philosophy (Popper, 2005). All conceptual models are wrong (Box, 1979) and they are not verifiable in the sense that the true conceptual model is never possible to be proven (Konikow and Bredehoeft, 1992; Oreskes et al., 1994). On the other hand, a hypothesis can be corroborated (confirmed) or falsified (refuted) depending on wether or not it is in agreement with the current scientific knowledge (Popper, 2005). If a hypothesis predicts the observed data much more poorly than the other hypothesis, then it is falsified, otherwise, it is retained and tested against new data and hypotheses. This is one view of how science progresses.

Some important aspects should be kept in mind when comparing different conceptual models.

Purpose. The formulation of the research question is the crucial starting point. It should address the final practical purpose or effectively inform the decision makers and it should drive hard thinking about prior knowledge, conceptual model-building and data collection. *"A good answer to a poor question [...] is little better than a poor answer to a poor question"* (Burnham and Anderson, 2003). Most of the real investigations have an *inferential* purpose, that is, obtaining the most reliable inferences about the quantities of interest. In these cases, a set of carefully defined conceptual models may be used. On the other hand, if very little prior knowledge (e.g., geological structure, model parameters, governing equations) is available for the system under study, the comparison of alternative conceptual models may be used for *exploratory* purposes (Burnham and Anderson, 2003) and for guiding the conceptual model-building process based on falsifications. In such a case, the conceptual models should differ as much as possible from each other. For both exploratory and inferential purposes, the set of competing conceptual models should be ideally as large as possible (while staying within the limits of computational constraints).

Importance of data. Data should be of high quality and they should carry information that is relevant specifically to the practical purposes at hand and to the process of decision-making. Therefore, conceptual model selection is useful only if they are compared using such informative data sets. Scheidt et al. (2018) stress the need for more data of higher quality in groundwater management.

Consistency. If enough data are available and if the true conceptual model is part of the set of the competing conceptual models, then Bayesian model selection guarantees the selection of the true model (Berger et al., 2001).

Interpretation of the "best" conceptual model in the set. The conceptual model that performs the best in the set does not mean that it is the one that best represent the full reality; it just suggests that it is the hypothesis in the set that is the most supported by the information in the data (Refsgaard and Henriksen, 2004). Moreover, since the Bayesian approach to model selection naturally honour the principle of parsimony (Section 1.3.2), the "best" model will not be too simple (underfitting) and not too complex (overfitting) and, consequently, estimates and predictions based on this model will not be too overly optimistic (Berger et al., 2001).

As shown in the following sections, Bayesian model selection based on Bayes factors consists in applying Bayesian inference at two different levels: at the level of the model parameters and at the level of the conceptual model.

1.3.1 Bayesian inference

Bayesian inference is the process of drawing conclusions from data in terms of probability statements about quantities of interest that are not directly observed. This is accomplished by Bayes' theorem, a learning rule that expresses how prior knowledge about the system under study is updated by a data-dependent term, called the likelihood function, resulting in a posterior pdf that depends both on the prior state of knowledge as well as the data.

Assume that we want to apply Bayesian model selection on a set of m conceptual models, $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_m\}$ and that each model η_k, with $k = 1, \ldots, m$, is described by a vector of parameters $\boldsymbol{\theta}_k$. At the first level of inference, we infer what the model parameters, $\boldsymbol{\theta}_k$, of the conceptual model η_k might be, given n data, $\widetilde{\mathbf{Y}} = \{\widetilde{y}_1, \ldots, \widetilde{y}_n\}$. Applying Bayes' theorem, we obtain the posterior pdf, $p(\boldsymbol{\theta}_k|\widetilde{\mathbf{Y}}, \eta_k)$, of the parameters of interest $\boldsymbol{\theta}_k$ as:

$$p(\boldsymbol{\theta}_k|\widetilde{\mathbf{Y}}, \eta_k) = \frac{p(\boldsymbol{\theta}_k|\eta_k)\,p(\widetilde{\mathbf{Y}}|\boldsymbol{\theta}_k, \eta_k)}{p(\widetilde{\mathbf{Y}}|\eta_k)}. \tag{1.1}$$

The prior pdf, $p(\boldsymbol{\theta}_k|\eta_k)$, quantifies probabilistically the initial state of knowledge about what values the model parameters might take before considering the observed data. The likelihood function, $L(\boldsymbol{\theta}_k, \eta_k|\widetilde{\mathbf{Y}}) \equiv p(\widetilde{\mathbf{Y}}|\boldsymbol{\theta}_k, \eta_k)$, quantifies the plausibility of the model parameters given the data. Bayesian inference can be performed with any type of likelihood function. However, a Gaussian likelihood function is often used (out of convenience) by assuming uncorrelated and normally distributed data errors with constant standard deviation, $\sigma_{\widetilde{\mathbf{Y}}}$,

$$L(\boldsymbol{\theta}_k, \eta_k|\widetilde{\mathbf{Y}}) = \left(\sqrt{2\pi\sigma^2_{\widetilde{\mathbf{Y}}}}\right)^{-n} \exp\left[-\frac{1}{2}\sum_{h=1}^{n}\left(\frac{\mathscr{F}_h(\boldsymbol{\theta}_k) - \widetilde{y}_h}{\sigma_{\widetilde{\mathbf{Y}}}}\right)^2\right]. \tag{1.2}$$

The larger the likelihood the better the forward model, $\mathscr{F}(\boldsymbol{\theta}_k)$, predicts the observed data, $\widetilde{\mathbf{Y}}$. The evidence, $p(\widetilde{\mathbf{Y}}|\eta_k)$, also called marginal likelihood, evaluates the support provided by the observed data to the conceptual model, η_k. The evidence is the normalising constant in Bayes' theorem and, in the case of discrete $\boldsymbol{\theta}_k$, is defined as $p(\widetilde{\mathbf{Y}}|\eta_k) = \sum_{\boldsymbol{\theta}_k} p(\boldsymbol{\theta}_k|\eta_k)L(\boldsymbol{\theta}_k, \eta_k|\widetilde{\mathbf{Y}})$ where the sum is over all possible values of $\boldsymbol{\theta}_k$. However, in most applications, $\boldsymbol{\theta}_k$ is continuous and, therefore, the evidence is defined as the (multidimensional) integral of the likelihood function over the prior distribution,

$$p(\widetilde{\mathbf{Y}}|\eta_k) = \int p(\boldsymbol{\theta}_k|\eta_k)L(\boldsymbol{\theta}_k, \eta_k|\widetilde{\mathbf{Y}})d\boldsymbol{\theta}_k. \tag{1.3}$$

In other words, we can define the evidence as the average (integral) over the parameter space of the likelihood function weighted by the prior pdf. In the first level of inference that focuses on parameter values, the posterior pdf of the model parameters of each conceptual model are inferred and the evidence is neglected because it is solely a normalising constant.

At the second level of inference, the Bayes' theorem is applied a second time in order to infer which conceptual model in the set is most plausible given the data. We obtain the posterior probability of the conceptual model η_k as:

$$p(\eta_k|\widetilde{\mathbf{Y}}) = \frac{p(\eta_k)p(\widetilde{\mathbf{Y}}|\eta_k)}{\sum_{i=1}^{m} p(\eta_i)p(\widetilde{\mathbf{Y}}|\eta_i)}. \tag{1.4}$$

The prior probability, $p(\eta_k)$, of the conceptual model η_k quantifies the prior plausibility that we assign to η_k before considering the data. However, specifying the prior probability for each conceptual model is often not necessary and we assume here that all the competing conceptual models in the set have the same prior probability. This implies that the conceptual model ranking is based merely on the evidence estimates. The denominator is the normalising factor of Equation 1.4 and the sum is over all the competing conceptual models in the set, $\boldsymbol{\eta}$. It is clear from Equation 1.4 that the evidence is how the observed data update our prior beliefs about a conceptual model.

1.3.2 Bayes Factors and Evidence

The usefulness of a conceptual model to predict data relative to other hypotheses is assessed by Bayes factors (Jeffreys, 1935, 1939; Kass and Raftery, 1995; Morey et al., 2016). If we want to compare two conceptual models of the set, η_1 and η_2, the Bayes' theorem in Equation 1.4 can be rewritten in terms of posterior and prior odds of η_1 compared to η_2:

$$\frac{p(\eta_1|\widetilde{\mathbf{Y}})}{p(\eta_2|\widetilde{\mathbf{Y}})} = \frac{p(\eta_1)p(\widetilde{\mathbf{Y}}|\eta_1)}{p(\eta_2)p(\widetilde{\mathbf{Y}}|\eta_2)}. \tag{1.5}$$

The posterior odds, $p(\eta_1|\widetilde{\mathbf{Y}})/p(\eta_2|\widetilde{\mathbf{Y}})$, is the ratio of the posterior probability of η_1 and η_2 and it quantifies the degree of belief in favour of η_1 over η_2 after observing the data. The prior odds, $p(\eta_1)/p(\eta_2)$, is the ratio of the prior probability of η_1 and η_2 and it quantifies the degree of belief a-priori (before considering the data) in favour of η_1 over η_2. We are interested in comparing the performance of η_1 against η_2 and this information is carried by the ratio of the evidence of the two competing model (last ratio in Equation 1.5) and it is termed the Bayes factor. The Bayes factor of conceptual model η_1 with respect to conceptual model η_2 is defined as the ratio between the posterior and prior odds of η_1 compared to η_2:

$$B_{(\eta_1,\eta_2)} = \frac{p(\eta_1|\widetilde{\mathbf{Y}})}{p(\eta_2|\widetilde{\mathbf{Y}})} \bigg/ \frac{p(\eta_1)}{p(\eta_2)} = \frac{p(\widetilde{\mathbf{Y}}|\eta_1)}{p(\widetilde{\mathbf{Y}}|\eta_2)}, \tag{1.6}$$

and it expresses how well the observed data were predicted by η_1 compared to η_2.

Jeffreys (1939) and Kass and Raftery (1995) proposed a scale to interpret the Bayes factor (Table 1.1). If Bayes factors, $B_{(\eta_1,\eta_i)}$, of η_1 compared to each competing conceptual model, η_i, are all > 150, we may select η_1 as the best conceptual model in the set; otherwise, if all the $B_{(\eta_1,\eta_i)}$ are < 1, we may discard η_1 from the set. The scale proposed by Jeffreys (1939) and Kass and Raftery (1995) is helpful and it can be used as a guideline, however, the interpretation may depend on the context and on the judgement of the modeller based on the practical purposes addressed.

Table 1.1 – Interpretation of Kass and Raftery (1995) (slightly different from the original interpretation of Jeffreys (1939)) for the Bayes factor of two conceptual models η_1 and η_2.

$2\log B_{(\eta_1,\eta_2)}$	$B_{(\eta_1,\eta_2)}$	Evidence for η_1
< 0	< 1	negative (supports η_2)
0 to 2	1 to 3	barely worth mentioning
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
> 10	> 150	very strong

The evidence is the cornerstone of Bayesian model comparison, model ranking, model selection and model averaging. Bayesian model ranking consists in listing the conceptual models of the set from the best one to the worst one according to the decreasing value of the evidence. Bayesian model selection chooses one single (the "best") model from the set that is the one with the highest evidence and uses that conceptual model for inferring the parameter of interest. Bayesian model selection may be justified when Bayes factors in favour of the best conceptual model compared to the other conceptual models in the set are all larger than 150 (Table 1.1). If this is not the case and the evidence values are within two orders of magnitude, then Bayesian model averaging is preferred. Bayesian model averaging (Hoeting et al., 1999) retains all the conceptual models and derives a composite estimation of a quantity of interest, ϕ, as:

$$p(\varphi|\widetilde{\mathbf{Y}}) = \sum_{k=1}^{m} p(\varphi|\widetilde{\mathbf{Y}},\eta_k)p(\eta_k|\widetilde{\mathbf{Y}}), \tag{1.7}$$

that is, the average of the posterior probability of φ under each of the conceptual models considered, $p(\theta|\widetilde{\mathbf{Y}},\eta_k)$, weighted by the posterior probability of each conceptual model, $p(\eta_k|\widetilde{\mathbf{Y}})$ (Equation 1.4). Bayesian model averaging provides a rigorous assessment of conceptual uncertainty. However, keeping all the conceptual models may not be practical for communication or descriptive purposes (Berger et al., 2001) and, in these cases, selecting the "best" conceptual model of the set may be useful (Clyde and George, 2004).

A powerful property of the evidence is that it intrinsically honours the Occam's razor principle (Jefferys and Berger, 1992; MacKay, 2003). The Occam's razor (Thorburn, 1918) refers to what William of Occam suggested in the fourteenth century, *"shave away all that is unnecessary"*, that is the principle of parsimony. When comparing and ranking alternative hypotheses, it is advisable to honour the concept of parsimony, which can be seen as a trade-off between model complexity and goodness of fit. For instance, if two (or more) conceptual models, η_1 and η_2, fit (almost) equally well the observed data, $\widetilde{\mathbf{Y}}$, then the simplest one, say η_1, is preferred over the more complex one, η_2, thereby, avoiding the well known problem of overfitting. The complexity of a conceptual model is not easily definable (van der Linde, 2012; Guthke, 2017; Höge et al., 2018), however, for the sake of simplicity, we may refer to it as the

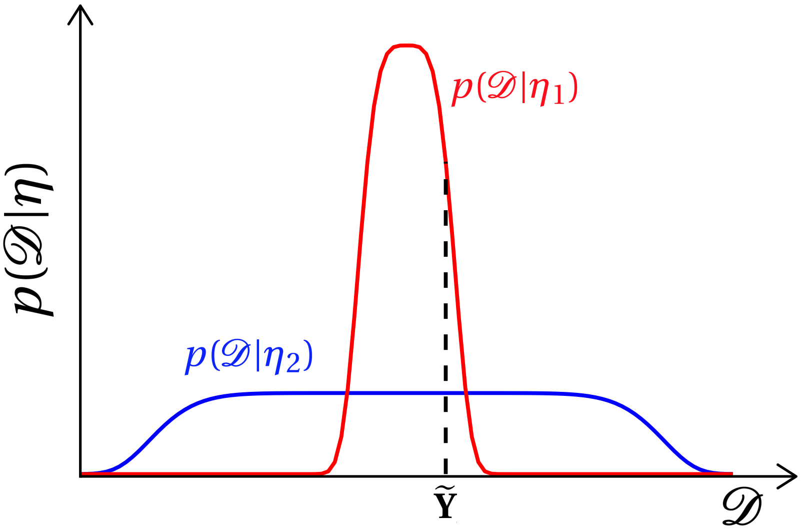


Figure 1.2 – The principle of parsimony in Bayesian model selection. The evidence is here interpreted as the normalised probability distribution, $p(\mathscr{D}|\eta)$, on the space of all possible data sets of fixed size n, $\mathscr{D}$. A more simple conceptual model, η_1, spreads the probability distribution, $p(\mathscr{D}|\eta_1)$ (red line), over a smaller range of data sets than a more complex model, η_2, (blue line). Given the observed data set, $\widetilde{\mathbf{Y}}$ (dotted black line), then it is clear that the simpler model η_1 receives a higher evidence than η_2 provided that the two conceptual models fit (almost) equally well $\widetilde{\mathbf{Y}}$ and they have equal prior probability. (Figure inspired by the figures from MacKay (2003); Ghahramani (2013).

number of degrees of freedom (i.e., number of independent model parameters). A more complex conceptual model has more adjustable parameters that allows for a larger range of predictions; vice-versa, simpler models generate a more narrow range of predictions. As a consequence, if η_1 and η_2 fit (almost) equally well the observed data and they have equal prior probability, then the more complex model, η_2, will not predict the observed data set as strongly as η_1, and η_1 will be favoured (higher evidence) over η_2 (Figure 1.2). One challenge in using the evidence for Bayesian model selection is that its computation is difficult and expensive. In most applications of interest, the parameter space is rather high-dimensional and the computation of the integral that defines the evidence (Equation 1.3) has not an analytical solution. Different methods exist to estimate the evidence and we detail some of them in the next section.

1.3.3 Evidence computation

Model selection can either be performed based on analytical expressions based on strong mathematical approximations or numerical estimations of the evidence. The most common approximations are: Akaike's information criterion (AIC) (Akaike, 1973), deviance information criterion (DIC) (Spiegelhalter et al., 2002; Steininger et al., 2014), Bayesian information criterion (BIC) (Schwarz et al., 1978), Kashyap's information criterion (KIC) (Kashyap, 1982) and Laplace-Metropolis method (LM) (Lewis and Raftery, 1997) that is closely related to KIC (i.e., KIC=-2logLM). All these mathematical approximations are easy to implement and fast to compute, but when the forward model is non-linear and the conceptual models are described by many parameters, most of them (AIC, DIC, BIC) provide poor and inconsistent results (e.g., Lu et al. (2011); Schöniger et al. (2014); Pooley and Marion (2018)).

The LM method evaluated at the maximum a-posteriori (MAP) point is the one that performs the best among the mathematical approximations listed above, (Schöniger et al., 2014). The LM method estimates the evidence by approximating the integrand in Equation 1.3 with a quadratic Taylor series expansion around the MAP estimate, θ^*:

$$p_{\mathrm{LM}}(\widetilde{\mathbf{Y}}|\eta) \approx (2\pi)^{d/2} |\mathbf{H}(\theta^*)|^{1/2} p(\theta^*|\eta) L(\theta^*, \eta|\widetilde{\mathbf{Y}}), \tag{1.8}$$

where d is the number of parameters in the conceptual model η and $|\mathbf{H}(\theta^*)|$ is the determinant of minus the inverse Hessian matrix evaluated at θ^*. The LM is computationally fast but it is built on the strong assumption that the posterior distribution of each parameter of interest is well approximated by a Gaussian distribution that peaks on the corresponding MAP estimate. This underlying assumption should be kept in mind when interpreting the results provided by the LM method.

An alternative to the mathematical approximation of the evidence is to use numerical evaluations such as the Brute-force Monte Carlo sampling (BFMC) (Hammersley and Handscomb, 1964), Harmonic mean estimator (HM) (Newton and Raftery, 1994), importance sampling (Hammersley and Handscomb, 1964), thermodynamic integration (TH) (also called path sampling) (Gelman and Meng, 1998; Friel and Pettitt, 2008), stepping stone sampling (SS)(Xie et al., 2011) and nested sampling (Skilling, 2004; Skilling et al., 2006).

The BFMC method consists in drawing randomly N samples of the model parameters, θ, from their corresponding prior distributions and, as the sum of the prior probabilities of the samples equals 1 in this case, the integral of Equation 1.3 can be simply approximated by the average of the likelihood of their prior samples:

$$p_{\mathrm{BFMC}}(\widetilde{\mathbf{Y}}|\eta) \approx \frac{1}{N} \sum_{i=1}^{N} L(\theta_i, \eta|\widetilde{\mathbf{Y}}). \tag{1.9}$$

The accuracy of the BFMC method is ensured by the law of large numbers and the central limit theorem. The BFMC method is an accurate estimator if a very large number of samples is considered and if a large fraction of the prior range has a significant likelihood. For most real-world applications, it suffers from the curse of dimensionality meaning that the computational cost prevent the use of an appropriate number of samples, which leads to underestimation of the evidence. The computational requirements of the BFMC method, thereby, becomes rather impractical for parameter-rich models as many millions or even billions of model evaluations are required to average the likelihood surface.

As opposed to the BFMC, the harmonic mean estimator draws samples from the posterior distribution (e.g., through MCMC methods) and approximates the evidence as the harmonic mean of the likelihood of the posterior samples. This estimator has been proven unreliable (e.g., Newton and Raftery (1994); Liu et al. (2016)). In particular, as this estimator relies only on samples from the posterior distribution (area of high likelihoods) it tends to overestimate the evidence.

Another numerical approach to evidence estimation is importance sampling. The basic idea of importance sampling is to generate samples from an importance distribution instead of drawing them from the prior distributions as done by the BFMC method. The aim of importance sampling is to focus the sampling to the regions of the distribution that are of most "importance". The resulting bias in the sampling is corrected by attributing a weight to each sample. However, the choice of the importance distribution is not straightforward and it influences the efficiency and robustness of the evidence estimates (Perrakis et al., 2014). Gaussian mixture importance sampling (GMIS) (Volpi et al., 2017) uses as importance distribution an optimal mixture of normal distributions that fit the posterior distribution. The evidence is estimated as a weighted average based on N samples drawn from this importance distribution:

$$p_{\mathrm{GMIS}}(\widetilde{\mathbf{Y}}|\eta) \approx \frac{1}{N}\sum_{r=1}^{N} \frac{p(\theta_r^{\mathrm{imp}}|\eta)L(\theta_r^{\mathrm{imp}},\eta|\widetilde{\mathbf{Y}})}{q(\theta_r^{\mathrm{imp}})}, \tag{1.10}$$

where $q(\theta_r^{\mathrm{imp}})$ are the importance probability pdf. The GMIS method addresses the problem concerning the choice of an appropriate importance distribution in importance sampling at the expense of an increased computational time. The GMIS works well even in presence of complex multimodal distributions.

A recent approach to evidence estimation that is not based on MCMC algorithms is nested sampling that considers $p(\theta)d\theta$ in Equation 1.3 as equal to the element of prior mass, dX. In this method, a transformation is made from θ to the prior mass X, thereby, reducing Equation 1.3 into a one-dimensional integral over unit range in the likelihood space:

$$p(\widetilde{\mathbf{Y}}|\eta) = \int_0^1 L(X)dX. \tag{1.11}$$

The estimation procedure consists in drawing $i = 1,\dots,N$ samples from the prior distribution under the constraint of a lower bound of the log-likelihood function that increases with time. The one-dimensional integral in Equation 1.11 is then approximated by the weighted mean $w_i L(X_i)$ where the weights are defined as $w_i = X_i - X_{i+1}$. Nested sampling is well suited for high-dimensional parameter spaces and complex multimodal distributions. However, the exploration of the parameter space based on the likelihood constraint imposed by nested sampling is less efficient than MCMC methods based on the Metropolis-Hastings rule (Section 1.5, Equation 1.20).

Recent studies in hydrology suggest that nested sampling is less accurate and stable than thermodynamic integration (Liu et al., 2016; Zeng et al., 2018). Thermodynamic integration (path sampling) and stepping-stone sampling are based on sampling from a sequence of so-called power posterior distributions, $p_\beta(\theta|\widetilde{\mathbf{Y}})$, that create a path in the probability density space connecting the prior to the posterior distribution:

$$p_\beta(\theta|\widetilde{\mathbf{Y}}) \propto p(\theta)L(\theta|\widetilde{\mathbf{Y}})^\beta. \tag{1.12}$$

The power coefficient, β, varies between 0 and 1. For β=0, the prior distribution is sampled and for β=1, the posterior distribution is sampled. The normalising constant of Equation 1.12

is:

$$p(\widetilde{\mathbf{Y}}|\eta,\beta) = \int p(\theta)L(\theta|\widetilde{\mathbf{Y}})^{\beta}d\theta. \tag{1.13}$$

Assuming a proper prior, Equation 1.13 evaluated at β=0 is the integral of the prior distribution and it is equal to 1. Equation 1.13 evaluated at β=1 correspond to the evidence. As a consequence, the thermodynamic integration approach estimates the log-evidence as the integral of the expectations of the log-likelihoods over the interval [0,1] with respect to β:

$$\log p(\widetilde{\mathbf{Y}}|\eta) = \frac{p(\widetilde{\mathbf{Y}}|\eta,\beta=1)}{p(\widetilde{\mathbf{Y}}|\eta,\beta=0)} = \int_0^1 E_{\theta|\widetilde{\mathbf{Y}},\beta}\left[\log L(\theta|\widetilde{\mathbf{Y}},\eta)\right]d\beta. \tag{1.14}$$

The one-dimensional integral in Equation 1.14 is then approximated by quadature rule (e.g., composite trapezoidal rule). Stepping-stone sampling is based on a different idea, that is, using importance sampling to acurately approximate the ratio in Equation 1.14. In this context the evidence is estimated as:

$$p(\widetilde{\mathbf{Y}}|\eta) = \prod_{k=2}^{K}\frac{1}{N}\sum_{j=1}^{N} L(\theta_{k-1,j}|\widetilde{\mathbf{Y}})^{(\beta_k-\beta_{k-1})}, \tag{1.15}$$

where K is the number of the power coefficients β. The evidence estimators based on power posteriors are quite easy to implement. However, their accuracy is strongly influenced by the discretisation scheme used for the β values. The general idea is to place most of them close to zero where the log-likelihood increases the most.

1.4 Petrophysical models

In hydrogeophysics, the effective use of geophysical data for hydrogeological investigations is strongly linked to the reliability of the underlying relationship between estimated geophysical attributes (e.g., permittivity, electrical conductivity, bulk density) and the hydrogeological properties and states variables of interest (e.g., porosity, hydraulic conductivity, water content). The definition of a proper petrophysical relationship is one of the main challenges in hydrogeophysics (Binley et al., 2015) because they are often non-unique and non-stationary, that is, their parameter values and their analytical form can vary drastically between different types of lithologies (Hubbard and Rubin, 2005). Indeed, petrophysical relationships are often treated as site-specific because they depend on the geological structures of the subsurface and this implies that corresponding parameter values need to be calibrated for each site under study. Many relationships have been explored for hydrogeological studies (Mavko et al., 1998; Lesmes and Friedman, 2005; Pride, 2005). The petrophysical relationships can be physically or empirically-based (Linde et al., 2006b). In this thesis, we link GPR data to porosity and hydraulic conductivity properties using a physically and empirically based petrophysical relationship, respectively. The physically based relationship is built using volume-averaging to link the effective relative permittivities, $\boldsymbol{\varepsilon}$ [-], to porosity values, $\boldsymbol{\Phi}$ [-], and to radar slownesses, **s** [s/m], (Pride, 1994):

$$\begin{cases} \boldsymbol{\varepsilon} = \boldsymbol{\Phi}^{m}\,[\boldsymbol{\varepsilon_w} + (\boldsymbol{\Phi}^{-m} - 1)\varepsilon_{\mathrm{s}}] \\ \boldsymbol{\varepsilon} = \mathbf{s}^{2} c^{2} \end{cases} \tag{1.16}$$

where ε_{w} [-] and ε_{s} [-] are the relative permittivities of water and mineral grains, respectively; m [-] is the cementation index and $c = 3 \cdot 10^{8}$ [m/s] is the speed of light in vacuum. The slowness, **s**, is defined as the inverse of the velocity, **v**.

Combining the equations in 1.16, the petrophysical relationship reduces to:

$$\mathbf{v} = \sqrt{\boldsymbol{\Phi}^{-m} c^{2} [\varepsilon_{\mathrm{w}} + (\boldsymbol{\Phi}^{-m} - 1)\varepsilon_{\mathrm{s}}]^{-1}} \tag{1.17}$$

The empirically-based relationships are obtained by fitting polynomial functions. In our case, we test linear and quadratic petrophysical relationships to link the GPR velocities, **v** [m/s], to the natural logarithm of the hydraulic conductivities, $\mathscr{K} = \log \mathbf{K}$ [log(m/h)]:

$$\mathbf{v} = a_0 + a_1 \mathscr{K} \tag{1.18}$$

$$\mathbf{v} = a_0 + a_1 \mathscr{K} + a_2 \mathscr{K}^{2} \tag{1.19}$$

where a_0, a_1 and a_2 are the polynomial coefficients.

The petrophysical model consist in the definition of a functional form for the petrophysical relationship and the parameter values needed to describe such relationship. The petrophysical parameter values (e.g., m and ε_s in Equation 1.17 and a_0, a_1 and a_2 in Equations 1.18-1.19) may be inferred within the Bayesian inversion.

1.5 Markov chain Monte Carlo

How can we evaluate the posterior distribution of Equation 1.1? In most applications, the posterior distribution cannot be analytically estimated and sampling schemes are needed to numerically approximate it. The MCMC method (Gilks et al., 1995) provides a means to sample high-dimensional and very complicated posterior distributions by combining random sampling (Monte Carlo integration) with a "clever" search within the parameter space by building Markov chains. The resulting Markov chain is a sequence of random variables, $\{\theta_0, \theta_1, \theta_2, ...\}$, that are drawn from the model parameter space proportionally to the posterior distribution such that, at each iteration t, the probability of θ_{t+1} depends only on the value of θ_t. This lack of memory is the Markov property. The posterior distribution is approximated by the stationary distribution of the Markov chain, that is, the chain will gradually "forget" its initial state and converge to a unique and stationary distribution (i.e., which does not change with t). This property is ensured by the fulfilment of the ergodicity and

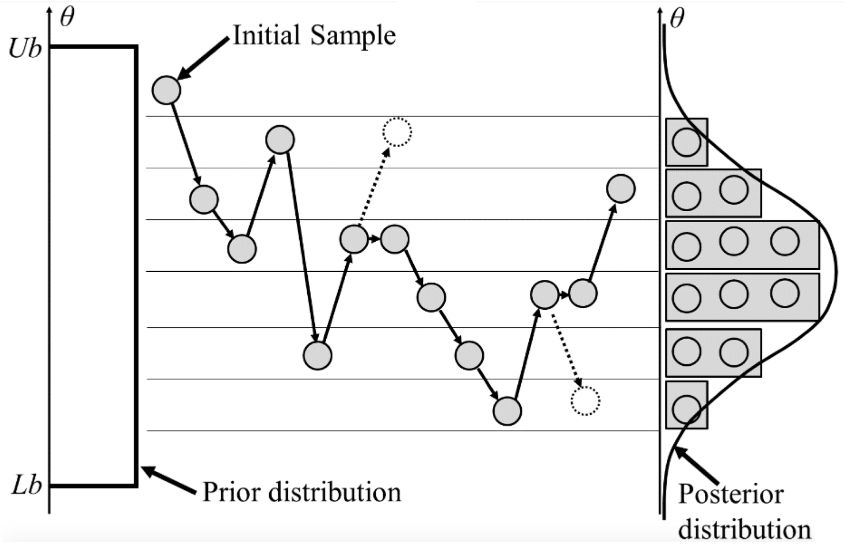


Figure 1.3 – Simplified representation of a Markov chain for a model parameter, θ, drawn from an uniform prior distribution $\mathcal{U}(L_b, U_b)$. The circles indicate the states of the chain and the arrows indicate the moves from one state to another. The dotted arrows and circles represent the proposed moves and samples that were not accepted based on the acceptance criterion, α (Figure modified from Lee et al. (2015)).

detailed balance (reversibility) conditions. The states of the chain drawn from this stationary distribution are samples of the posterior distribution, $p(\theta|\widetilde{\mathbf{Y}}, \eta)$.

Most MCMC algorithms perform the following steps to build a Markov chain (Figure 1.3):

1. Start the chain at an initial position in the model parameter space, θ_0, and define the desired (possibly large) number of posterior samples, T, that we want to obtain
2. Set iteration $t = 1$
3. Set $\theta_{\text{curr}} = \theta_{\boldsymbol{t-1}}$
4. Propose a move in the model parameter space based on a proposal distribution that generates θ_{prop}
5. Randomly draw a value u from the uniform distribution between 0 and 1, $\mathcal{U}(0,1)$
6. Accept or reject θ_{prop} based on an acceptance criterion, α: if $u < \min\{1, \alpha\}$, then θ_{prop} is accepted and the chain moves to the new position, $\theta_{\text{t}} = \theta_{\text{prop}}$; otherwise, θ_{prop} is rejected and the chain does not move, $\theta_{\text{t}} = \theta_{\text{curr}}$
7. Set $t = t + 1$
8. If $t < T$ return to step 3; otherwise, stop the chain

The Markov chains can be constructed based on different acceptance criteria (Step 6). For the popular Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953), we have:

$$\alpha = \min\left\{1, \frac{L(\theta_{\text{prop}}, \eta|\widetilde{\mathbf{Y}})\, p(\theta_{\text{prop}}|\eta)\, Q(\theta_{\text{curr}}|\theta_{\text{prop}})}{L(\theta_{\text{curr}}, \eta|\widetilde{\mathbf{Y}})\, p(\theta_{\text{curr}}|\eta)\, Q(\theta_{\text{prop}}|\theta_{\text{curr}})}\right\}. \tag{1.20}$$

The proposal distribution $Q(\cdot|\cdot)$ generates, at each iteration, the model perturbation of Step 4. The proposal distribution can have different forms and may be asymmetric but they are often defined as multivariate normal distributions. The choice of the proposal distribution and its scale (e.g, standard deviation in a multivariate normal proposal distribution) is critical. A proposal distribution that generates too small moves, $\theta_{\text{prop}} - \theta_{\text{curr}}$, will lead to a high acceptance rate and very slow mixing (i.e., moving around the model parameter space). A proposal distribution that generates large steps will often result in a too low acceptance rate and slow mixing. A number of algorithms has been proposed in the literature to allow for an automatic tuning of the scale of the proposal distribution, such as, the DREAM family of algorithms (Vrugt, 2016).

In this thesis, we will focus on two special cases of the Metropolis-Hastings algorithm: the Metropolis rule (Metropolis et al., 1953) and the Extended Metropolis rule (Mosegaard and Tarantola, 1995; Hansen et al., 2012). The Metropolis rule is a Metropolis-Hastings algorithm that considers symmetric proposal distributions, that is $Q(\theta_{\text{curr}}|\theta_{\text{prop}}) = Q(\theta_{\text{prop}}|\theta_{\text{curr}})$. According to the Metropolis rule, the proposed sample θ_{prop} is accepted with probability:

$$\alpha = \min\left\{1, \frac{L(\theta_{\text{prop}}, \eta|\widetilde{\mathbf{Y}})\,p(\theta_{\text{prop}}|\eta)}{L(\theta_{\text{curr}}, \eta|\widetilde{\mathbf{Y}})\,p(\theta_{\text{curr}}|\eta)}\right\}. \tag{1.21}$$

Both the Metropolis-Hastings and the Metropolis algorithms require the evaluation of the prior pdf at each iteration. However, in the case of conceptual models and model parameters generated from complex prior information (e.g., geologically-realistic patterns) based on MPS (Section 1.2), the prior density of a given model proposal is not described by a formula and, therefore, cannot be quantified. For solving this issue, Mosegaard and Tarantola (1995) suggest a proposal distribution chosen to simulate directly the prior distribution and detailed balance is fulfilled if:

$$\frac{Q(\theta_{\text{curr}}|\theta_{\text{prop}})}{Q(\theta_{\text{prop}}|\theta_{\text{curr}})} = \frac{p(\theta_{\text{curr}})}{p(\theta_{\text{prop}})}. \tag{1.22}$$

This equation reduces the Metropolis rule to a simple likelihood ratio and the corresponding method is in geophysics often referred to as the Extended Metropolis rule:

$$\alpha = \min\left\{1, \frac{L(\theta_{\text{prop}}, \eta|\widetilde{\mathbf{Y}})}{L(\theta_{\text{curr}}, \eta|\widetilde{\mathbf{Y}})}\right\}. \tag{1.23}$$

Different aspects need to be considered when running MCMC algorithms, such as, how many chains to run or how to assess convergence to the stationary distribution. Whenever possible, running multiple chains is preferred rather then running one single chain. Using multiple chains allows to better explore the parameter space, to avoid being stuck in certain areas of the parameter space and to better detect a lack of convergence. A lack of convergence of the Markov chains to the stationary distribution can be assed by (i) visual inspection of plots (i.e., the parameter value as a function of number of iterations), (ii) high autocorrelation between the states of the chain indicating slow mixing and, therefore, slow convergence, (iii) acceptance rates that are too high or too low (Gelman et al., 1996), (iv) quantitative diagnostics such as the Gelman-Rubin statistic for multiple chains (Gelman and Rubin, 1992) or the Geweke method for one chain (Geweke, 1992).

The number of iterations needed before drawing a representative sample of the stationary distribution defines the so-called burn-in period. The samples in the burn-in are discarded before approximating the posterior distribution. As soon as a sufficiently large number of posterior samples is collected, in Bayesian inference, it is common to visualise a few individual posterior realisations and to summarise the posterior distribution in terms of means, standard deviations and credible intervals.

It should be pointed out that the MCMC method is not only used for the evaluation of the posterior distribution, but it is also an important component of the techniques employed herein for evidence estimation (Section 1.3.3).

1.6 Objectives

Geophysics has contributed to important advances in hydrological sciences in the last 20 years (National Research Council, 2012). However, numerous challenges need to be addressed to take full advantage of the potential offered by hydrogeophysical studies (National Research Council, 2012; Hubbard and Rubin, 2002; Binley et al., 2015, 2010; Linde, 2014). Bayesian model selection relying on evidence computation and Bayes factors provides a valuable tool to account for conceptual uncertainty in hydrogeological systems and, therefore, to inform and increase the reliability of subsurface modelling and management. In this thesis, we will investigate the use of Bayesian model selection in hydrogeophysics and hydrogeology by answering the following research questions:

1. Are geophysical data suitable for guiding model selection in hydrogeology (Chapter 2)?
2. Can petrophysical uncertainty, including its spatial structure, be inferred in hydrogeophysical studies and how does it impact Bayesian inversion and model selection (Chapter 3)?
3. How can we achieve model selection when targeting geologically-realistic hydrogeological conceptual models represented by training images (Chapter 4)?

Even if conceptual uncertainty is often the predominant source of uncertainty in model predictions (Section 1.1,1.3), most hydrogeophysical and hydrogeological studies ignore it. One approach to account for conceptual uncertainty is to implement Bayesian model selection based on evidence estimation and Bayes factors (Section 1.3.2,1.3.3). The first objective is to explore to which extent geophysical data can effectively be used to discriminate among alternative hydrogeological conceptual models through Bayesian model selection. For this purpose, we will explore different approaches to estimate the evidence in hydrogeophysical settings (Chapter 2, Appendix A).

A common goal in hydrogeophysics is to infer quantitative hydrogeological models from geophysical data. This implies the use of petrophysical relationships (Section 1.4) that are often uncertain and poorly known. The uncertainty associated with these relationships need to be accounted for in hydrogeophysical inversions in order to properly assess the capability of geophysical data to provide reliable information about hydrogeological properties. Therefore, the second objective is to investigate the possibility to infer spatially-correlated uncertainty associated with petrophysical relationships and how this source of uncertainty impacts hydrogeological parameters and Bayes factors (Chapter 3).

Geologically-realistic conceptual models are often essential for reliable subsurface system studies. Such conceptual models can be built from training images and prior realisations of them can be sampled using concepts from multiple point statistics (Section 1.2). When considering conceptual models in the form of training images, the prior pdf cannot be computed as it does not have a parametric form and, therefore, many MCMC-based methods for evidence estimation cannot be used for Bayesian model selection. Hence, the third objective is to propose a methodology for Bayesian model selection among conceptual models built from complex spatial priors (training images) using concepts from MPS (Chapter 4).

These objectives will be addressed using a Bayesian approach (Section 1.1) to uncertainty quantification, parameter inference and model selection and the work rely on Markov chain Monte Carlo algorithms (Section 1.5). The objectives will be explored in light of synthetic and field case studies with the purpose of characterising spatially-distributed porosity or hydraulic conductivity fields in aquifers. We will compare various evidence computation methods applied to very different conceptualisation of the subsurface hydrogeological heterogeneity.

1.7 Outline

We present work that has been published in peer-reviewed journals (Chapters 2-3) or will soon be submitted (Chapter 4).

Chapter 2 presents a first comparative study of Bayesian hydrogeophysical model selection in the context of a synthetic example and a real case study of aquifer characterisation at the South Oyster Bacterial Transport site, Virginia (USA). We compare the evidence estimates provided by three methods: Brute-force Monte Carlo, Laplace-Metropolis and Gaussian-mixture importance sampling. The case-studies considered focus on the estimation of the spatial porosity distribution using first-arrival travel time data from crosshole GPR.

Chapter 3 proposes a methodology to account for and infer the spatially-correlated petrophysical prediction uncertainty in hydrogeophysical inversion and model selection. Results in the context of synthetic examples and a real case study of aquifer characterisation at the South Oyster Bacterial Transport site are presented. The case-studies considered in this chapter focus on porosity and hydraulic conductivity estimation using first-arrival travel time data from crosshole GPR.

Chapter 4 proposes a new full Bayesian methodology for performing Bayesian model selection among conceptual hydraulic conductivity models with high geological realism that are represented by training images. A comparison is made between different published conceptual models of the heterogeneous alluvial aquifer at the Macrodispersion Experiment (MADE) site, Mississippi (USA). We consider a small-scale tracer test (MADE-5) and its multilevel solute concentration data.

Chapter 5 concludes with a summary of this thesis, some remarks on current limitations and an outlook. Further details on the settings used in each chapter are listed in Table 1.2.

Appendix A consists in a report that will not be published but that brings worthwhile results that motivate our choice of not using nested sampling for evidence computations. The study in Appendix A explores the potential of the POLYCHORD (PC) algorithm to provide reliable evidence estimates based on nested sampling as compared to LM and GMIS methods. A synthetic example and a real case study of aquifer characterisation at the South Oyster Bacterial Transport site, Virginia (USA) are considered. First-arrival travel time data from crosshole GPR are used to infer the spatial porosity distribution.

Table 1.2 – Overview of technical details used in the different chapters and in the appendix of the thesis, such as, the type and number of data; the type and number of unknown parameters; the size of the grid cell (resolution) and the model size; the type of spatial parameterisation as described in Section 1.2; the method used to estimate the evidence as explained in Section 1.3.3.

	Chapter 2	**Chapter 3**	**Chapter 4**	**Appendix A**
Data type	Geophysical	Geophysical, Hydrogeological	Hydrogeological	Geophysical
N° data	100 up to 3248	100 up to 936	266	100 up to 3248
Parameter of interest	porosity	porosity, hydraulic conductivity	hydraulic conductivity	porosity
N° unknown parameters	16 up to 105	102 up to 211	MPS	103
Resolution [m]	0.04	0.04	0.1	0.04
Model domain [m]	7.2 × 7.2	7.2 × 7.2	8.1 × 6.3	7.2 × 7.2
Type of spatial parameterisation	zonation, variogram	zonation	variogram	MPS
Method for evidence estimation	BFMC, LM, GMIS	GMIS	TH, SS	LM, GMIS, PC

Chapter 2

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

Carlotta Brunetti, Jasper A. Vrugt and Niklas Linde.

Published[1] in *Advances in Water Resources*

[1] Brunetti, C., Linde, N. and Vrugt, J. A., (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

2.1 Abstract

Geophysical data can help to discriminate among multiple competing subsurface hypotheses (conceptual models). Here, we explore the merits of Bayesian model selection in hydrogeophysics using crosshole ground-penetrating radar data from the South Oyster Bacterial Transport Site in Virginia, USA. Implementation of Bayesian model selection requires computation of the marginal likelihood of the measured data, or evidence, for each conceptual model being used. In this paper, we compare three different evidence estimators, including (1) the brute force Monte Carlo method, (2) the Laplace-Metropolis method, and (3) the numerical integration method proposed by Volpi et al. (2017). The three types of subsurface models that we consider differ in their treatment of the porosity distribution and use (a) horizontal layering with fixed layer thicknesses, (b) vertical layering with fixed layer thicknesses and (c) a multi-Gaussian field. Our results demonstrate that all three estimators provide equivalent results in low parameter dimensions, yet in higher dimensions the brute force Monte Carlo method is inefficient. The isotropic multi-Gaussian model is most supported by the travel time data with Bayes factors that are larger than 10^{100} compared to conceptual models that assume horizontal or vertical layering of the porosity field.

