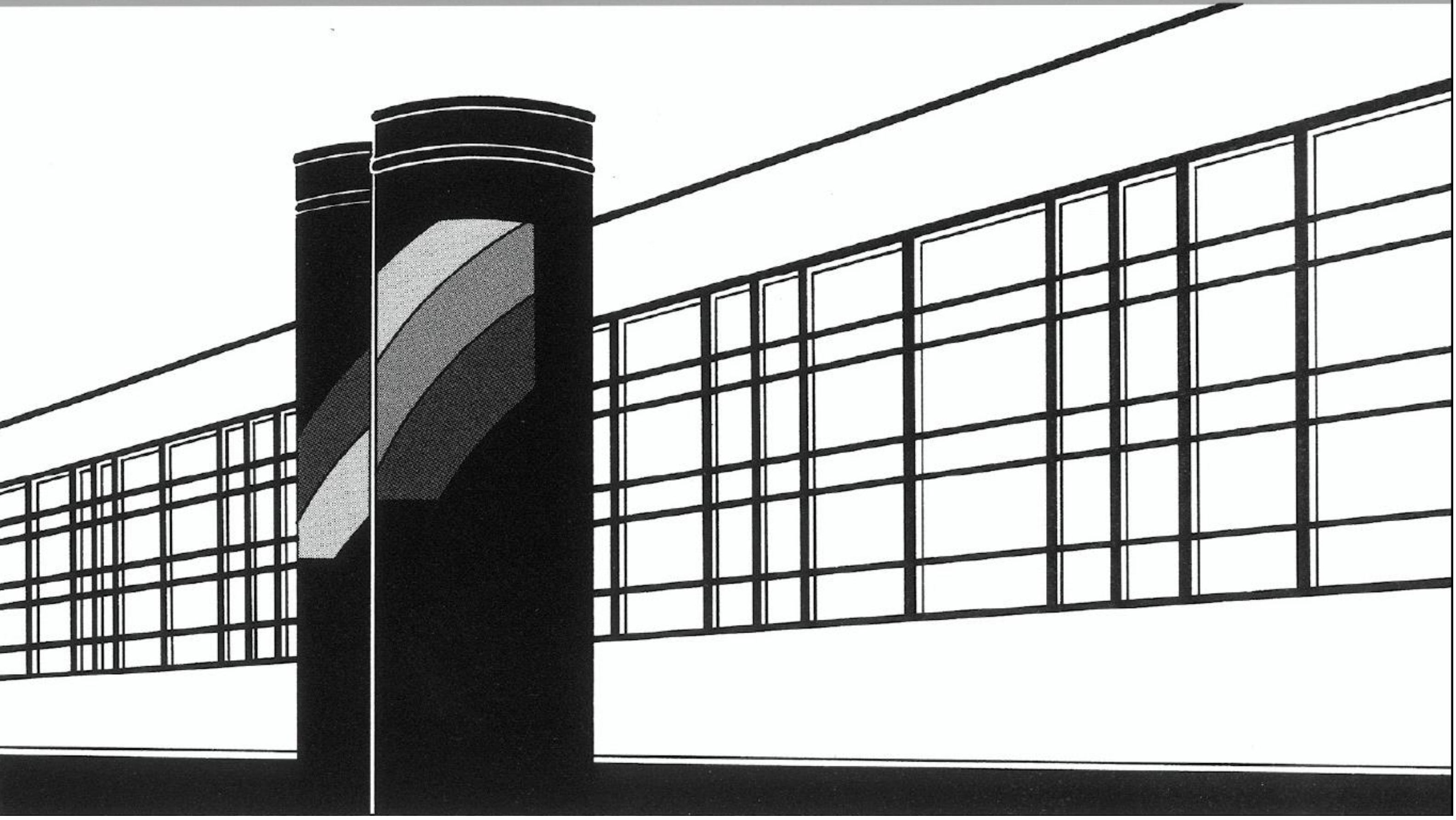
Universität Stuttgart

Institut für Wasser- und Umweltsystemmodellierung

Mitteilungen



Heft 285 Sebastian Reuschen

Bayesian Inversion and Model Selection of Heterogeneities in Geostatistical Subsurface Modeling

Bayesian Inversion and Model Selection of Heterogeneities in Geostatistical Subsurface Modeling

von der Fakultät Bau- und Umweltingenieurwissenschaften der Universität Stuttgart und des Stuttgarter Zentrums für Simulationswissenschaften zur Erlangung der Würde eines Doktor-Ingenieurs (Dr.-Ing.) genehmigte Abhandlung

vorgelegt von

Sebastian Reuschen

aus Nürtingen

Hauptberichter:	Prof. Dr.-Ing. Wolfgang Nowak
Mitberichter:	Prof. Dr. rer. nat. Dr.-Ing. habil. András Bárdossy
	Prof. Dr. Niklas Linde

Tag der mündlichen Prüfung:	2.12.2021

Institut für Wasser- und Umweltsystemmodellierung
der Universität Stuttgart
2021

Heft 285

Bayesian Inversion and Model Selection of Heterogeneities in Geostatistical Subsurface Modeling

von
Dr.-Ing.
Sebastian Reuschen

Eigenverlag des Instituts für Wasser- und Umweltsystemmodellierung
der Universität Stuttgart

D93 Bayesian Inversion and Model Selection of Heterogeneities in Geostatistical Subsurface Modeling

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://www.d-nb.de abrufbar

Sebastian Reuschen:
Bayesian Inversion and Model Selection of Heterogeneities in Geostatistical Subsurface Modeling, Universität Stuttgart. - Stuttgart: Institut für Wasser- und Umweltsystemmodellierung, 2021

(Mitteilungen Institut für Wasser- und Umweltsystemmodellierung, Universität Stuttgart: H. 285)
Zugl.: Stuttgart, Univ., Diss., 2021
ISBN 978-3-942036-89-4

NE: Institut für Wasser- und Umweltsystemmodellierung <Stuttgart>: Mitteilungen

Herausgegeben 2021 vom Eigenverlag des Instituts für Wasser- und Umweltsystemmodellierung

Druck: DCC Kästl e.K., Ostfildern

Danksagung

Hiermit möchte ich mich bei allen denjenigen bedanken, die mich während meiner Promotion begleitet haben. Ein herzliches Dankeschön geht an Wolfgang Nowak, der mir die Möglichkeit gab, an seinem Lehrstuhl zu forschen. Ich danke dir für die fachliche und persönliche Unterstützung, welche stets gut und ehrlich war. Außerdem danke ich meinen Mitberichtern András Bárdossy und Niklas Linde für die exzellenten Diskussionen, welche sehr geholfen haben, die Richtung dieser Arbeit zu festigen.

Großer Dank geht an meine Kollegen vom LS3. Ich habe mich in dieser Gruppe sehr wohl gefühlt und bin froh, dass ich so viel Glück mit so tollen Kollegen hatte. Besonderer Dank geht hierbei an Aline, Anneli, Jannik, Micha, Sinan, Teng und Ute, die immer für fachliche und persönliche Gespräche offen waren und deren Meinung mir sehr viel bedeutet.

Außerdem danke ich meinen SimTech-Kommilitonen, welche mich während der letzten 10 Jahre an der Uni Stuttgart begleitet haben. Es waren sehr schöne Jahre und ihr habt daran einen großen Beitrag geleistet. Besonders danke ich dabei meiner D&D Gruppe, bestehend aus Anika, Basti, Claudius, David, Dome, Etienne, Jan und Rapha, welche in den letzten 8 Jahren jede Woche für den richtigen Ausgleich gesorgt haben.

Abseits der Uni danke ich meinen Eltern, Christine und Rolf, dafür, dass sie immer hinter mir standen und mich in allem, was ich machte, unterstützten. Abschließend will ich besonders meiner Freundin Isi danken. Ich danke dir für die interessanten wissenschaftlichen Diskussionen und die tolle Unterstützung in guten und in schlechten Zeiten.

Ich danke euch allen!

Contents

List of Figures

Nomenclature

Selected Acronyms

BME	Bayesian model evidence
BMS	Bayesian model selection
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hasting
MPS	Multiple-point statistics
pCN-MCMC	Pre-conditioned Crank Nicolson MCMC

Symbols

α	Acceptance probability
$\mathbf{d}$	Data
$\overline{\alpha}$	Acceptance rate (percentage of accepted proposals)
$\boldsymbol{\theta}_i$	Parameter at i-th MCMC step
$\boldsymbol{\theta}_j$	Parameter at j-th (typically i+1) MCMC step
$\boldsymbol{\theta}$	Parameter
$\boldsymbol{\theta}_m$	Parameter of model m

π	Target distribution
$\widetilde{\mathbf{d}}$	artificially generated data
$h(.,.)$	Transition kernel
$L(\boldsymbol{\theta})$	Likelihood function ($p(\mathbf{d}\|\boldsymbol{\theta}_m, M_m)$)
M_G	Gaussian random field model
M_m	Model m
M_{MPS}	MPS prior model
$p(\mathbf{d}\|\boldsymbol{\theta}_m, M_m)$	Likelihood function
$p(\mathbf{d}\|M_m)$	Bayesian model evidence
$P(\boldsymbol{\theta})$	Prior distribution
$p(\boldsymbol{\theta}\|M_m)$	Prior distribution under model m
$p(\boldsymbol{\theta}\|\mathbf{d}, M_m)$	Posterior (parameter) distribution
$p(M_k\|M_i)$	Average posterior model weight of model k given data from model i
$P(M_m)$	Prior model weight of model m
$p(M_m\|\mathbf{d})$	Posterior model weight of model m
$q(.,.)$	Proposal distribution
T	Temperature

Abstract

Motivation The planning of drinking water supply and the evaluation of subsurface CO_2 and nuclear waste storage sites are just a few examples that require knowledge about the subsurface structure. While drinking water wells should be built in regions with highly permeable soils, nuclear waste storage sites should be built in impermeable salt rock. In this example, the hydraulic conductivity can be used to describe the permeability of soils. Determining such subsurface parameters is difficult because they are heterogeneous and only scarce data is available. Consequently, the spatial distribution of subsurface parameters is uncertain. To overcome this challenge, parameter inference and uncertainty quantification can be used to (1) improve predictions and (2) quantify remaining uncertainties in various applications.

This thesis aims to improve the inference and uncertainty quantification of relevant subsurface parameters. To reach this goal, a two-step approach is taken. First, Bayesian inversion, also known as Bayesian inference, is used for probabilistic predictions of subsurface parameter fields. Here, Bayesian inversion defines a so-called posterior distribution which is the updated belief of the prior parameter model given data. Analytical calculation of the posterior is often impossible. Instead, numerical sampling methods can be used to approximate the posterior. *Markov chain Monte Carlo* (MCMC) methods have shown great potential in doing so. Depending on the amount of available data and type of prior field, MCMC methods still have two drawbacks. In some problem classes, existing MCMC methods are too slow to be used commonly. In other problem classes, no MCMC methods exist to perform Bayesian inversion. This thesis closes these gaps by developing novel MCMC methods to enable or speed up Bayesian inversion. Second, all MCMC methods assume some prior (geostatistical) models. Selecting appropriate prior geostatistical models can be done using Bayesian model selection (BMS). This thesis shows that BMS results can differ based on where measurement noise is modeled and gives guidance on how to model noise correctly.

My contributions This thesis comprises a total of five journal articles and manuscripts. In Reuschen et al. [2021b], the *sequential pCN-MCMC* is developed which combines the ideas of the current state-of-the-art methods *sequential Gibbs sampling* and *preconditioned Crank Nicolson MCMC* (pCN-MCMC) for Gaussian prior fields. It is shown that the combined method is up to 6.5 times more efficient than the mentioned state-of-the-art alternatives for Gaussian prior fields and weakly informative data.

For highly informative data and Gaussian priors, Bayesian inversion and posterior sampling are more challenging due to more likelihood-dominated posteriors. Xu et al. [2020] introduces the *pCN-PT* MCMC, which enables Bayesian inversion with highly informative data by combining *parallel tempering* with the *pCN-MCMC* approach.

As the next step, the *pCN-PT* is generalized to Gaussian priors with uncertain variogram statistics. Xiao et al. [2021] summarizes how the variogram hyperparameters, e.g., the lengthscale or variance, can be inferred jointly with the parameter field. Further, it gives insights on which type of measurements are suitable for hyperparameter approximation.

While the first three contributions focus on Gaussian prior models, Reuschen et al. [2020] presents an MCMC that can perform Bayesian inversion on hierarchical prior models. This enables a more realistic representation of nature. The presented MCMC is applicable to highly and weakly informative data, and the differences in the resulting posterior distributions are analyzed.

The second large challenge besides posterior approximation via numerical sampling methods is the selection of appropriate prior geostatistical models using BMS. Reuschen et al. [2021a] shows that the results of BMS heavily depend on where (in model and/or data) noise is modeled. Further, guidance on how noise should be modeled in BMS and which scientific question is answered based on this choice is explained and discussed. Possible pitfalls are pointed out to prevent errors in future BMS studies.

Summary With these contributions, this thesis enhances the existing methods to infer, and quantify the uncertainty of, subsurface parameter fields. The presented methods enable a broad range of users in scientific and practical applications to achieve better predictions with less computational effort. Further, this thesis sheds light on the theoretical implications of measurement noise modeling. Altogether, these contributions enable better planning of subsurface-related applications such as drinking water supply or nuclear waste storage.

Zusammenfassung

Motivation Die Planung der Trinkwasserversorgung, sowie die Evaluierung von unterirdischen CO_2-Speichern und Atommüll-Endlagern sind nur einige Beispiele, bei denen Kenntnisse über die Beschaffenheit des Untergrunds erforderlich sind. Während Trinkwasserbrunnen in Regionen mit stark durchlässigen Böden gebaut werden sollten, sollten Atommüll-Endlager in undurchlässigen Salzstöcken errichtet werden. Dabei kann beispielsweise die hydraulische Leitfähigkeit als Parameter genutzt werden, um die Durchlässigkeit von Böden zu beschreiben. Die Bestimmung solcher Untergrundparameter ist schwierig, da sie heterogen sind und nur wenige Daten zur Verfügung stehen. Folglich ist die räumliche Verteilung der Untergrundparameter unsicher. Dies ist eine Herausforderung in vielen Anwendungen. Parameterinferenz und Unsicherheitsquantifizierung können eingesetzt werden, um (1) die Vorhersagen zu verbessern und (2) die verbleibenden Unsicherheiten zu quantifizieren.

Ziel dieser Arbeit ist es, die Inferenz und Unsicherheitsquantifizierung von relevanten Untergrundparametern zu verbessern. Dieses Ziel wird in zwei Schritten angegangen: Im ersten Schritt wird die Bayes'sche Inversion, auch bekannt als Bayes'sche Inferenz, für probabilistische Vorhersagen von Untergrundparameterfeldern verwendet. Ziel der Bayes'schen Inversion ist es, die sogenannte A-posteriori-Verteilung, die an die Daten assimilierte A-priori Verteilung, zu finden. Eine analytische Berechnung der A-posteriori-Verteilung ist oft nicht möglich. Stattdessen können numerische Stichprobenverfahren verwendet werden, um die A-posteriori-Verteilung zu approximieren. *Markov chain Monte Carlo* (MCMC) Methoden sind eine Klasse von Methoden, die in der Literatur sehr erfolgreich zum Approximieren der A-posteriori-Verteilung genutzt werden. Abhängig von der Menge der verfügbaren Daten und der Art des A-priori-Feldes haben MCMC Methoden dennoch zwei Nachteile: In einigen Problemklassen sind die vorhandenen MCMC Methoden zu rechenintensiv, um auch außerhalb der

Wissenschaft praktikabel zu sein. In anderen Problemklassen existieren keine passenden MCMC Methoden. In dieser Arbeit gehe ich diese Herausforderungen an, indem ich neue MCMC Methoden entwickle, die eine Bayes'sche Inversion entweder ermöglichen oder beschleunigen. Im zweiten Schritt wird das Problem betrachtet, dass alle MCMC Methoden (geostatistische) A-priori-Modelle voraussetzen. Die Auswahl geeigneter geostatistischer A-priori-Modelle kann mittels Bayes'scher Modellauswahl (eng. Bayesian model selection, BMS) erfolgen. Für diese Modellauswahl werden Messdaten benötigt. Diese Arbeit zeigt, dass sich die BMS Ergebnisse je nach Modellierung des Messrauschens unterscheiden können und gibt Hinweise zur korrekten Modellierung des Messrauschens.

Meine Beiträge Diese Arbeit umfasst insgesamt fünf Publikationen und Manuskripte. In Reuschen et al. [2021b] wird die *sequential pCN-MCMC* Methode entwickelt. Diese kombiniert die *sequential Gibbs sampling* Methode mit der *pre-conditioned Crank Nicolson MCMC* (pCN-MCMC) Methode für Gauß'sche A-priori-Felder. Es wird gezeigt, dass die kombinierte Methode für Gauß'sche A-priori-Felder und schwach informative Daten bis zu $6,5$ mal effizienter ist als die derzeitig besten Alternativen.

Für hochinformative Daten und Gauß'sche A-priori-Felder sind die Bayes'sche Inversion und das A-posteriori-Sampling anspruchsvoller, da die A-posteriori-Verteilung hauptsächlich von der Likelihood-Funktion abhängt. Xu et al. [2020] präsentiert den *pCN-PT* MCMC, welcher die Bayes'sche Inversion mit hochinformativen Daten ermöglicht. Dafür werden in dieser Publikation die Ideen von *parallel tempering* und des *pCN-MCMC* Ansatz kombiniert.

Im nächsten Schritt wird die *pCN-PT* Methode auf Gauß'sche A-priori-Felder mit unsichererem Variogramm verallgemeinert. Xiao et al. [2021] fasst zusammen, wie die Variogramm-Hyperparameter, z. B. die Längenskala oder die Varianz, gemeinsam mit dem Parameterfeld geschätzt werden können. Außerdem gibt diese Publikation Einblicke darauf, welche Art von Messungen sich für Rückschlüsse auf die Hyperparameter am besten eignen.

Während sich die ersten drei Beiträge auf Gauß'sche A-priori-Modelle konzentrieren, wird in Reuschen et al. [2020] eine MCMC Methode vorgestellt, welche die Bayes'sche Inversion mit hierarchischen A-priori-Modellen durchführt. Dies ermöglicht eine realistischere Darstellung der Natur. Die vorgestellte MCMC Methode ist mit hoch- und

schwach-informativen Daten anwendbar und die Unterschiede in den resultierenden A-posteriori-Verteilungen werden analysiert.

Die zweite große Herausforderung, neben der A-posteriori-Approximation mittels MCMC Methoden, ist die Auswahl geeigneter geostatistischer A-priori-Modelle unter Verwendung von BMS. Reuschen et al. [2021a] zeigt, dass die BMS Ergebnisse stark davon abhängen, an welcher Stelle (im Modell und/oder in den Daten) Messrauschen modelliert wird. Außerdem wird erläutert und diskutiert, wie Messrauschen in BMS modelliert werden sollte und welche philosophische Frage durch diese Wahl beantwortet wird. Abschließend wird auf mögliche Fehler in der BMS Berechnung hingewiesen, um diese in zukünftigen BMS Studien zu vermeiden.

Zusammenfassung Diese Arbeit stellt neue Methoden zur Inferenz und Unsicherheitsquantifizierung von Untergrundparameterfeldern vor. Die vorgestellten Methoden ermöglichen einem breiten Anwenderkreis, in Wissenschaft und Praxis, bessere Vorhersagen bei geringerem Rechenaufwand. Weiterhin zeigt diese Arbeit die theoretischen Implikationen der Modellierung von Messrauschen auf. Insgesamt ermöglichen diese Beiträge damit eine bessere Planung von untergrundbezogenen Anwendungen wie der Trinkwasserversorgung oder der Atommüll-Endlagerung.

1 Introduction

The importance of subsurface parameters

Drinking water supply, subsurface CO_2 storage, nuclear waste storage, saltwater intrusion, and water management in mining are only a few applications where subsurface characteristics are highly relevant for decision making. The challenges in these applications are twofold. On the one hand, having a good estimate for the spatial distribution of decision-relevant parameters (e.g., hydraulic conductivity) is crucial because it enables good predictions of the expected system behavior. On the other hand, knowing the uncertainty of subsurface parameters is equally important in many applications. Picture the planning of a long-term nuclear waste storage. Here, it is crucial to give probabilities of failure, which are only accessible with uncertainty quantification. Hence, we, as a society, need characterizations and uncertainty quantification of subsurface parameters.

This thesis aims to enhance existing methods to infer and quantify the uncertainty of key characteristics of aquifers, i.e., the spatial distribution of subsurface parameters. But why is this difficult? Unlike in many other disciplines of science, direct measurements of subsurface parameters are difficult, expensive, or even impossible to get due to different reasons. First, most measurements require a borehole (or a sample from a borehole), which is expensive to drill. Second, each borehole changes the characteristics of the subsurface due to the drilling itself. As a result, one will never (with classical methods) be able to measure the whole spatial distribution of parameters without altering them. Third, measurements can not be taken at arbitrary positions. Legal constraints, e.g., private properties or nature reserves, and physical constraints, e.g., mountain ranges or water bodies, reduce the number of permitted and feasible drilling locations significantly. As a result, often only a few sparse measurements are available.

These measurements can be used to predict the spatial distribution of parameters and their uncertainty. To generalize from point measurements to spatial predictions, prior assumptions on the structure of subsurface parameters need to be made. These a priori assumptions on the structure of parameters are called prior models in the remainder of this thesis.

Inference of spatial parameter fields

Statistical prior models incorporate the structural information of any spatially heterogeneous system before taking measurements. A large variety of prior models exist in various scientific applications. In subsurface modeling, multi-Gaussian random fields [e.g. Kitanidis, 1997], multiple-point geostatistics using training images [Strebelle, 2002, Mariethoz et al., 2010], level set methods [Iglesias and McLaughlin, 2011, Iglesias et al., 2014, 2016] and object-based methods [Jussel et al., 1994, Bennett et al., 2019] are some prominent examples. For an overview of subsurface modeling approaches, I refer to Koltermann and Gorelick [1996]. Here, each prior model incorporates different specific assumptions as prior knowledge.

After taking measurements, these prior models are calibrated to data. This calibration with the goal of predicting the parameter distribution is called inversion or inference. Bayesian statistics (i.e., Bayes theorem) provides a rigorous framework to combine prior knowledge with measurement information. This process is called *Bayesian inversion* or *Bayesian inference*. This process yields a so-called posterior parameter distribution, which consists of predictions and uncertainties of the parameter fields.

Calculating or estimating the posterior distribution can be challenging. For direct measurements and multi-Gaussian prior and likelihood, the Kriging procedure (or Gaussian process regression) can be used to analytically calculate the posterior distribution [e.g. Kitanidis, 1997]. Most geostatistical inverse problems use indirect measurements because parameter fields occur as factors in the governing differential equations [e.g. Kitanidis, 1997]. As a result, Kriging is often not applicable and numerical sampling methods or linearization-based methods are used to approximate the posterior distribution. Ensemble Kalman filters [e.g. Evensen, 2009] are sampling methods that linearize forward models (relating parameters to observable data), which enables them to exploit analytical approximations for multi-Gaussian priors and likelihoods. Purely linearization-based approaches like the Quasi-Linear Geostatistical Approach (QLGA) [Kitanidis, 1995] and the Successive Linear Estimator (SLE) [Yeh et al., 1996] also only converge towards approximations of the true posterior. For convergence towards the exact posterior, rejection sampling [e.g. Gelman et al., 1995, chapter 10.2] is efficient, but only for low-dimensional prior distributions or very weakly informative data. For high-dimensional distributions, iterative numerical estimators such as the smooth bootstrap filter [Smith and Gelfand, 1992] or Markov chain Monte Carlo (MCMC) methods are applicable but computationally expensive. In this thesis, I focus on MCMC methods because they are universally applicable for Bayesian inversion with non-linear measurements [e.g. Qian et al., 2003].

The computational effort of most common MCMC methods, e.g., the Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970], increases drastically with higher dimensions. In the context of Bayesian inversion, high-dimensional distributions are the result of discretization refinements. Hence, acquiring posterior parameter distributions with a high spatial resolution is unfeasible with most MCMC methods.

Hamiltonian MCMC methods [e.g. Betancourt, 2018] can sample high-dimensional posterior distributions but use derivatives for fast convergence. The derivatives of the posterior are not analytically available in subsurface Bayesian inversion and a numerical approximation of them is numerically expensive. As a result, Hamiltonian MCMCs are impracticable for posterior estimation. Other popular MCMC-based methods rely on concepts such as spectral parameterization [Laloy et al., 2015], pilot point methods [e.g. Jardani et al., 2013] and Karhunen-Loeve expansions [e.g. Mondal et al., 2014]. These are dimension reduction approaches to reduce the computational effort at the cost of only converging to an approximate solution of the true posterior. This thesis focuses on exact approximations of the posterior only, and hence does not apply these approaches.

On the contrary, specialized MCMC methods exist that can sample high-resolution spatial posterior parameter distributions. Existing methods include the pre-conditioned Crank Nicolson MCMC (pCN-MCMC) [Beskos et al., 2008, Cotter et al., 2013] and the Gibbs approach [e.g. Gelman et al., 1995, chapter 11.3]. The latter achieves a discretization-independent efficiency by iteratively modifying different subsets of parameters [Fu and Gómez-Hernández, 2008, 2009a,b, Hansen et al., 2012]. This modification respects the conditional prior distribution given by the prior covariance structure and the surrounding fixed values of parameters. Both approaches are well suited for weakly informative data and have dimension-independent convergence rates for Gaussian prior fields. This reduces the computational burden for highly discretized fields. However, they are not appealing to many researchers due to their still high computational costs.

This thesis tackles this challenge in the first four contributions by developing new efficient MCMC methods for accurate, high-dimensional Bayesian inversion with nonlinear measurements. Here, the focus lies on MCMC methods for two prior types: (1) For Gaussian priors (contributions one, two, and three) due to their popularity in geoscience and (2) for hierarchical priors (contribution four) because of their good representation of nature.

Model selection

Calibrating a prior model to data yields posterior predictions of the spatial parameter distribution. The remaining challenge is to choose a good prior model. In this thesis, the terms *prior model* and *model* will be used interchangeably for shorter notation. In many practical applications, various competing models exist and modelers need to decide which one to use. Bayesian statistics and Bayes theorem offer a rigorous framework called *Bayesian model selection* (BMS) [e.g. Wasserman et al., 2000] to choose between competing models.

In BMS, *posterior model weights* are calculated that predict the probabilities of models given data [e.g. Höge, 2019]. This calculation has two downsides. First, it is computationally expensive because it requires to estimate the so-called *Bayesian model evidence* (BME) [Schöniger et al., 2014], which is a high-dimensional integral. Hence, the BME is approximated using Monte Carlo sampling [Schöniger et al., 2014, 2015] or MCMC methods using thermodynamic integration [Lartillot and Philippe, 2006], which are both computationally expensive. Second, the posterior model weights can be indecisive by resulting in similar weights for all models. For these scenarios, the so-called *model justifiability analysis* [Schöniger et al., 2015] reveals whether weakly informative data or similarity between the candidate models is the reason for this occurrence. Based on that information, decision-makers can either decide to take more informative measurements (to make the model selection more decisive) or to use any one of the proposed models due to their high similarity.

One big remaining challenge in BMS is the presence of measurement noise. All measurements in real-world experiments are noisy. So far, the BMS literature misses a detailed study of where measurement noise should be modeled in BMS. The fifth contribution of this thesis aims to close this gap. It shows that the way where measurement noise is modeled can change the results of BMS drastically and gives guidance on where to model noise correctly.

Objectives and structure of thesis

The overarching goal of this thesis is to improve the prediction and uncertainty quantification of subsurface parameter fields. To reach this goal, I enhance MCMC methods in Bayesian inversion and give guidance on how to handle noisy data in BMS. A summary of common prior models and the state of the art in Bayesian inversion, BMS, and MCMC methods are presented in Chapter 2. In Chapter 3, the objectives and contributions of this thesis are summarized. The corresponding publications are presented in Appendix A-E. Chapter 4 draws conclusions and gives an outlook on potential future research.

2 State of the art

This chapter gives an overview of state of the art methods in subsurface modeling. The statistical Bayesian framework is presented in Chapter 2.1. In Chapter 2.2, the prior probability fields used within this thesis are summarized. Finally, Chapter 2.3 introduces state of the art MCMC methods for sampling the posterior in Bayesian inversion.

2.1 Bayesian framework

There are two perspectives to think of probabilities: The frequentist's and the Bayesianist's perspective. For frequentists, a probability is the frequency of occurrence of a particular event. As a result, only probabilities of repeating processes, e.g., the result of a dice roll, can be defined.

On the contrary, Bayesianists see probabilities as a representation of the state of knowledge. As a result, Bayesian statistics can define probabilities of processes that only occur once or of the state of knowledge of some unknown parameter. Describing the state of knowledge of subsurface parameters using probability distributions is the goal of this thesis.

To formulate probabilities for singular events or states (such as parameter values in aquifers), Bayesian statistics takes the following approach: First, a prior distribution is selected to define the prior belief of a parameter $\boldsymbol{\theta}$. Maximum entropy priors (Maximum entropy probability distributions) [Jaynes, 1957] can be used to reflect the state of knowledge for one-dimensional distributions correctly. In most geoscience applications, one is not interested in a single parameter at some position but rather in the spatial distribution of parameters. Hence, specific priors that account for the correlation between spatially neighboring parameters are used. Chapter 2.2 summarizes popular prior models for spatially distributed parameters. Dependent on the chosen prior model M_m, the prior distribution $p(\boldsymbol{\theta}_m|M_m)$ of parameters varies. Here, the subscript m denotes the m-th prior model in a potential set of models.

Next, this prior belief is updated using data $\mathbf{d}$ as shown in Chapter 2.1.1. This update leads to the final prediction of parameters which is called *posterior distribution* or just *posterior* $p(\boldsymbol{\theta}_m|\mathbf{d}, M_m)$. In reality, often several prior models are available. Hence, the best prior model (here: prior probability distribution $p(\boldsymbol{\theta}_m|M_m)$) needs to be chosen from the ensemble of possible models. Bayesian model selection (BMS), as presented in Chapter 2.1.2, shows how to do this consistently. After that, Chapter 2.1.3 introduces the Model justifiability analysis, which helps to interpret the obtained BMS results.

Throughout this thesis, I solely use the (probabilistic) Bayesian framework which does not give deterministic predictions. Instead, probabilities of models $p(M_m|\mathbf{d})$ (discrete) and probability distributions of parameters $p(\boldsymbol{\theta}_m|\mathbf{d}, M_m)$ (mostly continuous) given data $\mathbf{d}$ are calculated as shown in the following.

2.1.1 Bayesian inversion

The goal of Bayesian inversion (also called Bayesian inference) is to update the prior knowledge with data $\mathbf{d}$ using the Bayes' theorem

$$p(\boldsymbol{\theta}_m|\mathbf{d}, M_m) = \frac{p(\mathbf{d}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)}{p(\mathbf{d}|M_m)} \propto p(\mathbf{d}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m) \tag{2.1}$$

to get the *posterior* $p(\boldsymbol{\theta}_m|\mathbf{d}, M_m)$. This *posterior distribution* is the updated belief on the parameter $\boldsymbol{\theta}_m$ given the prior knowledge $p(\boldsymbol{\theta}_m|M_m)$, data $\mathbf{d}$ and model M_m.

The likelihood $p(\mathbf{d}|\boldsymbol{\theta}_m, M_m)$ defines how likely the data $\mathbf{d}$ is, given the parameter $\boldsymbol{\theta}_m$ and model M_m. A popular approach is to use a normal distributed likelihood function

$$p(\mathbf{d}|\boldsymbol{\theta}_m, M_m) = \mathrm{N}\Big(\big(\mathbf{y}(\boldsymbol{\theta}_m, M_m) - \mathbf{d}\big), \mathbf{R}\Big) \tag{2.2}$$

$$= (2\pi)^{\frac{N_d}{2}} |\mathbf{R}|^{\frac{1}{2}} \exp\left(\frac{1}{2}\Big(\mathbf{y}(\boldsymbol{\theta}_m, M_m) - \mathbf{d}\Big)^T \mathbf{R}^{-1}\Big(\mathbf{y}(\boldsymbol{\theta}_m, M_m) - \mathbf{d}\Big)\right) \tag{2.3}$$

where N_d is the number of data points and $\mathbf{y}(\boldsymbol{\theta}_m, M_m)$ are the deterministic predictions of the measurement values $\mathbf{d}$ given the parameters $\boldsymbol{\theta}_m$ and model M_m. $\mathbf{y}(\boldsymbol{\theta}_m, M_m) = \boldsymbol{\theta}_m$ means that the parameters are measured directly and is referred to as *direct measurements*. In this thesis, I focus on *indirect measurements* where $\mathbf{y}(\boldsymbol{\theta}_m, M_m)$ can be any function or simulation that connects the parameter $\boldsymbol{\theta}$ to the measured values $\mathbf{d}$. Here, I use the groundwater simulation framework MODFLOW [McDonald and Harbaugh, 1988, Harbaugh et al., 2000], to calculate predictions $\mathbf{y}(\boldsymbol{\theta}_m, M_m)$ of measured hydraulic head (pressure) values $h(x, y, t)$, given the uncertain isotropic hydraulic conductivity field $K(x, y)$ ($\boldsymbol{\theta} = K(x, y)$). To make this prediction, MODFLOW numerically solves the saturated groundwater flow equation

$$\nabla\left[K(x,y)\nabla h(x,y,t)\right] = \eta(x,y) + S_0\frac{\partial h(x,y,t)}{\partial t}, \tag{2.4}$$

where η encapsulates all source and sink terms and the specific storage S_0.

For many sampling approaches (e.g., MCMC methods), it is sufficient to evaluate $p(\mathbf{d}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)$ because it is proportional to $p(\boldsymbol{\theta}_m|\mathbf{d}, M_m)$. The Bayesian model evidence (BME) $p(\mathbf{d}|M_m)$ is a normalizing constant which ensures that the posterior $p(\boldsymbol{\theta}_m|\mathbf{d}, M_m)$ integrates to one, which can be accounted for implicitly by many methods. In some applications, however, the BME $p(\mathbf{d}|M_m)$ of model m is explicitly needed and defined as

$$p(\mathbf{d}|M_m) = \int p(\mathbf{d}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)\mathrm{d}\boldsymbol{\theta}_m\ . \tag{2.5}$$

The BME can be interpreted as the probability of the data given the model M_m.

2.1.2 Bayesian model selection

In reality, often several models are available and one has to choose between them. Bayesian model selection [e.g. Höge, 2019] is a rigorous framework to choose one model out of a set of models. Here, posterior model weights $p(M_m|\mathbf{d})$ define the probability of the model given the data. They are calculated using Bayes' theorem with discrete models instead of continuous parameters:

$$p(M_m|\mathbf{d}) = \frac{p(\mathbf{d}|M_m)p(M_m)}{\sum_i p(\mathbf{d}|M_i)p(M_i)} \ . \tag{2.6}$$

Here, $p(M_m)$ represents the prior probability of model m. Typically, a uniform prior distribution ($p(M_1) = p(M_2) = p(M_3)...$) between all models is assumed. This uniform prior is updated using the BME $p(\mathbf{d}|M_m)$ of each model. The resulting posterior model weights $p(M_m|\mathbf{d})$ are the probabilities of each model being the data-generating model and they sum up to one.

These posterior model weights can be used in several ways [e.g. Hoeting et al., 1999, Höge et al., 2019]. First, they are used in Bayesian model selection (BMS) to choose the best model. Here, the model with the highest posterior weight is selected and used for parameter inference as presented in Chapter 2.1.1. Second, posterior model weights are used to combine several models to one multi-model framework using Bayesian model averaging. Third, Bayesian model ranking ranks the models from highest to lowest posterior weight.

2.1.3 Model justifiability analysis

Independent of the purpose, the interpretation of posterior model weights is difficult because all results are always conditional on the considered set of models, which do not reflect the infinitely large space of possible models. As a result, the weights are always dependent on the very limited set of candidate models and do not reflect model probabilities on a universal, absolute scale. The *model justifiability analysis* [Schöniger et al., 2015] was introduced to help with the interpretation of posterior model weights. It shows whether a lack of confidence in model choice is due to weakly informative data or high similarity between models.

The *model justifiability analysis* samples all candidate models to produce many artificial data sets $\widetilde{\mathbf{d}}_{ij}$. Here, i denotes the i-th model while j denotes the j-th randomly generated artificial data sample. Then, the posterior model weights of all models given data $\widetilde{\mathbf{d}}_{ij}$ are calculated using Equation 2.6. Averaging the posterior weights of model k over the N_r random realizations of data of model i leads to the average posterior model weights

$$p(M_k|M_i) = \sum_{j=1}^{N_r} \frac{p(M_k|\widetilde{\mathbf{d}}_{ij})}{N_r} \ . \tag{2.7}$$

In principle, $p(M_k|M_i)$ can be seen as a similarity measure between models. Large $p(M_k|M_i)$ indicate high similarities of models, while low $p(M_k|M_i)$ indicate strong discrepancies between models i and k. Consequently, $p(M_k|M_k) \geq p(M_k|M_i)$ holds true for all i.

This analysis can be further used to compare the so-called self identification $p(M_k|M_k)$ with the posterior model weight $p(M_k|\mathbf{d})$. Assuming that model M_k is a perfect representation of nature, $p(M_k|M_k)$ would be the expected value of $p(M_k|\mathbf{d})$ because it calculates the average posterior model weight given data from model M_k. In turn, similar $p(M_k|M_k)$ and $p(M_k|\mathbf{d})$ indicate that model M_k represents nature reasonably well. On the contrary, distinct $p(M_k|M_k)$ and $p(M_k|\mathbf{d})$ are an indication that model M_k is not a good representation of nature. This holds true, even if $p(M_k|\mathbf{d})$ is the highest posterior model weight out of a set of models. In this case, none of the models represents nature well.

2.2 Prior geostatistical models

Several geostatistical prior models exist in the literature. I focus on two fundamentally different approaches. In Chapter 2.2.1, Gaussian random fields are presented. They define a covariance between any two spatial parameters. While they lack geological realism, they are mathematically convenient. In contrast, Chapter 2.2.2 gives an overview of multiple-point statistical approaches using training images that use more than two parameters to characterize the spatial parameter structure. They are computationally cumbersome to handle, yet offer more geological realism than Gaussian random fields.

2.2.1 Two-point statistics: Gaussian random fields

Random fields describe random functions over arbitrary domains. In geostatistics, Gaussian random fields are commonly used to describe the prior distribution of parameters. In general, Gaussian random fields are defined on continuous domains (e.g. $\mathbb{R}^2$). To handle them numerically, they are discretized on a grid. This results in a multivariate distribution. Each discretization point is presented by one random variable which is correlated to other random variables. Gaussian random fields have the property that the resulting multivariate distribution is multivariate Gaussian. Hence, a discretization of Gaussian random fields prior $p(\boldsymbol{\theta}_G|M_G)$ can be written as

$$p(\boldsymbol{\theta}_G|M_G) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.8}$$

$$= (2\pi)^{\frac{N_v}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \exp\left(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right). \tag{2.9}$$

In this formula, M_G explicitly denotes that a Gaussian prior is used and N_v represents the number of discretization points. Let $\boldsymbol{\theta}_G^i$ be the i-th entry of the vector $\boldsymbol{\theta}_G$. Then, each $\boldsymbol{\theta}_G^i$ corresponds to the probability distribution of the Gaussian random field at some position in space. Assuming a two-dimensional domain, each variable $\boldsymbol{\theta}_G^i(\boldsymbol{x})$ corresponds to one position $\boldsymbol{x} = (x^1, x^2)$.

To fully describe a Gaussian random field, only the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ need to be defined. Although it is commonly not written explicitly, each entry of $\boldsymbol{\mu}(\boldsymbol{x})$ corresponds to some position $\boldsymbol{x}$. Here $\boldsymbol{\mu}(\boldsymbol{x})$ is the expected value of $\boldsymbol{\theta}(\boldsymbol{x})$ at position $\boldsymbol{x}$. Often, a constant $\boldsymbol{\mu}(\boldsymbol{x})$ is used because it reflects the state of knowledge best. However, if more information about the expected values is available, any function $\boldsymbol{\mu}(\boldsymbol{x}) = f(\boldsymbol{x})$ can be used to describe them.

$\boldsymbol{\Sigma}(\boldsymbol{x_1}, \boldsymbol{x_2})$ defines the covariance between $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$. Any positive semi-definite symmetric square matrix can be used to define a Gaussian random field. In geostatistics and other fields of science, the so-called *stationary models* using variogram statistics [Kitanidis, 1997] are used commonly. Here, a constant known mean function $\boldsymbol{\mu}$ and a stationary covariance function is assumed. Latter means, that the value of covariance matrix entries $\boldsymbol{\Sigma}_{12}$ are solely dependent on the distance between points:

$$\boldsymbol{\Sigma}_{12}(\boldsymbol{x_1}, \boldsymbol{x_2}) = R(h) = R(|\boldsymbol{x}_1 - \boldsymbol{x}_2|) \ . \tag{2.10}$$

Different definitions of the covariance function $R(h)$ exist and lead to different random field structures. A summary of popular covariance functions can be found in Kitanidis [1997]. Note, that $\boldsymbol{\Sigma}(\boldsymbol{x_1}, \boldsymbol{x_2})$ only defines the point-wise covariance between any two points. Therefore, this approach is called a two-point statistical approach. Taken together with the stationarity assumption represented by Equation 2.10, the covariance (correlation) of points are only dependent on pairs of parameters and their distance to each other. In the following, the exponential and Gaussian covariance functions are presented.

Exponential covariance

The exponential covariance function is given by

$$R(h) = \sigma^2 \left(1 - exp\left(\frac{h}{\ell}\right)\right) \tag{2.11}$$

where σ^2 is the variance of the field and ℓ is a lengthscale parameter. Figure 2.1 (top row) visualizes Equation 2.11 and shows two samples from Gaussian random fields using an exponential variogram with different lengthscales. One can see that small ℓ lead to low covariances of neighboring parameters and hence a quick succession of high and low values of the parameter. A large lengthscale leads to higher covariance and larger areas of similar values.

Gaussian covariance

The Gaussian or squared exponential covariance function is given by

$$R(h) = \sigma^2 \left(1 - exp\left(\frac{h^2}{\ell^2}\right)\right) \tag{2.12}$$

with σ^2 being the variance of the field and the lengthscale parameter ℓ. Figure 2.1 (bottom row) visualizes 2.12 and shows two samples of the Gaussian random fields

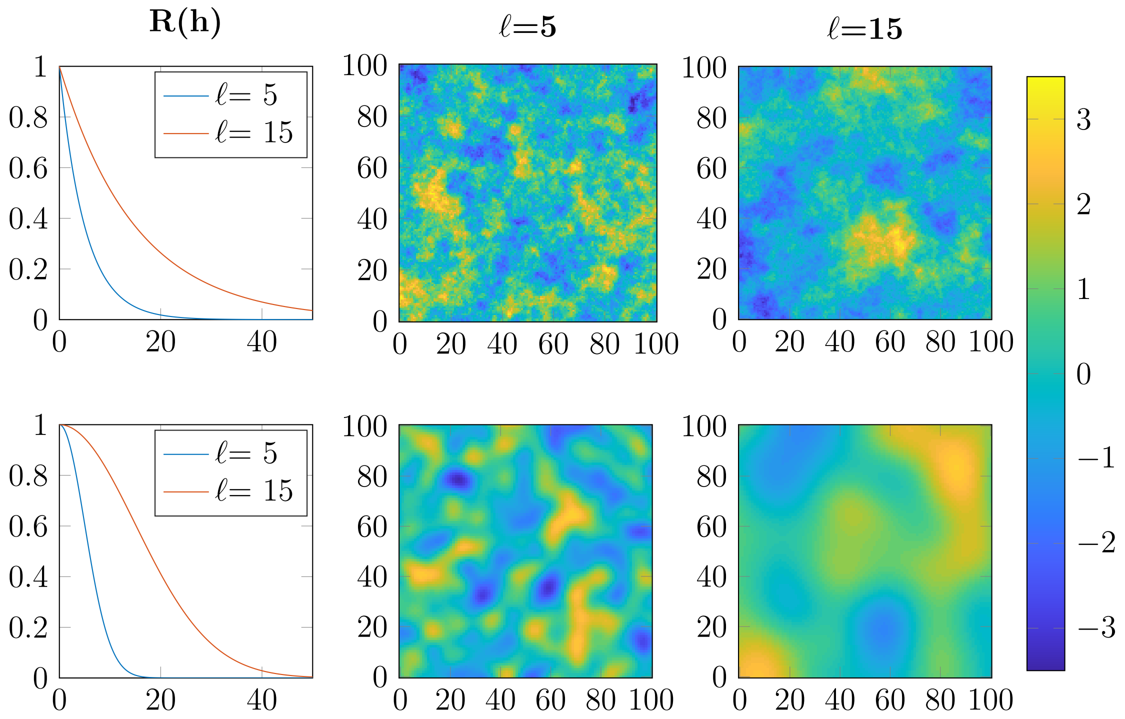


Figure 2.1: Visualization of covariance function $R(h)$ of Equation 2.11 and 2.12 (left) and samples from the two-dimensional random field with the respective covariance structure (center, right). The differences in field structure of exponential model (top row) and Gaussian model (bottom row) as well as the differences with different lengthscale parameter ℓ can be observed. All samples are created with $\boldsymbol{\mu} = 0$ and $\sigma = 1$.

with different lengthscales. Again, small values of ℓ lead to small areas of similar values whereas large values of ℓ lead to highly correlated fields consisting of large areas with similar values.

The Gaussian covariance function and the exponential covariance function lead to different structures of the resulting samples. The Gaussian covariance field (Figure 2.1 bottom) is smooth, whereas the exponential covariance field (Figure 2.1 bottom) is jagged and not differentiable. Both are possible priors and encapsulate different prior knowledge.

The lengthscale ℓ, the variance σ^2, and the mean $\boldsymbol{\mu}$ have a big influence on the field structure in both cases. They can be either chosen as distinct values that are assumed to be known or they can be treated as additional unknown so-called *hyperparameters*. Consequently, a prior distribution of ℓ, σ^2 and $\boldsymbol{\mu}$ can be defined. Then, the goal of parameter inference is to jointly infer the parameters of the field and the *hyperparameters*.

2.2.2 Multiple-point statistics: utilizing training images

In many subsurface applications, the geological object and structures of interest have a complex geometry, e.g., sand channels from ancient river beds, that can not be modeled with the (two-point) variogram statistics presented in the previous chapter. Instead, multiple-point statistics (MPS) are used. In MPS methods, not only the information between any two points is used to define a random field. Instead, the whole geometry of a so so-called training image (TI) is used. A TI (see Figure 2.2, left) is an image of the expected parameter field structure. Different MPS algorithms were proposed in the literature [Strebelle, 2002, Mariethoz et al., 2010] that can use the spatial information of the TI and randomly generate samples with similar structure. The SNESIM algorithm [Strebelle, 2002] is one popular example and samples from that algorithm are shown in Figure 2.2 (center and right).

The prior $p(\boldsymbol{\theta}_{MPS}|M_{MPS})$ of MPS algorithms is not defined explicitly as compared to the explicit definition of Gaussian fields in Equation 2.8. Instead, each MPS algorithm (e.g. SNESIM) combined with a TI (e.g. Figure 2.2, left) implicitly defines

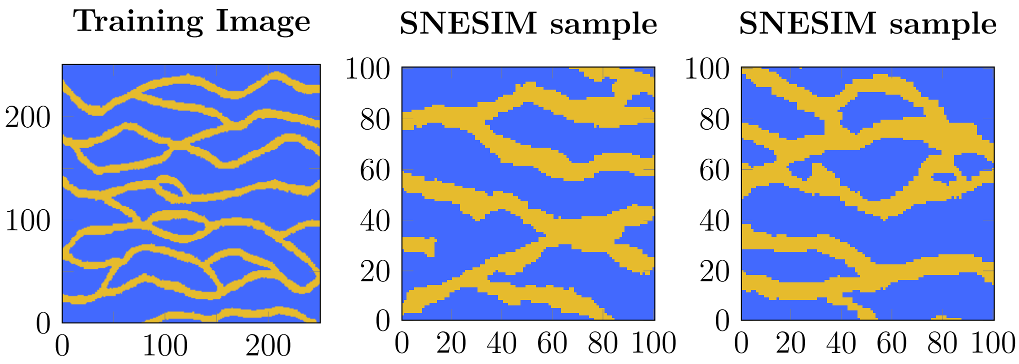


Figure 2.2: Visualization of training-image (left) and samples form the two-dimensional training-image-based random fields (center right). Both samples were created using the SNESIM algorithm [Strebelle, 2002]. In this example, two facies (e.g. sand channels and shale aquifers) are shown as examples (represented with yellow and blue colors).

a prior $p(\boldsymbol{\theta}_{MPS}|M_{MPS})$ by the samples (e.g. Figure 2.2, center and right) it (randomly) produces. As a result, it is impossible to calculate the probability density $p(\boldsymbol{\theta}_{MPS}|M_{MPS})$ for a given parameter $\boldsymbol{\theta}_{MPS}$. Only sampling from the probability distribution $p(\boldsymbol{\theta}_{MPS}|M_{MPS})$ is possible. This sampling property is used in MCMC methods with asymmetric proposals.

Figure 2.2 shows a categorical TI with two categories. Alternatively, TIs with more categories or even continuous TIs can be used that will lead to samples with many categories or continuous samples. Independent of the type of TI, the main challenge of MPS approaches is to get a good TI for the problem at hand. In 2D, images of a slice of an aquifer can be used as TI. However, acquiring a slice of an aquifer is often tricky and expensive. Instead, hand-drawn images of some *field expert* are often used as training image. In 3D, the task of getting a TI is even more challenging. While 3D tomography can be used for 3D binary TI, this is really expensive already for small domains and almost impossible for larger domains [e.g. Li et al., 2019, Wu et al., 2020]. In a nutshell, MPS approaches with TI are a powerful tool to generate random samples that replicate complex structures. However, obtaining a good TI is a challenge.

2.3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are popular approaches to sample from any probability distribution π. Sampling from a distribution means to produce samples randomly drawn from π, i.e., with a frequency proportional to the corresponding probability density. As a result, the histogram over the MCMC samples converges towards π as shown in the center of the top row in Figure 2.3. The bottom row of Figure 2.3 shows three independent visualizations of a plain vanilla random walk MCMC. All MCMC methods have the same structure as defined in Algorithm 1.

Algorithm 1: MCMC

```
Input  : Probability distribution $\pi(\boldsymbol{\theta})$ that can be evaluated for any
         parameter $\boldsymbol{\theta}$
Output: Samples $\boldsymbol{\theta}_i$ from distribution $\pi(\boldsymbol{\theta})$
Set $i = 1$
Draw $\boldsymbol{\theta}_0$
while $i \leq N_{\max}$ do
    Propose $\boldsymbol{\theta}_j = g(\boldsymbol{\theta}_i)$, given by e.g. Equation 2.20.
    Compute $\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$, given by e.g. Equation 2.22.
    Draw $r \sim U(0,1)$
    if $r \leq \alpha$ (with probability $\alpha$) then
        $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_j$
    else
        $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$
    end
    $i = i + 1$
end
```

First, all MCMC methods start with a random sample (e.g. $\boldsymbol{\theta}_0 = 0$ in Figure 2.3). Second, a new sample $\boldsymbol{\theta}_j$ is proposed given the current sample $\boldsymbol{\theta}_i$. Third, the new sample $\boldsymbol{\theta}_j$ gets accepted, i.e. becomes the new current samples $\boldsymbol{\theta}_{i+1}$, with probability $\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$. With probability $1 - \alpha$, $\boldsymbol{\theta}_j$ is rejected, and the old sample $\boldsymbol{\theta}_i$ becomes the next current sample $\boldsymbol{\theta}_{i+1}$. Then, steps two and three are repeated until the predefined number of MCMC samples (e.g. $N_{\max} = 2000$ in Figure 2.3) is reached. Using this procedure, MCMC methods can sample from any $\pi(\boldsymbol{\theta})$ without making any assumptions about its form. The only differences between different MCMC methods are their proposal functions and their expressions to obtain the acceptance probabilities.

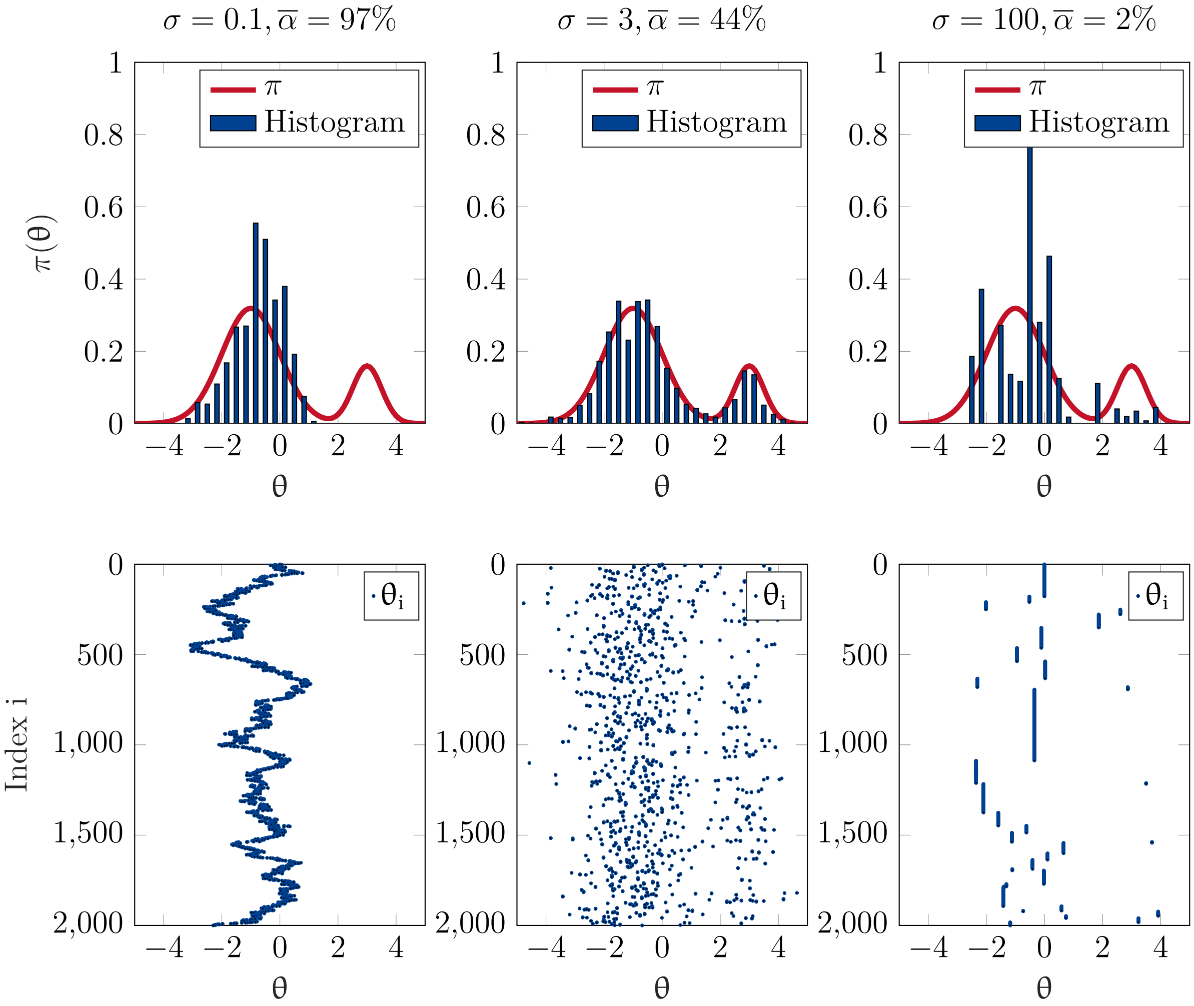


Figure 2.3: Visualization of random walk Metropolis-Hastings MCMCs with too small proposals (left), good proposals (center), and too large proposals (right). The bottom row shows the value of each MCMC sample (x-axis) over the 2000 iterations of the chain (y-axis). The top row shows the target distribution π and the (scaled) histograms of MCMC samples (shown in the bottom row).

As the basis for my contributions, I name all properties that an MCMC method needs to fulfill to converge to the exact distribution π and then derive the formulas for different methods. Afterwards, I focus on the random walk Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970] with symmetric proposals in Chapter 2.3.1 and on asymmetric proposal MCMCs in Chapter 2.3.2. Chapter 2.3.3 concludes with a short summary of parallel tempering for MCMC methods.

MCMC methods converge to an arbitrary distribution π (at the limit of infinite runtime) if and only if irreducibility, aperiodicity, and the detailed balance are fulfilled [Smith and Roberts, 1993]. Irreducibility is the condition that the MCMC can move from any point (with $\pi(\boldsymbol{\theta}) > 0$) to any other point (with $\pi(\boldsymbol{\theta}) > 0$) in the parameter space within a finite number of steps with a positive probability. Aperiodicity is the condition that the Markov chain can not get trapped in periodic loops. These conditions are almost always fulfilled for continuous parameter spaces and continuous proposal distributions with the same linear span. Hence, I focus on the detailed balance from now on. It is defined as

$$\pi(\boldsymbol{\theta}_i)h(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \pi(\boldsymbol{\theta}_j)h(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i) \tag{2.13}$$

with the transition kernel h, which is usually defined as

$$h(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \ . \tag{2.14}$$

Here, $q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ is called the proposal distribution and $\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ is the so-called acceptance probability. Inserting Equation 2.14 into Equation 2.13 leads to

$$\pi(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \pi(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)\alpha(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i). \tag{2.15}$$

This equation can be restructured to find

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \frac{\pi(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)}\alpha(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i). \tag{2.16}$$

The property of $0 \leq \alpha \leq 1$ can be added to this equation to find that

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = min\left[\frac{\pi(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)}, 1\right]. \tag{2.17}$$

For any distribution π and any proposal distribution q, Equation 2.17 provides an α such that the detailed balance is fulfilled. Further, only the fraction $\frac{\pi(\boldsymbol{\theta}_j)}{\pi(\boldsymbol{\theta}_i)}$ is used when calculating α. Hence, any multiple of π can be used for MCMC methods because multiplicative constants cancel out. In many applications, using multiples of π makes the numerical computation faster, specifically when π is given only by a proportionality as it is the case in most applications of Bayesian inference. Then, the normalizing constant in Equation 2.1 can be neglected.

MCMC in Bayesian Inversion

In this thesis, I want to use MCMCs to sample posterior distributions of Bayesian inverse problems as defined in Equation 2.1. Thus, the target distribution π is defined as

$$\pi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})L(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_m|M_m)p(\mathbf{d}|\boldsymbol{\theta}_m, M_m) \propto p(\boldsymbol{\theta}_m|\mathbf{d}, M_m) \tag{2.18}$$

Here, I define the prior $P(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_m|M_m)$ and likelihood $L(\boldsymbol{\theta}) = p(\mathbf{d}|\boldsymbol{\theta}_m, M_m)$ for shorter notation within this chapter. While $p(\boldsymbol{\theta}_m|\mathbf{d}, M_m)$ would be a properly normalized distribution, $\pi(\boldsymbol{\theta})$ is not, which is no problem for MCMC methods as explained above. Inserting Equation 2.18 into Equation 2.17 leads to

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = min \left[\frac{P(\boldsymbol{\theta}_j)L(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)}{P(\boldsymbol{\theta}_i)L(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)}, 1 \right]. \tag{2.19}$$

Assuming that irreducibility and aperiodicity are fulfilled, we can construct an MCMC with arbitrary proposal distributions q. This yields the question of how to choose q for fast convergence?

In general, different MCMC proposals are better suited for different problem classes. Suitable MCMC methods stand out by their ability to explore the parameter space fast [Gelman et al., 1996]. It is desirable to propose large changes in the parameter space and accept them with a high probability [Gelman et al., 1996]. In practice, these two conditions contradict each other. Proposing small changes in $\boldsymbol{\theta}$ usually results in similar $\pi(\boldsymbol{\theta}_j)$ and $\pi(\boldsymbol{\theta}_i)$. This results in α close to 1 and the fact that most proposals get accepted. In contrast, proposing large changes in $\boldsymbol{\theta}$ results in distinct $\pi(\boldsymbol{\theta}_j)$ and $\pi(\boldsymbol{\theta}_i)$. Hence, α will often assume very small values and only a few proposals get accepted. As a result, a trade-off between the magnitude of the proposed change and the acceptance rate needs to be found [Gelman et al., 1996].

Figure 2.3 visualizes a plain vanilla random walk Metropolis-Hastings (MH) algorithm that shows this trade-off. The goal of an MCMC method is to sample the posterior distribution. Visually speaking, this means that the histogram of the MCMC sample should resemble the target distribution π. The magnitude of change in the proposals is defined by a scale parameter σ (resembling the standard deviation of proposed jumps, see Chapter 2.3.1) in the MH algorithm. Small values of σ are synonymous with small changes and lead to high acceptance rates $\overline{\alpha}$. The acceptance rates $\overline{\alpha}$ denote the percentages of accepted proposals. Figure 2.3 shows, that small changes (left) lead to very correlated samples along the chain. As a result, the chain cannot explore the whole parameter space within the first 2000 samples. On the contrary, too large proposed jumps (right column of Figure 2.3) lead to a low $\overline{\alpha}$. Only very few proposals get accepted, which also prevents the chain from exploring the parameter space. Finding a good proposal as shown in the center of Figure 2.3 (which is neither too large nor too small) is not trivial.

This example only shows the MH algorithm and how modifications of σ change its behavior. Other MCMC approaches design smart proposals that make distant jumps with high acceptance rates. To realize this, additional information about the target distribution π is needed. Chapter 1 discussed several MCMC approaches that use knowledge of the derivative of π. In Bayesian inversion, derivatives are routinely not available. Instead, Equation 2.19 and the knowledge that $\pi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})L(\boldsymbol{\theta})$ can be used to create asymmetric proposal distributions as discussed in Chapter 2.3.2.

2.3.1 Symmetric proposal distributions

The simplest realization of an MCMC method is the Metropolis-Hasting (MH) [Metropolis et al., 1953, Hastings, 1970] random walk algorithm. It assumes a symmetric proposal function

$$g(\boldsymbol{\theta}_i) = \boldsymbol{\theta}_i + \epsilon,\ \epsilon \sim N(0, \sigma^2)\ . \tag{2.20}$$

This function $g(\boldsymbol{\theta}_i)$ fulfills

$$q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)\ . \tag{2.21}$$

because the normal distribution $N(0, \sigma)$, with mean $\mu = 0$ and standard deviation σ, is symmetric. Inserting this into Equation 2.19 yields that

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = min\left[\frac{P(\boldsymbol{\theta}_j)L(\boldsymbol{\theta}_j)}{P(\boldsymbol{\theta}_i)L(\boldsymbol{\theta}_i)}, 1\right] = min\left[\frac{\pi(\boldsymbol{\theta}_j)}{\pi(\boldsymbol{\theta}_i)}, 1\right], \tag{2.22}$$

i.e. the proposal distribution cancels out. As discussed earlier, Figure 2.3 visualizes the random walk MH algorithm with different σ. For an extensive explanation of the MH algorithm, I refer to Chib and Greenberg [1995].

The downside of this algorithm is, that the acceptance rate of the proposal function (Equation 2.22) depends on both the prior and the likelihood. This leads to a fast decrease of α for increasing σ, especially in high-dimensional problems [Roberts and Rosenthal, 2002]. This can be improved as shown in the following.

2.3.2 Asymmetric proposal distributions

The core idea of the algorithms presented in this work is to use the knowledge that $\pi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})L(\boldsymbol{\theta})$ to increase the efficiency of MCMC methods. In typical geoscience problems, the prior $P(\boldsymbol{\theta})$ is usually complex and high-dimensional. Here, high-dimensional does not refer to physical dimensionality but to a mathematical dimensionality, i.e., the number of uncertain parameters to be inferred. As a result, the acceptance rate α of the MH algorithm mostly depends on the prior fraction of subsequent samples.

Hence, it is reasonable to define the acceptance probability α independent of the prior $P(\boldsymbol{\theta})$ and to enforce the prior within the proposal density explicitly. Changing the proposal distribution to

$$q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \frac{P(\boldsymbol{\theta}_j)}{P(\boldsymbol{\theta}_i)} q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i), \tag{2.23}$$

achieves this behavior [Mosegaard and Tarantola, 1995]. Inserting Equation 2.23 into Equation 2.19 results in [e.g. Tarantola, 2005]

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = min \left[\frac{L(\boldsymbol{\theta}_j)}{L(\boldsymbol{\theta}_i)}, 1 \right] . \tag{2.24}$$

This idea was termed ”extended Metropolis sampling” by Hansen et al. [2012] and I named it ”sampling from the prior distribution” in my contributions. With this technique, new proposed samples are only rejected due to the likelihood ratio and not due to the prior. This makes ”sampling from the prior” MCMCs converge faster because they accept proposals with a higher probability.

In a nutshell, the presented asymmetric proposal distributions implicitly enforce the prior (Equation 2.23). Here, the asymmetry of the proposals cancels out, even with non-uniform prior distributions. As a result, the acceptance step does only depend on the likelihood fraction (Equation 2.24). In the special case of uniform prior distributions ($P(\boldsymbol{\theta}_j) = P(\boldsymbol{\theta}_i) \forall \boldsymbol{\theta}_i, \boldsymbol{\theta}_j$), the so-constructed asymmetric proposal distributions degenerate to symmetric proposal distributions. This can be easily verified by inserting $P(\boldsymbol{\theta}_j) = P(\boldsymbol{\theta}_i)$ into the presented equations. In the following, I present the pCN-MCMC and the Gibbs approach as two prominent existing asymmetric methods that sample from the prior.

pCN-MCMC

The preconditioned Crank-Nicolson algorithm (pCN-MCMC) fulfills Equation 2.23 for multi-Gaussian priors [Beskos et al., 2008, Cotter et al., 2013]. It proposes

$$g(\boldsymbol{\theta}_i) = \sqrt{(1-\beta^2)}\,(\boldsymbol{\theta}_i - \boldsymbol{\mu}) + \beta\boldsymbol{\xi} + \boldsymbol{\mu}, \quad \boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.25}$$

with $\boldsymbol{\Sigma}$ being the covariance matrix of the prior as defined in Equation 2.10. The proposed Equation in 2.25 fulfills Equation 2.23 for multi-Gaussian priors. Consequently, the resulting acceptance probability α is given by Equation 2.19 and only depends on the likelihood. Analogously to σ in the MH algorithm, the tuning-parameter β is used to specify the change between subsequent samples in the pCN-MCMC. The main difference is, that β is restricted to values between zero and one in the pCN approach. For $\beta = 1$, subsequent samples are independent of each other. The smaller β gets, the higher the similarity between subsequent samples up to the theoretical limit of identical samples for $\beta = 0$.

Gibbs sampling

The Gibbs approach fulfills the property in Equation 2.23 by conditionally resampling parts of the parameter space [Geman and Geman, 1984]. Assuming a parameter vector

$\boldsymbol{\theta}$ of size $N_p \times 1$ (N_p denotes the number of parameters) and some permutation matrix $\mathbf{M}$ (also called P_π in the literature), the parameter can be ordered into two parts

$$\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix} = \mathbf{M}\boldsymbol{\theta} \text{ with size } \begin{bmatrix} q \times 1 \\ (N_p - q) \times 1 \end{bmatrix}, \tag{2.26}$$

where $\boldsymbol{\theta}_1$ incorporates the parameters which will be resampled conditionally on the parameters $\boldsymbol{\theta}_2$. The number of resampled parameters is given by q and the remaining non-resampled parameters are named $\mathbf{r}$. The Gibbs approach defines a new proposal as

$$g(\boldsymbol{\theta}) = \mathbf{M}^T \begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\theta}_2 = \mathbf{r} \end{bmatrix}, \quad \boldsymbol{\xi} \sim p_P(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2 = \mathbf{r}) \ . \tag{2.27}$$

This proposal can be used in any application where conditional samples from the prior ($\boldsymbol{\xi} \sim p_P(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2 = \mathbf{r})$) can be drawn without the need to evaluate the prior $P(\boldsymbol{\theta})$ explicitly. This enables Gibbs-based approaches to also sample from prior distributions with unknown closed form such as MPS random field generators that use training images [e.g. Strebelle, 2002]. Several authors [Fu and Gómez-Hernández, 2008, 2009a,b, Hansen et al., 2012] have shown the effectiveness of the Gibbs approach to resample boxes in the parameter space with multi-Gaussian or MPS-induced priors. This sequential resampling of boxes is called *sequential Gibbs approach.*

2.3.3 Parallel tempering

In this thesis, I distinguish between weakly and highly informative data. The difference is, that $L(\boldsymbol{\theta})$ is narrow for highly informative data and broad for weakly informative data. As a result, the posterior of highly informative data mainly depends on the likelihood function and can become complex and multi-modal. In contrast, weakly informative data lead to prior-dominated posteriors, which are easier to sample with ”prior sampling MCMCs”.

Sampling from the posterior with highly informative data is difficult for MCMC methods for two reasons. First, because they suffer from long burn-in times (the period in which the MCMC converges towards the final range of values). Second, because the MCMC chain can get stuck in one mode of the distribution (i.e. in one local optimum). Parallel tempering can solve both these problems [Earl and Deem, 2005, Geyer and Thompson, 1995].

Laloy et al. [2016] applied parallel tempering to categorical high-dimensional geostatistical MCMCs and showed that it increases ”convergence towards appropriate data misfit and [the] sampling diversity”. Further, they confirmed that it reduces the risk of getting stuck in one local optimum of the posterior.

The idea of parallel tempering [e.g. Earl and Deem, 2005] is to run N_{pt} chains with different temperatures $T = [T_1, ...T_{N_{pt}}]$ with $1 = T_1 < T_2 < ...T_{N_{pt}}$. The so-called tempered posterior density at temperature T is defined by

$$p(\boldsymbol{\theta}, T|\mathbf{d}) \propto P(\boldsymbol{\theta})L(\boldsymbol{\theta})^{\frac{1}{T}}. \tag{2.28}$$

Large temperatures T flatten the posterior towards the prior and thereby make the problem simpler. For $T \to \infty$, the tempered distribution $p(\boldsymbol{\theta}, T|\mathbf{d})$ converges towards the prior distribution $p(\boldsymbol{\theta})$. The other limit ($T = 1$) leads to $p(\boldsymbol{\theta}, T|\mathbf{d})$ being equal to the actual posterior distribution $p(\boldsymbol{\theta}|\mathbf{d})$. As a result, only the $T = 1$ chain can be used for posterior sampling. All remaining chains are used to help the first (productive) chain explore the posterior.

To realize this, we take the likelihood fraction given in Equation 2.24 to the power of $\frac{1}{T}$, which brings the acceptance probability of tempered chains

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = min\left[\left(\frac{L(\boldsymbol{\theta}_j)}{L(\boldsymbol{\theta}_i)}\right)^{\frac{1}{T}}, 1\right] \tag{2.29}$$

closer to one. This enables hot chains to make farther jumps than colder chains with similar acceptance rates because they run on a simplified problem. Consequently, they explore the posterior distribution faster.

To communicate ”good” regions from one chain to another, between-chain swaps are proposed. Therefore, after a few in-chain MCMC steps (e.g. Equation 2.25), a between-chain swap is proposed. In this swap, the parameters $\boldsymbol{\theta}_i$ of chain i with temperature T_i are swapped with parameters $\boldsymbol{\theta}_j$ of chain j with temperature T_j with probability

$$\alpha_s(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = min \left[\frac{L(\boldsymbol{\theta}_j)}{L(\boldsymbol{\theta}_i)}^{\left(\frac{1}{T_i} - \frac{1}{T_j}\right)}, 1 \right] \ . \tag{2.30}$$

Parallel tempering can be computationally expensive, because several chains are run, only to support the sampling from the first chain. This computational effort is not as significant as it might seem at first glance because parallel tempering MCMC can be easily parallelized. Each chain can run independently on a different worker and only between-chain jumps need to be communicated. Hence, the total runtime only increases dramatically if the number of chains exceeds the number of workers (cores) of the used computing system.

3 Objectives & Contributions

The overarching goal of this thesis is to improve the prediction and uncertainty quantification of subsurface parameter fields. To reach this goal, I enhance MCMC methods in Bayesian inversion and give guidance on how to handle noisy data in BMS. A fully Bayesian approach is chosen in this thesis because it can quantify the uncertainties in model parameters, model structures, and data accurately as discussed in Chapter 2.1. My goal is to find and approximate (i.e., to sample) the full posterior parameter probability distribution.

A two-step approach is taken to reach this goal. In contributions 1-4 of this thesis, I assume that the prior model structure is known and I focus on the uncertainties in spatially distributed model parameters. MCMC methods are chosen to sample from the posterior parameter distribution because they can perform Bayesian inversion without making assumptions. However, MCMC methods are time-intensive. Therefore, speeding up MCMC methods is important to make them more attractive to the community. I focus on building new and faster MCMC methods for different types of prior parameter field models.

Furthermore, in contribution five of my thesis, the model structure is assumed to be unknown and BMS is used to select the best prior model. To do so, measurement data is needed. However, the available data in subsurface modeling has a lot of measurement noise. I show that BMS results change, depending on where noise is considered in BMS and give guidance on how to model noise correctly.

As discussed in Chapter 2.3.3, I distinguish between *weakly* and *highly* informative data. Weakly informative data refers to data from few measurements that result in a broad likelihood function. As a result, the resulting posterior mainly depends on the prior. On the contrary, highly informative data refers to the case where the likelihood function dominates the resulting posterior.

To improve the prediction and uncertainty quantification of subsurface parameter fields, the following research questions are addressed in this thesis:

1. How can the efficiency of MCMC methods in Bayesian inversion be increased for multi-Gaussian priors and *weakly* informative data?
2. How can the efficiency of MCMC methods in Bayesian inversion be increased for multi-Gaussian priors and *highly* informative data?
3. How can hyperparameters of the parameter field be included in the inversion with MCMC methods?
4. How can we model nature more realistically and still find an appropriate MCMC method?
5. How should we handle noisy data in BMS?

The following five contributions address these research questions one by one.

Contribution 1: Speeding up Bayesian inversion for multi-Gaussian priors with *weakly* informative data (sequential pCN-MCMC)

Gaussian random fields are commonly used as a prior for subsurface parameters. Sequential Gibbs sampling and the pCN-MCMC are both popular methods that sample from the prior, and hence, have shown great performance in Bayesian inversion. Both methods are applicable to Gaussian random field priors and work well for weakly informative data. For highly informative data in non-linear inversion problems, both methods run the risk of getting stuck in one local mode of the posterior as discussed in the next contribution. In Reuschen et al. [2021b], I focus on weakly informative data and combine the ideas of *sequential Gibbs* and the *pCN-MCMC* to create a more efficient method with faster convergence. I named the combined method the *sequential pCN-MCMC*. This method uses two tuning parameters β and κ. Optimizing these tuning parameters is difficult, and classical approaches, e.g., tuning the acceptance rate, do not work. Instead, I develop a novel approach that optimizes the tuning parameters during the runtime of the algorithm. Finally, the *sequential pCN-MCMC* is tested on several test cases with weakly informative measurements.

Contribution 2: Speeding up Bayesian inversion for multi-Gaussian priors with *highly* informative data (pCN-PT)

Using *highly* informative data (many measurements) in non-linear Bayesian inversion can lead to complex (e.g. multi-modal) posterior distributions. One challenge of common MCMC methods is that they get stuck in local optima or are not able to get to high-likelihood regions of such complex posteriors. Previous work showed that parallel tempering can solve this issue. However, it has not yet been combined with the highly efficient pCN-MCMC. In Xu et al. [2020], we used parallel tempering with the pCN-MCMC to sample efficiently from multi-Gaussian priors with *highly* informative data. This combined method was named *pCN-PT* and was tested on different types of Bayesian inverse problem. With this, we show that we gain a higher efficiency compared to state-of-the-art alternatives.

Contribution 3: Enable joint Bayesian inversion of multi-Gaussian fields and hyperparameters (extended pCN-PT)

All previous contributions assume that the hyperparameters of the Gaussian random fields (e.g. σ and ℓ in Equation 2.12) are known. This is seldom the case in real-world applications. Instead, the hyperparameters can be treated as unknowns with respective prior distributions. Logically, the next challenge arises: Can we infer the hyperparameters together with the subsurface parameters? This poses a broader, harder, and more realistic problem. In Xiao et al. [2021], we show how the pCN-PT can be extended to infer the hyperparameters and the parameters jointly. To reach this goal, the original parameters are decomposed into hyperparameters (mean, standard derivation, and correlation lengths) and white noise (a standard normal random vector). This makes the white noise independent from the hyperparameters, which allows running the pCN-PT on white noise and hyperparameters independently. Finally, the original parameters are recovered by recombining white noise with hyperparameters. We test the extended pCN-PT in different scenarios, e.g., with direct and indirect measurements, and report the varying results of different priors and types of data.

Contribution 4: Enable Bayesian inversion for hierarchical geostatistical models (parallel-tempering sequential Gibbs MCMC)

Gaussian random fields are often not a good prior representation of complex subsurface structures. Instead, *hierarchical* geostatistical models can be used as prior to reflect the complex structure of nature. In this contribution, I use a hierarchical joint model that accounts for two (and possibly more) categories using an MPS tool and heterogeneities inside each category using Gaussian random fields. Bayesian inversion of such combined models is challenging and has only been performed using approximate methods such as Ensemble Kalman filters. These methods do not converge towards the true posterior distribution. In Reuschen et al. [2020], I present the first efficient MCMC method that can perform Bayesian inversion of these combined models and converges towards the true posterior. Here, the categorical fields (which facies is present) and the heterogeneity fields within each facies are inferred jointly. This enables realistic inversion of, e.g., channelized flow in the subsurface. To show the performance of the presented *parallel-tempering sequential Gibbs MCMC*, I test it in a synthetic channelized flow scenario with *weakly* and *highly* informative data.

Contribution 5: How to handle measurement noise in BMS (Model confusion analysis)

All previous contributions showed how to calibrate parameters for a given prior (geo-) statistical model. In the next step, I choose between competing models to select the best one using BMS. In Reuschen et al. [2021a], I discuss where and for which reasons measurement noise should be treated in BMS. I distinguish between four cases which represent four ways to model measurement noise that differ philosophically and mathematically. Only two of them are logically consistent and answer two different research questions: (1) ”Which model is best in modeling the pure physics?” and (2) ”Which model is best in predicting the data-generating process (i.e., physics plus noise)?”. Using synthetic scenarios and real-world test cases, I show that the choice of research question significantly impacts BMS results. Hence, the decision of where to model measurement noise can not be neglected. Furthermore, I show that the practical implementation of BMS to answer question (1) is challenging or even impossible. However, exploiting the other three cases can help to approximate the BMS results of question (1).

4 Conclusions & Outlook

This thesis advances statistical methods to predict subsurface parameters in groundwater modeling. To reach this goal, contributions 1-4 enhance existing and develop new MCMC methods for a more efficient Bayesian geostatistical inversion. Going one step further, contribution five focuses on BMS to select the best prior model. The main findings of this thesis are summarized in the following.

Contribution 1: Combining the pCN-MCMC with sequential Gibbs sampling

First, I focused on Bayesian inversion of multi-Gaussian priors with *weakly* informative data. I combined the ideas of the two state-of-the-art methods *pCN-MCMC* and *sequential Gibbs sampling* and named the combination *sequential pCN-MCMC*. Here, the *pCN-MCMC* makes global proposals and slightly changes all parameters in one proposal step. On the contrary, the *sequential Gibbs sampling* makes local proposals, which means, that it makes large changes in a small spatial domain. The presented *sequential pCN-MCMC* is designed to make semi-global proposals with medium changes.

All three methods were tested on several test cases. Interestingly, I observed that the methods show different behavior based on the type of measurements used for the inversion. For local measurements, such as head or direct measurements that convey local information around the measurement position, the *sequential Gibbs sampling* with its local proposals showed a good performance. On the contrary, the *pCN-MCMC*, with global proposals, showed good performance for global, e.g., transport-related measurements. Hence, local proposals show good performance for local measurements and global proposals are suitable for global measurements.

The presented *sequential pCN-MCMC* always chooses the best trade-off between the two. This is done by optimizing its two tuning parameters during the run time of the algorithm. This leads to a speedup of $1-5.5$ over the *pCN-MCMC* and $1-6.5$ over *sequential Gibbs sampling*. Hence, this contribution reduced the computational cost of Bayesian inversion with Gaussian priors and few measurements by up to 85% dependent on the test case. These reduced computational costs enable more researchers and decision-makers to use the presented MCMC instead of using (fast) approximate methods as, e.g., ensemble Kalman filters, which make approximation errors by design. This leads to better prediction and uncertainty quantification with *weakly* informative data.

Contribution 2: The benefit of parallel tempering in combination with the pCN-MCMC

In contrast to the previous contribution, my second contribution focuses on handling multi-Gaussian priors with *highly* informative data. *Highly* informative data makes the posterior more dependent on the likelihood and less on the prior. Consequently, the *pCN-MCMC* is combined with parallel tempering to obtain fast convergence with *highly* informative data. The combination is called *pCN-PT*.

