Year : 2015

Error models in hydrogeology applications

JOSSET Laureline

Faculté des géosciences et de l'environnement

Institut des sciences de la Terre

Error models in hydrogeology applications

Thèse de doctorat

Présentée à la

Faculté des géosciences et de l'environnement de

l'Université de Lausanne

par

Laureline JOSSET

Master of Science in Physics

Ecole Polytechnique de Lausanne (Switzerland)

Jury

Prof. Dr. Eric Verrecchia, Président

Prof. Dr. Klaus Holliger, Rapporteur

Prof. Dr. Ivan Lunati, Directeur de thèse

Prof. Dr. Mikhail Kanevski, Expert

Prof. Dr. Philippe Renard, Expert

Lausanne, 2015

To my extraordinary family,

and in particular to Grany, Xavier and Wilfried,

for their awe-inspiring thirst for knowledge

Contents

Résumé

Notre consommation en eau souterraine, en particulier comme eau potable ou pour l'irrigation, a considérablement augmenté au cours des années. De nombreux problèmes font alors leur apparition, allant de la prospection de nouvelles ressources à la remédiation des aquifères pollués. Indépendamment du problème hydrogéologique considéré, le principal défi reste la caractérisation des propriétés du sous-sol. Une approche stochastique est alors nécessaire afin de représenter cette incertitude en considérant de multiples scénarios géologiques et en générant un grand nombre de réalisations géostatistiques. Nous rencontrons alors la principale limitation de ces approches qui est le coût de calcul dû à la simulation des processus d'écoulements complexes pour chacune de ces réalisations.

Dans la première partie de la thèse, ce problème est investigué dans le contexte de propagation de l'incertitude, où un ensemble de réalisations est identifié comme représentant les propriétés du sous-sol. Afin de propager cette incertitude à la quantité d'intérêt tout en limitant le coût de calcul, les méthodes actuelles font appel à des modèles d'écoulement approximés. Cela permet l'identification d'un sous-ensemble de réalisations représentant la variabilité de l'ensemble initial. Le modèle complexe d'écoulement est alors évalué uniquement pour ce sous-ensemble, et, sur la base de ces réponses complexes, l'inférence est faite. Notre objectif est d'améliorer la performance de cette approche en utilisant toute l'information à disposition. Pour cela, le sous-ensemble de réponses approximées et exactes est utilisé afin de construire un modèle d'erreur, qui sert ensuite à corriger le reste des réponses approximées et prédire la réponse du modèle complexe. Cette méthode permet de maximiser l'utilisation de l'information à disposition sans augmentation perceptible du temps de calcul. La propagation de l'incertitude est alors plus précise et plus robuste.

La stratégie explorée dans le premier chapitre consiste à apprendre d'un sous-ensemble de réalisations la relation entre les modèles d'écoulement approximé et complexe. Dans la seconde partie de la thèse, cette méthodologie est formalisée mathématiquement en introduisant un modèle de régression entre les réponses fonctionnelles. Comme ce problème est mal-posé, il est nécessaire d'en réduire la dimensionnalité. Dans cette optique, l'innovation du travail présenté provient de l'utilisation de l'analyse en composantes principales fonctionnelles (ACPF), qui non seulement effectue la réduction de dimensionnalités tout en maximisant l'information retenue, mais permet aussi de diagnostiquer la qualité du modèle d'erreur dans cet espace fonctionnel. La méthodologie proposée est appliquée à un problème de pollution par une phase liquide non-aqueuse et les résultats obtenus montrent que le modèle d'erreur permet une forte réduction

du temps de calcul tout en estimant correctement l'incertitude. De plus, pour chaque réponse approximée, une prédiction de la réponse complexe est fournie par le modèle d'erreur.

Le concept de modèle d'erreur fonctionnel est donc pertinent pour la propagation de l'incertitude, mais aussi pour les problèmes d'inférence bayésienne. Les méthodes de Monte Carlo par chaîne de Markov (MCMC) sont les algorithmes les plus communément utilisés afin de générer des réalisations géostatistiques en accord avec les observations. Cependant, ces méthodes souffrent d'un taux d'acceptation très bas pour les problèmes de grande dimensionnalité, résultant en un grand nombre de simulations d'écoulement gaspillées. Une approche en deux temps, le "MCMC en deux étapes", a été introduite afin d'éviter les simulations du modèle complexe inutiles par une évaluation préliminaire de la réalisation. Dans la troisième partie de la thèse, le modèle d'écoulement approximé couplé à un modèle d'erreur sert d'évaluation préliminaire pour le "MCMC en deux étapes". Nous démontrons une augmentation du taux d'acceptation par un facteur de 1.5 à 3 en comparaison avec une implémentation classique de MCMC.

Une question reste sans réponse : comment choisir la taille de l'ensemble d'entrainement et comment identifier les réalisations permettant d'optimiser la construction du modèle d'erreur. Cela requiert une stratégie itérative afin que, à chaque nouvelle simulation d'écoulement, le modèle d'erreur soit amélioré en incorporant les nouvelles informations. Ceci est développé dans la quatrième partie de la thèse, où cette méthodologie est appliquée à un problème d'intrusion saline dans un aquifère côtier.

Abstract

Our consumption of groundwater, in particular as drinking water and for irrigation, has considerably increased over the years and groundwater is becoming an increasingly scarce and endangered resource. Nowadays, we are facing many problems ranging from water prospection to sustainable management and remediation of polluted aquifers. Independently of the hydrogeological problem, the main challenge remains dealing with the incomplete knowledge of the underground properties. Stochastic approaches have been developed to represent this uncertainty by considering multiple geological scenarios and generating a large number of realizations. The main limitation of this approach is the computational cost associated with performing complex flow simulations in each realization.

In the first part of the thesis, we explore this issue in the context of uncertainty propagation, where an ensemble of geostatistical realizations is identified as representative of the subsurface uncertainty. To propagate this lack of knowledge to the quantity of interest (e.g., the concentration of pollutant in extracted water), it is necessary to evaluate the flow response of each realization. Due to computational constraints, state-of-the-art methods make use of approximate flow simulation, to identify a subset of realizations that represents the variability of the ensemble. The complex and computationally heavy flow model is then run for this subset based on which inference is made. Our objective is to increase the performance of this approach by using all of the available information and not solely the subset of exact responses. Two error models are proposed to correct the approximate responses following a machine learning approach. For the subset identified by a classical approach (here the distance kernel method) both the approximate and the exact responses are known. This information is used to construct an error model and correct the ensemble of approximate responses to predict the "expected" responses of the exact model. The proposed methodology makes use of all the available information without perceptible additional computational costs and leads to an increase in accuracy and robustness of the uncertainty propagation.

The strategy explored in the first chapter consists in learning from a subset of realizations the relationship between proxy and exact curves. In the second part of this thesis, the strategy is formalized in a rigorous mathematical framework by defining a regression model between functions. As this problem is ill-posed, it is necessary to reduce its dimensionality. The novelty of the work comes from the use of functional principal component analysis (FPCA), which not only performs the dimensionality reduction while maximizing the retained information, but

also allows a diagnostic of the quality of the error model in the functional space. The proposed methodology is applied to a pollution problem by a non-aqueous phase-liquid. The error model allows a strong reduction of the computational cost while providing a good estimate of the uncertainty. The individual correction of the proxy response by the error model leads to an excellent prediction of the exact response, opening the door to many applications.

The concept of functional error model is useful not only in the context of uncertainty propagation, but also, and maybe even more so, to perform Bayesian inference. Monte Carlo Markov Chain (MCMC) algorithms are the most common choice to ensure that the generated realizations are sampled in accordance with the observations. However, this approach suffers from low acceptance rate in high dimensional problems, resulting in a large number of wasted flow simulations. This led to the introduction of two-stage MCMC, where the computational cost is decreased by avoiding unnecessary simulation of the exact flow thanks to a preliminary evaluation of the proposal. In the third part of the thesis, a proxy is coupled to an error model to provide an approximate response for the two-stage MCMC set-up. We demonstrate an increase in acceptance rate by a factor three with respect to one-stage MCMC results.

An open question remains: how do we choose the size of the learning set and identify the realizations to optimize the construction of the error model. This requires devising an iterative strategy to construct the error model, such that, as new flow simulations are performed, the error model is iteratively improved by incorporating the new information. This is discussed in the fourth part of the thesis, in which we apply this methodology to a problem of saline intrusion in a coastal aquifer.

Chapter 1

Introduction

1.1 Preamble

Groundwater is one of the most crucial resources constituting approximately 26% of the world's total water abstraction and 43% of all irrigation water (UNESCO, 2012). It is a privileged source of drinking water due to its excellent quality: Worldwide, it accounts for nearly half of the total drinking water. In Switzerland, that fraction grows up to 80%, half of which does not necessitate any purification (FOEN, 2009). For 75% of the African population, groundwater is the main source of drinking water.