2.2 Introduction

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

Many different approaches have been suggested in the statistical literature to help select the "best" model among a group of competing hypotheses. This includes frequentist and Bayesian solutions. The application of such approaches to geophysical studies has its own special challenges. For instance, a parameter-rich, but geologically-unrealistic model may fit the data equally well or perhaps even better than a more parsimonious model with more appropriate conceptualization of the subsurface (Rosenkrantz, 1977). What is more, the decision about which model is favoured, is also heavily influenced by the choice of the models' prior parameter distribution, even for geophysical data comprised of many different measurements. With the use of an inappropriate prior the model can be made to fit the data arbitrarily poorly, changing fundamentally our opinion about which model should be favoured, a phenomenon known as the Jeffreys-Lindley paradox (Jeffreys, 1939; Lindley, 1957).

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

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

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

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

The purpose of this study is twofold. In the first place, we investigate to what extent evidence estimates and Bayes factors derived for moderately high parameter dimensionalities (i.e., up to 105 unknowns) can be used to perform Bayesian model selection in synthetic and real-world case studies. For this purpose, we compare evidence estimates computed by (1) the brute force Monte Carlo method (Hammersley and Handscomb, 1964), (2) the Laplace-Metropolis method (Lewis and Raftery, 1997) and (3) the Gaussian mixture importance sampling (GMIS) estimator of Volpi et al. (2017). This latter method approximates the evidence by importance sampling from a Gaussian mixture model fitted to a large sample of posterior solutions generated with the DREAM$_{(ZS)}$ algorithm (Vrugt, 2016; Vrugt et al., 2008; Laloy and Vrugt, 2012). Then, we present an application of Bayesian model selection to subsurface modeling using geophysical data from the South Oyster Bacterial Transport Site in

Virginia (USA) (Chen et al., 2001, 2004; Hubbard et al., 2001; Linde et al., 2008; Linde and Vrugt, 2013). These data consist of travel time observations made by crosshole ground-penetrating radar (GPR), and exhibit small measurement errors typical of most near-surface geophysical sensing methods.

2.3 Theory and Methods

2.3.1 Bayesian inference with MCMC

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

$$p(\theta|\widetilde{\mathbf{Y}}) = \frac{p(\theta)p(\widetilde{\mathbf{Y}}|\theta)}{p(\widetilde{\mathbf{Y}})}, \tag{2.1}$$

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

$$L(\theta|\widetilde{\mathbf{Y}}, \sigma_{\widetilde{\mathbf{Y}}}) = \left(\sqrt{2\pi\sigma^2_{\widetilde{\mathbf{Y}}}}\right)^{-n} \exp\left[-\frac{1}{2}\sum_{h=1}^{n}\left(\frac{\mathscr{F}_h(\theta) - \widetilde{y}_h}{\sigma_{\widetilde{\mathbf{Y}}}}\right)^2\right], \tag{2.2}$$

where $\sigma_{\widetilde{\mathbf{Y}}}$ denotes the standard deviation of the measurement data error. This entity can be fixed a-priori by the user if deemed appropriate, or alternatively, the measurement data error can be treated as nuisance variable and the value of $\sigma_{\widetilde{\mathbf{Y}}}$ is inferred jointly with the d-vector of model parameters, θ. The Gaussian likelihood function of Eq. (2.2) has found widespread application and use in the field of geophysics, nevertheless it is important to stress that the error residuals hardly ever satisfy the rather restrictive assumptions of normality, constant variance, and lack of serial correlation. The Gaussian likelihood in Eq.(2.2) is sufficient, though, to illustrate the power and usefulness of Bayesian model selection.

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

2.3.2 Evidence and Bayes factor

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

$$B_{(\eta_1,\eta_2)} = \frac{p(\widetilde{\mathbf{Y}}|\eta_1)}{p(\widetilde{\mathbf{Y}}|\eta_2)}, \tag{2.3}$$

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

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

$$p(\widetilde{\mathbf{Y}}|\eta_l) = \int L(\theta_l,\eta_l|\widetilde{\mathbf{Y}})p(\theta_l|\eta_l)d\theta_l \qquad l = 1,2. \tag{2.4}$$

In practice, it is often not necessary to integrate over the entire prior distribution, as large portions of this space are made up of areas with a negligible posterior density whose contributions to the integral of Eq. (2.4) are negligibly small. Instead, we can restrict our attention to those areas of the parameter space occupied by the posterior distribution.

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

The Bayes factor is a sufficient statistic for hypothesis testing, yet renders necessary the definition of "formal" guidelines on how to interpret its value before we can proceed with model selection. Table 2.1 articulates an interpretation of the Bayes factor as advocated by Kass and Raftery (1995). This interpretation differentiates four (increasing) levels of support for proposition η_1 relative to η_2. In general, the evidence in favor of η_1 increases with the value of its Bayes factor. Thus, the larger the value of $B_{(\eta_1,\eta_2)}$, the more the data $\widetilde{\mathbf{Y}}$ supports the hypothesis η_1 relative to η_2, and the easier it becomes to reject this alternative hypothesis. It is suggested that the Bayes factor must be larger than 3 (or smaller than 1/3) to discriminate positively among two competing hypotheses.

Table 2.1 – Interpretation of Kass and Raftery (1995) for the Bayes factor of two conceptual models η_1 and η_2.

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

Unfortunately, the integral in Eq. (2.4) cannot be derived by analytic means nor by analytic approximation, and we therefore resort to numerical methods to calculate the evidence of each conceptual model. In the next section, we review briefly three different methods for estimating the evidence, including the brute force Monte Carlo method (BFMC), the Laplace-Metropolis (LM) method and the Gaussian mixture importance sampling (GMIS) approach recently developed by Volpi et al. (2017).

Brute force Monte Carlo method

The BFMC method (Hammersley and Handscomb, 1964) approximates the evidence in Eq. (2.4) as an average of the likelihoods of N different samples drawn randomly from the (multivariate) prior distribution (Kass and Raftery, 1995)

$$p_{\mathrm{BFMC}}(\widetilde{\mathbf{Y}}) \approx \frac{1}{N}\sum_{i=1}^{N} L(\theta_i|\widetilde{\mathbf{Y}}). \tag{2.5}$$

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

Laplace-Metropolis method

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

$$p_{\mathrm{LM}}(\widetilde{\mathbf{Y}}) \approx (2\pi)^{d/2}|\mathbf{H}(\theta^*)|^{1/2}p(\theta^*)L(\theta^*|\widetilde{\mathbf{Y}}), \tag{2.6}$$

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

$$\mathbf{H}(\theta^*) = \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j - \theta^*)^T(\theta_j - \theta^*). \tag{2.7}$$

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

The KIC (Kashyap, 1982)

$$\mathrm{KIC}_{\theta^*} = -2\log(p_{\mathrm{LM}}(\widetilde{\mathbf{Y}})) \tag{2.8}$$

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

Gaussian mixture importance sampling

As third and last method we consider the GMIS evidence estimator developed recently by Volpi et al. (2017). This method uses multidimensional numerical integration of the posterior parameter distribution via bridge sampling (a generalization of importance sampling) of a mixture distribution fitted to samples of the target derived from MCMC simulation with the DREAM algorithm (Vrugt, 2016). This approach has elements in common with the BFMC method, yet draws samples directly from the posterior distribution, rather than the prior distribution (as in BFMC) to approximate the evidence. One would therefore expect a much higher sampling efficiency of the GMIS method. The use of a Gaussian mixture distribution allows GMIS to approximate as closely and consistently as possible the actual posterior target distribution. Indeed, this distribution can be multimodal, truncated, and "quasi-skewed" - properties that can be emulated with a mixture model if a sufficient number of normal components is used. The Expectation-Maximization (EM) algorithm is used to construct the Gaussian mixture distribution (Dempster et al., 1977; Hoogerheide et al., 2012). Let us assume that MCMC simulation with DREAM has produced J realizations, $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_J\}$, of the d-variate posterior parameter distribution under hypothesis, η_1. We approximate these samples' probability density function, $p(\boldsymbol{\theta}|\widetilde{\mathbf{Y}})$, with a mixture distribution

$$q(\boldsymbol{\theta},K) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{\theta};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \tag{2.9}$$

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

The EM algorithm (Dempster et al., 1977; Hoogerheide et al., 2012) is used to determine the values of the d_{mix}-variables of the mixture distribution, $\boldsymbol{\Phi} = \{\alpha_1,\dots,\alpha_K,\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_K\}$, where each $\boldsymbol{\beta}_k = \{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}$ characterizes the mean and covariance matrix of a different normal density of the mixture, and $k = \{1,\dots,K\}$. This algorithm maximizes the log-likelihood, $\log\{L(\boldsymbol{\Phi}|\boldsymbol{\Theta},K)\}$, of the mixture density

$$\log\{L(\boldsymbol{\Phi}|\boldsymbol{\Theta},K)\} = \sum_{j=1}^{J} \log\left\{\sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{\theta}_j;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}, \tag{2.10}$$

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

$$\text{BIC}(K) = -2\log\{L(\boldsymbol{\Phi}|\boldsymbol{\Theta},K)\} + d_{\text{mix}}(K)\log(J). \tag{2.11}$$

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

$$\widehat{K} = \arg\min_{K\in\mathbb{N}_+} \text{BIC}(K), \tag{2.12}$$

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

The optimal mixture distribution now serves as importance density, $q(\boldsymbol{\theta},\widehat{K})$, in GMIS to estimate the marginal likelihood, $p_{\text{GMIS}}(\widetilde{\mathbf{Y}})$. To this end, we draw at random from the importance distribution, $Q(\boldsymbol{\theta},\widehat{K})$, a total of N different samples, $\{\boldsymbol{\theta}_1^{\text{imp}},\dots,\boldsymbol{\theta}_N^{\text{imp}}\}$. We then evaluate each of these N parameter vectors in our hypothesis (conceptual model), and calculate their unnormalized posterior densities, $p(\boldsymbol{\theta}_r^{\text{imp}})L(\boldsymbol{\theta}_r^{\text{imp}}|\widetilde{\mathbf{Y}})$, where $r = \{1,\dots,N\}$. The evidence, $p_{\text{GMIS}}(\widetilde{\mathbf{Y}})$, is now computed by GMIS as a weighted mean of the ratios of the samples' unnormalized posterior densities and corresponding importance densities (Perrakis et al., 2014)

$$p_{\text{GMIS}}(\widetilde{\mathbf{Y}}) \approx \frac{1}{N}\sum_{r=1}^{N} \frac{p(\boldsymbol{\theta}_r^{\text{imp}})L(\boldsymbol{\theta}_r^{\text{imp}}|\widetilde{\mathbf{Y}})}{q(\boldsymbol{\theta}_r^{\text{imp}})}. \tag{2.13}$$

This concludes our description of the GMIS estimator. We refer interested readers to Volpi et al. (2017) for a more detailed treatment and explanation of the theory, concepts, and main building blocks of GMIS. This paper also documents a diverse set of case studies (up to $d = 100$) which evaluate and benchmark the performance of GMIS against other commonly used evidence estimation methods.

2.3.3 Evidence estimation in practice

The posterior distribution and the MAP solution that is used by the LM and GMIS methods (Section 2.3.2) are derived from MCMC simulation using the DREAM$_{(\text{ZS})}$ algorithm (Laloy and Vrugt, 2012; Vrugt, 2016; Vrugt et al., 2008). This multi-chain method creates proposals on the fly from an historical archive of past states using a mix of parallel direction and snooker updates. We refer the reader to Linde and Vrugt (2013); Lochbühler et al. (2014a, 2015); Rosas-Carbajal et al. (2013, 2015) for various geophysical case-studies in which this algorithm is used. For the actual field application, we use a hierarchical Bayesian formulation, in which the data error, $\sigma_{\tilde{\mathbf{Y}}}$ in Eq. (2.2) is jointly estimated with the model parameters (e.g., Rosas-Carbajal et al. (2013)). For numerical reasons we work with a log-likelihood formulation of Eq. (2.2). A total of four chains were deemed sufficient for 25 parameters, five chains were used for model dimensions between 26 and 64, and eight chains for models with more than 65 parameters. The number of generations varied between 200000 and 500000 depending on the dimensionality of the target distribution. The scaling factor, β_0 of the jump rate was tuned to achieve an adequate acceptance rate and the univariate $\widehat{R}$-diagnostic (Gelman and Rubin, 1992) was used to judge when convergence had been achieved of the DREAM$_{(\text{ZS})}$ algorithm to a limiting distribution.

We report the evidence estimates of the BFMC method using three different sample sizes involving $N = 10^5$, $N = 10^6$ and $N = 10^7$ samples in Eq. (2.5). In GMIS, we use a total of $N = 10^5$ importance samples (Eq. (2.13)). We repeat each of these two numerical experiments ten times, and summarize the mean evidence and associated range in the results section. Lastly, in the case of the LM method, we report the evidence computed as the mean of the estimates on the different Markov chains (Van Haasteren, 2013) together with the range.

2.3.4 Conceptual subsurface models

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

Petrophysics and forward modelling

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

$$s = \sqrt{\phi^m c^{-2}[\varepsilon_\mathrm{w} + (\phi^{-m} - 1)\varepsilon_\mathrm{s}]}, \tag{2.14}$$

where $\varepsilon_\mathrm{w} = 81$ (-) denotes the relative permittivity of water, $c = 3 \cdot 10^8$ (m/s) is the speed of light in a vacuum, ε_s (-) signifies the relative permittivity of the mineral grains and m is a unitless cementation index. We use the non-linear 2D travel time solver (*time 2d*) of Podvin and Lecomte (1991) to compute first-arrival travel times from slowness fields obtained by applying the petrophysical relationship of Eq. (2.14) to each porosity field.

2.4 Illustrative toy example

To benchmark the different evidence estimators of Section 2.3.2, we first consider an illustrative example involving a simple crosshole GPR experiment. A total of 10 transmitter and receiver antennas are placed at multiple different depths (uniform intervals) in boreholes located in the left and right side of the domain, respectively (see Fig. 2.1a). This results in a total of 100 different transmitter-receiver antenna pairs. The spatial domain that necessitates porosity characterization covers an area of 7.2 m × 7.2 m. To warrant accurate model simulations, a spatial discretization of 0.04 × 0.04 m is considered. We contaminate the $n = 100$ first-arrival travel time data with Gaussian white noise using a measurement error, $\sigma_{\tilde{\mathbf{Y}}} = 2$ ns. This comparatively high error level was chosen to facilitate comparison with the BFMC method, which is known to work better in the presence of large measurement errors. This leads to a likelihood function that is less peaked, and, consequently, a posterior distribution that is more dispersed as it will distribute more evenly the probability mass over the parameter space. The "true" porosity field of the subsurface is made up of four different layers of equal thickness with porosity values of 0.3, 0.45, 0.35 and 0.4, in the downward direction, respectively (see Fig. 2.1a). We varied the number of horizontal layers of constant thickness from $d = 1$ to $d = 16$, and assume a uniform prior distribution for the porosity, ϕ, of each respective layer using upper and lower bound values of 0.25 and 0.50, respectively. The petrophysical parameters of Eq. (2.14) are assumed fixed using values of $m = 1.5$ and $\varepsilon_s = 5$, respectively.

Figure 2.1b-e presents the posterior mean porosity field derived from the DREAM$_{(ZS)}$ algorithm for four different model conceptualizations. The two layer model (Fig. 2.1b) is an overly simplistic representation of the true porosity field which is, by construction, perfectly described by the conceptual model with four layers shown in Fig. 2.1c. The posterior mean porosity field of the six layers model presented in Fig. 2.1d exhibits a relatively poor agreement with the reference porosity field. Finally, the porosity values for the eight layer model (Fig. 2.1e) correspond rather closely with their counterparts of the reference field (Fig. 2.1a). The bottom panel, in Fig. 2.1f-i, display the posterior standard deviation of the porosity estimates for the different layers of our four model conceptualizations. As expected, the uncertainty of the porosity estimates increases with the number of layers that are used in our subsurface characterizations.

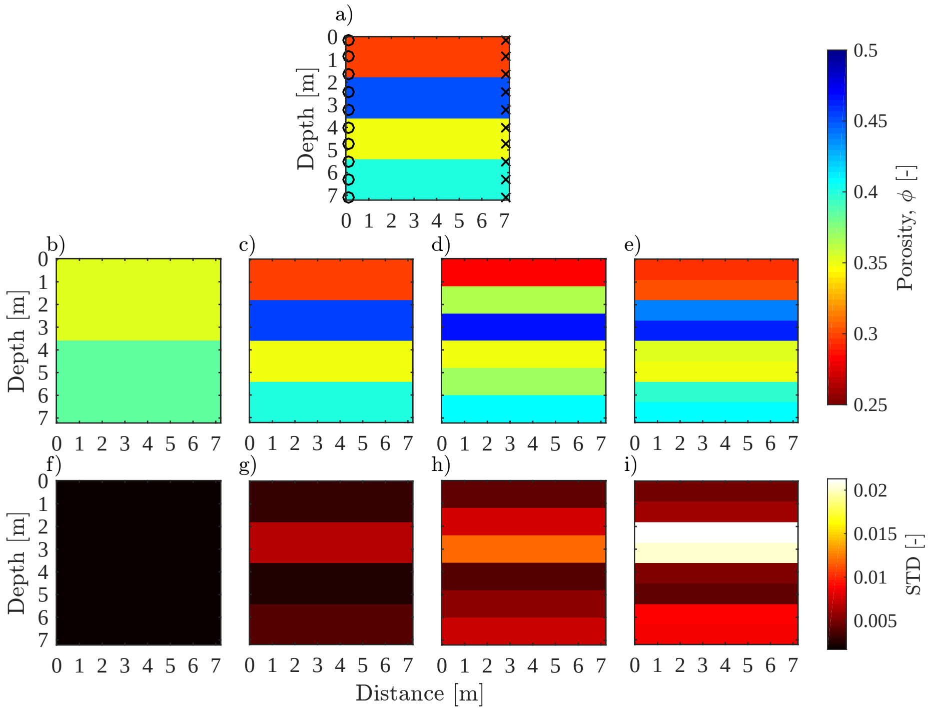


Figure 2.1 – a) The "true" subsurface porosity model used in our synthetic crosshole-GPR experiment. The different measurement depths of the transmitter antenna (black crosses) and receiver antenna (black circles) are separately indicated. Mean porosity fields of the posterior distribution derived from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using four different conceptualizations of the subsurface involving (b) two, (c) four, (d) six, and (e) eight horizontal layers. The corresponding posterior standard deviations of the porosity estimates for the four different conceptualizations of the subsurface are shown in (f), (g), (h) and (i), respectively.

Now we calculate the marginal likelihood of each hypothesis using the BFMC, LM, and GMIS estimators. The results of this analysis are presented in Fig. 2.2 using at the left hand-side a plot of the mean evidence computed by each method against model complexity, and at the right-hand-side a graph of the associated uncertainty of each estimator. We consider subsurface models with up to $d = 16$ horizontal porosity layers of equal thickness. To simplify graphical interpretation of the results, we plot $\log_{10}$ transformed values of the evidence, and refer to this entity as $\mathscr{P}(\widetilde{\mathbf{Y}})$. Colour coding is used to differentiate between the results of the three different methods. The results highlight several important findings. In the first place, the evidence estimates confirm that the model with four different porosity layers, that is $d = 4$, is most supported by the available data (Fig. 2.2a). This finding is not surprising as this model uses the exact same layering of the porosity field as used in the synthetic GPR experiment that was used to create the "measured" travel time data. Secondly, the BFMC (black), the LM (blue) and the GMIS (red) estimators are in excellent agreement and provide nearly identical values of the evidence for conceptual models with just a few parameters (horizontal layers)(Fig. 2.2a). Thirdly, the BFMC starts to deviate from the LM and GMIS methods at seven model dimensions and substantial differences appear for models with more than nine layers (Fig. 2.2a). This behavior is explained by the fact that the BFMC estimates did not converge for model dimensions higher than six. The convergence analysis was performed by a bootstrap analysis with 1000 bootstrap estimates (results not shown herein). In the fourth place, notice in Fig. 2.2b that the LM and GMIS estimators exhibit a negligible uncertainty compared to the range of evidence values considered and that the upper and lower bound values of the evidence derived from both methods appear rather similar. Evidence estimates derived from the BFMC method, on the contrary, exhibit a much larger uncertainty due to the fact that the BFMC does not reach convergence for model dimensions higher than six. This provides further support for the claim that, in our implementation and algorithmic settings, the BFMC method is inefficient when applied to models of high dimensionality since large numbers of samples (implying prohibitively large CPU-costs) are needed to properly characterize the likelihood surface and obtain reliable results.

We now investigate in more detail the discrepancies between the results of the three estimators, and plot in Fig. 2.3 the differences between the logarithmic values of the marginal likelihoods, $\mathscr{P}(\widetilde{\mathbf{Y}})$, computed by the methods for the competing models used in this study. The solid black line depicts the difference in the mean evidence estimates derived by comparing each pairs of methods, and the grey shaded region quantifies the range associated with the differences in evidence estimates (i.e., the upper and lower boundaries of the grey shaded region are, respectively, the maximum and minimum difference in evidence estimate computed by each pairs of methods). Note, we use $N = 10^7$ in the BFMC method and report results for subsurface models with number of horizontal porosity layers (equal thickness) that ranges from $d = 1$ to $d = 16$.

The results in Fig. 2.3 provide further evidence for our earlier conclusions. Indeed, the three methods provide rather similar evidence values (Fig. 2.3a) for the simpler subsurface models (i.e., up to $d = 6$ different porosity layers). For larger model complexities the LM and the GMIS estimators differ a bit from each other - but this difference is very small in comparison to their mean estimates. It is now evident that the difference in the evidence estimates derived from LM and GMIS increases with model complexity. Note that the maximum deviation between both methods is on the order of 0.7 unit in $\mathscr{P}(\widetilde{\mathbf{Y}})$ space, which, with mean estimates on the order of one-hundred (see Fig. 2.2a), equates to a difference smaller than 1%. However, it

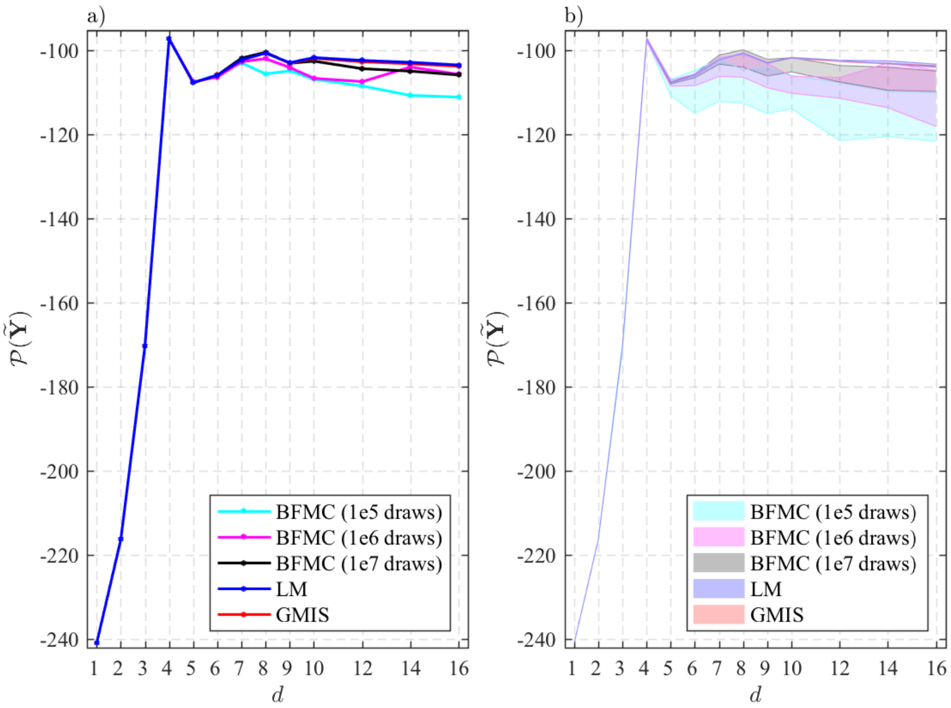


Figure 2.2 – Mean values of the evidence in $\log_{10}$ space, $\mathscr{P}(\widetilde{\mathbf{Y}})$ (a: left graph), and their associated uncertainty (b: right graph) derived from the BFMC, LM, and GMIS estimators for each model complexity, d used herein. Color coding is used to differentiate among the different methods. The evidence estimates of the LM and GMIS estimators are in excellent agreement and their uncertainty is negligibly small.

is important to stress here that there is no reason to expect that the two methods provide equivalent results since they are based on very different assumptions (details in Section 2.3.2). Results from Fig. 2.3 also confirm that the evidence values derived from the BFMC method start to deviate from the other two methods for model dimensions higher than six since the method does not reach convergence for those models (Fig. 2.3b-c). These differences grow as large as 6-7% in $\mathscr{P}(\widetilde{\mathbf{Y}})$ space for the most complex subsurface models with $d = 14$ and $d = 16$ porosity layers. It is worth noting that we are primarily interested in an accurate model ranking, while the accuracy of the evidence estimates themselves are of secondary importance. In light of this, we find that the differences in the evidence estimates obtained by the three different estimators do not have an impact on which models are ranked first and second in the presented synthetic example.

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

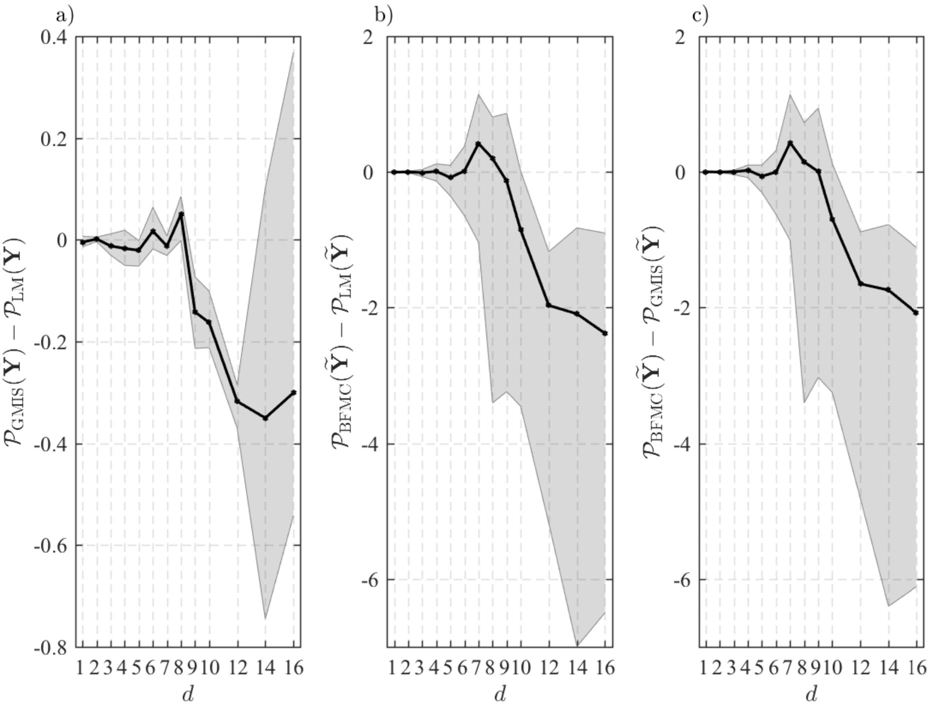


Figure 2.3 – Difference in the evidence estimates derived from different pairs of two methods as function of model complexity, (a) GMIS and LM, (b) BFMC and LM, and (c) BFMC and GMIS. The solid black line in each graph portrays the difference in the mean evidence estimates, and the grey shaded region quantifies the range associated with the difference in the mean evidence estimates of each method. Note, we use $\log_{10}$ transformed value of the evidence estimates.

and carry forward to the next case study the LM and GMIS estimators that are relatively CPU-efficient.

2.5 Field example

2.5.1 Field site and available data

We now focus our attention on the South Oyster Bacterial Transport Site in Virginia, USA, and use geophysical data measured at this experimental site to determine which model of the subsurface is preferred statistically. The geological characteristics of the South Oyster Bacterial Transport Site are described in Hubbard et al. (2001). GPR travel time data were measured at the S14-M13 borehole transect using a PulseEKKO 100 system with a 100-MHz nominal-frequency antenna. A domain of 7.2 × 7.2 m was measured with a total of 57 sources and 57 receivers, leading to a data set of 3248 observations of first-arrival travel times (one value is missing). We assume the measurement errors of the travel time to be uncorrelated and normally distributed with constant standard deviation, $\sigma_{\tilde{\mathbf{Y}}}$. A relatively fine spatial discretization consisting of square cells with length 0.04 m was used in our forward simulations with the non-linear 2D travel time solver (*time 2d*) of Podvin and Lecomte (1991) to compute the first-arrival travel times for the 7.2 × 7.2 m domain of interest. The models used in this study differ in their conceptual representation of the subsurface, and use horizontal and vertical layering of the porosity. The numbers of porosity layers (equal thickness) is varied between 1 to 60, thereby providing a large array of competing hypotheses. Table 2.2 lists the parameters of both spatial porosity configurations which are subject to inference with the DREAM$_{(ZS)}$ algorithm. This includes, the porosity, ϕ, of each individual layer, and the values of m, ε_s and $\sigma_{\tilde{\mathbf{Y}}}$. We list their symbol, unit, range, type of prior distribution, and respective number of unknowns.

Table 2.2 – Parameters of the conceptual subsurface models with horizontal and vertical porosity layering. The last three columns summarize the range, prior distribution, and number, of each parameter, respectively as used in our MCMC inversion with the DREAM$_{(ZS)}$ algorithm. The variable n_{layer} defines the number of layers that is used in each conceptual model.

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

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

The use of horizontal and vertical layering of the porosity is perhaps convenient computationally, but might not describe properly the subsurface of an actual field site. Indeed, the subsurface might exhibit much more complex porosity structure. We therefore augment the ensemble of hypotheses with a model that assumes a multi-Gaussian porosity field. This field is generated over a regular 2D grid of size 180 × 180 with geostatistical properties and spatial structure described with the Matérn variogram. Fortunately, the values of the integral scales in the x and z-direction, I_x and I_z, respectively, anisotropy angle, φ, and smoothness parameter, ν, of this variogram have been reported in the literature for the South Oyster Bacterial Transport Site (Chen et al., 2001; Hubbard et al., 2001). Their values are listed in the second column of Table 2.3, and assume horizontal anisotropy of the porosity field, that is $\varphi = 90°$. This model is referred to as MGha. For completeness, we also consider herein a multi-Gaussian model with vertical anisotropy, $\varphi = 0°$ (third column), coined MGva, and include an isotropic description of the porosity (fourth column), hereafter referred to as MGis, and enforced by setting I_x and I_z equal to the geometric mean of the integral scales of the first two multi-Gaussian models. We fix the value of $\nu = 0.5$ in the Matérn variogram, as we expect an exponential variogram model. Interested readers are referred to Laloy et al. (2015) for a more detailed description of the Matérn variogram.

We now focus our attention to the "unknown" parameters in each model (see Table 2.4), which are subject to inference using the observed travel time data. In our MCMC inversions we infer jointly the petrophysical parameters, ε_s and m of Eq. (2.14), mean porosity, $\overline{\phi}$, and

Table 2.3 – Integral scales in x- and z-direction, I_x and I_y, respectively, anisotropy angle, φ, and smoothness parameter, ν for the multi-Gaussian model with horizontal anisotropy (second column, MHha), vertical anisotropy (third column, MGva), and isotropy (last column, MGis).

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

its associated variance, v, the measurement data error, $\sigma_{\tilde{\mathbf{Y}}}$, of the travel time data, and 100 dimensionality reduction variables, **DR** (details in Section 2.3.4).

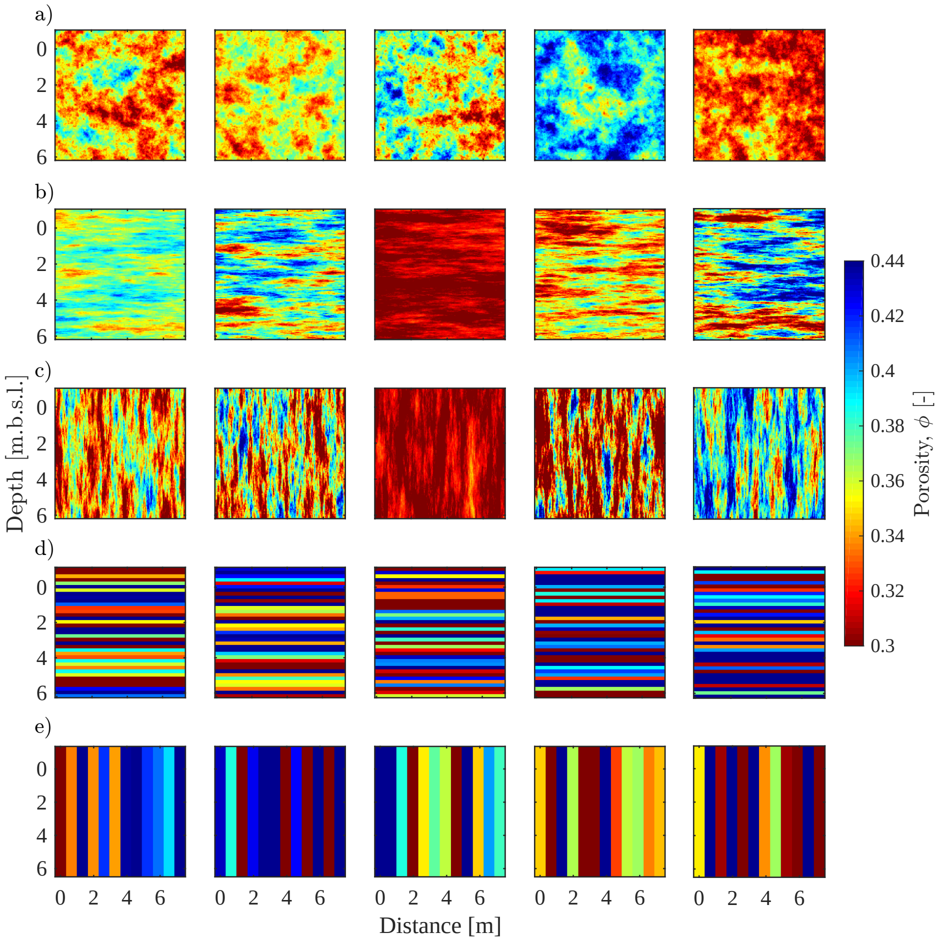
Table 2.4 – Parameters of multi-Gaussian models (first column) and their respective units (second column), range (third column), prior distribution (fourth column), and number (last column).

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

2.5.2 Results

We first display in Fig. 2.4 five realizations of the prior porosity field (columns) for each of the conceptual models (different rows) used in this case study. This includes the three multi-Gaussian models with (a) isotropy, (b) horizontal anisotropy, and (c) vertical anisotropy, and more simplistic models that assume (d) horizontal and (e) vertical layering of the porosity values. It is evident that these five model types provide very different descriptions of the porosity field of the subsurface at the experimental site. The multi-Gaussian models exhibit most spatial diversity with realizations that differ substantially in their mean porosity and associated variance. The porosity values of the layered models change abruptly from one depth to the next.

We now move on to our inversion results and present in Fig. 2.5 for each model of the ensemble (different rows), four different draws of the posterior distribution (first four columns), the posterior mean porosity field (fifth column) and the associated standard deviation (last column) derived from the DREAM$_{\text{(ZS)}}$ algorithm. The order of the presentation matches



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

Figure 2.4 –]

exactly Fig. 2.4, that is, the first three rows presents the results of the multi-Gaussian models with (a) isotropy, (b) horizontal anisotropy, and (c) vertical anisotropy of the porosity values, and the bottom two rows illustrate the results of the models with (d) horizontal and (e) vertical layering. The different conceptual models provide quite different characterizations of the porosity field. Some commonalities can be observed, though. For instance, the isotropic multi-Gaussian model, the multi-Gaussian model with horizontal anisotropy and the horizontally layered model (Fig. 2.5a-b-d) all depict the presence of a low-porosity zone just below the surface and at a depth of 4-5 m. They also demonstrate high-porosity zones at depths of 2 m and 6 m, and at 3 m below the ground surface a small high-porosity area is also visible, although this is not so evident for the isotropic multi-Gaussian model. The porosity fields parametrized by these three conceptual models are estimated with relatively low uncertainties (i.e., maximum of posterior standard deviations equals to or less than ±0.01), especially, in the case of the horizontal layering. Also, the conceptual subsurface model with vertically oriented porosity structures (i.e., the vertically layered model and the multi-Gaussian model with vertical anisotropy) exhibit more variation in their porosity values (first four columns in Fig. 2.5c-e) and characterized by larger uncertainties (last column in Fig. 2.5c-e) than the other models.

Note that the posterior mean porosity field of the multi-Gaussian model with horizontal anisotropy (fifth column in Fig.2.5b) is in good agreement with the velocity field obtained by Linde et al. (2008) and Linde and Vrugt (2013) for the exact same data set.

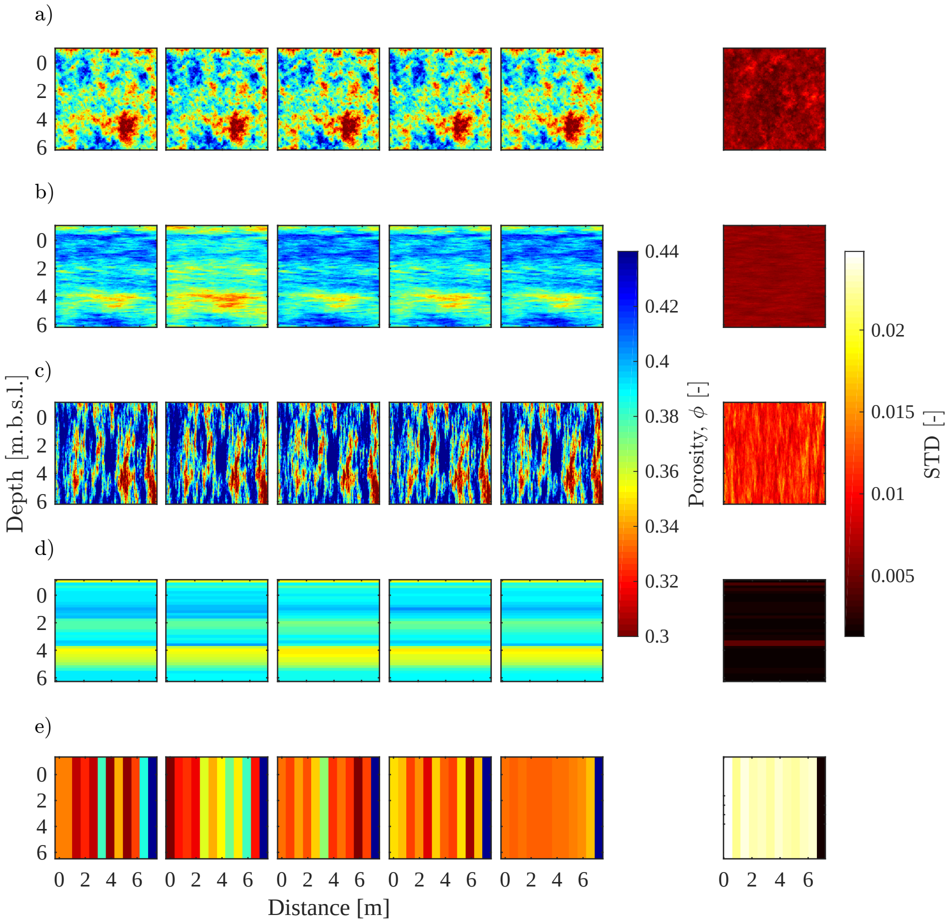


Figure 2.5 – Four realizations drawn randomly from the posterior distribution (first four columns), the posterior mean porosity field (fifth column) and the standard deviations of the posterior porosity estimates (last column) for the (a) isotropic multi-Gaussian model, (b) multi-Gaussian model with horizontal anisotropy, (c) multi-Gaussian model with vertical anisotropy, (d) horizontally layered model with 37 layers of equal thickness, and (e) vertically layered model with 12 layers of equal thickness.

To provide more insights into the posterior parameter distributions of each model, Fig. 2.6 plots histograms of the marginal distributions of the cementation index, m (first column), the relative permittivity of the mineral grains, ε_s (second column), and the inferred data error, $\sigma_{\tilde{\mathbf{Y}}}$ (third column) for the multi-Gaussian (top three rows) and layered (bottom two rows) subsurface models. The prior distribution is separately indicated in each plot with the red line. Note, to simplify graphical notation, the density of all the distributions was scaled to be between 0 and 1. This figure highlights several interesting findings. In the first place, notice that the three parameters appear to be well defined in each of the five conceptual models with posterior distributions that occupy only a small portion of their respective prior distributions. This is particularly true for the marginal distribution of $\sigma_{\tilde{\mathbf{Y}}}$, the measurement error of the travel time data. Secondly, notice that the use of a vertically layered porosity (Fig. 2.6e) results in truncated histograms of the parameters m and ε_{s} and a large inferred data error, $\sigma_{\tilde{\mathbf{Y}}} > 1.5$ ns. These are possible signs of model malfunctioning, a claim that we will investigate next by looking in detail at the evidence estimates of each model, but supported thus far by the much larger posterior values of $\sigma_{\tilde{\mathbf{Y}}}$ for the vertically layered model than the other four competing subsurface models. Thirdly, notice that the histograms of the petrophysical parameters m and ε_{s} differ quite substantially between the conceptual models. These parameters probably compensate in different ways for imperfections in each model's porosity structure. The histograms of the nuisance parameter $\sigma_{\tilde{\mathbf{Y}}}$ appear almost similar with the exception of the model with vertically layered porosity values. Altogether, the lowest value of the measurement data error, $\sigma_{\tilde{\mathbf{Y}}} = 0.457$ ns, is found for the isotropic multi-Gaussian model (Fig. 2.6a), which should suggest that this model most closely matches the observed travel time data.

We now turn our attention to the evidence of each model. Fig. 2.7 presents the results of this analysis using a $\log_{10}$ transformation of the evidence values. The left graph (Fig. 2.7a) displays the results for the three multi-Gaussian models with isotropy (circle), horiziontal anisotropy (square) and vertical anisotropy (triangles), respectively, using a single complexity involving $d = 105$ parameters. The graph in the middle (Fig. 2.7b) and on the right (Fig. 2.7c) depict the results for the conceptual models with horizontal and vertical layering, respectively, using between 1 to 60 different porosity layers. Colour coding is used in all the three plots to differentiate between the LM (blue) and GMIS (red) estimators. The vertical bars in Fig. 2.7a and shaded regions in Fig. 2.7b-c depict the uncertainty of the evidence estimates derived from the different trials with the LM and GMIS methods.

