The Prior and the Likelihood:

Accommodating Uncertain Priors and Model Defects in Groundwater Modelling for Decision Support

By

Tomas Opazo (MSc)

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy

College of Science and Engineering

6 Jun 2025

Contents

List of Figures

List of Tables

Summary

Quantification of uncertainty is a fundamental task in groundwater modelling to ensure reliable predictions of aquifer behaviour, facilitate effective water resource management, and support robust decision-making. However, misspecification of prior uncertainties in model parameters, along with model imperfections, can introduce bias in decision-critical predictions and lead to an underestimation of predictive uncertainty. Consequently, the reliability of the predictions and their associated uncertainties can be compromised, which can result in suboptimal management decisions. This research aims to address these challenges by developing and evaluating methodologies that explicitly account for potential misspecification in the prior characterization of parameter uncertainties and incorporate estimates of model structural errors into both history matching and predictive uncertainty quantification processes, ultimately enhancing the reliability and robustness of groundwater models.

This thesis pursues four primary objectives: (1) to perform a focused review of existing history matching and uncertainty quantification methods, (2) to develop a methodology for addressing prior-data conflict and updating prior uncertainties using empirical Bayesian inference, (3) to introduce a hierarchical parametrization scheme to represent nonstationary priors, and (4) to develop techniques for quantifying and integrating model structural errors into history matching and predictive uncertainty quantification.

The research employs advanced modelling techniques, including empirical Bayesian inference to update prior uncertainties, a hierarchical two-level parametrization scheme that integrates spatially variable geostatistical hyperparameters with spatially distributed parameters, and a structural error model that is incorporated into the history matching and predictive uncertainty quantification process. These methodologies are tested using synthetic two-dimensional groundwater models that reflect real-world situations, which results are evaluated based on their ability to reduce predictive bias and provide more reliable uncertainty estimates.

Key findings from this research include the following: (1) acknowledging and updating uncertain priors using empirical Bayesian inference yields conservative and robust predictive uncertainty estimates, (2) the hierarchical parametrization scheme effectively manages nonstationary priors, thereby allowing for a more realistic heterogeneity representation and achieving both reasonable fits to the data and acceptable predictive uncertainty es-

timates, and (3) incorporating structural errors into history matching reduces predictive bias and provides more conservative uncertainty estimates. Based on these findings, it is demonstrated that the proposed methodologies improve the reliability and robustness of groundwater modelling for decision support.

This thesis advances both history matching and predictive uncertainty quantification in groundwater modelling by providing practical frameworks for quantifying and integrating uncertainties related to prior parameter uncertainties and model structural defects. The contributions encompass novel methodologies for updating prior uncertainties via empirical Bayesian inference, managing nonstationary priors, and quantifying and integrating structural errors into the history matching and predictive uncertainty quantification process. These findings have substantial practical implications for improving the reliability of groundwater modelling for decision support, and also lead to new research directions, some of which are identified and recommended for future research.

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

06/06/2025

.....

Tomas Opazo

Acknowledgments

This research was undertaken as part of the Groundwater Modelling Decision Support Initiative (GMDSI), funded by BHP and Rio Tinto, and coordinated by the National Centre for Groundwater Research and Training (NCGRT), from Flinders University. I greatly appreciate the financial support provided, which allowed me to focus on my research and to attend conferences and workshops.

I would like to thank my supervisor, Dr. John Doherty, for his technical guidance, support, and encouragement throughout my PhD. These years have been a period of intense learning and personal development, and I am grateful for the opportunity to work with him. To this day, I remain impressed by his ability to generate new ideas and articulate them clearly and concisely. I still have much to learn. I would also like to thank my former co-supervisor, Dr. Craig Simmons, and my current co-supervisor, Dr. Ilka Wallis, for their support and advice. Finally, I extend my sincere appreciation to Dr. Catherine Moore and Dr. Jeremy White for their insights as assessors of my PhD journey. Their enthusiasm for research and their constructive feedback have been invaluable.

Preface

Groundwater modelling practitioners operate at the intersection of hydrogeology, numerical methods, and decision-making. Knowledge of hydrogeology is essential to understand the physical processes that govern the flow and transport of water and contaminants in the subsurface. Although understanding the mathematical proofs behind the numerical methods used in hydrogeology is not essential, gaining insight into how the prior is integrated with the model and data for predictions requires familiarity with the underlying concepts and assumptions of existing methods for history matching and uncertainty quantification Therefore, some level of mathematical knowledge, especially in probability theory, statistics, and linear algebra, is required to have the ability to use and customize the numerical methods. Geoscientists with expertise in both hydrogeological and computational approaches are exceptionally rare yet critically needed to adopt a holistic approach to today's pressing water resource challenges. With a background in geology and hydrogeology, I have developed a strong interest in mathematics and computer programming to support history matching and uncertainty quantification in groundwater modelling. While I am not a mathematician or a computer scientist, I hope that my skills can help bridge the gap created by the shortage of geoscientists capable of integrating hydrogeology, numerical methods, and decision-making.

Before starting this PhD, I naively believed that studying uncertainty would lead me down a nebulous path of unknowns and subjectivity, ultimately revealing profound insights into predictive uncertainty in groundwater modelling. I was completely mistaken. Instead, I discovered that uncertainty quantification is firmly grounded in mathematics, with subjectivity arising only from the modeller's assumptions and the methods employed. Thus, investigating these assumptions not only offers personal perspective but also provides valuable insights and, hopefully, contributes to advancing the field. This thesis focuses on critically examining key assumptions: the knowledge of the prior distribution of model parameters and the implicit belief that the model is a perfect representation of the real system. By challenging these assumptions, this research aims to enhance understanding of their influence on history matching and uncertainty quantification, ultimately contributing to more robust groundwater modelling approaches.

Chapter 1

Introduction

1.1 Research Motivation

Decisions on water management actions rely on the uncertain future behaviour of groundwater systems, which are increasingly affected by human activities and climate change, often under conditions that differ from historical trends. One or more predictions of interest may be required for a specific management action, where certain outcomes could result in environmental, social, or economic costs. Decision-makers must assess the risk level associated with implementing a management action, while accounting for an acceptable level of risk tolerance.

Risk is quantified by combining the probability of occurrence of identified unwanted events (or probability of failure), with the associated consequences (Aven, 2010). Forecast modelling, and particularly groundwater modelling, facilitates the evaluation of the likelihood of the occurrence of certain unwanted events (Freeze et al., 1990). A typical example of an unwanted event is the potential exceedance of predefined thresholds for groundwater levels, which may negatively affect water availability to groundwater-dependent ecosystems, under a proposed groundwater extraction strategy. As a key component of the scientific method, groundwater modelling for decision support is a process of hypothesis testing. To reject a hypothesis, the probability of failure must be estimated. Therefore, model predictions must be accompanied by uncertainty estimates (Doherty and Simmons, 2013; Doherty and Moore, 2020). Predictive uncertainty quantification is then a critical component of the decision-making process, as it provides the basis for assessing the risk associated with the management action. The estimation of predictive uncertainty is the result of history matching, the process of data assimilation that allows the use of a numerical model to both represent the historical behaviour of a system and to produce a feasible range of predictions.

Both history matching and predictive uncertainty quantification are embedded within a Bayesian framework, whereby the current understanding of a range of possible future

behaviours of a system is updated as new information becomes available. In this framework, the probabilities of possible outcomes are updated from a prior to a posterior state. If a physical system is represented by a model, the physical quantities that characterize the system can be aggregated into model parameters. Then, the possible outcomes for future system behaviour are generated by applying the model to these parameters and to external model forcings, resulting in model predictions. This is a mapping process, akin to the application of a function to a set of inputs to produce a set of outputs. In this sense, model predictions are derived through a filtering process that maps a set of possible parameter values and model forcings onto a set of possible predictions. If a subset of these predictions corresponds to available measurements of the system's behaviour, the prior parameter uncertainties represented by probability density functions, can be updated to new probability distributions by propagating samples through the model, thereby generating simulated quantities consistent with the measurements. Consequently, the range of possible outcomes of future system behaviour is refined, based on their linkage to the measurements via the model, thereby reducing predictive uncertainty. Based on the mapping process described, it becomes clear that parameters only make sense in the context of the model, and the model only makes sense in the context of the data (Gelman et al., 2017).

With the consolidated usage of history matching and uncertainty quantification techniques in aquifer and reservoir modelling, new challenges have emerged, especially related to the reliability of model predictions, which is the cornerstone of informed decision-making (Caers, 2018). For model predictions to be reliable, two main requirements must be met. First, model uncertainty quantification should provide realistic uncertainty estimates. Second, model predictions should exhibit minimal departure from the true, although unknown, future system behaviour. These two metrics are related to the linkage between parameters, the model, and data. In other words, they are associated with the definition of the prior and the likelihood, and the combination of the two, which are the two components of Bayes' equation. How these components are defined and combined lead to a handful of challenges that must be addressed to provide reliable predictions and uncertainty estimates. Some of these challenges motivate the research presented in this thesis.

Defining the prior probability distribution of model parameters is the first problem. When the definition of model parameters is motivated by physical properties, an informative probability distribution can be elicited to represent the uncertainty in the parameter values prior to data assimilation. An informative prior merely indicates that the distribution is not uniform, but it is not necessarily accurate. Given that in this case the probability distribution is based on expert knowledge, it is a subjective task that depends on personal judgement, and is therefore prone to biases and errors (Sprenger, 2018). This is particularly true for subsurface hydraulic properties, whose heterogeneity, anisotropy,

and spatial discontinuity make them difficult to characterize. In the subsurface, materials may be intersected by linear or sinuous features of structural or erosional origin that may expedite or impede the flow of water and contaminants, with their locations, properties, and even existence only vaguely known. Therefore, the null hypothesis that the prior is incorrect cannot be rejected. Instead, it is important to evaluate potential actions to update the prior if evidence suggests its inaccuracy. Furthermore, in the absence of certainty regarding the correct prior, the prior must be treated as uncertain. In this way, a greater potential to represent heterogeneity informed or not by data is achieved, mitigating, to the extent possible, the underestimation of predictive uncertainty (Doherty, 2015; White et al., 2014). Embracing uncertainty in the prior requires exploring methods that integrate this source of uncertainty into the history-matching and predictive uncertainty quantification process, facilitating the mapping from parameters to predictions while incorporating data constraints.

Second, even with a perfectly specified prior, the action of the model on its parameters may yield incorrect predictions. This is because the model inevitably represents a simplified abstraction of the real system; hence, it is inherently imperfect. Any numerical model necessarily omits some details of the system that may be critical to linking parameters, the model, and the data. Thus, mapping perfectly-defined model parameters through an imperfect model does not necessarily guarantee reliable predictions. Consequently, even a perfect prior may require modification to be effectively mapped through the model to generate reliable simulations of future system behaviour (Mathews and Vial, 2017), resulting in a degree of parameter abstraction that may be challenging to interpret and communicate. A more honest approach is to acknowledge the model's imperfections, even if they cannot be precisely identified, accepting that the data can only be partially assimilated, to a level of noise not necessarily commensurate with the measurement errors in the data. In other words, the posterior range of possible outcomes for past and future system behaviour, as derived from the model, may be inadequately linked to the data due to the model's inability to represent the system accurately. A better but more complicated task is to identify and incorporate the model's structural errors into the history matching and uncertainty quantification process, to improve the reliability of model predictions. Model structural error (Beven, 2005) is a broad term that includes all imperfections in the model that may lead to discrepancies between observed data and model outputs beyond what can be attributed to measurement error. Measurement errors can be characterized by a probability distribution, which typically assumes uncorrelated Gaussian noise. Structural error is more challenging to characterize, as it may be correlated, non-Gaussian, and may vary spatially and temporally. Some effort has been made to address this issue in the literature, including the generation of covariance matrix of structural error prior to perform model calibration (Cooley, 2004; Cooley and Christensen, 2006) and updating the initial uncorrelated covariance matrix of measurement noise during the history match-

ing process (Oliver and Alfonzo, 2018; Alfonzo and Oliver, 2020; Evensen, 2021; Lu and Chen, 2020). However, additional research is needed to develop more robust methods for estimating and incorporating model structural error into the history matching and uncertainty quantification process. This constitutes the second challenge motivating this work.

This thesis focuses on groundwater modelling to support decision-making. In this context, most—if not all—of the problems presented in this document relate to predictions of management interest and their associated uncertainties. The research is developed within a Bayesian framework, where the prior and the likelihood are the main components of the uncertainty quantification process. The fundamental assumptions underlying these two Bayesian components are analysed and discussed in the context of predictive uncertainty quantification. New methodologies are proposed that allow updating the prior when evidence suggests its inaccuracy, and that accommodate uncertain priors and model defects by explicitly integrating these issues into the uncertainty quantification process. The proposed methods are tested using simple numerical examples to enhance the understanding of the problem and the efficacy of the proposed solutions. These examples are, however, inspired by real problems faced by the author during his professional career. Such problems include the estimation of drawdown in aquifers in locations where groundwater sustains ecosystems, lithium extraction from brines in salars, and groundwater inflows to open pit mines. Thus, in addition to the conclusions related to the methods themselves, further insights can be gained regarding the utility of groundwater modelling and uncertainty quantification in supporting decision-making in real-world problems, and the potential for improvement by integrating new methods—such as the ones proposed in this thesis—into routine decision-support modelling.

1.2 Research Objectives

The primary objectives of this thesis centre on the prior and the likelihood. They are as follows:

- To perform a focused review of existing history matching and uncertainty quantification methods.
- To assess the importance of critically examining prior and likelihood assumptions in the uncertainty quantification process of groundwater modelling.
- To identify approaches for evaluating prior-data compatibility, and integrating uncertain priors in history matching.
- To analyse the impact of model defects and explore methods for incorporating them into the history matching and uncertainty quantification process.

Additionally, the following secondary objectives are pursued:

- To investigate the assumptions, benefits, and limitations of the existing history matching methods.
- To explore strategies for improving existing methods to address challenges posed by nonlinearities.
- To evaluate the performance of these new strategies in history matching and predictive uncertainty quantification.

1.3 Research contributions

The main contributions of this thesis are as follows:

- A critical and focused review of existing methods for history matching and uncertainty quantification in groundwater modelling and related fields, presented within a unified mathematical framework. This unification ensures consistency in the terminology and formulation, facilitating systematic comparisons. Two numerical examples are used to illustrate the application of the methods and to compare their performance. Notably, the subspace ensemble smoother (SEnRML) is applied to groundwater modelling for the first time, representing an additional contribution of this thesis.
- The integration and testing of ensemble-based methods with localization and local updating to address challenges posed by highly nonlinear problems.
- Development of a novel method for updating the prior probability distribution of parameters using calibration results, introduced as a form of empirical prior Bayesian inference.
- Extension of the non-centred parameterization method proposed by Oliver (1995) to accommodate nonstationary priors, thereby expanding its applicability.
- Development a new methodology for estimating model structural error and incorporating it into the history matching and uncertainty quantification process, thereby improving the reliability of model predictions.

1.4 Thesis Outline

The thesis is structured as follows:
Chapter 2 is a review of the literature on the topic of uncertainty quantification in groundwater modelling and related fields, presenting the state of the art and a comparison of methods using numerical examples. Chapter 3 introduces a new method for updating the prior probability distribution of the parameters in a groundwater model using calibration results, within an empirical Bayesian framework. Chapter 4 presents an extended version of the non-centred parameterization method proposed by Oliver (1995) to accommodate nonstationary priors in the history matching process. Chapter 5 presents a new methodology for estimating model structural error, and incorporates it into the history matching and uncertainty quantification process. Finally, Chapter 6 summarizes the main findings and suggests future research directions.
This thesis is written as a collection of four manuscripts for publication in peer-reviewed journals. The manuscripts correspond to chapters Chapter 2, Chapter 3, Chapter 4, and Chapter 5. Consequently, this work is presented concisely and in a focused manner to provide a clear and coherent narrative, and each of the thesis chapters can be read independently as a stand-alone piece of research.
The four manuscripts that are intended for submission to peer-reviewed journals are the following:

- **Paper 1:** *Opazo, T., Doherty, J. Selected Methods for History Matching and Uncertainty Quantification in Groundwater Modelling: A Focused Review and Comparison. To be submitted to Computational Geosciences.*
- **Paper 2:** *Opazo, T., Doherty, J. Prior Inference from Groundwater Model Calibration: Empirical Bayesianism to Improve Predictive Uncertainty. To be submitted to Journal of Hydrology.*
- **Paper 3:** *Opazo, T., Doherty, J. Accommodating Uncertain and Nonstationary Priors in History Matching and Predictive Uncertainty Quantification for Groundwater Models. To be submitted to Advances in Water Resources.*
- **Paper 4:** *Opazo, T. Quantifying Model Structural Errors in History Matching and Predictive Uncertainty Quantification in Groundwater Modelling. To be submitted to Water Resources Research.*

Chapter 2

Selected Methods for History Matching and Uncertainty Quantification in Groundwater Modelling: A Focused Review and Comparison

Author contributions

T. Opazo: Conceptualization 100%, Realization 100%, Writing 100%.

Manuscript in preparation for submission to Computational Geosciences: Opazo, T. Selected Methods for History Matching and Uncertainty Quantification in Groundwater Modelling: A Focused Review and Comparison.

Abstract

This work presents a focused review and a unified mathematical framework for selected methods used in history matching and uncertainty quantification in groundwater modelling. The methods discussed include Markov Chain Monte Carlo (MCMC), Iterative Ensemble Smoothers (IES), Data Space Inversion (DSI), and regularized inversion. The review highlights the theoretical foundations, advantages, and limitations of each method, providing a consistent comparison. Additionally, localization techniques are examined to address challenges associated with high-dimensional problems in ensemble methods. Numerical examples, including a simple one-parameter nonlinear problem and a complex one-dimensional unsaturated flow problem, are used to evaluate the performance of these selected methods. The results confirm that ensemble methods are efficient but

often exhibit slow convergence in nonlinear problems. Among the ensemble methods, the Levenberg-Marquardt Ensemble Randomized Maximum Likelihood (LM-EnRML) method shows the best data fitting performance but is prone to predictive bias. This work presents the first application of the Subspace Ensemble Randomized Maximum Likelihood (SEnRML) method to groundwater modelling. The method, without the best data fit, provides reasonable estimates of predictive uncertainty.

2.1 Introduction

In groundwater modelling for decision support, it is essential to determine the feasible range of n model parameters $\mathbf{x} \in \mathbb{R}^n$ and the corresponding model predictions—dependent on these parameters—that lie within the bounds of hydrogeological knowledge and conform to the observed data (i.e., the measured state of the system). Here the term 'model parameters' is used broadly, as a 'parameter' can refer to model hydraulic properties, initial conditions, boundary conditions, or any other uncertain input of the model. It is acknowledged that, in classical system theory, parameters are typically defined as time-invariant quantities. However, in the context of Bayesian inference for environmental modelling, it is common to adopt a broader definition and treat all uncertain model inputs—including temporally variable quantities such as initial and boundary conditions—as parameters to be inferred. In this sense, there is no explicit separation between system state and model parameters (in the classical sense) if both quantities are uncertain. Fundamentally, this approach constitutes a Bayesian inference problem, where the prior knowledge of the model parameters is updated as data are assimilated, represented in a probabilistic way through Bayes' equation:

$$f(\mathbf{x}|\mathbf{d}) = \frac{f(\mathbf{d}|\mathbf{x})f(\mathbf{x})}{f(\mathbf{d})}, \tag{2.1}$$

where $f(\mathbf{x}|\mathbf{d})$ represents the posterior probability density function (pdf) of the model parameters $\mathbf{x}$ given the data $\mathbf{d} \in \mathbb{R}^m$, $f(\mathbf{d}|\mathbf{x})$ the likelihood of the system state given the model parameters, $f(\mathbf{x})$ the prior pdf of the model parameters, and $f(\mathbf{d})$ the marginal pdf of the data. This pdf acts as a normalization factor so that the posterior pdf integrates to one. The immediate goal of history matching is to obtain an approximate representation of the posterior probability of model parameters given the data $f(\mathbf{x}|\mathbf{d})$. The ultimate objective in groundwater modelling for decision support is to quantify the posterior uncertainty of one or several predictions of interest, by applying the posterior probability of the model parameters to a forward model.

As Evensen et al. (2022) points out, Bayes' equation is a mathematical framework for updating the prior knowledge of the model parameters (or states) as data are assimilated; therefore it is a forward problem rather than an inverse problem. Bayes' equation provides

an intuitive approach to data assimilation, as it is based on the idea of updating knowledge as new data become available. In theory, a correct sampling of the posterior pdf of the model parameters can be obtained by random sampling of the prior and likelihood distributions followed by applying Bayes' equation. This can be done albeit inefficiently, through rejection sampling or sequential realization (Tarantola, 2005). However, these methods are only feasible for low-dimensional problems. A more efficient strategy to sample the posterior pdf is to use Markov Chain Monte Carlo (MCMC) methods, such as the famous random walk Metropolis (RWM) algorithm (Metropolis et al., 1953). However, MCMC methods are also computationally expensive and are rarely feasible for high-dimensional problems in real-world applications.

Ensemble methods are data assimilation and history matching techniques, all derived from Bayes' equation, that use a finite number of parameter and model realizations to approximate the posterior pdf by model inversion, making these methods much more computationally efficient than MCMC methods. Iterative ensemble smoothers (IES) are a group of methods developed during the last 15 years in the petroleum community, originating from a series of filtering and smoothing methods developed in the atmospheric and oceanographic sciences. These include the Kalman filter (KF) (Kalman, 1960), extended Kalman filter (EKF), ensemble Kalman filter (EnKF) (Evensen, 1994), ensemble smoother (ES) (van Leeuwen and Evensen, 1996), and ensemble Kalman smoother (EnKS) (Evensen and van Leeuwen, 2000). A review of these methods is outside the scope of this work, and the reader is referred to Evensen et al. (2022) for a comprehensive review. All IES methods have their roots in the randomized maximum likelihood (RML) approach (Kitanidis, 1995; Oliver et al., 1996), in which the posterior pdf of the model parameters is approximated by an ensemble sampled from the prior pdf and optimized through an inversion process that minimizes a cost function. The RML method was introduced in the petroleum community by Gu and Oliver (2007), who developed an iterative form of the ensemble Kalman filter called ensemble randomized maximum likelihood filter (EnRML) using an average sensitivity matrix estimated from the ensemble when updating the model parameters in the Gauss-Newton inversion process. A batch and iterative ensemble smoother version of EnRML with Gauss-Newton (GN) formulation (GN-EnRML) was introduced by Chen and Oliver (2012), and a Levenberg-Marquardt version called LM-EnRML presented by Chen and Oliver (2013) revolutionized the petroleum and groundwater community. Raanes et al. (2019) revised the iterative ensemble smoother formulation of Chen and Oliver (2012, 2013) and proposed a subspace version of the EnRML (here called SEnRML), further developed by Evensen et al. (2019), where instead of searching for the solution in the full parameter space, the inversion is carried out in the ensemble space by solving for weighted combinations of the initial parameter ensemble members.

Although ensemble methods are computationally efficient, they have limitations, such as

the potential for parameter and predictive bias, and parameter ensemble collapse. These issues can theoretically occur even when using a perfect model.

In some cases, it is valuable to look for a minimum error variance solution to the inverse problem, as is done with regularized inversion methods. Doherty (2015) presented a comprehensive review of the theory and practice of regularized inversion, predictive error variance, and linear and nonlinear uncertainty analysis in the context of groundwater modelling. The algorithms described by Doherty (2015) are implemented in the PEST suite of software (Doherty, 2023), which is widely used in the groundwater community. Parameter estimation performed with PEST searches for the minimum error variance solution to the inverse problem. As such, parameter and predictive uncertainties are not part of the inversion process but can be calculated after using linear or nonlinear approaches.

A recently developed method called Data Space Inversion (DSI) (Sun and Durlofsky, 2017; Sun et al., 2017) performs inversion in the data space (model output space), using a number N of model output realizations. This approach avoids the need to update the model parameters, and therefore the requirement of estimating a sensitivity matrix, as in IES methods and regularized inversion. The cost of DSI is only the number of prior model runs, N. History matching the DSI model incurs a minimum cost, as 'the model' is a linear correlation between the model outputs that pertain to the past and those that pertain to the future. By constraining model outputs to observed data, the method can be used to estimate the posterior pdf of model predictions, and therefore their uncertainties.

In this work a non-exhaustive selection of existing methods of uncertainty quantification for groundwater modelling, reservoir engineering, hydrology, and geophysics are reviewed. They range from Markov Chain Monte Carlo (MCMC), passing through ensemble methods (Iterative Ensemble Smoothers) and regularized inversion, and ending with Data Space Inversion (DSI). Some of the methods that exist but are not discussed here (at least not in detail) include the ensemble Kalman filter (EnKF) (Evensen, 1994), ensemble smoother multiple data assimilation (ES-MDA) (Emerick and Reynolds, 2013), hybrid iterative ensemble smoother (hybrid IES) (Oliver, 2022), weighted randomized maximum likelihood (weighted RML) (Ba and Oliver, 2023), and importance weighting hybrid iterative ensemble smoother (Ba and Oliver, 2024), to name a few. The equations here presented are general, with a focus on the understanding of the methods rather than providing detailed mathematical proofs. Therefore, some aspects of the methods are simplified or omitted for the sake of clarity. In these cases, the reader is referred to the original papers for further details. Localization methods (Luo et al., 2018; Luo and Bhakta, 2020; Luo et al., 2023; Silva Neto et al., 2021) and local updating (Zhang et al., 2018) are also discussed, as they are important for the application of ensemble methods in high-dimensional problems. To date, there is no groundwater literature that presents

existing methods in a unified mathematical framework. Therefore, it is hoped that this chapter provides a clear summary of existing knowledge and will serve as a reference for future research in the field of uncertainty quantification.
The chapter is organized as follows: First, a description of the methods is presented, including their mathematical formulation, advantages, and limitations. The definitions and equations are presented in a unified mathematical framework, allowing for a consistent comparison of the methods. Second, some methods are compared using two illustrative examples. The first example is a simple one-parameter problem with a polynomial equation used as a model. This example is presented by Chen and Oliver (2013) to illustrate the behaviour of iterative ensemble smoothers with a highly nonlinear problem. The second example is a one-dimensional highly-parameterized and nonlinear unsaturated flow problem. After presenting the results of these examples, a discussion highlights the advantages and limitations of the methods, and the implications of the results for groundwater modelling for decision support. The chapter ends with conclusions and a summary of its main findings.

2.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms that generate samples from the posterior distribution using a Markov chain that randomly moves through the parameter space, repeatedly visiting solutions with frequencies consistent with a stationary distribution (Vrugt, 2016). Among all practical methods of posterior sampling, MCMC methods are deemed as the most reliable in terms of posterior pdf representation. However, they are computationally expensive and not commonly used for applied cases, as they require many samples to obtain convergence to the true posterior pdf. In this work, whenever possible, MCMC is used as the reference method to sample the true posterior pdf and compare with other methods.

2.2.1 Random walk Metropolis (RWM) algorithm

The RWM algorithm (Metropolis et al., 1953) is one of the original MCMC methods. It starts with an initial state of the Markov chain, $\mathbf{x_t}$, and generates a new candidate state, $\mathbf{x_p}$,using a proposal distribution (also called the jumping distribution) centred at the current state $\mathbf{x_t}$. A probability of acceptance ($P_A(\mathbf{x_p})$) is calculated as the minimum between one and the ratio of the posterior pdf of the new proposed state $p(\mathbf{x_p})$ to the posterior pdf of the current state $p(\mathbf{x_t})$ (assuming the proposal distribution is symmetric), formulated as follows:

$$P_A(\mathbf{x_p}) = min(1, \frac{p(\mathbf{x_p})}{p(\mathbf{x_t})}). \tag{2.2}$$

The new candidate state is accepted if the acceptance probability P_A is greater than a random number drawn from a uniform distribution between 0 and 1. The process is repeated until a sufficient number of samples are obtained.

Generally, a burn-in period is used to allow the Markov chain to reach the stationary distribution before samples are collated. Also, the process can be parallelized using more than one Markov chain. This can be particularly useful when sampling multimodal and multidimensional distributions, allowing the use of diagnostics to check the convergence of the Markov Chain, such as the Gelman-Rubin statistic (Gelman and Rubin, 1992).

The choice of the proposal distribution is crucial for the efficiency of the RWM method. Generally, a multivariate normal distribution centred at the current state is used as the proposal distribution, with a unit covariance matrix scaled by the scaling factor $s_d = 2.38^2/d$ (Roberts et al., 1997), where d is the number of dimensions of the problem. As Vrugt (2016) points out, there are several factors that impact the efficiency of the RWM method, the most important being the choice of the proposal distribution, specifically its scale and orientation. If the proposal distribution is too wide, meaning the scale is too large, the Markov chain will take a long time to explore posterior parameter space, as the acceptance probability will be low. On the contrary, if the proposal distribution is too narrow, the Markov chain will evolve through very short steps leading to long convergence times. Naturally, the RWM method is also limited by the dimensionality of the problem, as the acceptance probability decreases with the number of dimensions. In these cases, the orientation of the proposal distribution becomes important, as the Markov Chain will take longer to explore the parameter space if the proposal distribution is not aligned with the principal directions of the posterior pdf (which is unknown for obvious reasons).

As a result of the limitations of the RWM method described above, and given that the optimal shape of the proposal distribution cannot be known a priori, several adaptive MCMC methods have been developed for single chain (such as adaptive proposal (AP) from Haario and Saksman (1998), adaptive Metropolis (AM) from Haario et al. (2001), delayed rejection adaptive metropolis (DRAM) from Haario et al. (2006)) and multiple-chain (such as the Differential Evolution-Markov Chain (DE-MC) from Ter Braak (2006), and the Differential Evolution Adaptive Metropolis (DREAM) from Vrugt et al. (2008, 2009)). These methods share the common approach of modifying the scale and orientation of the proposal distribution during the burn-in period, using information from the sample history.

2.2.2 Differential Evolution Adaptive Metropolis (DREAM) algorithm

One of the most successful multiple-chain adaptive MCMC methods is the Differential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al., 2008, 2009). The

method is originally based on the Differential Evolution Markov Chain (DE-MC) method developed by Ter Braak (2006). DE-MC uses a population $\mathbf{X}$ of N chains from the past state of the Markov Chain to generate a new proposal state for each chain i, $\mathbf{x_p^i}$, by applying differential evolution:

$$\mathbf{x_p}^i = \mathbf{x_t}^i + \gamma(\mathbf{X^a} - \mathbf{X^b}) + \mathbf{e}, \tag{2.3}$$

where $\mathbf{X^a}$ and $\mathbf{X^b}$ are two different chains randomly selected from the population $\mathbf{X}_{-i}$ (excluding chain i),γ is the jump rate (with an optimal value of $2.38/\sqrt{2d}$), and $\mathbf{e}$ is a random vector drawn from a normal distribution with zero mean and small variance. The state update using Equation 2.3 is also known as parallel direction sampling (ter Braak and Vrugt, 2008). The jump rate is set to one ($\gamma = 1$) every 10 generations to allow jumps between disconnected modes of the posterior pdf. The proposal is accepted if the probability of acceptance calculated from Equation 2.2 is greater than a random number drawn from a uniform distribution between 0 and 1. Compared to the RWM method, DE-MC does not require a proposal distribution, as the proposal is generated from the population of chains. In practice, DE-MC is not efficient for high-dimensional problems, as it requires $N = 2d$ chains to properly sample the posterior (Ter Braak, 2006). Vrugt et al. (2008, 2009) extended the DE-MC method to improve convergence to the target distribution and reduce the number of chains, by incorporating the use of multiple (more than two) chain pairs, self-adaptive randomized subspace sampling, and outlier chain detection. Randomized subspace sampling is performed with the help of a crossover operator (a detailed explanation can be found in Vrugt et al. (2009)) that constantly includes new directions where the chain can jump outside the current subspace. Crossover probabilities are tuned adaptively during the burn-in period. This allows the use of $N < d$ chains to sample the posterior pdf, compared to the $N = 2d$ required when using DE-MC. A population subset $\mathbf{A}$ of size d^* (reduced parameter dimensionality) is randomly selected from the population $\mathbf{X}$, and the proposal increment for each chain $\mathbf{dx}^i$, or jump, is calculated as follows:

$$\begin{aligned} \mathbf{dx}_A^i &= (1_{d^*} + \lambda_{d^*})\,\gamma(\delta, d^*) \sum_{j=1}^{\delta} \left(\mathbf{x}_A^{a_j} - \mathbf{x}_A^{b_j}\right) + \mathbf{e}_{d^*} \\ \mathbf{dx}_{\neq A}^i &= 0, \end{aligned} \tag{2.4}$$

where $\mathbf{x}_A^{a_j}$ and $\mathbf{x}_A^{b_j}$ are two different chains randomly selected from the matrix $\mathbf{A}_{-i}$, δ is the maximum number of chain pairs, and λ_{d^*} is a sample from a uniform distribution bounded by $[-b, b]$ (b being small compared to the target distribution). During each step and for each chain, the number of chain pairs is randomly selected from the set $[1, 2, \ldots, \delta]$. The jump rate γ is calculated for every subspace sampling, as $2.38/\sqrt{2\delta d^*}$

and is set to one every 5 generations. The new proposal state, $\mathbf{x_p^i}$, is calculated as the sum of the current state $\mathbf{x_t^i}$ and the proposal increment $\mathbf{dx}^i$. Convergence is checked using the Gelman-Rubin statistic (Gelman and Rubin, 1992), using the last 50% of the samples.

The DREAM algorithm has been extended since its creation to improve efficiency, including DREAM$_{(ZS)}$ and MT-DREAM$_{(ZS)}$ (Laloy and Vrugt, 2012). In DREAM$_{(ZS)}$, the chain population is replaced by a matrix $\mathbf{Z}$ that contains thinned history of past states of each of the N chains. The proposal increment is calculated as follows:

$$\begin{aligned} \mathbf{dx}_A^i &= (1_{d^*} + \lambda_{d^*})\,\gamma(\delta, d^*) \sum_{j=1}^{\delta} \left(\mathbf{z}_A^{a_j} - \mathbf{z}_A^{b_j}\right) + \mathbf{e}_{d^*} \\ \mathbf{dx}_{\neq A}^i &= 0, \end{aligned} \tag{2.5}$$

where $\mathbf{z}_A^{a_j}$ and $\mathbf{z}_A^{b_j}$ are two different chains randomly selected from the matrix $\mathbf{A}_{-i}$. In this case $\mathbf{A}_{-i}$ is a random subspace sample (reduced parameter dimensionality d^*) of the matrix $\mathbf{Z}$. The recommended initial size of the matrix $\mathbf{Z}$ and the thinning rate (saving frequency of samples to the $\mathbf{Z}$) are $10d$, and 10, respectively. Although using past states violates the Markov property of the chain (this is minimized as $\mathbf{Z}$ grows, as shown by ter Braak and Vrugt (2008)), it has been shown that the method is extremely efficient only requiring a few chains (typically $N = 3$ will be sufficient). The "S" in DREAM$_{(ZS)}$ stands for "Snooker", because this algorithm also implements Snooker updates in combination with parallel direction updates to diversity the jumping possibilities (ter Braak and Vrugt, 2008). For each generation iteration and for each chain, a decision is made to use the Snooker update, with a probability of 0.1. Laloy and Vrugt (2012) extended the DREAM$_{(ZS)}$ algorithm to the MT-DREAM$_{(ZS)}$ algorithm, implementing multiple-try Metropolis sampling to improve efficiency for very high-dimensional problems.

The family of DREAM algorithms has been implemented in MATLAB (Vrugt, 2016) and Python (pyDREAM) (Shockley et al., 2017), to name a few. In this work, DREAM$_{(ZS)}$ implemented in pyDREAM is used to sample the posterior pdf of the model parameters when applying MCMC, unless otherwise stated.

2.3 Iterative Ensemble Smoothers (IES)

2.3.1 Batch Ensemble Randomized Maximum Likelihood (GN-EnRML)

Many ensemble-based history matching techniques, including the Ensemble Randomized Maximum Likelihood (EnRML), originate from the field of state estimation, where the goal is to estimate the time-dependent system state from sequential measurements, while

parameters are of secondary interest. In contrast, in history matching the focus is on updating model parameters, here defined by the vector $\mathbf{x}$, to fit observed data, rather than estimating the system state directly. However, as parameters are defined in a Bayesian context, the vector $\mathbf{x}$ may include initial (system state) and boundary conditions, which could lead to a joint state–parameter estimation problem.

The Batch Ensemble Randomized Maximum Likelihood (GN-EnRML) method was developed by Chen and Oliver (2012), and as the name suggests, it is a batch method, meaning that all data are assimilated at once. This is a different approach from the sequential methods used in the EnKF, where data are assimilated sequentially.

If model parameters are defined by the vector $\mathbf{x}$, the vector of model outputs $\mathbf{y} \in \mathbb{R}^m$ associated with measurements $\mathbf{d}$ can be mapped through a function $\mathbf{g}$, as follows:

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) \tag{2.6}$$

The function $\mathbf{y} = \mathbf{g}(\mathbf{x})$ is a general function (linear or nonlinear) that includes the forward model and any other transformation needed to convert model outputs to measurements. A set of measurements $\mathbf{d}$ can then be simulated by $\mathbf{y}$:

$$\mathbf{d} = \mathbf{y} + \mathbf{e}, \tag{2.7}$$

where $\mathbf{e}$ is a random vector representing measurement noise. From Bayes' equation (Equation 2.1), the posterior pdf of the model parameters is proportional to the product of the likelihood and the prior pdf. Assuming that the prior and the likelihood are Gaussian, the posterior pdf is given by the following equation:

$$f(\mathbf{x}|\mathbf{d}) \propto \exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{g}(\mathbf{x}))^T {\mathbf{C_d}}^{-1}(\mathbf{d}-\mathbf{g}(\mathbf{x})) - \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^T {\mathbf{C_x}}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right), \tag{2.8}$$

where $\mathbf{C_d}$ is the measurement error covariance matrix, $\mathbf{C_x}$ is the prior error covariance matrix, and $\overline{\mathbf{x}}$ is the prior mean of $\mathbf{x}$. The value that maximizes the likelihood of the posterior pdf, i.e, the maximum a posteriori estimate (MAP), is obtained by maximizing the exponent of Equation 2.8, which is equivalent to minimizing the following cost function:

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{d}-\mathbf{g}(\mathbf{x}))^T {\mathbf{C_d}}^{-1}(\mathbf{d}-\mathbf{g}(\mathbf{x})) + \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^T {\mathbf{C_x}}^{-1}(\mathbf{x}-\overline{\mathbf{x}}). \tag{2.9}$$

Equation 2.8 and Equation 2.9 form the basis of variational data assimilation approaches (e.g., 3D-Var or 4D-Var), where the aim is to find the parameter or state configuration that minimizes a suitably defined cost function combining data misfit and prior information.

It is important to mention that if the function $\mathbf{g}(\mathbf{x})$ is not linear, the posterior pdf will not be Gaussian, and it may have multiple modes. In this case, searching for the maximum

of the posterior pdf is not trivial, and the solution may depend on the initial guess of the model parameters used in iterative methods.

The first term in the cost function $J(\mathbf{x})$ is called the data mismatch or data misfit, and the second term is called the model mismatch or model misfit (Oliver et al., 2008). The data mismatch term represents the difference between the model outputs and the measurements, weighted by the measurement error covariance matrix $\mathbf{C_d}$. The model mismatch term represents the difference between the model parameters and the prior mean or first guess, weighted by the prior error covariance matrix $\mathbf{C_x}$. This second term naturally regularizes the inversion process, as it penalizes parameter values that are far from the prior mean. As pointed out by Chen and Oliver (2012), the model mismatch term is an important term that if it is omitted in the cost function and therefore in the parameter upgrade equation that is derived from it, it can lead to parameter ensemble collapse. Some mitigation measures are needed to avoid this situation, such as the use of a few iterations or the inclusion of a prior penalization term to the objective function, or only updating parameters whose total objective function value is reduced between iterations.

As the function $\mathbf{g}(\mathbf{x})$ can be nonlinear, the solution that minimizes the cost function $J(\mathbf{x})$ can be iteratively obtained using the Gauss-Newton method. Starting from a first guess $\mathbf{x}^f$ that is usually equal to the prior mean $\overline{\mathbf{x}}$, the solution at each iteration l is obtained by the following equation (Evensen et al., 2022):

$$\mathbf{x}^{l+1} = \mathbf{x}^l - \gamma^l(\mathbf{C_x}^{-1} + \mathbf{G}^{l^T}\mathbf{C_d}^{-1}\mathbf{G}^l)^{-1}(\mathbf{C_x}^{-1}(\mathbf{x}^l - \mathbf{x}^f) + \mathbf{G}^{l^T}\mathbf{C_d}^{-1}(\mathbf{g}(\mathbf{x}^l) - \mathbf{d})), \quad (2.10)$$

where γ^l is the step length at iteration l (it can change during iterations), and $\mathbf{G}^l$ is the sensitivity matrix (linearization of $\mathbf{g}(\mathbf{x})$ around $\mathbf{x}^l$). The term $(\mathbf{C_x}^{-1} + \mathbf{G}^{l^T}\mathbf{C_d}^{-1}\mathbf{G}^l)$ is the approximate Hessian (third and higher orders are discarded), equal to the posterior parameter covariance matrix $\mathbf{C}'_\mathbf{x}$ (Tarantola, 2005; Doherty, 2015). Given that in highly parameterized problems the number of model parameters is much larger than the number of measurements ($n >> m$), the inversion of this term can be computationally expensive. For this reason, the Woodbury corollaries (Koch, 1999) can be used to rewrite Equation 2.10 and solve the update in measurement space (as presented in Chen and Oliver (2012)):

$$\mathbf{x}^{l+1} = \gamma^l\mathbf{x}^f + (1-\gamma^l)\mathbf{x}^l - \gamma^l\mathbf{C_x}\mathbf{G}^{l^T}(\mathbf{C_d} + \mathbf{G}^l\mathbf{C_x}\mathbf{G}^{l^T})^{-1}(\mathbf{g}(\mathbf{x}^l) - \mathbf{d} + \mathbf{G}^l(\mathbf{x}^l - \mathbf{x}^f)). \quad (2.11)$$

Note that due to this matrix manipulation the Hessian approximation is not explicitly shown in Equation 2.11.

If now an ensemble of vectors $\mathbf{x}_i$ of size N is sampled from the prior pdf, and an ensemble of perturbed measurements $\mathbf{d}_i$ is generated by sampling the measurement error pdf (with

mean $\mathbf{d}$ and covariance $\mathbf{C_d}$), an ensemble of cost functions $J_i(\mathbf{x})$ is obtained, where each cost function is calculated as follows (Kitanidis, 1995; Oliver et al., 1996):

$$J_i(\mathbf{x}) = \frac{1}{2}(\mathbf{d}_i - \mathbf{g}(\mathbf{x}_i))^T {\mathbf{C_d}}^{-1}(\mathbf{d}_i - \mathbf{g}(\mathbf{x}_i)) + \frac{1}{2}(\mathbf{x}_i - {\mathbf{x}_i}^f)^T {\overline{\mathbf{C}}_\mathbf{x}}^{-1}(\mathbf{x}_i - {\mathbf{x}_i}^f), \tag{2.12}$$

where $\overline{\mathbf{C}}_\mathbf{x}$ is the ensemble covariance matrix of model parameters, and ${\mathbf{x}_i}^f$ is the first guess for each ensemble member. The iterative solution that minimizes the cost function $J_i(\mathbf{x})$ of each ensemble member can be obtained by the following equation:

$$\mathbf{x}_i^{l+1} = \mathbf{x}_i^l - \gamma^l({\overline{\mathbf{C}}_\mathbf{x}}^{-1} + {\overline{\mathbf{G}}^l}^T {\mathbf{C_d}}^{-1}\overline{\mathbf{G}}^l)^{-1}({\overline{\mathbf{C}}_\mathbf{x}}^{-1}(\mathbf{x}_i^l - \mathbf{x}_i^f) + {\overline{\mathbf{G}}^l}^T {\mathbf{C_d}}^{-1}(\mathbf{g}(\mathbf{x}_i^l) - \mathbf{d}_i)). \tag{2.13}$$

Equivalently, the solution can be obtained in measurement space as follows:

$$\mathbf{x}_i^{l+1} = \gamma^l \mathbf{x}_i^f + (1-\gamma^l)\mathbf{x}_i^l - \gamma^l \overline{\mathbf{C}}_\mathbf{x} {\overline{\mathbf{G}}^l}^T (\mathbf{C_d} + \overline{\mathbf{G}}^l \overline{\mathbf{C}}_\mathbf{x} {\overline{\mathbf{G}}^l}^T)^{-1}(\mathbf{g}(\mathbf{x}_i^l) - \mathbf{d}_i + \overline{\mathbf{G}}^l(\mathbf{x}_i^l - \mathbf{x}_i^f)). \tag{2.14}$$

This is the Randomized Maximum Likelihood (RML) method. As discussed by Evensen et al. (2022), if the function $\mathbf{g}(\mathbf{x})$ is linear and the prior and likelihood are Gaussian, minimizing the cost functions of Equation 2.12 will sample the posterior pdf. However, if the function $\mathbf{g}(\mathbf{x}_i)$ is nonlinear, this is not necessarily the case.

Although the sensitivity matrix for the Randomized Maximum Likelihood (RML) method can be calculated for each ensemble, note that $\overline{\mathbf{G}}^l$ is not subscripted in Equation 2.13, as it is an average sensitivity matrix calculated from the ensemble and used for all ensemble members. In Chen and Oliver (2012), the authors showed that the sensitivity matrix $\overline{\mathbf{G}}^l$ can be calculated from the ensemble anomalies of model outputs and model parameters. Following the notation of Evensen et al. (2019), if the matrix of model output anomalies is normalized by $\sqrt{N-1}$, $\mathbf{Y}^l$ is defined as:

$$\mathbf{Y}^l = \mathbf{g}(\mathbf{X}^l)\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1}, \tag{2.15}$$

where $\mathbf{X}^l \in \mathbb{R}^n x \mathbb{R}^N$ is the ensemble of model parameters, $\mathbf{1}$ is a vector of ones, and $\mathbf{I}_N$ is the identity matrix of size N, and the matrix of model parameter anomalies normalized by $\sqrt{N-1}$, $\mathbf{A}^l$, is defined as:

$$\mathbf{A}^l = \mathbf{X}^l\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1}, \tag{2.16}$$

then the sensitivity matrix $\overline{\mathbf{G}}^l$ can be calculated as:

$$\overline{\mathbf{G}}^l = \mathbf{Y}{\mathbf{A}^l}^+, \tag{2.17}$$

where ${\mathbf{A}^l}^+$ is the Moore-Penrose pseudo-inverse of $\mathbf{A}^l$. Chen and Oliver (2012) demon-

strated with a two-dimensional reservoir model that the average sensitivity matrix $\overline{\mathbf{G}}^l$ estimated from the ensemble is quite noisy, but the product $\overline{\mathbf{C}}_\mathbf{x}\overline{\mathbf{G}}^{l^T}$ is not. Additionally, the authors showed that $\overline{\mathbf{C}}_\mathbf{x}\overline{\mathbf{G}}^{l^T}$ can be estimated from the cross-covariance between model parameters and model outputs, $\overline{\mathbf{C}}_\mathbf{xg}$, as proven in the following equation:

$$\begin{aligned}\overline{\mathbf{C}}_\mathbf{x}\overline{\mathbf{G}}^{l^T} &= E[(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^T]\overline{\mathbf{G}}^{l^T} \\ &= E[(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^T\overline{\mathbf{G}}^{l^T}] \\ &\approx E[(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{g}(\mathbf{x})-\mathbf{g}(\overline{\mathbf{x}}))^T] \\ &= \overline{\mathbf{C}}_\mathbf{xg}.\end{aligned} \tag{2.18}$$

Iteration superscripts are omitted for clarity, but note that the approximation in Equation 2.18 occurs at line 3 when the term $(\mathbf{x}-\overline{\mathbf{x}})^T\overline{\mathbf{G}}^{l^T}$ is approximated by $(\mathbf{g}(\mathbf{x})-\mathbf{g}(\overline{\mathbf{x}}))^T$. This is only true at the first iteration when both $(\mathbf{x}-\overline{\mathbf{x}})^T$ and $(\mathbf{g}(\mathbf{x})-\mathbf{g}(\overline{\mathbf{x}}))^T$ are calculated from the prior. In following iterations, the term $(\mathbf{x}-\overline{\mathbf{x}})^T$ changes, then it does not represent the prior. However, it is not clear what the repercussions of this approximation are in the inversion process.

Looking at Equation 2.14, the average sensitivity matrix is always multiplied by $\overline{\mathbf{C}}_\mathbf{x}$ except for the last term $\overline{\mathbf{G}}^l(\mathbf{x}_i^l-\mathbf{x}_i^f)$. Chen and Oliver (2012) advocated that this term is small assuming the correction to the model parameters is small. This is a strong assumption, as it is not necessarily true that optimized model parameters will be close to the prior first guess, or the prior mean. However, in the GN-EnRML algorithm this term is included in the update equation (not discarded). It follows that the explicit computation of $\overline{\mathbf{G}}^l$, along with its corresponding noise, cannot be avoided.

2.3.2 Levenberg-Marquardt EnRML (LM-EnRML)

Later on, Chen and Oliver (2013) developed a Levenberg-Marquardt (LM) version of the EnRML method, called LM-EnRML. There are two main differences between the GN-EnRML and the LM-EnRML methods. The first difference is that the LM-EnRML method uses the LM algorithm to damp the update of model parameters. The second difference is that the LM-EnRML method avoids the explicit calculation of the average sensitivity matrix $\overline{\mathbf{G}}^\mathbf{l}$. As it will be shown below, this is done by including additional approximations in the parameter update equation.

Equation 2.13 can be rewritten as follows:

$$\delta\mathbf{x} = -(\overline{\mathbf{C}}_\mathbf{x}{}^{-1} + \mathbf{G}^{l^T}\mathbf{C}_\mathbf{d}{}^{-1}\mathbf{G}^l)^{-1}(\overline{\mathbf{C}}_\mathbf{x}{}^{-1}(\mathbf{x}^l-\mathbf{x}^f) + \mathbf{G}^{l^T}\mathbf{C}_\mathbf{d}{}^{-1}(\mathbf{g}(\mathbf{x}^l)-\mathbf{d})), \tag{2.19}$$

where $\delta\mathbf{x} = \mathbf{x}^{l+1}-\mathbf{x}^l$, and $\gamma = 1$. Note that from now on, the ensemble superscripts are omitted for clarity.

The LM algorithm includes an additional λ scalar to the inverse of the parameter covariance matrix in the Hessian term of Equation 2.19:

$$\delta\mathbf{x} = -((1+\lambda^l)\overline{\mathbf{C}}_\mathbf{x}^{-1} + \mathbf{G}^{l^T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}^l)^{-1}(\overline{\mathbf{C}}_\mathbf{x}^{-1}(\mathbf{x}^l - \mathbf{x}^f) + \mathbf{G}^{l^T}\mathbf{C}_\mathbf{d}^{-1}(\mathbf{g}(\mathbf{x}^l) - \mathbf{d})). \quad (2.20)$$

Equivalently, the solution can be obtained in measurement space as follows (Chen and Oliver, 2013):

$$\delta\mathbf{x} = -\frac{(\mathbf{x}^l - \mathbf{x}^f)}{1+\lambda^l} - \overline{\mathbf{C}}_\mathbf{x}\overline{\mathbf{G}}^{l^T}\left((1+\lambda^l)\mathbf{C}_\mathbf{d} + \overline{\mathbf{G}}^l\overline{\mathbf{C}}_\mathbf{x}\overline{\mathbf{G}}^{l^T}\right)^{-1}\left(\mathbf{g}(\mathbf{x}^l) - \mathbf{d} + \frac{\overline{\mathbf{G}}^l(\mathbf{x}^l - \mathbf{x}^f)}{1+\lambda^l}\right) \quad (2.21)$$

As the algorithm of Chen and Oliver (2013) performs singular value decomposition on the matrices of model output anomalies $\mathbf{Y}^l$ and initial model parameter anomalies $\mathbf{A}^0$, these matrices require additional scaling to properly reflect the variability of each data and parameter type, for the inversion process to be well-conditioned. The scaling is performed as follows:

$$\mathbf{Y}^l = \mathbf{S}_\mathbf{y}^{-1/2}\mathbf{g}(\mathbf{X}^l)\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1}, \quad (2.22)$$

$$\mathbf{A}^l = \mathbf{S}_\mathbf{x}^{-1/2}\mathbf{X}^l\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1}, \quad (2.23)$$

where $\mathbf{S}_\mathbf{y}$ and $\mathbf{S}_\mathbf{x}$ are diagonal matrices with diagonal elements equal to the variance of data noise and the prior variance of model variables, respectively.
The estimate of the average sensitivity matrix $\overline{\mathbf{G}}^l$ can be calculated as:

$$\overline{\mathbf{G}}^l = \mathbf{S}_\mathbf{y}^{1/2}\mathbf{Y}^l\mathbf{A}^{l^+}\mathbf{S}_\mathbf{x}^{-1/2}. \quad (2.24)$$

Now to avoid the explicit calculation of the average sensitivity matrix $\overline{\mathbf{G}}^l$, Chen and Oliver (2013) proposed to approximate the ensemble estimate of the prior covariance matrix of model parameters $\overline{\mathbf{C}}_\mathbf{x}$ in the Hessian term of Equation 2.20 by another matrix $\overline{\mathbf{P}}_\mathbf{x}^\mathbf{l}$ that is calculated from the updated ensemble (that changes every iteration) as

$$\overline{\mathbf{P}}_\mathbf{x}^\mathbf{l} = \mathbf{S}_\mathbf{x}^{1/2}\mathbf{A}^l\mathbf{A}^{l^T}\mathbf{S}_\mathbf{x}^{1/2}. \quad (2.25)$$

Then, Equation 2.20 can be formulated as follows:

$$\delta\mathbf{x} = -((1+\lambda^l)\overline{\mathbf{P}}_\mathbf{x}^\mathbf{l}{}^{-1} + \mathbf{G}^{l^T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}^l)^{-1}(\overline{\mathbf{C}}_\mathbf{x}^{-1}(\mathbf{x}^l - \mathbf{x}^f) + \mathbf{G}^{l^T}\mathbf{C}_\mathbf{d}^{-1}(\mathbf{g}(\mathbf{x}^l) - \mathbf{d})). \quad (2.26)$$

Using the Woodbury corollaries (see equation 1.115 in Koch, 1999), Equation 2.26 can

be rewritten as

$$\begin{aligned}\delta\mathbf{x} = &-\left((1+\lambda^l)\overline{\mathbf{P}}_{\mathbf{x}}^{\mathbf{l}\ -1} + \mathbf{G}^{l^T}{\mathbf{C}_\mathbf{d}}^{-1}\mathbf{G}^l\right)^{-1}{\overline{\mathbf{C}}_\mathbf{x}}^{-1}(\mathbf{x}^l - \mathbf{x}^f) \\ &-\overline{\mathbf{P}}_{\mathbf{x}}^{\mathbf{l}}\mathbf{G}^{l^T}\left((1+\lambda^l)\mathbf{C}_\mathbf{d} + \mathbf{G}^l\overline{P}_x^l\mathbf{G}^{l^T}\right)^{-1}\left(\mathbf{g}(\mathbf{x}^l) - \mathbf{d}\right).\end{aligned} \tag{2.27}$$

The terms $\overline{\mathbf{P}}_{\mathbf{x}}^{\mathbf{l}}\overline{\mathbf{G}}^{l^T}$ and $\overline{\mathbf{G}}^l\overline{\mathbf{P}}_{\mathbf{x}}^{\mathbf{l}}\overline{\mathbf{G}}^{l^T}$ can be calculated using Equation 2.24 and Equation 2.25 as

$$\overline{\mathbf{P}}_{\mathbf{x}}^{\mathbf{l}}\overline{\mathbf{G}}^{l^T} = {\mathbf{S}_\mathbf{x}}^{1/2}\mathbf{A}^l\mathbf{Y}^{l^T}{\mathbf{S}_\mathbf{y}}^{1/2}, \tag{2.28}$$

and

$$\overline{\mathbf{G}}^l\overline{\mathbf{P}}_{\mathbf{x}}^{\mathbf{l}}\overline{\mathbf{G}}^{l^T} = {\mathbf{S}_\mathbf{y}}^{1/2}\mathbf{Y}^l\mathbf{Y}^{l^T}{\mathbf{S}_\mathbf{y}}^{1/2}, \tag{2.29}$$

and the term ${\mathbf{G}^l}^T{\mathbf{C}_\mathbf{d}}^{-1}\mathbf{G}^l$ can be calculated from Equation 2.24 and approximating the measurement error covariance matrix $\mathbf{C}_\mathbf{d}$ by the ensemble model output anomalies $\mathbf{Y}^l$:

$${\mathbf{G}^l}^T{\mathbf{C}_\mathbf{d}}^{-1}\mathbf{G}^l = {\mathbf{S}_\mathbf{x}}^{-1/2}{\mathbf{A}^{l+}}^T\mathbf{A}^{l+}{\mathbf{S}_\mathbf{x}}^{-1/2}. \tag{2.30}$$

Finally, the parameter prior covariance matrix $\overline{\mathbf{C}}_\mathbf{x}$ can be calculated from initial model parameter anomalies as follows:

$$\overline{\mathbf{C}}_\mathbf{x} = {\mathbf{S}_\mathbf{x}}^{1/2}\mathbf{A}^0{\mathbf{A}^0}^T{\mathbf{S}_\mathbf{x}}^{1/2}. \tag{2.31}$$

Inserting the terms of Equation 2.28, Equation 2.29, Equation 2.30,and Equation 2.31, into Equation 2.27, the parameter update equation for the LM-EnRML method is obtained as follows:

$$\begin{aligned}\delta\mathbf{x} = &-{\mathbf{S}_\mathbf{x}}^{-1/2}\mathbf{A}^l\left((1+\lambda^l)\mathbf{I}_n + {\mathbf{Y}^l}^T\mathbf{Y}^l\right)^{-1}{\mathbf{A}^l}^T{\mathbf{A}^0}^{-T}{\mathbf{A}^0}^{-1}{\mathbf{S}_\mathbf{x}}^{-1/2}(\mathbf{x}^l - \mathbf{x}^f) \\ &-{\mathbf{S}_\mathbf{x}}^{1/2}\mathbf{A}^l\mathbf{Y}^{l^T}\left((1+\lambda^l)\mathbf{I}_m + \mathbf{Y}^l\mathbf{Y}^{l^T}\right)^{-1}{\mathbf{S}_\mathbf{y}}^{-1/2}\left(\mathbf{g}(\mathbf{x}^l) - \mathbf{d}\right)\end{aligned} \tag{2.32}$$

Equation 2.32 is the update equation for the LM-EnRML method, and is equivalent to equation (18) of Chen and Oliver (2013). To perform the matrix inversions in Equation 2.32, the initial model parameter anomalies $\mathbf{A}^0$ and the model output anomalies $\mathbf{Y}^l$ are subject to truncated singular value decomposition (SVD). A certain level of energy is defined to calculate the number of singular values to preserve in the inversion process. Chen and Oliver (2013) included an additional simplification by discarding the model mismatch term (first term) in the update equation. This method is called the LM-EnRML(approx) method, and the update equation is simplified as

$$\delta\mathbf{x} = -{\mathbf{S}_\mathbf{x}}^{1/2}\mathbf{A}^l\mathbf{Y}^{l^T}\left((1+\lambda^l)\mathbf{I}_m + \mathbf{Y}^l\mathbf{Y}^{l^T}\right)^{-1}{\mathbf{S}_\mathbf{y}}^{-1/2}\left(\mathbf{g}(\mathbf{x}^l) - \mathbf{d}\right). \tag{2.33}$$

This is equivalent to equation (19) in Chen and Oliver (2013). It is important to note that the LM-EnRML(approx) method is a maximum likelihood method, as it does not include the model mismatch term in the update equation. Caution is then needed to avoid parameter ensemble collapse, as the model mismatch term is an important term that regularize the inversion process.
The LM-EnRML method of Chen and Oliver (2013) has been extensively used in the groundwater community after its implementation in the software PESTPP-IES by White (2018).