This contribution revealed how large the difference between weakly and highly informative data is. For weakly informative data, we show that the *pCN-MCMC* and the *pCN-PT* have similar efficiency, although the latter method needs a multiple of computational power (due to parallel tempering). Hence, it is better to use the *pCN-MCMC* for weakly informative data. On the contrary, for highly informative data, the *pCN-PT* outperforms the *pCN-MCMC*, despite the additional computational costs of parallel tempering. As a result, the *pCN-PT* enables efficient Bayesian inversion of multi-Gaussian priors with highly informative data.

Currently, there is no Bayesian inversion benchmark with highly-informative data. The presented method is designed to be able to produce exact high-resolution reference solutions for Bayesian inversion benchmarks. We plan to write a follow-up paper that defines benchmarks and makes their reference solution (produced by the *pCN-PT*) available to the community. The long-term goal of this contribution is that all future proposed Bayesian inversion methods can be tested on these benchmarks to enable a quantitative comparison between methods.

Contribution 3: A joint approach to infer parameters and hyperparmeters

The third contribution extends the approach of my second contribution to infer the hyperparameters jointly together with the parameters. The presented approach is tested in different scenarios, and we present the results of Bayesian inversion with uncertain hyperparameters.

This contribution illustrates how significant the influence of hyperparameters is. While they are often assumed to be known, small changes in hyperparameters change the inverse solution drastically. Hence, modeling them is important but also challenging. Especially with indirect measurements, it is almost impossible to narrow down the hyperparameters. This leads to higher posterior uncertainties in parameters and hyperparameters.

In a nutshell, the presented MCMC method allows for efficient Bayesian inversion of multi-Gaussian priors with *highly* informative data and uncertain hyperparameters. In doing so, it enables a more realistic inverse modeling of subsurface parameters with appropriate uncertainty quantification. Similar to contribution two, this method will be used to create exact high-resolution reference solutions for Bayesian inversion benchmarks with highly informative data and uncertain hyperparameters. This will enable a quantitative comparison of inversion methods, which is currently impossible due to the lack of inverse reference solutions.

Contribution 4: The benefit of parallel tempering in combination with sequential Gibbs sampling

Fourth, I have investigated how to model nature more realistically. Here, I used hierarchical geostatistical models for a better approximation of nature. The prior and posterior of these models are highly multi-modal. As a result, most MCMC methods have trouble sampling the whole parameter space because they get stuck in local modes. To counteract this, I used parallel tempering and combined it with the *sequential Gibbs sampling* method. The latter resamples boxes of the parameter fields (categorical field and heterogeneity fields jointly) with increasing box size for larger temperatures. I named this approach *parallel-tempering sequential Gibbs MCMC* and showed convergence with different amounts of available data.

Interestingly, this work showed different behavior of the posterior distribution dependent on the amount of measurement data. For *highly* informative measurements, the structure of the artificial test aquifer was reconstructed with only a few uncertainties. With *weakly* informative data, the uncertainty of the predictions increased drastically and resulted in a multi-modal posterior where the position of subsurface structures was uncertain. Visualization of such distributions with only mean and variance does not help to understand the posterior. To remedy this issue, I used non-supervised clustering to visualize and quantify each mode separately.

To summarize, this contribution presents the *parallel-tempering sequential Gibbs MCMC*, which lays the groundwork to model and infer subsurface parameters more realistically. It is not restricted to solely using Gaussian random fields or MPS-induced priors but can perform Bayesian inversion with hierarchical prior models consisting of both. In the long run, this will lead to better, i.e., more realistic models of nature.

Contribution 5: The strong influence of measurement noise in BMS

Fifth, I discussed how and where measurement noise should be modeled in BMS. My results show that BMS results can change drastically depending on where noise is modeled and that these results correspond to different research questions that could be posed.

If the measurement noise is modeled in data and models, one answers the question ”which model is best in predicting the data-generating process (i.e., physics plus noise)?”. Performing BMS to answer this question is straightforward. However, I argue that most researchers are not interested in this question. Instead, ”which model is best in modeling the pure physics?” is the question most researchers want to pose. To answer this question, measurement noise needs to be removed from models and data. This comes with two challenges: First, noise-free data is not available with real-world measurements and the numerical computation of the BME is challenging. I show how to circumvent both challenges and approximate the BMS results. Interestingly, it is revealed that the easier ”data-generation question” is not a good proxy for the ”physics question”. Instead, an inconsistent case that uses noisy data and noise-free models can better approximate the ”physics questions”. Here, the model justifiability analysis can be used to evaluate how trustworthy the approximation is.

This contribution revealed and discussed the fundamental challenges of measurement noise in BMS. It demonstrates that ignoring this issue can lead to wrong results. To prevent mistakes, this contribution gives guidelines on how to handle measurement noise in BMS to avoid pitfalls that would distort the BMS results. This will lead future researchers to make the correct decision when selecting between models. As a result, more suitable models will be selected and used in practice, which gives decision-makers a better understanding of subsurface processes.

Summary of contributions

This thesis showed that the amount and uncertainty of measurements is an essential factor when inferring subsurface parameters. On the one hand, the amount of measurements, and hence the information content of the data, is vital to choose the appropriate MCMC method for Bayesian Inversion. On the other hand, the way measurement noise is handled in Bayesian model selection greatly influences the results. With my contributions, I enhanced the existing methods to infer and quantify the uncertainty of subsurface parameter fields.

I hope that future generations use the presented methods to infer the spatial distributions of subsurface parameters better. This helps to make better (or at least more informed) decisions when deciding over possible drinking water extraction wells or nuclear waste storage sites. In particular, the contributions of this thesis can prevent unexpected contamination of drinking water supply or wrong uncertainty quantification of leakages in nuclear waste storage sites.

Remaining research questions

The first three contributions introduced efficient MCMC methods for Bayesian inversion with multi-Gaussian priors. In contribution one, the pCN-MCMC is combined with Gibbs sampling, while contributions two and three show the effectiveness of a combination of the pCN-MCMC with parallel tempering. Future research can combine both ideas and develop a combination of sequential pCN-MCMC (from contribution one) and parallel tempering. I assume that the resulting approach would be even more efficient than both developed approaches for posterior sampling with *weakly* and *highly* informative data. The main challenge in developing such an MCMC is that the method introduced in contribution one, to optimize the tuning parameters β and κ during the algorithm's runtime, does not work in combination with parallel tempering. Further, each chain deployed in parallel tempering needs its own tuning parameters. As a result, systematic testing of β and κ is impossible, and a novel intelligent solution to optimize tuning parameters needs to be found.

The fourth contribution introduced a hierarchical framework to model subsurface structures. This framework was only tested with two categories and two Gaussian fields corresponding to each of the categories. I believe that the presented MCMC will perform well with training images with more than two categories and more Gaussian random fields. However, this was not proven, and the computational effort increases at least linearly with the number of categories used.

The fifth contribution discusses that the research question ”which model is best in modeling the pure physics?” can not be answered with real data using BMS. Instead, the answer to this question can only be approximated. Furthermore, I present one scheme of how this approximation can be done. Discussing and comparing different approximation schemes and giving general guidelines for when to use which scheme is an important task for the future.

Bibliography

J. P. Bennett, C. P. Haslauer, M. Ross, and O. A. Cirpka. An open, object-based framework for generating anisotropy in sedimentary subsurface models. *Groundwater*, 57(3):420–429, 2019. doi: 10.1111/gwat.12803.

A. Beskos, G. Roberts, A. Stuart, and J. Voss. MCMC methods for diffusion bridges. *Stochastics and Dynamics*, 08(03):319–350, 2008. doi: 10.1142/s0219493708002378.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434v2*, 2018.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995.

S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013. doi: 10.1214/13-STS421.

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

G. Evensen. *Data assimilation: the ensemble Kalman filter*. Springer, 2009.

J. Fu and J. J. Gómez-Hernández. Preserving spatial structure for inverse stochastic simulation using blocking Markov chain Monte Carlo method. *Inverse Problems in Science and Engineering*, 16(7):865–884, 2008. doi: 10.1080/17415970802015781.

J. Fu and J. J. Gómez-Hernández. A blocking markov chain Monte Carlo Method for Inverse Stochastic Hydrogeological Modeling. *Mathematical Geosciences*, 41(2): 105–128, 2009a. doi: 10.1007/s11004-008-9206-0.

J. Fu and J. J. Gómez-Hernández. Uncertainty assessment and data worth in groundwater flow and mass transport modeling using a blocking Markov chain Monte Carlo method. *Journal of Hydrology*, 364(3-4):328–341, 2009b. doi: 10.1016/j.jhydrol.2008.11.014.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

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

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984.

C. J. Geyer and E. A. Thompson. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *Journal of the American Statistical Association*, 90: 909–920, 1995. doi: 10.1080/01621459.1995.10476590.

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

B. A. W. Harbaugh, E. R. Banta, M. C. Hill, and M. G. Mcdonald. MODFLOW-2000 , The U .S . Geological Survey modular ground-water model — User guide to modularization concepts and the ground-water flow process. *U.S. Geological Survey*, page 130, 2000. doi: 10.1029/2006WR005839.

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

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

M. Höge. *Bayesian Multi-Model Frameworks-Properly Addressing Conceptual Uncertainty in Applied Modelling*. PhD thesis, Universitätsbibliothek Tübingen, 2019.

M. Höge, A. Guthke, and W. Nowak. The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, 572:96 – 107, 2019. doi: 10.1016/j.jhydrol.2019.01.072.

M. A. Iglesias and D. McLaughlin. Level-set techniques for facies identification in reservoir modeling. *Inverse Problems*, 27(3):035008, 2011.

M. A. Iglesias, K. Lin, and A. M. Stuart. Well-posed Bayesian geometric inverse problems arising in subsurface flow. *Inverse Problems*, 30(11):114001, 2014. doi: 10.1088/0266-5611/30/11/114001.

M. A. Iglesias, Y. Lu, and A. M. Stuart. A Bayesian level set method for geometric inverse problems. *Interfaces and Free Boundaries*, 18(2):181–217, 2016.

A. Jardani, A. Revil, and J. P. Dupont. Stochastic joint inversion of hydrogeophysical data for salt tracer test monitoring and hydraulic conductivity imaging. *Advances in Water Resources*, 52:62–77, 2013. doi: 10.1016/j.advwatres.2012.08.005.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957. doi: 10.1103/PhysRev.106.620.

P. Jussel, F. Stauffer, and T. Dracos. Transport modeling in heterogeneous aquifers: 1. statistical description and numerical generation of gravel deposits. *Water Resources Research*, 30(6):1803–1817, 1994. doi: 10.1029/94WR00162.

P. K. Kitanidis. Quasi-linear geostatistical theory for inversing. *Water Resources Research*, 31(10):2411–2419, 1995. doi: 10.1029/95WR01945.

P. K. Kitanidis. *Introduction to Geostatistics: Applications in Hydrogeology.* Cambridge university press, 1997.

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

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

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

N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207, 04 2006. doi: 10.1080/10635150500433722.

Y. Li, P. Wu, W. Liu, X. Sun, Z. Cui, and Y. Song. A microfocus X-ray computed tomography based gas hydrate triaxial testing apparatus. *Review of Scientific Instruments*, 90(5):055106, 2019. doi: 10.1063/1.5095812.

G. Mariethoz, P. Renard, and J. Straubhaar. The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, 46(11):1–14, 2010. doi: 10.1029/2008WR007621.

M. G. McDonald and A. W. Harbaugh. *A modular three-dimensional finite-difference ground-water flow model.* US Geological Survey, 1988.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953. doi: 10.1063/1.1699114.

A. Mondal, B. Mallick, Y. Efendiev, and A. Datta-Gupta. Bayesian Uncertainty Quantification for Subsurface Inversion Using a Multiscale Hierarchical Model. *Technometrics*, 56(3):381–392, 2014. doi: 10.1080/00401706.2013.838190.

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

S. S. Qian, C. A. Stow, and M. E. Borsuk. On Monte Carlo methods for Bayesian inference. *Ecological Modelling*, 159(2):269–277, 2003. doi: 10.1016/S0304-3800(02) 00299-5.

S. Reuschen, T. Xu, and W. Nowak. Bayesian inversion of hierarchical geostatistical models using a parallel-tempering sequential Gibbs MCMC. *Advances in Water Resources*, 141:103614, 2020. doi: 10.1016/j.advwatres.2020.103614.

S. Reuschen, A. Guthke, and W. Nowak. The four ways to consider measurement noise in Bayesian model selection—And which one to choose. *Water Resources Research*, 57(11), 2021a. doi: 10.1029/2021WR030391.

S. Reuschen, F. Jobst, and W. Nowak. Efficient discretization-independent Bayesian inversion of high-dimensional multi-Gaussian priors using a hybrid MCMC. *Water Resources Research*, 57(8), 2021b. doi: 10.1029/2021WR030051.

G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2002. doi: 10.1214/ss/1015346320.

A. Schöniger, T. Wöhling, L. Samaniego, and W. Nowak. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, 50(12):9484–9513, 2014. doi: 10.1002/2014WR016062.

A. Schöniger, W. A. Illman, T. Wöhling, and W. Nowak. Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. *Journal of Hydrology*, 531:96 – 110, 2015. doi: 10.1016/j.jhydrol.2015.07.047.

A. Smith and G. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of Royal Statistical Society: Part B*, 55(1):3–23, 1993.

A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992. doi: 10.1080/00031305.1992.10475856.

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

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

L. Wasserman et al. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

P. Wu, Y. Li, X. Sun, W. Liu, and Y. Song. Pore-scale 3d morphological modeling and physical characterization of hydrate-bearing sediment based on computed tomography. *Journal of Geophysical Research: Solid Earth*, 125(12):e2020JB020570, 2020. doi: 10.1029/2020JB020570.

S. Xiao, T. Xu, S. Reuschen, W. Nowak, and H.-J. Hendricks Franssen. Bayesian inversion of multi-Gaussian log-conductivity fields with uncertain hyperparameters: an extension of preconditioned Crank-Nicolson Markov chain Monte Carlo with parallel tempering. *Water Resources Research*, 57(9), 2021. doi: 10.1029/2021WR030313.

T. Xu, S. Reuschen, W. Nowak, and H.-J. Hendricks Franssen. Preconditioned Crank-Nicolson Markov chain Monte Carlo coupled with parallel tempering: an efficient method for Bayesian inversion of multi-Gaussian log-hydraulic conductivity fields. *Water Resources Research*, 56(8), 2020. doi: 10.1029/2020WR027110.

T.-C. J. Yeh, M. Jin, and S. Hanna. An iterative stochastic inverse method: Conditional effective transmissivity and hydraulic head fields. *Water Resources Research*, 32(1): 85–92, 1996. doi: 10.1029/95WR02869.

List of Publications

Contribution 1: Reuschen, S., Jobst, F., & Nowak, W. (2021). Efficient discretization-independent Bayesian inversion of high-dimensional multi-Gaussian priors using a hybrid MCMC. *Water Resources Research*, 57(8), DOI: 10.1029/2021WR030051

Contribution 2: Xu, T., Reuschen, S., Nowak, W., & Hendricks Franssen, H. J. (2020). Preconditioned Crank-Nicolson Markov chain Monte Carlo coupled with parallel tempering: an efficient method for Bayesian inversion of multi-Gaussian log-hydraulic conductivity fields. *Water Resources Research*, 56(8), DOI: 10.1029/2020WR027110 *

Contribution 3: Xiao, S., Xu, T., Reuschen, S., Nowak, W., & Hendricks Franssen, H.J. (2021). Bayesian inversion of multi-Gaussian log-conductivity fields with uncertain hyperparameters: an extension of preconditioned Crank-Nicolson Markov chain Monte Carlo with parallel tempering. *Water Resources Research*, 57(9), DOI: 10.1029/2021WR030313

Contribution 4: Reuschen, S., Xu, T., & Nowak, W. (2020). Bayesian inversion of hierarchical geostatistical models using a parallel-tempering sequential Gibbs MCMC. *Advances in Water Resources*, 141, 103614, DOI: 10.1016/j.advwatres.2020.103614 †

Contribution 5: Reuschen, S. and Guthke, A., & Nowak, W (2021). The four ways to consider measurement noise in Bayesian model selection—And which one to choose. *Water Resources Research*, 57(11), 10.1029/2021WR030391

* Republished with permission of John Wiley & Sons - Books; permission conveyed through Copyright Clearance Center, Inc.

† Republished with permission of Elsevier Science & Technology Journals; permission conveyed through Copyright Clearance Center, Inc.

A Contribution 1: Efficient discretization-independent Bayesian inversion of high-dimensional multi-Gaussian priors using a hybrid MCMC

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR030051

Efficient Discretization-Independent Bayesian Inversion of High-Dimensional Multi-Gaussian Priors Using a Hybrid MCMC

Sebastian Reuschen[1], **Fabian Jobst**[1], **and Wolfgang Nowak**[1]

[1]Department of Stochastic Simulation and Safety Research for Hydrosystems, University of Stuttgart, Stuttgart, Germany

Key Points:

- A hybrid MCMC between the sequential Gibbs and the pCN-MCMC approach is presented
- This hybrid MCMC is more efficient than both special cases
- We present an adaptive algorithm to optimize the hyper-parameters of this hybrid MCMC during burn-in

Correspondence to:

S. Reuschen,
sebastian.reuschen@iws.uni-stuttgart.de

Citation:

Reuschen, S., Jobst, F., & Nowak, W. (2021). Efficient discretization-independent Bayesian inversion of high-dimensional multi-Gaussian priors using a hybrid MCMC. *Water Resources Research*, *57*, e2021WR030051. https://doi.org/10.1029/2021WR030051

Received 25 MAR 2021
Accepted 8 JUL 2021

Abstract

In geostatistics, Gaussian random fields are often used to model heterogeneities of soil or subsurface parameters. To give spatial approximations of these random fields, they are discretized. Then, different techniques of geostatistical inversion are used to condition them on measurement data. Among these techniques, Markov chain Monte Carlo (MCMC) techniques stand out, because they yield asymptotically unbiased conditional realizations. However, standard Markov Chain Monte Carlo (MCMC) methods suffer the curse of dimensionality when refining the discretization. This means that their efficiency decreases rapidly with an increasing number of discretization cells. Several MCMC approaches have been developed such that the MCMC efficiency does not depend on the discretization of the random field. The preconditioned Crank Nicolson Markov Chain Monte Carlo (pCN-MCMC) and the sequential Gibbs (or block-Gibbs) sampling are two examples. This paper presents a combination of both approaches with the goal to further reduce the computational costs. Our algorithm, the sequential pCN-MCMC, will depend on two tuning-parameters: the correlation parameter β of the pCN approach and the block size κ of the sequential Gibbs approach. The original pCN-MCMC and the Gibbs sampling algorithm are special cases of our method. We present an algorithm that automatically finds the best tuning-parameter combination (κ and β) during the burn-in-phase of the algorithm, thus choosing the best possible hybrid between the two methods. In our test cases, we achieve a speedup factors of 1–5.5 over pCN and of 1–6.5 over Gibbs. Furthermore, we provide the MATLAB implementation of our method as open-source code.

1. Introduction

The heterogeneity of soil parameters is a key control on subsurface flow and transport. Geostatistical methods are usually used to characterize these heterogeneities (e.g., Refsgaard et al., 2012). In general, all soil parameters can be described by random functions. In this work, we focus on soil parameters which can be (a priori) described by Gaussian processes. A Gaussian process is a stationary random function, in which any finite collection of variables can be described by a multivariate normal distribution. Such a distribution is fully described by a mean vector and a covariance matrix.

The goal of Bayesian inversion is to predict (and give uncertainties) of parameters given measurements. The probability distribution of parameters before measurements is called prior probability distribution, whereas the conditional probability (after measurements) is called a posterior probability distribution. If the parameters are measured directly, the Kriging (also called Gaussian process regression) procedure allows us to calculate the posterior probability distributions of all parameters analytically (e.g., Kitanidis, 1997). However, if the parameters are not measured directly (here: measure the hydraulic head and infer the hydraulic conductivity), Kriging is not applicable.

Instead, sampling methods can be used to solve this problem. Examples are rejection sampling (e.g., Gelman et al., 1995, Chapter 10.2), which is applicable to low-dimensional prior distributions or weak data, Ensemble Kalman filters (e.g., Evensen, 2009), which are used to linearize forward models for multi-Gaussian posteriors, and more. Here, we will focus on Markov chain Monte Carlo (MCMC) methods which are universally applicable for Bayesian inference (e.g., Qian et al., 2003) but computationally expensive.

In MCMC approaches, the random function is discretized to enable numerical computations. The main problem of most common MCMC methods, for example, the Metropolis-Hastings algorithm (Hastings, 1970;

© 2021. The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Metropolis et al., 1953), is, that a refinement of discretization leads to worse convergence speed of the methods (Cotter et al., 2013). Different approaches have been presented in the literature to overcome this challenge. In the following, we present two approaches.

The key idea of the (sequential) Gibbs approach (e.g., Gelman et al., 1995, Chapter 11.3) is to randomly modify only one (or a subset of) parameter(s). This modification respects both the prior distribution and the surrounding values of the parameters that stay fixed. In subsequent steps, different parameter(s) get modified. This leads to a discretization-independent efficiency for arbitrary prior distributions (e.g., Fu & Gómez-Hernández, 2008). The limitations of the Gibbs approach are that the conditional sampling from the prior needs to be possible and computationally cheap (e.g., Fu & Gómez-Hernández, 2008). In most applications, however, the forward simulation is the computational bottleneck. One specific version of this approach was proposed by Fu and Gómez-Hernández (2008, 2009a, 2009b), who resampled boxes of the multi-Gaussian parameter field. Hansen et al. (2012) applied this idea to resample boxes of the parameter field in a binary classification problem. A combination of these approaches for binary classification problems with multi-Gaussian heterogeneity was presented by Reuschen et al. (2020).

The second discretization-independent approach we discuss is the preconditioned Crank Nicolson MCMC (pCN-MCMC) (Beskos et al., 2008; Cotter et al., 2013). It is easy to implement and computationally fast, as demonstrated in the respective original papers and in a recent effort to construct reference solutions algorithms for geostatistical inversion benchmarks (e.g., Xu et al., 2020). In fact, pCN-MCMC has been derived for inverting random *functions*, so that the numerical discretization of the random field does not matter for its convergence speed by construction. However, the pCN-MCMC can only be used for multi-Gaussian priors. The reason for this restriction is that the pCN proposed modifications to random fields by a small-magnitude random field to a dampened version of the current field, thus resembling an autoregressive process of order one along the chain (Beskos et al., 2008). While this way of proposing new solutions is highly effective and independent in convergence speed of the spatial discretization, it can only be constructed when the prior is multi-Gaussian.

Other alternatives using spectral parametrization (Laloy et al., 2015), Karhunen-Loeve expansions (e.g., Mondal et al., 2014) or pilot point methods (e.g., Jardani et al., 2013) use dimension reduction approaches for fast convergence. The consequences of these approaches are twofold: On the one hand, dimension reduction approaches can reduce computational cost. On the other hand, they only converge toward approximate solutions of the true posterior. In this work, we focus on methods that converge to the true posterior.

Another direction of research uses derivatives of the posterior distribution to increase the efficiency of MCMC methods (e.g., Hamiltonian MCMC as summarized in Betancourt, 2017). Analytical derivatives of the likelihood function are possible in some scenarios, but in most hydraulic forward models, they are impossible. Numerical approximation of gradients is an alternative. However, numerical differentiation is computationally expensive and often negates the advantage of these methods. Hence, we focus on methods that do not require gradient information.

Reducing the computational costs of state-of-the-art MCMC methods is an integral part to make MCMC methods more attractive to a broad community. We combine the sequential Gibbs idea with the pCN-MCMC idea and create a hybrid method called sequential pCN-MCMC to reduce this computational burden. However, the hybrid method comes with two tuning-parameters. The standard way of finding the optimal tuning-parameters in MCMC algorithms for high-dimensional inverse problems is to tune them for an acceptance rate equal to 23.4% (see Gelman et al., 1996). In our hybrid method, this is not possible anymore because infinitely many tuning-parameter combinations lead to the same acceptance rate. Hence, we refrain to the efficiency defined in Gelman et al. (1996) to find the optimal tuning-parameter combination.

Overall, the novelty of our paper is a combination of the sequential Gibbs MCMC and the pCN-MCMC, which we call sequential pCN-MCMC. Here, the pCN-MCMC and sequential Gibbs are special cases of the sequential pCN-MCMC. Our hypothesis is that the most efficient method is neither of the special cases.

We compare the new hybrid method to the two original algorithms for Bayesian inversion of fully saturated groundwater flow. Here, we use different scenarios where we alter the prior information, discretization, and measurement type to test and confirm our hypothesis. A variety of different measurement types are used for geostatistical Bayesian inversion (e.g., Butera & Soffia, 2017; Ezzedine & Rubin, 1996; Gutjahr et al., 1994; Zimmerman et al., 1998) and optimal choices can be made (Nowak et al., 2010). Here, however, we use a typical benchmarking setup of Xu et al. (2020) with hydraulic head or concentration measurements, because it ensures intercomparability of results. However, our method is not restricted to this choice of data. The performance of all methods is investigated using the acceptance rate, efficiency (Gelman et al., 1996), Kullback-Leibner divergence (Kullback & Leibler, 1951), and R-statistic (Gelman & Rubin, 1992). The MATLAB implementation of our code is available at https://bitbucket.org/Reuschen/sequential-pcn-mcmc.

The paper is structured as follows: Section 2gives a definition of the inverse problem, an overview over existing methods and introduces metrics to evaluate the performance of algorithms. In Section 3, we present our proposed sequential pCN-MCMC method. After that, we introduce our test cases in Section 4. Our results are shown in Section 5 and discussed in Section 6. Finally, Section 7 concludes the most important findings in a short summary.

2. Methods

In this section, we briefly recall existing MCMC methods for multi-Gaussian priors. We focus on those without dimensionality reduction and without derivatives. First, we give the definition of the problem class in Section 2.1. In Section 2.2, we introduce the generic MCMC approach. After that, we recall the Metropolis-Hastings approach in Section 2.3 and discuss the differences to the so-called prior sampling methods in Section 2.4. Sections 2.5 and 2.6 introduce the existing algorithms pCN-MCMC and sequential Gibbs sampling, respectively, which are both instances of prior sampling methods. Finally, we present metrics to evaluate the presented methods in Section 2.7.

2.1. Bayesian Inference

Let

$$\mathbf{d} = F(\boldsymbol{\Theta}) + \mathbf{e} \tag{1}$$

be the stochastic representation of a forward problem. $F(\boldsymbol{\Theta})$ is an error-free deterministic forward model. Equation 1 describes the relation between the unknown and uncertain parameters $\boldsymbol{\Theta}$ and the measurements $\mathbf{d}$. The noise term $\mathbf{e}$ aggregates all error terms. The goal of Bayesian inversion is to infer the posterior parameter distribution of $\boldsymbol{\Theta}$ based on prior knowledge of $\boldsymbol{\Theta}$ and the data $\mathbf{d}$ under the model F.

We use the parameters $\boldsymbol{\theta}$ to refer to realizations of the random variable $\boldsymbol{\Theta}$ with some prior distribution $p(\boldsymbol{\Theta})$ and a posterior distribution $p(\Theta|\mathbf{d})$. The resulting posterior density can be evaluated for each realization $\boldsymbol{\theta}$ as

$$p(\theta|\mathbf{d}) = \frac{p(\boldsymbol{\theta})p(\mathbf{d}|\theta)}{p(\mathbf{d})} \propto p(\boldsymbol{\theta})p(\mathbf{d}|\theta) = P(\boldsymbol{\theta})L(\theta|\mathbf{d}). \tag{2}$$

In this paper, we define the prior distribution $P(\boldsymbol{\theta}) := p(\boldsymbol{\theta})$ and the likelihood $L(\theta) := p(\mathbf{d}|\boldsymbol{\theta})$ for shorter notation. This likelihood assumes that the data $\mathbf{d}$ do not change during the runtime of the algorithm. The challenge in high-dimensional Bayesian inversion is to sample efficiently from the posterior distribution $p(\theta|\mathbf{d})$.

2.2. Generic Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a popular, accurate, but typically time-intensive algorithm to sample from the posterior distribution. In contrast to other methods, it only needs the unnormalized posterior density

$$\pi(\theta) = P(\theta)L(\theta) \propto p(\theta|\mathbf{d}) \tag{3}$$

to sample from the posterior distribution $p(\theta|\mathbf{d})$.

In the following, we name all properties that an MCMC method needs to fulfill to converge to the exact posterior distribution. Based on these, we derive the formulas of our proposed MCMC. A general introduction to MCMC can be found in Chib and Greenberg (1995).

MCMC methods converge to π (as presented in Equation 3) at the limit of infinite runtime (Smith & Roberts, 1993) if and only if irreducibility, aperiodicity and the detailed balance are fulfilled. Irreducibility and aperiodicity are fulfilled for multi-Gaussian proposals (see below) in continuous problems, which are typically used for engineering purposes. Consequently, we focus on the detailed balance in the following. Given any two parameter sets $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$, the detailed balance is defined as

$$\pi(\theta)h(\theta,\tilde{\theta}) = \pi(\tilde{\theta})h(\tilde{\theta},\theta) \tag{4}$$

with the transition kernel h, which is defined as

$$h(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})\alpha(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}). \tag{5}$$

The term $q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})$ refers to the proposal distribution and $\alpha(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})$ to the acceptance probability. Here, $\tilde{\boldsymbol{\theta}}$ is proposed based on the current parameter $\boldsymbol{\theta}$. Equations 3–5 can be combined to

$$\alpha(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = min\left[\frac{P(\tilde{\boldsymbol{\theta}})L(\tilde{\boldsymbol{\theta}})q(\tilde{\boldsymbol{\theta}},\boldsymbol{\theta})}{P(\boldsymbol{\theta})L(\boldsymbol{\theta})q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})},1\right]. \tag{6}$$

Equation 6 provides an α such that the detailed balance is fulfilled. This holds for any prior P, any likelihood L and any proposal distribution q. Consequently, an infinite number of possible proposal distributions q exist (Gelman et al., 1995, Chapter 11.5). This raises the questions of how to choose q for fast convergence in a given problem class.

Fast convergence (after burn-in) is mainly a question of low autocorrelation of successive samples (Gelman et al., 1996). This is achieved by large changes in the parameter space. As a result, it is desirable to propose far jumps for new candidate points $\tilde{\boldsymbol{\theta}}$ via the proposal function q and still hope to accept them with a high probability α. However, in most practical cases, these two properties contradict each other: making large changes in $\boldsymbol{\theta}$ results in distinct $P(\tilde{\boldsymbol{\theta}})L(\tilde{\boldsymbol{\theta}})$ and $P(\boldsymbol{\theta})L(\boldsymbol{\theta})$, which results in a small α. In contrast, small changes in $\boldsymbol{\theta}$ results in similar $P(\tilde{\boldsymbol{\theta}})L(\tilde{\boldsymbol{\theta}})$ and $P(\boldsymbol{\theta})L(\boldsymbol{\theta})$ (if the prior and the likelihood function are smooth), which results in α close to 1. Thus, Gelman et al. (1996) stated that a trade-off between the size of the change and the acceptance rate needs to be found.

2.3. Metropolis Hastings

The Metropolis-Hastings (MH) algorithm (Hastings, 1970; Metropolis et al., 1953) can be used with arbitrary proposal functions. Here, we present the random walk MH algorithm. It assumes a symmetric proposal distribution

$$q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = q(\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}). \tag{7}$$

Inserting this into Equation 6, it follows that

$$\alpha(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = \min\left[\frac{P(\tilde{\boldsymbol{\theta}})L(\tilde{\boldsymbol{\theta}})}{P(\boldsymbol{\theta})L(\boldsymbol{\theta})},1\right] = \min\left[\frac{\pi(\tilde{\boldsymbol{\theta}})}{\pi(\boldsymbol{\theta})},1\right]. \tag{8}$$

The MH algorithm samples from any parameter distribution $\pi(\boldsymbol{\theta})$. The specific proposal function q of the so-called random walk MH is given by

$$\tilde{\theta} = \theta + \beta\xi, \quad \xi \sim N(0,\mathbb{1}). \tag{9}$$

Here, the parameter β controls how big the change between successive parameters $\boldsymbol{\theta}$ is. The proposal function q of the random walk MH algorithm fulfills Equation 7 because the proposal step (Equation 9) is symmetric per definition.

The main weakness of the Metropolis-Hastings algorithm is that the acceptance rate in Equation 7 decreases rapidly for increasing β, especially in high-dimensional problems (Roberts & Rosenthal, 2002). This can be improved by using the additional information that $\pi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})L(\boldsymbol{\theta})$. This enables us to make the acceptance rate α only dependent on $L(\boldsymbol{\theta})$ in the next section.

2.4. Prior Sampling

Bayesian inversion methods often exploit the knowledge that the posterior distribution follows by construction from $\pi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})L(\boldsymbol{\theta})$ to increase the efficiency (see Section 2.7.2 for the definition of efficiency). To exploit this situation, the a priori knowledge contained in the prior distribution $P(\boldsymbol{\theta})$ is used to define a tailored proposal distribution $q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})$ (which is only efficient for the respective prior distribution). Mathematically, this is realized by defining $q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})$ such that it fulfills

$$q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = \frac{P(\tilde{\boldsymbol{\theta}})}{P(\boldsymbol{\theta})}q(\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}). \tag{10}$$

Combining Equations 6 and 10 results in (e.g., Tarantola, 2005)

$$\alpha(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = \min\left[\frac{L(\tilde{\boldsymbol{\theta}})}{L(\boldsymbol{\theta})},1\right]. \tag{11}$$

Many problem classes, for example, high-dimensional geoscience problems, have a complex prior $P(\boldsymbol{\theta})$. As a result, the acceptance rate α is almost exclusively dependent on the prior. This leads to decreasing efficiencies of the MCMC (Roberts & Rosenthal, 2002). Equations 10 and 11 enable us to circumvent that problem by making the acceptance rate α only dependent on the likelihood, because the prior is already considered in the proposal distribution. This makes it possible to have high acceptance rates α even for far jumps in the parameter space, which is synonymous with a high efficiency (see Section 2.7.2). We call this approach “sampling from the prior distribution” (Reuschen et al., 2020).

In the following, we will present the preconditioned Crank Nicolson MCMC (pCN-MCMC) and the block-Gibbs MCMC algorithms. The proposal functions of both methods fulfill Equation 10. Based on them, we propose our new sequential pCN-MCMC algorithm, which combines the approaches of pCN and Gibbs.

2.5. pCN-MCMC

The idea of the preconditioned Crank Nicolson MCMC (pCN-MCMC) was first introduced by Beskos et al. (2008), who called it a Langevin MCMC approach. In 2013, Cotter et al. (2013) revived the idea and named it pCN-MCMC.

The pCN-MCMC takes the assumption that the prior $P(\boldsymbol{\theta})$ is multi-Gaussian ($\boldsymbol{\theta} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$). For these priors, the proposal step of the pCN-MCMC

$$\tilde{\boldsymbol{\theta}}^{(i)} = \sqrt{(1-\beta^2)}\left(\boldsymbol{\theta}^{(i)} - \boldsymbol{\mu}\right) + \beta\boldsymbol{\xi}^{(i)} + \boldsymbol{\mu}, \quad \boldsymbol{\xi}^{(i)} \sim N(\mathbf{0},\boldsymbol{\Sigma}) \tag{12}$$

fulfills Equation 10. Hence, the acceptance probability α is only depended on the likelihood as denoted in Equation 6. The tuning-parameter β of the pCN-MCMC specifies the change between subsequent samples. For $\beta = 1$, subsequent samples are independent of each other. For lower β, the similarity of samples increases up to the theoretical limit of $\beta = 0$ where subsequent samples are identical. In most applications, similar samples lead to similar likelihoods and a high acceptance rate. Hence, the tuning-parameter β can be used to adjust the acceptance rate of pCN-MCMC algorithms (large β lead to low acceptance rates and vice versa). A pseudocode of the pCN-MCMC is presented in the Appendix A.

2.6. Sequential Gibbs Sampling

In 1987, Geman and Geman (1984) introduced Gibbs sampling as a specific instance of Equation 10. The basic concept of Gibbs sampling is to resample parts of the parameter space $\boldsymbol{\theta}$. In the geostatistical context, this typically means to select a random box within the parameter field, and then to generate a new random field within that box while keeping the parts outside the box fixed. The new random part is sampled from the prior, but under the condition that it must match (e.g., by conditional sampling) with the outside part. For illustrative examples on Gibbs sampling, we refer to (Gelman et al., 1995).

Assuming a random parameter vector $\boldsymbol{\theta}$ of size $N_p \times 1$ (N_p denotes number of parameters) and some permutation matrix $\mathbf{M}$ (usually called P_π in the literature), we can order the random variables into two parts

$$\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix} = \mathbf{M}\boldsymbol{\theta} \text{ with size} \begin{bmatrix} q \times 1 \\ (N_p - q) \times 1 \end{bmatrix}, \tag{13}$$

where $\boldsymbol{\theta}_1$ incorporates all parameters which will be resampled conditionally on $\boldsymbol{\theta}_2$. The number of resampled parameters is given by q. A new proposal is defined as

$$\tilde{\boldsymbol{\theta}} = \mathbf{M}^{-1}\begin{bmatrix} \tilde{\boldsymbol{\theta}}_1 \\ \tilde{\boldsymbol{\theta}}_2 \end{bmatrix} = \mathbf{M}^T\begin{bmatrix} \tilde{\boldsymbol{\theta}}_1 \\ \tilde{\boldsymbol{\theta}}_2 \end{bmatrix} = \mathbf{M}^T\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\theta}_2 = \mathbf{r} \end{bmatrix}, \quad \boldsymbol{\xi} \sim p_\pi(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2 = \mathbf{r}). \tag{14}$$

Here, the values $\mathbf{r}$ of $\boldsymbol{\theta}_2$ remain constant, whereas the first part of the parameter space $\boldsymbol{\theta}_1$ gets resampled conditional on $\boldsymbol{\theta}_2$. This approach is applicable to any probability distribution for which the conditional probability distribution $p_\pi(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2 = \mathbf{r})$ can be sampled.

In this work, we follow the approach of Fu and Gómez-Hernández (2008, 2009a, 2009b) to resample boxes in a parameter space representing a two-dimensional domain. Let $\boldsymbol{\theta}$ be a discretization of some parameter field (e.g., hydraulic conductivity). Here $\boldsymbol{\theta}(x, y)$ is the value of the parameter $\boldsymbol{\theta}$ at the spatial position $(x, y) \in ([0, l_x], [0, l_y])$. Hereby, let l_x and l_y be the length of the investigated domain in x and y direction. To determine M, we use a parameter and $\kappa \in (0,1]$ that defines the size of the resampled box as defined in Equation 15, where a larger κ corresponds to a larger resampling box. To include the dependence of $\mathbf{M}$ on κ, we will denote it as $\mathbf{M}_\kappa$ in the following.

Based on a randomly chosen center point (x^*, y^*) (which is rechosen every MCMC step), we can choose $\mathbf{M}_\kappa$ such that

$$\begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix} = \begin{bmatrix} \left\{\boldsymbol{\theta}(x, y)\right\}, \text{with} \left|\dfrac{x - x^*}{l_x}\right| \leq \kappa \text{ and} \left|\dfrac{y - y^*}{l_y}\right| \leq \kappa \\ \left\{\boldsymbol{\theta}(x, y)\right\}, \text{with} \left|\dfrac{x - x^*}{l_x}\right| > \kappa \text{ or} \left|\dfrac{y - y^*}{l_y}\right| > \kappa \end{bmatrix} = \mathbf{M}_\kappa\boldsymbol{\theta}. \tag{15}$$

This means that all parameters $\boldsymbol{\theta}(x, y)$ with a distance smaller than κ to the centerpoint (x, y) are part of the parameter set $\boldsymbol{\theta}_1$ and all $\boldsymbol{\theta}(x, y)$ with a distance larger than κ are part of the parameter set $\boldsymbol{\theta}_2$. Pseudocode for computing $\mathbf{M}_\kappa$ is shown in the Appendix A.

Following Fu and Gómez-Hernández (2008, 2009a, 2009b), this work will focus on multi-Gaussian priors. In a multi-Gaussian prior setting, the prior probability distribution is only based on the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. According to Equation 13, we portion $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{with sizes} \begin{bmatrix} q \times 1 \\ (N_p - q) \times 1 \end{bmatrix}, \tag{16}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{with sizes} \begin{bmatrix} q \times q & q \times (N_p - q) \\ (N_p - q) \times q & (N_p - q) \times (N_p - q) \end{bmatrix}. \tag{17}$$

Here, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ signify the covariance matrices of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ whereas $\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{21}$ denote the cross-covariance matrices between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. With that, we can express the resampled parameter distribution for $\tilde{\boldsymbol{\theta}}_1$ as $P(\tilde{\boldsymbol{\theta}}_1 \mid \boldsymbol{\theta}_2 = \mathbf{r}) \sim N(\widetilde{\boldsymbol{\mu}_1, \tilde{\boldsymbol{\Sigma}}}_{11})$ using the Kriging theory.

The Kriging (or Gaussian progress regression) theory (e.g., Rasmussen & Williams, 2006) states that the conditional probability is multi-Gaussian with mean

$$\tilde{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{r} - \boldsymbol{\mu}_2) \tag{18}$$

and the covariance matrix

$$\tilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{19}$$

After combining Equation 14, Equations 18 and 19, we arrive at the proposal distribution

$$\tilde{\boldsymbol{\theta}} = \mathbf{M}^T\begin{bmatrix}\tilde{\boldsymbol{\theta}}_1\\ \boldsymbol{\theta}_2\end{bmatrix}, \quad \tilde{\boldsymbol{\theta}}_1 \sim N(\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_{11}) \tag{20}$$

which fulfills Equation 10.

The tuning-parameter κ specifies the size of the resampling box and therefore the change between subsequent samples. Thereby, smaller κ will lead to more similarity of subsequent samples and hence, to higher acceptance rates. The Appendix A includes a pseudocode of the sequential Gibbs sampling method.

2.7. Metrics

The quality of MCMC methods can be quantified using different metrics. An overview over such metrics can be found in Cowles and Carlin (1996) or Roy (2020). We use the following four test metrics.

2.7.1. Acceptance Rate $\bar{\alpha}$

The acceptance rate $\bar{\alpha}$ is the fraction of proposals that get accepted divided by the total number of proposals. Gelman et al. (1996) showed empirically that $\bar{\alpha} = 0.234$ is optimal for normal target distributions. This value of $\bar{\alpha}$ is often used to optimize the tuning-parameter (e.g., β) of MCMC runs because it is easy to implement.

2.7.2. Efficiency

The efficiency of one parameter j within a MCMC chain is defined as (e.g., Gelman et al., 1996)

$$eff_j = \frac{1}{1 + 2\cdot\sum_{i=1}^{\inf}\rho_i} \tag{21}$$

where ρ_i is the autocorrelation of the chain with lag i. With this, the effective sample size (ESS) (Robert & Casella, 2013) is defined as

$$ESS_j = eff_j \cdot N \tag{22}$$

with N being the total number of MCMC samples. The ESS represents the number of independent samples equivalent (i.e., having the same error) to a set of correlated MCMC samples. Hence, the efficiency (or ESS) can be used to estimate the number of MCMC samples needed to get a certain number of independent samples.

In the following, we aggregate the individual efficiencies of all parameters to one combined efficiency. Therefore, we define the efficiency of several parameters as

$$eff = \frac{1}{1 + 2\cdot\frac{1}{N_p}\sum_{j=1}^{N_p}\sum_{i=1}^{\inf}\rho_{i,j}} \tag{23}$$

with N_p being the number of parameters and $\rho_{i,j}$ being the autocorrelation of length i of the jth parameter.

2.7.3. R-Statistic

The potential scale reduction factor $\sqrt{\hat{R}}$ introduced by Gelman and Rubin (1992) is a popular method for MCMC diagnostics. It measures the similarity of posterior distributions, generated by different independent MCMC chains, by comparing their first two moments. Similarity between posterior distributions suggests convergence of the chains. This enables a convergence test in the absence of reference solutions. Gelman et al. (1995) stated that $\sqrt{\hat{R}} \le 1.2$ signifies acceptable convergence.

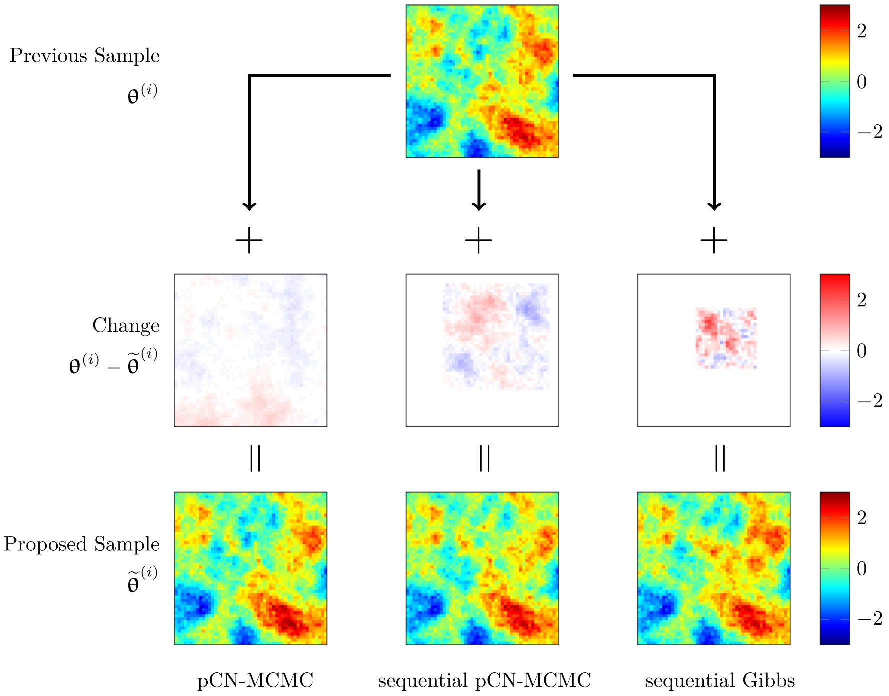


Figure 1. Proposal step of pCN-MCMC, sequential pCN-MCMC and sequential Gibbs sampling. The pCN-MCMC makes a small global change. The sequential Gibbs sampling makes a large local change. The sequential pCN-MCMC makes a medium change in a medium-sized area. This figure is only for visualization. Realistic problems lead to smaller changes in all three algorithms.

2.7.4. Kullback-Leibler Divergence

The Kullback-Leibler divergence (Kullback & Leibler, 1951) is a measure to compare probability density functions. In this paper, we estimate the marginal density of each parameter and compute the KL-divergence of the MCMC chain to a reference solution. This leads us to as many KL-divergences as we have parameters. To aggregate the KL-divergences over all parameters, we only report the mean value.

The KL-divergence is used solely as a postprocessing metric in our work. The advantage in comparison to the R-statistic is twofold in our case: first, we use the R-statistic to show that two independent chains converge to the same distribution. In contrast, we use the KL-divergence to show convergence to a previously calculated reference distribution. Second, the R-statistic shows convergence in the first two moment whereas the KL-divergence shows convergence in the entire distribution.

3. Sequential pCN-MCMC

In this section, we present our proposed sequential pCN-MCMC. To understand the underlying idea, let us look at the different MCMCs from a conceptual point of view. On the one hand, the proposal method of the pCN-MCMC makes global, yet small, changes that sample from the prior (left column of Figure 1). On the other hand, the sequential Gibbs method makes local, yet large, changes that also sample from the prior (right column of Figure 1). We want to combine these two approaches to make medium changes in a medium-sized area which again sample from the prior (center column of Figure 1).

We take the same preparatory steps as in the sequential Gibbs approach (Equations 13–19). However, we propose a new sample within the resampling box based on the pCN approach (Equation 12)

$$\tilde{\boldsymbol{\theta}}_1 = \sqrt{(1-\beta^2)}\left(\boldsymbol{\theta}_1^{(i)} - \tilde{\boldsymbol{\mu}}_1\right) + \beta\boldsymbol{\xi} + \tilde{\boldsymbol{\mu}}_1, \quad \boldsymbol{\xi} \sim N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{11}). \tag{24}$$

$\tilde{\boldsymbol{\mu}}_1$ and $\tilde{\boldsymbol{\Sigma}}_{11}$ have been defined in Equation 18 and Equation 19, respectively. Consequently, the proposal distribution is defined as

$$\tilde{\boldsymbol{\theta}} = \mathbf{M}_{\kappa}^{T}\begin{bmatrix}\tilde{\boldsymbol{\theta}}_1\\ \boldsymbol{\theta}_2\end{bmatrix}, \quad \tilde{\boldsymbol{\theta}}_1 \sim \text{see equation 24}\,. \tag{25}$$

This allows for sequential pCN-MCMC proposal in blocks of the parameter space.

Any combination of the tuning-parameter β of the pCN-MCMC approach and the tuning-parameter κ of the Gibbs approach can be chosen. Alike the pure cases, increasing κ and β will lead to larger changes in subsequent samples and lower acceptance rates. Both, the sequential Gibbs and the pCN-MCMC are special cases of the proposed sequential pCN-MCMC. The sequential Gibbs method is the special case for $\beta = 1$ and $\kappa = 1$ leads to the pCN-MCMC approach.

3.1. Adaptive Sequential pCN-MCMC

Section 2.7.1 stated that the acceptance rate $\overline{\alpha}$ is often used to tune the tuning-parameters of MCMC methods. This tuning does only work for one tuning-parameter. In our proposed method, we have two tuning-parameter, namely the box size κ of the sequential Gibbs method and the pCN parameter β. The presence of two tuning-parameters destroys the uniqueness of the optimum ($\overline{\alpha} = 0.234$). Hence, we need to find the optimal (or good enough) parameter combination in another way.

We propose an adaptive version of the sequential pCN-MCMC, which finds the optimal tuning-parameter during its burn-in period. For this, we take a gradient descent approach. As starting values, we choose random values between 10^{-4} and 10^{0} for β and κ. We evaluate the performance of tuning-parameters by running the MCMC for N_{hp} steps, for example, 10^3 steps, with the same tuning-parameters. Then we evaluate the produced subsample based on the efficiency (see Section 2.7.2). The efficiency is used because it can be evaluated, unlike the R-statistic or KL-divergence, during the runtime of the algorithm. The R-statistic and the KL-divergence need several independent chains to assess performance. The presented approach uses one chain, so it cannot build its optimization on the R-statistic or KL-divergence (while it can use the efficiency), but we use them as independent and complementary checks.

However, the autocorrelation, on which the efficiency is based on, is normalized by the total variance of the sample. Hence, the efficiency is independent of the total variance of the sample. To favor tuning-parameters that explore the posterior as much as possible, even in a small subsample, we add a standard derivation term to the objective function. For an increasing size of the subsamples, the effect of the standard derivation diminishes as the standard derivation of all subsamples converges to the standard derivation of the posterior. This leads to the objective function (which should be maximized) of a subsample

$$f(\beta,\kappa) = \frac{1}{N_p}\sum_{j=1}^{N_p} eff_j \cdot s_j \tag{26}$$

where s_j is the current standard derivation of the jth parameter.

The adaptive sequential pCN-MCMC starts with a random (or expert-guess) tuning-parameter combination β_0,κ_0. Then, the derivatives of the objective functions are approximated numerically by

$$\frac{\partial f}{\partial \beta} = \frac{f(\beta_k^{+},\kappa_k) - f(\beta_k^{-},\kappa_k)}{\beta_k^{+} - \beta_k^{-}} \tag{27}$$

$$\frac{\partial f}{\partial \kappa} = \frac{f(\beta_k,\kappa_k^{+}) - f(\beta_k,\kappa_k^{-})}{\kappa_k^{+} - \kappa_k^{-}} \tag{28}$$

with

$$\beta_k^{+} = \beta_k \cdot \delta_\beta, \;\; \beta_k^{-} = \beta_k \cdot \frac{1}{\delta_\beta}, \tag{29}$$

$$\kappa_k^+ = \kappa_k \cdot \delta_\beta, \; \kappa_k^- = \kappa_k \cdot \frac{1}{\delta_\beta}, \tag{30}$$

where $\delta_\beta = \delta_\kappa = \sqrt{2}$ in our implementation. This selection of evaluation points leads to an equal spacing of evaluation points in the log-space. Next, the algorithm moves a predefined distance (in the loglog-space) toward the steepest descent (here: rise, because we maximize the objective function) and restarts evaluating the tuning-parameters.

Two things are important here. First, we use the loglog-space because the results (Figures 3 and 4) suggest that the efficiency does not have sudden jumps in the loglog-space which makes the optimization easy. Optimizing in the "normal" space would lead to a more complex optimization problem due to a more complex structure of good values (banana shaped instead of a straight line) and high derivatives for small κ and β. Second, the predefined distance is important because the objective function is stochastic, that is, starting it twice with the same parameters will not lead to the same result f. Not predefining the distance (which is done in vanilla steepest descent methods) leads to some, randomly happening, high derivatives that prevent convergence of the algorithm.

4. Testing Cases and Implementation

4.1. Testing Procedure

We test our method by inferring the hydraulic conductivity of a confined aquifer based on measurements of hydraulic heads in a fully saturated, steady state, 2-D groundwater flow model. The data for inversion are generated synthetically with the same model as used for inversion. We are interested in several different cases: First, we test our method in a coarse-grid resolution [50×50] cells. Here, we systematically test tuning-parameter (β,κ) combinations to find the optimal parameter combination. Further, we developed the adaptive sequential pCN-MCMC in this case.

Second, we used the same reference solution with more informative measurement data and conducted the same systematic testing of tuning-parameters as in test Case 1. We test the adaptive sequential pCN-MCMC on this new test case on which it was not developed. This enables us to make (more or less) general statements about the performance of this tool. After that, we try different variants of the original model to test our algorithm in different conditions and at higher resolutions.

In all test cases, we run the sequential pCN-MCMC methods with 2 million samples of which we save every 200th sample. We discard the first half of each run due to burn-in and calculate the metrics presented in Section 2.7 based on the second half. Hence, each metric is calculated using 5,000 samples. This is done three times for each tuning-parameter combination in test case 1–4 and we report the mean value in the following. In test Case 5, we test each adaptive method (adaptive sequential Gibbs, adaptive pCN-MCMC, adaptive sequential pCN-MCMC) three times and report the mean values of these runs. Here, the adaptive sequential Gibbs (or adaptive sequential pCN-MCMC) corresponds to the case where we optimize κ (or β) as proposed in Section 3.1 and set $\beta = 1$ (or $\kappa = 1$).

To evaluate the KL-divergence, a reference solution is calculated using the best tuning-parameter combination of each test case and running the sequential pCN-MCMC for 10 million samples. We save every 200th sample and remove the first million samples as burn-in.

4.2. Description of Test Cases

4.2.1. Base Case

We consider an artificial steady state groundwater flow in a confined aquifer test case as proposed in Xu et al. (2020). It has a size of $5{,}000 \times 5{,}000[m]$ with a constant depth of 50 m and is discretized into 50×50 cells as shown in Figure 2.

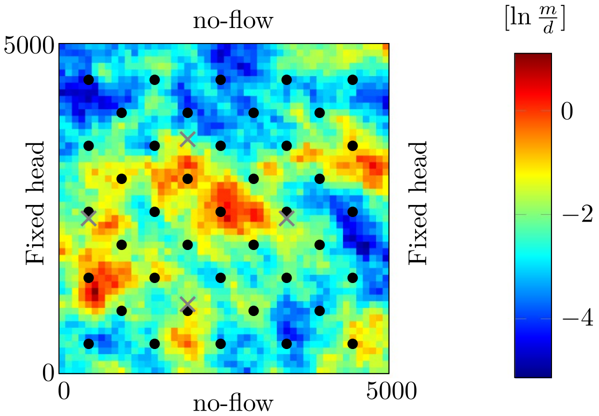


Figure 2. Log-conductivity field of the synthetic reference aquifer with a fixed head boundary condition on the left side and right side, no-flow boundary conditions at top and bottom and groundwater extraction wells marked with gray crosses. The positions of the measurement wells are marked in black.

We assume a multi-Gaussian prior model with mean $-2.5\left[\frac{m}{d}\right]$ and variance equal to 1. Further, we assume an anisotropic exponential variogram with lengthscale parameters [1,500, 2,000] rotated by 135°. The higher value of the lengthscale is pointing from the bottom left to the top right (see Figure 2).

We assume no-flow boundary conditions at the top and bottom boundary, a fixed head boundary condition with $h = 20$ m at the left and $h = 0$ m at the right side. Further, we assume four groundwater extraction wells as shown in Table 1.

Figure 2 shows the hydraulic conductivity distribution of the artificial true aquifer and the 41 measurement locations marked in black. We corrupt each of the 41 simulated (with the hydraulic conductivity of the artificial true aquifer) head values with a variate drawn at random from a zero-mean normal distribution with variance of 0.05 [m] to obtain synthetic data for the inversion

The flow in the domain can be described by the saturated groundwater flow equation

$$\nabla \cdot \left[K(x,y)\nabla h(x,y,t)\right] = \eta(x,y), \tag{31}$$

where K is the isotropic hydraulic conductivity, h is the hydraulic head and η encapsulates all source and sink terms. We solve the equation using the flow solver described in Nowak (2005) and Nowak et al. (2008) numerically.

4.2.2. Test Case 2

The sole difference between test Case 2 and the base case is that a standard derivation of the measurement error of 0.02 m is used. This leads to a more likelihood-dominated Bayesian inverse problem. Hence, slightly changing parameters results in higher differences in the corresponding likelihoods. This leads to a smaller acceptance rate which is dependent on the quotient of subsequent likelihoods. To keep a constant acceptance rate, a smaller jump width is needed. Hence, we expect the optimal κ and β to decrease for a more likelihood-dominated posterior.

4.2.3. Test Case 3

The only difference between test Case 3 and the base case is that only 16 instead of 41 measurement positions were used. This leads to a less likelihood-dominated Bayesian inverse problem and reverses the variation done in test Case 2.

4.2.4. Test Case 4

Test Case 4 has different reference solution with a Matern covariance with $\nu = 2.5$ and isotropic lengthscale parameter $\lambda = 1{,}000$. All other parameters are identical to the base case setup. We do this variation to test the influence of the prior covariance structure on our proposed method.

Table 1
Position and Pumping Strength of Groundwater Extraction Wells

Position x [m]	Position y [m]	Pump strength $[\frac{m^3}{d}]$
500	2,350	120
3,500	2,350	70
2,000	3,550	90
2,000	1,050	90

4.2.5. Test Case 5

Test Case 5 uses a refined reference solution with a [100,100] discretization grid. Here, the higher discretization in the inversion makes the problem numerically more expensive, serving to demonstrate the efficiency and applicability of our method.

4.2.6. Test Case 6

Test Case 6 uses the artificial true aquifer of the base case. However, instead of inferring the hydraulic conductivity with head measurements, we infer the hydraulic conductivity with concentration measurements.

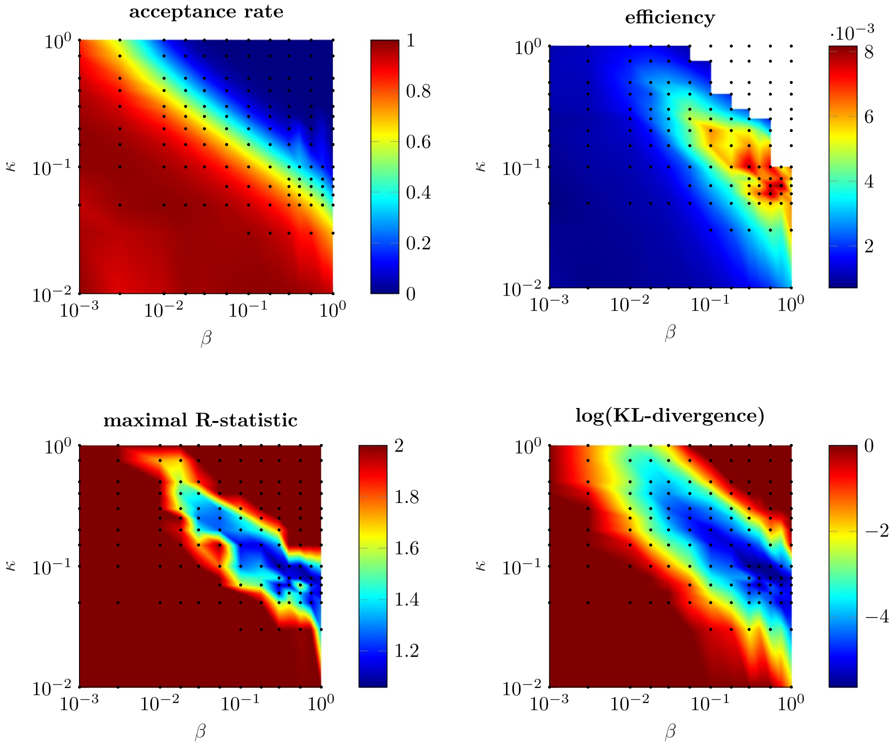


Figure 3. Results of base case. The sequential pCN-MCMC was tested with all combinations marked by black points. All in-between values are interpolated. The efficiency was not calculated in the top right corner (white area) because the MCMC did not finish the burn-in during runtime. The pCN-special case is at the top with $\kappa = 1$ and the Gibbs special case is at the right with $\beta = 1$. The acceptance rate shows no unique optimum ($\alpha = 0.234$), but rather a line of optimal tuning-parameters combinations. The efficiency, the R-statistic and the KL-divergence indicate similar optima. The optimal efficiency is at $\beta = 0.75$ and $\kappa = 0.07$.

We assume a dissolved, conservative, nonsorbing tracer transported by the advection dispersion equation in steady state flow

$$\nabla \cdot (\boldsymbol{v}c - \boldsymbol{D}\nabla c) = 0 \tag{32}$$

with the seepage velocity $\boldsymbol{v}$ and the local dispersion tensor $\boldsymbol{D}$. We assume that some concentration c is constantly entered in the center third of the left boundary by setting the boundary condition to $\frac{c}{c_0} = 1$ there, where c_0 is for example, a solubility limit. Then, we calculate the steady state solution of Equation 32 under time-constant boundary conditions and measure the concentration at the 5 upmost right measurement locations shown in Figure 2. Here, we assume a standard derivation of the measurement error equal to 0.05. This test case shows the influence that different measurements types have to our results. We use the flow and transport solver described in Nowak (2005) and Nowak et al. (2008) to solve the equations numerically. This solver uses a streamline upwind Petrov Galerkin finite element method with bilinear elements on a regular grid with 50 × 50 cells and cell-wise constant K values.

5. Results

5.1. Base Case

The acceptance rate of the MCMC, the efficiency, the R-statistic and the (log of the) KL-divergence to a reference solution are visualized in Figure 3. As discussed in Section 2.7, we aim for an R-statistic equals 1, a high efficiency and a low log KL conductivity which corresponds to an acceptance rate equals 23%. We focus on two things in this plot. First, we see that all metrics have a similar appearance. Hence, a tuning-parameter combination that performs well in one metric also performs well in the other two metrics

Table 2
Test Metrics of Algorithms With Optimal Tuning-Parameters

	Efficiency			Maximal R-statistic			KL-divergence			Optimal	
	se. pCN	pCN	Gibbs	se. pCN	pCN	Gibbs	se. pCN	pCN	Gibbs	β	κ
Base case	0.0082	0.0016	0.0065	1.0572	2.0697	1.1222	0.0036	0.0660	0.0066	0.75	0.07
Case 2	0.0061	0.0011	0.0061	1.2888	5.4102	1.9264	0.0326	0.5923	0.0419	1	0.06
Case 3	0.0087	0.0029	0.0063	1.0521	1.2661	1.0854	0.0051	0.0212	0.0072	0.3	0.1
Case 4	0.0019	0.0012	0.0018	1.8374	3.9433	2.5057	0.0694	0.6710	0.1103	0.3	0.1
Case 5[a]	0.0065	0.0019	0.0065	1.1315	2.25	1.1315	—	—	—	1	0.0579
Case 6[a]	0.1711	0.1711	0.0264	1.002	1.002	1.024	—	—	—	0.1741	1

[a] No systematic testing was performed. Optimal tuning-parameters were found using the adaptive sequential pCN-MCMC.

and vice versa. Second, the optimal tuning-parameter combination (β,κ) is around $\beta = 0.75$ and $\kappa = 0.07$ for all norms. Hence, neither the pCN ($\kappa = 1$) nor the sequential Gibbs special case ($\beta = 1$) is optimal. The sequential pCN-MCMC has a better performance than the special cases. Further, the efficiency plot (or table 2) indicates a speedup of approximately 5.1 over pCN-MCMC and of 1.3 over sequential Gibbs sampling.

5.2. Test Case 2

Using the second test case, we visualize the acceptance rate of the MCMC, the efficiency, the R-statistic and the KL-divergence to a reference solution in Figure 4. As discussed in Section 5.2, we expect smaller optimal tuning-parameters compared to test Case 1. Comparing Figures 3 and 4, we see that our expectations are partly met. The light blue line of acceptance rates of approximately 40% is moving to the bottom left, to smaller κ and β, as expected. However, the optimal values change from $\beta = 0.75$ and $\kappa = 0.07$ in the first test case to $\beta = 1$ and $\kappa = 0.06$ in the second case. Hence, the optimal tuning-parameters tend more toward the Gibbs approach (no pCN correlation at $\beta = 1$, instead a smaller window). In fact, we find that the sequential Gibbs approach is as good as the sequential pCN-MCMC approach because the optimal parameter β equals 1.

5.3. Further Test Cases

Next, we test our algorithm with fewer measurement locations (Case 3), with a Matern variogram model (Case 4) and with a finer discretization (Case 5) as summarized in Table 2. In short, the results indicate that the sequential pCN-MCMC has at least the same performance as the Gibbs or pCN-MCMC approach. We achieve a speedup (measured by the ratio of efficiencies) of 1–5.5 over pCN and of 1–6.5 over Gibbs.

In test cases 1–4, structured testing as shown in Sections 5.1 and 5.2 was performed. For the high-dimensional test Case 5 and the transport test Case 6 this testing procedure was too computationally expensive. Hence, the previous tested adaptive sequential pCN-MCMC was used to find the best parameter distribution.

5.4. Adaptive Sequential pCN-MCMC

Gelman et al. (1996) stated that a wide range of tuning-parameters are satisfactory (close to optimal) for MCMCs with one tuning-parameter. Figure 3 shows that this holds for this two tuning-parameter MCMC as well. The efficiency, the R-statistic and the KL-divergence have a broad area of near-optimal tuning-parameters. Hence, the adaptive sequential pCN-MCMC only needs to find some point in this good area to choose near-optimal parameters. Figure 5 shows five paths of the adaptive sequential pCN-MCMC during burn-in with random start tuning-parameters for the first and second test cases. Each path consists of all tested

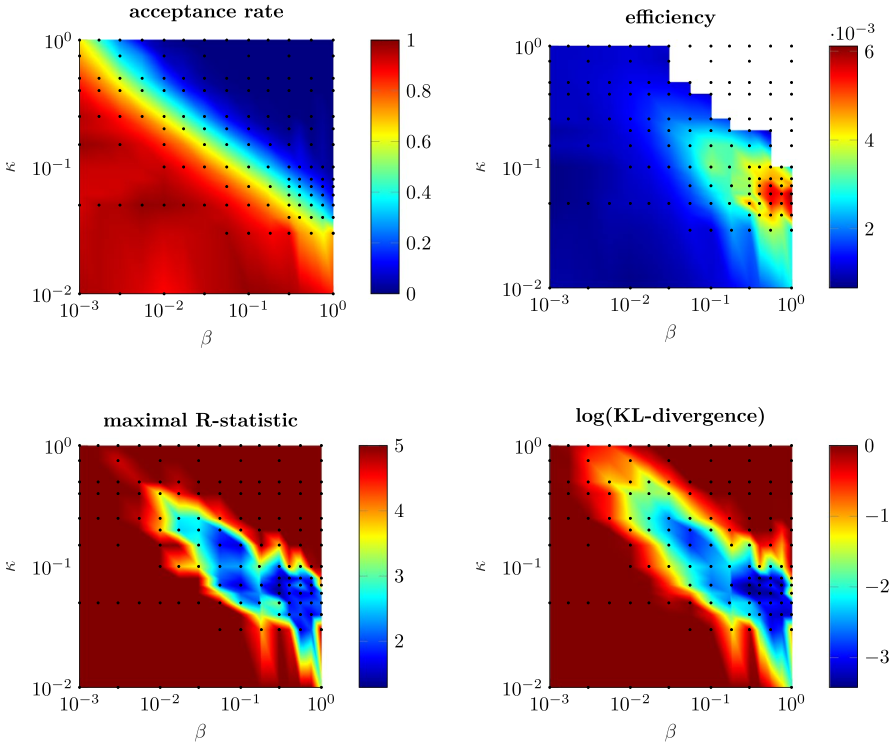


Figure 4. Results of test Case 2. The sequential pCN-MCMC was tested with all combinations marked by black points. All in-between values are interpolated. The efficiency was not calculated in the top right corner (white area) because the MCMC did not finish the burn-in during runtime. The pCN-special case is at the top with $\kappa = 1$ and the Gibbs special case is at the right with $\beta = 1$. The acceptance rate shows no unique optimum ($\alpha = 0.234$), but rather a line of optimal tuning-parameters combinations. The efficiency, the R-statistic and the KL-divergence indicate similar optima. The optimal efficiency is at $\beta = 1$ and $\kappa = 0.06$.

tuning-parameter combinations. Figure 5 shows that all paths converge to the targeted area of near-optimal tuning-parameters.

We note here, that many tuning-parameter optimization steps N_{hp} are needed to achieve these results. Using fewer MCMC steps per tuning-parameter iteration leads to less exact tuning-parameter tuning. Although the tuning is less exact with smaller N_{hp}, we find in additional experiments that the adaptive sequential pCN-MCMC converges to acceptable tuning-parameters with small N_{hp} due to broad high-efficiency areas.

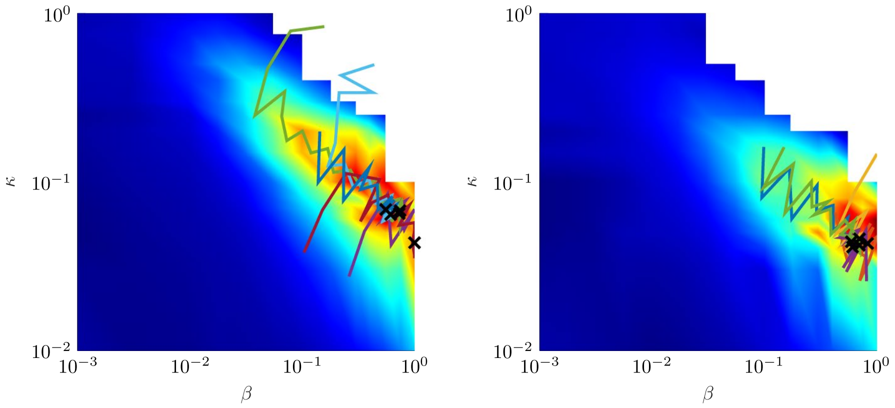


Figure 5. Convergence of adaptive sequential pCN-MCMC in test Case 1 (left) and test Case 2 (right). The chosen parameter for production is marked with a cross. The efficiency is shown in the background.

The needed size of N_{hp} depends on the amount of measurement information. Having more information leads, even for near-optimal tuning-parameters, to a smaller step size (of the MCMC) and lower efficiency. Hence, we need more samples for good MCMC results and simultaneously for each tuning-parameter tuning iteration.

6. Discussion

6.1. Global Versus Local Proposal Steps

The results in Table 2 indicate that the pure (local) Gibbs approach is superior to the pure (global) pCN approach when used with head measurements. We did further testing using a simple Kriging (measuring the parameters directly) example and found that the Gibbs approach is superior to the pCN approach in that case as well. In transport scenarios (test Case 6), the (global) pCN approach is superior to the (local) Gibbs approach.

We explain this behavior in the following way: direct (and head measurements) typically yield us local (or relatively local) information of the aquifer (Rubin, 2003). It only lets us infer the hydraulic conductivity in a small area around the measurement location because the influence of hydraulic conductivity on the measurement decreases rapidly with distance (Rubin, 2003). This leads to the conclusion, that measurements with localized information (head, direct measurements) work better with local updating schemes (Gibbs), whereas measurements with global information (transport) work better with global updating schemes (pCN). Note, that we only tested these algorithms on a few geostatistical problems and encourage researchers to compare global and local proposal steps and endorse or oppose our findings.

6.2. Limits of Sequential pCN-MCMC

In our test cases, the sequential pCN-MCMC and the sequential Gibbs approach have higher efficiencies than the pCN-MCMC. However, this speedup comes at the increased cost of the proposal step. Computing the conditional probability (Equations 18 and 19) is time consuming due to the computation of $\mathbf{\Sigma}_{22}^{-1}$. In most applications, the forward simulation (i.e., the calculation of the likelihood) is much more expensive than the inversion of the matrix $\mathbf{\Sigma}_{22}$ and the time difference in the proposal step can be neglected. However, with simple forwards problems, this might make the pCN approach a viable alternative to the sequential pCN-MCMC because all norms discussed in this paper neglect this time difference.

Fu and Gómez-Hernández (2008) discuss different schemes on how the conditional sampling can be performed without the need to compute $\mathbf{\Sigma}_{22}^{-1}$. The downside of these schemes is, that they do not sample from $N(\tilde{\boldsymbol{\mu}}_1, \tilde{\mathbf{\Sigma}}_{11})$ directly but from some approximation of it. On regular, equispaced grids, FFT-related methods and sparse linear algebra methods (Fritz et al., 2009; Nowak & Litvinenko, 2013) offer exact and very fast solutions.

The sequential pCN-MCMC is not designed to handle multimodal posteriors. However, applying parallel tempering approaches (e.g., Laloy et al., 2016) can solve this challenge. They can be applied straightforward to our method but come with one downside. The adaptive sequential pCN-MCMC will not work in a parallel tempering setup because the efficiency depends on the autocorrelation. In parallel tempering, the autocorrelation is dominated by between-chain swaps and hence will not be a good estimator for the performance of MCMC. Finding another way to tune the tuning-parameters during burn-in will be the big challenge in generalizing the sequential pCN-MCMC to parallel tempering.

6.3. Limits of Optimizing the Acceptance Rate

Our results suggest that acceptance rate values of 10–60%, (Case 1: 37%, Case 2: 10%, Case 3: 59%, Case 4: 57%) are optimal. This is in conflict with the literature, especially Gelman et al. (1996), stating that acceptance rates equal 23.4% are optimal for multi-Gaussian settings. The reason for this is the synthetic setting of Gelman et al. (1996). As a consequence, when using the acceptance rate α for optimizing the jump width, researchers should be aware that $\alpha = 23.4\%$ is not always optimal. Apart from that, we endorse Gelman et al. (1996) that a wide area of acceptance rates leads to near-optimal results. Hence, the error by tuning for the wrong acceptance rate might be neglectable. We cannot give a solution to this challenge but only point out that the suggested optimal acceptance rate of 23.4% might not be the one you should always aim for.

6.4. Transfer to Multipoint Geostatistics

The idea of building a hybrid between global and local jumps in parameter distributions can be applied to training image-based sampling methods in multipoint geostatistics as well. Both global (resampling a percentage of parameters scattered over the domain, e.g., Mariethoz et al., 2010) and local approaches (resampling a box of parameters, e.g., Hansen et al., 2012) exist and a combination might speed up the convergence for training image-based approaches as well. Thereby, a hybrid method should resample a higher percentage of scattered parameters in a larger box.

7. Conclusion

We presented the sequential pCN-MCMC approach, a combination of the sequential Gibbs and the pCN-MCMC approach. All approaches have discretization-independent convergence rates, which means that their efficiency does not decrease with higher resolutions of the inferred parameter field. We show that the (local) Gibbs approach is better for local measurements (head measurements) and the (global) pCN approach is better for global measurements (tracer experiments).

The presented sequential pCN-MCMC can choose the best trade-off between the two existing methods. To do so, it has two tuning-parameters, the parameter β of the pCN approach and κ of the Gibbs approach. Setting either one of them to 1 makes the algorithm collapse to either the pCN or the Gibbs approach. We show that the proposed method is as efficient or more efficient than the sequential Gibbs and pCN-MCMC methods by testing all possible tuning-parameters of the sequential pCN-MCMC method. To be more precise, a speedup of 1–5.5 over the pCN-MCMC method and 1–6.5 over the sequential Gibbs method is observed.

Using more than one tuning-parameter has the downside that finding the optimal tuning-parameters is difficult. We presented the adaptive sequential pCN-MCMC to find good tuning-parameters during the burn-in of the algorithm. This work can be extended to parallel tempering easily. However, the presented approach of finding the optimal tuning-parameters during burn-in needs to be adapted to fit the challenges of multiple chains.

For practical applications, the presented adaptive sequential pCN-MCMC is a fast and easy to handle MCMC method for Bayesian inversion. It requires no manual adjustment of tuning-parameters because the method optimizes them automatically during burn-in. Further, it is at least as fast and usually faster than the state-of-the-art alternatives on all tested data types. Hence, we recommend using the open-source implementation of our method on your inversion problem.

APPENDIX A

The pseudocodes of the algorithms discussed in this work are shown in the following.

Algorithm 1: sequential Gibbs sampling

```
Input      : Prior probability density function $P(\boldsymbol{\theta}), \boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
             Likelihood function $L(\boldsymbol{\theta})$
Output     : Samples $\boldsymbol{\theta}^{(i)}$ from posterior distribution $\pi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})L(\boldsymbol{\theta})$
Parameter: $\kappa \in (0, 1]$
Set $i = 0$
Draw $\boldsymbol{\theta}^{(0)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from prior
while true do
    $[\mathbf{M}^{(i)}, q^{(i)}] = \text{GetBoxAlgorithm}(\kappa)$
    $\begin{bmatrix} \boldsymbol{\theta}_1^{(i)} \\ \boldsymbol{\theta}_2^{(i)} \end{bmatrix} = \mathbf{M}^{(i)}\boldsymbol{\theta}^{(i)}, \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \mathbf{M}^{(i)}\boldsymbol{\mu},$
    $\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \mathbf{M}^{(i)}\boldsymbol{\Sigma}\mathbf{M}^{(i)^T}$
    $\widetilde{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\mu}_2)$
    $\widetilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$
    Propose $\widetilde{\boldsymbol{\theta}}_1^{(i)} \sim N(\widetilde{\boldsymbol{\mu}}_1, \widetilde{\boldsymbol{\Sigma}}_{11})$
    Set $\widetilde{\boldsymbol{\theta}}^{(i)} = \mathbf{M}^{(i)^T} \begin{pmatrix} \widetilde{\boldsymbol{\theta}}_1^{(i)} \\ \boldsymbol{\theta}_2^{(i)} \end{pmatrix}$
    Compute $\alpha(\boldsymbol{\theta}^{(i)}, \widetilde{\boldsymbol{\theta}}^{(i)}) = min\left[\frac{L(\widetilde{\boldsymbol{\theta}}^{(i)})}{L(\boldsymbol{\theta}^{(i)})}, 1\right]$
    Draw $r \sim U(0, 1)$
    if $r \leq \alpha$ (with probability $\alpha$) then
        $\boldsymbol{\theta}^{(i+1)} = \widetilde{\boldsymbol{\theta}}^{(i)}$
    else
        $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$
    end
    $i = i + 1$
end
```

Algorithm 2: get box algorithm

```
Input      : vectors of grid points positions $\boldsymbol{x}, \boldsymbol{y}$
             lengths of domain $l_x, l_y$
             number of grid points $N_p$
Output     : permutation matrix $\mathbf{M}$
             $q$, the size of all vectors with subscript 1 (e.g. $\boldsymbol{\theta}_1$ )
Parameter: $\kappa \in (0, 1]$
Notation   : $\boldsymbol{x}(k)$ is the k-th element of vector $\boldsymbol{x}$
Draw midpoint $x^*, y^* \sim U(0, 1)$
$\mathbf{M} = \mathbf{0}$ (of size $N_p \times N_p$)
$q = 0,\ r = 0,\ k = 0$
while $k < N_P$ do
    $k = k + 1$
    if $\left|\frac{\boldsymbol{x}(k) - x^*}{l_x}\right| \leq \kappa$ and $\left|\frac{\boldsymbol{y}(k) - y^*}{l_y}\right| \leq \kappa$ then
        $q = q + 1$
        $\mathbf{M}(q, k) = 1$
    else
        $r = r + 1$
        $\mathbf{M}(N_p - r + 1, k) = 1$
    end
end
```

Algorithm 3: pCN-MCMC

```
Input      : Prior probability density function $P(\boldsymbol{\theta}), \boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
             Likelihood function $L(\boldsymbol{\theta})$
Output     : Samples $\boldsymbol{\theta}^{(i)}$ from posterior distribution $\pi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})L(\boldsymbol{\theta})$
Parameter: $\beta \in (0, 1]$
Set $i = 0$
Draw $\boldsymbol{\theta}^{(0)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from prior
while *true* do
    Propose $\widetilde{\boldsymbol{\theta}}^{(i)} = \sqrt{(1-\beta^2)}\left(\boldsymbol{\theta}^{(i)} - \boldsymbol{\mu}\right) + \beta\xi^{(i)} + \boldsymbol{\mu}, \xi^{(i)} \sim N(0, \boldsymbol{\Sigma})$
    Compute $\alpha(\boldsymbol{\theta}^{(i)}, \widetilde{\boldsymbol{\theta}}^{(i)}) = min\left[\frac{L(\widetilde{\boldsymbol{\theta}}^{(i)})}{L(\boldsymbol{\theta}^{(i)})}, 1\right]$
    Draw $r \sim U(0, 1)$
    if $r \leq \alpha$ (with probability $\alpha$) then
        $\boldsymbol{\theta}^{(i+1)} = \widetilde{\boldsymbol{\theta}}^{(i)}$
    else
        $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$
    end
    $i = i + 1$
end
```

```
Algorithm 4: sequential pNC-MCMC
Input      : Prior probability density function $P(\boldsymbol{\theta}), \boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
             Likelihood function $L(\boldsymbol{\theta})$
Output     : Samples $\boldsymbol{\theta}^{(i)}$ from posterior distribution $\pi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})L(\boldsymbol{\theta})$
Parameter: $\kappa \in (0, 1]$
             $\beta \in (0, 1]$
Set $i = 0$
Draw $\boldsymbol{\theta}^{(0)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from prior
while true do
    $[\mathbf{M}^{(i)}, q^{(i)}] = \text{GetBoxAlgorithm}(\kappa)$
    $\begin{bmatrix} \boldsymbol{\theta}_1^{(i)} \\ \boldsymbol{\theta}_2^{(i)} \end{bmatrix} = \mathbf{M}^{(i)}\boldsymbol{\theta}^{(i)}, \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \mathbf{M}^{(i)}\boldsymbol{\mu},$
    $\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \mathbf{M}^{(i)}\boldsymbol{\Sigma}\mathbf{M}^{(i)T}$
    $\tilde{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\theta}_2^{(i)} - \boldsymbol{\mu}_2)$
    $\tilde{\boldsymbol{\Sigma}}_{11} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$
    Propose $\tilde{\boldsymbol{\theta}}_1^{(i)} = \sqrt{(1-\beta^2)}\left(\boldsymbol{\theta}_1^{(i)} - \tilde{\boldsymbol{\mu}}_1\right) + \beta\xi^{(i)} + \tilde{\boldsymbol{\mu}}_1, \xi^{(i)} \sim N(0, \tilde{\boldsymbol{\Sigma}}_{11})$
    Set $\tilde{\boldsymbol{\theta}}^{(i)} = \mathbf{M}^{(i)T}\begin{pmatrix} \tilde{\boldsymbol{\theta}}_1^{(i)} \\ \boldsymbol{\theta}_2^{(i)} \end{pmatrix}$
    Compute $\alpha(\boldsymbol{\theta}^{(i)}, \tilde{\boldsymbol{\theta}}^{(i)}) = min\left[\frac{L(\tilde{\boldsymbol{\theta}}^{(i)})}{L(\boldsymbol{\theta}^{(i)})}, 1\right]$
    Draw $r \sim U(0, 1)$
    if $r \leq \alpha$ (with probability $\alpha$) then
        $\boldsymbol{\theta}^{(i+1)} = \tilde{\boldsymbol{\theta}}^{(i)}$
    else
        $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$
    end
    $i = i + 1$
end
```

Data Availability Statement

The implementation of the adaptive sequential pCN-MCMC is available at https://bitbucket.org/Reuschen/sequential-pcn-mcmc.