The most important conclusions are as follows. In the first place, the evidence estimates derived from both methods appear similar for model complexities with less than 30 (unknown) parameters. Beyond this, the difference between the marginal likelihoods derived from both methods grows up to 2% in $\log_{10}$ space for $d = 105$. Secondly, the evidence estimates derived from the different trials are quite similar, particularly for the GMIS method. Thirdly, the use of a larger number of layers in the two layered models does not necessarily increase the statistical support for this model. Indeed, the value of the evidence is maximized when using 37 horizontal porosity layers or 15 vertical porosity layers. Beyond this number of porosity layers, the evidence values deteriorate slowly but with the exception of a sudden increase in $\mathscr{P}(\tilde{\mathbf{Y}})$ at $d = 40$ for the vertically layered model. This spike is observed in the empirical $\mathscr{P}(\tilde{\mathbf{Y}})$ functions of both evidence estimators (LM and GMIS), inspiring confidence in their results. Notice that the GMIS estimator produces a secondary peak at $d = 63$ (sixty layers), which causes the LM and GMIS methods to diverge in the rightmost part of their $\mathscr{P}(\tilde{\mathbf{Y}})$ curves. Since it is not particularly clear which of the two estimators is at fault, we further test this case with GMIS by using 10^6 instead of 10^5 posterior realizations to construct the $d = 63$-variate importance distribution. The results (not shown herein) confirm the presence of the peak at $d = 63$ which suggests that the secondary peak is real. Fortunately, this does not affect at all

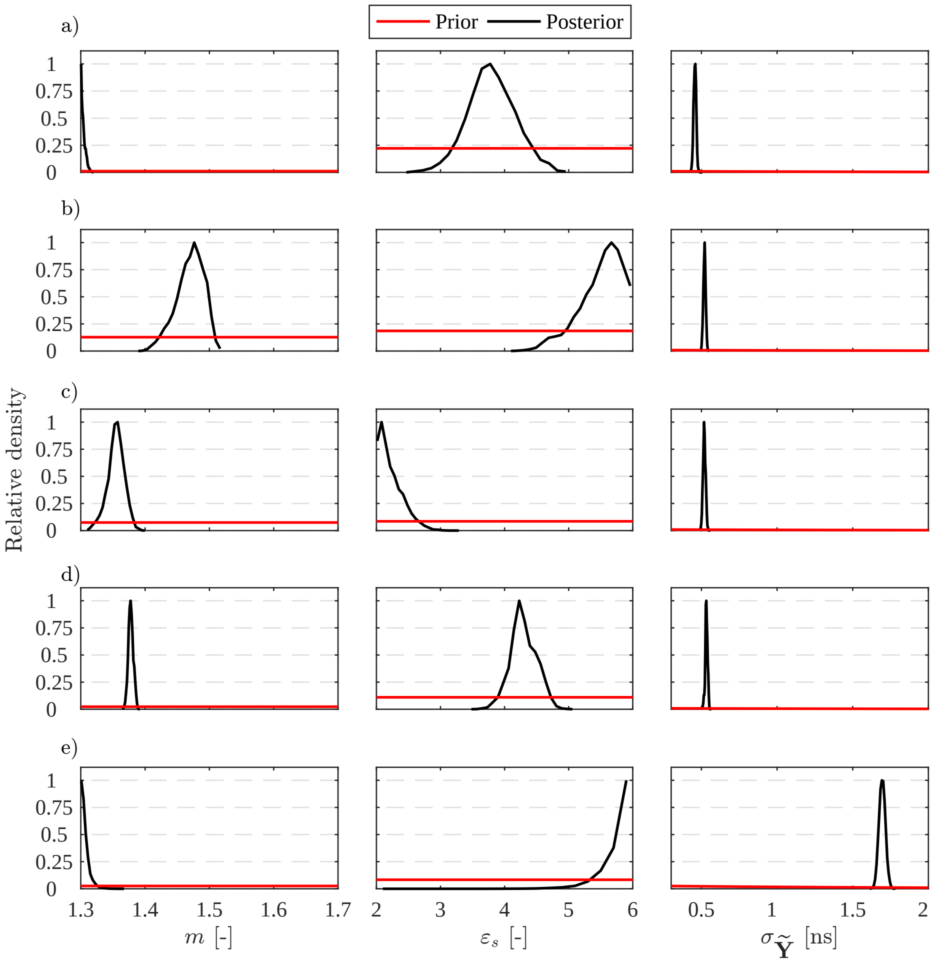


Figure 2.6 – Marginal posterior distributions of the inferred cementation index, m (first column), the relative permittivity of the mineral grains ε_s (second column), and the inferred data error, $\sigma_{\tilde{\mathbf{Y}}}$ (third column) for the multi-Gaussian models with (a) isotropy, (b) horizontal anisotropy, and (c) vertical anisotropy of the porosity values, and the two models with (d) horizontal and (e) vertical layering. The prior distribution is indicated separately in each plot using the red lines. The densities in each plot are normalized so that they all share the units of the y-axis on the left.

model ranking as the evidence values of the vertically layered porosity model are many orders of magnitude smaller than their counterparts of the multi-Gaussian models. These results illustrate the importance of hypothesis testing and highlight the need for (statistical) methods that help us to determine, in an efficient and robust manner, an appropriate model complexity. In fact, the marginalization approach that is used to determine the model evidence can be viewed as a formalization of Occam's razor and leads to a subsurface characterization that is not too simple nor too complex. Furthermore, and perhaps most important from the perspective of the present paper, the isotropic multi-Gaussian model receives the largest evidence values. This is true for both methods. Note, also that the vertically layered model exhibits very low evidence values. Indeed, the best vertically layered model has an evidence in $\log_{10}$ units of about -2757, much lower than the values of approximately -1025 and -1178 for the multi-Gaussian and horizontally layered models, respectively. This latter result confirms our earlier conclusion that the vertically layered model is deficient and inadequate.

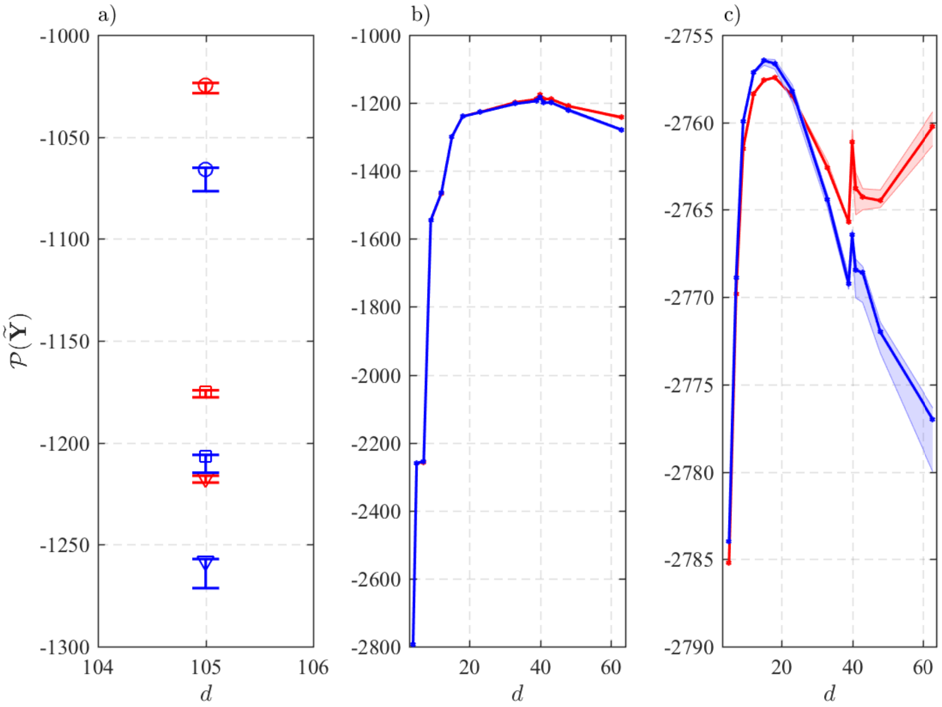


Figure 2.7 – Mean values of the evidence in $\log_{10}$ space, $\mathscr{P}(\widetilde{\mathbf{Y}})$, derived from the LM (blue) and GMIS (red) methods for (a) the multi-Gaussian models with isotropy (circles), horizontal anisotropy (squares), and vertical anisotropy (triangles), and the two models with (b) horizontal, and (c) vertical layering of the porosity. The error bars in (a) and the shaded areas in (b) and (c) summarize the ranges of the evidence estimates as derived from the different independent trials with both methods.

Table 2.5 shows the five top-ranking conceptual models based on their evidence estimates derived from the LM (first column), and GMIS (second column) methods. The conceptual model that is most supported by the experimental data appears on top of the list (first row). For completeness, we also present in the third column the ranking of the models using as metric the posterior values of the measurement data error, $\sigma_{\widetilde{\mathbf{Y}}}$. All three rankings demonstrate conclusively that the isotropic multi-Gaussian model is preferred. This model receives the highest evidence with both estimators and lowest value of the measurement data error, $\sigma_{\widetilde{\mathbf{Y}}} = 0.457$ ns. Note, that the LM and GMIS methods disagree in their assessment of the second best model. The more approximate LM method achieves the second highest support for the horizontally layered model with 37 layers ($d = 40$), whereas GMIS favours instead the multi-Gaussian model with horizontal anisotropy.

We now calculate the Bayes factor ("odds") for the best model (isotropic multi-Gaussian) of the ensemble in relationship to each conceptual model. The "odds" of the isotropic multi-Gaussian model are on the order of 10^{118} and 10^{151} relative to the second best model of the ensemble according to the LM and GMIS estimators (Table 2.5; Fig. 2.8). Figure 2.8a depicts

Table 2.5 – Ranking of the different conceptual models for the South Oyster Bacterial Transport Site based on evidence estimates derived from the LM (first column) and GMIS (second column) methods. The third column ranks the models based on their respective values of the measurement data error inferred from MCMC simulation using the DREAM$_{(ZS)}$ algorithm.

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

twice the natural logarithm of the Bayes factors with respect to the multi-Gaussian model with horizontal anisotropy (square symbol), and vertical anisotropy (triangle symbol), and Fig. 2.8b-c displays the same entity with respect to the horizontally and vertically layered models, respectively. Colour coding is used to differentiate between the LM (blue) and GMIS (red) evidence estimators. It is evident that the isotropic multi-Gaussian model receives most support by the data - the values listed on the y-axis in each plot are all larger than 600, which according to Table 2.1 suggests that there is very strong evidence against each of these alternative hypotheses.

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

2.5.3 A synthetic experiment

To shed some more light on the selection of the isotropic multi-Gaussian model, we proceed with a synthetic experiment. We use the exact same domain (7.2 × 7.2 m) and setup as in our real-world study (Section 2.5.1), and simulate first-arrival travel times for a multi-offset GPR experiment with 57 transmitter and 57 receiver antennas using as reference porosity a multi-Gaussian field with horizontal anisotropy. This "true" porosity field is constructed without the use of dimensional reduction using values of the integral scales and smoothness parameter listed in Table 2.3. The mean of this porosity field is, $\overline{\phi} = 0.39$ and the variance is, $\nu = 2 \cdot 10^{-4}$. The $57 \times 57 = 3249$ simulated travel times are corrupted with Gaussian white noise using $\sigma_{\widetilde{\mathbf{Y}}} = 0.5$ ns, and these distorted values are now used for numerical inversion using the DREAM$_{(ZS)}$ algorithm.

Table 2.6 presents the evidence estimates of the LM (first row) and GMIS (bottom row) methods using as competing hypotheses multi-Gaussian models with horizontal anisotropy (second column), isotropy (third column) and vertical anisotropy (right column). The numer-

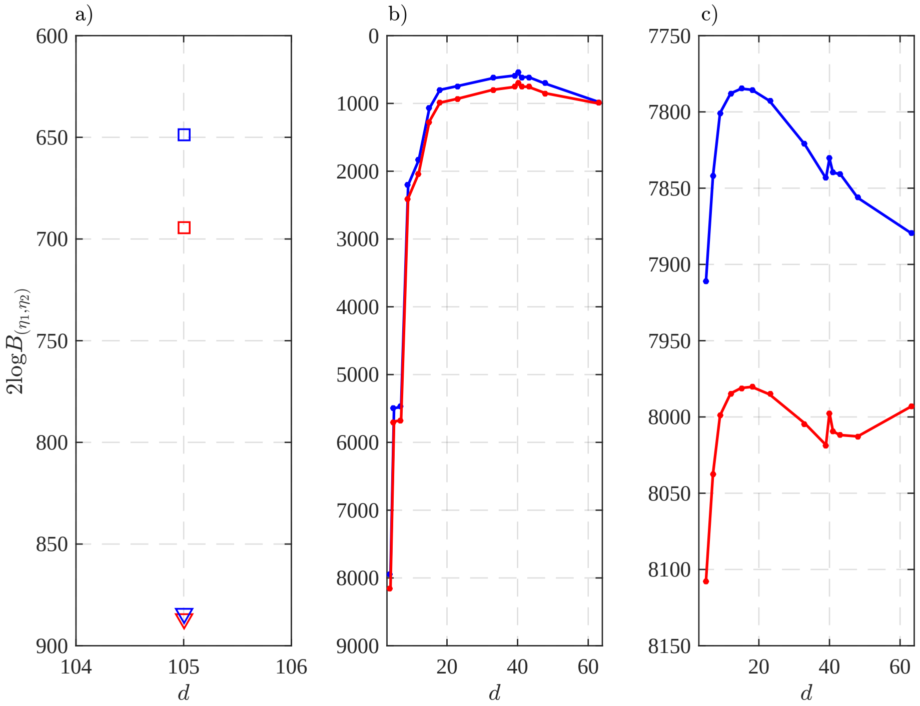


Figure 2.8 – Twice the natural logarithm of the Bayes factors of the best model (isotropic multi-Gaussian) of the ensemble with respect to the (a) multi-Gaussian model with horizontal anisotropy (squares) and vertical anisotropy (triangles), and the two conceptual models with (b) horizontal and (c) vertical layering of the porosity. Results are shown for the LM (blue) and the GMIS (red) methods.

ical setup of these three conceptual models follows exactly Tables 2.3 and 2.4. The results of Table 2.6 demonstrate that both evidence estimators provide a similar ranking of the three subsurface models. As is to be expected, the most support is found for the multi-Gaussian model with horizontal anisotropy (second column). This is followed by the isotropic multi-Gaussian model (third column) and the multi-Gaussian model with vertical anisotropy (last column). This latter model, though, receives rather low evidence values. These results illustrate that both evidence estimators correctly identify the "best" model of the ensemble. We thus feel confident with the main conclusions of our real-world experiment, that the porosity field of the subsurface at the South Oyster Bacterial Transport Site is best described with an isotropic multi-Gaussian model. This conclusion is different from Chen et al. (2001) and Hubbard et al. (2001) whose results favoured the use of a multi-Gaussian model with horizontal anisotropy. These works considered the geophysical tomogram as data within a geostatistical analysis. Possible reasons for this discrepancy is that previous studies relied on forward modeling with straight ray paths and geophysical tomograms with inversions that did not consider an explicit underlying geostatistical model.

Table 2.6 – Synthetic experiment: Evidence estimates derived from the LM and GMIS methods for the multi-Gaussian models with isotropy (MGis), horizontal anisotropy (MGha) and vertical anisotropy (MGva).

	MGha	MGis	MGva
$\mathscr{P}_{\text{LM}}(\widetilde{\mathbf{Y}})$	-1325.39	-1413.53	-1562.47
$\mathscr{P}_{\text{GMIS}}(\widetilde{\mathbf{Y}})$	-1293.94	-1371.91	-1516.72

2.6 Discussion

The transdimensional (or reversible jump) MCMC algorithm (Green, 1995) is not suitable for comparing conceptual models that are based on completely different model parameterizations (e.g., layered vs. multi-Gaussian). In this study, we investigated to what extent evidence estimates with BFMC (Hammersley and Handscomb, 1964), LM (De Bruijn, 1970) and GMIS (Volpi et al., 2017) can be used to perform Bayesian model selection in the context of synthetic and real-world case studies. This is the first comparative study of evidence estimation in hydrogeophysics and we consider realistically high parameter dimensions (i.e., up to 105), large data sets (several thousands) and small data errors.

The BFMC method is known to provide the most reliable and unbiased evidence estimates in the limit of infinite sample sizes. Schöniger et al. (2014, 2015a,b) found reliable evidence estimates with the BFMC method for different case-studies in hydrology. For our set-up with small errors and high data and model dimensions, we found that reliable evidence estimation with the BFMC method would need prohibitive computation times. If the assumption of a multi-Gaussian posterior density is fulfilled (a reasonable assumption in our test cases), the LM method should provide reliable evidence estimates (see also case-studies by Schöniger et al. (2014)). This is confirmed in our synthetic study in Section 2.4 by the strong agreement at low model dimensions between BFMC and LM estimates evaluated around the MAP estimate. The comparison of the LM and the more general (but more time-consuming) GMIS method shows that evidence estimates are similar for simpler subsurface conceptual models but that the difference between them increases with model complexity. Indeed, we do not expect to obtain equivalent results since the two methods are built on different assumptions (see details in Section 2.3.2). For instance, the LM method is built on the assumption that a Gaussian model can properly describe the posterior distribution. This is different for GMIS (or BFMC for that matter) that is based on importance sampling within the prior parameter bounds. It is clear then that the more the posterior distributions are far from being Gaussian, the more the LM and GMIS methods will provide different estimates.

In our application to the South Oyster Bacterial Transport Site (Section 2.5), we found that the isotropic multi-Gaussian model has the highest evidence (Fig. 2.7a). The corresponding Bayes factors (Eq. (2.3)), computed with respect to each tested conceptual models, are all larger than 10^{100}. This result is in agreement with the findings by Schöniger et al. (2014): one decisive winning conceptual model is often obtained when using large data sets and small data errors. We also considered the field example described in Section 2.5.1, but using less data (i.e., $n = 224$ instead of $n = 3248$) and we found (results not shown) that: (1) the isotropic multi-Gaussian model is still the winner, (2) all the evidence estimates are much larger (e.g., in the case of the isotropic multi-Gaussian model, the evidence increases from about 10^{-1000} to 10^{-100}) and that (3) the Bayes factors are much smaller (e.g., when comparing the multi-Gaussian model with vertical anisotropy and the one with isotropy, the Bayes factor decreases approximately from 10^{190} to 10^{10}). Hence, even if we can still identify one clear winning conceptual model, the magnitudes of the Bayes factors have been drastically decreased.

Among the layered models, the GMIS and the LM method both suggest that the conceptual model with 37 layers has the highest evidence (Fig. 2.7b). Moreover, the model type with the least expected geological realism (i.e., vertically layered model) has, by far, the lowest evidences (Fig. 2.7c).

Based on previous geostatistical analysis at the South Oyster Bacterial Transport Site (Chen et al., 2001; Hubbard et al., 2001) one would expect that the multi-Gaussian model with horizontal anisotropy would be the one with the highest evidence. To better understand why the isotropic multi-Gaussian model has a higher evidence than the one with horizontal anisotropy, we performed a synthetic example (Section 2.5.3) in which the true porosity field is described by a multi-Gaussian model with horizontal anisotropy. We found that this conceptual model had the highest evidence, which suggests that the LM and GMIS methods allow us to identify the right conceptual model (Table 2.6). This suggests that this field-site might display less anisotropy than previously thought or that modeling (e.g., ray-based modeling instead of waveform modeling) and geometrical (e.g., uncertainties in borehole and antenna positions) errors bias the evidence estimates.

Below, we outline three avenues for future research:

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

in the statistical literature (e.g. see Lindley's paradox), but comparatively little work has been done on high-dimensional priors as frequently encountered in subsurface characterization and geophysical inference.

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

2.7 Conclusions

Hydrogeophysical methods are well suited to guide the critical choice of the most suitable conceptual subsurface hydrological model. Despite its importance, this topic has largely been ignored in the hydrogeophysical literature to date. We have performed a first comparative study of evidence estimation in hydrogeophysical settings. We consider realistically high model dimensions (i.e., about 100 unknowns), large data sets and small data errors that typify hydrogeophysical investigations. In the context of an illustrative synthetic example, we find that the brute force Monte Carlo method provides reliable estimates at low model dimensions but, when applied to higher model dimensions (i.e., in our case, higher than 6), the BFMC method is inefficient since a prohibitively large number of samples (and thus CPU-time) is required to obtain accurate results. This implied that the brute force Monte Carlo method was unsuitable to address our field example from the South Oyster Bacterial Transport Site (Virginia, USA). We find that the Laplace-Metropolis and the recent Gaussian mixture importance sampling estimator by Volpi et al. (2017) provide overall consistent relative evidence estimates and with rather small errors in both the synthetic cases where simple and low-dimensional (Section 2.4) and more complex and high-dimensional conceptual models (Section 2.5.3) were considered. Application of the Laplace-Metropolis and the Gaussian mixture importance sampling estimator to conceptual subsurface models of the South Oyster Bacterial Transport Site in Virginia, USA, revealed that the isotropic multi-Gaussian model was most supported by the available GPR travel time data. This model had the largest evidence and its Bayes factors were all larger than 10^{100} relative to all other plausible conceptualizations of the subsurface. Finally, the model with the least geological realism (i.e., vertically layered model) has extremely low evidence values for all of its discretizations (i.e., more than 10^{1500} times smaller than the evidences computed for the horizontally layered or multi-Gaussian models). Future research will focus on including the statistical nature of petrophysical relationships, model errors, and more realistic conceptual models of the subsurface.

Chapter 3

Impact of petrophysical uncertainty on Bayesian hydrogeophysical inversion and model selection

Carlotta Brunetti and Niklas Linde.

Published[1] in *Advances in Water Resources*

[1] Brunetti, C. and Linde, N., (2017). Impact of petrophysical uncertainty on Bayesian hydrogeophysical inversion and model selection. *Advances in Water Resources*, **111** 346-359, doi:10.1016/j.advwatres.2017.11.028

3.1 Abstract

Quantitative hydrogeophysical studies rely heavily on petrophysical relationships that link geophysical properties to hydrogeological properties and state variables. Coupled inversion studies are frequently based on the questionable assumption that these relationships are perfect (i.e., no scatter). Using synthetic examples and crosshole ground-penetrating radar (GPR) data from the South Oyster Bacterial Transport Site in Virginia, USA, we investigate the impact of spatially-correlated petrophysical uncertainty on inferred posterior porosity and hydraulic conductivity distributions and on Bayes factors used in Bayesian model selection. Our study shows that accounting for petrophysical uncertainty in the inversion (I) decreases bias of the inferred variance of hydrogeological subsurface properties, (II) provides more realistic uncertainty assessment and (III) reduces the overconfidence in the ability of geophysical data to falsify conceptual hydrogeological models.

3.2 Introduction

A primary goal in hydrogeophysical studies is often to infer quantitative hydrogeological models from geophysical and any available hydrogeological data. Unfortunately, petrophysical relationships describing links between geophysical properties and hydrogeological parameters and state variables are uncertain and the information content of hydrogeophysically-inferred estimates is significantly affected by their predictive power. We distinguish here between three types of uncertainty in petrophysical (also called rock physics) models: (1) *petrophysical model uncertainty* refers to uncertainty about the most appropriate parametric form (e.g., Archie's law, time propagation model, Wyllie's formula), (2) *petrophysical parameter uncertainty* relates to uncertainty about the most appropriate parameter values (e.g., cementation index, saturation exponent), and (3) *petrophysical prediction uncertainty* describes the scatter and bias around the calibrated petrophysical model (e.g., dispersion around predictions based on Archie's law). These three types of uncertainty are clearly not independent of each other. For instance, petrophysical prediction uncertainty is described by the residuals between the actual prediction quantity (e.g., porosity, hydraulic conductivity) and the predictions for a given petrophysical model and parameter values.

To date, most focus in hydrogeophysical inversion has been on petrophysical parameter uncertainty (e.g., Kowalsky et al. (2005); Lochbühler et al. (2014a)) with the petrophysical parameter values being inferred (deterministically or probabilistically) as a part of the inversion process. However, ignoring the other two types of uncertainty may lead to biased estimates and unrealistically low uncertainty estimates. For instance, Brunetti et al. (2017) suggest that ignoring petrophysical prediction uncertainty when using Bayesian model selection to discriminate among conceptual hydrogeological models will likely lead to over confidence in the ability of geophysical data to falsify and discriminate between alternative conceptual hydrogeological models (Linde, 2014). Furthermore, it also implies that ad hoc data weighting schemes are needed when jointly inverting geophysical and hydrogeological data (e.g., Lochbühler et al. (2013) in which each data type was given an equal weight in the objective function).

One approach to partly circumvent these issues is to avoid the use of explicit petrophysical relationships altogether. For instance, this can be achieved using structural approaches to joint inversion (Haber and Oldenburg, 1997). The cross-gradient method of Gallardo and Meju (2003) is a widely employed approach to penalize structural dissimilarity between any two parameter fields (defined as the cross-product of the spatial gradients of two parameter fields). Hydrogeophysical adaptations and applications of this method can be found in Doetsch et al. (2010); Linde et al. (2006a, 2008); Lochbühler et al. (2013). Unfortunately, minimizing the cross-gradient function is an inappropriate approach when both hydrogeological properties and state variables vary (e.g., Doetsch et al. (2010); Linde et al. (2006a)). Among a multitude of cluster-based approaches, we highlight the works by Sun and Li (2016, 2017) who develop a multidomain joint clustering inversion method that uses the fuzzy c-means clustering technique to constrain the statistical behaviour of inverted physical property values in the parameter domain. This approach overcomes the problem of determining a priori the appropriate petrophysical model as it is allowed to exhibit different forms in different regions of the model domain. For time-lapse applications, Vasco et al. (2014) circumvent the use of an explicit petrophysical model by relating the time at which a significant change in geophysical data occurs to the time of a saturation and/or pressure change within a reservoir or aquifer. Alternative approaches are presented by Hermans et al. (2016) and Oware et al. (2013). They link geophysical properties to hydrogeological parameters by physically-based regularization operators or direct multivariate statistical models but, unlike other methods, they adopt an explicit petrophysical relationship to create a prior set of subsurface model realizations or training images. This is done to ensure geologically realistic results.

Explicit petrophysical relationships can be integrated in hydrogeophysical inversions using two types of work flows: two-step (or sequential) inversion approaches (Chen et al., 2001; Copty et al., 1993; Doyen, 1988, 2007; Rubin et al., 1992) and coupled inversion approaches (Hinnell et al., 2010; Kowalsky et al., 2005).

The two-step inversion approach consists of two sequential steps: first, the geophysical properties (e.g., electrical permittivity) are inferred from geophysical data (e.g., first-arrival ground-penetrating radar (GPR) travel times) through deterministic or stochastic inversions; second, petrophysical relationships are used to classify and map the inferred geophysical properties into probability density functions (Mukerji et al., 2001) or deterministic estimates of hydrogeological or reservoir properties. This is achieved by different statistical techniques, such as, co-kriging, discriminant analysis, neural networks and Bayesian classification/estimation. In reservoir geophysics, the two-step inversion approach has been favoured in conjunction with sophisticated statistical rock physics models. For instance, Shahraeeni and Curtis (2011); Shahraeeni et al. (2012) use neural networks to map inferred seismic wave impedances into posterior distributions of porosity, clay content, and water saturation. Grana and Della Rossa (2010); Grana et al. (2012) sample the posterior distribution of reservoir properties using the Monte Carlo method for a given seismic model. They conceptualize petrophysical prediction uncertainty as Gaussian random fields with zero mean and a covariance matrix estimated by comparing predictions with well-log data. In hydrogeophysics, the Bayesian two-step approaches are also used, for instance, by Chen et al. (2001, 2004) to estimate hydraulic conductivity conditioned to GPR velocity, GPR attenuation, and seismic velocity tomograms. In hydrogeophysics, the two-step approach has been criticized as it can lead to inconsistent estimates (apparent mass loss) and spatially-dependent bias (Day-Lewis et al., 2005).

The coupled inversion approach is often formulated within a Bayesian framework in which hydrogeological properties are estimated by inversion of geophysical and, possibly, hydrogeological data. A pioneering work on coupled inversion is Bosch (1999) who develops a formal Bayesian procedure, referred to as lithological tomography or lithological inversion. In this approach, Markov chain Monte Carlo (MCMC) is used to integrate geophysical data, geological concepts and uncertain petrophysical relationships. The coupled inversion approach is well suited to integrate multiple geophysical datasets and arbitrary petrophysical relationships. Also, when confronted with non-linear physics and non-linear petrophysical relationships, the coupled inversion approach is preferable to a two-step inversion approach (Bosch, 2004). Most hydrogeophysical works based on coupled inversion approaches assume that the petrophysical relationship is perfect with known or unknown parameter values (Chen et al., 2006; Kowalsky et al., 2005; Lochbühler et al., 2015). When petrophysical parameter values are unknown, they are inverted for simultaneously with the hydrogeological properties of interest. Petrophysical prediction uncertainty has received less attention in coupled inversion. In the rare circumstances it is included at all, it is commonly conceptualized with a multivariate Gaussian distribution with known mean and covariance matrix (Bosch, 2004; Bosch et al., 2009; Bosch, 2016; Chen and Dickens, 2009). The petrophysical prediction uncertainty is then typically sampled using the brute force Monte Carlo method by adding random multivariate Gaussian realizations to the petrophysical model outputs at each iteration of the MCMC inversion.

In this study, we address the following research questions using a coupled Bayesian hydrogeophysical inversion approach:

1. How can we efficiently incorporate petrophysical prediction uncertainty in MCMC inversions?
2. What are the consequences of ignoring or making incorrect assumptions on petrophysical prediction uncertainty (including its correlation structure) on inferred posterior distributions of interest?
3. Can we reliably infer a geostatistical model of petrophysical prediction uncertainty within the inversion?
4. What are the impacts of petrophysical uncertainty on Bayesian model selection results?

After introducing the theory and method (Section 3.3), we start out by exploring the above-mentioned research questions by means of porosity estimation using synthetic crosshole GPR travel time data and an explicit well-known petrophysical relationship with known parameters (Section 3.4). We then present a field case-study (Section 3.5) aiming at hydraulic conductivity estimation from GPR travel time and hydraulic conductivity (flowmeter) data measured at the South Oyster Bacterial Transport site in Virginia, USA (Chen et al., 2001; Hubbard et al., 2001; Scheibe et al., 2011). Here, we solely assume to know the parametric form of the petrophysical relationship and we infer for its petrophysical parameters (i.e., the petrophysical parameter uncertainty is considered in addition to petrophysical prediction uncertainty).

3.3 Theory and method

3.3.1 Bayesian inference and model selection

We present below a short summary of Bayesian inference and model selection.

Given n measurements, $\widetilde{\mathbf{Y}} = \{\widetilde{y}_1, \ldots, \widetilde{y}_n\}$, and a d-dimensional vector of model parameters, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_d\}$, Bayes' theorem defines the posterior probability density function (pdf) of the model parameters, $p(\boldsymbol{\theta}|\widetilde{\mathbf{Y}})$, as

$$p(\boldsymbol{\theta}|\widetilde{\mathbf{Y}}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\widetilde{\mathbf{Y}})}{p(\widetilde{\mathbf{Y}})}. \tag{3.1}$$

The posterior pdf describes the state of knowledge about the model parameters given the observed data and prior knowledge. The prior pdf, $p(\boldsymbol{\theta})$, quantifies the initial state of knowledge about the model parameters before considering the observed data. We consider a likelihood function, $L(\boldsymbol{\theta}|\widetilde{\mathbf{Y}})$, that is Gaussian in shape by imposing uncorrelated and normally distributed measurement errors with constant standard deviation, $\sigma_{\widetilde{\mathbf{Y}}}$,

$$L(\boldsymbol{\theta}|\widetilde{\mathbf{Y}}) = \left(\sqrt{2\pi\sigma_{\widetilde{\mathbf{Y}}}^2}\right)^{-n} \exp\left[-\frac{1}{2}\sum_{h=1}^{n}\left(\frac{\mathscr{F}_h(\boldsymbol{\theta}) - \widetilde{y}_h}{\sigma_{\widetilde{\mathbf{Y}}}}\right)^2\right]. \tag{3.2}$$

The larger the likelihood, the lower is the data misfit between the simulated forward responses, $\mathscr{F}(\boldsymbol{\theta})$, and the data, $\widetilde{\mathbf{Y}}$. The evidence, $p(\widetilde{\mathbf{Y}})$, evaluates the support provided by the observed data to a given model parametrization and prior pdf (conceptual model), η, and it is defined as the (multidimensional) integral of the likelihood function over the prior distribution,

$$p(\widetilde{\mathbf{Y}}|\eta) = \int L(\boldsymbol{\theta},\eta|\widetilde{\mathbf{Y}})p(\boldsymbol{\theta}|\eta)d\boldsymbol{\theta}. \tag{3.3}$$

Computing the evidence is challenging as, in general, the integral in Eq. (3.3) can not be evaluated analytically and it must be approximated by numerical means.

The evidence is used to calculate Bayes factors and is, thus, the cornerstone of Bayesian model selection (Kass and Raftery, 1995). Bayesian model selection (Jeffreys, 1935, 1939) aims at determining the competing conceptual model that is the most supported by the observed data while honouring the principle of Occam's razor. This implies that if multiple conceptual models fit the data nearly equally well, then the simplest model (e.g., with the least number of unknown parameters or the smallest prior parameter ranges) is favoured over more complex ones (Gull, 1988; Jeffreys, 1939; Jefferys and Berger, 1992; MacKay, 1992). Conceptual models could refer to different spatial parametrizations of the subsurface (e.g., multi-Gaussian fields with isotropy or vertical anisotropy) or alternative petrophysical relationships. Bayes factors are simply the ratio of the evidences of two competing conceptual models, η_1 and η_2. For instance, the Bayes factor of η_1 with respect to η_2, or $B_{(\eta_1,\eta_2)}$, is defined as

$$B_{(\eta_1,\eta_2)} = \frac{p(\widetilde{\mathbf{Y}}|\eta_1)}{p(\widetilde{\mathbf{Y}}|\eta_2)}. \tag{3.4}$$

Subsurface 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 the prior pdf. This implies that there is no guarantee that the "correct" conceptual model will be favoured if a simpler model allows for similar degrees of data misfit.

In this work, we perform coupled Bayesian hydrogeophysical inversion based on MCMC sampling (Robert and Casella, 2013) using the DREAM$_{(ZS)}$ algorithm (Laloy and Vrugt, 2012; Vrugt, 2016) to estimate $p(\theta|\widetilde{\mathbf{Y}})$. This multi-chain method creates symmetric model proposals from an historical archive of past states and automatically tunes the scales and orientation of the proposal distribution on the fly to the target posterior distribution. Each proposal is accepted or rejected based on the Metropolis acceptance ratio (Hastings, 1970; Metropolis et al., 1953). If the proposal is accepted, the chain moves to the new location, otherwise the chain remains at its current location. Acceptance ratios between 15% - 40% usually indicate good performance of the MCMC simulation (Gelman et al., 1996). The convergence to the target posterior distribution is monitored with the analysis of variance by Gelman and Rubin (1992). Approximate convergence is declared when the variance between the different chains is lower than the variance within each single chain (Gilks et al., 1995).

For purposes of Bayesian model selection, we estimate the evidence with the Gaussian mixture importance sampling approach recently developed by Volpi et al. (2017). This approach allows for four different sampling methods: reciprocal importance sampling, importance sampling and bridge sampling with geometric and optimal bridge. Following Brunetti et al. (2017), we rely on importance sampling from a Gaussian mixture model that is fitted to the estimated posterior probability density function.

3.3.2 MC and MCMC sampling of petrophysical prediction uncertainty

As mentioned in Section 3.2, in the rare cases when petrophysical prediction uncertainty is included in coupled inversion, it is sampled through the brute force Monte Carlo (MC) method (Hammersley and Handscomb, 1964) while the inference of model parameters of interest is achieved through MCMC. This method draws independent samples from the (multivariate) prior distribution of petrophysical prediction uncertainty and we refer to it as MC-within-MCMC. In Section 3.4.1, we will demonstrate that the MC-within-MCMC method turns out to be very inefficient because of acceptance rates that are prohibitively low. As an alternative, we make use of the DREAM$_{(ZS)}$ proposal mechanism (see details in Laloy and Vrugt (2012); Vrugt (2016)) to infer the petrophysical prediction uncertainty together with the other parameters by MCMC (full MCMC). In essence, this implies that petrophysical prediction uncertainty is parameterized and treated in the same way as the other unknowns that are inferred in the MCMC inversion. Both the MC-within-MCMC and the full MCMC approaches should converge to the same result. An alternative to such explicit treatments of petrophysical prediction uncertainty as "nuisance" parameters is to incorporate their effects in the likelihood function. However, efficient and theoretically-consistent ways to achieve this for non-linear problems remains an open research question (see Section 5.2 in Linde et al. (2017)).

3.3.3 Petrophysical relationships and geophysical forward model

We consider synthetic test cases for known and theoretically-based petrophysical relationships for which petrophysical prediction uncertainty is comparatively low. For the field study, we consider an unknown, empirically-based and comparatively weak petrophysical relationship. The synthetic example concerns predictions of the porosity field and the field study aims at predicting hydraulic conductivity. These two types of problems were chosen to span typical applications, as well as different strengths and types of petrophysical relationships.

The synthetic examples (Section 3.4) used in this study rely on the following petrophysical relationship to link GPR velocities, $\mathbf{v}$ [m/s], to porosities, $\mathbf{\Phi}$ [-]:

$$\mathbf{v} = \sqrt{\mathbf{\Phi}^{m} c^{-2}[\varepsilon_{\mathrm{w}} + (\mathbf{\Phi}^{-m} - 1)\varepsilon_{\mathrm{s}}]}^{-1}, \tag{3.5}$$

where $\varepsilon_{\mathrm{w}} = 81$ [-] and $c = 3 \cdot 10^{8}$ [m/s] are the relative permittivity of water and the speed of light in vacuum, respectively. We assume the relative permittivity of the mineral grains, ε_{s} [-], equal to 5 and the cementation index, m [-], equal to 1.5. In order to incorporate the petrophysical prediction uncertainty, Eq. (3.5) is computed in three steps. The effective relative permittivities, $\boldsymbol{\varepsilon}$, are first found for a given porosity model (Pride, 1994):

$$\text{Step 1} : \boldsymbol{\varepsilon} = \varepsilon_{\mathrm{s}} + \mathbf{\Phi}^{m}\varepsilon_{\mathrm{w}} - \mathbf{\Phi}^{m}\varepsilon_{\mathrm{s}}, \tag{3.6}$$

then the petrophysical prediction errors, $\Delta\mathbf{p}$, describing the residual for each model cell are added

$$\text{Step 2} : \boldsymbol{\varepsilon}' = \boldsymbol{\varepsilon} + \Delta\mathbf{p}, \tag{3.7}$$

and the corresponding GPR velocities are derived

$$\text{Step 3} : \mathbf{v} = \sqrt{c^{-2}\boldsymbol{\varepsilon}'}^{-1}. \tag{3.8}$$

In the context of the field study (Section 3.5) at the South Oyster Bacterial Transport Site, we compare linear and quadratic petrophysical relationships to link the GPR velocities, $\mathbf{v}$ [m/s], to the natural logarithm of the hydraulic conductivities, $\mathscr{K} = \log \mathbf{K}$ [log(m/h)]:

$$\text{Step 1} : \mathbf{v}' = a_0 + a_1\mathscr{K} \tag{3.9}$$

or

$$\text{Step 1} : \mathbf{v}' = a_0 + a_1\mathscr{K} + a_2\mathscr{K}^2 \tag{3.10}$$

where a_0, a_1 and a_2 are the polynomial coefficients. We then add $\Delta\mathbf{p}$:

$$\text{Step 2} : \mathbf{v} = \mathbf{v}' + \Delta\mathbf{p}. \tag{3.11}$$

Chen et al. (2001) and Hubbard et al. (2001) demonstrate at the South Oyster Bacterial Transport Site that the GPR velocities inferred by linear tomographic inversion are correlated to the logarithm of hydraulic conductivities with a correlation coefficient of 0.68. This suggests that the true underlying correlation is equal or stronger than this value. However, we stress that any relationship between GPR velocity and hydraulic conductivity is site-specific and typically weak.

The spatial model domain of interest covers an area of 7.2 m × 7.2 m below the ground surface. We consider multi-Gaussian models of porosity, hydraulic conductivity and petrophysical prediction uncertainty over a regular 2D grid of size 180 × 180. We use the non-linear 2D traveltime solver (*time 2d*) of Podvin and Lecomte (1991) to compute first-arrival travel times from velocity fields obtained by applying the petrophysical relationships of Eqs. (3.5), (3.8) and (3.9)-(3.11) to each porosity or hydraulic conductivity field.

3.3.4 Model parameterisation

We generally describe the petrophysical prediction uncertainty, $\Delta\mathbf{p}$, the porosity, $\mathbf{\Phi}$, and the log-hydraulic conductivity, $\mathscr{K}$, fields as multi-Gaussian random fields. The only exception is the illustrative synthetic example of Section 3.4.1, in which the $\mathbf{\Phi}$ and $\Delta\mathbf{p}$ fields correspond to independent horizontal layers. We parameterise our multi-Gaussian fields using the method by Laloy et al. (2015). This method generates stationary multi-Gaussian fields by employing circulant embedding of the covariance matrix. To decrease the number of unknowns inferred during the inversion process, the dimensionality is reduced by resampling two low-dimensional vectors of standard normal random numbers to the original size of the model using the one-dimensional Fast Fourier Transform interpolation. We refer to Laloy et al. (2015) for more details. In our case, we generate each vector with 50 dimensionality reduction (**DR**) variables (i.e., 100 instead of 32400 unknowns), which substantially decrease the MCMC computational cost. The multi-Gaussian model is described by the Matérn variogram model and associated geostatistical parameters, including the mean and the variance, the integral scale along the major axis of anisotropy, I, the anisotropy angle, φ, the ratio of the integral scales along the minor and major axis of anisotropy, R, and the shape parameter of the Matérn variogram model, ν. We jointly infer the geostatistical parameters and the **DR** variables describing the hydrogeological properties (i.e., porosity or hydraulic conductivity) with the corresponding parameters and variables characterising the petrophysical prediction uncertainty.

3.4 Synthetic examples

3.4.1 Toy example: MC-within-MCMC versus full MCMC sampling

Historically (see Section 3.3.2), petrophysical prediction uncertainty has been addressed by drawing independent proposals of $\Delta\mathbf{p}$ from the prior while parameters of interest have been inferred by MCMC (MC-within-MCMC). As an alternative, petrophysical prediction uncertainty is here parameterized and inferred as any other parameter in the MCMC inversion (full MCMC). We consider a toy example to demonstrate the advantage of using an appropriate model proposal distribution to infer the petrophysical prediction uncertainty (full MCMC) when considering moderately large or large data sets with high signal-to-noise-ratios. The set-up of this simple synthetic example consists of 10 GPR transmitters and 10 receivers placed at uniform depth intervals on the right and left side of the model domain, respectively (Fig. 3.1a). Considering all possible transmitter-receiver pairs yields 100 first-arrival travel time data. The true porosity field is characterized by four layers of equal thickness with values of 0.3, 0.45, 0.35 and 0.4 starting from the ground surface (Fig. 3.1a). We consider synthetic travel time data that are contaminated with uncorrelated and normally distributed measurement errors with standard deviation, $\sigma_{\tilde{\mathbf{Y}}}$, equal to 0.5 ns (i.e., typical of crosshole GPR) and 2 ns, respectively. We consider a uniform prior distribution of porosity in the range [0.25,0.50] and the prior distribution of the petrophysical prediction uncertainty, $\Delta\mathbf{p}$, is Gaussian with zero-mean and standard deviation of 0.8, chosen according to the experimental study of Roth et al. (1990). The $\Delta\mathbf{p}$ values are added following Eq. 3.7 and integrated in the inversion with the MC-within-MCMC and the full MCMC methods (see Section 3.3.2). The latter draws the parameters from the DREAM$_{(ZS)}$ proposal distribution that gradually update $\Delta\mathbf{p}$.