2.3.3 Subspace EnRML (SEnRML)

In order to avoid the approximations made in the LM-EnRML method, the SEnRML method was developed by Raanes et al. (2019) and derived in a clearer manner by Evensen et al. (2019). In this method, the solution is searched in the ensemble subspace as a linear combination of the initial ensemble anomalies and the first guess of the model parameters (Evensen et al., 2019),

$$\mathbf{x}^l = \mathbf{x}^f + \mathbf{A}\mathbf{w}^l, \tag{2.34}$$

where $\mathbf{x}^f$ and $\mathbf{x}^l$ are the first guess and updated model parameters, respectively, $\mathbf{A}$ is the matrix of initial ($l = 0$) model parameter ensemble anomalies as defined in Equation 2.16, and $\mathbf{w}^l \in \mathbb{R}^N$ is the vector of weights (Note that the ensemble subscripts are omitted for clarity). This equation shows that the optimized parameters, deemed to approximately sample the posterior pdf, will be a linear combination of the initial parameter realizations. Solving the problem in this way, the inversion process is naturally regularized.
The cost function presented in Equation 2.12 can be rewritten in terms of $\mathbf{w}^l$ as follows:

$$J(\mathbf{w}^l) = \frac{1}{2}(\mathbf{d} - \mathbf{g}(\mathbf{x}^f + \mathbf{A}\mathbf{w}^l))^T {\mathbf{C_d}}^{-1}(\mathbf{d} - \mathbf{g}(\mathbf{x}^f + \mathbf{A}\mathbf{w}^l)) + \frac{1}{2}{\mathbf{w}^l}^T\mathbf{w}^l. \tag{2.35}$$

The iterative solution that minimizes the cost function $J(\mathbf{w}^l)$ can be obtained using Equation 2.10 as

$$\mathbf{w}^{l+1} = \mathbf{w}^l - \gamma^l(\mathbf{I} + (\mathbf{G}^l\mathbf{A})^T{\mathbf{C_d}}^{-1}(\mathbf{G}^l\mathbf{A}))^{-1}(\mathbf{w}^l + (\mathbf{G}^l\mathbf{A})^T{\mathbf{C_d}}^{-1}(\mathbf{g}(\mathbf{x}^f + \mathbf{A}\mathbf{w}^l) - \mathbf{d})). \tag{2.36}$$

This equation can be rewritten in measurement space as

$$\mathbf{w}^{l+1} = \mathbf{w}^l - \gamma\left(\mathbf{w}^l - (\mathbf{G}^l\mathbf{A})^T(\mathbf{G}^l\mathbf{A}(\mathbf{G}^l\mathbf{A})^T + \mathbf{C_d})^{-1}(\mathbf{G}^l\mathbf{A}\mathbf{w}^l + \mathbf{d} - \mathbf{g}(\mathbf{x}^f + \mathbf{A}\mathbf{w}^l))\right). \tag{2.37}$$

As the equations show, the SEnRML uses the Gauss-Newton algorithm to iteratively update the model parameters. Again, instead of calculating the sensitivity matrix for each weight vector, an average sensitivity matrix $\overline{\mathbf{G}}^l$ is calculated from the ensemble.

Using Equation 2.17 it can be shown that

$$\overline{\mathbf{G}}^l\mathbf{A} = \mathbf{Y}^l\mathbf{A}^{l+}\mathbf{A}. \tag{2.38}$$

The matrix of updated anomalies $\mathbf{A}^l$ can be related to the matrix of initial anomalies $\mathbf{A}$ and the vector of updated weights $\mathbf{w}^l$ as follows:

$$\begin{aligned}
\mathbf{A}^l &= \mathbf{X}^l\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1} \\
&= \left(\mathbf{X}^f + \mathbf{A}\mathbf{W}^l\right)\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1} \\
&= \mathbf{A} + \mathbf{A}\mathbf{W}^l\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1} \\
&= \mathbf{A}\left(\mathbf{I}_N + \mathbf{W}^l\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1}\right) \\
&= \mathbf{A}\boldsymbol{\Omega}^l,
\end{aligned} \tag{2.39}$$

where $\boldsymbol{\Omega}^l$ it is defined as:

$$\boldsymbol{\Omega}^l = \left(\mathbf{I}_N + \mathbf{W}^l\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)/\sqrt{N-1}\right). \tag{2.40}$$

Note that $\boldsymbol{\Omega}^l$ is always full rank (Evensen et al., 2019). It follows that

$$\begin{aligned}
\overline{\mathbf{G}}^l\mathbf{A} &= \mathbf{Y}^l\mathbf{A}^{l+}\mathbf{A} \\
&= \mathbf{Y}^l\mathbf{A}^{l+}\mathbf{A}^l\boldsymbol{\Omega}^{l-1}.
\end{aligned} \tag{2.41}$$

The projection $\mathbf{A}^{l+}\mathbf{A}^l$ can be discarded when $n \geq N-1$ (this is generally the case in highly parameterized inversion) or if the model is linear (for further details on the demonstration the reader is referred to Evensen et al. (2019)), leading to the following equation:

$$\mathbf{S}^l = \overline{\mathbf{G}}^l\mathbf{A} = \mathbf{Y}^l\boldsymbol{\Omega}^{l-1}, \tag{2.42}$$

where $\mathbf{S}^l$ is defined as the matrix of predicted and deconditioned ensemble anomalies (Evensen et al., 2019). It is important to note that for Equation 2.42 to generally hold, $\mathbf{Y}^l$ needs to be multiplied by the term $\mathbf{A}^{l+}\mathbf{A}$ for the cases when $n < N-1$ or the model is nonlinear. With this definition, Equation 2.37 can be rewritten (in matrix form) as

$$\mathbf{W}^{l+1} = \mathbf{W}^l - \gamma\left(\mathbf{W}^l - \mathbf{S}^{l^T}(\mathbf{S}^l\mathbf{S}^{l^T} + \mathbf{C_d})^{-1}\mathbf{H}^l\right), \tag{2.43}$$

where $\mathbf{H}^l$ is the 'innovation' term (Evensen et al., 2019) defined as

$$\mathbf{H}^l = \mathbf{S}^l\mathbf{W}^l + \mathbf{D} - \mathbf{g}(\mathbf{X}^f + \mathbf{A}\mathbf{W}^l). \tag{2.44}$$

Equation 2.43 is the update equation for the SEnRML method, solved in measurement space. The same equation can be solved in ensemble subspace as follows:

$$\mathbf{W}^{l+1} = \mathbf{W}^{l} - \gamma \left(\mathbf{I}_N + \mathbf{S}^{l^T} \mathbf{C_d}^{-1} \mathbf{S}^{l} \right)^{-1} \left(\mathbf{W}^{l} + \mathbf{S}^{l^T} \mathbf{C_d}^{-1} (\mathbf{g}(\mathbf{X}^{f} + \mathbf{A}\mathbf{W}^{l}) - \mathbf{D}) \right). \quad (2.45)$$

To solve Equation 2.43, the matrix $\mathbf{S}^{l}\mathbf{S}^{l^T} + \mathbf{C_d}$ is the only term that requires inversion. Evensen et al. (2019) presented four alternatives to solve this inversion: 1. Direct inversion, 2. Exact inversion, 3. Ensemble subspace inversion using full $\mathbf{C_d}$, and 4. Ensemble subspace inversion using low-rank $\mathbf{C_d}$. For a full description of these matrix inversion options, the reader is referred to Evensen et al. (2019). The following is the solution for each of the four inversion options:

1. Direct inversion:

   In this case, the matrix $\mathbf{C} = \mathbf{S}^{l}\mathbf{S}^{l^T} + \mathbf{C_d}$ is directly inverted using SVD as

   $$\mathbf{C}^{-1} = \mathbf{Z}\mathbf{\Lambda}^{+}\mathbf{Z}^{T}, \quad (2.46)$$

   where $\mathbf{Z}$ is the matrix of right singular vectors of $\mathbf{C}$, and $\mathbf{\Lambda}^{+}$ is the diagonal matrix of the inverse of the singular values of $\mathbf{C}$.

2. Exact inversion:

   In this case, using the Woodbury corollaries, the updated weights can be calculated as

   $$\mathbf{W}^{l+1} = \mathbf{W}^{l} - \gamma \left(\mathbf{W}^{l} - (\mathbf{S}^{l^T} \mathbf{C_d}^{-1} \mathbf{S}^{l} + \mathbf{I}_N)^{-1} \mathbf{S}^{l^T} \mathbf{C_d}^{-1} \mathbf{H} \right). \quad (2.47)$$

   In this case it is generally assumed that $\mathbf{C_d}$ is equal to the identity matrix, which is obtained by scaling a diagonal measurement error covariance matrix by the measurement noise variance, and Equation 2.47 can be simplified as

   $$\mathbf{W}^{l+1} = \mathbf{W}^{l} - \gamma \left(\mathbf{W}^{l} - (\mathbf{S}^{l^T} \mathbf{S}^{l} + \mathbf{I}_N)^{-1} \mathbf{S}^{l^T} \mathbf{H} \right). \quad (2.48)$$

   The inversion of the matrix $\mathbf{S}^{l}\mathbf{S}^{l} + \mathbf{I}_N$ is performed using SVD of $\mathbf{S}^{l}$. It is important to note that this approach is not valid if $\mathbf{C_d}$ has a more complex structure.

3. Ensemble subspace inversion using full $\mathbf{C_d}$:

   This matrix inversion method elegantly identifies that, as the optimized parameters are the result of linear combinations of the initial parameters, the measurement noise covariance matrix $\mathbf{C_d}$ can be approximated by its projection into the predicted ensemble anomalies subspace $\mathbf{S}^{l}$. The matrix inversion is performed as

   $$(\mathbf{S}^{l}\mathbf{S}^{l^T} + \mathbf{C_d})^{-1} = \left(\mathbf{U}\mathbf{\Sigma}^{+^T}\mathbf{Z} \right) (\mathbf{I}_N + \mathbf{\Lambda})^{-1} \left(\mathbf{U}\mathbf{\Sigma}^{+^T}\mathbf{Z} \right)^{T}, \quad (2.49)$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD of $\mathbf{S}^l$, and $\mathbf{Z}$ and $\mathbf{\Lambda}$ are the SVD result of

$$\mathbf{\Sigma}^+\mathbf{U}^T\mathbf{C_d}\mathbf{U}\mathbf{\Sigma}^{+T} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T. \tag{2.50}$$

4. Ensemble subspace inversion using low-rank $\mathbf{C_d}$:

   This option is equivalent to the previous ensemble subspace inversion but instead of using full $\mathbf{C_d}$, a low-rank approximation $\mathbf{E}$ is used such that $\mathbf{C_d} \approx \mathbf{E}\mathbf{E}^T$. The inversion is performed as

$$(\mathbf{S}^l\mathbf{S}^{lT} + \mathbf{E}\mathbf{E}^T)^{-1} = \left(\mathbf{U}\mathbf{\Sigma}^{+T}\mathbf{Z}\right)(\mathbf{I}_N + \mathbf{\Lambda})^{-1}\left(\mathbf{U}\mathbf{\Sigma}^{+T}\mathbf{Z}\right)^T, \tag{2.51}$$

   and

$$\mathbf{\Sigma}^+\mathbf{U}^T\mathbf{E}\mathbf{E}^T\mathbf{U}\mathbf{\Sigma}^{+T} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T. \tag{2.52}$$

   This approach reduces the computational cost of the inversion process if $N < m$, and it does not require to explicitly define the structure of the measurement noise covariance matrix $\mathbf{C_d}$. This opens the opportunity to include more complex structures in the measurement noise, especially when it is combined with structural noise.

   It is noted that performing SVD on Equation 2.52 is equivalent to performing SVD on $\mathbf{\Sigma}^+\mathbf{U}^T\mathbf{E}$ and squaring the singular values to obtain $\mathbf{\Lambda}$. It is also important to mention that before performing SVD on $\mathbf{S}^l$, this matrix should be scaled to the variability of the observations. This can be done through simply generating a diagonal scaling matrix similar as presented in Chen and Oliver (2013), or obtain the Cholesky decomposition $\mathbf{L}$ of the matrix $\mathbf{E}\mathbf{E}^T$, and calculate the scaling matrix as $\mathbf{L}^{-1/2}$, as presented by Emerick and Reynolds (2012).

2.3.4 Levenberg-Marquardt subspace EnRML (LM-SEnRML)

Although there is no formal literature presenting a Levenberg-Marquardt version of the SEnRML method, it is easy to derive it. Looking at Equation 2.45, the LM algorithm can be included by adding a scalar λ^l to the inverse of the matrix $\mathbf{I}_N + \mathbf{S}^{lT}\mathbf{C_d}^{-1}\mathbf{S}^l$ as

$$\mathbf{W}^{l+1} = \mathbf{W}^l - \left((1+\lambda^l)\mathbf{I}_N + \mathbf{S}^{lT}\mathbf{C_d}^{-1}\mathbf{S}^l\right)^{-1}\left(\mathbf{W}^l + \mathbf{S}^{lT}\mathbf{C_d}^{-1}(\mathbf{g}(\mathbf{X}^f + \mathbf{A}\mathbf{W}^l) - \mathbf{D})\right). \tag{2.53}$$

Using the Woodbury corollaries, Equation 2.45 with the added λ scalar can be rewritten similarly to Equation 2.43 as

$$\mathbf{W}^{l+1} = \mathbf{W}^l - \left(\frac{\mathbf{W}^l}{1+\lambda^l} - \mathbf{S}^{lT}(\mathbf{S}^l\mathbf{S}^{lT} + (1+\lambda^l)\mathbf{C_d})^{-1}\mathbf{H}^l\right), \tag{2.54}$$

where $\mathbf{H}^l$ is the 'innovation' term defined as

$$\mathbf{H}^l = \frac{\mathbf{S}^l\mathbf{W}^l}{1+\lambda^l} + \mathbf{D} - \mathbf{g}(\mathbf{X}^f + \mathbf{A}\mathbf{W}^l). \tag{2.55}$$

It should be noted that when using the low-rank $\mathbf{C_d}$, i.e., $\mathbf{C_d} \approx \mathbf{E}\mathbf{E}^T$, the matrix $\mathbf{E}$ should be scaled by the square root of the scalar $1+\lambda^l$ before performing SVD of Equation 2.52. It is also worth mentioning that to my knowledge this approach has not been widely tested in the literature, and it is presented here as an alternative to the use of the step-length parameter γ in the SEnRML method.

2.3.5 Iterative Local Updating Ensemble Smoother (ILUES)

The iterative local updating ensemble smoother (ILUES) was developed by Zhang et al. (2018) as an improvement to the iterative ensemble smoother method with the purpose of sampling multi-modal posterior distributions. However, the ILUES method is also useful to sample unimodal posterior distributions for highly nonlinear problems. The method, as its name suggests, is an iterative ensemble smoother, that updates a subset of parameter realizations, with grouping based on a combination of parameter similarities and their goodness of fit to the observations. The following is a brief description of the ILUES method. For a more detailed explanation, the reader is referred to Zhang et al. (2018).

A normalized total mismatch function is divided into a normalized data mismatch and a normalized model mismatch functions as

$$\mathbf{J}_n = \mathbf{J_d}/{\mathbf{J_d}}^{max} + \mathbf{J_x}/{\mathbf{J_x}}^{max}, \tag{2.56}$$

where $\mathbf{J_d}$ and $\mathbf{J_x}$ are the data and model mismatch functions, respectively, and ${\mathbf{J_d}}^{max}$ and ${\mathbf{J_x}}^{max}$ are the maximum values of both functions. This normalization is applied to avoid the dominance of the data mismatch function over the model mismatch function. Then, a local ensemble $\mathbf{X}_L$ of size $N_L = \alpha N$ is selected from the full ensemble $\mathbf{X}$ based on the best N_L realizations of the normalized total mismatch function $\mathbf{J}_n$. This evaluation is repeated for each $\mathbf{x}_i \in \mathbf{X}$. Therefore, there will be N local ensembles, one for each ensemble member. There is no restriction that the ensemble members should belong to a unique local ensemble, as this process is iterative and independent. Once a local ensemble is defined, the model parameters can be updated using any of the methods described above, such as the LM-EnRML, SEnRML, or LM-SEnRML methods. Moreover, localization can be applied to the local ensemble to avoid spurious correlations and ensemble collapse.

For each $\mathbf{x}_i$ there will be N_l updated model parameters in the updated local ensemble $\mathbf{X}_L$. Naturally, only one ensemble member $\mathbf{X}_L$ has to be selected to update the model parameter vector $\mathbf{x}_i$ in the full ensemble $\mathbf{X}$. The selection process can be based on several

criteria. Zhang et al. (2018) proposed to randomly select one realization from the updated local ensemble $\mathbf{X}_L$. Another simpler option is to select the first realization, that in theory should be the updated realization associated with $\mathbf{x}_i$. The reason being that the model mismatch cost function for this realization is zero.

2.3.6 Localization

When using ensemble methods such as those described above, the use of a reduced ensemble size N compared to the number of model parameters n leads to a low-rank representation of the parameter covariance matrix. This can generate three main issues (Evensen et al., 2022):

1. Poor representation of the model parameter covariance matrix $\mathbf{C}_\mathbf{x}$ that can lead to spurious correlations and thereby promote parameter ensemble collapse.

2. A solution that is confined to the ensemble subspace defined by the prior realizations.

3. Projection of measurements into the ensemble subspace, limiting assimilation of data that cannot be represented by the ensemble of model outputs.

For these reasons, the use of localization techniques is essential. The term 'localization' refers to the auxiliary technique used in model inversion that limits the effect of observations on parameters to a certain local domain. The local domain can have a physical interpretation, such as time and distance, or it can be based on statistical correlation.
Distance-based localization schemes are localization techniques that use a tapering function, such as the Gaspari-Cohn function (Gaspari and Cohn, 1999) to penalize the effect of observations on model parameters as the distance between them increases. The problem with this approach is that all observations and parameters require a geographical location, which is not always possible. As a matter of fact, in groundwater modelling, it is common to have ambiguous locations of observations and model parameters, such as groundwater inflows to a river or the bulk hydraulic conductivity of an aquifer.
A localization scheme that is not dependent on geographical or temporal locations is more appealing for groundwater modelling. Luo et al. (2018) and Luo and Bhakta (2020) proposed a correlation-based adaptive and automatic localization scheme that generates a tapering function applied to the Kalman gain term in the parameter update equation. In Luo et al. (2018), the tapering function is an indicator matrix (0 or 1) $\mathbf{L} \in \mathbb{R}^{n \times m}$ that eliminates the correlations that are below a certain threshold:

$$\mathbf{L} = I(abs(\overline{\boldsymbol{\rho}}) > \boldsymbol{\theta}), \tag{2.57}$$

where $\overline{\boldsymbol{\rho}} \in \mathbb{R}^{n \times m}$ is the correlation matrix between $\mathbf{X}$ and $\mathbf{Y}$ calculated from the ensemble, $\boldsymbol{\theta} \in \mathbb{R}^{n \times m}$ is the noise correlation threshold, and I is the indicator function that returns 1 if the condition is true and 0 otherwise. The correlation noise $\boldsymbol{\epsilon}$ is calculated using a high-pass filter, and the threshold is estimated as a multiple of the standard deviation of the noise $\boldsymbol{\sigma_\epsilon}$, as follows (Donoho and Johnstone, 1994):

$$\boldsymbol{\theta} = \boldsymbol{\sigma_\epsilon} \times \sqrt{2 \ln(N_\epsilon)}, \tag{2.58}$$

where N_ϵ is the number of noise elements. In turn, the noise standard deviation $\boldsymbol{\sigma_\epsilon}$ is calculated using the median absolute deviation (MAD) (Donoho and Johnstone, 1995) as

$$\boldsymbol{\sigma_\epsilon} = 1.4826 \times MAD(\boldsymbol{\epsilon}). \tag{2.59}$$

Luo and Bhakta (2020) proposed two improvements to the localization scheme of Luo et al. (2018). First, they proposed the random shuffle approach as a simple method to estimate the noise $\boldsymbol{\epsilon}$ in the correlation matrix estimated from the ensemble. As presented in their work, assuming that the ensemble members in $\mathbf{X}$ are independent and identically distributed (which is true for the prior ensemble given the ensemble generation process), and if the predicted ensemble members in $\mathbf{Y}$ are shuffled assuring that no member repeats its original position, in theory the correlation matrix between $\mathbf{X}$ and the shuffled version of $\mathbf{Y}$ should tend to zero as $N \to \infty$. Given that the ensemble size is limited, the correlation matrix between $\mathbf{X}$ and the shuffled version of $\mathbf{Y}$ will not be zero, and the correlation values can be treated as estimates of the noise. The second improvement proposed by Luo and Bhakta (2020) is the use of the Gaspari-Cohn function as a continuous tapering function instead of the indicator function:

$$f_{GC}(z) = \begin{cases} -\frac{1}{4}z^5 + \frac{1}{2}z^4 + \frac{5}{8}z^3 - \frac{5}{3}z^2 + 1 & \text{if } 0 \leq z \leq 1, \\ \frac{1}{12}z^5 - \frac{1}{2}z^4 + \frac{5}{8}z^3 + \frac{5}{3}z^2 - 5z + 4 - \frac{2}{3}z^{-1} & \text{if } 1 < z \leq 2, \\ 0 & \text{if } z > 2, \end{cases} \tag{2.60}$$

where z is a dummy variable representing a pseudo distance, defined as

$$z = \frac{1 - abs(\overline{\rho})}{1 - \theta}. \tag{2.61}$$

In this last equation, correlation and threshold indices are omitted for clarity. It is important to note that the decision on how many and what parameter types are used to calculate the noise threshold is a subjective choice. In the extreme case one could estimate the noise threshold for each pair of model parameters and observations, requiring repetition of the application of the random shuffle method many times (Ranazzi et al., 2022; Luo et al., 2023). On the contrary, as done by Luo et al. (2018) and Luo and Bhakta

(2020), one could estimate a single noise threshold for a group of model parameters for each observation. Luo et al. (2023) added an optional alternative to estimate the noise threshold for global parameters, such as the bulk hydraulic conductivity of an aquifer, using the asymptotic estimate $c/\sqrt{N}$, where c is an arbitrary number between 3 and 4 (Luo et al., 2023) and N is the ensemble size.

It is recognized that the correlation noise threshold estimated with the universal rule leads to aggressive localization. This is particularly true when the noise samples for each pair of model parameters and observations are generated by repeating the random shuffle method. In this case, increasing the number of correlation noise samples does not reduce the standard deviation of correlation noise, and therefore applying the universal rule of Donoho and Johnstone (1994) will only increase the noise threshold. For this reason, in this work the noise threshold is estimated as a multiple (1.0 by default) of the standard deviation of the noise, calculated through the random shuffle method, as done in the PESTPP-IES software (White, 2018).

As discussed by Silva Neto et al. (2021) and Ranazzi et al. (2022), the pseudo distance dummy variable z of Luo and Bhakta (2020) generates undesired results. One of the issues is that the tapering function is suboptimal asymptotically, as it does not approach one as $N \to \infty$ (Ranazzi et al., 2022). In other words, tapering values equal to one are only obtained for correlation values that are also one. To address this issue, Silva Neto et al. (2021) proposed a new dummy z variable to use in the Gaspari-Cohn function, arbitrarily defined as

$$z = \max(1.67 - \frac{0.67|\overline{\rho}|}{\theta}, 0). \tag{2.62}$$

Another recent localization approach was developed by Ranazzi et al. (2022), where they combine the correlation-based localization method of Luo et al. (2018); Luo and Bhakta (2020) with the pseudo-optimal localization (POL) method of Furrer and Bengtsson (2007). In the POL method, a localization matrix is defined as

$$l_{i,j} = \frac{c_{i,j}^2}{c_{i,j}^2 + (c_{i,j}^2 + c_{i,i}c_{j,j})/N}, \tag{2.63}$$

where $l_{i,j}$ is the tapering value of the localization matrix $\mathbf{L}$ and $c_{i,j}$ is the true covariance between the i and j, respectively. It is noted that the tapering value tends to 1 as $N \to \infty$. Furrer and Bengtsson (2007) proposed to replace the true covariance with the ensemble estimate, leading to pseudo-optimality of the method. They also suggested adding sparseness to the localization matrix by replacing small values of $l_{i,j}$ with zeros, using a threshold for the cross-correlation $c_{i,j}$ as

$$|c_{i,j}| < \epsilon\sqrt{c_{i,i}c_{j,j}}, \tag{2.64}$$

where ϵ is a small value, typically between 0.01 and 0.001.
If the ensemble covariance estimates are used in Equation 2.63, it can be argued that there will be errors in the cross-covariance terms, but not necessarily in the diagonal terms (Ranazzi et al., 2022). Following this rationale, Ranazzi et al. (2022) proposed a modification to the POL method, where the tapering values of the localization matrix are calculated as

$$l_{i,j} = \frac{c_{i,j}^2}{c_{i,j}^2 + (c_{i,i}c_{j,j})/N + \beta_{i,j}^2}, \tag{2.65}$$

where $\beta_{i,j}^2$ is the error term, or penalty factor. The authors proposed to use the random shuffle method of Luo and Bhakta (2020) to estimate a covariance error threshold value $\theta_{i,j}$ from which the penalty factor is calculated as

$$\beta_{i,j} = F\theta_{i,j}, \tag{2.66}$$

where F is a general function that potentially depends on the covariance values and θ. Ranazzi et al. (2022) proposed four options for F, and here the two simplest are presented:

$$\begin{aligned} F_1 &= 1.0, \\ F_2(c_{i,j}, c_{i,i}, c_{j,j}) &= 1 - \frac{c_{i,j}^2}{c_{i,i}c_{j,j}}. \end{aligned} \tag{2.67}$$

The localization matrix $\mathbf{L}$ obtained from any of the aforementioned method, can be used to perform local analysis (Evensen et al., 2022). In this way, localization can be easily added to any of the ensemble-based methods, where instead of updating parameters to all observations in one step, a subset of parameters (or even each parameter) is independently updated to a significantly correlated subset of observations. In this way, localization is scheme-independent. Additionally, Silva Neto et al. (2021) stated that, similar to the approach of Chen and Oliver (2017), the localization matrix can taper the ensemble anomalies $\mathbf{S}$ and the innovation $\mathbf{H}$ terms, as follows:

$$\begin{aligned} \hat{\mathbf{S}}^l &= \mathbf{S}^l \cdot \mathbf{L}^{1/2}, \\ \hat{\mathbf{H}}^l &= \hat{\mathbf{S}}^l \mathbf{W}^l + \left(\mathbf{D} - \mathbf{g}(\mathbf{X}^f + \mathbf{A}\mathbf{W}^l)\right) \cdot \mathbf{L}^{1/2}. \end{aligned} \tag{2.68}$$

In this way lower correlations are penalized. If this approach is discarded, the localization matrix will only help to select the subset of observation that will be used to update the parameters. In this work it has been found that applying the localization matrix to both ensemble anomalies and innovation terms only works well for the LM-EnRML method. For the SEnRML and LM-SEnRML methods, the tapering leads to similar convergence behaviour compared to the non-localized cases only if it is applied to the innovation term. It is speculated this may be the case because subspace methods project the innovations into the ensemble anomalies space, and therefore the tapering of the ensemble anomalies

is redundant.

2.4 Regularized Inversion and Linear Uncertainty Analysis

2.4.1 Least Squares, SVD and Tikhonov Regularization

Suppose the following nonlinear inverse problem:

$$\mathbf{d}^* = \mathbf{g}(\mathbf{x}^*) + \boldsymbol{\epsilon}, \tag{2.69}$$

where $\mathbf{d}^* \in \mathbb{R}^m$ is the vector of observations, $\mathbf{x}^* \in \mathbb{R}^n$ is the vector of model parameters, $\mathbf{g}(\cdot)$ is the forward model, and $\boldsymbol{\epsilon} \in \mathbb{R}^m$ is the vector of measurement errors. A linearization of Equation 2.69 can be written as

$$\mathbf{d} = \mathbf{G}\mathbf{x} + \boldsymbol{\epsilon}, \tag{2.70}$$

where $\mathbf{G} \in \mathbb{R}^{m \times n}$ is the sensitivity matrix of the forward model evaluated at the first guess $\mathbf{x}^f$, or prior estimate, and $\mathbf{x}$ and $\mathbf{d}$ are departures from the first guess $\mathbf{x}^f$ and its model evaluation $\mathbf{g}(\mathbf{x}^f)$, as follows:

$$\begin{aligned} \mathbf{d} &= \mathbf{d}^* - \mathbf{g}(\mathbf{x}^f), \\ \mathbf{x} &= \mathbf{x}^* - \mathbf{x}^f. \end{aligned} \tag{2.71}$$

For an overdetermined problem, the matrix product $\mathbf{G}^T\mathbf{G}$ is positive definite (therefore it has an inverse), and the solution to the linearized problem discarding noise is given by

$$\mathbf{x} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d}. \tag{2.72}$$

Under the presence of measurement noise, the solution to the linear problem can be solved by minimizing the following least square problem (Tarantola, 2005):

$$J_{\mathbf{d}}(\mathbf{x}) = (\mathbf{d} - \mathbf{G}\mathbf{x})^T {\mathbf{C}_\mathbf{d}}^{-1} (\mathbf{d} - \mathbf{G}\mathbf{x}), \tag{2.73}$$

where $\mathbf{C}_\mathbf{d}$ is the measurement error covariance matrix. The solution $\underline{\mathbf{x}}$ in this case is an estimate of the true $\mathbf{x}$ solution (Doherty, 2015), and given by

$$\underline{\mathbf{x}} = (\mathbf{G}^T {\mathbf{C}_\mathbf{d}}^{-1} \mathbf{G})^{-1} \mathbf{G}^T {\mathbf{C}_\mathbf{d}}^{-1} \mathbf{d}. \tag{2.74}$$

For an underdetermined problem, the matrix product $\mathbf{G}^T\mathbf{G}$ or $\mathbf{G}^T {\mathbf{C}_\mathbf{d}}^{-1} \mathbf{G}$ is not positive definite, and therefore Equation 2.72 or Equation 2.74 cannot be directly used to solve

the linear problem. In this case, as there are infinite solutions to the linear problem, some regularization is required find a unique solution.
The simplest form of regularization is Singular Value Decomposition (SVD). The SVD of the expression $\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G}$ is given by

$$\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G} = \mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^T, \tag{2.75}$$

where $\mathbf{Z}$ is the matrix of eigenvectors, and $\mathbf{\Sigma}$ is the diagonal matrix of the eigenvalues. The solution to the regularized least squares problem is given by

$$\underline{\mathbf{x}} = \mathbf{Z}_1\mathbf{\Sigma}^{-1}\mathbf{Z}_1^T\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{d}, \tag{2.76}$$

where $\mathbf{Z}_1$ is the matrix of the first $k < m$ columns of the right singular vectors, and $\mathbf{\Sigma}_1$ is the diagonal matrix of the first k singular values. Another common form of regularization is Tikhonov regularization, which adds a regularization term to the least squares problem. A typical regularization function is departures from prior values normalized by their prior uncertainties, as follows:

$$J_\mathbf{t}(\mathbf{x}) = \frac{1}{2}J_\mathbf{d}(\mathbf{x}) + \frac{1}{2}J_\mathbf{x}(\mathbf{x}) = \frac{1}{2}(\mathbf{d}-\mathbf{G}\mathbf{x})^T\mathbf{C_d}^{-1}(\mathbf{d}-\mathbf{G}\mathbf{x}) + \frac{\mu^2}{2}(\mathbf{x}-\mathbf{x}_f)^T\mathbf{C_x}^{-1}(\mathbf{x}-\mathbf{x}_f), \tag{2.77}$$

where $\mathbf{C_x}$ is the model parameter covariance matrix, and μ is the regularization weight factor (Doherty, 2015). If data uncertainty and parameter can be represented by one single variance σ_d^2 and σ_x^2, respectively, the regularization weight factor can be set as $\mu = \sigma_d/\sigma_x$ (Oliver et al., 2008).
The solution to the regularized least squares problem is given by (Moore and Doherty, 2006; Doherty, 2015):

$$\underline{\mathbf{x}} = (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G} + \mu^2\mathbf{C_x}^{-1})^{-1}\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{d}, \tag{2.78}$$

If the regularization parameter $\mu = 1$, the solution to the regularized least squares problem is the following (Tarantola, 2005):

$$\underline{\mathbf{x}} = (\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{G} + \mathbf{C_x}^{-1})^{-1}\mathbf{G}^T\mathbf{C_d}^{-1}\mathbf{d}. \tag{2.79}$$

Using the Woodbury corollaries, Equation 2.79 can be rewritten as

$$\underline{\mathbf{x}} = \mathbf{C_x}\mathbf{G}^T(\mathbf{G}\mathbf{C_x}\mathbf{G}^T + \mathbf{C_d})^{-1}\mathbf{d}. \tag{2.80}$$

A more general and flexible form of regularization is presented in Doherty (2015), with

the following estimate of the model parameters:

$$\underline{\mathbf{x}} = (\mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{G} + \mu^2 \mathbf{G}_r^T {\mathbf{C_x}}^{-1}\mathbf{G}_r)^{-1}(\mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{d} + \mathbf{G}_r^T {\mathbf{C_x}}^{-1}\mathbf{d}_r), \tag{2.81}$$

where $\mathbf{G}_r$ is the regularization sensitivity matrix, $\mathbf{d}_r$ is the regularization observation vector, and in this case ${\mathbf{C_x}}^{-1}$ is the regularization weight matrix (it may not be necessarily derived from inverting a regular parameter covariance matrix). Note that Equation 2.81 is equal to Equation 2.78 when $\mathbf{G}_r = \mathbf{I}$ and $\mathbf{d}_r = \mathbf{0}$. This occurs when each parameter is regularized by its prior value, and regularization uncertainty is equal to the prior parameter uncertainty.

PEST (Doherty, 2023) uses Levenberg-Marquardt optimization to solve the regularized least squares problem. Therefore, Equation 2.81 can be rewritten as

$$\underline{\mathbf{x}} = (\mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{G} + \mu^2 \mathbf{G}_r^T {\mathbf{C_x}}^{-1}\mathbf{G}_r + \lambda\mathbf{I})^{-1}(\mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{d} + \mathbf{G}_r^T {\mathbf{C_x}}^{-1}\mathbf{d}_r), \tag{2.82}$$

where λ is the Levenberg-Marquardt parameter.

The iterative solution to the regularized least squares problem implemented in PEST minimizes the regularization function $J_\mathbf{r}(\mathbf{x})$ while achieving a certain predefined level of data misfit, that is adjusted for each iteration as a percentage of the initial data misfit $J_\mathbf{d}(\mathbf{x})$ (Doherty, 2015). Singular value decomposition adds numerical stability to the inversion process.

2.4.2 Linear Uncertainty and Error Variance

Linear uncertainty analysis is based on linear and Gaussian assumptions, which are implicit in the equations above by defining the least squares problem and by linearizing the forward model. As these assumptions are not valid for most cases in groundwater modelling, the estimation of linear parameter and predictive uncertainty is only an approximation. However, it is significantly cheaper computationally if the sensitivity matrix is already available from the inversion process (or from the first inversion iteration), compared to nonlinear methods. Solving the least squares problem of Equation 2.77 with $\mu = 1$, the posterior covariance matrix of the model parameters is given by (Tarantola, 2005; Doherty, 2015):

$$\mathbf{C}'_\mathbf{x} = (\mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{G} + {\mathbf{C_x}}^{-1})^{-1}. \tag{2.83}$$

Using the Woodbury corollaries, the posterior covariance matrix can be rewritten as

$$\mathbf{C}'_\mathbf{x} = \mathbf{C_x} - \mathbf{C_x}\mathbf{G}^T(\mathbf{G}\mathbf{C_x}\mathbf{G}^T + \mathbf{C_d})^{-1}\mathbf{G}\mathbf{C_x}. \tag{2.84}$$

Equation 2.83 is more computationally efficient than Equation 2.84 when the number of model parameters is smaller than the number of observations, and Equation 2.84 is more efficient when the opposite is true.
A resolution operator $\mathbf{R}$ can be defined as follows (Tarantola, 2005):

$$\mathbf{R} = \mathbf{C_x}\mathbf{G}^T(\mathbf{G}\mathbf{C_x}\mathbf{G}^T + \mathbf{C_d})^{-1}\mathbf{G}, \tag{2.85}$$

Assuming perfect data (no noise), the posterior covariance matrix can be rewritten as

$$\mathbf{C}'_\mathbf{x} = (\mathbf{I} - \mathbf{R})\mathbf{C_x}. \tag{2.86}$$

According to Tarantola (2005), this equation shows that, under the absence of measurement noise, if the resolution operator is close to the identity matrix, the posterior covariance matrix is close to zero, therefore the inverse problem is close to being completely determined.
Using the Woodbury corollaries, the resolution operation of Equation 2.85 can be rewritten as

$$\mathbf{R} = (\mathbf{G}^T{\mathbf{C_d}}^{-1}\mathbf{G} + {\mathbf{C_x}}^{-1})^{-1}\mathbf{G}^T{\mathbf{C_d}}^{-1}\mathbf{G}. \tag{2.87}$$

Note that the prior covariance matrix $\mathbf{C_x}$ appears in the resolution operator, which means that the resolution of the parameters is dependent on the prior uncertainty of the model parameters.
Assuming that a prediction $\mathbf{s}$ can be derived from the linearized forward model as

$$\mathbf{s} = \mathbf{g}^T\mathbf{x}, \tag{2.88}$$

where $\mathbf{g}$ is the sensitivity vector of the prediction, the posterior predictive uncertainty $\sigma'_\mathbf{s}$ is given by:

$$\sigma'_\mathbf{s} = \mathbf{g}^T\mathbf{C}'_\mathbf{x}\mathbf{g}. \tag{2.89}$$

Replacing Equation 2.83 in Equation 2.89, the posterior predictive uncertainty can be rewritten as (Doherty, 2015):

$$\sigma'_\mathbf{s} = \mathbf{g}^T(\mathbf{G}^T{\mathbf{C_d}}^{-1}\mathbf{G} + {\mathbf{C_x}}^{-1})^{-1}\mathbf{g} \tag{2.90}$$

A more computationally efficient version of Equation 2.90 when $n > m$ is given by:

$$\sigma'_\mathbf{s} = \mathbf{g}^T\mathbf{C_x}\mathbf{g} - \mathbf{g}^T\mathbf{C_x}\mathbf{G}^T(\mathbf{G}\mathbf{C_x}\mathbf{G}^T + \mathbf{C_d})^{-1}\mathbf{G}\mathbf{C_x}\mathbf{g}. \tag{2.91}$$

Replacing the resolution operator in Equation 2.91, the posterior predictive uncertainty can be rewritten as:

$$\sigma'_\mathbf{s} = \mathbf{g}^T(\mathbf{I} - \mathbf{R})\mathbf{C_x}\mathbf{g}. \tag{2.92}$$

Error variance and specifically predictive error variance equations are different from linear uncertainty analysis. First, error variance is defined as the variance of the difference between the true value and the predicted value, whereas uncertainty is defined as the variance of the predicted value. When performing regularized inversion, the objective is to minimize the parameter error variance, which is to minimize the propensity of the model parameters to depart from their unknown true values. This is propensity for bias, and it depends on the regularization strategy and parameter transformation (specifically Kahunen-Loeve transformation) (Doherty, 2015).
Parameter error can be defined as the difference between the true value and the estimated value. Using Equation 2.70 and Equation 2.80, parameter error can be derived as follows:

$$\begin{aligned}
\underline{\mathbf{x}} &= \mathbf{C_x G}^T(\mathbf{G C_x G}^T + \mathbf{C_d})^{-1}\mathbf{d} \\
&= \mathbf{C_x G}^T(\mathbf{G C_x G}^T + \mathbf{C_d})^{-1}(\mathbf{Gx} + \boldsymbol{\epsilon}) \\
&= \mathbf{C_x G}^T(\mathbf{G C_x G}^T + \mathbf{C_d})^{-1}\mathbf{Gx} + \mathbf{C_x G}^T(\mathbf{G C_x G}^T + \mathbf{C_d})^{-1}\boldsymbol{\epsilon} \\
\underline{\mathbf{x}} - \mathbf{x} &= \mathbf{Rx} + \mathbf{L}\boldsymbol{\epsilon} - \mathbf{x} \\
\underline{\mathbf{x}} - \mathbf{x} &= -(\mathbf{I} - \mathbf{R})\mathbf{x} + \mathbf{L}\boldsymbol{\epsilon},
\end{aligned} \tag{2.93}$$

where $\mathbf{L}$ is defined as follows:

$$\begin{aligned}
\mathbf{L} &= \mathbf{C_x G}^T(\mathbf{G C_x G}^T + \mathbf{C_d})^{-1} \\
&= (\mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{G} + {\mathbf{C_x}}^{-1})^{-1}\mathbf{G}^T {\mathbf{C_d}}^{-1}.
\end{aligned} \tag{2.94}$$

The second expression is similar to matrix $\mathbf{G}$ defined in Doherty (2015). The covariance matrix of the parameter error is given by (Doherty, 2015):

$$\mathbf{C}_{\underline{\mathbf{x}}-\mathbf{x}} = (\mathbf{I} - \mathbf{R})\mathbf{C_x}(\mathbf{I} - \mathbf{R})^T + \mathbf{L C_d L}^T. \tag{2.95}$$

The predictive error variance can be calculated from the difference between the true value and the predicted value as follows:

$$\begin{aligned}
\underline{\mathbf{s}} &= \mathbf{g}^T\underline{\mathbf{x}} \\
\underline{\mathbf{s}} - \mathbf{s} &= \mathbf{g}^T(\underline{\mathbf{x}} - \mathbf{x}) \\
\sigma^2_{\underline{s}-s} &= \mathbf{g}^T\mathbf{C}_{\underline{\mathbf{x}}-\mathbf{x}}\mathbf{g} \\
\sigma^2_{\underline{s}-s} &= \mathbf{g}^T(\mathbf{I} - \mathbf{R})\mathbf{C_x}(\mathbf{I} - \mathbf{R})^T\mathbf{g} + \mathbf{g}^T\mathbf{L C_d L}^T\mathbf{g}.
\end{aligned} \tag{2.96}$$

This equation is general. It depends on the regularization strategy through $\mathbf{R}$ and $\mathbf{L}$ (Doherty, 2015). The resolution operator $\mathbf{R}$ and the matrix operator $\mathbf{L}$ as defined in Equation 2.87 and Equation 2.94 are the result of Tikhonov regularization. A different definition will be provided for SVD regularization.
At least from a mathematical point of view, from comparing equations Equation 2.86

and Equation 2.95, it is clear that the posterior covariance matrix is not equal to the covariance matrix of parameter error. The same can be said about the predictive error variance and the posterior predictive uncertainty by comparing equations Equation 2.92 and Equation 2.96. Doherty (2015) demonstrated that the predictive error variance is greater than the posterior predictive uncertainty and that predictive uncertainty is immune to parameter transformation.

When using SVD as a regularization strategy, parameter and predictive error variance can be calculated from equations Equation 2.95 and Equation 2.96 but with the resolution operator $\mathbf{R}$ defined as (Moore and Doherty, 2005):

$$\mathbf{R} = \mathbf{Z}_1\mathbf{Z}_1^T, \tag{2.97}$$

and the matrix $\mathbf{L}$ is defined as

$$\mathbf{L} = \mathbf{Z}_1\boldsymbol{\Sigma}_1^{-1}\mathbf{Z}_1^T\mathbf{G}^T{\mathbf{C}_\mathbf{d}}^{-1}, \tag{2.98}$$

where $\boldsymbol{\Sigma}_1$ is the diagonal matrix of the first k singular values. The matrix $\mathbf{Z}_2$ contains the remaining $m-k$ columns of the right singular vectors. Then the expression $\mathbf{I}-\mathbf{R}$ can be expressed as

$$\mathbf{I} - \mathbf{R} = \mathbf{Z}_2\mathbf{Z}_2^T. \tag{2.99}$$

Predictive error variance can then be estimated as:

$$\sigma^2_{\underline{s}-s} = \mathbf{g}^T\mathbf{Z}_2\mathbf{Z}_2^T\mathbf{C}_\mathbf{x}\mathbf{Z}_2\mathbf{Z}_2^T\mathbf{g} + \mathbf{g}^T\mathbf{Z}_1\boldsymbol{\Sigma}_1^{-1}\mathbf{Z}_1^T\mathbf{g}. \tag{2.100}$$

Moore and Doherty (2005) showed the significance of the first and second terms in an equation similar to equations Equation 2.96 and Equation 2.100. The first term is the null space contribution to predictive error variance, i.e., the remaining uncertainty in the prediction that cannot be resolved by the data. When no singular values are used or Tikhonov regularization is maximized, the first term is maximized and equal to the prior predictive uncertainty. When the maximum number of singular values are used or Tikhonov regularization is minimized, the first term is minimized (although not necessarily zero). The second term is the solution space contribution to predictive error variance; this is the cost of measurement noise. As more parameters are adjusted to fit the data or Tikhonov regularization is dampened, there is greater propensity for increased predictive error variance, as measurement noise is amplified. As a result, the sum of the first term and the second term reaches a minimum that defines the optimum number of singular values or Tikhonov regularization parameter to use in the inversion process.

2.5 Data Space Inversion (DSI)

Data Space Inversion (DSI) is a method that, as the name suggests, performs inversion in the data space (model output space), using a number N of model output realizations. The simplest version of DSI is to apply the conditional expectation and covariance of model outputs given the data (see Koch, 1999), assuming Gaussian and linear conditions. This is the same as minimizing the least squares problem:

$$J_{\mathbf{DSI}}(\mathbf{o}) = \frac{1}{2}(\mathbf{d} - \mathbf{Ho})^T {\mathbf{C_d}}^{-1}(\mathbf{d} - \mathbf{Ho}) + \frac{1}{2}(\mathbf{o} - \mathbf{o}_f)^T {\mathbf{C_o}}^{-1}(\mathbf{o} - \mathbf{o}_f), \tag{2.101}$$

where $\mathbf{o}$ is the vector of model outputs, $\mathbf{H}$ is the mapping of model outputs to measurement space, $\mathbf{C_o}$ is the model output prior covariance matrix, and $\mathbf{o}_f$ is the prior estimate of the model outputs, which is generally assumed as the mean. The solution to the DSI problem can be easily obtained from Equation 2.80 by replacing $\mathbf{x}$ by $\mathbf{o}$, $\mathbf{G}$ by $\mathbf{H}$, and $\mathbf{C_x}$ by $\mathbf{C_o}$, as follows:

$$\underline{\mathbf{o}} = \mathbf{C_o}\mathbf{H}^T(\mathbf{HC_o}\mathbf{H}^T + \mathbf{C_d})^{-1}\mathbf{d}, \tag{2.102}$$

Note that $\mathbf{d}$ and $\underline{\mathbf{o}}$ are defined as departures from $\mathbf{Ho}_f$ and $\mathbf{o}_f$, respectively. Using the Woodbury corollaries, Equation 2.102 can be rewritten as

$$\underline{\mathbf{o}} = (\mathbf{H}^T {\mathbf{C_d}}^{-1}\mathbf{H} + {\mathbf{C_o}}^{-1})^{-1}\mathbf{H}^T {\mathbf{C_d}}^{-1}\mathbf{d}. \tag{2.103}$$

The posterior covariance matrix of the model outputs is given by:

$$\mathbf{C'_o} = (\mathbf{H}^T {\mathbf{C_d}}^{-1}\mathbf{H} + {\mathbf{C_o}}^{-1})^{-1}. \tag{2.104}$$

This equation can be written as

$$\mathbf{C'_o} = \mathbf{C_o} - \mathbf{C_o}\mathbf{H}^T(\mathbf{HC_o}\mathbf{H}^T + \mathbf{C_d})^{-1}\mathbf{HC_o}. \tag{2.105}$$

It is noted that model outputs might also include predictions. With this, posterior predictive uncertainty can be easily obtained from Equation 2.104, by extracting the diagonal elements of the matrix. Both the prior mean and covariance matrix of the model outputs can be obtained from an ensemble of model output realizations.

Sun and Durlofsky (2017) proposed an extended DSI procedure for non-Gaussian cases, that includes transformation of the model outputs to a space that is approximately Gaussian, and reparametrization of the data space using principal component analysis (PCA). First the re-parameterization strategy will be presented, and then transformation options of the model outputs will be discussed.

Let $\mathbf{Y}$ be the matrix of N model output anomalies (departures from the mean) including

matrix scaling, as defined in equation Equation 2.22. The matrix $\mathbf{Y}$ can be decomposed using PCA as follows:

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \boldsymbol{\Phi}\mathbf{V}^T. \tag{2.106}$$

Now model outputs can be modelled as a linear model of the mean model output $\mathbf{o}_f$ plus linear combination of $\boldsymbol{\Phi}$, the square root of the model output covariance matrix $\mathbf{C_o}$, as follows:

$$\mathbf{o} = \mathbf{o}_f + \boldsymbol{\Phi}\mathbf{z}, \tag{2.107}$$

where $\mathbf{z}$ is the vector of standard normal random variables, i.e., with a mean of zero and a variance of one. The size of the vector $\mathbf{z}$ is equal to the number of principal components used in the model, which can be defined by an eigenvalue or level of energy threshold. Now model outputs can be simulated with this simple linear model, and the least squares problem of Equation 2.101 can be solved by any history matching algorithm. For example, Lima et al. (2020) used the ensemble smoother with multiple data assimilation (ES-MDA) algorithm to solve the DSI problem, including localization to remove spurious correlations and increase degrees of freedom in the inversion process. In particular, as the model of Equation 2.107 is fast, the posterior of vector $\mathbf{z}$ can be fully sampled using a Markov Chain Monte Carlo (MCMC) algorithm. The posterior of the model outputs can then be obtained by applying Equation 2.107. Within the model output vector $\mathbf{o}$, one or more predictions of interest can be included, so posterior predictive uncertainty can be quantified.

The effectiveness of DSI depends on how proximate the model outputs are to a Gaussian distribution, and how well the principal components represent the model outputs. For this reason, Sun and Durlofsky (2017), Sun et al. (2017),and Jiang et al. (2021) have proposed transformations of the model outputs to a space that is approximately Gaussian. Among them, the histogram transformation of Sun et al. (2017) is the most straightforward strategy. This method, performs an inverse Gaussian anamorphosis procedure using the empirical cumulative distribution function (CDF) of the model outputs compared to prior realizations of the model outputs obtained from Equation 2.107. In this way, any model output modelled by Equation 2.107 will only result in outputs within the range of the prior model outputs. The main limitation of this transformation is that it does not take into account the correlation between model outputs, as the empirical CDF is calculated independently for each model output.

2.6 Numerical Examples

Two examples are presented to compare a selection of methods described above, in terms of convergence and uncertainty quantification capacity. The first example is a simple one-parameter nonlinear problem presented in Chen and Oliver (2013) and the second

example pertains to a one-dimensional unsaturated groundwater flow model.
MCMC was implemented using the Python package pyDREAM (Laloy and Vrugt, 2012), and regularized inversion, when applied, was implemented using the PEST software suite (Doherty, 2023). The remaining ensemble and DSI methods were implemented using Python codes developed by the author. Correlation-based localization as explained in subsection 2.3.6 was also tested for ensemble methods.

2.6.1 One parameter nonlinear problem

The objective of this example is to compare the performance of some of the ensemble methods discussed above, and verify equivalence with the results presented by Chen and Oliver (2013). The latter provides validation of the numerical implementation performed in this work. PEST was used to verify consistency with the ensemble methods results.
A parameter x has a Gaussian prior distribution with a mean of -2.0 and variance of 1.0. An ensemble of 1000 parameter realizations sampled from the prior was used for all ensemble methods. The forward model is a one-parameter function defined as follows:

$$g(x) = \frac{7}{12}x^3 - \frac{7}{2}x^2 + 8x. \tag{2.108}$$

Although this equation does not have a physical meaning, it may be reflective of a groundwater model with a nonlinear relationship between the parameter and the model output. A measurement $d = 48$ has a Gaussian measurement noise with a variance of 16.0. Although the true posterior distribution of the parameter x can be calculated analytically, it was obtained using pyDREAM with 50,000 samples and 3 chains with a burn-in of 25,000 samples.
For the ensemble methods that use the GN algorithm (batch-EnRML and SEnRML), the initial step length was set to 0.7 and the step length factor was set to 2.0. For the ensemble methods that use the LM algorithm (LM-EnRML and LM-EnRML), the initial lambda was set to 1.0 and the lambda factor was set to 4.0. A level of energy threshold of 0.99 was used for all ensemble methods. PEST was configured in regularization mode for native parameters with a regularization weight factor estimated to achieve 10% of the data mismatch at each iteration. The option to continue iterations was activated to minimize the model mismatch while achieving a data mismatch near 1.0.
The maximum number of iterations was set to 25. A data mismatch relative reduction is used as a stopping criterion if this value falls below 0.01 after 3 consecutive iterations. For all ensemble methods the data mismatch relative reduction is calculated for the ensemble mean and standard deviation at each iteration, and both statistics must meet the criterium. A parameter maximum change of 0.001 during 3 consecutive iterations was also set as a stopping criterion for all methods.

Figure 2.1 shows the data mismatch box plot for four ensemble methods except LM-IES (approx). It can be observed that all methods converge with a mean data mismatch near 1.0. Generally there is a quick convergence of the data mismatch for methods GN-EnRML and SEnRML, with a mean data mismatch near 1.0 after 3 iterations. In particular, convergence for these two methods is equal, suggesting that solving the inverse problem in the ensemble subspace is equivalent to solving it in parameter space, in this case. The method LM-SEnRML is also quick to converge, with a mean data mismatch near 1.0 after 5 iterations, approximately. Of all methods, LM-EnRML has the slowest convergence rate, and it only improves the data mismatch after 9 iterations. This result is consistent with the results obtained by Chen and Oliver (2013) (see Figure 1 in their paper). This may be due to the fact that the prior ensemble is replaced by the updated ensemble at each iteration, whereas the other methods use the prior ensemble in the Hessian term of the parameter update equations. Also, the LM-EnRML complied with the stopping criterion only after 23 iterations due to non-stabilization of the data mismatch standard deviation.

A comparison of the convergence of the data mismatch for the methods LM-EnRML and LM-EnRML (approx) after 10 iterations is shown in Figure 2.2. It can be observed that the LM-EnRML (approx) method converges faster than the LM-EnRML method, with a mean data mismatch near 1.0 after 7 iterations. However, given that the model term was discarded in the parameter update equation, the data mismatch decreases monotonically below the target data mismatch, generating overfitting. Interestingly, the data mismatch is reduced abruptly below the target data mismatch between iterations 7 and 8, suggesting that the approximation of the model term in the parameter update equation is not adequate. It can be inferred that it would be difficult to judge what iteration to choose as the best estimate of the parameter.

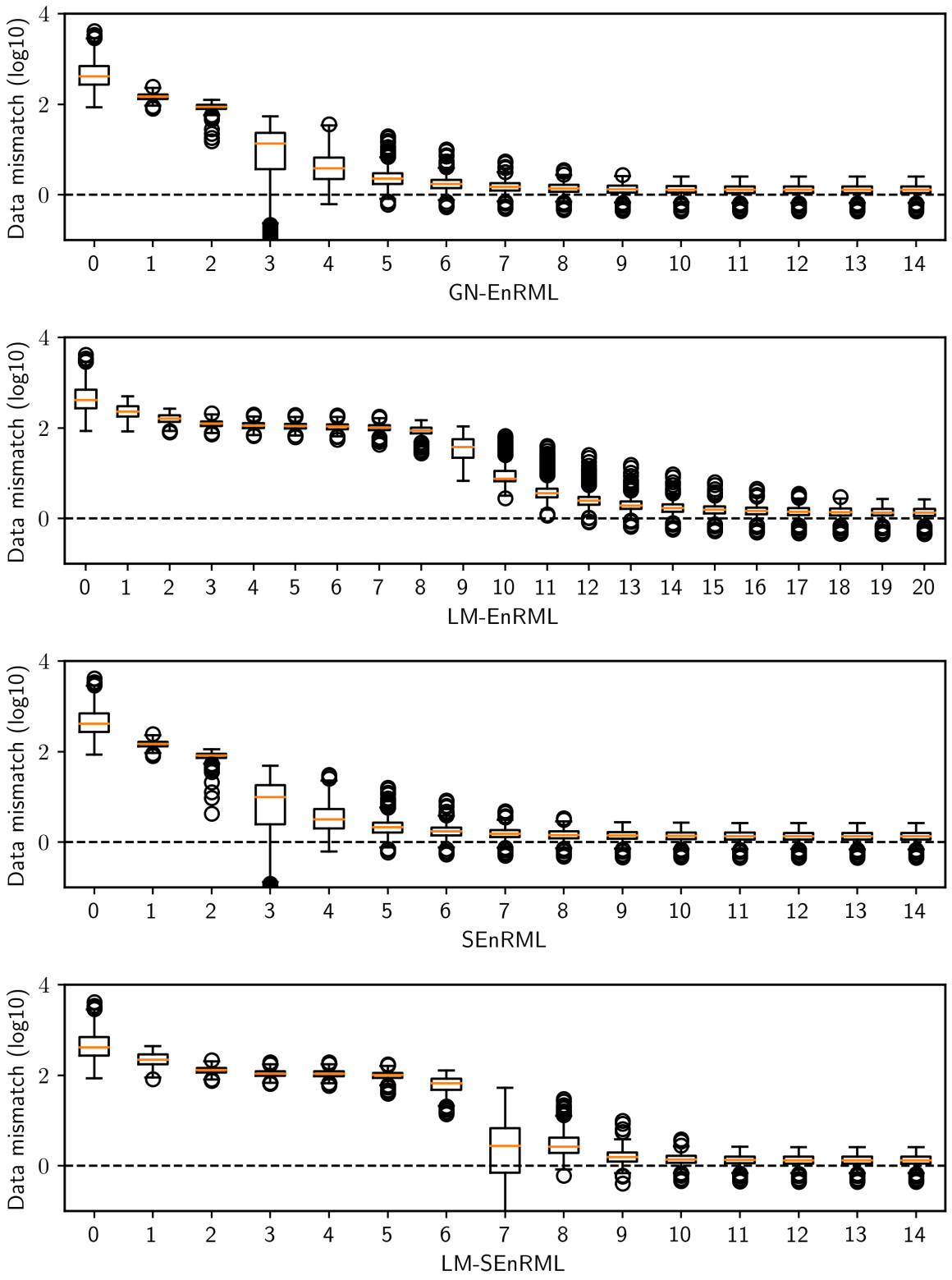

Figure 2.1: Data mismatch box plot for the one-parameter nonlinear problem. The boxes are built using the 25th and 75th percentiles, and the whiskers represent the 5th and 95th percentiles. The horizontal line inside the box represents the median. The black circles represent the outliers. The dashed horizontal line represents the target data mismatch of 1.0 (number of observations). Iteration 0 represents the initial data mismatch.