Unfortunately, this important resource is becoming increasingly scarce and endangered due to population growth and over-exploitation. The world abstraction rate has tripled over the past 50 years (UNESCO, 2012) and around 20% of the aquifers are being over-exploited (UNESCO, 2014), which results both in a decreased availability and in the degradation of the water quality. Depletions of groundwater tables have reached a dramatic level, particularly in large urban areas such as Bangkok where a drawdown of 65m has been observed (Phien-Wej *et al.*, 2006). It has in turn caused land subsidences in several cities, such as in Tianjin (China, 1.5m), Jakarta (36cm/year), or in the dramatic case of Mexicocity (9m of subsidence) (Foster *et al.*, 1998). In addition, over-abstraction of groundwater increases the concentration of both natural and anthropogenic contaminant. In areas boarded by the sea, the excessive exploitation of coastal aquifers leads to saltwater intrusion. It has been estimated that 53 out of 126 groundwater coastal areas in Europe present an excess in salt (UNESCO, 2012).

Preserving this resource is a key issue for the future of the planet. In order to achieve a sustainable management, it is important to be able to predict the dynamics of the groundwater systems and their response to anthropogenic and environmental forcing. To this end, numerical models are developed and used for numerous and varied objectives ranging from the estimation of water supply to the design of remediation strategies.

1.2 Groundwater modelling

Flow in porous media, or more specifically groundwater flow, can be described by Darcy's law, which relates the water flux to the pressure gradient, i.e.,

$$\mathbf{u} = \frac{Q}{A} = -\frac{k}{\mu}\nabla P, \tag{1.1}$$

where $\mathbf{u}$ is the Darcy velocity, Q the volumetric flux, A the cross-sectional area of the medium, μ the viscosity of the fluid, P the pressure, and k the permeability of the medium. This equation expresses a simplified momentum balance in which acceleration is neglected. In order to describe more complex physical phenomena, the equation has to be extended. For multi-phase problems, for instance, we have a generalized Darcy's law for each phase α, i.e.,

$$\mathbf{u}_\alpha = -\frac{k_{R_\alpha}(s_\alpha)k}{\mu_\alpha}\Big(\nabla P_\alpha - \rho_\alpha \mathbf{g}\Big), \tag{1.2}$$

where k_{R_α} is the relative permeability, which is a function of the saturation s_α; μ_α, ρ_α and P_α the viscosity, the density and the pressure of the phase α, respectively; $\mathbf{g}$ the gravity acceleration. In case of single-phase flow in presence of a solute, whose concentration, c, can affect the density, we have

$$\mathbf{u} = -\frac{k}{\mu}\Big(\nabla P - \rho(c)\mathbf{g}\Big). \tag{1.3}$$

In addition to an appropriate form of the Darcy equation, subsurface models are based on mass conservation equations. Focusing on single-phase flow of an incompressible fluid, for instance, the mass balance equation is simply $\nabla\cdot\mathbf{u} = w$, where w is a source term. Substituting the Darcy velocity $\mathbf{u}$ of eq. 1.3 leads to an elliptic equation for the pressure P. The mass conservation equation of the solute mass is

$$\frac{\partial}{\partial t}(\phi c) - \nabla\cdot\Big(c\mathbf{u} - (D_m + \mathbf{D}_d)\nabla c\Big) = q \tag{1.4}$$

where ϕ is the porosity of the medium, D_m the molecular diffusion coefficient and $\mathbf{D}_d$ is the hydrodynamic dispersion tensor, and q a source term. Additional details regarding transport and two-phase flow models can be found in appendix 3.7.1.

For practical purposes, most physical processes occurring in the subsurface are fairly well described by these equations. However, major challenges arise because they take place in the underground, whose properties cannot be measured systematically.

The characterization of the permeability field of the rock formation is particularly arduous; owing to the fact that it varies in space by several orders of magnitude over short distances ($< 1m$). Moreover, small scale features can play a crucial role on the hydraulic connectivity of the geological structures and a realistic description of the medium is necessary to reproduce the observations (Kerrou *et al.*, 2008; Refsgaard *et al.*, 2012; Renard and Allard, 2013). In the recent years, many techniques have been developed to represent subsurface heterogeneity as a function of geological environment and improve the realism of the geological models. They can be classified into three categories: object-based models, which offer a parametrized representation of the main observed structures for which analog values can be used (Lantuéjoul, 2002; Huber *et al.*, 2015); pseudo-genetic models, which aim at mimicking the formation of the aquifer by simulating a crude representation of the depositional processes (Deviese, 2010); and multipoint statistics models, which generate geostatistical realizations based on a conceptual representation of the subsurface via a training image (Strebelle *et al.*, 2001; Mariethoz *et al.*, 2010; Mariethoz and Caers, 2014). Recent approaches at the cross-roads of these three categories have been developed to further improve the realism of the realizations, in particular to represent braided-river depositional system: Pirot *et al.* (2014) make the use of a multipoint statistics technique within a pseudo-genetic model.

Despite the availibility of many geophysical or hydrogeological techniques, only indirect (e.g., hydraulic heads) or integrated observations (e.g., pumping test) of the properties (typically permeability) are possible. No matter how many measurements are taken, most of the properties will remain unknown and this requires additional work to incorporate this data into the model (see section 1.3.2). Note that the permeability field is not the only source of uncertainty. Aquifer geometry, storativity, dispersivity, reactive properties, boundary conditions and many other parameters have to be assessed and they are all subject to uncertainty.

1.3 Stochastic methods

The most common approach to deal with uncertainties is to opt for a stochastic description of the problem: the physics of the phenomena is considered deterministic, whereas the model parameters (such as the permeability field) are treated as random variables to reflect the uncertainty. While practitioners may not systematically follow this approach, Monte Carlo techniques offer the simplest solution to uncertainty analysis (Kennedy and O'Hagan, 2001; Dagan, 2002; Rubin, 2004; Carrera *et al.*, 2005; Renard, 2007; Wu and Zeng, 2013).

1.3.1 Monte Carlo simulations

Monte Carlo approaches are based on the fundamental property that a sample converges to the true distribution when we iteratively increase the number of elements in the sample. In mathematical terms, if we are interested in the expected value of $g(m)$, which depends on the random variable m that follows a distribution p, then an estimation of the expectation can be obtained by repeatedly sampling p and evaluate g for the sampled values m_i, i.e.,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} g(m_i) = \int g(m) p(m) dm, \tag{1.5}$$

where $\{m_i\}_{i=1,...,n} \sim p$.

In groundwater problems, Monte Carlo methods are typically used to propagate the uncertainty on the subsurface parameters to the quantity of interest (for instance, the concentration of pollutant in drinking water). The unknown aquifer properties are assumed to follow a distribution p from which a sample can be obtained using geostatistical techniques. As we are interested in how this translates to the quantity of interest, for each realization of the sample the flow model g relating the input parameters m to the flow response is evaluated. Any summary statistic based on high dimensional integrals can then be estimated using the sampled responses (for instance, the expectation can be approximated by the sample mean as in eq. 1.5) (Yustres *et al.*, 2012).

1.3.2 Bayesian inference

The process of identifying the input parameters that match the observed data is referred to as the inverse problem (Carrera *et al.*, 2005; Zhou *et al.*, 2014). In a Bayesian framework, if $p(m)$ is the prior distribution of the input parameters m, our objective is to update this knowledge

and find the distribution of parameters that best fit the observed data d. A rigorous way to incorporate the new information is to use the Bayes rule,

$$p(m|d) = \frac{p(d|m)p(m)}{\int p(d|m)p(m)dm},$$
(1.6)

where $p(m|d)$ is the posterior distribution conditioned on d, and $p(d|m)$ the likelihood function, which measures how likely is the data d given a model m.

Except for simplistic representations, analytical formulation of the posterior from the prior is impossible as the likelihood entails complex forward simulations of the processes to compare modeled quantities to the observed data (e.g. hydraulic heads measurements, tracer tests). Most importantly, the denominator in eq. 1.6 is generally unknown. These issues are resolved by Monte Carlo Markov Chains (MCMC) techniques, the most famous strategy being the Metropolis-Hastings algorithm. Because the algorithm makes use of ratios of densities, it allows us to sample from unnormalized distributions and obtain samples from the posterior distribution (see section 4.4.2). We refer to Robert and Casella (2004) for the theoretical background and to Zhou *et al.* (2014) and reference therein for reviews of applications to groundwater problems.

The choice of prior distribution is a critical step. Rojas *et al.* (2009) have shown that, in a multi-model framework, the posterior plausibility of the various conceptual models heavily depends on the prior model probabilities. Ideally, an extremely large datasets could be acquired to eliminate the influence of the choice of prior. This is unrealistic for groundwater applications due to the large number of uncertain parameters and to the sparsity of data. A possible strategy is to elicit experts' opinion to try to identify a reliable prior. However, incorporation of soft data is not without danger as Freni and Mannina (2010) have shown in the context of storm water quality: incorporation of *weak information (mostly coming from literature or other model applications)* may lead *to wrong estimations of uncertainty in modelling results*. The preferred approach remains to choose a non-informative prior (such as uniform distribution) arguing that the incorporation of the data will dominate the posterior distribution (Kavetski *et al.*, 2006).

The likelihood function is defined by the error model selected for the analysis (Kavetski

et al., 2006). Yustres *et al.* (2012) remark that *the correct construction of the likelihood function* and *the identification of the sources of uncertainty [...] are essential to carry out a meaningful analysis.* A metric is required to quantify the discrepancy (or misfit) between simulated and observed data. The most common choice is to assume that the misfit dominates other sources of errors and that errors are Gaussian (see section 1.5).