We obtain appropriate acceptance rates of 20% (with $\sigma_{\tilde{\mathbf{Y}}} = 0.5$ ns) and 22% (with $\sigma_{\tilde{\mathbf{Y}}} = 2.0$ ns) when considering full MCMC (Table 3.1). For MC-within-MCMC, the acceptance ratio is 0.002% when $\sigma_{\tilde{\mathbf{Y}}} = 0.5$ ns and 0.31% when $\sigma_{\tilde{\mathbf{Y}}} = 2.0$ ns. Convergence to the target distribution is consequently much faster for full MCMC than for MC-within-MCMC, especially when $\sigma_{\tilde{\mathbf{Y}}} = 0.5$ ns (i.e., $5 \cdot 10^3$ forward simulations needed instead of $9.5 \cdot 10^6$, Table 3.1). That is, the MCMC-derived method allows for an almost 2000-fold decrease in sampling time with respect to the MC-within-MCMC method. This ratio grows further when using smaller $\sigma_{\tilde{\mathbf{Y}}}$ and more data.

Table 3.1 – First column, method used to sample $\Delta\mathbf{p}$; second column, standard deviation of the measurement errors used to contaminate the data; third column, average acceptance rate; fourth column, number of iterations needed to reach convergence.

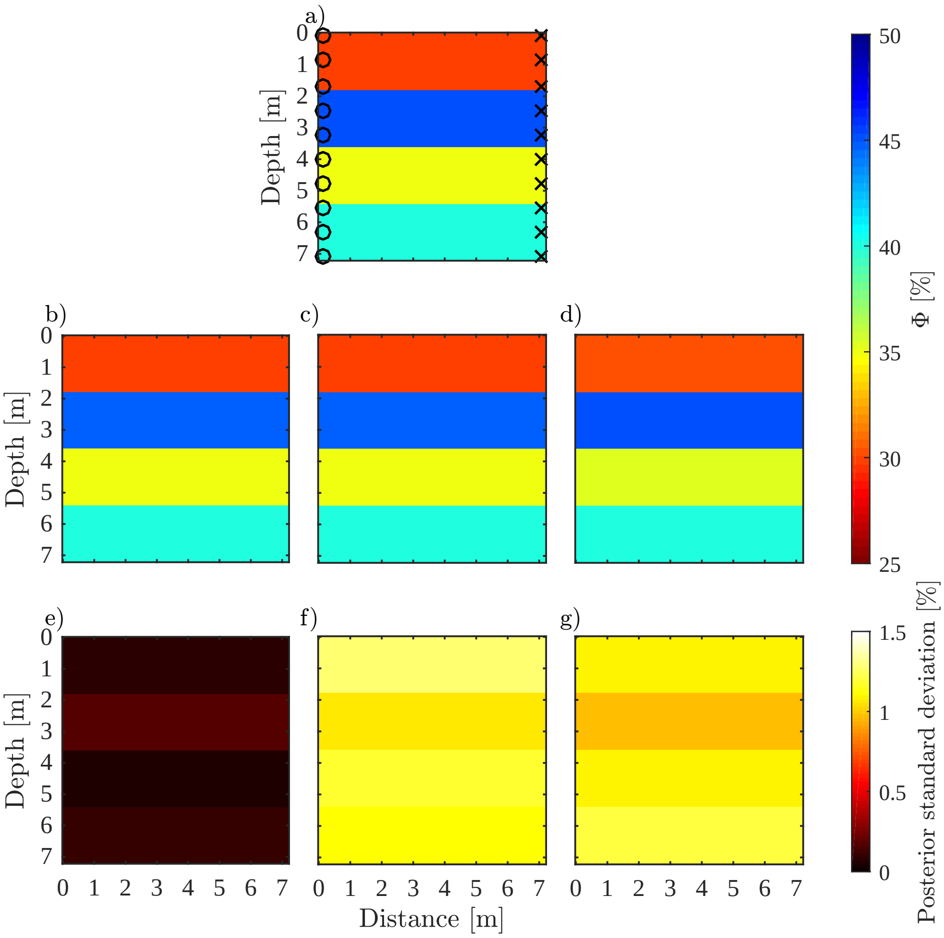
Method	$\sigma_{\tilde{\mathbf{Y}}}$ **[ns]**	**AR [%]**	**T [-]**
Full MCMC	0.5	20.1	$5.0 \cdot 10^3$
	2.0	21.9	$4.0 \cdot 10^3$
MC-within-MCMC	0.5	0.002	$9.5 \cdot 10^6$
	2.0	0.31	$9.6 \cdot 10^4$

For the case of $\sigma_{\tilde{\mathbf{Y}}} = 0.5$ ns, we compare the posterior mean porosity fields and associated standard deviations obtained when ignoring $\Delta\mathbf{p}$ (Fig. 3.1b and 3.1e), when using the full MCMC (Fig. 3.1c and 3.1f) and the MC-within-MCMC estimated $\Delta\mathbf{p}$ (Fig. 3.1d and 3.1g). The posterior mean porosity fields obtained in the three cases (Fig. 3.1b-d) are very similar and agree very well with the true porosity field shown in Fig. 3.1a. The incorporation of the petrophysical prediction uncertainty results in a standard deviation (Fig. 3.1f-g) that is ten times higher than for the case without petrophysical prediction uncertainty (Fig. 3.1e). These results suggest that petrophysical prediction uncertainty has a strong effect on the inferred model uncertainty and that the full MCMC approach is much more efficient than MC-within-MCMC. In the following, we will only present results obtained by the full MCMC approach and recommend it over MC-within-MCMC.

3.4.2 The forward problem: impact of petrophysical prediction uncertainty

For a given study area, geological facies and properties change in space (e.g., porosity, specific surface area, tortuosity) such that the optimal parameters describing any petrophysical relationship are likely to vary in space. This implies that, when relying on the common assumption of a stationary petrophysical relationship (i.e., the parameter values are the same everywhere), the petrophysical prediction uncertainty is likely to have a spatially-correlated structure at a scale similar to the geological variability.

In this section, we investigate the impact of spatially-correlated petrophysical prediction uncertainty on data residuals by considering forward responses obtained with and without spatially-correlated petrophysical errors. In this section, we do not perform any inversion, but simply demonstrate the impact of the correlation scale of petrophysical prediction uncertainty. We consider 841 synthetic crosshole GPR travel times that are related to the porosity field in Fig. 3.2a. The porosity field is described by a multi-Gaussian field with horizontal anisotropy with: $\varphi = 90°$, mean, $\overline{\Phi} = 0.39$, variance, $\sigma^2_{\mathbf{\Phi}} = 2 \cdot 10^{-4}$, integral scale, $I_{\mathbf{\Phi}} = 1.5$ m, integral scales ratio, $R_{\mathbf{\Phi}} = 0.13$ and the shape parameter, $\nu_{\mathbf{\Phi}} = 0.5$ that corresponds to an exponential variogram. In the absence of any petrophysical prediction uncertainty, we obtain the velocity field by applying Eq. 3.5 with known petrophysical parameters. After calculating the corresponding forward response (Section 3.3.3), we add uncorrelated Gaussian observational noise with $\sigma_{\tilde{\mathbf{Y}}} = 0.5$ ns, which leads to a root mean square error (RMSE) of 0.5 ns. For the case of uncorrelated petrophysical prediction errors, we apply Eq. (3.6), (3.7) and (3.8) and draw $\Delta\mathbf{p}$ realizations from an uncorrelated Gaussian distribution with $\sigma_{\Delta\mathbf{p}} = 0.8$. On the resulting simulated travel time data, we add the same observational noise realization. This yields a RMSE of 0.64 ns (Fig. 3.2b); a comparatively small increase in RMSE compared with the previous case. We then describe the petrophysical prediction uncertainty with zero-mean isotropic ($R_{\Delta\mathbf{p}} = 1$) multi-Gaussian models with $\sigma_{\Delta\mathbf{p}} = 0.8$ and $\nu_{\Delta\mathbf{p}} = 0.5$. To assess the impact of the spatial correlation of the petrophysical prediction uncertainty, we draw $\Delta\mathbf{p}$ realizations for isotropic multi-Gaussian distributions with increasing integral scales. For the corresponding forward responses, we observe a sharp increase of RMSE with increasing integral scales (Fig. 3.2b). For example, it is higher than 1.20 ns for an integral scale of 1.5 m. The RMSE reaches a plateau slightly above 1.36 ns when the integral scale



.

Figure 3.1 – (a) The "true" subsurface porosity model used in our toy example with the different measurement depths of the GPR transmitters (black crosses) and receivers (black circles) indicated. Mean porosity fields of the posterior distribution derived from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using a conceptual model with four layers in the case where (b) the petrophysical prediction uncertainty is not taken into account, (c) the petrophysical prediction uncertainty is sampled by MCMC and (d) the petrophysical prediction uncertainty is sampled by MC-within-MCMC. The corresponding posterior standard deviations of the porosity estimates are shown in (e), (f) and (g), respectively. All these plots were obtained with $\sigma_{\tilde{\mathbf{Y}}} = 0.5$ ns

approaches the size of the model domain (7.2 m). These results suggests that uncorrelated petrophysical prediction uncertainty (i.e., described by a nugget model) will have a relatively weak impact on inversion results when considering finely-discretized models. However, we suspect petrophysical prediction uncertainty to be spatially-correlated and this correlation increase the effect on the observed data. If these effects are ignored in the inversion, one would expect negative impacts on the inversion results. This is studied in the following section.

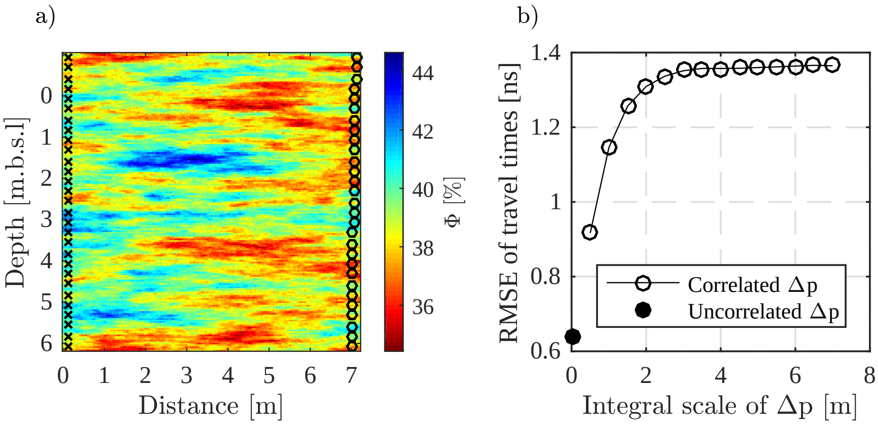


Figure 3.2 – (a) The true porosity model used in our synthetic examples. The 29 GPR transmitter (black crosses) and 29 receiver (black circles) locations are indicated. (b) Root mean square error of GPR travel time data as a consequence of observational errors and petrophysical prediction uncertainty with increasing correlation. In the absence of petrophysical prediction uncertainty, the RMSE is 0.5 ns.

3.4.3 The inverse problem: impact of assumptions on petrophysical prediction uncertainty

In this section, we investigate the consequences of making incorrect assumptions about petrophysical prediction uncertainty when inferring posterior distributions and Bayesian model selection. We consider the same "true" porosity field (Fig. 3.2a) as in Section 3.4.2 and 841 first-arrival GPR travel time data contaminated with uncorrelated and normally-distributed measurement errors with standard deviation, $\sigma_{\tilde{\mathbf{Y}}} = 0.5$ ns. In the MCMC inversions, we infer multi-Gaussian porosity fields with horizontal anisotropy and $\mathbf{DR}_{\mathbf{\Phi}}$, $\overline{\Phi}$, $\sigma^2_{\mathbf{\Phi}}$ being "unknown" parameters drawn from the associated prior distributions listed in Table 3.2, while all the other geostatistical parameters affecting the porosity structure are kept fixed. The petrophysical prediction uncertainty (if considered) is described as a zero-mean multi-Gaussian field with horizontal anisotropy and known geostatistical parameters (i.e., only $\mathbf{DR}_{\Delta\mathbf{p}}$ variables are inferred in the inversion, see Table 3.2). As before, the standard deviation, $\sigma_{\Delta\mathbf{p}}$, was set equal to 0.8 according to the experimental study of Roth et al. (1990). The addition of $\mathbf{DR}_{\Delta\mathbf{p}}$ leads to a decrease in the magnitude of the correlation coefficient (from -1 to -0.81) between the "true" porosity and the "true" GPR velocity values.

We consider four cases: $\Delta\mathbf{p}$ is not present in the data (i.e., it is not used to generate the synthetic data) and it is not inferred in the MCMC inversion (Case 1); $\Delta\mathbf{p}$ is inferred but it is not present in the data used for inversion (Case 2); $\Delta\mathbf{p}$ is present in the data, but not inferred (Case 3); $\Delta\mathbf{p}$ is present in the data and inferred (Case 4). Cases 1 and 4 represent situations where the assumptions are consistent with the "field" situation, while Cases 2 and 3 are based on inconsistent assumptions. We suggest that Case 3 represent the most common situation in the hydrogeophysics literature (i.e., petrophysical prediction uncertainty exists, but it is ignored).

Table 3.2 – Geostatistical parameters of the multi-Gaussian models subject to inference (first column), their respective units (second column), range (third column), prior distribution (fourth column), and number (last column). Dimensionality reduction variables, $\mathbf{DR}_{\boldsymbol{\Phi}}$, mean, $\overline{\Phi}$, and variance, $\sigma^2_{\boldsymbol{\Phi}}$, of the porosity field; dimensionality reduction variables, $\mathbf{DR}_{\Delta\mathbf{p}}$, of the petrophysical prediction errors.

Parameter	**Units**	**Prior range**	**Prior**	**No.**
$\mathbf{DR}_{\boldsymbol{\Phi}}$	-	-	Normal	100
$\overline{\Phi}$	-	$[0.3, 0.5]$	Uniform	1
$\sigma^2_{\boldsymbol{\Phi}}$	-	$[10^{-4}, 2.5 \cdot 10^{-3}]$	Log-uniform	1
$\mathbf{DR}_{\Delta\mathbf{p}}$	-	-	Normal	100

All cases considered provide accurate estimates of the mean porosity (Fig. 3.3a), but only the consistent cases (Case 1 and 4) give significant probability to the actual variance (i.e., sill) describing the porosity field (Fig. 3.3b), with (as expected) Case 4 providing less precise estimates (i.e., parameter uncertainty is higher). For the inconsistent cases, we find for Case 2 that the standard deviation of the porosity field is greatly underestimated, while it is overestimated in Case 3 (Fig. 3.3b).

We now consider the resulting mean porosity fields and the standard deviations for the consistent cases. For Case 1, we find a mean porosity field (Fig. 3.4a) that is very close to the true field (Fig. 3.2a). The standard deviation is low (Fig. 3.4e), the scatter between the mean model and the true model follows the 1:1 trend line (Fig. 3.4i) and the correlation coefficient is high (0.9). For Case 4, we find a slightly less precise mean model (Fig. 3.4d), which is reflected in the standard deviation being twice as large (Fig. 3.4h). Nevertheless, the corresponding scatter plot (Fig. 3.4l) indicates that there is no bias (the scatter falls on the 1:1 trend line) and the correlation coefficient is 0.75.

We now turn our attention to the inconsistent cases. When considering Case 2, we find a less variable mean field (Fig. 3.4b) and standard deviations that are in-between the two consistent cases (Fig. 3.4f). The correlation coefficient is high (0.88), but the estimates are biased as they do not follow the 1:1 trend line (Fig. 3.4j). For Case 3, we find an overly variable mean field (Fig. 3.4c), rather small standard deviations (Fig. 3.4g) and a moderate correlation coefficient (0.75) with a scatter plot above the 1:1 trend line (Fig. 3.4k). These results suggest different outcomes. First, including a known petrophysical prediction uncertainty in the inversion leads to consistent estimates, but a wider posterior distribution than if petrophysical prediction uncertainty is absent. Second, the correlation coefficient with the true model is mainly determined by the petrophysical prediction uncertainty. Third, the estimated petrophysical prediction uncertainty (that does not exist) in Case 2 accounts for some of the variability due to porosity variations, which leads to a too smooth mean porosity field. Lastly, ignoring actual petrophysical prediction uncertainty in the inversion process (Case 3; the common case) leads to overly variable fields in order to accommodate data variability caused by both porosity variations and petrophysical prediction uncertainty. From these first inversion examples, we conclude that ignoring petrophysical prediction uncertainty leads to overly confident parameter inference and that some of the estimated parameters might be biased.

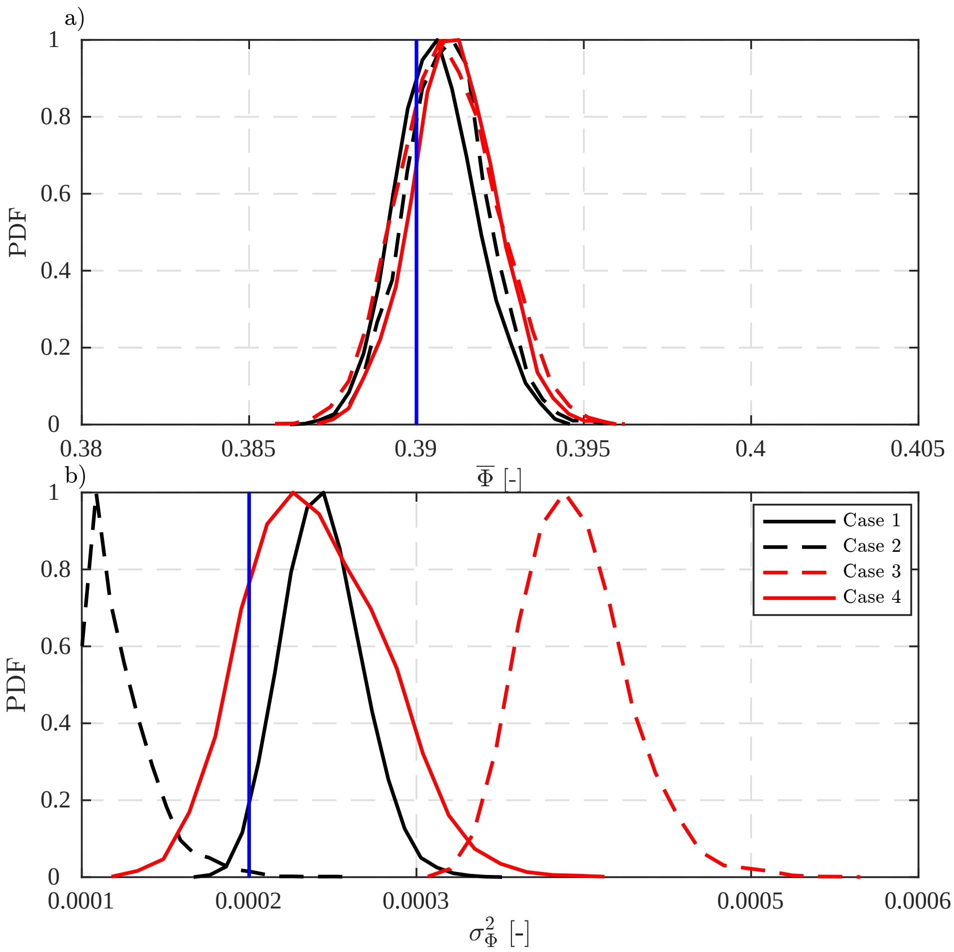


Figure 3.3 – (a) Posterior distributions of the inferred mean of the porosity field. (b) Posterior distributions of the inferred variance (i.e., sill) of the porosity field. The vertical blue lines depict the values of the true model. The posterior distributions are derived from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using 8 chains with 2.5·10^5 iterations.

We now focus our attention on Bayesian model selection. For each of the four cases, we also use the data to infer porosity fields assuming (erroneously) a multi-Gaussian conceptual model with isotropy or vertical anisotropy. We compute the evidence for each of these conceptual models (the case of the true horizontal anisotropy and the incorrect cases of isotropy and vertical anisotropy) by approximating the integral in Eq. (3.3) with the Gaussian mixture importance sampling estimator (Section 3.3.1). For each case, we use a total of 10^5 importance samples and repeat the evidence computation 10 times. The mean evidences and associated ranges are presented in Fig. 3.5.

We find that the ranking of the different conceptual models is the same for all cases. As expected, the multi-Gaussian model with horizontal anisotropy (true conceptual model) has the largest evidence followed by the isotropic model (Fig. 3.5a). The evidence values are the largest when no petrophysical prediction uncertainty is present in the data or in the inversion (Case 1, Fig. 3.5a). When we include $\Delta\mathbf{p}$ in the inversion, the evidence estimates (Case 2, Fig. 3.5a) decrease drastically with respect to Case 1. For instance, we find a 29

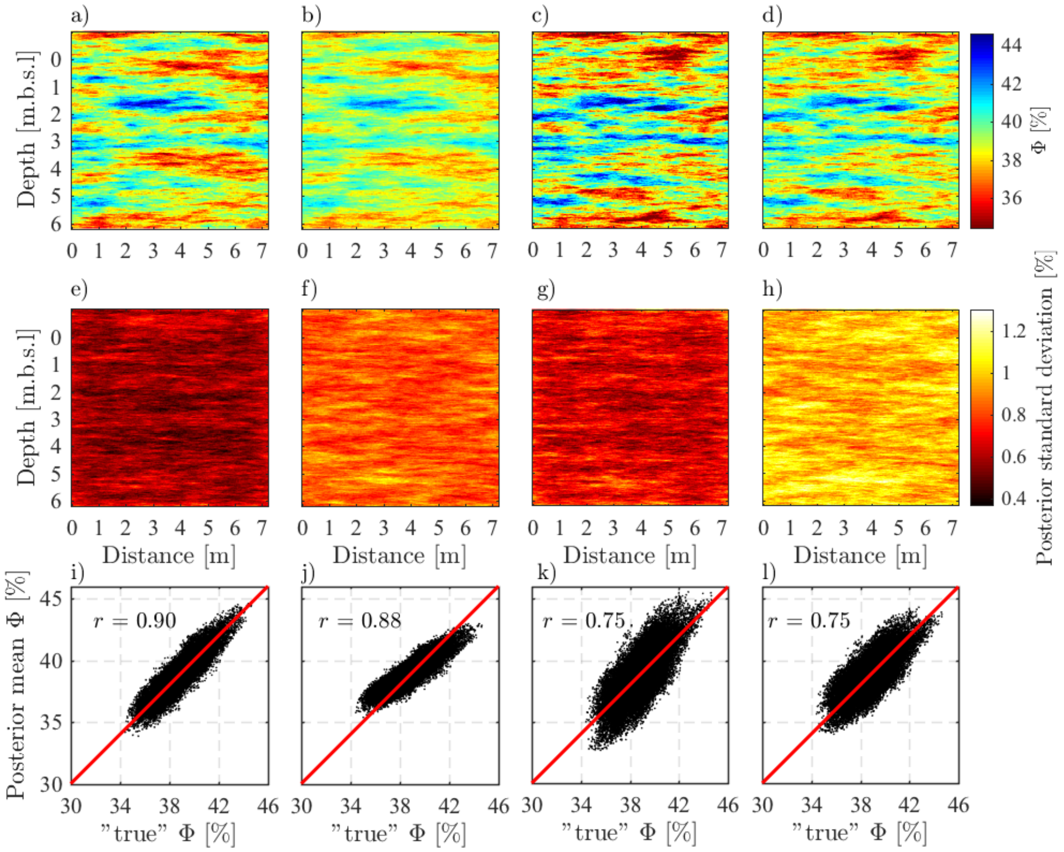


Figure 3.4 – (a-d) Mean porosity fields of the posterior distribution derived from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using 8 chains with 2.5·10^5 iterations for Cases 1-4, respectively. The corresponding posterior standard deviations of the porosity estimates for the four different cases are shown in (e-h), respectively. From (j) to (l), scatter plots of the "true" porosity values versus the mean posterior porosity estimates obtained in the four cases. In each plot, from (j) to (l), the Pearson correlation coefficients, *r*, are reported and the red lines depict the theoretical 1:1 trend line (i.e., Pearson correlation coefficient equal to 1).

orders of magnitude decrease of the evidence estimates for the best model (multi-Gaussian model with horizontal anisotropy). When petrophysical prediction uncertainty is absent in the data (Cases 1 and 2), we find thus that Bayesian model selection clearly indicates that the conceptual model with horizontal anisotropy and no petrophysical prediction uncertainty is superior (the consistent case). Note that this is the case despite the fact that we find the highest log-likelihoods for Case 2 (black dotted lines in Fig. 3.5b-d). The addition of 100 "unnecessary" degrees of freedom in Case 2 leads to a much decreased ability to differentiate among the different geostatistical models. The error bars of the evidence estimates overlap for Case 2 and the Bayes factors (Table 3.3) are much smaller than for Case 1, which imply that it is much more difficult to judge which geostatistical model is preferred statistically.

We have seen above that the Bayesian model selection clearly favours the consistent Case 1 when comparing Cases 1 and 2. Unfortunately, this is not the case when comparing Cases 3 and 4. The consistent Case 4 (petrophysical prediction error in data and model parameterization) has a much lower evidence (Fig. 3.5a) for the multi-Gaussian model with horizontal anisotropy and much lower Bayes factors (Table 3.3) than the inconsistent Case 3 (petrophysical prediction errors in the data only). The reason for this is that Case 3 has similar log-likelihoods (i.e., data misfit) as Case 4 (Fig. 3.5b), but half as many model parameters. The ability to fit the data so well with this inconsistent model is probably a consequence of the petrophysical prediction uncertainty having the same geostatistical model as the porosity field. This implies that formal Bayesian model selection will favour a lower-dimensional model parameterization that fits the data well, regardless of if it is the "correct" model or not. This is a characteristic of Bayesian model selection (e.g., Schöniger et al. (2015b)). Additional tests were performed (not shown) with conditioning to 17 porosity values along each borehole. This decreased the evidence for Case 3 somewhat and increased it for Case 4. However, Case 3 was still strongly favoured when calculating the corresponding Bayes factor.

Table 3.3 – Bayes factors in $\log_{10}$ space of the best conceptual model, MGha, (horizontal anisotropy) with respect to the isotropic one, MGis, (first column) and to the vertically anisotropic one, MGva (last column).

Cases	$\log_{10} B_{(\text{MGha,MGis})}$	$\log_{10} B_{(\text{MGha,MGva})}$
Case 1	18.36	29.09
Case 2	0.58	1.94
Case 3	35.65	78.38
Case 4	10.19	15.37

3.4.4 Inference of petrophysical prediction uncertainty

We have shown (Section 3.4.3) that ignoring petrophysical prediction uncertainty in MCMC inversions leads to over confident parameter estimates and biased estimates of geostatistical properties (e.g., the sill). In practical field situations, it is difficult to determine a priori the appropriate geostatistical model that governs petrophysical prediction uncertainty. In this section, we explore to which extent it is possible to infer for both $\Delta\mathbf{p}$ and its underlying geostatistical model. We consider the same overall setting as in Sections 3.4.2 and 3.4.3 and the same "true" porosity field (Fig. 3.2a). Here, the true petrophysical prediction uncertainty is a zero-mean isotropic multi-Gaussian field with $\sigma_{\Delta\mathbf{p}}$ = 0.8, $I_{\Delta\mathbf{p}}$=0.8 m, $R_{\Delta\mathbf{p}}$=1, and $\nu_{\Delta\mathbf{p}}$=0.5. We then infer for the mean and variance of the porosity field and for all the geostatistical parameters of $\Delta\mathbf{p}$ described above and the corresponding $\mathbf{DR}_{\Delta\mathbf{p}}$ variables. The corresponding prior distributions of these "unknown" parameters are listed in Tables 3.2 and 3.4. The petrophysical relationship used is Eq. (3.5) and the petrophysical prediction uncertainty is accounted for following Eq. (3.7).

The inferred posterior distributions of the mean (Fig. 3.6a) and variance (Fig. 3.6b) of the porosity field are in general quite well recovered, even if they show a slight tendency to underestimate the true values. Overall, the geostatistical properties of the reference

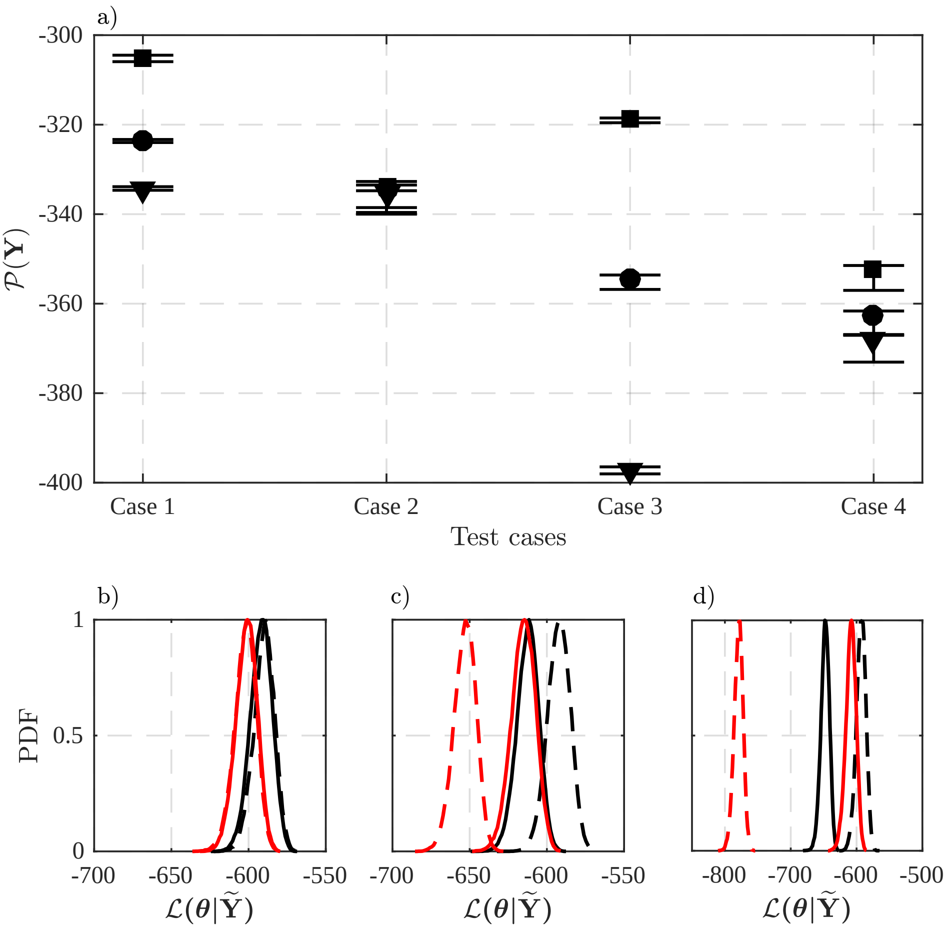


Figure 3.5 – (a) Mean values of the evidence in $\log_{10}$ space, $\mathscr{P}(\widetilde{\mathbf{Y}})$, and corresponding uncertainty (error bars) derived from the Gaussian mixture importance sampling method for the multi-Gaussian conceptual models with horizontal anisotropy (squares), isotropy (circles) and vertical anisotropy (triangles). Posterior distribution of the log-likelihood, $\mathscr{L}(\theta|\widetilde{\mathbf{Y}})$, for the multi-Gaussian model with (b) horizontal anisotropy, (c) isotropy and (d) vertical anisotropy in Case 1 (black solid line), Case 2 (black dotted line), Case 3 (red dotted line) and Case 4 (red solid line).

petrophysical prediction uncertainty field are captured in the sense that the corresponding "true" values are included in the posterior distributions (Fig. 3.6c-g). However, some of the parameters are poorly recovered. For instance, the inferred standard deviation of $\Delta\mathbf{p}$ is centered on the value of 1 instead of 0.8 (Fig. 3.6c) and the inferred shape parameter of the Matérn variogram peaks on a value that is half of the corresponding "true" value (Fig. 3.6g). The anisotropy angle is poorly estimated, which is a consequence of the "true" $\Delta\mathbf{p}$ field being isotropic (Fig. 3.6e). The integral scale along the major axis of anisotropy and the ratio of the integral scales peak on the "true" values, but their posterior distributions are relatively wide (Fig. 3.6d and 3.6f).

The dominant structures in the reference porosity field (Fig. 3.7a), such as the low-porosity zones at a depth of 0.5 m, 4 m and 6 m, are well represented by the posterior mean porosity

Table 3.4 – Geostatistical parameters of the multi-Gaussian model of the petrophysical prediction uncertainty subject to inference (first column), their respective units (second column), range (third column), prior distribution (fourth column), and number (last column). Standard deviation, $\sigma_{\Delta\mathbf{p}}$, integral scale along the major axis of anisotropy, $I_{\Delta\mathbf{p}}$, anisotropy angle, $\varphi_{\Delta\mathbf{p}}$, ratio of the integral scales, $R_{\Delta\mathbf{p}}$, and shape parameter of the Matérn variogram, $\nu_{\Delta\mathbf{p}}$, of the petrophysical prediction uncertainty field.

Parameter	**Units**	**Prior range**	**Prior**	**No.**
$\sigma_{\Delta\mathbf{p}}$	-	$[0.2, 3.6]$	Log-uniform	1
$I_{\Delta\mathbf{p}}$	m	$[0.6, 3]$	Uniform	1
$\varphi_{\Delta\mathbf{p}}$	°	$[0, 180]$	Uniform	1
$R_{\Delta\mathbf{p}}$	-	$[0.05, 1]$	Uniform	1
$\nu_{\Delta\mathbf{p}}$	-	$[0.1, 5]$	Log-uniform	1

field (Fig. 3.7b). The posterior standard deviations on the inferred porosity field span a range between 0.6% and 1% (Fig. 3.7c). We find that the inferred mean petrophysical prediction uncertainty field (Fig. 3.7d) and the "true" field (Fig. 3.7e) have a rather low correlation coefficient (0.55). The posterior standard deviations of $\Delta\mathbf{p}$ span a range between 0.6 and 1 (Fig. 3.7f). These large uncertainties are also reflected in the $\Delta\mathbf{p}$ posterior realizations (Fig. 3.8) that appear to be rather isotropic but with integral scales that vary significantly. Overall, the structural features of the GPR velocity field are well inferred even if their values span a wider range than the reference field (Fig. 3.7g-h). In particular, the high-velocity zone in the bottom right corner of the model domain are enhanced and characterized by large uncertainties (Fig. 3.7i).

We performed also a test with the petrophysical prediction uncertainty field conceptualized by a multi-Gaussian field with anisotropy at 45° (not shown). For this case, we find a significant improvement in the ability to infer for the standard deviation, angle of anisotropy and the shape parameter of $\Delta\mathbf{p}$. These results suggest that $\Delta\mathbf{p}$ is best resolved when its geostatistical properties are markedly different from the underlying porosity field. However, Bayesian model selection between the two conceptual models that include and not include $\Delta\mathbf{p}$ in the inversion still favours the case in which petrophysical prediction uncertainty errors are ignored (not shown).

3.5 Field example

3.5.1 Field site and available data

We now focus our attention on field data from the South Oyster Bacterial Transport Site in Virginia, USA (Hubbard et al., 2001). In Section 3.4, we considered a well known and strong petrophysical relationship, while here we consider a case of an unknown and only moderately strong petrophysical relationship. A PulseEKKO 100 GPR system with a 100-MHz nominal-frequency antenna was used and we consider 841 crosshole GPR first-arrival travel time data between 29 transmitter and 29 receiver locations in boreholes S14 and M3, respectively. A

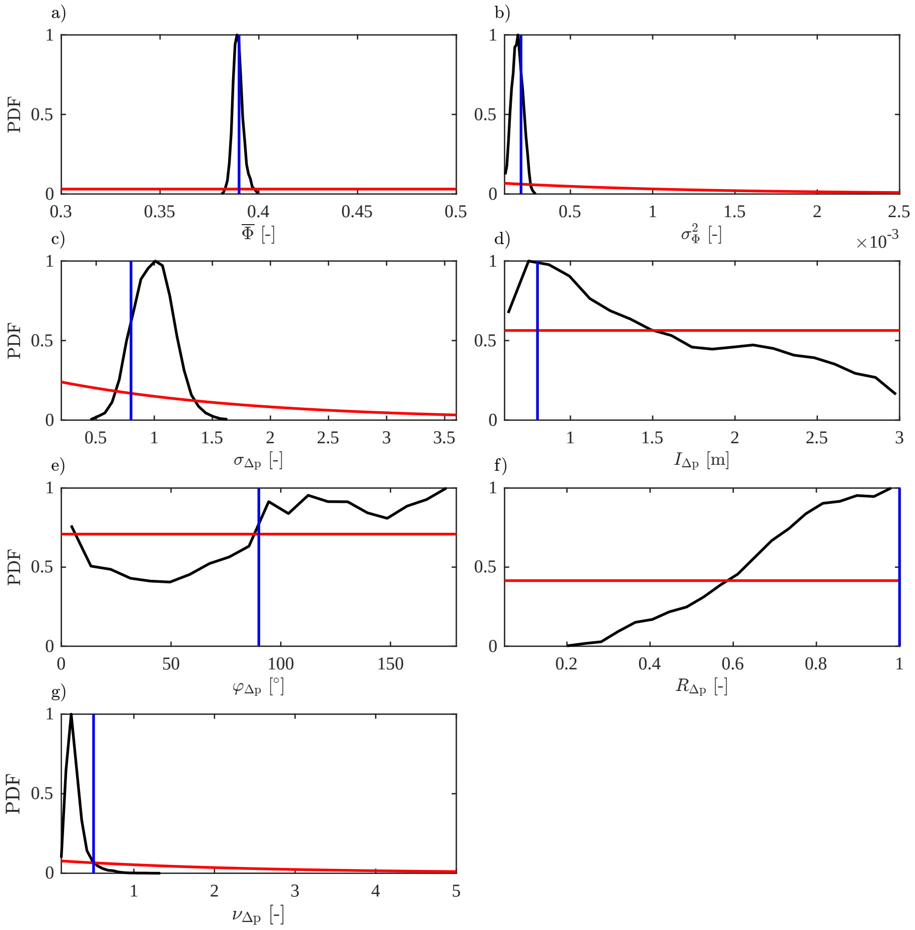


Figure 3.6 – Posterior distributions (black lines) derived from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using 8 chains with 2.5·10^5 iterations of the (a) inferred mean, $\overline{\Phi}$, and (b) variance, σ^2_Φ, of the porosity field and of the geostatistical parameters of the petrophysical prediction uncertainty field: (c) standard deviation, $\sigma_{\Delta \mathbf{p}}$, (d) integral scale along the major axis of anisotropy, $I_{\Delta \mathbf{p}}$, (e) anisotropy angle, $\varphi_{\Delta \mathbf{p}}$, (f) ratio of the integral scales along the minor and major axis of anisotropy, $R_{\Delta \mathbf{p}}$, and (g) shape parameter of the Matérn variogram, $\nu_{\Delta \mathbf{p}}$. The red and blue lines depict the corresponding prior distributions and values of the reference field, respectively. The densities in each plot are normalized.

total of 95 hydraulic conductivity estimates along boreholes S14, T2 and M13 obtained from an electromagnetic flowmeter were used for point conditioning following the methodology outlined by Laloy et al. (2015). We use the GPR data to infer the underlying log-hydraulic conductivity field, $\mathcal{K}$, assuming a multi-Gaussian model with horizontal anisotropy. Its integral scales, the anisotropy angle, and the shape parameter of the Matérn variogram are set based on previous investigations at the site (Chen et al., 2001; Hubbard et al., 2001). These fixed parameters include, $I_{\mathcal{K}}$ = 1.5 m, $\varphi_{\mathcal{K}}$ = 90°, $R_{\mathcal{K}} \approx 0.13$ and $\nu_{\mathcal{K}} = 0.5$. The dimensionality reduction variables, $\mathbf{DR}_{\mathcal{K}}$, the mean, $\overline{\mathcal{K}}$, and standard deviation, $\sigma_{\mathcal{K}}$, of the log-hydraulic conductivity field are subject to inference and the corresponding prior ranges are listed in Table 3.5. The prior range on $\sigma_{\mathcal{K}}$ is set to include the 0.42 log(m/h) standard deviation of the available flowmeter data. The petrophysical prediction uncertainty is described by a

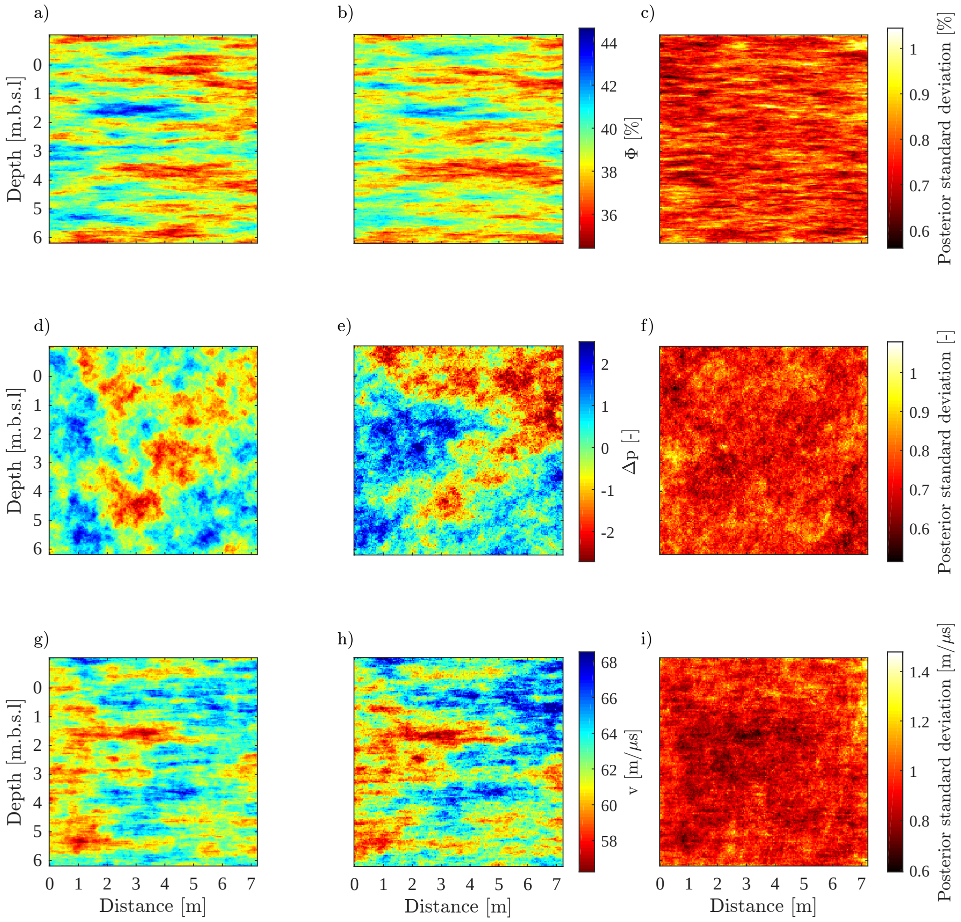


Figure 3.7 – (a) The "true" subsurface porosity model used in our synthetic example; (b) mean porosity field of the posterior distribution derived from MCMC simulation and the corresponding (c) standard deviations. (d) The "true" petrophysical prediction uncertainty model; (e) mean petrophysical prediction uncertainty field of the posterior distribution derived from MCMC simulation and the corresponding (f) standard deviations. (g) The "true" GPR velocity model; (h) mean velocity field of the posterior distribution derived from MCMC simulation and the corresponding (i) standard deviations. The mean fields are obtained from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using 8 chains with $2.5{\cdot}10^5$ iterations.

zero-mean multi-Gaussian field with prior distributions outlined in Table 3.5. The upper bound on the prior range of $\sigma_{\Delta\mathbf{p}}$ is chosen such that the resulting correlation coefficient between GPR velocities and log-hydraulic conductivities is equal or stronger than 0.68, which corresponds to the value reported by Chen et al. (2001) and Hubbard et al. (2001). We also jointly infer the petrophysical parameters a_0, a_1 and a_2 in Eqs. (3.9)-(3.10) and the standard deviation of the measurement errors, $\sigma_{\tilde{\mathbf{Y}}}$ (Table 3.5). The overall number of parameters subject to inference is 211.