Acknowledgments
The authors thank Niklas Linde who was a valuable discussion partner with great ideas regarding the transfer to multipoint geostatistics and Sinan Xiao for revising the Manuscript. We thank the three anonymous reviewers for their constructive comments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 327154368—SFB 1313.

References

Beskos, A., Roberts, G., Stuart, A., & Voss, J. (2008). MCMC methods for diffusion bridges. *Stochastics and Dynamics*, *8*(3), 319–350. https://doi.org/10.1142/s0219493708002378

Betancourt, M. (2017). *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv preprint, arXiv:1701.02434.

Butera, I., & Soffia, C. (2017). Cokriging transmissivity, head and trajectory data for transmissivity and solute path estimation. *Ground Water*, *55*(3), 362–374. https://doi.org/10.1111/gwat.12483

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, *49*(4), 327–335. https://doi.org/10.1080/00031305.1995.10476177

Cotter, S. L., Roberts, G. O., Stuart, A. M., & White, D. (2013). MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science*, *28*(3), 424–446. https://doi.org/10.1214/13-sts421

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*(434), 883–904. https://doi.org/10.1080/01621459.1996.10476956

Evensen, G. (2009). *Data assimilation: The ensemble Kalman filter*. Springer Science & Business Media.

Ezzedine, S., & Rubin, Y. (1996). A geostatistical approach to the conditional estimation of spatially distributed solute concentration and notes on the use of tracer data in the inverse problem. *Water Resources Research*, *32*(4), 853–861. https://doi.org/10.1029/95WR02285

Fritz, J., Neuweiler, I., & Nowak, W. (2009). Application of FFT-based algorithms for large-scale universal kriging problems. *Mathematical Geosciences*, *41*(5), 509–533. https://doi.org/10.1007/s11004-009-9220-x

Fu, J., & Gómez-Hernández, J. J. (2008). Preserving spatial structure for inverse stochastic simulation using blocking Markov chain Monte Carlo method. *Inverse Problems in Science and Engineering*, *16*(7), 865–884. https://doi.org/10.1080/17415970802015781

Fu, J., & Gómez-Hernández, J. J. (2009a). A blocking Markov chain Monte Carlo method for inverse stochastic hydrogeological modeling. *Mathematical Geosciences*, *41*(2), 105–128. https://doi.org/10.1007/s11004-008-9206-0

Fu, J., & Gómez-Hernández, J. J. (2009b). Uncertainty assessment and data worth in groundwater flow and mass transport modeling using a blocking Markov chain Monte Carlo method. *Journal of Hydrology*, *364*(3–4), 328–341. https://doi.org/10.1016/j.jhydrol.2008.11.01410.1016/j.jhydrol.2008.11.014

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

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

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

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741. https://doi.org/10.1109/tpami.1984.4767596

Gutjahr, A., Bullard, B., Hatch, S., & Hughson, L. (1994). Joint conditional simulations and the spectral approach for flow modeling. *Stochastic Hydrology and Hydraulics*, *8*(1), 79–108. https://doi.org/10.1007/bf01581391

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

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

Jardani, A., Revil, A., & Dupont, J. P. (2013). Stochastic joint inversion of hydrogeophysical data for salt tracer test monitoring and hydraulic conductivity imaging. *Advances in Water Resources*, *52*, 62–77. https://doi.org/10.1016/j.advwatres.2012.08.00510.1016/j.advwatres.2012.08.005

Kitanidis, P. K. (1997). *Introduction to geostatistics: Applications in hydrogeology*. Cambridge University Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Laloy, E., Linde, N., Jacques, D., & 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. https://doi.org/10.1016/j.advwatres.2016.02.008

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

Mariethoz, G., Renard, P., & Caers, J. (2010). Bayesian inverse problem and optimization with iterative spatial resampling. *Water Resources Research*, *46*, 1–17. https://doi.org/10.1029/2010WR009274

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092. https://doi.org/10.1063/1.1699114

Mondal, A., Mallick, B., Efendiev, Y., & Datta-Gupta, A. (2014). Bayesian uncertainty quantification for subsurface inversion using a Multiscale Hierarchical Model. *Technometrics*, *56*(3), 381–392. https://doi.org/10.1080/00401706.2013.838190

Nowak, W., & (2005). *Geostatistical methods for the identification of flow and transport parameters in the subsurface* (Unpublished doctoral dissertation). https://doi.org/10.18419/opus-201

Nowak, W., De Barros, F. P., & Rubin, Y. (2010). Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain. *Water Resources Research*, *46*, W03535. https://doi.org/10.1029/2009WR008312

Nowak, W., & Litvinenko, A. (2013). Kriging and spatial design accelerated by orders of magnitude: Combining low-rank covariance approximations with FFT-techniques. *Mathematical Geosciences*, *45*(4), 411–435. https://doi.org/10.1007/s11004-013-9453-6

Nowak, W., Schwede, R. L., Cirpka, O. A., & Neuweiler, I. (2008). Probability density functions of hydraulic head and velocity in three-dimensional heterogeneous porous media. *Water Resources Research*, *44*, W08452. https://doi.org/10.1029/2007WR006383

Qian, S. S., Stow, C. A., & Borsuk, M. E. (2003). On Monte Carlo methods for Bayesian inference. *Ecological Modelling*, *159*(2), 269–277. https://doi.org/10.1016/s0304-3800(02)00299-5

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

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

Reuschen, S., Xu, T., & Nowak, W. (2020). Bayesian inversion of hierarchical geostatistical models using a parallel-tempering sequential Gibbs MCMC. *Advances in Water Resources*, *141*, 103614. https://doi.org/10.1016/j.advwatres.2020.103614

Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods* (Vol. 42). Springer Science & Business Media. https://doi.org/10.2307/1270959

Roberts, G. O., & Rosenthal, J. S. (2002). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, *16*(4), 351–367. https://doi.org/10.1214/ss/1015346320

Roy, V. (2020). Convergence diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and its Application*, *7*, 387–412. https://doi.org/10.1146/annurev-statistics-031219-041300

Rubin, Y. (2003). *Applied stochastic hydrogeology*. Oxford University Press.

Smith, A., & Roberts, G. (1993). Bayesian computation via the Gibbs Sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society Part B*, *55*(1), 3–23. https://doi.org/10.1111/j.2517-6161.1993.tb01466.x

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics.

Xu, T., Reuschen, S., Nowak, W., & Hendricks Franssen, H. J. (2020). Preconditioned Crank-Nicolson Markov Chain Monte Carlo coupled with parallel tempering: An efficient method for Bayesian inversion of multi-Gaussian log-hydraulic conductivity fields. *Water Resources Research*, *56*, e2020WR027110. https://doi.org/10.1029/2020WR027110

Zimmerman, D. A., de Marsily, G., Gotway, C. A., Marietta, M. G., Axness, C. L., Beauheim, R. L., et al. (1998). A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. *Water Resources Research*, *34*(6), 1373–1413. https://doi.org/10.1029/98WR00003

B Contribution 2: Preconditioned Crank-Nicolson Markov chain Monte Carlo coupled with parallel tempering: an efficient method for Bayesian inversion of multi-Gaussian log-hydraulic conductivity fields

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR027110

Preconditioned Crank-Nicolson Markov Chain Monte Carlo Coupled With Parallel Tempering: An Efficient Method for Bayesian Inversion of Multi-Gaussian Log-Hydraulic Conductivity Fields

Teng Xu[1,2], Sebastian Reuschen[2], Wolfgang Nowak[2], and Harrie-Jan Hendricks Franssen[3]

[1]State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China, [2]Department of Stochastic Simulation and Safety Research for Hydrosystems (IWS/LS3), University of Stuttgart, Stuttgart, Germany, [3]Institute of Bio- and Geosciences (IBG-3): Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany

Key Points:

- We proposed an efficient MCMC method called pCN-PT for Bayesian inversion of multi-Gaussian log-hydraulic conductivity fields
- We demonstrated that pCN-PT can estimate and conditionally sample multi-Gaussian random fields in high-dimensional linear and nonlinear problems
- Parallel tempering can improve the performance of pCN-MCMC

Correspondence to:
T. Xu,
teng.xu@hhu.edu.cn

Citation:
Xu, T., Reuschen, S., Nowak, W., & Hendricks Franssen, H.-J. (2020). Preconditioned Crank-Nicolson Markov chain Monte Carlo coupled with parallel tempering: An efficient method for Bayesian inversion of multi-Gaussian log-hydraulic conductivity fields. *Water Resources Research*, *56*, e2020WR027110. https://doi.org/10.1029/2020WR027110

Received 12 JAN 2020
Accepted 27 JUL 2020
Accepted article online 3 AUG 2020

©2020. American Geophysical Union. All Rights Reserved.

Abstract

Geostatistical inversion with quantified uncertainty for nonlinear problems requires techniques for providing conditional realizations of the random field of interest. Many first-order second-moment methods are being developed in this field, yet almost impossible to critically test them against high-accuracy reference solutions in high-dimensional and nonlinear problems. Our goal is to provide a high-accuracy reference solution algorithm. Preconditioned Crank-Nicolson Markov chain Monte Carlo (pCN-MCMC) has been proven to be more efficient in the inversion of multi-Gaussian random fields than traditional MCMC methods; however, it still has to take a long chain to converge to the stationary target distribution. Parallel tempering aims to sample by communicating between multiple parallel Markov chains at different temperatures. In this paper, we develop a new algorithm called pCN-PT. It combines the parallel tempering technique with pCN-MCMC to make the sampling more efficient, and hence converge to a stationary distribution faster. To demonstrate the high-accuracy reference character, we test the accuracy and efficiency of pCN-PT for estimating a multi-Gaussian log-hydraulic conductivity field with a relative high variance in three different problems: (1) in a high-dimensional, linear problem; (2) in a high-dimensional, nonlinear problem and with only few measurements; and (3) in a high-dimensional, nonlinear problem with sufficient measurements. This allows testing against (1) analytical solutions (kriging), (2) rejection sampling, and (3) pCN-MCMC in multiple, independent runs, respectively. The results demonstrate that pCN-PT is an asymptotically exact conditional sampler and is more efficient than pCN-MCMC in geostatistical inversion problems.

1. Introduction

Good predictions of subsurface dynamics, subsurface environmental risk assessment, and good design of remediation actions are hard to achieve without proper characterization of subsurface properties. In almost all applications, there are not enough direct (hard) data (e.g., hydraulic conductivity and porosity) to construct subsurface property fields. On the contrary, indirect (soft) data (e.g., piezometric head and contaminant concentration) are easier to collect. How to properly characterize subsurface properties from indirect data is still an important and active research topic in many disciplines, such as, petroleum engineering, mining engineering, meteorology, or hydrology. Methods to characterize the desired properties from the information contained in indirect data, while quantifying the remaining uncertainties, are called stochastic inverse modeling.

Many different stochastic inverse modeling approaches have been developed, such as the Representer method (e.g., Franssen et al., 2009; Valstar et al., 2004), the ensemble Kalman filter (Chen & Zhang, 2006; Hendricks Franssen & Kinzelbach, 2008; Li et al., 2011; Xu et al., 2013; Zhou et al., 2014), the ensemble smoother (e.g., Bailey et al., 2012; Crestani et al., 2013; Evensen & Van Leeuwen, 2000), the Pilot Points method (e.g., Alcolea et al., 2006; Christensen & Doherty, 2008; RamaRao et al., 1995), the inverse sequential simulation method (e.g., Xu & Gómez-Hernández, 2015a, 2015b), the sequential self-calibration (e.g., Franssen et al., 2003; Gómez-Hernánez et al., 1997; Wen et al., 2002), and the Markov chain Monte Carlo method (MCMC) (e.g., Fu & Gómez-Hernández, 2009; Laloy et al., 2013; Oliver et al., 1997). Among these methods, MCMC is commonly regarded as a classical but computationally expensive method, though it is good for dealing with complex construction of subsurface properties and nonlinearity (Saley et al., 2016). All other methods mentioned above rely on some sort of linearization or first-order, second-moment approximation. They are often much faster than MCMC but are only approximations that get more and more inaccurate and instable with increasing nonlinearity of the problems. Thus, for problems that require high accuracy and for critically testing the advancement of approximate methods, asymptotically exact methods such as MCMC are the only solution.

To handle the computational cost problem of MCMC, many advanced versions have been proposed in the literature. For example, Duane et al. (1987) proposed the Hybrid Monte Carlo algorithm (also called Hamiltonian Monte Carlo algorithm) that can exploit gradient information, and Saley et al. (2016) applied it for the characterization of hydraulic conductivity; Cui et al. (2011) developed an adaptive delayed acceptance Metropolis Hasting algorithm that was shown to greatly improve the computational efficiency in a geothermal application; Martin et al. (2012) presented a Stochastic Newton method to accelerate MCMC by building a local Gaussian approximation based on local gradient and Hessian information; Andrieu et al. (2010) proposed a combination of MCMC and sequential Monte Carlo methods, in which efficient proposal distributions are built by using sequential MCMC schemes; Scott et al. (2016) proposed a consensus MCMC algorithm (a scalable MCMC method), which runs separate MCMC samplers independently for each data subset on each machine, and averages individual MCMC samplers from each subset across machines. When we have high-performance computing available, the most direct method to accelerate MCMC is to design parallel architectures and run many MCMC chains in parallel (e.g., Angelino et al., 2014; Calderhead, 2014; Terenin et al., 2015).

A remaining key problem is the low efficiency obtained with MCMC in geostatistical inverse problems: Most samples proposed within MCMC are rejected because they are impossible against the (prior) geostatistical model. To overcome this problem, Cotter et al. (2013) proposed a preconditioned Crank-Nicolson Markov chain Monte Carlo method (pCN-MCMC) by combining the preconditioned Crank-Nicolson method with MCMC to make MCMC faster for Bayesian inversion of multi-Gaussian hydraulic conductivity fields. This algorithm was shown to be capable for Bayesian inversion of multi-Gaussian subsurface properties in many applications (e.g., Cotter et al., 2013; Hu et al., 2017; Iglesias et al., 2012). Still, when pCN-MCMC explores the target posterior distribution, it may still need a long chain and it is still difficult to explore a large potential model space since the MCMC simulation proceeds by local jumps in the vicinity of the current solutions (Robert et al., 2018). Parallel tempering can handle the dilemma with exploring large model spaces and can improve the efficiency of exploring the target posterior by exchange swaps between cold chains and hot chains. The hot chains mainly explore the solution and colder chains exploit the found high-likelihood regions (e.g., Earl & Deem, 2005; Geyer, 1991). Due to this merit, the technique has received a lot of attention and applications in recent years. For example, Laloy et al. (2016) merged parallel tempering with sequential geostatistical resampling to improve posterior exploration of subsurface categorical conductivity fields; Blatter et al. (2016) coupled the technique with a reversible-jump MCMC method for Bayesian inversion of 2-D models from airborne transient EM data; Bertrand et al. (2001) coupled it with MCMC to solve the magnetoencephalography inverse problem.

Parallel Tempering is a meta-algorithm that can be integrated into many existing MCMC methods (Sambridge, 2013). Therefore, in this work, we combine the parallel tempering method with pCN-MCMC for Bayesian inversion of multi-Gaussian random fields. To the best of our knowledge, we are the first to use this combination for Bayesian inversion of continuous multi-Gaussian hydraulic conductivity fields. While improving an algorithm is always a valuable scientific endeavor on its own right, we bear two specific motivations in mind: (1) to provide a method for application cases that require high-accuracy solutions in improved time, and (2) to provide a method that can offer high-accuracy reference solutions for critically testing improvement of much faster, approximate techniques such as quasi-linear ones or first-order second-moment methods. To demonstrate the high-accuracy reference character, we test the accuracy and efficiency of pCN-PT for estimating a multi-Gaussian log-hydraulic conductivity field with a relative high variance against analytical solutions (kriging) in a high-dimensional, linear problem (based on direct observations of log-hydraulic conductivity), against rejection sampling in a high-dimensional, nonlinear problem (based on observations of piezometric head linked to hydraulic conductivity through a groundwater flow equation) with only few measurements, and against pCN-MCMC in multiple, independent runs in a high dimensional, nonlinear problem with sufficient measurements. Note that we have the first two cases to build confidence in the algorithm. In the second case we use only few measurements so that it is possible with rejection sampling to calculate the true solution. Moreover, the new method is tested for high-dimensional problems and a relative high variance of log-hydraulic conductivity, which results in a strong statistical nonlinearity. Dealing with these kinds of problems is challenging for first-order second-moment methods.

The remainder of the paper is structured as follows. First, we introduce the algorithms of all relevant methods (pCN-MCMC, parallel tempering, pCN-PT, kriging, and rejection sampling), and then compare pCN-PT with simple kriging, rejection sampling, and pCN-MCMC on a synthetic confined aquifer. The paper ends with a summary and discussion.

2. Methodology

2.1. MCMC and pCN-MCMC

pCN-MCMC is an MCMC-based method proposed by coupling the pCN technique with the MCMC method. It draws samples from a chain that automatically honors the prior distribution in multi-Gaussian problems and therefore is more computationally efficient than classical MCMCs when dealing with highly resolved geostatistical problems (e.g., Cotter et al., 2013; Gilks et al., 1995). Hence, before we introduce the algorithm of pCN-MCMC, we briefly recall the classical Metropolis-Hastings MCMC algorithm, and the reader can refer to the literature (e.g., Geyer, 1992; Gilks et al., 1995; Hastings, 1970; Metropolis et al., 1953) for a more detailed description.

According to the Bayesian theorem, the posterior distribution $p(u \mid y)$ of a parameter vector u conditioned on observations y is dependent on the likelihood $L(u)$ and prior distribution of $p(u)$ (in this study, u and y correspond to hydraulic conductivity K and piezometric head H in our later application, respectively)

$$p(u|y) \propto L(u) * p(u) \tag{1}$$

MCMC algorithms sample from $p(u \mid y)$ by sequentially proposing jumps

$$u_{i+1} = u_i + \triangle u_i \tag{2}$$

where $\triangle u_i$ is randomly drawn (with certain rules) from a jump distribution. In these jumps, the most important rule is that the so-called detailed balance condition should be satisfied. This means that for any consecutive two states u_i and u_j, the jumps $u_i \rightarrow u_j$ and $u_j \rightarrow u_i$ are equally probable. In the Bayesian context, this can be expressed as

$$L(u_i) * p(u_i) * h(u_i, u_j) = L(u_j) * p(u_j) * h(u_j, u_i) \tag{3}$$

where h denotes the so-called transition kernel, and it is defined as

$$h(u_i, u_j) = q(u_i, u_j) * a(u_i, u_j) \tag{4}$$

where $q(u_i, u_j)$ is a proposal density and $a(u_i, u_j)$ is a corresponding acceptance probability. Hence, Equation 3 can be rewritten as

$$L(u_i) * p(u_i) * q(u_i, u_j) * a(u_i, u_j) = L(u_j) * p(u_j) * q(u_j, u_i) * a(u_j, u_i) \tag{5}$$

where, in order to enforce the detailed balance, the acceptance probability can be expressed as

$$a(u_i, u_j) = \min\left\{1, \frac{L(u_j) * p(u_j) * q(u_j, u_i)}{L(u_i) * p(u_i) * q(u_i, u_j)}\right\} \tag{6}$$

As was mentioned before, pCN-MCMC is a combination of the pCN technique and MCMC, specifically for multi-Gaussian prior (Cotter et al., 2013), and changes Equation 3 to

$$p(u_j) * q(u_j,\, u_i) = p(u_i) * q(u_i,\, u_j) \tag{7}$$

so that the acceptance probability Equation 6 can be reexpressed as

$$a(u_i,\, u_j) = \min\left\{1,\, \frac{L(u_j)}{L(u_i)}\right\} \tag{8}$$

In comparison with the acceptance probability in MCMC (Equation 6), we can see the acceptance probability in pCN-MCMC (Equation 8) is only dependent on the likelihood; but no longer on the prior $p(u)$. This contributes to a relatively large decrease of computational cost, because there will be no rejections due to the prior. In the following, without loss of generality, we will consider multi-Gaussian random fields discretized on some fine meshes, with the mean removed. Hence, we can write $p(u) = N(0,C)$, where C is the covariance matrix of the random field. Overall, the procedure of the pCN-MCMC algorithm can be described as follows:

1. Generate an initial realization u^0, $u^0 \sim N(0,C)$.
2. Generate a pCN proposal v^k at the kth sampling iteration according to the proposal function

$$v^k = \sqrt{1-\beta^2}u^k + \beta\varepsilon^k,\ \varepsilon^k \sim N(0,\, C) \tag{9}$$

where u^k is the sampling realization at the kth sampling iteration, ε is colored noise following the same distribution as the prior, and β denotes a jumping factor.

3. Set the new sampling realization $u^{k+1} = v^k$ with acceptance probability $a(u^k, v^k)$,

$$a(u^k,\, v^k) = \min\left\{1,\, \frac{L(v^{k|y})}{L(u^{k|y})}\right\} \tag{10}$$

4. Otherwise, set $u^{k+1} = u^k$.
5. $k \rightarrow k+1$.

If all observation errors and modeling errors follow Gaussian distributions, the log-likelihood $\phi(u) = \ln L(u \mid y)$ is

$$\phi(u) = \ln L(u|y) = \ln\left\{(2\pi)^{-\frac{m}{2}}\|C_y\|^{-\frac{1}{2}}exp\left[-\frac{1}{2}(y-y^o)^T C_y^{-1}(y-y^o)\right]\right\} \tag{11}$$

$$y = g(u)+\eta,\ \eta \sim N(0,\, C_n) \tag{12}$$

where y and y^o are simulated and measured observations, respectively; $g(\cdot)$ is a model parameterized by u (in our later application, the steady-state groundwater flow model); η is the measurement-and-model error; C_n is the covariance matrix of these errors. Then, Equation 8 can be rewritten as

$$a(u^k,\, v^k) = \min\{1,\, exp[\phi(v^k) - \phi(u^k)]\} \tag{13}$$

Note that the proposal density in Equation 9 tries to construct a multivariate autoregressive process of order one (AR1), which by design is multi-Gaussian at each step k and with distribution $N(O,C)$. The term $\sqrt{1-\beta^2}$ is the correlation coefficient along the chain. Then, $\beta = 1$ leads to independent sampling, while $\beta \rightarrow 0$ leads to small steps with high autocorrelation along the chain.

2.2. Parallel Tempering

Parallel tempering employs multiple parallel Markov chains at different temperatures to work with samples from multiple tempered posterior distributions (e.g., Earl & Deem, 2005; Geyer, 1991; Swendsen & Wang, 1986). In the temperature system, cold chains with precise sampling may become trapped in local models

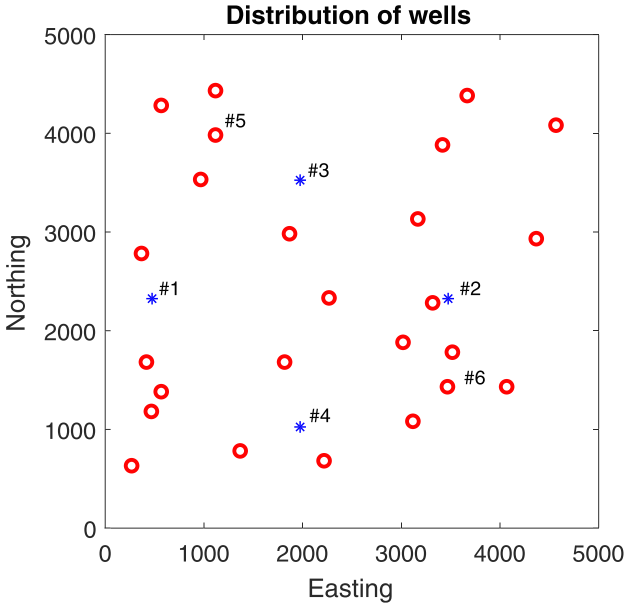


Figure 1. Distribution of pumping and observation wells used in computational experiments. Red circles denote observation wells; blue asterisks denote pumping wells.

and/or progress too slowly because they must use narrow proposal distributions to achieve a meaningful acceptance rate. In contrast, the hot chains are capable of sampling in a large model space due to their flatter and broader likelihood, which allows them to use wide proposal distributions. Therefore, parallel tempering can help the cold chain with unit temperature (also called the target chain) achieve good sampling by swapping solutions between cold chains and hot chains. This leads to improved mixing, and hence to better exploration. Many studies have proven that parallel tempering is superior to simple Monte Carlo and simulated annealing in the reconstruction of random fields (e.g., Earl & Deem, 2005; Makrodimitris et al., 2002; Moreno et al., 2003; Wang et al., 2015). The procedure of the algorithm can be described as below:

1. First, design a temperature ladder $T_1 < T_2 < ... < T_i ... < T_n$, with $T_1 = 1$, then the posterior $\pi_t(u)$ gets flattened toward the prior by temperatures

$$\pi_t(u) \propto L(u)^{\frac{1}{T}} p(u) \tag{14}$$

2. Then, for the ith chain with temperature T_i, the acceptance probability $a(u_i^k, v_i^k)$ in Equation 6 turns into

$$a(u_i^k, v_i^k) = \min\left\{1, \left(\frac{L(v_i^k)}{L(u_i^k)}\right)^{\frac{1}{T_i}} * \frac{p(v_j) * q(v_j, u_i)}{p(u_i) * q(u_i, v_j)}\right\} \tag{15}$$

3. Last, for any two chains with temperatures T_i and T_j, at intervals of $\frac{1}{d}$ steps (where d is a swap proposal frequency), swaps are proposed between cold chains and hot chains, with swap acceptance probability $a_s(u_i, u_j)$.

$$a_s(u_i, u_j) = \min\left\{1, \frac{L(u_j)^{\frac{1}{T_i}} * p(u_j) * q(u_j, u_i) * L(u_i)^{\frac{1}{T_j}} * p(u_i) * q(u_i, u_j)}{L(u_i)^{\frac{1}{T_j}} * p(u_i) * q(u_i, u_j) * L(u_j)^{\frac{1}{T_i}} * p(u_j) * q(u_j, u_i)}\right\} \tag{16}$$

2.3. pCN-PT

In this work, we take advantage of pCN-MCMC in efficiently dealing with highly resolved multi-Gaussian problems and explore parallel tempering for improving efficiency by faster mixing. Therefore, we blend pCN-MCMC and parallel tempering to develop a new fast MCMC algorithm named pCN-PT. On the basis of the above algorithms, the pCN-PT algorithm can be summarized as follows:

1. Set initial realizations u_i^0, $u_i^0 \sim N(0, C)$, a temperature ladder $T_1 < T_2 < ... < T_i < ... < T_n$, with $T_1 = 1$, a jumping factor ladder $\beta_1 < \beta_2 < ... < \beta_i < ... < \beta_n$, with $\beta_n < 1$, where i is the number of chain, $i \in (1,n)$, and a swap proposal frequency d. Note that, to make an acceptance rate close to optimal for each chain, the jumping factor should increase with increasing temperature.
2. Generate a pCN proposal v_i^k at the kth sampling iteration and for all chains $i = 1, ... , n$,

$$v_i^k = \sqrt{1 - \beta_i^2} u_i^k + \beta_i \varepsilon_i^k, \ \varepsilon_i^k \sim N(0, C) \tag{17}$$

3. For each chain i, set $u_i^{k+1} = v_i^k$ with acceptance probability $a(u_i^k, v_i^k)$,

$$a(u_i^k, v_i^k) = \min\left\{1, \exp\left[\frac{\phi(v_i^k) - \phi(u_i^k)}{T_i}\right]\right\} \tag{18}$$

4. Otherwise, set $u_i^{k+1} = u_i^k$.

Table 1
Pumping Wells Used for Computational Experiments

Well number	Grid position	Pumping rate (m^3/day)
#1	(10,47)	120
#2	(70,47)	70
#3	(40,71)	90
#4	(40,21)	90

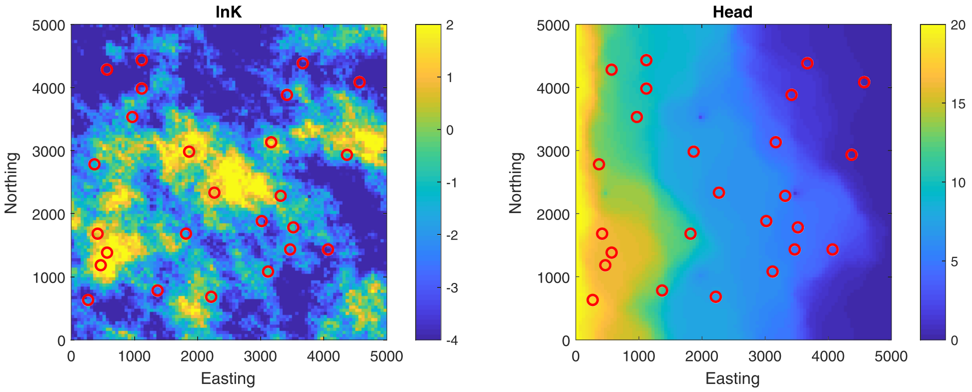


Figure 2. Reference (left) lnK field and reference piezometric (right) head solution.

5. For any pairs of chains, for example, ith and jth chain, if $\frac{k}{d} = integer$, swap values between pairs of chains $u_i^k \leftrightharpoons u_j^k$ with swap acceptance probability $a_s(u_i^k, u_j^k)$,

$$a_s(u_i^k, u_j^k) = \min\left\{1, \exp\left[\left(\phi(u_j^k) - \phi(u_i^k)\right) * \left(\frac{1}{T_i} - \frac{1}{T_j}\right)\right]\right\} \tag{19}$$

6. $k \rightarrow k + 1$.

The tuning parameters of pCN-PT are the number n of chains, the temperature ladder $T_1, \ldots, T_n$, the jumping factor ladder $\beta_1, \ldots, \beta_n$, and the swap proposal frequency d. How to handle these tuning parameter values and the settings in our later application can be found in section 3.3.

2.4. Simple Kriging

In this work, we will use simple kriging to construct an analytical test case. Simple kriging is a best linear unbiased estimator assuming the first moment over the entire domain to be a known constant and the covariance to be stationary (e.g., Deutsch & Journel, 1998; Remy et al., 2009). The main difference between simple kriging and ordinary kriging is that the mean is known for simple kriging, whereas it is unknown for ordinary kriging.

Simple kriging is a linear estimator, meaning the difference between the estimate $\hat{u}_0$ and mean m is a weighted linear combination of the difference between the neighboring data u_i and mean m,

$$\hat{u}_0 = m + \sum_{i=1}^{n} \lambda_i * (u_i + \varepsilon_i - m),\ i = 1, 2, \ldots, n \tag{20}$$

where λ_i and ε_i denote the kriging weight and measurement data error for the ith data position, respectively; m is the mean. As can be seen from the presence of ε_i, we adopt the version for imprecise data, so that a comparison with a method that assumes a likelihood (our MCMC) is meaningful.

The weights are found by solving the simple kriging equation system:

$$\sum_{j=1}^{n} \lambda_j * C_{i,j} + \lambda_i \sigma_\varepsilon^2 = C_{i,0} \tag{21}$$

where $C_{i,j}$ corresponds to the covariance between the pairs of random variables at positions i and j (i and j are measurement locations); $C_{i,0}$ corresponds to the covariance between the pairs of random variables at positions i and position 0 (0 is a position for which an estimation is required); σ_ε^2 is the variance of data error.

The estimation variance is then given by

Table 2
Parameters of the Random Functions Describing the Heterogeneity of lnK

	Mean	Variogram type	$\lambda_{\max}$	$\lambda_{\min}$	Std.dev	Angle
Reference	−2.5	Exponential	2,000	1,500	2	135

Note. $\lambda_{\max}$ and $\lambda_{\min}$ are the correlation ranges in the x and y directions.

Table 3
Definition of Scenarios

			Method			
	Scenario	Sampling iterations	Kriging	Rejection-sampling	pCN-PT	pCN-MCMC
25 lnK measurements (std.dev = 0.02)	S1	400,000	√			
	S2	400,000			√	
Two head measurements (std.dev = 1.5)	S3	2,000,000		√		
	S4	400,000			√	
	S5	400,000				√
25 head measurements (std.dev = 0.05)	S6	800,000			√	
	S7	800,000				√

$$\sigma_R^2 = C_{0,0} + \sum_{j=1}^{n}\sum_{i=1}^{n}\lambda_i\lambda_j C_{i,j} - 2\sum_{i=1}^{n}\lambda_i C_{i,0} + \sum_{i=1}^{n}\lambda_i^2\sigma_\varepsilon^2 \tag{22}$$

where $C_{0,0}$ is the variance of the random variables.

In our current context, the kriging estimate is the exact posterior mean, and the estimated variance is the exact posterior variance of the inverse problem stated in Equation 1, if the prior is multivariate Gaussian with constant, known mean, if the likelihood is independent and identically distributed Gaussian with variance σ_n^2, and if the data are direct point observations of the unknown u.

2.5. Rejection Sampling

Rejection sampling is a basic technique used to draw samples from almost arbitrary distributions by selecting, accepting, or rejecting samples from a proposal distribution. In the Bayesian setting of Equation 1, it is a simple but computationally expensive method that can provide independent and unbiased samples from a posterior distribution (e.g., Tarantola, 2005; von Neumann, 1951). Rejection sampling is very efficient in many low-dimensional problems, but may be difficult in high-dimensional problems. This is because proposal distributions typically become less efficient in higher dimensions, which can make the acceptance rate dramatically low (Tarantola, 2005). For Bayesian updating, the algorithm can be described as follows:

1. Generate an ensemble of samples u^k, $k \in (1,n)$ from the prior $p(u)$ as proposals.
2. Calculate the acceptance probability $P(u^k)$, which is a function of the ratio of the likelihood $L(u^k)$ to the supremum $L_{\max}$ of the sampled likelihoods,

$$P(u^k) = \frac{L(u^k)}{L_{\max}} \tag{23}$$

3. Accept u^k with probability $P(u^k)$, otherwise reject.

We will use rejection sampling for comparison as reference solution, because it is unbiased and provides independent samples, while MCMC is only asymptotically unbiased and can only provide dependent samples along its domain. To escape the problems that rejection sampling can have in high dimensions, we will have to limit the corresponding test case to only a few and weakly informative data, such that the prior distribution is still an efficient proposal distribution for the posterior.

Table 4
(Base) Jumping Factor and (Base) Temperature Used to Construct Ladder in pCN-PT (S2, S4, S6) and pCN-MCMC (S5, S7)

Scenario	Jumping factor	Temperature
S2	0.765	1.73
S4	0.97	1.06
S5	0.53	
S6	0.78	1.76
S7	0.005	

2.6. Testing Criteria

As testing criteria, we use (1) the closeness of posterior samples u^k or posterior mean fields to a synthetic reference field u^{ref}, (2) the attained log-likelihoods for the closeness to the synthetic data, and (3) a convergence metric for MCMC chains.

First, we evaluate the goodness of the generated sampling fields using the root-mean-square error (RMSE) against the synthetic reference field,

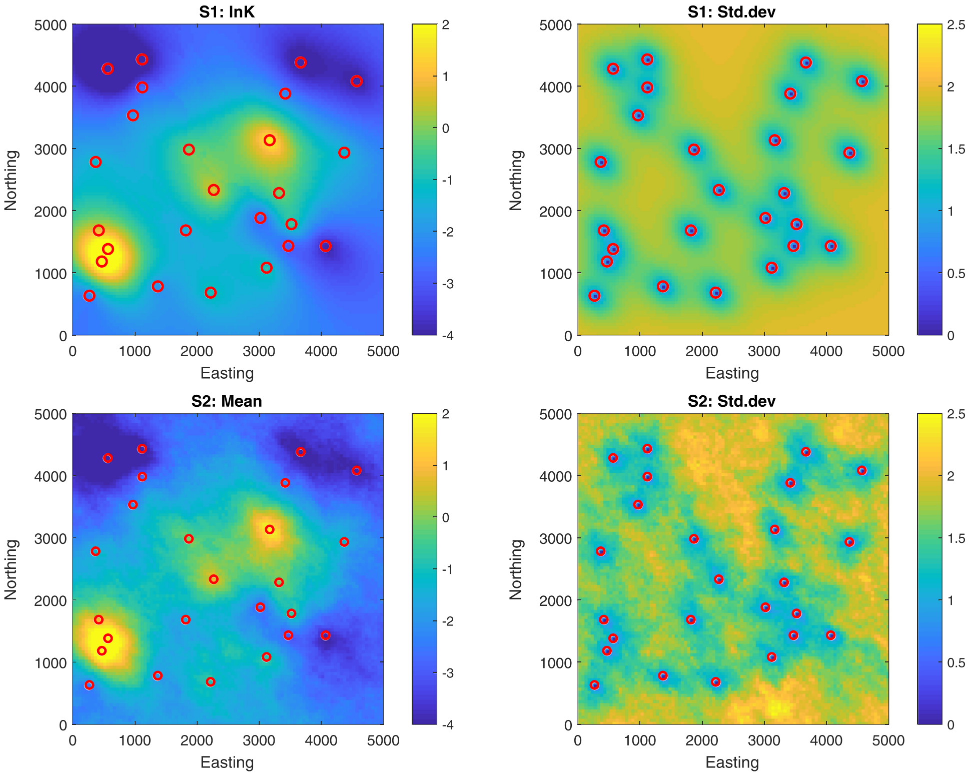


Figure 3. The maps show the estimates and std.dev obtained analytically by simple kriging (top row, S1) and obtained by pCN-PT (bottom row, S2). The locations of the 25 lnK measurements are also displayed on the maps.

$$RMSE^k = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(u_j^{ref} - u_j^k\right)^2} \tag{24}$$

where N is the number of model gridblocks; u_j^{ref} is the value of the reference field at the jth gridblock; u_j^k is the value of the sampling field at the jth gridblock in the kth iteration. As a global matrix independent of k, we can also use Equation 24 with the posterior mean field. When comparing samples to the kriging solution, we will use the posterior mean fields from MCMC and assess the RMSE against the kriging solution.

Second, for log-likelihood, we simply take the log-likelihood from Equation 11.

Third, for convergence, a major consideration for any MCMC chain is whether it is long enough to converge and fully explore the target posterior. Here, we evaluate the convergence of multiple independent MCMC runs using the Gelman-Rubin convergence diagnostic (Gelman et al., 2013; Gelman & Rubin, 1992). For each model parameter, the convergence is quantified using the potential scale reduction factor $\hat{R}$ by comparing the estimated between-chain and within-chain variances. The reader is referred to the literature (Brooks & Gelman, 1998; Gelman & Rubin, 1992) for a detailed description of the method:

$$\hat{R} = \sqrt{\frac{V}{W}} \tag{25}$$

where V is the estimated variance of the stationary distribution as a weighted average of the within-chain (W) and between-chain (B) variance

$$V = 1 - \frac{1}{n} W + \frac{1}{n} B \tag{26}$$

where

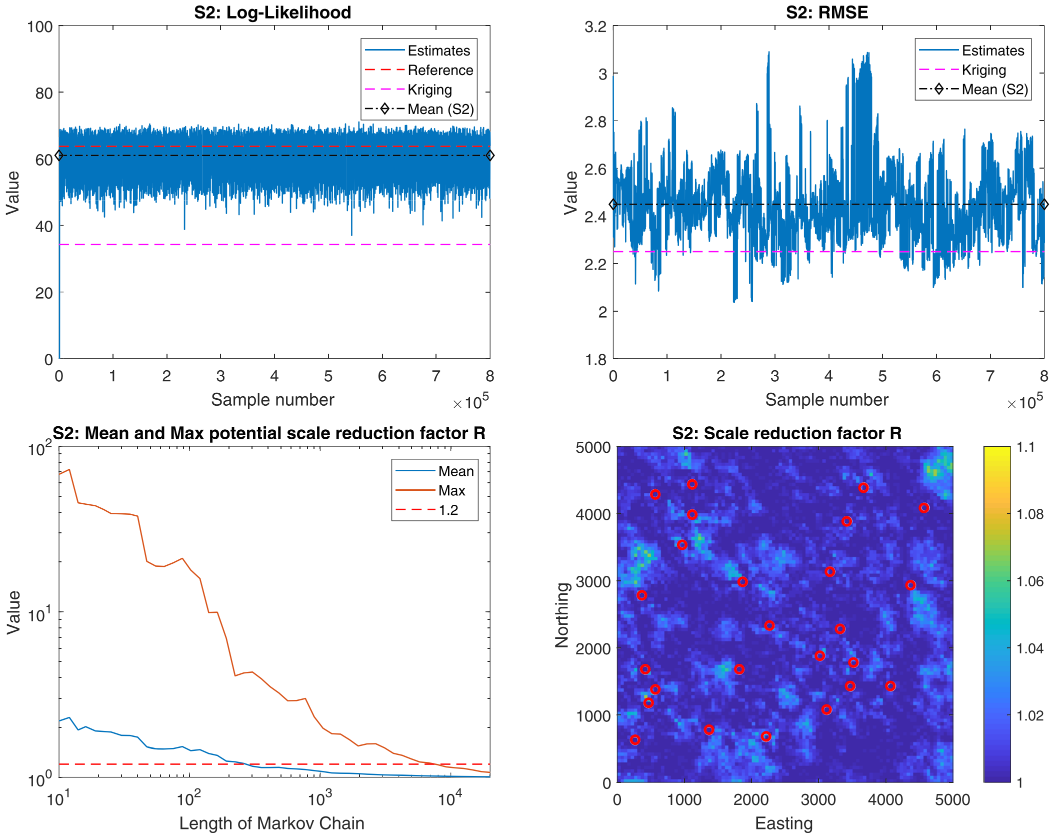


Figure 4. The 25 lnK measurements. (top left) Log-likelihood, (top right) RMSE, (bottom left) the evolution of mean (blue) and maximum (red) potential scale reduction factor, and (bottom right) potential scale reduction factor field obtained by pCN-PT (S2). The red dashed line, pink dashed line, and black dash-diamond line in the figure of the log-likelihood correspond to the log-likelihood of the reference and kriging, and the mean of the log-likelihood for pCN-PT (S2), respectively; the pink dashed line and black dash-diamond line in the figure of the RMSE correspond to the RMSE of kriging and the mean of the RMSE for pCN-PT (S2); the red dashed line, the blue line, and the brown line in the evolution of mean and maximum potential scale reduction factor correspond to the value 1.2, the mean value and the maximum value for posterior samples, respectively.

$$W = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{k=1}^{n} (u_i^k - \bar{u}_i)^2 \tag{27}$$

$$B = \frac{n}{m-1} \sum_{i=1}^{m} \left(\bar{u}_i - \frac{1}{m} \sum_{i=1}^{m} \bar{u}_i \right)^2 \tag{28}$$

where m is the number of multiple independent chains; n is the number of samples in each chain; $\bar{u}_i$ is the mean of samples for the ith chain. Note that the samples we mention here are the second half of samples in each chain, already discarding the burn-in period.

3. Application

3.1. Setup

For testing our proposed pCN-PT method, we consider fully saturated, steady-state groundwater flow in a synthetic confined, two-dimensional aquifer in a 5,000 m by 5,000 m domain with 50 m thickness. It is discretized into 100 by 100 by 1 cells, where each cell is 50 m by 50 m by 50 m. The west and the east boundaries are specified head boundaries, with heads fixed at 20 and 0 m, respectively; the north and the south boundaries are impermeable. Four pumping wells are located in the domain, as can be seen in Figure 1. The exact locations and pumping rates are provided in Table 1. A reference log-conductivity field is generated following a multi-Gaussian distribution with mean −2.5 ln [m/day] and standard deviation (std.dev) 2 ln [m/day]

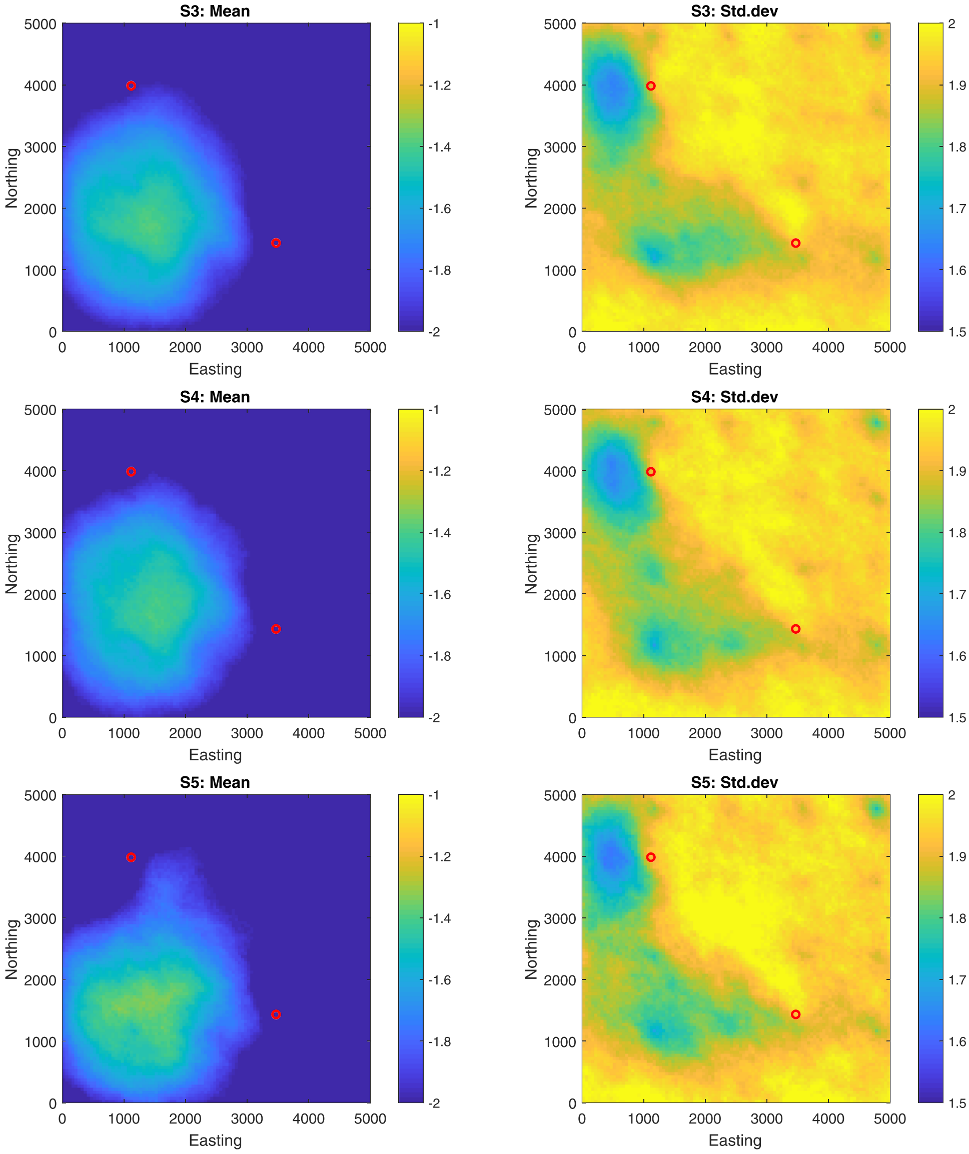


Figure 5. Two head measurements. Mean and std.dev of 10,728 accepted lnK realizations obtained by rejection sampling (top row, S3), lnK realizations obtained by pCN-PT (middle row, S4), and pCN-MCMC (bottom row, S5).

(Figure 2), using a sequential Gaussian simulation module (SGSIM) of the GSLIB software (Deutsch & Journel, 1998). The details of the parameters used to generate the reference field are provided in Table 2.

The steady-state groundwater flow Equation 29 is solved by using the groundwater flow simulator MODFLOW (McDonald & Harbaugh, 1988):

$$\nabla \cdot (K\nabla H) + q = 0 \tag{29}$$

where $\nabla \cdot$ denotes the divergence operator; K is the hydraulic conductivity [m/day]; ∇ is the Nabla operator; H denotes piezometric head [m], and q is the volumetric injection flow rate per unit volume of aquifer [1/day].

3.2. Testing Scenarios

In this work, seven scenarios are designed to test the pCN-PT method for Bayesian inference of multi-Gaussian random fields in different situations. They can be grouped into three categories: (1) a

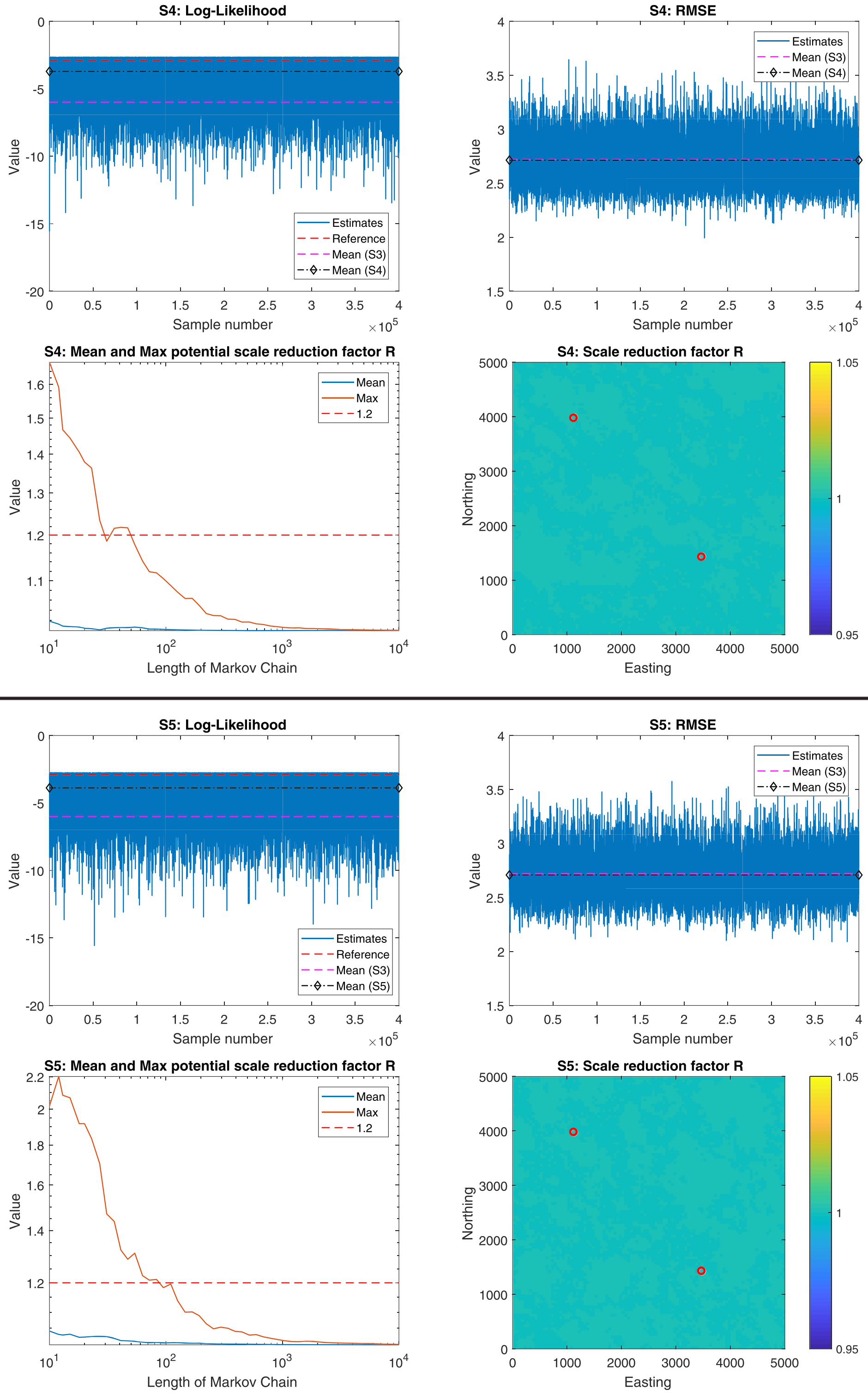


Figure 6. Two head measurements. Log-likelihood, RMSE, the evolution of mean (blue) and maximum (red) potential scale reduction factor and potential scale reduction factor field obtained by pCN-PT (top 2 rows, S4) and pCN-MCMC (bottom 2 rows, S5). The red dashed line, pink dashed line, and black dash-diamond line in the figure of the log-likelihood correspond to the log-likelihood of the reference, and the mean of the log-likelihood for rejection sampling (S3), pCN-PT (S4), or pCN-MCMC (S5), respectively; the pink dashed line and black dash-diamond line in the figure of the RMSE correspond to the mean of the RMSE for rejection sampling (S3), pCN-PT (S4), or pCN-MCMC (S5); the red dashed line, the blue line, and the brown line in the evolution of mean and maximum potential scale reduction factor correspond to the value 1.2, the mean value, and the maximum value for posterior samples, respectively.

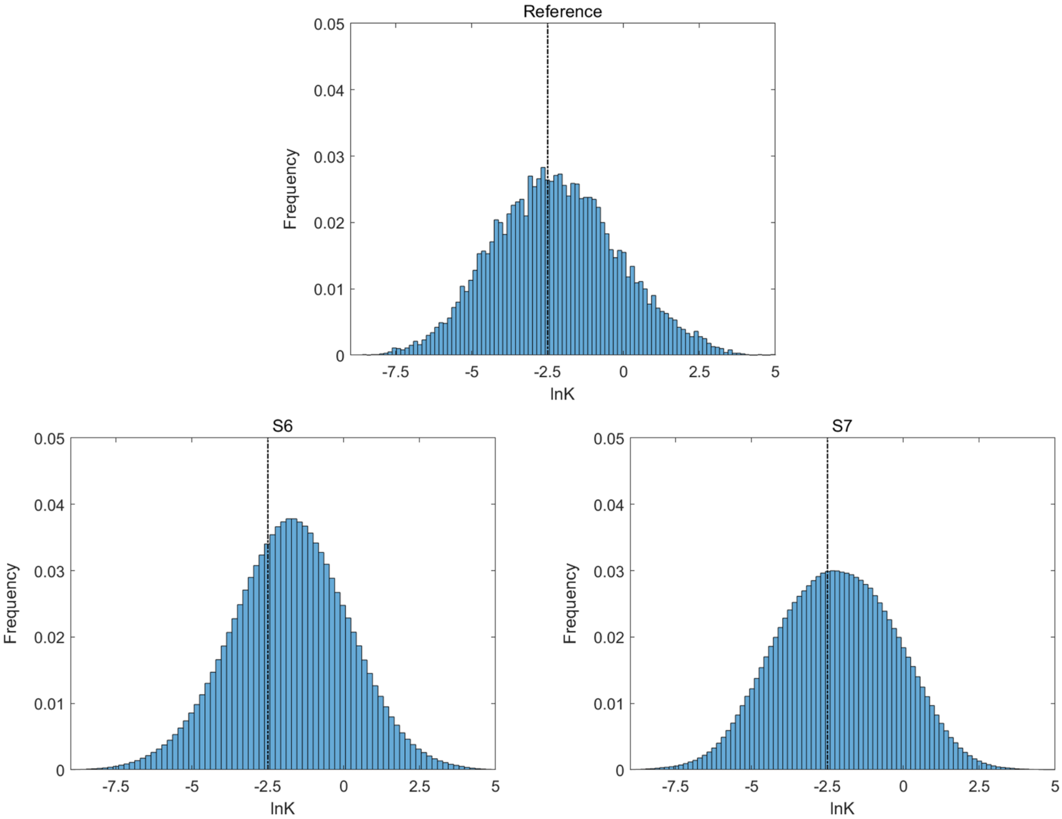


Figure 7. Histogram. The histogram of the reference lnK field (top row), the histogram obtained by pCN-PT (bottom left, S6), and pCN-MCMC(bottom right, S7). The vertical dashed line indicates the mean of the random functions used in the SGSIM.

high-dimensional, linear problem; (2) a high-dimensional, nonlinear problem with only few measurements; (3) and a high-dimensional, nonlinear problem with many measurements. The first category serves to compare the performance of pCN-PT (scenario S2) against an accurate analytical solution obtained by simple Kriging (scenario S1). Therefore, we only consider 25 direct measurements of lnK (no head data) at the positions marked with red circles in Figure 1. The second category is designed to allow a high-fidelity reference solution by rejection sampling (scenario S3). Here, we can test pCN-PT (scenario S4) and pCN-MCMC (scenario S5) against rejection sampling, but only in a low-information regime with two head measurements (labeled as positions #5 and #6 in Figure 1) and with large measurement errors that lead to a wide likelihood function. Finally, the last category serves to test pCN-PT (scenario S6) against pCN-MCMC (scenario S7) on a problem with 25 more accurate head measurements. The details of scenarios can be found in Table 3.

For scenarios S2 and S6 and S7 conditioned to 25 measurements, 800,000 MCMC realizations are generated and saved every 20 realizations. For scenarios S4 and S5 conditioned to two head measurements, 400,000 MCMC realizations are generated, again saved every 20 realizations. For scenario S3 performed by rejection sampling, we generate 2,000,000 sampling realizations due to its dramatically low acceptance rate even with a high std.dev for the measurement error. There are only 10,728 realizations accepted. For all scenarios tested by pCN-PT or pCN-MCMC (S2 and S4–S7), we discard the first half of all sampling realizations as a burn-in period, and the calculation of mean, std.dev, and scale reduction factor R rests on the second half of sampling realizations. Besides, the potential scale reduction factor R is calculated for each grid of a posterior sampler.

3.3. Algorithmic Settings

Many studies have indicated that the optimal acceptance rate for MCMC-based methods is close to 23.4% and that acceptance rates between 10% and 40% perform close to optimal (e.g., Gelman et al., 1996; Roberts et al.,

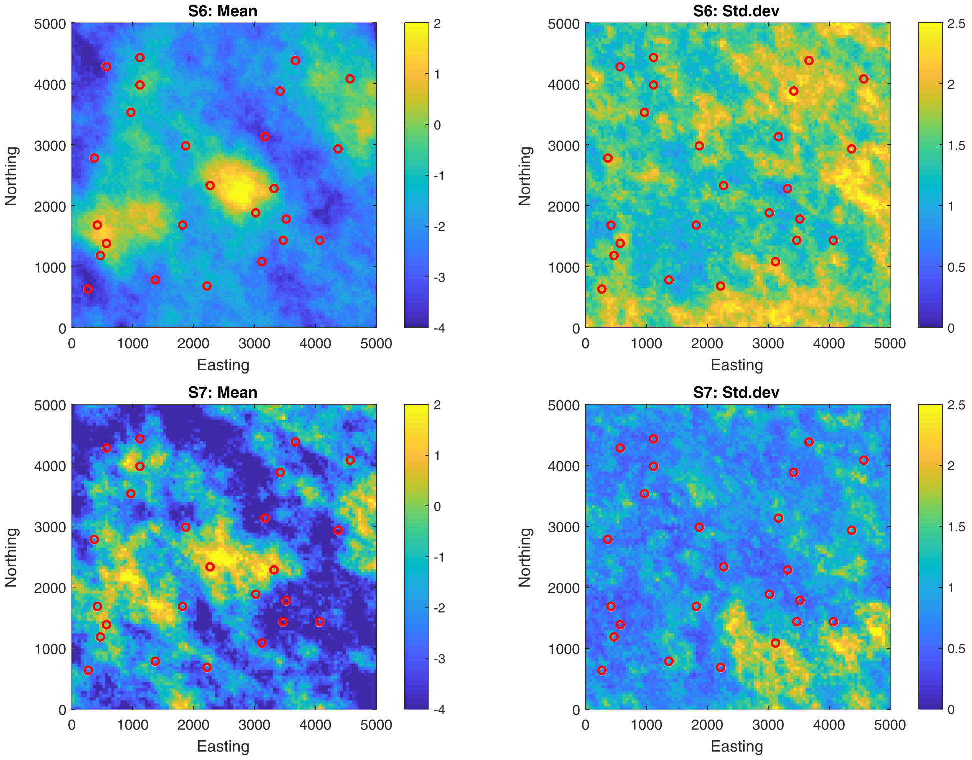


Figure 8. The maps show the mean and std.dev of lnK realizations obtained by pCN-PT (top row, S6) and pCN-MCMC (bottom row, S7). The locations of the 25 head measurements are also displayed on the maps.

1997, 2001). There are also many analyses for the optimal acceptance swap rate for parallel tempering. However, there is no well-recognized optimal setting. The optimal acceptance swap rate differs with respect to the specific applications, such as, Rathore et al. (2005) found an optimal acceptance swap rate at 20%, whereas Predescu et al. (2005) and Laloy et al. (2016) mentioned optimal acceptance swap rates at 39% and 8%, respectively. In this work, we control the acceptance rate in a range from 20% to 30% and the swap acceptance rate in a range from to 10% to 30% by adjusting the jumping factor and temperature ladders. For scenarios S2, S4, and S6 using pCN-PT, we run 20 parallel chains for each scenario. Each chain runs independently on a dedicated computing node, and data commutation only occurs for between-chain swaps. Hence, if the number of sampling iterations is the same, the computational wall clock time for pCN-PT and pCN-MCMC is almost the same. For constructing the temperature ladder, the temperatures increase exponentially with a base temperature (larger than 1) and powers increasing by integers from 0 to 19, while the jumping factors increase exponentially with a base jumping factor (less than 1) and powers decrease by integers from 20 to 1. The base jumping factors and base temperatures for scenarios S2, S4, and S6 as well as the jumping factors for scenarios S5 and S7 can be found at Table 4.

4. Results

4.1. Comparison to Kriging (S1 and S2)

Figure 3 shows the estimated multi-Gaussian lnK field and the corresponding std.dev obtained by simple kriging (top row, S1) and the ensemble mean and std.dev of the last 200,000 sampling lnK realizations obtained by pCN-PT (bottom row, S2). We can see, obviously, that both results capture the main features of the reference field and look very similar. Additionally, the std.dev around the conditioning points is close to zero. Even so, the area with small std.dev for scenario S2 is a bit larger than that for scenario S1. This demonstrates qualitatively that pCN-PT is capable of estimating a multi-Gaussian random field in a high-dimensional linear problem.

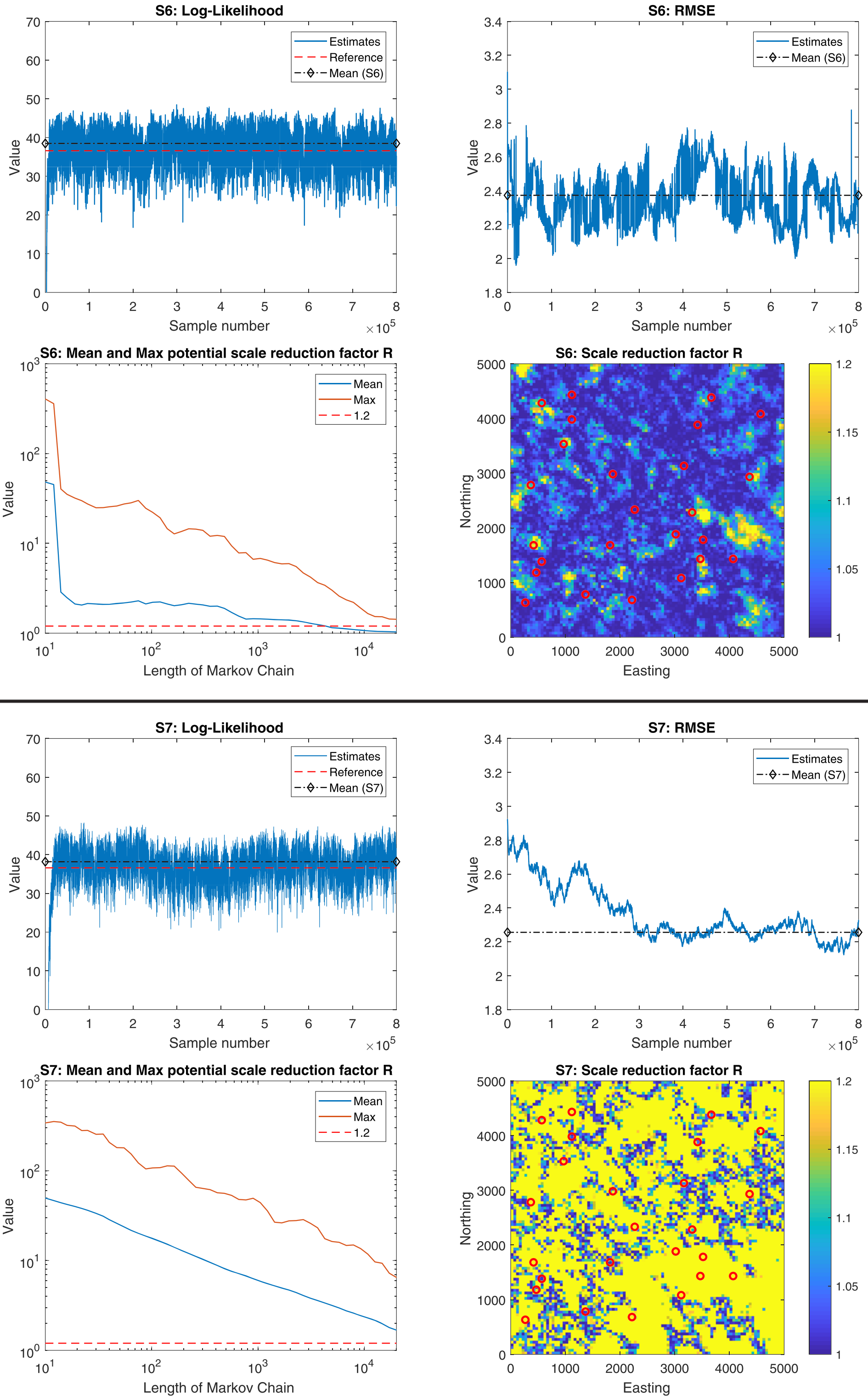


Figure 9. The 25 head measurements. RMSE, the evolution of mean (blue) and maximum (red) potential scale reduction factor and potential scale reduction factor field obtained by pCN-PT (top 2 rows, S6) and pCN-MCMC (bottom 2 rows, S7). The red dashed line and black dash-diamond line in the figure of the log-likelihood correspond to the log-likelihood of the reference and the mean of the log-likelihood, respectively; the black dash-diamond line in the figure of the RMSE corresponds to the mean of the RMSE; the red dashed line, the blue line, and the brown line in the evolution of mean and maximum potential scale reduction factor correspond to the value 1.2, the mean value, and the maximum value for posterior samples, respectively.

Figure 4 shows the log-likelihood (upper left), RMSE (upper right), the mean and maximum potential scale r-eduction factor R (lower left), and the spatially resolved potential scale reduction factor R field (lower right) for the target chain obtained by pCN-PT (S2). The log-likelihood and RMSE for the kriging estimates and the mean of the log-likelihood and RMSE for pCN-PT are also included. From the log-likelihood and RMSE results, we can see that the burn-in and convergence to the reference log-likelihood is very fast, the mean of the log-likelihood for pCN-PT is close to that of the reference, and the log-likelihood and RMSE for kriging are smaller than the mean of those for pCN-PT. Besides, the potential scale reduction factor R keeps decreasing and drops below the recommended value of 1.2 already at around 135,060 sampling iterations. It means that a stationary distribution can be achieved after around 135,060 sampling iterations.

4.2. Comparison to Rejection Sampling (S3, S4, and S5)

To make rejection sampling accessible as feasible reference solution algorithm, we selected only two head data values, and we widened the likelihood by manual trial and error. We enlarged the std.dev of the measurement error to 1.5, and the acceptance rate of rejection sampling increased to 0.5364% with 10,728 sampling realizations accepted in 2,000,000 iterations. This is a compromise between a suitably large number of accepted realizations versus an unrealistically large measurement error. Figure 5 shows the ensemble mean and std.dev of 10,728 accepted realizations obtained by rejection sampling (top row, S3), and the ensemble mean and std.dev obtained by pCN-PT (middle row, S4) and pCN-MCMC (bottom row, S5), respectively. As can be expected when measurement error is large, the mean of the estimates is smooth with a large uncertainty in most parts of the domain, where the std.dev is close to the prior std.dev of 2. Figure 6 shows the log-likelihood, RMSE, the mean and maximum potential scale reduction factor R, and spatially resolved potential scale reduction factor R field for the target chain obtained by pCN-PT (top 2 rows, S4) and pCN-MCMC (bottom 2 rows, S5). We can see from Figures 5 and 6, that the estimates are similar for the three methods, pCN-PT and pCN-MCMC burn in fast. The mean of the log-likelihood for the three methods is smaller than the reference, but the mean of the log-likelihood for pCN-PT and pCN-MCMC is much closer to the reference. Plus, the mean of the RMSE for the rejection sampling overlaps that for pCN-PT and pCN-MCMC. When comparing the convergence between pCN-PT and pCN-MCMC (Figure 6), we can see that parallel tempering really helps to improve pCN-MCMC with better mixing (the potential scale reduction factor R of pCN-PT drops below the value of 1.2 around 940 sampling iterations, while the potential scale reduction factor R of pCN-MCMC needs 2,200 sampling iterations to drop below the value of 1.2). It demonstrates that pCN-PT can deal with high-dimensional, nonlinear problems as accurate as rejection sampling, but more efficient than rejection sampling and pCN-MCMC. We also show these results for scenarios S4a and S5a in Figures A1 and A2 in Appendix A. The setting of scenarios S4a and S5a is the same as that of scenarios S4 and S5, respectively, except that the std.dev of measurement error is 0.05 instead of 1.5. The figures in the appendix display that for the case of the std.dev of the error of 0.05, the estimates get closer to the reference. However, comparison for rejection sampling is impossible, so we refer to the next section for performance assessment in realistic cases.

4.3. Convergence in a Realistic Case (S6 and S7)

Figure 7 shows the histogram of the reference lnK field and the histogram for pCN-PT and pCN-MCMC. We can observe that the Gaussianity for both methods is preserved and similar to the reference one. Figure 8 shows the mean, std.dev of lnK sampling realizations for pCN-PT (S6) and pCN-MCMC (S7). We can see that both methods can capture the main features of the reference lnK field, and the mean obtained by pCN-PT (S6) is a bit smoother. Figure 9 shows the log-likelihood, RMSE, the mean and maximum potential scale reduction factor R, and spatially resolved potential scale reduction factor R field for the target chain of the two methods for comparison. As anticipated in the visual analysis, the log-likelihood for pCN-PT (S6) converges faster and is closer to the reference than the one for pCN-MCMC (S7) and the mean of the log-likelihood for both methods is close to the reference. Moreover, the RMSE for pCN-PT (S6) decreases much faster and converges more quickly. Overall, pCN-PT shows a better mixing and faster burn-in than pCN-MCMC. Besides, as we can see in Figure 9, the potential scale factor R for pCN-PT (S6) keeps decreasing with the increase of the length of the chain, and after 800,000 sampling iterations, the potential scale factor R can reduce to less than 1.2 in most parts of the domain. However, for pCN-MCMC (S7), although the potential scale factor R also decreases as the length of the chain increases, it is still larger than 1.2 in most parts of the domain at the end of our chain solution runtime. This shows that pCN-PT can arrive at the stationary distribution faster than pCN-MCMC. Given this analysis, we can conclude that pCN-PT can handle high-dimensional nonlinear problems better than pCN-MCMC.

5. Summary and Discussion

In this study, a pCN-PT algorithm has been developed and we have demonstrated that it works more efficiently than rejection sampling and pCN-MCMC for Bayesian inversion of multi-Gaussian hydraulic conductivity fields. Besides, we have also demonstrated it can be used to deal with multi-Gaussian random fields in both high-dimensional linear problems and high-dimensional nonlinear problems. Note that the simulation problem in this paper is relatively high dimensional compared to the number of parameters which has been estimated until now in other MCMC-applications in groundwater hydrology. For real-world problems we will often have to estimate more uncertain parameters, as the number of grid cells is larger and there are more unknown variables.

Especially, the main focus of this study is to evaluate the accuracy and efficiency of pCN-PT in the estimation of multi-Gaussian random fields. In our follow-up research, we plan to design and implement a suite of well-defined benchmark scenarios for stochastic inversion and will invite interested researchers in testing their approaches (such as pCN-PT, MCMC-based methods; e.g., DREAM, Vrugt et al., 2008, 2009), the ensemble Kalman filter, and the inverse sequential simulation for joint scenarios. This will allow to evaluate and discuss the drawbacks and benefits of the approaches based on the comparisons.

In terms of the results, parallel tempering is clearly helpful for improving the performance of pCN-MCMC. How to design an optimal temperature configuration is still an ongoing discussion (e.g., Atchadé et al., 2011; Carter & White, 2013). Here, we choose an exponential distribution for temperatures and the details can be found in section 3.3. There may be other optimal temperature configurations. If others exist, the configuration in this work can be treated as a benchmark for the comparison.

In this study, we implemented the algorithms running in a hybrid parallel environment. Although this parallel computer architecture is helpful to improve the computational efficiency, there is a tradeoff between including more processors and increasing the efficiency of the calculations, which is related to the increasing communication cost between different processors for problems where many processors are used.

Appendix A

In the above context, we have showed the results of scenarios S4 and S5 with high std.dev of the measurement error in Figures 5 and 6. But the high uncertainty of the measurement error induces a large uncertainty

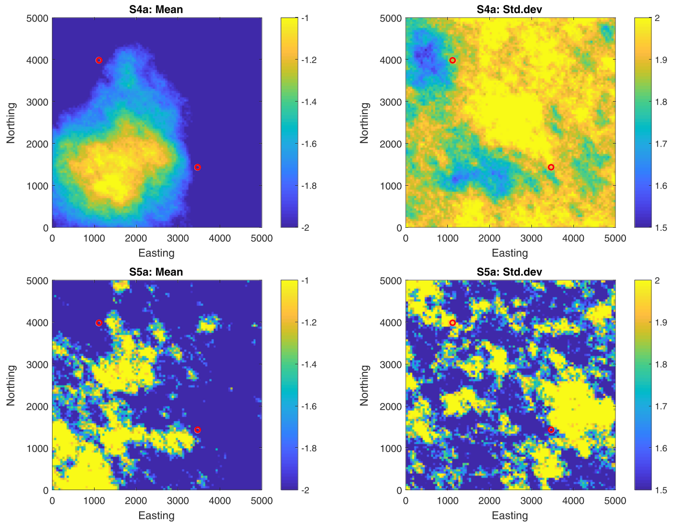


Figure A1. Two head measurements. Mean and std.dev of lnK realizations for scenarios (top row) S4a and (bottom row) S5a.

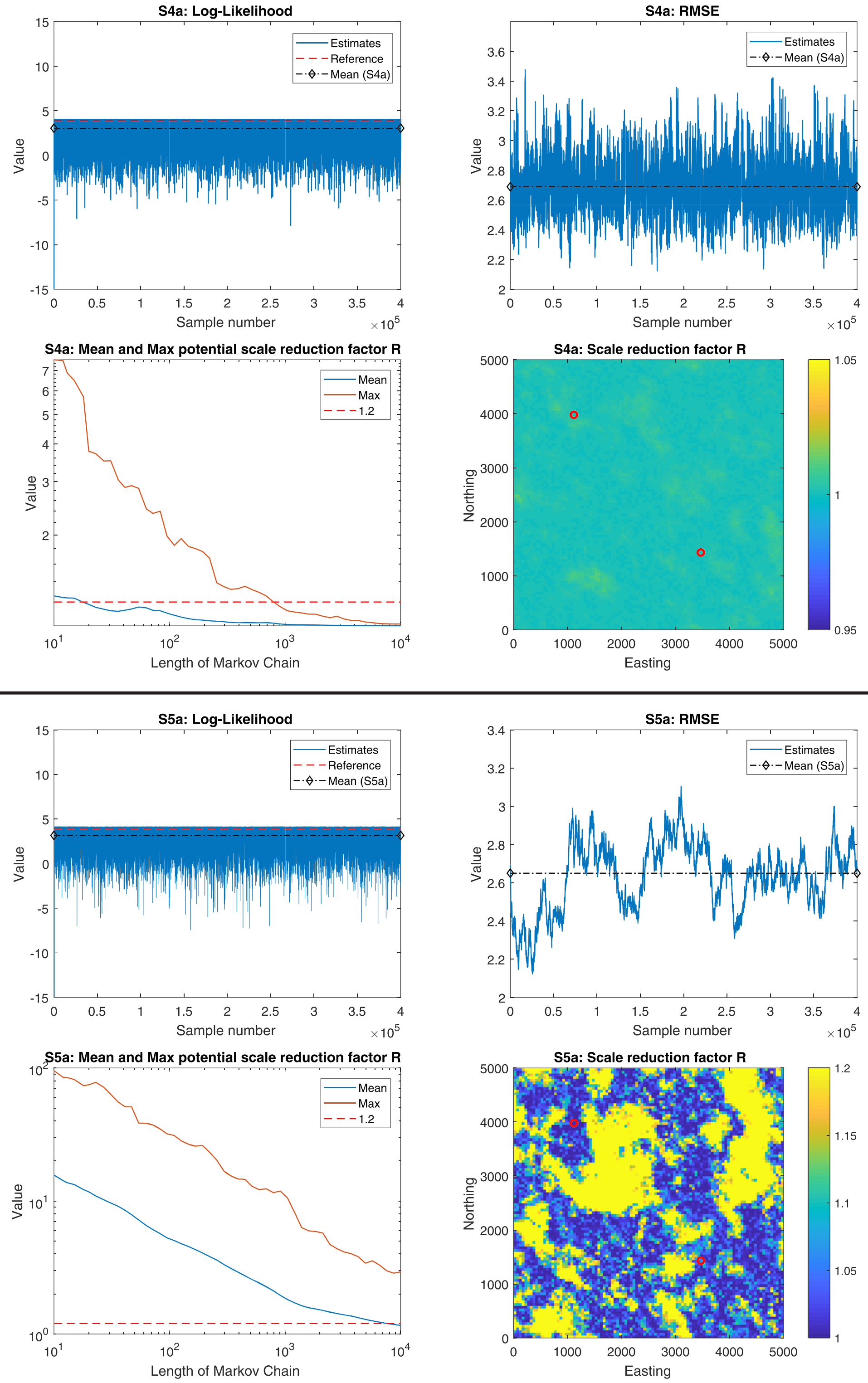


Figure A2. Two head measurements. Log-likelihood, RMSE, the evolution of mean (blue) and maximum (red) potential scale reduction factor, and potential scale reduction factor field obtained by pCN-PT (top 2 rows, S4a) and pCN-MCMC (bottom 2 rows, S5a). The red dashed line and black dash-diamond line in the figure of the log-likelihood correspond to the log-likelihood of the reference and the mean of the log-likelihood, respectively; the black dash-diamond line in the figure of the RMSE corresponds to the mean of the RMSE; the red dashed line, the blue line, and the brown line in the evolution of mean and maximum potential scale reduction factor correspond to the value 1.2, the mean value, and the maximum value for posterior samples, respectively.

for Bayesian inversion of log Gaussian hydraulic conductivity fields. And there is no big difference between pCN-PT and pCN-MCMC from the three figures. Here, we show the results of two new scenarios S4a and S5a in Figures A1 and A2, which have the same settings as scenarios S4 and S5, respectively, just replacing the std.dev of the measurement error 1.15 with a low value of 0.05.