Beside the problems related to the choice of prior and likelihood distributions, the main limitation of Monte Carlo approaches stems from the computational cost of simulating the flow responses particularly in presence of complex physical processes. While the convergence of Monte Carlo (or Monte Carlo Markov Chains) is guaranteed only asymptotically, in practice MC can be expected to reach convergence after $10^3 - 10^4$ iterations depending on the difficulty of the problem (Ballio and Guadagnini, 2004), and MCMC after 10^5 iterations (e.g. Hassan *et al.*, 2009). The evaluation of the flow solver is necessary at each iterations (i.e for each generated geostatistical realizations) resulting in an intractable computational cost.

1.4 How to limit the cost of the flow simulations

Several strategies have been developed to reduce the computational cost of flow simulations. A classical approach is to reduce the size of the problem by using upscaling methods (Wen and Gómez-Hernández, 1996; Renard and de Marsiliy, 1997; Christie, 1996; Durlofsky, 2005). Instead of describing the porous media properties at fine-scale, the properties are averaged to obtain a description at a coarser scale, hence leading to a smaller number of unknowns for which the flow and transport equations have to be solved. However, the coarsening strategy has to be devised with care to accurately model the behavior of the media. More sophisticated methods were introduced to face this issue, such as the Multi-scale Finite Volume method (MsFV, Jenny *et al.* 2003), in which a set of small local problems with fine scale resolution are coupled through a global coarse-scale problem. This approach, however, is subject to approximation errors. Two different strategies can be devised to reduce the potential bias: add an iterative correction to reduce the localization error and converge to the true solution (Lunati *et al.*, 2011) or construct error models to correct the bias (see chapter 2).

Another strategy is to opt for an approximate flow solver (or proxy) in place of a more complete physical description of the processes (see, e.g., streamline simulations, Aarnes *et al.*,

2005). If simplification are too crude, the approximate responses cannot be used directly to perform inference. In the context of uncertainty propagation, the approximate solver is most often used to identify a subset of realizations for which the complex model is run. This is the case, for instance, in dynamic ranking methods (McLennan and Deutsch, 2005), or distance kernel methods (Scheidt and Caers 2009a, more details in 2.4.2). However, as the identification of the subset of realizations is based solely on the proxy responses, the informativeness of the proxy should be assessed in order to prevent a biased subset selection if the proxy is weakly informative of the sought responses, or, on the contrary, to avoid a loss in information if the quality of the proxy is good (see chapter 2).

The last strategy consists in using surrogate models (Razavi *et al.*, 2012), where, from a learning set of realizations for which the flow responses are known, an analytical model is devised to find the mapping between the input parameters and the quantity of interest (such as the misfit or the flow response). The most common example is polynomial chaos expansion (PCE, Xiu and Karniadakis, 2002; Zeng *et al.*, 2012; Elsheikh *et al.*, 2014). An issue with surrogate models is the parameterization of the permeability field, because an analytical model using all grid cells permeability values as inputs would be under-determined. A solution is to perform a Karhunen-Loève expansion of the field ensemble and to project each realization on this basis (Elsheikh *et al.*, 2014). While this may work well for lognormal random fields, it is not appropriate for more realistic representation of the permeability. A second limitation is that surrogate models often consider the prediction of scalar quantities such as the response surface (e.g., oil in place in petroleum engineering applications, or more directly the misfit between observed and generated data), rather than time dependent quantities (e.g., oil or water extraction rates). In addition they are only appropriate when the scalar response varies smoothly in the parameter space. In a Bayesian inference context, PCE is useful to provide a first estimate of the generated realizations at virtually zero computational cost once the prediction model has been devised. Recent application can be found in Zeng *et al.* (2012) for the determination of a contaminant source in a MCMC framework and in Elsheikh *et al.* (2014) where a generalized polynomial chaos is coupled with Nested Sampling.

The previously mentioned strategies are all subjected to limitations and prone to errors: upscaling methods might smooth permeability loosing resolution to describe preferential paths or flow barriers; approximate models may be too simplistic to enable inference; and surrogate

models are appropriate only for relatively smooth problems. To obtain a reliable prediction of the uncertainty it is necessary to combine approximate models, which allows the evaluation of numerous realizations, with error models that correct the bias introduced by the approximations.

1.5 Model errors and error models

We first define the type of errors that we are considering.

1.5.1 Model errors

Many examples of model errors can be found in the literature (e.g., Kennedy and O'Hagan, 2001; Carrera *et al.*, 2005; Kavetski *et al.*, 2006; Yustres *et al.*, 2012; Zhou *et al.*, 2014). Here, we follow the terminology introduced by Christie *et al.* (2005), who identifies three types of errors: inaccuracy in the input data (e.g., permeability values), inaccuracy in the physical description (e.g., perfect tracer transport assumptions) and errors stemming from the solution strategies (e.g., numerical and discretization errors).

Another source of uncertainty is the conceptual uncertainty (Yustres *et al.*, 2012), which can be seen as arising from erroneous conceptual models. For instance, rather than the uncertainty associated with the distribution of the width of a channel, conceptual uncertainty relates to establishing whether channels are actually present. Another example would be to determine the number of facies necessary to describe the geology of the aquifer. Taking into account the conceptual uncertainty is of tremendous importance to avoid the risk of wasting generated realizations that will rarely or never match the data. The Bayesian framework offers a mean to compare the different designed scenarios $S^1, S^2, ...$ corresponding to different conceptualizations. The probability of explaining the observed phenomenon can be quantified by the evidence of each scenario,

$$Z_{S^i} = \int p(d|m_{S^i})p(m_{S^i})dm_{S^i}, \tag{1.7}$$

where the integral is taken over all the possible models following the scenario S^i, and $\frac{Z_{S^i}}{\sum_i Z_{S^j}}$ is the probability associated with the scenario S^i. The evaluation of Z_{S^i} in eq. 1.7 requires a Monte Carlo approach to compute the integrals, and many flow simulations have to be per-

formed.

The last type of errors is experimental. While not directly associated to the simulation process, no simulation can be done without comparison to actual measurements that are prone to errors. Whether the data is used as input or for model calibration, these errors have to be taken into account.

1.5.2 Review of error models

The objective of error modeling is to provide a thorough analysis of the errors and establish confidence in the predictions. Error models are crucial in balancing accuracy and computational cost. While this is widely acknowledged (Kennedy and O'Hagan, 2001; Carrera *et al.*, 2005; Kavetski *et al.*, 2006; Yustres *et al.*, 2012; Zhou *et al.*, 2014; Christie *et al.*, 2005), few studies address the question explicitly. However, inverse modelling (Bayesian or not) always requires a measure to compare simulated to observed data, and it is never assumed that the model is exempt of inaccuracies.

Error models aims at describing and quantifying the errors that are due to the modeling and data acquisition. More specifically, we refer here to the error model underlying the definition of likelihood. A complete description of the errors would require separating the different sources of error. This is arduous as an accurate evaluation of parameter error requires knowledge of perfect model and vice versa (Gaganis and Smith, 2001). Additionally, due to non-linearities of many groundwater problems, an evaluation of the joint effects of the source terms is necessary. To overcome these difficulties, it has been suggested to consider that *modelling errors can be treated as measurement errors* (see argumentation in Beven, 2006). The general trend seems to assume Gaussianity and additivity of the sources. In a MCMC set-up, the likelihood is, to our knowledge, systematically defined assuming Gaussian errors and choosing the parameters according to the quality of the measurements, hence supposing that measurement errors dominate (Zeng *et al.*, 2012). Recent investigations have shown that simulation errors may not follow a normal distribution, especially when drastic approximations of the physical processes are introduced (Köpke *et al.*, 2014). In addition, conceptual errors may be more significant than measurement errors, and yet not described by Gaussian processes (Keating *et al.*, 2010). This has been pointed out by Beven (2006), who advocates that model errors can

9

be assimilated to measurement errors *assuming that the chosen distributional form is correct.*

In recent years, error modeling has gone a step further and a correction of the error has been proposed. When CPU limitations impose a crude simplification of the model, error models are necessary to correct the bias introduced by the approximate physics or the solution errors. Hereafter, we detail four examples that illustrate the various strategies. We do not aim at providing an exhaustive review of correcting error model, but we remark that few examples can be found in the literature.

O'Sullivan *et al.* (2005) and O'Sullivan and Christie (2006a) have studied a test case in petroleum engineering, in which the objective is to calibrate the parameter of the fluid (the oil viscosity) to the production data. They highlight that the exclusive use of an upscaled model results in a biased estimation of the oil viscosity. They devised an error model by computing the fine-grid and the upscaled solutions for several values of viscosity to estimate the error. The error is then linearly interpolated to intermediate viscosity values and used to correct new upscaled simulations.

Another example in petroleum engineering is introduced by Scheidt *et al.* (2010) to account for upscaling errors during calibration of the permeability field. A distance metric is used to cluster reservoir realizations on the basis of the upscaled response. Fine-scale simulations are run for a realization subset and the observed errors are then applied to correct the approximation of each realization in the cluster. Due to the rather crude results, the error model is used only to guide the exploration of the parameter space, while the calibration is still being performed using fine-scale simulations.