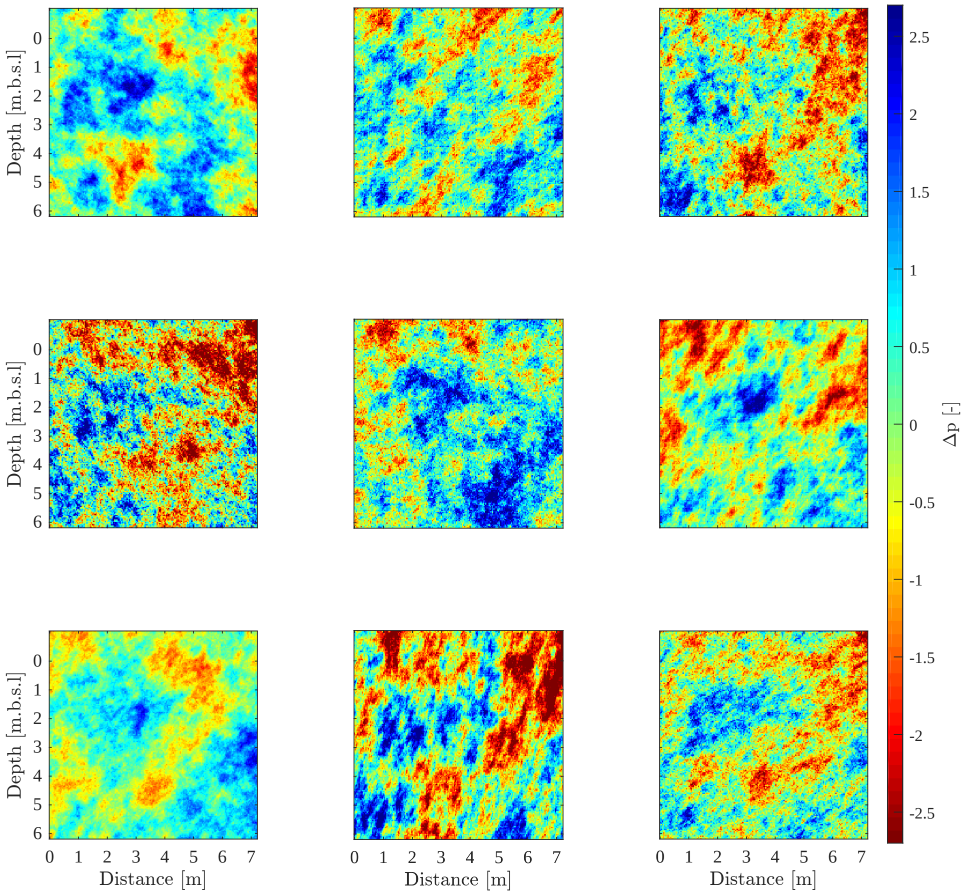


Figure 3.8 – Nine realizations of the petrophysical prediction uncertainty field drawn randomly from the posterior distribution obtained from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using 8 chains with 2.5·10^5 iterations. The petrophysical prediction uncertainty is conceptualized by a multi-Gaussian field with isotropy.

3.5.2 Results at the South Oyster Bacterial Transport Site

In Section 3.4, we considered a synthetic example and a known petrophysical relationship. In the present field example, we only assume to know the parametric form of the petrophysical relationship and we estimate its petrophysical parameters. We infer the underlying log-hydraulic conductivity field and compare the results obtained by assuming three different petrophysical models: a perfect linear petrophysical relationship (Eq. (3.9)) in which the petrophysical prediction uncertainty is ignored (Model 1), a linear petrophysical relationship with scatter $\Delta\mathbf{p}$ taken into account by following Eqs. (3.9) and (3.11) (Model 2), and a quadratic petrophysical relationship with scatter $\Delta\mathbf{p}$ accounted for as in Eqs. (3.10)-(3.11) (Model 3).

After MCMC inversion, we obtain similar posterior distributions of the mean log-hydraulic conductivity when using a perfect linear (-1.58 log(m/h)) and a scattered linear (-1.57 log(m/h)) petrophysical relationship and a slightly lower value (-1.68 log(m/h)) when using a scattered quadratic petrophysical relationship (Fig. 3.9a). When ignoring $\Delta\mathbf{p}$, the inferred

Table 3.5 – Parameters subject to inference at the South Oyster Bacterial Transport Site (first column), their respective units (second column), range (third column), prior distribution (fourth column), and number (last column). Dimensionality reduction variables, $\mathbf{DR}_{\mathcal{K}}$, mean, $\overline{\mathcal{K}}$, and standard deviation, $\sigma_{\mathcal{K}}$, of the natural log-hydraulic conductivity field; dimensionality reduction variables, $\mathbf{DR}_{\Delta\mathbf{p}}$, standard deviation, $\sigma_{\Delta\mathbf{p}}$, integral scale along the major axis of anisotropy, $I_{\Delta\mathbf{p}}$, anisotropy angle, $\varphi_{\Delta\mathbf{p}}$, ratio of the integral scales, $R_{\Delta\mathbf{p}}$, and shape parameter of the Matérn variogram, $\nu_{\Delta\mathbf{p}}$, of the petrophysical prediction uncertainty field; standard deviation of the measurement errors on the travel time data, $\sigma_{\tilde{\mathbf{Y}}}$, and polynomial coefficients of the constant, a_0, the linear, a_1, and quadratic, a_2, terms used to describe linear or a quadratic petrophysical relationships.

Parameter	**Units**	**Prior range**	**Prior**	**No.**
$\mathbf{DR}_{\mathcal{K}}$	-	-	Normal	100
$\overline{\mathcal{K}}$	log(m/h)	$[-2,-1]$	Uniform	1
$\sigma_{\mathcal{K}}$	log(m/h)	$[0.4, 0.5]$	Log-uniform	1
$\mathbf{DR}_{\Delta\mathbf{p}}$	-	-	Normal	100
$\sigma_{\Delta\mathbf{p}}$	m/µs	$[0, 0.8]$	Uniform	1
$I_{\Delta\mathbf{p}}$	m	$[0.6, 3]$	Uniform	1
$\varphi_{\Delta\mathbf{p}}$	°	$[0, 180]$	Uniform	1
$R_{\Delta\mathbf{p}}$	-	$[0.05, 1]$	Uniform	1
$\nu_{\Delta\mathbf{p}}$	-	$[0.1, 5]$	Log-uniform	1
$\sigma_{\tilde{\mathbf{Y}}}$	ns	$[0.3, 2]$	Log-uniform	1
a_0	m/µs	$[40, 100]$	Uniform	1
a_1	$\log(h/m)\cdot$ m/µs	$[0, 80]$	Uniform	1
a_2	$\log(h^2/m^2)\cdot$ m/µs	$[0, 5]$	Uniform	1

standard deviation of the log-hydraulic conductivity field peaks close to the upper bound (black line, Fig. 3.9b). When using a scattered linear or quadratic petrophysical relationship, the inferred posterior distribution of the standard deviation is truncated on the lower bound of the prior range (green and blue lines, Fig. 3.9b). The highest inferred standard deviation of the measurement errors, 0.56 ns, is obtained when ignoring $\Delta\mathbf{p}$ in the inversion (black line, Fig. 3.9c). When considering the scattered linear or quadratic petrophysical relationship, the corresponding estimates are 0.37 ns and 0.36 ns, respectively (Fig. 3.9c).

The parameters describing the three petrophysical relationships are well defined (Fig. 3.9d-e-f). The inferred standard deviation of the petrophysical prediction uncertainty peak on the upper bound of the prior range (Fig. 3.9g). The other geostatistical parameters describing the $\Delta\mathbf{p}$ field have similar posterior distributions regardless of if a linear (green lines) or a quadratic (blue lines) petrophysical relationship is used (Fig. 3.9h-k). In particular, we find that the petrophysical prediction uncertainty field is characterized by an integral scale along the major axis of anisotropy centred around 2.4 m (Fig. 3.9h), an almost horizontal anisotropy (Fig. 3.9i) and a ratio of the integral scales of 0.30 (Fig. 3.9j). The posterior distribution of the Matérn shape parameter is truncated by the upper bound, thereby, suggesting a smooth field (Fig. 3.9k).

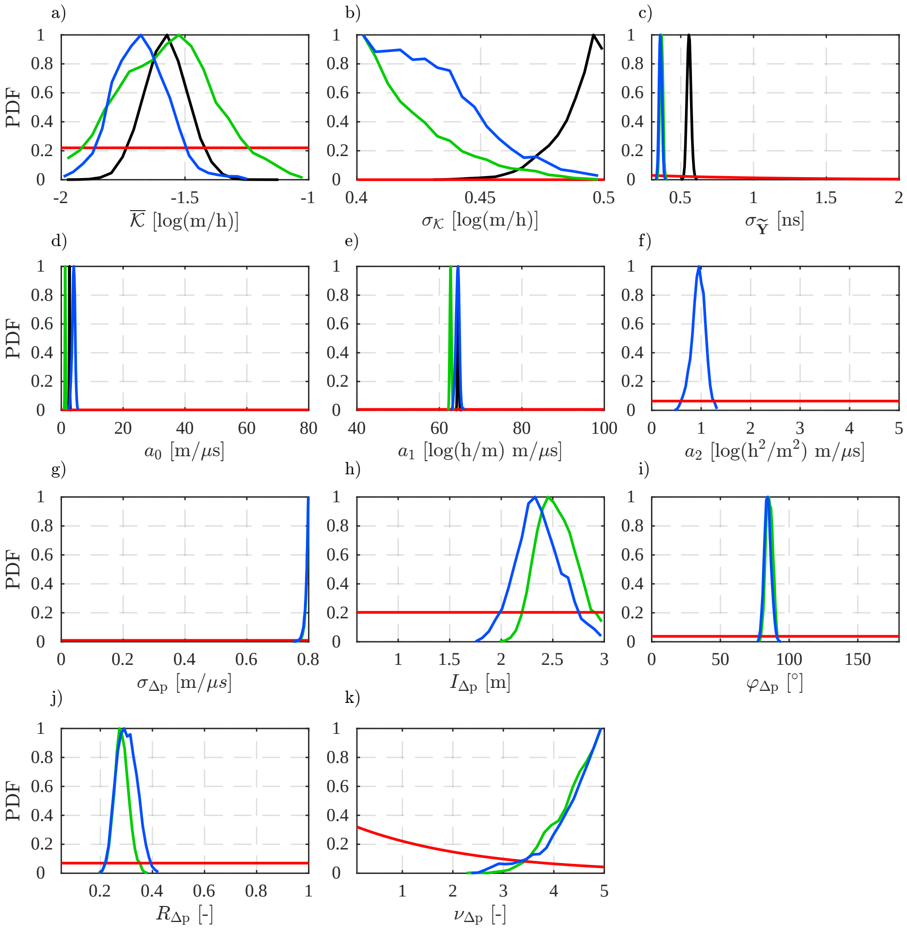


Figure 3.9 – Posterior distributions derived from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using 8 chains with $5{\cdot}10^5$ iterations of the (a) mean, $\overline{\mathcal{K}}$, and (b) standard deviation, $\sigma_{\mathcal{K}}$, of the log-hydraulic conductivity field. Posterior distributions of the (c) standard deviation of the measurement errors, $\sigma_{\tilde{\mathbf{Y}}}$, and the polynomial coefficients of the (d) constant, a_0, (e) linear, a_1 and (f) quadratic, a_2, terms describing the petrophysical relationships of Eqs. (3.9)-(3.11). Posterior distributions of the geostatistical parameters of the petrophysical prediction uncertainty field: (g) standard deviation, $\sigma_{\Delta\mathbf{p}}$, (h) integral scale along the major axis of anisotropy, $I_{\Delta\mathbf{p}}$, (i) anisotropy angle, $\varphi_{\Delta\mathbf{p}}$, (j) ratio of the integral scales along the minor and major axis of anisotropy, $R_{\Delta\mathbf{p}}$, and (k) shape parameter of the Matérn variogram, $\nu_{\Delta\mathbf{p}}$. The results for the perfect linear, scattered linear and scattered quadratic petrophysical relationship are depicted with black, green and blue lines, respectively. The red lines indicate the corresponding prior distributions. The densities in each plot are normalized.

In Fig. 3.10a-c, we display the mean posterior hydraulic conductivity fields in linear scale. The three fields show similar values close to the boreholes where flowmeter data are available but, away from these locations, the different petrophysical models lead to different subsurface structures and estimates (e.g., within the first meter below the ground surface and between borehole T2 and M3, Fig. 3.10a-c). Nevertheless, all the three hydraulic conductivity mean models depict a low-hydraulic conductivity zone at a depth of 1-2 m.b.s.l. and at 5-6 m.b.s.l. (Fig. 3.10a-c). When the petrophysical prediction uncertainty is ignored, the inferred hydraulic conductivity (Fig. 3.10a) and GPR velocity (Fig. 3.10g) fields are characterized by a high variability. On average, the standard deviations of the posterior hydraulic conductivity estimates are higher when petrophysical prediction uncertainty is accounted for (Fig. 3.10d-f).

We observe similarities between the corresponding posterior GPR mean velocities (Fig. 3.10g-i). For instance, they all show a low-velocity zone within the first 2 m.b.s.l, at 3 m.b.s.l. and at 5-6 m.b.s.l and a high-velocity zone at 4-5 m.b.s.l. As expected, the inferred velocity fields derived from scattered petrophysical relationships (Fig. 3.10h-i) are smoother than the case in which this uncertainty is ignored (Fig. 3.10g). The mean posterior fields of the petrophysical prediction uncertainty distributions (Fig. 3.10k-l) are very similar and correlated with the posterior velocity means.

The red lines in Fig. 3.11a-c depict the inferred mean petrophysical relationships and the scatter (black dots) around them represents the inferred mean petrophysical prediction uncertainty. The GPR velocity range appears to be overestimated whether $\Delta\mathbf{p}$ is ignored (Fig. 3.11a) or accounted for together with a quadratic petrophysical model (Fig. 3.11c), while a scattered linear petrophysical relationship (Fig. 3.11b) provides a velocity range in agreement with previous studies (Hubbard et al., 2001; Chen et al., 2001; Linde et al., 2008; Linde and Vrugt, 2013; Brunetti et al., 2017).

We now turn our attention to the Bayesian model selection results. We find that Model 2 (scattered linear relationship) has the largest evidence value (-260.20 in $\log_{10}$ units) and Model 1 ($\Delta\mathbf{p}$ are ignored) has the lowest one (-361.00) (Fig. 3.12). The Bayes factor for the "best" petrophysical model (Model 2) with respect to Model 1 and Model 3 is $10^{100.80}$ and $10^{9.38}$, respectively. These results confirm that the perfect petrophysical model (Model 1) is erroneous. Furthermore, the results suggest that the use of a more complex petrophysical relationship is not necessarily favoured. Even if predictions based on the quadratic petrophysical model (Model 3) fits the data slightly better than the linear petrophysical model (Model 2) (Fig. 3.9c), the highest evidence is found for Model 2. This is a consequence of the trade-off between parsimony and goodness of fit typical of the Occam's razor principle on which Bayesian model selection is based.

3.6 Discussion

Our coupled Bayesian hydrogeophysical inversion approach with explicit inference of spatially-correlated petrophysical prediction uncertainty leads to less bias (e.g., in the inferred variance of the inferred hydrogeological property field), more realistic uncertainty quantification and less over confident model selection compared to the common choice of ignoring this type of uncertainty. Even if our approach to infer petrophysical prediction uncertainty doubles the number of parameters in the inversion problem, we observe dramatic gains in sampling efficiency compared to MC-within-MCMC (e.g., Bosch (1999, 2016)). Moreover, DREAM$_{(ZS)}$ allows for parallel evaluation of the different Markov chains and, therefore, enables feasible computational times even in high (e.g., in our case, more than 200) model dimensions. Our synthetic and field-based case-studies suggest that it is not always possible to independently

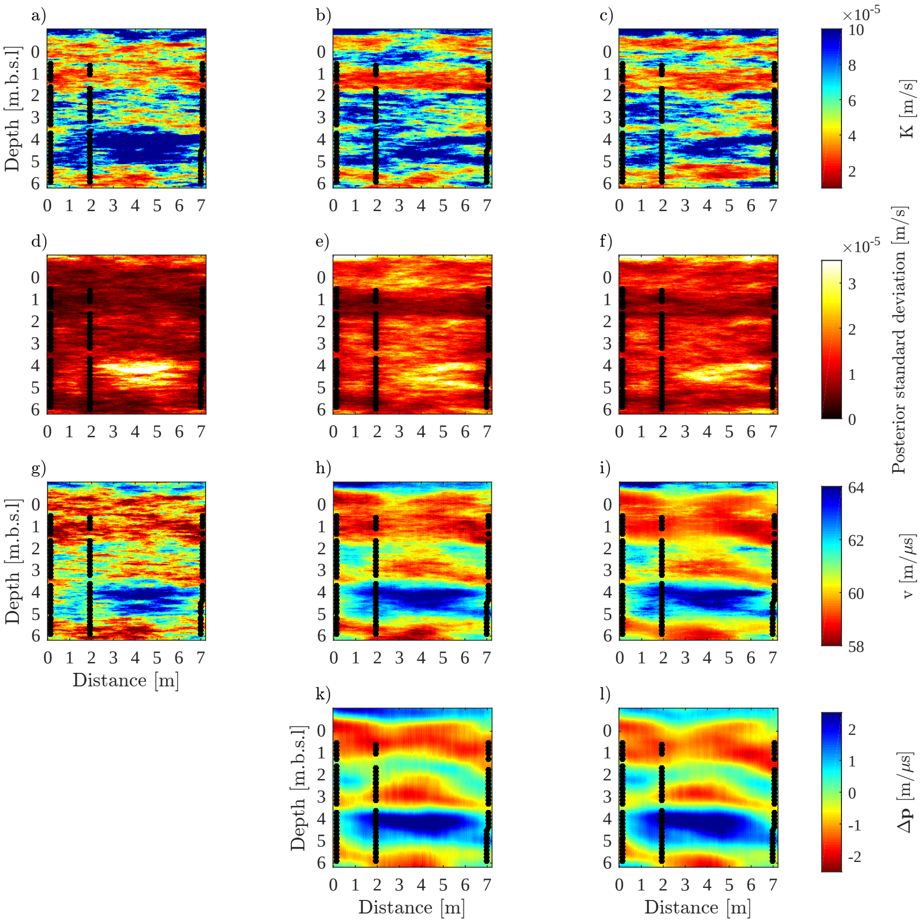


Figure 3.10 – Mean of the posterior hydraulic conductivity, *K*, realizations obtained using a (a) perfect linear, (b) scattered linear and (c) scattered quadratic petrophysical relationship with the corresponding (d)-(f) standard deviation of the posterior hydraulic conductivity estimates, respectively. Mean of the posterior GPR velocity realizations obtained using (g) perfect linear, (h) scattered linear and (i) scattered quadratic petrophysical relationships. Mean of the posterior petrophysical prediction uncertainty estimates for the (k) linear and (l) quadratic petrophysical relationship. The different measurement depths of the flowmeter data (black points) are indicated for boreholes S14 (on the left), T2 (in the middle) and M3 (on the right). The posterior distributions are computed from MCMC simulation with the DREAM$_{(ZS)}$ algorithm using 8 chains with $2.5{\cdot}10^5$ iterations.

constrain hydrogeological and petrophysical properties. This trade-off is particularly acute when the petrophysical prediction errors have similar geostatistical properties (e.g., orientations and integral scales) as the hydrogeological property field of interest (Fig. 3.7). A manifestation of this trade-off is given by the field application at the South Oyster Bacterial Transport Site, for which it was necessary to constrain the standard deviation of petrophysical prediction uncertainty and the standard deviation of the logarithm of hydraulic conductivity. Without such constraints, the inversion yields largely uncorrelated log-hydraulic conductivity and GPR velocity fields, results that are inconsistent with previous studies (Chen et al., 2001; Hubbard et al., 2001; Linde et al., 2008). This suggests that a careful petrophysical

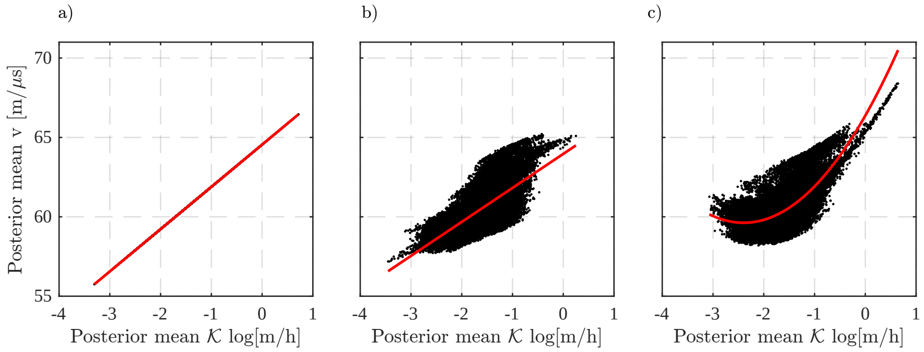


Figure 3.11 – Scatter plots of the mean posterior hydraulic conductivity estimates against the mean posterior GPR velocity estimates assuming a (a) perfect linear, (b) scattered linear and (c) scattered quadratic petrophysical relationship. The red lines depict the inferred mean petrophysical relationship, while the scatter represents the inferred mean petrophysical prediction uncertainty.

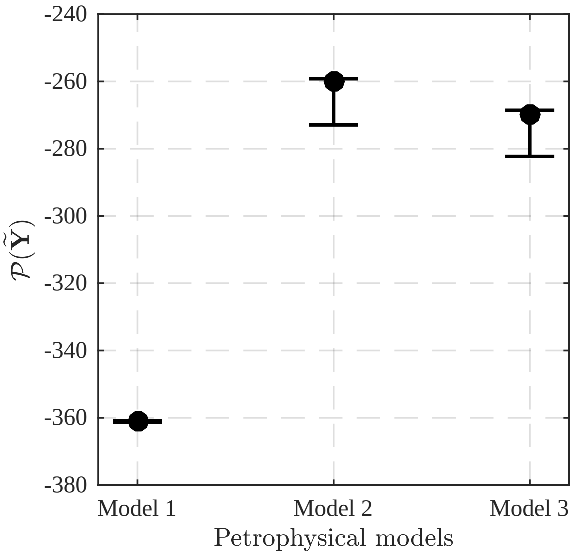


Figure 3.12 – Mean values of the evidence in $\log_{10}$ space, $\mathscr{P}(\widetilde{\mathbf{Y}})$, and corresponding uncertainty (error bars) derived from the Gaussian mixture importance sampling method for (Model 1) a perfect linear petrophysical relationship as shown in Eq. (3.9), (Model 2) scattered linear petrophysical relationship such that $\Delta\mathbf{p}$ is taken into account as shown in Eqs. (3.9) and (3.11), (Model 3) scattered quadratic petrophysical relationship where $\Delta\mathbf{p}$ is taken into account as shown in Eqs. (3.10)-(3.11).

analysis involving borehole data or literature reviews are needed to define constraining prior information when performing coupled hydrogeophysical inversion of field data.

In a previous study on Bayesian hydrogeophysical inversion model selection that ignored petrophysical prediction uncertainty (Brunetti et al., 2017), it was found that the typically large data sets encountered in geophysics and the assumption of small uncorrelated data errors (Gaussian likelihood) lead to very strong confidence in the ability of geophysical data to discriminate between alternative conceptual hydrogeological models. By including spatially-correlated petrophysical prediction uncertainty, we find for a synthetic example (Fig. 3.5) that the magnitude of the Bayes factor of the "best" conceptual model relative to the worse one decreases by 63 orders of magnitude. Nevertheless, the comparison between Case 3 (petrophysical prediction errors ignored) and Case 4 (petrophysical prediction errors accounted for) in Fig. 3.5a and Table 3.3 still indicates high Bayes factors and a practically-speaking unique ability of geophysical data to find the most appropriate conceptual hydrogeological model among a set of candidates. In the future, one should also account for the effect of modelling errors (i.e., the discrepancy between actual physical responses and those simulated with simplified physics; here, a ray-based approximation in the present study instead of a full solution of the Maxwell's equations). A number of promising approaches to address modelling errors are available (Brynjarsdóttir and O'Hagan, 2014; Hansen et al., 2014; Xu and Valocchi, 2015). Accounting for modelling errors is an essential next step to achieve reliable Bayesian hydrogeophysical model selection; we anticipate that this will further decrease the range of Bayes factors.

Bayesian model selection at the South Oyster Bacterial Transport Site (Section 3.5) demonstrates clearly that the relationship between log-hydraulic conductivity and GPR velocity is not a perfect relationship. That is, the petrophysical model with a scattered linear relationship has a much higher evidence than results obtained by assuming a perfect linear relationship. However, contrasting results were obtained in the synthetic example of Section 3.4.3 that did not involve any hydrogeological point measurements. In that case, formal Bayesian model selection erroneously favoured a conceptual model that ignored petrophysical prediction uncertainty. This happens because this conceptual model has fewer parameters and is still able to fit the data well, albeit with a porosity model with biased variance. At the South Oyster Bacterial transport Site, we condition all model proposals to point data (flowmeter estimates of hydraulic conductivity) and it is then not possible to propose a biased model close to the boreholes. Hence, the scattered petrophysical relationship is preferred. However, even if the inclusion of point conditioning in the synthetic example (not shown) decreased the Bayes factor, the model selection still favoured the wrong conceptual model. In the synthetic example, we considered boreholes at the left and right sides of the model domain, and the relative petrophysical prediction uncertainty was much smaller than for the field example. This could explain why the inconsistency between point data and GPR data is more evident for the field example, which led the Bayesian model selection to favour a model with petrophysical prediction uncertainty. These findings suggest that MCMC inversion and model selection is not always able to identify the "right" model and that their outputs need to be treated with some caution. The more prior information that is available (e.g., on petrophysical prediction uncertainty in terms of variance and correlation scale), the more reliable are the results. Indeed, Bayesian model selection is built on the principle of Occam's razor and a problem-specific and conceptual-model specific level of informative data is needed to overcome this tendency to favour a simpler, but erroneous conceptual model (e.g., Schöniger et al. (2015a)).

In this study, we have made the choice to infer for petrophysical prediction uncertainty, instead of accounting for its effects in the likelihood function. For linear theory, it is indeed possible to propagate the impact of (multi-Gaussian) petrophysical errors and add the corresponding covariance matrices to the data covariance matrix (Bosch (2004); Bosch et al. (2009); Bosch (2016); Chen and Dickens (2009)). This is not possible for non-linear theory, as the resulting impact of petrophysical uncertainty on the data leads to model-dependent non-Gaussian distributions. The corresponding problem formulation and ways to address this problem was recently discussed by Linde et al. (2017) in their Section 5.2. In the future, it would be interesting to compare these two approaches (i.e., inferring for petrophysical uncertainty (this study) or accounting for the effect of petrophysical uncertainty in the likelihood function).

3.7 Conclusions

We have demonstrated the importance of accounting for petrophysical prediction uncertainty in coupled hydrogeophysical inversion and highlighted the critical role played by its spatial correlation. As MCMC inversions are primarily performed to enable accurate uncertainty quantification, we suggest that petrophysical prediction uncertainty should be accounted for in future hydrogeophysical studies. In this work, we parameterize the petrophysical prediction uncertainty as a multi-Gaussian field that is inferred together with hydrogeological target properties. To decrease model dimensionality, future work should also focus on developing computationally efficient and accurate approaches to account for this uncertainty in the likelihood function.

Inferring petrophysical prediction uncertainty with MCMC leads to dramatic performance gains compared to previous work, in which it has been accounted for by Monte Carlo sampling. In our examples, we show that ignoring petrophysical prediction uncertainty and (above all) its spatial correlation causes bias in the inferred variance of the hydrogeological properties, which implies overly variable fields. Accounting for this error source allows for consistent hydrogeological estimates and widens the estimated posterior distributions. However, the geostatistical model describing petrophysical prediction uncertainty is only partially recoverable by the inversion. When performing Bayesian model selection, accounting for petrophysical prediction uncertainty reduces overconfidence in the ability of geophysical data to discriminate between conceptual hydrogeological models of the subsurface. When considering geophysical data alone, there is a risk that Bayesian hydrogeophysical model selection will favour a model parameterization that ignores petrophysical prediction uncertainty provided that the resulting overly variable hydrogeological estimates can explain the geophysical data well. This highlights the importance of including constraining prior information about petrophysical prediction uncertainty and the value of combining geophysical and hydrogeological data in the inversion.

Chapter 4

Hydrogeological model selection among complex spatial priors with application to the MADE-5 tracer experiment

Carlotta Brunetti, Marco Bianchi, Guillaume Pirot and Niklas Linde.

Submitted [1] to *Water Resources Research*

[1] Brunetti, C., Bianchi, M., Pirot, G., and Linde, N., (2018). Hydrogeological model selection among complex spatial priors with application to the MADE-5 tracer experiment. *Water Resources Research*

4.1 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 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. 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.

4.2 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. (2015b); 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., 2015b; 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. (2015b)). 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, 2008), stepping-stone (Xie et al., 2011) and nested sampling (Skilling, 2004; Skilling et al., 2006). 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.

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 tracer test (MADE-5) (Bianchi et al., 2011a).

4.3 Theory

4.3.1 Bayesian inference and model selection

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

$$p(\boldsymbol{\theta}|\widetilde{\mathbf{Y}}) = \frac{p(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\widetilde{\mathbf{Y}})}{p(\widetilde{\mathbf{Y}})}. \tag{4.1}$$

The prior pdf, $p(\boldsymbol{\theta})$, quantifies all the information that is available about the model parameters before considering the observed data. Typically, $p(\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$ 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 $p(\boldsymbol{\theta})$. The likelihood function, $L(\boldsymbol{\theta}|\widetilde{\mathbf{Y}})$, 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, $\sigma_{\widetilde{\mathbf{Y}}}$,

$$L(\boldsymbol{\theta}|\widetilde{\mathbf{Y}}) = \left(\sqrt{2\pi\sigma^2_{\widetilde{\mathbf{Y}}}}\right)^{-n} \exp\left[-\frac{1}{2}\sum_{h=1}^{n}\left(\frac{\mathscr{F}_h(\boldsymbol{\theta}) - \widetilde{y}_h}{\sigma_{\widetilde{\mathbf{Y}}}}\right)^2\right]. \tag{4.2}$$

As the residuals between the simulated forward responses, $\mathscr{F}_h(\boldsymbol{\theta})$, and the observed data, $\widetilde{y}_h$, becomes smaller, the likelihood increases. The denominator in Bayes' theorem is the evidence (or marginal likelihood), $p(\widetilde{\mathbf{Y}})$, and it is the cornerstone quantity in Bayesian model selection problems. 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, η, with model parameters, $\boldsymbol{\theta}$, and prior distribution, $p(\boldsymbol{\theta}|\eta)$, has generated the data $\widetilde{\mathbf{Y}}$,

$$p(\widetilde{\mathbf{Y}}|\eta) = \int L(\boldsymbol{\theta}, \eta|\widetilde{\mathbf{Y}}) p(\boldsymbol{\theta}|\eta) d\boldsymbol{\theta}. \tag{4.3}$$

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

$$B_{(\eta_1,\eta_2)} = \frac{p(\widetilde{\mathbf{Y}}|\eta_1)}{p(\widetilde{\mathbf{Y}}|\eta_2)}. \tag{4.4}$$

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. 4.3 must be approximated by numerical means. In this work, the different conceptual models represent alternative spatial representations of hydraulic conductivity in the subsurface.

4.3.2 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. 4.3) 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, 2008). The power posterior distribution is proportional to the prior pdf multiplied by the likelihood function raised to the power of $\beta \in [0,1]$:

$$p_\beta(\theta|\widetilde{\mathbf{Y}}) \propto p(\theta)L(\theta|\widetilde{\mathbf{Y}})^\beta. \tag{4.5}$$

Decreasing β has the effect of flattening the likelihood function. For $\beta = 1$, the posterior distribution is sampled, $p_1(\theta|\widetilde{\mathbf{Y}}) \propto p(\theta)L(\theta|\widetilde{\mathbf{Y}})$; for $\beta = 0$, the prior distribution is sampled, $p_0(\theta|\widetilde{\mathbf{Y}}) \propto p(\theta)$. In thermodynamic integration and stepping-stone sampling, a sequence of β-values are defined (see Section 4.3.2). 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 (2008). 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 4.3.2-4.3.2).

Thermodynamic integration

Thermodynamic integration reduces the multi-dimensional integral of Eq. 4.3 into a one-dimensional integral of the expectation of the log-likelihood, $\mathscr{L}(\theta|\widetilde{\mathbf{Y}}) \equiv \log L(\theta|\widetilde{\mathbf{Y}})$, as:

$$\log p(\widetilde{\mathbf{Y}}|\eta) = \int_0^1 \mathrm{E}_{\theta|\widetilde{\mathbf{Y}},\beta}\left[\mathscr{L}(\theta|\widetilde{\mathbf{Y}},\eta)\right] d\beta. \tag{4.6}$$

For the derivation of Eq. 4.6, we refer to Friel and Pettitt (2008) and Lartillot and Philippe (2006). The integral in Eq. 4.6 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 $\ell_j \equiv \mathrm{E}_{\theta|\widetilde{\mathbf{Y}},\beta_j}\left[\mathscr{L}(\theta|\widetilde{\mathbf{Y}},\eta)\right]$ and their corresponding variances as $\sigma_j^2 \equiv \mathrm{V}_{\theta|\widetilde{\mathbf{Y}},\beta_j}\left[\mathscr{L}(\theta|\widetilde{\mathbf{Y}},\eta)\right]$. In this work, we use the corrected composite trapezoidal rule:

$$\log p(\widetilde{\mathbf{Y}}|\eta) \approx \sum_{j=2}^{J} \frac{(\beta_j - \beta_{j-1})}{2}(\ell_j + \ell_{j-1}) - \sum_{j=2}^{J} \frac{(\beta_j - \beta_{j-1})^2}{12}(\sigma_j^2 - \sigma_{j-1}^2), \tag{4.7}$$

which provides more accurate estimates compared with the classical composite trapezoidal rule (first term in Eq. 4.7) as it also considers the second-order correction term (second term in Eq. 4.7). 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, their number, J, (details provided in Section 4.3.2) and the number and the correlation of samples, N, of the power posteriors obtained by MCMC. These uncertainties 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:

$$N_{\mathrm{eff},j} = \frac{N_j}{1 + 2\sum_{z=1}^{\infty} \rho_j(z)}, \tag{4.8}$$

where $\rho_j(z)$ is the autocorrelation at lag z. Applying the rules for uncertainty propagation to the first leading term in Eq. 4.7 and assuming the errors of ℓ_j to be independent of those associated to ℓ_{j-1}, the sampling error is:

$$\sigma_s^2 = \sum_{j=2}^{J} \frac{(\beta_j - \beta_{j-1})^2}{4}\left(\frac{\sigma_j^2}{N_{\mathrm{eff},j}} + \frac{\sigma_{j-1}^2}{N_{\mathrm{eff},j-1}}\right). \tag{4.9}$$

Discretisation errors arise as the continuous integral of Eq. 4.6 is estimated using a finite number of evaluation points (Eq. 4.7). 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. 4.6 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:

$$\sigma_d^2 = \sum_{j=2}^{J} \frac{(\beta_j - \beta_{j-1})^2}{4} (\ell_j - \ell_{j-1})^2. \tag{4.10}$$

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

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):

$$r = \frac{p(\widetilde{\mathbf{Y}}|\eta, \beta = 1)}{p(\widetilde{\mathbf{Y}}|\eta, \beta = 0)}. \tag{4.11}$$

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:

$$r = \prod_{j=2}^{J} r_{j-1} = \prod_{j=2}^{J} \frac{p(\widetilde{\mathbf{Y}}|\eta, \beta_j)}{p(\widetilde{\mathbf{Y}}|\eta, \beta_{j-1})}. \tag{4.12}$$

Then, importance sampling is applied to the numerator and denominator of Eq. 4.12 using the power posterior $p_{\beta_{j-1}}(\theta|\widetilde{\mathbf{Y}})$ as the importance distribution:

$$r_{j-1} = \frac{1}{N} \sum_{i=1}^{N} L(\theta_{j-1,i}|\widetilde{\mathbf{Y}})^{\beta_j - \beta_{j-1}} \tag{4.13}$$

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

$$\log p(\widetilde{\mathbf{Y}}|\eta) = \sum_{j=2}^{J} \log r_{j-1} = \sum_{j=2}^{J} \log \left\{ \frac{1}{N} \sum_{i=1}^{N} \exp\left[(\beta_j - \beta_{j-1}) \cdot \mathscr{L}(\theta_{j-1,i}|\widetilde{\mathbf{Y}}) \right] \right\}. \tag{4.14}$$

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:

$$\widehat{\text{Var}} \log p(\widetilde{\mathbf{Y}}|\eta) = \sum_{j=2}^{J} \frac{1}{N_{\text{eff},j-1} \cdot N} \sum_{i=1}^{N} \left(\frac{L(\theta_{j-1,i}|\widetilde{\mathbf{Y}})^{\beta_j - \beta_{j-1}}}{r_{j-1}} - 1 \right)^2. \tag{4.15}$$

The derivation of Eq. 4.14 and 4.15 appears in Xie et al. (2011), and interested readers are referred to this publication for further details. The only difference in our Eq. 4.15 is that we consider the effective sample size as defined in Eq. 4.8.

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. This rapid change is enhanced when large and informative data sets are used. As a consequence, the accuracy of the evidence estimates increases when placing most of the β-values close to 0. 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. 4.6). 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 (2008) are the first to employ a discretisation scheme of β-values that follows a power law spacing as:

$$\beta_j = \left(\frac{j-1}{J-1}\right)^c \quad \text{with} \quad j = 1, 2 \ldots, J. \tag{4.16}$$

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).

4.4 Method

4.4.1 General framework

It is common to sample the unnormalised posterior pdf of Eq. 4.1 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:

$$\alpha = min\left\{1, \frac{L(\theta_{\mathbf{prop}}|\widetilde{\mathbf{Y}})}{L(\theta_{\mathbf{cur}}|\widetilde{\mathbf{Y}})}\right\}. \tag{4.17}$$

To sample the power posteriors, we simply modify the extended Metropolis acceptance criteria by raising the likelihoods in Eq. 4.17 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.