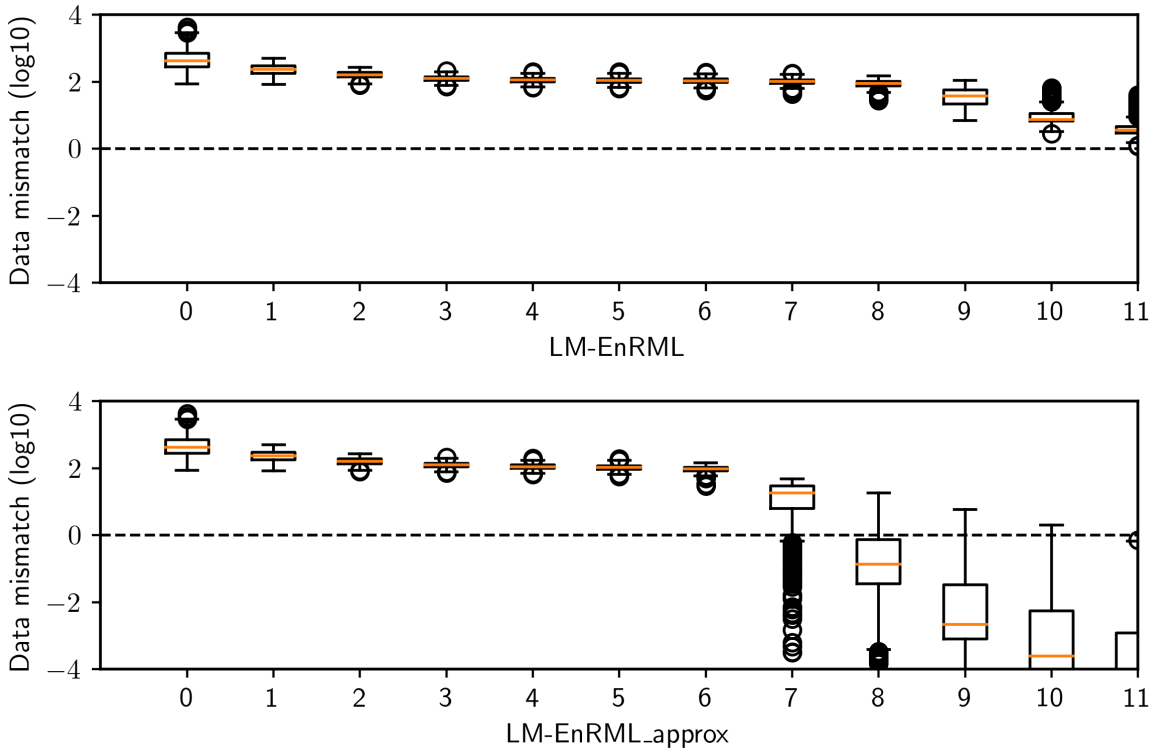

Figure 2.2: Data mismatch comparison of methods LM-EnRML and LM-EnRML (approx) for the one-parameter nonlinear problem after 10 iterations. The boxes are built using the 25th and 75th percentiles, and the whiskers represent the 5th and 95th percentiles. The horizontal line inside the box represents the median. The black circles represent the outliers. The horizontal line represents the target data mismatch of 1.0 (number of observations). Iteration 0 represents the initial data mismatch.

The posterior distribution of the parameter x derived from the ensemble methods (except LM-EnRML (approx)) is shown in Figure 2.3. It is observed that all methods provide a posterior distribution that matches the true posterior distribution obtained with MCMC. The true value of the parameter x is 6.0, and the posterior distribution is centred around 5.85, approximately. The same value was obtained with PEST. The difference, although minor, is likely due to how the the prior distribution was defined, centred around -2.0 and with a variance of 1.0, with a minimum support of the true value.

A comparison of the posterior distribution of the parameter x derived from the LM-EnRML and LM-EnRML (approx) methods is shown in Figure 2.4. It is clear that the posterior distribution of the parameter x derived from the LM-EnRML (approx) method is centred around the true value, and is not consistent with the true posterior distribution. This is an outcome of not including the model term in the parameter update equation.

Overall, these results validate the numerical implementation of the ensemble methods, at least for the one-parameter nonlinear problem presented, as they are consistent with the results obtained by Chen and Oliver (2013).

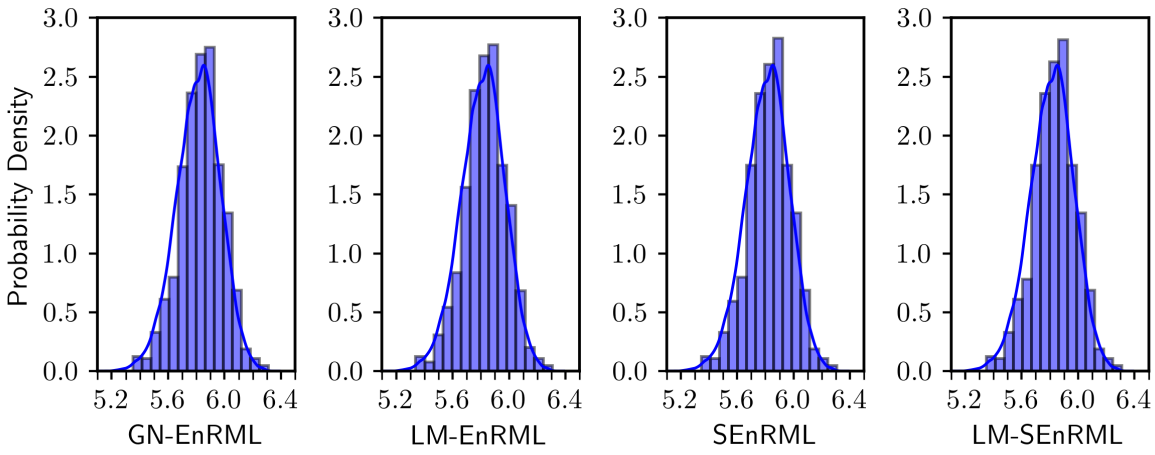

Figure 2.3: Distribution of ensemble of realizations at the end of the inversion process compared to the true posterior distribution of the parameter x (blue line).

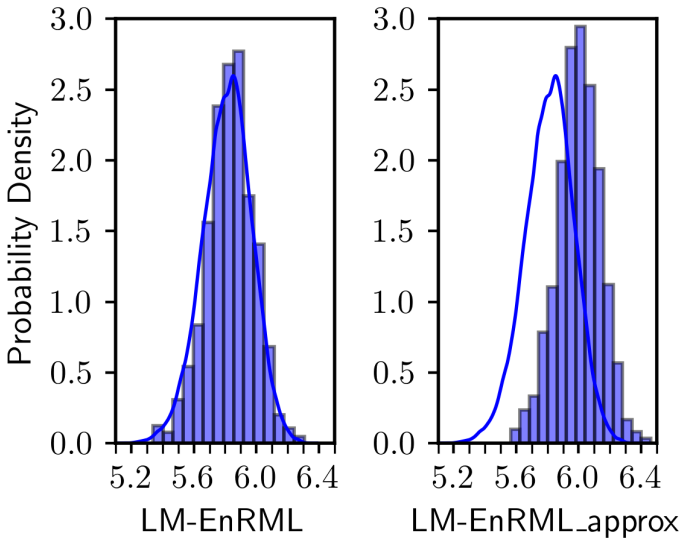

Figure 2.4: Distribution of ensemble of realizations at the end of the inversion process compared to the true posterior distribution of the parameter x (blue line).

2.6.2 1D unsaturated flow problem

The objective of this example is to compare the performance of the ensemble methods in a more complex and high-dimensional problem. The problem complexity originates from the nonlinearity of the Richards equation, which is used to simulate unsaturated flow. In particular, the relations between suction and effective saturation, and hydraulic conductivity and effective saturation, are nonlinear. The additional complexity comes from the high dimensionality of the problem, which is defined by the number of cells in the 1D domain and the number of parameters that define the hydraulic properties.

The 1D model represents the advance of a wetting front in a 100 cm long heterogeneous soil profile, discretized in 1 cm cells, and modelled using MODFLOW-USG transport (Panday, 2024; Panday et al., 2013). A 3-day precipitation event is simulated, with a constant rainfall rate of 5.0 cm/day. Free drainage is assumed at the bottom of the domain, using the specified gradient second-type boundary condition (McCord, 1991). The initial condition is a dry soil profile, which is the result of a 50-day drying period starting from a fully saturated soil. After the precipitation event, the soil profile is allowed to drain for 10 days, and the total cumulative infiltration simulated by the model is the prediction of interest. A total of 27 measurements of pressure head h are taken at 20.5, 40.5, and 60.5 cm from the bottom of the domain, every 0.3 days, from 0.5 days of the beginning of the precipitation event until day 3.

The relative permeability of the soil is simulated using the Brooks-Corey model (Brooks and Corey, 1966), as:

$$k_r = S_e^{bc}, \tag{2.109}$$

where k_r is the relative permeability, S_e is the effective saturation, and bc is the Brooks-Corey exponent. The effective saturation is calculated as:

$$S_e = \frac{\theta(h) - \theta_r}{\theta_s - \theta_r} = \frac{S(h) - S_r}{1 - S_r}, \tag{2.110}$$

where $\theta(h)$ is the volumetric water content at pressure head h, θ_r is the residual water content, θ_s is the saturated water content, $S(h)$ is the total saturation at pressure head h, and S_r is the residual saturation.

The effective saturation is related to pressure head by the van Genuchten model (van Genuchten, 1980), as:

$$S_e = \begin{cases} (1 + (\alpha h)^n)^{-m} & h < 0 \\ 1 & h \geq 0 \end{cases}, \tag{2.111}$$

where α (cm^{-1}) is the inverse of the air entry pressure, n is the van Genuchten exponent, and $m = 1 - 1/n$ is the inverse of the pore size distribution. It is assumed that the soil is heterogeneous to the cell level for the parameters θ_r, θ_s, α, n, and K_s (saturated

hydraulic conductivity). An exponential variogram model with a range of 15 cells is used to simulate the spatial correlation of the parameters, for each parameter type. A covariance matrix was generated using this exponential variogram model.

As discussed by Scharnagl et al. (2011), van Genuchten parameters and saturated hydraulic conductivity are a-priori correlated. For conceptual consistency, a correlation matrix for these parameters was generated using ROSETTA (Zhang and Schaap, 2017), a neural network-based model that predicts unsaturated soils hydraulic parameters from soil texture data such as percentage of sand, silt, and clay, and bulk density. A total of 1000 samples of percentages of sand, silt, and clay, were generated using a Dirichlet distribution, from which the van Genuchten parameters and saturated hydraulic conductivity were sampled using ROSETTA version 3 (Zhang and Schaap, 2017), available through rosetta-soil python library (Skaggs, 2024). A correlation matrix was then generated using the Pearson correlation coefficient, and presented in Table 2.1.

Table 2.1: Correlation matrix for van Genuchten parameters and saturated hydraulic conductivity.

	θ_r	θ_s	$\log_{10}(\alpha)$	$\log_{10}(n)$	$\log_{10}(K_s)$
θ_r	1.00				
θ_s	0.75	1.00			
$\log_{10}(\alpha)$	-0.19	-0.18	1.00		
$\log_{10}(n)$	-0.84	-0.50	-0.25	1.00	
$\log_{10}(K_s)$	-0.62	-0.18	0.15	0.71	1.00

It is recognized that this approach is simplistic, and that the correlation between parameters obtained from random samples of different soil textures is not necessarily correct, as correlations depend on the soil texture itself. In a real case, one could generate samples of soil textures from a multivariate distribution centred around the mean soil texture representative of the site.

A multi-Gaussian prior distribution was defined for each $\log_{10}$-transformed parameter type, except for the Brooks-Corey exponent bc for which a constant value of 4.0 was assumed. Each parameter consists of 100 samples, one for each cell in the 1D domain. To preserve the relation between residual water content θ_r and saturated water content θ_s, residual water content was calculated as a factor of saturated water content f_{θ_r}, as follows:

$$\theta_r = f_{\theta_r}\theta_s. \tag{2.112}$$

Then, a prior distribution was defined for the $\log_{10}$-transformed factor f_{θ_r} from which the residual water content was calculated. The mean and standard deviation of the prior distribution for each parameter are presented in Table 2.2.

The cross-correlation between parameter types was obtained by applying the Cholesky decomposition of the correlation matrix to the multi-Gaussian samples. A prior parameter

Table 2.2: Prior distribution for the parameters of the 1D unsaturated flow problem.

Parameter	Mean	Standard deviation
$\log_{10}(f_{\theta_r})$	-1.00	0.15
$\log_{10}(\theta_s)$	-0.64	0.1
$\log_{10}(\alpha)$	-1.0	0.5
$\log_{10}(n)$	0.2	0.1
$\log_{10}(K_s)$	0.2	0.5

ensemble of 300 samples was generated for a total of 500 parameters. One parameter set was selected as the true parameter set, and the forward model outputs were used as the measurements to be history-matched. The true parameter set was chosen so that the simulated suctions at the measurement locations were extreme values, to test the capacity of the ensemble methods under these difficult conditions.

Figure 2.5 shows pairwise scatter plots of the prior parameter samples and also the true values. It can be observed that some parameter correlations are evident.

Prior realizations of model outputs were generated using the parameter prior ensemble, and the suction outputs for the selected measurement locations are presented in Figure 2.6. It is noted that the suction values go below zero for some observation points. This adds another level of complexity to the problem, as once the pressure head goes above zero, van Genuchten parameters are not correlated with heads in saturated conditions, for obvious reasons.

The histogram of predicted cumulative infiltration throughout the simulation time, derived from the prior runs, including the true value resulted from the model run with the true parameter values, is shown in Figure 2.7. As previously stated, this is the prediction of interest.

History matching of the 27 observations with Gaussian noise of 1.0 cm was performed using the ensemble methods LM-EnRML, SEnRML, and LM-SEnRML. In this case, the prior parameter ensemble size was 200. Increasing the ensemble size to 300 did not significantly improve the results (not shown here). To later evaluate the effect of ensemble size on correlation noise, an additional parameter ensemble of size 2000 was also generated. To facilitate model inversion in this highly nonlinear problem, history matching parameters where defined as standard deviates of each parameter type, for each model cell. Spatial correlation between parameters of the same type was generated by transforming the standard deviates to the original parameter space using the following equation:

$$\mathbf{x}_i = \overline{\mathbf{x}}_i + \mathbf{E}\mathbf{F}^{1/2}\mathbf{z}, \tag{2.113}$$

where $\mathbf{x}_i$ is the parameter vector of the parameter type i, $\overline{\mathbf{x}}_i$ is the mean of the parameter type i, $\mathbf{z}$ is a vector of standard deviates (of size 100), and $\mathbf{E}$ and $\mathbf{F}$ are the result of SVD on the covariance matrix of the parameter type i, i.e., $\mathbf{C}_{\mathbf{x}i} = \mathbf{E}\mathbf{F}\mathbf{E}^T$.

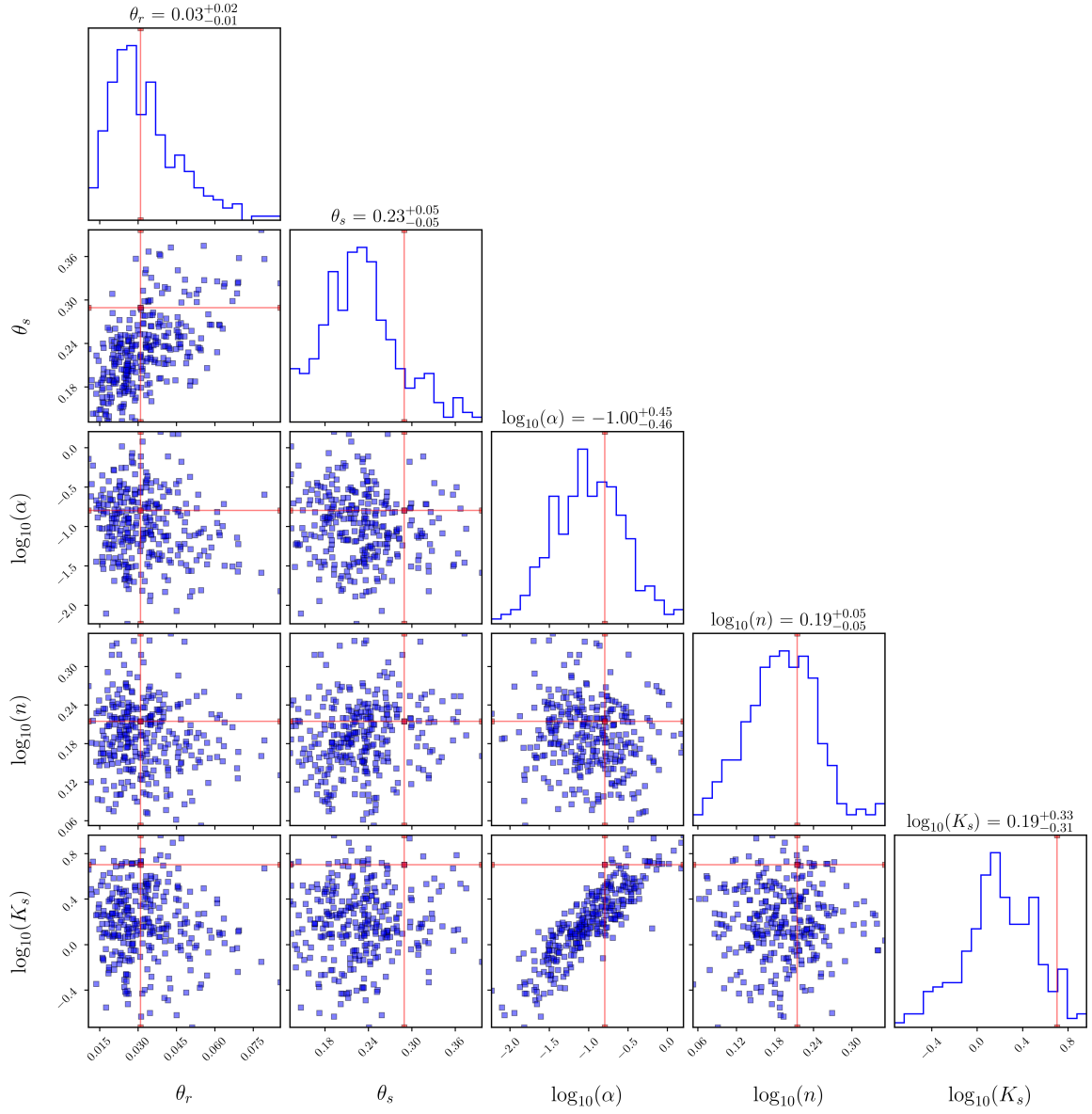

Figure 2.5: Pairwise scatter plots of the prior samples of the parameters of the 1D unsaturated flow problem at z = 60.5 cm. The red points represent the true values of the parameters.

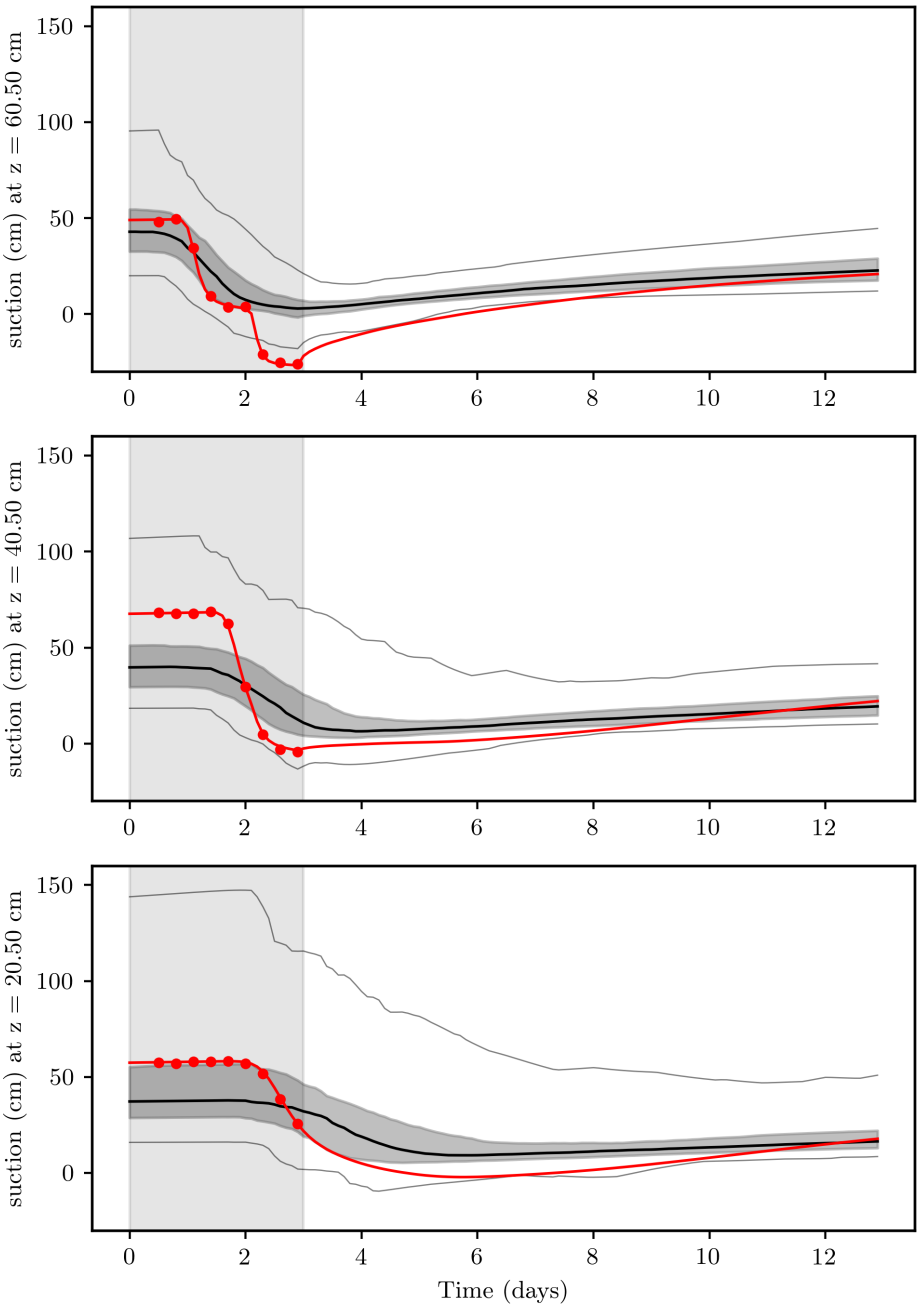

Figure 2.6: Suction and pressure head outputs for the selected measurement locations of the 1D unsaturated flow problem. The red lines with solid circles represent the true values of the pressure head, the solid black line is the median, the grey-shaded area is the P25-P75 percentile region, and the external black lines are the P5 and P95 percentiles.

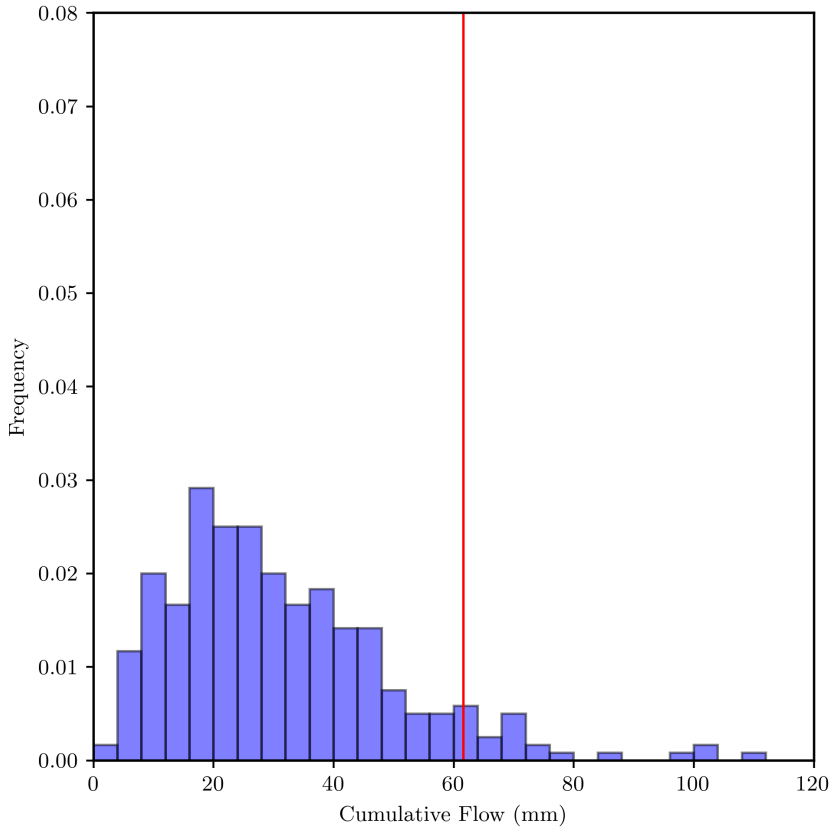

Figure 2.7: Histogram of cumulative infiltration throughout the simulation time derived from the prior runs of the 1D unsaturated flow problem.

At this point parameter types are not correlated. The correlation between parameter types was then obtained by applying the Cholesky decomposition of the correlation matrix, presented above. The result of this process is a correlated parameter ensemble used in the numerical model.

For the LM-EnRML and LM-SEnRML method, the initial Levenberg-Marquardt λ was calculated as (following Chen and Oliver (2013)):

$$\lambda = 10^{\text{Floor}(\log_{10}(\overline{J}/2m))}, \tag{2.114}$$

where $\overline{J}$ is the mean data mismatch, and m is the number of observations. For this particular problem, the resulting initial lambda is 100. The λ factor is reduced by a factor of 4.0 at each iteration if the data mismatch mean and standard deviation improve with respect to the previous iteration. If only the mean data mismatch improves between iterations, the λ factor is not changed. If the data mismatch mean does not improve, the λ factor is increased by a factor of 5.0. A maximum lambda value of 5.0x10^4 was set for the LM-EnRML and LM-SEnRML methods.

For the SEnRML method, the initial step length was set to 0.5, with minimum and maximum values of 0.01 and 0.6. If the data mismatch mean and standard deviation improve with respect to the previous iteration, the step length is increased according to the following equation:

$$\gamma = \gamma + (\gamma_{max} - \gamma) * 2^{-l/(\delta-1)}, \tag{2.115}$$

where γ_{max} is the maximum step length, l is the iteration number, and δ is a decay parameter. It is recognized that this is a simple heuristic to increase the step length, and that a different approach could be used. As explained by Evensen et al. (2022), the step length evolution over iterations influences the convergence of the method. Same as for the EnRML method, the γ damping factor is unchanged if only the mean data mismatch improves between iterations. If the data mismatch mean does not improve, the step length is decreased by a factor of 2.0.

The subset of ensemble members that improve the data mismatch and standard deviation respect to the previous iteration are accepted, even if the data mismatch mean does not improve, as implemented in PESTPP-IES (White, 2018). Iterations are stopped before reaching the total of 20 maximum iterations if the relative improvement of the data mismatch mean is less than 1x10^{-3}, or if the damping factors overcome their maximum or minimum values (depending on the method).

History matching results for the ensemble methods are compared with results obtained using DREAM accessed through the python library pyDREAM. In this case, the pyDREAM algorithm was configured with 20,000 samples and 3 chains, with a burn-in of 10,000 samples. The Gelman-Rubin convergence diagnostic was used to verify the convergence of the chains, and the results were considered valid if the diagnostic was less

than 1.2.

Table 2.3 presents the data mismatch means and standard deviations, and the number of iterations of the ensemble methods. As discussed by Chen and Oliver (2013), realizations are expected to achieve a data mismatch similar to the number of observations, with an upper bound as follows:

$$Sd \leq m + 5\sqrt{2m}, \tag{2.116}$$

where m is the number of observations. It is important to note that this is just a reference value given that it is based on Gaussian assumptions and linear models, and assumes that the model mismatch is negligible compared to the data mismatch. In any case, the estimated upper bound using Equation 2.116 is 63.7 for this problem. It is then observed that none of the ensemble methods achieved a data mismatch mean near the expected value. This is a consequence of the nonlinearity of the problem, the limited ensemble size, and the fact that the same average sensitivity matrix is used for all ensemble members. The LM-EnRML method achieved the best data mismatch mean, with a value of 283.8, and the LM-SEnRML method achieved the worst data mismatch mean, with a value of 1390.6.

Table 2.3: Data mismatch mean, standard deviation, and number of iterations of ensemble methods, resulted from history matching of the 1D unsaturated flow problem.

Method	Mean	Standard deviation	N Iterations
LM-EnRML	284	261	13
SEnRML	993	1104	12
LM-SEnRML	1391	1184	15

Figure 2.8 shows the data mismatch box plot for the ensemble methods. The slow convergence rate is apparent for all methods, where the data mismatch mean appears to stabilize after 10 iterations, approximately. The dispersion around the mean data mismatch is high, showing the issues in the performance of these ensemble methods to achieve good fits, for all the ensemble members.

Notwithstanding the high data mismatch mean, the posterior uncertainty of suction at the selected measurement locations is qualitatively consistent with the MCMC benchmark posterior, as shown in Figure 2.9. The reasonableness criterion is based on visual comparison of the posterior distributions obtained from each ensemble method against those from MCMC, which is used as a reference approximation of the true posterior. This is especially true when considering the number of forward model runs required to obtain the posterior distribution using ensemble methods, that vary between 2,400 and 3,000, compared to the MCMC method that required 60,000 runs (20,000 runs per chain) to achieve a fit commensurate with measurement noise (Figure 2.9). Among the ensemble methods, the LM-EnRML method achieved the best posterior distribution consistent with the data mismatch analysis, with a median that closely matches the observed val-

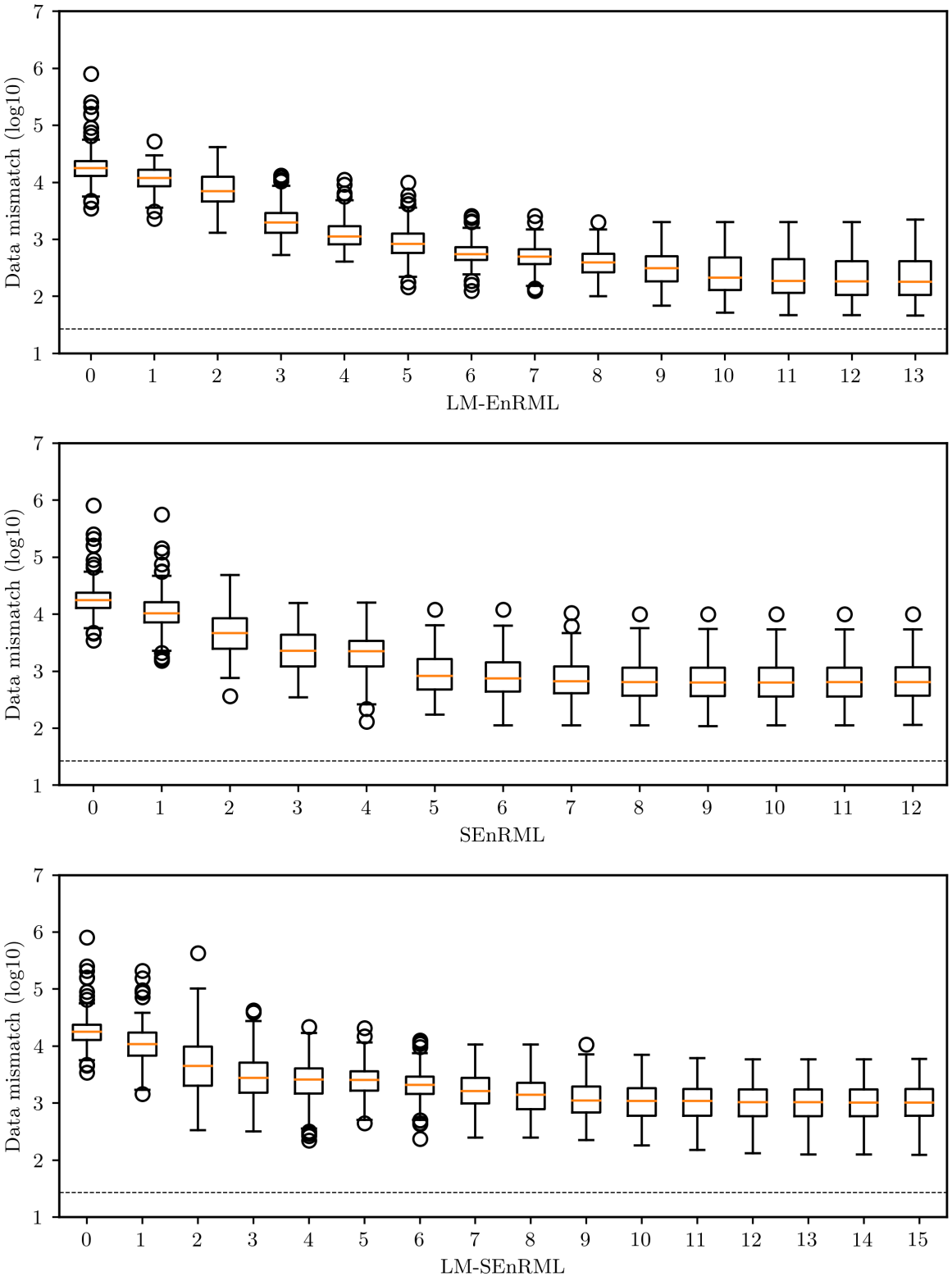

Figure 2.8: Data mismatch box plot evolution from different ensemble methods for the history matching of the 1D unsaturated flow problem. The horizontal line represents the target data mismatch of 27.0 (number of observations). Iteration 0 represents the initial data mismatch.

ues. On the contrary, the SEnRML and LM-SEnRML methods exhibit a wider posterior distribution, with a median that is offset from the observed values. This is interpreted as a consequence of the fact that SEnRML and LM-SEnRML methods perform history matching in the ensemble subspace, constrained by the prior parameter realizations. In the case of the LM-EnRML method, the replacement of the prior covariance matrix by the updated parameter ensemble in the hessian term of the parameter update equation and the use of the LM damping factor, may lead to at least two interrelated consequences: 1) the ensemble members are updated in a way that they are not totally constrained by the prior parameter ensemble, as the hessian term is updated with the updated anomaly matrix $\mathbf{A}_l$, that is a function of the residuals between $\mathbf{g}(\mathbf{X})$ and $\mathbf{d}$, and 2) exploration of new directions in the parameter space, may lead to a better fit to the data.

A practical evidence from the discussion above can be gathered using the results derived from this example. If the updated ensemble anomaly matrix $\mathbf{A}_l$ is a linear combination of the prior ensemble anomaly matrix $\mathbf{A}_0$, its projection onto the prior span should be the equal to the original matrix $\mathbf{A}_l$, i.e, $\mathbf{R} = (\mathbf{I} - \mathbf{U}_0\mathbf{U}_0^T)\mathbf{A}_l \approx 0$, where $\mathbf{R}$ is the parameter residual matrix, and $\mathbf{U}_0$ is the result of SVD on the prior parameter ensemble anomaly matrix $\mathbf{A}_0$. Equivalently, the norm of the parameter residual matrix to the total parameter anomaly matrix $\mathbf{A}_l$ should be small, i.e., $\mathbf{R}_r = \|\mathbf{R}\|/\|\mathbf{A}_l\| \approx 0$. The calculated values of $\mathbf{R}_r$ for the LM-EnRML and SEnRML yielded 0.04 and $4.52x10^{-9}$, respectively. This suggests that the LM-EnRML method explores new directions in the parameter space, while the SEnRML method is constrained by the prior parameter ensemble.

Even though the goodness of fit varies greatly between ensemble members, the posterior predictive uncertainty of cumulative infiltration estimated from the tested ensemble methods is reasonable, as shown in Figure 2.10, compared to the 'true' predictive uncertainty estimated from MCMC.

Of all the ensemble methods, the SEnRML method achieved the best posterior predictive uncertainty, with a median that closely matches the true value. On the contrary, the methods LM-EnRML and LM-SEnRML show more bias but still provide reasonable uncertainty estimates. These results can be initially thought of as counterintuitive. However, by looking at Figure 2.9, it is clear that the ensemble methods exhibit a good fit to the lower suction and higher pressure head values, which are most sensitive to cumulative infiltration.

Localization was tested for the ensemble methods. Before applying localization, a correlation noise analysis is presented. An ensemble of 2000 parameter realizations was generated, from which a corresponding ensemble of model outputs was obtained using the forward model. A correlation matrix was calculated using these results, herein referred as the true correlation matrix. One realization of correlation noise can be obtained by calculating the difference between the true correlation matrix and the correlation matrix estimated from the ensemble of limited size ($N = 200$). This is the ideal approach

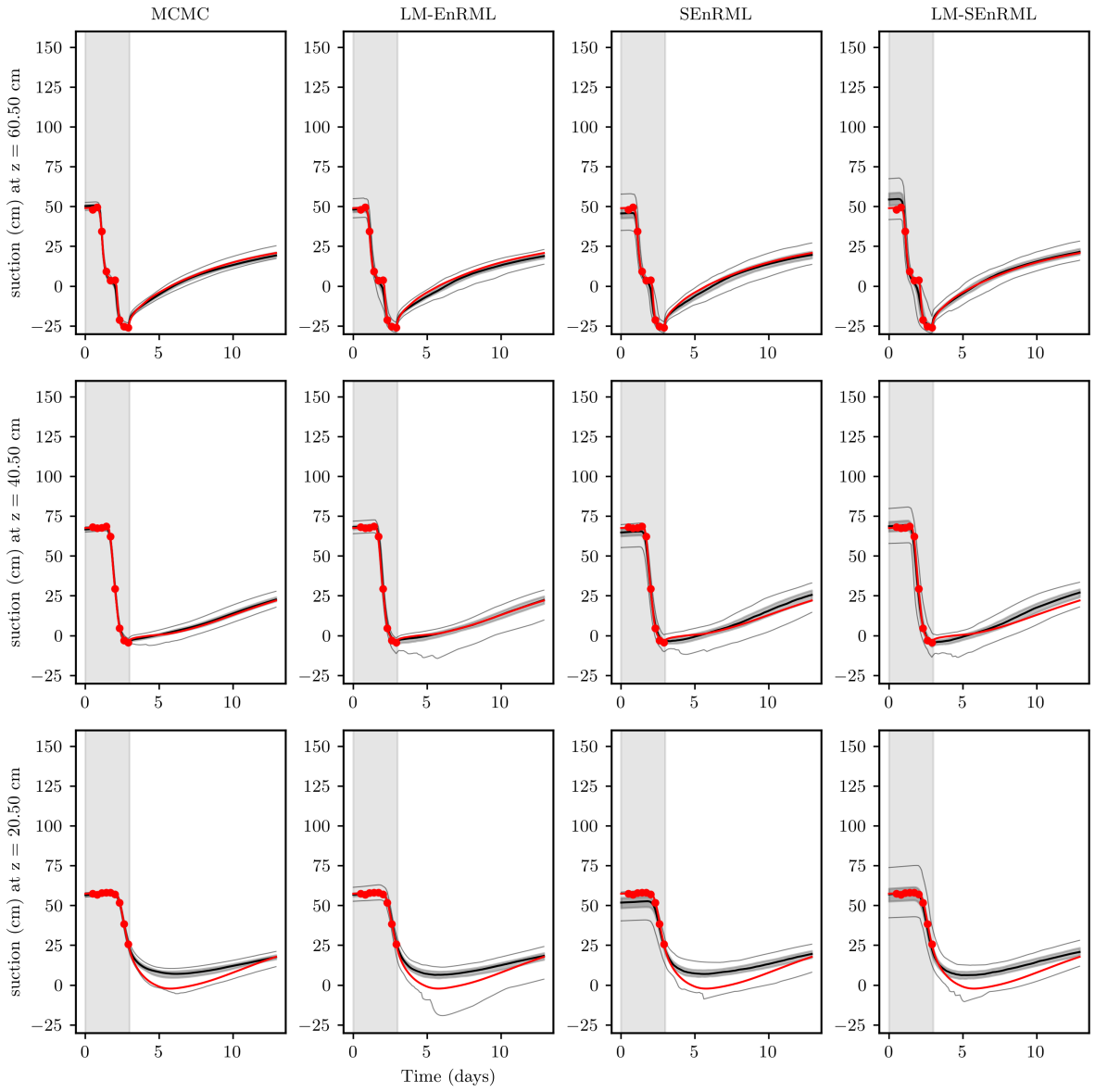

Figure 2.9: Posterior distribution of suction at the selected measurement locations of the 1D unsaturated flow problem, obtained with MCMC and ensemble methods. The solid circles represent the observed values of the pressure head and the red line extends them into predictive times. The solid black line is the median, the grey-shaded area is the P25-P75 percentile region, and the external grey lines are the P5 and P95 percentiles. The period of the precipitation event is represented by the grey-shaded area.

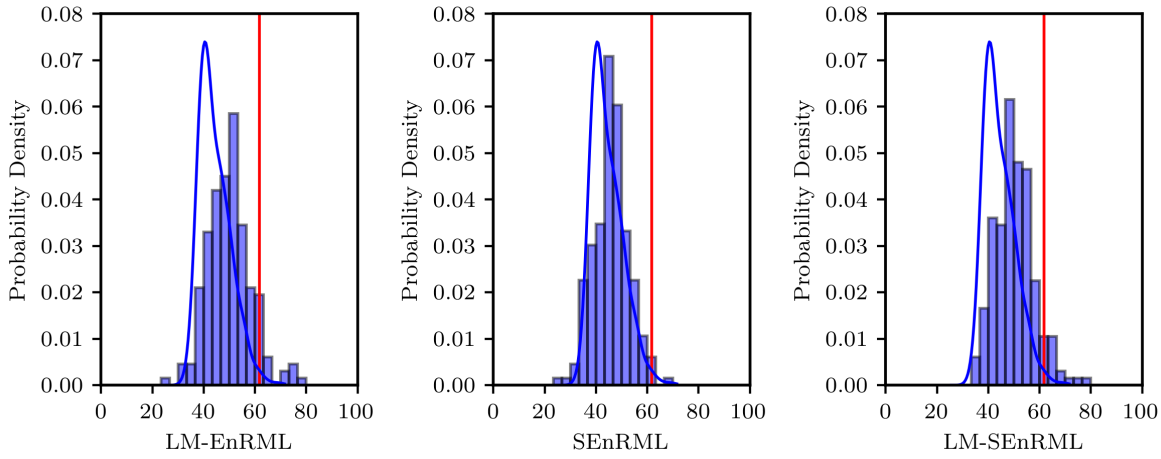

Figure 2.10: Histograms of posterior predictive uncertainty of total cumulative infiltration of the 1D unsaturated flow problem, obtained with ensemble methods. The true predictive uncertainty obtained from MCMC is represented by a blue pdf. The predictive cumulative infiltration derived from the true parameter set is represented by a red line.

of Luo and Bhakta (2020). Figure 2.11 presents a matrix block of the first 20 elements (parameters and observations) of the estimated and true correlation matrices, and the correlation noise (difference between them), for saturated hydraulic conductivity (Ksat), and alpha and beta Van Genuchten parameter types. The true correlation matrix shows that the first 3 to 5 parameters present relatively high correlation with the observations, with a low correlation (mostly near zero) for the remaining parameters. It can be observed that the estimated correlation matrix does capture these high correlations but with correlation noise.

As discussed by Luo and Bhakta (2020) and others, generating the true correlation matrix to estimate the noise of the ensemble correlation matrix adds computational burden that may be prohibitive for large problems. Because of this reason, the statistical properties of correlation noise should be estimated by other means, such as applying the random shuffle method of Luo and Bhakta (2020). Figure 2.12 presents realizations of correlation noise obtained from the ideal approach and the random shuffle method, for the first 20 matrix elements, for saturated hydraulic conductivity (Ksat), and alpha and beta Van Genuchten parameters.

It can be observed that the realization of correlation noise obtained from the random shuffle method visually shares the statistical properties of the correlation noise obtained from the ideal approach, consistent with the results presented by Luo and Bhakta (2020). As this is just a visual comparison, a more robust statistical verification could be performed by generating multiple samples of correlation noise using both methods. However, this goes beyond the scope of this work.

Using the random shuffle approach to generate samples of correlation noise, several local-

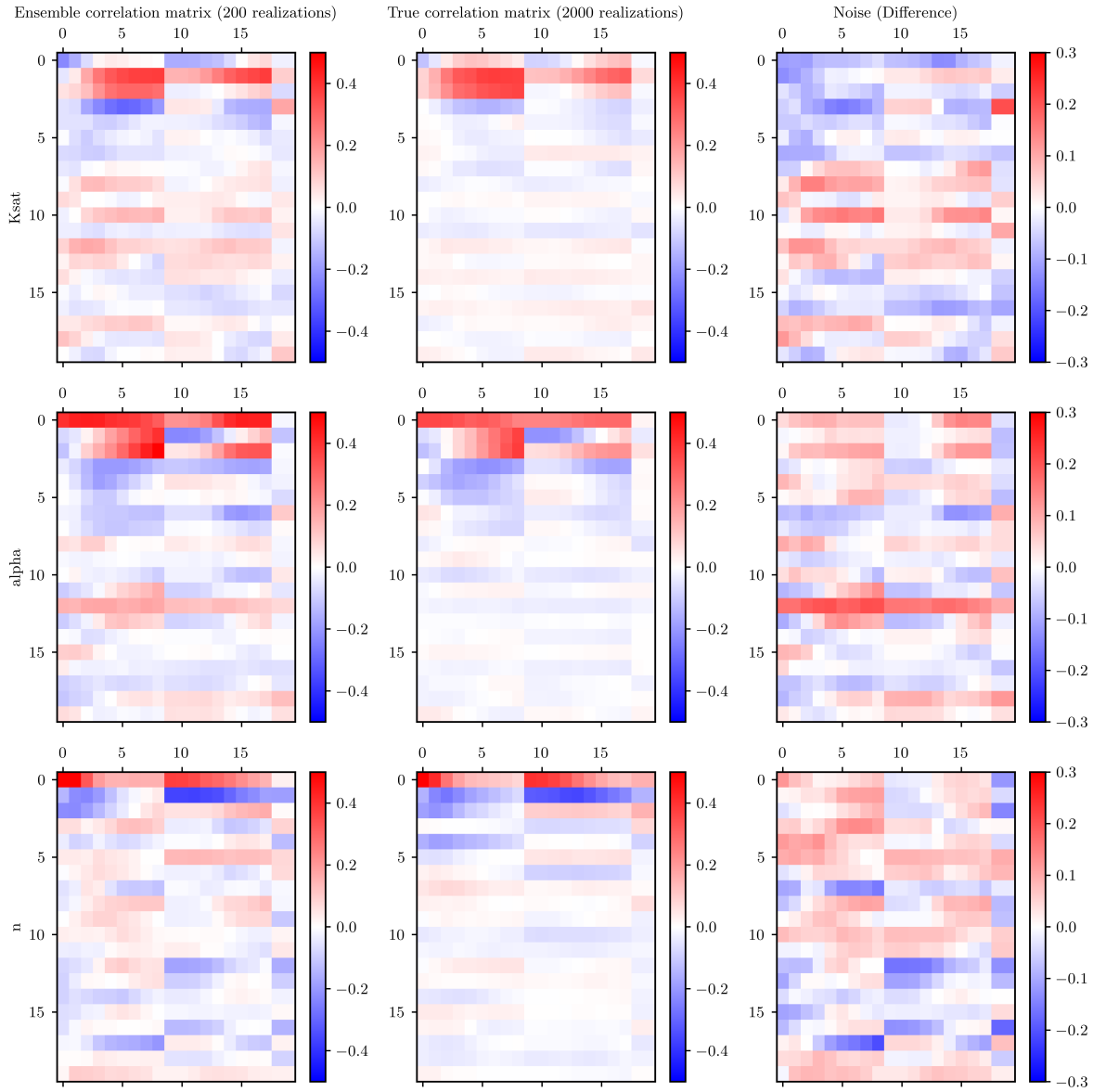

Figure 2.11: Comparison of the estimated and true correlation matrices, and the correlation noise for the first 20 elements (parameters and observations) for three selected parameter types: saturated hydraulic conductivity (Ksat), alpha and n Van Genuchten parameters.

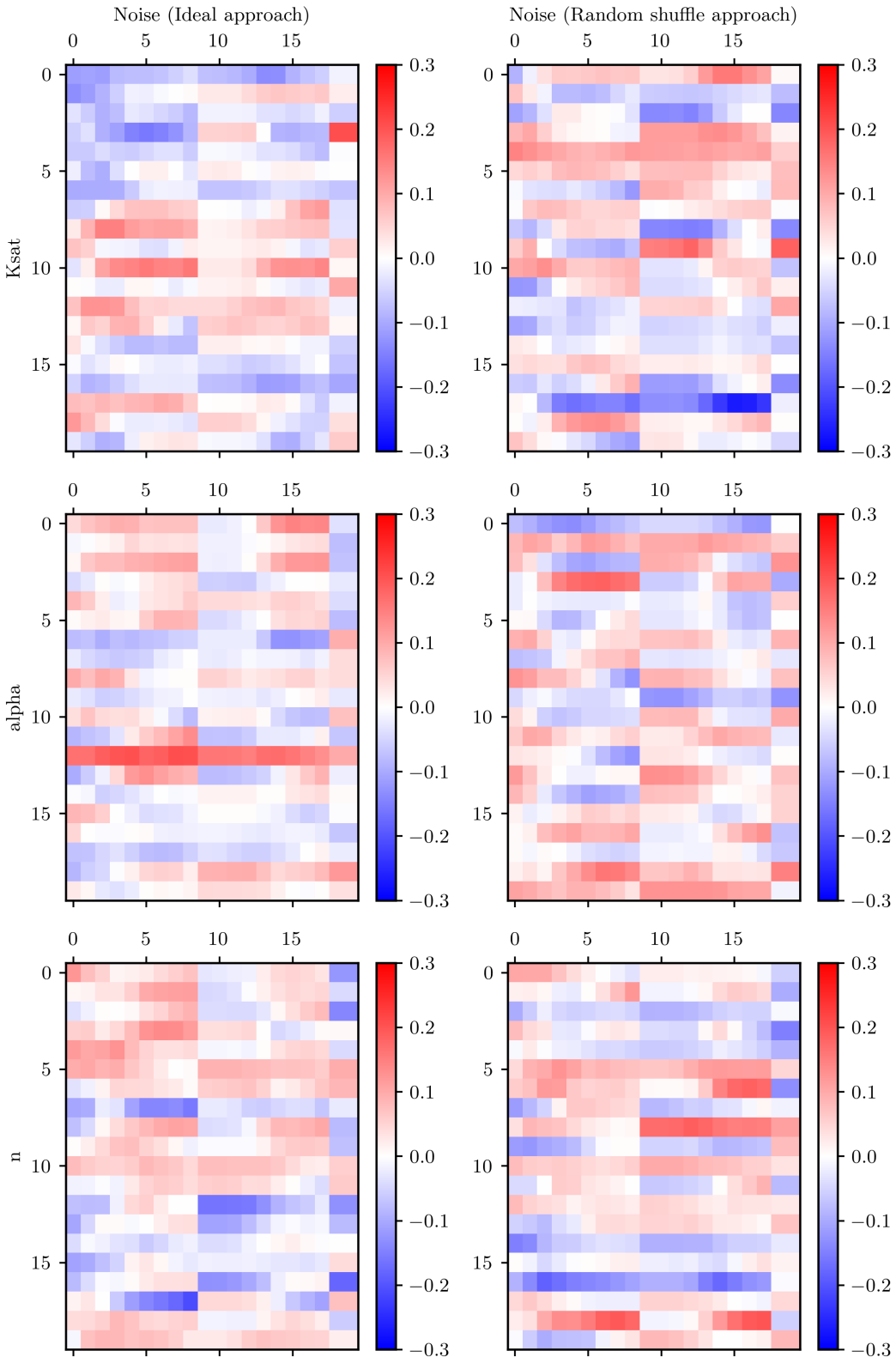

Figure 2.12: Comparison of the correlation noise obtained from the ideal approach and the random shuffle method for the first 20 matrix elements, for saturated hydraulic conductivity (Ksat), and alpha and beta Van Genuchten parameters.

ization schemes previously described were tested to generate the localization matrix, that can taper the estimated correlation matrix to reduce the correlation noise. The following alternatives were tested:

1. Random shuffle approach and GC function using the z dummy variables as in:
   (a) Luo and Bhakta (2020)
   (b) Silva Neto et al. (2021)
2. Pseudo-optimal localization of Ranazzi et al. (2022) using the following general function:
   (a) constant value of F1 = 1.0 (first function of Equation 2.67)
   (b) function F2 of Equation 2.67

100 realizations of correlation noise were generated to estimate the correlation noise threshold for each pair of parameter and observation. Figure 2.13 shows the correlation matrices corrected by localization, using the methods listed above, compared to the ensemble correlation matrix and the true correlation matrix. It can be observed that, in general all localization methods reduce the correlation noise, and the correlation matrices corrected by localization look more similar to the true correlation matrix. However, there appears to be generalized tapering of low correlation values, which may taper correlations that are not necessarily spurious.
A simple quantification of the difference between the estimated correlation matrices and the true correlation matrix can be performed using the Frobenius norm, which is presented in Table 2.4. It is verified from the table that all localization methods reduce the correlation noise, with the random shuffle and GC function of Silva Neto et al. (2021) achieving the lowest Frobenius norm value. Caution must be taken when interpreting these results, as the Frobenius norm is a simple measure of the difference between two matrices. In the author's view, for the objectives of the current work, this norm acts as a metric to verify the ability of the localization methods to reduce correlation noise, but it is not possible to infer which method is the best. In this respect, it may occur that a Pearson correlation coefficient obtained from a limited ensemble size is different from the correlation coefficient obtained from a larger ensemble size, due to the nonlinearity of the problem and not necessarily due to the correlation noise. Then, by applying adaptive localization, real correlations could be treated as noise and be tapered, which is not desirable.

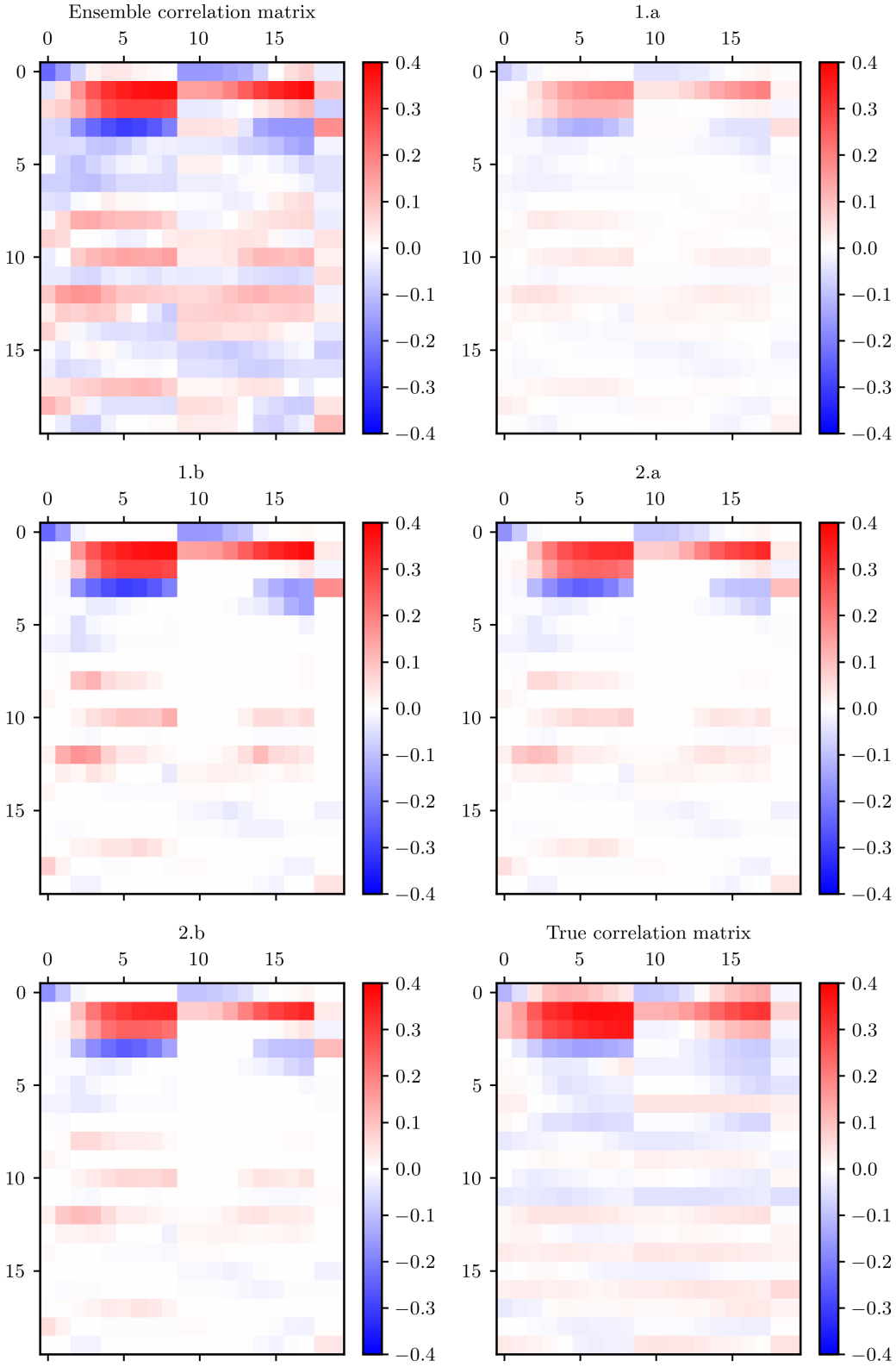

Figure 2.13: Comparison of the correlation matrices corrected by localization, obtained with 1. Random shuffle approach and GC function with Luo and Bhakta (2020) z dummy variable (1.a) and Silva Neto et al. (2021) z dummy variable (1.b), and 2. Pseudo-optimal localization with constant value of F1 = 1.0 (2.a) and function F2 of Equation 2.67 (2.b).

Table 2.4: Frobenius norm of the difference between the estimated correlation matrices and the true correlation matrix.