Following a suggestion by Cui *et al.* (2011), Laloy *et al.* (2013) used a surrogate model (generalized Polynomial Chaos, gPC) together with an error model to reduce the computational cost of MCMC. Their error model is rather simple: the discrepancy between the approximated and exact misfit at the previous iteration is used to correct the result of the successive gPC prediction. This methodology is limited by the strong assumptions on the smoothness of the problem, both for the error model and gPC construction.

A very promising approach has been proposed by Ginsbourger *et al.* (2013), who used a proxy, together with a surrogate model (or "metamodel"), in a model calibration context. The metamodel relates the proxy response to the misfit using kriging method. While this does not allow the correction of the proxy response per se, the kriging metamodel provides an estimate of the misfit and its variance. This allows the computation of the Expected Improvement

criterion, which determines the regions of the parameter space that are of interest.

1.6 Objective

The general objective of the present study is to suggest a new framework to reduce the computational costs required by flow simulations in a stochastic context. We do not aim at devising alternatives to the methods proposed in section 1.4, which focus on limiting of the cost of the individual simulation, but rather at facilitating stochastic approach by reducing the CPU cost of the ensemble of flow simulations. Combining simplified and surrogate models with error modeling, we propose a purpose oriented approach in a machine learning context.

The thesis consists of four chapters. In chapter 2, the central concepts of the methodology are laid out: the use of the information from both the proxy and exact flow simulation is maximized to increase both accuracy and robustness of uncertainty propagation techniques. In chapter 3, this strategy is formalized in a more rigorous mathematical framework by making use of functional data analysis tools (Ramsay, 2006) to build a regression model between proxy and exact solvers as error model. In chapter 4, we present a new application of the error model in a Bayesian inference context by combining the error model with a two-stage MCMC algorithm (Efendiev *et al.*, 2005; Christen and Fox, 2005). Finally, in chapter 5, we present an attempt of devising a fully automated procedure by proposing an iterative construction of the error model. Throughout the thesis, various applications are considered as test cases. For instance, solute transport (chapter 2), a two-phase contamination (chapters 3 and 5), an history matching (chapter 4), and a saline intrusion problem (chapter 5). The thesis concludes with an overview of the potential outlooks for the proposed methodology.

Chapter 2

Local and Global Error Models to improve uncertainty quantification

Laureline Josset, Ivan Lunati

published in *Mathematical Geosciences*[1]

---

[1]L. Josset, I. Lunati (2013). Local and Global Error Models to improve uncertainty quantification. *Math. Geosci.*, 45:601–620.

2.1   Abstract

In groundwater applications, Monte Carlo methods are employed to model the uncertainty on geological parameters. However, their brute-force application becomes computationally prohibitive for highly detailed geological descriptions, complex physical processes, and a large number of realizations. The Distance Kernel Method (DKM) overcomes this issue by clustering the realizations in a multidimensional space based on the flow responses obtained by means of an approximate (computationally cheaper) model; then, the uncertainty is estimated from the exact responses that are computed only for one representative realization per cluster (the medoid). Usually, DKM is employed to reduce the space of uncertainty and decrease the size of the ensemble of realizations that are considered. We propose to use the subset of exact solutions provided by DKM to construct an error model and correct the potential bias of the approximate model (here the Multiscale Finite Volume method (MsFV) is considered). We devise two error models that employ the difference between approximate and exact medoid solutions, but differ in the way medoid errors are interpolated to correct the whole ensemble of realizations. The Local Error Model (LEM) rests upon the clustering defined by DKM and can be seen as a natural way to account for intra-cluster variability; the Global Error Model (GEM) employs a linear interpolation of all medoid errors regardless of the cluster to which the single realization belongs. These error models are evaluated for an idealized pollution problem in which the uncertainty of the breakthrough curve needs to be estimated. For this numerical test case, we demonstrate that the error models improve the uncertainty quantification provided by the DKM algorithm and are effective in correcting the bias of the estimate computed solely from the MsFV results. The framework presented here is not specific to the methods considered and can be applied to other combinations of approximate models and techniques to select a subset of realizations.

2.2   Introduction

In groundwater applications one has to deal with an incomplete characterization of the aquifer: only sparse and uncertain measurements of the properties dictating the flow response is usually available. To account for this partial information, Monte Carlo methods are employed (Dagan 2002), which treat aquifer parameters, and in particularly the permeability (or equivalently the hydraulic conductivity), as stochastic variables. Several realizations of the permeability

field, conditioned on the available data, are generated and the uncertainty is estimated from the variability of the responses obtained from different realizations. Despite the conceptual simplicity of this approach, the geostatistical representation of the uncertainty is rarely sufficient for realistically complex problems due to the large number of realizations required and the consequent prohibitive computational costs.

One possible strategy to overcome this issue is to employ approximate models that are less computationally expensive. Since in many applications large geological models are considered to describe the aquifer with high spatial resolution, one of the most effective techniques is to upscale the permeability on a coarser grid and solve reduced models. Several classical techniques exist at this end (Wen and Gómez-Hernández 1996; Renard and de Marsiliy 1997; Christie 1996; Durlofsky 2005); more modern multiscale approaches have been developed in the last decade that allow a better representation of the fine-scale details of the permeability field which are described by means of local numerical solution (Hou and Wu 1997; Arbogast 2002; Aarnes *et al.* 2005; Jenny *et al.* 2003).

The Multiscale Finite Volume (MsFV) method (Jenny *et al.* 2003) belongs to the latter group and has demonstrated great flexibility in modeling physically complex flows (Jenny *et al.* 2006; Lunati and Jenny 2006, 2007, 2008; Hajibeygi and Jenny 2009; Jenny and Lunati 2009; Künze and Lunati 2012a). The accuracy of the MsFV method has been studied in a deterministic context and evaluated in terms of the ability to mimic the solution provided by the exact model in a single realization. This has fostered the development of several iterative strategies aimed at reducing these differences, which might be large in case of particularly challenging problems (Hajibeygi *et al.* 2008; Lunati *et al.* 2011; Zhou and Tchelepi 2012; Künze and Lunati 2012a; Hajibeygi *et al.* 2012). In a stochastic context, however, a high level of accuracy might not be necessary because the goal is not to model each realization exactly, but simply to represent the variability of the ensemble of solutions (Chen and Durlofsky 2008; Chen *et al.* 2011; Aarnes and Efendiev 2008). As all methods that provide an approximate and relative inexpensive solution, the MsFV method is well suited to be applied in a stochastic context.

Another strategy to limit the computational cost of Monte Carlo approaches is to reduce the number of realizations for which the exact model is solved to estimate the uncertainty. Several methods exist to determine an optimal subset of realizations and coarsen the stochastic space. Some ranking methods classify the realizations based on static criteria such as geostatistical measures of connectivity or conductivity (McLennan and Deutsch 2005). As they do not

exploit information about the flow response, these methods are extremely efficient in terms of computational costs but have limited accuracy, which may result in a biased estimate of the uncertainty. Accuracy can be improved by using methods that sort the realizations based on a measure that depends on the flow response, such as in dynamic ranking methods (Ballin *et al.* 1992) or in the Distance Kernel Method (DKM) (Scheidt and Caers 2009a,b). While those approaches lead to much better results as they can be tailored to the question of interest, the problem remains of being able to inexpensively compute the dynamic measure.

In this paper, the MsFV method and the DKM are combined. However, rather than simply employing the MsFV method as approximate model to compute the dynamic measure in the DKM, the approximate MsFV solutions are used to obtain a first estimate of the uncertainty. The DKM selects a subset of realizations for which the exact model is solved; then, an error model to correct the potential bias of the MsFV estimate is constructed from the difference between the exact and the approximate solutions, which are available for the subset. Here, the ranking technique is used not solely to reduce the number of flow simulations, but rather to provide a representative subset of exact solutions to be compared to the approximate solutions. Note that whereas ranking techniques, or methods like DKM, make in general no direct use of the dynamic measure, in our approach this information is further exploited to construct an error model with negligible extra costs.

The paper is organized as follows: after a brief problem statement, we review the MsFV method and the DKM; then we present two error models that are devised by combining MsFV and DKM; finally, we present a thorough evaluation of the error models for a numerical test case that is representative of fluvio-glacial aquifers. The paper ends with some concluding remarks and perspectives for future development.

2.3   Problem statement

Here we consider the problem of predicting the breakthrough curve of a contaminant, which behaves as an ideal tracer (i.e., it does not alter the density and the viscosity of the fluid). The evolution of the contaminant concentration in the aquifer, c, is described by the following system of equations:

$$\nabla \cdot (K \nabla h) = 0 \tag{2.1}$$

$$\phi \frac{\partial c}{\partial t} + \nabla \cdot (c\boldsymbol{u} - \mathbf{D} \nabla c) = 0 \tag{2.2}$$

where

$$\boldsymbol{u} = -K\nabla h, \tag{2.3}$$

is the Darcy velocity; K the hydraulic conductivity (which is obtained dividing the permeability by the water viscosity); ϕ the porosity; and $\mathbf{D}$ the hydromechanical-dispersion tensor, which includes the effects of molecular diffusion and dispersion. When appropriate boundary and initial conditions are assigned, the system above can be solved and the breakthrough curve at the location of interest can be computed as a function of time, $C(t)$.