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 $\beta$
  distributed according to Eq. 4.16; a training image
Output: $\Lambda_j$, 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. 4.2);
for $j = 1,\ldots, J$ do
    for $t = 2,\ldots, T$ 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. 4.2);
        Accept $\theta_{\mathbf{prop}}$ with probability, $\alpha = min\left\{1, \frac{L(\theta_{\mathbf{prop}}|\widetilde{\mathbf{Y}})^{\beta_j}}{L(\theta_{\mathbf{cur}}|\widetilde{\mathbf{Y}})^{\beta_j}}\right\}$;
        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 $\Lambda_j$;
        Set $t$=$t$+1;
    end
end
Compute $\log p(\widetilde{\mathbf{Y}}|\eta)$ (Eqs. 4.7 and 4.14) and corresponding variances (Eqs. 4.9-4.10 and
  4.15) using the information stored in $\Lambda_j$ after the removal of the burn-in period.
```

4.4.2 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. (2017b,a). The main steps in the Graph Cuts algorithm are depicted in Figure 4.1. Basically, a section of the same size as the model domain, $\theta_{\mathbf{new}}$ (Figure 4.1b), is randomly drawn from the training image and the absolute difference between $\theta_{\mathbf{new}}$ and the current model realisation, $\theta_{\mathbf{cur}}$ (Figure 4.1a), is computed and raised to the power of the cost power, δ_{cp}, (Pirot et al., 2017a) to obtain the cost image, $\boldsymbol{\delta} = |\theta_{\mathbf{cur}}\text{-}\theta_{\mathbf{new}}|^{\delta_{cp}}$ (Figure 4.1d). Two distinct regions of high cost, similar size and containing at least p pixels are randomly selected (Figure 4.1e). To choose these terminals, Pirot et al. (2017b) 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 4.1f) and the new model proposal, $\theta_{\mathbf{prop}}$ (Figure 4.1c), is built by cutting the patch from $\theta_{\mathbf{new}}$ and replacing the corresponding area in $\theta_{\mathbf{cur}}$.

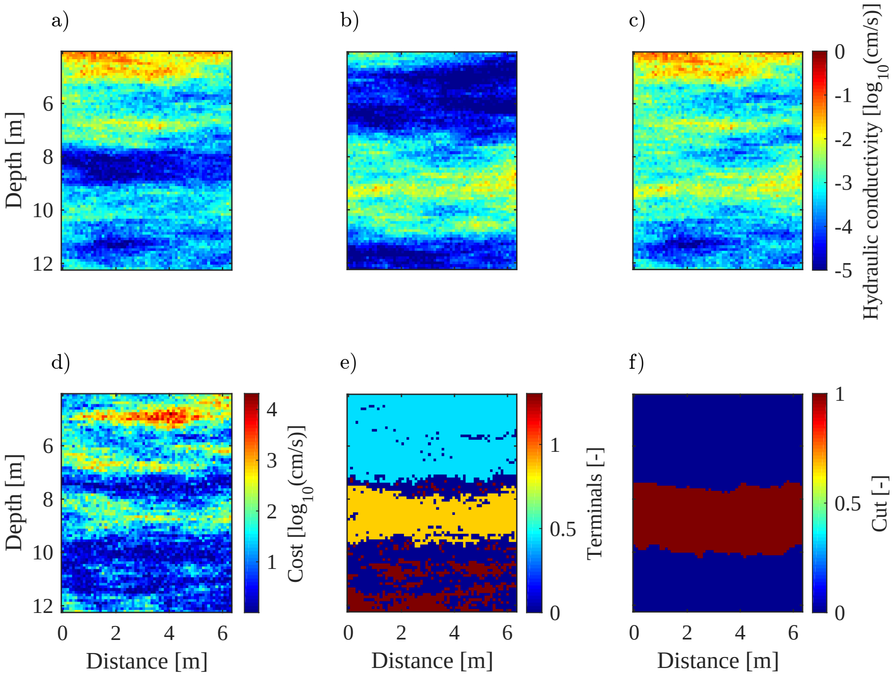


Figure 4.1 – Illustration of how model proposals are obtained using the Graph Cuts algorithm. (a) Current model realisation, $\theta_{\mathbf{cur}}$, (b) section drawn randomly from the training image, $\theta_{\mathbf{new}}$, and (c) the resulting model proposal, $\theta_{\mathbf{prop}}$. This model 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) $\theta_{\mathbf{prop}}$.

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.

4.4.3 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 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). The structure is thought to be made up of a highly permeable network of 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). Before tracer injection, a steady-state dipole flow field was established by injecting clean water into a well and by simultaneously abstracting groundwater from another well located 6 m apart form the injection well. 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. Between the injection and extraction wells, two multi-level sampler (MLS) wells are installed for monitoring the temporal and spatial evolution of the tracer plume. In particular, bromide concentrations 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). The forward model used to simulate the flow and transport during the MADE-5 experiment and a simple 3D to 2D transformation of the data is described in Section 4.8.

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 4.2). The multi-Gaussian training image in Figure 4.2a 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 4.1) 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 4.2b-d were generated following Linde et al. (2015a). The highly conductive and connected channels in an homogeneous matrix (Figure 4.2b) 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 4.2c 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 4.2.

The training image in Figure 4.2d 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. (2015a) and

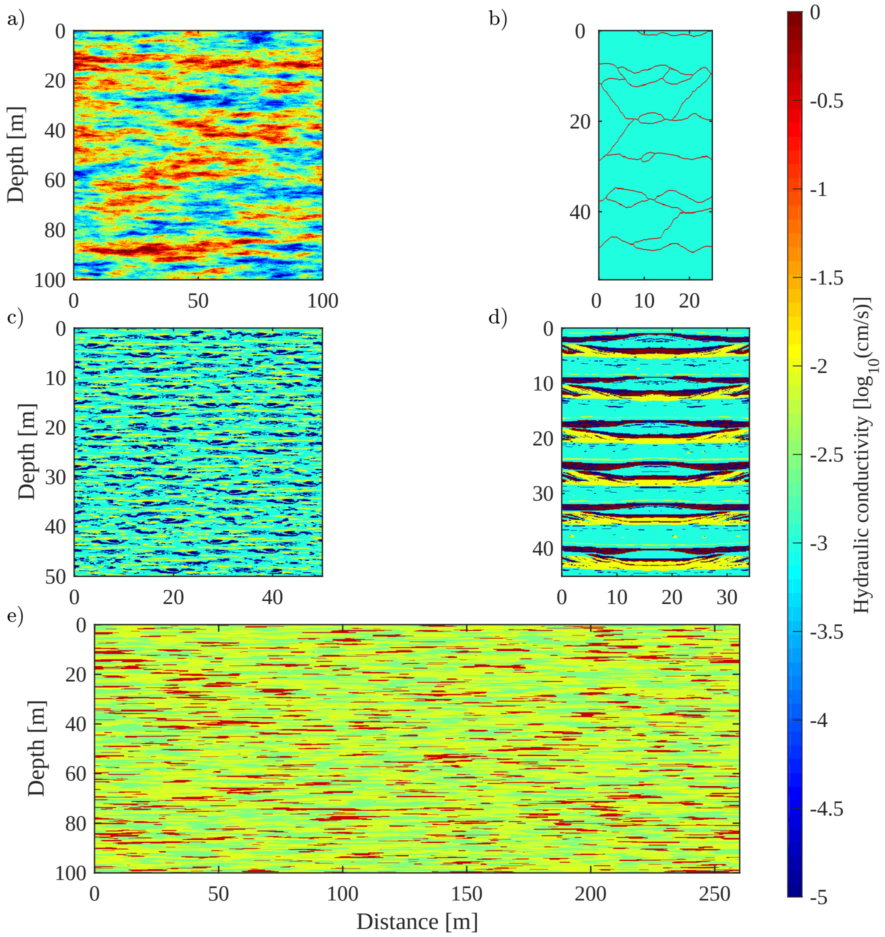


Figure 4.2 – Training images used in the MPS-based inversion to represent spatial hydraulic conductivity of the MADE site: (a) multi-Gaussian field (Bianchi et al., 2011a), (b) highly conductive channels in an homogeneous matrix (Strebelle, 2002; Ronayne et al., 2010; Linde et al., 2015a), (c) model based on a mapping study of a MADE outcrop (Rehfeldt et al., 1992; Linde et al., 2015a), (d) model based on a mapping study at the Herten site in Germany (Bayer et al., 2011; Comunian et al., 2011; Linde et al., 2015a) featuring representative alluvial deposit structures and (e) model based on lithological borehole data collected at the MADE site (Bianchi and Zheng, 2016).

Lochbühler et al. (2014b) used the training image of Figure 4.2d 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 4.2).

The training image of Figure 4.2e is built based on five hydrogeological facies identified from lithological borehole data at the MADE site (Bianchi and Zheng, 2016) and reported in Table 4.3. This training image is a stochastic unconditional realisation that was generated following Bianchi and Zheng (2016).

Table 4.1 – Geostatistical parameters of the multi-Gaussian training image (Figure 4.2a) 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 4.2 – 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 4.2c-d.

Facies	$\log_{10}$ K [cm/s]
Open framework gravel	$-6.83\cdot10^{-4}$
Sand	-2.00
Undifferentiated sandy gravel	-3.00
Sandy, clayey gravel	-5.00

Table 4.3 – 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 4.2e.

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

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 4.2b) 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 4.2a; (2) *hybrid* that consists of the highly conductive channels of Figure 4.2b overlaid on the multi-Gaussian background of Figure 4.2a; (3) *outcrop-based* built from the training image in Figure 4.2c; (4) *analog-based* built from the training image in Figure 4.2d; (5) *lithofacies-based* built from the training image in Figure 4.2e. 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. Instead, the focus is on a versatile methodology that enables comparison of widely different conceptual models.

4.4.4 Evidence estimation in practice

We discretise the power coefficients β using the commonly used power law of Eq. 4.16 (Grzegorczyk et al., 2017; Höhna et al., 2017; Baele and Lemey, 2013; Xie et al., 2011; Calderhead and Girolami, 2009; Friel and Pettitt, 2008). 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. 4.7), as well as with the stepping-stone sampling method (Eq. 4.14). In order to correctly estimate the uncertainty of the evidence estimates, the effective sample size (Eq. 4.8) in each chain needs to be assessed. When evaluating Eq. 4.8, 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

Table 4.4 – 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 [%]			$\sigma_{\tilde{\mathbf{Y}}}$ [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

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, SE = $\sqrt{\widehat{\mathrm{Var}} \log p(\tilde{\mathbf{Y}}|\eta)}$ with corresponding 95% confidence intervals. The variances $\widehat{\mathrm{Var}} \log p(\tilde{\mathbf{Y}}|\eta)$ are computed using Eqs. 4.9-4.10 for the thermodynamic integration and using Eq. 4.15 for the stepping-stone sampling method.

4.5 Results for the MADE-5 case study

4.5.1 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 4.3). Each set of prior realisations shows considerable spatial variability and is in broad agreement with the original training image (Figure 4.2). This is valid for both continuous (Figure 4.3b), categorical (Figures 4.3c-e) and hybrid conceptual models (Figure 4.3a).

The posterior distributions (i.e., $\beta = 1$) are obtained by assuming that the standard deviation of the measurement errors, $\sigma_{\tilde{\mathbf{Y}}}$ [mg/L], follows a log-uniform prior distribution in the range [1,10] mg/L (last column of Table 4.4). The lowest mean of the inferred $\sigma_{\tilde{\mathbf{Y}}}$ is obtained for the hybrid conceptual model (5.8 mg/L) suggesting that this model enables the best match with the data. The highest $\sigma_{\tilde{\mathbf{Y}}}$ is found for the outcrop-based model (9.4 mg/L). The acceptance rates are lower (second column in Table 4.4) 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 4.4) as described in Section 4.4.4.

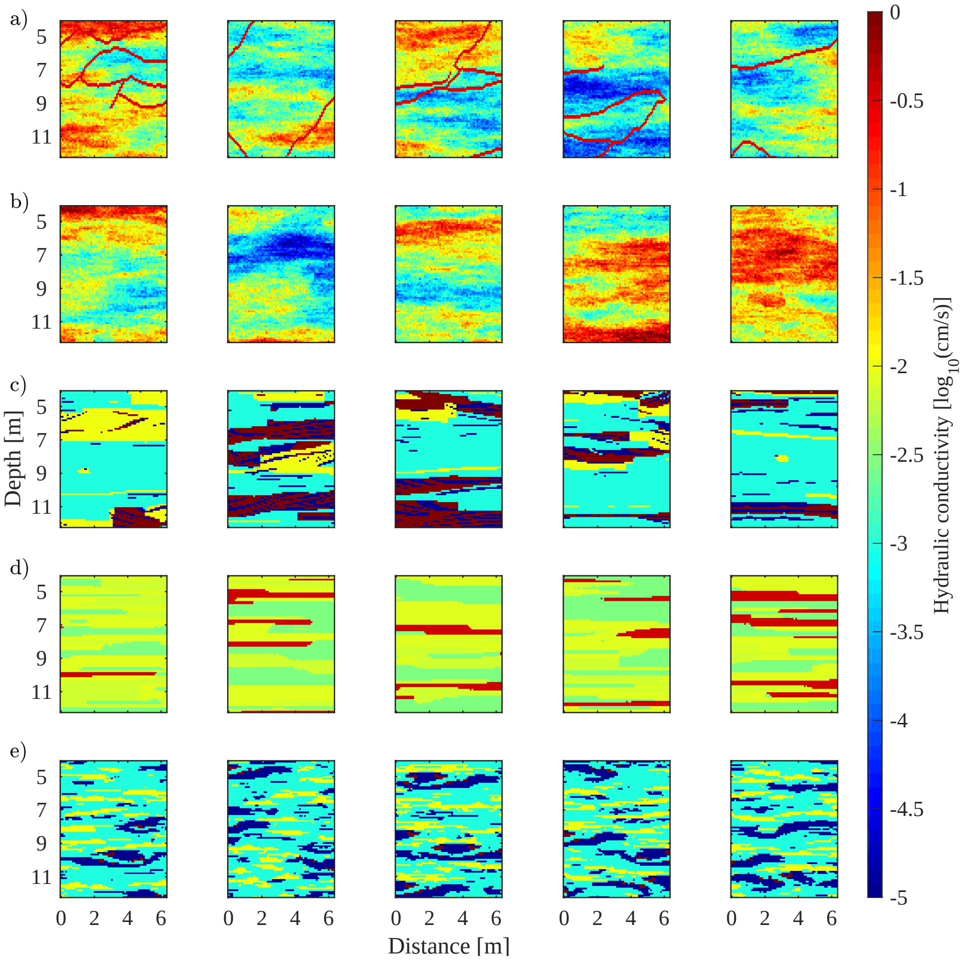


Figure 4.3 – Five prior realisations of hydraulic conductivity fields generated from the training images of Figure 4.2 with the Graph Cuts algorithm for the (a) hybrid, (b) multi-Gaussian, (c) analog-based, (d) lithofacies-based and (e) outcrop-based conceptual model of the MADE site.

The different conceptual models provide quite different posterior distributions of the hydraulic conductivity field (Figure 4.4), 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 4.4). The analog- and outcrop-based conceptual models exhibit more variability in the inferred hydraulic conductivity values (Figures 4.4c and 4.4e) with respect to the others and the lithofacies-based conceptual model is characterised by the smallest posterior standard deviations (Figure 4.4d). 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 4.4 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.

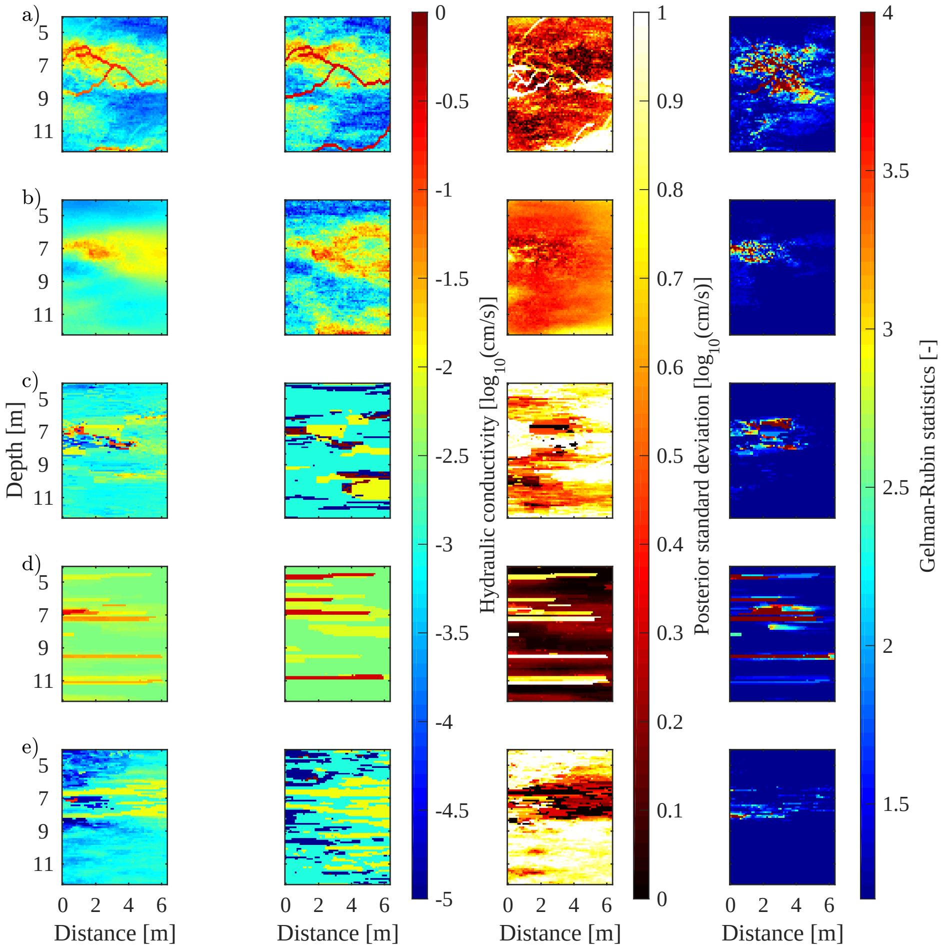


Figure 4.4 – Mean (first column), maximum a-posteriori (second column), and standard deviation (third column) of the posterior hydraulic conductivity realisations for the (a) hybrid, (b) multi-Gaussian, (c) analog-based, (d) lithofacies-based and (e) outcrop-based conceptual model at the MADE site. In 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.

In Figure 4.5, 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 4.5a) and MLS-2 (Figure 4.5b) 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 4.5c) and MLS-2 (Figure 4.5d). Note that the range of measured concentration values spans two orders of magnitude (Figure 4.5). 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 4.5d. Corresponding plots at all measurement locations are found in Section 4.9. 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.

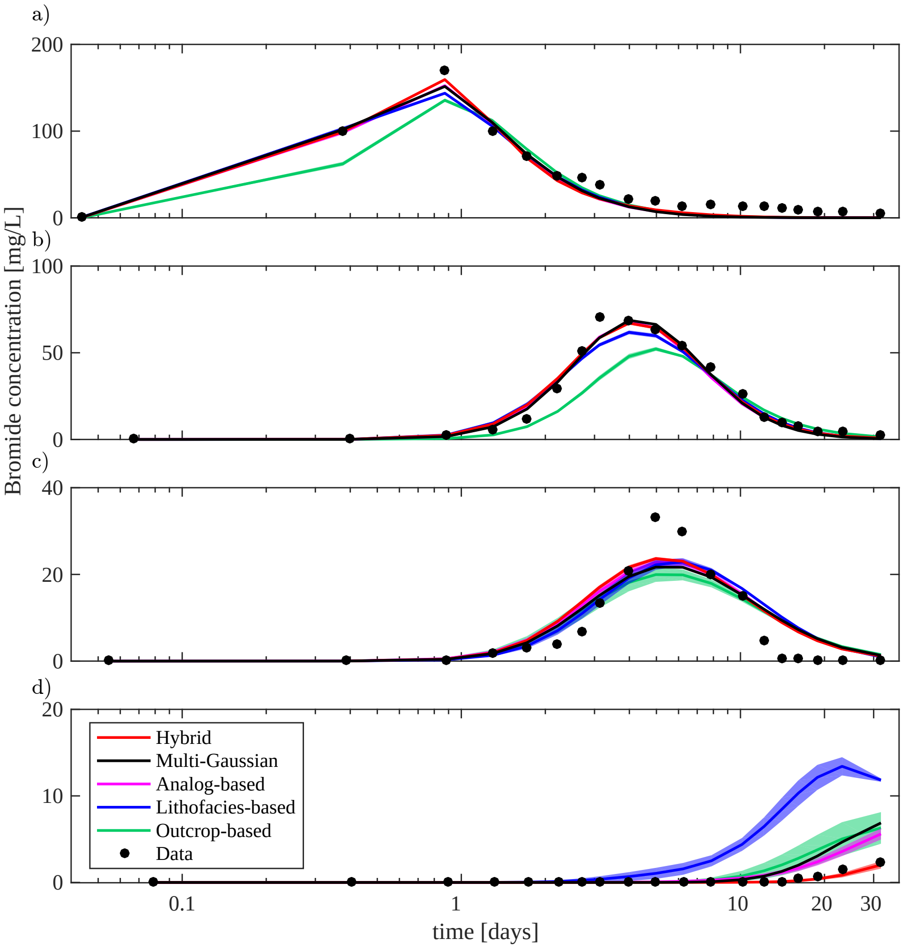


Figure 4.5 – Simulated (solid lines) and measured (black dots) bromide breakthrough curves from the MADE-5 experiment in the two monitoring wells MLS-1 and MLS-2 at a depth of 7 m (a-b) and 11 m (c-d), respectively. The simulated breakthrough curves are summarised by the mean of the posterior realisations (solid lines) and their 95% confidence intervals (shaded areas).

4.5.2 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 4.6). 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 4.6). 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 4.6 and Table 4.5). 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 4.8). For this reason, the width of the confidence intervals obtained by thermodynamic integration does not reduce significantly with increasing numbers of MCMC iterations (Figure 4.6).

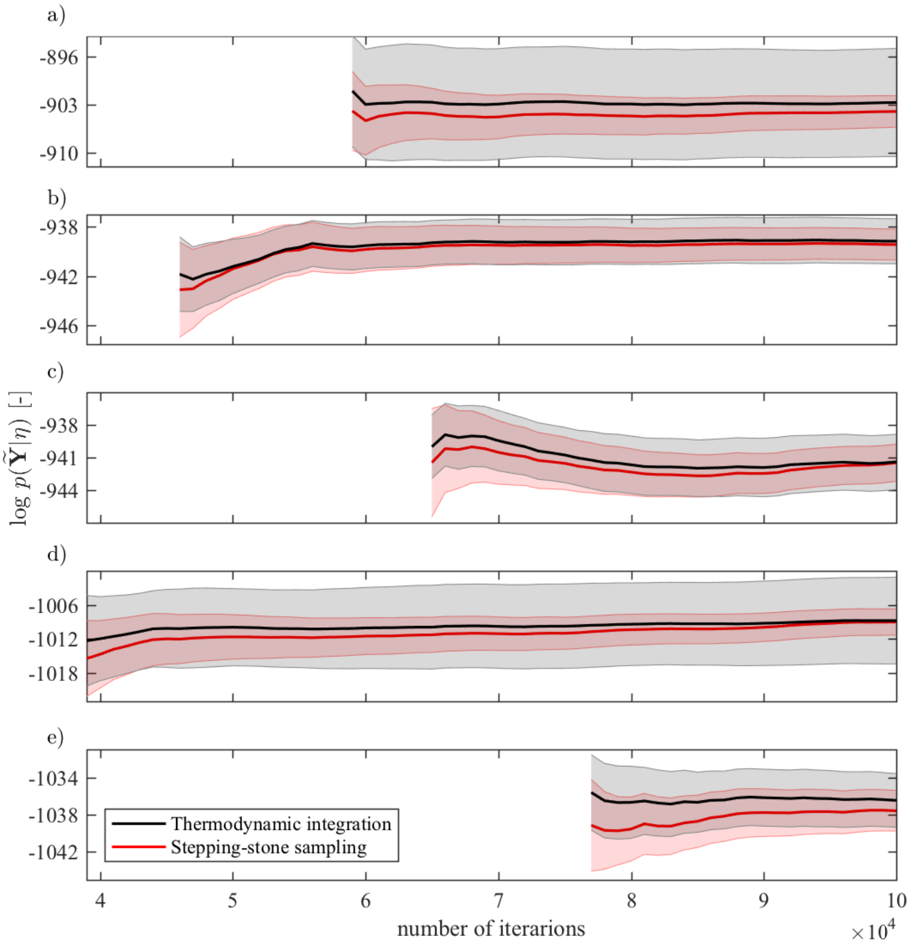


Figure 4.6 – Natural logarithm of the evidence estimates, $\log p(\widetilde{\mathbf{Y}}|\eta)$, as a function of the number of MCMC iterations. Evidences are computed with the stepping-stone sampling method (red line) and the thermodynamic integration method based on the corrected trapezoidal rule (black 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).

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 4.5). The multi-Gaussian and the analog-based conceptual models have very similar evidence estimates and they are the second-best performing conceptual models (Table 4.5).

Table 4.5 – 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

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 4.7). 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 4.7) 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 4.7), 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 4.7).

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 4.8). 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 4.8c) and stepping-stone sampling (Figure 4.8d). 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. 4.16) ensures that the discretisation errors for small β are relatively small (i.e., between 10^{-10} and 10^{-3}, Figure 4.8b). The pronounced fluctuations of the discretisation errors especially for $\beta > 0.1$ (Figure 4.8b) 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

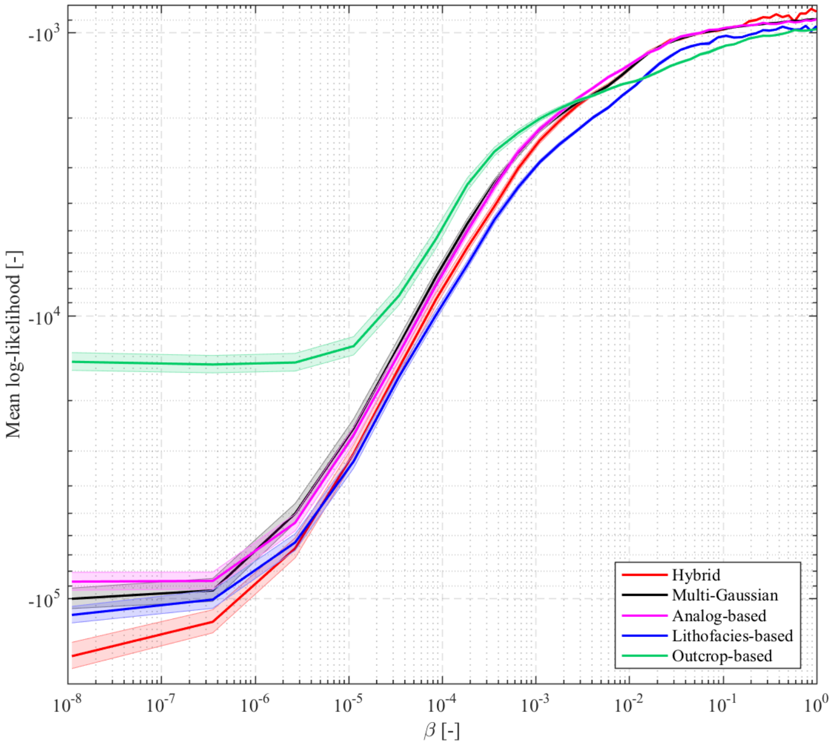


Figure 4.7 – Mean (line) of the natural logarithm of the likelihood functions, ℓ, computed for each β value and the 95% confidence interval of the ℓ-estimates (shaded areas). Note that the x- and y-axes are in $\log_{10}$ scale.

proposed by Kass and Raftery (1995) and we present twice the natural logarithm of the Bayes factors (Figures 4.9a-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 4.9a-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 4.9b) than the ones based on thermodynamic integration (Figure 4.9a). 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.

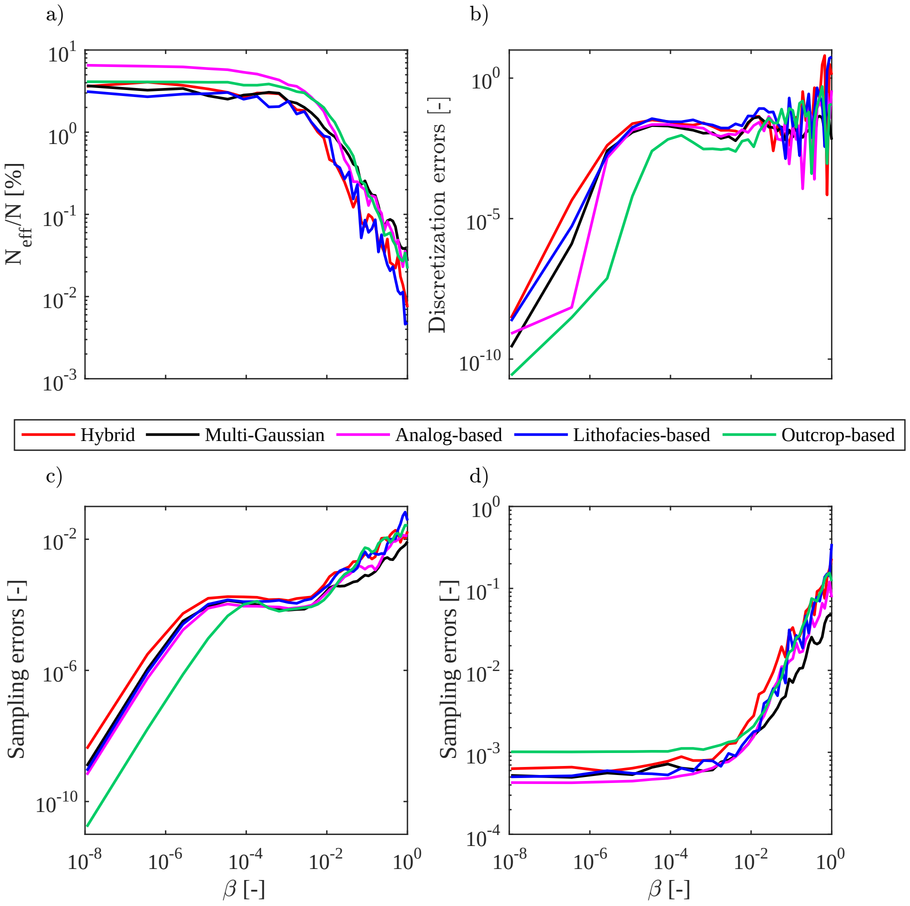


Figure 4.8 – (a) Percentage ratio between the effective and the total number of MCMC samples, (b) discretisation errors in the thermodynamic integration method (square root of Eq. 4.10), (c) sampling errors in the thermodynamic integration method (square root of Eq. 4.9) and (d) sampling errors in the stepping-stone sampling method (square root of Eq. 4.15) as a function of β-values. Note that all the x- and y-axes are in $\log_{10}$ scale.

4.6 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 4.3-4.4). Moreover, the Graph Cuts algorithm can include point conditioning (Li et al., 2016) even if this is not considered herein. We find that the hybrid conceptual model allows for the best fit of the observed breakthrough curves (Figure 4.5). 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

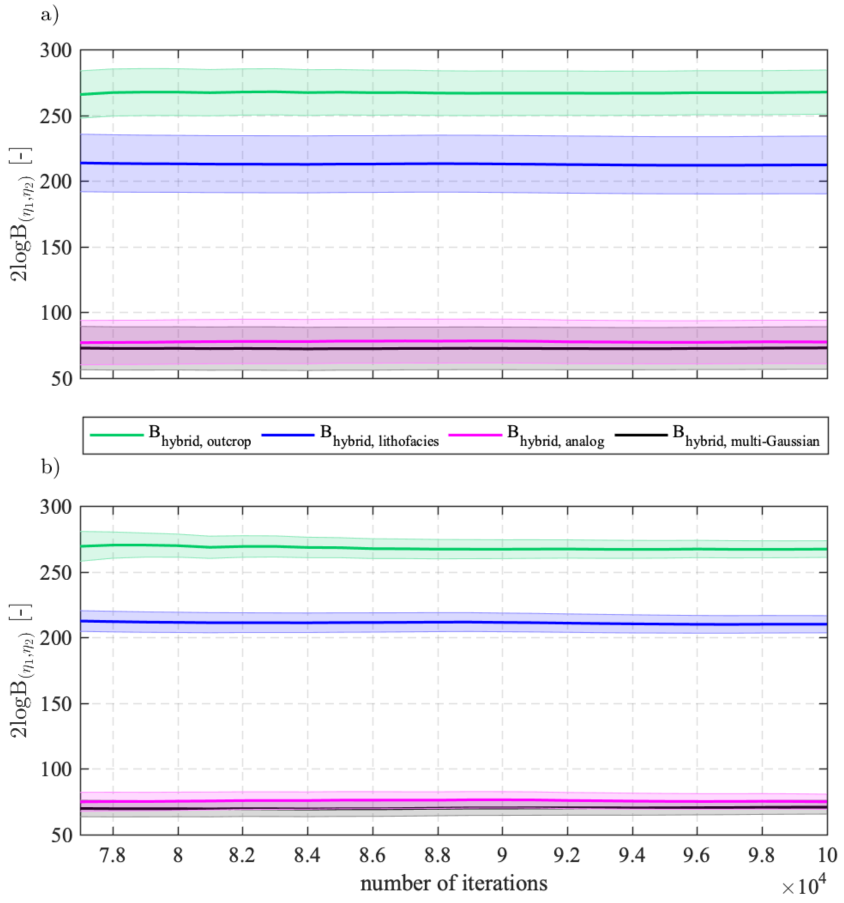


Figure 4.9 – Twice the natural logarithm of the Bayes factors of the "best model" (hybrid) with respect to the outcrop-based (green line), lithofacies-based (blue line), analog-based (magenta line) and multi-Gaussian (black line) conceptual model at the MADE site. Results are shown for (a) the thermodynamic 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.

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 4.5). However, all conceptual models have difficulties in fitting the observed BTCs. This is probably related to the fact that we ignore 3D heterogeneity. 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 4.4, Figure 4.6, Table 4.5). Linde et al. (2015a) 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. (2015a) 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, $\sigma_{\tilde{\mathbf{Y}}}$ (Table 4.4), 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 4.7) 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. It is also worth noting that the lithofacies-based conceptual model provided an excellent description of a large-scale tracer experiment (MADE-2) (Bianchi and Zheng, 2016) but did not perform equally to describe the small-scale heterogeneity involved in the MADE-5 test.

We find that stepping-stone sampling almost always provides slightly lower evidence estimates than thermodynamic integration (Table 4.5). 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. 4.7 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 4.7). We also highlight that the comparison between the uncertainty estimates of the evidence values provided by thermodynamic integration and stepping-stone sampling (Figure 4.6) 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. 4.6 with a rectangular rule.

Future work should better account for model errors caused by the 3D to 2D flow and transport approximation described in Section 4.8. 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 4.8a).

4.7 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 most suitable conceptual model of the MADE aquifer using a small-scale tracer test (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. Among the conceptual models considered for the MADE site, we find that the hybrid (highly conductive channels in a multi-Gaussian background) conceptual model is the best-performing one, followed by the multi-Gaussian and the analog-based conceptual model that is built based on outcrop studies at the Herten site in Germany. 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 MADE-5 tracer test, 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 more formal treatment of model errors due to the considered 3D to 2D approximation needs to be considered.

4.8 Appendix 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 4.10a-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 4.10a-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 4.10c). 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 4.10c) 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 4.10c).

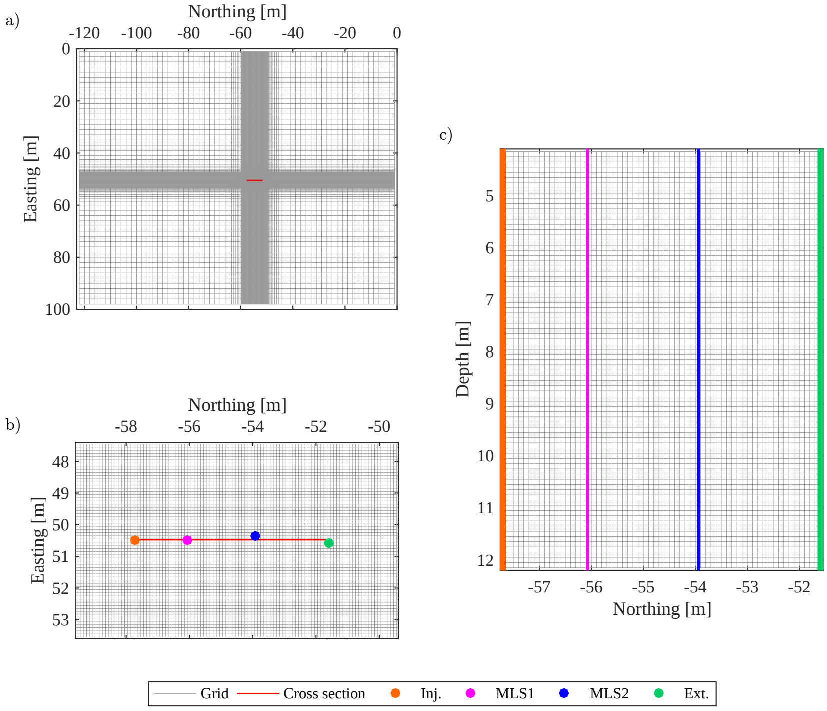


Figure 4.10 – (a) Aerial view of the 3D grid used for simulations with MODFLOW/MT3DMS; (b) zoom in the tracer test area, in which the grid size was refined to 0.1 m; (c) cross section used for simulations with MaFloT. The width of the lines in (c) is representative of the diameter of the four wells.

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 4.11). 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 4.11b-c) are removed (Figures 4.11d-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.

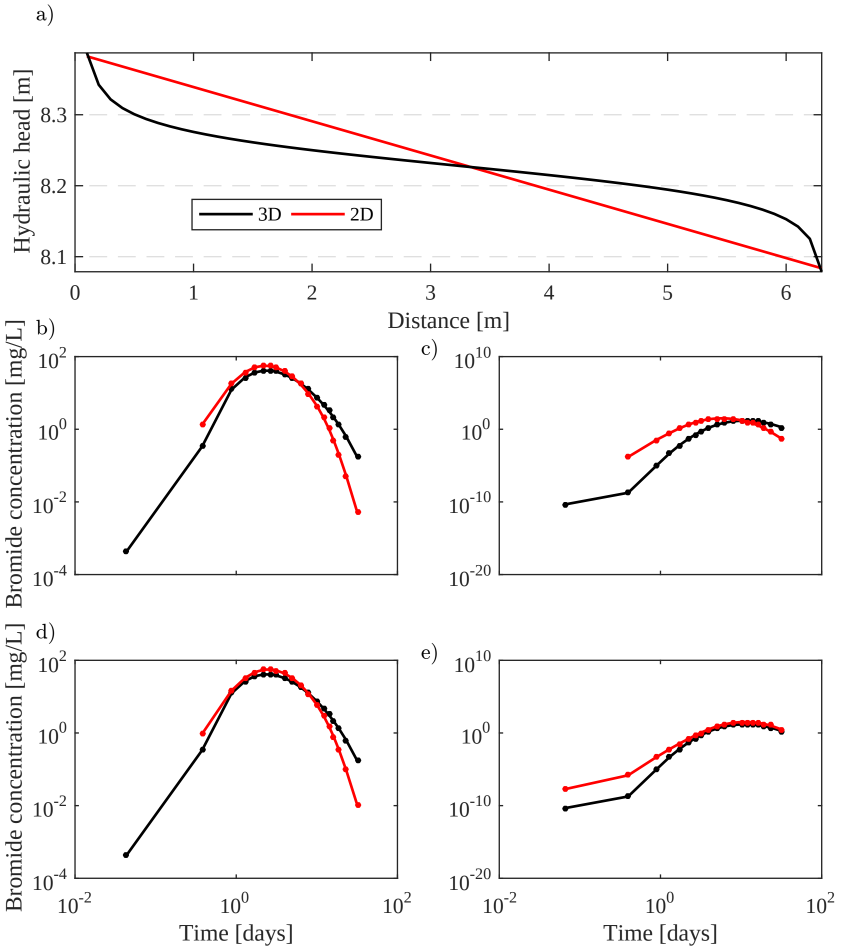


Figure 4.11 – (a) Hydraulic head profiles between the injection and extraction wells arising from 2D and 3D flow simulations in a uniform hydraulic conductivity field. Simulated breakthrough curves at 7 m depth in (b) MLS-1 and (c) MLS-2 without corrections. The shifts in the 2D simulations are removed when (d-e) applying the correction factors.

4.9 Supporting information

This supporting information provides the same figures as Figure 4.5 in the main article but at all measurement locations in the two monitoring wells MLS-1 and MLS-2. They show the simulated and measured bromide breakthrough curves from the MADE-5 experiment.

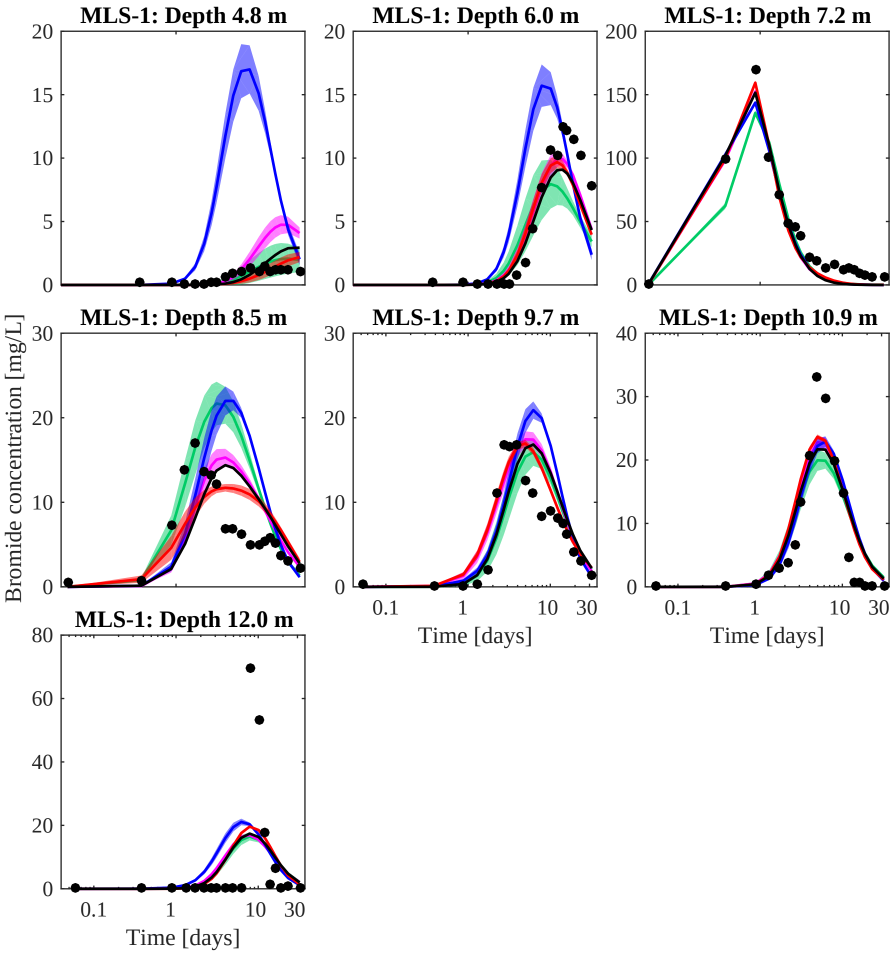


Figure 4.12 – Simulated (solid lines) and measured (black dots) bromide breakthrough curves from the MADE-5 experiment in the monitoring well MLS-1 at different depths. The simulated breakthrough curves are summarised by the mean of the posterior realisations and their thicknesses depict the 95% confidence intervals for the hybrid (red line), multi-Gaussian (black line), analog-based (magenta line), lithofacies-based (blue line) and outcrop-based conceptual model (green line).

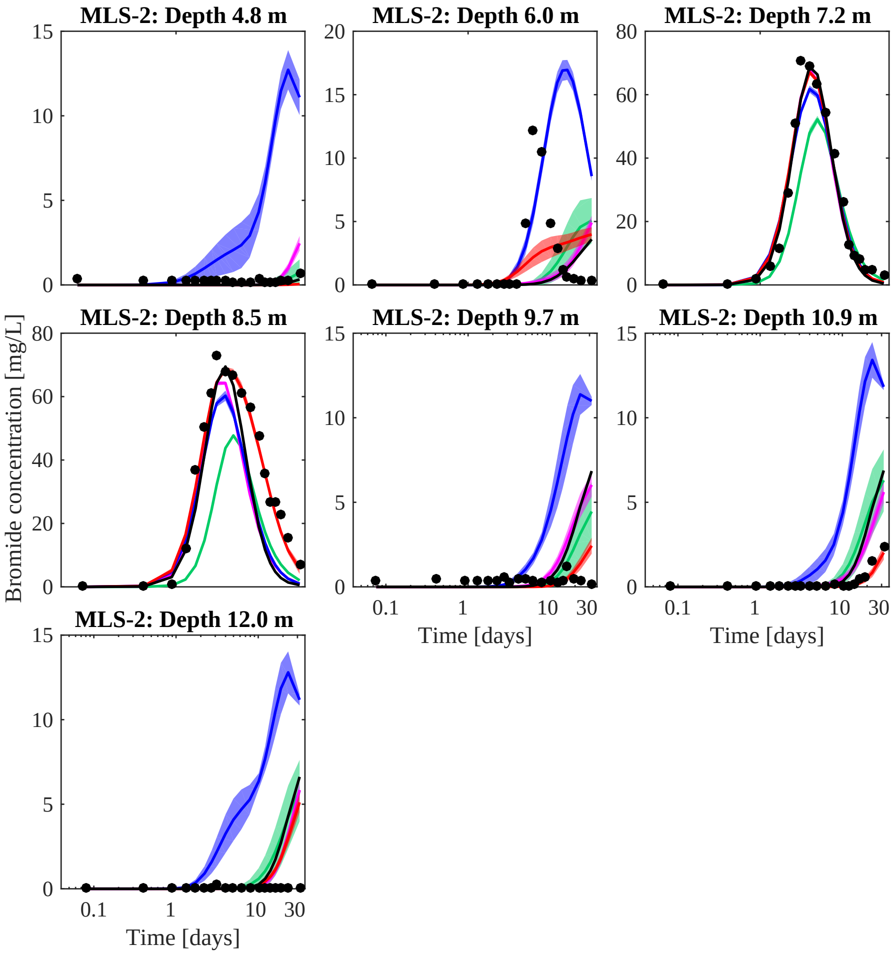


Figure 4.13 – Simulated (solid lines) and measured (black dots) bromide breakthrough curves from the MADE-5 experiment in the monitoring well MLS-2 at different depths. The simulated breakthrough curves are summarised by the mean of the posterior realisations and their thicknesses depict the 95% confidence intervals for the hybrid (red line), multi-Gaussian (black line), analog-based (magenta line), lithofacies-based (blue line) and outcrop-based conceptual model (green line).

Chapter 5

Conclusions

Bayesian model selection based on evidence computation and subsequent computation of Bayes factors provides a valuable tool to account for and minimise conceptual uncertainty in Bayesian inference and, therefore, to inform and increase the reliability of subsurface systems modelling and management. In this thesis, we have explored the potential of Bayesian model selection in hydrogeophysics and hydrogeology.

We conclude that much more reliable Bayesian uncertainty quantification, parameter inference and model selection results are obtained in hydrogeophysical and hydrogeological inversions based on MCMC when (i) combining informative geophysical and hydrogeological data; (ii) accounting for the petrophysical prediction uncertainty and its spatial correlation; (iii) considering geologically-realistic conceptual models represented by training images. These three aspects can enhance the fidelity of the subsurface characterisation that is fundamental for safe and sustainable management of groundwater resources, reliable assessment of water policies and effective support to decision-making.

In Chapter 2, we find that geophysical methods can be valuable in providing guidance about which hydrogeological representation of the subsurface is the most supported by the available data given a set of competing conceptual models. Our first comparative study of evidence estimation in hydrogeophysical settings suggests that the Brute-force Monte Carlo (BFMC) method cannot be used because it is too computationally expensive when confronted with many parameters and data. We find that the Laplace-Metropolis (LM) approximation and the Gaussian mixture importance sampling (GMIS) method provide overall consistent evidence estimates with rather small errors. When these two latter evidence estimators were applied to conceptual subsurface models of the South Oyster Bacterial Transport Site in Virginia (USA), we find that the isotropic multi-Gaussian model is the most supported by the GPR travel time data.

In Chapter 3, we demonstrate the critical role of spatial correlation in petrophysical errors and how sampling this uncertainty with a full MCMC technique allow for dramatic gains in sampling efficiency compared with previously published implementations. Our approach in which we explicitly infer spatially-correlated petrophysical prediction uncertainty leads to less bias, more realistic uncertainty quantification and less overconfident model selection compared to the common choice of altogether ignoring this type of uncertainty. In particular, ignoring petrophysical prediction uncertainty leads to bias in the form of too high variance in the inferred hydraulic conductivity fields.

In Chapter 4, we propose a new methodology for performing Bayesian inversion and model selection with geologically-realistic priors (training images). We find that thermodynamic integration and the stepping-stone sampling provide consistent rankings of the competing conceptual models regardless of the number of MCMC iterations after burn-in. For practical purposes, both thermodynamic integration and stepping-stone sampling are suitable evidence estimators. We suggest to use the stepping-stone sampling method because it is not affected by discretisation errors and its uncertainty (sampling errors) can be significantly decreased with increasing numbers of MCMC iterations. This is not the case for the thermodynamic integration because its discretisation errors dominate over the sampling errors. Thermodynamic integration and stepping-stone sampling applied to conceptual subsurface models of the Macrodispersion Experiment Site in Mississippi (USA) suggest that, among the conceptual models considered, the hybrid (highly conductive channels in a multi-Gaussian background) model is the most supported by the MADE-5 solute concentration data.

The methods explored and developed in this thesis have been applied to aquifer characterisation but they also offer the potential of being used in other fields with different types of data and scales. Notably, many of these methods can handle both continuous and categorical conceptual models (e.g., geologic facies image).

5.1 Limitations and outlook

Even if hydrogeophysical methods can provide valuable guidance about the selection of the conceptual subsurface model, they also reveal some limitations. Hydrogeophysical investigations are typically characterised by large geophysical data sets (several thousands) and small uncorrelated data errors (Gaussian likelihood function) with the consequence that the likelihoods for each data residual are multiplied together. If a proposed conceptual model performs only slightly better (worse) on average, then the total likelihood and the evidence will be remarkably higher (lower) than the one of the other competing models. This effect grows when increasing the size of the data set. In such a case, Bayesian model selection results in a ranking where the best-performing conceptual model is strongly supported by the data and it may suggests that considering only the "best" model is worthwhile for future studies (Chapter 2) and all the other competing conceptual models in the set can be rejected. This very marked preference for one conceptual model should not be interpreted only in term of performance of the selected model but it may rather highlight that significant sources of uncertainty are ignored in the formulation of the problem. Indeed, this overconfidence in the ability of geophysical data to falsify and discriminate between alternative conceptual hydrogeological models can be decreased, for instance, by properly accounting for uncertainty in the petrophysical relationship and the model errors.

In Chapter 3 in which we account for the uncertainty on petrophysical relationships, we identify mainly three issues: (i) the geostatistical model describing petrophysical prediction uncertainty is only partially recoverable by the inversion, especially when the petrophysical prediction uncertainty has similar geostatistics as the hydrogeological property field of interest; (ii) prior constrains on the standard deviations of the hydrogeological and petrophysical fields are needed to avoid inferring geologically-unrealistic hydrogeological fields; (iii) the geophysical data alone may favour a simpler (102 instead of 205 unknowns in our case), but erroneous model that ignores petrophysical errors over the one that accounts for petrophysical errors. Future work should focus on developing a computationally efficient and accurate approach to account for petrophysical uncertainty in the likelihood function. In this way, the geostatistical model describing petrophysical errors requires less parameters, the model dimension is reduced and, thereby, Bayesian model selection would be more efficient.

In Chapters 2 and 3, model errors are ignored in the inversion. They arise from the use of the 2D ray-based approximation instead of a full solution of the Maxwell's equations in the three spatial directions when computing first-arrival GPR travel times from velocity fields. In Chapter 4, model errors are related to the 3D to 2D approximation of the flow and transport of a solute in a porous medium and we partially account for them in the inversion. We emphasise the need to account for and describe model errors, but it is a very challenging task for problems involving many data, high-dimensional parameter spaces and non-linear forward solvers.

Bayesian inversion and model selection based on MCMC are very time-consuming and this is a main reason behind the limited use of Bayes factors in hydrogeophysics and hydrogeology. Indeed, hydrogeophysical and hydrogeological investigations attempt to infer spatially-distributed hydrogeological properties of the subsurface for problems that may involve many thousands of unknowns and high-dimensional parameter spaces, for which the likelihood function is very peaky. In Chapter 2 and 3, we resort to model reduction of multi-Gaussian fields (100 instead of 32400 unknowns) and we use the DREAM$_{(ZS)}$ algorithm that efficiently explores high-dimensional space. In Chapter 4, the inversion is made tractable by the use of sequential geostatistical resampling based on MPS. All these choices allow us to achieve feasible computational times and to successfully perform Bayesian inversion and model selection in moderately challenging hydrogeophysical and hydrogeological settings. However, we needed a computational cluster and to run the Markov chains in parallel. In Chapter 4, the sampling can possibly be improved by applying parallel tempering and using the resulting chains for evidence computation with methods based on power posteriors.

A possible alternative could be to explore model selection with approximate Bayesian computation (ABC) methods. These algorithms replace the definition of the likelihood function (pair-wise comparisons of the observed and simulated data) with summary statistics such as the variance of the data. This could offer advantages: the assumption about uncorrelated Gaussian measurement errors is relaxed, less computational time is needed, the sensitivity to model errors in the inversion is decreased.

Future work need to be done for better elucidate and understand the relationship between the choice of the priors and the evidence estimate for a given conceptual model, especially in the case of comparatively high-dimensional priors. The sensitivity of the Bayes factors to the choice of different prior distributions and ranges need to be investigated and assessed. In the ideal case, the ranking of the conceptual models do not change when using slightly different prior ranges.

A potential extension of this work would be to explore the impact of Bayesian model selection outcomes within an integrated modelling framework for groundwater management. The prediction about a parameter of interest and its uncertainty provided by the "best" conceptual model selected (Bayesian model selection) or by a combination of the competing conceptual models (Bayesian model averaging) could be propagated within, for instance, socio-economic and climatic models and it may significantly alter the process of decision-making.

Bibliography

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

Backus, G. E. and Gilbert, J. (1967). Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal International*, 13(1-3):247–276.

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

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

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

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

Bazargan, H. and Christie, M. (2017). Bayesian model selection for complex geological structures using polynomial chaos proxy. *Computational Geosciences*, 21(3):533–551.

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

Berger, J. O., Pericchi, L. R., Ghosh, J., Samanta, T., De Santis, F., Berger, J., and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison. *Lecture Notes-Monograph Series*, 38:135–207.

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

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

Bianchi, M., Zheng, C., Wilson, C., Tick, G. R., Liu, G., and Gorelick, S. M. (2011b). Spatial connectivity in a highly heterogeneous aquifer: From cores to preferential flow paths. *Water Resources Research*, 47(5):503–514.

Binley, A., Cassiani, G., and Deiana, R. (2010). Hydrogeophysics: opportunities and challenges. *Bollettino di Geofisica Teorica ed Applicata*, 51(4):267–284.