Data Availability Statement

Codes and related data are available from this site (https://data.mendeley.com/datasets/9zphw6c8xx/ draft?a=34ae2274-f505-4bff-84ed-305d22c3b752).

Acknowledgments

Financial support to carry out this work was received from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through COMPUFLOW-Projects 359880532 and 327154368-SFB 1313 and Germany's Excellence Strategy-EXC-2075-390740016. Teng Xu also acknowledges the financial support from the Fundamental Research Funds for the Central Universities (B200201015) and Jiangsu Specially-Appointed Professor Program (B19052).

References

Alcolea, A., Carrera, J., & Medina, A. (2006). Pilot points method incorporating prior information for solving the groundwater flow inverse problem. *Advances in Water Resources*, *29*(11), 1678–1689.

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(3), 269–342.

Angelino, E., Kohler, E., Waterland, A., Seltzer, M., & Adams, R. P. (2014). Accelerating MCMC, via parallel predictive prefetching. *arXiv preprint arXiv*, *1403*, 7265.

Atchadé, Y. F., Roberts, G. O., & Rosenthal, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, *21*(4), 555–568.

Bailey, R. T., Baù, D. A., & Gates, T. K. (2012). Estimating spatially-variable rate constants of denitrification in irrigated agricultural groundwater systems using an ensemble smoother. *Journal of Hydrology*, *468*, 188–202.

Bertrand, C., Ohmi, M., Suzuki, R., & Kado, H. (2001). A probabilistic solution to the MEG inverse problem via MCMC methods: The reversible jump and parallel tempering algorithms. *IEEE Transactions on Biomedical Engineering*, *48*(5), 533–542.

Blatter, D. B., Key, K., & Ray, A. (2016). Bayesian inversion of 2D models from airborne transient EM data. In *AGU Fall Meeting Abstracts*. San Francisco, CA.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Calderhead, B. (2014). A general construction for parallelizing Metropolis- Hastings algorithms. *Proceedings of the National Academy of Sciences*, *111*(49), 17,408–17,413.

Carter, J., & White, D. (2013). History matching on the Imperial College fault model using parallel tempering. *Computational Geosciences*, *17*(1), 43–65.

Chen, Y., & Zhang, D. (2006). Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Advances in Water Resources*, *29*(8), 1107–1122.

Christensen, S., & Doherty, J. (2008). Predictive error dependencies when using pilot points and singular value decomposition in groundwater model calibration. *Advances in Water Resources*, *31*(4), 674–700.

Cotter, S. L., Roberts, G. O., Stuart, A. M., & White, D. (2013). MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science*, *28*(3), 424–446.

Crestani, E., Camporese, M., Baú, D., & Salandin, P. (2013). Ensemble Kalman filter versus ensemble smoother for assessing hydraulic conductivity via tracer test data assimilation. *Hydrology and Earth System Sciences*, *17*(4), 1517.

Cui, T., Fox, C., & 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*, W10521. https://doi.org/10.1029/2010WR010352

Deutsch, C., & Journel, A. (1998). *GSLIB: Geostatistical software library and user's guide*. New York: Oxford University Press.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222.

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

Evensen, G., & Van Leeuwen, P. J. (2000). An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review*, *128*(6), 1852–1867.

Franssen, H. H., Alcolea, A., Riva, M., Bakr, M., Van der Wiel, N., Stauffer, F., & Guadagnini, A. (2009). A comparison of seven methods for the inverse modelling of groundwater flow. Application to the characterisation of well catchments. *Advances in Water Resources*, *32*(6), 851–872.

Franssen, H.-J. H., Gómez-Hernández, J., & Sahuquillo, A. (2003). Coupled inverse modelling of groundwater flow and mass transport and the worth of concentration data. *Journal of Hydrology*, *281*(4), 281–295.

Fu, J., & Gómez-Hernández, J. J. (2009). Uncertainty assessment and data worth in groundwater flow and mass transport modeling using a blocking Markov chain Monte Carlo method. *Journal of Hydrology*, *364*(3), 328–341.

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

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics*, *5*(599-608), 42.

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

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Proceedings of the 23rd Symposium on the Interface* (pp. 156–163). Boca Raton, FL.

Geyer, C. J. (1992). Practical Markov chain Monte Marlo. *Statistical Science*, 473–483.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman and Hall/CRC.

Gómez-Hernánez, J., Sahuquillo, A., & Capilla, J. (1997). Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data—1. Theory. *Journal of Hydrology (Amsterdam)*, *203*(1), 167–174.

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

Hendricks Franssen, H., & Kinzelbach, W. (2008). Real-time groundwater flow modeling with the ensemble Kalman filter: Joint estimation of states and parameters and the filter inbreeding problem. *Water Resources Research*, *44*(9).

Hu, Z., Yao, Z., & Li, J. (2017). On an adaptive preconditioned Crank-Nicolson MCMC algorithm for infinite dimensional Bayesian inference. *Journal of Computational Physics*, *332*, 492–503.

Iglesias, M., Law, K., & Stuart, A. (2012). MCMC for the evaluation of Gaussian approximations to Bayesian inverse problems in groundwater flow. *AIP Conference Proceedings* (Vol. 1479, pp. 920–923 AIP).

Laloy, E., Linde, N., Jacques, D., & 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., Rogiers, B., Vrugt, J. A., Mallants, D., & Jacques, D. (2013). Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. *Water Resources Research*, *49*, 2664–2682. https://doi.org/10.1002/wrcr.20226

Li, L., Zhou, H., Franssen, H., & Gómez-Hernández, J. (2011). Groundwater flow inverse modeling in non-multiGaussian media: Performance assessment of the normal-score ensemble Kalman filter. *Hydrology & Earth System Sciences Discussions*, *8*(4).

Makrodimitris, K., Papadopoulos, G., Philippopoulos, C., & Theodorou, D. (2002). Parallel tempering method for reconstructing isotropic and anisotropic porous media. *The Journal of Chemical Physics*, *117*(12), 5876–5884.

Martin, J., Wilcox, L. C., Burstedde, C., & Ghattas, O. (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, *34*(3), A1460–A1487.

McDonald, M. G., & Harbaugh, A. W. (1988). *A modular three-dimensional finite-difference ground-water flow model* (Vol. 6). Reston: U.S. Geological Survey.

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

Moreno, J., Katzgraber, H. G., & Hartmann, A. K. (2003). Finding low-temperature states with parallel tempering, simulated annealing and simple Monte Carlo. *International Journal of Modern Physics C*, *14*(03), 285–302.

Oliver, D., Cunha, L., & Reynolds, A. (1997). Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology*, *29*(1), 61–91.

Predescu, C., Predescu, M., & Ciobanu, C. V. (2005). On the efficiency of exchange in parallel tempering Monte Carlo simulations. *The Journal of Physical Chemistry B*, *109*(9), 4189–4196.

RamaRao, B., LaVenue, A., De Marsily, G., & Marietta, M. (1995). Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 1. Theory and computational experiments. *Water Resources Research*, *31*(3), 475–493.

Rathore, N., Chopra, M., & de Pablo, J. J. (2005). Optimal allocation of replicas in parallel tempering simulations. *The Journal of Chemical Physics*, *024*(2), 111.

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

Robert, C. P., Elvira, V., Tawn, N., & Wu, C. (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(5), e1435.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120.

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, *16*(4), 351–367.

Saley, A. D., Jardani, A., Ahmed, A. S., Raphael, A., & Dupont, J.-P. (2016). Hamiltonian Monte Carlo algorithm for the characterization of hydraulic conductivity from the heat tracing data. *Advances in Water Resources*, *97*, 120–129.

Sambridge, M. (2013). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, *196*(1), 357–374.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., & McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, *11*(2), 78–88.

Swendsen, R. H., & Wang, J.-S. (1986). Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, *57*(21), 2607.

Tarantola, A. (2005). Inverse problem theory and methods for model parameter estimation. Society of Industrial and Applied Mathematics (SIAM).

Terenin, A., Simpson, D., & Draper, D. (2015). Asynchronous Gibbs, sampling. arXiv 1509.08999.

von Neumann, J. (1951). Various techniques used in connection with random digits. Monte Carlo methods. *National Bureau of Standards AMS*, *12*, 36–38.

Valstar, J. R., McLaughlin, D. B., Te Stroet, C., & van Geer, F. C. (2004). A representer-based inverse method for groundwater flow and transport applications. *Water Resources Research*, *40*, W05116. https://doi.org/10.1029/2003WR002922

Vrugt, J. A., Braak, C. J. F. T., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*, W00B09. https://doi.org/10.1029/2007WR006720

Vrugt, J. A., Ter Braak, C. G., Diks, C. G. H., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, *10*, 271–288.

Wang, W., Machta, J., & Katzgraber, H. G. (2015). Comparing Monte Carlo methods for finding ground states of Ising spin glasses: Population annealing, simulated annealing, and parallel tempering. *Physical Review E, 92(1)*, 013303.

Wen, X., Deutsch, C., & Cullick, A. (2002). Construction of geostatistical aquifer models integrating dynamic flow and tracer data using inverse technique. *Journal of Hydrology*, *255*(1), 151–168.

Xu, T., & Gómez-Hernández, J. J. (2015a). Inverse sequential simulation: Performance and implementation details. *Advances in Water Resources*, *86*, 311–326.

Xu, T., & Gómez-Hernández, J. J. (2015b). Inverse sequential simulation: A new approach for the characterization of hydraulic conductivities demonstrated on a non-Gaussian field. *Water Resources Research*, *51*(4), 2227–2242.

Xu, T., Gómez-Hernández, J. J., Li, L., & Zhou, H. (2013). Parallelized ensemble Kalman filter for hydraulic conductivity characterization. *Computers & Geosciences*, *52*, 42–49.

Zhou, H., Gómez-Hernández, J. J., & Li, L. (2014). Inverse methods in hydrogeology: Evolution and recent trends. *Advances in Water Resources*, *63*, 22–37.

C Contribution 3: Bayesian inversion of multi-Gaussian log-conductivity fields with uncertain hyperparameters: an extension of preconditioned Crank-Nicolson Markov chain Monte Carlo with parallel tempering

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR030313

Key Points:

- The efficient preconditioned Crank-Nicolson Markov chain Monte Carlo with parallel tempering method is extended for Bayesian geostatistical inversion with uncertain hyperparameters of multi-Gaussian fields
- The model parameters are decomposed into hyperparameters and white noise to utilize the dimension robustness of pCN algorithm
- Hyperparameters cannot be constrained well with head data and direct data are important

Correspondence to
T. Xu and W. Nowak,
teng.xu@hhu.edu.cn;
wolfgang.nowak@iws.uni-stuttgart.de

Citation:
Xiao, S., Xu, T., Reuschen, S., Nowak, W., & Hendricks Franssen, H.-J. (2021). Bayesian inversion of multi-Gaussian log-conductivity fields with uncertain hyperparameters: An extension of preconditioned Crank-Nicolson Markov chain Monte Carlo with parallel tempering. *Water Resources Research*, *57*, e2021WR030313. https://doi.org/10.1029/2021WR030313

Received 18 MAY 2021
Accepted 11 AUG 2021

© 2021. The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Bayesian Inversion of Multi-Gaussian Log-Conductivity Fields With Uncertain Hyperparameters: An Extension of Preconditioned Crank-Nicolson Markov Chain Monte Carlo With Parallel Tempering

Sinan Xiao[1,2], Teng Xu[3], Sebastian Reuschen[1], Wolfgang Nowak[1], and Harrie-Jan Hendricks Franssen[2]

[1]Department of Stochastic Simulation and Safety Research for Hydrosystems, Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Stuttgart, Germany, [2]Institute of Bio- and Geosciences (IBG-3): Agrosphere, Forschungszentrum Jülich GmbH, Jülich, Germany, [3]State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China

Abstract In conventional Bayesian geostatistical inversion, specific values of hyperparameters characterizing the prior distribution of random fields are required. However, these hyperparameters are typically very uncertain in practice. Thus, it is more appropriate to consider the uncertainty of hyperparameters as well. The preconditioned Crank-Nicolson Markov chain Monte Carlo with parallel tempering (pCN-PT) has been used to efficiently solve the conventional Bayesian inversion of high-dimensional multi-Gaussian random fields. In this study, we extend pCN-PT to Bayesian inversion with uncertain hyperparameters of multi-Gaussian fields. To utilize the dimension robustness of the preconditioned Crank-Nicolson algorithm, we reconstruct the problem by decomposing the random field into hyperparameters and white noise. Then, we apply pCN-PT with a Gibbs split to this "new" problem to obtain the posterior samples of hyperparameters and white noise, and further recover the posterior samples of spatially distributed model parameters. Finally, we apply the extended pCN-PT method for estimating a finely resolved multi-Gaussian log-hydraulic conductivity field from direct data and from head data to show its effectiveness. Results indicate that the estimation of hyperparameters with hydraulic head data is very challenging and the posterior distributions of hyperparameters are only slightly narrower than the prior distributions. Direct measurements of hydraulic conductivity are needed to narrow more the posterior distribution of hyperparameters. To the best of our knowledge, this is a first accurate and fully linearization free solution to Bayesian multi-Gaussian geostatistical inversion with uncertain hyperparameters.

1. Introduction

The characterization of hydraulic properties of aquifers and soils is essential to better predict water flow in the subsurface and the transport of heat or solutes. Typically, not enough direct data (e.g., hydraulic conductivity) are available to characterize the heterogeneous subsurface. Thus, additional indirect data (e.g., hydraulic heads) are important for improving characterization and, in turn, predictions by subsurface flow and transport models. In the geostatistical context, the resulting inverse problem for subsurface problems is typically underdetermined. Therefore, approaches were developed which limited the number of independent parameters to be estimated, either by defining a limited number of zones with constant parameters (Carrera & Neuman, 1986) or parameterizing the spatially variable parameter field by a geostatistical function with a few unknown parameters (Kitanidis & Vomvoris, 1983). Later, methods were formulated to estimate a series of equally likely solutions to the groundwater inverse problem, either by an ensemble-based variational data assimilation approach (Gómez-Hernández et al., 1997) or by a sequential data assimilation approach for an ensemble of random parameter fields (Chen & Zhang, 2006). In summary, the combination of regularization and casting the problem in a stochastic framework, helped to tackle groundwater inverse problems.

Bayesian inversion has been widely used for model parameter inference (Bui-Thanh et al., 2013; Buland & Omre, 2003; Cotter et al., 2009; Mariethoz et al., 2010; Stuart, 2010). However, it is often not trivial to define the prior distribution, for example, there may not be enough information to confidently determine the prior mean and prior variance (Malinverno & Briggs, 2004). In addition, in subsurface hydrological problems, we are typically dealing with heterogeneous parameter fields (e.g., hydraulic conductivity) to be estimated, and we need to define a spatial covariance function which characterizes this spatial variability. The hyperparameters that define the spatial covariance function (e.g., nugget, range, and sill) are typically very uncertain. This uncertainty roots back to the limited amount of direct measurements of the parameter of interest, to very incomplete geological information and to the absence of geophysical surveys that could provide indirect information on spatial correlation structures. Nevertheless, the large majority of geostatistical inversion studies does not consider hyperparameter uncertainty.

A seminal cornerstone to tackle this problem was the quasi-linear geostatistical approach by Kitanidis (1995), where parameters that govern the covariance function are jointly iterated with the spatial heterogeneous field of transmissivity. This approach actually includes the restricted maximum likelihood algorithm to update the hyperparameters for the spatial covariance model. It provides a first-order estimate of hyperparameters, together with a first-order variance to specify the remaining hyperparameter uncertainty. However, most of the published studys using the quasi-linear geostatistical approach did not implement the estimation of the hyperparameters, with only few exceptions (Malinverno & Briggs, 2004; Nagel & Sudret, 2016; Nowak & Cirpka, 2006; Woodbury & Ulrych, 2000). Another geostatistical approach, the successive linear estimator (Yeh et al., 1996; Zha et al., 2018), updates the spatial covariance using the posterior covariance, which is sort of similar to the update of hyperparameters, if the inverted data are strong enough to dominate over the prior. Recently, Zhao and Luo (2021a) also used the posterior covariance for the correction of biased prior hyperparameters.

Many hydrogeological practitioners may just want the best-estimate solution to an inverse problem, not its uncertainty. Yet, as the hyperparameters govern the structure of the best estimate, it is still important to estimate them well (i.e., linearization-free). This is important due to several reasons: (a) the hyperparameters control the regularization of the best-estimate inversion result by the prior covariance, and thus should be chosen well; (b) they are the only information available about the spatial variability unresolved by the best estimate, which appears both in the estimation variance and in the structure of conditional realizations; (c) both the resolved structure and the unresolved heterogeneity are highly important for predicting scale-sensitive processes such as contaminant transport. This third reason becomes apparent in classical studies on macrodispersion (Dagan, 1988), on effective dispersion and contaminant dilution (Dentz et al., 2000), and in the fact that transport simulations on best-estimate fields clearly lack dispersion (Cirpka & Nowak, 2003; Nowak & Cirpka, 2006). One may argue whether one needs to know the values of hyperparameters as such. But one must know the structure, and the structure is governed by the hyperparameters.

Once acknowledging the relevance of inferring the structure (i.e., the hyperparameters), the joint estimation problem of hyperparameters and random field is an extended Bayesian inversion, which is also known as "hierarchical Bayes" (Gelman et al., 2013), where the hyperparameters are also uncertain. Now, the final result of Bayesian inversion will be the joint posterior distribution of model parameters and hyperparameters that govern the spatial covariance function.

When sufficient data are available, the joint posterior distribution will be dominated by the data. Several studies did synthetic test cases to show that enough information can override a "wrong fixed" choice of hyperparameters (if not too far off), especially EnKF-based studies (Li et al., 2012). However, in real-world applications, this situation will hardly appear: Research sites such as MADE (Boggs et al., 1992) or Borden Airforce Base (Sudicky, 1986) have a dense enough instrumentation, but practical sites (e.g., in remediation Troldborg et al., 2012) do not.

In most cases, it is not possible to get an analytical expression of the posterior distribution in hierarchical Bayes, especially for a nonlinear inverse problem. The above studies all rest on linearization concepts and then utilize an analytical first-order expression in an iterative scheme. At large variances and for strongly nonlinear problems, these methods have limitations. A widely used method to approximate the inverse solution is the sampling method Markov chain Monte Carlo (MCMC). Within MCMC, the random field is discretized to enable numerical computation. However, a refinement of the discretization (or an increase of the problem size), resulting in more parameters to be estimated, will usually lead to slower convergence rates of plain MCMC methods, such as the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953). In the literature, different approaches have been presented to overcome this challenge. The pre-conditioned Crank Nicolson MCMC (pCN-MCMC) (Cotter et al., 2013) is one of these approaches. Its main advantage is that the acceptance probability for proposed solutions only depends on the proposal's likelihood to match with the data, not on the proposal's position in the prior distribution. However, when pCN-MCMC is applied to nonlinear inverse problems with multi-modal posterior distributions, it may still need a long chain and a large potential space may still be missed since the MCMC simulation proceeds by local jumps in the vicinity of the current state (Robert et al., 2018). To deal with this problem, parallel tempering (Altekar et al., 2004; Earl & Deem, 2005) is a good candidate, which runs multiple chains with different temperatures simultaneously. The hot chains can more easily explore the whole parameter space, while the cold chains perform precise sampling in high-likelihood regions of the parameter space.

In a previous study (Xu et al., 2020), pCN-MCMC was combined with parallel tempering (pCN-PT) for Bayesian inversion of multi-Gaussian fields with fixed hyperparameters. The focus was to gain efficiency by combining these two ideas, while still focusing on the simpler problem with fixed hyperparameters. In this study, we will extend pCN-PT to Bayesian inversion with uncertain hyperparameters of multi-Gaussian fields. That means, our focus now is to extend the applicability of the highly efficient pCN-PT to a harder, wider and more realistic problem. If we apply pCN-MCMC or pCN-PT to the extended Bayesian inversion directly, the acceptance probability would still depend on the prior (Malinverno, 2002), which destroys the efficiency trick of pCN. To let the acceptance probability only depend on the likelihood, we will reconstruct the problem, that is, we decompose the random field model parameters into hyperparameters and white noise. The latter, after coloring through the covariance function, represents the subsurface field of hydraulic properties. Then, we have a "new" problem with hyperparameters and white noise as the primary inversion parameters. Thanks to this reconstruction, the acceptance probability is now only depending on the likelihood when using pCN-MCMC or pCN-PT. This allows to approach the hyperparameter inversion and the white noise inversion with specialized sub-algorithms within a joint iteration. In addition, this will also allow us to assess the feasibility to estimate hyperparameters. As indicated, hyperparameters are associated with significant uncertainty, and constraining this uncertainty by inversion is important. Previous inversion studies relied on linearization and it remains unclear to which degree measurement data like hydraulic conductivity and hydraulic heads carry enough information to reduce uncertainty with respect to hyperparameters. This study will also give more insight on the value of hydraulic conductivity and hydraulic head data for constraining hyperparameters with possible implications for the design of monitoring network.

The rest of this study is organized as follows. Section 2 gives the definition of the extended Bayesian inversion problem and the extended pCN-PT method (specifically in Section 2.4). In Section 3, we introduce a model setup, algorithmic settings, test cases for demonstration and testing criteria. Section 4 shows the results. Finally, Section 5 gives the conclusion.

2. Methodology

2.1. Bayesian Geostatistical Inversion With Uncertain Hyperparameters

Our forward problem can be formulated as follows:

$$\boldsymbol{d} = M(\boldsymbol{\theta}) + \boldsymbol{e}, \tag{1}$$

where M is a deterministic forward model (e.g., for groundwater flow), $\boldsymbol{\theta}$ denotes the uncertain (hydraulic) parameters for the (groundwater flow) model, $\boldsymbol{d}$ represents the measurements, and $\boldsymbol{e}$ contains measurement errors. Parameter $\boldsymbol{\theta}$ is a vector containing the discretized values of the random space function. The purpose of Bayesian geostatistical inversion is to infer the posterior distribution of uncertain parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{d}$ and prior knowledge about $\boldsymbol{\theta}$. Then, we can make a better estimation of the uncertain parameter $\boldsymbol{\theta}$.

The posterior distribution of the uncertain parameters $\boldsymbol{\theta}$ can be obtained through Bayes' rule as (Congdon, 2003)

$$p(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{d}|\boldsymbol{\theta})}{p(\boldsymbol{d})} \propto p(\boldsymbol{\theta})p(\boldsymbol{d}|\boldsymbol{\theta}), \tag{2}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of the unknown (hydraulic) parameters, which describes the prior knowledge of $\boldsymbol{\theta}$ independently of the data $\boldsymbol{d}$; $p(\boldsymbol{d}|\boldsymbol{\theta})$ is the likelihood function, which quantifies how probable are the measurement data $\boldsymbol{d}$ for a given realization of uncertain parameters $\boldsymbol{\theta}$; and $p(\boldsymbol{d})$ is the marginal likelihood.

The marginal likelihood $p(\boldsymbol{d})$ is a normalizing factor for the posterior distribution. Since the marginal likelihood is not a function of parameters $\boldsymbol{\theta}$, it is often ignored when making inference of parameters. This allows writing the posterior distribution as proportional to the product of prior and likelihood, as reflected in the last part of Equation 2 (Duijndam, 1988; Cary & Chapman, 1988).

The prior distribution of parameters $\boldsymbol{\theta}$ is dependent on some hyperparameters $\boldsymbol{h}$, such as the mean and parameters that govern a covariance function. In a conventional Bayesian geostatistical inversion, fixed values are given to $\boldsymbol{h}$. Generally, we can also let the hyperparameters have their own uncertainty and assume a prior distribution for them to reflect our prior knowledge of hyperparameters (Kitanidis, 1995; Woodbury & Ulrych, 2000). Then, we can similarly update this prior distribution and get the corresponding posterior distribution. This is an extended Bayesian inversion (Malinverno & Briggs, 2004), and we can get the joint posterior distribution of model parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{h}$ through Bayes' rule as

$$p(\boldsymbol{\theta},\boldsymbol{h}|\boldsymbol{d}) = \frac{p(\boldsymbol{\theta},\boldsymbol{h})p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h})}{p(\boldsymbol{d})} \propto p(\boldsymbol{\theta},\boldsymbol{h})p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h}), \tag{3}$$

where $p(\boldsymbol{\theta},\boldsymbol{h})$ is the joint prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{h}$, $p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h})$ is the likelihood and $p(\boldsymbol{d})$ is the marginal likelihood.

Using the definition of conditional distribution, we can write the joint prior and posterior distributions as

$$p(\boldsymbol{\theta},\boldsymbol{h}) = p(\boldsymbol{\theta}|\boldsymbol{h})p(\boldsymbol{h}), \tag{4}$$

$$p(\boldsymbol{\theta},\boldsymbol{h}|\boldsymbol{d}) = p(\boldsymbol{\theta}|\boldsymbol{h},\boldsymbol{d})p(\boldsymbol{h}|\boldsymbol{d}), \tag{5}$$

where $p(\boldsymbol{h})$ and $p(\boldsymbol{h}|\boldsymbol{d})$ are the prior and posterior distributions of $\boldsymbol{h}$ separately, while $p(\boldsymbol{\theta}|\boldsymbol{h})$ and $p(\boldsymbol{\theta}|\boldsymbol{h},\boldsymbol{d})$ are the prior and posterior distributions of $\boldsymbol{\theta}$ separately for a given choice of $\boldsymbol{h}$. We can see that the distribution of model parameters $\boldsymbol{\theta}$ depends on hyperparameters $\boldsymbol{h}$. Based on the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{h}$, we can get the marginal prior and posterior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{h}$ by marginalization. In general, the integral required for marginalization cannot be solved in closed form. This is the problem solved by Kitanidis (1995) by restricted maximum likelihood in successive linearization. Sampling methods, such as Markov chain Monte Carlo, can avoid this need for an explicit calculation of the integral, the need for restrictions and the need for linearization.

2.2. Markov Chain Monte Carlo

MCMC is an interesting method to get samples from a target probability distribution. It has been found to be well suited for Bayesian inversion problems (Besag et al., 1995; Gelman & Rubin, 1992; Grandis et al., 1999; Mosegaard & Tarantola, 1995; Schott et al., 1999; Tierney, 1994). Basically, MCMC algorithms construct a Markov chain, that is, a specific sequence of realizations of the solution where the next realization in the sequence only depends on the previous realization. For sampling, the equilibrium distribution of that chain is equal to the target distribution. To guarantee this property, the most important rule for the transition from the current realization to the next is that the so-called detailed balance condition has to be satisfied. After an initial "burn-in" period that is, influenced by the initialization of the chain (e.g., van Ravenzwaaij et al., 2018), the Markov chain samples the target distribution, that is, it provides a set of realizations following the target distribution. For Bayesian inversion problems, the posterior distribution is the target distribution. We can sample it with an MCMC and then we can easily estimate any desired properties of the posterior distribution, such as mean and variance or relevant probabilities. An important property of MCMC is its memory mechanism, which can make the Markov chain stay in the parameter space with high posterior probability (Malinverno, 2002). This property makes MCMC more efficient compared to other sampling approaches such as rejection sampling, but only if the MCMC is adapted well to the problem at hand.

At each stage with the current sample $\boldsymbol{\theta}$, one needs to propose a candidate sample $\tilde{\boldsymbol{\theta}}$ based on a proposal distribution $q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$. Then, one needs to decide whether to accept the candidate as the next sample with the following acceptance probability (Hastings, 1970)

$$\begin{aligned}\alpha &= \min\left\{\frac{p(\tilde{\boldsymbol{\theta}}|\boldsymbol{d})q(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta}|\boldsymbol{d})q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})},1\right\}\\ &= \min\left\{\frac{p(\tilde{\boldsymbol{\theta}})p(\boldsymbol{d}|\tilde{\boldsymbol{\theta}})q(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta})p(\boldsymbol{d}|\boldsymbol{\theta})q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})},1\right\}.\end{aligned} \tag{6}$$

If the candidate sample $\tilde{\boldsymbol{\theta}}$ is accepted, the next sample of the chain will be changed into $\tilde{\boldsymbol{\theta}}$. Otherwise, the chain will stay at $\boldsymbol{\theta}$.

For the case of extended Bayesian inversion, we need to extend Equation 6 from $\boldsymbol{\theta}$ to $(\boldsymbol{\theta},\boldsymbol{h})$. Then, the corresponding acceptance probability can be written as

$$\alpha = \min\left\{\frac{p(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{h}})p(\boldsymbol{d}|\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{h}})q(\boldsymbol{\theta},\boldsymbol{h}|\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{h}})}{p(\boldsymbol{\theta},\boldsymbol{h})p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h})q(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{h}}|\boldsymbol{\theta},\boldsymbol{h})},1\right\}. \tag{7}$$

Since the model parameters $\boldsymbol{\theta}$ depend on the hyperparameters $\boldsymbol{h}$, it is intuitive to divide the proposal of a candidate into two steps: first propose a candidate hyperparameter set $\tilde{\boldsymbol{h}}$ based on the current hyperparameter set $\boldsymbol{h}$, then propose a candidate model parameter field $\tilde{\boldsymbol{\theta}}$ based on $\tilde{\boldsymbol{h}}$ and on the current model parameter field $\boldsymbol{\theta}$. Thus, the proposal distribution can be written as (Malinverno, 2002)

$$q(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{h}}|\boldsymbol{\theta},\boldsymbol{h}) = q(\tilde{\boldsymbol{h}}|\boldsymbol{h})q(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{h}},\boldsymbol{\theta}). \tag{8}$$

Then, the acceptance probability can be written as

$$\alpha = \min\left\{\frac{p(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{h}})p(\tilde{\boldsymbol{h}})p(\boldsymbol{d}|\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{h}})q(\boldsymbol{h}|\tilde{\boldsymbol{h}})q(\boldsymbol{\theta}|\boldsymbol{h},\tilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta}|\boldsymbol{h})p(\boldsymbol{h})p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h})q(\tilde{\boldsymbol{h}}|\boldsymbol{h})q(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{h}},\boldsymbol{\theta})},1\right\}. \tag{9}$$

Generally, we can choose (almost) any proposal distribution q. However, different proposal distributions will affect the convergence in a given problem. Therefore, it is important to choose a proper q to have a fast convergence.

2.3. pCN-MCMC With Fixed Hyperparameters

pCN-MCMC (Cotter et al., 2013) combines the pCN proposal with the MCMC algorithm. It is an approach where the proposal automatically samples from the prior distribution if the prior is (multi-)Gaussian. A significant feature of the pCN-MCMC algorithm is its dimension robustness (Hairer et al., 2014), which makes it interesting for high-dimensional sampling problems, such as finely resolved multi-Gaussian random fields.

In pCN-MCMC, one assumes that the prior $p(\boldsymbol{\theta})$ follows a multivariate normal distribution, that is, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. For these priors, the candidate sample in pCN-MCMC can be written as

$$\tilde{\boldsymbol{\theta}} = \sqrt{1-\beta^2}(\boldsymbol{\theta}-\boldsymbol{\mu}) + \beta\xi + \boldsymbol{\mu}, \xi \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}). \tag{10}$$

The parameter β is a jumping factor that can be chosen freely (or optimized for statistical efficiency), and it follows the constraint $0 < \beta < 1$. The corresponding proposal can also be written as $q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \mathcal{N}(\sqrt{1-\beta^2}(\boldsymbol{\theta}-\boldsymbol{\mu}) + \boldsymbol{\mu}, \beta^2\boldsymbol{\Sigma})$. It fulfills the following condition

$$\frac{q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})} = \frac{p(\tilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta})}, \tag{11}$$

which confirms that it samples from the prior per construction. Inserting Equation 11 into Equation 6, the acceptance probability α can be simplified as

$$\alpha = \min\left\{\frac{p(\boldsymbol{d}|\tilde{\boldsymbol{\theta}})}{p(\boldsymbol{d}|\boldsymbol{\theta})},1\right\}. \tag{12}$$

Compared to the acceptance probability in conventional MCMC (Equation 6), the acceptance probability α with the pCN proposal (Equation 12) is only dependent on the likelihood. This contributes to a relatively large increase of efficiency, because there will be no rejections due to the prior.

2.4. Extending pCN-MCMC for Uncertain Hyperparmaeters

Now, we extend, as our key contribution, the pCN-MCMC to the case of uncertain hyperparameters while maintaining the highly efficient acceptance probability as in Equation 12. For the case of extended Bayesian inversion, we can also use the pCN proposal for $q(\tilde{\boldsymbol{h}}|\boldsymbol{h})$ and $q(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{h}},\boldsymbol{\theta})$. The pCN proposal of hyperparameters can be written as

$$q(\tilde{\boldsymbol{h}}|\boldsymbol{h}) = \mathcal{N}\left(\sqrt{1-\beta_h^2}\,(\boldsymbol{h}-\boldsymbol{\mu}_{\boldsymbol{h}}) + \boldsymbol{\mu}_{\boldsymbol{h}}, \beta_h^2\boldsymbol{\Sigma}_{\boldsymbol{h}}\right), \tag{13}$$

where $\boldsymbol{\mu}_{\boldsymbol{h}}$ is the mean of $\boldsymbol{h}$, $\boldsymbol{\Sigma}_{\boldsymbol{h}}$ is the covariance matrix of $\boldsymbol{h}$ and $0 < \beta_h < 1$ is the jumping factor. For Equation 13, we still have the following condition

$$\frac{q(\tilde{\boldsymbol{h}}|\boldsymbol{h})}{q(\boldsymbol{h}|\tilde{\boldsymbol{h}})} = \frac{p(\tilde{\boldsymbol{h}})}{p(\boldsymbol{h})}. \tag{14}$$

The pCN proposal for the model parameters can be written as

$$q(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{h}},\boldsymbol{\theta}) = \mathcal{N}(\sqrt{1-\beta_\theta^2}\,(\boldsymbol{\theta}-\tilde{\boldsymbol{h}}_\mu) + \tilde{\boldsymbol{h}}_\mu, \beta_\theta^2\tilde{\boldsymbol{h}}_{\boldsymbol{\Sigma}}), \tag{15}$$

where $\tilde{\boldsymbol{h}}_\mu$ denotes the mean components in the hyperparameters $\tilde{\boldsymbol{h}}$, $\tilde{\boldsymbol{h}}_{\boldsymbol{\Sigma}}$ denotes covariance components in the hyperparameters $\tilde{\boldsymbol{h}}$ and $0 < \beta_\theta < 1$ is the jumping factor. However, due to the difference between $\boldsymbol{h}$ and $\tilde{\boldsymbol{h}}$, we have

$$\frac{q(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{h}},\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{h},\tilde{\boldsymbol{\theta}})} \neq \frac{p(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{h}})}{p(\boldsymbol{\theta}|\boldsymbol{h})}. \tag{16}$$

Therefore, the acceptance probability will still depend on the prior. One can find more details in Malinverno (2002). To let the acceptance probability stay independent of the prior just as in Equation 12, we will do the following reconstruction.

For the model parameters $\boldsymbol{\theta}$ following a conditional multivariate Gaussian distribution (for given values of the hyperparameter $\boldsymbol{h}$), that is, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the hyperparameters), we can rewrite (Alabert, 1987; Davis, 1987)

$$\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{w}, \tag{17}$$

where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$, $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{h})$ is a matrix square root of $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{h})$, and $\boldsymbol{w}$ is a standard normal random vector ($\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, $\boldsymbol{I}$ is the identity matrix). We will denote $\boldsymbol{w}$ as white noise from here onwards. We can see that the model parameters are decomposed into their hyperparameters $\boldsymbol{h}$ (i.e., to define $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$(or $\boldsymbol{A}$)) and white noise $\boldsymbol{w}$. Then, the forward problem from Equation 1 can be rewritten as

$$\boldsymbol{d} = M(\boldsymbol{\theta}(\boldsymbol{h},\boldsymbol{w})) + \boldsymbol{e}. \tag{18}$$

Now, we can argue that $\boldsymbol{h}$ and $\boldsymbol{w}$ are the "fundamental model parameters" in Equation 18. Then, we can infer the posterior distribution of hyperparameters $\boldsymbol{h}$ and white noise $\boldsymbol{w}$ at first, and then reconstruct the posterior distribution of the original model parameters $\boldsymbol{\theta}$ based on $\boldsymbol{h}$ and $\boldsymbol{w}$ through Equation 17. In this fashion, the extended Bayesian inversion problem can also be rewritten as

$$p(\boldsymbol{h},\boldsymbol{w}|\boldsymbol{d}) \propto p(\boldsymbol{h},\boldsymbol{w})p(\boldsymbol{d}|\boldsymbol{h},\boldsymbol{w}). \tag{19}$$

Since the hyperparameters $\boldsymbol{h}$ and white noise $\boldsymbol{w}$ are independent, the joint prior distribution $p(\boldsymbol{h},\boldsymbol{w})$ can be written as

$$p(\boldsymbol{h},\boldsymbol{w}) = p(\boldsymbol{h})p(\boldsymbol{w}), \tag{20}$$

and the proposal can also be written as

$$q(\tilde{\boldsymbol{h}},\tilde{\boldsymbol{w}}|\boldsymbol{h},\boldsymbol{w}) = q(\tilde{\boldsymbol{h}}|\boldsymbol{h})q(\tilde{\boldsymbol{w}}\,|\boldsymbol{w}). \tag{21}$$

Therefore, the acceptance probability can be written as

$$\alpha = \min\left\{\frac{p(\tilde{\boldsymbol{h}})p(\tilde{\boldsymbol{w}})p(\boldsymbol{d}|\tilde{\boldsymbol{h}},\tilde{\boldsymbol{w}})q(\boldsymbol{h}|\tilde{\boldsymbol{h}})q(\boldsymbol{w}|\tilde{\boldsymbol{w}})}{p(\boldsymbol{h})p(\boldsymbol{w})p(\boldsymbol{d}|\boldsymbol{h},\boldsymbol{w})q(\tilde{\boldsymbol{h}}|\boldsymbol{h})q(\tilde{\boldsymbol{w}}|\boldsymbol{w})},1\right\}. \tag{22}$$

Now, we can use the pCN proposal for $\boldsymbol{h}$ and $\boldsymbol{w}$ separately. The pCN proposal for $\boldsymbol{h}$ is the same as in Equation 13. The pCN proposal for $\boldsymbol{w}$ can be written as

$$q(\tilde{\boldsymbol{w}}|\boldsymbol{w}) = \mathcal{N}(\sqrt{1-\beta_w^2}\,\boldsymbol{w},\beta_w^2\boldsymbol{I}), \tag{23}$$

Now, we will have

$$\frac{p(\tilde{\boldsymbol{h}})}{p(\boldsymbol{h})} = \frac{q(\tilde{\boldsymbol{h}}|\boldsymbol{h})}{q(\boldsymbol{h}|\tilde{\boldsymbol{h}})},\frac{p(\tilde{\boldsymbol{w}})}{p(\boldsymbol{w})} = \frac{q(\tilde{\boldsymbol{w}}|\boldsymbol{w})}{q(\boldsymbol{w}|\tilde{\boldsymbol{w}})}, \tag{24}$$

and then, the acceptance probability can again be simplified as

$$\alpha = \min\left\{\frac{p(\boldsymbol{d}|\tilde{\boldsymbol{h}},\tilde{\boldsymbol{w}})}{p(\boldsymbol{d}|\boldsymbol{h},\boldsymbol{w})},1\right\}. \tag{25}$$

In Equations 14 and 24, we assumed that the prior distributions of hyperparameters and white noise follow multivariate Gaussian distributions. This is true for white noise since it is i.i.d. standard normal (the simplest case of multivariate Gaussian). For the hyperparameters $\boldsymbol{h} = (h_1,h_2,\ldots,h_k)^{\mathrm{T}}$, the prior distribution may not be a Gaussian. Then, assuming that the hyperparameters are independent from each other, we can do an isoprobabilistic transformation at first, that is, $v_i = \Phi^{-1}(F_i(h_i)), i = 1,2,\ldots,k$. Here, F_i is the cumulative distribution function of h_i and Φ is the cumulative distribution function of the standard normal distribution. Then, v_i will be a random variable following the standard normal distribution, and we can just perform the pCN procedure on $\boldsymbol{v} = (v_1,v_2,\ldots,v_k)^{\mathrm{T}}$, which is also i.i.d. standard normal. The final posterior samples of hyperparameters $\boldsymbol{h}$ can be obtained based on the samples of $\boldsymbol{v}$ through the isoprobabilictic back-transformation, that is, $h_i = F_i^{-1}(\Phi(v_i))$. In the following, we assume that hyperparameters follow standard normal distribution just for simple expression.

2.5. Parallel Tempering

Parallel tempering (Altekar et al., 2004; Earl & Deem, 2005; Hukushima & Nemoto, 1996), also called Metropolis coupled MCMC, is a method for improving traditional MCMC algorithms for multi-modal distributions. In parallel tempering, multiple Markov chains are simulated at different temperatures simultaneously. The lower-temperature chains perform precise sampling in high-density regions of the parameter space, but they could easily get stuck in local modality. The higher-temperature chains can more easily explore the whole parameter space due to their flatter and broader likelihood. These different chains will regularly swap their members in a way that preserves the detailed balance condition. Therefore, the hotter chains will make sure the coldest chain with unit temperature (target chain) can access the desired regions of the parameter space. Several studies have also shown the superiority of parallel tempering compared to simple Monte Carlo, simulated annealing and simple MCMC (Moreno et al., 2003; Xu et al., 2020).

To implement parallel tempering, a temperature ladder $T_1 < T_2 < \cdots < T_n$, with $T_1 = 1$, needs to be designed at first. Then, we can get a series of tempered posterior distributions corresponding to different temperatures:

$$p_i(\boldsymbol{h},\boldsymbol{w}|\boldsymbol{d}) \propto p(\boldsymbol{h},\boldsymbol{w})p(\boldsymbol{d}|\boldsymbol{h},\boldsymbol{w})^{\frac{1}{T_i}}, i = 1,\ldots,n. \tag{26}$$

Apparently, the likelihood is taken to the power of the inverse temperature. Therefore, temperature $T_1 = 1$ corresponds to the original target posterior distribution. With increasing the temperature, the corresponding likelihood will become flatter, with a fully flat likelihood (i.e., just the prior) at the limit of $T_i \to \infty$.

Then, for the ith chain with temperature T_i, if we use the pCN proposal, the acceptance probability in Equation 25 would turn into

$$\alpha = \min\left\{\left(\frac{p(\boldsymbol{d}|\tilde{\boldsymbol{h}},\tilde{\boldsymbol{w}})}{p(\boldsymbol{d}|\boldsymbol{h},\boldsymbol{w})}\right)^{\frac{1}{T_i}},1\right\}. \tag{27}$$

For two different chains with temperatures T_i and T_j, there would be a potential swap between the states of them, that is,

$$\{(\boldsymbol{h}_i,\boldsymbol{w}_i,T_i),(\boldsymbol{h}_j,\boldsymbol{w}_j,T_j)\} \rightarrow \{(\boldsymbol{h}_j,\boldsymbol{w}_j,T_i),(\boldsymbol{h}_i,\boldsymbol{w}_i,T_j)\}, \tag{28}$$

where $(\boldsymbol{h}_i,\boldsymbol{w}_i)$ and $(\boldsymbol{h}_j,\boldsymbol{w}_j)$ are the states of two chains corresponding to the temperatures T_i and T_j before the potential swap. The acceptance probability of this potential swap can be written as

$$\alpha_s = \min\left\{\left(\frac{p(\boldsymbol{d}|\boldsymbol{h}_j,\boldsymbol{w}_j)}{p(\boldsymbol{d}|\boldsymbol{h}_i,\boldsymbol{w}_i)}\right)^{\left(\frac{1}{T_i}-\frac{1}{T_j}\right)},1\right\}. \tag{29}$$

Two candidate chains need to be determined at first to implement the potential swap. A common choice is restricting them to the neighboring chains (Earl & Deem, 2005), which will be used in this work.

2.6. Final Algorithm: Extended pCN-PT With Gibbs Split

In a recent study (Xu et al., 2020), pCN-PT (pCN-MCMC and parallel tempering) was used for the conventional Bayesian inversion with fixed values of hyperparameters. In this work, we will extend it to the extended Bayesian inversion problem with uncertain hyperparameters in Equation 19.

Since hyperparameters and white noise play different roles for generating a Gaussian random field and are a priori mutually independent (see Equation 20), we can update them separately during the MCMC process. This corresponds to a classical Gibbs sampler (Casella & George, 1992), blocked together to split between $\boldsymbol{h}$ and $\boldsymbol{w}$. It allows fine-tuning the corresponding MCMC parameters separately (here: β_h and β_w) for improved efficiency. Considering the independence between hyperparameters $\boldsymbol{h}$ and white noise $\boldsymbol{w}$, the pCN-PT algorithm can be summarized as follows:

1. Initialize a temperature ladder $T_1 < T_2 < \cdots < T_n$ with $T_1 = 1$, two pCN jumping factor ladders $\beta_{h,1},\beta_{h,2},\ldots,\beta_{h,n}$ and $\beta_{w,1},\beta_{w,2},\ldots,\beta_{w,n}$ corresponding to hyperparameters and white noise separately. Then, set initial realizations $\boldsymbol{h}_i^{(0)}$ and $\boldsymbol{w}_i^{(0)}$, where $i \in \{1,2,\ldots,n\}$ is the number of the chain in the parallel tempering.
2. Generate a pCN proposal $\tilde{\boldsymbol{h}}_i^{(k)}$ of hyperparameters at the kth sampling iteration for each chain $i = 1,2,\ldots,n$,

$$\tilde{\boldsymbol{h}}_i^{(k)} = \sqrt{1-\beta_{h,i}^2}\,\boldsymbol{h}_i^{(k)} + \beta_{h,i}\xi_h, \xi_h \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}). \tag{30}$$

3. For each chain i, accept or reject $\tilde{\boldsymbol{h}}_i^{(k)}$:

$$\boldsymbol{h}_i^{(k+1)} = \begin{cases} \tilde{\boldsymbol{h}}_i^{(k)} & \text{with probability } \alpha(\boldsymbol{h}_i^{(k)},\tilde{\boldsymbol{h}}_i^{(k)}) \\ \boldsymbol{h}_i^{(k)} & \text{otherwise} \end{cases}, \tag{31}$$

where

$$\alpha(\boldsymbol{h}_i^{(k)},\tilde{\boldsymbol{h}}_i^{(k)}) = \min\left\{\left(\frac{p(\boldsymbol{d}|\tilde{\boldsymbol{h}}_i^{(k)},\boldsymbol{w}_i^{(k)})}{p(\boldsymbol{d}|\boldsymbol{h}_i^{(k)},\boldsymbol{w}_i^{(k)})}\right)^{\frac{1}{T_i}},1\right\}. \tag{32}$$

4. Generate a pCN proposal $\tilde{\boldsymbol{w}}_i^{(k)}$ of white noise at the kth sampling iteration for each chain $i = 1,2,\ldots,n$,

$$\tilde{\boldsymbol{w}}_i^{(k)} = \sqrt{1-\beta_{w,i}^2}\,\boldsymbol{w}_i^{(k)} + \beta_{w,i}\xi_w, \xi_w \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}). \tag{33}$$

Table 1
Locations and Pumping Rates of the Four Pumping Wells

Well number	Position x (m)	Position y (m)	Pumping rate (m^3/day)
#1	500	2,350	120
#2	3,500	2,350	70
#3	2,000	3,550	90
#4	2,000	1,050	90

5. For each chain i, accept or reject $\tilde{\boldsymbol{w}}_i^{(k)}$:

$$\boldsymbol{w}_i^{(k+1)} = \begin{cases} \tilde{\boldsymbol{w}}_i^{(k)} & \text{with probability } \alpha\left(\boldsymbol{w}_i^{(k)}, \tilde{\boldsymbol{w}}_i^{(k)}\right) \\ \boldsymbol{w}_i^{(k)} & \text{otherwise} \end{cases}, \tag{34}$$

where

$$\alpha\left(\boldsymbol{w}_i^{(k)}, \tilde{\boldsymbol{w}}_i^{(k)}\right) = \min\left\{\left(\frac{p\left(\boldsymbol{d} \mid \boldsymbol{h}_i^{(k+1)}, \tilde{\boldsymbol{w}}_i^{(k)}\right)}{p\left(\boldsymbol{d} \mid \boldsymbol{h}_i^{(k+1)}, \boldsymbol{w}_i^{(k)}\right)}\right)^{\frac{1}{T_i}}, 1\right\}. \tag{35}$$

6. For the neighboring chains i and j, swap values between them $\left(\boldsymbol{h}_i^{(k)}, \boldsymbol{w}_i^{(k)}\right) \rightleftharpoons \left(\boldsymbol{h}_j^{(k)}, \boldsymbol{w}_j^{(k)}\right)$ with swap acceptance probability

$$\alpha_s = \min\left\{\left(\frac{p\left(\boldsymbol{d} \mid \boldsymbol{h}_j^{(k)}, \boldsymbol{w}_j^{(k)}\right)}{p\left(\boldsymbol{d} \mid \boldsymbol{h}_i^{(k)}, \boldsymbol{w}_i^{(k)}\right)}\right)^{\left(\frac{1}{T_i} - \frac{1}{T_j}\right)}, 1\right\}. \tag{36}$$

7. $k \rightarrow k+1$ and restart at step 2.

The tuning parameters of the pCN-PT algorithm are the number of chains, the temperature ladder and the two jumping factor ladders (the latter for both $\boldsymbol{h}$ and $\boldsymbol{w}$, separately). Details about dealing with these tuning parameters are shown in Section 3.2.

As intuitive illustration, step 2 uses a change in covariance to morph the given field $\boldsymbol{\theta}$, step 3 tests the morphed field, step 4 tries out a new field with the same covariance, and step 5 tests the new field. The overall chain stores a new realization after an attempted morph and innovation. This also provides an option to extend other MCMC algorithms for the extended Bayesian inversion.

3. Application

3.1. Model Setup

We consider fully saturated, steady-state groundwater flow as test case, which is an extension of the test case of Xu et al. (2020). The flow equation can be written as

$$\nabla \cdot (K \nabla H) + S = 0, \tag{37}$$

where $\nabla \cdot$ denotes the divergence operator, K is the hydraulic conductivity (m/day), ∇ is the Nabla operator, H denotes hydraulic head (m), and S denotes the source/sink term as volumetric injection flow rate per unit volume of aquifer (1/day). The flow equation is solved with the groundwater flow simulator MODFLOW (McDonald & Harbaugh, 1988).

We consider a synthetic confined, two-dimensional aquifer in a 5,000 m × 5,000 m domain with 50 m thickness. The domain is discretized into 100 × 100 × 1 cells with cell sizes of 50 m × 50 m × 50 m. The west and east boundaries are specified head boundaries with heads fixed to 20 and 0 m, respectively. The north and the south boundaries are impermeable. There are four pumping wells in the domain. The corresponding locations and pumping rates are listed in Table 1. A reference ln-conductivity field is generated based on a multi-Gaussian random field with the parameters listed in Table 2 (Xu et al., 2020). The reference ln-conductivity field and hydraulic head solution (obtained by MODFLOW) are shown in Figure 1. We will consider the ln-conductivity as the model parameter $\boldsymbol{\theta}$, which has a prior multivariate Gaussian distribution (conditionally on $\boldsymbol{h}$) with exponential covariance function. The following hyperparameters $\boldsymbol{h}$ will also be considered as uncertain: (a) mean μ; (b) standard deviation σ; (c) correlation length λ_1; (d) correlation length λ_2. The angle of anisotropy is fixed as in Table 2. This extends the test cases already used by Xu et al. (2020) toward uncertain hyperparameters, while using the same reference field.

Table 2
Parameters of the Gaussian Random Field Generating the Reference Ln-Conductivity Field

Variogram type	μ (m/day)	σ (m/day)	λ_1 (m)	λ_2 (m)	Angle (deg)
Exponential	−2.5	2	2,000	1,500	135

Note. μ is the mean, σ is the standard deviation, λ_1 and λ_2 are the correlation lengths in the x and y directions.

3.2. Algorithmic Settings and Adaptivity

Previous studies have shown that the optimal acceptance rate for MCMC-based methods is close to 23.4% and that acceptance rates between 10% and 40% still perform close to optimal (Gelman et al., 1996; Roberts & Rosenthal, 2001; Roberts et al., 1997). The parameter β in the pCN proposal affects the acceptance rate of each chain. Basically, high (low) values of β will lead to low (high) acceptance rate. Therefore, to obtain an acceptance rate close to the optimal one, we will use an adaptive way to adjust the values of β during the "burn-in" period. First, we set an initial value for each β. Then, we estimate the acceptance rate by running the Markov chain for N_a steps without changing β. If the acceptance rate is too high (low), we will try to increase (decrease) the value of β a little. Thus, we can set an initial value, for example, 0.5, for all jumping factors. Then, through repeating this procedure several times, we can get an acceptance rate close to the optimal one. Here, the β values for hyperparameters and white noise are adjusted based on their own acceptance rates separately. To adjust β values adaptively, we set $N_a = 10^3$. The first 10^5 steps in each Markov chain are used for adjusting, which means the adjusting procedure above is repeated 100 times.

In addition, the acceptance swap rate is also an important feature for parallel tempering. The optimal acceptance swap rate differs with respect to the specific applications (Laloy et al., 2016; Predescu et al., 2005; Rathore et al., 2005). In this study, we will control the acceptance swap rate between 10% and 30% just as in the previous work (Xu et al., 2020). The temperature ladder will affect the swap acceptance rate. A smaller (larger) distance between the neighboring temperatures will lead to a higher (lower) swap acceptance rate. Therefore, we will also adjust the temperature ladder adaptively similar to that for adjusting β during the "burn-in" period. First, we set an initial temperature ladder and run the Markov chain for N_a steps. Then we can estimate the current acceptance rate. If it is too high (low), we will try to increase (decrease) the distance between the neighboring temperatures a little. After repeating this procedure several times, we can obtain an acceptance swap rate between 10% and 30%. Similarly, we set $N_a = 10^3$ and use the first 10^5 steps in each Markov chain for adjusting.

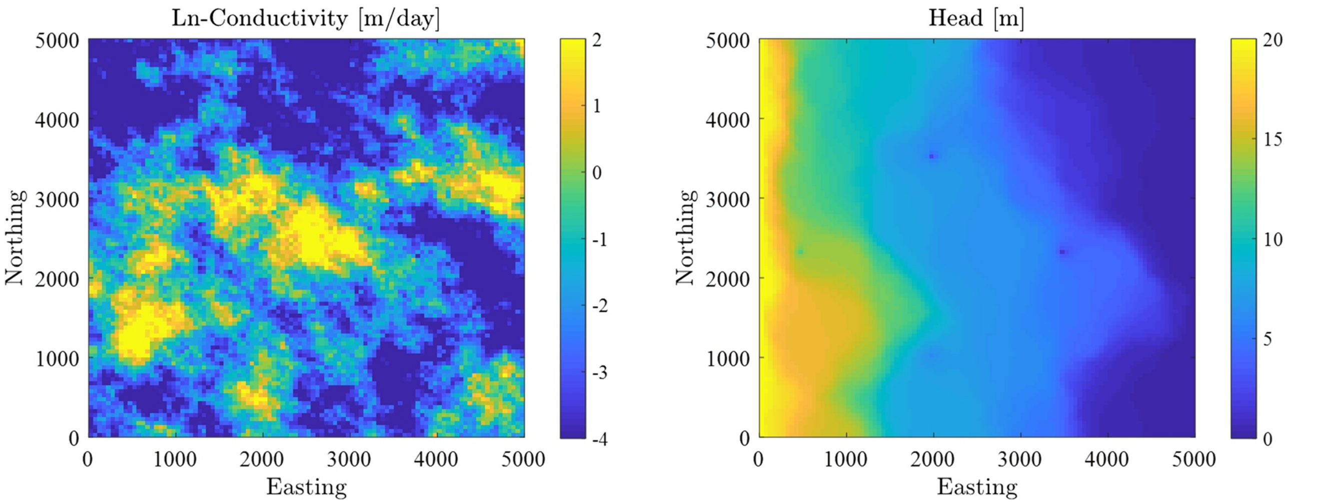


Figure 1. Reference ln-conductivity field (left) and reference hydraulic head solution (right).

Table 3
Computational Time to Generate One Realization of a Gaussian Random Field (Zero Mean, Unit Variance, Exponential Covariance Function) on a Domain Discretized Into 100 × 100 Cells

Method	Cholesky decomposition	Circulant embedding
Time (s)	4.542	0.128

For each test case presented in the upcoming section, we will run 20 parallel chains. Each chain runs independently on a dedicated computing node, and data commutation only occurs for between-chain swaps.

For each pair of hyperparameters $\boldsymbol{h}$ and white noise $\boldsymbol{w}$, we need to combine them to get the original model parameters (Equation 17), and then we can calculate the model response and likelihood. A traditional way to achieve this purpose is the Cholesky decomposition of the covariance matrix (Alabert, 1987; Davis, 1987), which is simple and can handle any grid structure. However, it is restricted to moderate numbers of discrete cells for the random field. For problems with finer resolution (100 × 100 in this work), the Cholesky decomposition will be very slow. For a case with n discrete cells, the best algorithm for Cholesky decomposition has a complexity of $\mathcal{O}(n^2)$ (Dietrich & Newsam, 1997).

In this study, we will use the circulant embedding approach (Dietrich & Newsam, 1993, 1997) as implemented in Fritz et al. (2009). Although this method is only suitable for regularly meshed grids, it is more computationally efficient compared to the Cholesky decomposition. The computational complexity of the circulant embedding approach is only $\mathcal{O}(n \log n)$. An example of the computational cost of these two approaches is shown in Table 3, which is performed with MATLAB 2018a in a computer with Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz and 32 GB RAM. We can see that the circulant embedding approach is more than 30 times faster than the Cholesky decomposition approach in our setup. For larger fields, this advantage will grow due to the almost-linear complexity of the FFT. A recent study (Nowak & Litvinenko, 2013) has even upgraded FFT-based geostatistics to safely handle up to 1 million voxels per direction in 3D space (i.e., 10^{18} cells).

Options for irregular grids include, for example, the Karhunen-Loeve decomposition (which relies on a numerical eigendecomposition of the auto-covariance matrix without the circulant embedding and hence without FFT assistance). While this sounds much slower, the Karhunen-Loeve decomposition is usually truncated, such as in the dimension-reduced approach by Zhao and Luo (2021b), and it could be truncated early if desired. Yet another option is to re-parameterize, on a coarser grid, the random field through pilot points (Doherty et al., 2010), followed by interpolation or conditional simulation in between (Keller et al., 2021).

3.3. Test Cases

In this section, we will consider four different test cases. For all test cases, we use the following likelihood

$$p(\boldsymbol{d}|\boldsymbol{\theta}(\boldsymbol{h},\boldsymbol{w})) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{d} - M(\boldsymbol{\theta}(\boldsymbol{h},\boldsymbol{w})))^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{e}}^{-1}(\boldsymbol{d} - M(\boldsymbol{\theta}(\boldsymbol{h},\boldsymbol{w})))\right)}{\sqrt{(2\pi)^{n_d}\det(\boldsymbol{\Sigma}_{\mathbf{e}})}}, \tag{38}$$

where $\boldsymbol{\Sigma}_{\mathbf{e}}$ is the covariance matrix of the error term $\boldsymbol{e}$, and n_d is the number of measurements. In this study, the measurement standard deviation is set to $\sigma_e = 0.05$, that is, $\boldsymbol{\Sigma}_{\mathbf{e}}$ is a diagonal matrix with constant diagonal elements σ_e^2. All test cases use the exponential model for the covariance function with hyperparameters μ (for the mean), σ^2 (for the variance), and λ_1, λ_2 (for the length scales). Further comments on the problem of covariance-model choice are provided in out last section.

In test cases 1, 2, and 3, we will directly use ln-conductivity data, which will lead to a high-dimensional linear estimation problem. For this situation, we can analytically get the marginal likelihood function $p(\boldsymbol{d}|\boldsymbol{h})$ for the direct inference of hyperparameters. Thus, we can easily get posterior samples of the hyperparameters with MCMC at first. Then, we can use kriging to interpolate the ln-conductivity field for each given posterior sample of hyperparameters and finally obtain the average result quasi-analytically. These results can be used for very precise comparison. The details can be found in Appendix A. For test Case 4, we will use hydraulic head data, which will lead to a high-dimensional nonlinear inverse estimation problem closer to reality.

Table 4
Settings for Each Test Case

Test case	Data type	Iterations	Iterations for parameter tuning	Methods
T1	25 ln-conductivity	1,200,000	100,000	pCN-PT & AML-Kriging
T2	25 ln-conductivity	1,200,000	100,000	pCN-PT & AML-Kriging
T3	500 ln-conductivity	1,600,000	100,000	pCN-PT & AML-Kriging
T4	25 head	1,200,000	100,000	pCN-PT

Abbreviation: AML, analytical marginal likelihood; pCN-PT, preconditioned Crank-Nicolson-parallel tempering.

The data is taken from the reference ln-conductivity field and hydraulic head solution in Figure 1 plus an artificial error (random value from a normal distribution with mean zero and standard deviation 0.05) at the predefined locations. A summary of the settings for each test case can be found in Table 4.

3.3.1. Test Case 1

In the first test case, we consider 25 direct measurements of ln-conductivity (no head data) at the positions marked with red circles in Figure 2. Among these 25 measurements, 16 measurements are uniformly distributed and the remaining nine measurements are randomly distributed. For the hyperparameters, we assume the following prior distributions

$$\begin{aligned} \mu &\sim \mathcal{U}(-10,10), \\ \sigma &\sim \mathcal{U}(0,20), \\ \lambda_1 &\sim \mathcal{U}(0,5000), \\ \lambda_2 &\sim \mathcal{U}(0,5000). \end{aligned} \tag{39}$$

A constraint for the anisotropic ratio between the two correlation lengths λ_1 and λ_2 is also considered as $0.1 < \lambda_1/\lambda_2 < 10$. In this test case, we use uniform distributions for the hyperparameters over a broad interval to assume that we only have weak prior information about the hyperparameters. This tests the exploration capabilities of our proposed algorithm, while allowing to compare to a quasi-analytical solution.

3.3.2. Test Case 2

In the second test case, we still consider the same measurements of ln-conductivity as in test Case 1 so that we can again compare to quasi-analytical results. But for the hyperparameters, we now assume the following prior distributions

$$\begin{aligned} \mu &\sim \mathcal{N}(-4,1), \\ \log_2 \sigma &\sim \mathcal{N}(2,0.5), \\ \log_{10} \lambda_1 &\sim \mathcal{N}(3,0.5), \\ \log_{10} \lambda_2 &\sim \mathcal{N}(3,0.5). \end{aligned} \tag{40}$$

The same constraint $0.1 < \lambda_1/\lambda_2 < 10$ for the anisotropic ratio is adopted. In this test case, we use (log)normal distributions for the hyperparameters to assume that we have more prior information about the hyperparameters. This allows testing the exploitation capability of our proposed algorithm in a much more restricted prior setting.

3.3.3. Test Case 3

In the third test case, we consider 500 direct measurements of ln-conductivity at the positions marked with red circles in Figure 3. Among these 500 measurements, 400 measurements are uniformly distributed and the remaining 100 measurements are randomly distributed. The prior distributions of the hyperparameters are the same as those in test Case 2 with the same constraint for the anisotropic ratio. The large number of measurement data enhances test Case 2 by enforcing a highly restrictive likelihood function. This is the last test case against quasi-analytical results.

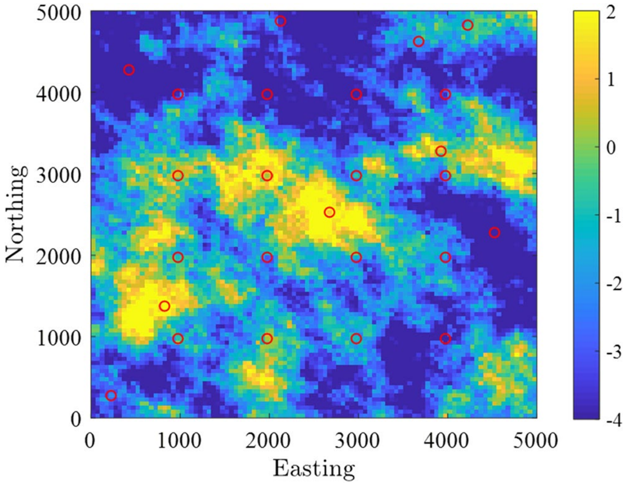


Figure 2. Twenty-five ln-conductivity measurement locations (red circles) used in test cases 1 and 2.

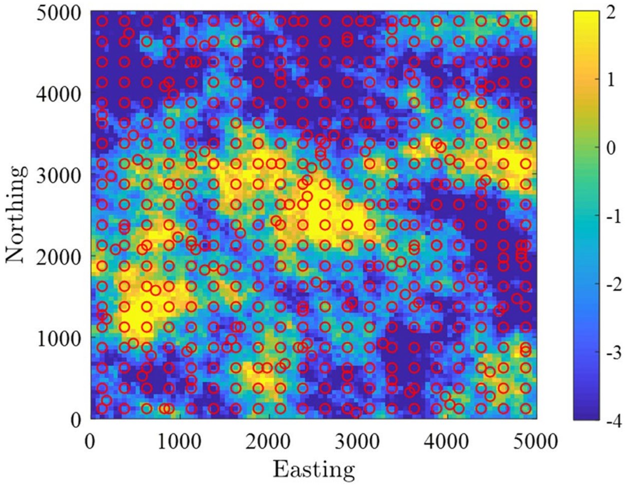


Figure 3. Five hundred ln-conductivity measurement locations (red circles) used in test Case 3.

3.3.4. Test Case 4

In the fourth test case, we consider 25 head measurements as used in Xu et al. (2020). The positions (red circles) of these measurements are shown in Figure 4. The prior distributions of the hyperparameters are still (almost) the same as those in test Case 2 with the same constraint for the anisotropic ratio. The only difference is that we truncate the log-normal distributions of the correlation lengths λ_1 and λ_2 at $\lambda_1 = 4500$ m and $\lambda_2 = 4500$ m. The reason is that the log-normal distribution has no upper bound, head data are very weakly informative for scales due to their diffuse character, and numerical problems could arise within the FFT-based algorithm (Fritz et al., 2009) at correlation length values close to (or above) the domain size. As no quasi-analytical reference is available for test Case 4, we run two independent repetitions and assess convergence by comparison between these repetitions based on the potential scale reduction factor R.

3.4. Testing Criteria

We compare the posterior samples or posterior mean to the synthetic reference. For the test cases 1, 2 and 3, the posterior mean and standard deviation of the ln-conductivity field will also be compared to the quasi-analytical kriging results, and the posterior distributions of hyperparameters will also be compared to the results obtained by MCMC with analytical marginal likelihood. We will also evaluate the difference between the posterior samples of the ln-conductivity field and the synthetic reference field through the root mean square error (RMSE)

$$RMSE^{(i)} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{\theta}_j^{(i)} - \boldsymbol{\theta}_j^{ref}\right)^2}, \tag{41}$$

where N is the number of cells in the entire domain, $\boldsymbol{\theta}_j^{(i)}$ is the value of posterior estimate of ln-conductivity at the jth cell in the ith iteration, $\boldsymbol{\theta}_j^{ref}$ is the value of ln-conductivity at the same grid cell for the reference field.

Second, we will use ln-likelihood to measure the closeness of the posterior samples to the synthetic data, and we will just take $\ln(p(\boldsymbol{d}|\boldsymbol{\theta}(\boldsymbol{h},\boldsymbol{w})))$ based on Equation 38.

Third, we will evaluate the convergence of the Markov chain by the potential scale reduction factor R introduced by Gelman and Rubin (1992). A R value less than 1.2 is considered as an acceptable convergence (Brooks & Gelman, 1998). The factor R is defined for a set of m Markov chains, each of which has n samples. The within-chain variance (for each pixel separately) is estimated as

$$W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\boldsymbol{\theta}_j^{(i)} - \overline{\boldsymbol{\theta}}_j\right)^2, \tag{42}$$

where $\boldsymbol{\theta}_j^{(i)}$ is the ith sample of the jth chain and $\overline{\boldsymbol{\theta}}_j$ is the mean of the samples in the jth chain (for each pixel separately). The pixel-wise between-chain variance is estimated as

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\overline{\boldsymbol{\theta}}_j - \frac{1}{m}\sum_{j=1}^{m}\overline{\boldsymbol{\theta}}_j\right)^2. \tag{43}$$

Then, the estimated variance V is a weighted average of the within-chain variance and between-chain variance:

$$V = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B \tag{44}$$

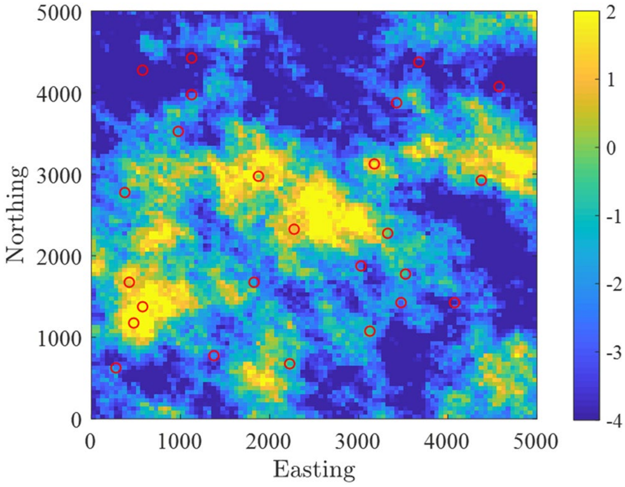


Figure 4. Twenty five head measurement locations (red circles) used in test Case 4.

Finally, the pixel-wise potential scale reduction factor R is defined as

$$R = \sqrt{\frac{V}{W}}. \tag{45}$$

We compute R for ln-conductivity in each model cell and for all hyperparameters. For ln-conductivity, hence, we can show a map of R, and summarize the map by its mean and maximum.

3.5. Obtaining the (Reference) Solutions

For test cases 1, 2, and 3, when using the analytical marginal likelihood (AML) for direct inference of hyperparameters as reference solution, we only have a 4-dimensional small problem. Therefore, we generate 1 million posterior samples of hyperparameters by pCN-MCMC (running only on the Gaussian-transformed hyperparameters $\boldsymbol{h}$) and assume these samples are enough to obtain an accurate estimation of the posterior distribution as reference solution. These posterior samples are saved every 10th iteration and the first half of them are discarded as a "burn-in" period. The second half of these samples are used to obtain the average (conditional) kriging results of the ln-conductivity field (denoted as AML-Kriging) and to plot the posterior distribution of hyperparameters (denoted as pCN-AML).

For the proposed pCN-PT method, 1.2 million posterior samples are generated for test cases 1, 2, and 4, 1.6 million posterior samples are generated for test Case 3. Similarly, these samples are saved every 10th iteration and the first half are discarded as a "burn-in" period. The second half of these samples are used to calculate the mean and standard deviation of the ln-conductivity field, plot the posterior distribution of hyperparameters, and estimate the potential scale reduction factor R.

4. Results

4.1. Test Case 1 (T1)

Figure 5 shows the posterior mean and standard deviation of the ln-conductivity field for test Case 1 obtained by pCN-PT (top row) and AML-Kriging (bottom row). We can see that both results look very similar and the mean of the results captures the main features of the reference field. Additionally, the standard deviation around the measurement points is close to zero and the area with small standard deviation obtained by pCN-PT is only a little larger than that obtained by kriging. To quantitatively measure the closeness between these two results, we calculate the RMSE between them, see Table 5. The RMSE between these two results is 0.195 and 0.135 for mean and standard deviation of the ln-conductivity field. We also calculate the root mean square of the mean and standard deviation of the ln-conductivity field obtained by AML-Kriging, which is 2.607 and 2.185. Thus, we can see that the RMSE between the two results is small compared to the root mean square of the reference solution. These results show that pCN-PT is capable of estimating the hyperparameters and (conditionally) multi-Gaussian ln-conductivity field in a high-dimensional linear problem with limited prior information on hyperparameters.

Figure 6 shows the marginal posterior and prior distributions of the hyperparameters for test Case 1 obtained by pCN-PT and pCN-AML. We can see that the posterior distributions of hyperparameters μ and σ obtained by pCN-PT are very close to the reference distributions obtained by pCN-AML. For hyperparameters λ_1 and λ_2, although there are still some differences between the posterior distributions obtained by pCN-PT and pCN-AML, their shapes are similar, that is, the posterior PDF increases significantly for small values of λ_1 (λ_2) and decreases gradually until the upper bound of λ_1 (λ_2). In addition, we also use the Kolmogorov-Smirnov statistic (maximum difference between the cumulative distribution functions in [0,1]) to quantitatively compare the marginal posterior distributions of hyperparameters obtained by pCN-PT and pCN-AML, and the results are shown in Table 6. It also shows small differences between these two results. Based on these results, we can see that pCN-PT can obtain a reliable estimate of the posterior distribution of the hyperparameters in a high-dimensional linear problem when there is only little prior information.

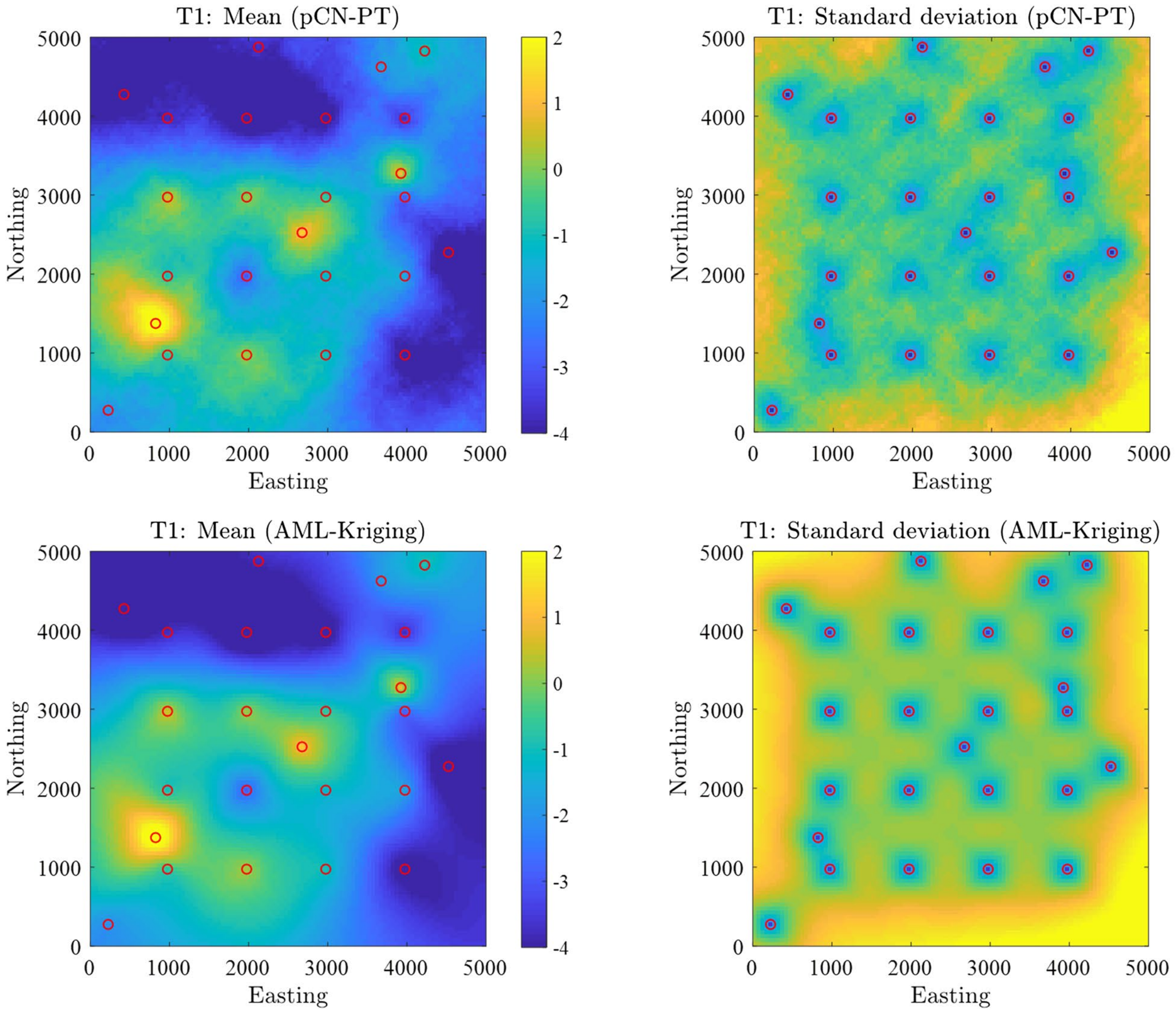


Figure 5. Posterior mean and standard deviation of the ln-conductivity field for test Case 1. Top row: results obtained by preconditioned Crank-Nicolson Markov-parallel tempering. Bottom row: results obtained by analytical marginal likelihood-Kriging.

Figure 7 shows the ln-likelihood, RMSE, and potential scale reduction factors *R* of the ln-conductivity realizations and hyperparameters for test Case 1 obtained by pCN-PT along the MCMC chain. According to the results of ln-likelihood and RMSE, we can see that the 'burn-in' and convergence to the reference is very fast. Observing the potential scale reduction factors *R*, we can see that these factor values decrease as the length of the Markov chain increases and finally drop below the recommended value (Brooks & Gelman, 1998) of 1.2. We can also see that all the final pixel-wise potential scale reduction factor values on the map are close to 1. This shows a good convergence of the results obtained by pCN-PT.