The solution strongly depends on the structure of permeability and porosity fields (Lunati and Kinzelbach 2004), which are usually not fully characterized on the basis of experimental observations. To model the uncertainty on these parameters, N_r realizations are generated, $\{K_i, \phi_i\}_{i=1,2,\ldots,N_r}$, which represent the variability of the properties due to the limited characterization of the aquifer. To evaluate the propagation of this uncertainty to the quantity of interest, flow and transport problems are solved in each realization and the breakthrough curve is computed, $C_i(t)$. (Here, initial and boundary conditions are treated as deterministic variables). The set of curves, $\{C_i(t)\}_{i=1,2,\ldots,N_r}$, obtained by these procedures, allows a characterization of the uncertainty on the breakthrough curve conditioned to the set of realizations that have been generated. In the following we are concerned with the problem of reducing the computational cost of these procedures, which can become prohibitive in presence of many geological realizations containing a large number of cells and involving complex physical processes.

2.4 Methodology

There are two natural strategies to overcome this issue: one is to use an approximate model that reduces the cost of computing a set of (approximate) curves $\{C_i^a(t)\}_{i=1,2,\ldots,N_r}$; the other is to reduce the dimensionality of the stochastic space and consider only a subset of $N_s < N_r$ realizations with breakthrough curves, $\{C_i(t)\}_{i=1,2,\ldots,N_s}$. Both strategies, however, might lead to biased predictions of the uncertainty.

The main idea of the present work is that the bias can be reduced by a combination of these two approaches. In the DKM, for instance, approximate models are used only to select the subset of realizations, $\{K_i, \phi_i\}_{i=1,2,\ldots,N_s}$, on the basis of their flow response. However,

these approximate solutions can be used to estimate the variability neglected by the subset selection. On the other hand, the exact-model responses calculated for the selected realizations can be used to construct an error model and reduce the bias of the uncertainty estimated by the approximate model. In this paper we are precisely concerned with the problem of devising a methodology which allows an optimal exploitation of the information contained in the two sets of curves, that is $\{C_i^a(t)\}_{i=1,2,\dots,N_r}$ and $\{C_i(t)\}_{i=1,2,\dots,N_s}$.

2.4.1   The Multiscale Finite Volume (MsFV) method

The approximate model employed in this study is the MsFV method, which has been devised to efficiently solve the flow problem, Eq. (2.1), and deliver an approximate but fully conservative velocity field that can be used in the transport equation without introducing mass-balance errors (Jenny *et al.* 2003; Lunati and Jenny 2006). Although extensions of the MsFV method have been proposed in the past to solve the transport problem (Lee *et al.* 2009; Künze and Lunati 2012a), here the MsFV method is employed only to solve the flow problem, whereas the transport problem is solved exactly.

We use the operator formulations employed in Lunati and Lee (2009) to briefly present the MsFV method. First, we introduce the discrete form of Eq. (2.1)

$$\mathbf{Ah} = \mathbf{r}, \tag{2.4}$$

where $\mathbf{h}$ is the vector of the unknown hydraulic heads; $\mathbf{A}$ is the coefficient matrix, which depends on the hydraulic conductivity K; and $\mathbf{r}$ is the vector containing the information about the boundary conditions. In addition to the fine-scale grid introduced to define Eq. (2.4), the MsFV method employs two auxiliary coarse grids: a (primary) coarse grid and the corresponding dual (coarse) grid, which are represented in Fig. 2.1.

The main idea of the MsFV method is to approximate the hydraulic head by means of a set of interpolators, which are local numerical solutions computed on the cells of the dual grid, that is

$$\mathbf{h} \approx \mathbf{h}^{ms} = \mathbf{Bh}_n + \mathbf{Cr}, \tag{2.5}$$

where $\mathbf{B}$ is the basis-function operator, whose columns interpolate the hydraulic head, $\mathbf{h}_n$, at the node of the dual grid (which are at the centers of the coarse grid, see Fig. 2.1) to the fine-scale grid; $\mathbf{C}$ is the correction function operator, which accounts for the local effects of $\mathbf{r}$
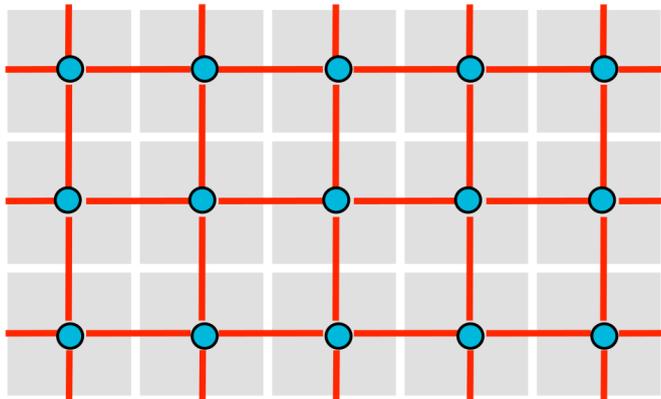
Figure 2.1: Representation of the auxiliary coarse grids used in the MsFV method. The dual (coarse) grid (red lines) is used to construct a set of local interpolators, which are local numerical solutions, whereas the cells of the (primary) coarse grid (white lines) serve as control volumes to build a coarse problem that defines the coarse-scale unknown at the nodes of the dual grid, or centers of the coarse grid (blue circles). Once the coarse solution is obtained, the interpolator can be used to obtain an approximate fine-scale solution.

and can be regarded as a source-term interpolator. In the MsFV method errors are introduced by the localization assumptions that are required to assign the boundary conditions of the local problems and compute basis and correction functions. Depending on flow conditions and on medium heterogeneity, localization might prevent a faithful description of long-correlation structures as channels or flow barriers (Lunati and Jenny 2004, 2007; Lunati *et al.* 2011).

The node hydraulic head, $\mathbf{h}_n$, is solution of the coarse equation

$$\mathbf{M}_{nn}\mathbf{h}_n = (\boldsymbol{\chi}\mathbf{AB})\mathbf{h}_n = \boldsymbol{\chi}(\mathbf{I} - \mathbf{AC})\mathbf{r}, \tag{2.6}$$

which is obtained by imposing the mass balance on the cells of the coarse grid (which serve as control volumes), that is by applying to $\mathbf{Ah}^{ms} = \mathbf{r}$ the summation operator, $\boldsymbol{\chi}$, which sums up all fine-cell values belonging to the same coarse cell and is the discrete analogous of control-volume integration. The computational advantage of the MsFV method stems from the fact that a large problem, Eq. (2.4), is split into a set of small local problems (which are solved to construct $\mathbf{B}$ and $\mathbf{C}$), and a coarse problem, Eq. (2.6), whose coefficient matrix, $\mathbf{M}_{nn} = \boldsymbol{\chi}\mathbf{AB}$, is smaller than the original matrix, $\mathbf{A}$.

Once the approximate pressure solution, $\mathbf{h}^{ms}$, is obtained, a fine-scale conservative velocity field is constructed by solving a second set of local problems on the cells of the coarse grid and used in the transport equation. We refer to the existing literature for further details on the

MsFV method (Lunati and Lee 2009 and references therein). Here, we simply remark that this framework offers great flexibility to implement several adaptive strategies: the MsFV method can be seen as a numerical upscaling procedure, if the fine-scale velocity is not reconstructed and the transport is solved on the coarse grid (Lee *et al.* 2009; Künze and Lunati 2012a); as an iterative linear solver, if a procedure is introduced to iteratively correct the boundary conditions of the localized problems (Hajibeygi *et al.* 2008; Lunati *et al.* 2011; Zhou and Tchelepi 2012); or as a downscaling method, if the original grid is taken as the coarse grid (Künze and Lunati 2012a). Here, we use the MsFV method (with construction of a conservative velocity) as approximate model to compute a velocity field in each geostatistical realization; then, the MsFV approximate velocity is used in the transport equation, Eq. (2.2), to obtain a set of approximate breakthrough curves $\{C_i^{ms}(t)\}_{i=1,2,...,N_r}$, which can be used to estimate the uncertainty.

2.4.2 Distance Kernel Methods (DKM)

DKM (Scheidt and Caers 2009a,b) is an alternative to traditional ranking techniques to select a subset of realizations that preserves the uncertainty spread of the sample. Dynamic ranking techniques (Ballin *et al.* 1992) sort realizations based on the responses of an approximate model and solve the exact model only for a subset of realizations that correspond to the desired quantiles. DKM, instead, employs the approximate information to quantify similarities between geostatistical models and selects a subset aiming at reproducing the same statistics as the full set of realizations. The first step is to compute a distance matrix $\mathbf{d}$ (a square matrix of size $N_r \times N_r$), which measures dissimilarity between realizations from the approximate flow responses. Here, the distance between two realizations, i and j, is defined as the l_2-distance between their breakthrough curves

$$d_{ij} = \sqrt{\sum_{t=1}^{n_t}[C_i^{ms}(t) - C_j^{ms}(t)]^2} \tag{2.7}$$

where $C_i^{ms}(t)$ is the curve obtained using MsFV as approximate model, and the sum is taken over all n_t discrete times at which the concentration is recorded (in our case the n_t time steps of the simulation). Eq. (2.7) naturally defines a multidimensional space, $\mathcal{S}$, where each realization is represented by a point and the distance between points is proportional to their dissimilarity in term of breakthrough response. It is natural to attempt to coarsen the space of

uncertainty by grouping the realizations into N_s clusters based on their distances and assume that each cluster, Γ_k, can be represented by a representative realization (e.g., the medoid) weighted by the number of realizations in the cluster, N_{Γ_k}.