Binley, A., Hubbard, S. S., Huisman, J. A., Revil, A., Robinson, D. A., Singha, K., and Slater, L. D. (2015). The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. *Water Resources Research*, 51(6):3837–3866.

Bodin, T. and Sambridge, M. (2009). Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, 178(3):1411–1436.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K., and Rawlinson, N. (2012). Transdimensional inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research-Solid Earth*, 117(B2):1–24.

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

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

Bosch, M. (1999). Lithologic tomography: From plural geophysical data to lithology estimation. *Journal of Geophysical Research-Solid Earth*, 104(B1):749–766.

Bosch, M. (2004). The optimization approach to lithological tomography: Combining seismic data and petrophysics for porosity prediction. *Geophysics*, 69(5):1272–1282.

Bosch, M. (2016). Inference Networks in Earth Models with Multiple Components and Data. In *Integrated Imaging of the Earth: Theory and Applications*, chapter 3, pages 29–47. John Wiley & Sons, Inc.

Bosch, M., Carvajal, C., Rodrigues, J., Torres, A., Aldana, M., and Sierra, J. (2009). Petrophysical seismic inversion conditioned to well-log data: Methods and application to a gas reservoir. *Geophysics*, 74(2):O1–O15.

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier.

Box, G. E. and Draper, N. R. (1987). *Empirical model-building and response surfaces*, volume 424. Wiley New York.

Boykov, Y. and Kolmogorov, V. (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.

Brunetti, C., Linde, N., and Vrugt, J. A. (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.

Brynjarsdóttir, J. and O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30(11):114007.

Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Caers, J. and Zhang, T. (2004). Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models. In Grammer, G. M., Harris, P. M., and Eberli, G. P., editors, *Integration of Outcrop and Modern Analogs in Reservoir Modeling*, chapter 18, pages 383–394. American Association of Petroleum Geologists.

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

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

Chen, J. and Dickens, T. A. (2009). Effects of uncertainty in rock-physics models on reservoir parameter estimation using seismic amplitude variation with angle and controlled-source electromagnetics data. *Geophysical Prospecting*, 57(1):61–74.

Chen, J., Hubbard, S., Peterson, J., Williams, K., Fienen, M., Jardine, P., and Watson, D. (2006). Development of a joint hydrogeophysical inversion approach and application to a contaminated fractured aquifer. *Water Resources Research*, 42(6):W06425.

Chen, J., Hubbard, S., and Rubin, Y. (2001). Estimating the hydraulic conductivity at the South Oyster Site from geophysical tomographic data using Bayesian techniques based on the normal linear regression model. *Water Resources Research*, 37(6):1603–1613.

Chen, J., Hubbard, S., Rubin, Y., Murray, C., Roden, E., and Majer, E. (2004). Geochemical characterization using geophysical data and Markov Chain Monte Carlo methods: A case study at the South Oyster bacterial transport site in Virginia. *Water Resources Research*, 40(12):W12412.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical science*, 19(1):81–94.

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

Copty, N., Rubin, Y., and Mavko, G. (1993). Geophysical-hydrological identification of field permeabilities through Bayesian updating. *Water Resources Research*, 29(8):2813–2825.

Cui, T., Fox, C., and O'sullivan, M. (2011). Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm. *Water Resources Research*, 47(10):W10521.

Day-Lewis, F. D., Singha, K., and Binley, A. M. (2005). Applying petrophysical models to radar travel time and electrical resistivity tomograms: Resolution-dependent limitations. *Journal of Geophysical Research-Solid Earth*, 110(B8):1–17.

De Bruijn, N. G. (1970). *Asymptotic methods in analysis*, volume 4. Dover Publications.

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

de Pasquale, G. and Linde, N. (2017). On structure-based priors in Bayesian geophysical inversion. *Geophysical Journal International*, 208(3):1342–1358.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38.

Dettmer, J., Dosso, S. E., and Holland, C. W. (2009). Model selection and Bayesian inference for high-resolution seabed reflection inversion. *The Journal of the Acoustical Society of America*, 125(2):706–716.

Dettmer, J., Dosso, S. E., and Osler, J. C. (2010). Bayesian evidence computation for model selection in non-linear geoacoustic inference problems. *The Journal of the Acoustical Society of America*, 128(6):3406–3415.

Doetsch, J., Linde, N., Coscia, I., Greenhalgh, S. A., and Green, A. G. (2010). Zonation for 3D aquifer characterization based on joint inversions of multimethod crosshole geophysical data. *Geophysics*, 75(6):G53–G64.

Doyen, P. M. (1988). Porosity from seismic data: A geostatistical approach. *Geophysics*, 53(10):1263–1275.

Doyen, P. M. (2007). Seismic reservoir characterization: An Earth Modelling Perspective. *EAGE publications*, 2:255.

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

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

Emery, X. and Lantuéjoul, C. (2014). Can a training image be a substitute for a random field model? *Mathematical Geosciences*, 46(2):133–147.

European Commission (2012). A blueprint to safeguard Europe's water resources. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Brussels, 14.11.2012.*

Feehley, C. E., Zheng, C., and Molz, F. J. (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.

Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society. Series B*, 70(3):589–607.

Gallardo, L. A. and Meju, M. A. (2003). Characterization of heterogeneous near-surface materials by joint 2D inversion of dc resistivity and seismic data. *Geophysical Research Letters*, 30(13):1–4.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B*, 56(3):501–514.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.

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

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

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

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

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A*, 371(1984):20110553.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

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

Grana, D. and Della Rossa, E. (2010). Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion. *Geophysics*, 75(3):O21–O37.

Grana, D., Pirrone, M., and Mukerji, T. (2012). Quantitative log interpretation and uncertainty propagation of petrophysical properties and facies classification from rock-physics modeling and formation evaluation analysis. *Geophysics*, 77(3):WA45–WA63.

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

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

Guardiano, F. B. and Srivastava, R. M. (1993). Multivariate geostatistics: beyond bivariate moments. In Soares, A., editor, *Geostatistics Tróia ’92*, pages 133–144. Springer.

Gull, S. F. (1988). Bayesian inductive inference and maximum entropy. In *Maximum-entropy and Bayesian methods in Science and Engineering*, volume 31-32, pages 53–74. Springer.

Guthke, A. (2017). Defensible model complexity: A call for data-based and goal-oriented model choice. *Groundwater*, 55(5):646–650.

Haber, E. and Oldenburg, D. (1997). Joint inversion: A structural approach. *Inverse Problems*, 13(1):63.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo methods*, volume 1. Springer Netherlands.

Handley, W., Hobson, M., and Lasenby, A. (2015a). POLYCHORD: nested sampling for cosmology. *Monthly Notices of the Royal Astronomical Society: Letters*, 450(1):L61–L65.

Handley, W., Hobson, M., and Lasenby, A. (2015b). POLYCHORD: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, 453(4):4384–4398.

Hansen, T., Cordua, K., Jacobsen, B., and Mosegaard, K. (2014). Accounting for imperfect forward modeling in geophysical inverse problems - Exemplified for crosshole tomography. *Geophysics*, 79(3):H1–H21.

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

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

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

Hermans, T., Oware, E., and Caers, J. (2016). Direct prediction of spatially and temporally varying physical properties from time-lapse electrical resistance data. *Water Resources Research*, 52(9):7262–7283.

Hinnell, A., Ferré, T., Vrugt, J., Huisman, J., Moysey, S., Rings, J., and Kowalsky, M. (2010). Improved extraction of hydrologic information from geophysical data through coupled hydrogeophysical inversion. *Water Resources Research*, 46(4):W00D40.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401.

Höge, M., Wöhling, T., and Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, 54(3):1688–1715.

Höhna, S., Landis, M. L., and Huelsenbeck, J. P. (2017). Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics. *bioRxiv*, pages 1–7.

Hoogerheide, L., Opschoor, A., and Van Dijk, H. K. (2012). A class of adaptive importance sampling weighted EM algorithms for efficient and robust posterior and predictive simulation. *Journal of Econometrics*, 171(2):101–120.

Hu, L. and Chugunova, T. (2008). Multiple-point geostatistics for modeling subsurface heterogeneity: A comprehensive review. *Water Resources Research*, 44(11):W11413.

Hubbard, S. and Linde, N. (2011). Hydrogeophysics. In Wilderer, P., editor, *Treatise on Water Science*, pages 401–434. Elsevier.

Hubbard, S. and Rubin, Y. (2002). Study institute assesses the state of hydrogeophysics. *Eos, Transactions American Geophysical Union*, 83(51):602–606.

Hubbard, S. S., Chen, J., Peterson, J., Majer, E. L., Williams, K. H., Swift, D. J., Mailloux, B., and Rubin, Y. (2001). Hydrogeological characterization of the South Oyster Bacterial Transport Site using geophysical data. *Water Resources Research*, 37(10):2431–2456.

Hubbard, S. S. and Rubin, Y. (2005). Introduction to Hydrogeophysics. In Hubbard, S. S. and Rubin, Y., editors, *Hydrogeophysics*, pages 3–21. Springer.

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

James, F. (1980). Monte Carlo theory and practice. *Reports on Progress in Physics*, 43(9):1145–1189.

Jefferys, W. H. and Berger, J. O. (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.

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

Jiménez Cisneros, B., Oki, T., Arnell, N., Benito, G., Cogley, J., Döll, P., Jiang, T., and Mwakalila, S. (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, book section Freshwater resources, pages 229–269. Cambridge University Press.

Jordan, M. (2011). What are the open problems in Bayesian statistics. *The ISBA Bulletin*, 18(1):568.

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

Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE T Pattern Anal*, PAMI-4(2):99–104.

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

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

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

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

Konikow, L. F. and Bredehoeft, J. D. (1992). Ground-water models cannot be validated. *Advances in Water Resources*, 15(1):75–83.

Kowalsky, M. B., Finsterle, S., Peterson, J., Hubbard, S., Rubin, Y., Majer, E., Ward, A., and Gee, G. (2005). Estimation of field-scale soil hydraulic and dielectric parameters through joint inversion of GPR and hydrological data. *Water Resources Research*, 41(11):W11425.

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

Laloy, E., Hérault, R., Jacques, D., and Linde, N. (2018). Training-image based geostatistical inversion using a Spatial Generative Adversarial Neural Network. *Water Resources Research*, 54(1):381–406.

Laloy, E., Linde, N., Jacques, D., and Mariethoz, G. (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.

Laloy, E., Linde, N., Jacques, D., and Vrugt, J. A. (2015). Probabilistic inference of multi-Gaussian fields from indirect hydrological data using circulant embedding and dimensionality reduction. *Water Resources Research*, 51(6):4224–4243.

Laloy, E. and Vrugt, J. A. (2012). High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{\mathrm{ZS}}$ and high-performance computing. *Water Resources Research*, 48(1):W01526.

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

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

Lee, J., Sung, W., and Choi, J.-H. (2015). Metamodel for efficient estimation of capacity-fade uncertainty in Li-Ion batteries for electric vehicles. *Energies*, 8(6):5538–5554.

Lesmes, D. P. and Friedman, S. P. (2005). Relationships between the electrical and hydrogeological properties of rocks and soils. In Hubbard, S. S. and Rubin, Y., editors, *Hydrogeophysics*, pages 87–128. Springer.

Lewis, S. M. and Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655.

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

Li, X. and Tsai, F. T.-C. (2009). Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod. *Water Resources Research*, 45(9):W09403.

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

Linde, N., Binley, A., Tryggvason, A., Pedersen, L. B., and Revil, A. (2006a). Improved hydrogeophysical characterization using joint inversion of cross-hole electrical resistance and ground-penetrating radar traveltime data. *Water Resources Research*, 42(12):W12404.

Linde, N., Chen, J., Kowalsky, M. B., and Hubbard, S. (2006b). Hydrogeophysical parameter estimation approaches for field scale characterization. In Vereecken, H., Binley, A., Cassiani, G., Revil, A., and Titov, K., editors, *Applied Hydrogeophysics*, pages 9–44. Springer.

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

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

Linde, N., Renard, P., Mukerji, T., and Caers, J. (2015b). Geological realism in hydrogeological and geophysical inverse modeling: A review. *Advances in Water Resources*, 86:86–101.

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

Linde, N. and Vrugt, J. A. (2013). Distributed soil moisture from crosshole ground-penetrating radar travel times using stochastic inversion. *Vadose Zone Journal*, 12(1):1–21.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1-2):187–192.

Liu, G., Zheng, C., Tick, G. R., Butler, J. J., and Gorelick, S. M. (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):W03524,.

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

Lochbühler, T., Breen, S. J., Detwiler, R. L., Vrugt, J. A., and Linde, N. (2014a). Probabilistic electrical resistivity tomography of a CO_2 sequestration analog. *Journal of Applied Geophysics*, 107:80–92.

Lochbühler, T., Doetsch, J., Brauchler, R., and Linde, N. (2013). Structure-coupled joint inversion of geophysical and hydrological data. *Geophysics*, 78(3):ID1–ID14.

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

Lochbühler, T., Vrugt, J. A., Sadegh, M., and Linde, N. (2015). Summary statistics from training images as prior information in probabilistic inversion. *Geophysical Journal International*, 201(1):157–171.

Lu, D., Ye, M., and Neuman, S. P. (2011). Dependence of Bayesian model selection criteria and Fisher information matrix on sample size. *Mathematical Geosciences*, 43(8):971–993.

MacKay, D. J. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

Maliva, R. G. (2016). *Aquifer Characterization Techniques.* Springer.

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

Mariethoz, G., Renard, P., and Caers, J. (2010a). Bayesian inverse problem and optimization with iterative spatial resampling. *Water Resources Research*, 46(11):W11530.

Mariethoz, G., Renard, P., and Straubhaar, J. (2010b). The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, 46(11):W11536.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Marshall, L., Nott, D., and Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, 41(10):W10422.

Mavko, G., Mukerji, T., and Dvorkin, J. (1998). The rock physics handbook.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Moorkamp, M., Lelièvre, P. G., Linde, N., and Khan, A. (2016). *Integrated Imaging of the Earth: Theory and Applications*, volume 218. John Wiley & Sons.

Morey, R. D., Romeijn, J.-W., and Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72:6–18.

Mosegaard, K. and Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, 100(B7):12431–12447.

Mukerji, T., Jørstad, A., Avseth, P., Mavko, G., and Granli, J. (2001). Mapping lithofacies and pore-fluid probabilities in a North Sea reservoir: Seismic inversions and statistical rock physics. *Geophysics*, 66(4):988–1001.

National Research Council (2012). *Challenges and opportunities in the hydrologic sciences*. National Academies Press.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 31(3):705–741.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B*, 56(1):3–48.

Nilsson, B., Højberg, A., Refsgaard, J., and Troldborg, L. (2006). Uncertainty in geological and hydrogeological data. *Hydrology and Earth System Sciences Discussions*, 3(4):2675–2706.

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

Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641–646.

Oware, E., Moysey, S., and Khan, T. (2013). Physically based regularization of hydrogeophysical inverse problems for improved imaging of process-driven systems. *Water Resources Research*, 49(10):6238–6247.

Perrakis, K., Ntzoufras, I., and Tsionas, E. G. (2014). On the use of marginal posteriors in marginal likelihood estimation via importance sampling. *Computational Statistics & Data Analysis*, 77:54–69.

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

Pirot, G., Linde, N., Mariethoz, G., and Bradford, J. H. (2017b). Probabilistic inversion with graph cuts: Application to the Boise Hydrogeophysical Research Site. *Water Resources Research*, 53(2):1231–1250.

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

Podvin, P. and Lecomte, I. (1991). Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *Geophysical Journal International*, 105(1):271–284.

Pooley, C. and Marion, G. (2018). Bayesian model evidence as a practical alternative to deviance information criterion. *Royal Society Open Science*, 5(3):171519.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Pride, S. (1994). Governing equations for the coupled electromagnetics and acoustics of porous media. *Physical Review B*, 50(21):15678–15696.

Pride, S. R. (2005). Relationships between seismic and hydrological properties. In Hubbard, S. S. and Rubin, Y., editors, *Hydrogeophysics*, pages 253–290. Springer.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

Randle, C. H., Bond, C. E., Lark, R. M., and Monaghan, A. A. (2018). Can uncertainty in geological cross-section interpretations be quantified and predicted? *Geosphere*, 14(3):1087–1100.

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

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

Refsgaard, J. C., Van der Sluijs, J. P., Brown, J., and Van der Keur, P. (2006). A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources*, 29(11):1586–1597.

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

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

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

Renard, P., Demougeot-Renard, H., and Froidevaux, R. (2005). *Geostatistics for Environmental Applications*. Springer.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Rojas, R., Feyen, L., and Dassargues, A. (2008). Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research*, 44(12):W12418.

Ronayne, M. J., Gorelick, S. M., and Zheng, C. (2010). Geological modeling of submeter scale heterogeneity and its influence on tracer transport in a fluvial aquifer. *Water Resources Research*, 46(10):W10519.

Rosas-Carbajal, M., Linde, N., Kalscheuer, T., and Vrugt, J. A. (2013). Two-dimensional probabilistic inversion of plane-wave electromagnetic data: Methodology, model constraints and joint inversion with electrical resistivity data. *Geophysical Journal International*, 196(3):1508–1524.

Rosas-Carbajal, M., Linde, N., Peacock, J., Zyserman, F., Kalscheuer, T., and Thiel, S. (2015). Probabilistic 3-D time-lapse inversion of magnetotelluric data: application to an enhanced geothermal system. *Geophysical Journal International*, 203(3):1946–1960.

Rosenkrantz, R. D. (1977). *Inference, method and decision: Towards a Bayesian philosophy of science*, volume 115. Springer.

Roth, K., Schulin, R., Fluhler, H., and Attinger, W. (1990). Using a composite dielectric approach. *Water Resources Research*, 26(10):2267–2273.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

Rubin, Y. and Hubbard, S. S. (2005). *Hydrogeophysics*. Springer.

Rubin, Y., Mavko, G., and Harris, J. (1992). Mapping permeability in heterogeneous aquifers using hydrologic and seismic data. *Water Resources Research*, 28(7):1809–1816.

Ruggeri, P., Irving, J., and Holliger, K. (2015). Systematic evaluation of sequential geostatistical resampling within MCMC for posterior sampling of near-surface geophysical inverse problems. *Geophysical Journal International*, 202(2):961–975.

Sambridge, M., Gallagher, K., Jackson, A., and Rickwood, P. (2006). Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International*, 167(2):528–542.

Scheibe, T. D., Hubbard, S. S., Onstott, T. C., and DeFlaun, M. F. (2011). Lessons learned from Bacterial Transport Research at the South Oyster Site. *Groundwater*, 49(5):745–763.

Scheidt, C., Li, L., and Caers, J. (2018). *Quantifying Uncertainty in Subsurface Systems*, volume 236. John Wiley & Sons.

Schöniger, A., Illman, W. A., Wöhling, T., and Nowak, W. (2015a). Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. *Journal of Hydrology*, 531:96–110.

Schöniger, A., Wöhling, T., and Nowak, W. (2015b). A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. *Water Resources Research*, 51(9):7524–7546.

Schöniger, A., Wöhling, T., Samaniego, L., and Nowak, W. (2014). Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, 50(12):9484–9513.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shahraeeni, M. S. and Curtis, A. (2011). Fast probabilistic nonlinear petrophysical inversion. *Geophysics*, 76(2):E45–E58.

Shahraeeni, M. S., Curtis, A., and Chao, G. (2012). Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data. *Geophysics*, 77(3):O1–O19.

Skilling, J. (2004). Nested sampling. In *AIP Conference Proceedings*, volume 735, pages 395–405. AIP.

Skilling, J. (2012). Bayesian computation in big spaces: Nested sampling and Galilean Monte Carlo. *AIP Conference Proceedings*, 1443(1):145–156.

Skilling, J. et al. (2006). Nested sampling for general Bayesian computation. *Bayesian analysis*, 1(4):833–859.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B*, 64(4):583–639.

Steininger, G., Dosso, S. E., Holland, C. W., and Dettmer, J. (2014). Estimating seabed scattering mechanisms via Bayesian model selection. *The Journal of the Acoustical Society of America*, 136(4):1552–1562.

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

Sun, J. and Li, Y. (2016). Joint inversion of multiple geophysical data using guided fuzzy c-means clustering. *Geophysics*, 81(3):ID37–ID57.

Sun, J. and Li, Y. (2017). Joint inversion of multiple geophysical and petrophysical data using generalized fuzzy clustering algorithms. *Geophysical Journal International*, 208(2):1201–1216.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*, volume 89. siam.

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

Thorburn, W. M. (1918). The myth of Occam's razor. *Mind*, 27(107):345–353.

Tsai, F. T.-C. and Li, X. (2008). Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window. *Water Resources Research*, 44(9):W09434.

van der Linde, A. (2012). A Bayesian view of model complexity. *Statistica Neerlandica*, 66(3):253–271.

Van Haasteren, R. (2013). *Gravitational Wave Detection and Data Analysis for Pulsar Timing Arrays*, chapter Marginal likelihood calculation with MCMC methods, pages 99–120. Springer Science & Business Media.

Vanrolleghem, P. A. (2010). *Modelling Aspects of Water Framework Directive Implementation.* IWA Publishing.

Vasco, D., Daley, T. M., and Bakulin, A. (2014). Utilizing the onset of time-lapse changes: A robust basis for reservoir monitoring and characterization. *Geophysical Journal International*, 197:542–556.

Vereecken, H., Binley, A., Cassiani, G., Revil, A., and Titov, K. (2006). *Applied hydrogeophysics*, volume 71. Springer.

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

Volpi, E., Schoups, G., Firmani, G., and Vrugt, J. (2017). Sworn Testimony of the Model Evidence: Gaussian Mixture Importance (GAME) sampling. *Water Resources Research*, 53:6133–6158.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, 75:273–316.

Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., and Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, 44(12):W00B09.

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

Xu, T. and Valocchi, A. (2015). A Bayesian approach to improved calibration and prediction of groundwater models with structural error. *Water Resources Research*, 51(11):9290–9311.

Ye, M., Pohlmann, K. F., Chapman, J. B., Pohll, G. M., and Reeves, D. M. (2010). A model-averaging method for assessing groundwater conceptual model uncertainty. *Groundwater*, 48(5):716–728.

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

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

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

Zheng, C., Bianchi, M., and Gorelick, S. M. (2011). Lessons learned from 25 years of research at the MADE site. *Groundwater*, 49(5):649–662.

Zheng, C. and Gorelick, S. M. (2003). Analysis of solute transport in flow fields influenced by preferential flowpaths at the decimeter scale. *Groundwater*, 41(2):142–155.

Zinn, B. and Harvey, C. F. (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):1051,.

Appendix A

Evidence estimation with POLYCHORD for hydrogeophysical applications

Carlotta Brunetti and Niklas Linde.

Internal report

A.1 Abstract

Bayesian model selection requires computation of the marginal likelihood of the measured data, or evidence, for each conceptual model being considered. This task is not particularly easy for subsurface models, as the evidence is, in general, a high-dimensional integral of the posterior parameter distribution without analytic solution. The purpose of this study is to investigate the potential of the POLYCHORD algorithm to provide reliable evidence estimates in high dimensions in the context of hydrogeophysical case studies. We compare the evidence estimated by POLYCHORD with the ones computed with the Gaussian mixture importance sampling estimator (GMIS) and the Laplace-Metropolis (LM) method for both a synthetic case and a real-world case that uses crosshole ground-penetrating radar data of the South Oyster Bacterial Transport Site in Virginia, USA. The main finding is that the POLYCHORD algorithm is faster in evaluating one single forward simulation than GMIS and LM methods but it requires a number of forward simulations that is at least one order of magnitude larger than the ones used by GMIS and LM methods for evaluating the evidence to a similar level of accuracy. For this reason, we can not fully benefit from the potential of POLYCHORD.

A.2 Introduction

Computing the evidence by traditional means implies integration over a high-dimensional parameter space. Large portions of this space are made up of areas with a negligible posterior density whose contributions to the integral are negligibly small. Algorithms need to be able to quickly focus exploration in the parameter space from the prior onto the posterior. Nested sampling (Skilling et al., 2006) is a recent methodology for computing evidences and posterior distributions simultaneously. Nested sampling has been popular to compute the evidence in high-dimensional parameter spaces in cosmology and astroparticle physics. Adaptation of this methodology has been implemented in algorithms such as POLYCHORD (Handley et al., 2015b,a). The POLYCHORD algorithm utilises slice sampling (Neal, 2003) at each iteration to draw a new point from the prior subject to the hard likelihood constraint of nested sampling. We compare the evidence estimated by POLYCHORD (PC) with the ones computed with the Gaussian mixture importance sampling (GMIS) estimator (Volpi et al., 2017) and the Laplace-Metropolis (LM) method (De Bruijn, 1970) in the context of a synthetic and a real-world case study that uses crosshole ground-penetrating radar data from the South Oyster Bacterial Transport Site in Virginia, USA. We explore parameter spaces of up to 100 dimensions. In Section A.3, we summarize the main features of POLYCHORD. A detailed description of the algorithm can be found in Handley et al. (2015a).

A.3 Theory

A.3.1 Nested sampling

Nested sampling maintains a population of n_{live} live points within a region of the parameter space. These points are sequentially updated so that the region that they occupy contracts around the peak(s) of the posterior. In the following, the nested sampling algorithm is explained in steps.

Step 1: n_{live} points are drawn uniformly from the prior distribution, $p(\theta)$.

At each iteration i, the following steps are performed:

Step 2: The likelihoods of each live point are evaluated and the lowest value, L_i, is recorded.

Step 3: The fraction of prior volume, X_i, covering all likelihoods greater than L_i is computed. Initially, the prior volume is 1 and then it decreases exponentially, tending towards 0 as $X_i = exp(-i/n_{live})$.

Step 4: The weights, w_i, are computed as $w_i = X_{i-1} - X_i$.

Step 5: The discarded points (i.e., the points with minimum likelihood) that are named dead points are used to update the evidence as a weighted sum, $p(\tilde{\mathbf{Y}}) \approx \sum_{i \in dead} w_i L_i$. The remaining posterior mass left in the live points is estimated as $p(\tilde{\mathbf{Y}})_{live} \approx \overline{L}_{live} X_i$, where $\overline{L}_{live}$ is the mean of the likelihoods of the live points. The algorithm terminate when $p(\tilde{\mathbf{Y}})_{live}$ is some small fraction of $p(\tilde{\mathbf{Y}})$. Providing that this small fraction is less than 1, this should not have any appreciable effect on results.

Step 6: The point with the lowest likelihood, L_i, is deleted and then replaced by a new point drawn from the prior, subject to the constraint that its likelihood is greater than L_i. The new live point is generated by slice sampling (Section A.3.2).

The nested sampling algorithm (Skilling et al., 2006) can be summarized as:

Algorithm 2: Nested sampling