Table 5
Root Mean Square Error (RMSE) Between the Results (Mean and Standard Deviation of Ln-Conductivity Field) Obtained by Preconditioned Crank-Nicolson-Parallel Tempering and Analytical Marginal Likelihood-Kriging for Test Cases 1, 2, and 3

Test cases	RMSE of mean	RMSE of standard deviation
T1	0.195	0.135
T2	0.163	0.122
T3	0.403	0.233

4.2. Test Case 2 (T2)

Figure 8 shows the posterior mean and standard deviation of the ln-conductivity field for test Case 2 obtained by pCN-PT (top row) and AML-Kriging (bottom row). We can see that both solutions have a very similar appearance. The RMSE between the two results is 0.163 and 0.122, while the root mean squares of the results obtained by AML-Kriging (reference solution) are 2.605 and 2.315. Again, this indicates small

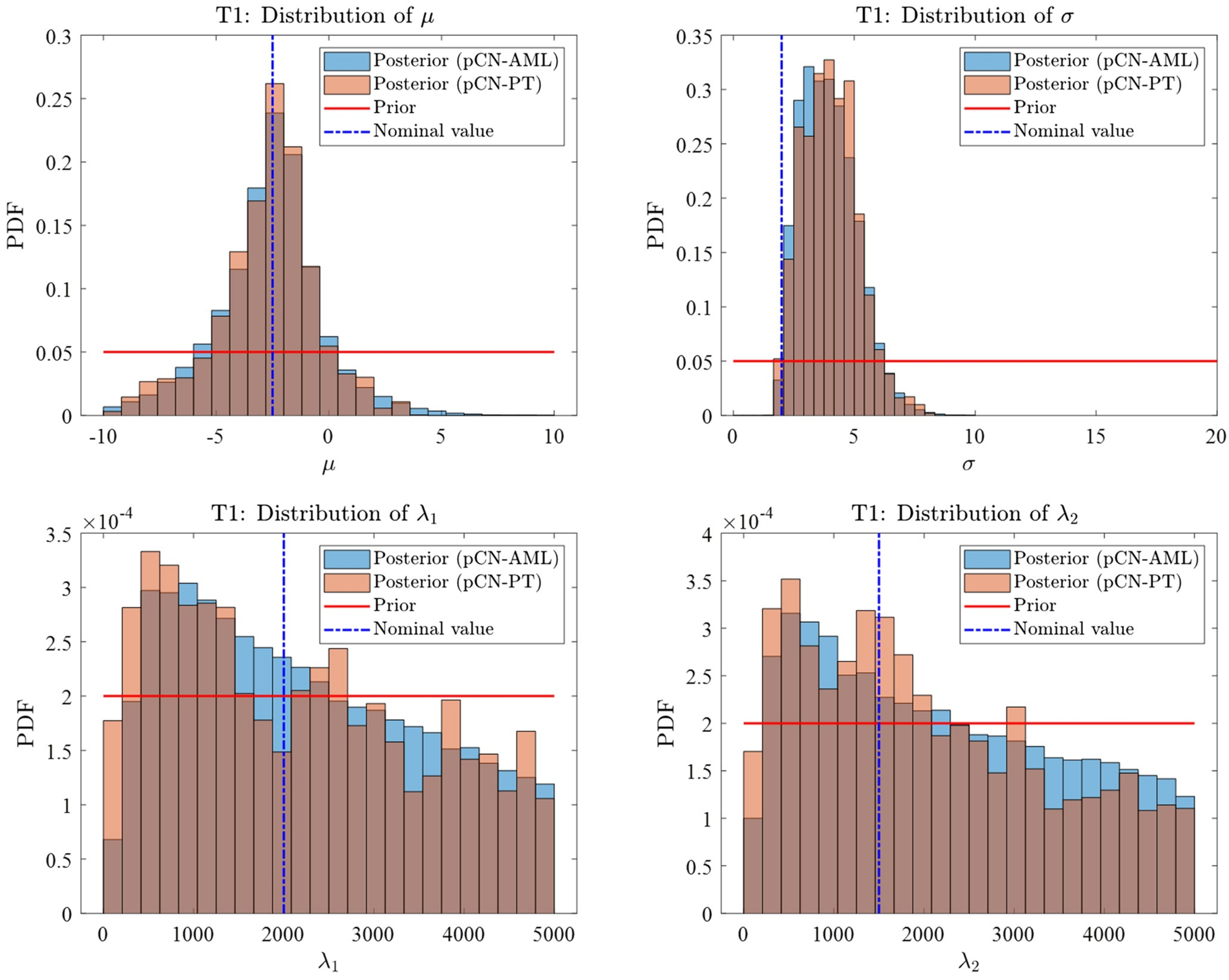


Figure 6. Marginal posterior (colored bars) and prior (horizontal lines) distributions of the hyperparameters for test Case 1. The vertical dot-dashed lines denote the nominal values of the hyperparameters used to generate the synthetic data.

RMSE and shows the similarity between these two results. Now, we see that pCN-PT is also able to estimate the multi-Gaussian ln-conductivity field in a high-dimensional linear problem with more prior information about the hyperparameters. In addition, these results are also very close to the results of test Case 1. This indicates that, although different prior knowledge of hyperparameters is assumed in test Case 1 and 2, we can still get similar results for the ln-conductivity field based on the same data set.

Table 6
Quantitative Comparison of Marginal Posterior Distributions of Hyperparameters Obtained by Preconditioned Crank-Nicolson-Parallel Tempering and Preconditioned Crank-Nicolson-Analytical Marginal Likelihood Through Kolmogorov-Smirnov Statistic

	Kolmogorov-Smirnov statistic			
Test cases	μ	σ	λ_1	λ_2
T1	0.023	0.046	0.055	0.065
T2	0.042	0.015	0.070	0.091
T3	0.217	0.255	0.195	0.211

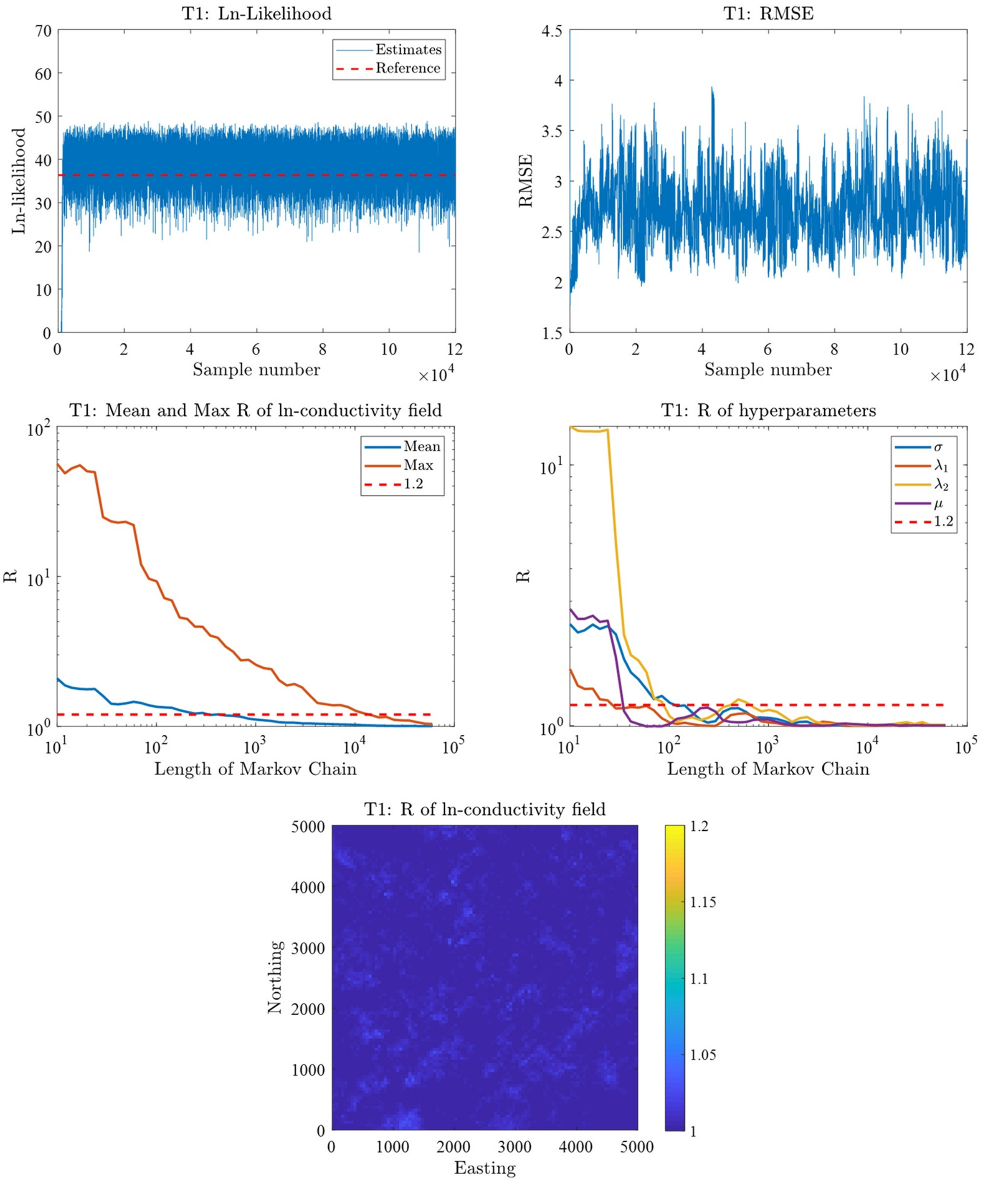


Figure 7. Testing criteria for test Case 1: Ln-likelihood of the posterior samples of the ln-conductivity field, root mean square error between the posterior samples of the ln-conductivity field and the synthetic reference field, the evolution of mean and max potential scale reduction factor of the ln-conductivity field, the evolution of the potential scale reduction factor of the hyperparameters, and the final potential scale reduction factor of the ln-conductivity field obtained by preconditioned Crank-Nicolson-parallel tempering. The dashed line in the figure of ln-conductivity corresponds to the ln-likelihood of the reference. The dashed lines in the evolution of the potential scale reduction factor denote the value of 1.2.

The marginal posterior and prior distributions of hyperparameters for test Case 2 obtained by pCN-PT and pCN-AML are shown in Figure 9. It confirms that pCN-PT provides similar results compared to the reference results obtained by pCN-AML also in this case. The Kolmogorov-Smirnov statistic in Table 6 again confirms the similarity between these two results. This denotes that pCN-PT can also estimate reliably the posterior distribution of the hyperparameters and conductivity field in problems with more prior information.

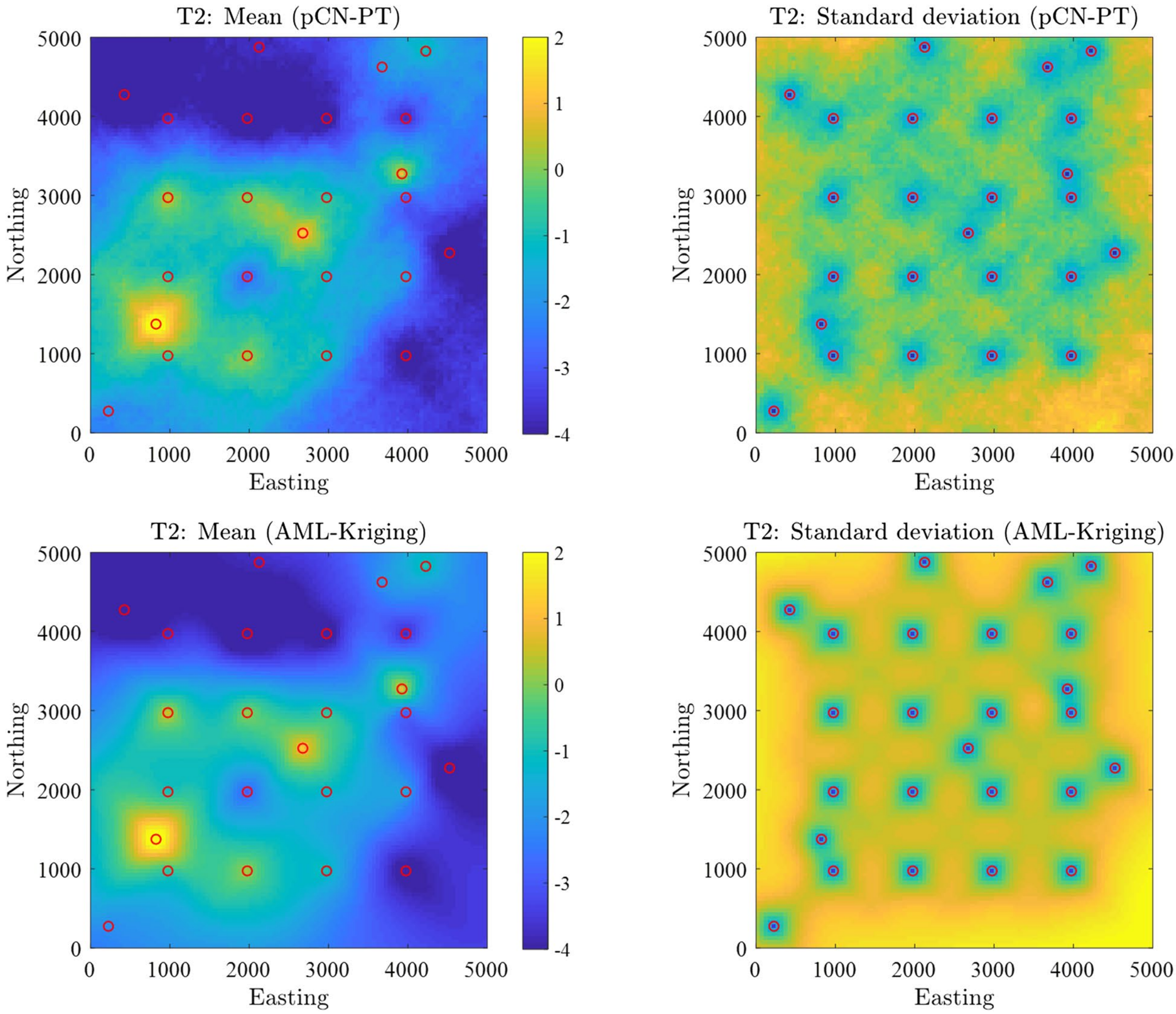


Figure 8. Posterior mean and standard deviation of the ln-conductivity field for test Case 2. Top row: results obtained by preconditioned Crank-Nicolson-parallel tempering. Bottom row: results obtained by analytical marginal likelihood-Kriging.

Figure 10 shows the ln-likelihood, RMSE, and pixel-wise potential scale reduction factors R of the ln-conductivity field and the hyperparameters for test Case 2 obtained by pCN-PT. We can see that the Markov chain converges very fast based on the results of ln-likelihood and RMSE. The potential scale reduction factor values also decrease as the length of Markov chain increases. We can see that all the final potential scale reduction factor values are very close to 1, which indicates a good convergence of the results obtained by pCN-PT also for test Case 2 with a more constraining prior.

4.3. Test Case 3 (T3)

Figure 11 shows the posterior mean and standard deviation of the ln-conductivity field for test Case 3 obtained by pCN-PT (top row) and AML-Kriging (bottom row). Compared to test cases 1 and 2, more (500) measurements are used in test Case 3. First, we can see that both results are very close to each other. The RMSE in Table 5 also confirms the similarity between these two results (the root mean square of the results obtained by kriging is 2.7757 and 0.8657). This indicates that pCN-PT still has a good performance with a large number of data (highly restrictive likelihood function). Second, we can see that the posterior mean is also very close to the reference ln-conductivity field. This follows the plain expectation that a fully Bayesian

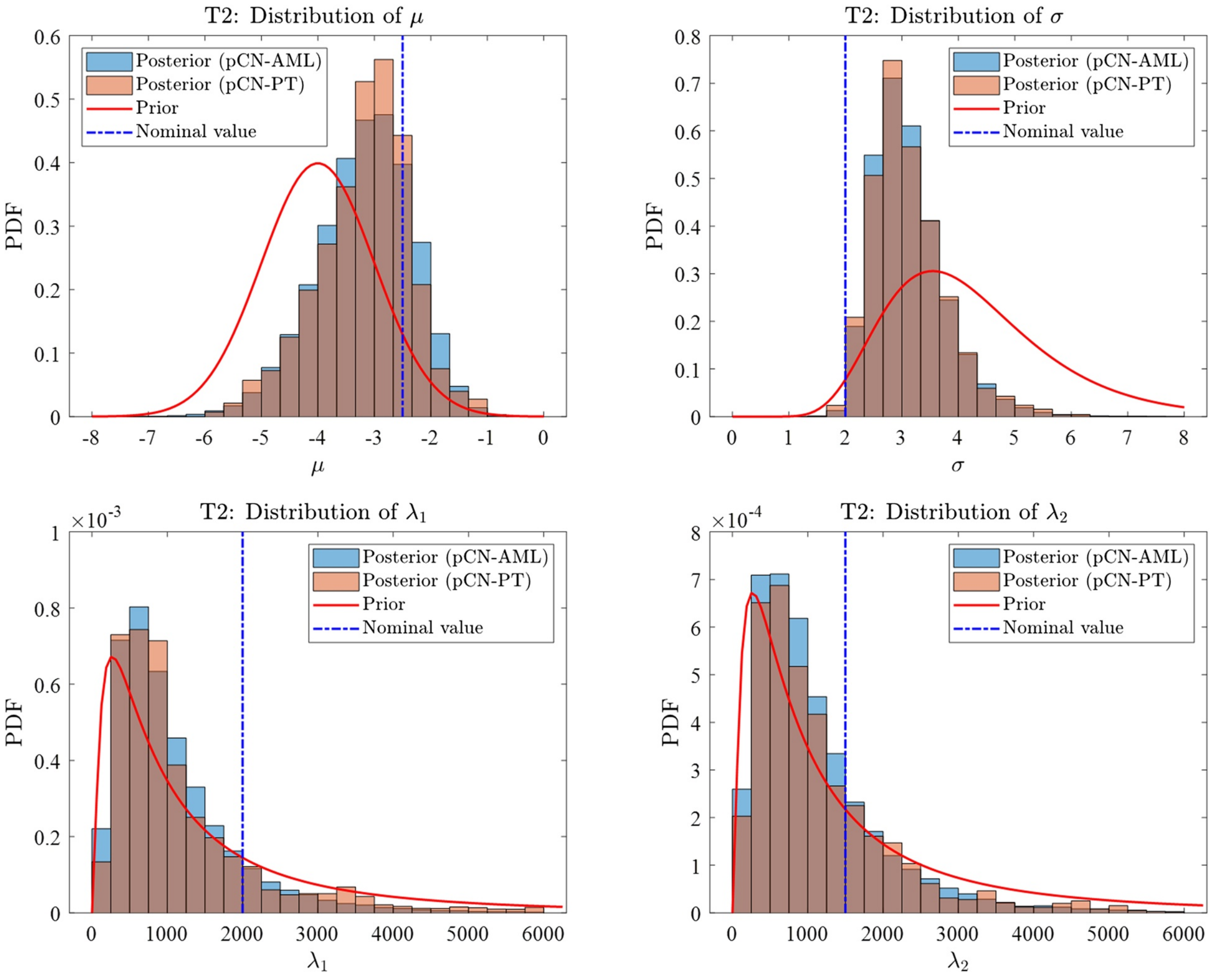


Figure 9. Marginal posterior (colored bars) and prior (red curves) distributions of the hyperparameters for test Case 2. The vertical dot-dashed lines denote the nominal values of the hyperparameters.

inversion executed via pCN-PT can recover the multi-Gaussian ln-conductivity field with a large number of (direct) measurements.

Figure 12 shows the marginal posterior and prior distributions of hyperparameters for test Case 3 obtained by pCN-PT and pCN-AML. We find again a large overlap between the posterior distributions obtained by pCN-PT and pCN-AML. The Kolmogorov-Smirnov statistic between these two results is also shown in Table 6. Compared to test cases 1 and 2, the Kolmogorov-Smirnov statistic is a little larger for test Case 3. One reason is that parallel tempering does not provide significant help in test Case 3. In this test case, the contribution of the likelihood function is much larger than for test cases 1 and 2 (see Figures 7, 10 and 13). To ensure enough swap between the neighboring chains, the distance between the neighboring temperatures has to be chosen very small. Therefore, with 20 chains, the hottest chain still does not have a very flat likelihood and cannot explore the whole parameter space very easily. Thus, more chains are needed to let the parallel tempering provide more help, which will also need more computational resources. However, the current results of pCN-PT still look good, especially for the peak values of the marginal posterior distributions.

Figure 13 shows the ln-likelihood, RMSE, and pixel-wise potential scale reduction factors R of the ln-conductivity field and the hyperparameters for test Case 3 obtained by pCN-PT. We can see that the ln-likelihood and RMSE become stable very fast, which shows a short "burn-in" period and fast convergence of the

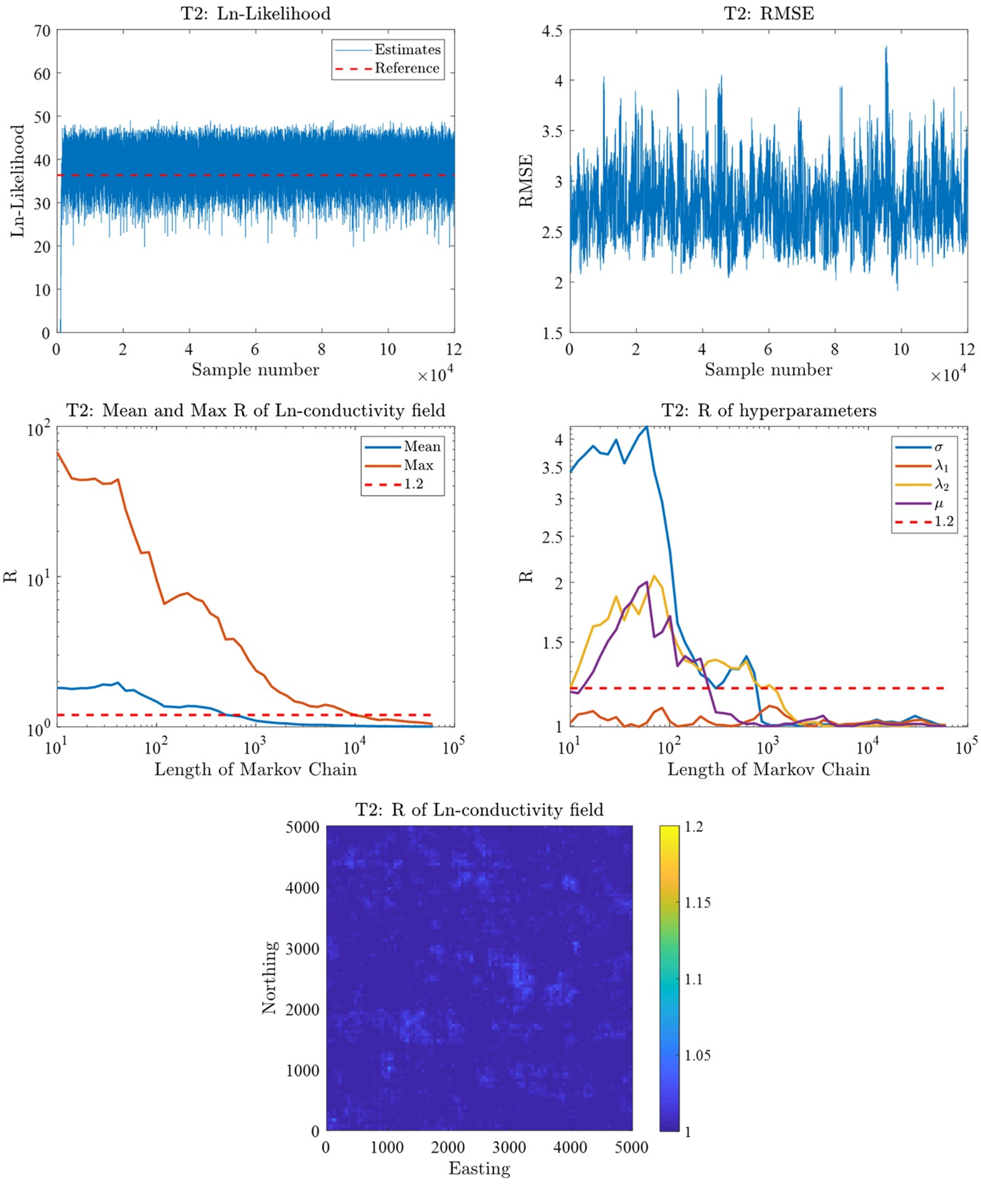


Figure 10. Testing criteria for test Case 2: Ln-likelihood of the posterior samples of the ln-conductivity field, root mean square error between the posterior samples of the ln-conductivity field and the synthetic reference field, the evolution of mean and max potential scale reduction factor of the ln-conductivity field, the evolution of the potential scale reduction factor of the hyperparameters, and the final potential scale reduction factor of the ln-conductivity field obtained by preconditioned Crank-Nicolson-parallel tempering. The dashed line in the figure of ln-likelihood corresponds to the ln-likelihood of the reference. The dashed lines in the evolution of the potential scale reduction factor denote the value of 1.2.

Markov chain. The overall potential scale reduction factor R of the ln-conductivity field keeps decreasing as the length of the Markov chain increases and finally becomes less than 1.2 in most parts of the domain. The potential scale reduction factor R of hyperparameters stays around 1.2 and is finally close to 1. This also shows a good convergence of the results obtained by pCN-PT.

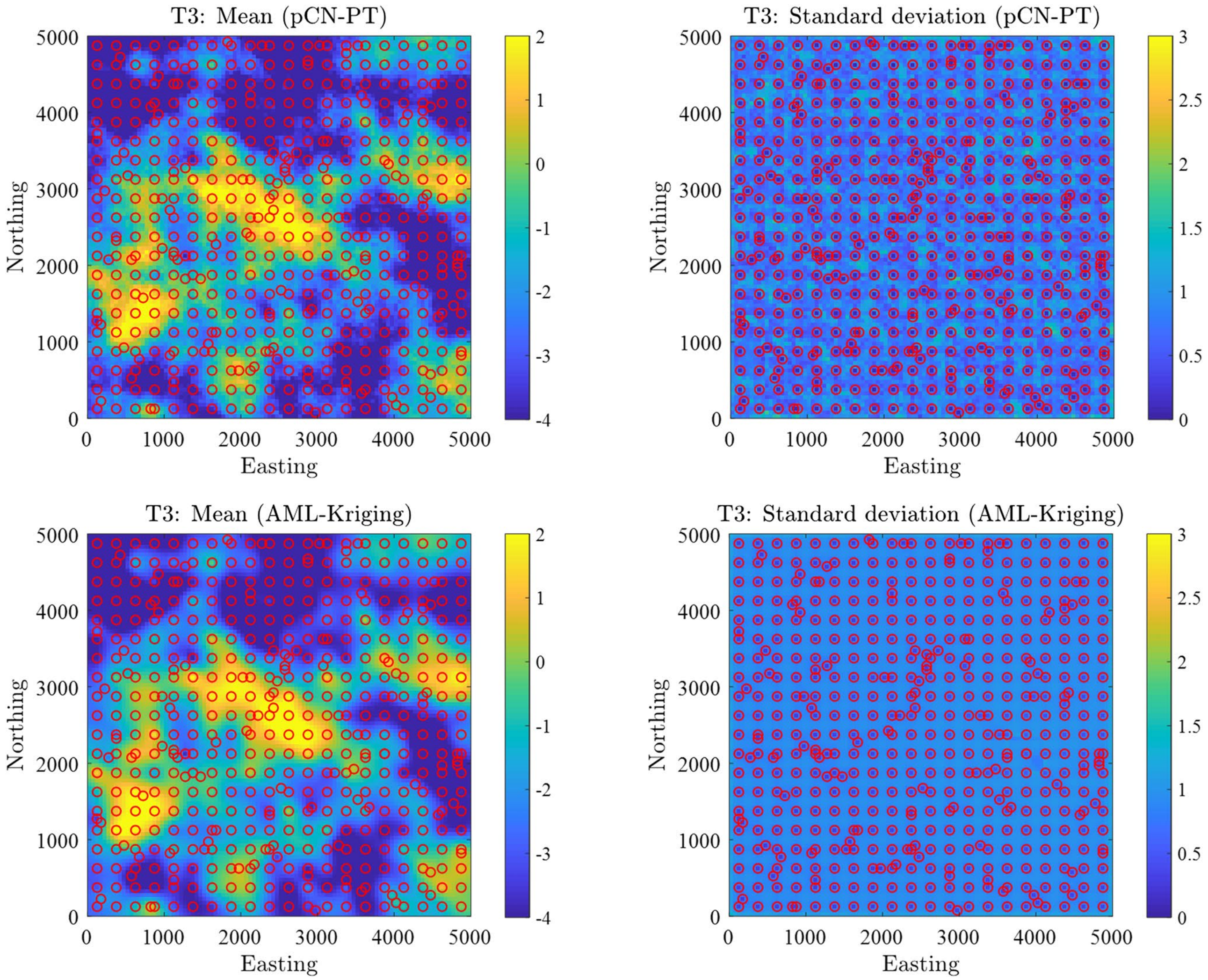


Figure 11. Posterior mean and standard deviation of the ln-likelihood field for test Case 3. Top row: results obtained by preconditioned Crank-Nicolson-parallel tempering. Bottom row: results obtained by analytical marginal likelihood-Kriging.

4.4. Test Case 4 (T4)

For test Case 4 with hydraulic head data, we only have the results obtained by pCN-PT. Figure 14 shows the posterior mean and standard deviation of the ln-conductivity field. Compared to the reference ln-conductivity field, we can see that the mean of the results captures the main features in the high conductivity areas. The marginal posterior and prior distributions of the hyperparameters for test Case 4 are shown in Figure 15. We can find that all the posterior distributions have similar shapes compared to the prior distributions and have a slight shift to the nominal values of the hyperparameters, which can also be seen from Table 7. This means the data helps to calibrate the hyperparameters a little, although head values are less informative for hyperparameters compared to direct data. Figure 16 shows the testing criteria for test Case 4. The ln-likelihood once again shows a fast convergence of the Markov chain. Although the RMSE is a little larger for some realizations, most realizations in the second half still have a small RMSE. For the ln-conductivity field, the potential scale reduction factor R decreases as the length of Markov chain increases. Finally, most parts of the ln-conductivity field have a pixel-wise R value less than 1.2. The potential scale reduction factor R of the hyperparameters also decreases and finally is less than 1.2. Overall, pCN-PT still gets a good result for this test case with a high-dimensional nonlinear problem.

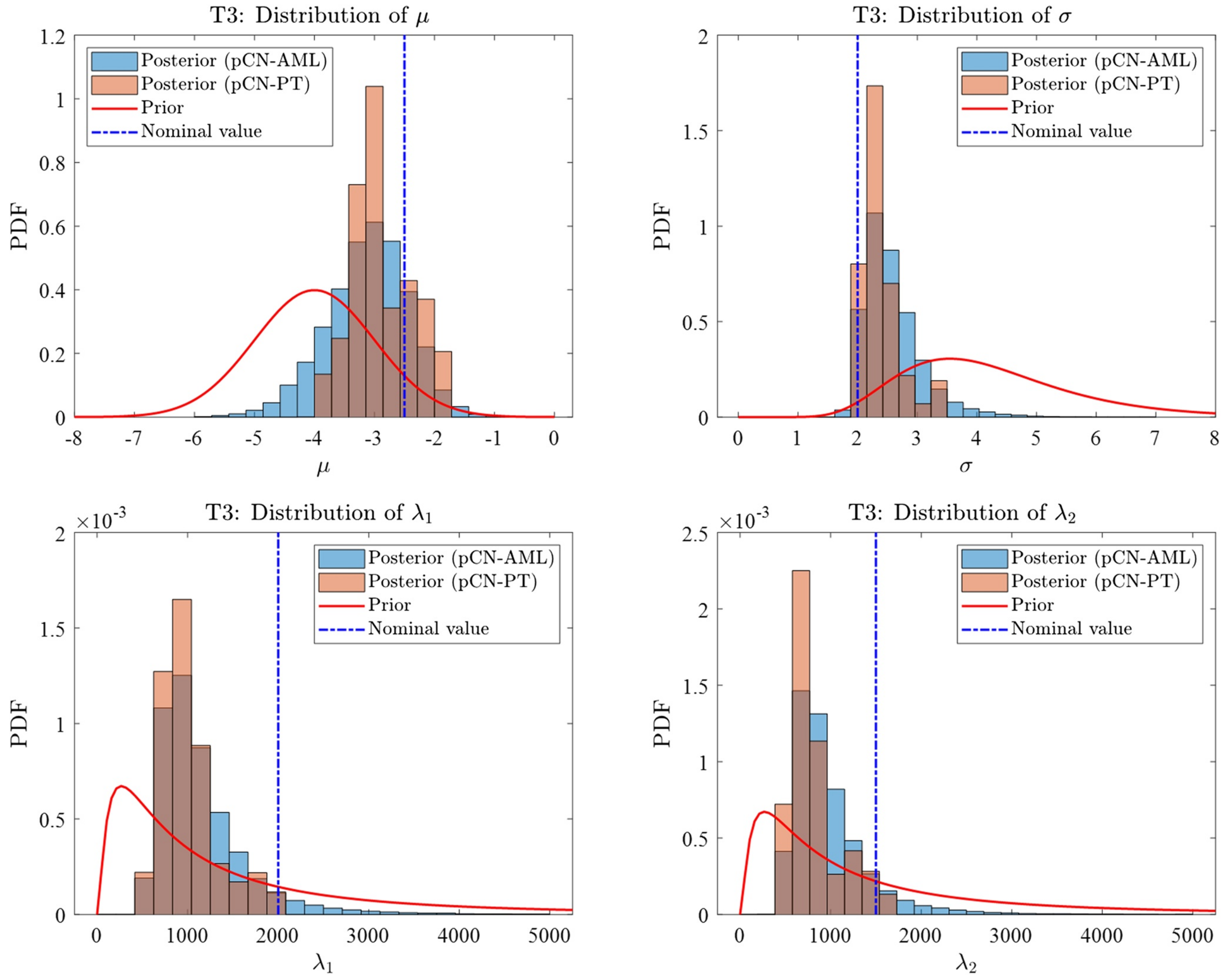


Figure 12. Marginal posterior (colored bars) and prior (red curves) distributions of the hyperparameters for test Case 3. The vertical dot-dashed lines denote the nominal values of the hyperparameters.

5. Conclusion, Final Discussion, and Outlook

In this work, we extend the highly efficient pCN-PT algorithm for geostatistical inversion and estimation of unknown spatially variable hydraulic conductivity fields to the extended Bayesian inversion problem with the estimation of uncertain hyperparameters of multi-Gaussian fields besides the hydraulic conductivities. This extended Bayesian inversion poses a harder, wider and more realistic problem since the values of hyperparameters are not fixed a priori, but one performs a formal joint Bayesian inference of hyperparameters together with the geostatistical field. To keep the high efficiency of the pCN-PT algorithm, we first reconstruct the original problem by decomposing the original model parameters (ln-conductivity) into hyperparameters (mean, standard deviation, correlation lengths) and white noise (a standard normal random vector). Then, we perform the Bayesian inversion with pCN-PT by considering hyperparameters and white noise as the primary inversion parameters. With this approach, the acceptance probability is still only dependent on the likelihood when using pCN-PT. Finally, the posterior samples of original model parameters are recovered by combining the posterior samples of hyperparameters and white noise.

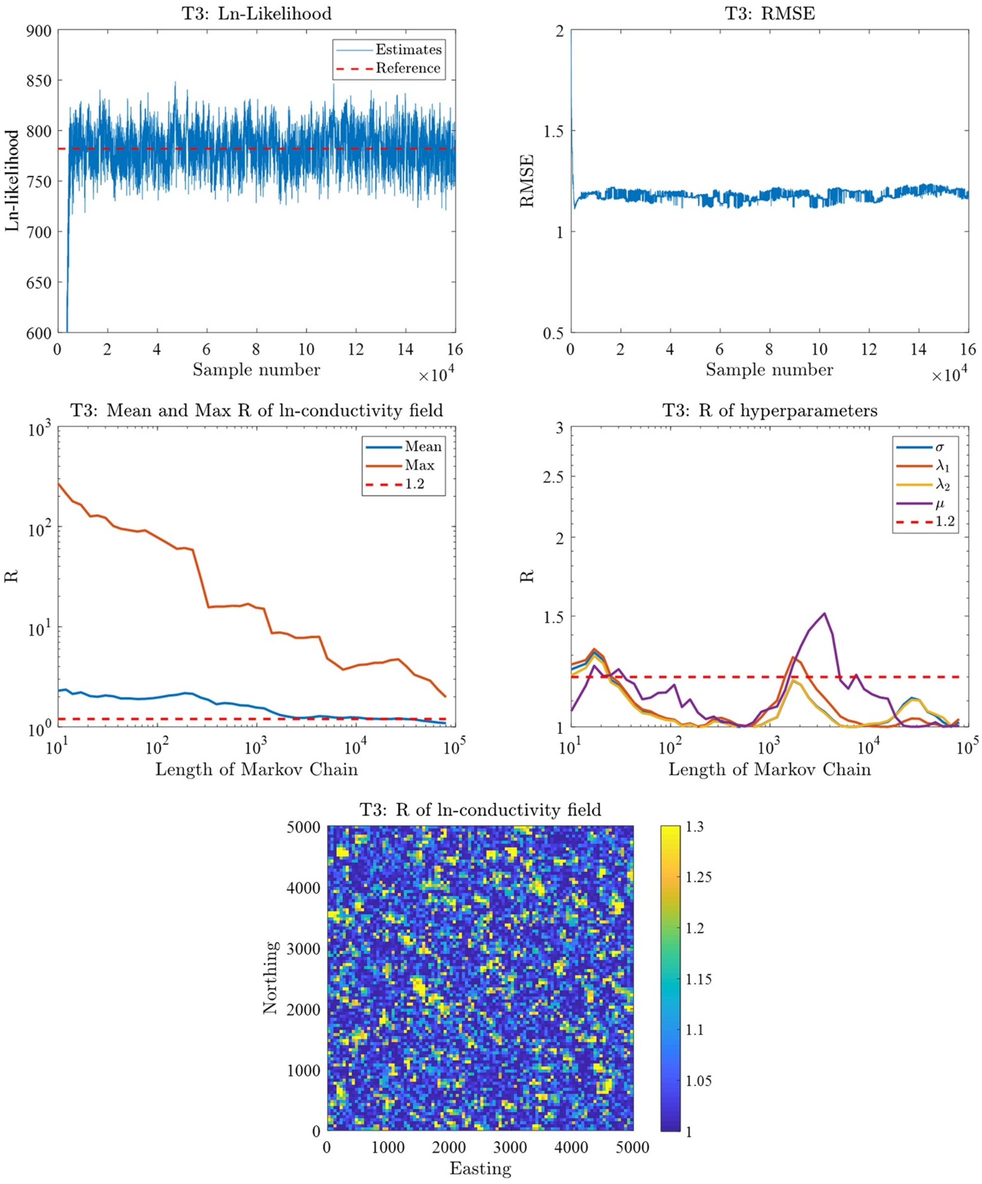


Figure 13. Testing criteria for test Case 3: Ln-likelihood of the posterior samples of the ln-conductivity field, root mean square error between the posterior samples of the ln-conductivity field and the synthetic reference field, the evolution of mean and max potential scale reduction factor of the ln-conductivity field, the evolution of the potential scale reduction factor of the hyperparameters, and the final potential scale reduction factor of the ln-conductivity field obtained by preconditioned Crank-Nicolson-parallel tempering. The dashed line in the figure of ln-likelihood corresponds to the ln-likelihood of the reference. The dashed lines in the evolution of the potential scale reduction factor denote the value of 1.2.

In this extended pCN-PT algorithm, we update hyperparameters and white noise separately during the MCMC process since they can be considered as independent to each other. This leads to the classical Gibbs sampler and allows a better fine-tuning of the corresponding MCMC parameters. There are mainly two kinds of tuning parameters, that is, a jumping factor β in the pCN proposal and temperatures T for the parallel chains. We adjust the jumping factor β adaptively based on the recent acceptance rate in each chain. The temperature T is adjusted based on the recent swap acceptance rate between neighboring chains. This adaptive parameter tuning can help us easily find proper values of the MCMC parameters.

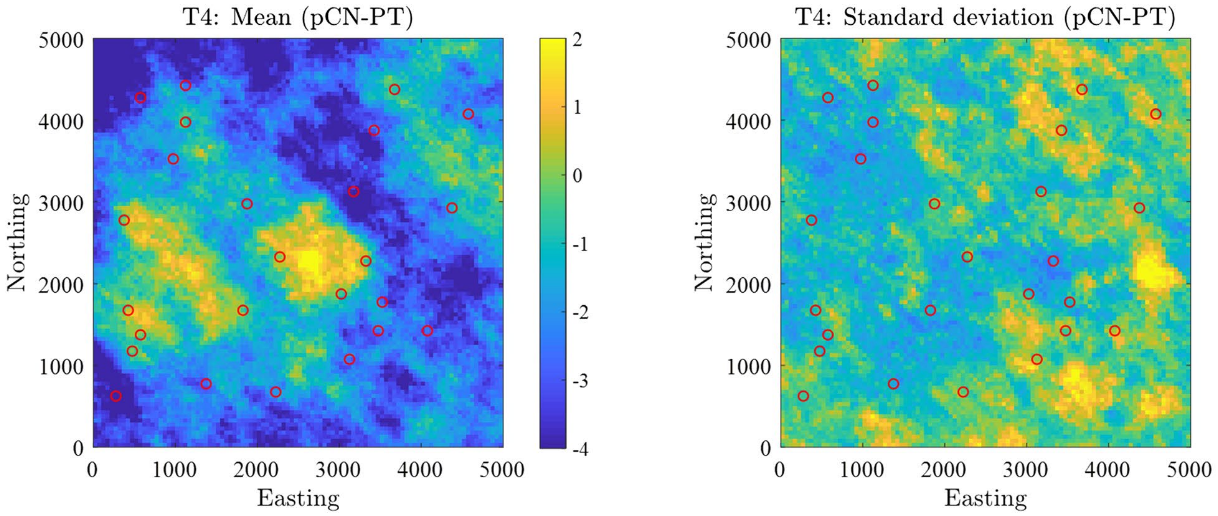


Figure 14. Posterior mean and standard deviation of the ln-conductivity field for test Case 4 obtained by preconditioned Crank-Nicolson-parallel tempering.

Based on the results of the test cases, we see that the extended pCN-PT algorithm is applicable for the extended Bayesian inversion with uncertain hyperparameters of multi-Gaussian random fields in both high-dimensional linear problems and high-dimensional nonlinear problems. We also find that it is difficult to constrain hyperparameters, especially with a relative small set of hydraulic head data. The posterior distribution of the hyperparameters is a bit narrower than the prior distribution for such a case, and approaches the reference value, but considerable uncertainty is left. This points to the importance of direct data (hydraulic conductivity) and other data sources which give information on spatial structures of hydraulic conductivity, like data from geological maps and geophysical surveys. The use of these data sources in this inversion framework requires further research.

There are many other data types that could be included in the inverse problem. In this study, we just use head and conductivity data for demonstration on a well-known benchmark case. Whether other data types (e.g., hydraulic tomography data, borehole dilution or tracer data) would make hyperparameters almost certain is another research question. This question can be answered only after one has an accurate algorithm, as the one proposed in the current study.

In fact, since our test case is not too large to be ergodic, some degree of uncertainty will always remain in the hyperparameters. However, we should still do our best to let the data speak about the structural assumptions of geostatistical models, and have the algorithms at hand to see the remaining structural uncertainty while doing the inversion.

Our test cases all used the exponential covariance function, and inferred its hyperparameters (i.e., mean, variance, and scale). However, smoothness and distribution-across-scales is also an important issue, which means the type of covariance function is known to be equally important. This could be included either as a Bayesian model selection problem or by using the Matérn covariance function (Handcock & Stein, 1993). The latter includes an additional shape parameter that controls the differentiability and scale-distribution of the covariance function. In fact, the exponential, Gaussian and power-law variograms are included as special cases and/or limit cases in the Matérn model. Therefore, it was successfully applied to parameterize the choice across covariance models (Leube et al., 2012; Nowak et al., 2010).

Such extensions toward more complicated (nonlinear) data types and additional structural uncertainty will make the problem harder to solve. In the current work, we used 20 chains, and more chains may become necessary to optimally explore and exploit the solution space. Additionally, larger domains and transient 3D

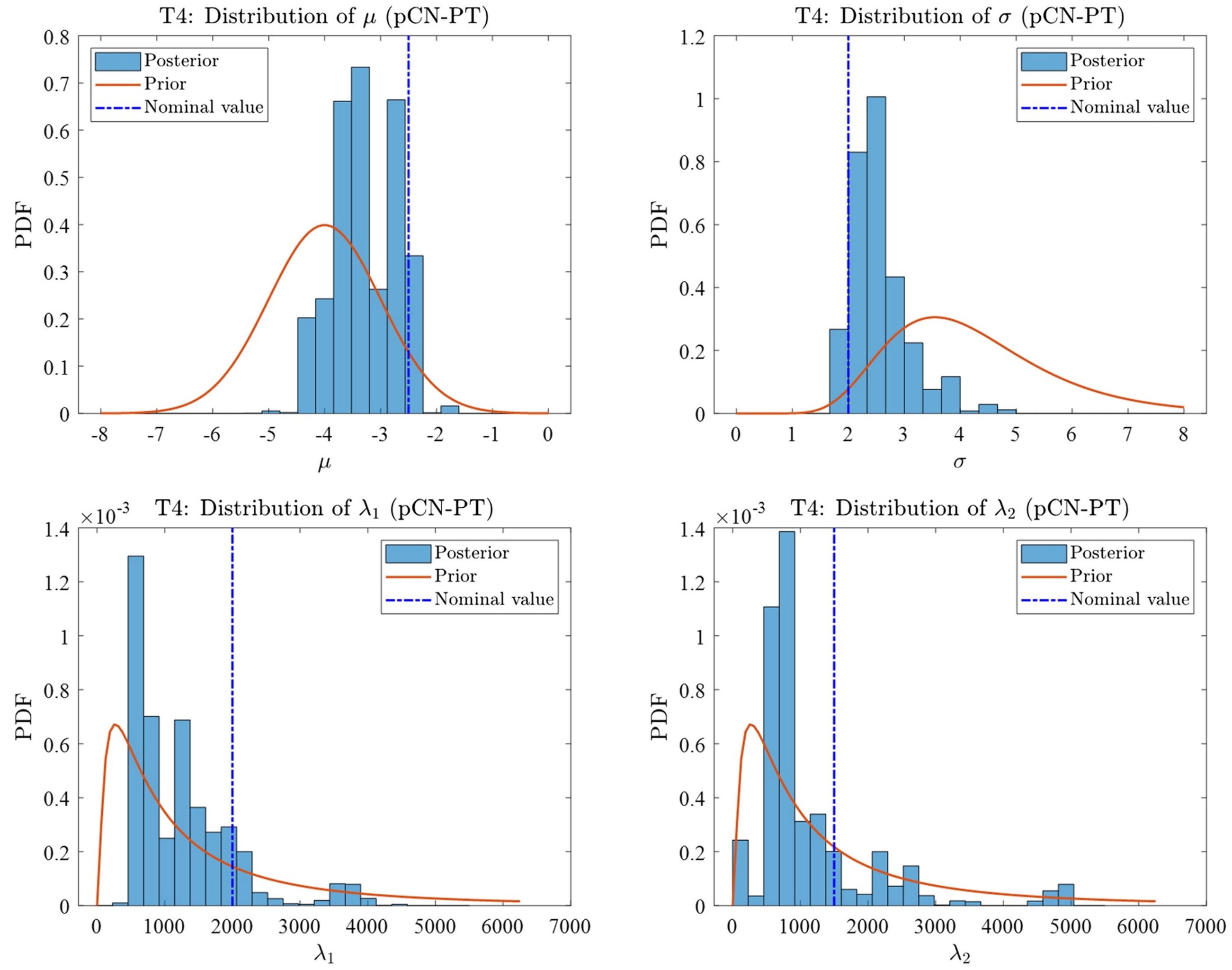


Figure 15. Marginal posterior and prior distributions of the hyperparameters for test Case 4 obtained by preconditioned Crank-Nicolson-parallel tempering. The vertical dot-dashed lines denote the nominal values of the hyperparameters.

problems will consume more computing time. Then, there will be the ubiquitous trade-off between computational resources and computational accuracy.

The core contribution of our work, that is, the decomposition of the problem into white noise at the basis of our pCN-MCMC extension, can be used to extend other geostatistical MCMC algorithms for uncertain hyperparameters. For example, one could extend the blocking MCMC approach as used in Fu and Gómez-Hernández (2009). The idea would be to repeat a random simulation (via sequential simulation

Table 7
Difference Between the Prior (Posterior) Mean and the Nominal Value of Hyperparameters

	∣ prior mean—Nominal value∣	∣ posterior mean—Nominal value∣
μ	1.50	1.05
σ	2.36	1.19
λ_1	666.48	527.05
λ_2	166.48	144.30

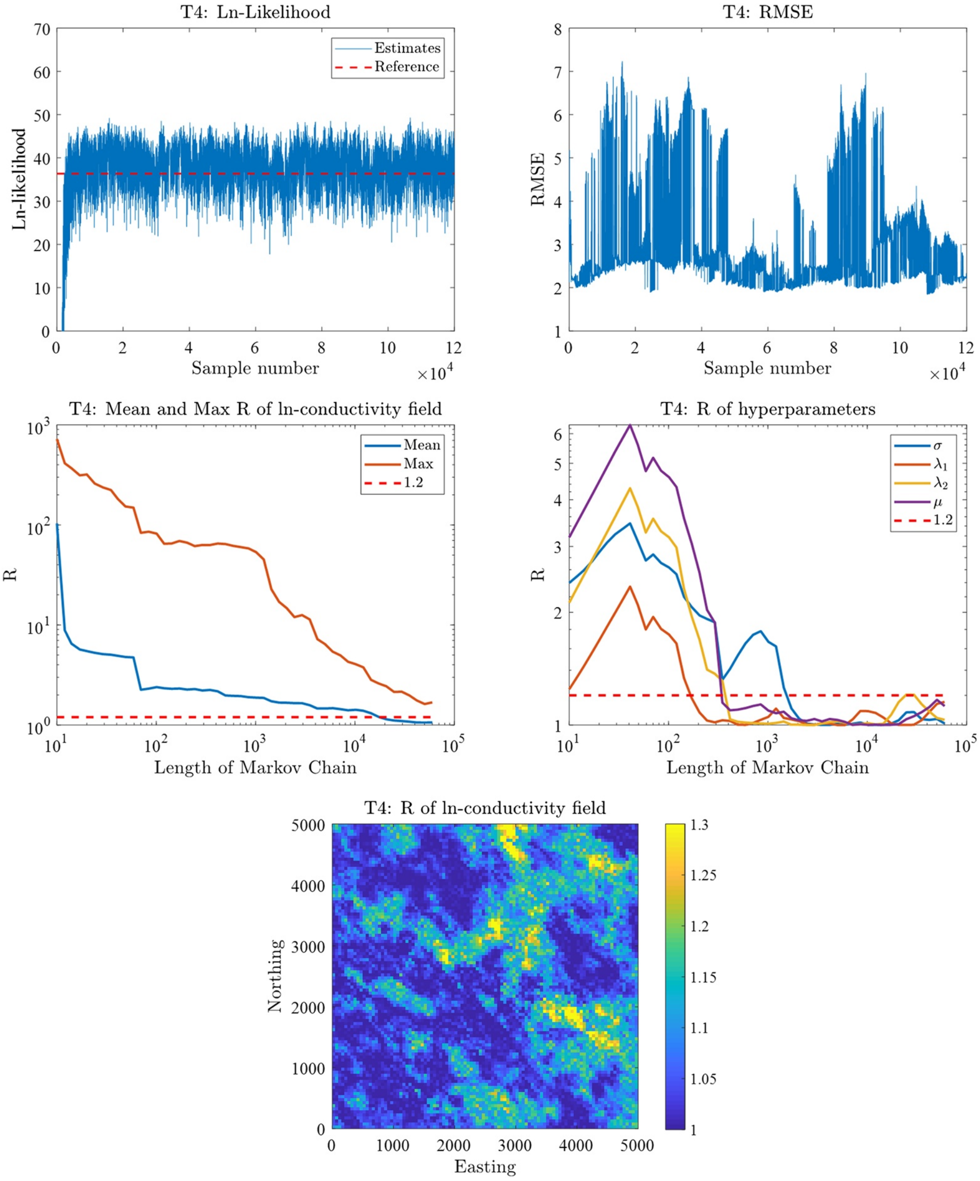


Figure 16. Testing criteria for test Case 4: Ln-likelihood of the posterior samples of the ln-conductivity field, root mean square error between the posterior samples of the ln-conductivity field and the synthetic reference field, the evolution of mean and max potential scale reduction factor of the ln-conductivity field, the evolution of the potential scale reduction factor of the hyperparameters, and the final potential scale reduction factor of the ln-conductivity field obtained by preconditioned Crank-Nicolson-parallel tempering. The dashed line in the figure of ln-likelihood corresponds to the ln-likelihood of the reference. The dashed lines in the evolution of the potential scale reduction factor denote the value of 1.2.

algorithms) with the same random seed decisions (white noise and sequential simulation path), but with a new covariance to do the morphing step, Then, one can use a conventional blocking MCMC step (which also uses sequential simulation) to achieve a change in the parameter field.

Appendix A

The posterior distribution $p(\boldsymbol{h}|\boldsymbol{d})$ of hyperparameters $\boldsymbol{h}$ can be obtained through marginalization based on the joint posterior distribution $p(\boldsymbol{\theta},\boldsymbol{h}|\boldsymbol{d})$ obtained through Equation 3. However, it can also be obtained directly through Bayes' rule as (Malinverno & Briggs, 2004)

$$p(\boldsymbol{h}|\boldsymbol{d}) \propto p(\boldsymbol{h})p(\boldsymbol{d}|\boldsymbol{h}), \tag{A1}$$

where the marginal likelihood function $p(\boldsymbol{d}|\boldsymbol{h})$ can be calculated by

$$p(\boldsymbol{d}|\boldsymbol{h}) = \int_{\Theta} p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h})p(\boldsymbol{\theta}|\boldsymbol{h})\mathrm{d}\boldsymbol{\theta}, \tag{A2}$$

where $p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h})$ is the likelihood function in Equation 3.

For the test cases with ln-conductivity data only, the forward problem transforms into a simple linear problem according

$$\boldsymbol{d} = \boldsymbol{\theta} + \boldsymbol{e}. \tag{A3}$$

For this simple linear problem, the likelihood function $p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h})$ can be written as

$$p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{h}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{d}-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{e}}^{-1}(\boldsymbol{d}-\boldsymbol{\theta})\right)}{\sqrt{(2\pi)^{nd}\det(\boldsymbol{\Sigma}_{\mathbf{e}})}}. \tag{A4}$$

Here, $\boldsymbol{\theta}$ denotes the model parameters at the measurement locations, that is, the dimension of $\boldsymbol{\theta}$ is equal to the number of data points.

Since we assume a priori a conditionally multivariate Gaussian distribution for model parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{h})$ can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{h}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^{nd}\det(\boldsymbol{\Sigma})}}. \tag{A5}$$

Here, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix of model parameters $\boldsymbol{\theta}$ at the measurement locations, with $\boldsymbol{\mu} = \boldsymbol{u}(\boldsymbol{h})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{h})$.

Utilizing the properties of Gaussian distributions, the marginal likelihood function $p(\boldsymbol{d}|\boldsymbol{h})$ in Equation A2 can be calculated analytically as (Tarantola, 2005, Section 6.21)

$$p(\boldsymbol{d}|\boldsymbol{h}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_{\mathbf{e}})^{-1}(\boldsymbol{d}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^{nd}\det(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_{\mathbf{e}})}}. \tag{A6}$$

In this study, we have four hyperparameters, that is, mean μ, standard deviation σ and two correlation lengths λ_1 and λ_2. In Equation A6, the mean vector $\boldsymbol{\mu}$ is constructed based on the hyperparameter mean μ, and the covariance matrix $\boldsymbol{\Sigma}$ is constructed based on the hyperparameters standard deviation σ and correlation lengths λ_1,λ_2. Then, we can easily sample the posterior distribution $p(\boldsymbol{h}|\boldsymbol{d})$ based on Equations A1 and A6 through MCMC (e.g., Metropolis-Hastings algorithm or pCN-MCMC).

For a given posterior sample of hyperparameters, we can use kriging to obtain a best linear unbiased estimator m_K of ln-conductivity over the entire domain and obtain the corresponding variance v_K of the estimator. Then, for all posterior samples of hyperparameters, we can get a set of kriging estimators $m_K^{(i)}$ $(i = 1,\ldots,n$, where n is the number of posterior samples of hyperparameters) and the corresponding variances $v_K^{(i)}$ $(i = 1,\ldots,n)$. The final statistical mean of the ln-conductivity field based on all realizations of hyperparameters is estimated by the average of all kriging estimators as

$$m_K^f = \frac{1}{n}\sum_{i=1}^{n} m_K^{(i)}. \tag{A7}$$

The final statistical variance of the ln-conductivity field across all realizations of hyperparameters is estimated based on the law of total variance as

$$v_K^f = \frac{1}{n}\sum_{i=1}^{n} v_K^{(i)} + \frac{1}{n-1}\sum_{i=1}^{n}\left(m_K^{(i)} - m_K^f\right)^2. \tag{A8}$$

In this study, kriging is implemented with the FFT-based algorithm in Fritz et al. (2009).

Data Availability Statement

The code and related date are available from this site https://data.mendeley.com/datasets/mtk879vsst/draft?a=a9a048b5-f7de-4a37-907b-27ecb1159d5f.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 359880532, and Project Number 327154368-SFB 1313 and Germany's Excellence Strategy—EXC 2075-390740016. T. Xu also acknowledges the financial support from the Fundamental Research Funds for the Central Universities (B200201015) and Jiangsu Specially Appointed Professor Program (B19052).

References

Alabert, F. (1987). The practice of fast conditional simulations through the LU decomposition of the covariance matrix. *Mathematical Geology*, *19*(5), 369–386. https://doi.org/10.1007/bf00897191

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., & Ronquist, F. (2004). Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, *20*(3), 407–415. https://doi.org/10.1093/bioinformatics/btg427

Besag, J., Green, P., Higdon, D., & Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, *10*(1), 3–41. https://doi.org/10.1214/ss/1177010130

Boggs, J. M., Young, S. C., Beard, L. M., Gelhar, L. W., Rehfeldt, K. R., & Adams, E. E. (1992). Field study of dispersion in a heterogeneous aquifer: 1. overview and site description. *Water Resources Research*, *28*(12), 3281–3291. https://doi.org/10.1029/92wr01756

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational & Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Bui-Thanh, T., Ghattas, O., Martin, J., & Stadler, G. (2013). A computational framework for infinite-dimensional Bayesian inverse problems part I: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, *35*(6), A2494–A2523. https://doi.org/10.1137/12089586x

Buland, A., & Omre, H. (2003). Bayesian linearized AVO inversion. *Geophysics*, *68*(1), 185–198. https://doi.org/10.1190/1.1543206

Carrera, J., & Neuman, S. P. (1986). Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resources Research*, *22*(2), 199–210. https://doi.org/10.1029/wr022i002p00199

Cary, P. W., & Chapman, C. H. (1988). Automatic 1-D waveform inversion of marine seismic refraction data. *Geophysical Journal*, *93*(3), 527–546. https://doi.org/10.1111/j.1365-246x.1988.tb03879.x

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174. https://doi.org/10.1080/00031305.1992.10475878

Chen, Y., & Zhang, D. (2006). Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Advances in Water Resources*, *29*(8), 1107–1122. https://doi.org/10.1016/j.advwatres.2005.09.007

Cirpka, O. A., & Nowak, W. (2003). Dispersion on kriged hydraulic conductivity fields. *Water Resources Research*, *39*(2). https://doi.org/10.1029/2001WR000598

Congdon, P. (2003). *Applied Bayesian modelling*. John Wiley & Sons, Ltd.

Cotter, S. L., Dashti, M., Robinson, J. C., & Stuart, A. M. (2009). Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, *25*(11), 115008. https://doi.org/10.1088/0266-5611/25/11/115008

Cotter, S. L., Roberts, G. O., Stuart, A. M., & White, D. (2013). MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science*, *28*(3), 424–446. https://doi.org/10.1214/13-sts421

Dagan, G. (1988). Time-dependent macrodispersion for solute transport in anisotropic heterogeneous aquifers. *Water Resources Research*, *24*(9), 1491–1500. https://doi.org/10.1029/WR024i009p01491

Davis, M. W. (1987). Production of conditional simulations via the LU triangular decomposition of the covariance matrix. *Mathematical Geology*, *19*(2), 91–98. https://doi.org/10.1007/bf00897749

Dentz, M., Kinzelbach, H., Attinger, S., & Kinzelbach, W. (2000). Temporal behavior of a solute cloud in a heterogeneous porous medium: 1. Point-like injection. *Water Resources Research*, *36*(12), 3591–3604. https://doi.org/10.1029/2000wr900162

Dietrich, C. R., & Newsam, G. N. (1993). A fast and exact method for multidimensional Gaussian stochastic simulations. *Water Resources Research*, *29*(8), 2861–2869. https://doi.org/10.1029/93wr01070

Dietrich, C. R., & Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, *18*(4), 1088–1107. https://doi.org/10.1137/s1064827592240555

Doherty, J. E., Fienen, M. N., & Hunt, R. J. (2010). Approaches to highly parameterized inversion: Pilot-point theory, guidelines, and research directions. *US Geological Survey scientific investigations report*, *5168*, 36.

Duijndam, A. J. W. (1988). Bayesian estimation in seismic inversion. Part I: Principles. *Geophysical Prospecting*, *36*(8), 878–898. https://doi.org/10.1111/j.1365-2478.1988.tb02198.x

Earl, D. J., & Deem, M. W. (2005). Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, *7*, 3910–3916. https://doi.org/10.1039/b509983h

Fritz, J., Neuweiler, I., & Nowak, W. (2009). Application of FFT-based algorithms for large-scale universal Kriging problems. *Mathematical Geosciences*, *41*(5), 509–533. https://doi.org/10.1007/s11004-009-9220-x

Fu, J., & Gómez-Hernández, J. J. (2009). A blocking Markov chain Monte Carlo method for inverse stochastic hydrogeological modeling. *Mathematical Geosciences*, *41*(2), 105–128. https://doi.org/10.1007/s11004-008-9206-0

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

Gelman, A., Roberts, G., & Gilks, W. (1996). Efficient metropolis jumping rules. *Bayesian Statistics*, *5*, 599–607.

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

Gómez-Hernández, J. J., Sahuquillo, A., & Capilla, J. E. (1997). Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data—I. theory. *Journal of Hydrology*, *203*(1), 162–174.

Grandis, H., Menvielle, M., & Roussignol, M. (1999). Bayesian inversion with Markov chains—I. The magnetotelluric one-dimensional case. *Geophysical Journal International*, *138*(3), 757–768. https://doi.org/10.1046/j.1365-246x.1999.00904.x

Hairer, M., Stuart, A. M., & Vollmer, S. J. (2014). Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Annals of Applied Probability*, *24*(6), 2455–2490. https://doi.org/10.1214/13-aap982

Handcock, M. S., & Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, *35*(4), 403–410. https://doi.org/10.1080/00401706.1993.10485354

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

Hukushima, K., & Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, *65*(6), 1604–1608. https://doi.org/10.1143/jpsj.65.1604

Keller, J., Hendricks Franssen, H.-J., & Nowak, W. (2021). Investigating the pilot point ensemble Kalman filter for geostatistical inversion and data assimilation. *Advances in Water Resources*, 104010. https://doi.org/10.1016/j.advwatres.2021.104010

Kitanidis, P. K. (1995). Quasi-linear geostatistical theory for inversing. *Water Resources Research*, *31*(10), 2411–2419. https://doi.org/10.1029/95wr01945

Kitanidis, P. K., & Vomvoris, E. G. (1983). A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations. *Water Resources Research*, *19*(3), 677–690. https://doi.org/10.1029/wr019i003p00677

Laloy, E., Linde, N., Jacques, D., & 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. https://doi.org/10.1016/j.advwatres.2016.02.008

Leube, P., Geiges, A., & Nowak, W. (2012). Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resources Research*, *48*, W02501. https://doi.org/10.1029/2010WR010137

Li, L., Zhou, H., Hendricks Franssen, H. J., & Gómez-Hernández, J. J. (2012). Groundwater flow inverse modeling in non-Multigaussian media: Performance assessment of the normal-score ensemble Kalman filter. *Hydrology and Earth System Sciences*, *16*(2), 573–590. https://doi.org/10.5194/hess-16-573-2012

Malinverno, A. (2002). Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, *151*(3), 675–688. https://doi.org/10.1046/j.1365-246x.2002.01847.x

Malinverno, A., & Briggs, V. A. (2004). Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes. *Geophysics*, *69*(4), 1005–1016. https://doi.org/10.1190/1.1778243

Mariethoz, G., Renard, P., & Caers, J. (2010). Bayesian inverse problem and optimization with iterative spatial resampling. *Water Resources Research*, *46*(11), W11530. https://doi.org/10.1029/2010wr009274

McDonald, M. G., & Harbaugh, A. W. (1988). *A modular three-dimensional finite-difference ground-water flow model (Tech. Rep.)*. U.S. Geological Survey. Retrieved from http://pubs.er.usgs.gov/publication/twri06A1

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Moreno, J. J., Katzgraber, H. G., & Hartmann, A. K. (2003). Finding low-temperature states with parallel tempering, simulated annealing and simple Monte Carlo. *International Journal of Modern Physics C*, *14*(03), 285–302. https://doi.org/10.1142/s0129183103004498

Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research*, *100*(B7), 12431–12447. https://doi.org/10.1029/94jb03097

Nagel, J. B., & Sudret, B. (2016). A unified framework for multilevel uncertainty quantification in Bayesian inverse problems. *Probabilistic Engineering Mechanics*, *43*, 68–84. https://doi.org/10.1016/j.probengmech.2015.09.007

Nowak, W., & Cirpka, O. A. (2006). Geostatistical inference of hydraulic conductivity and dispersivities from hydraulic heads and tracer data. *Water Resources Research*, *42*(8), W08416. https://doi.org/10.1029/2005wr004832

Nowak, W., de Barros, F. P. J., & Rubin, Y. (2010). Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain. *Water Resources Research*, *46*, W03535. https://doi.org/10.1029/2009WR008312

Nowak, W., & Litvinenko, A. (2013). Kriging and spatial design accelerated by orders of magnitude: Combining low-rank covariance approximations with FFT-techniques. *Mathematical Geosciences*, *45*(4), 411–435. https://doi.org/10.1007/s11004-013-9453-6

Predescu, C., Predescu, M., & Ciobanu, C. V. (2005). On the efficiency of exchange in parallel tempering Monte Carlo simulations. *The Journal of Physical Chemistry B*, *109*(9), 4189–4196. https://doi.org/10.1021/jp045073+

Rathore, N., Chopra, M., & de Pablo, J. J. (2005). Optimal allocation of replicas in parallel tempering simulations. *The Journal of Chemical Physics*, *122*(2), 024111. https://doi.org/10.1063/1.1831273

Robert, C. P., Elvira, V., Tawn, N., & Wu, C. (2018). Accelerating MCMC algorithms. *WIREs Computational Statistics*, *10*(5), e1435. https://doi.org/10.1002/wics.1435

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, *7*(1), 110–120.

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, *16*(4), 351–367. https://doi.org/10.1214/ss/1015346320

Schott, J.-J., Roussignol, M., Menvielle, M., & Nomenjahanary, F. R. (1999). Bayesian inversion with Markov chains—II. The one-dimensional DC multilayer case. *Geophysical Journal International*, *138*(3), 769–783. https://doi.org/10.1046/j.1365-246x.1999.00905.x

Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica*, *19*, 451–559. https://doi.org/10.1017/s0962492910000061

Sudicky, E. A. (1986). A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in the dispersion process. *Water Resources Research*, *22*(13), 2069–2082. https://doi.org/10.1029/wr022i013p02069

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*(4), 1701–1728. https://doi.org/10.1214/aos/1176325755

Troldborg, M., Nowak, W., Lange, I. V., Santos, M. C., Binning, P. J., & Bjerg, P. L. (2012). Application of Bayesian geostatistics for evaluation of mass discharge uncertainty at contaminated sites. *Water Resources Research*, *48*(9), W09535. https://doi.org/10.1029/2011wr011785

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov chain Monte-Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. https://doi.org/10.3758/s13423-016-1015-8

Woodbury, A. D., & Ulrych, T. J. (2000). A full-Bayesian approach to the groundwater inverse problem for steady state flow. *Water Resources Research*, *36*(8), 2081–2093. https://doi.org/10.1029/2000wr900086

Xu, T., Reuschen, S., Nowak, W., & Hendricks Franssen, H.-J. (2020). Preconditioned Crank-Nicolson Markov chain Monte Carlo coupled with parallel tempering: An efficient method for Bayesian inversion of multi-Gaussian log-hydraulic conductivity fields. *Water Resources Research*, *56*(8), e2020WR027110. https://doi.org/10.1029/2020wr027110

Yeh, T.-C. J., Jin, M., & Hanna, S. (1996). An iterative stochastic inverse method: Conditional effective transmissivity and hydraulic head fields. *Water Resources Research*, *32*(1), 85–92. https://doi.org/10.1029/95wr02869

Zha, Y., Yeh, T.-C. J., Illman, W. A., Zeng, W., Zhang, Y., Sun, F., & Shi, L. (2018). A reduced-order successive linear estimator for geostatistical inversion and its application in hydraulic tomography. *Water Resources Research*, *54*(3), 1616–1632. https://doi.org/10.1002/2017wr021884

Zhao, Y., & Luo, J. (2021a). Bayesian inverse modeling of large-scale spatial fields on iteratively corrected principal components. *Advances in Water Resources*, *151*, 103913. https://doi.org/10.1016/j.advwatres.2021.103913

Zhao, Y., & Luo, J. (2021b). A quasi-newton reformulated geostatistical approach on reduced dimensions for large-dimensional inverse problems. *Water Resources Research*, *57*(1), e2020WR028399. https://doi.org/10.1029/2020wr028399

D Contribution 4: Bayesian inversion of hierarchical geostatistical models using a parallel-tempering sequential Gibbs MCMC

Advances in Water Resources 141 (2020) 103614

Contents lists available at ScienceDirect

Advances in Water Resources

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

Bayesian inversion of hierarchical geostatistical models using a parallel-tempering sequential Gibbs MCMC

Sebastian Reuschen*, Teng Xu, Wolfgang Nowak

Department of Stochastic Simulation and Safety Research for Hydrosystems, University of Stuttgart, Stuttgart, Germany

ARTICLE INFO

Keywords:
Bayesian inversion
MCMC
Parallel tempering
Multiple-point statistics
Sequential geostatistical resampling
Training image

ABSTRACT

The feasibility of probabilistic Bayesian inversion strongly depends on the dimensionality and complexity of the statistical prior model. Most geostatistical inversion approaches assume multi-Gaussian fields, and some assume (non-Gaussian) categorical fields, e.g., via multiple-point geostatistics. We combine these two into one hierarchical joint problem, which accounts for two (and possibly more) categories as well as heterogeneities inside each category. Recent works developed the conditional probability field method based on the Ensemble Kalman filter (EnKf) for this scenario. However, EnKf-type approaches take implicit linearity and (trans-)Gaussian assumptions, which are not feasible in weak-information regimes. Therefore, we develop a tailored Gibbs sampler, a kind of Markov chain Monte Carlo (MCMC) method. It can do this inversion without assumptions. Our algorithm extends an existing Gibbs sampler with parallel tempering for categorical fields to account for multi-Gaussian internal heterogeneity. We show our key idea and derive our algorithm from the detailed balance, required for MCMC algorithms. We test our algorithm on a synthetic channelized flow scenario for different levels of data available: A highly informative setting (transient flow data) where the synthetic truth can be recovered and a weakly informative setting (steady-state data only) where the synthetic truth cannot be recovered. Instead, we obtain a multi-modal posterior. For the proper testing of convergence, we use the scale reduction factor by Gelman and Rubin. Overall, the test illustrates that our algorithm performs well in both settings.

1. Introduction

Heterogeneity of hydraulic parameters is a key control on subsurface flow, transport and energy transfer processes. Characterizing these heterogeneities is difficult because we cannot measure subsurface parameters directly and at sufficient resolution.

Bayesian geostatistical inversion is one way to obtain spatially variable estimates of parameters. Bayesian inversion combines prior knowledge of the system with the likelihood of observation data. The resulting estimate is obtained as a posterior probability distribution. In this process, a model of the prior knowledge for spatial heterogeneity is needed. Conventionally, multi-Gaussian fields (e.g., Matheron, 1975) or categorical fields (e.g., Hansen et al., 2012; Laloy et al., 2016; Strebelle, 2002) are used. Following Iglesias et al. (2014), Xu and Gómez-Hernández (2015) and Mo et al. (2020), we combine these two approaches and create a hierarchical model consisting of categorical fields with internal Gaussian fields.

Formulating an analytical solution of the posterior based on available data is not possible for such an inverse hierarchical model. Instead, the problem is solved numerically. For this purpose, two different approaches can be taken: approximate or exact methods. Examples for approximate methods were proposed by Xu and Gómez-Hernández (2015) and Mo et al. (2020) who used an ensemble Kalman filter and an ensemble smoother approach, respectively. Both converge to an approximate solution of the Bayesian inverse problem, because they take implicit linearity and (trans-)Gaussian assumptions as discussed for geostatistical inversion by Nowak (2009). Especially the Gaussian assumption is not reasonable for multi-facies systems. This can be addressed using normal score transformations (Zhou et al., 2011; Schöniger et al., 2012). However, the normal score transformation only removes the multi-modality of the parameters. The non-linear forward models still introduce an error. Consequently, ensemble Kalman filters converge (with increasing sample size) to an implicitly linearized approximation of the true Bayesian solution.

On the contrary, Iglesias et al. (2014) presented a Markov chain Monte Carlo (MCMC) method that converges to the exact solution of the Bayesian inverse problem. However, their work assumes that the categorical field can be parameterized by a low (here: 5) number of geometrical parameters. This is a strong assumption on the structure of the categorical field and is often not reasonable in applied geostatistical and geological setups. Hence, our work will present an approach that assumes the same spatial discretization (here: 50×50) for the categorical

* Corresponding author.
E-mail address: sebastian.reuschen@iws.uni-stuttgart.de (S. Reuschen).

https://doi.org/10.1016/j.advwatres.2020.103614
Received 12 December 2019; Received in revised form 4 May 2020; Accepted 6 May 2020
0309-1708/© 2020 Published by Elsevier Ltd.

and the multi-Gauss fields and avoids a low-dimensional parametrization. This enables solutions to be more appropriate and more flexible representations of geological reality. The difference in dimensionality leads Iglesias et al. (2014) to use a Metropolis Hastings proposal for the geometrical parameters. Instead, we use a parallel tempering sequential Gibbs approach tailored to the high-dimensional situation.

Many MCMC methods have been presented in the literature to solve Bayesian inverse problems. In this work, we use a Gibbs-based approach. This means to resample one (or several) parameter(s) conditional (with respect to the posterior distribution) on the remaining (unmodified) parameters. For geostatistical inversion, Hansen et al. (2012) proposed a sequential Gibbs sampler that resamples different spatial blocks of the geostatistical parameter field. With this algorithm, they were able to sample from categorical fields efficiently. We extend their approach (of resampling blocks of the parameter field) to hierarchical models, where the parameter field consists of a categorical field with internal multi-Gaussian heterogeneity per category. To speed up convergence, we adopt the parallel tempering approach (Geyer and Thompson, 1995). Parallel tempering is a method in which several MCMC chains run on similar problems with increasing hardness. The chains communicate their results with each other to improve the exploitation of possible solutions. Laloy et al. (2016) showed that this enables faster and more efficient computation on geostatistical problems.

We test our proposed algorithm in two fundamentally different information regimes for a fully saturated, two-dimensional groundwater flow. First, we feature a high-information regime (many measurements) where accurate inversion is challenging. Xu and Gómez-Hernández (2015) showed that Ensemble Kalman filters can get good approximations in this regime because the implicit multi-Gaussian assumption of EnKfs holds well over narrow unimodal posterior distributions. Second, we feature a weak-information regime (few measurements) that results in a multi-modal posterior. We show that our proposed method can reliably find and quantify all modes. The key contribution of our paper is to develop a method that can handle both weakly and highly informative regimes for hierarchical geostatistical models with multiple-point geostatistics and internal heterogeneity.

In Section 2 we derive our MCMC algorithm. Section 3 presents the test cases and implementation of our model. In Section 4 we show and discuss our results. Finally, in Section 5 we conclude the most important findings with a short summary.

2. Methods

In this section we give an overview over related MCMC approaches and present how our algorithm extends them. First, we present our problem in the framework of Bayesian inference (Section 2.1). Then, we present our key idea in Section 2.2 and present the details of our Gibbs-based MCMC algorithm in Section 2.3. Finally, we extend it in Section 2.4 by parallel tempering for increased efficiency.

2.1. Bayesian inference

We assume a stochastic representation of a forward problem

$$F(\theta) = \boldsymbol{d} + \boldsymbol{e} \qquad (1)$$

where $F(\theta)$ is an error-free deterministic forward model that describes the relation between the measurement data $\boldsymbol{d}$ and the unknown parameters θ. The noise term $\boldsymbol{e}$ condenses all possible error terms. Our goal is to infer the parameters θ based on the data $\boldsymbol{d}$ and the prior knowledge of θ.

The parameters θ are viewed as random variables with some prior distribution $p(\theta)$ and a posterior distribution $p(\theta|\boldsymbol{d})$. The posterior is given as

$$p(\theta|\boldsymbol{d}) = \frac{p(\theta)p(\boldsymbol{d}|\theta)}{p(\boldsymbol{d})} \propto p(\theta)p(\boldsymbol{d}|\theta) = P(\theta)L(\theta|\boldsymbol{d}). \qquad (2)$$

We define the likelihood $L(\theta|\boldsymbol{d}) := p(\boldsymbol{d}|\theta)$ and the prior distribution $P(\theta)$ as $P(\theta) := p(\theta)$ for a clearer notation in the next subsection. The probability density of the data $p(\boldsymbol{d}) = \int p(\boldsymbol{d}|\theta)p(\theta)d\theta$, also called Bayesian Model Evidence, can be obtained by numerical integration over the parameter space. However, this integration is difficult and is not required for the parameter inference when the parameter dimensionality is fixed (Laloy et al., 2016). Therefore, we use the unnormalized density

$$\pi(\theta) = P(\theta)L(\theta|\boldsymbol{d}) \propto p(\theta|\boldsymbol{d}), \qquad (3)$$

assuming some fixed data $\boldsymbol{d}$, where the unnormalized posterior probability π equals prior P times likelihood L. We will sample from $\pi(\theta)$ in the following.

The likelihood $L(\theta|\boldsymbol{d})$ can often assume values, close to machine precision. In order to avoid numerical underflow error, it is convenient to use the log-likelihood $l(\theta|\boldsymbol{d}) := \log(L(\theta|\boldsymbol{d}))$ instead. Assuming an uncorrelated, normally distributed error term $\boldsymbol{e}$ with standard deviation σ_e, the log-likelihood is

$$l(\theta|\boldsymbol{d}) = N \cdot \log\left(\frac{1}{\sqrt{2\pi\sigma_e^2}}\right) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{N}\left(d_i - F_i(\theta)\right)^2 \qquad (4)$$

where $F_i(\theta)$ are the simulated equivalents to the measured data d_i and N is the number of measurements. However, any other distribution of errors is possible as well.

To simplify our notation, we define the likelihood $L(\theta) := L(\theta|\boldsymbol{d})$ (and $l(\theta) := l(\theta|\boldsymbol{d})$) independent of the data $\boldsymbol{d}$ because we assume constant $\boldsymbol{d}$ during the run-time of the algorithm.

2.2. Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a popular, accurate, yet sometimes inefficient algorithm to solve Bayesian inverse problems. Most modern MCMC methods are based or inspired by the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). We name all properties an MCMC method needs to fulfill to have proven convergence to the exact distribution. Based on these, we derive the formulas needed for our proposed MCMC algorithm. For a general introduction to MCMC methods, we point to Chib and Greenberg (1995).

MCMC methods converge to π presented in Eq. 3 (at the limit of infinite runtime) if and only if irreducibility, aperiodicity and the detailed balance are fulfilled (Smith and Roberts, 1993). The first two are almost always fulfilled. Hence, we focus on the detailed balance from now on. It is defined as

$$\pi(\theta_i)h(\theta_i, \theta_j) = \pi(\theta_j)h(\theta_j, \theta_i) \qquad (5)$$

with the transition kernel h, which is usually defined as

$$h(\theta_i, \theta_j) = q(\theta_i, \theta_j)\alpha(\theta_i, \theta_j). \qquad (6)$$

Here, $q(\boldsymbol{\theta_i}, \boldsymbol{\theta_j})$ is the so-called proposal distribution and $\alpha(\boldsymbol{\theta_i}, \boldsymbol{\theta_j})$ is called the acceptance probability.

Combining Eqs. 3, 5 and 6 (see Appendix for derivation) leads to

$$\alpha(\theta_i, \theta_j) = min\left[\frac{P(\theta_j)L(\theta_j)q(\theta_j, \theta_i)}{P(\theta_i)L(\theta_i)q(\theta_i, \theta_j)}, 1\right]. \qquad (7)$$

For any prior P, any likelihood L and any proposal distribution q, Eq. 7 provides an α such that the detailed balance is fulfilled. Hence, we can construct an MCMC with (almost) any proposal distribution q. The only restriction is that irreducibility and aperiodicity are not always fulfilled for arbitrarily chosen proposal distributions. This yields the question of how to choose q for fast convergence for a given problem class.

The convergence rate of the MCMC algorithm depends on how fast it can explore the parameter space. The faster it moves through the parameter space, the faster it converges (Gelman et al., 1996). Hence, it is desirable to make large changes to the parameter set and accept them with a high probability (Gelman et al., 1996). In practice, however, these two things contradict each other: Making small changes in θ results in similar $P(\theta_j)L(\theta_j)$ and $P(\theta_i)L(\theta_i)$ (if the prior and the likelihood function are smooth), so that α is around 1. Making large changes in θ results in distinct $P(\theta_j)L(\theta_j)$ and $P(\theta_i)L(\theta_i)$, which results in a small α. Thus, a trade-off between the size of the change and the acceptance rate needs to be found (Gelman et al., 1996). Other approaches, e.g., Hamiltonian MCMC (Betancourt, 2017) construct clever proposal distributions to make far jumps with high acceptance rates. However they are not applicable to our problem class, because they use they use derivatives of the prior distribution, which are not attainable for hierarchical models.