In DKM the clustering is not applied directly in the original multidimensional space, $\mathcal{S}$, but a kernel expansion is used to project the points onto a new space (the feature space $\mathcal{F}$) in the attempt to linearize the space of uncertainty. Although the expansion is associated with a kernel function of the form $\kappa[C_i^{ms}(t), C_j^{ms}(t)] = \langle \varphi[C_i^{ms}(t)], \varphi[C_j^{ms}(t)] \rangle$, where φ is the mapping function from $\mathcal{S}$ to $\mathcal{F}$, the distance matrix in the feature space, $\mathbf{d}^{\mathcal{F}}$, can be computed without an explicit definition of φ by using only the scalar product computed by κ. Then the distance in the feature space is written as

$$d_{ij}^{\mathcal{F}} = \sqrt{K_{ii} + K_{jj} - 2K_{ij}} \tag{2.8}$$

where $\mathbf{K}$ is the kernel matrix associated to the kernel function. Among the many possible choices of the kernel matrix, we use a standard gaussian kernel of the form

$$K_{ij} = \exp\left\{\frac{-d_{ij}^2}{2\sigma^2}\right\} \tag{2.9}$$

where σ is the kernel width parameter.

Based on $\mathbf{d}^{\mathcal{F}}$, a k-medoid clustering algorithm (Hastie *et al.* 2009) is applied to find the many-to-one mapping, f, that assigns each curve, $C_i^{ms}(t)$, to a cluster (i.e., $f(i) = k$ if $C_i^{ms}(t) \in \Gamma_k$). The mapping corresponds to an optimization procedure, which finds

$$f = \arg\min_f \sum_{i,j:f(i)=f(j)} \mathbf{d}_{ij}^{\mathcal{F}} \tag{2.10}$$

and minimizes the average intra-cluster distances. In parallel to the definition of clusters, the algorithm identifies the medoids as the realizations that satisfies

$$i_k = \arg\min_{i:f(i)=k} \sum_{j:f(j)=k} \mathbf{d}_{ij}^{\mathcal{F}}. \tag{2.11}$$

The main advantage of k-medoids over k-means is that it does not require to explicitly compute points in the feature space and employs only the distance matrix in that space (Hastie *et al.*

2009). Moreover, k-medoids is not limited to Euclidean distances as k-means. This gives some freedom in defining the choice of the dissimilarity measure, which can be adapted to the question of interest.

The medoids define a subset of realizations, $\{K_{i_k}, \phi_{i_k}\}_{k=1,2,\ldots,N_s}$, for which the exact flow model is solved and a subset of exact curves $\{C_{i_k}(t)\}_{k=1,2,\ldots,N_s}$ is obtained. Classical DKM uses solely $\{C_i(t)\}_{i=1,2,\ldots,N_s}$ to compute experimental quantiles (Scheidt and Caers 2009a,b, 2010). This is done by assuming that all the realizations behave as the medoid realization, which leads to compute the experimental quantiles by weighting the medoid curves by the number of realizations in their cluster (or in other words, by considering a multiset of medoid curves, each having multiplicity equal to the number of cluster elements).

2.4.3 Error models

With the techniques described above, two sets of curves can be used to estimate the uncertainty of the predicted breakthrough curve that is $\{C_i^{ms}(t)\}_{i=1,2,\ldots,N_r}$ and $\{C_{i_k}(t)\}_{k=1,2,\ldots,N_s}$. In both cases, a sample of N_r realizations

$$\{C_i^*(t)\}_{i=1,2,\ldots,N_r}, \tag{2.12}$$

is used to compute experimental quantiles. If one choose to use only the approximate curves

$$\text{MsFV} \quad : \qquad\qquad C_i^*(t) = C_i^{ms}(t), \tag{2.13}$$

the MsFV uncertainty estimation is obtained. Employing the standard DKM is equivalent to choose

$$\text{DKM} \quad : \qquad\qquad C_i^*(t) = C_{i_k}(t), \quad \text{with} \quad k = f(i), \tag{2.14}$$

which construct a multiset where each medoid has multiplicity equal to the number of realizations in its cluster.

When the DKM is employed, information from approximate and exact responses is available and can be combined to improve uncertainty quantification at almost zero additional costs. On one hand, the information contained in the approximate curves can be used to estimate the intra-cluster variability, which is completely neglected by Eq. (2.14): the variability of cluster can be represented by the differences between each approximate curve and the approximate curve of its medoid, $C_i^{ms}(t) - C_{i_k}^{ms}(t)$. On the other hand, the exact curves of the medoids can
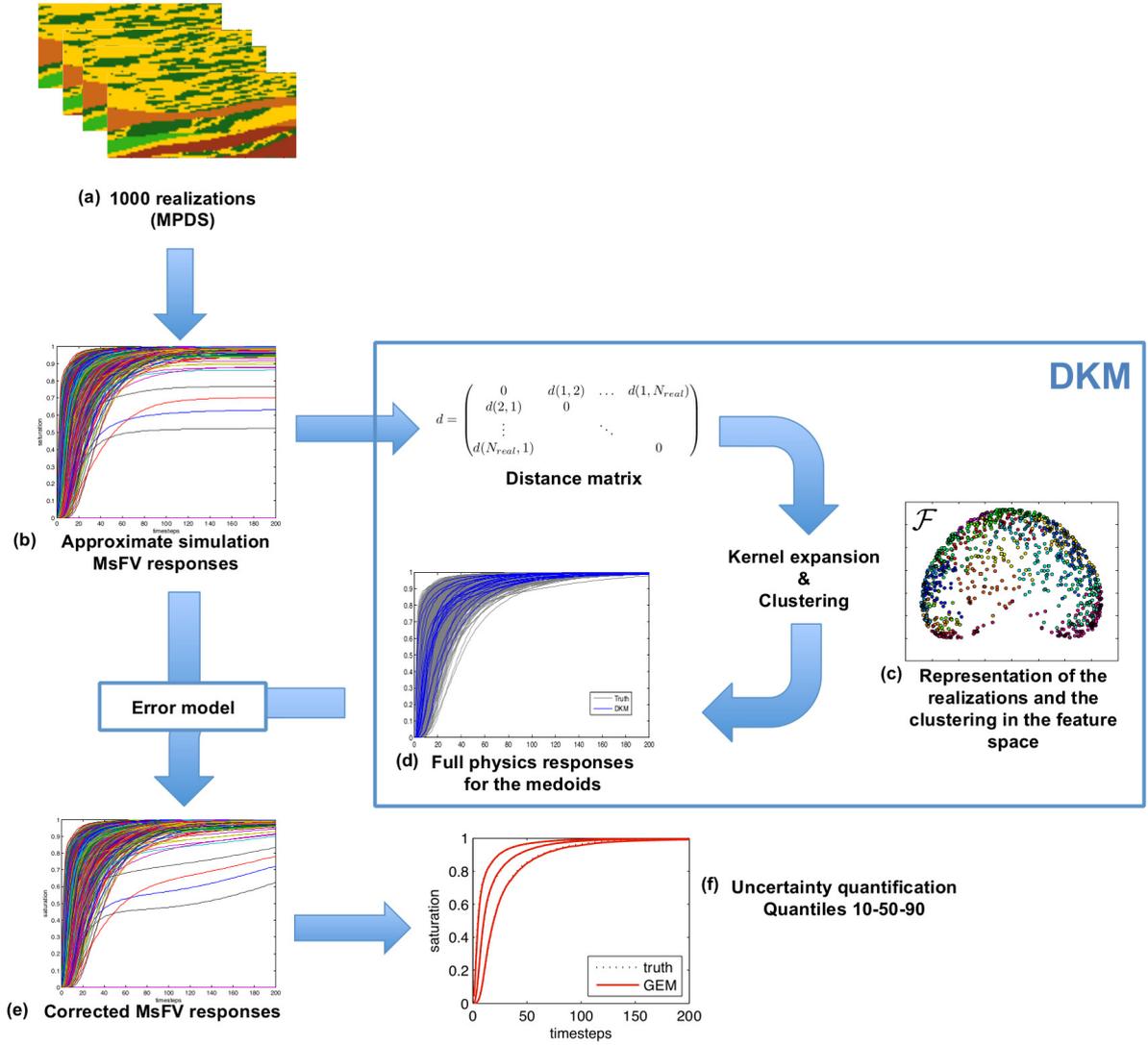
Figure 2.2: Starting from an set of geostatistical realizations $\{K_i, \phi_i\}$, MsFV simulations are run to compute a set of breakthrough curves, $\{C_i^{ms}(t)\}$. To select a subset of realizations, the euclidean distance between the curves, Eq. (2.7), is interpreted as a measure of dissimilarity. After the distance matrix $\mathbf{d}$ (Eq. 2.7) is constructed, a kernel method is used to compute a new distance in a feature space, $\mathbf{d}^{\mathcal{F}}$, Eq. (2.8). Based on $\mathbf{d}^{\mathcal{F}}$ the k-medoid algorithm, Eqs. (2.10) and (2.11), is used to cluster the realizations and finds a representative realization for each cluster (the medoid). After exact breakthrough curves are obtained for the medoids, the error model is constructed and generates the corrected curves, $\{C_i^*(t)\}$, which are used to compute the experimental quantiles.

be used to construct an error model aimed at reducing potential biases of the MsFV estimate: the difference between the exact and the approximate curves of the medoids, $C_{i_k}(t) - C_{i_k}^{ms}(t)$, can be used to correct all the curves in the cluster. These conceptually different approaches

lead to exactly the same corrected curves

$$C_i^*(t) = C_{i_k}(t) + [C_i^{ms}(t) - C_{i_k}^{ms}(t)] = C_i^{ms}(t) + [C_{i_k}(t) - C_{i_k}^{ms}(t)] \tag{2.15}$$

with $k = f(i)$.