Method	Frobenius norm
Original ensemble correlation matrix	8.68
1.a Random shuffle and GC function (Luo and Bhakta, 2020)	3.66
1.b Random shuffle and GC function (Silva Neto et al., 2021)	3.44
2.a Pseudo-optimal localization, F1 = 1.0 (Ranazzi et al., 2022)	3.73
2.b Pseudo-optimal localization, function F2 (Ranazzi et al., 2022)	3.75

The data mismatch mean, standard deviation, and number of iterations of the ensemble methods resulting from the localization tests, are presented in Table 2.5. Comparing these results with the original cases (see Table 2.3), it is observed that none of the methods improved the data mismatch mean, except for the LM-SEnRML method with the GC function approach of Luo and Bhakta (2020). A slight reduction in the data mismatch standard deviation was observed for the LM-EnRML and LM-SEnRML methods with the GC function approach of Luo and Bhakta (2020).

Table 2.5: Mean / Standard deviation

Method	1.a		1.b		2.a		2.b	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
LM-EnRML	375	190	361	174	403	438	461	246
SEnRML	1480	1292	1873	1729	1856	1719	1585	1534
LM-SEnRML	1127	986	5982	4615	2748	2861	2839	2758

Figure 2.14 shows the posterior predictive uncertainty of cumulative infiltration estimated from the localization tests. It is observed that the localization methods do not improve the predictive uncertainty estimates, and predictive bias increased compared to the original cases, except for LM-EnRML which exhibits bias in the original case.

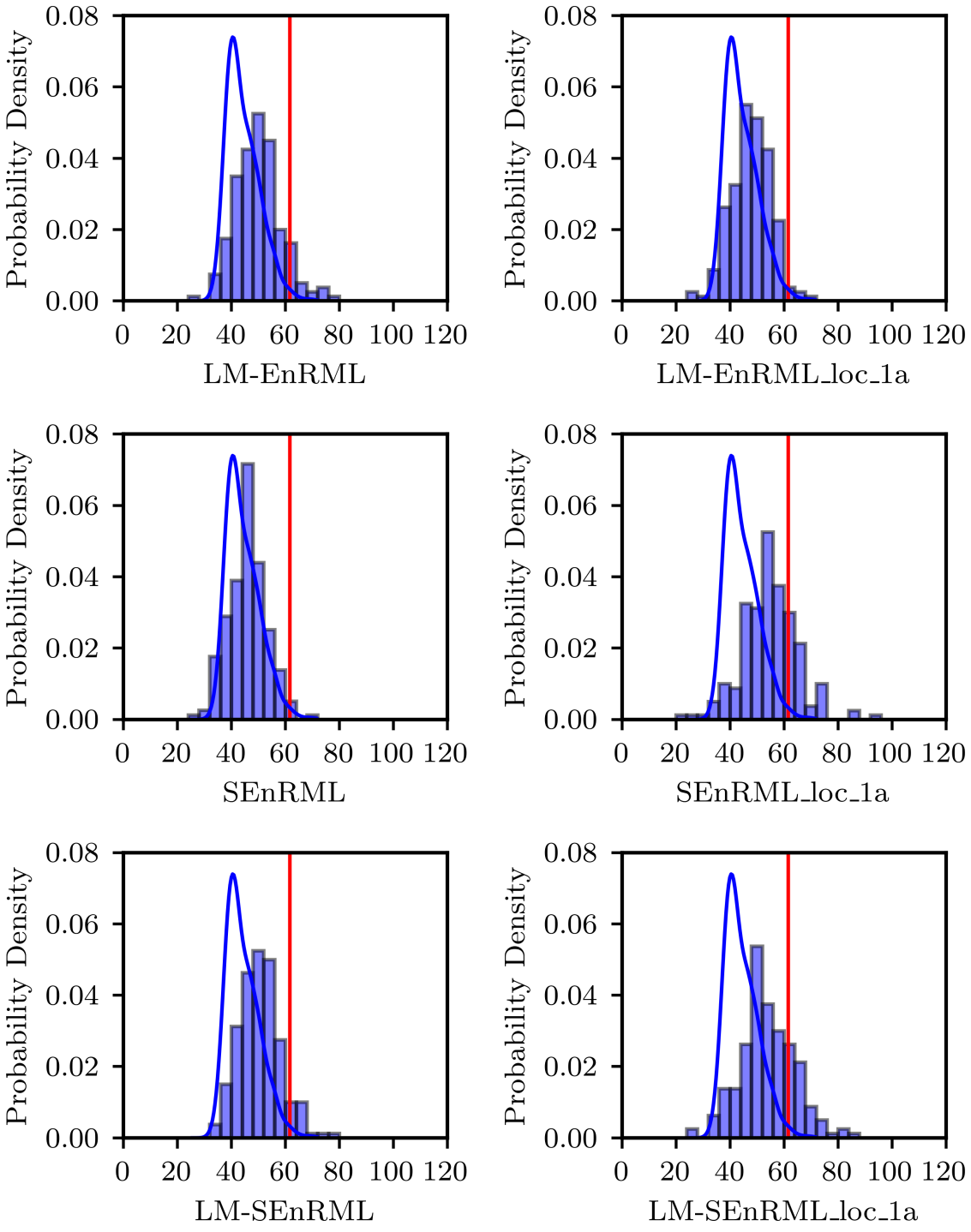

Figure 2.14: Histograms of posterior predictive uncertainty of total cumulative infiltration of the 1D unsaturated flow problem, obtained with ensemble methods and localization case 1.a. The true predictive uncertainty obtained from MCMC is represented by a blue pdf. The predictive cumulative infiltration derived from the true parameter set is represented by a red line.

2.7 Discussion

When using available methods for history matching, it is important to consider their limitations and assumptions. This is particularly important when history-matched models are used for predictive purposes that will ultimately support a decision. In this work, the mathematical framework of some existing methods was unified, allowing a consistent comparison between them. It is possible to recognize that they share the same or very similar theoretical foundations, and that the differences are mostly related to implementation details. By evaluating the performance of some of these methods in simple (but nonlinear) and complex problems, it is possible to identify strengths and weaknesses of each method.

It has been amply demonstrated (for example, Chen and Oliver, 2012, 2013; Evensen et al., 2019), that ensemble methods are a viable and efficient alternative to MCMC methods for history matching, especially when the computational cost of the forward model is high. However, it is important to realize that slow convergence may be more the rule than the exception. In the cases presented, ensemble methods required at least 10 iterations to achieve convergence. This has been previously shown by the rich literature on the subject (Chen and Oliver, 2012, 2013; Emerick and Reynolds, 2012, 2013). In the examples presented herein, the methods struggled to achieve a data mismatch close the expected value, which is a consequence of the nonlinearity of the problems, the limited ensemble size, and the usage of the same average sensitivity matrix for all ensemble members (Tarantola, 2005).

Among the ensemble methods, LM-EnRML appears to be the most efficient for history matching nonlinear cases. This may be due to the fact that this method uses a LM damping factor and approximates the prior covariance matrix in the hessian term of the parameter update equation using the updated parameter ensemble, allowing the exploration of directions outside the initial parameter ensemble subspace. This may contribute to overcome the limitations of the other methods whose solution is constrained to the initial parameter ensemble space (unless localization is implemented). Although this has not been theoretically proven in this work, the calculated values of the parameter residual matrix to the total parameter anomaly matrix $\mathbf{A}_l$ for the LM-EnRML and SEnRML methods, suggest that the LM-EnRML method explores new directions in the parameter space, compared to the SEnRML method. Consequences of this exploration are that the LM-EnRML method may achieve a better fit to the data, with the risk of generating bias, underestimation of parameter uncertainty, and therefore underestimation of predictive uncertainty. In fact, from the results of the numerical examples, it was shown that the LM-EnRML method, although attained the best fit to the data, incurred in more predictive bias compared to SEnRML. Because it is not possible to know the true posterior probability distribution of parameters or predictions when applying ensemble methods

to nonlinear problems (Evensen et al., 2022), interpretations of predictive bias and incorrectness of posterior uncertainties can only be speculative, and limited by exploration of example problems such as those that are presented herein.

The results of the 1D unsaturated flow problem show that the true posterior probability distribution of the cumulative infiltration is best represented by the SEnRML method, followed by the LM variant, neither of which achieved a good fit to the data. Overall, predictive uncertainty calculated by all three methods is reasonable, covering the true value of the cumulative infiltration. This suggest that it is not necessary to achieve a perfect fit to the data to obtain a good estimate of predictive uncertainty, but only to that aspect of the data that host the information that is relevant to the prediction and therefore has a similar character to it (Doherty and Christensen, 2011). In the example presented above, the cumulative infiltration is dependent on the sub-saturated to saturated portions of the soil; hence obtaining a good fit to large soil tension values may be not as important as getting a good fit to small tension values.

Finally, the localization tests show that all of the tested methods reduce correlation noise in the estimated correlation matrix. These results are consistent with the findings of Luo and Bhakta (2020), who showed that the ensemble correlation matrix can be noisy, and that localization can reduce this noise. However, it is not possible to infer which method is the best. Furthermore, because of the number of heuristics and assumptions made in design of the localization methods, they probably perform differently depending on the problem. In the above example, implementing localization in ensemble methods does not improve the data mismatch mean and standard deviation significantly, but it does generate more predictive bias compared to the original cases. This is an interesting outcome that suggests that by artificially increasing the degrees of freedom of the parameter ensemble space through localization, more predictive bias, and not necessarily a better fit to the data, may result. Therefore, in cases where localization is necessary to efficiently assimilate data, practitioners should be aware of these potential repercussions.

2.8 Conclusions

This chapter presents a unified mathematical framework for discussion of inverse and ensemble methods that are used for history matching and uncertainty quantification. This allows a consistent comparison of the methods to be made. It is hoped that this unification will help others to clarify the theoretical foundations of the methods, and to identify the strengths and weaknesses of each of them. Some of these methods were tested in a simple and complex problem.

The results of the numerical examples show that ensemble methods are very efficient, but slow convergence is more the rule than the exception in nonlinear problems. This needs to be considered when using these methods in practice. The LM-EnRML method

is the most efficient in history matching data in nonlinear cases; however it appears to be prone to predictive bias. It was also found that methods that do not achieve a good fit to the data, such as the subspace iterative ensemble smoother (SEnRML), can still provide reasonable estimates of predictive uncertainty. To the author's knowledge this is the first time that this method has been tested in a highly nonlinear groundwater problem. This work has shown that the method is a viable alternative to the LM-EnRML method for history matching and predictive uncertainty analysis in groundwater modelling, and that it can provide reasonable predictive uncertainty estimates.

It is hoped that the presented review of inverse and ensemble methods provides a consolidated framework of reference, and that the numerical examples expose their strengths and weaknesses when deployed in everyday modelling circumstances. Practitioners need to be aware of these strengths and weaknesses when using these methods in practice for providing predictive uncertainty estimates in support of a decision.

Chapter 3

Prior Inference from Groundwater Model Calibration: Empirical Bayesianism to Improve Predictive Uncertainty

Author contributions

T. Opazo: Conceptualization 50%, Realization 100%, Writing 100%. J. Doherty: Conceptualization 50%, Review 100%.

Manuscript in preparation for submission to Journal of Hydrology: Opazo, T., Doherty, J. Prior Inference from Groundwater Model Calibration: Empirical Bayesianism to Improve Predictive Uncertainty.

Abstract

Groundwater model calibration can lead to surprising patterns of parameter heterogeneity that challenges the prior probability distribution of model parameters. If predictive uncertainty estimates are based on an incorrect prior, they may be underestimated or biased, potentially resulting in poor decision-making. This work presents a methodology for addressing prior-data conflict to update the uncertainty in the prior for predictive uncertainty quantification in groundwater modelling. The approach evaluates the compatibility of the prior with the calibrated parameter field obtained from regularized inversion and performs empirical Bayesian inference of the prior hyperparameters using the calibrated parameter field. The methodology is tested on a synthetic 2D groundwater model simulating drawdown due to pumping, where the prior is treated as uncertain and updated using the calibrated parameter vector. Results suggest that recognizing uncer-

tainty in the prior and sequentially performing model calibration followed by predictive uncertainty may lead to more conservative predictive uncertainty estimates. This approach helps mitigate the underestimation of predictive uncertainty, which is essential in groundwater modelling for decision support.

3.1 Introduction

In a seminal and highly commented study, Capen (1976) demonstrated that the less we know about something the more likely we are to construct a narrow probability interval that does not contain the truth. For this and other reasons, Robust Bayesian Analysis (Good, 2018; Berger, 1990) treats the prior as uncertain, in recognition of the fact that the choice of the prior is open to criticism as any other modelling assumption (Sprenger, 2018). It is considered that uncertainty in the prior is irrelevant only if the range of posterior probabilities that results from evaluating different possible priors is small (Berger, 1990), deeming the analysis robust.

There is a high chance of constructing an erroneous prior when defining a history matching problem in a Bayesian framework. This seems to apply particularly well to groundwater systems, where there is an incomplete knowledge of the subsurface, and prior probabilities are decoded from a set of hydrogeological judgements derived from available information and expert knowledge. One pragmatic evidence that the prior may be misspecified is presented to the modeller when, as a result of history matching, surprising patterns of spatial parameter heterogeneity emerge, or the extreme values that parameters adopt in order to assimilate data gain little to no support from the prior. Assuming that the numerical model is adequate, it is said that the data is surprising given the prior, a situation referred to as prior-data conflict (Evans and Moshonov, 2006). On one hand, by gathering more data, the prior may become progressively less important or even irrelevant (Evans and Moshonov, 2006; Gelman et al., 2017). This implicitly solves the problem of prior-data conflict. On the other hand, as more data are obtained, more complex processes are often identified and required to be (parametrically) represented in the models if they are relevant to predictions of interest. As a result, more complex likelihoods are generated (Gelman et al., 2017) potentially increasing the relevance of the prior in the overall inference problem. Another option is to make the prior less informative, using methods such as mixture or e-contaminated priors (Berger, 1990; Egidi et al., 2022), or alternative prior evaluation (Evans and Jang, 2011), to name a few. Some of these methods are referred to as being empirical Bayesian, as they use data to modify the prior, therefore violating the precepts of Bayes' equation. In any case, the way of solving prior-data conflict, once it is detected, is not straightforward, especially when working in a high-dimensional parameter space (Gelman et al., 2017). To make matters worse, prior-data conflict, or the fact that the prior is misspecified, might not be identified at all when

performing history matching, but may have an impact on the estimation of predictive uncertainty.

Although there has been recognition of the importance of including uncertainty in the prior for environmental modelling (Reichert, 1997; Doherty and Moore, 2020), except for a few studies (Woodbury and Ulrych, 2000; Rojas et al., 2009; Shen et al., 2014; Hoffmann et al., 2019) no relevant research has been published in the groundwater literature that explore ways for checking prior-data conflict, prior updating, and evaluating the impact of an incorrect prior on predictive uncertainty quantification. Also, while related field of petroleum reservoir modelling has been active in researching on model diagnostics, model error and observation bias (Oliver and Alfonzo, 2018; Alfonzo and Oliver, 2019, 2020), it has placed less emphasis to uncertain priors (Oliver, 2022; Mioratina and Oliver, 2023). Given the increasing use of Bayesian techniques in groundwater modelling for both history matching and uncertainty analysis, this is matter that requires urgent attention.

In this work, a new workflow is proposed that embraces uncertain priors. First, prior-data conflict is indirectly evaluated by comparing the minimum error variance solution of a model calibration process with expectations from the prior. Second, the prior is updated using the calibrated parameters as data, resembling some forms of empirical Bayes analysis.

Using a synthetic but realistic groundwater model case, the workflow is tested. As an outcome, it is shown that by recognizing uncertainty in the prior and performing model calibration first and uncertainty analysis later, predictive uncertainty is not underestimated. Avoidance of uncertainty underestimation is a fundamental tenet of robust decision support modelling (Doherty and Simmons, 2013).

This chapter is organized as follows: the theory behind the proposed workflow is presented in the next section, followed by a test case description, methodology, and results. The chapter ends with a discussion and conclusions.

3.2 Theory

This section reiterates key elements of the mathematical framework presented in Chapter 2. While some overlap is unavoidable, the repetition is intended to ensure self-containment and clarity, particularly given the distinct application focus of this chapter. Readers seeking full mathematical details and theoretical background are referred to Chapter 2.

Let $\mathbf{x}$ denote a random vector that parameterizes a physical system and $\mathbf{d}$ the observed data vector of system state, i.e, a set of measurements. To make the following equations more tractable, the vector $\mathbf{x}$ is characterized using a probability density function (pdf) with a zero mean and covariance matrix $\mathbf{C}_\mathbf{x}$. A linear relationship between model parameters and observed data can be expressed through the action of a sensitivity matrix

$\mathbf{G}$, as follows:

$$\mathbf{d} = \mathbf{G}\mathbf{x} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{\epsilon}$ is a vector of random errors with a zero mean and covariance matrix $\mathbf{C_d}$. As stated above, for the sake of numerical simplicity $\mathbf{x}$ in Equation 3.1 represents departures from its prior mean values. Let $\underline{\mathbf{x}}$ represent the estimated parameter vector that calibrates the model, i.e., the minimum error variance solution to Equation 3.1. Using regularized inversion $\underline{\mathbf{x}}$ can be estimated as:

$$\underline{\mathbf{x}} = \mathbf{L}\mathbf{d}, \tag{3.2}$$

where $\mathbf{L}$ is the generalized inverse that is used to obtain a unique solution to the likely ill-posed inverse problem. If Tikhonov regularization is used, $\mathbf{L}$ is defined as follows (Moore and Doherty, 2006):

$$\mathbf{L} = \left(\mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{G} + \mu^2 {\mathbf{G}_r}^T {\mathbf{C_x}}^{-1}\mathbf{G}_r\right)^{-1} \mathbf{G}^T {\mathbf{C_d}}^{-1} \tag{3.3}$$

where μ is a regularization weight factor, estimated through the model calibration process, and $\mathbf{G}_r$ is the regularization matrix that operates on the parameters to enforce regularization constraints. In Equation 3.3 it is also assumed for simplicity that the regularization observations are equal to zero, reflecting preferred-value regularization (Doherty, 2015). If Equation 3.1 is substituted into equation Equation 3.2 the following equation is obtained:

$$\underline{\mathbf{x}} = \mathbf{L}\mathbf{G}\mathbf{k} + \mathbf{L}\boldsymbol{\epsilon}. \tag{3.4}$$

Now if the resolution matrix $\mathbf{R}$ is defined as

$$\mathbf{R} = \mathbf{L}\mathbf{G}, \tag{3.5}$$

or, equivalently, as

$$\mathbf{R} = \left(\mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{G} + \mu^2 {\mathbf{G}_r}^T {\mathbf{C_x}}^{-1}\mathbf{G}_r\right)^{-1} \mathbf{G}^T {\mathbf{C_d}}^{-1}\mathbf{G}, \tag{3.6}$$

then

$$\underline{\mathbf{x}} = \mathbf{R}\mathbf{x} + \mathbf{L}\boldsymbol{\epsilon}. \tag{3.7}$$

If measurement noise is discarded, Equation 3.7 is simplified to

$$\underline{\mathbf{x}} = \mathbf{R}\mathbf{x}, \tag{3.8}$$

which defines the relationship between the estimated parameter vector $\underline{\mathbf{x}}$ and the true parameter vector $\mathbf{x}$ through the resolution matrix $\mathbf{R}$. This equation shows that, for an ill-posed inverse problem, each estimated parameter value of $\underline{\mathbf{x}}$ is a linear combination of

true parameter values of $\mathbf{x}$ (Moore and Doherty, 2006). This only happens over many realizations of reality.

The covariance matrix of the estimated parameter vector $\underline{\mathbf{x}}$, i.e., $\mathbf{C}_{\underline{\mathbf{x}}}$, can be related to the covariance matrix of the true parameter vector $\mathbf{x}$, i.e., $\mathbf{C}_{\mathbf{x}}$, through the resolution matrix $\mathbf{R}$ as follows (Koch, 1999):

$$\mathbf{C}_{\underline{\mathbf{x}}} = \mathbf{R}\mathbf{C}_{\mathbf{x}}\mathbf{R}^T. \tag{3.9}$$

Equation 3.9 has important interpretations as discussed by Moore and Doherty (2006). First, it links the covariance matrix of the estimated parameter vector $\underline{\mathbf{x}}$, with the covariance matrix of the true parameter vector $\mathbf{x}$, through the action of the resolution matrix $\mathbf{R}$. If the inverse problem is well-posed, the resolution matrix $\mathbf{R}$ is equal to the identity matrix; under these circumstances the covariance matrix $\mathbf{C}_{\underline{\mathbf{x}}}$ is equal to $\mathbf{C}_{\mathbf{x}}$. Hence, $\underline{\mathbf{x}}$ is a sample of the true covariance matrix. Therefore, it provides information about the true spatial variability of the system. If the inverse problem is ill-posed, the resolution matrix is not the identity matrix. Then $\mathbf{C}_{\underline{\mathbf{x}}}$ is not equal to $\mathbf{C}_{\mathbf{x}}$, and therefore $\underline{\mathbf{x}}$ is only providing information about $\mathbf{C}_{\underline{\mathbf{x}}}$, a 'projection' of the true covariance matrix onto a lower-dimensional space. This is the reason why a calibrated parameter field cannot be directly used to infer the true spatial variability of a system. Nevertheless, the estimated parameter vector $\underline{\mathbf{x}}$ does provide information about an altered, regularized (by the resolution matrix) version of $\mathbf{C}_{\mathbf{x}}$.

The discussion above motivates a methodology for indirectly testing for prior-data conflict. This is done by evaluating the compatibility of the estimated parameter vector $\underline{\mathbf{x}}$ with an estimated covariance matrix $\mathbf{C}_{\underline{\mathbf{x}}}$ that is calculated from the true prior covariance matrix $\mathbf{C}_{\mathbf{x}}$ and the resolution matrix according to Equation 3.9. Additionally, a family of candidate priors $\mathbf{C}_{\mathbf{x}}$ can be conditioned by the necessity for $\underline{\mathbf{x}}$ to be a sample of $\mathbf{C}_{\underline{\mathbf{x}}}$ calculated using Equation 3.9. This can be done by applying Bayesian inference.

If the true but uncertain prior covariance matrix $\mathbf{C}_{\mathbf{x}}$ is characterized by a certain geostatistical structure, this structure can be hyper-parameterized. Let $\boldsymbol{\theta}$ define the parameter vector containing hyperparameters that define this structure. They can be for example, a variogram sill, range, and anisotropy. A Bayesian inference problem on the posterior uncertainty of $\boldsymbol{\theta}$ given the calibrated parameter vector $\underline{\mathbf{x}}$, $f(\boldsymbol{\theta}|\underline{\mathbf{x}})$, can be formulated as

$$f(\boldsymbol{\theta}|\underline{\mathbf{x}}) = \frac{f(\boldsymbol{\theta})\, f(\underline{\mathbf{x}}|\boldsymbol{\theta})}{f(\underline{\mathbf{x}})}, \tag{3.10}$$

where $f(\boldsymbol{\theta})$ is the prior pdf of $\boldsymbol{\theta}$, $f(\underline{\mathbf{x}}|\boldsymbol{\theta})$ is the likelihood function, and $f(\underline{\mathbf{x}})$ is the pdf of the estimated parameter vector $\underline{\mathbf{x}}$ among all possible values of $\boldsymbol{\theta}$. It is important to note that the parameter vector $\underline{\mathbf{x}}$, the minimum error variance solution of the inverse problem, is considered to be data for the inference problem presented in Equation 3.10. It is

recognized that using the minimum error variance solution as data to infer the posterior of hyperparameters that define the geostatistical structure of the prior is not a purely Bayesian approach, therefore it is considered a form of empirical Bayesian analysis.

The likelihood function in the hyperparameter inference problem can be represented by a model of affinity between a 'projected' covariance matrix candidate $\mathbf{C}_{\underline{\mathbf{x}}}$ and the calibrated parameter vector $\underline{\mathbf{x}}$. The matrix $\mathbf{C}_{\underline{\mathbf{x}}}$ is in turn calculated by applying Equation 3.9 to a candidate covariance matrix $\mathbf{C}_{\mathbf{x}}$ derived from a random realization of the hyperparameter vector $\boldsymbol{\theta}$. The closer the affinity, the higher the value of the likelihood function must be. One option of likelihood function for a multidimensional case is the Mahalanobis distance (Koch, 1999), generally used to identify outliers or to classify if an observation belongs to a certain population. However, for ill-posed inverse problems this becomes cumbersome due for two main reasons: first, this calculation would require to invert the covariance matrix $\mathbf{C}_{\underline{\mathbf{x}}}$ which is not possible as it is rank-deficient after applying Equation 3.5. A pseudo inverse would need to be calculated. Second, the estimated parameter vector $\underline{\mathbf{x}}$ is not unique, hence treating it directly as observation data would be misleading.

Both issues presented above can be overcome by working in lower-dimensional space and define a likelihood function within that space. This is done by calculating a new parameter vector $\mathbf{y}$ from $\underline{\mathbf{x}}$ and $\mathbf{C}_{\underline{\mathbf{x}}}$, and therefore $\mathbf{C}_{\mathbf{x}}$. Let $\mathbf{C}_{\underline{\mathbf{x}}}$ be the candidate projected covariance matrix obtained from applying Equation 3.9 to a candidate covariance matrix $\mathbf{C}_{\mathbf{x}}$. In turn, $\mathbf{C}_{\mathbf{x}}$ is derived from a random realization of the hyperparameter vector $\boldsymbol{\theta}$. Let perform Singular Value Decomposition (SVD) on $\mathbf{C}_{\underline{\mathbf{x}}}$, the candidate estimated covariance matrix obtained from applying Equation 3.9:

$$\mathbf{C}_{\underline{\mathbf{x}}} = \mathbf{E}\mathbf{F}\mathbf{E}^T, \tag{3.11}$$

where $\mathbf{E}$ is the eigenvector matrix and $\mathbf{F}$ is a diagonal matrix of singular values. Now, if the mean of $\underline{\mathbf{x}}$ is defined as zero according to Equation 3.8, after discarding measurement noise, the following parameter transformation can be performed:

$$\mathbf{y} = \mathbf{F}^{-1/2}\mathbf{E}^T\underline{\mathbf{x}}. \tag{3.12}$$

Given that it cannot be guaranteed that $\mathbf{C}_{\underline{\mathbf{x}}}$ is positive definite, SVD is truncated to the number $m < n$ of non-zero singular values. Note that $\mathbf{y}$ is not only a function of $\underline{\mathbf{x}}$, but also of $\mathbf{C}_{\underline{\mathbf{x}}}$, which is in turn a function of $\mathbf{C}_{\mathbf{x}}$ and therefore $\boldsymbol{\theta}$. Then, $\mathbf{y}$ is a random vector, contrary to $\underline{\mathbf{x}}$ which is the calibrated parameter vector, i.e., the minimum error variance solution of the inverse problem. A key aspect of this random vector is that it will be compatible with a multivariate standard normal distribution if the candidate covariance matrix $\mathbf{C}_{\underline{\mathbf{x}}}$ is compatible with the calibrated parameter vector $\underline{\mathbf{x}}$.

Two metrics have been chosen to evaluate the affinity of the vector $\mathbf{y}$ to a multivariate

standard normal distribution, the Kolmogorov-Smirnov test (KS) (Massey, 1951) and the χ^2 test. Both tests are translated into likelihood functions, combined as follows:

$$L = f_{KS}(D) f_{\chi^2_m}(S), \tag{3.13}$$

where $f_{KS}(D)$ is the Kolmogorov-Smirnov likelihood function, and $f_{\chi^2_m}(S)$ is the χ^2_m likelihood function of $S = \sum_{i=1}^{m} y_i^2$. The KS likelihood function $f_{KS}(D)$ is defined as a half-normal distribution with standard deviation equals to $D_\alpha(m)$:

$$f_{KS}(D) = \frac{1}{D_\alpha(m)} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{D^2}{2D_\alpha(m)^2}\right), \tag{3.14}$$

where $D_\alpha(m)$ is the critical distance for a sample of size m and a significance level α. The scalar D is the maximum absolute distance between the cumulative step-function distribution generated from each y_i in $\mathbf{y}$ and the standard normal cumulative distribution, calculated as

$$D = \max \left| F_0(y_i) - S_m(y_i) \right|, \tag{3.15}$$

where $F_0(y_i)$ is the population cumulative distribution (standard normal in this case) and $S_m(y_i)$ is the cumulative step-function distribution of the y_i, i.e, $S_m(y_i) = \frac{i}{m}$. In the standard KS test, if the maximum calculated distance D is greater than the critical distance $D_\alpha(m)$, the null hypothesis that the $\mathbf{y}$ vector is a sample of a multivariate standard normal distribution is rejected, for a significance level α. In this work, the critical distance is used as the standard deviation of a half-normal distribution, which is used as a pseudo-likelihood function, as shown in Equation 3.14. Then, a sample $\mathbf{y}$ that is closer to a standard normal distribution will have a maximum absolute distance D closer to zero, and therefore a higher likelihood value.
Additionally, if $\mathbf{y}$ is a sample of the standard normal distribution, the sum of ${y_i}^2$ should follow a χ^2 distribution of m degrees of freedom. Therefore, the second likelihood function $f_{\chi^2_m}(S)$ is directly calculated from the χ^2_m distribution.

3.3 Numerical Example

3.3.1 Methodology

The principles discussed in the previous sections are illustrated through a workflow that tests the theory, quantifying predictive uncertainty with a synthetic 2D groundwater model that simulates drawdown due to pumping. The correlation structure of the uncertain prior covariance matrix $\mathbf{C_x}$ is assumed known and characterized by an exponential decay with distance. However, the sill and effective range of the variogram are assumed uncertain, with independent, and uncorrelated parameters grouped in the hyperparam-

eter vector $\boldsymbol{\theta}$. The prior uncertainty of $\boldsymbol{\theta}$ is defined using log-Gaussian probability distributions, as presented in Table 3.1, with a mode value of 0.25 and 2500 m for the sill and range, respectively. A sill value of 1.0 and a range value of 5000 m are assigned to the true covariance matrix $\mathbf{C}_{\mathbf{x}true}$. Based on the prior pdf of $\boldsymbol{\theta}$, these values have a low support in the prior, meaning that the probability of having a large sill and range values is low, from a prior perspective. Defining the problem in this way adds complexity to the inference problem. Then, one realization $\mathbf{x}_{true}$ of the random parameter vector $\mathbf{x}$ is generated from a multi-Gaussian distribution with mean $\overline{\mathbf{x}}$ and covariance matrix $\mathbf{C}_{\mathbf{x}true}$. The parameter vector $\mathbf{x}_{true}$ corresponds to hydraulic conductivity values that are used to run the groundwater model, and obtain model outputs used as synthetic observations. In turn, the hyperparameter vector $\boldsymbol{\theta}_{mode}$ with values of 0.25 and 2500 m for the sill and range, respectively, is used to generate a 'wrong' prior covariance matrix $\mathbf{C}_{\mathbf{x}wrong}$.

Parameter	**Transform**	**Pdf**	**Parameters**
Sill	Log	Normal	$\mu = -0.52, \sigma = 0.2$
Range	Log	Normal	$\mu = 3.42, \sigma = 0.1$

Table 3.1: Details of the sill and range of the prior Gaussian probability density functions.

First, history matching and predictive uncertainty quantification are carried out using the iterative ensemble smoother method (IES) (Chen and Oliver, 2013) with the 'wrong' prior $\mathbf{C}_{\mathbf{x}wrong}$. Given that the measurement dataset is generated from a parameter realization $\mathbf{x}_{true}$ obtained from the true covariance matrix $\mathbf{C}_{\mathbf{x}true}$, and that history matching is performed using the 'wrong' prior $\mathbf{C}_{\mathbf{x}wrong}$, the estimated predictive uncertainty is expected to be underestimated.

Having the luxury of knowing the true value of the prediction, the adequacy of the predictive uncertainty estimation is evaluated using the metric defined by Doherty and Simmons (2013). It is checked if the true value of the prediction is within the estimated predictive uncertainty limits; this is done with IES. The model is then subject to regularized inversion to obtain a minimum error variance solution $\underline{\mathbf{x}}$, using PEST (Doherty, 2023). The calibrated parameter vector $\underline{\mathbf{x}}$ serves first to check its compatibility with the 'estimated' wrong prior $\mathbf{C}_{\mathbf{x}wrong}$ (after applying Equation 3.5) and then to perform hyperparameter inference of the hyperparameter vector $\boldsymbol{\theta}$ using the likelihood function defined in Equation 3.13. This in turn results in estimation of the posterior uncertainty of the unknown covariance matrix $\mathbf{C}_{\mathbf{x}}$.

Because model calibration yields a covariance matrix, the resolution matrix $\mathbf{R}$ is estimated using Equation 3.5, using the PEST-optimized regularization weight factor μ^2. Here, the covariance matrix $\mathbf{C}_{\mathbf{d}}$ of measurement noise is a diagonal matrix with a standard deviation commensurate with measurement noise. The prior parameter covariance matrix is $\mathbf{C}_{\mathbf{x}wrong}$, and the $\mathbf{Z}_R$ is the preferred value regularization matrix, which in this case is the identity matrix.

Treating $\underline{\mathbf{x}}$ as an observation and using the resolution matrix $\mathbf{R}$ obtained from model calibration (using Equation 3.5), an inference problem is defined for $\boldsymbol{\theta}$ to estimate its posterior pdf $f(\boldsymbol{\theta}|\underline{\mathbf{x}})$ according to Equation 3.10. The likelihood function is defined based on Equation 3.13. In this case, the dimensionality of the estimated covariance matrix $\mathbf{C}_{\underline{\mathbf{x}}}$ once projected by the resolution matrix $\mathbf{R}$ is reduced to $m = 8$. Then, the standard deviation used in the pseudo-likelihood function $f_{KS}(D)$ is equal to 0.457, which is equal to the critical difference $D_\alpha(8)$ for a significance level of 0.05 (Massey, 1951).
The hyperparameter inference problem is performed with Markov Chain Monte Carlo (MCMC) using the (Multiple-Try) Differential Evolution Adaptive Metropolis (MT-DREAM$_{(ZS)}$) algorithm implemented in pyDREAM (Laloy and Vrugt, 2012). The workflow of the MCMC algorithm is detailed below:

1. Take a random sample from the prior pdf of $\boldsymbol{\theta}$.
2. Build a candidate covariance matrix C_x from a variogram using the sample of $\boldsymbol{\theta}$ and obtain the estimated covariance matrix $\mathbf{C}_{\underline{\mathbf{x}}}$ from $\mathbf{C}_{\mathbf{x}}$ and $\mathbf{R}$ using Equation 3.9.
3. Perform truncated-SVD on $\mathbf{C}_{\underline{\mathbf{x}}}$ and obtain the transformed parameter vector $\mathbf{y}$ using Equation 3.12. Define the number m of y_i scalars to use, based on the dimensionality of $\mathbf{C}_{\underline{\mathbf{x}}}$.
4. Estimate the KS-maximum distance D between the empirical cumulative distribution of $\mathbf{y}$ for all $i \leq m$ and the standard normal cumulative distribution, and calculate the pseudo-likelihood function $f_{KS}(D)$ using Equation 3.14.
5. Obtain the sum S of ${y_i}^2$ for all $i \leq m$, and calculate the likelihood function $f_{\chi^2_m}(S)$.
6. Calculate the total likelihood function $L = f_{KS}(D) f_{\chi^2_m}(S)$.

Once the posterior pdf of the hyperparameter vector $\boldsymbol{\theta}$ has been explored in this manner, the model is again history-matched with IES using a parameter ensemble sampled from a family of priors characterized by an exponential variogram, using posterior samples of the hyperparameter vector, i.e, $f(\boldsymbol{\theta}|\underline{\mathbf{x}})$. It is expected that the estimated predictive uncertainty is more conservative and robust, as the prior has been updated using the calibrated parameter vector $\underline{\mathbf{x}}$. Ideally the true value of the prediction should therefore lie between the updated predictive uncertainty limits. This is the adopted metric for evaluating the success of the proposed workflow.

3.3.2 Model Description

A synthetic 2D steady-state groundwater flow model using MODFLOW 6 (Langevin et al., 2017) was built to simulate pumping from a 100-m thick confined aquifer, where

the prediction of interest is drawdown at a distant location from the pumping wells. This numerical setting resembles a typical layout of a water-supply well field or a lithium extraction mining operation, that may be environmentally constrained by drawdown thresholds.

The model domain extends to 10,000 m x 10,000 m, discretized into 100 by 100 rows and columns. Along the top and bottom limits of the model, head-dependent flux boundary conditions are defined with a reference head of 100 m and a conductance of 25 m^2/d. Along the left and right model boundaries no-flow conditions are applied. Pumping is simulated using two consecutive steady-state stress periods, representing existing and future conditions. Two model cells are used to represent pumping at two separated areas, with a constant-flux boundary condition of 4320 m^3/d (50 L/s). One cell represents the existing extraction location, and the second cell represents the future extraction location. During the first stress period, only the existing extraction location is active. During the second stress period, additional pumping is simulated at the future extraction location.

A 'true' hydraulic conductivity field was generated using Unconditional Sequential Gaussian Simulation (USGSIM) with a log mean value of zero and an exponential variogram with a sill of 1.0 and an effective range of 5000 m. These are the hyperparameter values of the true hyperparameter vector $\boldsymbol{\theta}_{true}$.

The forward model was run using the true hydraulic conductivity field to obtain synthetic drawdown values at 9 observation wells located in the vicinity of the existing extraction location. These observed values vary between 12.6 m and 19.6 m, and a Gaussian noise with a standard deviation of 0.01 m was added to each observation value before using them as measurement data for history matching. A prediction of drawdown simulated in the second stress period is evaluated at an observation point located to the right of the new extraction location (Figure 3.1). By simulating the future pumping conditions using the true hydraulic conductivity field, a maximum drawdown of 22.8 m was obtained. This is the true value of the prediction of interest.

The model is parameterized with one bulk hydraulic conductivity parameter x_b representing the mean of the parameter field, and 400 pilot points multipliers x_{pp} spaced at 500 m (one each 5 model cells). Kriging interpolation is used to populate all cells of the model using an exponential variogram with a range of 1000 m (twice the distance between pilot points).

The model was subject to history matching using IES (Chen and Oliver, 2013) implemented in the PEST++ software (White, 2018). A pilot points parameter ensemble of size 300 was generated using a log-transformed multi-Gaussian prior with a mean of zero and a prior covariance matrix. History matching was carried out twice, first using a parameter ensemble generated from the 'wrong' prior covariance matrix $\mathbf{C}_{\mathbf{x}wrong}$, and then using a second parameter ensemble derived from a family of prior covariance matrices $\mathbf{C}_{\mathbf{x}}$ built from the posterior distribution of the hyperparameter vector $\boldsymbol{\theta}$. A standard

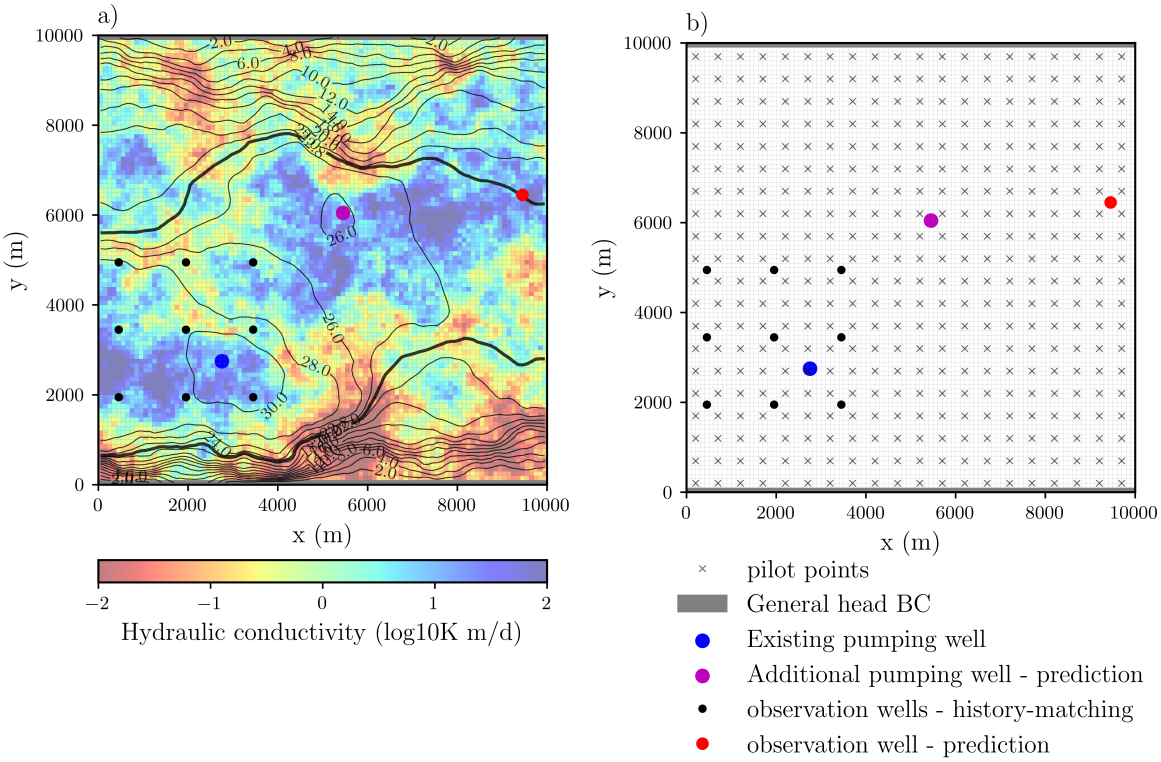

Figure 3.1: 2D model map view showing the distribution of hydraulic conductivity, boundary conditions, pumping and observation wells: (a) true hydraulic conductivity field with drawdown contours for the second stress period (b) model domain and discretization, with locations of pilot points. The thick black line represents the 22.8 m drawdown contour.

deviation of 0.2 was assumed for the mean of the parameter field x_b. For both cases an observation ensemble was generated assuming an uncorrelated Gaussian noise with a mean of zero and standard deviation of 0.01 m. For the IES optimization process, the maximum number of iterations was set to 10. Five randomly-selected parameter realizations were used for Levenberg-Marquardt lambda testing. Several convergence criteria were used to terminate the IES optimization process, including a relative mean objective function reduction of 0.005 for four consecutive iterations, four iterations without a reduction in the mean objective function, and four iterations without realizations in which the smallest objective function is less than 1.05 times its previous value.

A minimum error variance solution $\underline{\mathbf{x}}$ was obtained using regularized inversion with PEST (Doherty, 2023). Initial log-transformed parameters were assigned a value of zero (log-transformed from their mean of 1.0). Preferred value (equal to zero) regularization was implemented with a weighting matrix derived from the inverse of the pilot points' covariance matrix $\mathbf{C}_{\mathbf{x}wrong}$. Nine drawdown observations were used as calibration targets, setting their individual weights to 100, this being the inverse of the standard deviation of the measurement noise. A target measurement objective function equal to the number of observations was defined. The convergence criterion on the measurement objective function was set to 2% higher than the target measurement objective function. The same convergence criteria that was used for IES was implemented for PEST calibration, except for the last criterion. The resolution matrix $\mathbf{R}$ was calculated at the end of the calibration process using Equation 3.5.

According to the workflow described above, the posterior uncertainty of the hyperparameter vector $\boldsymbol{\theta}$ was estimated using MCMC with the prior of $\boldsymbol{\theta}$ defined in Table 3.1, and the likelihood function defined in Equation 3.13. The minimum error variance solution $\underline{\mathbf{x}}$ was used as data, and the resolution matrix $\mathbf{R}$ was used to project each covariance matrix candidate to obtain $\mathbf{C}_{\underline{\mathbf{x}}}$.

3.3.3 Results

Figure 3.2 shows the prior and posterior predictive uncertainty of drawdown at the observation well of interest, derived from the IES history matching process using the 'wrong' prior. It can be observed that the prior predictive uncertainty covers the true value of the prediction. However, the true value of the prediction lies outside the posterior predictive uncertainty range calculated with IES. It could be argued that the model is not adequate to history-match the observed data. However, by examining the prior drawdown uncertainty range of the 9 observation wells it is verified that the observed drawdowns are within the prior range except for one well for which the observed value lies at the edge of the prior range(lower left plot in Figure 3.3). Then there is no clear evidence of prior-data conflict, based on the comparison between the measurement dataset and the

prior realizations of counterpart model outputs.

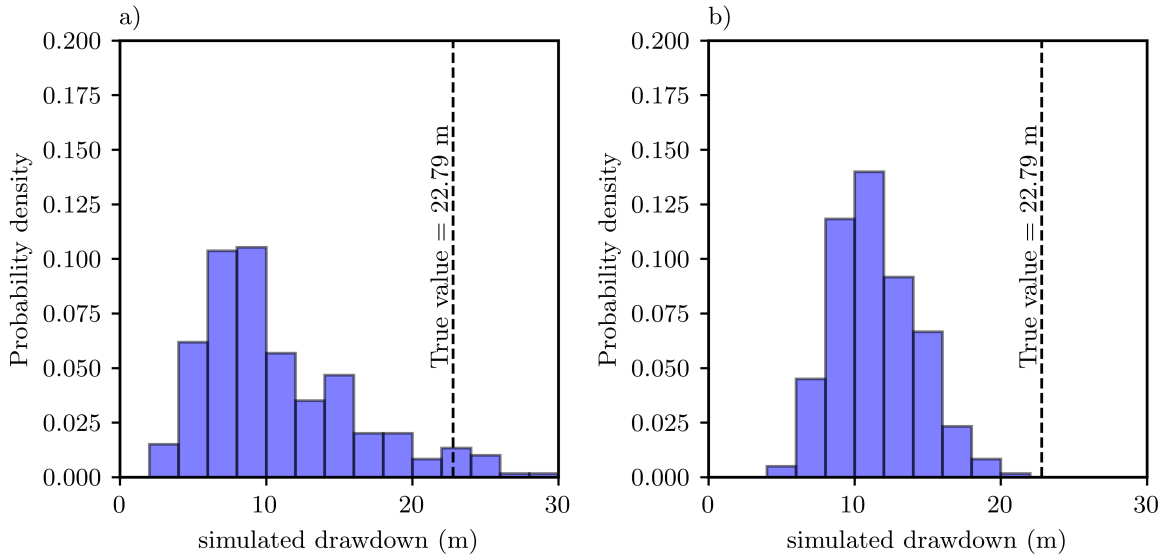

Figure 3.2: Prior (a) and posterior (b) predictive uncertainty of drawdown at the observation well of interest, derived from the IES history matching process using the 'wrong' prior. The true value of the prediction is shown as a black dashed line.

An examination of four random IES history-matched parameter fields (Figure 3.4) reveals a limited spatial continuity of either low permeability or high permeability zones, and less variance compared to the true field (Figure 3.1). In particular, this is observed in the area where the additional pumping well and the observation well of the prediction of interest are located. This results in a drawdown cone extension that is mostly constrained to the area of the pumping well, underestimating drawdown at the observation well. This is consistent with the underestimation of the predictive uncertainty observed in Figure 3.2. The minimum error variance solution $\underline{\mathbf{x}}$ obtained from PEST calibration is presented in Figure 3.5. This calibrated parameter field results in a measurement objective function less than 2% higher than the target measurement objective function. The calibrated bulk hydraulic conductivity parameter x_b is 0.02, which is near zero. Meanwhile, the minimum and maximum calibrated pilot point values are -1.2 and 1.1 (in log10 scale), respectively. In general, the calibrated parameter field shows a spatial disposition of variability that is similar to the parameter fields generated from the posterior samples obtained with IES. However, it is observed that the calibrated parameter field shows heterogeneity in the area of the observation wells, and homogeneity outside this area. This is the result of regularization constraints applied in the inversion process. Regularization is subdued in the area where there is drawdown data that informs parameters. In contrast, the calibrated parameter values for pilot points located away from the observation wells remain approximately the same as their initial preferred values as the observation dataset is lacking in any information to the contrary. As a result, the simulated maximum

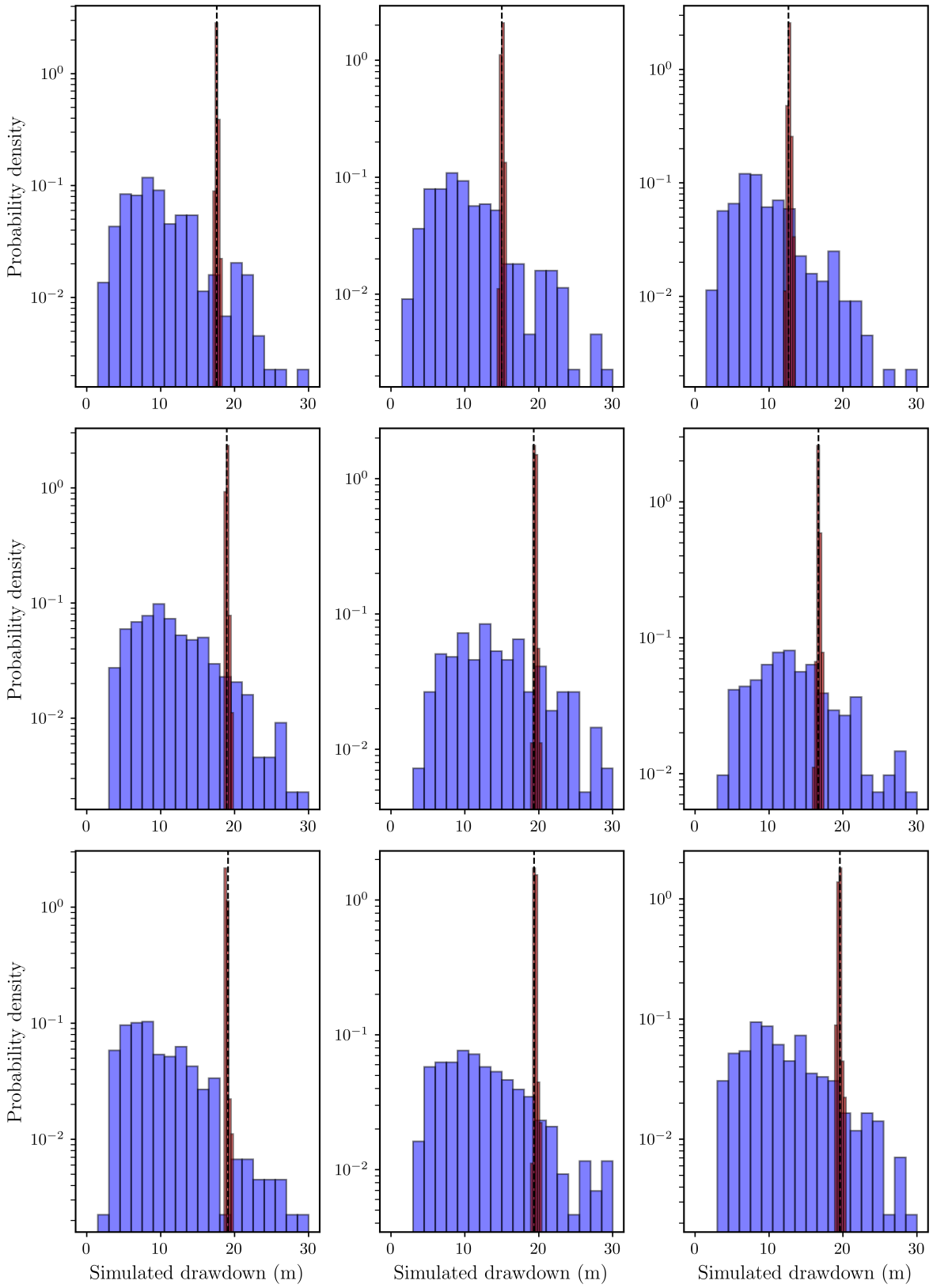

Figure 3.3: Prior (blue) and posterior (red) simulated drawdown histograms at the 9 observation wells that comprise the history matching dataset.

drawdown for future extraction at the observation well of interest is 10.77 m, similar to the mode of posterior predictive uncertainty obtained with IES using the 'wrong' prior (Figure 3.2). Recall that the true value of the prediction is 22.8 m.

Now assuming an uncertain prior, the posterior uncertainty of the hyperparameter vector $\boldsymbol{\theta}$ was estimated using MCMC. The calibrated pilot point parameter vector $\underline{\mathbf{x}}_{pp}$ was used for the hyperparameter inference. The resolution matrix was calculated using the regularization weight factor μ^2 of 4.25, obtained at the end of the PEST calibration process. Using the wrong prior, the calculated maximum absolute distance D between the empirical cumulative probability distribution and the theoretical cumulative probability distribution is equal to 0.5821, which is greater than the critical distance of 0.457. Also, the calculated sum of squares S using the wrong prior is 35.8648 which has a very low likelihood value according to the χ^2_8 pdf. Hence, the hypothesis that the calibrated parameter vector $\underline{\mathbf{x}}_{pp}$ is a sample of a multi-Gaussian distribution with a mean of zero and a covariance matrix $\mathbf{C}_{\mathbf{x}wrong}$ can be rejected, based on both metrics.

A set of 1000 realizations and 5 chains were used for MCMC, obtaining chain convergence according to the Gelman-Rubin statistic (values of 1.003 and 1.004,for the sill and the range, respectively). Figure 3.6 shows the prior and posterior probability distribution of sill and the effective correlation range. The posterior histograms were generated from samples obtained from MCMC inference. It can be observed that the sill posterior histogram is shifted towards higher values compared to the prior, with a maximum likelihood value of 0.6, approximately. The sill value of 0.25 used to generate the 'wrong' prior is outside the sill posterior distribution. On the contrary, the posterior histogram of the correlation range is similar to the prior, suggesting lack of information regarding this hyperparameter in the calibration parameter field.

The sill and correlation range joint prior (contours) and posterior (filled contours) probability density function, smoothed by kernel density estimation (KDE), is presented in Figure 3.7. The prior mode and true values are also shown in the figure. It can be observed that the posterior pdf is only shifted towards the true value of the sill. This suggests there is hyperparameter learning from the data, and the posterior probability distribution of the hyperparameter vector $\boldsymbol{\theta}$ appears more data-compatible than its prior. The correlation range does not show any apparent change, indicating that this hyperparameter is not informed by the data, in this case. Interesting enough is the shape of probability density distribution around the maximum a posteriori (MAP) value of the sill and the range, where a certain positive correlation between the two hyperparameters is observed. This is consistent with the true values of the hyperparameters, where a high sill value is associated with a high correlation range.

With the posterior uncertainty of the hyperparameter vector $\boldsymbol{\theta}$, the model is again subject to history matching using IES, starting from a prior parameter ensemble of size 300 sampled from a family of prior covariance matrices $\mathbf{C}_{\mathbf{x}}$. Figure 3.8 shows the prior and

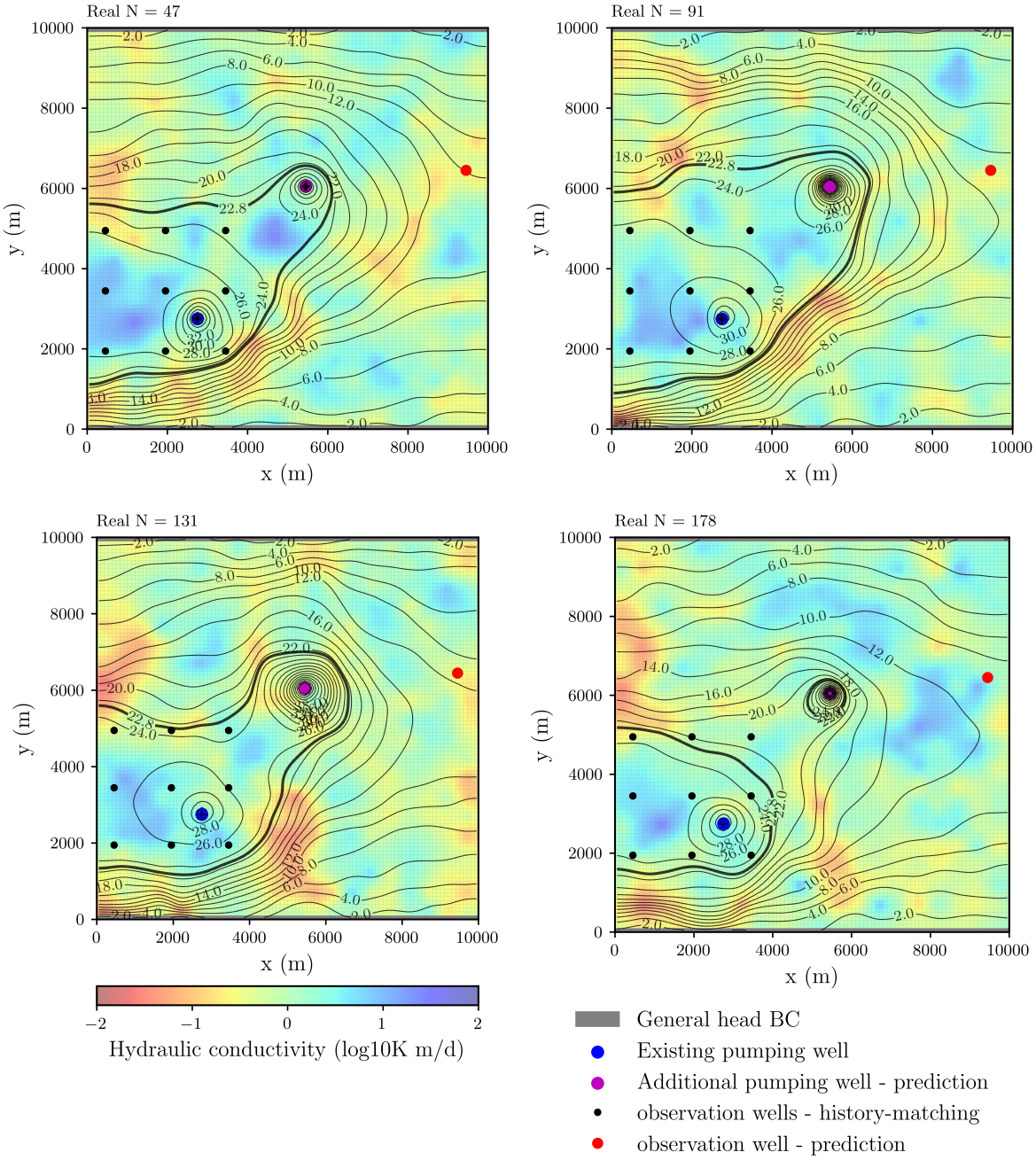

Figure 3.4: Random IES history-matched parameter fields using the 'wrong' prior. The thick black line represents the 22.8 m drawdown contour.