2.2.1. Metropolis-Hasting

The standard Metropolis-Hasting algorithm (Metropolis et al., 1953; Hastings, 1970) assumes a symmetric proposal distribution

$$q(\theta_i, \theta_j) = q(\theta_j, \theta_i). \tag{8}$$

Inserting this into Eq. 7 yields that

$$\alpha(\theta_i, \theta_j) = min\left[\frac{P(\theta_j)L(\theta_j)}{P(\theta_i)L(\theta_i)}, 1\right] = min\left[\frac{\pi(\theta_j)}{\pi(\theta_i)}, 1\right]. \tag{9}$$

The Metropolis Hasting algorithm can sample from $\pi(\theta)$ without taking any assumptions about its form. A standard realization of this approach is the random walk proposal function

$$g(\theta_i) = \theta_i + \epsilon,\ \epsilon \sim N(0, \sigma). \tag{10}$$

This function $g(\theta_i)$ fulfills Eq. 8 because the normal distribution $N(0, \sigma)$, with mean $\mu = 0$ and standard deviation σ, is symmetric. However, the acceptance rate (Eq. 8) depends on the prior and the likelihood, which leads to a fast decrease of α for increasing σ especially in high-dimensional problems (Roberts and Rosenthal, 2002). We want to improve this by using all available information about $\pi(\theta)$, to increase α and speed up convergence.

2.2.2. Sampling from the prior

The basic idea of many Bayesian inversion methods is to use the knowledge that $\pi(\theta) = P(\theta)L(\theta)$. Using this information, the performance of MCMC methods can be increased. In many problem classes, especially in high-dimensional geoscience problems, the prior $P(\theta)$ is complex. Hence, the acceptance rate α often depends almost exclusively on the prior if a standard Metropolis-Hastings algorithm is used. Furthermore, there are cases where the prior $P(\theta)$ can not be evaluated for any given θ because no closed-form is known. Examples are multiple-point geostatistics tools that use training images (Strebelle, 2002) or any other prior $P(\theta)$, which is implicitly defined by some random (field) generator. In our work, we use training images.

In these cases, it is more reasonable to have an acceptance rate α independent of the prior $P(\theta)$, but explictily enforcing the prior within the proposal density. Through changing the proposal distribution $q(\theta_i, \theta_j)$ to

$$q(\theta_i, \theta_j) = \frac{P(\theta_j)}{P(\theta_i)} q(\theta_j, \theta_i), \tag{11}$$

this can be achived (Mosegaard and Tarantola, 1995). Inserting Eq. 11 into Eq. 7 results in (e.g. Tarantola, 2005)

$$\alpha(\theta_i, \theta_j) = min\left[\frac{L(\theta_j)}{L(\theta_i)}, 1\right] = min\left[e^{l(\theta_j)-l(\theta_i)}, 1\right]. \tag{12}$$

This idea was termed “extended Metropolis sampling” by Hansen et al. (2012). This strategy can be nicknamed “sampling from the prior distribution” for easy understanding. New proposed values are only rejected based on the likelihood ratio and not based on the prior. Sampling from the prior distribution makes the algorithm converge faster because we can make larger changes with similar acceptance rates. This leads to the conclusive question: *How can we find a proposal distribution which satisfies Eq. 11 for a given problem class?*

Different algorithms in the literature satisfy Eq. 11. In the following, we name two approaches. First, the preconditioned Crank-Nicolson algorithm (pCN-MCMC) fulfills this property for multi-Gaussian priors (Beskos et al., 2008; Cotter et al., 2013). Second, the Gibbs approach fulfills this property by conditional resampling parts of the parameter space (Geman and Geman, 1984). In Hansen et al. (2012), this approach was used for resampling boxes in the parameter space of a categorical field. We are extending this approach and make it applicable to hierarchical models in the next section.

2.3. Sequential box resampling

In this subsection, we first describe how we express categorical geostatistical fields with internal heterogeneity as a hierarchical model. Then, we present our novel MCMC algorithm, which fulfills Eq. 11.

We explain our procedure with the help of Fig. 1 throughout this subsection. Let us start by looking at the first column and defining our model. We assume a hierarchical model that consists of several multi-Gaussian fields. Therefore we assume three things: First, that we have an indicator field θ^c, which decides which category (facies) is present at which location. Second, that each category (facies) θ^{f_i} is internally multi-Gaussian. Third, that N_f categories (facies) exist in our domain (two categories in Fig. 1).

We assume a fixed discretization of the considered domain. In our example in Fig. 1, we have 50 × 50 elements which results in a total of 2500 elements. We call this the number of elements N_e. With that, the categorical (or indicator) field $\theta^c \in \{1, 2, \dots N_f\}^{N_e}$ is a vector of size N_e which takes integer values. The internal heterogeneity of each category $\theta^{f_i} \in \mathbb{R}^{N_e}$ are vectors of the same size and take real values.

For shorter notation, we define θ as a combination of θ^c and θ^{f_i}

$$\theta = \begin{pmatrix} \theta^c \\ \theta^{f_1} \\ \vdots \\ \theta^{f_{N_f}} \end{pmatrix}. \tag{13}$$

Here, the parameters θ^c and θ^{f_i} are vectors where each element $\theta^{c,k}$ and $\theta^{f_i,k}$ is the parameter at a specific spatial position k. Thus, θ is a matrix of parameters where θ^k is the vector containing $\theta^{c,k}$ and all $\theta^{f_i,k}$. Given the categorical field θ^c and the internal heterogeneity θ^{f_i} we define the quantity of interest s (the log-conductivity in our application) as

$$s(\theta) = \sum_{i=1}^{N_f} \delta_{i,\theta^c} \cdot \theta^{f_i} \tag{14}$$

with Kronecker delta $\delta_{i,j}$. In this formula, θ^c is an indicator which category is present at which spatial location. δ_{i,θ^c} (which is $\delta_{i,\theta^{c,k}}$) is a vector of zeros and ones, indicating whether category number i is present at location k. Taking the element-wise product $(\delta_{i,\theta^c} \cdot \theta^{f_i})$ sets the quantity of interest s equals to θ^{f_i} if and only if $i = \theta^c$ at that position. This step is visualized in the first column of Fig. 1.

Next, we want to sample new parameters θ_i. In the following, we use lower indices (e.g. θ_i) to declare different samples and upper indices (e.g., θ^c) to declare a part of a sample. We define the random proposal function $g(\theta)$ with proposal distribution $q(\theta_i, \theta_j)$. Hereby, $q(\theta_i, \theta_j)$ is the probability density that $g(\theta_i) = \theta_j$ under the probilistic proposal function g. We take a block-Gibbs approach as presented in Hansen et al. (2012) and do this in two steps. First, we decide which parameters to keep which one to delete, and second we conditionally resample the deleteted parameters.

To do so, we define a 'box' of parameters, the set $\mathbf{\Gamma} \subseteq \{1, 2, \dots, N_e\}$, to resample, based on two arguments: A center point and a diameter of the box. The center point is chosen randomly and independently of θ. This independence is needed for convergence (see Appendix). In our implementation, each position has the same probability of being the center point.

Next, we need to fix the size of the box. The size of the box defines how much the resulting field $s(\theta)$ changes in one step (similar to the σ in the Metropolis-Hastings proposal in Eq. 10). A large box diameter leads

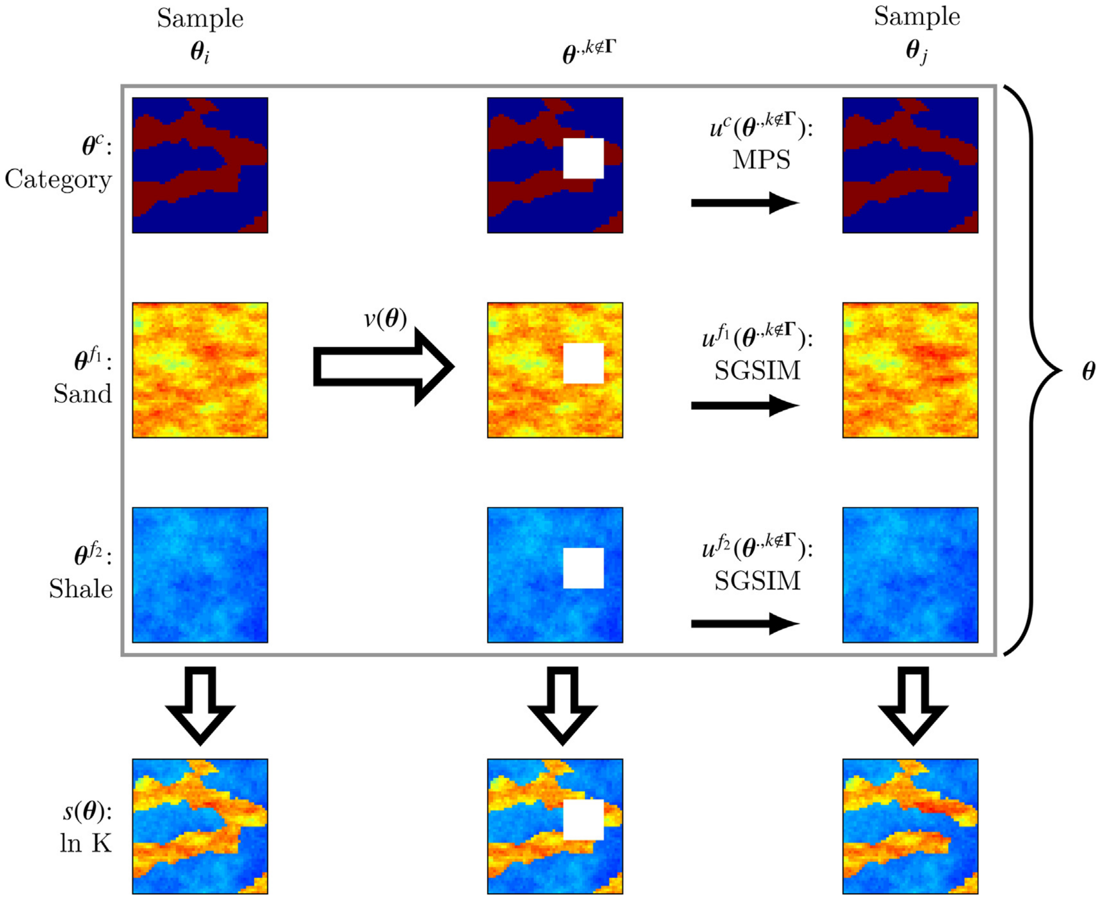

Sample
$\boldsymbol{\theta}_i$
$\boldsymbol{\theta}^{\cdot,k\notin\boldsymbol{\Gamma}}$
Sample
$\boldsymbol{\theta}_j$
$\boldsymbol{\theta}^c$: Category
$u^c(\boldsymbol{\theta}^{\cdot,k\notin\boldsymbol{\Gamma}})$: MPS
$v(\boldsymbol{\theta})$
$\boldsymbol{\theta}^{f_1}$: Sand
$u^{f_1}(\boldsymbol{\theta}^{\cdot,k\notin\boldsymbol{\Gamma}})$: SGSIM
$\boldsymbol{\theta}$
$\boldsymbol{\theta}^{f_2}$: Shale
$u^{f_2}(\boldsymbol{\theta}^{\cdot,k\notin\boldsymbol{\Gamma}})$: SGSIM
$s(\boldsymbol{\theta})$: ln K

Fig. 1. Proposal step of sequential box resampling method. The quantity of interest $s(\theta)$ is constructed by combining a categorical field and multi-Gaussian fields (Eq. 14, first column). Next, a box is deleted in the categorical and all gaussian fields (second column) and conditionally resampled (third column).

to large changes in each proposal and a low acceptance rate, whereas a small diameter leads to small changes and a high acceptance rate.

One could resample a random set of points in the parameter space instead of a box at a specific position (e.g., Mariethoz et al., 2010). However, we found that this is less efficient compared to the box approach.

Let us first explain some notation. We define the set $\theta^{c,k\notin\boldsymbol{\Gamma}}$ which inherits all parameters of the categorical field θ^c which are not resampled. Respectively, we define $\theta^{f_i,k\notin\boldsymbol{\Gamma}}$ as the set of parameters of the internal heterogeneity fields θ^{f_i} which are not resampled. For shorter notation, we define $\theta^{\cdot,k\notin\boldsymbol{\Gamma}}$ as the union of $\theta^{c,k\notin\boldsymbol{\Gamma}}$ and $\theta^{f_i,k\notin\boldsymbol{\Gamma}}$ (see Eq. 15). Furthermore, we define the resampled parameters of the categorical and the internal heterogeneity fields as $\theta^{c,k\in\boldsymbol{\Gamma}}$ and $\theta^{f_i,k\in\boldsymbol{\Gamma}}$, respectively.

With that, we define the selection function $v(\theta)$

$$v(\theta) := \theta^{\cdot,k\notin\boldsymbol{\Gamma}} := \begin{pmatrix} \bigcup_{i=1}^{N_e} \theta^{c,k\notin\boldsymbol{\Gamma}} \\ \bigcup_{i=1}^{N_e} \theta^{f_1,k\notin\boldsymbol{\Gamma}} \\ \vdots \\ \bigcup_{i=1}^{N_e} \theta^{f_{N_e},k\notin\boldsymbol{\Gamma}} \end{pmatrix}. \tag{15}$$

to choose the persistent parameters $\theta^{\cdot,k\notin\boldsymbol{\Gamma}}$ we want to keep for the resampling. Note, that the same set $\boldsymbol{\Gamma}$ is used for the categorical field θ^c as well as the multi-gaussian fields θ^{f_i}. This step is visualized in the first two columns of Fig. 1.

Now we need to conditionally resample the chosen box $\theta^{\cdot,k\in\boldsymbol{\Gamma}}$. Therefore, we need to sample from the conditional probabilities $p_q(\theta^{c,k\in\boldsymbol{\Gamma}}|\theta^{c,k\notin\boldsymbol{\Gamma}})$ for the categorical field and $p_q(\theta^{f_i,k\in\boldsymbol{\Gamma}}|\theta^{f_i,k\notin\boldsymbol{\Gamma}})$ for the internal heterogeneity field. Here, p_q is the conditional probability of the prior. Let us start with the resampling of the categorical field θ^c. Any conditional sampling method $u^c(\theta^{c,k\notin\boldsymbol{\Gamma}})$ which can sample from $p_q(\theta^{c,k\in\boldsymbol{\Gamma}}|\theta^{c,k\notin\boldsymbol{\Gamma}})$ can do this job for the categorical field. Many multiple point geostatistic (MPS) methods exist in the literature for this part and we use the SNESIM (Strebelle, 2002) algorithm.

Next, each multi-Gaussian field θ^{f_i} is repopulated for the internal heterogeneity. Therefore, we need a conditional sampler $u^{f_i}(\theta^{f_i,k\notin\boldsymbol{\Gamma}})$ which is able to sample from $p_q(\theta^{f_i,k\in\boldsymbol{\Gamma}}|\theta^{f_i,k\notin\boldsymbol{\Gamma}})$. Different sequential Gaussian simulation (SGSIM) tools exist which are capable of doing so. We use the SGSIM algorithm of the GSLIB library described in Deutsch and Journel (1992). We chose the SNESIM algorithm and the GSLIB library because they are widely used and freely available online. The last two columns in Fig. 1 show how the conditional sampling method is resampling the deleted parameter box. We can rewrite the proposal function $g(\theta)$ as one function and get

$$\theta_j = g(\theta_i) = \begin{pmatrix} u^c(v(\theta_i)) \\ u^{f_1}(v(\theta_i)) \\ \vdots \\ u^{f_{N_f}}(v(\theta_i)) \end{pmatrix} \tag{16}$$

in which $v(\theta_i)$ decides the position of the resampling box, and we use the conditional resampling functions $u^c(\theta^{c,k\notin\boldsymbol{\Gamma}})$ and $u^{f_i}(\theta^{f_i,k\notin\boldsymbol{\Gamma}})$ discussed above. In the Appendix, we show a proof that, if u^c and u^{f_i} are chosen correctly, we fulfill the detailed balance as in Eq. 11.

2.4. Parallel tempering MCMC

The problem specified in Section 2.3 is high-dimensional and multimodal. This leads to two challenges for MCMC techniques: Long burn-in times (the period in which the MCMC chain converges towards the final range of values) and the risk of getting stuck in one mode (i.e. one local optima).

Laloy et al. (2016) showed that parallel tempering solves both these problems. First, it increases efficiency in high-dimensional geostatistical inversion. They state that parallel tempering increases “convergence towards appropriate data misfit and [the] sampling diversity”. Second, it reduces the risk of only finding one mode (local optima) in the posterior. Especially for complex multimodal problems, not using parallel tempering results in being trapped in local optima. This phenomenon is less likely with parallel tempering (Laloy et al. (2016)).

The idea of parallel tempering (e.g. Earl and Deem, 2005) is to run several chains on different temperatures $T = [T_1, \dots T_n]$ with $1 = T_1 < T_2 < \dots T_n$. Each temperature defines the posterior density at temperature T

$$p(\boldsymbol{\theta}, T|\boldsymbol{d}) \propto p(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\boldsymbol{d})^{1/T}. \tag{17}$$

Increasing the temperature T flattens the posterior towards the prior. In the limit of $T \to \infty$, the tempered distribution $p(\boldsymbol{\theta}, T|\boldsymbol{d})$ becomes equal to the prior distirbution $p(\boldsymbol{\theta})$. In the other limit, $T = 1$ yields $p(\boldsymbol{\theta}, T|\boldsymbol{d})$ equal to the real posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d})$. Thus, only the chain with a temperature of $T = 1$ can be used for posterior sampling. The remaining chains are constructed to help the first (productive) chain in exploring the posterior distribution. In the meantime, the first chain exploits the “good” regions found by the other chains. Hot chains can be built to make farther jumps (due to the smoother tempered likelihood function) than the colder chains while accepting a similar percentage of proposals. The farther jumps, in our context, mean larger resampling boxes of hotter chains. A chain at $T \to \infty$ always accepts all proposals from the prior when using a “sampling from the prior strategy” as in Eq. 11.

To make use of all chains, the chains need to communicate with each other. Therefore, between every few in-chain MCMC steps, a between-chain swap is proposed, which gets accepted with probability

$$\alpha_s(\theta_i, \theta_j) = min\left[\frac{L(\theta_j)^{\left(\frac{1}{T_i} - \frac{1}{T_j}\right)}}{L(\theta_i)}, 1\right]. \tag{18}$$

where T_i and T_j are the temperatures of the chain of θ_i and θ_j. If accepted, the parameters of these two chains get swapped.

3. Test cases and implementation

3.1. Testing procedure

As application test, we infer the hydraulic conductivity of a confined aquifer based on hydraulic head data using a groundwater flow model for fully saturated conditions. We chose this problem because it is a typical problem in geoscience which is challenging due to dimensionality (as a result of the spatial discretization). We focus on channelized flow consisting of two different heterogeneous porous media (here: sand and shale).

We are interested in two different test cases. First, a steady-state test case with weakly informative data (25 measurements once in time). Second, a transient highly informative test case (25 measurements at ten different time steps; 250 measurements in total). In a highly informative case, the main challenge lies in *finding* a suitable parameter set. In contrast, in a weakly informative case with many possible outcomes (due to the limited available data), *exploring* the possibly multi-modal posterior is challenging.

For the latter case, we use clustering algorithms to show the different parameter modes and to quantify how likely they are. This visualization is an enrichment to only showing mean and variance because latter statistics cannot visualize the multi-modality of distributions.

Next, we show that the algorithm convergences during runtime. We do so by independently restarting our algorithm five times and then computing the potential scale reduction factor $\sqrt{\hat{R}}$ introduced by Gelman and Rubin (1992). $\sqrt{\hat{R}}$ measures how similar the results of different runs are. By showing that the results of different runs are similar, we can conclude that convergence is likely. This way, we can asses convergence without having a reference solution, which cannot be produced in our problems due to the high complexity of the model. Gelman et al. (1995) proposed that $\sqrt{\hat{R}} \leq 1.2$ signifies acceptable convergence. We try to reach that value for all parameters i.e., for each pixel of the random field. A complete introduction to $\sqrt{\hat{R}}$ can be found in Gelman et al. (1995).

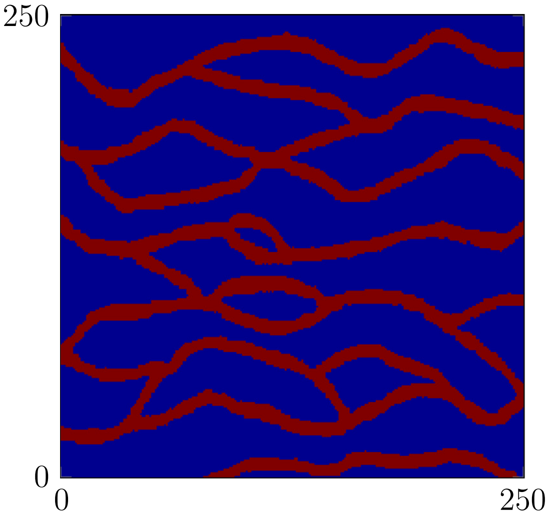


Fig. 2. Training image used in the SNESIM algorithm (Strebelle, 2002).

Table 1
Parameter of the variograms, modeling the heterogenities in the facies.

Facies	Proportions	Mean [$ln(m/d)$]	Variogramm type	k_x	k_y	Sill
Sand	0.35	2.3	Exponential	48	24	1
Shale	0.65	-3.5	Exponential	24	24	0.35

3.2. Setup and test cases

In this section, we give a short overview of the test cases and describe them shortly. We use the benchmark proposed in Xu and Gómez-Hernández (2015). It is a synthetic confined aquifer which is 50m × 50m × 5m large. It is discretized into 50 × 50 × 1 cells. It is composed of 65% low-conductivity shale and 35% high-conductivity sand. The spatial sand and shale distributions are characterized by the training image by Strebelle (2002) shown in Fig. 2. The hydraulic log-conductivity inside each facies follows a multi-Gaussian distribution with exponential variograms. The parameters for these variograms are shown in Table 1. For simplicity, the specific storage S_0 is homogenous with $S_0 = 0.1\ m^{-1}$.

Inside the domain, flow can be described using the saturated groundwater flow equation

$$\nabla[K(x,y)\nabla h(x,y,t)] = \eta(x,y) + S_0 \frac{\partial h(x,y,t)}{\partial t}, \tag{19}$$

where $K(x, y)$ is the isotropic hydraulic conductivity and η encapsulates all source and sink terms. This equation can be solved for the hydraulic head $h(x, y, t)$. Fig. 3 shows the synthetic reference conductivity field $K(x, y)$ and the used boundary conditions. It consists of a spatially distributed conductivity field with a so-called general head boundary condition (Harbaugh et al., 2000; Xu and Gómez-Hernández, 2015) on the left side. Further, we assume no-flow boundary conditions on the top and bottom and fixed outflow on the right side at the positions marked in Fig. 3. As an initial condition for the transient case, we assume a constant head of 8 m.

As mentioned earlier, we define a highly informative (transient flow) and a weakly informative (steady-state flow) test case. In the transient

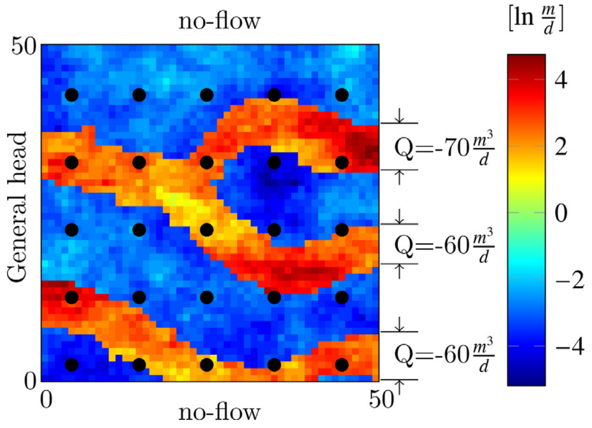


Fig. 3. Log-conductivity field of the synthetic reference aquifer with a general head boundary condition (Harbaugh et al., 2000) on the left side, no-flow boundary conditions at top and bottom and fixed outflow at the right side. The positions of the measurement wells are marked in black.

Table 2
Parameters of the transient test case.

Temperature T	1.7^0	1.7^1	1.7^2	1.7^3	...	1.7^{19}
Box size w	6	7	8	9	...	25

flow case, the head values change from the initial conditions towards a steady-state solution. We assume 25 measurements at the marked positions in Fig. 3 over time. In practice, we save the computed heads at these points after [10, 21, 34, 49, 65, 83, 104, 127, 153, ∞] days and add normal distributed noise with standard deviation 0.05 m to simulate real-world measurements. In the second scenario, we assume that only the steady-state measurements (after ∞ days) are available.

3.3. Implementation

This section specifies all parameters of the used algorithms. We first introduce our forward solver, then our conditional samplers and then focus on parallel tempering. Finally, we report the machine used for the numerical experiments.

The groundwater equation is solved using MODFLOW (McDonald and Harbaugh, 1988; Harbaugh et al., 2000). We decided to use this solver because it is widely used in the literature. The SNESIM resampling is done using the training image shown in Fig. 2. The SGSIM algorithm is run with the parameters presented in Table 1.

The parameters of the MCMC are heavily influencing the result. The algorithm only converges fast if the right parameters are chosen. Choosing these parameters is complicated and is broadly discussed in the literature (Gelman et al., 1996; Roberts et al., 1997; Roberts and Rosenthal, 2002). A target acceptance rate in the range of $10\% - 50\%$ is generally recommended. We tried to get an acceptance rate of approximately 23.4% (Gelman et al., 1996) for all chains. We recommend reading Gelman et al. (1996) for a good introduction on how to chose these parameters for one chain and Laloy et al., (2016) for parallel tempering.

The box size can be chosen adaptively (Hansen et al., 2012; Laloy et al., 2016) during burn-in. We refrain from doing that, because that would lead to different box sizes in each test run. This makes it ambiguous whether differences between independent MCMC runs occur due to different box sizes or because of slow convergence of the algorithm. Instead, we did a manual tuning in smaller test runs and used identical settings in each independent MCMC run.

In the transient case, we use 20 parallel chains with the parameters shown in Table 2. In the steady-state case, we use 12 chains with the parameters shown in Table 3. We use different box sizes for each chain because preliminary test runs showed that this leads to better results than using the same box size for all chains. Assuming some box size w (number of pixels for the edge length of a square), the number of resampled parameters is always smaller or equal than w^2. It can be smaller than w^2 if the center of the box is close to the border of the domain. In both scenarios, we randomly propose swaps between neighboring chains (T_i and T_{i+1}) after 10 in-chain MCMC steps.

Table 3
Parameters of the steady-state test case.

Temperature T	2^0	2^1	2^2	2^3	...	1.7^{11}
Box size w	9	11	13	15	...	31

We run our experiments on a high-performance cluster where each node has two Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz processors with 10 cores each. On each node, we run independent repetitions of our algorithm to compute $\sqrt{\hat{R}}$. With our implementation, each MCMC step takes around 2 seconds in the transient and 1.2 seconds in the steady-state case. This time mainly consists of the forward simulation (groundwater solver) and the conditional resampling. For 1 million samples, the algorithm runs for 23 days in the transient and 14 days in the steady-state case.

Our MATLAB implementation of the algorithm and all data of the MCMC runs are available at https://doi.org/10.18419/darus-741.

4. Results and discussion

4.1. Preparatory investigations

We define the L_2 error of one sample

$$L_2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \underbrace{(d_i - F_i(\theta))^2}_{e_i}} = \sqrt{l(\theta|d)\cdot\frac{2\sigma_e{}^2}{N}} \tag{20}$$

where $F_i(\theta)$ are the simulated equivalents to the measured data d_i, e_i are the residuals and N is the number of measurements. This L_2 error can be seen as an averaged error of predictions and is a slight variant of the log-likelihood $l(\theta|d)$ defined in Eq. 4.

The L_2 error converges roughly to σ_e (total error standard derivation, here: 0.05) independent of the number of measurements N used if $N \gg 1$ (it converges exactly to σ_e for N $\to \infty$). Fig. 4 shows the L_2 error of all experiments over the number of iterations. The L_2 error, in the first iteration, is high in all cases and converges towards 0.05. The burn-in contains samples with extremely low likelihoods (high L_2 errors), which would distort all further investigations. Thus, we delete all burn-in samples in all further investigations. Detecting the length of the burn-in was done manually and chosen to be 10,000 samples in the weakly informative case and 100,000 samples in the highly informative case. The long burn-in time in the highly informative test case shows that finding samples with a good fit to the data is harder (compared to the weakly informative test case). The remaining 990,000 samples (out of 1,000,000) in the weakly informative case and 1,000,000 samples (out of 1,100,000) in the highly informative case are used for all statistics below.

A second observation in Fig. 4, besides the burn in effect, is that the L_2 error fluctuates less in the highly informative test case. The key explanation is that the number of data over which the squared residuals e_i^2 are averaged in Eq. 20 is larger (250 versus 25). This provides a more stable L_2 simply by more stable sample statistics.

The signal to noise ratio (SNR) is defined as

$$\text{SNR} = \frac{\text{average } L_2 \text{ error of prior samples}}{\text{measurement error}}. \tag{21}$$

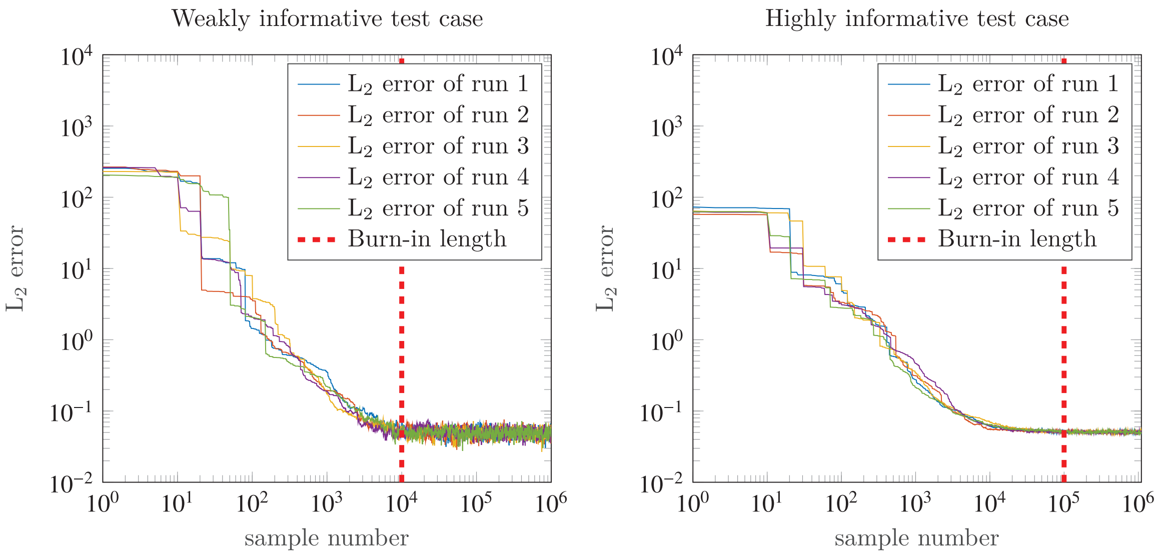


Fig. 4. L_2 error over sample number of the coldest chain (T=1). The five conducted runs per test case are shown. The manually chosen burn-in length is marked in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The transient test case has a SNR of approximately 650 and the steady state test case has a SNR of approximately 2000. The SNR in the transient case is lower because the head at t=0 is fixed, leading to a smaller prior variance of the heads. This SNR suggests that the sampling problem is reasonably hard in both cases.

4.2. Test case 1: Highly informative data

First, we have a look at the highly informative test case. Fig. 5 shows independent samples of the posterior sampled by our MCMC algorithm. All samples are alike and similar to the synthetic solution.

Next, we want to look at the whole ensemble. Fig. 6 shows the mean and standard deviation of the whole ensemble with burn-in removed. We can see two things: First, the MCMC can find the spatial position of the sand channel. Second, it is uncertain about the exact position of it. The latter is expressed by the high standard deviation at the borders of the channel. In conclusion, the algorithm can produce results that are similar to the synthetic truth.

Next, we want to make sure that we achieve this behavior every time we restart the algorithm. Thus, we want to compare different independent test runs and show that all converge to the same posterior. We do that in a two-step approach. First, we have a look at several independent mean and standard deviation fields to get a better understanding of where weaknesses could lie. Second, we use the scale reduction factor $\sqrt{\hat{R}}$ to quantify the convergence of our test runs.

Fig. 7 shows two mean and standard deviation fields of two independent test runs. At first sight, the results look mainly similar, although we also see some differences (e.g., in the top right corner). This observation emphasizes the question of how crucial these differences are.

The scale reduction factor $\sqrt{\hat{R}}$ can answer this question and is shown in Fig. 8. The left side of Fig. 8 shows the spatial distribution of $\sqrt{\hat{R}}$. We see that it did not converge to values lower than 1.2 everywhere. We see that the top right corner is an area of concern and we should try to improve predictions in this area. The right side of Fig. 8 shows the mean and maximum scale reduction factor $\sqrt{\hat{R}}$ over the length of the Markov chains. This indicates how much longer we need to run the MCMC algorithm until the maximum $\sqrt{\hat{R}}$ gets smaller than the 1.2 treshold. Furthermore, it shows that the mean $\sqrt{\hat{R}}$ value has reached the 1.2 mark after $2 \cdot 10^5$ iterations.

4.3. Test case 2: Weakly informative data

Let us have a look at the weakly informative test case. Fig. 9 shows individual samples of the posterior, sampled by our MCMC algorithm. One can see that these samples neither look alike nor similar to the synthetic reference solution. Nevertheless, all these samples are valid solutions. The reason they look different is that the likelihood is less restrictive (weakly informative). This leads to a broader posterior (many different possible solutions).

To investigate this aspect further, we have a look at the mean and variance of the whole ensemble. Fig. 10 shows the mean and standard deviation of one test run. The posterior has a low uncertainty concerning the position of the sand channel at the right and left boundary of the domain (similar to the highly informative test case). However, in the middle of the domain, the inversion results suggest that the connection of the sand channel from left to right is unknown. Hence, based on our (limited) measurements, we do not know the spatial course of the sand channel between the left and right boundary.

Only looking at mean and standard deviation is not informative for two reasons: First, the mean does not look similar to individual samples due to the spatial smoothing that occurs in the ensemble average. Second, the standard deviation is remarkably high, which could indicate a multi-modal posterior in such a non-linear and non-Gaussian problem. Thus, we need to do further analysis and investigate the mean and standard deviation of each potential mode. From a machine learning point of view, each mode can be represented by one cluster. Hence, we can find modes by clustering the posterior.

We used k-means clustering with 4 clusters and a euclidean distance. An introduction to k-means clustering can be found in Hastie et al. (2005). We chose 4 clusters because it produced the best results for our test case. Because the cluster algorithms are susceptible to the input data, we want to emphasize that the cluster look remarkably different when produced for various test runs. Furthermore, other norms (instead of the euclidean distance) and different clustering algorithms change the results as well. However, the discussed conclusions are not affected. Fig. 11 shows the mean and standard deviation of the posterior of a representative clustering example.

On the top row of Fig. 11, we see the mean fields of the clusters and the respective probability of the clusters. The probability is defined as the percentage of samples that lie inside the respective cluster. We see that 48% of samples are similar to the synthetic reference solution

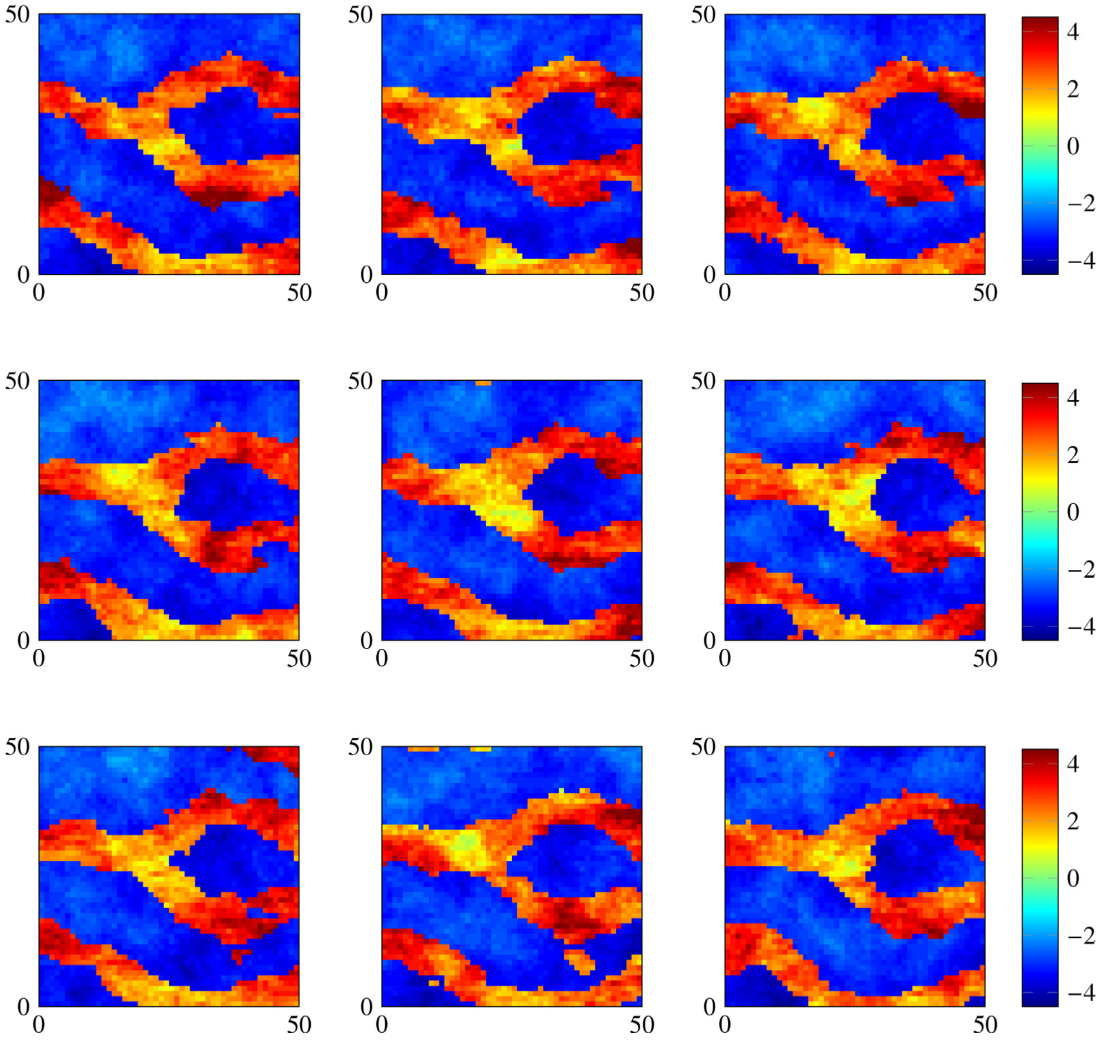


Fig. 5. Independent samples in the highly informative test case.

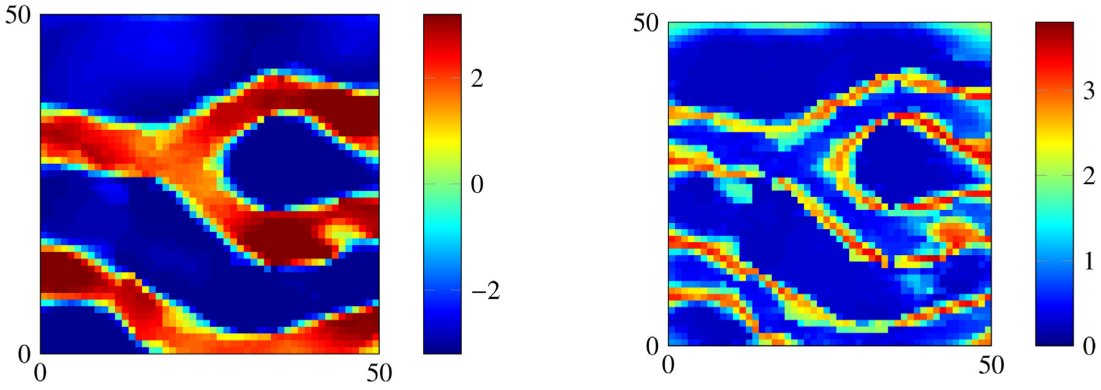


Fig. 6. Expected values (left) and standard deviation (right) in the highly informative test case.

(cluster 1). However, based on the available information, we see that three other clusters are also possible. To exclude these other clusters in the inversion, one would need more (informative) measurement data. We show this example to emphasize clustering as a possible tool to investigate multi-modal distributions by splitting them into more homogeneous sub-distributions. This clustering resembles a non-parametric version of Gaussian mixture models for the posterior.

Next, we consider the standard deviation within the clusters. We can see that the standard deviation of the clusters is significantly smaller than the total standard deviation in Fig. 10. This observation indicates that clustering reduced the uncertainty around the cluster-wise mean fields shown in Fig. 11 drastically when compared to the non-clustered field in Fig. 10. Hence, the mean fields of the clusters are more reliable and should be used for further investigations.

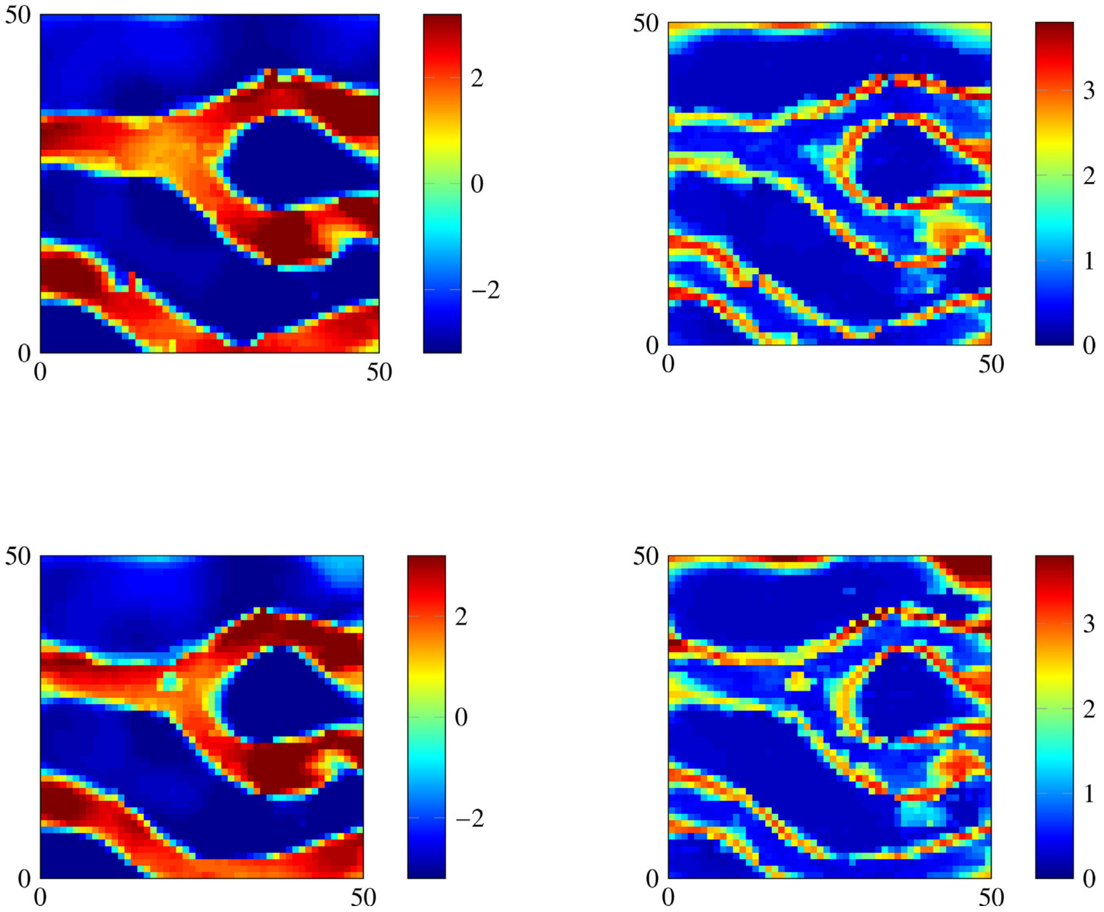


Fig. 7. Expected values (left) and standard deviation (right) in the highly informative test case for two independent test runs (top and bottom).

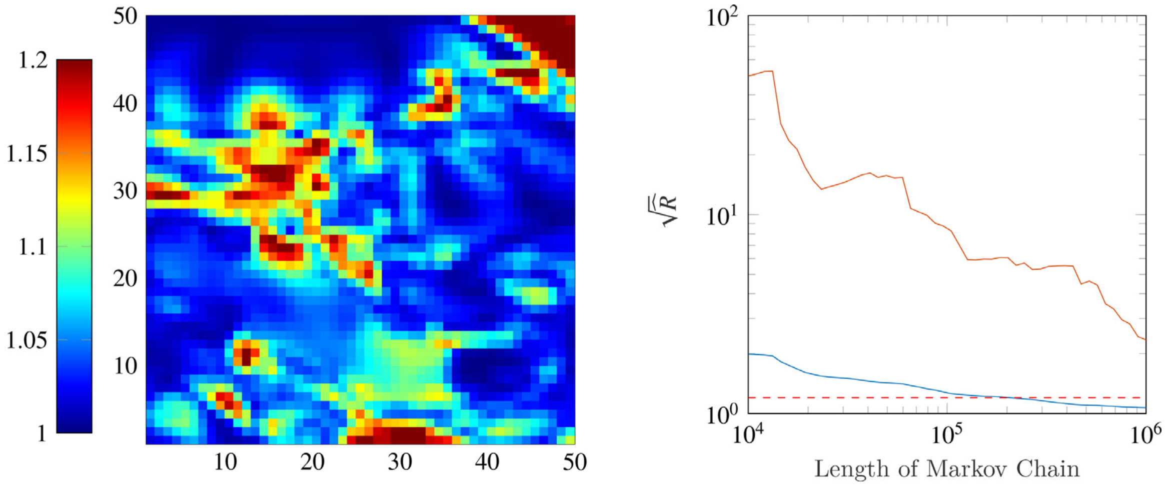


Fig. 8. Scale reduction factor $\sqrt{\hat{R}}$ in the highly informative test case. On the left the spatial distribution of the scale reduction factor is illustrated. On the right the spatial average (blue) and maximum (red) for different lengths of Markov chains is shown. The red dotted line signifies $\sqrt{\hat{R}} = 1.2$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Next, we show the convergence of our algorithm. Fig. 12 shows the mean of two different test runs. We can see that the mean fields look similar. To quantify this similarity, we have a look at the scale reduction factor $\sqrt{\hat{R}}$ in Fig. 13. We see that it is lower than 1.2 everywhere in the domain. This indicates that the MCMC algorithm converged sufficiently well. Furthermore, we can see that this criterion is reached after approximately $2 \cdot 10^5$ MCMC steps.

This looks nice but, is the scale reduction factor $\sqrt{\hat{R}}$ the right statistic to use? The scale reduction factor $\sqrt{\hat{R}}$ checks the convergence of the mean value and not of the distribution. Hence, $\sqrt{\hat{R}}$

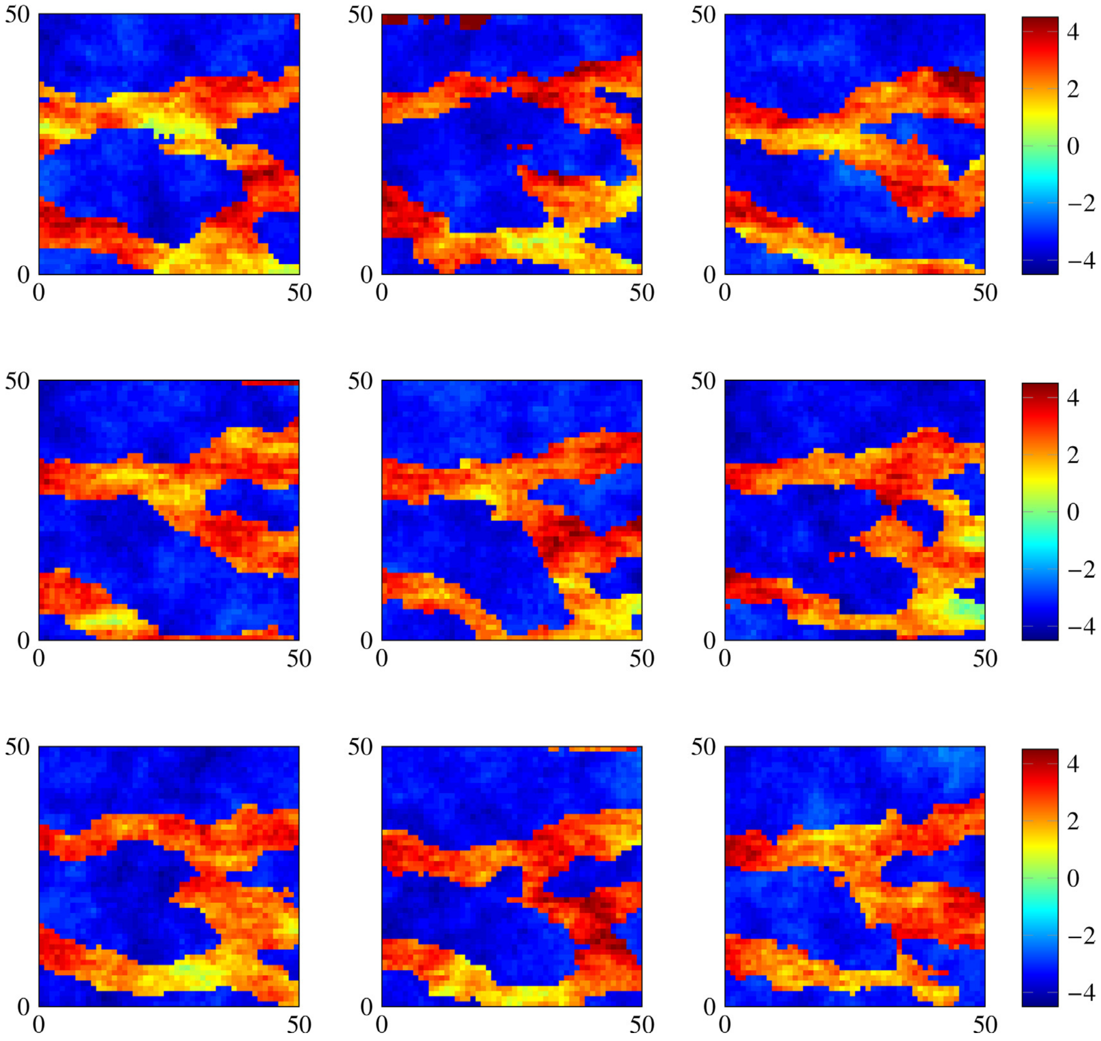


Fig. 9. Independent samples in the weakly informative test case.

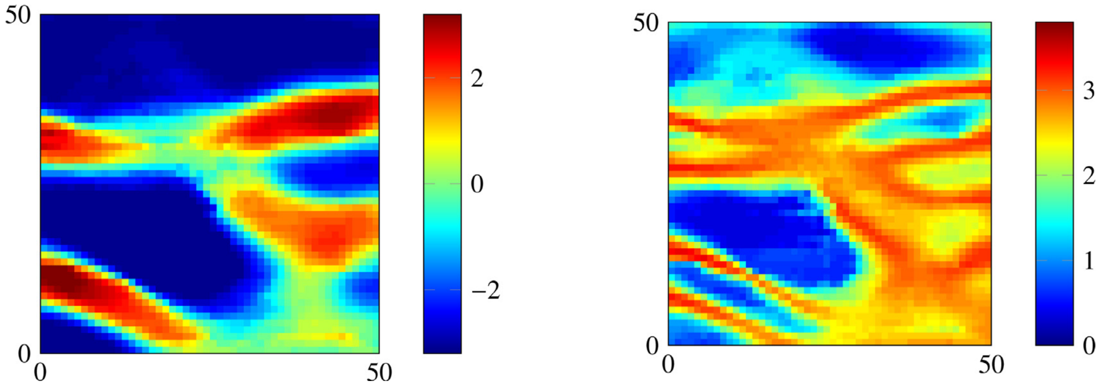


Fig. 10. Expected values (left) and standard deviation (right) in the weakly informative test case.

might be the wrong norm in multi-modal applications, such as the current test case. Other measures like the Kullback-Leibler divergence (Kullback and Leibler, 1951) might be more suitable. However, we used $\sqrt{\hat{R}}$ because it is widely used in the literature.

4.4. Summary

The proposed algorithm converges to the posterior in the weakly and highly informative test case settings. The convergence is measured using the scale reduction factor $\sqrt{\hat{R}}$. The highly informative test case reaches the benchmark value of $\sqrt{\hat{R}} < 1.2$ in 92.76% of the parameter cells and the weakly informative test case reaches it everywhere.

The highly informative test case shows a posterior that is uni-modal and similar to the reference solution. The weakly informative test case shows a posterior that is multi-modal and can be split using a clustering algorithm. We visualize all possible scenarios and their respective probabilities.

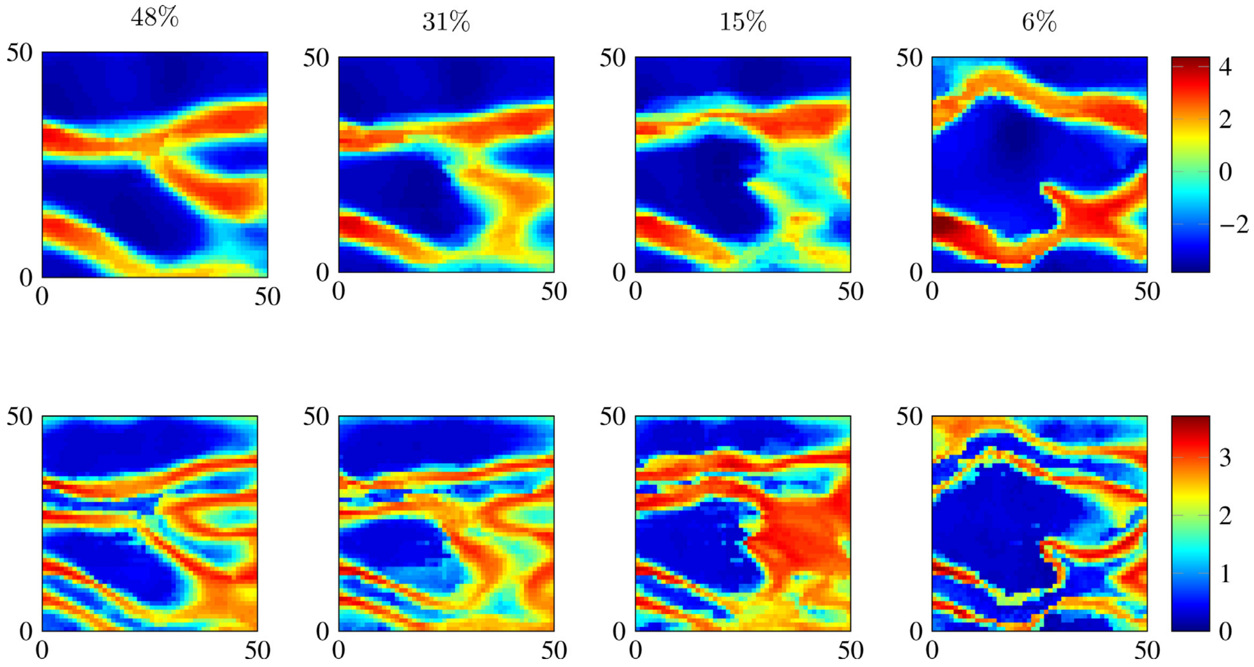


Fig. 11. Clusters of the posterior ensemble. Top: The mean value of the 4 different clusters. Bottom: The respective standard deviation in the clusters. The percentage of samples in each cluster is noted over the respective column.

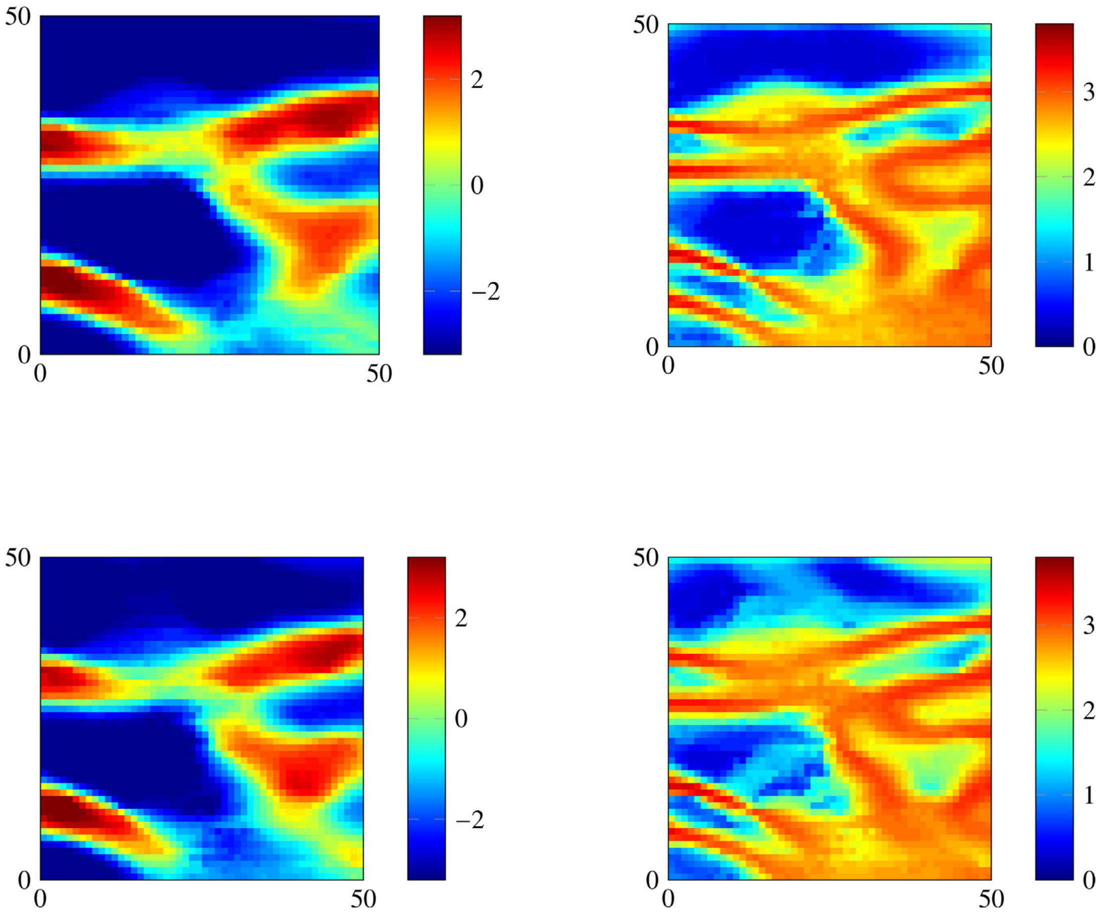


Fig. 12. Expected values (left) and standard deviation (right) in the weakly informative test case for two independent test runs (top and bottom).

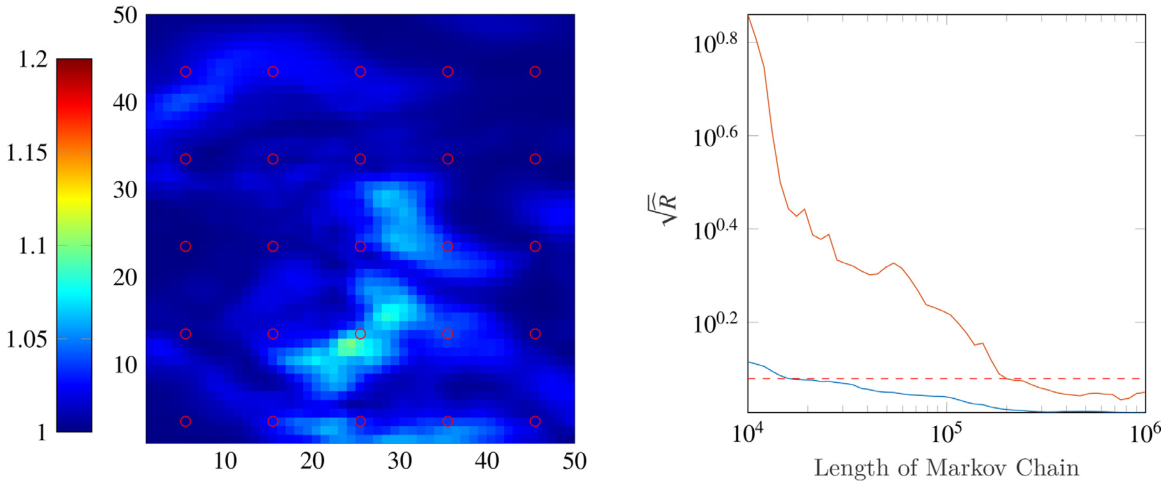


Fig. 13. Scale reduction factor R in the weakly informative test case. On the left the spatial distribution of the scale reduction factor is illustrated. On the right the spatial average (blue) and maximum (red) for different lengths of Markov chains is shown. The red dotted line signifies $\sqrt{\hat{R}} = 1.2$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusion

This work enables realistic inversion of channelized flow in the subsurface. Thereby it solves the categorical decision (which facies is present) and the heterogeneity within each facies in a hierarchical framework. To achieve this goal, we propose a novel MCMC algorithm that combines parallel tempering and sequential resampling.

This algorithm is an extension of Laloy et al. (2016), who only treated categorical fields. Compared to existing solution methods such as the EnKF approach in Xu and Gómez-Hernández (2015), it converges to the true solution by design, not just to an implicit quasi-linearized solution.

We test our algorithm on a highly- and weakly-informative test case and it converges in both cases. The MCMC converges although the posterior is extremely narrow in the highly informative test case and broad and multi-modal in the weakly informative test case. This shows the general applicability of our method.

Declaration of Competing Interest

The authors declare that they do not have any financial or nonfinancial conflict of interests

CRediT authorship contribution statement

Sebastian Reuschen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Teng Xu:** Software, Resources, Writing - review & editing, Supervision. **Wolfgang Nowak:** Conceptualization, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project Number 327154368 - SFB 1313 and Project Number 359880532 - COMPU-FLOW.

Appendix A

In the following, we show the derivation of the detailed balance. The detailed balance is defined as

$$\pi(\theta_i)h(\theta_i,\theta_j) = \pi(\theta_j)h(\theta_j,\theta_i) \tag{A.1}$$

with the transition kernel h, which is usually defined as

$$h(\theta_i,\theta_j) = q(\theta_i,\theta_j)\alpha(\theta_i,\theta_j). \tag{A.2}$$

Combining these equations, the detailed balance can be written as

$$\pi(\theta_i)q(\theta_i,\theta_j)\alpha(\theta_i,\theta_j) = \pi(\theta_j)q(\theta_j,\theta_i)\alpha(\theta_j,\theta_i). \tag{A.3}$$

Inserting Eq. (3), it follows that

$$P(\theta_i)L(\theta_i)q(\theta_i,\theta_j)\alpha(\theta_i,\theta_j) = P(\theta_j)L(\theta_j)q(\theta_j,\theta_i)\alpha(\theta_j,\theta_i). \tag{A.4}$$

After resorting this equation, we find that

$$\alpha(\theta_i,\theta_j) = \frac{P(\theta_j)L(\theta_j)q(\theta_j,\theta_i)}{P(\theta_i)L(\theta_i)q(\theta_i,\theta_j)}\alpha(\theta_j,\theta_i). \tag{A.5}$$

Combining this equation with the property of $0 \leq \alpha \leq 1$, one obtains that

$$\alpha(\theta_i,\theta_j) = min\left[\frac{P(\theta_j)L(\theta_j)q(\theta_j,\theta_i)}{P(\theta_i)L(\theta_i)q(\theta_i,\theta_j)}, 1\right]. \tag{A.6}$$

Appendix B

In the following, we show a proof that our algorithm fulfills the proposal distribution as specified in Eq. (11) (and hence fulfills the detailed balance), assuming that the functions u^c and u^{f_i} are able to sample from the distribution $p_q({\theta_j}^c|\theta^{\cdot,k\notin\Gamma})$ and $p_q({\theta_j}^{f_i}|\theta^{\cdot,k\notin\Gamma})$ respectively.

First, we partion the parameter space into

$$\theta_i = [\theta_i^{\cdot,k\notin\Gamma}, \theta_i^{\cdot,k\in\Gamma}] \tag{B.1}$$

with the persistent parameters $\theta_i^{\cdot,k\notin\Gamma}$ and the resampled parameters $\theta_i^{\cdot,k\in\Gamma}$. Eq. (16) defines the proposal function to be

$$\theta_j = g(\theta_i) = \begin{pmatrix} u^c(v(\theta_i)) \\ u^{f_1}(v(\theta_i)) \\ \vdots \\ u^{f_{N_f}}(v(\theta_i)) \end{pmatrix} \tag{B.2}$$

Evidently, if $\theta_i^{\cdot,k\notin\Gamma} \neq \theta_j^{\cdot,k\notin\Gamma}$ it follows that $q(\theta_i,\theta_j) = q(\theta_j,\theta_i) = 0$ which fulfills Eq. (11).

Thus, we know that $\theta_i^{\cdot,k\notin\Gamma} = \theta_j^{\cdot,k\notin\Gamma} = \theta^{\cdot,k\notin\Gamma}$. We can resample the rest ${\theta_i}^r$ of the parameter space based on fixed part $\theta^{\cdot,k\notin\Gamma}$. In the given setting, we can express $q(\theta_i, \theta_j)$ as

$$q(\theta_i,\theta_j) = p_q(\theta_j|\theta^{\cdot,k\notin\Gamma})p_w(\theta^{\cdot,k\notin\Gamma}|\theta_i)\ \forall\theta^{\cdot,k\notin\Gamma} \tag{B.3}$$

where $p_q(\theta_j|\theta^{\cdot,k\notin\Gamma})$ is the probability density of θ_j based on $\theta^{\cdot,k\notin\Gamma}$ given by some conditional re-sampling method and $p_w(\theta^{\cdot,\,k\notin\Gamma}|\theta_i)$ is the probability of the persistent data based on the previous sample. Assuming that we choose the parameter box indepently of the θ_i, θ_j, we know that

$$p(\theta^{\cdot,k\notin\Gamma}|\theta_i) = p(s) \tag{B.4}$$

where $p(s)$ is the probability that $\theta^{\cdot,k\notin\Gamma}$ is chosen. One could imagine it as the probability to place the box at a certain position such that it excludes exactly $\theta^{\cdot,k\notin\Gamma}$. Assuming non-zero probabilities $p(s)$ and $p(\theta^{\cdot,k\notin\Gamma})$ it follows that

$$\frac{q(\theta_i,\theta_j)}{q(\theta_j,\theta_i)} = \frac{p(\theta_j|\theta^{\cdot,k\notin\Gamma})p(s)}{p(\theta_i|\theta^{\cdot,k\notin\Gamma})p(s)} = \frac{p(\theta_j|\theta^{\cdot,k\notin\Gamma})}{p(\theta_i|\theta^{\cdot,k\notin\Gamma})} = \frac{p(\theta_j|\theta^{\cdot,k\notin\Gamma})p(\theta^{\cdot,k\notin\Gamma})}{p(\theta_i|\theta^{\cdot,k\notin\Gamma})p(\theta^{\cdot,k\notin\Gamma})} = \frac{p(\theta_j)}{p(\theta_i)} \tag{B.5}$$

which is equivalent to Eq. (11).

$p(s) = 0$ or $p(\theta^{\cdot,k\notin\Gamma}) = 0$ leads to $q(\theta_i,\theta_j) = q(\theta_j,\theta_i) = 0$ which fulfills Eq. (11) as well. □

References

Beskos, A., Roberts, G., Stuart, A., Voss, J., 2008. MCMC Methods for diffusion bridges. Stochastics Dyn. 8 (3), 319–350. https://doi.org/10.1142/s0219493708002378.

Betancourt, M., 2017. A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv: 1701.02434.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings Algorithm. Am. Stat. 49 (4), 327–335.

Cotter, S.L., Roberts, G.O., Stuart, A.M., White, D., 2013. MCMC Methods for functions: modifying old algorithms to make them faster. Stat. Sci. 28 (3), 424–446. https://doi.org/10.1214/13-STS421.

Deutsch, C.V., Journel, A.G., 1992. Geostatistical software library and user's guide, New York.

Earl, D. J., Deem, M. W., 2005. Parallel tempering: Theory, applications, and new perspectives. arXiv:0508111v2. doi:10.1039/b509983h.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 1995. Bayesian data analysis. Chapman and Hall/CRC.

Gelman, A., Roberts, G.O., Gilks, W.R., 1996. Efficient metropolis jumping rules. Bayesian Stat. 5, 599–608.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7 (4), 457–472.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721–741.

Geyer, C.J., Thompson, E.A., 1995. Annealing Markov chain Monte Carlo with applications to ancestral inference. J Am Stat Assoc 90 (September 1995), 909–920. https://doi.org/10.1080/01621459.1995.10476590.

Hansen, T.M., Cordua, K.S., Mosegaard, K., 2012. Inverse problems with non-trivial priors: efficient solution through sequential Gibbs sampling. Comput. Geosci. 16 (3), 593–611. https://doi.org/10.1007/s10596-011-9271-1.

Harbaugh, B.A.W., Banta, E.R., Hill, M.C., Mcdonald, M.G., 2000. MODFLOW-2000, The U.S. Geological survey modular graound-water model – User guide to modularization concepts and the ground-water flow process. U.S. Geologic. Surv. 130. https://doi.org/10.1029/2006WR005839.

Hastie, T., Tibashirani, R., Friedman, J. (Eds.), 2005. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

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

Iglesias, M.A., Lin, K., Stuart, A.M., 2014. Well-posed Bayesian geometric inverse problems arising in subsurface flow Well-posed Bayesian geometric inverse problems arising in subsurface flow. Inverse Probl. 30. https://doi.org/10.1088/0266-5611/30/11/114001.

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

Laloy, E., Linde, N., Jacques, D., Mariethoz, G., 2016. Merging parallel tempering with sequential geostatistical resampling for improved posterior exploration of high-dimensional subsurface categorical fields. Adv. Water Resour. 90, 57–69. https://doi.org/10.1016/j.advwatres.2016.02.008.

Mariethoz, G., Renard, P., Caers, J., 2010. Bayesian inverse problem and optimization with iterative spatial resampling. Water Resour. Res. 46 (11), 1–17. https://doi.org/10.1029/2010WR009274.

Matheron, G., 1975. Random sets and integral geometry.

McDonald, M., Harbaugh, A., 1988. A modular three-dimensional finite difference ground-water flow model. Techniques of Water-Resources Investigations, book 6 588. https://doi.org/10.1016/0022-1694(70)90079-X.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092. https://doi.org/10.1063/1.1699114.

Mo, S., Zabaras, N., Shi, X., Wu, J., 2020. Integration of adversarial autoencoders with residual dense convolutional networks for estimation of non-Gaussian hydraulic conductivities. Water Resour. Res. 56 (2), 1–24. https://doi.org/10.1029/2019WR026082.

Mosegaard, K., Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems. J. Geophys. Res.: Solid Earth 100, 12431–12447.

Nowak, W., 2009. Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator. Water Resour. Res. 45 (4), 1–17. https://doi.org/10.1029/2008WR007328.

Roberts, G.O., Gelman, A., Gilks, W.R., 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. 7 (1), 110–120. https://doi.org/10.1214/aoap/1034625254.

Roberts, G.O., Rosenthal, J.S., 2002. Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. 16 (4), 351–367. https://doi.org/10.1214/ss/1015346320.

Schöniger, A., Nowak, W., Hendricks Franssen, H.J., 2012. Parameter estimation by ensemble Kalman filters with transformed data: approach and application to hydraulic tomography. Water Resour. Res. 48 (4), 1–18. https://doi.org/10.1029/2011WR010462.

Smith, A., Roberts, G., 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. R. Stat. Soc. 55 (1), 3–23.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. 34 (1), 1–21.

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

Xu, T., Gómez-Hernández, J.J., 2015. Probability fields revisited in the context of ensemble Kalman filtering. J. Hydrol. (Amst) 531, 40–52. https://doi.org/10.1016/j.jhydrol.2015.06.062.

Zhou, H., Gómez-Hernández, J.J., Hendricks Franssen, H.J., Li, L., 2011. An approach to handling non-Gaussianity of parameters and state variables in ensemble Kalman filtering. Adv. Water Resour. 34 (7), 844–864. https://doi.org/10.1016/j.advwatres.2011.04.014.

E Contribution 5: The four ways to consider measurement noise in Bayesian model selection—And which one to choose

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR030391

Key Points:

- We distinguish four cases of how to consider measurement noise in models and/or data within Bayesian model selection (BMS)
- We demonstrate with synthetic and real-world scenarios that the choice of the case significantly affects the outcome of Bayesian model weights
- Finally, we recommend a BMS procedure that ensures philosophically consistent and numerically stable results under noise

Correspondence to:
A. Guthke,
anneli.guthke@simtech.uni-stuttgart.de

Citation:
Reuschen, S., Nowak, W., & Guthke, A. (2021). The four ways to consider measurement noise in Bayesian model selection—And which one to choose. *Water Resources Research*, *57*, e2021WR030391. https://doi.org/10.1029/2021WR030391

Received 11 MAY 2021
Accepted 19 OCT 2021

Author Contributions:
Conceptualization: Sebastian Reuschen, Anneli Guthke
Data curation: Sebastian Reuschen
Formal analysis: Sebastian Reuschen
Funding acquisition: Wolfgang Nowak
Investigation: Sebastian Reuschen
Methodology: Sebastian Reuschen, Anneli Guthke
Project Administration: Wolfgang Nowak
Software: Sebastian Reuschen
Supervision: Wolfgang Nowak, Anneli Guthke
Validation: Sebastian Reuschen, Anneli Guthke
Visualization: Sebastian Reuschen
Writing – original draft: Sebastian Reuschen, Anneli Guthke

© 2021. The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

The Four Ways to Consider Measurement Noise in Bayesian Model Selection—And Which One to Choose

Sebastian Reuschen[1], Wolfgang Nowak[1], and Anneli Guthke[1,2]

[1]Department of Stochastic Simulation and Safety Research for Hydrosystems, University of Stuttgart, Stuttgart, Germany, [2]Now at Stuttgart Center for Simulation Science, University of Stuttgart, Stuttgart, Germany

Abstract Bayesian model selection (BMS) is a statistically rigorous approach to assess the plausibility of competing models. It naturally accounts for uncertainties in models and data. In this study, we discuss the role of measurement noise in BMS deeper than in past literature. We distinguish between four cases, accounting for noise in models and/or data: (1) no-no, (2) no-yes, (3) yes-no, and (4) yes-yes. These cases differ mathematically and philosophically. Only two out of these four cases are logically consistent, and they represent two potentially conflicting research questions: "Which model is best in modeling the pure physics?" (Case 1) and "which model is best in predicting the data-generating process (i.e., physics plus noise)?" (Case 4). If we are interested in the "pure physics question," we face two practical challenges: First, we would need noise-free data, which is impossible to obtain; and second, the numerical approximation of Bayesian model evidence can be hard when neglecting noise. We discuss how to address both challenges and reveal that a fallback to the easier "data-generation question" as a proxy for the "physics question" is not appropriate. We demonstrate on synthetic scenarios and a real-world hydrogeological case study that the choice of the case has a significant impact on the outcome of posterior model weights, and hence on results of the model ranking, model selection, model averaging, model confusion analysis, and uncertainty quantification. Reality might force us to use a different case than philosophy would suggest, and we provide guidance on how to interpret model probabilities under such conditions.

1. Introduction

Models are used to predict and/or investigate and explain phenomena in nature. Often, many hypotheses exist for these two tasks. Naturally, the question arises, which of the competing modeling approaches predicts or explains nature best. Bayesian model selection (BMS, e.g., Wasserman, 2000) is a statistical method that uses observed data to select between competing models. BMS is settled in a rigorous probabilistic framework and follows the scheme of Bayesian updating: A prior belief about the plausibility of each candidate model is updated to a posterior model weight in the light of measured data (i.e., the probability of the model to have generated the data, given the model set). Posterior model weights are then used as a basis for Bayesian model ranking, selection, or averaging (BMA, Hoeting et al., 1999).

To help with the interpretation of posterior model weights, the so-called model confusion matrix (MCM) has been introduced by Schöniger, Illman, et al. (2015). It reveals whether a lack of confidence in model choice is due to similarity between the candidate models or due to weakly informative data. The MCM is a purely synthetic analysis that can be used as a scale of reference for model weights obtained from real data. Schäfer Rodrigues Silva et al. (2020) have recently extended the MCM analysis to identify the best *surrogate* model from a set of candidates to replace an expensive full-complexity model in stochastic analysis.

Technically, the Bayesian updating procedure requires calculating the so-called Bayesian model evidence (BME). BME is the likelihood of a model to have generated the data, integrated over its whole parameter space and all involved probability distributions. While the likelihood accounts for uncertainty in measured data, the integration considers parameter uncertainty, and potentially also uncertainty in model drivers or boundary conditions. In some cases, the integration even accounts for statistical representations of model errors (Leube et al., 2012; Nowak et al., 2012), which is perceived by many studies to be part of the likelihood.

Due to its statistical rigor and its elegance in accounting for uncertainty, BMS has become popular in water resources research. It has been applied in various different contexts, such as evaluation of hydrologi-

Writing – review & editing: Sebastian Reuschen, Wolfgang Nowak, Anneli Guthke

cal models (Marshall et al., 2005), frequency analysis of hydrological extremes (Laio et al., 2009), climate change impact studies (Najafi et al., 2011), model complexity analysis (Höge et al., 2018; Schöniger, Illman, et al., 2015), optimal design for model choice (Nowak & Guthke, 2016), as well as hydrogeophysical (Brunetti et al., 2017), hydro-morphodynamic (Mohammadi et al., 2018), and groundwater transport modeling (Elshall & Ye, 2019).

While the ability to account for uncertainties is the primary reason for the popularity of the Bayesian framework in hydrological multi-modeling analyses, the theoretical basis of how to treat measurement noise in BMS has rarely been in the focus. Lu et al. (2013) have investigated the effect of error covariance structure in the likelihood function on posterior weights for transport models. Schöniger, Wöhling, and Nowak (2015) have demonstrated that posterior weights of soil-plant-atmosphere models can vary significantly under random outcomes of measurement noise in the observed data and that this variability can even change model ranking and corresponding modeling conclusions. What is missing so far is a theoretical dissection of how and for which reasons measurement noise should be treated in BMS. We wish to make modelers aware of the impact this treatment has on the final outcome of model ranking results and dependent conclusions.

For our discussion, we assume a deterministic "physics model" $f_k(\mathbf{u}_k, \mathbf{v}_k)$ that produces probabilistic output, for example, due to uncertain parameters $\mathbf{u}_k$ and/or inputs $\mathbf{v}_k$. The modeled output is potentially complemented by a statistical model error term ε_{M_k}. The result is a prediction of the true system state (without measurement noise) and we call the corresponding hypothetical observation-noise-free data $\mathbf{d}_0$. To make predictions of real (measurement noise-corrupted) noisy data $\hat{\mathbf{d}}_0$, a measurement noise term ε_{d_0} is added to the physics prediction:

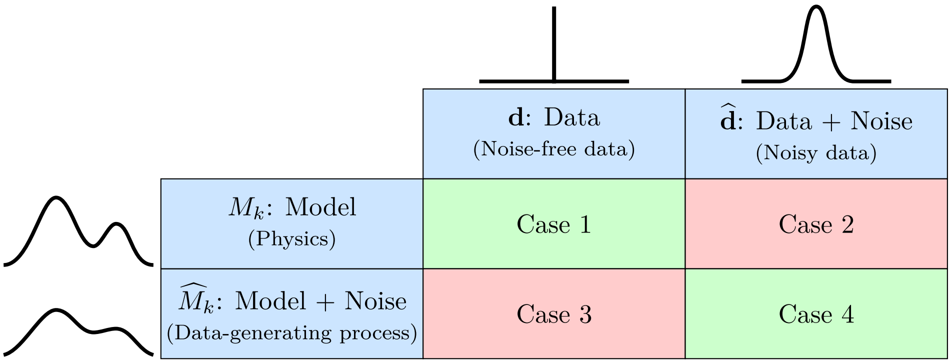
$$\hat{\mathbf{d}}_0 = \overbrace{\underbrace{\overbrace{f_k(\mathbf{u}_k, \mathbf{v}_k) + \varepsilon_{M_k}}^{\text{Physics:}\quad M_k}}_{\mathbf{d}_0} + \varepsilon_{d_0}}^{\text{Data-generating}\quad\text{process:}\quad \hat{M}_k} . \tag{1}$$

Equation 1 shows that we can either predict physics only (denoted as model M_k), or we can predict the data-generating process (physics + noise, denoted as model $\hat{M}_k$). We argue that equipping models with noise ($\hat{M}_k$) and comparing them to noisy data ($\hat{\mathbf{d}}_0$) is of course logically consistent, but answers a research question that is not at the core scientific interest: "which model is best in predicting the data-generating process (i.e., physics plus noise)?" The more intriguing question would be "which model is best in modeling the pure underlying physics?" (in the spirit of Bayesian hypothesis testing), and this question could only be addressed by comparing noise-free model (M_k) predictions with noise-free data ($\mathbf{d}_0$). This is hindered by two practical challenges: First, we do not have noise-free observation data (we call this the "data-availability barrier"); and second, numerical approximation of BME can become an intractable problem with the likelihood function collapsing to a Dirac delta function (we call this the "numerical-approximation barrier"). We comment on these challenges in Sections 3.4 and 3.5.