An error model of this form has been proposed in Scheidt *et al.* (2011) to estimate an upscaling error that is assumed to be the same for all realizations in the same cluster. In Scheidt *et al.* (2011), however, the corrected curves are used to generate realizations constrained to dynamic data. Notice that, if applied directly, Eq. 2.15 might lead to corrected curves that are unphysical and not constrained between zero and one. This is a severe limitation if the corrected curves are used to obtain an estimate of the uncertainty. To avoid this problem, the breakthrough curves are not corrected directly: first a logistic transformation is applied to all curves, $\hat{C}_i = \text{logit}^{-1}(C_i)$; then the transformed curves are corrected, $\hat{C}_i^*$; and finally, the corrected curves are transformed back via logit transformation, $C_i^* = \text{logit}(\hat{C}_i^*)$. This yields the Local Error Model (LEM)

$$\text{LEM:} \qquad C_i^*(t) = C_i^{lem}(t) = \text{logit}\left\{\hat{C}_i^{ms}(t) + [\hat{C}_{i_k}(t) - \hat{C}_{i_k}^{ms}(t)]\right\}, \tag{2.16}$$

which delivers corrected curves that lay between zero and one.

The error model above, which considers only intra-cluster information, can be readily extended by considering a set of linear combinations of corrected curves

$$C_i^*(t) = \sum_k^{N_s} \beta_{ik}\{C_i^{ms}(t) + [C_{i_k}(t) - C_{i_k}^{ms}(t)]\}, \tag{2.17}$$

where the weights, β_{ik}, might be chosen to enforce that the corrected curves have some desired characteristics (e.g., that they are constrained between zero and one, or that they are monotonic). Although the choice of the weighting function might be critical, here we chose a simple weighting function that depends exclusively on the distance in the feature space

$$\beta_{ik} = \frac{\exp(-d_{ii_k}^{\mathcal{F}})}{\sum_k^{N_s} \exp(-d_{ii_k}^{\mathcal{F}})}. \tag{2.18}$$

The underlying assumption is that realizations that are closer in the feature space have more similar errors. As for the LEM, to guarantee concentration values constrained between zero and one the logistic transformation used before applying the GEM and the corrected transformed

curves are then transformed back via a logit transformation. This yields the Global Error Model (GEM)

$$\text{GEM:} \qquad C_i^*(t) = C_i^{gem}(t) = \text{logit} \left\{ \hat{C}_i^{ms}(t) + \sum_k^{N_s} \beta_{ik}[\hat{C}_{i_k}(t) - \hat{C}_{i_k}^{ms}(t)] \right\}, \qquad (2.19)$$

where it is assumed that $\sum_k \beta_{ik} = 1$, and observed that $C_i^{ms}(t)$ is independent of k.

Eq. 2.19 can be interpreted as an error model for the MsFV method. The exact curves computed for the N_s medoids are compared with the approximate curves of the medoids, and their difference is used to correct the approximate solution for each realizations i. Note that for an arbitrary weight, β_{ik}, all the medoid differences are used to correct each approximate curve. If $\beta_{ik} = \delta_{i,f(i)}$, the GEM reduces to the LEM and only intra-cluster information is used. If the constraint $\sum_k \beta_{ik} = 1$ is relaxed, the MsFV estimate of the uncertainty can be obtained by choosing $\beta_{ik} = 0$. A flowchart of the uncertainty analysis proposed here (which combines MsFV, DKM, and an error model) is presented in Fig. 2.2.

2.5 Numerical results

2.5.1 An idealized pollution problem

The methodology described above is applied to an idealized pollution problem in which the breakthrough curve of a contaminant has to be predicted. We consider a two-dimensional section of a confined aquifer of length 10.8 m and depth 5.1 m. The conductivity field, K, is inspired by the geology of a sedimentary aquifer, typical of braided river deposits. A vertical section acquired at the Herten site (Germany) (Bayer *et al.* 2011) is used as an input training image in the Direct Sampling method (MPDS) (Mariethoz *et al.* 2010) to perform multiple point geostatistical simulations and generate 1000 synthetic realizations. The 10 facies of the original data (Bayer *et al.* 2011) are reduced to 5 facies by grouping similar lithofacies. The porosity and of hydraulic conductivity values are reported in Fig. 2.3, together with the facies distribution of four realizations and the corresponding breakthrough curves.

No-flow conditions are applied at the upper and lower boundary of the domain, whereas two types of boundary conditions are considered for the left and right boundaries: prescribed incoming flux (BCF), or prescribed hydraulic-head difference (BCH). The contaminant is released at the left boundary with normalized concentration $c = 1$, and the breakthrough curves are computed by averaging the concentration of the outcoming fluxes at the right bound-
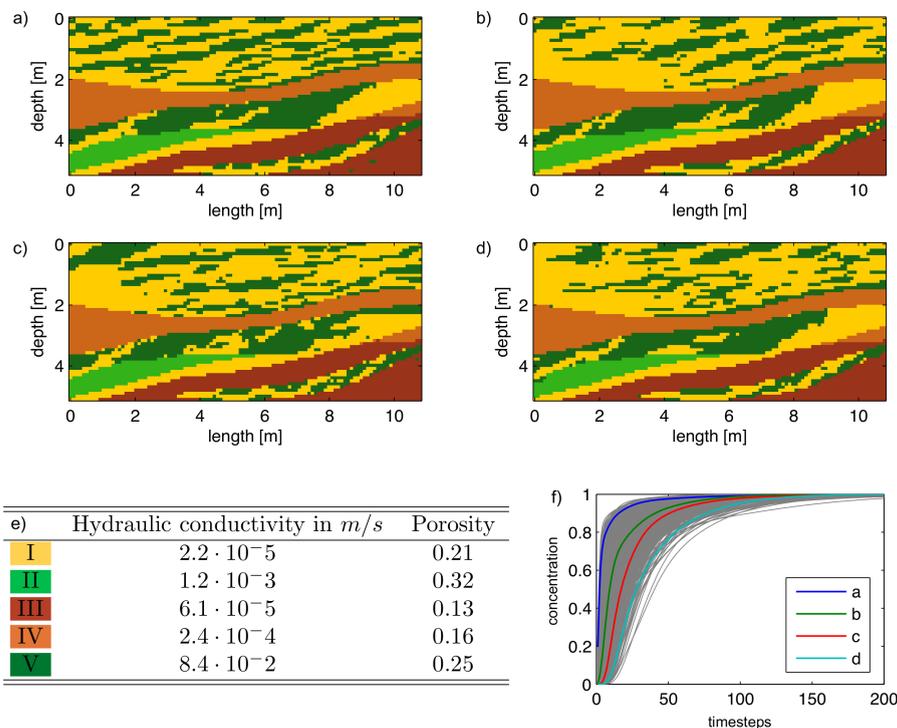
Figure 2.3: (a-d) Examples of stochastic fields generated by DS (Mariethoz *et al.* 2010); (e) Table of the hydraulic conductivity and the porosity of the 5 lithofacies; (f) Breakthrough curves of the whole set of realization (grey) and of the four fields depicted in a, b, c and d (colors).

ary. In accordance with realistic natural gradient conditions simulations in which contaminant transport is dominated by advection (Péclet number Pe > 50) are run.

2.5.2 Application of the methodology

In this section, the methodology outlined in Fig. 2.2 is applied to the idealized pollution problem. Simulations with the exact model are performed on the full set of realizations and the variability of the responses, $\{C_i(t)\}_{i=1,2,...,N_r}$ (Fig. 2.4(a)), is taken has the reference uncertainty to evaluate the performance of the error models. Estimates provided by MsFV and DKM are also computed to illustrate the improvement achieved by LEM and GEM.