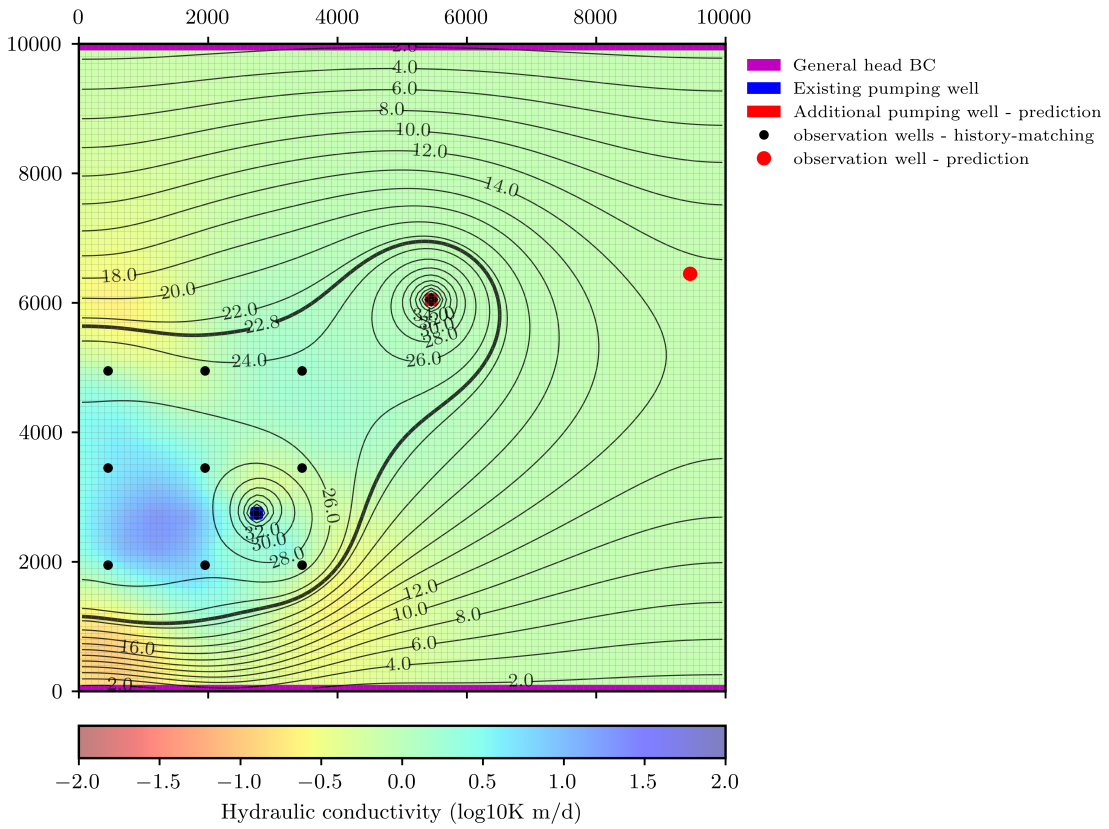

Figure 3.5: Calibrated parameter field obtained from PEST calibration using the 'wrong' prior. The thick black line represents the 22.8 m drawdown contour.

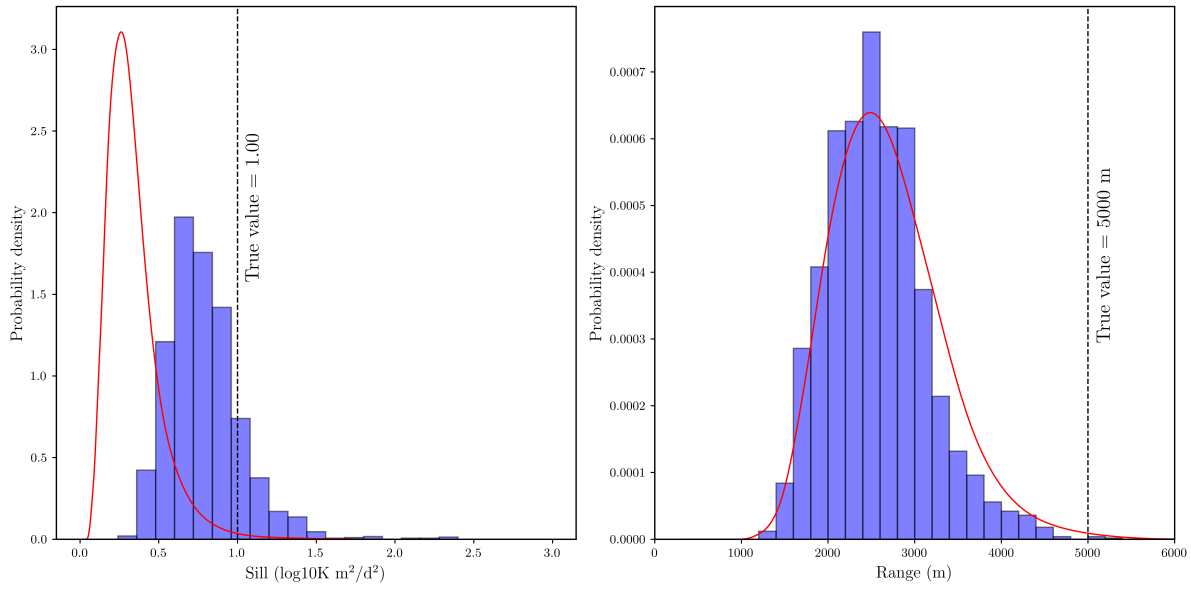

Figure 3.6: Posterior histograms of the sill (a) and the effective range (b) compared to their priors.

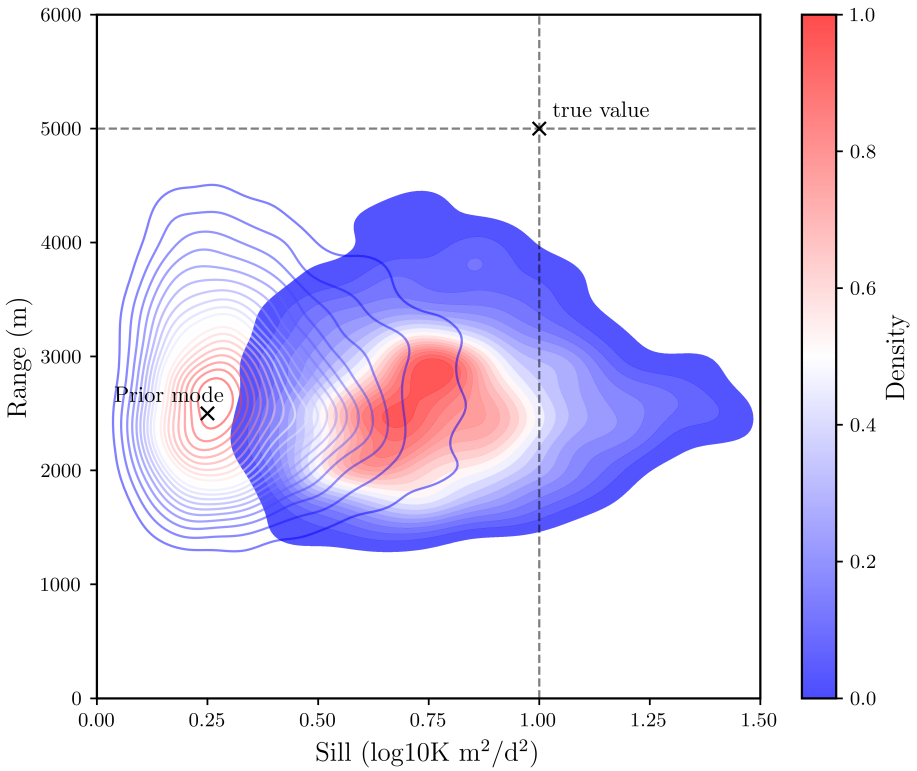

Figure 3.7: Sill and correlation range joint prior (contours) and posterior (filled contours) probability density functions.

posterior predictive uncertainty of drawdown at the observation well of interest, derived from the IES history matching process using the updated ensemble. It can be observed that the posterior predictive uncertainty covers the true value of the prediction. Also, comparing the prior and posterior predictive uncertainty, there is no significant reduction in predictive uncertainty. This is interpreted as a result of the prior being treated as uncertain. It is worth clarifying that the posterior predictive uncertainty may cover the true value of the prediction even with a wrong prior. However, what is demonstrated here is that, once it is assumed the prior is uncertain, its uncertainty can be made compatible with the calibrated parameter field, prior to performing predictive uncertainty quantification.

The examination of random IES history-matched parameter fields (Figure 3.9) using the posterior parameter ensemble derived from the IES history matching with the uncertain prior reveals more similarity with the true field (Figure 3.1). Also, the spatial continuity between both low permeability or high permeability zones is more pronounced compared to the parameter fields resulting from the previous history matching process. This leads to a drawdown cone extension that is more consistent with the true value of the prediction.

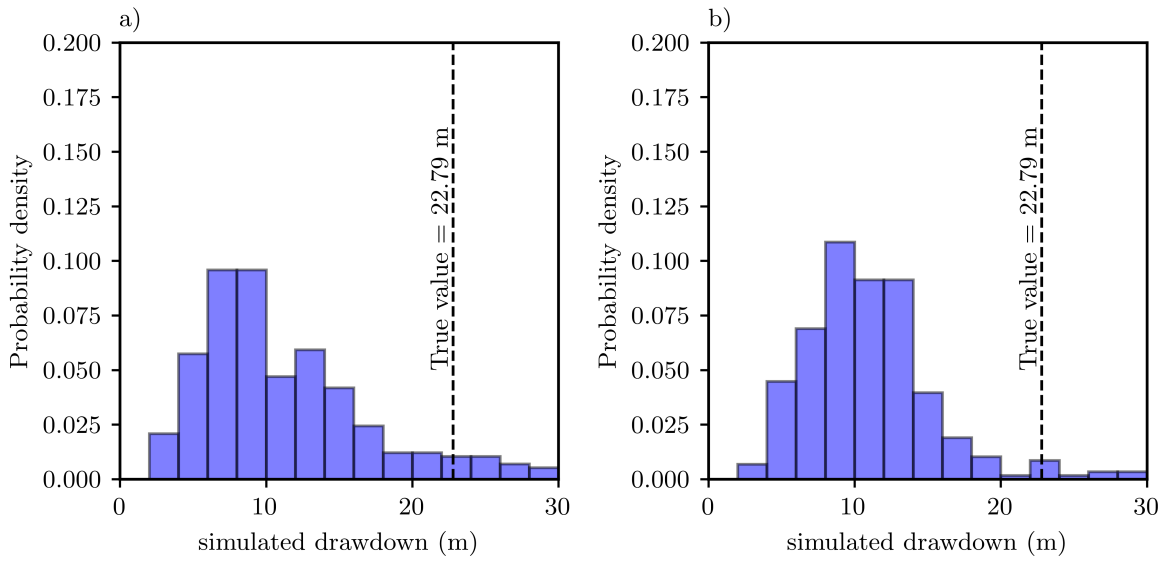

Figure 3.8: Prior (a) and posterior (b) predictive uncertainty of drawdown at the observation well of interest, derived from the IES history matching process with uncertain prior.

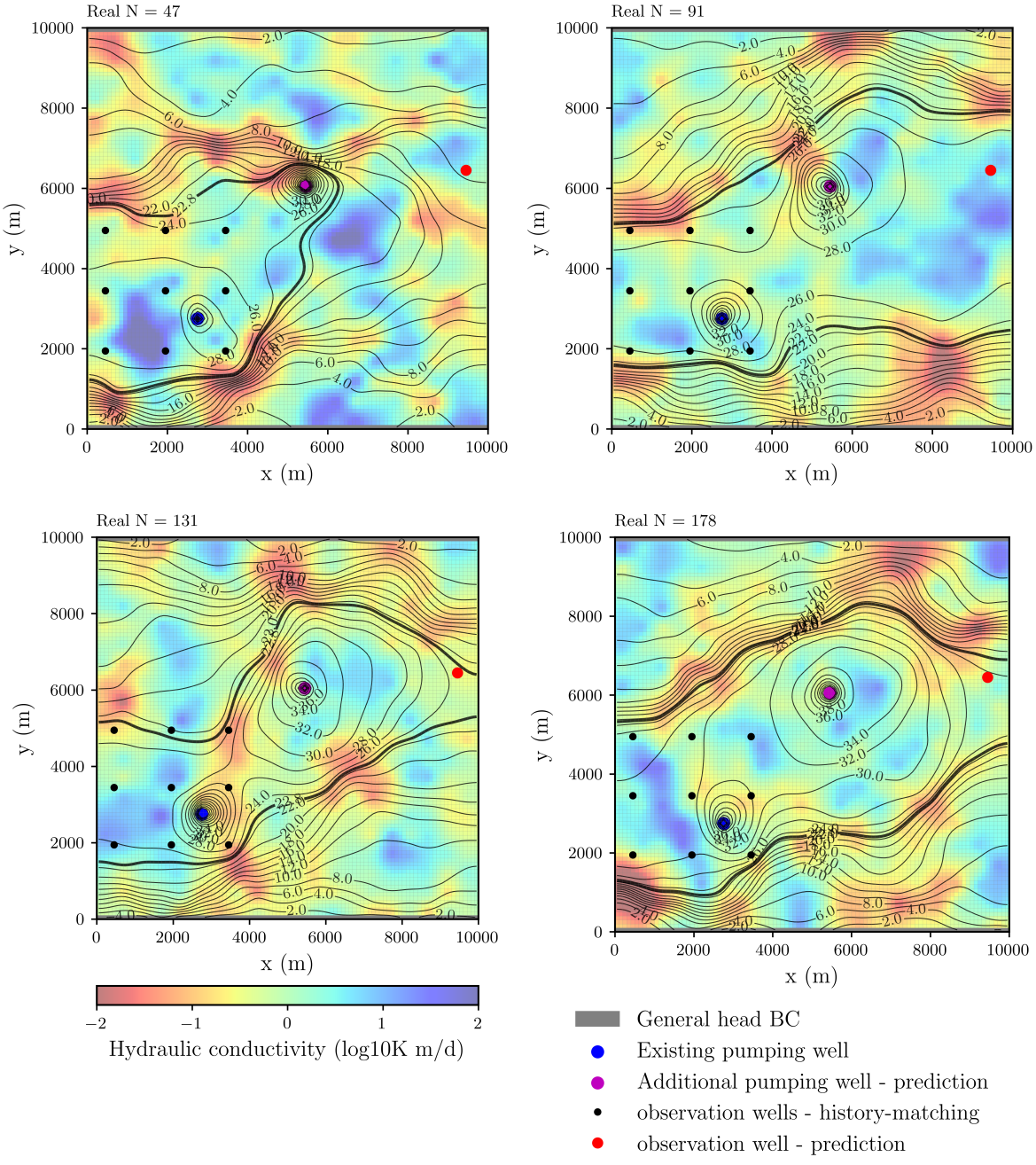

Figure 3.9: Random IES history-matched parameter fields using the updated ensemble.

3.4 Discussion

Given the inherent uncertainties of the distribution of hydraulic properties in the subsurface, it is likely that the way the prior is defined will be most of the time misspecified in real world modelling. This is especially the case for groundwater modelling, where data is limited. It becomes apparent from the test case results that treating the prior as uncertain helps minimize the underestimation of predictive uncertainty. Moreover, a new methodology was shown to be effective in updating the prior uncertainty of the hyperparameter vector $\boldsymbol{\theta}$ using the result of a minimum error variance solution obtained from regularized inversion. In this way, not only is it assumed that the prior is uncertain, but assurance is gained that its uncertainty is compatible with the calibrated parameter field. The predictive uncertainty obtained from history matching using a wrong and fixed prior is underestimated, as the true value of the prediction is outside the posterior predictive uncertainty limits. This is an outcome that in real-world groundwater modelling is impossible to verify, as the true value of the prediction is unknown. It is also demonstrated with the test case presented that the calibrated parameter field obtained from regularized inversion is not compatible with the prior, according to likelihood functions defined in the hyperparameter inference problem. Assuming the prior is uncertain, it is possible to update its uncertainty, represented by a hyperparameter vector $\boldsymbol{\theta}$, using the calibrated parameter field. An updated prior parameter ensemble generated from sampling several priors used to perform history matching with IES, resulted in an estimated posterior uncertainty covering the true value of the prediction.

Although regularized inversion is accompanied by an increased computational burden compared to ensemble methods, the resultant parameter field could surprise the modeller and lead to questioning the prior. Moreover, the calibrated parameter vector using regularized inversion is the best solution that minimize the deviation from the prior. The heterogeneity that appears in the calibrated parameter field is required to fit the data to the level of measurement noise. This is not the case when using ensemble methods. The fact that a calibrated parameter field surprises a modeller even when prior-data conflict is not evident from a comparison of model outputs with field measurements, may inform a modeller that the resulting parameter field lies in a probability region where the prior has low support. This is a sign of prior-data conflict, as defined by Evans and Moshonov (2006). The meaning of the prior in respect to the likelihood (model) can be put into question, especially if the model is a simplification of the real system (which is always the case). This lends support to the notion that the prior can only be understood in the context of the likelihood (Gelman et al., 2017).

Some limitations of the presented methodology are worth mentioning. It is acknowledged that the proposed methodology does not demonstrate the requirement of priors compatible with the calibrated parameter field for a successful estimation of predictive

uncertainty. In other words, there may be cases whereby only assuming uncertain priors, will lead to conservative estimates of predictive uncertainty. Although the application of the methodology to a synthetic numerical example has shown that the posterior predictive uncertainty covers the true value of the prediction, this is not a generalized conclusion. What the proposed methodology provides is an increased level of confidence in the predictive uncertainty estimation, using uncertain priors that are compatible with the calibrated parameter field, as performed in Empirical Bayesian methods (Robert, 2007).

It is also important to recognize that the selection of the likelihood functions for the hyperparameter inference problem is a subjective choice. Converting the Kolmogorov-Smirnov test into a pseudo likelihood function is a crude approximation. However, the way likelihood functions were defined were effective in sampling the posterior of the hyperparameter vector $\boldsymbol{\theta}$. Other more sophisticated methods could be used. These are left for future work.

As demonstrated in this work, there is an added value of identifying prior-data conflict (at least indirectly) and updating the prior before performing predictive uncertainty quantification. When using ensemble methods, and without the application of the proposed methodology, history matching and predictive uncertainty quantification are part of the same task. The history-matched parameter fields will not necessarily show evidence of conflict with the prior, especially when using ensemble methods whose posterior samples are linear combinations of the prior ensemble members (Evensen et al., 2019), derived from an incorrect prior. This suggests that there may be circumstances where performing regularized inversion first and predictive uncertainty quantification later is a good idea.

3.5 Conclusions

In this work, a methodology is proposed to identify prior-data conflict and update the uncertainty of the prior, represented by hyperparameters, using a minimum error variance solution obtained from regularized inversion.

The proposed methodology is tested using a synthetic 2D groundwater model simulating drawdown due to pumping. The results show that by using an uncertain prior that is compatible with the calibrated parameter field, the posterior predictive uncertainty covers the true value of the prediction. Moreover, the predictive uncertainty is more conservative compared to the case where an incorrect prior is used.

It is concluded that, in certain circumstances, separating history matching from predictive uncertainty quantification may be beneficial. By obtaining a minimum error variance solution to the regularized inversion, the modeller can assess the compatibility of the prior with the calibrated parameter field, embrace its uncertain nature, and constraint its uncertainty using the calibrated parameter vector, before performing predictive uncertainty quantification. Using the updated uncertain prior, it is hoped that a more conservative

estimate of predictive uncertainty will encompass the true value of the prediction, leading to increased confidence in groundwater modelling as a decision-support tool.

Chapter 4

Accommodating Uncertain and Nonstationary Priors in History Matching and Predictive Uncertainty Quantification for Groundwater Models

Author contributions

T. Opazo: Conceptualization 90%, Realization 100%, Writing 70%. J. Doherty: Conceptualization 10%, Writing 30%, Review 100%.

Manuscript in preparation for submission to Advances in Water Resources: Opazo, T., Doherty, J. Accommodating Uncertain and Nonstationary Priors in History Matching and Predictive Uncertainty Quantification for Groundwater Models.

Abstract

Ensemble methods are efficient ways of estimating predictive uncertainty in groundwater modelling. However, their results are constrained by the prior parameter probability distribution, or the prior, for short. If the prior is misspecified, predictive uncertainty estimates can be biased and underestimated. This is expected to be the rule rather than the exception, as the subsurface hydraulic properties can be highly heterogeneous and nonstationary. This study presents a novel methodology for accommodating uncertain and nonstationary priors in history matching and predictive uncertainty quantification of groundwater models. The approach improves the decision-support utility of groundwater models by relaxing the assumption of geostatistical stationarity and enhancing

their ability to generate realistic patterns of heterogeneity. The methodology employs a hierarchical two-level parameterization scheme that integrates spatially variable geostatistical hyperparameters with spatially distributed parameters. Two numerical examples are used to test the methodology: a 2D aquifer hydraulic conductivity model and a 2D flow and transport model simulating solute extraction and reinjection. Three history matching methods are compared: Subspace Ensemble Randomized Maximum Likelihood (SEnRML), Levenberg-Marquardt Ensemble Randomized Maximum Likelihood (LM-EnRML), and Data Space Inversion (DSI). Results show that the proposed methodology effectively handles uncertain and nonstationary priors, achieving reasonable fits to the data and acceptable predictive uncertainty estimates. While ensemble methods such as SEnRML and LM-EnRML face challenges in highly nonlinear problems, DSI provides a computationally efficient alternative for predictive uncertainty quantification. However, DSI lacks the ability to generate physical parameter fields that can be assessed for geological realism.

4.1 Introduction

When performing maximum a posteriori (MAP) estimation (i.e. history matching and uncertainty quantification), a prior probability distribution, or prior, must be assigned to model parameters that represent subsurface hydraulic properties, boundary conditions, and any other uncertain model inputs. This is an important step, especially in contexts where the history matching dataset is information-poor, and/or where decision-critical model predictions are sensitive to parameter components that are relatively unconstrained by history matching. Often, however, the prior is itself uncertain and therefore possibly wrong. This is because the prior is a subjective expression of expert knowledge based on limited and sparse data.

Attainment of MAP or stochastic solutions to groundwater history matching problems generally requires manipulation of large numbers of parameters. This applies particularly to models whose predictions of interest are sensitive to hydraulic property and/or hydraulic process detail. Inclusion in a model of parameterization complexity that reflects (as best extent it can) site hydraulic property complexity reduces the likelihood of history-match-induced predictive bias and increases the likelihood that predictive uncertainty is not underestimated (Doherty, 2015; White et al., 2014). Unless a model is equipped with the capacity to undertake adjoint sensitivity calculations, ensemble methods and variants thereof provide computationally feasible means of accommodating appropriate parameterization complexity in the history matching process. Their efficiency relies on the way they perform history matching and uncertainty analysis, as realizations drawn from the prior parameter probability distribution are simultaneously optimized until they approximate the posterior parameter probability distribution. See Evensen et al. (2022)

for a summary of the extensive literature and numerous variants of highly-parameterised, ensemble-based Bayesian history matching.

Naturally, the numerical attractiveness of ensemble methods is accompanied by certain disadvantages. Their computational economy relies on inclusion in any ensemble of only a moderate number of parameter field realizations, often of the order of a few hundred. This can hamper their numerical capacity to solve inverse problems that are characterized by a highly nonlinear relationship between model outputs and parameters. Samples of the posterior parameter probability distribution that are calculated using ensemble methods are linear combinations of samples of the prior parameter probability distribution that are used to initiate the ensemble adjustment process, or nearly so, depending on the ensemble-based history matching methodology that is employed. Moreover, assimilating data from a history matching dataset using ensemble methods requires projection of that data onto a model range space that is limited by the dimensions of the ensemble (Evensen et al., 2022). Therefore, the success of the history matching process depends on the ability of the prior ensemble to span the posterior parameter probability distribution, and whether, or not, observations lie in the model range space. If these conditions are not met, ensemble-based history matching can induce predictive bias, ensemble collapse and predictive uncertainty underestimation. Sometimes, these problems can be ameliorated through adoption of an appropriate localization strategy; see, for example, Furrer and Bengtsson (2007); Chen and Oliver (2017); Luo et al. (2018); Luo and Bhakta (2020), to name a few studies. Nevertheless, where the relationship between history-match-pertinent model outputs and parameters is highly nonlinear, uncertainties that are evaluated using ensemble methods should be treated as indicative only (Evensen, 2018). However, in highly parameterized contexts, where model run times are long and adjoint sensitivity methods are unavailable, there is no other option but to resort to some type of ensemble method for history matching and uncertainty quantification.

One option to partially mitigate some limitations of ensemble methods in particular, and Bayesian methods in general, is to assume the prior as uncertain. This can be incorporated into the history matching process in different ways. A simple approach is to generate an initial prior parameter ensemble by sampling several prior probability distributions and perform history matching using this ensemble. Emerick (2016) presented a method that, under an assumption of geostatistical stationarity, allows weight-based deployment of a few different prior probability distributions in propagating a prior ensemble to a posterior ensemble. He pointed out the challenges that accompany including multiple priors in history matching highly parameterized problems, and remarked on how few researchers had, up until that time, attempted to solve these kinds of problems using ensemble methods. A more complex approach is to include hyperparameters that describe the uncertainty of the prior in the analysis, and to adjust them during the history matching process. This is called hierarchical Bayes analysis (Robert, 2007). Oliver

(2022) formulated the hierarchical inverse problem for ensemble-based reservoir parameter and predictive uncertainty analysis using a non-centred parameterization of spatial stochastic variability; adoption of such a parameterization scheme is key to the success of hierarchical methods (Chada et al., 2018). Using this methodology, a spatially-correlated parameter field is constructed through spatial integration, or spatial averaging (Oliver, 1995), over a field of independent standard normal deviates. Oliver (2022) demonstrated how nonlinearity of the hierarchical inverse problem poses severe difficulties for conventional ensemble approaches because of the approximate nature of the Jacobian matrix that is used to update all parameter realizations at once. He ameliorated this problem by analytically calculating realization-specific Jacobian elements whose values are partially dependent on realisation-specific hyperparameter values, while the remainder of the parameter sensitivities are calculated from the ensemble. His work demonstrated that the numerical difficulties associated with history matching and predictive uncertainty quantification increase dramatically when uncertainties in the prior are admitted into the analysis.

Problems associated with an uncertain prior are exacerbated where the prior parameter probability distribution is nonstationary. In this case hyperparameters which characterize hydraulic property variability, as well as hydraulic properties themselves, are spatially variable. This approach has been explored by Chada et al. (2018), among others. As will be demonstrated, the nonlinearity of this high-dimensional inverse problem poses severe difficulties for posterior predictive uncertainty analysis. This is particularly the case for ensemble methods, on which reliance must be placed where parameter numbers are high and model run times are long.

Regardless of the numerical difficulties that attend it, solution of this type of inverse problem is a matter of some urgency. This is because it is typical of many, if not most, circumstances in which decision-support groundwater modelling is required. Inspired by the work of Oliver (1995, 2022), Higdon et al. (1999), Fuentes (2001), Paciorek and Schervish (2006) and Chada et al. (2018), this problem is approached by representing uncertain subsurface nonstationarity within a two-level hierarchical framework (Robert, 2007) through spatial averaging of uncorrelated random deviates. This aims to enhance the decision-support utility of groundwater model history matching and uncertainty quantification by relaxing the assumption of geostatistical stationarity, thereby increasing the capacity to generate patterns of heterogeneity which are realistic enough to reflect the properties of real geological media while maintaining history-match adjustability of these patterns. Ideally this flexibility reduces bias incurred by incorrect assumptions pertaining to the prior, at the same time as it ensures the integrity of posterior predictive probability distributions attained through Bayesian analysis. The performance of a number of ensemble and related methods in attempting to solve a problem that is posed in this way is investigated, by assessing their numerical efficiency, as well as the quality of the

predictive uncertainty estimates that these methods yield. Although this work builds on previous work, the methodology presented here is new and has not been previously applied to groundwater model history matching and uncertainty quantification.

This chapter is organized as follows. First, a brief outline of the theoretical background for spatial averaging as a mechanism for generation of stochastic fields is discussed. An extension of the Oliver (1995) methodology for applying spatial averaging in a hierarchical way that includes nonstationarity is then presented. Next, three history matching and predictive uncertainty quantification methods that are used for data assimilation and uncertainty quantification are described. These are the iterative ensemble smoother LM-EnRML of Chen and Oliver (2013), subspace iterative ensemble smoother SEnRML (Raanes et al., 2019; Evensen et al., 2019), and data space inversion (Sun and Durlofsky, 2017; Sun et al., 2017). These methods are tested using two numerical examples where hydraulic conductivity has a nonstationary distribution. After demonstrating and discussing the performance of these methods, some consequences for decision-support groundwater modelling are discussed. Of particular interest is whether the evaluation of predictive uncertainty requires the evaluation of parameter uncertainty. This question is prompted by the difficulties that are associated with the latter, and the high levels of model run efficiency that are attainable through methods such as data space inversion.

It is important to acknowledge that all numerical methods have their strengths and weaknesses. While there is a strong temptation to compare them, this is not the objective of this work. On the contrary, the aim is to evaluate the suitability of the proposed methodology in enhancing the history matching and uncertainty quantification capacity of these methods within a nonstationary framework and to examine and explain the difficulties that attend respect for the intricate complexity of the unknown subsurface through which groundwater flows, when undertaking numerical simulation. A discussion on the variety of alternatives that are available for addressing these difficulties is also offered.

4.2 Methodology

4.2.1 Spatial Averaging and Nonstationary Fields

Before embarking on history matching and uncertainty quantification, expert knowledge on model parameters must be expressed using a prior parameter probability density function, or 'prior' for short. Where parameters represent hydraulic properties, their spatial correlation is often characterized by covariance functions (sometimes encapsulated in variograms).

Zero-mean stochastic fields of a parameter vector $\mathbf{x}$ which exhibit spatial correlation can be generated by the method of moving averages according to the following convolution

integral (Oliver, 1995):

$$\mathbf{x}(\mathbf{y}) = \int_{-\infty}^{\infty} f(\mathbf{y} - \mathbf{s}) z(\mathbf{s}) d\mathbf{s}, \tag{4.1}$$

where Function f is referred to as the spatial averaging kernel. z represents Brownian motion, and $\mathbf{y}$ and $\mathbf{s}$ are locations in space. Where the stochastic field is discretized to a model grid, Equation 4.1 becomes a summation while z becomes a set of independent normal standard deviates.

Oliver (1995) derived analytical expressions for the spatial averaging kernel that results in several well-known covariance functions that characterize the stochastic field $\mathbf{x}$, these including exponential, spherical, and Gaussian. For example, an isotropic Gaussian covariance stochastic model is specified by the following function:

$$C(r) = \sigma^2 \exp\left(-\frac{r^2}{a^2}\right), \tag{4.2}$$

where $C(r)$ is the covariance between two points separated by distance r, σ^2 is the variance of the stochastic field, and a is the correlation length. The corresponding 2D spatial averaging kernel which yields this covariance function is the following:

$$f(r) = \sigma \sqrt{\frac{4\pi}{a^2}} \exp\left(-\frac{2r^2}{a^2}\right). \tag{4.3}$$

As is apparent from Equation 4.2 and Equation 4.3, the a value of a Gaussian covariance model is $\sqrt{2}$ times that used in the spatial averaging kernel. Also, as discussed by Oliver (1995) and Oliver (2022), the standard deviates over which integration takes place should be extended beyond the model grid to avoid edge effects. In this work, this is not considered, as boundary effects are not key for the synthetic models used in this study. The same assumption was made by Oliver (2022) in his study.

While stationary and isotropic hydraulic property fields may be useful expressions of prior uncertainty in some geological settings, stochastic subsurface hydraulic property variability is more likely to be nonstationary, and exhibit spatially-variable anisotropy, over the large domains that characterize many groundwater models. Higdon et al. (1999) showed that the spatial averaging kernel must be a function of location in order to generate nonstationary stochastic parameter fields using the moving average method. Use of a Gaussian spatial averaging kernel yields nonstationary spatial covariance functions that remain Gaussian, and that are therefore amenable to characterization using tractable expressions (Higdon et al., 1999; Paciorek and Schervish, 2006). Alternatively, use of an arbitrary spatial averaging kernel that is described by spatially varying hyperparameters yields stochastic fields that are characterized by spatial variability of covariance that is not amenable to simple analytic description. This is a matter of concern where local geostatistical characterization is undertaken in order to support interpolation and estimation

of interpolation uncertainty between measurement points; see, for example, Paciorek and Schervish (2006) and Fuentes (2001). However, it matters less where stochastic fields are generated in order to simulate patterns of hydraulic property heterogeneity in groundwater model domains.

In this work, an extended version of the spatial averaging method is proposed. The method consists of the generation of two levels of stochastic fields, in a hierarchical manner, that allows representation of nonstationary spatial variability of hydraulic properties. The hierarchical model is schematized in Figure 4.1.

In the first level (Level 1), a hyperparameter vector $\boldsymbol{\theta}_i$ is used to represent the spatial variability of the i-th hyperparameter type, such as variance, correlation range, anisotropy factor (the ratio of maximum to minimum correlation length), or anisotropy angle, that pertain to hydraulic properties that populate a model grid. The dimension of vector $\boldsymbol{\theta}_i$, is equal to n, the dimension of the model grid. For each spatially-varying hyperparameter type, a standard deviate parameter vector $\mathbf{z}_{\boldsymbol{\theta}_i} \in \mathbb{R}^m$ is defined, where m can be smaller than n. The smaller m the smoother the representation of the spatial variability of the hyperparameter type will be. In most cases, there is no need to have $m = n$, as the desired spatial variability of hyperparameters rarely requires the same level of detail as the hydraulic properties themselves. As shown in Equation 4.3, the hyperparameter kernel $f_{\boldsymbol{\theta}_i}$ is defined by a standard deviation $\sigma_{\boldsymbol{\theta}_i}$ and a hyperparameter correlation range $a_{\boldsymbol{\theta}_i}$, for each i-th hyperparameter type. Although these quantities can be also treated as uncertain and potentially spatially-variant, in this study, they are assumed known. In the case of the mean hyperparameter vector $\overline{\boldsymbol{\theta}}_i$, it is assumed to be spatially invariant over the model domain, but uncertain. Then, the vector $\boldsymbol{\theta}_i$, that holds all hyperparameter values over the model grid, is the result of the sum of the mean hyperparameter vector $\overline{\boldsymbol{\theta}}_i$ and the convolution of an isotropic and stationary 'Level 1' Gaussian spatial averaging kernel $f_{\boldsymbol{\theta}_i}$ with the standard deviate vector $\mathbf{z}_{\boldsymbol{\theta}_i}$, as

$$\boldsymbol{\theta}_i(\mathbf{y}) = \overline{\boldsymbol{\theta}}_i + \sum f_{\boldsymbol{\theta}_i}(\mathbf{y}^* - \mathbf{s}) \cdot \mathbf{z}_{\boldsymbol{\theta}_i}(\mathbf{s}) \cdot \Delta s, \tag{4.4}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the model grid location vector. Δs is the standard deviates' grid spacing, and $\mathbf{y}^*$ is the scaled model grid location vector. As part of this methodology, standard deviates are homogeneously distributed on an imaginary grid with unitary distance between them. Then the geometry measure Δs is equal to 1.0. The way standard deviates are arranged in the imaginary grid is not relevant, as long as they preserve the unitary distance between them. The model grid locations vector $\mathbf{y}$ is then scaled (and offset if required) to this imaginary grid, resulting in a new set of coordinates $\mathbf{y}^*$. Consistently with this transformation, the correlation range $a_{\boldsymbol{\theta}_i}$ is also scaled.

In the second level (Level 2) of the hierarchical model parameterization scheme, the objective is to generate model hydraulic properties that populate the model grid. Equivalent

to the first level, the model parameter vector $\mathbf{x}(\mathbf{y})$ is generated by the sum of a mean vector $\overline{\mathbf{x}}$ and the convolution of a Gaussian kernel $f_{\mathbf{x}}$ with a set of standard deviates $\mathbf{z}_{\mathbf{x}} \in \mathbb{R}^p$ as follows:

$$\mathbf{x}(\mathbf{y}) = \overline{\mathbf{x}} + \sum f_{\mathbf{x}}(\mathbf{y}^* - \mathbf{s}) \cdot \mathbf{z}_{\mathbf{x}}(\mathbf{s}) \cdot \Delta s. \tag{4.5}$$

Note that the dimension p of the standard deviate vector $\mathbf{z}_{\mathbf{x}}$ can be different from the dimension n of the model grid, and also different from the dimension m of the first level standard deviate vectors $\mathbf{z}_{\boldsymbol{\theta}_i}$. For this reason, the model grid locations vector $\mathbf{y}$ is scaled to the imaginary grid of standard deviates using a potentially different scale factor than the one used in Level 1. The added complexity at this level is that the kernel $f_{\mathbf{x}}$ is not isotropic nor stationary. Therefore, the application of Equation 4.3 is not straightforward. In this case, each parameter x_j at location $\mathbf{y}_j$ is still obtained from Equation 4.5, but an 'effective' distance r'_j in the kernel function $f_{\mathbf{x}}$ is calculated for each location j, as follows:

$$r_j = \sqrt{(\mathbf{y}_j^* - \mathbf{s}_j)^T \mathbf{H}_{\mathbf{j}} (\mathbf{y}_j^* - \mathbf{s}_j)}, \tag{4.6}$$

where $\mathbf{H}_{\mathbf{j}}$ is defined as

$$\mathbf{H}_j = \mathbf{R}_j^T \mathbf{S}_j^T \mathbf{S}_j \mathbf{R}_j. \tag{4.7}$$

Matrices $\mathbf{R}_j$ and $\mathbf{S}_j$ are the rotation and scaling matrices, respectively, constructed as

$$\mathbf{R}_j = \begin{bmatrix} \cos(\alpha_j) & \sin(\alpha_j) \\ -\sin(\alpha_j) & \cos(\alpha_j) \end{bmatrix}, \tag{4.8}$$

and

$$\mathbf{S}_j = \begin{bmatrix} 1.0 & 0 \\ 0 & \eta_j \end{bmatrix}, \tag{4.9}$$

where α_j is the anisotropy angle, and η_j is the anisotropy factor, at location j of the model grid. Note that the anisotropy angle is measured with respect to the x-axis, and is positive in the anticlockwise direction.

Stochastic population of a model grid in this way requires random sampling of $\mathbf{z}_{\boldsymbol{\theta}_i}$ for each i-th hyperparameter type, and $\mathbf{z}_{\mathbf{x}}$. As stated above, this is easily done as elements of these vectors are samples of independent standard normal distributions. Where a model is history-matched, $\mathbf{z}_{\boldsymbol{\theta}_i}$ and $\mathbf{z}_{\mathbf{x}}$ become history matching parameters. As they follow standard normal distributions, they are well suited to history matching using ensemble methods (whose Bayesian roots are based on the assumption that model parameters are Gaussian).

Hierarchical stochasticity of the kind discussed above supports generation of nonstationary parameter fields. At the same time, hierarchical parameterization of this stochasticity supports adjustment of parameter fields during history matching. Not only, therefore, can the locations of hydraulic property heterogeneity be estimated through history matching.

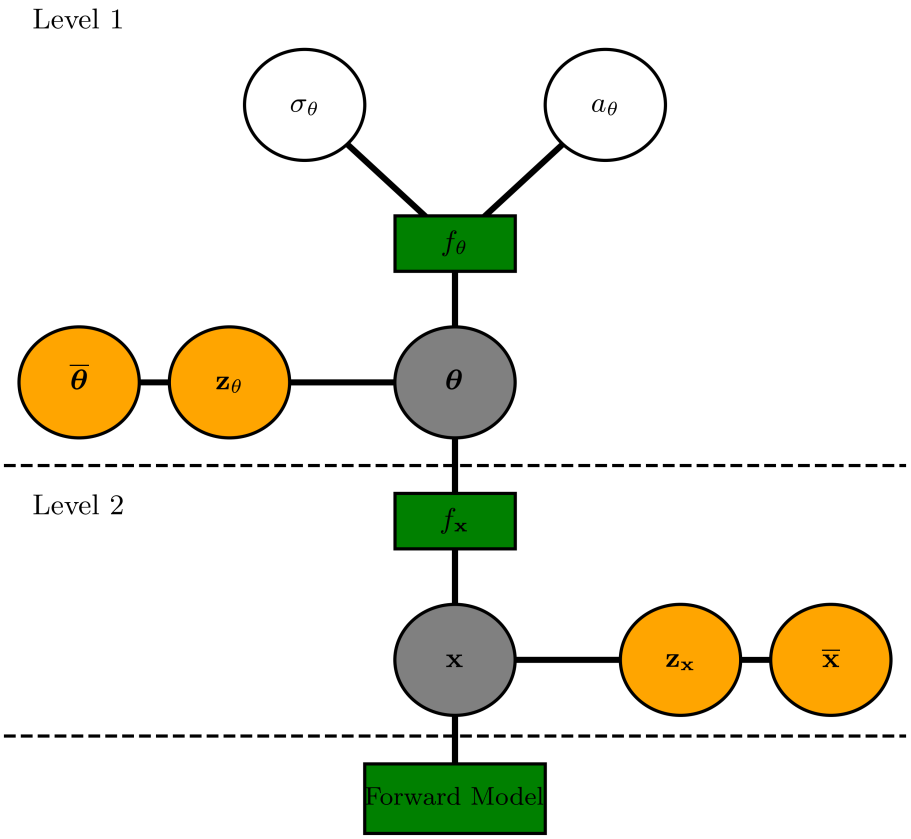

Figure 4.1: Hierarchical model for the generation of nonstationary stochastic fields. Hyperparameters assumed as uncertain are filled with an orange colour. The first level (Level 1) generates hyperparameters $\boldsymbol{\theta}$ using a Gaussian kernel f_θ and independent standard normal deviates $\mathbf{z}_{\boldsymbol{\theta}}$. The second level (Level 2) generates model hydraulic properties using a kernel f_x and independent standard normal deviates $\mathbf{z}_{\mathbf{x}}$.

Their patterns can also be subject to history matching constraints. This adds a high level of flexibility, but also complexity, to the history matching process.

4.2.2 History-Matching and Uncertainty Quantification

In this section, three methods for history matching and uncertainty quantification are summarized. These include two versions of ensemble randomized maximum likelihood (EnRML) methods (Chen and Oliver, 2013; Raanes et al., 2019; Evensen et al., 2019), and data space inversion (Sun and Durlofsky, 2017; Sun et al., 2017). For a more detailed explanation of these methods, the reader is referred to Chapter 2 of this thesis or to the original papers.

The LM-EnRML method, proposed by Chen and Oliver (2013), implements the Levenberg-Marquardt (LM) algorithm to damp the update of model parameters, and avoids the explicit calculation of the average sensitivity matrix, which is noisy when calculated from an ensemble (Chen and Oliver, 2012). The parameter update equation for the LM-EnRML method is obtained as follows:

$$\begin{aligned}\delta\mathbf{x} = &-\mathbf{S_x}^{-1/2}\mathbf{A}^l\left((1+\lambda^l)\mathbf{I}_n + \mathbf{Y}^{l^T}\mathbf{Y}^l\right)^{-1}\mathbf{A}^{l^T}\mathbf{A}^{0^{-T}}\mathbf{A}^{0^{-1}}\mathbf{S_x}^{-1/2}(\mathbf{x}^l - \mathbf{x}^f)\\ &-\mathbf{S_x}^{1/2}\mathbf{A}^l\mathbf{Y}^{l^T}\left((1+\lambda^l)\mathbf{I}_m + \mathbf{Y}^l\mathbf{Y}^{l^T}\right)^{-1}\mathbf{S_y}^{-1/2}\left(\mathbf{g}(\mathbf{x}^l) - \mathbf{d}\right)\end{aligned} \tag{4.10}$$

where $\mathbf{S_y}$ and $\mathbf{S_x}$ are diagonal scaling matrices with diagonal elements equal to the variance of data noise and the prior variance of model variables, respectively, $\mathbf{A}^0$ and $\mathbf{A}^l$ are the initial and updated (at iteration l) scaled model parameter anomalies matrices. $\mathbf{Y}^l$ is the scaled model output anomalies matrix, $\mathbf{x}^f$ is the initial model parameter vector, $\mathbf{x}^l$ is the updated model parameter vector at iteration l, $\mathbf{g}(\mathbf{x}^l)$ is the model output vector at iteration l, and $\mathbf{d}$ is the data vector. Finally, λ^l is the LM damping factor at iteration l. An approximate version of Equation 4.10 called LM-EnRML(approx) discards the first term of the right-hand side of the equation, which is the term that minimizes the parameter distance to the prior estimate. This approximation is not used in this study.

One of the key assumptions declared by Chen and Oliver (2013) is that the prior covariance matrix of model parameters $\overline{\mathbf{C}}_\mathbf{x}$ in the Hessian term of the original parameter update equation (not shown here for brevity) is replaced by another matrix $\overline{\mathbf{P}}_\mathbf{x}^\mathbf{l}$ that is calculated from the updated ensemble (that changes every iteration) as

$$\overline{\mathbf{P}}_\mathbf{x}^\mathbf{l} = \mathbf{S_x}^{1/2}\mathbf{A}^l\mathbf{A}^{l^T}\mathbf{S_x}^{1/2}. \tag{4.11}$$

The SEnRML method, proposed by Raanes et al. (2019), avoids some approximations made in the LM-EnRML method, especially the assumption of the prior covariance matrix

$\overline{\mathbf{C}}_{\mathbf{x}}$ in the Hessian term of the parameter update equation. In this method, the parameter solution is a linear combination of the initial ensemble anomalies and the first guess (Evensen et al., 2019),

$$\mathbf{X}^{l} = \mathbf{X}^{f} + \mathbf{A}\mathbf{W}^{l}, \tag{4.12}$$

where $\mathbf{X}^{f}$ and $\mathbf{X}^{l}$ are the first guess and updated model parameter ensemble realizations, respectively. Matrix $\mathbf{A}$ is the matrix of initial ($l = 0$) model parameter ensemble anomalies, and $\mathbf{W}^{l} \in \mathbb{R}^{N}xN$ is the matrix of weights. Solving the problem in this way, the inversion process is naturally regularized. The weights are iteratively updated as follows:

$$\mathbf{W}^{l+1} = \mathbf{W}^{l} - \gamma \left(\mathbf{W}^{l} - \mathbf{S}^{l^{T}} (\mathbf{S}^{l}\mathbf{S}^{l^{T}} + \mathbf{C}_{\mathbf{d}})^{-1} \mathbf{H}^{l} \right), \tag{4.13}$$

where γ is the Gauss-Newton step length, and $\mathbf{H}^{l}$ is the 'innovation' term (Evensen et al., 2019) defined as

$$\mathbf{H}^{l} = \mathbf{S}^{l}\mathbf{W}^{l} + \mathbf{D} - \mathbf{g}(\mathbf{X}^{f} + \mathbf{A}\mathbf{W}^{l}). \tag{4.14}$$

The only matrix that requires inversion in Equation 4.13 is $\mathbf{S}^{l}\mathbf{S}^{l^{T}} + \mathbf{C}_{\mathbf{d}}$, where $\mathbf{C}_{\mathbf{d}}$ is the covariance matrix of the data noise, and $\mathbf{S}^{l}$ is the matrix of predicted and 'deconditioned' ensemble anomalies at iteration l. There are several options for inverting this matrix as presented by Evensen et al. (2019). In this study, the low-rank inversion is used, which is defined as

$$(\mathbf{S}^{l}\mathbf{S}^{l^{T}} + \mathbf{E}\mathbf{E}^{T})^{-1} = \left(\mathbf{U}\mathbf{\Sigma}^{+^{T}}\mathbf{Z} \right) (\mathbf{I}_{N} + \mathbf{\Lambda})^{-1} \left(\mathbf{U}\mathbf{\Sigma}^{+^{T}}\mathbf{Z} \right)^{T}, \tag{4.15}$$

where $\mathbf{E}$ is the ensemble of data noise anomalies, $\mathbf{U}$, $\mathbf{\Sigma}^{+}$, are the eigenvector matrix and pseudo-inverse of singular values matrix, derived from SVD decomposition of $\mathbf{S}^{l}$. Matrices $\mathbf{Z}$ and $\mathbf{\Lambda}$ and are eigenvectors and singular values of the following:

$$\mathbf{\Sigma}^{+}\mathbf{U}^{T}\mathbf{E}\mathbf{E}^{T}\mathbf{U}\mathbf{\Sigma}^{+^{T}} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^{T}. \tag{4.16}$$

The final ensemble parameter update $\mathbf{X}^{l+1}$ is calculated as

$$\mathbf{X}^{l+1} = \mathbf{X}^{f}(\mathbf{I} + \mathbf{W}^{l+1}/\sqrt{N-1}). \tag{4.17}$$

Localization is a key aspect for both LM-EnRML and SEnRML methods, to minimize the impact of spurious correlations in the parameter update step, and to add degrees of freedom to the inversion process. In this study, the adaptive and automatic correlation-based localization is implemented as local analysis (parameter by parameter) for both ensemble methods. Among the several localization methods available, in this work localization is implemented as in Luo and Bhakta (2020) with a modified Gaspari-Cohn function after Silva Neto et al. (2021). Also, the iterative local updating ensemble smoother (ILUES)

from Zhang et al. (2018) is applied to both the LM-EnRML and SEnRML methods to improve the localization process. Some implementation details are worth mentioning.
For the second numerical example, it is necessary to apply local updating of the parameters with the iterative local updating ensemble smoother (ILUES), as none of the methods are able to converge to an acceptable solution using their standard implementation with localization. The ILUES method, as its name suggests, is an iterative ensemble smoother that updates a subset of parameter realizations grouped based on a combination of parameter similarities and their goodness of fit to the observations. For each parameter realization, the method finds the best $N_l = \alpha N$ realizations that are most similar to it and that have the best fit to the observations. This is calculated using a normalized mismatch measure as follows:

$$\mathbf{J}_n = \mathbf{J_d}/{\mathbf{J_d}}^{max} + \mathbf{J_x}/{\mathbf{J_x}}^{max}, \tag{4.18}$$

where $\mathbf{J_d}$ and $\mathbf{J_x}$ are the data and model mismatch functions, respectively, and ${\mathbf{J_d}}^{max}$ and ${\mathbf{J_x}}^{max}$ are the maximum values. Once the realization subset is selected, the problem is now defined as a local problem, where the subset of realizations is used to update the parameter realization. As the parameter update of each realization subset will be more than one parameter realization, a selection mechanism is required to choose one of the updated parameter realizations from the parameter ensemble subset. Zhang et al. (2018) proposed to use a random selection; for simplicity, in this work, the first realization of the subset is chosen. It is important to note that the parameter update method is independent of ILUES, although Zhang et al. (2018) used the ensemble smoother multiple data assimilation (ES-MDA) method.
In this work, the ILUES method is used to improve the localization process of the LM-EnRML method for the second numerical example. In this method, the ILUES implementation is straightforward, and no further explanation is required. However, incorporating ILUES into the SEnRML method is not easy, and requires reformulation of the method if it is to be used. Given the nature of SEnRML, and since ILUES generates a new ensemble subset for each parameter realization in each iteration, there will be numerical discontinuity in the weight matrix $\mathbf{W}$, unless a separate weight matrix is calculated for each parameter realization. This is cumbersome and computationally expensive, which is the opposite of the purpose of SEnRML. Possibly due to the lack of a clear mathematical rational for its implementation, the literature is absent in respect to the application of ILUES to the SEnRML method, and the author is not aware of any such implementation. Research on this topic is left for future work.
The third method tested is data space inversion (DSI), proposed by Sun and Durlofsky (2017); Sun et al. (2017), that performs history matching of model outputs to data in the data space, using a number N of model output realizations, and a statistical linear

correlation model. The result of the history matching process is a set of model outputs that are consistent with the data, and not a set of model parameters. Using an initial ensemble of model output realizations, derived from a prior ensemble of model parameters, a linear correlation model can be built, as follows:

$$\mathbf{o} = \mathbf{o}_f + \mathbf{\Phi z \Sigma Y}, \tag{4.19}$$

where $\mathbf{o}$ are model outputs simulated by the DSI model, $\mathbf{o}_f$ is the mean model output vector and $\mathbf{z}$ is a vector of standard normal random deviates. $\mathbf{Y}$ is the ensemble of model output anomalies normalized by its standard deviation. The matrix $\mathbf{\Phi}$ is derived by truncated singular value decomposition (SVD) of the ensemble of standarized model output anomalies $\mathbf{Y}$, as

$$\mathbf{Y} = \mathbf{U\Sigma V}^T = \mathbf{\Phi V}^T, \tag{4.20}$$

where $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$ are the left singular vectors, singular values, and right singular vectors of the ensemble of model output anomalies, respectively. In this work, truncation of the $\mathbf{\Phi}$ matrix is based on the energy criterion, as proposed by Sun et al. (2017). An energy threshold of 99% was used to determine the number of modes to be retained in the truncated SVD decomposition (this is the same threshold value used for truncated SVD in ensemble methods). The number of truncated singular values defines the dimension of the DSI model parameter vector, $\mathbf{z}$.

Sun et al. (2017) pointed out that direct use of Equation 4.19 may lead to unphysical predictions of model outputs which comprise time series. To overcome this problem, they proposed histogram transformation of $\mathbf{o}$ prior to construction of the DSI model. DSI model predictions are then back-transformed before use. This was done in implementation of DSI that is described below. More sophisticated transformations such as those described by Sun and Durlofsky (2017) and by Jiang et al. (2021) can also be employed.

4.2.3 Metrics

Groundwater modelling for decision support requires that modelling metrics be applied to predictions of management interest (nevertheless, in the first numerical experiment that is discussed below the prediction of interest is the permeability field itself). Ideally, model-quantified predictive uncertainties should span the true values of predictions of management interest. At the same time, these intervals should be as narrow as available information allows (Doherty and Simmons, 2013).

Predictive uncertainty is dependent on the model-to-measurement fit to the extent that the prediction is conditioned by data. This is quantified by the data mismatch or objective function. Given that, for all numerical examples presented in this chapter, observation realizations are generated by adding measurement noise, the same noise that was added

to the true model outputs, the expected value of the objective function is approximately equal to the number of observations. A metric of model-to-measurement misfit is the difference between the expected and obtained objective function values. While attainment of a good fit with field measurements should not be considered as a modelling end in itself, failure to fit field data reveals data assimilation shortcomings.
Another statistic that is of considerable interest is the number of model runs that is required to achieve an acceptable level of model-to-measurement fit. While this metric matters little for the test examples (due to their simplicity), it is of far greater importance in real-world decision-support modelling contexts where model run times are considerably longer. Because model run scalability is integral to all the methods discussed herein, their model run efficiency is likely to be transferrable to contexts where parameter numbers are considerably larger than for the present examples, and where model run times are considerably longer.

4.3 Numerical Example 1: 2D-Aquifer Hydraulic Conductivity Model

4.3.1 Model description

In the first numerical example, a 2D aquifer hydraulic conductivity field is history-matched using observations of hydraulic conductivity at 25 locations in the model grid. A 50×50 model grid of dimension 1.0×1.0 represents the spatial distribution of hydraulic conductivity of a 2D aquifer domain, generated by a two-level hierarchical model. The first hierarchical parameterization level is defined by a spatially varying anisotropy angle α. This is the only hyperparameter considered spatially-variant. The second hierarchical level is defined by the hydraulic conductivity field itself. Both parameterization levels are generated using the proposed methodology, as described in the previous section.
A true hydraulic conductivity field is first generated. At the first parameterization level, the hyperparameter vector $\boldsymbol{\alpha}_{true}$ represents the true spatial variability of α over the model grid, and is calculated from the sum of a mean value of 0.79 radians (or 45°) and the convolution of 100 independent standard normal deviates $\mathbf{z}_{\alpha-true}$ that populate a 10×10 grid with an isotropic Gaussian kernel $f_{\alpha-true}$. The kernel $f_{\alpha-true}$ has a standard deviation of 0.52 radians (or 30°), and a correlation range a_α of 0.5 (or 25 grid cells). At the second parameterization level, the model parameter vector $\mathbf{x}_{true}$ is generated by the sum of a mean value of 0.0 (in log10 scale) and the convolution of a set of 100 independent standard normal deviates $\mathbf{z}_{\mathbf{x}-\mathbf{true}}$ that populate a 50×50 grid, with a Gaussian kernel $f_{\mathbf{x}-\mathbf{true}}$ with standard deviation of 2.0 and a correlation range $a_{\mathbf{x}-true}$ of 0.5. When performing the convolution, the effective distance r'_j for each model grid location $\mathbf{y}_j$ is

calculated using Equation 4.6. The anisotropy factor is set to 5.0, and the anisotropy angle changes spatially from the results of the first hierarchical level. Figure 4.2 shows the true hydraulic conductivity field generated using this methodology with the specified settings. It can be seen that as a result of the spatial variability of the anisotropy angle, the hydraulic conductivity field is nonstationary, changing its spatial correlation properties (direction and range) as a function of location in the model grid. This field is treated as the true field for the history matching process. It is noted that more complex models can be generated by including more hyperparameters in the hierarchical model, such as the variance of the hydraulic conductivity field, or the anisotropy factor.

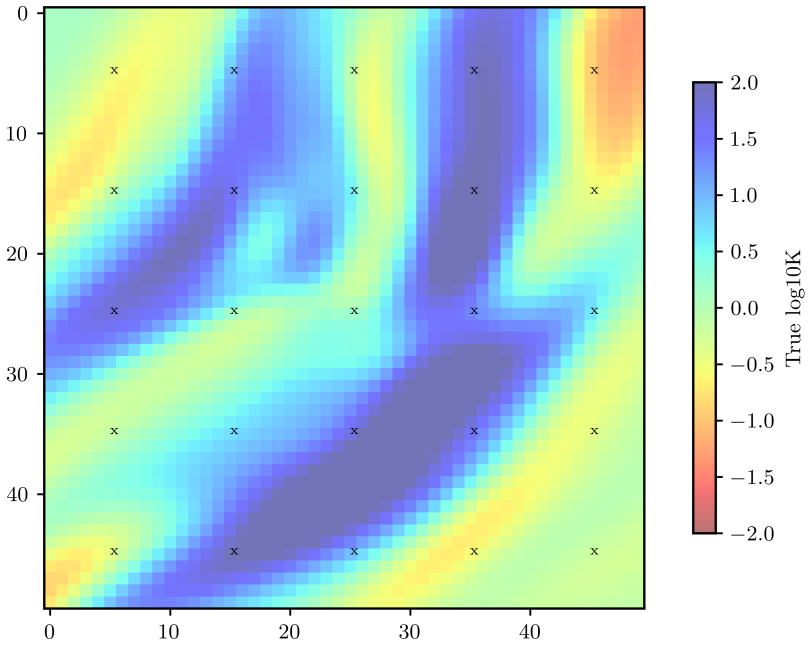

Figure 4.2: Stochastic field generated by the moving average method using spatially varying hyperparameters. Observation locations of hydraulic conductivity measurements are shown as black crosses.

History matching of this parameter field is performed using an observation dataset comprised of measurements of log10 of hydraulic conductivity at 25 evenly-spaced locations in the model grid, separated by 10 grid cells. Measurement noise of 0.05 is added to each synthetic measurement extracted from the true field. The forward model is then the hierarchical model described above. The model output is the simulated hydraulic conductivity field which is compared with observed data at the observation locations. This is a straightforward means of testing the methodology, without the intervention of a groundwater flow model. However, the forward model is still highly nonlinear, as it involves two levels of convolution of standard normal deviates and Gaussian kernel functions.