We put these issues into a consistent frame by enumerating all four combinatorially possible ways of how to treat noise in BMS: First, we can assume data to be noise-free, and predict with noise-free models (Case 1). Second, we may believe to have noisy data but choose not to simulate it with our models (Case 2). Third, we can add noise to our predictions but compare them to (allegedly) noise-free data (Case 3). And fourth, we can consider noise in both models and data (Case 4). This is illustrated by the "decision matrix" in Figure 1.

We shed light on the four cases from a philosophical, mathematical, and numerical perspective, and discuss ways to overcome or deal with these barriers. We focus on the two logically consistent cases (cf. Figure 1) that represent the two potentially conflicting research questions mentioned above and investigate whether we can exploit the inconsistent cases to approximate the pure-physics case.

We show that the theoretical differences between the four cases translate into differences in model weight outcomes. We test whether these differences matter in real-world applications (i.e., whether there is a risk that the model identified as the "most plausible physics model" is different from the identified "most plausible data-generating model"). To do so, we investigate in three analytical test case scenarios under which conditions the differences in model weights are most pronounced. We also derive a mathematical



	$\mathbf{d}$: Data (Noise-free data)	$\widehat{\mathbf{d}}$: Data + Noise (Noisy data)
M_k: Model (Physics)	🟩 Case 1	🟥 Case 2
$\widehat{M}_k$: Model + Noise (Data-generating process)	🟥 Case 3	🟩 Case 4

Figure 1. The four ways to compare a model with data, accounting for or ignoring the noise. Green cells represent consistent scenarios, red cells represent inconsistent scenarios. Distributions drawn in black indicate whether we are dealing with fixed values (Dirac delta function in the case of noise-free data) or random numbers with assigned distribution (noisy data, probabilistic physics predictions, and probabilistic data-generating process predictions).

formulation of model confusion weights for all four cases and use the MCM results as a tool to interpret the severity of differences between the cases. Finally, we demonstrate the implications of the four cases on a real-world hydrogeological case study that features data from a sandbox aquifer lab experiment by Illman et al. (2010).

We here investigate how choosing between our four cases impacts an example of BMS for a set of groundwater models in a brute-force Monte Carlo setup. Yet, our theoretical discussion is general enough to be also applicable (a) to other multi-model approaches such as BMA, pseudo-BMA, or Bayesian stacking (e.g., Höge et al., 2020), to (b) other numerical implementation schemes (e.g., Liu et al., 2016; Volpi et al., 2017) or mathematical approximations of model weights via information criteria (e.g., Schöniger et al., 2014; Ye et al., 2008), and (c) to arbitrary other fields of applications in water resources research and beyond.

The main contributions of this study are the following:

1. We discuss where and for which reasons measurement noise should be considered in BMS (four cases).
2. We distinguish between two potentially conflicting research questions and identify the cases in which they can be pursued in a logically consistent way.
3. We reveal practical hindrances and analyze which approximations can be taken to still approach the desired research question.
4. We finally provide a recipe for BMS that ensures philosophically consistent and numerically stable results under noise.

We summarize the existing mathematical framework of BMS, including the derivation of the MCM, in Section 2. In Section 3, we introduce the four different ways to handle measurement noise and discuss them from a philosophical, mathematical, and practical perspective. Section 4 demonstrates the differences between these cases on simplified 1D analytical examples. Then, we present the application of the four cases to a real-world hydrogeological case study in Section 5. Finally, we summarize findings and provide a recipe for BMS that ensures philosophically consistent and numerically stable results under noise in Section 6.

2. Bayesian Model Selection Framework

2.1. Bayesian Model Selection

BMS is a statistical framework to choose between competing models (Raftery, 1995). It compares the predictive distributions of the different models and ranks the models in their ability to predict the measured data. Besides the goodness-of-fit, BMS takes into account the complexity of the models. It, therefore, yields a model ranking that reflects a compromise between performance and parsimony. A general introduction to the BMS framework can be found in Höge et al. (2019).

When using the BMS framework, we implicitly assume that we are in an M-closed setting (Bernardo & Smith, 2009). That is, we assume the data was produced by one of the N_m considered models (and hence, model selection here means to identify the *true* model, not "the best"). The goal of BMS is to find the posterior probability that model M_k ($k = 1, \ldots, N_m$) produced the data $\mathbf{d}_0$. This probability $P(M_k|\mathbf{d}_0)$ is also called the posterior model weight of model k and is defined as

$$P(M_k|\mathbf{d}_0) = \frac{p(\mathbf{d}_0|M_k)P(M_k)}{\sum_{i=1}^{N_m} p(\mathbf{d}_0|M_i)P(M_i)}, \tag{2}$$

with $p(.)$ denoting probability densities of continuous random variables, and $P(.)$ denoting discrete probability distributions. The prior model weight $P(M_k)$ represents the prior belief in model M_k to be the true one. This prior belief is updated by the predicted density of the data, $p(\mathbf{d}_0|M_k)$, which is called Bayesian model evidence (BME). It is defined as the expected likelihood of the observed data integrated over the model's prior parameter space $\mathcal{U}_k$:

$$p(\mathbf{d}_0|M_k) = \int_{\mathcal{U}_k} p(\mathbf{d}_0|M_k, \mathbf{u}_k)p(\mathbf{u}_k|M_k) \ \ d\mathbf{u}_k \tag{3}$$

Here, the prior distribution of parameters $\mathbf{u}_k$ in model M_k is denoted by $p(\mathbf{u}_k|M_k)$. $p(\mathbf{d}_0|M_k, \mathbf{u}_k)$ represents the likelihood of the data $\mathbf{d}_0$ given the parameters $\mathbf{u}_k$ of model M_k.

In practice, we are missing a reference to compare the obtained Bayesian model weights to, in order to judge their decisiveness. One way to move forward is to use the so-called Bayes factor (Kass & Raftery, 1995) to determine the decisiveness of evidence in favor of one out of two models; however, the Bayes factor does not tell us how decisive this choice could *ideally* be under the given conditions (model set, uncertainty in models, and available data). This is where the model justifiability analysis proposed by Schöniger, Illman, et al. (2015) comes into play: it determines in a synthetic setting, how decisive model weights could be at most (i.e., if any of the models in the set was actually the true one). The core ingredient of the model justifiability analysis is the model confusion matrix, which will be presented in the next Section.

2.2. Model Confusion Matrix

The model confusion matrix (MCM) was proposed by Schöniger, Illman, et al. (2015) as the basis for a so-called model justifiability analysis. The MCM consists of posterior model weights for the models in the set under the condition that the data set was *actually* produced, in turn, by each one of the models. Since we assume model predictions to be probabilistic (due to, e.g., parameter and input uncertainty), there is not a single data set representative of each model, but instead, we have to consider the models' predictive distribution. That means, the given data set $\mathbf{d}_0$ from Equations 2 and 3 now turns into a random number $\tilde{\mathbf{d}}_\ell$ with distribution $p(\tilde{\mathbf{d}}|M_\ell)$, $\ell = 1 \ldots N_m$.

As we now fulfill the requirement of an M-closed setting, we obtain Bayesian probabilities under ideal conditions. Within this synthetic setting, we can exclude noise, model errors, and other annoyances from the analysis, and put the focus on the theoretical information content of the data set. This information is characterized by the type of data and the data set length; or, more specifically: it is given by the sensitivity of model choice to the experimental setup. The goal of the justifiability analysis is to identify an upper limit to the decisiveness of model weights, which can be used to judge the model ranking result obtained for the real data set $\mathbf{d}_0$.

The columns of the MCM represent the data-generating model, whereas the rows list the models to be evaluated, that is, that aim to predict the given data. To populate the k-th row and ℓ-th column of the MCM, we determine the expected weight for model M_k based on the data produced by model M_ℓ and call it $P(M_k|M_\ell)$. This expected posterior model weight is given by

$$\begin{aligned} P(M_k|M_\ell) &= \mathbb{E}_{\tilde{D}}\left[P(M_k|\tilde{\mathbf{d}})\right] = \int_{\tilde{D}} P(M_k|\tilde{\mathbf{d}})p(\tilde{\mathbf{d}}|M_\ell) \ \ d\tilde{\mathbf{d}} \\ &= \int_{D} \frac{p(\tilde{\mathbf{d}}|M_k)P(M_k)}{\sum_{i=1}^{N_m} p(\tilde{\mathbf{d}}|M_i)P(M_i)} p(\tilde{\mathbf{d}}|M_\ell) \ \ d\tilde{\mathbf{d}} \end{aligned} \tag{4}$$

where $p(\tilde{\mathbf{d}}|M_\ell)$ denotes the distribution of synthetic data generated by model M_ℓ. In the following, we assume that prior model weights are uniformly distributed (all $P(M_k)$ are identical), which simplifies Equation 4 to:

$$P(M_k|M_\ell) = \mathbb{E}_{\tilde{D}}\left[\frac{p(\tilde{\mathbf{d}}|M_k)}{\sum_{i=1}^{N_m} p(\tilde{\mathbf{d}}|M_i)}\right] = \int_D \frac{p(\tilde{\mathbf{d}}|M_k)}{\sum_{i=1}^{N_m} p(\tilde{\mathbf{d}}|M_i)} p(\tilde{\mathbf{d}}|M_\ell)\ \mathrm{d}\tilde{\mathbf{d}}. \tag{5}$$

The MCM provides two major insights. First, the main diagonal reveals the maximum *self-identification weights* given the experimental setup. These weights can be used as a reference for BMS results with real data: the diagonal values $P(M_k, M_k)$ tell us how large the model weight for model M_k would be if it perfectly represented nature. Hence, if our BMS analysis resulted in a posterior weight of 60% for model 1 and $P(M_1, M_1) = 0.6$, we would know that we could not get any higher weight (on average) under the given experimental setup. We would conclude that (a) based on the observed data model 1 most probably has produced the data and (b) our measurement data is not well suited to distinguish between the competing models with high confidence. One way to solve the latter issue is to take more or more informative (i.e., more precise or more sensitive) measurements, which will increase the values on the main diagonal of the MCM and decrease the values on the off-diagonals.

Second, the off-diagonal entries reveal the similarity between models. The value $P(M_k, M_\ell)$ indicates how similar model M_k and model M_ℓ are under the available data. Here, large values represent a high predictive similarity of the models. If the value is as large as the diagonal entry of, for example, $P(M_k, M_k)$, it follows that models M_ℓ and M_k cannot be distinguished based on this specific experimental setup. Again, taking more measurements might solve this problem. A value of zero on the off-diagonals represents perfectly distinguishable models with no similarity at all. The similarity of models is relevant, for example, identifying the best surrogate model for a (too) expensive high-fidelity model. To this end, the column with the data of the high-fidelity model is used to identify the most similar surrogate model, that is, the model showing the highest confusion with the high-fidelity model (Schäfer Rodrigues Silva et al., 2020).

3. The Four Ways to Consider/Ignore Noise in Bayesian Model Selection

As laid out in Sections 1 and 2, the Bayesian framework for model evaluation is designed to account for uncertainty in models and data. Here, we focus on measurement uncertainty. Typically, measurement noise is accounted for in the likelihood function. We will explain that this approach is equivalent to adding noise to model predictions and that this is only one out of four possible perspectives on the noise handling theme.

Here and in the following, we distinguish between hypothetical noise-free data $\mathbf{d}_0$ and real noisy data $\hat{\mathbf{d}}_0$. Noise-free data refers to data that represents the true system state. Noisy data, in contrast, refers to observations of the system state that suffer from measurement noise. Of course, “measurement noise” is a conceptual simplification with a frequentist interpretation: to us, repeated measurements of the same system state lead to a set of outcomes that center on the true system state (assuming an unbiased observation procedure) and that shows a certain level of variability (= noise). Similarly, our models M_k target the true system state, and we can add a noise description to obtain models $\hat{M}_k$ that aim to reproduce the expected value and the variability in the data.

We shed light on the four cases defined by Figure 1 from a scientific/philosophical perspective in Section 3.1. Then, we provide a rigorous mathematical derivation of posterior model weights and model confusion weights in the four cases in Sections 3.2 and 3.3, respectively. Section 3.4 discusses challenges in numerical implementation, and Section 3.5 discusses constraints posed by limited data availability. Finally, in Section 3.6, we combine philosophical, mathematical, and pragmatic considerations to briefly summarize implications for model selection in the presence of measurement noise.

3.1. Philosophical Perspective

The different cases are shown in Figure 1 model four different scenarios as schematically illustrated in Figure 2.

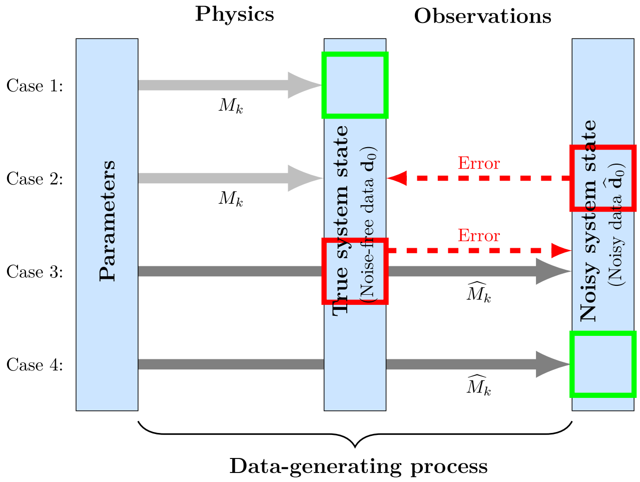


Figure 2. Schematic illustration of what is modeled (system physics only or complete data-generating process) and what is measured (hypothetical noise-free data or real noisy data) in the four cases. Boxes mark the data used for model evaluation. Red arrows represent errors made in the two inconsistent cases.

3.1.1. Case 1

Figure 2 illustrates that, in Case 1, the model simulates the physical processes that cause the (only theoretically observable) true system state; these processes are summarized as "physics." If this true system state was observable, we could compare our predictions with noise-free data to judge the model's quality. Such comparison would be fair and consistent, and very insightful if we wanted to learn about the true processes acting in the (hydrological) system. A typical application of Case 1 is to identify the most plausible physical model, that is, to increase system understanding by identifying dominant processes.

3.1.2. Case 2

In Case 2, the model shall represent the true system state by simulating physics only; but the model output is compared to real noisy data. This case is logically inconsistent because model output and data do not match: the "noise-generating" observation process is missing in the model formulation. We label this case as inconsistent because noise-free predictions are compared to noisy data. Hence, there is no philosophical reason to use this case.

3.1.3. Case 3

Case 3 models what we call the "data-generating process" in Figure 2 and in the following. This includes modeling both the physics and the (simplified) observation procedure that introduces measurement noise. Using noise-free data for evaluating such a model produces a conceptual error, and hence we label this case as inconsistent as well. Again, there is no philosophical reason to use this case.

3.1.4. Case 4

Case 4 models the data-generating process and uses noisy data for model-data comparison. This case is hence logically consistent. Similarly, Nearing et al. (2016) argue that a set of modeling hypotheses can only be tested as a whole, i.e., by modeling the full measurement (error) process. In our words, they argue that only the ”data-generating process” question can be answered. Hence, in Case 4, we cannot assess the plausibility of the physics model alone. A typical application of Case 4 could be, for example, predicting whether a certain legal threshold value will be exceeded - here it is only relevant to predict the combined signal of physics and observation process; the true system state cannot be uncovered and hence cannot be the subject of regulatory guidelines.

3.2. Mathematical Definition of Posterior Model Weights

In the specific context of Bayesian model selection, we wish to formalize the impact of the four introduced cases on the calculation of posterior model weights, assuming noise-free or noisy model predictions and given noise-free or noisy data.

In the noise-free data case, we simply evaluate posterior model weights once on the given data $\mathbf{d}_0$. In contrast, evaluating posterior model weights given real noisy data is more challenging. What we do in practice is to determine model weights once, given the observed noisy data $\hat{\mathbf{d}}_0$. This data contain a specific but unknown realization of noise ε_0 (note that we drop the subscript d when describing ε in the remainder of this work), with $\varepsilon_0 = \hat{\mathbf{d}}_0 - \mathbf{d}_0$. Due to the conceptualization of noise ε_0 as a realization of a random variable, also noisy data $\hat{\mathbf{d}}$ becomes a random variable $\hat{\mathbf{d}} = \mathbf{d}_0 + \varepsilon$. Note that we assign fixed given values a subscript $(\mathbf{d}_0, \hat{\mathbf{d}}_0, \varepsilon_0)$, while random numbers are written without subscript $(\hat{\mathbf{d}}, \varepsilon)$. The random variable $\hat{\mathbf{d}}$ reveals its randomness, e.g., under repetition of an observation or experiment. The distribution of $\hat{\mathbf{d}}$ is given by

$$p(\hat{\mathbf{d}}) = \delta_{\mathbf{d}_0} * p_\varepsilon \tag{6}$$

where $*$ denotes a convolution, $\delta_{\mathbf{d}_0}$ is the probability density function of $\mathbf{d}$ (a Dirac delta function centered at $\mathbf{d}_0$), and p_ε is the measurement noise distribution.

To investigate the general differences between the four cases, we integrate over all possible outcomes of random noisy data $\hat{\mathbf{d}}$. This means that we compare *expected* values of posterior model weights given noisy data in Cases 2 and 4 with *fixed* posterior model weights given noise-free data in Cases 1 and 3. For consistency, we formulate all results as expected values over $p_{\mathcal{E}}$: in Cases 1 and 3, the expected value is over a Dirac delta noise function centered at 0 ($p(\mathbf{d}) = \delta_{\mathbf{d}_0} * p_{\mathcal{E}} = \delta_{\mathbf{d}_0} * \delta_0 = \delta_{\mathbf{d}_0}$), which collapses to the result given $\mathbf{d}_0$; in Cases 2 and 4, the expectation is over the random part of $\hat{\mathbf{d}}$, that is, over ε.

3.2.1. Case 1

Case 1 compares predictions of the true system state by model M_k with noise-free data $\mathbf{d}_0$. Not modeling the noise in predictions and having noise-free data leads to posterior model weights

$$P(M_k|\mathbf{d}_0) = \frac{p(\mathbf{d}_0|M_k)}{\sum_{i=1}^{N_m} p(\mathbf{d}_0|M_i)}, \tag{7}$$

still assuming uniform prior model weights to simplify the equations.

As stated above, in the noise-free data case, the expected posterior model weight over $p(\mathbf{d})$ is equal to the posterior model weight given $\mathbf{d}_0$

$$\mathbb{E}_{\mathcal{E}}\left[P(M_k|\mathbf{d})\right] = P(M_k|\mathbf{d}_0). \tag{8}$$

3.2.2. Case 2

In Case 2, we compare predictions of the true system state by model M_k with given noisy data $\hat{\mathbf{d}}_0$ to obtain the posterior model weight $P(M_k|\hat{\mathbf{d}}_0)$:

$$P(M_k|\hat{\mathbf{d}}_0) = \frac{p(\hat{\mathbf{d}}_0|M_k)}{\sum_{i=1}^{N_m} p(\hat{\mathbf{d}}_0|M_i)}. \tag{9}$$

Since the random outcomes of $\hat{\mathbf{d}}$ are defined by the deterministic (but unknown) true $\mathbf{d}_0$ and random outcomes of noise ε, we can use the convolution formulation (cf. Equation 6) to express the expected posterior model weight as a function of $p_{\mathcal{E}}$:

$$\begin{aligned}\mathbb{E}_{\mathcal{E}}\left[P(M_k|\hat{\mathbf{d}})\right] &= \mathbb{E}_{\hat{D}}\left[P(M_k|\hat{\mathbf{d}})\right] \\ &= \mathbb{E}_{\hat{D}}\left[\frac{p(\hat{\mathbf{d}}|M_k)}{\sum_{i=1}^{N_m} p(\hat{\mathbf{d}}|M_i)}\right] = \int_{-\infty}^{\infty} \frac{p(\hat{\mathbf{d}}|M_k)}{\sum_{i=1}^{N_m} p(\hat{\mathbf{d}}|M_i)} p(\hat{\mathbf{d}})\ \mathrm{d}\hat{\mathbf{d}} \\ &= \int_{-\infty}^{\infty} \frac{p(\mathbf{d}_0+\varepsilon|M_k)}{\sum_{i=1}^{N_m} p(\mathbf{d}_0+\varepsilon|M_i)} p_{\mathcal{E}}(\varepsilon)\ \mathrm{d}\varepsilon = \frac{p(\mathbf{d}_0|M_k)}{\sum_{i=1}^{N_m} p(\mathbf{d}_0|M_i)} * p_{\mathcal{E}}.\end{aligned} \tag{10}$$

3.2.3. Case 3

In Case 3, we predict the noisy system state with models $\hat{M}_k$, but use noise-free data $\mathbf{d}_0$ for model evaluation. This leads to

$$P(\hat{M}_k|\mathbf{d}_0) = \frac{p(\mathbf{d}_0|\hat{M}_k)}{\sum_{i=1}^{N_m} p(\mathbf{d}_0|\hat{M}_i)} = \frac{p(\mathbf{d}_0|M_k) * p_{\mathcal{E}}}{\sum_{i=1}^{N_m} (p(\mathbf{d}_0|M_i) * p_{\mathcal{E}})}, \tag{11}$$

using the fact that the predictive PDF produced by $\hat{M}_k$ can be obtained by widening the noise-free predictive PDF by M_k via convolution with $p_{\mathcal{E}}$.

Like in Case 1, we formally consider the noise distribution defining the outcome of $\mathbf{d}$ as a Dirac delta function at $\varepsilon_0 = 0$, which leads to

$$\mathbb{E}_{\mathcal{E}}\left[P(\hat{M}_k|\mathbf{d})\right] = P(\hat{M}_k|\mathbf{d}_0). \tag{12}$$

One might think that considering noise either in the data (Case 2) or in the model (Case 3) would lead to the same results. This is true for the calculation of the expected BME ($\mathbb{E}_{\mathcal{E}}\left[p(\hat{\mathbf{d}}|M_k)\right] = p(\mathbf{d}|M_k) * p_{\mathcal{E}} = \mathbb{E}_{\mathcal{E}}\left[p(\mathbf{d}|\hat{M}_k)\right]$). However, when targeting expected posterior model weights, the expectation is not over BME. Rather, model weights are obtained by normalization with the sum of all BMEs in the denominator *before* the expectation is taken. This is why expected posterior model weights differ between Cases 2 and 3 ($\mathbb{E}_{\mathcal{E}}\left[P(M_k|\hat{\mathbf{d}})\right] \neq \mathbb{E}_{\mathcal{E}}\left[P(\hat{M}_k|\mathbf{d})\right]$). In Case 2 (Equation 10), the normalization happens *per* noise realization, while in Case 3 (Equation 11) each model's predictive PDF is widened by convolution with $p(\varepsilon)$ *before* the normalization.

3.2.4. Case 4

In Case 4, we simulate the noisy system state by model $\hat{M}_k$ and compare these predictions to given noisy data $\hat{\mathbf{d}}_0$. This leads to posterior model weights:

$$P(\hat{M}_k|\hat{\mathbf{d}}_0) = \frac{p(\hat{\mathbf{d}}_0|\hat{M}_k)}{\sum_{i=1}^{N_m} p(\hat{\mathbf{d}}_0|\hat{M}_i)} = \frac{p(\hat{\mathbf{d}}_0|M_k) * p_\varepsilon}{\sum_{i=1}^{N_m} p(\hat{\mathbf{d}}_0|M_i) * p_\varepsilon}. \tag{13}$$

Averaging this over all possible realization of noise leads to:

$$\mathbb{E}_{\mathcal{E}}\left[P(\hat{M}_k|\hat{\mathbf{d}})\right] = \frac{p(\mathbf{d}_0|M_k) * p_\varepsilon}{\sum_{i=1}^{N_m} \left(p(\mathbf{d}_0|M_i) * p_\varepsilon\right)} * p_\varepsilon. \tag{14}$$

3.3. Mathematical Definition of Model Confusion Weights

Next, we derive the model confusion weights for the four cases. The only difference to posterior model weights as defined in Section 3.2 is that the data $\mathbf{d}_0$ is not given in the form of observations anymore. Instead, we assume that model M_ℓ produced synthetic data $\tilde{\mathbf{d}}$ (noise-free) or $\tilde{\mathbf{d}} + \varepsilon$ (noise-perturbed). Hence, instead of a Dirac delta for $p(\mathbf{d})$, we deal with a distribution of possible synthetic data $p(\tilde{\mathbf{d}}|M_\ell)$ produced by model M_ℓ. When accounting for noise, model $\hat{M}_\ell$ produces noisy synthetic data $p(\tilde{\mathbf{d}}|\hat{M}_\ell) = p(\tilde{\mathbf{d}}|M_\ell) * p_{\mathcal{E}}$.

3.3.1. Case 1

In Case 1, we use models M_k to predict the noise-free system state and compare against noise-free synthetic data by model M_ℓ. Hence, the corresponding expected posterior model weight is given by:

$$P(M_k|M_\ell) = \int_{\tilde{D}} \frac{p(\tilde{\mathbf{d}}|M_k)}{\sum_{i=1}^{N_m} p(\tilde{\mathbf{d}}|M_i)} p(\tilde{\mathbf{d}}|M_\ell)\ \mathrm{d}\tilde{\mathbf{d}}. \tag{15}$$

This equation is identical to the general MCM formulation in Equation 5, with the only difference that we now explicitly specify $\tilde{\mathbf{d}}$ to be noise-free (in our general derivation in Section 2.2, data can be noisy or not).

3.3.2. Case 2

Now, while still keeping M_k noise-free, we sample from a model $\hat{M}_\ell$ that produces noisy data $\tilde{\mathbf{d}} + \varepsilon$ and obtain the posterior model confusion weight:

$$P(M_k|\hat{M}_\ell) = \int_{\tilde{D}} \frac{p(\tilde{\mathbf{d}}|M_k)}{\sum_{i=1}^{N_m} p(\tilde{\mathbf{d}}|M_i)} \cdot p(\tilde{\mathbf{d}}|M_\ell) * p_{\mathcal{E}}\ \mathrm{d}\tilde{\mathbf{d}}. \tag{16}$$

3.3.3. Case 3

When we evaluate model confusion weights for model $\hat{M}_k$ that considers noise, but do not add noise to the synthetic data produced by model M_ℓ, we obtain the following expression:

$$P(\hat{M}_k|M_\ell) = \int_{\tilde{D}} \frac{p(\tilde{\mathbf{d}}|M_k) * p_\varepsilon}{\sum_{i=1}^{N_m} p(\tilde{\mathbf{d}}|M_i) * p_\varepsilon} p(\tilde{\mathbf{d}}|M_\ell)\ \mathrm{d}\tilde{\mathbf{d}}. \tag{17}$$

As for posterior model weights, model confusion weights differ between Cases 2 and 3 due to the normalization in the denominator.

3.3.4. Case 4

In Case 4, we simulate the noisy system state and compare it with noise-perturbed synthetic data. Hence, model confusion weights are given by

10.1029/2021WR030391

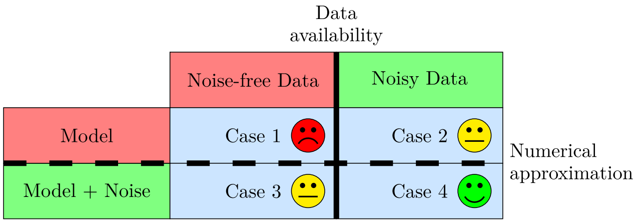


Figure 3. Barriers that hinder transitions between the cases: impervious data-availability barrier separating Cases 2 and 4 from 1 and 3 vertically, and semi-pervious numerical-approximation barrier separating Cases 1 and 2 from 3 and 4 horizontally. Red data column is impossible to populate in real data scenarios. Green model row is numerically straight-forward to obtain with satisfying accuracy; the red model row relies on rough numerical approximations.

$$P(\hat{M}_k|\hat{M}_\ell) = \int_{\tilde{D}} \frac{p(\tilde{\mathbf{d}}|M_k) * p_\varepsilon}{\sum_{i=1}^{N_m}\left(p(\tilde{\mathbf{d}}|M_i) * p_\varepsilon\right)} \cdot p(\tilde{\mathbf{d}}|M_\ell) * p_\varepsilon \;\; \mathrm{d}\tilde{\mathbf{d}}. \tag{18}$$

3.3.5. Symmetry

By comparing Equations 15–18, we see that some equations are symmetric. In Case 1,$P(M_k|M_\ell) = P(M_\ell|M_k)$, which results in a symmetric MCM. The same can be observed in Case 4 where $P(\hat{M}_k|\hat{M}_\ell) = P(\hat{M}_\ell|\hat{M}_k)$. In contrast, Case 2 and Case 3 lead to asymmetric MCMs because $P(M_k|\hat{M}_\ell) \neq P(M_\ell|\hat{M}_k)$ and $P(\hat{M}_k|M_\ell) \neq P(\hat{M}_\ell|M_k)$, respectively. Hence, the consistent Cases 1 and 4 produce symmetric MCMs, whereas the inconsistent Cases 2 and 3 lead to asymmetric MCMs.

3.4. Numerical Implementation

3.4.1. Bayesian Model Evidence

The numerical challenge lies in approximating the BME value per model, as a basis for posterior model weights and model confusion weights. If a model's predictive PDF is known analytically, we can determine BME easily as the PDF value at the given data value(s). For most real-world applications, however, BME cannot be evaluated analytically and the integral in Equation 3 needs to be approximated numerically or mathematically instead.

In our study, we will use brute-force Monte Carlo sampling (Gelman et al., 1995, chapter 10) of the prior distributions. As argued by Schöniger et al. (2014), Monte Carlo is superior to other numerical schemes in that it is an unbiased scheme that is known to converge to the correct limit, and its convergence can be easily monitored. In our chosen test cases, the computational burden of Monte Carlo is bearable; for computationally heavier practical applications, alternative numerical methods could be used to improve on computational efficiency, such as nested sampling (Elsheikh et al., 2014; Skilling, 2006), thermodynamic integration (Lartillot & Philippe, 2006; Liu et al., 2016), stepping stone sampling (Elshall & Ye, 2019; Xie et al., 2011), or Gaussian mixture importance sampling (Volpi et al., 2017), to name a few examples. However, these methods are less straightforward to implement and bear the risk of introducing biases into the BME estimation.

The Monte Carlo approximation of BME (exemplarily shown here for Case 3) is based on N_{MC} random samples (realizations with subscript r) of the model's prior parameter space $\mathcal{U}_k$:

$$p(\mathbf{d}_0|\hat{M}_k) \approx \sum_{r=1}^{N_{MC}} p(\mathbf{d}_0|\hat{M}_k, \mathbf{u}_{k_r}) = \sum_{r=1}^{N_{MC}} \left(p(\mathbf{d}_0|M_k, \mathbf{u}_{k_r}) * p_{\mathcal{E}}\right). \tag{19}$$

Assuming uncorrelated, unbiased measurement errors (i.e., $p_{\mathcal{E}}$ being a normal distribution with mean zero and standard deviation σ), the likelihood function $p(\mathbf{d}_0|\hat{M}_k, \mathbf{u}_{k_r})$ can be expressed as:

$$p(\mathbf{d}_0|\hat{M}_k, \mathbf{u}_{k_r}) = \frac{1}{\sqrt{(2\pi)^{N_o}|\mathbf{R}|}} \exp\left(-\frac{1}{2}(\mathbf{d}_0 - \mathbf{y}_{k_r})^T \mathbf{R}^{-1}(\mathbf{d}_0 - \mathbf{y}_{k_r})\right), \tag{20}$$

where $\mathbf{y}_{k_r}$ is the prediction of model M_k with parameters $\mathbf{u}_{k_r}$. The error covariance matrix $\mathbf{R}$ is of size N_o (number of observation points) × N_o, and features main diagonal entries of σ^2. Other choices of $p_{\mathcal{E}}$ are possible and would affect the results of individual BME values; investigating whether and how other likelihood functions aggravate the differences between the four cases is beyond the scope of this study.

Approximation of BME with Monte Carlo integration is hence straightforward if we understand the noise description as part of the model (Cases 3 and 4). In contrast, Cases 1 and 2 do not consider measurement noise in predictions. This could be interpreted as a standard deviation σ of zero in Equation 20. Mathematically, this leads to a Dirac delta likelihood function centered about the model prediction y_{k_r}, and practically all parameter realizations would receive a likelihood of zero. Hence, it is practically impossible to approximate BME with a likelihood-based numerical scheme in Cases 1 and 2. We call this the "numerical approximation barrier" and visualize it in Figure 3. Figure 3 illustrates the practical feasibility of the four cases. We distinguish between challenges regarding numerical approximation (discussed in this section) and regarding data availability (discussed in Section 3.5). The numerical approximation barrier horizontally separates Cases 1 and 2 from Cases 3 and 4. This barrier can be understood as semi-pervious - we can push the barrier toward Cases 1 and 2 by applying a numerical trick: we increase the theoretical standard deviation from $\sigma = 0$, representing zero noise, to a very small value > 0. This allows us to obtain a (biased) value of BME by using Monte Carlo integration according to Equation 19. An adaptive algorithm that automatically determines an optimal σ for this approximation would be handy and is left for future studies.

3.4.2. Posterior Model Weights

Once BME has been determined for all models in the set, the calculation of posterior model weights according to Equation 2 is straightforward. In Cases 2 and 4, we are interested in *expected* model weights over possible outcomes of $\hat{\mathbf{d}}$, that is, over the noise distribution $p_{\mathcal{E}}(\varepsilon)$. Assuming that we hypothetically know the true data value $\mathbf{d}_0$, we can approximate the expected value using brute-force Monte Carlo integration over random realizations of noise in three steps:

1. Sample from the measurement noise distribution $p_{\mathcal{E}}$. We draw N_d samples (realizations) and denote the jth sample as ε_j.
2. Determine the posterior model weight for each of these data sets: $P(M_k|\mathbf{d}_0 + \varepsilon_j) = \frac{p(\mathbf{d}_0+\varepsilon_j|M_k)}{\sum_{i=1}^{N_m} p(\mathbf{d}_0+\varepsilon_j|M_i)}$.
3. Report the average value of the posterior model weights over all data samples ε_j: $\mathbb{E}_{\mathcal{E}}\left[P(M_k|\hat{\mathbf{d}})\right] \approx \frac{1}{N_d}\sum_{j=1}^{N_d} P(M_k|\mathbf{d}_0 + \varepsilon_j)$.

This example shows the approximation for Case 2 where we use a noise-free model M_k. To calculate posterior model weights in Case 4, simply replace M_k by $\hat{M}_k$ in all three steps.

In real-world applications, we are typically given a single realization $\hat{\mathbf{d}}_0$, and $\mathbf{d}_0$ is unknown. Hence, we cannot average over repeated outcomes of ε, and rely on the model weight given $\hat{\mathbf{d}}_0$ instead.

3.4.3. Model Confusion Weights

For the calculation of MCM entries, we follow the approach proposed by Schöniger, Illman, et al. (2015) and approximate Equation 15 in the same three steps:

1. Sample from each predictive distribution $p(\tilde{\mathbf{d}}|M_\ell)$, $l = 1 \ldots N_m$. We draw N_d samples (realizations) per model and denote the jth sample as $\tilde{\mathbf{d}}_{l,j}$.
2. Determine the posterior model weight for each of these data sets: $P(M_k|\tilde{\mathbf{d}}_{l,j}) = \frac{p(\tilde{\mathbf{d}}_{l,j}|M_k)}{\sum_{i=1}^{N_m} p(\tilde{\mathbf{d}}_{l,j}|M_i)}$.
3. Report the average value of the posterior model weights over all data samples $\tilde{\mathbf{d}}_{l,j}$: $P(M_k|M_\ell) = \mathbb{E}_{\tilde{\mathbf{d}}_j}\left[P(M_k|\tilde{\mathbf{d}}_{l_j})\right] \approx \frac{1}{N_d}\sum_{j=1}^{N_d} P(M_k|\tilde{\mathbf{d}}_{l,j})$.

This example shows the approximation of Case 1. For Case 3, simply replace M_k by $\hat{M}_k$; for Cases 2 and 4, replace M_ℓ by $\hat{M}_\ell$, that is, additionally sample the noise distribution as described for posterior model weights above.

3.5. Constraints Due to Data Availability

For model selection given a real experimental or field data set, all we have is noisy data, that is, we are bound to the right column of the decision matrix in Figure 1: Cases 2 and 4. We call this constraint the "data-availability barrier" and illustrate its location in Figure 3. This barrier can be understood as impervious—there is no way to remove the noise from the observed data in order to transition to the left column of the decision matrix because we do not know the actual outcome of random measurement noise that is added to the true system state value (and hence, we are unable to identify the exact true system state).

3.6. Brief Synthesis of Philosophical, Mathematical, and Pragmatic Aspects

We claim that modelers always want to be in Case 1 or Case 4 because these cases are consistent and do not introduce conceptual errors. We carefully choose the word "want" instead of "should" here, because we might not always be able to choose these cases.

Table 1
Parameter Choices for PDFs in Scenario 1 (Predictive PDFs of Varied Location and Width)

Model	Distribution	μ	σ
Blue	Normal	0	0.2
Red	Normal	3	1
Yellow	Normal	– 2	2
Noise	Normal	0	0.3

Philosophically, the choice is clear: either one is interested in pure physics, then one chooses Case 1, or one is interested in simulating the data-generating process, then one move to Case 4. If one is interested in identifying the model that has most likely produced the noisy data (that mimics the data-generating process best), one simply has to pay attention to equip the model predictions with noise when calculating model confusion weights. Indeed, this is also computationally the most straightforward way to go.

Case 1 is only meaningful if noise-free data are available, for example, when aiming to identify the most suitable surrogate model for a given reference model. As soon as real noisy data are involved as the basis for determining model weights, we are faced with the data-availability barrier that forces us to stay with Cases 2 and 4.

In hydro(geo)logical applications, we are additionally faced with the numerical-approximation barrier, because model output PDFs are typically not given analytically. This leaves us with a straightforward implementation of Cases 3 and 4; and the barrier can be pushed to approximate Cases 1 and 2 by introducing a "small enough" noise level into the likelihood function.

From the mathematical formulation of the four cases, we notice that the MCMs of Case 1 and Case 4 are symmetric. It is interesting to note that both consistent cases are symmetric, while the inconsistent cases are not. Consequently, we always want to obtain a symmetric MCM; any asymmetry of the MCM indicates either conceptual errors due to the choice of an inconsistent case, or numerical approximation errors due to the choice of implementation scheme. Hence, the degree of symmetry in the calculated MCM can be understood as an indicator of "trustworthyness."

We summarize that the only safe choice seems to be Case 4 because it is logically consistent, it handles real noisy data, and it is numerically straightforward to implement. Scientifically, we would be more interested in obtaining results for Case 1. This raises two questions: (a) Is the difference in results for the four cases actually significant in practical applications? And, if yes, (b) can we find a reasonably good approximation via any of the other cases to the desired Case 1? We will investigate these questions in three analytical scenarios (Section 4) and a real-world hydrogeological case study (Section 5).

4. Didactic Illustration With Analytical Scenarios

To illustrate how the four ways to consider or ignore measurement noise affect the outcome of BMS, we first design three didactic examples. We choose these scenarios such that (a) the model output distribution is one-dimensional and therefore easy to visualize, (b) BME values can be obtained analytically and therefore all four cases can be determined accurately, and (c) typical challenges encountered in real-world studies are represented such as models of varying predictive uncertainty, bounded support of the output PDF, and the search for an appropriate surrogate model.

The setup and motivation of the three scenarios is explained in Section 4.1, and results (posterior model weights and model confusion weights) are presented in Section 4.2. In Cases 2 and 4 (noisy data), posterior model weights are reported as averaged values: the expected value over $N_d = 1,000$ possible realizations of noise ε is numerically approximated by Monte Carlo sampling as described in Section 3.4.

The source code of the analytical scenarios is available at https://bitbucket.org/Reuschen/measurement-noise-demo.

4.1. Setup of Scenarios

4.1.1. Scenario 1: Predictive PDFs of Varied Location and Width

In the first analytical example, we consider three models as three Gaussian predictive PDFs of varied location and width as defined in Table 1. They are visualized in Figure 4a. These PDFs shall represent model output simulating physics only (Cases 1 and 2). A real-world interpretation could be models of varying

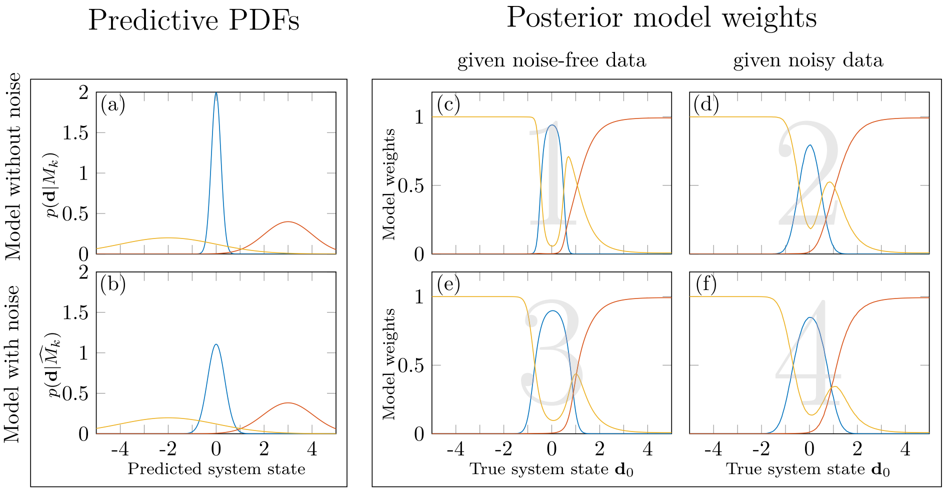


Figure 4. Predictive PDFs and posterior model weights of analytical scenario 1 (three predictive PDFs with varied location and width). Left column: Predictive PDFs of the three models (a) without noise, (b) with noise. Center and right column: Posterior model weights as a function of the true system state value in (c) Case 1—$P(M_k|\mathrm{d}_0)$, (d) Case 2—$\mathbb{E}_\mathcal{E}\left[P(M_k|\hat{\mathrm{d}})\right]$, (e) Case 3—$P(\hat{M}_k|\mathrm{d}_0)$, and (f) Case 4—$\mathbb{E}_\mathcal{E}\left[P(\hat{M}_k|\hat{\mathrm{d}})\right]$.

complexity and predictive skill that differ in their maximum a posteriori prediction of the true system and in their predictive uncertainty. Panel (b) shows the same three PDFs after convolution with the Gaussian measurement noise PDF and hence representing model predictions in Cases 3 and 4. The chosen error standard deviation σ is also listed in Table 1.

4.1.2. Scenario 2: Predictive PDFs With Bounded or Semi-Infinite Support

The second analytical scenario features a normal, exponential, and uniform PDF to represent model output (Figure 5a). PDF parameters are given in Table 2. The focus is here on the effect of stark contrasts in the support of the competing models/PDFs: the uniform PDF has bounded support and hence assigns values outside of the supported range a probability of zero, and the exponential distribution has semi-infinite support of only non-negative values. These hard constraints should be honored by the model selection result. This didactic scenario is hence designed to illustrate the effect of choosing one of the four ways to account

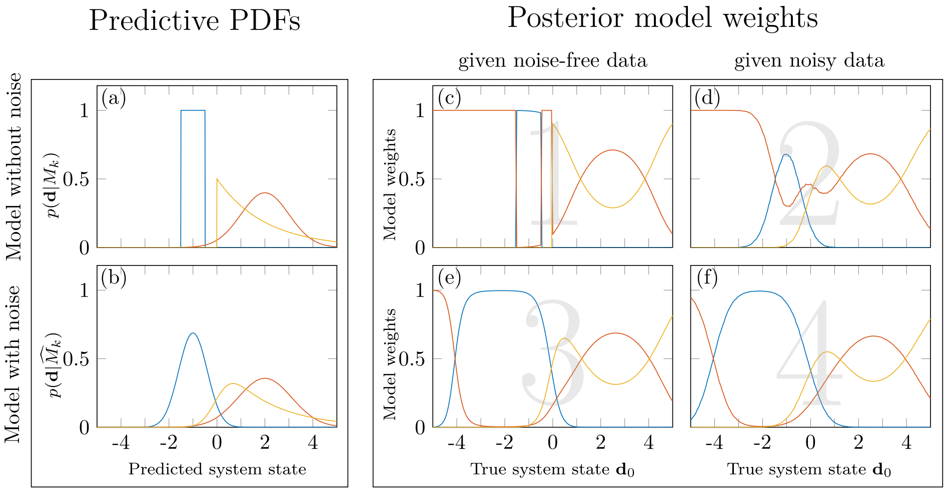


Figure 5. Predictive PDFs and posterior model weights of analytical scenario 2 (predictive PDFs with bounded or semi-infinite support). Left column: Predictive PDFs of the three models (a) without noise, (b) with noise. Center and right column: Posterior model weights as a function of the true system state value in (c) Case 1—$P(M_k|\mathrm{d}_0)$, (d) Case 2—$\mathbb{E}_\mathcal{E}\left[P(M_k|\hat{\mathrm{d}})\right]$, (e) Case 3—$P(\hat{M}_k|\mathrm{d}_0)$, and (f) Case 4—$\mathbb{E}_\mathcal{E}\left[P(\hat{M}_k|\hat{\mathrm{d}})\right]$.

Table 2
Parameter Choices for PDFs in Scenario 2 (Predictive PDFs With Bounded or Semi-Infinite Support)

Model	Distribution	μ	σ	Lower	Upper
Blue	Uniform			– 1.5	– 0.5
Red	Normal	2	1		
Yellow	Exponential	2			
Noise	Normal	0	0.5		

for noise in the presence of bounded or semi-infinite PDF support (data ranges where at least one of the competing PDFs is zero).

This scenario is highly relevant in real-world applications since many variables of natural systems are affected by processes that lead to such hard constraints (e.g., conservation of mass, monotonous increase in entropy, water flowing downhill only, non-negative concentration values, saturation falling between 0 and 1). Real data affected by measurement noise might even lie outside of the value range supported by the model (if only predicting physics), but such data points would *not* automatically reject the model's underlying hypotheses. Hence, in such realistic situations, it is even more important to carefully dissect whether models and data contain noise or not.

4.1.3. Scenario 3: Identification of a Surrogate Model

To complete the set of analytical scenarios, we wish to mimic the realistic case of identifying a simpler model as a suitable surrogate for a complex reference model. That means the model selection task is to choose the surrogate model that approximates the high-fidelity model best. This can be achieved using the MCM as proposed by Schäfer Rodrigues Silva et al. (2020). The PDFs considered in this scenario are defined by the parameters listed in Table 3, and are visualized in Figure 6a. We declare the blue model to be the high-fidelity model which should be approximated by either the red or yellow surrogate model.

4.2. Results and Discussion

For each of the three didactic scenarios, we first present posterior model weights as a function of measured data for the four different cases (Section 4.2.1), and then the full MCM (Section 4.2.2).

4.2.1. Posterior Model Weights

4.2.1.1. Scenario 1: Predictive PDFs of Varied Location and Width

Figure 4a shows the predictive PDFs for the true system state obtained by the three competing models; Figure 4b shows the predictive PDFs modeling the full data-generating process (including noise), which are therefore wider than the pure-physics PDFs. In those subplots, the BME value of model k corresponds to the height of the PDF of model k. For example, the blue model without noise in Figure 4a scores a BME value of 2 for $d_0 = 0$. Note that the data vector $\mathbf{d}_0$ is a scalar value d_0 in all three analytical scenarios, since we are investigating one-dimensional predictive distributions for simpler illustration.

Further, Figures 4c–4f show the posterior model weights obtained in the four different cases of how to consider noise. In these plots, the x-axis corresponds to a noise-free data value d_0 and the y-axis corresponds to the posterior model weight a model achieves given this data value. In Cases 2 and 4, these are averaged values, because we randomly sample the measurement error to be added to the noise-free data and then average over the respective posterior model weights to find the corresponding y-value for a given data value d_0. Hence, the plots are to be read as follows: the height of each model curve represents the average posterior model weight, given that the true system state equals d_0.

Table 3
Parameter Choices for PDFs in Scenario 3 (Identification of a Surrogate Model)

Model	Distribution	μ	σ	Lower	Upper
Blue	Uniform			– 1	0
Red	Uniform			0.2	1.2
Yellow	Normal	2	3		
Noise	Normal	0	0.5		

Apparently, the four cases yield different results. Let us investigate the data point $d_0 = 1$ in Figure 4 in more detail. The posterior model weights in Case 1 of models blue, red, and yellow are 0, 0.45, and 0.55, respectively. In Case 2, they are 0.05, 0.45 and 0.50; in Case 3, 0.16, 0.40 and 0.44, respectively. Finally, in Case 4, they are 0.26, 0.40, and 0.34. This means that, for a noise-free data value of $d_0 = 1$, the posterior weights per model vary between 0 to 0.26, 0.4 to 0.45, and 0.34 to 0.55, respectively, depending on the choice of how to treat noise. Especially the jump from 0 to 0.26 is dramatic in terms of Bayes factors and their linguistic interpretation (Jeffreys, 1961), where evidence is in favor of the best versus the worst

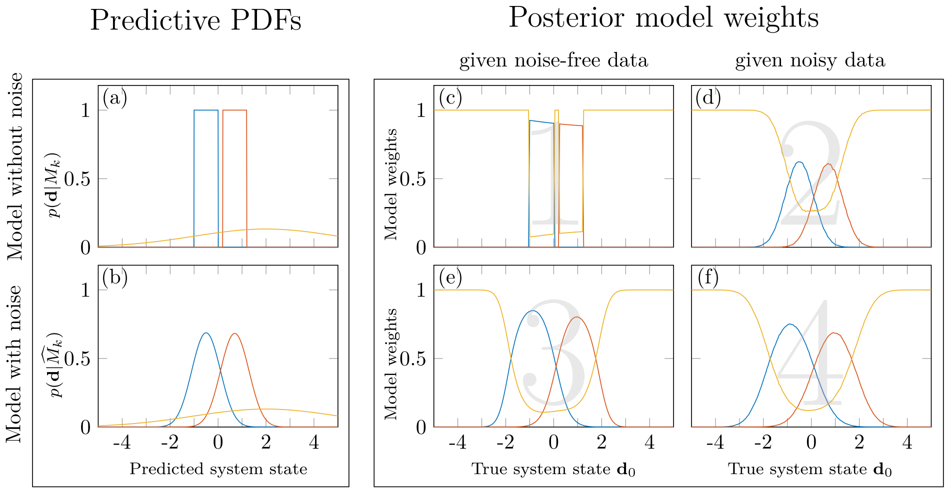


Figure 6. Predictive PDFs and posterior model weights of analytical scenario 3 (identification of a surrogate model). Left column: Predictive PDFs of the three models (a) without noise, (b) with noise. Center and right column: Posterior model weights as a function of the true system state value in (c) Case 1—$P(M_k|\mathrm{d}_0)$, (d) Case 2—$\mathbb{E}_{\mathcal{E}}\left[P(M_k|\hat{\mathrm{d}})\right]$, (e) Case 3—$P(\hat{M}_k|\mathrm{d}_0)$, and (f) Case 4—$\mathbb{E}_{\mathcal{E}}\left[P(\hat{M}_k|\hat{\mathrm{d}})\right]$.

model would jump from "decisive" to "not worth more than a bare mention." Importantly, not only the decisiveness in model selection (or rejection) changes but also the ranking itself: in Case 4, the red model wins, whereas in all other cases, the yellow one scores best. Hence, conclusions about model selection may change dramatically depending on the choice of the case to be in.

Further, we learn from Figure 4 that the differences between the four cases are smallest if one model clearly shows the highest predictive density and hence obtains a much higher BME value than the others (here, e.g., at the PDF modes $\mathrm{d}_0 = -2, 0, 3$). Differences are highest in the "transition zones" where BME values are not significantly different and model ranking is less decisive. Unfortunately, these are often the scientifically most interesting situations.

4.2.1.2. Scenario 2: Predictive PDFs of Bounded or Semi-Infinite Support

Figure 5 visualizes the predictive PDFs of the three competing models in Scenario 2 without noise in panel (a) and containing noise in panel (b). In this scenario, the difference is much more pronounced and important than in scenario 1: here, adding noise to the models eliminates the bounds of model blue (originally only defined on the interval $[-1; 1]$, now a symmetrical PDF with infinite support) and the semi-infinite support of model yellow (now a skewed PDF with infinite support). Hence, if we consider noise in the models, results look similar to scenario 1 (in cases 3 and 4, panels (e) and (f)), with differences between the two cases of using noise-free or noisy data for model evaluation being relatively small. As expected, using noisy data dilutes the decisiveness of model ranking to some degree, which of course depends on the level of measurement error standard deviation; a flip in model preference is not observed in this specific example.

If, instead, we model physics only (Cases 1 and 2, panels (c) and (d)), it matters a lot whether we use noise-free or noisy data for BME evaluation. If we use noise-free data (Case 1), we directly evaluate the predictive PDFs in panel (a) to read off BME values, and hence, the blue model scores a model weight of almost one in its bounded interval because it shows a much higher predictive density than the red model and because the yellow model's PDF is zero in this range. Outside of its interval, the blue model of course obtains a weight of zero. Similarly, the yellow model can only gain a weight larger than zero for non-zero values. Consequently, the red model scores a weight of one over those data value ranges where none of the other PDFs "lives," simply because it is defined there, not because it shows a competitively high predictive density. This scenario has the potential to provide a lot of insight about the system under study: if we compared three competing hypotheses and were (magically) able to consult noise-free data of the true system state, we could clearly reject the individual hypotheses outside of their scope, and hence, only the one hypothesis survives that is general enough to cover the full data range.

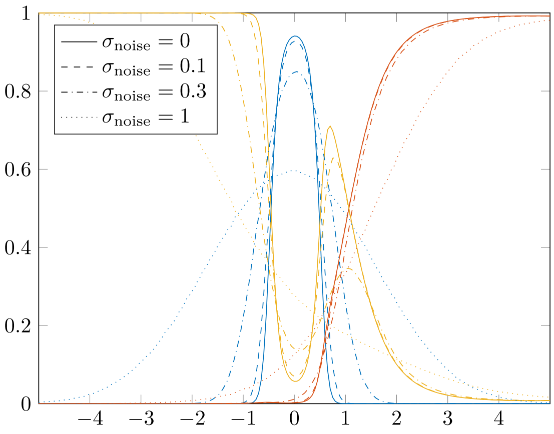


Figure 7. Posterior model weights as a function of measurement error standard deviation σ_{noise} on the example of Case 4 scenario 1. For a standard deviation of zero, Case 4 collapses to Case 1.

If we use noisy data (Case 2), the decisiveness in model choice is practically smoothed out. This is due to the averaging over random samples of noise: we evaluate the predictive PDFs in panel (a) at several locations *around* a specific noise-free data value d_0, and then average it to determine the posterior model weight. Averaging in a neighborhood of d_0 is dangerous in this scenario, since we possibly cross the boundaries of the support of the individual model PDFs. This is why the sharp transitions of model preference as seen in panel (c) (Case 1) are diluted in panel (d) (Case 2). In reality, however, this averaging procedure cannot be done, because we only have one noisy observation $\hat{\mathrm{d}}_0$. Consequently, we rely on determining $P(M_k|\hat{\mathrm{d}}_0)$ instead of $\mathbb{E}_{\mathcal{E}}\left[P(M_k|\hat{\mathrm{d}})\right]$.

4.2.1.3. Scenario 3: Identification of a Surrogate Model

In the surrogate scenario (Figure 6), we assume that the blue model is a high-fidelity model and the red and yellow models are possible surrogates, with the red model being an exact but shifted copy of the blue model (e.g., due to some simplification bias), and the yellow model being a much more uncertain description of the system with a bias in location. Under this setup, noise-free data *is* in fact available (by the high-fidelity model), so this is the only occasion where we can indeed freely choose between Case 1 (using noise-free data) and Case 4 (using noisy data).

The resulting weights are a mix of the results discussed for scenarios 1 and 2: only in Case 1, the bounded support of the blue and red model is honored in the posterior model weights. Cases 2, 3, and 4 show different types of smearing, with Case 2 being closest to Case 1 with respect to how far the high-weight region of each model extends (i.e., steepest slopes at the transition points), and Case 3 being closest to Case 1 with respect to decisiveness in model choice (i.e., highest weights for blue and red, but note that the high-weight region is shifted in comparison to Case 1). Section 4.2.2 presents how the MCM can be used to choose the best surrogate in scenario 3.

4.2.1.4. Influence of Measurement Error Standard Deviation on Differences Between the Four Cases

In general, the differences between Case 1 and Case 4 depend on the level of measurement noise. Here, we want to show how large the influence of measurement noise is. Figure 7 shows the posterior model weights obtained in Case 4 for different standard deviations of noise on the example of scenario 1 (PDFs of varied location and width). With a standard deviation of zero, all four cases are the same and collapse to Case 1 (noise-free model and data). Obviously, the difference between Case 4 and Case 1 increases with increasing measurement noise.

It might seem unlikely that the measurement error standard deviation exceeds the variability of the model prediction; however, in realistic cases, this might happen for strongly calibrated models that are underdispersive in their predictive distribution. To detect whether the present level of noise dominates the model selection result in a specific application, we recommend using the MCM as a diagnostic tool (Section 4.2.2).

4.2.1.5. Approximations of Case 1

Recall that Case 1 is scientifically most interesting but intractable in real-data scenarios due to the data-availability barrier depicted in Figure 3. Since large measurement noise leads to relevant differences between the four cases, we search for a suitable alternative to approximate the results of Case 1 as closely as possible with the right column of the decision matrix (Cases 2 and 4). Case 4 would be a preferable candidate because it is the only consistent case left; however, all three investigated scenarios have shown that results for Cases 1 and 4 significantly differ from each other. Especially scenario 2 teaches us to be very careful when analyzing Case 4 but interpreting it as a proxy for Case 1: there is a danger that we believe to have identified a plausible model although its core hypothesis about the functioning of the system (without considering noise) is plainly rejected (e.g., the blue model at $\mathrm{d}_0 = -2$ in Figure 5).

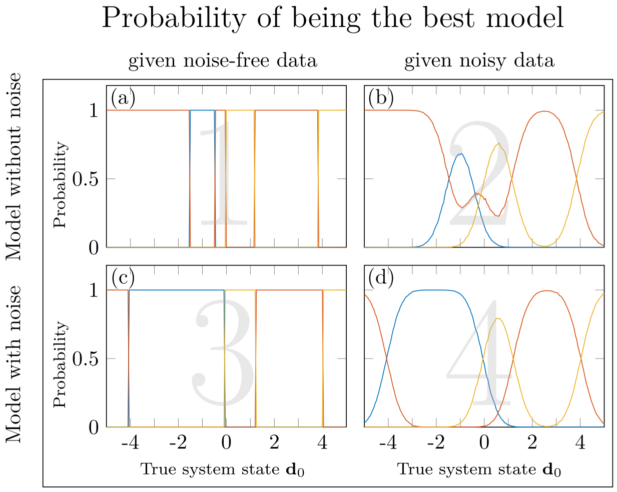


Figure 8. Probability of being the best model of analytical scenario 3 (identification of a surrogate model). Probability of having the highest posterior model weight as a function of the true system state value in (a) Case 1, (b) Case 2, (c) Case 3, and (d) Case 4 (scenario 2).

Rather, results suggest that, if we are interested in Case 1, we should use Case 2 to approximate Case 1. Although being logically inconsistent, posterior model weights produced under Case 2 generally show a higher agreement with those in Case 1 (Figures 4–6). Thinking of the decision matrix and its barriers (Figure 3), we cross one barrier (i.e., going from top-left to top-right, crossing the data-availability barrier) to find a good proxy for Case 1.

4.2.1.6. Sensitivity of Model Ranking Results to Actual Outcome of Noise

We showed that Case 2 can be used as a reasonable proxy for Case 1. When using this proxy, we should be aware that all Case 2 results are averaged results over random realizations of noise ε. In real-world applications we only have one realization of noise ε_0 and hence only one measurement $\hat{\mathbf{d}}_0 = \mathbf{d}_0 + \varepsilon_0$. Naturally the next question arises: How large are the differences in Case 2 between the averaged results ($P(M_k|\hat{\mathbf{d}})$) and the results based on one realization of noise ($P(M_k|\hat{\mathbf{d}}_0)$)?

The largest difference lies in the fact that $P(M_k|\hat{\mathbf{d}}_0)$ might be a overconfident posterior model weight if only one realization of data $\hat{\mathbf{d}}_0$ is used. We illustrate this with a simple example. Assume $\mathrm{d}_0 = -0.25$ in scenario 2 (Figure 5). The BMS analysis given this data value (Case 1) results in a posterior model weight of 1 for the red model and 0 for the other two (Figure 5c). Because the true system state is perturbed by random noise (Case 2), we will end up with one of three possibilities. First, ε_0 could be between − 0.25 and 0.25, yielding $\hat{\mathrm{d}}_0 = \mathrm{d}_0 + \varepsilon_0$ between − 0.5 and 0. For that range, Case 2 will report the same results as Case 1. However, this only happens with a probability of 38% according to the assumed distribution of measurement noise. Second, with 31% probability the measurement noise ε_0 is in the interval $[-1.25, -0.25]$ which leads to a noisy data value $\hat{\mathrm{d}}_0$ in the interval $[-1.5, -0.5]$. For these data values, Case 2 will report a posterior model weight of almost 1 for the blue (not the red) model. Third, the measurement noise ε_0 could be in $[0.25, 1.25]$ (31%), which results in the yellow model being most likely. Hence, depending on the actual outcome of noise ε_0, any of the three models could win the model ranking.

This uncertainty in which model scores highest is visualized in Figure 8. Instead of posterior model weights, we now show the probability of each model to win the model ranking (i.e., being the model with the highest posterior model weight) as a function of true system state d_0. We observe that, for noise-free data (left column), probabilities are binary (0 or 1), because being the best model is completely defined by the constellation of the predictive PDFs in the model set. For noisy data, the randomness in the actual outcome of noise ε leads to probabilities of being the best model smaller than one in the transition areas between high-density regions of the three model PDFs. The uncertainty about which model is best is highest in Case 2 for data values between –1.5 and 1.5, as discussed above.

To summarize, the shown posterior model weights of blue/red/yellow in Case 2 (e.g., 0.3/0.45/0.25 in Figure 5b) are a weighted average of the described three extreme outcomes of model ranking. If only a single data set $\hat{\mathrm{d}}_0$ is available, this average cannot be taken and only one of the three extreme cases will occur. Hence, in practical applications, we need to keep in mind that the obtained ranking might be overly decisive (see also a preliminary summary of findings in Section 4.2.3).

4.2.2. Model Confusion Weights

Building on the obtained posterior model weights, we now use the model confusion matrices (MCMs) of the four cases to further investigate the BMS results. We here exemplarily discuss the results of scenario 3. Figure 9 displays the MCM obtained for scenario 3. Panel (a) shows the MCM in Case 1, panel (b) shows the MCM in Case 2, panel (c) the MCM in Case 3, and panel (d) in Case 4. For the results of scenarios 1 and 2, we refer the reader to Figures A1 and A2 in Appendix A.

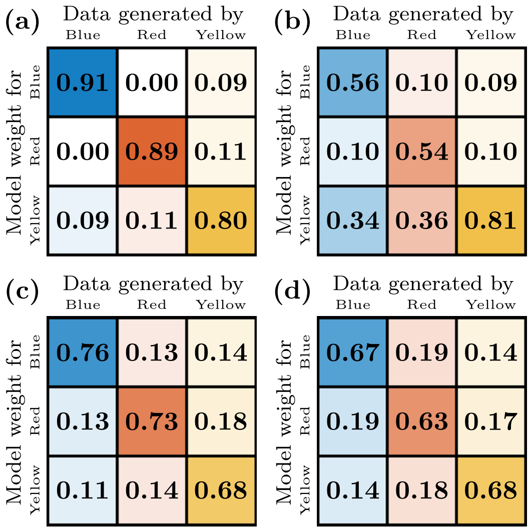


Figure 9. Model confusion matrix in (a) Case 1, (b) Case 2, (c) Case 3 and (d) Case 4 of scenario 3 (identification of surrogate model).

4.2.2.1. Self-Identification Potential

The MCM can show if less noisy measurements can lead to more decisive posterior model weights. To do so, the self-identification weights (diagonal entries) can be used to compare the (expected) self-identification of Case 1 and Case 4.

The self-identification potential of the (blue) physics alone (Case 1, 91%) is much higher than the self-identification potential of the data-generating process (Case 4, 67%). This is expected because measurement noise (when conceptually attached to the model predictions by convolution) smears out differences between the model PDFs and hence makes them more similar (i.e., there is more potential for confusion). Case 1 can therefore be understood as an upper limit to the self-identification potential in Case 4 with the level of noise approaching zero. Practically, this means that we can investigate whether (and to which degree) more precise measurements could help in model selection by comparing the self-identification weights of Cases 1 and 4.

4.2.2.2. Symmetry of the Model Confusion Matrix

As mentioned earlier, the MCM of the consistent cases (Case 1 and Case 4) should result in a symmetric MCM, whereas the inconsistent cases should not (at least not necessarily). Generally, we can trace back any asymmetry in the MCM to either numerical problems or to an inconsistent choice of case. Either way, the symmetry of the MCM gives us an indication of how much we can trust our BMS results. If the MCM is symmetric, we can assume, yet not guarantee (because errors might cancel out), that everything is all right. If the MCM is asymmetric, we know that either numerics or philosophy should be a source of headache. Either way, this should motivate us to change something in our approach and/or implementation.

From Figure 9, we see that symmetry almost holds true for this analytical example in Case 1 and Case 4; we only see very small asymmetry on the level of rounding errors in Case 4. This happens due to the numerical approximation of the expected value over possible realizations of noise ε by Monte Carlo sampling. However, this numerical asymmetry is much smaller than the theoretically expected asymmetry of the inconsistent cases 2 and 3 (Figures 9b and 9c).

In practice, we can use the asymmetry in the MCM of Case 2 as an indicator for the (lack of) quality in approximating posterior model weights in Case 1 with Case 2. The more symmetric Case 2 is, the better it approximates Case 1 and the smaller the approximation error by using noisy data instead of noise-free data.

4.2.2.3. Identification of Most Suitable Surrogate Model

The MCM can be used to determine the best suitable surrogate model (Schäfer Rodrigues Silva et al., 2020). As shown hereafter, the best suitable surrogate model can depend on the choice of case.

Let us evaluate the red and yellow models' potential as a surrogate model for the blue high-fidelity reference model. Considering that the red model has no overlap with the blue model when modeling physics only (Case 1, cf. Figure 6a), one would reject its use as a surrogate altogether. This is what we see in the MCM for Case 1: there is no confusion between the two models (confusion weight of 0%), and hence the yellow model would be preferred as a surrogate with a (still small) confusion weight of 9%.

If we instead aim to model the data-generating process (Case 4) as represented by the blue model, the red model is better suited because if the blue model produced the data, the model weight of red (19%) is slightly larger than the model weight of yellow (14%) (Figure 9d).

4.2.3. Summary and Implications of Findings From Analytical Scenarios

Our investigations on the three analytical scenarios have clearly demonstrated that the theoretical differences laid out in Section 3 yield significantly different results, for both posterior model weights, and for model confusion weights. Hence, we need to be aware of the assumptions underlying each case and make them match the specific modeling context.

The logically inconsistent cases generally lead to problems in interpretation: Using noisy data to test noise-free model predictions (Case 2) leads to smearing over potential bounds or jumps in the predictive PDFs, and hence can mask severe differences between the hypotheses about the modeled system. Predicting the data-generating process but testing the competing models with noise-free data (Case 3) can be seen as an incomplete assessment of Case 4, because only the expected value of noise (=zero) is tested, which neglects the fact that posterior model weights are a nonlinear function of noise.

The logically consistent Case 4 suffers from a lower degree of identifiability per model because model differences are smoothened by noise. Yet, for specific applications, this will be the right context to choose (e.g., predicting whether or not a certain threshold value will be exceeded in a measurement).

Case 1 is scientifically most interesting because it aims to compare the predictive model PDFs of the true system state, that is, it focuses on the way how the models explain the system under study. Unfortunately, it suffers from the data-availability barrier and can only be evaluated in the synthetic setup of the MCM, but not for model ranking given real observed data. We found that Case 2 is a better proxy for posterior model weights of Case 1 than Case 4, even though being logically inconsistent. The approximation error can then be estimated from the degree of asymmetry of the corresponding MCM.

One issue remains: model ranking may turn out overly decisive in favor of one model if only one measurement (or data set) is used, that is, if one cannot average over many repetitions of data noise as in most practical applications. Trusting this overly decisive model ranking is dangerous because it does not necessarily represent the ranking given the true system state.

5. Application to Real-World Hydrogeological Case Study

To demonstrate the impact of choosing between the four ways to account for measurement noise when comparing models to data under real-world conditions, we have implemented the model ranking and model confusion analysis for a hydrogeological case study. All scenarios discussed so far have featured illustrative, one-dimensional predictive PDFs. One-dimensional problems are beautiful, easy to visualize, and allow for analytical solutions. In real life, we typically have high-dimensional predictive distributions, which are expected to amplify the differences between the four cases that arose already in 1D. Further, we are not only facing the data-availability barrier, but also the numerical-approximation barrier, since analytical solutions typically do not hold in practical applications. Our hydrogeological case features 2D fully-saturated groundwater flow in a confined sandbox aquifer, with heterogeneity in hydraulic parameters being represented by four competing models. We first describe the case study setup in Section 5.1, then present and discuss results in Section 5.2.

5.1. Case Study Setup

5.1.1. Experimental Data

For our analysis, we will use the experimental data from Illman et al. (2010). They performed steady-state cross-hole pumping tests in a lab-scale sandbox aquifer. These drawdown data sets have been used to infer the spatial distribution of hydraulic conductivity in the sandbox via Bayesian updating in Illman et al. (2010) and Schöniger, Illman, et al. (2015).

The experimental setup is described in Illman et al. (2010). In the following, we will give a short summary. The sandbox is 1.93 m long, 0.826 m high and has a width of 0.0102 m. It is filled with natural sand layering as shown in Figure 10a.

48 horizontal hydraulic ports (”wells”) were installed to monitor the pressure loss or to perform pumping. The sandbox can be modeled as quasi-2D because the ports penetrate the entire width. In this study, we follow Schöniger, Illman, et al. (2015) and use the 36 monitoring ports that provided the largest signal-to-noise ratio during the experiment. We consider pumping in the six wells marked in blue in Figure 10a. For the model

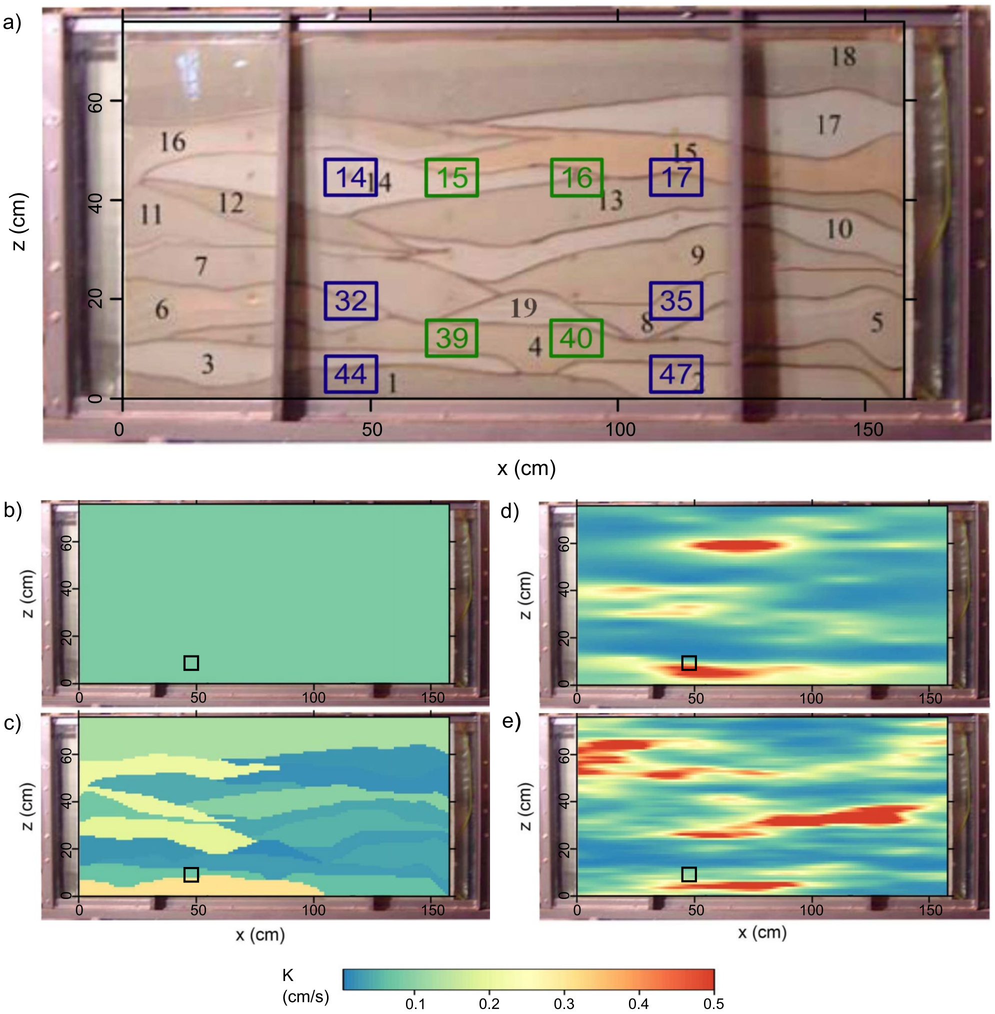


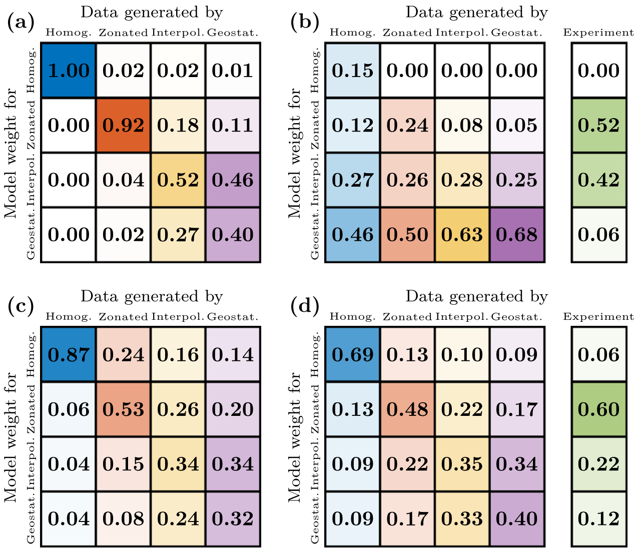
Figure 10. Illustration of experimental setup and model set for the hydrogeological case study. (a) 2D view of the synthetic sandbox aquifer (Schöniger, Illman, et al., 2015), with black numbers indicating different soil layers and blue squares indicating ports used for hydraulic tomography. (b–e) Model realizations conditioned to pumping test in port 44 (lower left): (b) homogeneous model, (c) zonated model, (d) interpolated model, and (e) geostatistical model.

selection task, we consider two data-availability scenarios: we first assume just a single pumping test has been performed (and to obtain representative results, we average resulting model weights of the six data sets corresponding to possible six well locations), and second we combine all six pumping tests into one data set.

5.1.2. Model Set

We here use the drawdown predictions of the four alternative groundwater models developed by Schöniger, Illman, et al. (2015). These models vary in their spatial representation of heterogeneity in hydraulic conductivity K. All four models use $Y = \ln(K(x))$ as a random field model, but differ in their assumption on the spatial (correlation) structure of $K(x)$: (a) homogeneous (effective) value, (b) zonation with a homogeneous value within each zone, using independent random variables for each zone, and the geometry of zones known from visual inspection of Figure 10a, (c) geostatistical interpolation between pilot points (e.g., RamaRao et al., 1995), and (d) fully geostatistical parameterization.

To get a visual impression of the models and their differences in the heterogeneity of K, conditional realizations of each model are shown in Figures 10b–10e. These realizations show the best fit to drawdown induced by a single pumping test in port 44, marked as a black rectangle in each plot. For a detailed description of the models, we refer to Schöniger, Illman, et al. (2015).



(a)

	Homog.	Zonated	Interpol.	Geostat.
Homog.	1.00	0.02	0.02	0.01
Zonated	0.00	0.92	0.18	0.11
Interpol.	0.00	0.04	0.52	0.46
Geostat.	0.00	0.02	0.27	0.40

(b)

	Homog.	Zonated	Interpol.	Geostat.	Experiment
Homog.	0.15	0.00	0.00	0.00	0.00
Zonated	0.12	0.24	0.08	0.05	0.52
Interpol.	0.27	0.26	0.28	0.25	0.42
Geostat.	0.46	0.50	0.63	0.68	0.06

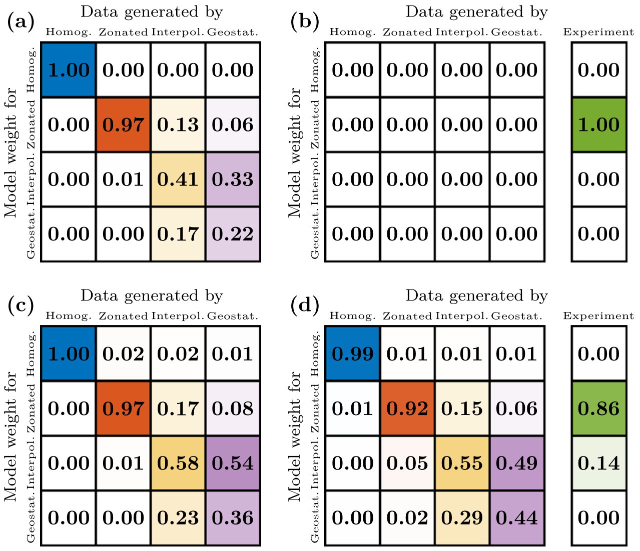
(c)

	Homog.	Zonated	Interpol.	Geostat.
Homog.	0.87	0.24	0.16	0.14
Zonated	0.06	0.53	0.26	0.20
Interpol.	0.04	0.15	0.34	0.34
Geostat.	0.04	0.08	0.24	0.32

(d)

	Homog.	Zonated	Interpol.	Geostat.	Experiment
Homog.	0.69	0.13	0.10	0.09	0.06
Zonated	0.13	0.48	0.22	0.17	0.60
Interpol.	0.09	0.22	0.35	0.34	0.22
Geostat.	0.09	0.17	0.33	0.40	0.12

Figure 11. MCM resulting from inversion with drawdown data from single pumping test, averaged over six pumping locations, in (a) Case 1, (b) Case 2, (c) Case 3, and (d) Case 4. Further, the posterior model weights in Case 2 ($P(M_k|\hat{\mathbf{d}}_0)$, top right) and Case 4 ($P(\hat{M}_k|\hat{\mathbf{d}}_0)$, bottom right) are shown.



(a)

	Homog.	Zonated	Interpol.	Geostat.
Homog.	1.00	0.00	0.00	0.00
Zonated	0.00	0.97	0.13	0.06
Interpol.	0.00	0.01	0.41	0.33
Geostat.	0.00	0.00	0.17	0.22

(b)

	Homog.	Zonated	Interpol.	Geostat.	Experiment
Homog.	0.00	0.00	0.00	0.00	0.00
Zonated	0.00	0.00	0.00	0.00	1.00
Interpol.	0.00	0.00	0.00	0.00	0.00
Geostat.	0.00	0.00	0.00	0.00	0.00

(c)

	Homog.	Zonated	Interpol.	Geostat.
Homog.	1.00	0.02	0.02	0.01
Zonated	0.00	0.97	0.17	0.08
Interpol.	0.00	0.01	0.58	0.54
Geostat.	0.00	0.00	0.23	0.36

(d)

	Homog.	Zonated	Interpol.	Geostat.	Experiment
Homog.	0.99	0.01	0.01	0.01	0.00
Zonated	0.01	0.92	0.15	0.06	0.86
Interpol.	0.00	0.05	0.55	0.49	0.14
Geostat.	0.00	0.02	0.29	0.44	0.00

Figure 12. MCM resulting from inversion with drawdown data from six pumping tests in (a) Case 1, (b) Case 2, (c) Case 3, and (d) Case 4. Further, the posterior model weights in Case 2 ($P(M_k|\hat{\mathbf{d}}_0)$, top right) and Case 4 ($P(\hat{M}_k|\hat{\mathbf{d}}_0)$, bottom right) are shown.

5.1.3. Numerical Implementation of Model Ranking and Confusion Analysis

In this work, we investigate two different scenarios of data availability as mentioned above. In the first scenario, we only consider drawdown data induced by a single pumping test. This results in a simple numerical problem because only a few conditioning data are available, with the high-likelihood region being rather spread out. In the second scenario, we combine data from all six pumping tests, which leads to a much stronger conditioning effect and a very peaked likelihood surface to be sampled. This makes the approximation of BME and hence the computation of posterior model weights and model confusion weights numerically much harder.