```
Start with $n_{live}$ points $\theta_1, ..., \theta_{n_{live}}$ from prior;
        initialise $p(\tilde{\mathbf{Y}}) = 0$, $X_0 = 1$.
Repeat for $i = 1, 2, ..., j$;
        record the lowest of the current likelihood values as $L_i$,
        set $X_i = exp(-i/n_{live})$,
        set $w_i = X_{i-1} - X_i$,
        increment $p(\tilde{\mathbf{Y}})$ by $L_i w_i$,
        then replace point of lowest likelihood by a new one drawn from within
        $L(\theta) > L_i$, in proportion to the prior, $p(\theta)$.
Increment $p(\tilde{\mathbf{Y}})$ by $n_{live}^{-1}(L(\theta_1) + ... + L(\theta_{n_{live}}))X_j$.
```

At each iteration i of the nested sampling algorithm, the new sample is drawn from the prior by inverse transform sampling since, in general, the priors are simple analytic functions, $f(\theta)$, (e.g., uniform and Gaussian distributions). This method is based on the concept that, if a random variable x has a uniform distribution in $[0,1]$ and if θ has a cumulative distribution function $F(\theta)$, then, the random variable $F^{-1}(x)$ has the same distribution as θ. In the general d-dimensional case, d uniform variables $\{x_k : k = 1,...,d\}$ drawn from the unit hypercube are transformed into $\{\theta_k : k = 1,...,d\}$ in the physical space distributed according to $f(\boldsymbol{\theta})$. As a consequence, the nested sampling is performed in the unit d-dimensional hypercube, $\mathbf{x} \in [0,1]^d$, with the likelihood redefined as $L(\theta) = L(\mathbf{F}^{-1}(\mathbf{x}))$.

A.3.2 Multi-dimensional slice sampling

POLYCHORD implements new features on the original Markov-Chain based procedure with slice sampling by Neal (2003). In particular, POLYCHORD performs slice sampling in the unit d-dimensional hypercube (Figure A.1) using information present in the live and phantom points (i.e., the points that constitutes the Markov chain before an independent point from the initial one is accepted).

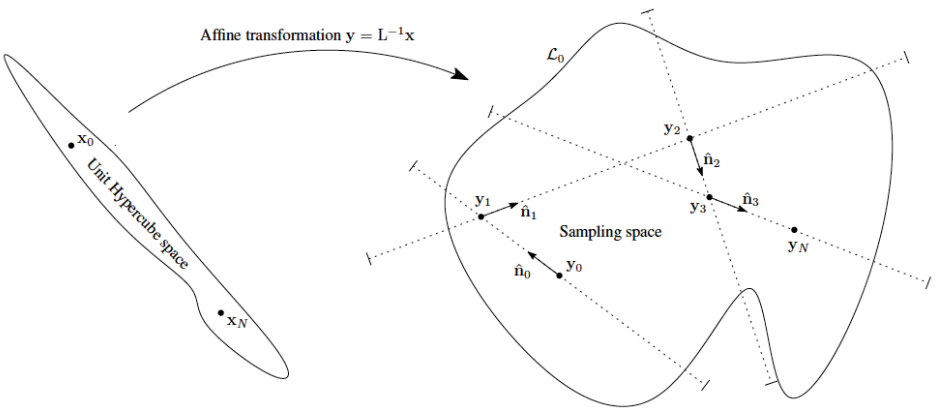


Figure A.1 – Slice sampling in d dimensions. The unit hypercube is whitened by linearly transforming a degenerate contour into one with dimensions $\sim \mathcal{O}(1)$ in all directions. (Figure from Handley et al. (2015a))

At each iteration i, the following slice sampling steps are performed:

Step 1: One of the live points is randomly choosen as start point for a new chain with hypercube coordinates $\mathbf{x}_0$.

Step 2: One-dimensional slice sampling is performed in a random direction $\hat{\mathbf{n}}_0$ chosen from a probability distribution $\mathrm{P}(\hat{\mathbf{n}})$ according to Figure A.2. This step generates the new point $\mathbf{x}_1$ which is uniformly sampled in the hypercube but is correlated with $\mathbf{x}_0$.

This procedure is appropriate if some optimal estimate of $\mathrm{P}(\hat{\mathbf{n}})$ and w is known. This information is provided in POLYCHORD by the sample covariance of the live and phantom points that are already uniformly sampled within the contour. The covariance matrix is used to

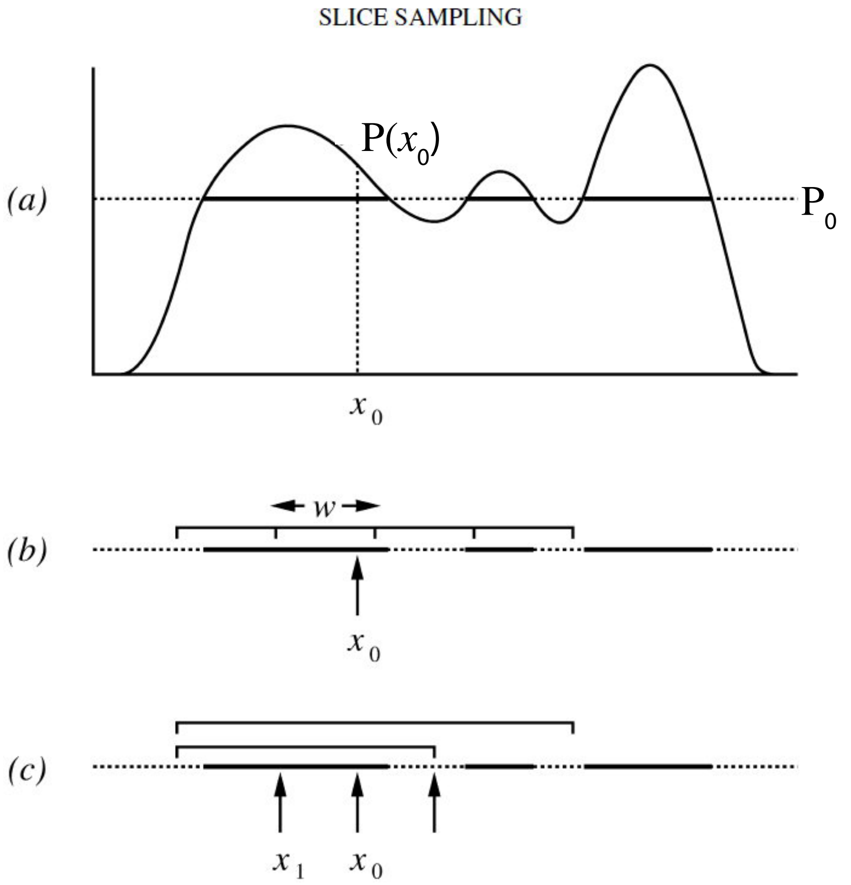


Figure A.2 – Illustration of slice sampling. a) Sample a P_0 slice (probability level) uniformly between 0 and P(x_0) and draw a horizontal line across the curve at this P_0 position. b) An interval of width w is randomly positioned around x_0, and then expanded in steps until both ends are outside the slice. Then, sample uniformly a point x_1 in w until a point within the slice is found. Points picked outside the slice are used to shrink the interval by replacing one of the interval ends with this point. Repeat the process using the new x_1 value. (Figure modified from Neal (2003))

construct an affine transformation which whitens the contour. Sampling is then performed in this whitened space called sampling space. In this space, the contour size is $\sim \mathcal{O}(1)$ in every direction so that w can be set equal to 1.

Step 3: Repeat Step 2 $n_{repeats}$ times; the length of the chain $n_{repeats}$ should be large enough so that the final point of the chain is decorrelated from the start point. This final point can be considered a new uniformly sampled point from the prior distribution subject to the hard likelihood constraint.

A.3.3 Tuning parameters

The POLYCHORD algorithm has mainly two tuning parameters: n_{live} and $n_{repeats}$.

n_{live} is the number of live points that are maintained during the algorithm and it defines the resolution of the results. Increasing n_{live} increases the accuracy of the inference of the evidence since the evidence error scales $\sim \mathcal{O}(n_{live}^{-1/2})$.

$n_{repeats}$ is the length of the slice sampling chain used to generate a new live point and it defines the reliability of the results. Increasing $n_{repeats}$ decreases the correlation between live points and increases the reliability of evidence estimations. For reliable posteriors $n_{repeats} \sim \mathcal{O}(d)$ is suggested and for reliable evidences $n_{repeats} \sim \mathcal{O}(5d)$. However, setting $n_{repeats} \sim \mathcal{O}(3d)$ is typically sufficient (Handley et al., 2015b).

A.4 Illustrative toy example

To compare the evidence estimated by POLYCHORD with the ones computed with the Laplace-Metropolis method and the Gaussian mixture importance sampling estimator, we first consider an illustrative example involving a simple crosshole GPR experiment. A total of 10 transmitter and receiver antennas are placed at multiple different depths (uniform intervals) in boreholes located in the left and right side of the domain, respectively (see Figure A.3). This results in a total of 100 different transmitter-receiver antenna pairs. The spatial domain that necessitates porosity characterization covers an area of 7.2 m × 7.2 m. To warrant accurate model simulations, a spatial discretization of 0.04 × 0.04 m is considered. We contaminate the $n = 100$ first-arrival travel time data with Gaussian white noise using a measurement error of the traveltime observations, $\sigma_{\tilde{\mathbf{Y}}} = 2$ ns. The "true" porosity field of the subsurface is made up of four different layers of equal thickness with porosity values of 0.3, 0.45, 0.35 and 0.4, in the downward direction, respectively (see Figure A.3). We varied the number of horizontal layers of constant thickness from $d = 1$ to $d = 16$, and assume a uniform prior distribution for the porosity, ϕ, of each respective layer using upper and lower bound values of 0.25 and 0.50, respectively.

Now we calculate the marginal likelihood of each subsurface conceptual model using the GMIS, LM and PC estimators. The evidences computed by PC are obtained using $n_{live} = 25d$ and $n_{repeats} = 5d$ as suggested by Handley et al. (2015b,a). The results of this analysis are presented in Figure A.4 using at the left hand-side a plot of the evidence computed by each method against model dimension, and at the right-hand-side a graph of the associated uncertainty of each estimator. We consider subsurface models with up to $d = 16$ horizontal porosity layers of equal thickness. To simplify graphical interpretation of the results, we plot $\log_{10}$ transformed values of the evidence, and refer to this entity as $\mathscr{P}(\tilde{\mathbf{Y}})$. Colour coding is used to differentiate between the results of the three different methods. The results highlight several findings. In the first place, the evidence estimates from the three methods confirm that the model with four different porosity layers, that is $d = 4$, is the most supported by the available data. This finding is not surprising as this model uses the exact same layering of the porosity field as used in the synthetic GPR experiment that was used to create the

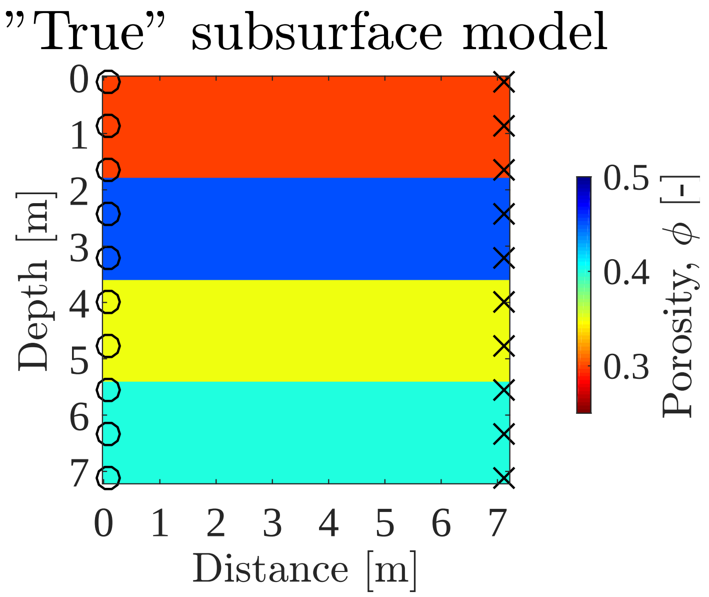


Figure A.3 – The "true" subsurface porosity model used in our synthetic crosshole-GPR experiment. The transmitter (black crosses) and receiver antennas (black circles) are indicated.

"measured" traveltime data. Secondly, all the three estimators are in excellent agreement and provide very similar values of the evidence for each of the conceptual models used. Thirdly, notice that all the three estimators exhibit a negligible uncertainty compared to the range of evidence values considered. The upper and lower bound values of the evidence derived from the three methods appear rather similar, demonstrating further the robustness of these three estimators.

We now investigate in more detail the discrepancies between the results of the three estimators, and plot in Figure A.5 the differences between the logarithmic values of the marginal likelihoods, $\mathscr{P}(\tilde{\mathbf{Y}})$, computed by the methods for the competing models used in this study. The solid black line depicts the difference in the evidence estimates derived from the three different methods and report results for subsurface models with number of horizontal porosity layers (equal thickness) that ranges from $d = 1$ to $d = 16$.

The results in Figure A.5 provide further evidence for our earlier conclusions. Indeed, the GMIS, LM and PC methods provide rather similar evidence values. In particular, we observe a stronger agreement between the PC and GMIS evidence (Figure A.5a) values at higher dimensions ($d > 8$) in comparison with the LM estimates (Figure A.5b). On the other hand, for the simpler subsurface models with up to $d = 8$ different porosity layers, the results from the GMIS and LM estimators are the most consistent with each other (Figure A.5c).

Different tests were performed to investigate the impact of the choice of n_{live} and $n_{repeats}$ on the PC evidence values considering that the number of forward simulations scales down linearly with decreasing n_{live} and $n_{repeats}$ (Figure A.6). All the different tests are listed in Table A.1.

In Figure A.7a and A.7b is shown respectively the effect of decreasing n_{live} (i.e., from $25d$ to d) and $n_{repeats}$ (i.e., from $5d$ to d) by plotting the differences of evidence estimates obtained

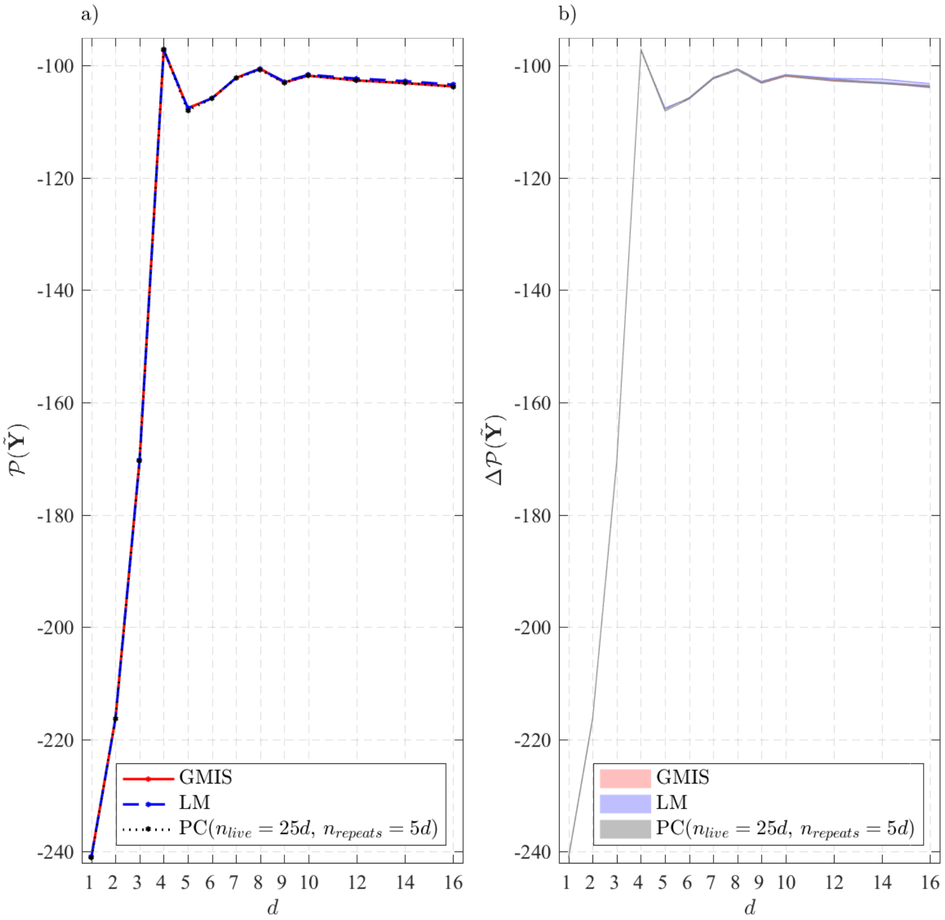


Figure A.4 – Evidence values in $\log_{10}$ space, $\mathscr{P}(\tilde{\mathbf{Y}})$ (a: left graph), and their associated uncertainty (b: right graph) derived from the GMIS, LM and PC estimators for each model dimension, d, used herein. Colour coding is used to differentiate among the different methods. The evidence estimates of the GMIS, LM and PC estimators are in excellent agreement and their uncertainty is small.

from tests T1 up to T5 with respect to the "best" case, B, for all the model dimensions considered. The results suggest that a decrease in n_{live} has an higher impact on the evidence estimates than a decrease in $n_{repeats}$. When decreasing n_{live}, we get differences in the evidence estimates of up to 1.5 $\log_{10}$ units and, when decreasing $n_{repeats}$, the differences are, at maximum, 0.5 $\log_{10}$ units. The evidence estimates start to significantly diverge from the "best" setting when decreasing n_{live} to $5d$ (case T2). Since we have changed n_{live} and $n_{repeats}$ proportionally to the model dimension, d, there is not any appreciable trend of the evidence differences with increasing model dimensions.

Even if we are mainly focused on evidence estimation rather than parameter estimation, it might be interesting to compare the posterior distributions obtained by MCMC and PC (Figure A.8). The posterior distributions of the inferred porosities obtained by the two methods match very nicely for both the cases where the layered model with two layers (Figure A.8a) and ten layers (Figure A.8b) are considered.

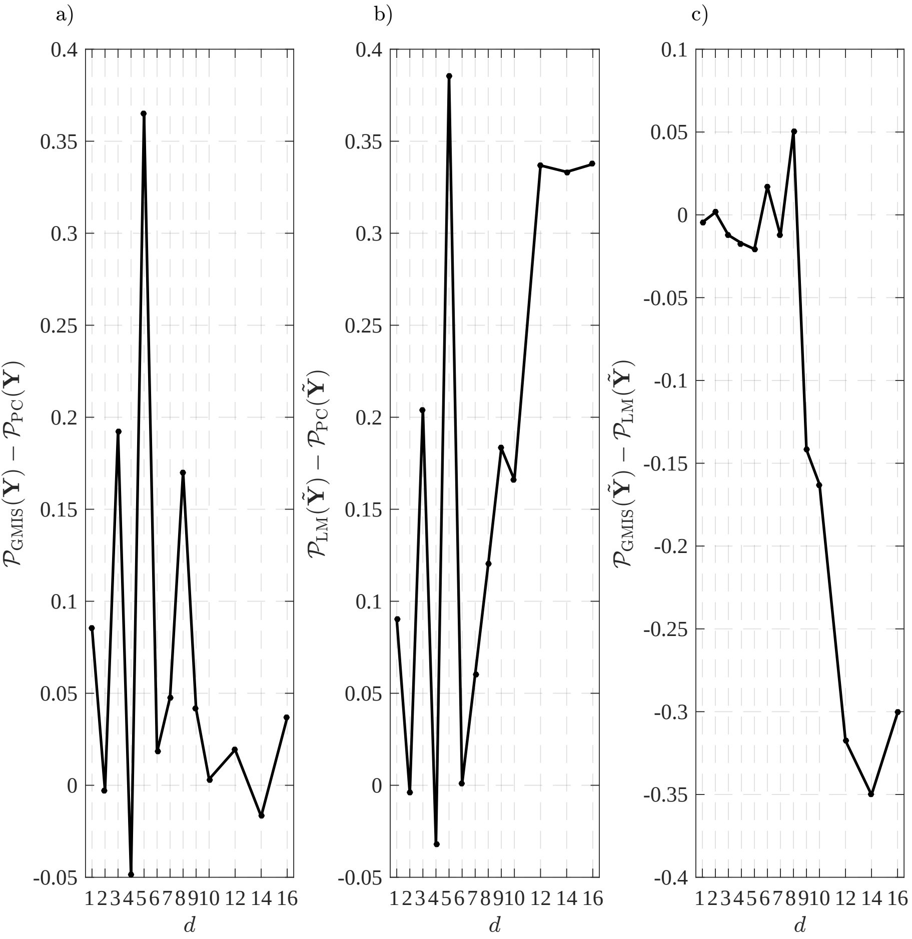


Figure A.5 – Difference in the evidence estimates derived from different pairs of two methods as function of model dimension, (a) GMIS and PC, (b) LM and PC, and (c) GMIS and LM. Note, we use $\log_{10}$ transformed value of the evidence estimates.

A.5 Field example

We now focus our attention on the South Oyster Bacterial Transport Site in Virginia, USA, and use geophysical data measured at this experimental site to determine the potential of POLYCHORD algorithm in providing reliable evidence values at high dimensions (i.e., $d \sim 100$) in a real case study. The geological characteristics of the South Oyster Bacterial Transport Site are described in (Hubbard et al., 2001). GPR traveltime data were measured at the S14-M13 borehole transect using a PulseEKKO 100 system with a 100-MHz nominal-frequency antenna. A domain of 7.2 × 7.2 m was measured with a total of 57 sources and 57 receivers, leading to a data set of 3248 observations of first-arrival traveltimes (one value is missing). We assume the measurement errors of the traveltime to be uncorrelated and normally distributed with constant standard deviation, $\sigma_{\tilde{\mathbf{Y}}}$. A relatively fine spatial discretization consisting of square cells with length 0.04 m was used in our forward simulations with the non-linear 2D traveltime solver (*time 2d*) of Podvin and Lecomte (1991) to compute the first-arrival traveltimes for the 7.2 × 7.2 m domain of interest.

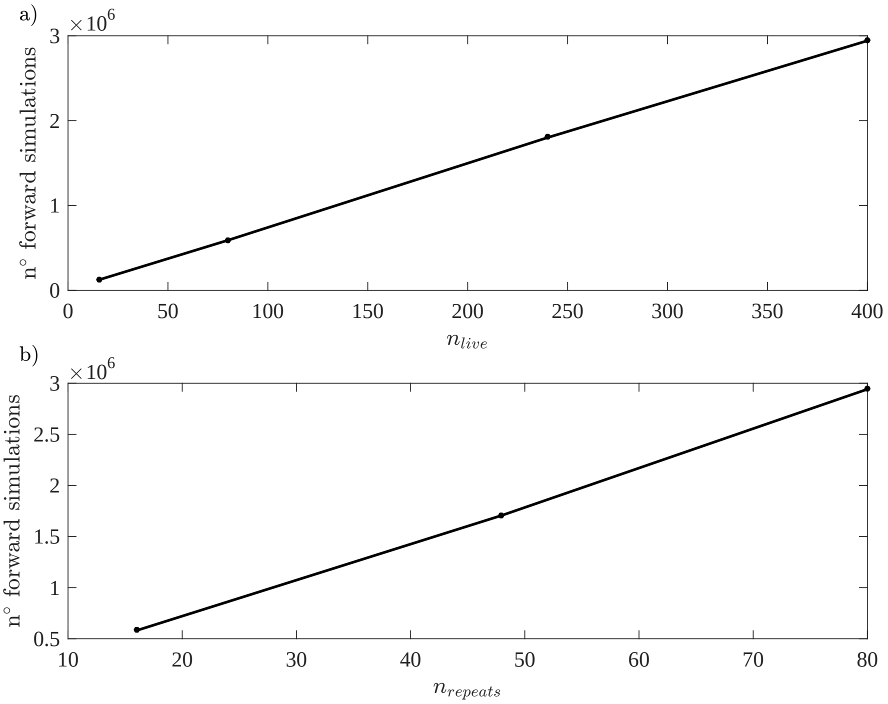


Figure A.6 – Linear relationship between the number of forward simulations and (a) n_{live} and (b) $n_{repeats}$ in the case of the layered model with 16 layers.

Table A.1 – In the column on the left, the abbreviations used to indicate the different tests are reported. d is the model dimension. Test B refers to the "best" case where n_{live} and $n_{repeats}$ are set as suggested by Handley et al. (2015b,a). Tests from T1 to T3 maintain a constant value of $n_{repeats}$ while decreasing n_{live} from $15d$ to d. Tests T4 and T5, instead, maintain n_{live} constant while decreasing $n_{repeats}$ up to d. Test W refers to the "worst" case in which both the n_{live} and $n_{repeats}$ are set quite low.

Test name	n_{live}	$n_{repeats}$
B	$25d$	$5d$
T1	$15d$	$5d$
T2	$5d$	$5d$
T3	d	$5d$
T4	$25d$	$3d$
T5	$25d$	d
W	d	d

A.5.1 Preliminary test

The conceptual subsurface model used in this preliminary test is a uniform 5 × 5 grid ($d = 28$) of the underlying porosity field at the South Oyster Bacterial Transport Site in Virginia. We estimate the evidence for this conceptual model by PC with different setting of n_{live} and $n_{repeats}$ as reported in Table A.2.

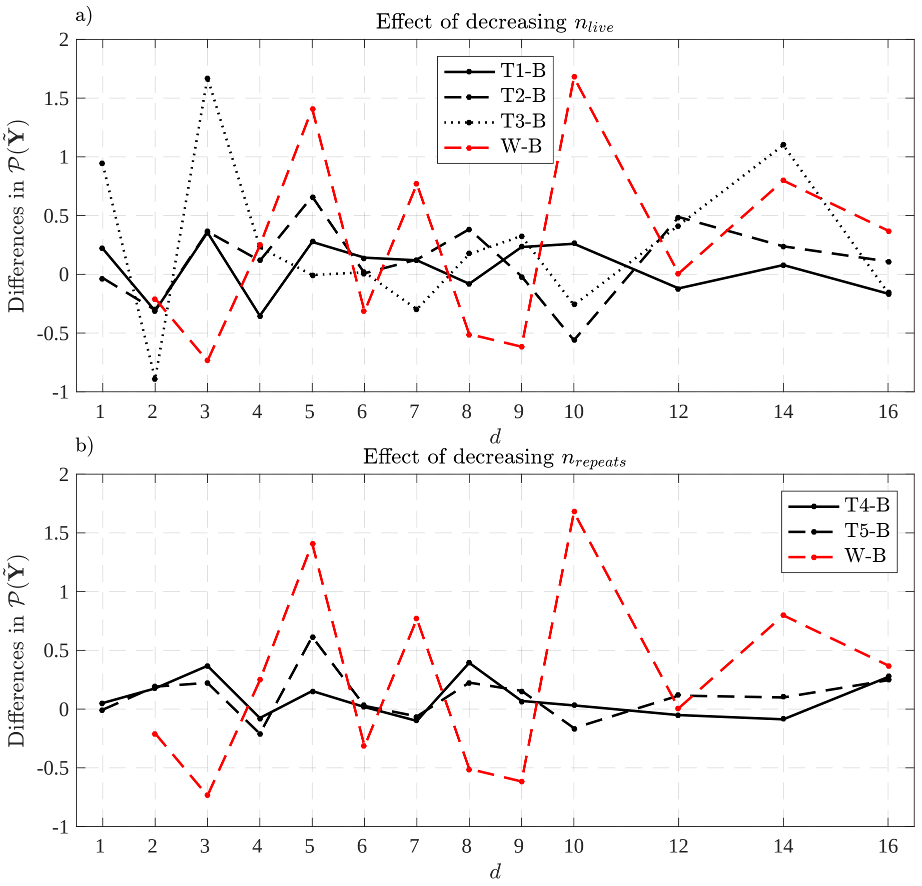


Figure A.7 – Difference in the evidence estimates derived from tests T1, T2, T3, T4, T5 (Table A.1) with respect to the "best" case, B, where n_{live} and $n_{repeats}$ are set as suggested by Handley et al. (2015b,a). (a) effect of decreasing n_{live} from $25d$ to d while keeping $n_{repeats}$ constant and (b) effect of decreasing $n_{repeats}$ from $5d$ to d while keeping n_{live} constant . The red line depicts the differences in evidence estimates between the "best" setting, B, and "worst" setting, W. Note, we use $\log_{10}$ transformed differences of the evidence estimates.

In Figure A.9, we show the effect of a decrease in n_{live} and $n_{repeats}$ on the evidence values. The earlier conclusions on the illustrative synthetic case are here reconfirmed: a decrease in n_{live} has an higher impact on the evidence estimates than a decrease in $n_{repeats}$. When decreasing n_{live}, we get a range (i.e., the difference between the largest and smallest evidence value) of about 107 $\log_{10}$ units (Figure A.9a) and, when decreasing $n_{repeats}$, the range is 30 $\log_{10}$ units (Figure A.9b).

In Figure A.9a, we observe that a large number of live points (i.e., at least $n_{live} = 20d$) is required to have negligible effect of the decrease of n_{live}. However, even if the evidence values appear to reach a plateau for $n_{live} \geq 20d$, the PC estimates do not approach the ones computed with the GMIS and LM. This is probably due to the fact that we set very low $n_{repeats} = 1$ and this introduces a bias on the evidence estimates as clearly shown in Figure A.9b. In the case of Figure A.9b, the plateau region where the evidence estimated with the

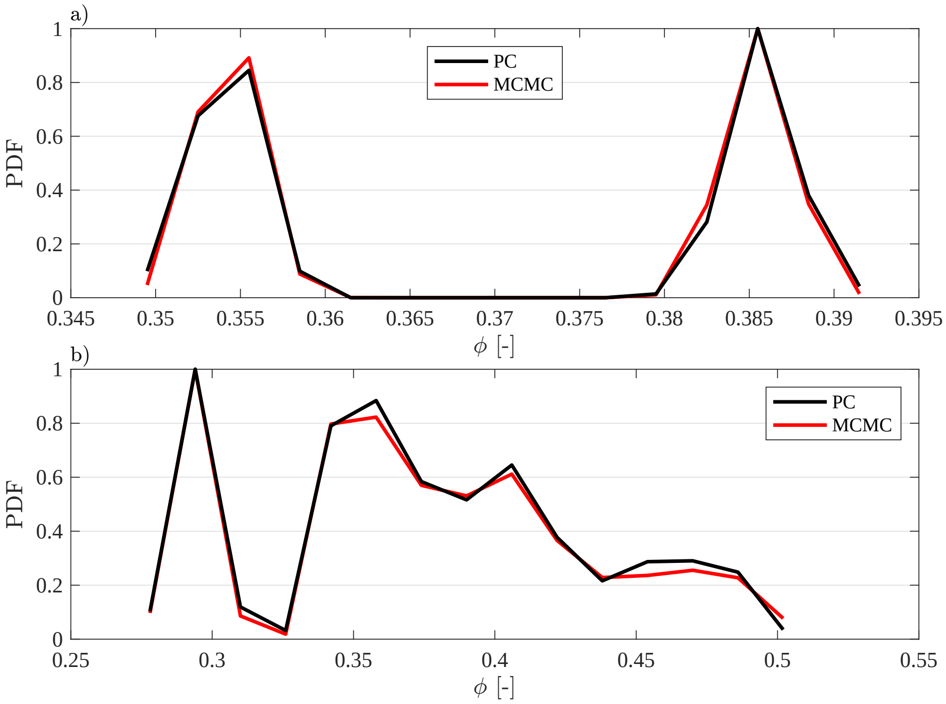


Figure A.8 – Marginal posterior distributions of the inferred porosity, ϕ, for the horizontally layered model with (a) 2 layers and (b) 10 layers. The posterior distributions inferred by PC refer to the setting with $n_{live} = 25d$ and $n_{repeats} = 5d$ (Test B). The densities in each plot are normalized and colour coding is used to differentiate among the different methods to infer the posterior distributions.

Table A.2 – In the column on the left, the abbreviations used to indicate the different tests are reported. d is the model dimension. Tests from R1 to R6 maintain a constant low value of $n_{repeats}$ while decreasing n_{live} from $25d$ to d. Tests from R7 to R11 and R5, instead, maintain n_{live} constant while decreasing $n_{repeats}$ from $5d$ to 1.

Test name	n_{live}	$n_{repeats}$
R1	$25d$	1
R2	$20d$	1
R3	$15d$	1
R4	$10d$	1
R5	$5d$	1
R6	d	1
R7	$5d$	$5d$
R8	$5d$	$4d$
R9	$5d$	$3d$
R10	$5d$	$2d$
R11	$5d$	d

three methods are quite similar is reached already in the case R11 where $n_{live} = 5d$ and $n_{repeats} = d$.

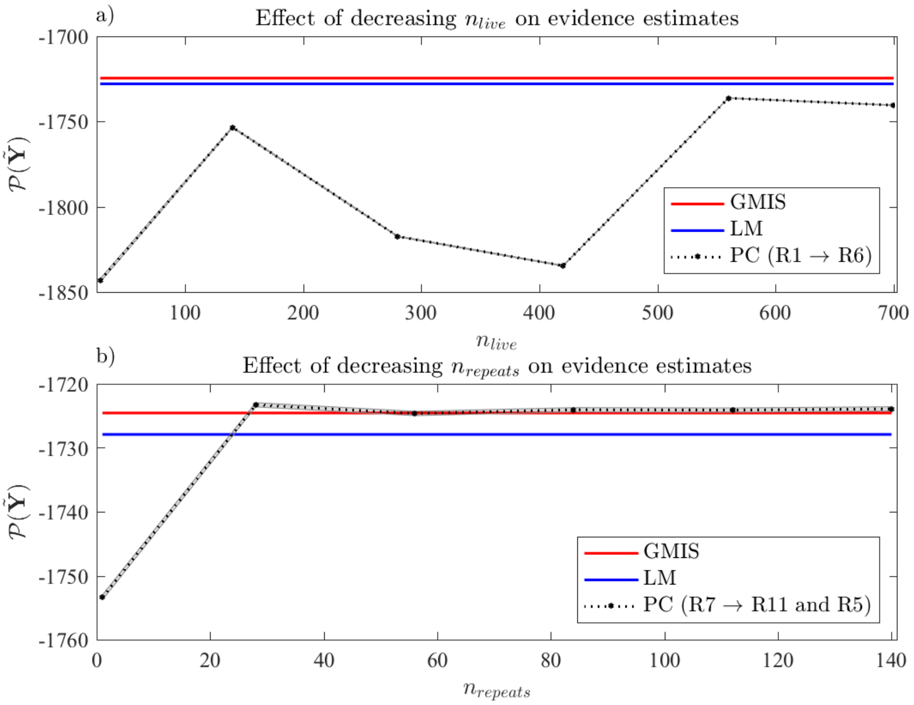


Figure A.9 – Evidence values in $\log_{10}$ space, $\mathscr{P}(\tilde{\mathbf{Y}})$, as function of (a) n_{live} and (b) $n_{repeats}$ for a uniform gridded conceptual model with 5 × 5 grid cells. The evidence estimated by GMIS and LM methods are indicated with the red line and the blue line, respectively.

A.5.2 Results

The conceptual models used in this study are uniform grids which differ in their discretization of the subsurface. The numbers of porosity grid cells is varied between 1 to 100, thereby providing a large array of competing hypotheses. In this real hydrogeological setting where we consider relatively high dimensions, large data sets and small measurement errors, the computation of the evidences by PC is limited by the high computational cost. For this reason, we could not perform the PC algorithm with the "best" choice of n_{live} and $n_{repeats}$ (i.e., $n_{live} = 25d$ and $n_{repeats} = 5d$), as previously done in the illustrative synthetic case. We compute, instead, the evidences with PC, setting n_{live} at least equal or greater than $4d$ and $n_{repeats} = d$ (Table A.3), since they are the smallest values that we can set for not having too degraded evidence values, as we show in the preliminary test in Section A.5.1. Chosing n_{live} at least $4d$ is quite low but we attempt to find a trade-off between n_{live} and $n_{repeats}$ that still allows us to get reliable and accurate evidence estimates in a feasible computational time.

Before showing the results of the evidence estimations, the posterior distributions obtained by MCMC and PC are compared in the case of the gridded model with 5 × 5 grid cells (Figure A.10). The posterior distributions of the inferred porosities obtained by the two methods match quite nicely (Figure A.10a). The posterior distributions obtained by PC and MCMC show some differences in the case of the inferred cementation index, m, (Figure A.10b): where the MCMC gives the highest probability the PC algorithm gives the lowest probability. However, both methods suggest that values of $m < 1.4$ are more probable than higher ones.

Table A.3 – The first column shows the number of grid cells of each conceptual model considered herein; the second column lists the model dimension of each conceptual model; in the third and fourth columns, the values of n_{live} and $n_{repeats}$ used for computing the evidence are listed.

n° grid cells	d	n_{live}	$n_{repeats}$
2 × 2	7	200	d
3 × 3	12	200	d
4 × 4	19	300	d
5 × 5	28	300	d
6 × 6	39	400	d
9 × 9	84	400	d
10 × 10	103	500	d

The overall trend of the posterior distributions of the inferred relative permittivity of the mineral grains, ϵ_s and the inferred data error, $\sigma_{\tilde{\mathbf{Y}}}$, is quite similar (Figure A.10c-d).

We now turn our attention to the evidence of each model. Figure A.11 depicts the results of this analysis using a $\log_{10}$ transformation of the evidence values of each uniform gridded conceptual model which uses between 4 to 100 different porosity grid cells. Colour coding is used to differentiate between the GMIS (red), LM (blue) and PC (black) estimators.

The evidence estimates derived from all three methods appear almost similar for model complexities with less than 39 (unknown) parameters. Beyond this, the marginal likelihoods derived from the three methods diverge from each others reaching differences of the order of 10^2 in $\log_{10}$ space for $d = 103$.

In general, it is not possible to know exactly the evidence value for a given conceptual model. However, comparing different methods, we can find the one that provides possibly the most correct evidence estimation. In order to asses this issue, we compare the evidence estimates obtained by different setting of the GMIS and PC methods and by the LM method (Figure A.12). We have already mentioned that in order to increase the accuracy and reliability of the evidence estimates by the PC algorithm, we need to increase as much as possible n_{live} and $n_{repeats}$. In the case of the GMIS method, we can get better results by increasing the number of importance samples, N, and the number of evidence estimates, N_{rep}, used to compute the mean evidence. We find that increasing the accuracy of the evidence estimates by increasing N and N_{rep} in the GMIS method and by increasing n_{live} and $n_{repeats}$ in the PC method, we obtain higher evidence values. This result lead us to assume that the method that provides lower evidence estimates (i.e., the LM method) is less accurate in comparison with the other methods. We also notice that the evidence estimates provided by PC with different settings are characterized by a much higher variability in comparison with the GMIS method for which, instead, the evidence estimates are quite similar even if the settings used are quite different.

We now investigate the reasons why the PC algorithm is computationally costly in our setting. If we focus on the number of forward simulations required for computing the evidence for

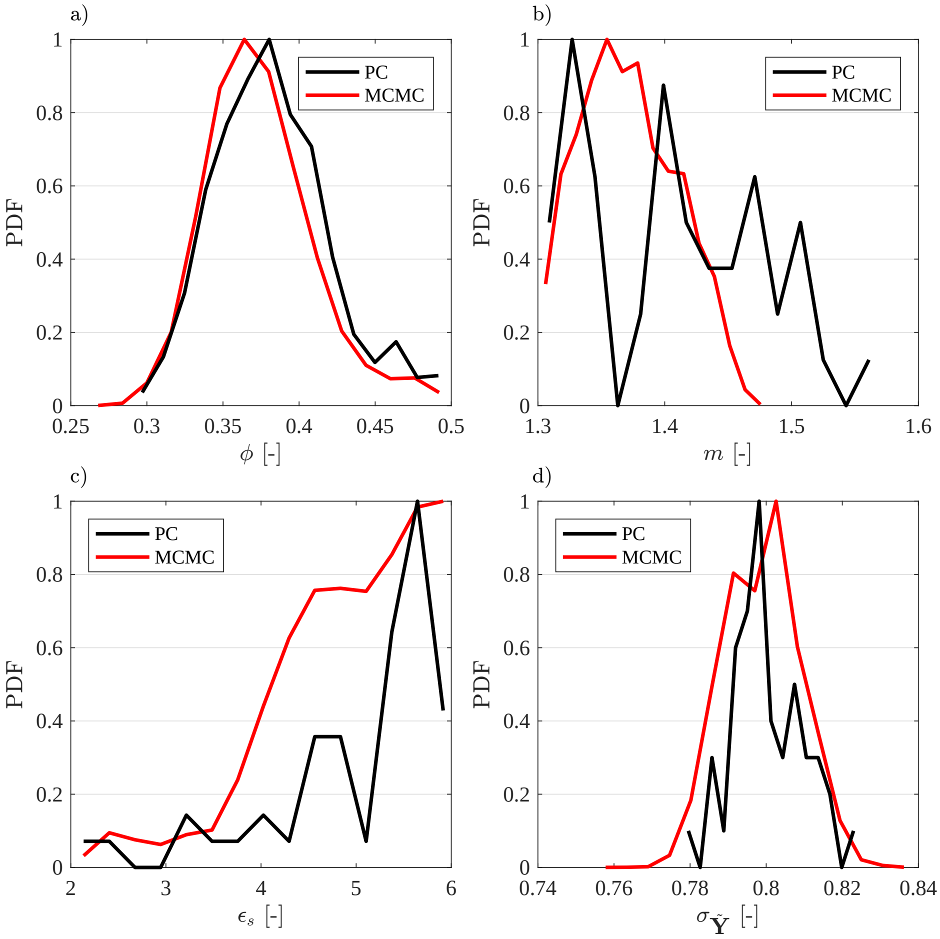


Figure A.10 – Marginal posterior distributions of (a) the inferred porosity, ϕ, (b) the inferred cementation index, m, (c) the inferred relative permittivity of the mineral grains, ϵ_s and (d) the inferred data error, $\sigma_{\tilde{\mathbf{Y}}}$, for the gridded model with 5 × 5 grid cells. The posterior distributions inferred by PC refer to the setting with $n_{live} = 300$ and $n_{repeats} = d$ (i.e., in this case d is equal to 28). The densities in each plot are normalized and colour coding is used to differentiate among the different methods used to infer the posterior distributions.

each model dimension, it is clear that the number of forward simulations performed with PC increase exponentially with model dimension. For $d \geq 84$, the number of forward simulations required by PC is at least one order of magnitude larger than the ones required by the other two estimators.

As a consequence, if we now look at the computational time of each of the three estimators for each model dimension considered herein (Table A.4), we observe that the time for estimating the evidence by PC may last up to 11 days for $d = 103$.

However, if we consider the time for performing one single forward simulation, all the three methods require comparable computational time but we notice that the PC is faster (i.e., 0.024 s) than the GMIS or LM (i.e., 0.034 s).

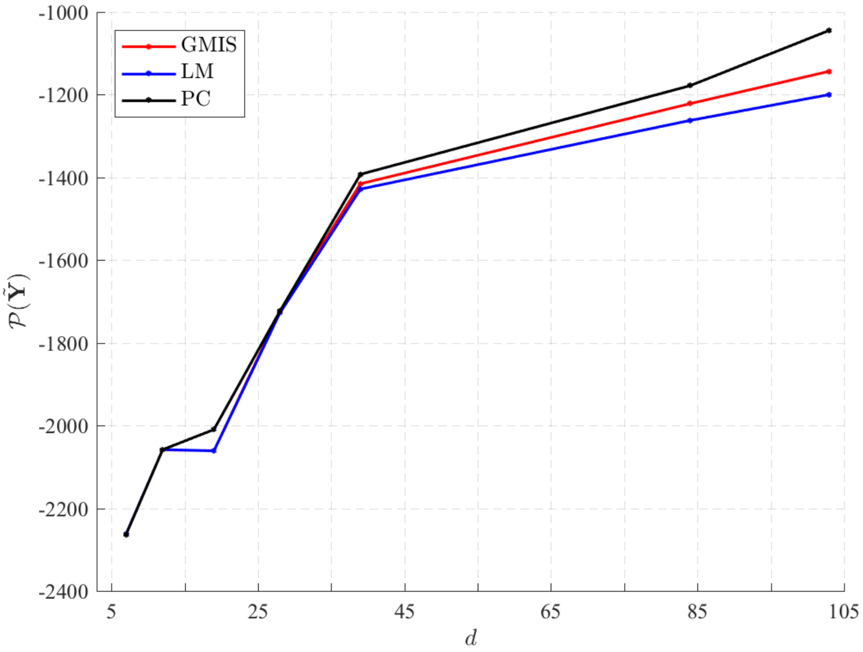


Figure A.11 – Evidence in $\log_{10}$ space, $\mathscr{P}(\tilde{\mathbf{Y}})$ derived from the GMIS, LM and PC estimators for each model dimension, d, used herein. Colour coding is used to differentiate among the different methods.

Table A.4 – Computational time expressed in hours for estimating the evidence by GMIS, LM and PC. The LM method uses the same computational time as the MCMC run and the GMIS method requires a MCMC run and sampling from the importance distribution which, in our setting, requires from 3 to 4 hours more.

Computational time				
n° grid cells	t_{PC} [h]	t_{MCMC} [h]	t_{LM} [h]	t_{GMIS} [h]
2 × 2	1.0	9.5	t_{MCMC}	$t_{\mathrm{MCMC}}+3$
3 × 3	10.0	10.0	t_{MCMC}	$t_{\mathrm{MCMC}}+3$
4 × 4	18.5	9.0	t_{MCMC}	$t_{\mathrm{MCMC}}+3$
5 × 5	22.0	9.5	t_{MCMC}	$t_{\mathrm{MCMC}}+3$
6 × 6	89.0	9.5	t_{MCMC}	$t_{\mathrm{MCMC}}+3$
9 × 9	288.5	36.0	t_{MCMC}	$t_{\mathrm{MCMC}}+4$
10 × 10	267.0	38.0	t_{MCMC}	$t_{\mathrm{MCMC}}+4$

A.6 Discussion

We investigate the potential of POLYCHORD algorithm in providing reliable evidence estimates in high model dimensions in the context of synthetic and real-world hydrogeophysical settings and the main observations are as follows.

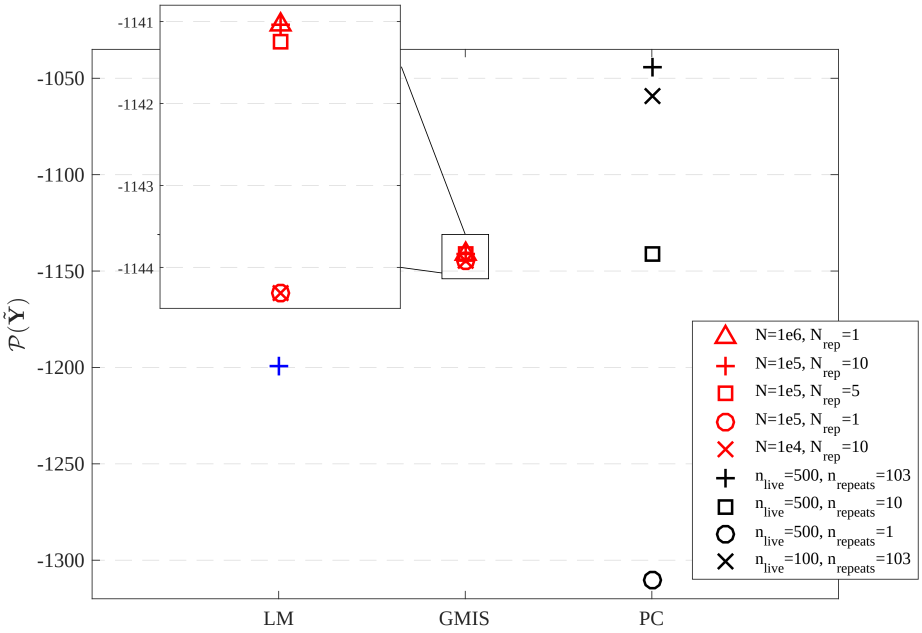


Figure A.12 – Evidence in $\log_{10}$ space, $\mathscr{P}(\tilde{\mathbf{Y}})$ derived from the GMIS, LM and PC estimators using different setting as shown in the legend. In the case of the GMIS method, N indicates the number of importance samples and N_{rep} is the number of evidence estimates used to compute the mean evidence. Colour coding is used to differentiate among the different methods.

In the illustrative synthetic case (Section A.4) where we use 100 GPR traveltimes data, relatively low model dimensions (from $d = 1$ to $d = 16$) and quite simple conceptual models (horizontally layered model), the results suggest that GMIS, LM and PC are in excellent agreement and provide nearly similar values of the evidence for each of the competing conceptual models considered herein. Moreover, all the three estimators exhibit a small uncertainty compared to the range of evidence values considered underlying the robustness of these three methods.

In both the illustrative synthetic case (Section A.4) and the preliminary test on the field case (Section A.5.1), we find that a decrease in n_{live} has an higher impact on the evidence estimates than a decrease in $n_{repeats}$. In particular, we find that in the field case study, setting $n_{repeats} = d$ is reasonably sufficient to achieve reliable evidence estimates with PC. On the other hand, it is clear that keeping a large number of live points is important for avoiding biased evidence values. Unfortunately, in the context of the real case study at the South Oyster Bacterial Transport Site in Virginia (Section A.5.2) where we consider many GPR traveltimes data (i.e., 3248), relatively high model dimensions (from $d = 7$ to $d = 103$) and more complex conceptual models (uniform grid models), it is not possible to set large n_{live} values due to the high cost of the computational time. As a consequence, the evidence estimates derived from GMIS, LM and PC appear almost similar for $d < 39$ but, beyond this, the marginal likelihoods derived from the three methods diverge from each other reaching differences of the order of 10^2 in log10 space for $d = 103$.

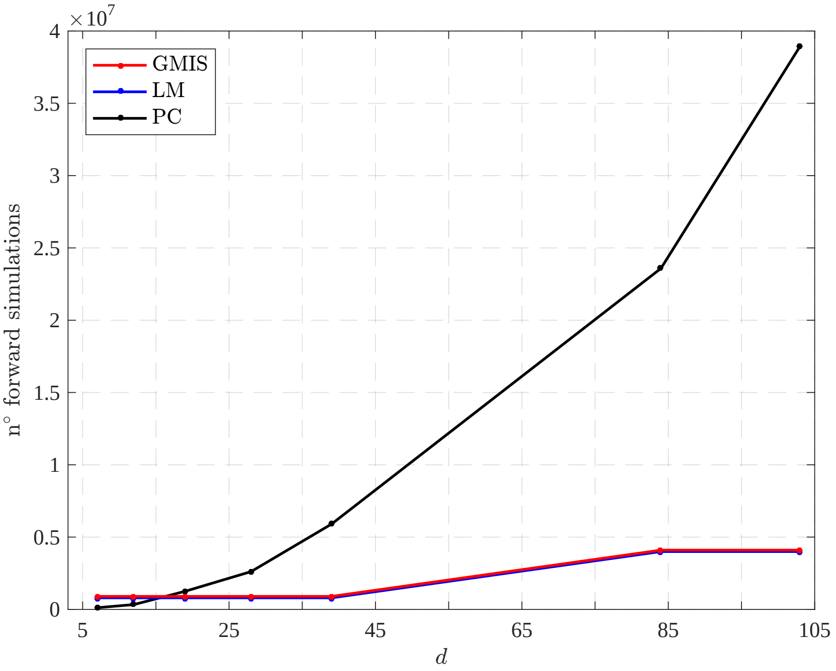


Figure A.13 – Number of forward simulations required for estimating the evidence by GMIS (red), LM (blue) and PC (black) as a function of model dimension. At high model dimensions, the number of forward simulations required by PC is at least one order of magnitude larger than the ones required by GMIS and LM estimators.

The high computational cost of PC is explained by the fact that, for $d \geq 84$, the number of forward simulations required by PC is at least one order of magnitude larger than the ones required by the other two estimators. However, PC results in almost 42% decrease in computational time for evaluating one single forward simulation with respect to the corresponding time required by the GMIS or LM methods.

We find that increasing the accuracy of the evidence estimates by increasing N and N_{rep} in the GMIS method and by increasing n_{live} and $n_{repeats}$ in the PC method, we obtain higher evidence values. This suggests that the method that provides higher evidence estimates (i.e., the PC algorithm) is more accurate in comparison with the other methods in providing evidence estimates.

A.7 Conclusions

When using POLYCHORD, the accuracy and reliability of the evidence estimates are easy to control by the user by setting appropriate n_{live} and $n_{repeats}$. However, in our real-world hydrogeophysical setting, POLYCHORD is computationally costly since it requires many more forward simulations than the other two methods. For this reason, we can not fully benefit from the potential of POLYCHORD since, to get feasible computational time, we have to set smaller n_{live} and $n_{repeats}$ than the values suggested by Handley et al. (2015b,a) which results in less accurate evidence estimates. However, we need to underline the fact that in order to obtain a robust model selection we do not necessarily need accurate absolute estimates of the evidence but we need a method which provides reliable relative evidence estimates among a competing set of conceptual models.