The number and statistical properties for the parameters that are part of the history

matching process are presented in Table 4.1. Three cases were defined, as shown in the table. Case 1 is the simplest, where the mean values of the anisotropy factor, anisotropy angle, and correlation range of the hydraulic conductivity field are fixed to the true values. Only the standard deviates $\mathbf{z_x}$ and $\mathbf{z}_\alpha$ are adjusted during the history matching process. The latter engenders the nonstationarity of the hydraulic conductivity field. In Case 2, all mean values of the hyperparameters are adjusted during the history matching process, but their priors are centred around the true values. In Case 3, the priors are centred around values that are different from the true values, increasing the complexity of the history matching process. It is a more realistic representation of the fact that any prior is likely to be wrong, but hopefully covers the true value.
Figure 4.3 shows the histograms of the prior hyperparameters for Case 2 and Case 3. The true values are also shown in the figure.

Parameter	Description	n	Mean / Scale		
			Case 1	Case 2	Case 3
$\bar{\eta}$	anisotropy factor	1	5.00 0.00	5.00 1.00	1.00 2.50
$\bar{\alpha}$	anisotropy angle (radians)	1	0.79 0.00	0.79 0.52	0.00 0.52
$\bar{a}$	log - correlation range of x	1	-0.69 0.00	-0.69 0.30	-0.60 0.30
z_α	z of α	100	0.00 1.00	0.00 1.00	0.00 1.00
z_x	z of x	400	0.00 1.00	0.00 1.00	0.00 1.00

Table 4.1: Parameters adjusted during the history matching process for three cases. The probability distribution of the anisotropy factor in Case 3 is half-normal. All hyperparameters (first three rows) represent mean values.

The history matching process was performed using the three methods described above, starting from a prior ensemble of 100 realizations. The LM-EnRML and SEnRML methods were implemented using python codes developed by the author, the same as used in the examples of Chapter 2 of this thesis. The DSI method was also implemented using a python code developed by the author, based on the original algorithm provided by Sun and Durlofsky (2017). MCMC for posterior analysis of the DSI model was implemented using the Python package pyDREAM (Shockley et al., 2017).
Some specification settings for the history matching process are required. For the LM-EnRML method, the initial Levenberg-Marquardt λ was calculated as (following Chen and Oliver (2013)):

$$\lambda = 10^{\text{Floor}(\log_{10}(\overline{J}/2m))}, \tag{4.21}$$

where $\overline{J}$ is the mean data mismatch, and m is the number of observations. For this particular problem, the resulted initial lambda is 100. The λ factor is reduced by a factor of 4.0 at each iteration if the data mismatch mean and standard deviation improve with respect to the previous iteration. If only the mean data mismatch improves between iterations, the λ factor is not changed. If the data mismatch mean does not improve,

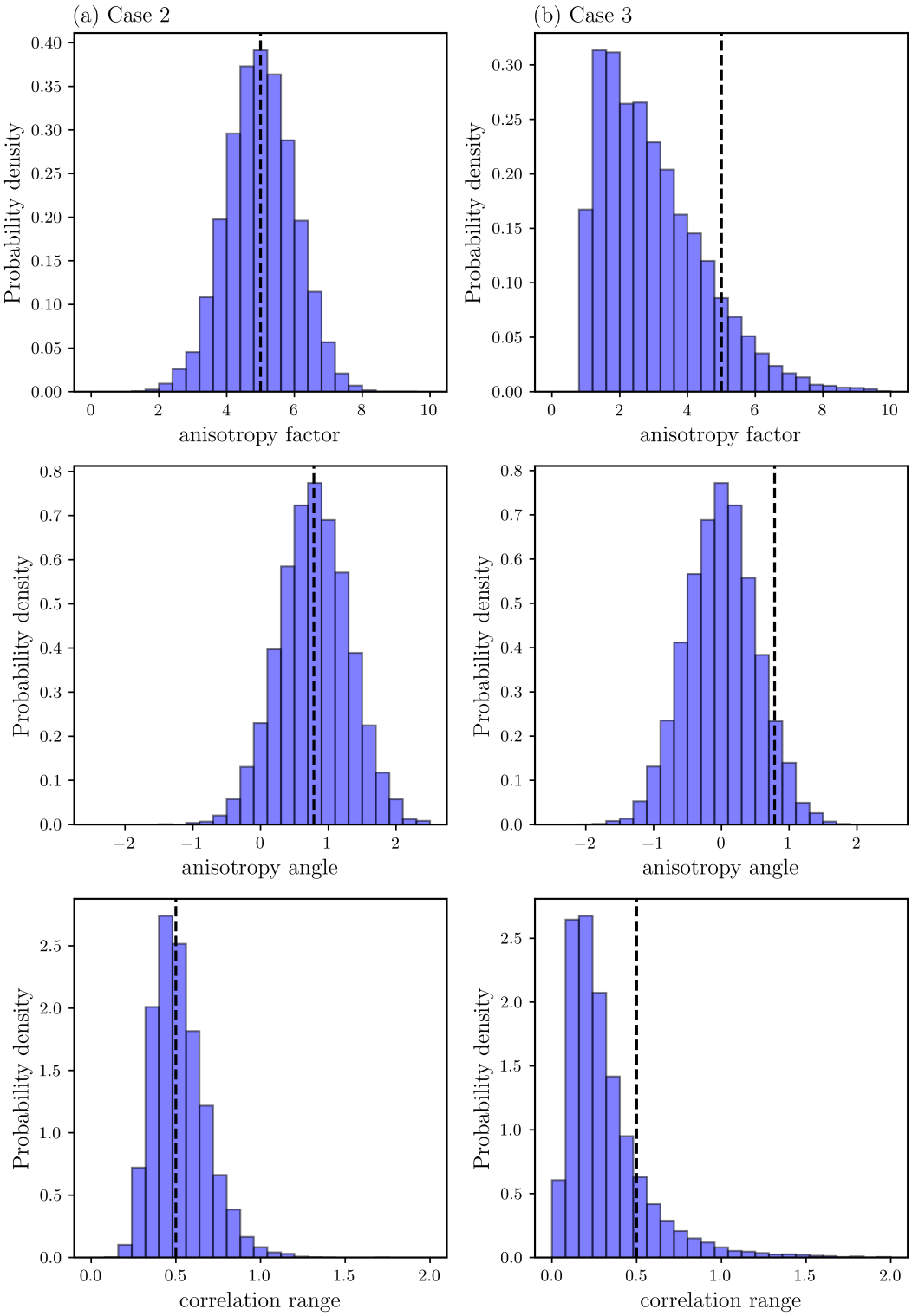

Figure 4.3: Hyperparameter priors for (a) Case 2 and (b) Case 3. The dashed vertical lines represent the true values of the hyperparameters.

the λ factor is increased by a factor of 5.0, with a maximum value of 5×10^{-4}. For the SEnRML method, the initial step length was set to 0.6, with a minimum and maximum values of 0.01 and 0.6. If the data mismatch mean and standard deviation improve respect to the previous iteration, the step length is increased according to the following equation:

$$\gamma = \gamma + (\gamma_{max} - \gamma) * 2^{-l/(\delta-1)}, \tag{4.22}$$

where γ_{max} is the maximum step length, l is the iteration number, and δ is a decay parameter. As for the EnRML method, the γ damping factor is unchanged if only the mean data mismatch improves between iterations. If the data mismatch mean does not improve, the step length is decreased by a factor of 2.0. The subset of ensemble members that improve the data mismatch and standard deviation with respect to the previous iteration are accepted, even if the data mismatch mean does not improve, as implemented in PESTPP-LM-EnRML (White, 2018). Iterations are stopped if they reach a total of 20 maximum iterations, or if the relative improvement of the data mismatch mean is less than 1x10^{-3}, or if the damping factors overcome their maximum or minimum values (depending on the method). Only for the EnRML, limits on the hyperparameter values were imposed to eliminate divergence issues encountered during the inversion process. The limits were set to 1.0 and 10.0 for the anisotropy factor, -2.5 and 2.5 for the anisotropy angle, and -2.0 and 1.0 for the log of the correlation range. This was not necessary for SEnRML; no limits on the hyperparameter values were imposed for this method.

4.3.2 Results

A summary of model mismatch and the number of model runs required to achieve an acceptable level of model-to-measurement fit is presented in Table 4.2. It is important to note that the expected value of the data mismatch is 25. However, it is not uncommon for ensemble methods to fail to achieve this level of data fit. Case 1 resulted in the best convergence behaviour, and the LM-EnRML method achieved the best data mismatch mean and standard deviation. As the problems become more complex (Case 2 and Case 3), the data mismatch mean and standard deviation increase. This is particularly true for the LM-EnRML method, appearing to struggle more with the nonstationarity of the hydraulic conductivity field. For Case 3, the LM-EnRML method converged to a data mismatch mean value that is over 400 times the expected value, compared to the SEnRML method that achieved a data mismatch mean value that is over 100 times the expected value. Although the SEnRML also shows a degradation in performance as the complexity of the problem increases, it is less pronounced than for the LM-EnRML method. That being said, it is acknowledged that these methods could be further improved by tuning some of their optimization parameters.

Similar observations can be made about the evolution of data mismatch during the history

Table 4.2: Data mismatch mean, standard deviation, and number of iterations of ensemble methods, for the 3 cases analysed.

Method	Mean / Std / N model runs		
	Case 1	Case 2	Case 3
SEnRML	1611 2865 1600	2731 880 1600	3294 1167 1400
LM-EnRML	678 185 2000	8779 2865 700	10776 5137 700

matching process, as shown in Figure 4.4. Overall, all methods show an improvement in the data mismatch mean, especially for the LM-EnRML method in Case 1. It can be observed that, when methods show a good convergence behaviour, the number of iterations required to achieve a stabilization of the data mismatch mean is more than 7. The exception is the LM-EnRML method in Case 2 and Case 3, which achieved a stabilization of the data mismatch mean in less than approximately 3 iterations, but with a high data mismatch mean value.

The best realizations of the history-matched hydraulic conductivity fields for each case and for each method are shown in Figure 4.5. The identification of the best realization for each method was not based on the data mismatch, but on the squared difference between the true field and the estimated field. Case 1 is shown in the first row, being the simplest case where only the standard deviates are adjusted, whereas the mean values of the hyperparameters are fixed at true values. It can be observed that for this case both methods, SEnRML and LM-EnRML, are able to reasonably reproduce the true field, which is expected given the simplicity of the case, but also an important verification as the history matching process involves the estimation of standard deviates at two hierarchical levels, which might be a challenging task. Visually comparing the estimated fields resultant from Case 2 and Case 3, it is apparent that difficulties arise impeding the methods from reproducing the true field. This is also an expected result as Case 2 includes uncertainty around the true mean values of the hyperparameters, and Case 3 adds more complexity by centring the hyperparameter priors around values that are different from the true values. Consistent with the data mismatch results, the LM-EnRML method struggled to reproduce the true field in Case 3.

The hyperparameter posterior distributions for LM-EnRML and SEnRML are shown in Figure 4.6 and Figure 4.7, respectively. For the LM-EnRML method, the posterior distributions of the mean hyperparameters do cover the true values, but their variances are large for some hyperparameters, and skewed for others. It appears that the posterior distribution of the correlation range is the result of incipient parameter ensemble collapse, as the posterior distribution is skewed to the left, at the imposed lower bound, for both cases. In contrast, for Case 2, the posterior distribution of the anisotropy angle shows a wider variance compared to the prior distribution. Case 3 has a generally better performance in hyperparameter estimation for method LM-EnRML compared to

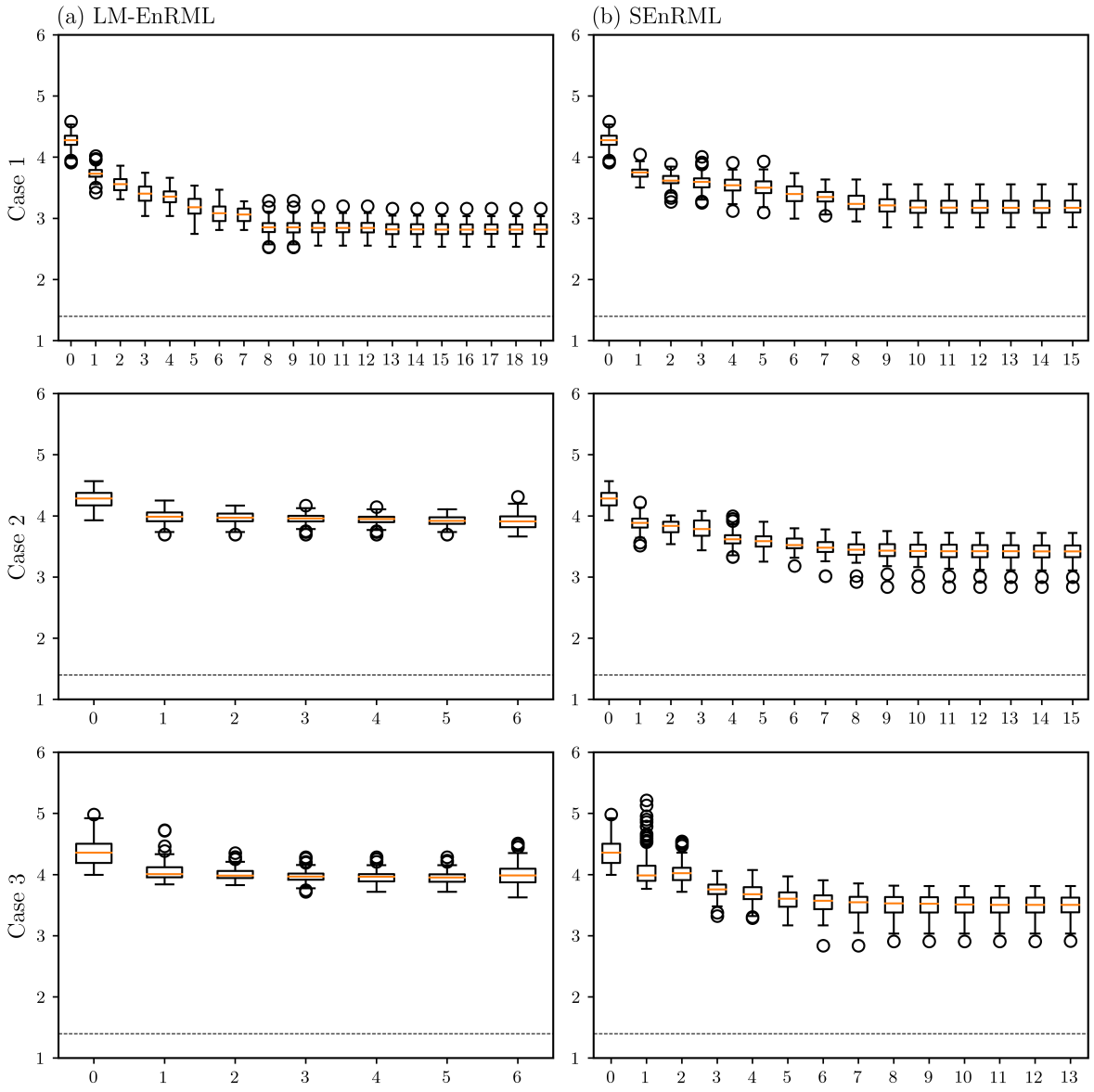

Figure 4.4: Evolution of data mismatch (log10) vs number of iterations during the history matching process for the 2D aquifer model using the (a) SEnRML and (b) LM-EnRML methods for the three cases defined. The boxes are built using the 25th and 75th percentiles, and the whiskers represent the 5th and 95th percentiles. The horizontal line inside the box represents the median. The black circles represent the outliers. The dashed horizontal line represents the target data mismatch of 25.0 (number of observations). Iteration 0 represents the initial data mismatch.

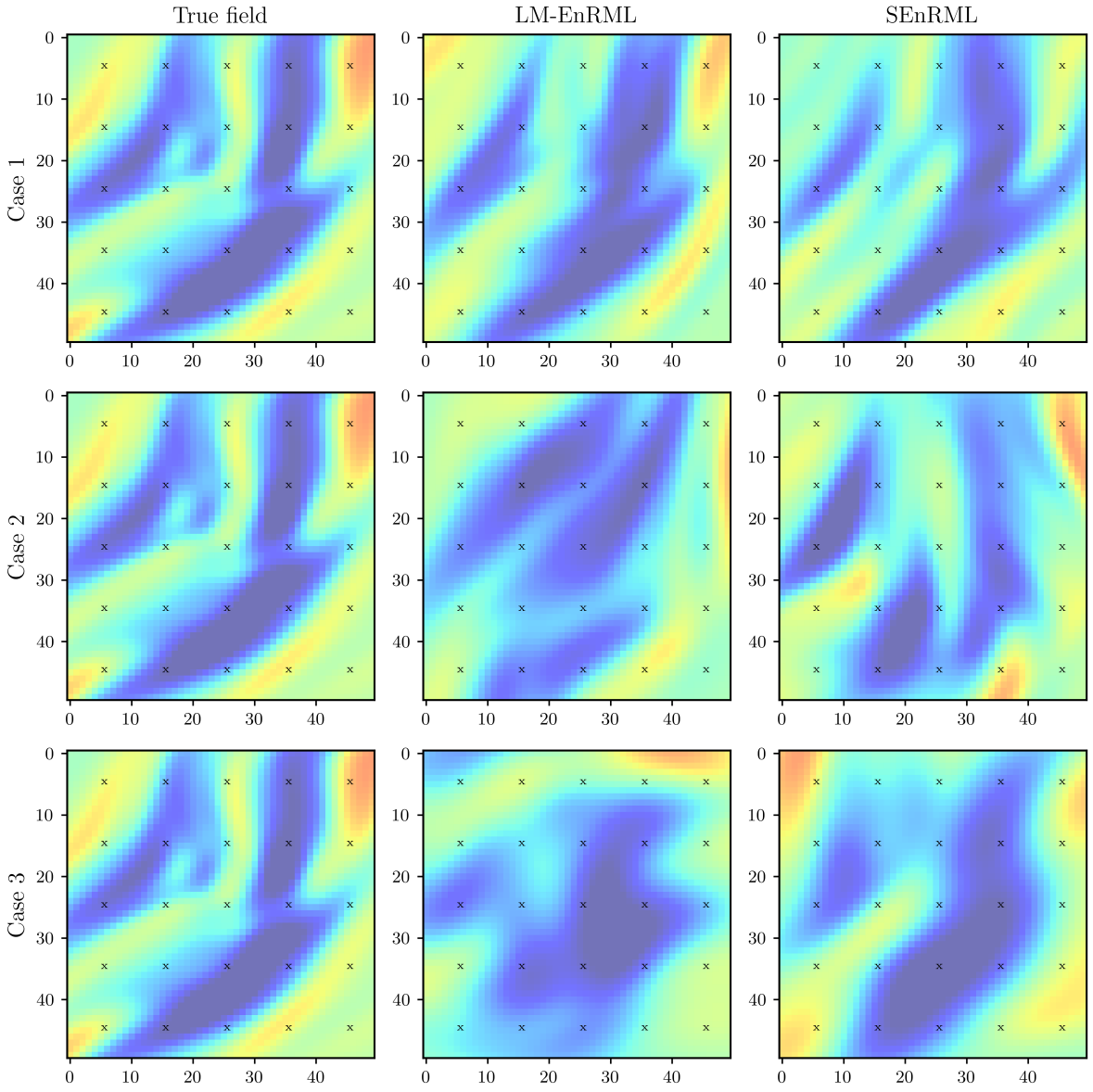

Figure 4.5: History-matched hydraulic conductivity fields for the 2D aquifer model using the SEnRML and LM-EnRML methods, for the three cases defined. The true field is shown in the first column for comparison.

Case 2. However, the posteriors do not significantly change from the priors, which is an indication that the method is adjusting the values of standard deviates to fit the data, but not the hyperparameters. The SEnRML method shows a better performance in the hyperparameter estimation for both cases, with the posterior distributions of the mean hyperparameters covering the true values (Figure 4.7), except for the anisotropy factor in Case 3. In general, the posteriors are in this case narrower than the priors, which shows that the method is adjusting the hyperparameters to fit the data, and not only the standard deviates. Case 2 shows an acceptable performance, although the posterior modes are shifted away from the true values for the anisotropy factor and the anisotropy angle, even though their priors were correctly defined. In contrast, for Case 3, the anisotropy angle exhibits a posterior mode that is close to the true value, and the correlation range has a posterior distribution that is shifted towards the true value, compared to the prior distribution.

It is noted that the number of standard deviates $\mathbf{z}$ for the hydraulic conductivity field was set to 400 (20×20), which is less than the number used in the true field generation. This was explicitly done to verify the effect on inferring the true hydraulic conductivity field with a smaller number of parameters. It was demonstrated that this approach worked, at least for the simplest case, and it is not clear if using fewer parameters is the reason why the performance of the ensemble methods was degraded for cases 2 and 3. Therefore, an additional test was performed with Case 3 using 100,200, and 900 standard deviates $\mathbf{z}$ for the hydraulic conductivity field, combined with a prior ensemble of size 100, 200, and 300. Figure 4.8 shows maps of the mean data mismatch normalized to the mean values obtained for Case 3 (Table 4.2) for the SEnRML and LM-EnRML methods for the different number of standard deviates and ensemble sizes. It can be observed that both methods improve their performance under different configurations, with a minimum relative data mismatch mean of 0.18 for the SEnRML method and 0.07 for the LM-EnRML method. The SEnRML method reaches its best performance with 100 standard deviates and 300 ensemble members, whereas the LM-EnRML method showed the best relative reduction in data mismatch mean with 900 standard deviates and 100 ensemble members. These are interesting and important results, as they suggest that some configurations may work better for some methods and not for others.

Figure 4.9 shows the posterior distributions of the mean values of the hyperparameters for Case 3, for the cases with the best combination of number of standard deviates and ensemble size for each method ((900,200) for LM-EnRML and (100,300) for SEnRML), following the results of Figure 4.8. It can be observed that the posterior distributions of the mean values of the hyperparameters are more acceptable in general, compared to the previous cases (Figure 4.6, Figure 4.7). In particular, the true values are within the sampled posterior distributions, except for the anisotropy factor in the LM-EnRML method, and the correlation range for both methods.

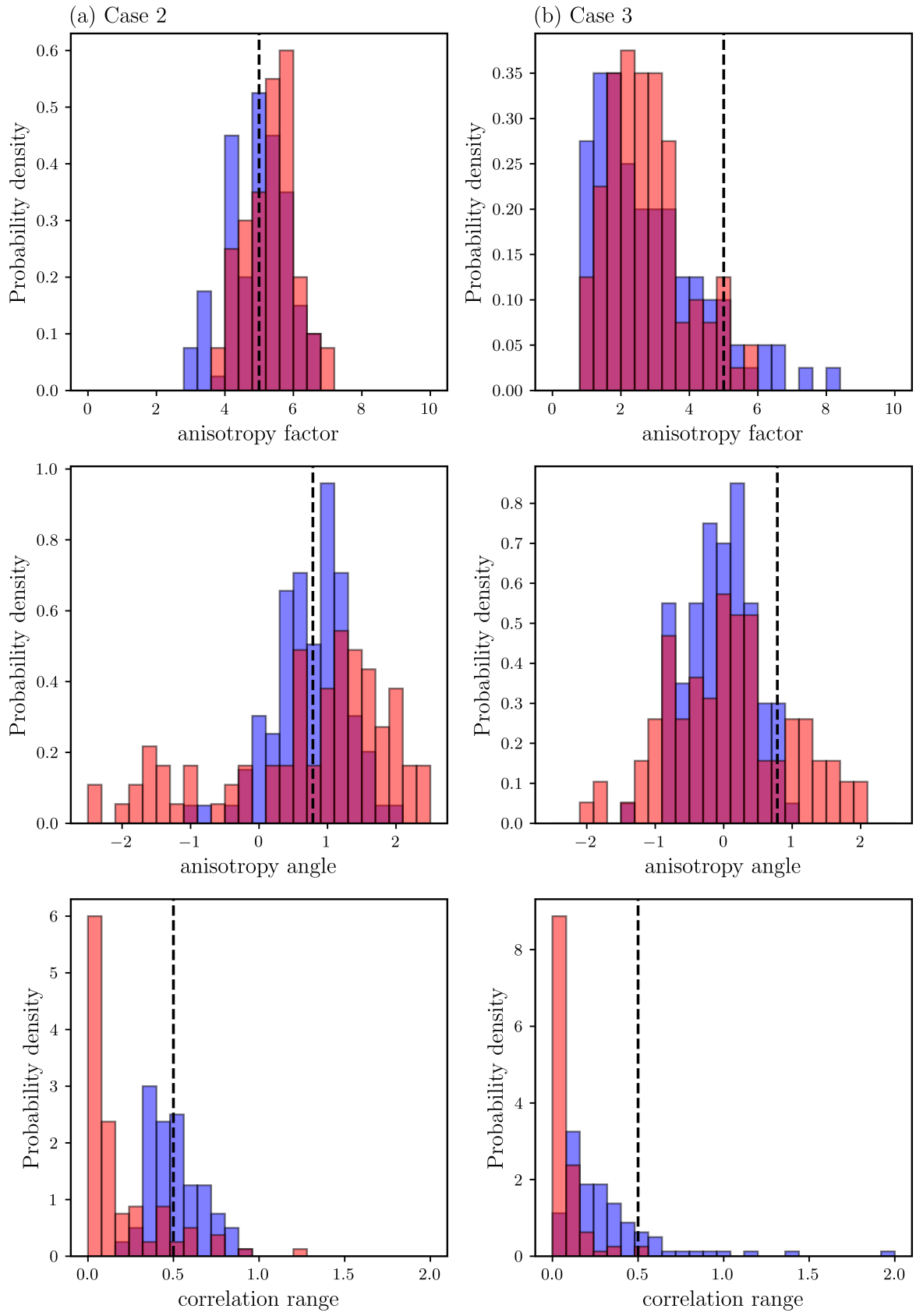

Figure 4.6: Comparison of prior (blue) and posterior (red) distributions of the mean values of the hyperparameters for the 2D aquifer model using the LM-EnRML method, for (a) Case 2 and (b) Case 3.

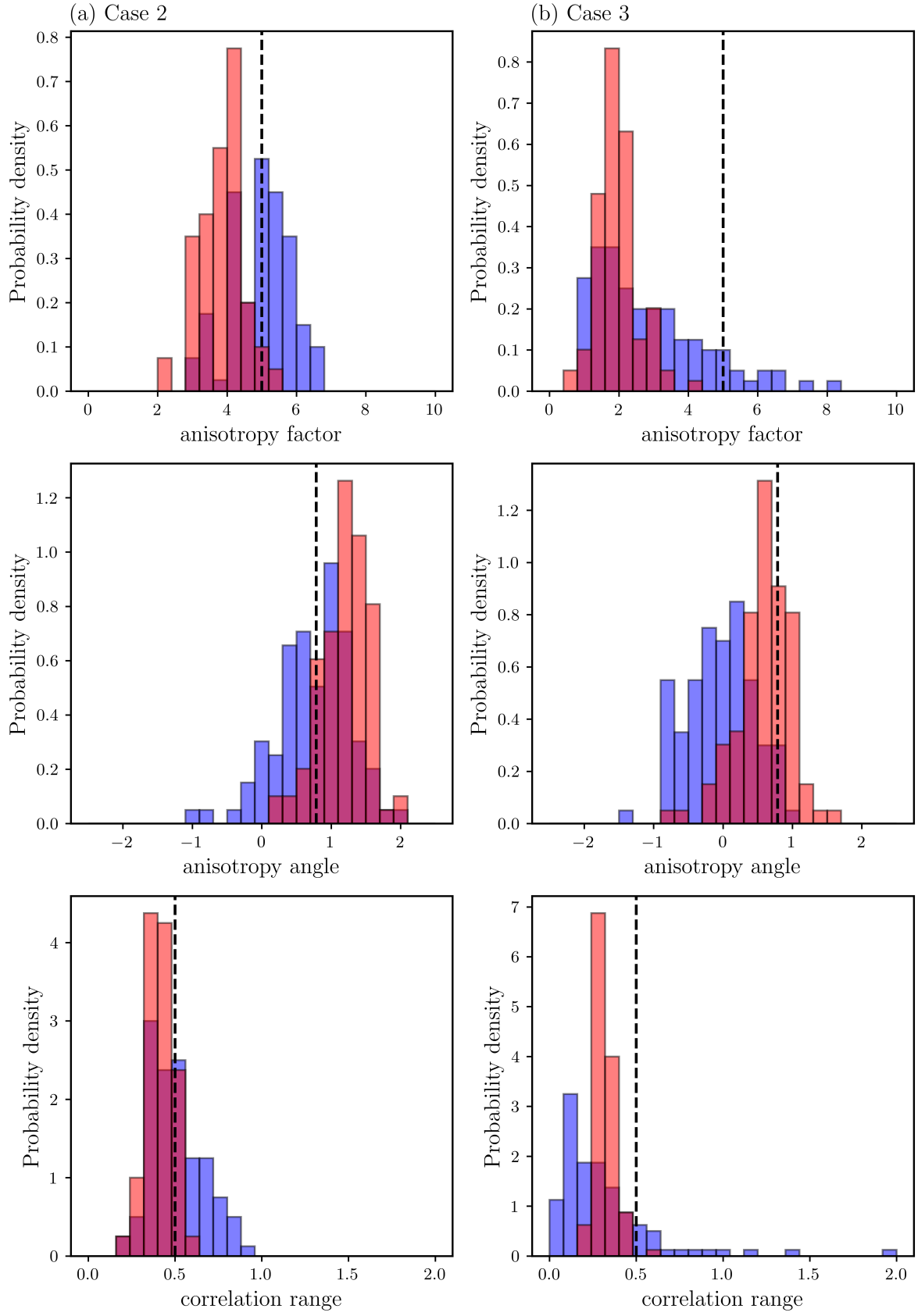

Figure 4.7: Comparison of prior (blue) and posterior (red) distributions of the mean values of the hyperparameters for the 2D aquifer model using the SEnRML method, for (a) Case 2 and (b) Case 3.

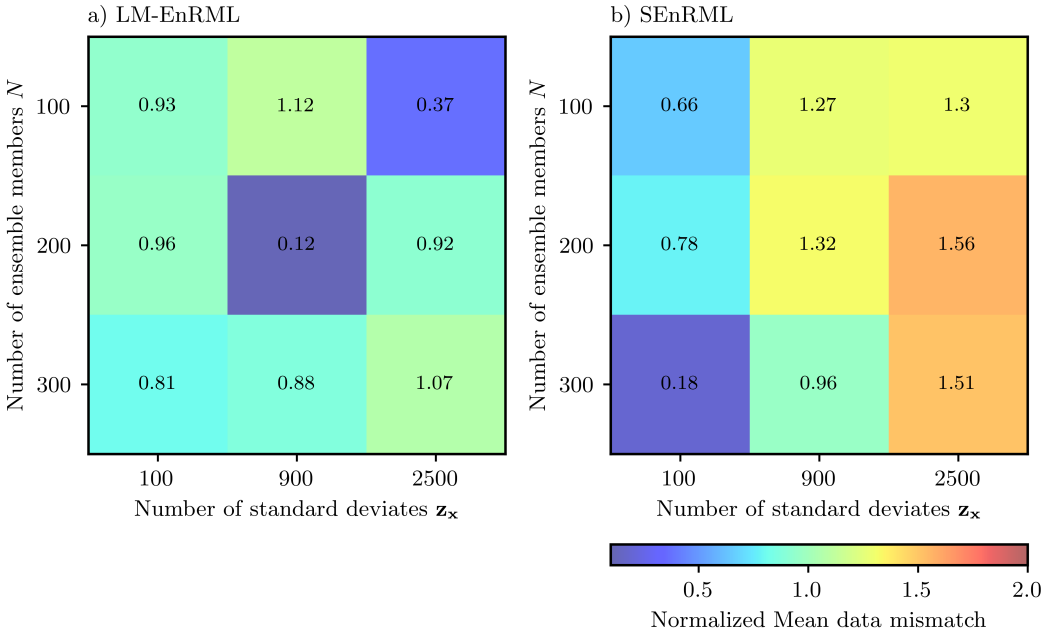

Figure 4.8: Normalized best mean data mismatch (relative to Case 3 results) for various parameter and ensemble sizes: (a) SEnRML method, (b) LM-EnRML method.

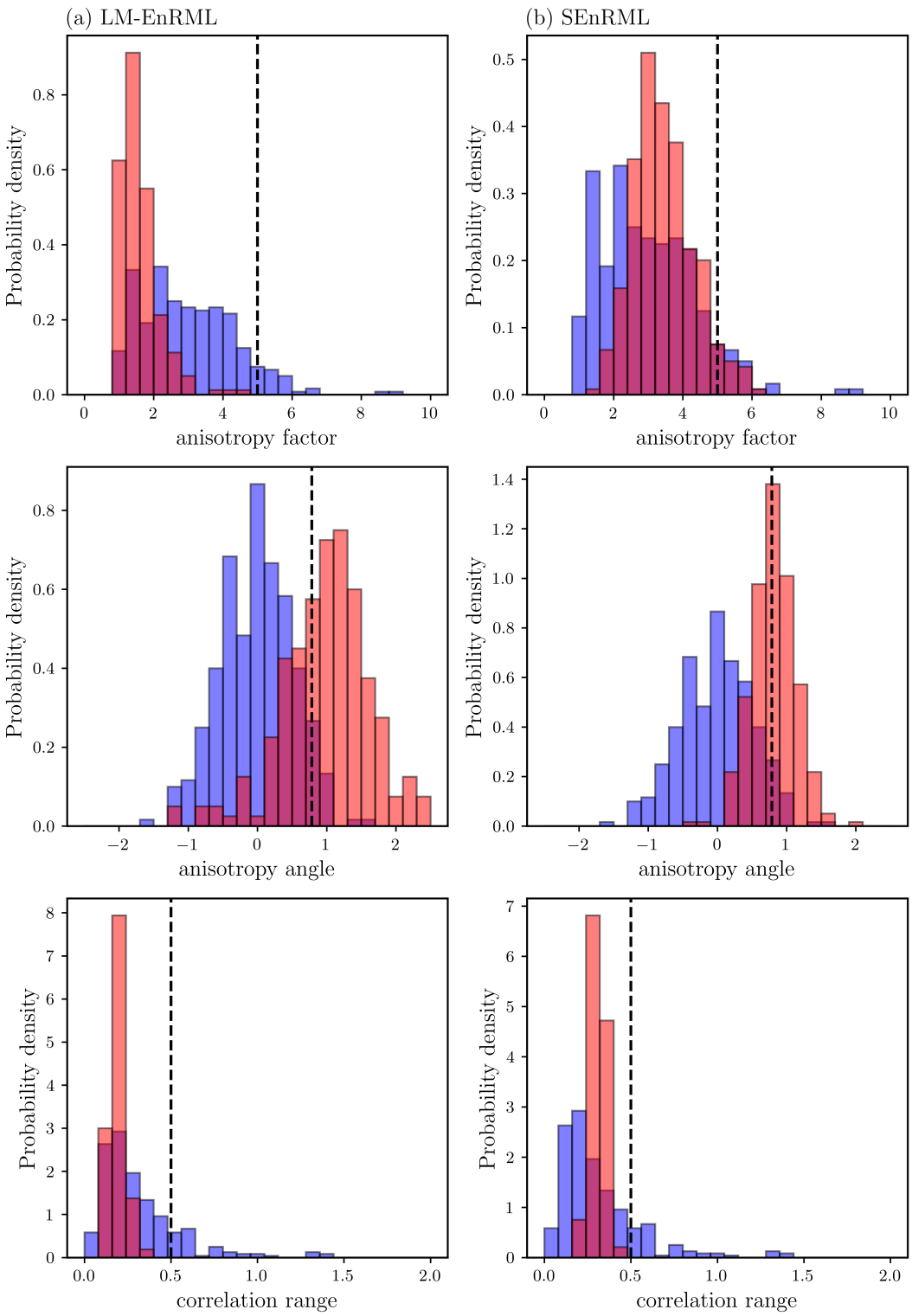

Figure 4.9: Prior (blue) and posterior (red) distributions of the mean values of the hyperparameters for Case 3, for the best combination of number of standard deviates and ensemble size for (a) LM-EnRML and (b) SEnRML methods.

Finally, a DSI model was implemented using a prior ensemble of model outputs that resulted from the prior ensemble of model parameters. In this case, given that the observations and the forward model are hydraulic conductivity values, the DSI model is a linear correlation model that mimics the hydraulic conductivity field. Therefore, the DSI model outputs are a set of hydraulic conductivity fields that are consistent with the data. The DSI model was history-matched to the observation dataset, using DREAM as the MCMC algorithm provided by the pyDREAM package (Shockley et al., 2017). The chain generation was set to 30000, with a 50% burn-in period.

The three cases previously presented were tested. The data mismatch mean, and standard deviation are presented in Table 4.3. It can be observed from the table that the DSI method achieved a data mismatch mean value that is close to the expected value of 25 for all cases, which is significantly better than the ensemble methods. The standard deviation of the data mismatch is also lower than the ensemble methods. The DSI method is also computationally cheaper, as it only requires 100 forward model runs, plus the 30000 DSI model runs. Running of the DSI model has a minimal computational cost as it is the implementation of a linear matrix on vector multiplication.

The advantages of the DSI method described above come with a cost, as there is no assurance that DSI model outputs are physically meaningful, as suggested by visual inspection of selected estimated hydraulic conductivity fields presented in Figure 4.10. As shown in the figure, the DSI method is able to reproduce certain aspects of the true field, but it lacks the spatial continuity that is present in the true field. Moreover, the estimated fields appear to be more noisy and heterogeneous than the true field, in particular for Case 2 and Case 3, where uncertainty on the mean values of the hyperparameters was introduced. This is an inevitable consequence of the simplifications made in the DSI model based on a limited number of forward model runs.

Table 4.3: Data mismatch mean, standard deviation, and number of iterations of the DSI method, resulting from history matching of the 2D-aquifer hydraulic conductivity model.

Method	Mean / Std / N model runs		
	Case 1	Case 2	Case 3
DSI	31 9 100	30 9 39	10 0 100

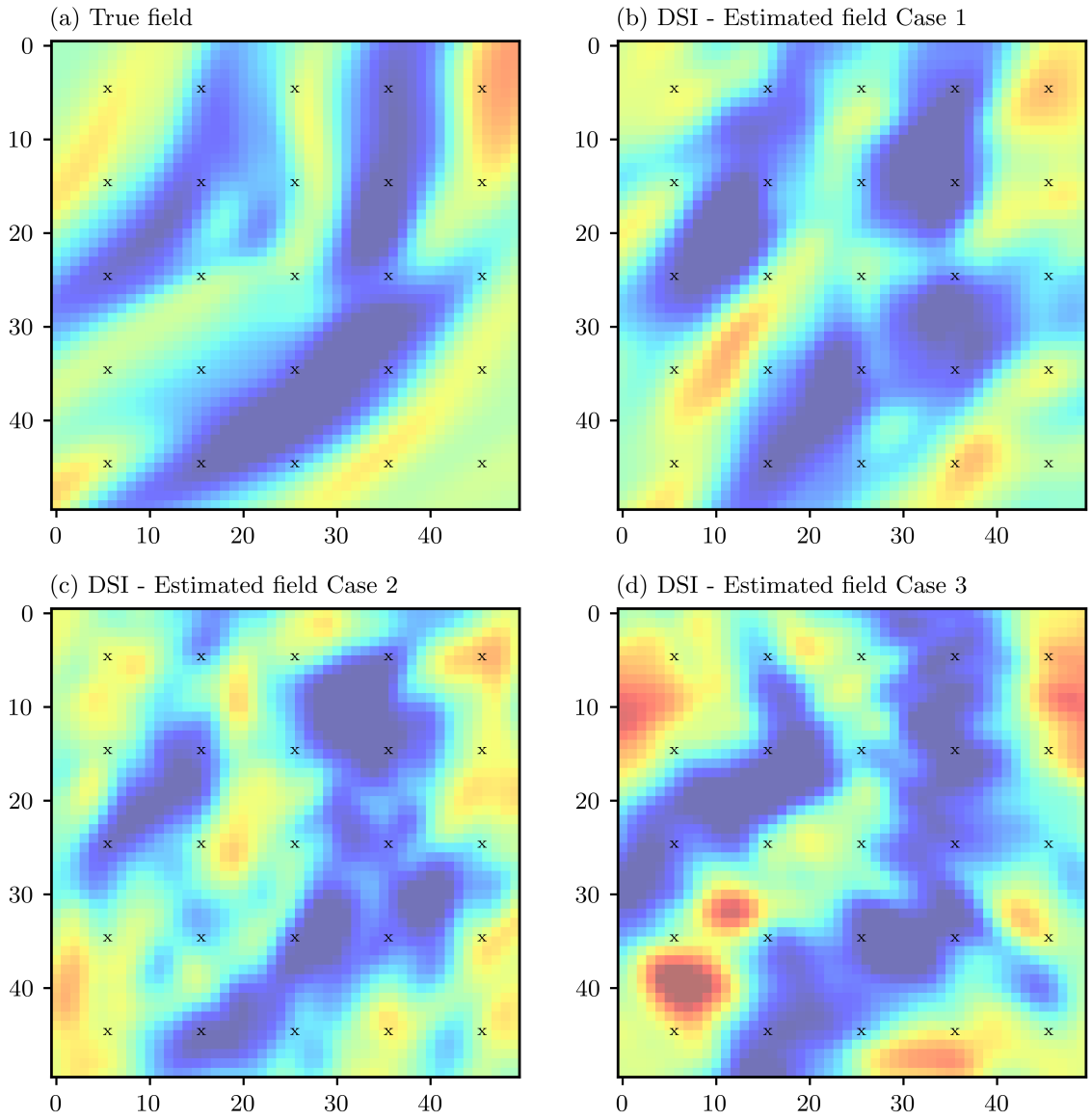

Figure 4.10: History-matched hydraulic conductivity fields for the 2D aquifer model using the DSI method, for the three cases defined. The true field is shown in the upper left corner for comparison.

4.4 Numerical Example 2: Flow and Transport 2D Model

4.4.1 Model description

The second numerical example expands on the first one by including a groundwater flow and transport model applied to the 2D aquifer case, simulating extraction and reinjection of a solute mass in the system. The model is inspired by Direct Lithium Extraction (DLE) from brines, where lithium-rich brine is pumped and reinjected once the mineral has been extracted in a processing plant. Pumping and reinjecting is also a common practice in the oil and gas industry, where water is injected in the reservoir to increase the pressure and enhance the production of hydrocarbons.

The distinctive challenge of this example is that the prediction of interest is the future depletion of lithium in the aquifer, which is a function of the extraction and reinjection rates, and the heterogeneity of the aquifer. Given known extraction and reinjection rates, and the depletion of the reinjected brine (assumed 1.0, i.e., complete), flow and transport modelling is used to quantify predictive uncertainty of solute depletion, which is conditioned by measurements of brine concentration in the aquifer.

The model domain, as in the first example, is a 50×50 2D grid of unitary dimension. The problem is worked in dimensionless units. The aquifer as described in the previous section, has nonstationary geostatistical properties for the distribution of hydraulic conductivity. The model simulates extraction and reinjection of brine for two simulation stress periods. The first stress period has a duration of 200 time units, while the second stress period has a duration of 400 time units. In the first of these periods, brine is extracted from 7 wells and injected into 2 wells at a total rate of 5×10^{-4} volume units per time. In the second period, the number of pumping wells increases to 10 while maintaining the same total pumping and injection rate, in order to represent well replacement due to solute depletion. Confined steady state flow and transient solute transport are simulated using MODFLOW 6 (Langevin et al., 2017). The effective porosity is 0.25, while longitudinal and transverse dispersivity are 5×10^{-3} and 5×10^{-4} length units, respectively. Figure 4.11 shows the (a) model configuration and contours the distribution of the depleted brine plume at the end of the historic period, and (b) the time series of solute depletion at pumping well locations. These results were obtained after running the forward model with the true hydraulic conductivity field, and adding Gaussian noise of 0.01 standard deviation to the model outputs. As inferred from the figure, the additional pumping wells introduced in the second stress period—pw8, pw9, and pw10—are located outside the depleted brine plume area observed at the end of the first stress period. It is assumed that no information is available at these locations during the history matching process.

A synthetic measurement is taken every 10 time units during the first simulation period

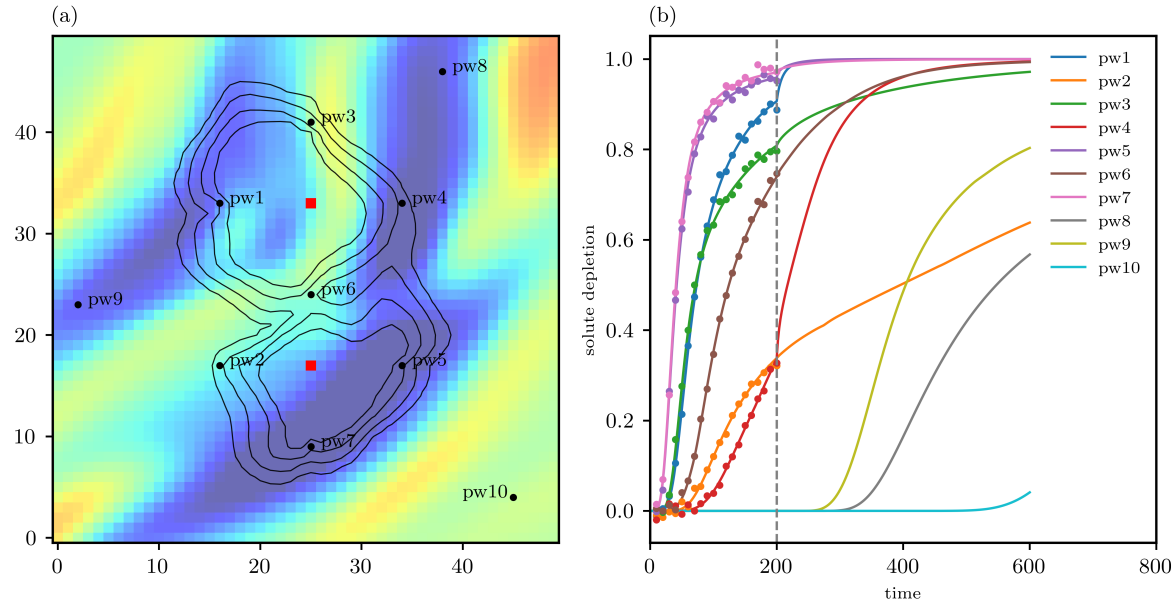

Figure 4.11: (a) Pumping well (back circles) and injection wells (red squares) locations, and contours of the depleted brine plume at the end of the historic period; (b) time series of brine depletion at pumping well locations. The dashed vertical line separates the historic and predictive periods.

at 7 pumping wells. Therefore, the history matching dataset is comprised of 140 measurements of solute depletion. A random realization of measurement noise with a standard deviation of 0.01 concentration units is added to each measurement. Note that, during the first stress period, depletions of 0.0 are measured at the sites of all pumping wells that are active during only the second stress period. The second simulation stress period is denoted as the "predictive period". During this period, predictions of depletion made by history-matched models (see below) can be compared with true depletions. The latter are calculated using the true hydraulic conductivity field, presented in the previous section. Model parameters that are adjusted are described in Table 4.1. Their priors correspond to Case 3 of the previous numerical experiment.

The history matching process was performed using LM-EnRML and DSI (an attempt to history-match the model using the SEnRML method was made, but the method did not converge to an acceptable level of data mismatch). Local updating (Zhang et al., 2018) and localization was required to get convergence for LM-EnRML, as previously explained. This is due to the strong nonlinearity of the problem, mainly derived from the nonstationary implementation of the hydraulic conductivity field. For DSI, prior model outputs were derived from a prior ensemble of model parameters of size 600, twice the size of the ensemble used for LM-EnRML. This was necessary to improve the predictive performance of DSI for a marginal increase in computational cost.

4.4.2 Results

The data mismatch evolution during the LM-EnRML history matching process is shown in Figure 4.12. As can be observed, the LM-EnRML method struggled to converge during at least the first 7 iterations, reaching full convergence after 20 iterations. Additional model runs would be required if lambda testing was implemented, as in the case of PESTPP-LM-EnRML (White, 2018). This could improve the convergence behaviour of the method, but it would also increase the computational cost. The observed slow convergence is the result of the strong nonlinearity of the problem. After 23 iterations, the data mismatch mean and standard deviation are 658, and 347, respectively, which are reasonable values given that the expected value is 140, the number of observations (note that this is just a reference number, as it does not apply to nonlinear problems).

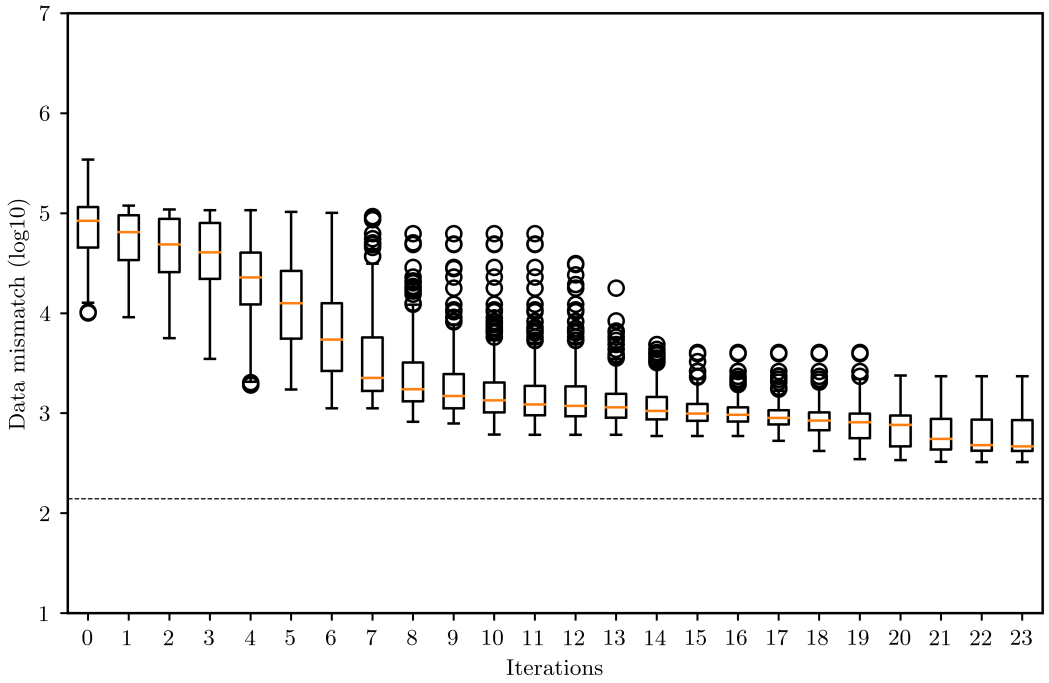

Figure 4.12: Evolution of the data mismatch during the history matching process for the 2D flow and transport model using the LM-EnRML method. The horizontal line represents the target data mismatch of 140.0 (number of observations). Iteration 0 represents the initial data mismatch. The boxes, lines, circles, and whiskers have the same meaning as in Figure 4.4.

With DSI, a very similar fit to the data was achieved after running MCMC with 12 chains of 50000 iterations each. As previously mentioned, the DSI method is computationally cheaper than the ensemble methods, as the inference of the posterior distribution of the DSI parameters is based on a linear correlation model. Once the correlation model is built, MCMC runs in a matter of minutes, compared to the LM-EnRML method that requires hours to converge. Comparing the number of model runs of this numerical example, LM-EnRML requires 6900 to achieve convergence (recall that during each iteration, 300

model runs are required). In contrast, DSI requires 600 model runs and the minimal cost of running MCMC with a linear model. That said, the DSI method is not free of challenges, as the result of the posterior inference is not a physical model, but a set of DSI parameters and DSI model outputs that are consistent with the data.

The observed and simulated depletion time series for the history-match dataset are shown in Figure 4.13. As shown in the figure, the fit is acceptable for both methods. It is expected that the DSI-MCMC method would achieve a better fit to the data, as it is based on a linear correlation model that is generally able to reproduce the data with a high level of accuracy. It is also expected that non-physical solutions were obtained with the DSI-MCMC method, with depletions reaching values above 1.0, for some wells. Looking at the predictive period, the uncertainty estimated by both methods is very narrow, and in most cases the true values are within the predictive uncertainty bounds. The exception is the depletion at well pw4, which is a well located inside the depleted plume and next to one of the reinjection wells. This well shows narrow uncertainty ranges that do not cover the true values. This suggests that in this highly nonlinear case, data assimilation has the potential to over-constrain, rather than under-constrain the uncertainties of predictions that bear a close relationship to observations, even where model-to-measurement misfit exceeds that which would be expected from measurement noise.

The simulated hydrographs of the wells that were not part of the history matching dataset are shown in Figure 4.14. As shown in the figure, both methods resulted in dissimilar predictive uncertainty bounds. The LM-EnRML method resulted in predictive bounds that, although narrow, are able to capture the true values of the depletion. The DSI method, on the other hand, resulted in wider predictive bounds, which are not fully able to capture the true values. This is especially true for well pw10, which is located outside the depleted plume area, and therefore the uncertainty is higher.

Finally, the posterior distributions of the mean values of the hyperparameters for the LM-EnRML method are shown in Figure 4.15. Parameter ensemble collapse is evident, and the posterior distributions do not cover the true values of the hyperparameters. This is a consequence of the strong nonlinearity of the problem, and the nature of the LM-EnRML method.

Although of secondary importance, the LM-EnRML history-matched hydraulic conductivity fields for the 2D aquifer model (Figure 4.16) do not reflect, to a reasonable extent, the nonstationarity nature of the true field. However, the estimated fields are able to reproduce some isolated zones of high and low hydraulic conductivity, which are consistent with the true field, and also the true anisotropy angle of the system. This shows that history matching using LM-EnRML can lead to a reasonable fit to the data, but not necessarily to a representative estimation of the true hydraulic parameter field.

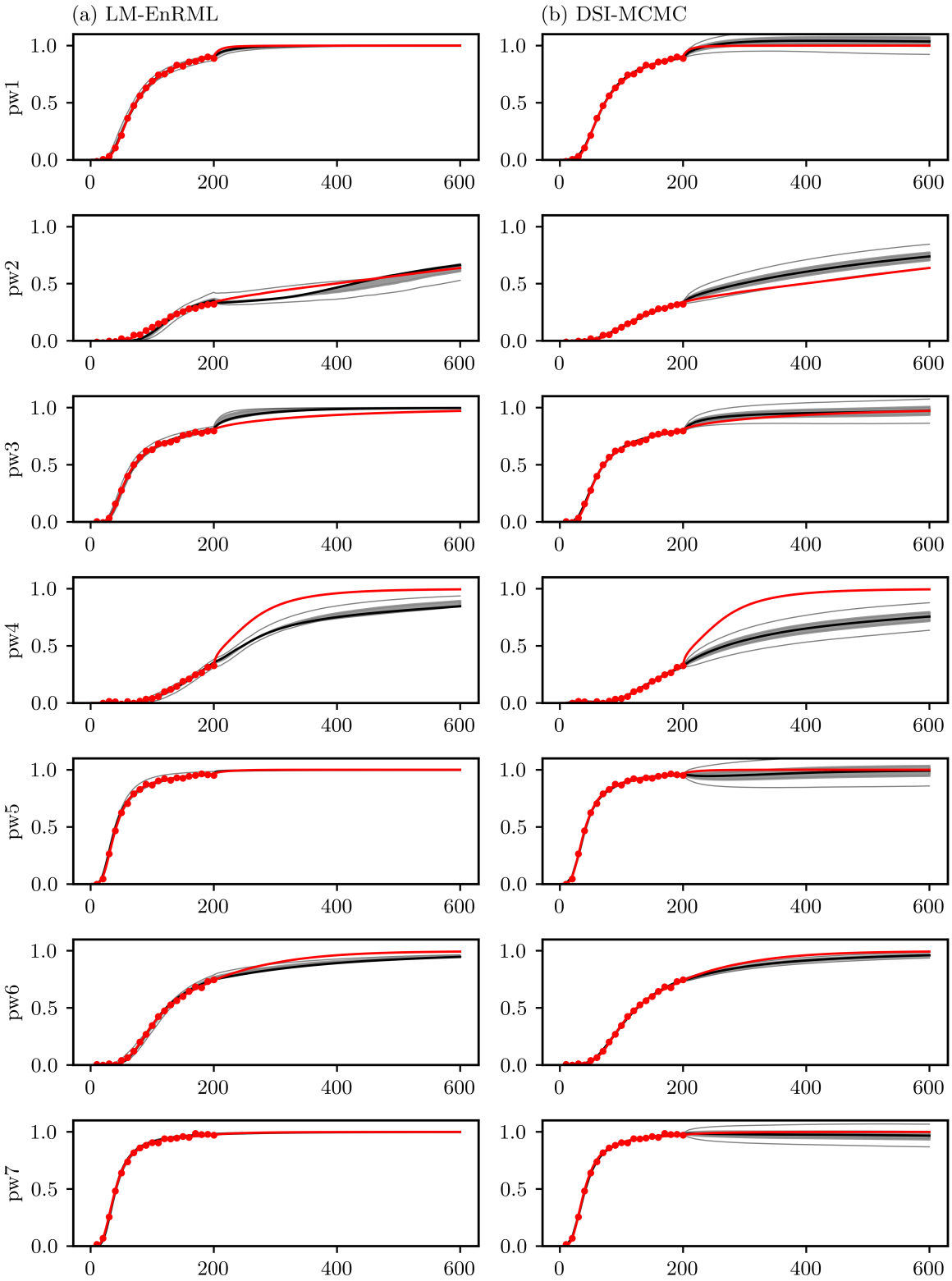

Figure 4.13: Measured and model-calculated depletion time series for wells part of the history matching dataset, For (a) LM-EnRML and (b) DSI-MCMC. The red lines with solid circles represent the true values, the solid black line is the median, the grey-shaded area is the P25-P75 percentile region, and the external black lines are the P5 and P95 percentiles of the simulated depletions.