We use $N_d = 1,000$ Monte Carlo realizations to sample the predictive distribution of the data-producing model. BME per model is approximated using $N_{MC} = 200,000$ Monte Carlo realizations of the homogeneous model and $N_{MC} = 1,000,000$ realizations for of each of the other models. As discussed in Section 3.4, the BME approximation is harder than the averaging over random model predictions, and hence needs more samples for convergence.

Further, as discussed in Section 3.4, we approximate the “zero noise in models” of Case 1 and Case 2 with a “small noise in models” of $\sigma = 0.001$ [m]. This is 4.6 times smaller than the actual level of measurement noise ($\sigma = 0.0046$ [m]).

5.2. Results and Discussion

In the previous analytical scenarios, we investigated the question of what the posterior model weight would be if the noise-free data d_0 be equal to some value. We can visualize the answer in one dimension (one measurement). Here, we have 35 drawdown measurements (excluding the one at the pumping port) or 210 measurements (when combining data from all six pumping tests). As a result, we cannot visualize all possible $\mathbf{d}_0$. We, therefore, focus on the posterior model weights for the noisy observed data $\hat{\mathbf{d}}_0$ (measured by Illman et al., 2010) and the model confusion weights in Figures 11 and 12.

5.2.1. Results Based on Individual Pumping Tests

The goal of BMS is to select the best model out of an ensemble of possible models. Here, we demonstrate how posterior model weights should be interpreted. To do so, let us first look at the (computationally simple) scenario with drawdown data from individual pumping tests in Figure 11.

5.2.1.1. Which Model is Best (in Case 4)?

Given the data-availability barrier for real data, we can obviously only report posterior model weights given experimental data for Cases 2 and 4 in Figure 11. When inspecting those weights for Case 4, we find a 60% probability of the zonated model being true. This is more than the self-identification potential of the zonated model (48%) in the MCM. It might seem surprising to score a higher posterior model weight than the reference value in the MCM. However, there are two causes that can lead to such results: First, it can happen by chance. The MCM shows the *expected* self-identification weight (48%) and we can score a higher one (or a lower one) just by coincidence. Second, we might have overestimated the level of measurement noise. Here, we chose the measurement noise standard deviation equal to the maximum observed measurement noise of 0.46 cm, despite the fact that lower noise levels were observed at many other ports (Schöniger, Illman, et al., 2015). We neglected the spatial distribution of noise in this study to reflect common simplified assumptions about measurement noise, but wish to remind the reader that the results of the BMS analysis are of course conditional to the choice of error description (and hence the choice of likelihood function).

5.2.1.2. Can Posterior Model Weights be Increased With Less Noisy Data?

We could increase our confidence in model selection by reducing the noise in measurements by more precise (expensive) experimentation. Figures 11a indicates that with measurement noise close to zero, a self-identification of 92% is possible.

5.2.1.3. Can We Trust the Results?

From the theoretical foundations laid out in Section 3.3, we expect that Cases 1 and 4 yield symmetric MCMs. Here, we find this confirmed for Case 4 but not for Case 1. This means that this MCM suffers from a numerical approximation error in either (a) sampling the synthetic data, in (b) evaluating BME, or in (c) approximating Case 1 with a small-but-not-zero noise level (cf. Section 3.4), and results should be interpreted with caution. These errors could be reduced by increasing the Monte Carlo ensemble size of both the BME approximation and random noise sampling; however, it is our intent to demonstrate that the level of asymmetry caused by potential numerical errors tends to be much smaller than the compete asymmetry seen in Cases 2 and 3 due to their logical inconsistency (potentially further aggravated by numerical errors).

5.2.1.4. Can We Use Case 2 as a Proxy for Case 1?

We could ask ourselves whether the posterior model weights for Case 2 could be used as an approximation to the ranking in Case 1 to identify true system behavior instead of the data-generating process. For our specific case study, we conclude that we should not interpret the weights from Case 2 in that direction, because the large asymmetry of the MCM of Case 2 reveals a large gap between the two cases.

5.2.1.5. How Can We Approximate the MCM of Case 1?

To determine the upper limit for self-identification we had a look at the MCM in Case 1. However, we may not be able to calculate the MCM in Case 1. As a result, we want to investigate which of the other cases is best in approximating the MCM of Case 1. Recall that we did not have to search for a proxy of the MCM in Case 1 in our analytical scenarios (Section 4) because BME could be calculated analytically.

Now, with the real-world example, we have to address the numerical approximation barrier. While Case 2 was found to be a reasonable proxy for posterior model weights, this does not hold for MCMs due to the role of variance in the synthetic data as explained below. Instead, we find that Case 3 yields, qualitatively, the most similar results to the MCM of Case 1. In fact, we approach Case 1 numerically by starting from Case 3 (with the noise level chosen a priori from the knowledge about the measurement process) and reducing it as much as numerically possible (here: standard deviation reduced by factor 4.6 to obtain the approximated result of Case 1).

5.2.1.6. Can We Explain the MCM Structure of the Inconsistent Cases?

From Figure 11, we observe that the MCM of Case 2 has higher values in the bottom left whereas the MCM of Case 3 has higher values in the top right. This does not happen by chance but because we sorted the models from low complexity (homogeneous) to high complexity (geostatistical). This observation shows that Case 2 overestimates the posterior model weight of complex models, whereas Case 3 underestimates the posterior model weight of complex models. This can be explained, again, with the variance of the models and the data: In Case 2, the variance in the data (columns) is much larger than in the tested models (rows), and hence, the model with the highest variance has the best chance of scoring non-zero likelihoods. The opposite is happening in Case 3: The data show less variance than the noise-free models, and hence, in the spirit of the bias-variance tradeoff implicitly performed by BMS (Schöniger, Illman, et al., 2015), the model with the smallest variance that still fits the data reasonably well will score the highest model weight.

5.2.2. Results Based on Six Pumping Tests

We will use this occasion to show that stronger data sets do not only increase numerical errors but make the inconsistent cases close to ridiculous.

5.2.2.1. Which model is best (in Case 4)?

Again, the data-availability barrier only allows us to report the posterior model weights given experimental data for Cases 2 and 4 in Figure 12. We find a 86% posterior model weight of the zonated model in Case 4. This weight is similar to the self-identification potential of the zonated model in the MCM (92%). Hence, we again are optimistic that the zonated model is the quasi-data-generating model and no ”not-in-the-set model” generated the data (although the MCM cannot prove that, it could only reject that hope).

5.2.2.2. Can Posterior Model Weights be Increased With Less Noisy Data?

We find that reducing measurement noise will probably only slightly increase our confidence in choosing the zonated model because Case 1 only produces a slightly higher expected self-identification weight of 97%.

5.2.2.3. Can We Trust the Results?

The second (computationally much more challenging) scenario based on six pumping tests shows even more difficulties in the resulting MCMs (Figure 12). The matrix of Case 4 is almost symmetric ($P(M_4|M_3) \approx P(M_3|M_4)$) and hence we conclude that also the posterior model weights given experimental data can be trusted, especially because the MCM in the row and column of the largest posterior model weight (86%) is almost symmetric. Case 1, instead, is asymmetric. Hence, the results of Case 1 are corrupted by numerical errors. This means, that the upper limit of 97%, found previously, might not be trustworthy.

In Case 2, numerics completely broke down and all BME values for all models and all data sets are always zero. This happens because the random noise added to the 210 measurements produces completely unlikely data values as seen through the noise-free predictive distributions. As a consequence, the likelihood of any data set, which equals the multiplication of 210 independent likelihoods of single observations, was smaller than machine precision, and hence calculating average posterior model weights failed.

5.2.2.4. Can We Use Case 2 as a Proxy for Case 1?

The posterior weights of Case 2 should not be used for model selection due to the completely “broken” MCM. This “broken” MCM indicates that the inconsistency of Case 2 is enormous as discussed earlier. The resulting posterior weights cannot be trusted and they certainly should not be used as an approximation of Case 1.

5.2.3. Summary and Implications of Findings From Real-World Case Study

The results of our hydrogeological case study showed that the difference between the four cases is even more pronounced under real-world conditions with high-dimensional data sets and more dominant measurement noise. Further, posterior model weights of Case 2 can only be obtained in scenarios where we obtain BME analytically. In this case study (and mostly in practice), Case 2 is not available, so we recommend approaching *posterior model weights* of Case 1 from Case 4 by reducing the noise level in the likelihood function. To find out whether better (noise-free) data can make model choice more decisive, we recommend approaching the *MCM* of Case 1 from Case 3 by reducing the noise level in the likelihood function.

6. Summary and Conclusions

In this study, we discuss where and for which reasons measurement noise should be considered in Bayesian model selection. We distinguish between four different cases (accounting for noise in models and/or data: (a) no-no, (b) no-yes, (c) yes-no, (d) yes-yes that differ conceptually as visualized in Figure 2). We have demonstrated on three analytical scenarios and a real-world case study that these conceptual differences result in significantly different outcomes of Bayesian model selection. Thus, knowing which case to use is of high practical relevance.

Cases 2 and 3 are logically inconsistent because noise is considered *either* in the models *or* in the data. Therefore, we focus on the two consistent cases 1 and 4. They answer the following research questions:

1. Case 1: Which model is best in modeling the physics?
2. Case 4: Which model is best in predicting the data-generating process (i.e., physics plus noise)?

Philosophically, the choice is clear: either you are interested in pure physics, then you choose Case 1, or you are interested in simulating the data-generating process, then you move to Case 4. We have shown that model selection results under those two modeling goals do not necessarily agree and that the differences are especially pronounced for

1. Large measurement noise
2. Large data sets
3. Sudden jumps in the predictive density of models, especially to zero probability

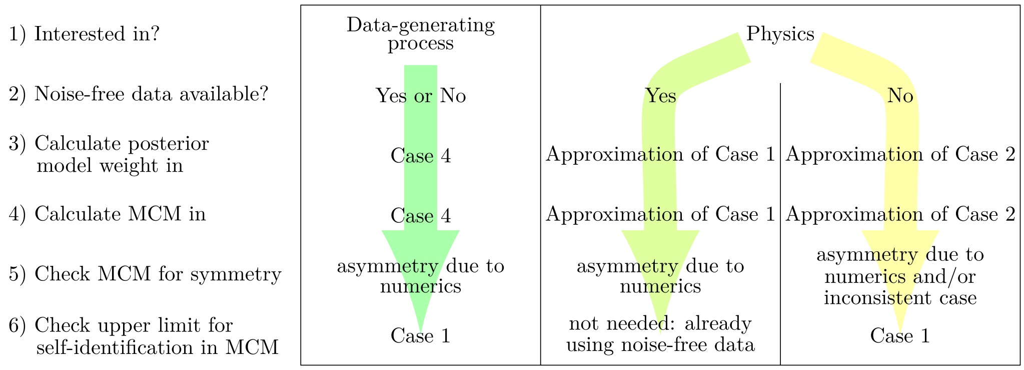
Practically, we face two challenges as visualized in Figure 3: First, in real-world applications, the data-availability barrier blocks us from using Case 1 and Case 3 for model selection because noise-free data is unavailable. One notable exception is the task of identifying a suitable surrogate model for an (expensive) complex model because this analysis does not involve real observed data. Second, if no analytical formulation of Bayesian model evidence (BME) is available, we can only numerically approximate Case 1 and Case 2 with bias (e.g., by using a narrow likelihood function instead of a Dirac delta function). We call this the numerical approximation barrier. As a result, only Case 4 can be evaluated straightforwardly.

To not give up on the scientifically more interesting Case 1, we investigated the potential of approximating Case 1 with any of the other three cases. From the analytical scenarios, we found that Case 2 is the best proxy of posterior model weights of Case 1. Hence, the inconsistent Case 2 is the best approximation of the "physics" case and a "default fall-back" to Case 4 is not optimal. Remember that Case 2 is only available if BME can be obtained analytically. In a general modeling context, Case 2, therefore, needs to be numerically approximated by reducing the noise level in the likelihood function of the model, effectively pushing the numerical-approximation barrier from Case 4 toward Case 2.

The model confusion matrix (MCM) can be used to evaluate how much we can trust the obtained posterior model weights. The MCM can be computed for all cases (with numerical challenges in Cases 1 and 2, as mentioned above). We have shown mathematically that the MCM of consistent cases is by design symmetric. Any asymmetry in the MCM, therefore, indicates that either numerical errors or logical inconsistencies exist. The degree of asymmetry can be understood as a warning of how much the (approximated) model selection results lack interpretability. Further, the MCM can be used to get insight on whether or not a better (almost) noise-free measurement device would make model selection more decisive: The self-identification potential of models in Case 1 provides an upper limit to decisiveness in model choice under minimal noise.

Based on our findings, we propose the following procedure for Bayesian model selection in the presence of measurement noise (visualized in Figure 13). First, the philosophical question (a) must be answered: are we interested in modeling physics or the data-generating process? Then, dependent on whether or not noise-free data is available (b), the posterior model weights of the corresponding case are calculated (c). Next, the corresponding MCM is calculated as a reference for interpreting the posterior model weight results (d). After that, we check the MCM for asymmetry, which reveals the degree of numerical errors and/or logical inconsistency (e). If the resulting MCM is strongly asymmetric, the posterior model weights and the MCM are deemed unreliable and one might be forced to change the research question. Finally, the MCM of Case 1 can be determined or approximated to find out whether better (noise-free) data can make a model choice more decisive (f).

With this recommended procedure, modelers are forced to reveal (and make themselves aware) of their motivation for model selection, and results are ensured to be as consistent as possible, given the practical limitations of real-world data availability and numerical implementation. We believe that understanding the four ways of how to treat measurement noise in Bayesian model selection is relevant to any modeling



1) Interested in?	Data-generating process	Physics	
2) Noise-free data available?	Yes or No	Yes	No
3) Calculate posterior model weight in	Case 4	Approximation of Case 1	Approximation of Case 2
4) Calculate MCM in	Case 4	Approximation of Case 1	Approximation of Case 2
5) Check MCM for symmetry	asymmetry due to numerics	asymmetry due to numerics	asymmetry due to numerics and/or inconsistent case
6) Check upper limit for self-identification in MCM	Case 1	not needed: already using noise-free data	Case 1

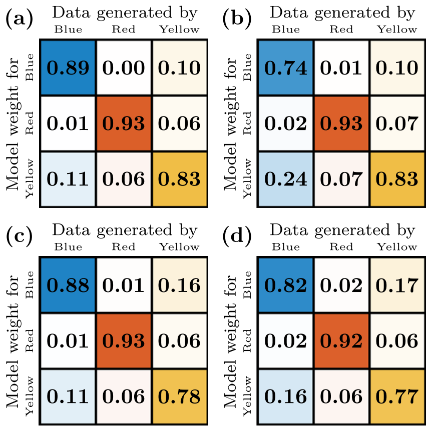
Figure 13. Recommended procedure to perform model selection. The color of the arrows indicates the number of approximations needed for model selection: green = few, yellow = many.

endeavor in water resources research and beyond, where measurement noise is significant enough to spoil a perfectly clear model choice.

Appendix A: Additional Results of Analytical Scenarios

A1 Model Confusion Weights of Scenario 1

Figure A1 displays the MCM obtained for scenario 1. Panel (a) shows the MCM in Case 1, panel (b) shows the MCM in Case 2, panel (c) the MCM in Case 3 and panel (d) in Case 4. Figure A1 reveals that the self-identification probability of the red model is almost independent of the chosen case. Differences are more pronounced for the blue and yellow model, with self-identification weights between 74% and 89% and between 77% and 83%, respectively. Cases 1 and 4 are practically symmetric; Cases 2 and 3 show some asymmetry between yellow/blue and blue/yellow.



(a)

Model weight for \ Data generated by	Blue	Red	Yellow
Blue	0.89	0.00	0.10
Red	0.01	0.93	0.06
Yellow	0.11	0.06	0.83

(b)

Model weight for \ Data generated by	Blue	Red	Yellow
Blue	0.74	0.01	0.10
Red	0.02	0.93	0.07
Yellow	0.24	0.07	0.83

(c)

Model weight for \ Data generated by	Blue	Red	Yellow
Blue	0.88	0.01	0.16
Red	0.01	0.93	0.06
Yellow	0.11	0.06	0.78

(d)

Model weight for \ Data generated by	Blue	Red	Yellow
Blue	0.82	0.02	0.17
Red	0.02	0.92	0.06
Yellow	0.16	0.06	0.77

Figure A1. Model confusion matrix obtained in (a) Case 1, (b) Case 2, (c) Case 3, and (d) Case 4 of scenario 1 (predictive PDFs of varied location and width).

A2 Model Confusion Weights of Scenario 2

Figure A2 displays the MCM obtained for scenario 2. Panel (a) shows the MCM in Case 1, panel (b) shows the MCM in Case 2, panel (c) the MCM in Case 3, and panel (d) in Case 4.

In Case 2 the red/blue (1%) and blue/red (36%) entries are completely different. This indicates that Case 2 is a bad approximation of Case 1 if we want to select between the red and blue model. And indeed, we can see

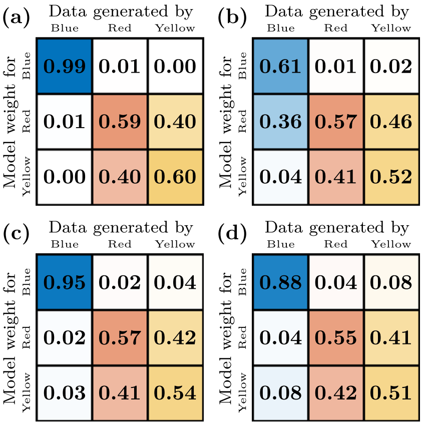


Figure A2. Model confusion matrix obtained in (a) Case 1, (b) Case 2, (c) Case 3, and (d) Case 4 of scenario 2 (predictive PDFs of bounded or semi-infinite support).

in Figure 5 that Case 1 and Case 2 differ a lot in the transition areas where the red and blue models show high posterior model weights.

In contrast, we can investigate the red and yellow models. Here, the off-diagonal entries (41% and 46%) are similar. Hence, we follow that Case 2 is a good approximation of Case 1 if we are interested in selecting between the red and yellow model. Again, Figure 5 confirms this because Case 1 and Case 2 have similar behavior in the area where the red and yellow models have high posterior model weights.

Data Availability Statement

The data, models, and ensembles of Schöniger, Wöhling, and Nowak (2015) were used for the real-world case study.

Acknowledgments
The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Collaborative Research Center SFB 1313 “Interface-Driven Multi-Field Processes in Porous Media—Flow, Transport and Deformation” (project number 327154368) and the Cluster of Excellence EXC 2075 ”Data-integrated Simulation Science (SimTech)” (project number 390740016). Open access funding enabled and organized by Projekt DEAL.

References

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian Theory* (Vol. *405*). John Wiley & Sons.

Brunetti, C., Linde, N., & 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. https://doi.org/10.1016/j.advwatres.2017.02.006

Elshall, A. S., & Ye, M. (2019). Making steppingstones out of stumbling blocks: A Bayesian model evidence stimator with application to groundwater transport model selection. *Water*, *11*(8), 1579. https://doi.org/10.3390/w11081579

Elsheikh, A. H., Wheeler, M. F., & Hoteit, I. (2014). Hybrid nested sampling algorithm for Bayesian model selection applied to inverse subsurface flow problems. *Journal of Computational Physics*, *258*, 319–337. https://doi.org/10.1016/j.jcp.2013.10.001

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (1995). *Bayesian data analysis*. CRC Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401. https://doi.org/10.1214/ss/1009212519

Höge, M., Guthke, A., & Nowak, W. (2019). The hydrologist’s guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, *572*, 96–107. https://doi.org/10.1016/j.jhydrol.2019.01.072

Höge, M., Guthke, A., & Nowak, W. (2020). Bayesian model weighting: The many faces of model averaging. *Water*, *12*(2), 309. https://doi.org/10.3390/w12020309

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, *54*(3), 1688–1715. https://doi.org/10.1002/2017WR021902

Illman, W. A., Zhu, J., Craig, A. J., & Yin, D. (2010). Comparison of aquifer characterization approaches through steady state groundwater model validation: A controlled laboratory sandbox study. *Water Resources Research*, *46*(4). https://doi.org/10.1029/2009WR007745

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford Univ. Press.

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

Laio, F., Di Baldassarre, G., & Montanari, A. (2009). Model selection techniques for the frequency analysis of hydrological extremes. *Water Resources Research*, *45*(7). https://doi.org/10.1029/2007WR006666

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

Leube, P., Geiges, A., & Nowak, W. (2012). Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resources Research*, *48*(2). https://doi.org/10.1029/2010WR010137

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

Lu, D., Ye, M., Meyer, P. D., Curtis, G. P., Shi, X., Niu, X.-F., & Yabusaki, S. B. (2013). Effects of error covariance structure on estimation of model averaging weights and predictive performance. *Water Resources Research*, *49*(9), 6029–6047. https://doi.org/10.1002/wrcr.20441

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, *41*(10). https://doi.org/10.1029/2004WR003719

Mohammadi, F., Kopmann, R., Guthke, A., Oladyshkin, S., & Nowak, W. (2018). Bayesian selection of hydro-morphodynamic models under computational time constraints. *Advances in Water Resources*, *117*, 53–64. https://doi.org/10.1016/j.advwatres.2018.05.007

Najafi, M., Moradkhani, H., & Jung, I. (2011). Assessing the uncertainties of hydrologic model selection in climate change impact studies. *Hydrological Processes*, *25*(18), 2814–2826. https://doi.org/10.1002/hyp.8043

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, *61*(9), 1666–1678. https://doi.org/10.1080/02626667.2016.1183009

Nowak, W., & Guthke, A. (2016). Entropy-based experimental design for optimal model discrimination in the geosciences. *Entropy*, *18*(11), 409. https://doi.org/10.3390/e18110409

Nowak, W., Rubin, Y., & de Barros, F. P. J. (2012). A hypothesis-driven approach to optimize field campaigns. *Water Resources Research*, *48*(6). https://doi.org/10.1029/2011WR011016

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

RamaRao, B. S., LaVenue, A. M., De Marsily, G., & Marietta, M. G. (1995). Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 1. Theory and computational experiments. *Water Resources Research*, *31*(3), 475–493. https://doi.org/10.1029/94WR02258

Schäfer Rodrigues Silva, A., Guthke, A., Höge, M., Cirpka, O. A., & Nowak, W. (2020). Strategies for simplifying reactive transport models-a Bayesian model comparison. *Water Resources Research*, *56*, e2020WR028100. https://doi.org/10.1029/2020WR028100

Schöniger, A., Illman, W. A., Wöhling, T., & Nowak, W. (2015). Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. *Journal of Hydrology*, *531*, 96–110. https://doi.org/10.1016/j.jhydrol.2015.07.047

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

Schöniger, A., Wöhling, T., Samaniego, L., & Nowak, W. (2014). Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, *50*(12), 9484–9513. https://doi.org/10.1002/2014WR016062

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

Volpi, E., Schoups, G., Firmani, G., & Vrugt, J. A. (2017). Sworn testimony of the model evidence: Gaussian mixture importance (game) sampling. *Water Resources Research*, *53*(7), 6133–6158. https://doi.org/10.1002/2016WR020167

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. https://doi.org/10.1006/jmps.1999.1278

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

Ye, M., Meyer, P. D., & Neuman, S. P. (2008). On model selection criteria in multimodel analysis. *Water Resources Research*, *44*(3). https://doi.org/10.1029/2008WR006803

Institut für Wasser- und Umweltsystemmodellierung
Universität Stuttgart

Pfaffenwaldring 61
70569 Stuttgart (Vaihingen)
Telefon (0711) 685 - 60156
Telefax (0711) 685 - 51073
E-Mail: iws@iws.uni-stuttgart.de
http://www.iws.uni-stuttgart.de

Direktoren
Prof. Dr. rer. nat. Dr.-Ing. András Bárdossy
Prof. Dr.-Ing. Rainer Helmig
Prof. Dr.-Ing. Wolfgang Nowak
Prof. Dr.-Ing. Silke Wieprecht

Vorstand *(Stand 21.05.2021)*
Prof. Dr. rer. nat. Dr.-Ing. A. Bárdossy
Prof. Dr.-Ing. R. Helmig
Prof. Dr.-Ing. W. Nowak
Prof. Dr.-Ing. S. Wieprecht
Prof. Dr. J.A. Sander Huisman
Jürgen Braun, PhD
apl. Prof. Dr.-Ing. H. Class
PD Dr.-Ing. Claus Haslauer
Stefan Haun, PhD
apl. Prof. Dr.-Ing. Sergey Oladyshkin
Dr. rer. nat. J. Seidel
Dr.-Ing. K. Terheiden

Emeriti
Prof. Dr.-Ing. habil. Dr.-Ing. E.h. Jürgen Giesecke
Prof. Dr.h.c. Dr.-Ing. E.h. Helmut Kobus, PhD

Lehrstuhl für Wasserbau und Wassermengenwirtschaft
Leiterin: Prof. Dr.-Ing. Silke Wieprecht
Stellv.: Dr.-Ing. Kristina Terheiden
Versuchsanstalt für Wasserbau
Leiter: Stefan Haun, PhD

Lehrstuhl für Hydromechanik und Hydrosystemmodellierung
Leiter: Prof. Dr.-Ing. Rainer Helmig
Stellv.: apl. Prof. Dr.-Ing. Holger Class

Lehrstuhl für Hydrologie und Geohydrologie
Leiter: Prof. Dr. rer. nat. Dr.-Ing. András Bárdossy
Stellv.: Dr. rer. nat. Jochen Seidel
Hydrogeophysik der Vadosen Zone
(mit Forschungszentrum Jülich)
Leiter: Prof. Dr. J.A. Sander Huisman

Lehrstuhl für Stochastische Simulation und Sicherheitsforschung für Hydrosysteme
Leiter: Prof. Dr.-Ing. Wolfgang Nowak
Stellv.: apl. Prof. Dr.-Ing. Sergey Oladyshkin

***VEGAS,* Versuchseinrichtung zur Grundwasser- und Altlastensanierung**
Leiter: Jürgen Braun, PhD
PD Dr.-Ing. Claus Haslauer

Verzeichnis der Mitteilungshefte

1 Röhnisch, Arthur: *Die Bemühungen um eine Wasserbauliche Versuchsanstalt an der Technischen Hochschule Stuttgart*, und
Fattah Abouleid, Abdel: *Beitrag zur Berechnung einer in lockeren Sand gerammten, zweifach verankerten Spundwand*, 1963

2 Marotz, Günter: *Beitrag zur Frage der Standfestigkeit von dichten Asphaltbelägen im Großwasserbau*, 1964

3 Gurr, Siegfried: *Beitrag zur Berechnung zusammengesetzter ebener Flächentrag-werke unter besonderer Berücksichtigung ebener Stauwände, mit Hilfe von Rand-wert- und Lastwertmatrizen*, 1965

4 Plica, Peter: *Ein Beitrag zur Anwendung von Schalenkonstruktionen im Stahlwasserbau*, und
Petrikat, Kurt: *Möglichkeiten und Grenzen des wasserbaulichen Versuchswesens*, 1966

5 Plate, Erich: *Beitrag zur Bestimmung der Windgeschwindigkeitsverteilung in der durch eine Wand gestörten bodennahen Luftschicht*, und
Röhnisch, Arthur; Marotz, Günter: *Neue Baustoffe und Bauausführungen für den Schutz der Böschungen und der Sohle von Kanälen, Flüssen und Häfen; Gestehungskosten und jeweilige Vorteile*, sowie
Unny, T.E.: *Schwingungsuntersuchungen am Kegelstrahlschieber*, 1967

6 Seiler, Erich: *Die Ermittlung des Anlagenwertes der bundeseigenen Binnenschiffahrtsstraßen und Talsperren und des Anteils der Binnenschiffahrt an diesem Wert*, 1967

7 *Sonderheft anläßlich des 65. Geburtstages von Prof. Arthur Röhnisch mit Beiträgen von* Benk, Dieter; Breitling, J.; Gurr, Siegfried; Haberhauer, Robert; Honekamp, Hermann; Kuz, Klaus Dieter; Marotz, Günter; Mayer-Vorfelder, Hans-Jörg; Miller, Rudolf; Plate, Erich J.; Radomski, Helge; Schwarz, Helmut; Vollmer, Ernst; Wildenhahn, Eberhard; 1967

8 Jumikis, Alfred: *Beitrag zur experimentellen Untersuchung des Wassernachschubs in einem gefrierenden Boden und die Beurteilung der Ergebnisse*, 1968

9 Marotz, Günter: *Technische Grundlagen einer Wasserspeicherung im natürlichen Untergrund*, 1968

10 Radomski, Helge: *Untersuchungen über den Einfluß der Querschnittsform wellenförmiger Spundwände auf die statischen und rammtechnischen Eigenschaften*, 1968

11 Schwarz, Helmut: *Die Grenztragfähigkeit des Baugrundes bei Einwirkung vertikal gezogener Ankerplatten als zweidimensionales Bruchproblem*, 1969

12 Erbel, Klaus: *Ein Beitrag zur Untersuchung der Metamorphose von Mittelgebirgsschneedecken unter besonderer Berücksichtigung eines Verfahrens zur Bestimmung der thermischen Schneequalität*, 1969

13 Westhaus, Karl-Heinz: *Der Strukturwandel in der Binnenschiffahrt und sein Einfluß auf den Ausbau der Binnenschiffskanäle*, 1969

14 Mayer-Vorfelder, Hans-Jörg: *Ein Beitrag zur Berechnung des Erdwiderstandes unter Ansatz der logarithmischen Spirale als Gleitflächenfunktion*, 1970

15 Schulz, Manfred: *Berechnung des räumlichen Erddruckes auf die Wandung kreiszylindrischer Körper*, 1970

16 Mobasseri, Manoutschehr: *Die Rippenstützmauer. Konstruktion und Grenzen ihrer Standsicherheit*, 1970

17 Benk, Dieter: *Ein Beitrag zum Betrieb und zur Bemessung von Hochwasserrückhaltebecken*, 1970

18 Gàl, Attila: *Bestimmung der mitschwingenden Wassermasse bei überströmten Fischbauchklappen mit kreiszylindrischem Staublech*, 1971, *vergriffen*

19 Kuz, Klaus Dieter: *Ein Beitrag zur Frage des Einsetzens von Kavitationserscheinungen in einer Düsenströmung bei Berücksichtigung der im Wasser gelösten Gase*, 1971, *vergriffen*

20 Schaak, Hartmut: *Verteilleitungen von Wasserkraftanlagen*, 1971

21 *Sonderheft zur Eröffnung der neuen Versuchsanstalt des Instituts für Wasserbau der Universität Stuttgart mit Beiträgen von* Brombach, Hansjörg; Dirksen, Wolfram; Gàl, Attila; Gerlach, Reinhard; Giesecke, Jürgen; Holthoff, Franz-Josef; Kuz, Klaus Dieter; Marotz, Günter; Minor, Hans-Erwin; Petrikat, Kurt; Röhnisch, Arthur; Rueff, Helge; Schwarz, Helmut; Vollmer, Ernst; Wildenhahn, Eberhard; 1972

22 Wang, Chung-su: *Ein Beitrag zur Berechnung der Schwingungen an Kegelstrahlschiebern*, 1972

23 Mayer-Vorfelder, Hans-Jörg: *Erdwiderstandsbeiwerte nach dem Ohde-Variationsverfahren*, 1972

24 Minor, Hans-Erwin: *Beitrag zur Bestimmung der Schwingungsanfachungsfunktionen überströmter Stauklappen*, 1972, *vergriffen*

25 Brombach, Hansjörg: *Untersuchung strömungsmechanischer Elemente (Fluidik) und die Möglichkeit der Anwendung von Wirbelkammerelementen im Wasserbau*, 1972, *vergriffen*

26 Wildenhahn, Eberhard: *Beitrag zur Berechnung von Horizontalfilterbrunnen*, 1972

27 Steinlein, Helmut: *Die Eliminierung der Schwebstoffe aus Flußwasser zum Zweck der unterirdischen Wasserspeicherung, gezeigt am Beispiel der Iller*, 1972

28 Holthoff, Franz Josef: *Die Überwindung großer Hubhöhen in der Binnenschiffahrt durch Schwimmerhebewerke*, 1973

29 Röder, Karl: *Einwirkungen aus Baugrundbewegungen auf trog- und kastenförmige Konstruktionen des Wasser- und Tunnelbaues*, 1973

30 Kretschmer, Heinz: *Die Bemessung von Bogenstaumauern in Abhängigkeit von der Talform*, 1973

31 Honekamp, Hermann: *Beitrag zur Berechnung der Montage von Unterwasserpipelines*, 1973

32 Giesecke, Jürgen: *Die Wirbelkammertriode als neuartiges Steuerorgan im Wasserbau*, und Brombach, Hansjörg: *Entwicklung, Bauformen, Wirkungsweise und Steuereigenschaften von Wirbelkammerverstärkern*, 1974

33 Rueff, Helge: *Untersuchung der schwingungserregenden Kräfte an zwei hintereinander angeordneten Tiefschützen unter besonderer Berücksichtigung von Kavitation*, 1974

34 Röhnisch, Arthur: *Einpreßversuche mit Zementmörtel für Spannbeton - Vergleich der Ergebnisse von Modellversuchen mit Ausführungen in Hüllwellrohren*, 1975

35 *Sonderheft anläßlich des 65. Geburtstages von Prof. Dr.-Ing. Kurt Petrikat mit Beiträgen von*: Brombach, Hansjörg; Erbel, Klaus; Flinspach, Dieter; Fischer jr., Richard; Gàl, Attila; Gerlach, Reinhard; Giesecke, Jürgen; Haberhauer, Robert; Hafner Edzard; Hausenblas, Bernhard; Horlacher, Hans-Burkhard; Hutarew, Andreas; Knoll, Manfred; Krummet, Ralph; Marotz, Günter; Merkle, Theodor; Miller, Christoph; Minor, Hans-Erwin; Neumayer, Hans; Rao, Syamala; Rath, Paul; Rueff, Helge; Ruppert, Jürgen; Schwarz, Wolfgang; Topal-Gökceli, Mehmet; Vollmer, Ernst; Wang, Chung-su; Weber, Hans-Georg; 1975

36 Berger, Jochum: *Beitrag zur Berechnung des Spannungszustandes in rotationssymmetrisch belasteten Kugelschalen veränderlicher Wandstärke unter Gas- und Flüssigkeitsdruck durch Integration schwach singulärer Differentialgleichungen*, 1975

37 Dirksen, Wolfram: *Berechnung instationärer Abflußvorgänge in gestauten Gerinnen mittels Differenzenverfahren und die Anwendung auf Hochwasserrückhaltebecken*, 1976

38 Horlacher, Hans-Burkhard: *Berechnung instationärer Temperatur- und Wärmespannungsfelder in langen mehrschichtigen Hohlzylindern*, 1976

39 Hafner, Edzard: *Untersuchung der hydrodynamischen Kräfte auf Baukörper im Tiefwasserbereich des Meeres*, 1977, ISBN 3-921694-39-6

40 Ruppert, Jürgen: *Über den Axialwirbelkammerverstärker für den Einsatz im Wasserbau*, 1977, ISBN 3-921694-40-X

41 Hutarew, Andreas: *Beitrag zur Beeinflußbarkeit des Sauerstoffgehalts in Fließgewässern an Abstürzen und Wehren*, 1977, ISBN 3-921694-41-8, *vergriffen*

42 Miller, Christoph: *Ein Beitrag zur Bestimmung der schwingungserregenden Kräfte an unterströmten Wehren*, 1977, ISBN 3-921694-42-6

43 Schwarz, Wolfgang: *Druckstoßberechnung unter Berücksichtigung der Radial- und Längsverschiebungen der Rohrwandung*, 1978, ISBN 3-921694-43-4

44 Kinzelbach, Wolfgang: *Numerische Untersuchungen über den optimalen Einsatz variabler Kühlsysteme einer Kraftwerkskette am Beispiel Oberrhein*, 1978, ISBN 3-921694-44-2

45 Barczewski, Baldur: *Neue Meßmethoden für Wasser-Luftgemische und deren Anwendung auf zweiphasige Auftriebsstrahlen*, 1979, ISBN 3-921694-45-0

46 Neumayer, Hans: *Untersuchung der Strömungsvorgänge in radialen Wirbelkammerverstärkern*, 1979, ISBN 3-921694-46-9

47 Elalfy, Youssef-Elhassan: *Untersuchung der Strömungsvorgänge in Wirbelkammerdioden und -drosseln*, 1979, ISBN 3-921694-47-7

48 Brombach, Hansjörg: *Automatisierung der Bewirtschaftung von Wasserspeichern*, 1981, ISBN 3-921694-48-5

49 Geldner, Peter: *Deterministische und stochastische Methoden zur Bestimmung der Selbstdichtung von Gewässern*, 1981, ISBN 3-921694-49-3, *vergriffen*

50 Mehlhorn, Hans: *Temperaturveränderungen im Grundwasser durch Brauchwassereinleitungen*, 1982, ISBN 3-921694-50-7, vergriffen

51 Hafner, Edzard: *Rohrleitungen und Behälter im Meer*, 1983, ISBN 3-921694-51-5

52 Rinnert, Bernd: *Hydrodynamische Dispersion in porösen Medien: Einfluß von Dichteunterschieden auf die Vertikalvermischung in horizontaler Strömung*, 1983, ISBN 3-921694-52-3, vergriffen

53 Lindner, Wulf: *Steuerung von Grundwasserentnahmen unter Einhaltung ökologischer Kriterien*, 1983, ISBN 3-921694-53-1, vergriffen

54 Herr, Michael; Herzer, Jörg; Kinzelbach, Wolfgang; Kobus, Helmut; Rinnert, Bernd: *Methoden zur rechnerischen Erfassung und hydraulischen Sanierung von Grundwasserkontaminationen*, 1983, ISBN 3-921694-54-X

55 Schmitt, Paul: *Wege zur Automatisierung der Niederschlagsermittlung*, 1984, ISBN 3-921694-55-8, vergriffen

56 Müller, Peter: *Transport und selektive Sedimentation von Schwebstoffen bei gestautem Abfluß*, 1985, ISBN 3-921694-56-6

57 El-Qawasmeh, Fuad: *Möglichkeiten und Grenzen der Tropfbewässerung unter besonderer Berücksichtigung der Verstopfungsanfälligkeit der Tropfelemente*, 1985, ISBN 3-921694-57-4, vergriffen

58 Kirchenbaur, Klaus: *Mikroprozessorgesteuerte Erfassung instationärer Druckfelder am Beispiel seegangsbelasteter Baukörper*, 1985, ISBN 3-921694-58-2

59 Kobus, Helmut (Hrsg.): *Modellierung des großräumigen Wärme- und Schadstofftransports im Grundwasser*, Tätigkeitsbericht 1984/85 (DFG-Forschergruppe an den Universitäten Hohenheim, Karlsruhe und Stuttgart), 1985, ISBN 3-921694-59-0, vergriffen

60 Spitz, Karlheinz: *Dispersion in porösen Medien: Einfluß von Inhomogenitäten und Dichteunterschieden*, 1985, ISBN 3-921694-60-4, vergriffen

61 Kobus, Helmut: *An Introduction to Air-Water Flows in Hydraulics*, 1985, ISBN 3-921694-61-2

62 Kaleris, Vassilios: *Erfassung des Austausches von Oberflächen- und Grundwasser in horizontalebenen Grundwassermodellen*, 1986, ISBN 3-921694-62-0

63 Herr, Michael: *Grundlagen der hydraulischen Sanierung verunreinigter Porengrundwasserleiter*, 1987, ISBN 3-921694-63-9

64 Marx, Walter: *Berechnung von Temperatur und Spannung in Massenbeton infolge Hydratation*, 1987, ISBN 3-921694-64-7

65 Koschitzky, Hans-Peter: *Dimensionierungskonzept für Sohlbelüfter in Schußrinnen zur Vermeidung von Kavitationsschäden*, 1987, ISBN 3-921694-65-5

66 Kobus, Helmut (Hrsg.): *Modellierung des großräumigen Wärme- und Schadstofftransports im Grundwasser*, Tätigkeitsbericht 1986/87 (DFG-Forschergruppe an den Universitäten Hohenheim, Karlsruhe und Stuttgart) 1987, ISBN 3-921694-66-3

67 Söll, Thomas: *Berechnungsverfahren zur Abschätzung anthropogener Temperaturanomalien im Grundwasser*, 1988, ISBN 3-921694-67-1

68 Dittrich, Andreas; Westrich, Bernd: *Bodenseeufererosion, Bestandsaufnahme und Bewertung*, 1988, ISBN 3-921694-68-X, vergriffen

69 Huwe, Bernd; van der Ploeg, Rienk R.: *Modelle zur Simulation des Stickstoffhaushaltes von Standorten mit unterschiedlicher landwirtschaftlicher Nutzung*, 1988, ISBN 3-921694-69-8, vergriffen

70 Stephan, Karl: *Integration elliptischer Funktionen*, 1988, ISBN 3-921694-70-1

71 Kobus, Helmut; Zilliox, Lothaire (Hrsg.): *Nitratbelastung des Grundwassers, Auswirkungen der Landwirtschaft auf die Grundwasser- und Rohwasserbeschaffenheit und Maßnahmen zum Schutz des Grundwassers*. Vorträge des deutsch-französischen Kolloquiums am 6. Oktober 1988, Universitäten Stuttgart und Louis Pasteur Strasbourg (Vorträge in deutsch oder französisch, Kurzfassungen zweisprachig), 1988, ISBN 3-921694-71-X

72 Soyeaux, Renald: *Unterströmung von Stauanlagen auf klüftigem Untergrund unter Berücksichtigung laminarer und turbulenter Fließzustände*,1991, ISBN 3-921694-72-8

73 Kohane, Roberto: *Berechnungsmethoden für Hochwasserabfluß in Fließgewässern mit übeströmten Vorländern*, 1991, ISBN 3-921694-73-6

74 Hassinger, Reinhard: *Beitrag zur Hydraulik und Bemessung von Blocksteinrampen in flexibler Bauweise*, 1991, ISBN 3-921694-74-4, vergriffen

75 Schäfer, Gerhard: *Einfluß von Schichtenstrukturen und lokalen Einlagerungen auf die Längsdispersion in Porengrundwasserleitern*, 1991, ISBN 3-921694-75-2

76 Giesecke, Jürgen: *Vorträge, Wasserwirtschaft in stark besiedelten Regionen; Umweltforschung mit Schwerpunkt Wasserwirtschaft*, 1991, ISBN 3-921694-76-0

77 Huwe, Bernd: *Deterministische und stochastische Ansätze zur Modellierung des Stickstoffhaushalts landwirtschaftlich genutzter Flächen auf unterschiedlichem Skalenniveau*, 1992, ISBN 3-921694-77-9, vergriffen

78 Rommel, Michael: *Verwendung von Kluftdaten zur realitätsnahen Generierung von Kluftnetzen mit anschließender laminar-turbulenter Strömungsberechnung*, 1993, ISBN 3-92 1694-78-7

79 Marschall, Paul: *Die Ermittlung lokaler Stofffrachten im Grundwasser mit Hilfe von Einbohrloch-Meßverfahren*, 1993, ISBN 3-921694-79-5, vergriffen

80 Ptak, Thomas: *Stofftransport in heterogenen Porenaquiferen: Felduntersuchungen und stochastische Modellierung*, 1993, ISBN 3-921694-80-9, vergriffen

81 Haakh, Frieder: *Transientes Strömungsverhalten in Wirbelkammern,* 1993, ISBN 3-921694-81-7

82 Kobus, Helmut; Cirpka, Olaf; Barczewski, Baldur; Koschitzky, Hans-Peter: *Versuchseinrichtung zur Grundwasser- und Altlastensanierung VEGAS, Konzeption und Programmrahmen*, 1993, ISBN 3-921694-82-5

83 Zang, Weidong: *Optimaler Echtzeit-Betrieb eines Speichers mit aktueller Abflußregenerierung*, 1994, ISBN 3-921694-83-3, vergriffen

84 Franke, Hans-Jörg: *Stochastische Modellierung eines flächenhaften Stoffeintrages und Transports in Grundwasser am Beispiel der Pflanzenschutzmittelproblematik*, 1995, ISBN 3-921694-84-1

85 Lang, Ulrich: *Simulation regionaler Strömungs- und Transportvorgänge in Karstaquiferen mit Hilfe des Doppelkontinuum-Ansatzes: Methodenentwicklung und Parameteridentifikation*, 1995, ISBN 3-921694-85-X, vergriffen

86 Helmig, Rainer: *Einführung in die Numerischen Methoden der Hydromechanik*, 1996, ISBN 3-921694-86-8, vergriffen

87 Cirpka, Olaf: *CONTRACT: A Numerical Tool for Contaminant Transport and Chemical Transformations - Theory and Program Documentation* -, 1996, ISBN 3-921694-87-6

88 Haberlandt, Uwe: *Stochastische Synthese und Regionalisierung des Niederschlages für Schmutzfrachtberechnungen*, 1996, ISBN 3-921694-88-4

89 Croisé, Jean: *Extraktion von flüchtigen Chemikalien aus natürlichen Lockergesteinen mittels erzwungener Luftströmung*, 1996, ISBN 3-921694-89-2, vergriffen

90 Jorde, Klaus: *Ökologisch begründete, dynamische Mindestwasserregelungen bei Ausleitungskraftwerken*, 1997, ISBN 3-921694-90-6, vergriffen

91 Helmig, Rainer: *Gekoppelte Strömungs- und Transportprozesse im Untergrund - Ein Beitrag zur Hydrosystemmodellierung-*, 1998, ISBN 3-921694-91-4, vergriffen

92 Emmert, Martin: *Numerische Modellierung nichtisothermer Gas-Wasser Systeme in porösen Medien,* 1997, ISBN 3-921694-92-2

93 Kern, Ulrich: *Transport von Schweb- und Schadstoffen in staugeregelten Fließgewässern am Beispiel des Neckars*, 1997, ISBN 3-921694-93-0, vergriffen

94 Förster, Georg: *Druckstoßdämpfung durch große Luftblasen in Hochpunkten von Rohrleitungen 1997*, ISBN 3-921694-94-9

95 Cirpka, Olaf: *Numerische Methoden zur Simulation des reaktiven Mehrkomponententransports im Grundwasser*, 1997, ISBN 3-921694-95-7, *vergriffen*

96 Färber, Arne: *Wärmetransport in der ungesättigten Bodenzone: Entwicklung einer thermischen In-situ-Sanierungstechnologie*, 1997, ISBN 3-921694-96-5

97 Betz, Christoph: *Wasserdampfdestillation von Schadstoffen im porösen Medium: Entwicklung einer thermischen In-situ-Sanierungstechnologie*, 1998, SBN 3-921694-97-3

98 Xu, Yichun: *Numerical Modeling of Suspended Sediment Transport in Rivers*, 1998, ISBN 3-921694-98-1, *vergriffen*

99 Wüst, Wolfgang: *Geochemische Untersuchungen zur Sanierung CKW-kontaminierter Aquifere mit Fe(0)-Reaktionswänden,* 2000, ISBN 3-933761-02-2

100 Sheta, Hussam: *Simulation von Mehrphasenvorgängen in porösen Medien unter Einbeziehung von Hysterese-Effekten*, 2000, ISBN 3-933761-03-4

101 Ayros, Edwin: *Regionalisierung extremer Abflüsse auf der Grundlage statistischer Verfahren*, 2000, ISBN 3-933761-04-2, *vergriffen*

102 Huber, Ralf: *Compositional Multiphase Flow and Transport in Heterogeneous Porous Media*, 2000, ISBN 3-933761-05-0

103 Braun, Christopherus: *Ein Upscaling-Verfahren für Mehrphasenströmungen in porösen Medien*, 2000, ISBN 3-933761-06-9

104 Hofmann, Bernd: *Entwicklung eines rechnergestützten Managementsystems zur Beurteilung von Grundwasserschadensfällen*, 2000, ISBN 3-933761-07-7

105 Class, Holger: *Theorie und numerische Modellierung nichtisothermer Mehrphasenprozesse in NAPL-kontaminierten porösen Medien,* 2001, ISBN 3-933761-08-5

106 Schmidt, Reinhard: *Wasserdampf- und Heißluftinjektion zur thermischen Sanierung kontaminierter Standorte*, 2001, ISBN 3-933761-09-3

107 Josef, Reinhold: *Schadstoffextraktion mit hydraulischen Sanierungsverfahren unter Anwendung von grenzflächenaktiven Stoffen*, 2001, ISBN 3-933761-10-7

108 Schneider, Matthias: *Habitat- und Abflussmodellierung für Fließgewässer mit unscharfen Berechnungsansätzen*, 2001, ISBN 3-933761-11-5

109 Rathgeb, Andreas: *Hydrodynamische Bemessungsgrundlagen für Lockerdeckwerke an überströmbaren Erddämmen*, 2001, ISBN 3-933761-12-3

110 Lang, Stefan: *Parallele numerische Simulation instätionärer Probleme mit adaptiven Methoden auf unstrukturierten Gittern*, 2001, ISBN 3-933761-13-1

111 Appt, Jochen; Stumpp Simone: *Die Bodensee-Messkampagne 2001, IWS/CWR Lake Constance Measurement Program 2001*, 2002, ISBN 3-933761-14-X

112 Heimerl, Stephan: *Systematische Beurteilung von Wasserkraftprojekten,* 2002, ISBN 3-933761-15-8, *vergriffen*

113 Iqbal, Amin: *On the Management and Salinity Control of Drip Irrigation,* 2002, ISBN 3-933761-16-6

114 Silberhorn-Hemminger, Annette: *Modellierung von Kluftaquifersystemen: Geostatistische Analyse und deterministisch-stochastische Kluftgenerierung,* 2002, ISBN 3-933761-17-4

115 Winkler, Angela: *Prozesse des Wärme- und Stofftransports bei der In-situ-Sanierung mit festen Wärmequellen,* 2003, ISBN 3-933761-18-2

116 Marx, Walter: *Wasserkraft, Bewässerung, Umwelt - Planungs- und Bewertungsschwerpunkte der Wasserbewirtschaftung,* 2003, ISBN 3-933761-19-0

117 Hinkelmann, Reinhard: *Efficient Numerical Methods and Information-Processing Techniques in Environment Water,* 2003, ISBN 3-933761-20-4

118 Samaniego-Eguiguren, Luis Eduardo: *Hydrological Consequences of Land Use / Land Cover and Climatic Changes in Mesoscale Catchments*, 2003, ISBN 3-933761-21-2

119 Neunhäuserer, Lina: *Diskretisierungsansätze zur Modellierung von Strömungs- und Transportprozessen in geklüftet-porösen Medien*, 2003, ISBN 3-933761-22-0

120 Paul, Maren: *Simulation of Two-Phase Flow in Heterogeneous Poros Media with Adaptive Methods*, 2003, ISBN 3-933761-23-9

121 Ehret, Uwe: *Rainfall and Flood Nowcasting in Small Catchments using Weather Radar*, 2003, ISBN 3-933761-24-7

122 Haag, Ingo: *Der Sauerstoffhaushalt staugeregelter Flüsse am Beispiel des Neckars - Analysen, Experimente, Simulationen -,* 2003, ISBN 3-933761-25-5

123 Appt, Jochen: *Analysis of Basin-Scale Internal Waves in Upper Lake Constance,* 2003, ISBN 3-933761-26-3

124 Hrsg.: Schrenk, Volker; Batereau, Katrin; Barczewski, Baldur; Weber, Karolin und Koschitzky, Hans-Peter: *Symposium Ressource Fläche und VEGAS - Statuskolloquium 2003, 30. September und 1. Oktober 2003,* 2003, ISBN 3-933761-27-1

125 Omar Khalil Ouda: *Optimisation of Agricultural Water Use: A Decision Support System for the Gaza Strip,* 2003, ISBN 3-933761-28-0

126 Batereau, Katrin: *Sensorbasierte Bodenluftmessung zur Vor-Ort-Erkundung von Schadensherden im Untergrund*, 2004, ISBN 3-933761-29-8

127 Witt, Oliver: *Erosionsstabilität von Gewässersedimenten mit Auswirkung auf den Stofftransport bei Hochwasser am Beispiel ausgewählter Stauhaltungen des Oberrheins*, 2004, ISBN 3-933761-30-1

128 Jakobs, Hartmut: *Simulation nicht-isothermer Gas-Wasser-Prozesse in komplexen Kluft-Matrix-Systemen*, 2004, ISBN 3-933761-31-X

129 Li, Chen-Chien: *Deterministisch-stochastisches Berechnungskonzept zur Beurteilung der Auswirkungen erosiver Hochwasserereignisse in Flussstauhaltungen,* 2004, ISBN 3-933761-32-8

130 Reichenberger, Volker; Helmig, Rainer; Jakobs, Hartmut; Bastian, Peter; Niessner, Jennifer: *Complex Gas-Water Processes in Discrete Fracture-Matrix Systems: Up-scaling, Mass-Conservative Discretization and Efficient Multilevel Solution,* 2004, ISBN 3-933761-33-6

131 Hrsg.: Barczewski, Baldur; Koschitzky, Hans-Peter; Weber, Karolin; Wege, Ralf: *VEGAS - Statuskolloquium 2004,* Tagungsband zur Veranstaltung am 05. Oktober 2004 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2004, ISBN 3-933761-34-4

132 Asie, Kemal Jabir: *Finite Volume Models for Multiphase Multicomponent Flow through Porous Media*. 2005, ISBN 3-933761-35-2

133 Jacoub, George: *Development of a 2-D Numerical Module for Particulate Contaminant Transport in Flood Retention Reservoirs and Impounded Rivers,* 2004, ISBN 3-933761-36-0

134 Nowak, Wolfgang: *Geostatistical Methods for the Identification of Flow and Transport Parameters in the Subsurface,* 2005, ISBN 3-933761-37-9

135 Süß, Mia: *Analysis of the influence of structures and boundaries on flow and transport processes in fractured porous media*, 2005, ISBN 3-933761-38-7

136 Jose, Surabhin Chackiath: *Experimental Investigations on Longitudinal Dispersive Mixing in Heterogeneous Aquifers,* 2005, ISBN: 3-933761-39-5

137 Filiz, Fulya: *Linking Large-Scale Meteorological Conditions to Floods in Mesoscale Catchments,* 2005, ISBN 3-933761-40-9

138 Qin, Minghao: *Wirklichkeitsnahe und recheneffiziente Ermittlung von Temperatur und Spannungen bei großen RCC-Staumauern*, 2005, ISBN 3-933761-41-7

139 Kobayashi, Kenichiro: *Optimization Methods for Multiphase Systems in the Subsurface - Application to Methane Migration in Coal Mining Areas*, 2005, ISBN 3-933761-42-5

140 Rahman, Md. Arifur: *Experimental Investigations on Transverse Dispersive Mixing in Heterogeneous Porous Media,* 2005, ISBN 3-933761-43-3

141 Schrenk, Volker: *Ökobilanzen zur Bewertung von Altlastensanierungsmaßnahmen,* 2005, ISBN 3-933761-44-1

142 Hundecha, Hirpa Yeshewatesfa: *Regionalization of Parameters of a Conceptual Rainfall-Runoff Model*, 2005, ISBN: 3-933761-45-X

143 Wege, Ralf: *Untersuchungs- und Überwachungsmethoden für die Beurteilung natürlicher Selbstreinigungsprozesse im Grundwasser,* 2005, ISBN 3-933761-46-8

144 Breiting, Thomas: *Techniken und Methoden der Hydroinformatik - Modellierung von komplexen Hydrosystemen im Untergrund*, 2006, ISBN 3-933761-47-6

145 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Müller, Martin: *Ressource Untergrund: 10 Jahre VEGAS: Forschung und Technologieentwicklung zum Schutz von Grundwasser und Boden,* Tagungsband zur Veranstaltung am 28. und 29. September 2005 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2005, ISBN 3-933761-48-4

146 Rojanschi, Vlad: *Abflusskonzentration in mesoskaligen Einzugsgebieten unter Berücksichtigung des Sickerraumes,* 2006, ISBN 3-933761-49-2

147 Winkler, Nina Simone: *Optimierung der Steuerung von Hochwasserrückhaltebecken-systemen,* 2006, ISBN 3-933761-50-6

148 Wolf, Jens: *Räumlich differenzierte Modellierung der Grundwasserströmung alluvialer Aquifere für mesoskalige Einzugsgebiete*, 2006, ISBN: 3-933761-51-4

149 Kohler, Beate: *Externe Effekte der Laufwasserkraftnutzung*, 2006, ISBN 3-933761-52-2

150 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhrmann, Matthias: *VEGAS-Statuskolloquium 2006,* Tagungsband zur Veranstaltung am 28. September 2006 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2006, ISBN 3-933761-53-0

151 Niessner, Jennifer: *Multi-Scale Modeling of Multi-Phase - Multi-Component Processes in Heterogeneous Porous Media*, 2006, ISBN 3-933761-54-9

152 Fischer, Markus: *Beanspruchung eingeerdeter Rohrleitungen infolge Austrocknung bindiger Böden,* 2006, ISBN 3-933761-55-7

153 Schneck, Alexander: *Optimierung der Grundwasserbewirtschaftung unter Berücksichtigung der Belange der Wasserversorgung, der Landwirtschaft und des Naturschutzes*, 2006, ISBN 3-933761-56-5

154 Das, Tapash: *The Impact of Spatial Variability of Precipitation on the Predictive Uncertainty of Hydrological Models,* 2006, ISBN 3-33761-57-3

155 Bielinski, Andreas: *Numerical Simulation of CO_2 sequestration in geological formations*, 2007, ISBN 3-933761-58-1

156 Mödinger, Jens: *Entwicklung eines Bewertungs- und Entscheidungsunterstützungssystems für eine nachhaltige regionale Grundwasserbewirtschaftung*, 2006, ISBN 3-933761-60-3

157 Manthey, Sabine: *Two-phase flow processes with dynamic effects in porous media - parameter estimation and simulation,* 2007, ISBN 3-933761-61-1

158 Pozos Estrada, Oscar: *Investigation on the Effects of Entrained Air in Pipelines,* 2007, ISBN 3-933761-62-X

159 Ochs, Steffen Oliver: *Steam injection into saturated porous media – process analysis including experimental and numerical investigations*, 2007, ISBN 3-933761-63-8

160 Marx, Andreas: *Einsatz gekoppelter Modelle und Wetterradar zur Abschätzung von Niederschlagsintensitäten und zur Abflussvorhersage, 2007,* ISBN 3-933761-64-6

161 Hartmann, Gabriele Maria: *Investigation of Evapotranspiration Concepts in Hydrological Modelling for Climate Change Impact Assessment*, 2007, ISBN 3-933761-65-4

162 Kebede Gurmessa, Tesfaye: *Numerical Investigation on Flow and Transport Characteristics to Improve Long-Term Simulation of Reservoir Sedimentation*, 2007, ISBN 3-933761-66-2

163 Trifković, Aleksandar: *Multi-objective and Risk-based Modelling Methodology for Planning, Design and Operation of Water Supply Systems*, 2007, ISBN 3-933761-67-0

164 Götzinger, Jens: *Distributed Conceptual Hydrological Modelling - Simulation of Climate, Land Use Change Impact and Uncertainty Analysis*, 2007, ISBN 3-933761-68-9

165 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhrmann, Matthias: *VEGAS – Kolloquium 2007,* Tagungsband zur Veranstaltung am 26. September 2007 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2007, ISBN 3-933761-69-7

166 Freeman, Beau: *Modernization Criteria Assessment for Water Resources Planning; Klamath Irrigation Project, U.S.*, 2008, ISBN 3-933761-70-0

167 Dreher, Thomas: *Selektive Sedimentation von Feinstschwebstoffen in Wechselwirkung mit wandnahen turbulenten Strömungsbedingungen*, 2008, ISBN 3-933761-71-9

168 Yang, Wei: *Discrete-Continuous Downscaling Model for Generating Daily Precipitation Time Series*, 2008, ISBN 3-933761-72-7

169 Kopecki, Ianina: *Calculational Approach to FST-Hemispheres for Multiparametrical Benthos Habitat Modelling,* 2008, ISBN 3-933761-73-5

170 Brommundt, Jürgen: *Stochastische Generierung räumlich zusammenhängender Niederschlagszeitreihen*, 2008, ISBN 3-933761-74-3

171 Papafotiou, *Alexandros: Numerical Investigations of the Role of Hysteresis in Heterogeneous Two-Phase Flow Systems*, 2008, ISBN 3-933761-75-1

172 He, Yi: *Application of a Non-Parametric Classification Scheme to Catchment Hydrology*, 2008, ISBN 978-3-933761-76-7

173 Wagner, Sven: *Water Balance in a Poorly Gauged Basin in West Africa Using Atmospheric Modelling and Remote Sensing Information*, 2008, ISBN 978-3-933761-77-4

174 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhrmann, Matthias; Schrenk, Volker: *VEGAS-Kolloquium 2008 Ressource Fläche III*, Tagungsband zur Veranstaltung am 01. Oktober 2008 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2008, ISBN 978-3-933761-78-1

175 Patil, Sachin: *Regionalization of an Event Based Nash Cascade Model for Flood Predictions in Ungauged Basins*, 2008, ISBN 978-3-933761-79-8

176 Assteerawatt, Anongnart: *Flow and Transport Modelling of Fractured Aquifers based on a Geostatistical Approach,* 2008, ISBN 978-3-933761-80-4

177 Karnahl, Joachim Alexander: *2D numerische Modellierung von multifraktionalem Schwebstoff- und Schadstofftransport in Flüssen*, 2008, ISBN 978-3-933761-81-1

178 Hiester, Uwe: *Technologieentwicklung zur In-situ-Sanierung der ungesättigten Bodenzone mit festen Wärmequellen,* 2009, ISBN 978-3-933761-82-8

179 Laux, Patrick: *Statistical Modeling of Precipitation for Agricultural Planning in the Volta Basin of West Africa,* 2009, ISBN 978-3-933761-83-5

180 Ehsan, Saqib: *Evaluation of Life Safety Risks Related to Severe Flooding,* 2009, ISBN 978-3-933761-84-2

181 Prohaska, Sandra: *Development and Application of a 1D Multi-Strip Fine Sediment Transport Model for Regulated Rivers,* 2009, ISBN 978-3-933761-85-9

182 Kopp, Andreas: *Evaluation of CO_2 Injection Processes in Geological Formations for Site Screening,* 2009, ISBN 978-3-933761-86-6

183 Ebigbo, Anozie: *Modelling of biofilm growth and its influence on CO_2 and water (two-phase) flow in porous media,* 2009, ISBN 978-3-933761-87-3

184 Freiboth, Sandra: *A phenomenological model for the numerical simulation of multiphase multicomponent processes considering structural alterations of porous media*, 2009, ISBN 978-3-933761-88-0

185 Zöllner, Frank: *Implementierung und Anwendung netzfreier Methoden im Konstruktiven Wasserbau und in der Hydromechanik*, 2009, ISBN 978-3-933761-89-7

186 Vasin, Milos: *Influence of the soil structure and property contrast on flow and transport in the unsaturated zone*, 2010, ISBN 978-3-933761-90-3

187 Li, Jing: *Application of Copulas as a New Geostatistical Tool*, 2010, ISBN 978-3-933761-91-0

188 AghaKouchak, Amir: *Simulation of Remotely Sensed Rainfall Fields Using Copulas*, 2010, ISBN 978-3-933761-92-7

189 Thapa, Pawan Kumar: *Physically-based spatially distributed rainfall runoff modelling for soil erosion estimation*, 2010, ISBN 978-3-933761-93-4

190 Wurms, Sven: *Numerische Modellierung der Sedimentationsprozesse in Retentionsanlagen zur Steuerung von Stoffströmen bei extremen Hochwasserabflussereignissen*, 2011, ISBN 978-3-933761-94-1

191 Merkel, Uwe: *Unsicherheitsanalyse hydraulischer Einwirkungen auf Hochwasserschutzdeiche und Steigerung der Leistungsfähigkeit durch adaptive Strömungsmodellierung*, 2011, ISBN 978-3-933761-95-8

192 Fritz, Jochen: *A Decoupled Model for Compositional Non-Isothermal Multiphase Flow in Porous Media and Multiphysics Approaches for Two-Phase Flow*, 2010, ISBN 978-3-933761-96-5

193 Weber, Karolin (Hrsg.): *12. Treffen junger WissenschaftlerInnen an Wasserbauinstituten*, 2010, ISBN 978-3-933761-97-2

194 Bliefernicht, Jan-Geert: *Probability Forecasts of Daily Areal Precipitation for Small River Basins,* 2011, ISBN 978-3-933761-98-9

195 Hrsg.: Koschitzky, Hans-Peter; Braun, Jürgen: *VEGAS-Kolloquium 2010 In-situ-Sanierung - Stand und Entwicklung Nano und ISCO -,* Tagungsband zur Veranstaltung am 07. Oktober 2010 an der Universität Stuttgart, Campus Stuttgart-Vaihingen*,* 2010*,* ISBN 978-3-933761-99-6

196 Gafurov, Abror: *Water Balance Modeling Using Remote Sensing Information - Focus on Central Asia,* 2010, ISBN 978-3-942036-00-9

197 Mackenberg, Sylvia: *Die Quellstärke in der Sickerwasserprognose: Möglichkeiten und Grenzen von Labor- und Freilanduntersuchungen,* 2010, ISBN 978-3-942036-01-6

198 Singh, Shailesh Kumar: *Robust Parameter Estimation in Gauged and Ungauged Basins,* 2010, ISBN 978-3-942036-02-3

199 Doğan, Mehmet Onur: *Coupling of porous media flow with pipe flow,* 2011, ISBN 978-3-942036-03-0

200 Liu, Min: *Study of Topographic Effects on Hydrological Patterns and the Implication on Hydrological Modeling and Data Interpolation,* 2011, ISBN 978-3-942036-04-7

201 Geleta, Habtamu Itefa: *Watershed Sediment Yield Modeling for Data Scarce Areas,* 2011, ISBN 978-3-942036-05-4

202 Franke, Jörg: *Einfluss der Überwachung auf die Versagenswahrscheinlichkeit von Staustufen,* 2011, ISBN 978-3-942036-06-1

203 Bakimchandra, Oinam: *Integrated Fuzzy-GIS approach for assessing regional soil erosion risks,* 2011, ISBN 978-3-942036-07-8

204 Alam, Muhammad Mahboob: *Statistical Downscaling of Extremes of Precipitation in Mesoscale Catchments from Different RCMs and Their Effects on Local Hydrology,* 2011, ISBN 978-3-942036-08-5

205 Hrsg.: Koschitzky, Hans-Peter; Braun, Jürgen: *VEGAS-Kolloquium 2011 Flache Geothermie - Perspektiven und Risiken,* Tagungsband zur Veranstaltung am 06. Oktober 2011 an der Universität Stuttgart, Campus Stuttgart-Vaihingen*,* 2011*,* ISBN 978-3-933761-09-2

206 Haslauer, Claus: *Analysis of Real-World Spatial Dependence of Subsurface Hydraulic Properties Using Copulas with a Focus on Solute Transport Behaviour,* 2011, ISBN 978-3-942036-10-8

207 Dung, Nguyen Viet: *Multi-objective automatic calibration of hydrodynamic models – development of the concept and an application in the Mekong Delta,* 2011, ISBN 978-3-942036-11-5

208 Hung, Nguyen Nghia: *Sediment dynamics in the floodplain of the Mekong Delta, Vietnam,* 2011, ISBN 978-3-942036-12-2

209 Kuhlmann, Anna: *Influence of soil structure and root water uptake on flow in the unsaturated zone,* 2012, ISBN 978-3-942036-13-9

210 Tuhtan, Jeffrey Andrew: *Including the Second Law Inequality in Aquatic Ecodynamics: A Modeling Approach for Alpine Rivers Impacted by Hydropeaking,* 2012, ISBN 978-3-942036-14-6

211 Tolossa, Habtamu: *Sediment Transport Computation Using a Data-Driven Adaptive Neuro-Fuzzy Modelling Approach,* 2012, ISBN 978-3-942036-15-3

212 Tatomir, Alexandru-Bodgan: *From Discrete to Continuum Concepts of Flow in Fractured Porous Media,* 2012, ISBN 978-3-942036-16-0

213 Erbertseder, Karin: *A Multi-Scale Model for Describing Cancer-Therapeutic Transport in the Human Lung,* 2012, ISBN 978-3-942036-17-7

214 Noack, Markus: *Modelling Approach for Interstitial Sediment Dynamics and Reproduction of Gravel Spawning Fish,* 2012, ISBN 978-3-942036-18-4

215 De Boer, Cjestmir Volkert: *Transport of Nano Sized Zero Valent Iron Colloids during Injection into the Subsurface*, 2012, ISBN 978-3-942036-19-1

216 Pfaff, Thomas: *Processing and Analysis of Weather Radar Data for Use in Hydrology*, 2013, ISBN 978-3-942036-20-7

217 Lebrenz, Hans-Henning: *Addressing the Input Uncertainty for Hydrological Modeling by a New Geostatistical Method*, 2013, ISBN 978-3-942036-21-4

218 Darcis, Melanie Yvonne: *Coupling Models of Different Complexity for the Simulation of CO_2 Storage in Deep Saline Aquifers*, 2013, ISBN 978-3-942036-22-1

219 Beck, Ferdinand: *Generation of Spatially Correlated Synthetic Rainfall Time Series in High Temporal Resolution - A Data Driven Approach*, 2013, ISBN 978-3-942036-23-8

220 Guthke, Philipp: *Non-multi-Gaussian spatial structures: Process-driven natural genesis, manifestation, modeling approaches, and influences on dependent processes*, 2013, ISBN 978-3-942036-24-5

221 Walter, Lena: *Uncertainty studies and risk assessment for CO_2 storage in geological formations*, 2013, ISBN 978-3-942036-25-2

222 Wolff, Markus: *Multi-scale modeling of two-phase flow in porous media including capillary pressure effects,* 2013, ISBN 978-3-942036-26-9

223 Mosthaf, Klaus Roland: *Modeling and analysis of coupled porous-medium and free flow with application to evaporation processes,* 2014, ISBN 978-3-942036-27-6

224 Leube, Philipp Christoph: *Methods for Physically-Based Model Reduction in Time: Analysis, Comparison of Methods and Application*, 2013, ISBN 978-3-942036-28-3

225 Rodríguez Fernández, Jhan Ignacio: *High Order Interactions among environmental variables: Diagnostics and initial steps towards modeling*, 2013, ISBN 978-3-942036-29-0

226 Eder, Maria Magdalena: *Climate Sensitivity of a Large Lake*, 2013, ISBN 978-3-942036-30-6

227 Greiner, Philipp: *Alkoholinjektion zur In-situ-Sanierung von CKW Schadensherden in Grundwasserleitern: Charakterisierung der relevanten Prozesse auf unterschiedlichen Skalen*, 2014, ISBN 978-3-942036-31-3

228 Lauser, Andreas: *Theory and Numerical Applications of Compositional Multi-Phase Flow in Porous Media*, 2014, ISBN 978-3-942036-32-0

229 Enzenhöfer, Rainer: *Risk Quantification and Management in Water Production and Supply Systems*, 2014, ISBN 978-3-942036-33-7

230 Faigle, Benjamin: *Adaptive modelling of compositional multi-phase flow with capillary pressure*, 2014, ISBN 978-3-942036-34-4

231 Oladyshkin, Sergey: *Efficient modeling of environmental systems in the face of complexity and uncertainty*, 2014, ISBN 978-3-942036-35-1

232 Sugimoto, Takayuki: *Copula based Stochastic Analysis of Discharge Time Series*, 2014, ISBN 978-3-942036-36-8

233 Koch, Jonas: *Simulation, Identification and Characterization of Contaminant Source Architectures in the Subsurface*, 2014, ISBN 978-3-942036-37-5

234 Zhang, Jin: *Investigations on Urban River Regulation and Ecological Rehabilitation Measures, Case of Shenzhen in China*, 2014, ISBN 978-3-942036-38-2

235 Siebel, Rüdiger: *Experimentelle Untersuchungen zur hydrodynamischen Belastung und Standsicherheit von Deckwerken an überströmbaren Erddämmen,* 2014, ISBN 978-3-942036-39-9

236 Baber, Katherina: *Coupling free flow and flow in porous media in biological and technical applications: From a simple to a complex interface description,* 2014, ISBN 978-3-942036-40-5

237 Nuske, Klaus Philipp: *Beyond Local Equilibrium — Relaxing local equilibrium assumptions in multiphase flow in porous media,* 2014, ISBN 978-3-942036-41-2

238 Geiges, Andreas: *Efficient concepts for optimal experimental design in nonlinear environmental systems*, 2014, ISBN 978-3-942036-42-9

239 Schwenck, Nicolas: *An XFEM-Based Model for Fluid Flow in Fractured Porous Media*, 2014, ISBN 978-3-942036-43-6

240 Chamorro Chávez, Alejandro: *Stochastic and hydrological modelling for climate change prediction in the Lima region, Peru*, 2015, ISBN 978-3-942036-44-3

241 Yulizar: *Investigation of Changes in Hydro-Meteorological Time Series Using a Depth-Based Approach*, 2015, ISBN 978-3-942036-45-0

242 Kretschmer, Nicole: *Impacts of the existing water allocation scheme on the Limarí watershed – Chile, an integrative approach*, 2015, ISBN 978-3-942036-46-7

243 Kramer, Matthias: *Luftbedarf von Freistrahlturbinen im Gegendruckbetrieb*, 2015, ISBN 978-3-942036-47-4

244 Hommel, Johannes: *Modeling biogeochemical and mass transport processes in the subsurface: Investigation of microbially induced calcite precipitation*, 2016, ISBN 978-3-942036-48-1

245 Germer, Kai: *Wasserinfiltration in die ungesättigte Zone eines makroporösen Hanges und deren Einfluss auf die Hangstabilität*, 2016, ISBN 978-3-942036-49-8

246 Hörning, Sebastian: *Process-oriented modeling of spatial random fields using copulas*, 2016, ISBN 978-3-942036-50-4

247 Jambhekar, Vishal: *Numerical modeling and analysis of evaporative salinization in a coupled free-flow porous-media system,* 2016, ISBN 978-3-942036-51-1

248 Huang, Yingchun: *Study on the spatial and temporal transferability of conceptual hydrological models*, 2016, ISBN 978-3-942036-52-8

249 Kleinknecht, Simon Matthias: *Migration and retention of a heavy NAPL vapor and remediation of the unsaturated zone*, 2016, ISBN 978-3-942036-53-5

250 Kwakye, Stephen Oppong: *Study on the effects of climate change on the hydrology of the West African sub-region*, 2016, ISBN 978-3-942036-54-2

251 Kissinger, Alexander: *Basin-Scale Site Screening and Investigation of Possible Impacts of CO_2 Storage on Subsurface Hydrosystems*, 2016, ISBN 978-3-942036-55-9

252 Müller, Thomas: *Generation of a Realistic Temporal Structure of Synthetic Precipitation Time Series for Sewer Applications*, 2017, ISBN 978-3-942036-56-6

253 Grüninger, Christoph: *Numerical Coupling of Navier-Stokes and Darcy Flow for Soil-Water Evaporation*, 2017, ISBN 978-3-942036-57-3

254 Suroso: *Asymmetric Dependence Based Spatial Copula Models: Empirical Investigations and Consequences on Precipitation Fields*, 2017, ISBN 978-3-942036-58-0

255 Müller, Thomas; Mosthaf, Tobias; Gunzenhauser, Sarah; Seidel, Jochen; Bárdossy, András: *Grundlagenbericht Niederschlags-Simulator (NiedSim3)*, 2017, ISBN 978-3-942036-59-7

256 Mosthaf, Tobias: *New Concepts for Regionalizing Temporal Distributions of Precipitation and for its Application in Spatial Rainfall Simulation,* 2017, ISBN 978-3-942036-60-3

257 Fenrich, Eva Katrin: *Entwicklung eines ökologisch-ökonomischen Vernetzungsmodells für Wasserkraftanlagen und Mehrzweckspeicher,* 2018, ISBN 978-3-942036-61-0

258 Schmidt, Holger: *Microbial stabilization of lotic fine sediments*, 2018, ISBN 978-3-942036-62-7

259 Fetzer, Thomas: *Coupled Free and Porous-Medium Flow Processes Affected by Turbulence and Roughness – Models, Concepts and Analysis*, 2018, ISBN 978-3-942036-63-4

260 Schröder, Hans Christoph: *Large-scale High Head Pico Hydropower Potential Assessment*, 2018, ISBN 978-3-942036-64-1

261 Bode, Felix: *Early-Warning Monitoring Systems for Improved Drinking Water Resource Protection*, 2018, ISBN 978-3-942036-65-8

262 Gebler, Tobias: *Statistische Auswertung von simulierten Talsperrenüberwachungsdaten zur Identifikation von Schadensprozessen an Gewichtsstaumauern*, 2018, ISBN 978-3-942036-66-5

263 Harten, Matthias von: *Analyse des Zuppinger-Wasserrades – Hydraulische Optimierungen unter Berücksichtigung ökologischer Aspekte*, 2018, ISBN 978-3-942036-67-2

264 Yan, Jieru: *Nonlinear estimation of short time precipitation using weather radar and surface observations*, 2018, ISBN 978-3-942036-68-9

265 Beck, Martin: *Conceptual approaches for the analysis of coupled hydraulic and geomechanical processes*, 2019, ISBN 978-3-942036-69-6

266 Haas, Jannik: *Optimal planning of hydropower and energy storage technologies for fully renewable power systems*, 2019, ISBN 978-3-942036-70-2

267 Schneider, Martin: *Nonlinear Finite Volume Schemes for Complex Flow Processes and Challenging Grids*, 2019, ISBN 978-3-942036-71-9

268 Most, Sebastian Christopher: *Analysis and Simulation of Anomalous Transport in Porous Media*, 2019, ISBN 978-3-942036-72-6

269 Buchta, Rocco: *Entwicklung eines Ziel- und Bewertungssystems zur Schaffung nachhaltiger naturnaher Strukturen in großen sandgeprägten Flüssen des norddeutschen Tieflandes*, 2019, ISBN 978-3-942036-73-3

270 Thom, Moritz: *Towards a Better Understanding of the Biostabilization Mechanisms of Sediment Beds*, 2019, ISBN 978-3-942036-74-0

271 Stolz, Daniel: *Die Nullspannungstemperatur in Gewichtsstaumauern unter Berücksichtigung der Festigkeitsentwicklung des Betons*, 2019, ISBN 978-3-942036-75-7

272 Rodriguez Pretelin, Abelardo: *Integrating transient flow conditions into groundwater well protection*, 2020, ISBN: 978-3-942036-76-4

273 Weishaupt, Kilian: *Model Concepts for Coupling Free Flow with Porous Medium Flow at the Pore-Network Scale: From Single-Phase Flow to Compositional Non-Isothermal Two-Phase Flow*, 2020, ISBN: 978-3-942036-77-1

274 Koch, Timo: *Mixed-dimension models for flow and transport processes in porous media with embedded tubular network systems*, 2020, ISBN: 978-3-942036-78-8

275 Gläser, Dennis: *Discrete fracture modeling of multi-phase flow and deformation in fractured poroelastic media*, 2020, ISBN: 978-3-942036-79-5

276 Seitz, Lydia: *Development of new methods to apply a multi-parameter approach – A first step towards the determination of colmation*, 2020, ISBN: 978-3-942036-80-1

277 Ebrahim Bakhshipour, Amin: *Optimizing hybrid decentralized systems for sustainable urban drainage infrastructures planning*, 2021, ISBN: 978-3-942036-81-8

278 Seitz, Gabriele: *Modeling Fixed-Bed Reactors for Thermochemical Heat Storage with the Reaction System $CaO/Ca(OH)_2$*, 2021, ISBN: 978-3-942036-82-5

279 Emmert, Simon: *Developing and Calibrating a Numerical Model for Microbially Enhanced Coal-Bed Methane Production*, 2021, ISBN: 978-3-942036-83-2

280 Heck, Katharina Klara: *Modelling and analysis of multicomponent transport at the interface between free- and porous-medium flow - influenced by radiation and roughness*, 2021, ISBN: 978-3-942036-84-9

281 Ackermann, Sina: *A multi-scale approach for drop/porous-medium interaction*, 2021, ISBN: 978-3-942036-85-6

282 Beckers, Felix: *Investigations on Functional Relationships between Cohesive Sediment Erosion and Sediment Characteristics*, 2021, ISBN: 978-3-942036-86-3

283 Schlabing, Dirk: *Generating Weather for Climate Impact Assessment on Lakes*, 2021, ISBN: 978-3-942036-87-0

284 Becker, Beatrix: *Efficient multiscale multiphysics models accounting for reversible flow at various subsurface energy storage sites*, 2021, ISBN: 978-3-942036-88-7

285 Reuschen, Sebastian: *Bayesian Inversion and Model Selection of Heterogeneities in Geostatistical Subsurface Modeling*, 2021, ISBN: 978-3-942036-89-4

Die Mitteilungshefte ab der Nr. 134 (Jg. 2005) stehen als pdf-Datei über die Homepage des Instituts: www.iws.uni-stuttgart.de zur Verfügung.