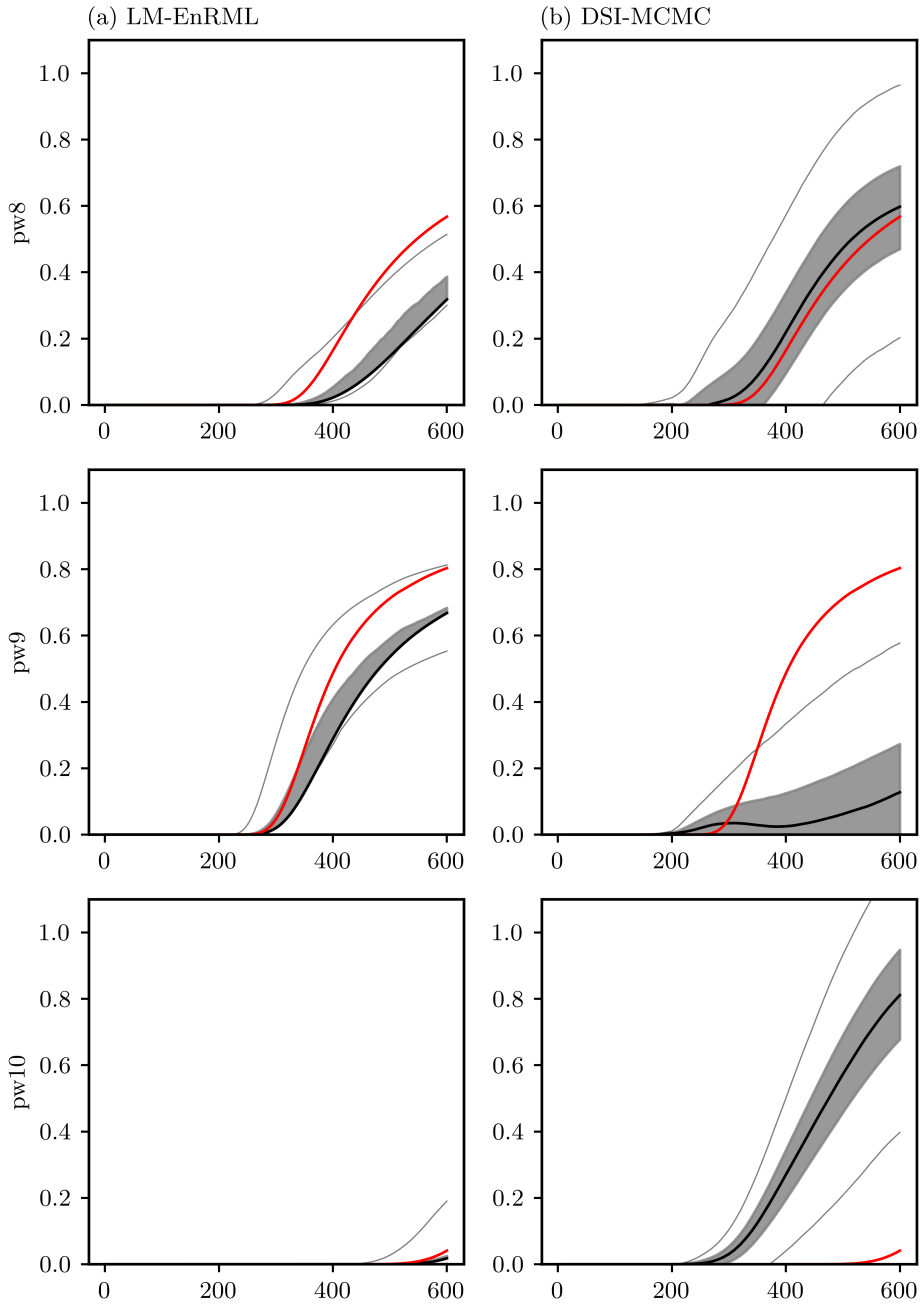

Figure 4.14: Model-calculated depletion time series for wells pw8, pw9, and pw10, for (a) LM-EnRML and (b) DSI-MCMC. The red lines with solid circles represent the true values, the solid black line is the median, the grey-shaded area is the P25-P75 percentile region, and the external black lines are the P5 and P95 percentiles of the simulated depletions.

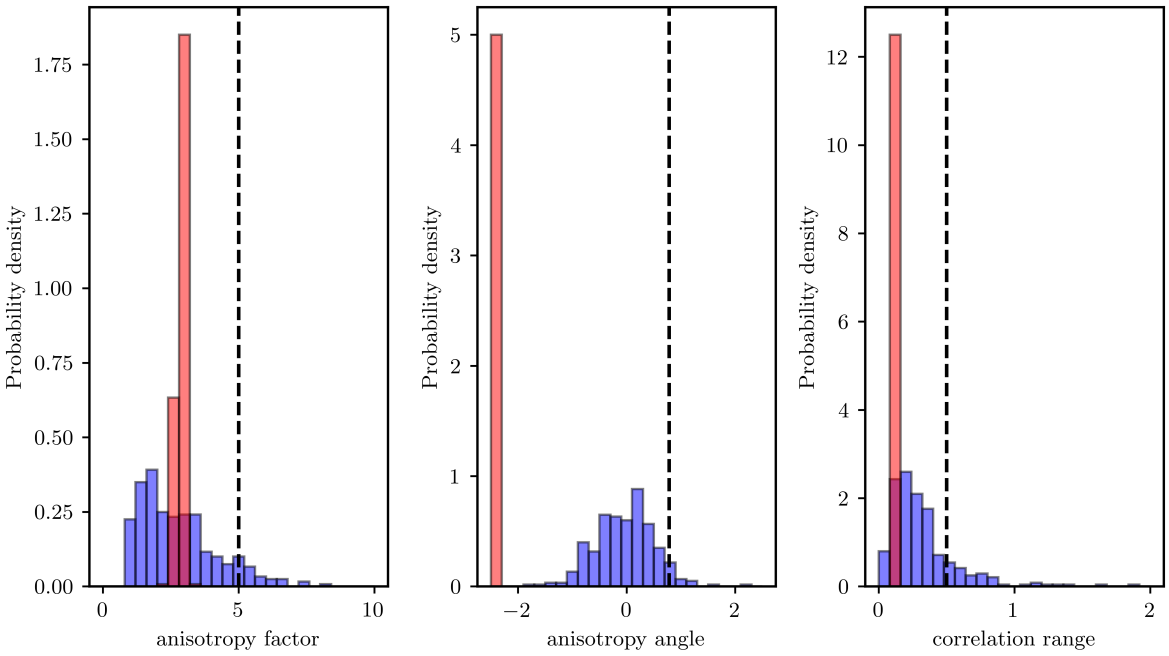

Figure 4.15: Prior (blue) and posterior (red) distributions of the mean values of the hyperparameters for the 2D flow and transport model using the LM-EnRML method.

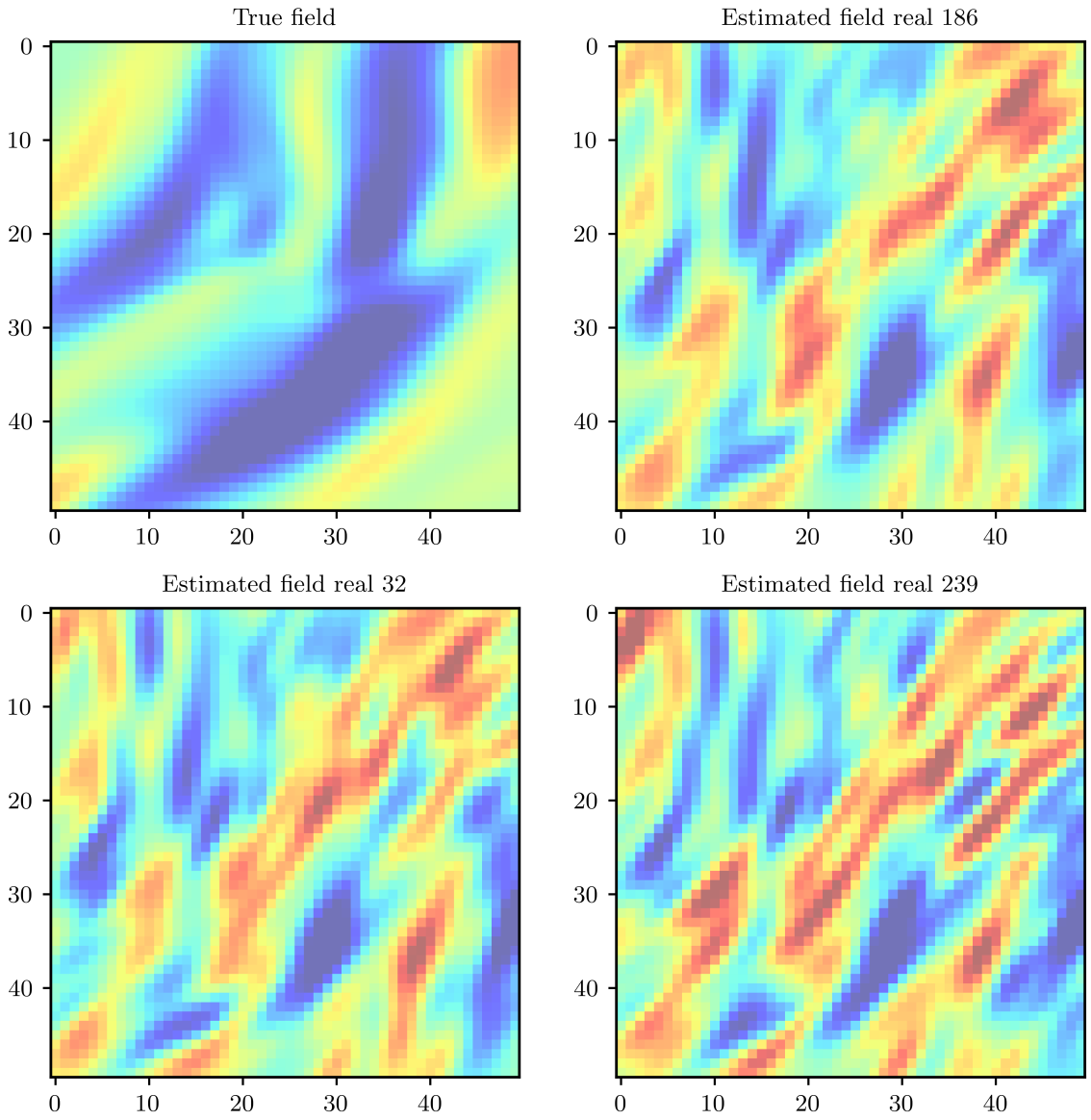

Figure 4.16: Selected realizations of history-matched hydraulic conductivity fields for the 2D aquifer model using the LM-EnRML method. The true field is shown in the upper left corner for comparison.

4.5 Discussion

The proposed methodology adopts a hierarchical two-level parameterization scheme to generate nonstationary fields. It was tested in two numerical examples using three different history matching methods (SEnRML, LM-EnRML, and DSI), demonstrating that it can handle uncertain and nonstationary priors, history-match nonlinear problems, and result in reasonable estimates of predictive uncertainty. The first example presented was focused on estimating a 2D hydraulic conductivity field from a set of hydraulic conductivity measurements, to evaluate the methodology in a case without requiring a groundwater model. The second example involved history matching and predicting solute depletion in a 2D aquifer model with nonstationary anisotropy angles, incorporating a groundwater flow and transport model. This is a more complex case as it requires the estimation of both parameters and hyperparameters from a set of solute depletion measurements. The nonlinearities in this case originate not only from the groundwater model but also from the process of generating the nonstationary fields. History matching convergence was achieved in both examples for most of the tested methods, except for SEnRML in the second example. In this latter case, iterative local updating and localization were required for LM-EnRML to converge. These techniques are more challenging to implement in SEnRML than in LM-EnRML, which limits the applicability of SEnRML for highly nonlinear problems like the second example.

The results from the first example showed reasonable history-matched nonstationary hydraulic conductivity fields, similar to the true field, for both the SEnRML and LM-EnRML methods when mean hyperparameter values were fixed at their true values. As more uncertainty was added to the problem, the performance of the SEnRML and LM-EnRML methods was degraded, and their difficulties in reproducing the true field became more evident. This was especially true for the LM-EnRML method, which appeared to prioritize the adjustment of the standard deviates over hyperparameters. Similar findings were presented by Oliver (2022) in his study. In contrast, the SEnRML method showed a better performance in the estimation of the hyperparameters, even when the mean values of the hyperparameters were uncertain. In the case of assuming uncertain hyperparameters centred on values different from the true values (Case 3), the LM-EnRML method struggled to converge to an acceptable fit and reproduce the true field. This was not the case for the SEnRML method, which produced a similar hydraulic conductivity spatial distribution compared to the true field (based on the best realization), even though both methods achieved a similar misfit. It appears that the SEnRML method preserves the prior ensemble structure better than the LM-EnRML method, which is expected given the nature of the method. It was also shown that both methods work best under distinct and specific configurations of the number of standard deviates and ensemble size. LM-EnRML appeared to perform better with an increased number of

standard deviates and a medium-sized (200) ensemble, whereas SEnRML preformed best with a reduced number of standard deviates and a larger ensemble size (300). This result aligns with the expected behaviour of ensemble methods, where the size of the ensemble constrains the number of degrees of freedom (Emerick and Reynolds, 2012). In the case of LM-EnRML, the best configuration led to a better fit to the data, but not necessarily to a better estimation of posterior probability distribution of the hyperparameters. In contrast, the best configuration for SEnRML resulted in improved estimation of the posterior hyperparameters as well as a better fit to the data.

The second example demonstrated that the LM-EnRML method was able to converge to an acceptable fit but at the cost of numerous iterations and, consequently, many model runs. Much cheaper computational cost was incurred for the DSI method to achieve a similar fit, but without the ability to provide history-matched parameter fields. With this, it is not possible to assess the physical meaning of the DSI model outputs. Having said that, LM-EnRML resulted in hyperparameter ensemble collapse and therefore history-matched parameter fields that are not representative of the true field. Both methods showed underestimation of predictive uncertainty for the solute depletion predictions at pumping well pw4 that were part of the history matching dataset. In the case of LM-EnRML, this underestimation does not appear to be the result of overfitting, as the method did not converge to the expected data mismatch value. On the contrary, as the combined number of parameters and hyperparameters was greater than the number of observations, the results suggest that underestimation of predictive uncertainty was due to the nature of the LM-EnRML method and the strong nonlinearity of the problem. On the other hand, the observed limited ability of DSI to capture the true uncertainty of predictions, may be due to the incomplete sampling of the prior ensemble parameter space and consequent inability to explore the full range of possible model outputs. Interestingly enough, predictions of solute depletion in wells that were not part of the history matching dataset (pw8, pw9, and pw10) showed narrow but minimally biased predictions for LM-EnRML, and wider but more biased predictions for DSI. In the case of LM-EnRML, the propensity of a limited ensemble size to result in underestimation of predictive uncertainty arises from limitations in its ability to explore the many dimensions of parameter space that are uninformed by a history matching dataset (i.e. the calibration null space) but to which predictions of interest may be sensitive (Kitlasten et al., 2022). As for DSI, the linear model makes strong assumptions about the structure of the groundwater flow and transport model, therefore it is expected that the predictions can be biased (Hastie et al., 2009).

Several general discussion points can be made based on the results of this study. Oliver (2022) suggested the need of a hybrid method that combines an analytical representation of the hyperparameter sensitivities with the ensemble-based model parameter sensitivities, in order to successfully fit the data in a hierarchical problem. This study, however,

showed that the application of the proposed methodology with ensemble methods works reasonably well in history matching a nonlinear problem, without the need of an analytical representation of the hyperparameter sensitivities (although in the second example, iterative local updating and localization were required to achieve convergence for LM-EnRML). It is also apparent that history-match adjustment of parameter ensembles is challenged when relationships between model outputs and parameters which must undergo change (including stochastic hyperparameters) are highly nonlinear (Evensen et al., 2022). This is unsurprising when it is considered that most methodologies that implement ensemble-based history matching are based on theory that assumes inverse problem linearity; nonlinearity is then accommodated through iterative parameter adjustment. Nonlinearity arises from two sources for the second example presented in this study. One of these is the composition of the measurement dataset; this comprises entirely concentration measurements. The other is the use of adjustable, and spatially-variable, stochastic hyperparameters that determine patterns of emergent hydraulic property heterogeneity. In the inverse problem that was posed in this study, these must be adjusted simultaneously with other parameters that determine the locations and magnitude of this heterogeneity. This is a challenging task, and appears to degrade the ability of ensemble methods in general and LM-EnRML in particular to estimate realistic parameters (and hyperparameters). In some cases, the resulting history-matched fields may not resemble any trait of the true field at all. However, making sense of these parameter fields is one of the aims of modellers to confirm that the model is a good representation of the system, model parameters still have a physical meaning, and that the model is able to make reliable predictions. It has been shown by several authors (for example, Clark and Vrugt, 2006; Doherty and Christensen, 2011) that parameters may adopt values that compensate for model structural defects, in order to fit the data. Based on the results of this study, it appears that parameters may also play 'numerical compensatory roles' when using ensemble methods in highly nonlinear problems, due to the nature of the method and the strong nonlinearity of the problem, even if the model is a perfect representation of the system. If this holds true, for some highly nonlinear problems, it may warrant dispensing with ensemble methods for history matching parameters. Instead, data space inversion (DSI) methods could be considered, as they may achieve a comparable fit to the data at a lower computational cost, without the need of adjusting parameters. In the present study, DSI was found to be the cheapest, but not completely effective, method of predictive uncertainty exploration. A feature of this method (which is both a strength and a weakness of it) is that it does not associate a parameter field with a prediction. Pessimistic predictions are therefore difficult to explain. This deficiency of DSI must be seen in context however, as the LM-EnRML method does not necessarily generate parameter fields that are particularly illuminating to a modeller. One of these benefits appears to be a reasonable level of confidence that predictive uncertainty is not grossly under-estimated.

It is worth remarking that experience gained in using the DSI methodology suggests that confidence in its evaluation of predictive uncertainty increases with the number of model runs that are dedicated to construction of the DSI model. Because DSI is so numerically cheap, its use does not preclude the deployment of alternative history matching and uncertainty analysis methods. The same parameter fields that are used for construction of a DSI model can be used as starting points for both of the ensemble-based parameter adjustment methods that are described herein.

Several limitations of this study should be noted. Although the proposed methodology can be deployed with more complex, more highly-parameterised, and slower-running models than the flow and transport model described herein, it remains to be demonstrated in real-world problems. Also, in the examples presented, a Gaussian kernel was used due to its simple formulation. Other kernels could be used and tested. One example is the exponential kernel, which is commonly used in geostatistics, but it would involve the use of modified Bessel functions (Oliver, 1995) which are more computationally expensive. Finally, the spatial distribution of standard deviates was performed in 2D and in a simplistic manner consistent with the spatial nature of the examples. More complex modelling problems involving the combination of zone-based and grid-based parameterization in a 3D domain, may challenge the applicability of the proposed methodology, due to the cumbersome nature of the parameterization. This is a topic for future research.

4.6 Conclusions

This work demonstrates the success of a new methodology that accommodate uncertain and nonstationary priors in history matching and predictive uncertainty quantification of groundwater models. It does so, by formulating the inverse problem in a hierarchical manner including spatially variable geostatistical hyperparameters that govern the shapes of emergent hydraulic property heterogeneity, in addition to spatially distributed parameters that govern its location. The spatial variability of hyperparameters and model hydraulic parameters results from a two-levels spatial averaging of history-match-adjustable standard normal deviates using a Gaussian kernel, with geostatistical properties that can be also history matched.

It was demonstrated that the proposed methodology can be deployed in history matching using ensemble methods, achieve a reasonable fit to the data, and provide acceptable predictive uncertainty estimates, assuming nonstationary and uncertain priors as additional sources of uncertainty. However, it is acknowledged that this is a highly nonlinear Bayesian inverse problem. As a result, ensemble-based parameter adjustment can be numerically inefficient and posterior parameter and predictive uncertainties can be underestimated as an outcome of both nonlinearity and parameter realization insufficiency. In contrast, data space inversion (DSI) was able to provide useable depictions of predictive

uncertainty with a numerical cost that is far smaller than LM-EnRML. Furthermore, its model run burden is immune to structural and parameterization complexity. It achieves these benefits by dispensing with the need for model parameter adjustment during the history matching process. Unfortunately, a modeller is not, therefore, able to view parameter fields that result in pessimistic predictions. This study suggests that this may not be an option anyway.

Chapter 5

Quantifying Model Structural Errors in History Matching and Predictive Uncertainty Quantification in Groundwater Modelling

Author contributions

T. Opazo: Conceptualization 100%, Realization 100%, Writing 100%.

Manuscript in preparation for submission to Water Resources Research: Opazo, T. Quantifying Model Structural Errors in History Matching and Predictive Uncertainty Quantification in Groundwater Modelling.

Abstract

Structural errors, arising from imperfections in the model, can introduce bias and lead to an underestimation of uncertainty. Accounting for these errors is essential to improve the reliability of predictions derived from groundwater models. This study presents a novel methodology to incorporate model structural errors in history matching and predictive uncertainty quantification. The proposed approach utilizes a complex model to generate synthetic observations, which are then used to calibrate a simplified model, enabling the identification and quantification of the statistical properties of structural error. A data space inversion (DSI) technique is employed to develop a correlation model between measurements and calibration residuals, termed the DSI-RES model. This model is conditioned on actual observations to generate realizations of structural error, which are integrated into the history matching process using the Subspace Ensemble Randomized Maximum Likelihood (SEnRML) method. The methodology is tested on a two-

dimensional numerical model that simulates groundwater inflows to an open pit. Results demonstrate that explicitly incorporating structural errors in the history matching process reduces predictive bias and produces more conservative predictive uncertainty estimates. The study highlights the importance of accounting for model structural errors to enhance the reliability of groundwater models for decision support. By providing a practical framework for quantifying and integrating structural errors, the proposed methodology improves predictive uncertainty quantification and supports more informed decision-making in groundwater management.

5.1 Introduction

History matching and predictive uncertainty quantification are two essential components of the groundwater modelling workflow in support of decision-making (Doherty and Moore, 2020). Model parameters are first adjusted to fit observed data. This is done to ensure that the model is capable of reproducing the observed system behaviour and to reduce parameter and predictive uncertainty, using methods that are based on the Bayesian framework (Tarantola, 2005). Within this framework, history matching requires the definition of prior uncertainties for model parameters and the measurement error in the data. More often than not, the outcomes of history matching are not satisfactory to the modeller, as the differences between observed data and their corresponding simulated outputs are not completely commensurate with measurement error. There are two primary causes for such mismatches: the prior is not correct, or the model is missing some important features of the system. When model imperfections are the most likely cause of the mismatch, adding complexity to the model may improve the fit to the data. However, this solution may not be applied *ad. infinitum* due to limited resources (Mathews and Vial, 2017), or because the model may become too difficult to run (due to model instabilities, long runtimes, or both) which limits its ability to perform history matching and predictive uncertainty quantification (Doherty, 2011). Moreover, model imperfections might not be even be visible to the modeller. This is the rule rather than the exception in groundwater modelling, acknowledging that the model will never be a perfect representation of a natural system. Hence, for the reasons presented above, imperfect models are unavoidable.

Model structural error (Beven, 2005) is a broad term that includes all imperfections in the model that may lead to discrepancies between observed data and model outputs beyond what can be attributed to measurement error. It produces two major impacts on uncertainty quantification: bias and the underestimation of uncertainty. Bias is defined as the statistical difference between the expected value of the model outputs and the true values (Hastie et al., 2009). Given a dataset of observations, bias can be quantified by comparing the history-matched model outputs and the observed data. This bias is there-

fore visible in the model residuals. Predictive bias in particular, is the difference between the expected value of a model prediction and the true unknown value, and is invisible for obvious reasons. One of the causes of predictive bias is the compensatory roles that parameters may play to fit the data (Clark and Vrugt, 2006; Doherty and Christensen, 2011), which occurs when using an imperfect model. This results in parameter values that can be significantly different from their prior expected values and potentially sensitive to the prediction of interest. However, even if the parameter values do not change, their prior expected values may be far from the values that minimize predictive bias in the context of an imperfect model. This is the case when the model is unable to capture the parameter or process details that significantly influence the predictions of interest (Doherty and Christensen, 2011). In fact, Mathews and Vial (2017) showed how a prior should be modified, under Gaussian assumptions, to minimize bias of a prediction of interest when using an imperfect model. Posterior predictive uncertainty, i.e., the variability of predictive outcomes that can be expected from the model after conditioning it to observed data, is also affected by model structural error. First, parameter uncertainties might be artificially reduced as a result of history matching an underparameterized model (an example of a structural model defect) to a relatively large dataset, potentially leading to the underestimation of predictive uncertainty. This occurs when a prediction of interest is sensitive to the parameters with underestimated uncertainties. Additionally, the lack of parameters and processes in a model that are important to observations may inhibit the history matching process from extracting the full information content of the data, potentially leading to an overestimation of predictive uncertainty (Doherty and Christensen, 2011). Finally, if there is predictive bias in the model, the posterior predictive uncertainty will be shifted away from the true probability distribution. This could limit the use of predictive uncertainty estimates for quantifying probabilities of occurrence of unwanted events.

There is a growing (but limited) body of literature in this area, some of which will be discussed here. Most of the studies focus on the identification of predictive bias, the statistical representation of model error, and its update during the history matching process. One of the strategies to quantify predictive bias is the use of paired simple and complex models, where the parameterization of the complex model can be informed by expert knowledge, and the simple model adopts a more parsimonious parameterization scheme. Several authors have used this approach, including Cooley (2004); Cooley and Christensen (2006). This methodology was extended by Doherty and Christensen (2011) and modified by Gosses and Wöhling (2019) to evaluate predictive bias induced by not only parameter simplifications but also other model structural errors, including factors such as the number of layers, boundary conditions, among others. In simple terms, the methodology proposed by Doherty and Christensen (2011) uses a random realization of a complex model (including any uncertain aspect of it) to generate a set of synthetic

outputs, which are then used to calibrate a simple model by applying regularized inversion. This process is repeated many times, resulting in a set of simple model and complex model predictions that can be compared to quantify predictive bias. Although this methodology is illustrative, it is computationally expensive, as it requires the generation of many complex model output realizations and the calibration of the simple model to each of these realizations. Another approach is to develop a statistical representation of model structural error that does not rely on the assumption of uncorrelated errors. Cooley (2004); Cooley and Christensen (2006) developed a methodology to estimate the covariance matrix of structural error induced by parameter simplification, by running many paired simple and complex models, and then apply it in the calibration of a parsimoniously parameterized model. Other authors have extended this methodology to evaluate and update the covariance matrix of structural error during the history matching process (Oliver and Alfonzo, 2018; Alfonzo and Oliver, 2020; Evensen, 2021; Lu and Chen, 2020). In their methods, the structural error is learned from the residuals obtained during the history matching process or at the end of it, without the need of a complementary complex model. Although this is a promising approach, it is not clear how the observed residuals are affected by the compensatory roles that parameters must adopt to fit the data, or even by the regularized/stochastic inversion process itself, all of which may lead to artificially low residuals on one hand or excessively noisy residuals on the other.

For groundwater modelling to still provide useful insights on the probability of occurrence of events with potentially detrimental economic, social, or environmental consequences, it is crucial to identify, evaluate, and assimilate model defects within the history matching and uncertainty quantification workflow. In this work, a new methodology is proposed to estimate the statistical properties of structural error using a correlation model of calibration residuals and observations. The correlation model is built using data space inversion (DSI) (Sun and Durlofsky, 2017), and relies on an ensemble of residuals and observations derived from the calibration of an imperfect model, referred to here as the simple model, to a set of synthetic observations generated by a more complex model. Thus, the use of paired complex-simple models is a prerequisite for applying the methodology. The simple model must be calibrated to the complex model outputs multiple times to generate the ensemble of residuals and observations. Conceptually, the complex model can be based on the simple model but incorporating additional features whose impact on predictive bias and predictive uncertainty can be tested. Alternatively, a modeller may choose to simplify a complex model to improve runtime efficiency or stability. In this scenario, the proposed methodology can be used to evaluate the impacts of these simplifications on model predictions, both in terms of bias and uncertainty. Before testing the methodology, an evaluation of predictive bias is performed using the complex-simple model approach proposed by Doherty and Christensen (2011). Two parameterization schemes are tested

for the simple model to assess their impact on misfit and predictive bias. An analysis of the results leads to a discussion of the advantages and disadvantages of each parameterization scheme, as well as whether structurally simple but parametrically complex models are better suited to quantify the uncertainty of predictions that are similar in nature to the observations. Finally, the effectiveness of the methodology is evaluated in the prediction of groundwater inflows to an open pit by comparing predictive uncertainty estimates obtained both with and without the incorporation of structural error in the history matching process of a zone-based parameterized simple model. It is demonstrated that by incorporating samples of structural error derived from the correlation model, predictive bias is reduced, and predictive uncertainty is estimated more conservatively.
The chapter is structured as follows: The next section describes the methodology of Doherty and Christensen (2011) to quantify predictive bias and presents the proposed methodology to develop a linear correlation model of structural error. A workflow is proposed that includes the generation of the ensemble of residuals and observations, the construction of the correlation model, the generation of realizations of structural error, and the use of these realizations in the history matching process. The subsequent section presents an illustrative numerical example, where the proposed methodology is tested. The chapter finishes with a discussion of the results and the drawing of conclusions.

5.2 Methodology

5.2.1 Predictive Bias Quantification

In this work, bias quantification is estimated using the complex-simple model approach proposed by Doherty and Christensen (2011). This approach is based on theory derived from subspace linear analysis, where the model parameter space is separated into a solution space and a null space. The following is a subspace linear analysis theory summarized from Doherty and Christensen (2011). Let the vector $\mathbf{d}$ be a set of measurements, simulated from a complex model as

$$\mathbf{d} = \mathbf{G}_c\mathbf{x}_c + \boldsymbol{\epsilon}, \tag{5.1}$$

where $\mathbf{G}_c$ is the sensitivity matrix of the complex model outputs associated with $\mathbf{d}$ to the complex model parameter vector $\mathbf{x}_c$, and $\boldsymbol{\epsilon}$ is the vector of measurement errors. The complex model is assumed to be a representation of the system as accurately as possible, including hydraulic parameters and boundary conditions.
Let the scalar s_c be a prediction made by a complex model, and derived from a linearized version of the model as

$$s_c = \mathbf{g}_c^T\mathbf{x}_c, \tag{5.2}$$

where $\mathbf{g}_c$ is the sensitivity vector of the prediction to complex model parameter vector $\mathbf{x}_c$. If the complex model is simplified, it can be assumed that the parameter vector $\mathbf{x}_c$ can be decomposed into two orthogonal components, as follows:

$$\mathbf{x}_c = \mathbf{x}_s + \mathbf{x}_e, \tag{5.3}$$

where $\mathbf{x}_s$ is the parameter vector of the simple model and $\mathbf{x}_e$ represents the vector of parameters that are excluded from the complex model to simplify it. With this decomposition, Equation 3.1 can be rewritten as

$$\mathbf{d} = \mathbf{G}_s\mathbf{x}_s + \mathbf{G}_e\mathbf{x}_e + \boldsymbol{\epsilon}, \tag{5.4}$$

where $\mathbf{G}_s$ is the model sensitivity matrix of the simple model outputs to simple model parameters $\mathbf{x}_s$, and $\mathbf{G}_e$ is the model sensitivity matrix of the complex model outputs to excluded model parameters $\mathbf{x}_e$.
Similarly, the complex model prediction s_c can be rewritten as

$$s_c = \mathbf{g}_s^T\mathbf{x}_s + \mathbf{g}_e^T\mathbf{x}_e, \tag{5.5}$$

where $\mathbf{g}_s$ is the sensitivity vector of the prediction, using to simple model parameter vector $\mathbf{x}_s$, and $\mathbf{g}_e$ is the sensitivity vector of the prediction to excluded model parameters $\mathbf{x}_e$. Let the scalar $\underline{s}_s$ be a prediction made by a calibrated simple model, and derived from a linearized version of the model as

$$\underline{s}_s = \mathbf{g}_s^T\underline{\mathbf{x}}_s. \tag{5.6}$$

The difference between the predictions of the complex and the simple calibrated model is a measure of predictive error, and can be written as

$$\underline{s}_s - s_c = \mathbf{g}_s^T\underline{\mathbf{x}}_s - \mathbf{g}_s^T\mathbf{x}_s - \mathbf{g}_e^T\mathbf{x}_e. \tag{5.7}$$

The systematic propensity of the simple model to predict higher or lower than the complex model is a measure of predictive bias.
When singular value decomposition (SVD) is used to calibrate the simple model, the calibrated parameter vector $\underline{\mathbf{x}}_s$ is estimated as

$$\underline{\mathbf{x}}_s = \mathbf{Z}_1\boldsymbol{\Sigma}_1^{-1}\mathbf{Z}_1^T\mathbf{d}, \tag{5.8}$$

where $\mathbf{Z}_1$ and $\boldsymbol{\Sigma}_1$ are the solution space matrices of left singular vectors and singular values, respectively, obtained from SVD of $\mathbf{G}_s$.
Replacing Equation 5.8 and Equation 5.4 into Equation 5.7, and further manipulating

and simplifying the equation, the predictive error can be written as

$$\underline{s}_s - s_c = -\mathbf{g}_s^T \mathbf{Z}_2 \mathbf{Z}_2^T \mathbf{x}_s + \mathbf{g}_s^T \mathbf{Z}_1 \boldsymbol{\Sigma}_1^{-1} \mathbf{Z}_1^T \boldsymbol{\epsilon} + \left(\mathbf{g}_s^T \mathbf{Z}_1 \boldsymbol{\Sigma}_1^{-1} \mathbf{Z}_1^T \mathbf{G}_e - \mathbf{g}_e^T \right) \mathbf{x}_e. \tag{5.9}$$

As discussed by Doherty and Christensen (2011), the three terms of Equation 5.9 are key to understanding the generation of predictive bias by the simple model. The first term is the contribution of the null space of the simple model to the predictive error. This term exists even if the simple model is a perfect representation of the complex model, or in other words, if the simple model is a perfect representation of the natural system. The expected difference of this term is zero, as it is assumed that the prior expected values of the simple model parameters are normalized to zero. Therefore, predictive bias is not expected to arise from this term (this does not mean there will not be error in the predictions). The second term is the contribution to the predictive error from misfit of the simple model to the data (Moore and Doherty, 2005). The more singular values, and therefore parameter combinations, are used to fit the data, the greater the risk of predictive error. However, this term does not necessarily lead to predictive bias, as the expected value of measurement error is zero. Finally, the third term is key. It is the contribution of the excluded model parameters to the predictive error. As it is a function of excluded model parameters, it may generate consistent predictive error, and therefore predictive bias. Even if the model is not calibrated (i.e., $\mathbf{g}_s^T \mathbf{Z}_1 \boldsymbol{\Sigma}_1^{-1} \mathbf{Z}_1^T \mathbf{G}_e = 0$), this term may be non-zero. Doherty and Christensen (2011) demonstrated that the only case where this term is zero is when the prediction is only sensitive to the solution space component of the complex (or real) model. This occurs when predictions tend to be similar in space and time to the observations used to calibrate the model.

The methodology proposed by Doherty and Christensen (2011) is based on the generation of a set of synthetic outputs from a complex model, which represents a wide range of possible outcomes. Parameters of the complex model may include any uncertain aspect of the model, including the number of layers, layer thickness, boundary conditions, geological units, among others. Moreover, several conceptual models can be tested, each representing a different hypothesis of the system. In this work, a complex model is parameterized to represent the system as accurately as possible, including hydraulic parameters and boundary conditions. A prior probability distribution is assigned to each parameter type, and the complex model is run multiple times using an ensemble of parameter realizations, to generate a set of synthetic outputs. A simple model, which presumably has structural defects, is calibrated against each of the complex model outputs, resulting in a set of residuals. Assuming one or several predictions are obtained from the complex model, counterpart predictions generated by the calibrated simple model can be compared to the complex model predictions. The difference between the complex model predictions and the simple model predictions is a sign of predictive bias and propensity

for predictive error. Plots of complex model prediction vs. simple model prediction ('s vs s' plots) can be generated as shown in Figure 5.1. Based on the theory presented above, conclusions can be drawn from these plots, where visual inspection of these plots can provide insights into the presence of bias and predictive error in the simple model predictions.

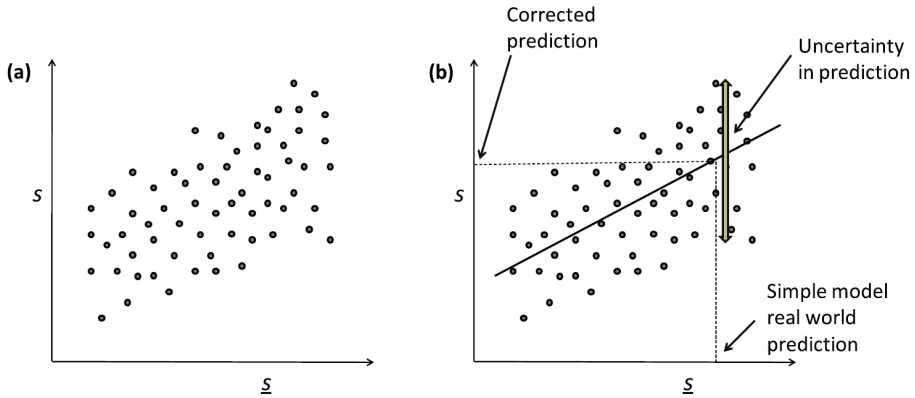

Figure 5.1: 's vs s' plot showing the bias in the predictions made by a simple model compared to a complex model (Doherty and Christensen, 2011).

5.2.2 Structural Model Error Estimation

In the original methodology proposed by Doherty and Christensen (2011), the primary goal was to generate a way of estimating unbiased predictive uncertainty estimates. While this is a valuable outcome, the main objective of this work is to estimate the statistical properties of structural error of the simple model, and incorporate these properties into the history matching process.

As in the original methodology, it is assumed that there is a complex and a simple model available. A total of N random realizations of the complex model outputs are generated, and the simple model is calibrated against these synthetic measurements. For each calibrated simple model, a set of residuals is obtained. It is evident that for some measurements, residuals will be larger than for others, suggesting the model is better suited to assimilate some data than others. This is a consequence of using a simplified model to represent a complex system. It is then expected that the residuals will contain information about the structural defects of the simple model. As more complex model outputs are generated, more information about the structural defects of the simple model can be obtained. Therefore, a statistical representation of the structural error in the simple model can be developed using an ensemble of measurements and calibration residuals.

Let $\mathbf{o}_i$ be the vector of measurements and corresponding residuals for the i-th simple

model calibration, from a set of N calibrations, as follows:

$$\mathbf{o}_i = \begin{bmatrix} \mathbf{d}_i \\ \mathbf{r}_i \end{bmatrix}, \tag{5.10}$$

where $\mathbf{d}_i$ is the vector of synthetic measurements (generated from running the complex model) and $\mathbf{r}_i$ is the vector of residuals. Let it be also assumed that an ensemble matrix $\mathbf{O}$ is built with the N vectors $\mathbf{o}_i$. A matrix of ensemble anomalies $\mathbf{Y}$ is obtained by centring $\mathbf{O}$ by its mean and normalizing it by its standard deviation $\boldsymbol{\sigma}_{\boldsymbol{Y}}$. Singular value decomposition (SVD) can be applied to $\mathbf{Y}$ as follows:

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T, \tag{5.11}$$

where $\mathbf{U}$, $\boldsymbol{\Sigma}$, and $\mathbf{V}$ are the left singular vectors, singular values, and right singular vectors of the ensemble of measurement-residual anomalies. If $\boldsymbol{\Phi}$ is defined as

$$\boldsymbol{\Phi} = \mathbf{U}^*\boldsymbol{\Sigma}^*, \tag{5.12}$$

where $\mathbf{U}^*$ and $\boldsymbol{\Sigma}^*$ are the left singular vectors and singular values truncated by an energy threshold (here set to 99%), then, a linear correlation model (DSI-RES model) can be defined as

$$\mathbf{o} = \mathbf{o}_f + \boldsymbol{\Phi} \cdot \mathbf{z} \cdot \boldsymbol{\sigma}_{\boldsymbol{Y}}, \tag{5.13}$$

where $\mathbf{o}$ are measurements and residuals simulated by the DSI-RES model, $\mathbf{o}_f$ is the mean vector of measurements and residuals, and $\mathbf{z}$ is a random vector of standard normal deviates (also called DSI-RES model parameters). Any realization of $\mathbf{o}$ can be then generated by sampling $\mathbf{z}$ from a standard normal distribution and applying Equation 5.13. This model is a surrogate model of the relationship between measurements and residuals, or in other words, the structural error in the simple model dependent on the measurements available for history matching. In other words, the structural error of the model will depend on its ability to assimilate the data, and is expected that this ability will vary depending on the measurements. Any random set of measurements and residuals can be generated by the DSI-RES model. The number of realizations N required to estimate model structural error is not clear, but given that DSI uses singular value decomposition, just a few realizations may be enough to capture some aspects of the structural error.
Let now be assumed that the simple model is to be history-matched to a set of real measurements $\underline{\mathbf{d}}$, through the use of ensemble methods, to estimate the posterior probability distribution of the model parameters and to quantify predictive uncertainty. Ensemble methods require the definition of a covariance matrix of the data noise $\mathbf{C}_\mathbf{d}$, or samples of noise derived from a probability density function (pdf) that uses $\mathbf{C}_\mathbf{d}$. In the presence of model structural error, samples of noise should not be necessarily drawn from a pdf

that assumes uncorrelated errors, as the residuals might be correlated. In this case, the DSI-RES model can be used to generate realizations of residuals.

Generating samples of residuals from the DSI-RES model to be used as samples of model structural error in the history matching process involves two steps: conditioning the DSI-RES model to the measurements $\underline{\mathbf{d}}$, and generating realizations of posterior residuals (hopefully the set of measurements is part of the output space from which the DSI-RES model was built). The first step is necessary as the unconditioned DSI-RES model is a correlation model of measurements and residuals within the broad range of possible measurements, generated from running the complex model multiple times; It is then required to constrain the DSI-RES correlation model to the specific set of measurements $\underline{\mathbf{d}}$, i.e., obtain a posterior probability distribution of residuals given the measurements. This can be done by applying any Bayesian-based method. Here, Markov chain Monte Carlo (MCMC) was implemented using the Python package pyDREAM (Shockley et al., 2017) to maximize the exploration of the posterior probability distribution of residuals, or posterior. The second step involves generating realizations of residuals from the posterior. Given that the results of the MCMC is an ensemble of samples representative of the posterior, subsamples can be drawn from this ensemble to be further used in the history matching process.

5.2.3 History Matching in the Presence of Structural Error

Once the DSI-RES model is built, realizations of structural model error conditioned to measurements can be generated, as explained in the previous section. There are several ensemble methods that can accommodate structural model error in the history matching process. The subspace iterative ensemble smoother (SEnRML) method (Evensen et al., 2019) is the method used in this work. In the low-rank implementation of the method, an ensemble matrix of model structural errors can be directly used in the inversion process, without the need of reconstructing the covariance matrix of noise. The method is summarized below.

The parameter solution of the SEnRML method is a linear combination of the initial ensemble anomalies $\mathbf{A}$,

$$\mathbf{X}^l = \mathbf{X}^f + \mathbf{A}\mathbf{W}^f, \tag{5.14}$$

where $\mathbf{X}^f$ and $\mathbf{X}^l$ are the first guess and updated model parameter ensemble realizations, respectively, and $\mathbf{W}^l \in \mathbb{R}^N xN$ is the matrix of weights. Solving the problem in this way, the inversion process is naturally regularized. The weights are iteratively updated as follows:

$$\mathbf{W}^{l+1} = \mathbf{W}^l - \gamma \left(\mathbf{W}^l - \mathbf{S}^{l^T} (\mathbf{S}^l \mathbf{S}^{l^T} + \mathbf{C_d})^{-1} \mathbf{H}^l \right), \tag{5.15}$$

where γ is the Gauss-Newton step length, and $\mathbf{H}^l$ is the 'innovation' term (Evensen et al.,

2019) defined as

$$\mathbf{H}^l = \mathbf{S}^l\mathbf{W}^l + \mathbf{D} - \mathbf{g}(\mathbf{X}^f + \mathbf{A}\mathbf{W}^l). \tag{5.16}$$

The only matrix that requires inversion in Equation 4.13 is $\mathbf{S}^l\mathbf{S}^{l^T} + \mathbf{C_d}$, where $\mathbf{C_d}$ is the covariance matrix of the data noise, and $\mathbf{S}^l$ is the matrix of predicted and 'deconditioned' ensemble anomalies at iteration l. There are several options for inverting this matrix as presented by Evensen et al. (2019). In the low-rank option, used in this work, the covariance matrix of the data noise (in this case the model structural error) is approximated by the ensemble of noise realizations $\mathbf{E}$, and the inversion of the referred term is approximated by the following:

$$(\mathbf{S}^l\mathbf{S}^{l^T} + \mathbf{E}\mathbf{E}^T)^{-1} = \left(\mathbf{U}\mathbf{\Sigma}^{+^T}\mathbf{Z}\right)(\mathbf{I}_N + \mathbf{\Lambda})^{-1}\left(\mathbf{U}\mathbf{\Sigma}^{+^T}\mathbf{Z}\right)^T, \tag{5.17}$$

where $\mathbf{U}$, $\mathbf{\Sigma}^+$, are the eigenvectors matrix and the pseudo-inverse of the matrix of singular values, derived from SVD decomposition of $\mathbf{S}^l$. Matrices $\mathbf{Z}$ and $\mathbf{\Lambda}$ are eigenvectors and singular values of the following:

$$\mathbf{\Sigma}^+\mathbf{U}^T\mathbf{E}\mathbf{E}^T\mathbf{U}\mathbf{\Sigma}^{+^T} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T. \tag{5.18}$$

The final ensemble parameter update $\mathbf{X}^{l+1}$ is calculated as

$$\mathbf{X}^{l+1} = \mathbf{X}^f(\mathbf{I} + \mathbf{W}^{l+1}/\sqrt{N-1}). \tag{5.19}$$

5.2.4 Workflow

The workflow of the proposed methodology is summarized below:

1. Build a simple and a complex model. It is expected that the complex model includes several aspects of the system, not represented in the simple model, that are candidates as potential causes of structural errors.

2. Run the complex model N times using random realizations of parameters (by parameters it is meant any uncertain aspect of the model that is of interest) to generate a set of synthetic measurements.

3. Calibrate the simple model to the complex model synthetic measurements, generating a set of residuals (difference between measurements and simple model outputs).

4. Build the ensemble matrix $\mathbf{O}$ integrating each vector $\mathbf{o}_i$ of measurements and residuals for each i-th calibration, and generate the DSI-RES correlation model.

5. Condition the DSI-RES model to the real measurements $\underline{\mathbf{d}}$ using MCMC (or any other numerical means), and obtain realizations of model structural error $\mathbf{E}$.

6. Use the matrix $\mathbf{E}$ in history matching the simple model to $\underline{\mathbf{d}}$ using the SEnRML method (Evensen et al., 2019) and quantify predictive uncertainty.

Two situations can be thought of in real world modelling that would allow the modeller to have a complex and a simple model available:

- The simple model is built first and predictive uncertainty has been quantified. Several hypotheses of potential causes of structural errors in the simple model that may affect predictions of interest are posed. A complex model is then built to include these aspects that are not represented in the simple model. It is assumed that, because of numerical instabilities or long runtimes, the complex model will not be used for history matching, but only to estimate the structural errors in the simple model. Still, it is necessary to run the complex model multiple times to generate a set of synthetic measurements. In any case, the number of model runs required should be significantly lower than the number of runs required to history match the complex model to the data. After estimating the simple model structural errors, they can be incorporated in the history matching process. After applying the methodology, the updated predictive uncertainty can be compared to the initial one, therefore assessing the impact of the structural errors in the simple model. This will provide insights into the sensitivity of the simple model predictions to the assumptions made on noise in the data.

- The complex model is built first, and the simple model is the result of a simplification process done to improve run time efficiency, stability, or to better assimilate data and quantify predictive uncertainty. The estimated model structural errors can then be incorporated in the history matching process of the simple model, and predictive uncertainty can be re-estimated. As the model is simplified, model structural errors may increase, leading to a reduced capacity of the simple model to extract information from the data, and therefore to quantify predictive uncertainty. However, assuming the simple model is a more efficient tool (in terms of runtime and stability) to perform history matching, a better exploration of uncertainty is foreseen. Perhaps prior to embarking on the last step, the modeller could evaluate if the level of structural error is acceptable, otherwise more complexity would be required. A trade-off between model structural errors and predictive uncertainty can then be evaluated.

5.3 Numerical Example

5.3.1 Model Description

Groundwater inflows to an open pit are important predictions for the design of dewatering systems, as they could lead to reduced operational efficiency and higher mining costs (Beale et al., 2014). Groundwater inflows are the result of the intersection of the slope with the saturated rock mass, due to the sequential excavation of a mineral deposit. Two numerical 2D models, simple and complex, were built to simulate groundwater inflows to a pit slope during 7 years over monthly stress periods, using MODFLOW-USG (Panday, 2024; Panday et al., 2013). The hydrogeological units are represented by three geological units: bedrock, overburden, and intrusive rocks, and two fault zones: subvertical and horizontal faults. The geological units are assumed to have contrasting hydraulic properties (see Table 5.1), where the overburden has the highest permeability, and the intrusive rocks have the lowest. Fault zones are assumed to have higher hydraulic conductivity values than the surrounding geological units, acting as preferential pathways for groundwater flow. The complex model has a total of 250 layers and 200 columns, resulting in a total of 50,000 cells. The lateral extension of the model is 500 m, and the vertical extension is 400 m, resulting in 2 m-width cells. This high level of refinement is assumed to be required to represent details on the distribution of geological units and faults, the pit excavation, and the temporal changes on hydraulic properties, due to the lithostatic unloading process. This process occurs as a result of the decompression of the rock mass due to the excavation of the pit, leading to the development of a disturbance zone (Hoek and Brown, 2019), which enhances the permeability and storage properties of the rock mass near the pit slope, potentially increases the groundwater inflows to the pit. The disturbance zone goes around 30% of the total excavation depth, as is modelled with the following equation:

$$\mathbf{x}_f = \mathbf{x}_{fmax} - \mathbf{x}_{fmax}\left(1 - \frac{\mathbf{z}_{pb} - \mathbf{z}}{0.3H}\right), \tag{5.20}$$

where $\mathbf{x}_f$ is the parameter multiplier, $\mathbf{x}_{fmax}$ is the maximum value of the parameter multiplier, $\mathbf{z}_{pb}$ is the elevation of the pit bottom, and $\mathbf{z}$ is the elevation of the cell. This is a simplification of the real process, as the disturbance zone is not a linear function of the depth, but it is a good approximation for the purpose of this study. Changes of hydraulic properties from their initial values are modelled as a function of the parameter multiplier $\mathbf{x}_f$, using the TVM package of MODFLOW-USG. This is applied for hydraulic conductivity, specific storage, and specific yield.

Time-varying recharge to the pit has also been included in the complex model, to add another layer of complexity to the groundwater inflows predictions. A view of the complex

model with the pit excavation at 4 selected years is shown in Figure 5.2, where it is possible to observe the disturbance zone around the pit slope, the three geological units, and the faults. The water table and piezometric contours as a result of running the model with contrasting parameter values are also shown for illustration purposes.

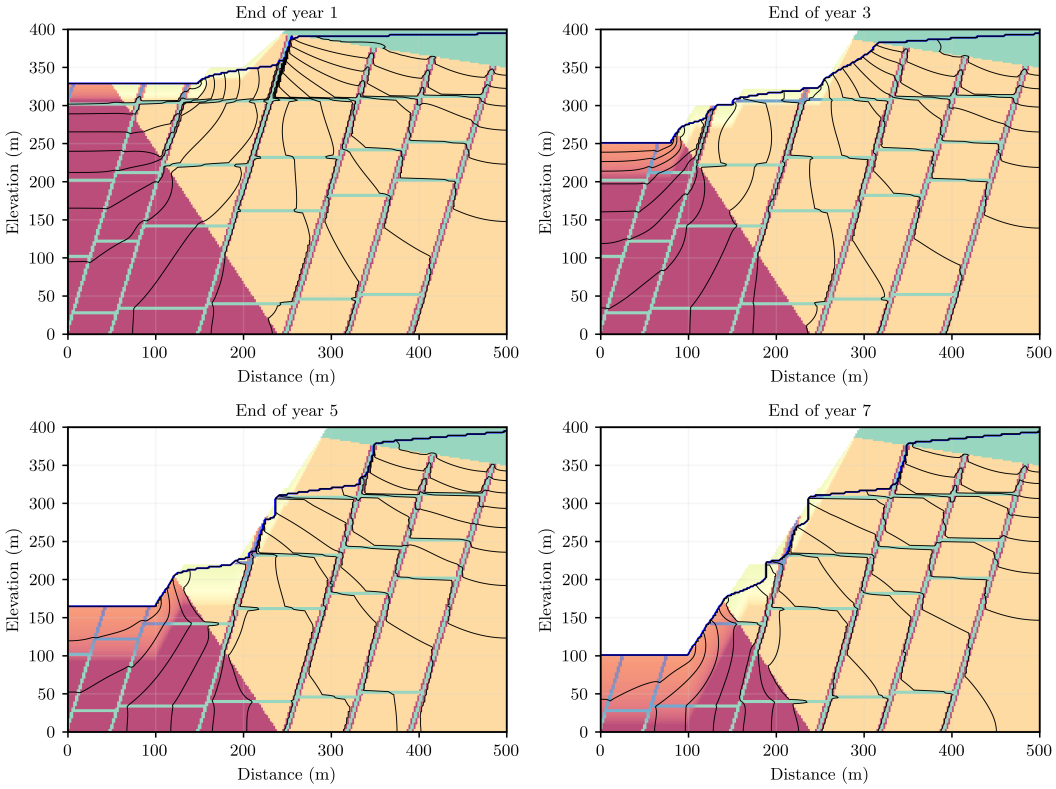

Figure 5.2: Cross-section views of the complex model at four selected years of pit excavation, showing the distribution of geological units, fault zones, lithostatic unloading zones, water table, and piezometric contours. The geological units, from top to bottom, are: overburden (green), bedrock (yellow), and intrusive rocks (red). Fault zones are represented in green, with damage zones shown in red. The lithostatic unloading zone is depicted by lighter-coloured cells around the excavation zone.

The simple model is a coarse representation of the complex model, with 16 layers and 13 columns, resulting in a total of 208 cells. This level of refinement precludes the model from representing geological faults, and the operational changes on hydraulic properties. Therefore, the model includes only the geological units and the pit advance (in a simplified fashion). Figure 5.3 shows the simple model with the pit excavation at 4 selected years, and the water table and piezometric contours.

The complex model is considered the true model and is used to estimate the structural errors of the simple model. It is run many times with different realizations of parameter sets. A grid-scale parameterization scheme is used for the complex model, for hydraulic conductivity, specific storage, and specific yield. The prior distribution of the parameters is assumed to be log-normal, with a covariance matrix derived from an exponential var-

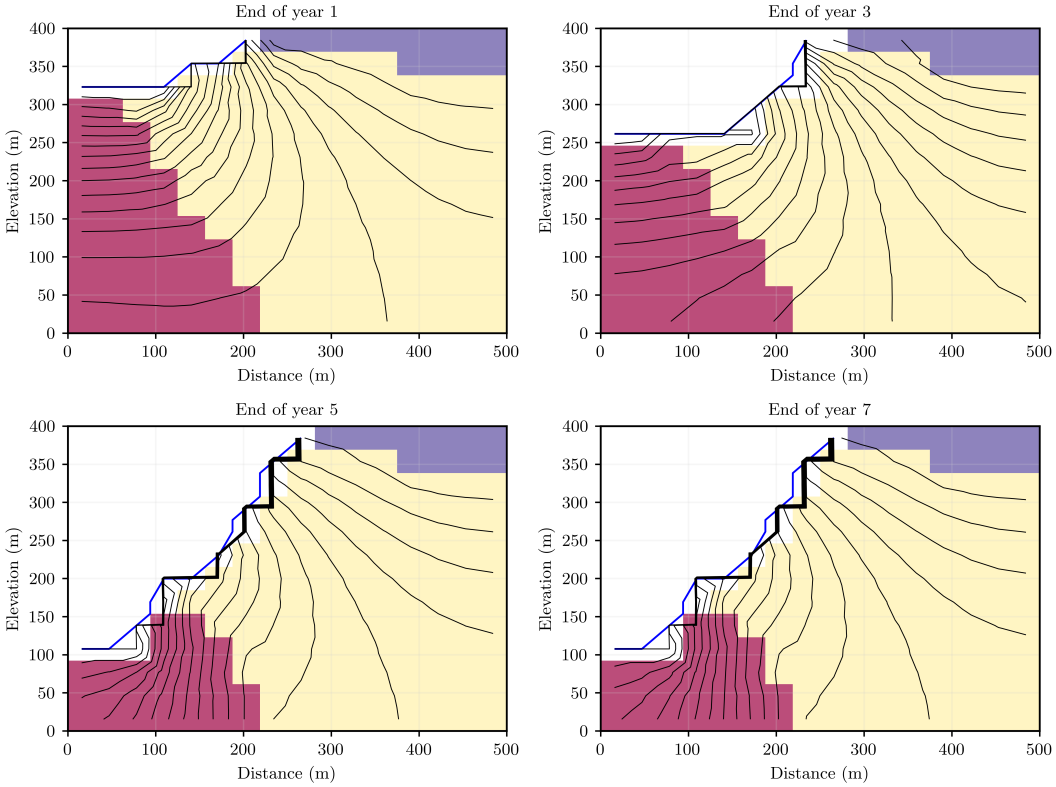

Figure 5.3: Cross-section views of the simple model at four selected years of pit excavation, illustrating the distribution of simplified geological units, water table, and piezometric contours, as described in Figure 5.2.

iogram model, with variable range. The hydraulic properties details are shown in Table Table 5.1.

Time-varying recharge factors for each stress period are also considered uncertain, and temporally correlated using an exponential variogram model with a range of 1 year. These factors are applied at each stress period to a synthetic recharge time series, that simulates the infiltration of water derived from sporadic rainfall events.

Table 5.1: Prior parameters for different properties and geological units of the complex model.

Property	Geological Unit	Mean	Min	Max	Sill	Range	Anisotropy Ratio
hk	Bedrock	5×10^{-3}	1×10^{-4}	1×10^{-2}	0.25	100.0	1.0
	Overburden	5×10^{-1}	1×10^{-2}	5×10^{0}	0.25	250.0	10.0
	Subvertical faults	1×10^{-1}	1×10^{-3}	1×10^{0}	0.25	250.0	1.0
	Subvertical fault damage zones	1×10^{-4}	1×10^{-5}	1×10^{-3}	0.25	250.0	1.0
	Horizontal faults	5×10^{-1}	1×10^{-2}	5×10^{0}	0.25	250.0	1.0
	Intrusive rocks	1×10^{-4}	1×10^{-5}	1×10^{-3}	0.25	100.0	1.0
vka	Bedrock	1.0	0.1	10.0	0.25	100.0	1.0
	Overburden	10.0	1.0	100.0	0.25	250.0	10.0
	Subvertical faults	1.0	0.1	10.0	0.25	250.0	1.0
	Subvertical fault damage zones	1.0	0.1	10.0	0.25	250.0	1.0
	Horizontal faults	1.0	0.1	10.0	0.25	250.0	1.0
	Intrusive rocks	1.0	0.01	100.0	0.25	100.0	1.0
ss	Bedrock	1×10^{-6}	1×10^{-7}	1×10^{-5}	0.25	100.0	1.0
	Overburden	1×10^{-5}	1×10^{-6}	1×10^{-4}	0.25	250.0	10.0
	Subvertical faults	1×10^{-7}	1×10^{-8}	1×10^{-6}	0.25	250.0	1.0
	Subvertical fault damage zones	1×10^{-7}	1×10^{-8}	1×10^{-6}	0.25	250.0	1.0
	Horizontal faults	1×10^{-7}	1×10^{-8}	1×10^{-6}	0.25	250.0	1.0
	Intrusive rocks	1×10^{-6}	1×10^{-7}	1×10^{-5}	0.25	100.0	1.0
sy	Bedrock	1.5×10^{-3}	1.0×10^{-4}	1.0×10^{-2}	0.09	100.0	1.0
	Overburden	5.0×10^{-2}	5.0×10^{-3}	2.5×10^{-1}	0.0144	250.0	10.0
	Subvertical faults	1.5×10^{-3}	1.0×10^{-4}	1.0×10^{-2}	0.09	250.0	1.0
	Subvertical fault damage zones	1.5×10^{-3}	1.0×10^{-4}	1.0×10^{-2}	0.09	250.0	1.0
	Horizontal faults	1.5×10^{-3}	1.0×10^{-4}	1.0×10^{-2}	0.09	250.0	1.0
	Intrusive rocks	1.5×10^{-3}	1.0×10^{-4}	1.0×10^{-2}	0.09	100.0	1.0

A total of 100 realizations of the prior distribution of parameters were generated, and the complex model was run for each of these realizations. Seven realizations did not result in physically plausible groundwater inflows to the pit, and were discarded. Figure 5.4 shows the groundwater inflows to the pit slope for the remaining 93 realizations. As can be seen in the figure, the first three years of the simulation are defined as the calibration period, and the remaining four years are the predictive period. A total of 15 discrete measurements of groundwater inflows, every 2 months, were extracted from each of the 93 realizations, and a measurement error with a standard deviation of 0.05 m^3/d was added to each of these measurements. One realization of these measurements is shown in Figure 5.4 as red squares. A separated predictive dataset was also extracted from the predictive period. For the sake of simplicity, only a selected subset of the predictive time series is used to discuss the results. These predictions are called O19, O23, O27, O31, and O39 and are shown in Figure 5.4.

The synthetic measurements were used as the observed data to calibrate the simple model using regularized inversion with the support of PEST software (Doherty, 2023).

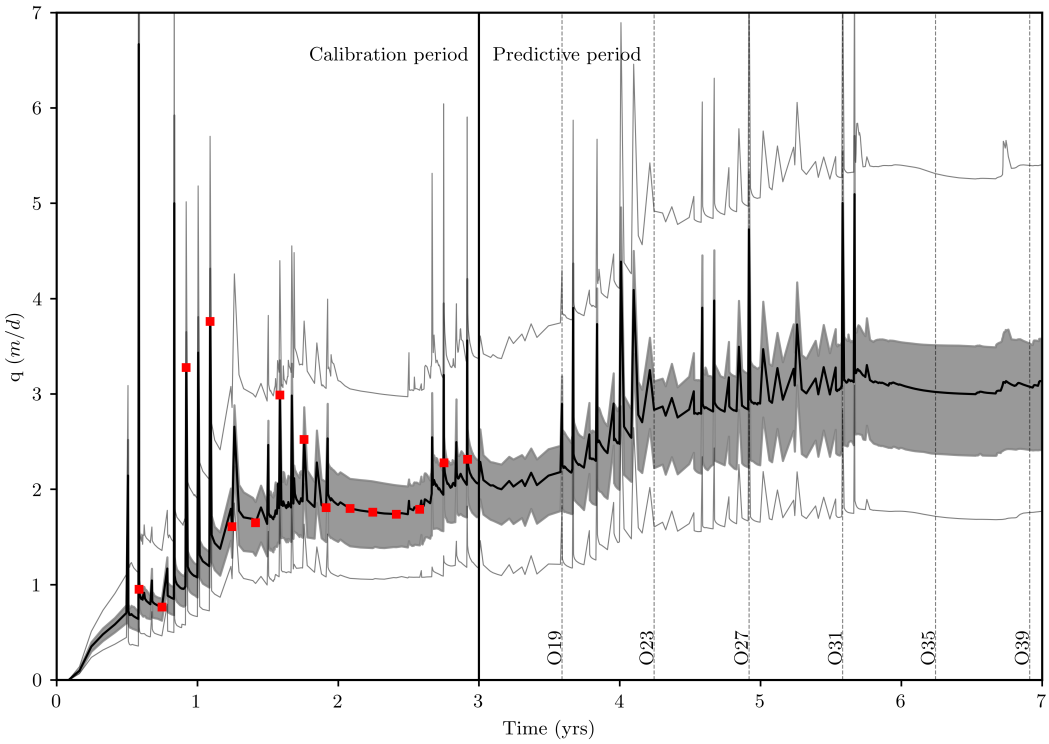

Figure 5.4: Complex model - prior simulated inflows to the pit. The solid black line is the mean, the grey-shaded area is the P25-P75 percentile region, and the external grey lines are the P2 and P98 percentiles of the simulated groundwater inflows. One realization of measurements extracted from the mean complex model are shown as red squares.

Four parameterization schemes were tested for the simple model, two of them using grid-scale parameterization, and other two using zones - zonal model - as the parameterization device. Two recharge variations were tested in these cases, one with a constant recharge to the pit, and another with time-varying recharge. Table 5.2 summarizes the cases tested in this work. The calibration included the parameters listed in Table 5.1, and the recharge factors for the time-varying recharge option. For the highly-parameterized model, two calibration schemes were tested, one with a target of goodness of fit commensurate with the noise level of the data, and another with a target comparable to five times more noise. The number of estimable parameters in this case was 1,008, including the 176 recharge factors. For the zonal model, two recharge parameterizations were tested, one with a constant recharge factor applied to the synthetic recharge time series, and another with 176 time-varying recharge factors, each applied to a different precipitation event. Then, the zonal model was parameterized with 13 estimable parameters in one case, and 189 in the other. Prior information, used for regularization purposes, was weighted as the inverse of the standard deviation, or the square root of the sill (see Table 5.1). The same geostatistical properties (sill and range) were used for the simple model (highly-parameterized case) as for the complex model.

Plots of predictive groundwater inflows made by the complex model and the simple model ('s vs s' plots) were generated to compare the predictions of the two models and identify predictive bias incurred by the simple model. These are the first results that will be presented in the next section. Calibration residuals of 'measured' groundwater inflows were also included in the 's vs s' plots and analysed to identify calibration-induced bias. Case (c) was selected to be further analysed in the presence of structural error. A DSI-RES correlation model was built using an ensemble $\mathbf{O}$ of 93 realizations of measurements extracted from complex model outputs and residuals derived from simple model calibrations. The DSI-RES model was conditioned to one specific measurement dataset $\underline{\mathbf{d}}$ selected from the complex model output realizations, using MCMC implemented in py-DREAM. Posterior residuals that represent model structural errors were extracted from the posterior ensemble $\mathbf{O}'$, obtained from the MCMC Bayesian inference. A subset of these realizations were used to build the ensemble matrix $\mathbf{E}$. The simple model of Case (c) was history matched twice: once with samples of noise drawn from a pdf that assumes uncorrelated errors with a standard deviation of 0.05 m^3/d, and another using the ensemble matrix $\mathbf{E}$, representing samples of model structural error. The low-rank version of SEnRML was used to history match the simple model to $\underline{\mathbf{d}}$.

5.3.2 Results

Figure 5.5 shows the 's vs s' plot for the case (a), where the simple model is highly-parameterized and calibrated to observed data. Before discussing the results, an expla-

Table 5.2: Simple model parameterization schemes and recharge options tested.

Case	Description
(a) Highly-parameterized model	Grid-based parameterization for hydraulic properties, time-varying recharge.
(b) Highly-parameterized, poorer fit	Increased target measurement objective function
(c) Zonal model	Cells are grouped into zones
(d) Zonal model + variable recharge	time-varying recharge

nation of the figure is necessary. First, measured versus simulated groundwater inflows are represented as red dots in the plot, along with the 1:1 line. Knowing that the number of measurements is 15, the red dots correspond to the 93 realizations of measured-simulated data pairs, each representing a different calibration of the simple model. It is useful to visualize the spread of these points along the 1:1 line, and compare it to the spread of predictive bias. Second, each of the six plots within the figure shows a distinct prediction of groundwater inflows (O19, O23, O27, O31, and O39), as previously described, with predictions becoming more temporally distant from the end of the calibration period. It is expected that predictions that are closer to the calibration period will exhibit less bias than those that are further away. Each of the black dots represents a pair of complex-simple model predictions for each of the 93 calibrated models. The distance between the best linear fit of the prediction pairs and the 1:1 line provides a measure of predictive bias of the simple model. The spread of these points quantifies the predictive error variance. As shown in Figure 5.5, the simple model of case (a) fits the data well, as evidenced by the red dots being close to the 1:1 line. However, there is a clear bias in the predictions made by the simple model, which tends to underestimate groundwater inflows to the pit. More notably, the bias is stronger for higher flows. As expected, bias increases as the prediction period extends further from the calibration period. For prediction O31, the simple model is strongly biased towards lower values, causing the line of best fit to deviate significantly from the 1:1 line. It is identified that prediction O31 corresponds to a peak flow derived from a recharge event. Although prediction O27 is also a peak flow, the bias is less pronounced in this case. When the target measurement objective function is increased, i.e., allowing for a poorer fit to the data (case (b)), the predictive bias did not change significantly relative to the case (a), as shown in Figure 5.6. This suggests that predictive bias is not necessarily caused by the adjustment of the parameters to all data, but rather by specific measurements that are more difficult to fit, particularly groundwater inflow peaks resulting from recharge events.

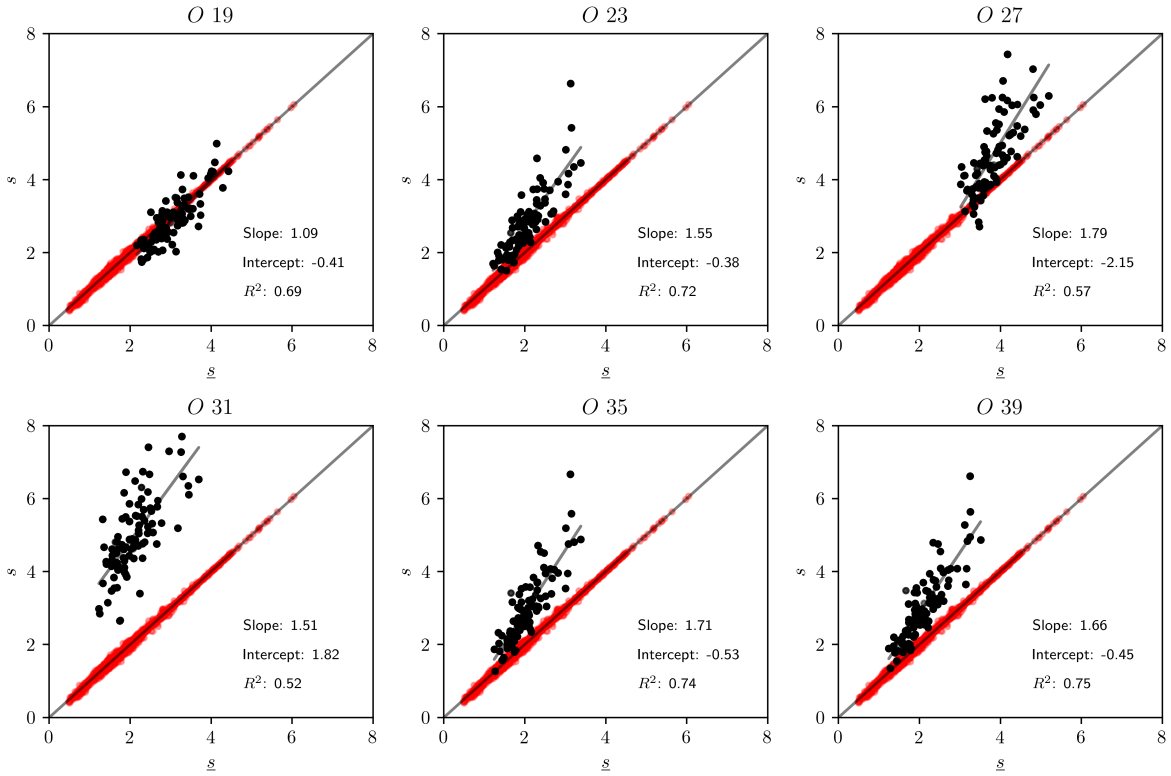

Figure 5.5: 's vs s' plot for case (a). Scatter points of measured vs simulated data are shown in red (repeated in every plot).

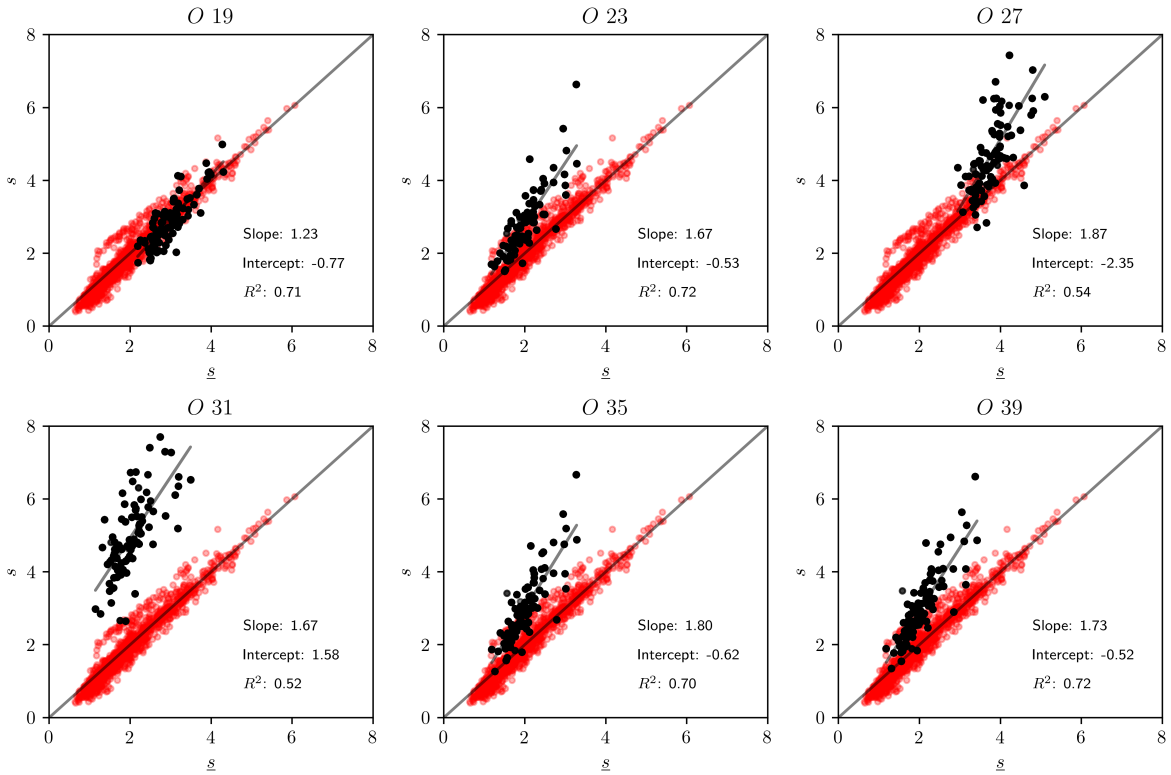

Figure 5.6: 's vs s' plot for case (b). Scatter points of measured vs simulated data are shown in red (repeated in every plot).

The zonal model (case (c) and case (d)) is tested next, and the results are shown in Figure 5.7 and Figure 5.8. It is apparent that the zonal model with only one recharge factor cannot fit the data to the level of measurement error. Moreover, its goodness of fit is similar to that of case (b), the highly-parameterized model with a poorer fit. The inability to fit the data is a consequence of the simplifications made in the model, but paradoxically translates into less predictive bias compared to the highly-parameterized model (cases (a) and (b)). In fact, the best linear estimate of the complex-simple model prediction pairs is closer to the 1:1 line, and the spread of the points along the y-axis is slightly reduced, as shown in Figure 5.7. Case (d) introduces 176 recharge factors to the zonal model to test their effect on the goodness of fit and predictive bias. Interestingly, the level of fit to the data improved significantly without a noticeable increase in predictive bias, as shown in Figure 5.8.

The visual comparison of the 's vs s' plots for the four cases can also be complemented by analysing of the best linear estimate of the complex-simple model prediction pairs. This fit is defined by the slope and intercept of the line that minimizes the sum of squared differences between the prediction pairs and the best linear estimate. The slope is a measure of how the bias changes with the magnitude of the prediction, and the intercept reflects the basal bias that is present for all predictions made by the simple model. A comparison of the slopes and intercepts for case (a) (highly-parameterized model) and

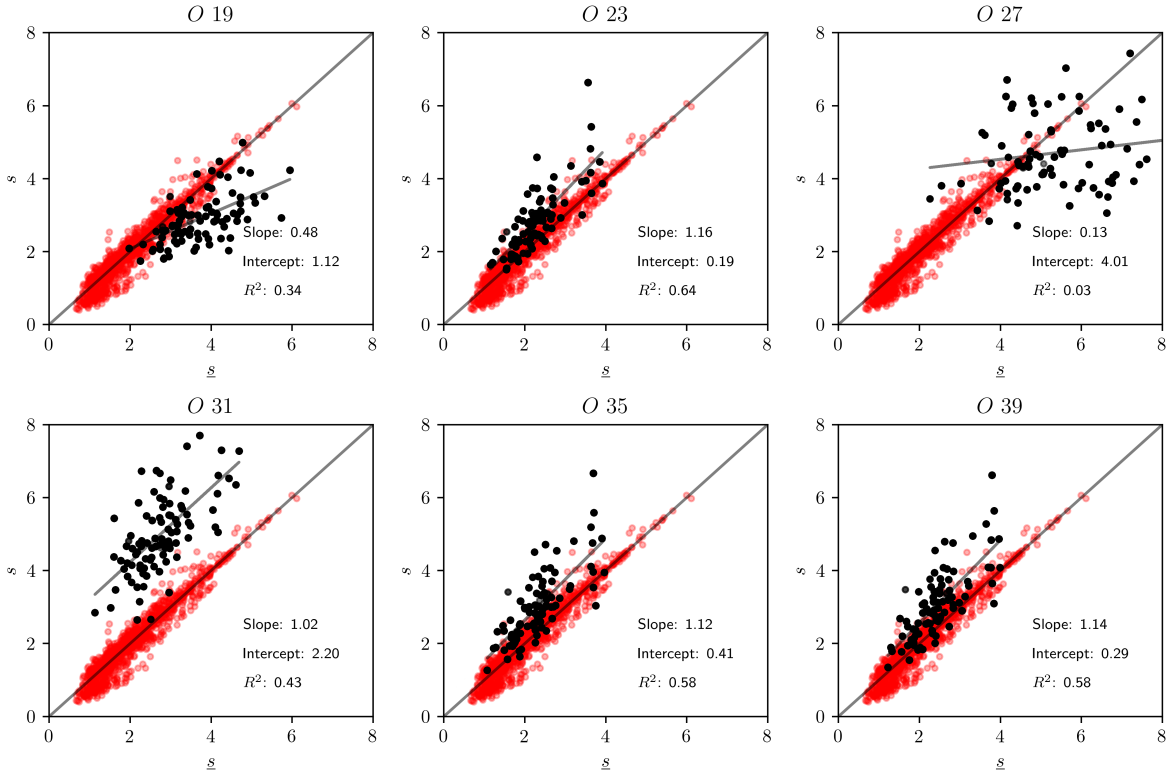

Figure 5.7: 's vs s' plot for case (c). Scatter points of measured vs simulated data are shown in red (repeated in every plot).

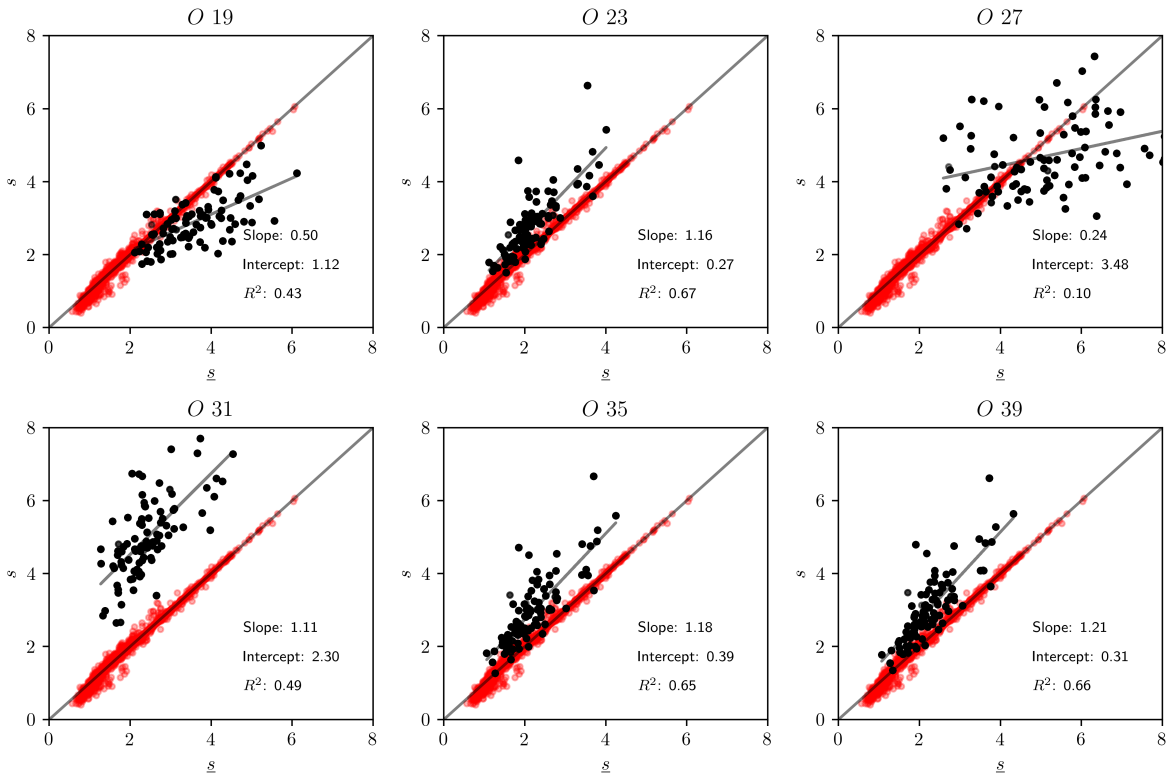

Figure 5.8: 's vs s' plot for case (d). Scatter points of measured vs simulated data are shown in red (repeated in every plot).

case (c) (zonal model) is provided in Table 5.3. First, the coefficient of determination R^2 is higher for the highly-parameterized model, indicating a clear relationship between the complex and simple model predictions. Also, the slope of the best linear estimate is always greater than 1.0 for the highly-parameterized model, with a maximum value of 1.79. This shows that the bias produced by the simple highly-parameterized model is more pronounced for higher flow predictions. Among the predictions, only O27 exhibits an intercept greater than 0, which is consistent with the visual analysis of the 's vs s' plot, where the bias is relatively constant across a wide range of predictions. For the zonal model, the slope of the best linear estimate varies, falling below 1.0 for some predictions and rising above 1.0 for others. The maximum value of the slope is 1.14, which is lower than the minimum slope of the highly-parameterized model. This provides a quantitative measure of the reduced predictive bias in the zonal model compared to the highly-parameterized case. However, reviewing the coefficients of determination reveals that the relationship between the complex and simple model predictions is less clear for the zonal model. This may be due to the fact that the zonal model was not able to fit the data to the level of measurement error, as previously discussed. By including 176 recharge factors in the zonal model, the slope of the best linear estimate remains largely unchanged, but the coefficient of determination increases slightly for all predictions.

Table 5.3: Comparison of the best linear estimated parameters between case (a) and case (c). R^2 is the coefficient of determination.

Prediction	(a) Highly Parameterized Model			(c) Zonal Model		
	Slope	Intercept	R^2	Slope	Intercept	R^2
O19	1.09	-0.41	0.69	0.48	1.12	0.34
O23	1.55	-0.38	0.72	1.16	0.19	0.64
O27	1.79	-2.15	0.57	0.13	4.01	0.03
O31	1.51	1.82	0.52	1.02	2.20	0.43
O35	1.71	-0.53	0.74	1.12	0.41	0.58
O39	1.66	-0.45	0.75	1.14	0.29	0.58

A DSI-RES model was constructed using the ensemble matrix $\mathbf{O}$, which corresponds to the set of 15 measurements and residuals obtained from the calibrating the simple model (case (c)) 93 times. Realizations of structural error were obtained after conditioning the DSI-RES model to a specific set of measurements (this is one realization selected from the complex model output realizations). First, it is insightful to compare the covariance matrix of noise under the assumption of uncorrelated errors (diagonal matrix, with a standard deviation of 0.05 m^3/d) to the covariance matrix of simple model structural error as shown in Figure 5.9. The latter is an empirical matrix built from posterior model residuals extracted from the posterior samples $\mathbf{O}'$ of the conditioned DSI-RES model. As shown in the figure, the covariance matrix of simple model structural error (right) exhibits spatial correlation, with greater variance and covariance between different

observations compared to the uncorrelated covariance matrix of noise (left). This better reflects the nature of the structural defects of the simple model when history matching is performed with a specific set of measurements.

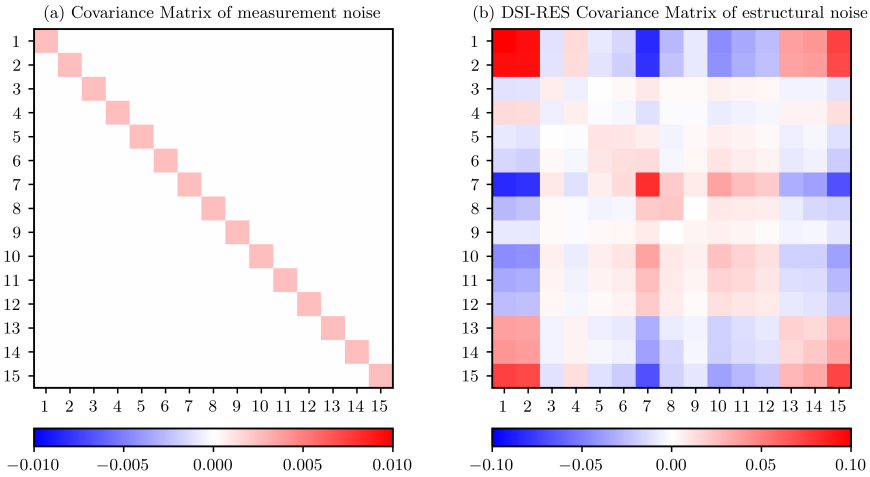

Figure 5.9: Covariance matrices of (a) measurement error and (b) simple model structural error, the latter generated with DSI-RES, for the set of 15 measurements that were used in the calibration process. Note the difference in the colour scale.

As previously explained, the simple model was history-matched using SEnRML on two occasions: first with the diagonal covariance matrix of error, and then using the ensemble matrix **E**, which represents samples of simple model structural error. The ensemble matrix **E** was built using a subset of the posterior residuals extracted from the posterior ensemble $\mathbf{O}'$. Figure 5.10 shows the history matching results for the simple model, using the two different approaches. The case that includes a diagonal covariance matrix for the noise is described as 'Diagonal Cd', and the case that includes the structural error is described as 'Structural Cd'. It is observed that the residuals are generally high for both cases and very similar to the regularized inversion results previously presented. On one hand, this reflects the inability of the simple model to history-match the data to the level of measurement error (0.05 m^3/d) when using a diagonal covariance matrix commensurate with measurement error. On the other hand, the similar result obtained when using the ensemble matrix **E** suggests that the structural error incorporated in the history matching process is of the same order of magnitude as the minimum misfit that can be achieved by the simple model, which is expected.
Figure 5.11 shows the predictive uncertainty ranges obtained for the six predictions of groundwater inflows to the pit, for both cases. As depicted in the figure, the predictive uncertainty ranges obtained using a diagonal covariance matrix fail to cover the true values for all predictions and have very limited coverage of the true values for three of the six predictions: O19, O27, and O31. In contrast, the predictive uncertainty ranges

obtained by incorporating structural error in the history matching process cover the true values for all predictions and exhibit reduced bias compared to the previous case. Furthermore, the predictive uncertainty ranges are wider due to the inclusion of structural error, which is greater than measurement error. It is also noted that lower groundwater inflows (predictions O23, O35, and O39) generally exhibit less predictive bias for both cases. Hence, including structural error in the history matching process did not make a significant difference in reducing predictive bias.

The results of predictive uncertainty ranges presented in Figure 5.11 and discussed above reveal that the simple model is not capable of extracting enough information from the data, and that there is predictive bias in the predictions. Although the inclusion of structural error in the history matching process did reduce predictive bias compared to the case where only measurement error was considered, it appears that predictive uncertainty is still overestimated for some predictions.

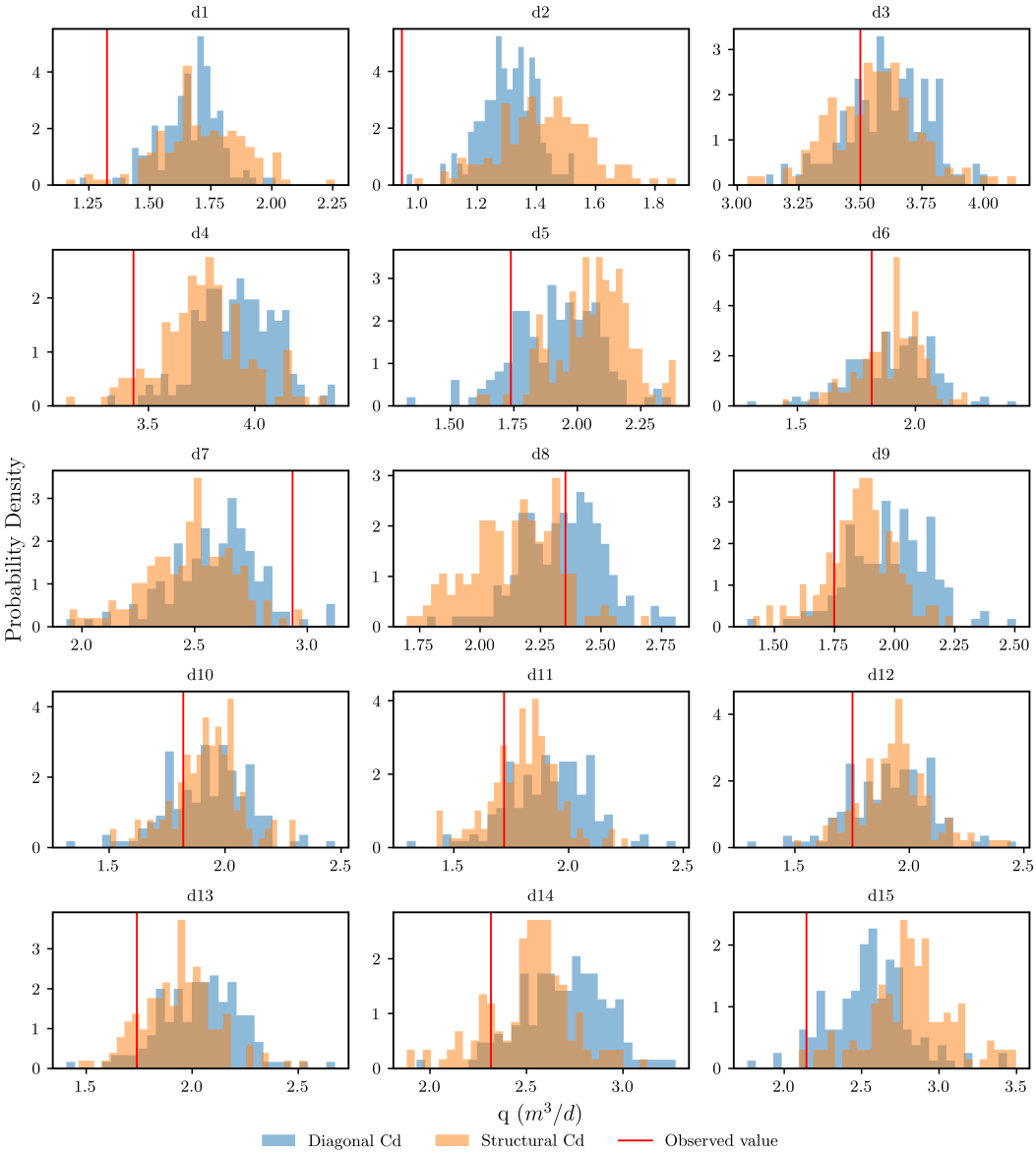

Figure 5.10: Histograms of history-matched groundwater inflows for the 15 measurements used in the calibration process.

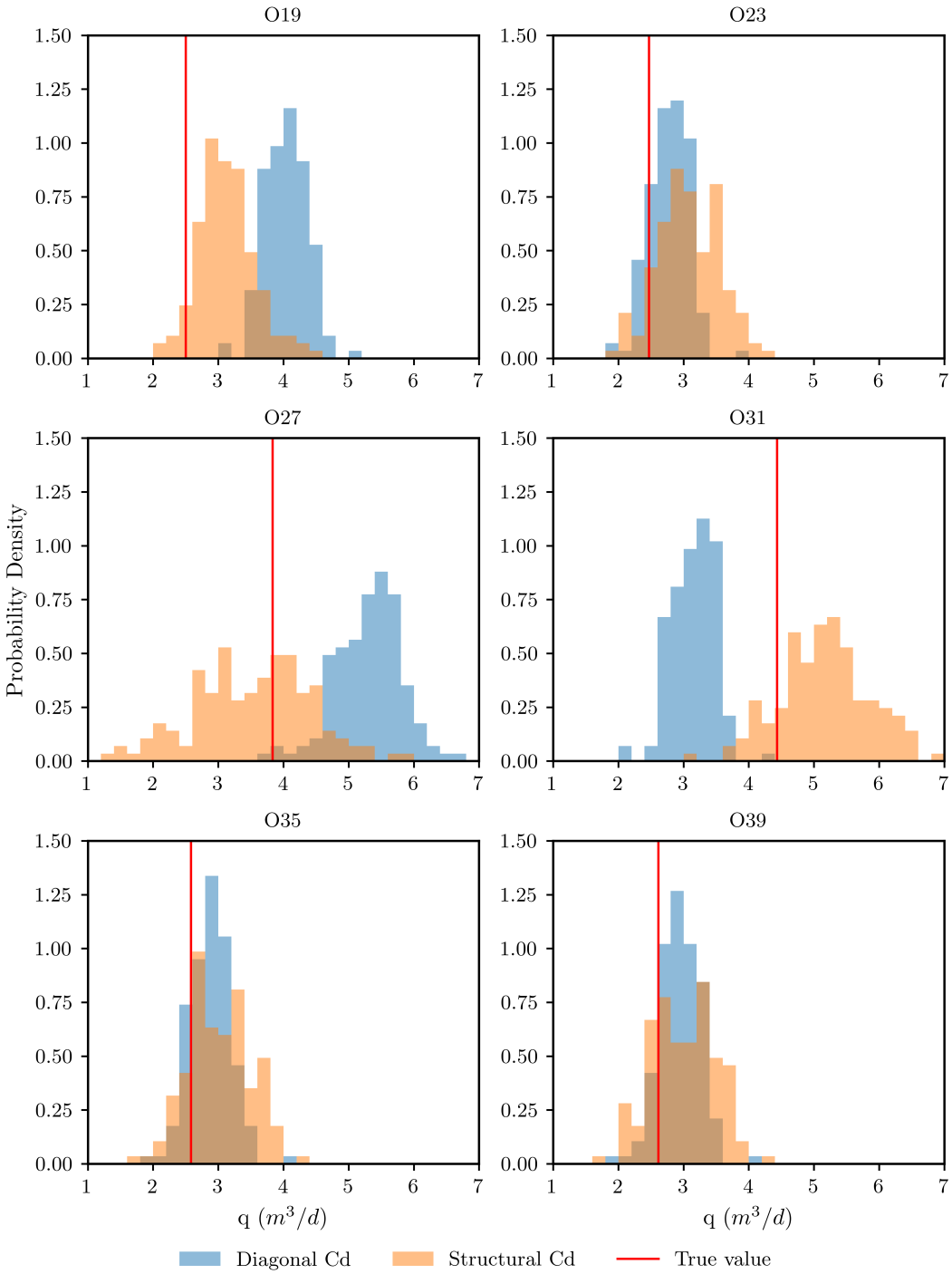

Figure 5.11: Predictive uncertainty ranges for the simple model.

5.4 Discussion

Several key findings of this study merit discussion. These include model simplification, data harvesting, predictive bias, and predictive uncertainty—each a critical aspect of groundwater modelling for decision support.

The results clearly show that the simple model, under an underparameterized zone-based scheme, could not fit the data to the level of measurement error. This is due to the limited capacity of the model to act as a receptacle for the information contained in the data. As a result, the compensatory roles of parameters are minimized with this parameterization scheme, and the misfit reflects these model deficiencies (White et al., 2014). Predictive uncertainty may be overestimated (although this is case-specific), and predictive bias may be partially mitigated by the inability of the parameters to adopt compensatory roles to fit the data. In contrast, the results for the simple model with a highly-parameterized scheme demonstrated that it could fit the data better than the zone-based model but incurred greater predictive bias. This outcome is a result of the expanded capacity of the parameters to adopt compensatory roles to fit the data, which is a well-known issue in groundwater modelling (Doherty and Christensen, 2011; White et al., 2014). However, a key finding of this study is that, in this case, the predictions of interest - groundwater inflows to the pit - were of the same nature as the data used for calibration. Predictive bias was found to be greater for the highly-parameterized model than for the zone-based model. It has been argued that in such cases, the damage to the predictive capacity of the model might be minimized, as the compensatory roles of the parameters in fitting the data could produce a similar compensatory effect in the predictions (Doherty and Christensen, 2011; White et al., 2014). This is characteristic of data-driven models. In line with this, some authors (for example, Knowling et al., 2019) have suggested that increasing the parameterization complexity of a simple model could reduce predictive bias, supporting the 'structurally simple but parametrically complex' paradigm. However, based on the results of this study, it can be argued that this is not always the case. Modellers cannot be fully aware of all the structural defects of a model relative to the unknown reality, nor they can discount the possibility that some of these unknown defects will contribute to predictive bias. This issue is inevitable, even when building a prediction-oriented and strategically abstract representation of a groundwater system (Doherty and Moore, 2020). Therefore, caution should be exercised when history matching a highly-parameterized simple model to the level of measurement error, as this may result in greater predictive bias than with a zone-based simple model that is considered poorly calibrated. Aware of this issue, modellers must consider the bias-variance trade-off (Hastie et al., 2009) when performing history matching and predictive uncertainty quantification.

The results of the history matching process of the simple model using an ensemble ma-

trix of structural error demonstrated that predictive bias was reduced and predictive uncertainty slightly increased compared to the case where only measurement error was considered. While it is recognized that wide predictive uncertainty ranges might not be desirable for decision-making, models with large predictive bias and narrow predictive uncertainty ranges might also be unsuitable for decision support, as their quantified predictive ranges might fail to encompass the true values of the predictions. Given that most groundwater models are imperfect representations of reality, it is advisable to include structural errors into the history matching process to minimize predictive bias. However, this comes at the necessary cost of increasing predictive uncertainty due to limited capacity of extracting information from the data. Achieving a balance between the model complexity and structural uncertainty is essential to minimize both predictive bias and uncertainty underestimation. The proposed methodology could be used to serve this purpose, but it is acknowledged that finding this balance is a monumental task.

The generation of random realizations of structural error estimates was achieved using the DSI-RES model, conditioned on a set of measurements. This represents an innovative application of data space inversion Sun and Durlofsky (2017) to a problem distinct from its traditional use. The numerical example presented in this study demonstrated the successful integration of structural errors into the history matching process of a simple groundwater model using the subspace ensemble randomized likelihood method (Raanes et al., 2019; Evensen, 2021). To the author's knowledge, this is the first application of this method, including structural errors, to a groundwater modelling problem.

Several limitations of the proposed methodology and the presented results should be acknowledged. First, implementing the methodology requires running the complex model multiple times to generate an ensemble of measurements and residuals. This can be computationally expensive, especially if model runtimes are long, even for a small number of realizations. Second, the statistical representation of the correlation model between measurements and residuals assumes linear relationships. If the model exhibits significant nonlinearity, the methodology may fail to capture the full characteristics of structural error, particularly with a limited number of realizations. Finally, the methodology was not tested on the highly-parameterized simple model, which could have provided insights into the effect of structural errors on predictive bias and uncertainty in these parameter settings. However, as the fit to the data improves, less information about structural errors can be extracted, potentially reducing the effectiveness of the methodology. This is left for future work.

5.5 Conclusions

In this work, two critical aspects of history matching and predictive uncertainty analysis in groundwater modelling were addressed.

First, the predictive bias of a simple model was evaluated using the 's vs s' plot approach of Doherty and Christensen (2011), comparing predictions of groundwater inflows to an open-pit operation made with a complex model to those made with a simple model calibrated against complex model outputs. Several parameterization strategies were tested for the simple model, including grid-based and zonal models, and time-varying recharge factors. It was found that the simple model, under a highly-parameterized scheme, could fit the data well but incurred greater predictive bias than the underparameterized zone-based model. This bias difference is due to the expanded capacity of parameters in the highly-parameterized model to adopt compensatory roles to fit the data, and predictive bias was increased notwithstanding the fact the data was of similar nature to the predictions.

Second, a new methodology was developed to estimate and incorporate structural errors into the history matching process of a simple model. The methodology involves using an ensemble of measurements and residuals obtained from calibrating a simple model to complex model outputs to build a linear statistical correlation model (DSI-RES model) using data space inversion in an innovative way, which is then conditioned on a specific set of measurements to generate realizations of structural error. These realizations are subsequently used as samples of structural error in the history matching process of the simple model with the aim of minimizing predictive bias and increasing predictive uncertainty. The methodology was demonstrated by history matching the simple model both with a diagonal covariance matrix of measurement error and with the ensemble matrix of structural error and comparing the predictive uncertainty ranges obtained for the two cases. It was found that predictive bias was reduced, and predictive uncertainty slightly increased, when structural errors were included in the history matching process. These are key metrics for effective groundwater modelling for decision support, as quantified uncertainty must accommodate all contributors to uncertainty, including those forthcoming from model simplification.

Chapter 6

Conclusions

6.1 Summary of findings

The following is a summary of the main findings of this research:

1. A comprehensive and mathematical unifying framework was developed to analyse the assumptions, benefits and limitations of existing methods of inverse modelling, history-matching and uncertainty quantification in groundwater modelling. Methods new to the groundwater community, such as SEnRML, and localization strategies developed and tested within the petroleum engineering community, were also included and tested in the groundwater context.

2. The value of separating calibration and uncertainty quantification was demonstrated, as it allows for assessment and updating of the prior. This can be performed using an empirical Bayesian approach, where the prior is updated using model calibration results. It was demonstrated that by implementing this approach, predictive uncertainty is not underestimated.

3. Acknowledging the uncertain nature of the subsurface, a new method was developed to accommodate nonstationary priors, that may act as a surrogate for the representation of discrete geological features that is amenable to adjustment during history-matching. The method was incorporated in history matching, confirming its potential to infer nonstationary aspects of the geological medium from the measurement dataset.

4. A method for generating realizations of structural error was developed using data space inversion, and the value of incorporating this information during history-matching was demonstrated reducing predictive bias and increasing the predictive uncertainty. Important conclusions were drawn about whether the use of many parameters to obtain a good model-to-measurement fit may, for some predictions,

be an inferior method for data assimilation than use of a parsimonious parameterization scheme accompanied by adequate representation of structural error in the inversion process.

6.2 Future work

The exploration of uncertainty in the prior and the accommodation of model structural defects in the likelihood is a complex problem that requires further research. In this work, an advance in the understanding of the problem has been made, and potential solutions have been proposed as part of new methodologies. However, several questions, some of which were partially explored during the PhD journey (not presented here), remain unanswered. The following is a list of potential future research directions:

1. Prior inference from model calibration results is a promising approach to update the prior probability distribution of the parameters in a groundwater model, before embarking on the predictive uncertainty quantification process. However, it is important to acknowledge that the prior will be always wrong not necessarily because of its misspecification, but because of the model structural defects. Research is required to evaluate how the prior must be updated to accommodate these model defects. An initial step in this direction was presented by Mathews and Vial (2017).

2. As part of this PhD it was demonstrated that the ensemble methods suffer from nonlinearity, limiting the ability to quantify predictive uncertainty. Research is needed to explore ways to overcome this limitation. Research opportunities include the development of new methods to perform ensemble filtering of the prior ensemble ahead to history-matching.

3. Also, as part of this PhD, model defects were identified using the paired simple-complex model approach. Then a statistical linear correlation model (DSI-RES model) of the structural model error was generated and used in the history-matching process. Part of this methodology can be improved by allowing the estimation of structural model error between iterations. This has been done in the past by Oliver and Alfonzo (2018); Lu and Chen (2020); the DSI-RES approach could be a potential alternative to this worth testing.

References

Alfonzo, M. and Oliver, D. S. (2019). Evaluating prior predictions of production and seismic data. *Computational Geosciences*, 23(6):1331–1347.

Alfonzo, M. and Oliver, D. S. (2020). Seismic data assimilation with an imperfect model. *Computational Geosciences*, 24(2):889–905.

Aven, T. (2010). On how to define, understand and describe risk. *Reliability Engineering and System Safety*, 95:623–631.

Ba, Y. and Oliver, D. (2023). Weighted rml using ensemble-methods for data assimilation.

Ba, Y. and Oliver, D. (2024). Importance weighting in hybrid iterative ensemble smoothers for data assimilation. *Mathematical Geosciences*, 56.

Beale, G., Milmo, P., Raynor, M., Price, M., and Donze, F. (2014). *Guidelines for Evaluating Water in Pit Slope Stability: Preparing a conceptual hydrogeological model*.

Berger, J. O. (1990). Robust bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25(3):303–328.

Beven, K. (2005). On the concept of model structural error. *Water Science and Technology*, 52(6):167–175.

Brooks, R. H. and Corey, A. T. (1966). Properties of porous media affecting fluid flow. *Journal of the Irrigation and Drainage Division*, 92(2):61–88.

Caers, J. (2018). *Bayesianism in the Geosciences*, pages 527–566. Springer International Publishing, Cham.

Capen, E. (1976). The difficulty of assessing uncertainty (includes associated papers 6422 and 6423 and 6424 and 6425). *Journal of Petroleum Technology*, 28(08):843–850.

Chada, N. K., Iglesias, M. A., Roininen, L., and Stuart, A. M. (2018). Parameterizations for ensemble kalman inversion. *Inverse Problems*, 34(5):055009.

Chen, Y. and Oliver, D. (2013). Levenberg-marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Computational Geosciences*, 17.

Chen, Y. and Oliver, D. (2017). Localization and regularization for iterative ensemble smoothers. *Computational Geosciences*, 21:1–18.

Chen, Y. and Oliver, D. S. (2012). Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, 44(1):1–26.

Clark, M. P. and Vrugt, J. A. (2006). Unraveling uncertainties in hydrologic model calibration: Addressing the problem of compensatory parameters. *Geophysical Research Letters*, 33(6).

Cooley, R. L. (2004). *A theory for modeling ground-water flow in heterogeneous media*. Number 1679. US Department of the Interior, US Geological Survey.

Cooley, R. L. and Christensen, S. (2006). Bias and uncertainty in regression-calibrated models of groundwater flow in heterogeneous media. *Advances in Water Resources*, 29(5):639–656.

Doherty, J. (2011). Modeling: Picture perfect or abstract art? *Groundwater*, 49(4):455–455.

Doherty, J. (2015). *Calibration and uncertainty analysis for complex environmental models*, volume 227. Watermark Numerical Computing Brisbane, Australia.

Doherty, J. (2023). Pest: Model-independent parameter estimation.

Doherty, J. and Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resources Research*, 47(12).

Doherty, J. and Moore, C. (2020). Decision support modeling: Data assimilation, uncertainty quantification, and strategic abstraction. *Groundwater*, 58(3):327–337.

Doherty, J. and Simmons, C. (2013). Groundwater modelling in decision support: Reflections on a unified conceptual framework. *Hydrogeology Journal*, 21.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

Egidi, L., Pauli, F., and Torelli, N. (2022). Avoiding prior–data conflict in regression models via mixture priors. *Canadian Journal of Statistics*, 50(2):491–510.

Emerick, A. and Reynolds, A. (2012). History matching time-lapse seismic data using the ensemble kalman filter with multiple data assimilations. *Computational Geosciences*, 16.

Emerick, A. and Reynolds, A. (2013). Investigation of the sampling performance of ensemble-based methods with a simple reservoir model. *Computational Geosciences*, 17.

Emerick, A. A. (2016). Towards a hierarchical parametrization to address prior uncertainty in ensemble-based data assimilation. *Computational Geosciences*, 20(1):35–47.

Evans, M. and Jang, G. H. (2011). Weak informativity and the information in one prior relative to another. *Statistical Science*, 26(3):423–439.

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162.

Evensen, G. (2018). Analysis of iterative ensemble smoothers for solving inverse problems. *Computational Geosciences*, 22(3):885–908.

Evensen, G. (2021). Formulating the history matching problem with consistent error statistics. *Computational Geosciences*, 25(3):945–970.

Evensen, G., Raanes, P. N., Stordal, A. S., and Hove, J. (2019). Efficient implementation of an iterative ensemble smoother for data assimilation and reservoir history matching. *Frontiers in Applied Mathematics and Statistics*, 5.

Evensen, G. and van Leeuwen, P. J. (2000). An ensemble kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128(6):1852 – 1867.

Evensen, G., Vossepoel, F., and van Leeuwen, P. (2022). *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem*. Springer Textbooks in Earth Sciences, Geography and Environment. Springer International Publishing.

Freeze, R. A., Massmann, J., Smith, L., Sperling, T., and James, B. (1990). Hydrogeological decision analysis: 1. a framework. *Groundwater*, 28(5):738–766.

Fuentes, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, 12(5):469–483.

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255.

Gaspari, G. and Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 – 472.

Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10).

Good, I. (2018). Probability and the weighing of evidence. *Royal Statistical Society. Journal. Series A: General*, 113(2):250–250.

Gosses, M. and Wöhling, T. (2019). Simplification error analysis for groundwater predictions with reduced order models. *Advances in Water Resources*, 125:41–56.

Gu, Y. and Oliver, D. S. (2007). An Iterative Ensemble Kalman Filter for Multiphase Fluid Flow Data Assimilation. *SPE Journal*, 12(04):438–446.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). Dram: efficient adaptive mcmc. *Statistics and Computing*, 16:339–354.

Haario, H. and Saksman, E. (1998). Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistics*, 14.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223 – 242.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Linear Methods for Regression*, pages 43–99. Springer New York, New York, NY.

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting June 6-10, 1998*. Oxford University Press.

Hoek, E. and Brown, E. (2019). The hoek–brown failure criterion and gsi – 2018 edition. *Journal of Rock Mechanics and Geotechnical Engineering*, 11(3):445–463.

Hoffmann, R., Dassargues, A., Goderniaux, P., and Hermans, T. (2019). Heterogeneity and prior uncertainty investigation using a joint heat and solute tracer experiment in alluvial sediments. *Frontiers in Earth Science*, 7.

Jiang, S., Hui, M.-H., and Durlofsky, L. (2021). Data-space inversion with a recurrent autoencoder for naturally fractured systems. *Frontiers in Applied Mathematics and Statistics*, 7:686754.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45.

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

Kitlasten, W., Moore, C. R., and Hemmings, B. (2022). Model structure and ensemble size: Implications for predictions of groundwater age. *Frontiers in Earth Science*, 10.

Knowling, M. J., White, J. T., and Moore, C. R. (2019). Role of model parameterization in risk-based decision support: An empirical exploration. *Advances in Water Resources*, 128:59–73.

Koch, K.-R. (1999). *Vector and Matrix Algebra*, pages 3–73. Springer Berlin Heidelberg, Berlin, Heidelberg.

Laloy, E. and Vrugt, J. A. (2012). High-dimensional posterior exploration of hydrologic models using multiple-try dream(zs) and high-performance computing. *Water Resources Research*, 48(1).

Langevin, C. D., Hughes, J. D., Banta, E. R., Niswonger, R. G., Panday, S., and Provost, A. M. (2017). Documentation for the modflow 6 groundwater flow model. Technical report, Reston, VA. Report.

Lima, M., Emerick, A., and Pico, C. (2020). Data-space inversion with ensemble smoother. *Computational Geosciences*, 24.

Lu, M. and Chen, Y. (2020). Improved estimation and forecasting through residual-based model error quantification. *SPE Journal*, 25(02):951–968.

Luo, X. and Bhakta, T. (2020). Automatic and adaptive localization for ensemble-based history matching. *Journal of Petroleum Science and Engineering*, 184:106559.

Luo, X., Bhakta, T., and Nævdal, G. (2018). Correlation-Based Adaptive Localization With Applications to Ensemble-Based 4D-Seismic History Matching. *SPE Journal*, 23(02):396–427.

Luo, X., Chalub, W., Zhang, X.-L., and Xiao, H. (2023). Hyper-parameter optimization for improving the performance of localization in an iterative ensemble smoother. *Geoenergy Science and Engineering*, 231:212404.

Massey, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Mathews, G. M. and Vial, J. (2017). Overcoming model simplifications when quantifying predictive uncertainty.

McCord, J. T. (1991). Application of second-type boundaries in unsaturated flow modeling. *Water Resources Research*, 27(12):3257–3260.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Mioratina, T. N. and Oliver, D. S. (2023). Quantifying prior model complexity for subsurface reservoir models. *Geoenergy Science and Engineering*, 227:211929.

Moore, C. and Doherty, J. (2005). Role of the calibration process in reducing model predictive error. *Water Resources Research*, 41(5).

Moore, C. R. and Doherty, J. (2006). The cost of uniqueness in groundwater model calibration. *Advances in Water Resources*, 29:605–623.

Oliver, D. (2022). Hybrid iterative ensemble smoother for history matching of hierarchical models.

Oliver, D. S. (1995). Moving averages for gaussian simulation in two and three dimensions. *Mathematical Geology*, 27(8):939–960.

Oliver, D. S. and Alfonzo, M. (2018). Calibration of imperfect models to biased observations. *Computational Geosciences*, 22(1):145–161.

Oliver, D. S., He, N., and Reynolds, A. C. (1996). Conditioning permeability fields to pressure data. In *5th European Conference on the Mathematics of Oil Recovery (ECMOR 5)*.

Oliver, D. S., Reynolds, A. C., and Liu, N. (2008). *Inverse Theory for Petroleum Reservoir Characterization and History Matching*. Cambridge University Press.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Panday, S. (2024). Usg-transport version 2.4.0: Transport and other enhancements to modflow-usg.

Panday, S., Langevin, C. D., Niswonger, R. G., Ibaraki, M., and Hughes, J. D. (2013). Modflow–usg version 1: An unstructured grid version of modflow for simulating groundwater flow and tightly coupled processes using a control volume finite-difference formulation. Report 6-A45, U.S. Geological Survey.

Raanes, P. N., Stordal, A. S., and Evensen, G. (2019). Revising the stochastic iterative ensemble smoother. *Nonlinear Processes in Geophysics*, 26(3):325–338.

Ranazzi, P., Luo, X., and Pinto, M. A. (2022). Improving pseudo-optimal kalman-gain localization using the random shuffle method. *Journal of Petroleum Science and Engineering*, 215.

Reichert, P. (1997). On the necessity of using imprecise probabilities for modelling environmental systems. *Water Science and Technology*, 36(5):149–156.

Robert, C. P. (2007). *Hierarchical and Empirical Bayes Extensions*, pages 457–506. Springer New York, New York, NY.

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

Rojas, R., Feyen, L., and Dassargues, A. (2009). Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modelling. *Hydrological Processes*, 23(8):1131–1146.

Scharnagl, B., Vrugt, J. A., Vereecken, H., and Herbst, M. (2011). Inverse modelling of in situ soil water dynamics: investigating the effect of different prior distributions of the soil hydraulic parameters. *Hydrology and Earth System Sciences*, 15(10):3043–3059.

Shen, S., Zeng, G., Liang, J., Li, X., Tan, Y., Li, Z., and Li, J. (2014). Markov chain monte carlo approach for parameter uncertainty quantification and its impact on groundwater mass transport modeling: Influence of prior distribution. *Environmental Engineering Science*, 31:487–495.

Shockley, E. M., Vrugt, J. A., and Lopez, C. F. (2017). PyDREAM: high-dimensional parameter inference for biological models in python. *Bioinformatics*, 34(4):695–697.

Silva Neto, G. M., Soares, R. V., Evensen, G., Davolio, A., and Schiozer, D. J. (2021). Subspace Ensemble Randomized Maximum Likelihood with Local Analysis for Time-Lapse-Seismic-Data Assimilation. *SPE Journal*, 26(02):1011–1031.

Skaggs, T. (2024). Rosetta-soil: A python library for estimating soil hydraulic properties. https://github.com/usda-ars-ussl/rosetta-soil. Version 0.1.2.

Sprenger, J. (2018). The objectivity of subjective bayesianism. *European Journal for Philosophy of Science*, 8(3):539–558.

Sun, W. and Durlofsky, L. (2017). A new data-space inversion procedure for efficient uncertainty quantification in subsurface flow problems. *Mathematical Geosciences*, 49.

Sun, W., Hui, M.-H., and Durlofsky, L. (2017). Production forecasting and uncertainty quantification for naturally fractured reservoirs using a new data-space inversion procedure. *Computational Geosciences*, 21.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics.

ter Braak, C. and Vrugt, J. (2008). Differential evolution markov chain with snooker updater and fewer chains. *Statistics and Computing*, 18(4):435–446.

Ter Braak, C. J. F. (2006). A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3):239–249.

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, 44(5):892–898.

van Leeuwen, P. J. and Evensen, G. (1996). Data assimilation and inverse methods in terms of a probabilistic formulation. *Monthly Weather Review*, 124(12):2898 – 2913.

Vrugt, J. (2016). Markov Chain Monte Carlo Simulation Using the DREAM Software Package: Theory, Concepts, and MATLAB Implementation. *Environmental Modelling & Software*, 75:273–316.

Vrugt, J. A., ter Braak, C. J., Diks, C., Robinson, B., Hyman, J., and Higdon, D. (2009). Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspce sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3):273–290.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with markov chain monte carlo simulation. *Water Resources Research*, 44(12).

White, J. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environmental Modelling and Software*, 109.

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resources Research*, 50(2):1152–1173.

Woodbury, A. D. and Ulrych, T. J. (2000). A full-bayesian approach to the groundwater inverse problem for steady state flow. *Water Resources Research*, 36(8):2081–2093.

Zhang, J., Lin, G., Li, W., Wu, L., and Zeng, L. (2018). An iterative local updating ensemble smoother for estimation and uncertainty assessment of hydrologic model parameters with multimodal distributions. *Water Resources Research*, 54:1716–1733.

Zhang, Y. and Schaap, M. G. (2017). Weighted recalibration of the rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (rosetta3). *Journal of Hydrology*, 547:39–53.