Experimental quantiles are calculated based on the approximate breakthrough curves, $\{C_i^{ms}(t)\}_{i=1,2,...,N_r}$ (Fig. 2.4(b)), and provide the MsFV estimate of the uncertainty. Then, a distance matrix is constructed using MsFV curves and DKM is applied to identified N_{Γ_k} clusters and select a subset of realizations. The number of clusters should be sufficient to capture the error and estimate the desired quantiles, but not too large in order to limit the
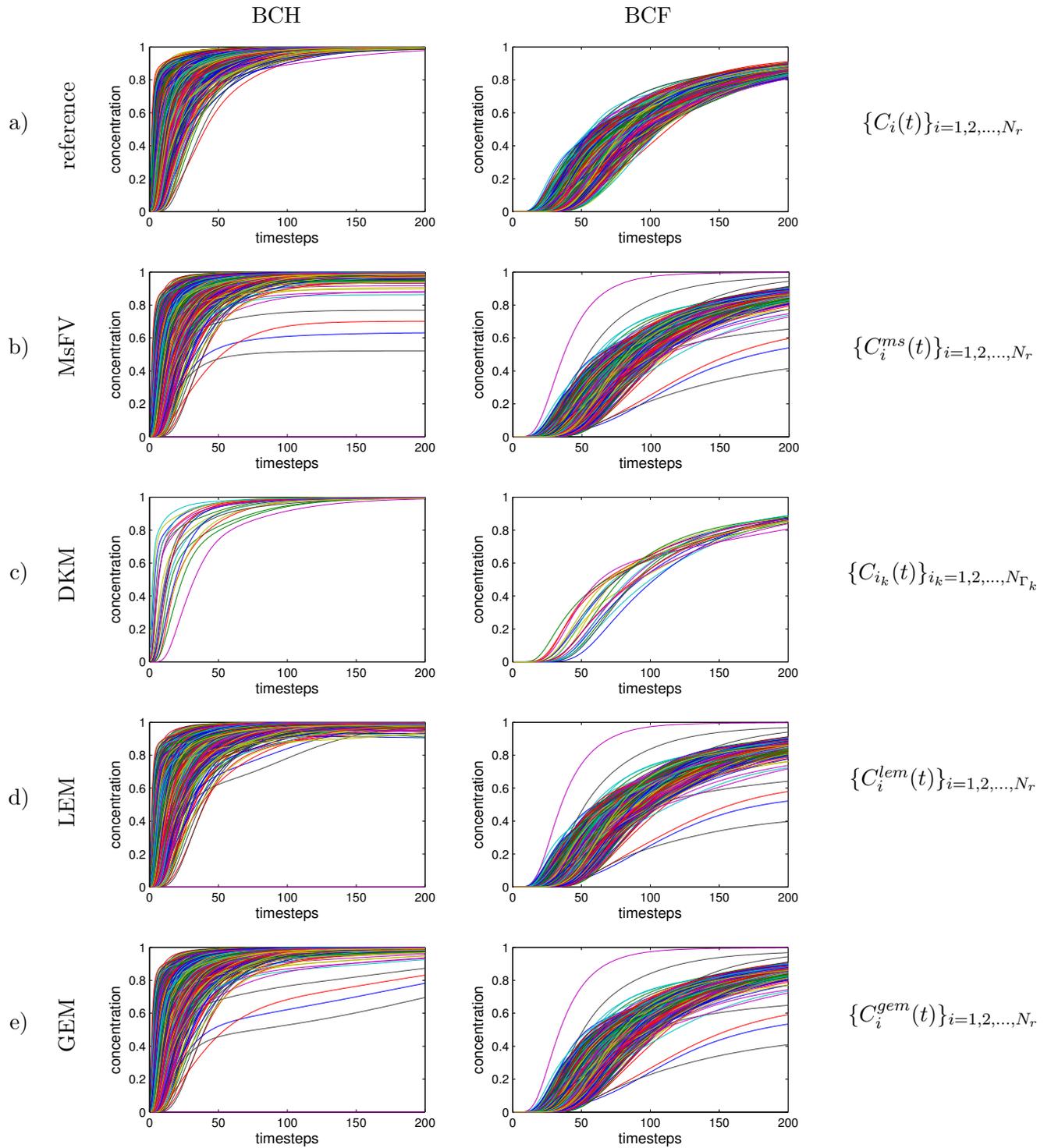
Figure 2.4: Ensemble of the breakthrough curves corresponding to each model, for the two types of boundary conditions: BCH (left), and BCF (right).

computational costs. Although a procedure could be devised to identify an optimal number, here we simply set $N_{\Gamma_k} = 20$, which corresponds to a coarsening factor of 50 for the uncertainty

space and allows computing the 10th and 90th percentiles (P10 and P90, respectively) by the DKM. The identification of the subset is performed in the feature space using a Gaussian-kernel expansion. After a sensitivity analysis, the width parameter is set equal to the standard deviation of the distance matrix, which 0.55 and 0.98 for BCF and BCH, respectively. The clustering is performed only on the base of the kernel matrix and does not require constructing the feature space explicitly. The k-medoids algorithm is used to identify N_{Γ_k} medoids for which the exact responses are computed, $\{C_{i_k}(t)\}_{k=1,2,\ldots,N_{\Gamma_k}}$.

A two-dimensional representation of the clustering in the feature space is shown in Fig. 2.5. The realizations seem continuously distributed rather than arranged in well separated clusters. Although this might be partially due to the two-dimensional visualization of the feature space, the fact that clusters are not well defined is confirmed by the instability of the clustering algorithm: different initializations of the algorithm (which require an initial guess on the N_{Γ_k} medoids) lead to different cluster repartitions and different uncertainty predictions, independently of the kernel width choice. A set of exact breakthrough curves obtained for one of the cluster repartitions, $\{C_{i_k}(t)\}_{k=1,2,\ldots,N_{\Gamma_k}}$, is shown in Fig. 2.4(c).

The approximate curves for the entire set of realizations and for the medoid exact response are then used to construct the error model. Here our approach differs from the standard DKM, which estimates the quantiles based exclusively on the subset of exact curves and does not make any direct use of the set of approximate curves. In contrast, we use the differences between the approximate and exact medoid responses to correct the entire set of approximate curves, which is then used to estimate the quantiles. In the LEM the responses are corrected using only local (intracluster) information and the set of curves $\{C_i^{lem}(t)\}_{i=1,2,\ldots,N_r}$ (Eq. (2.16), Fig. 2.4(d)) is used to compute the quantile. In the GEM the responses are corrected globally, regardless to the cluster to which they belong, and the set of curves $\{C_i^{gem}(t)\}_{i=1,2,\ldots,N_r}$ (Eq. (2.19), Fig. 2.4(e)) is obtained. Notice that few outliers are not effectively corrected due to the limited coverage of the extreme regions by the set of medoids. As it will be seen in the next section, this few outliers do not sensitively affect the estimate of P10, P50, and P90. However, in cases where uncertainty on extremes needs to be quantified, a different strategy has to be used to identify the subset of realizations used to construct the error models and extreme regions have to be more densely sampled. Note that due to the non-clear repartitions of the realizations into well defined clusters, this global model is more consistent with the data and it is expected to lead to more stable uncertainty estimations in terms of dependency on the initial medoids guess.
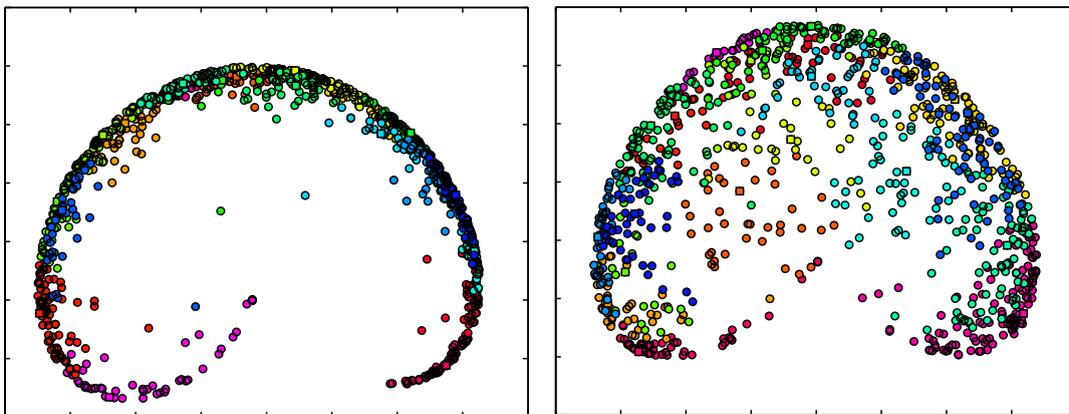
Figure 2.5: A two-dimensional representation of the feature space for BCH on the left and BCF on the right. Each dot is a realization, and each color represent a cluster. The realizations represented by a square are the medoids defined by the clustering algorithm, for which exact simulation are run. Note that this representation is obtained by the Multidimensional Scaling (Borg and Groenen 2005; Cox and Cox 2008; Scheidt and Caers 2009a), which is used here for visualization purposes only.

2.5.3 Comparison of quantile-curve estimates

In general, the characterization of uncertainty is done on the basis of a limited number of experimental quantiles; here we consider the 10th, 50th and 90th percentiles (P10, P50, and P90, respectively). Figs. 2.8 and 2.9 compare the three quantile curves obtained with the four models (MsFV, DKM, LEM and GEM) with the reference quantile curves for both sets of boundary conditions. Notice that due to the instability of the DKM algorithm, which depends on the initial guess on the medoids, very different quantile curves can be obtained with DKM, LEM, and GEM. Here we present the comparison for an initialization with yield an average performance, whereas the variability in model response due to the stability of DKM is investigated in the next section.

For BCF, MsFV provides a good measure of the statistical variability but tends to slightly underestimate contaminant concentration of about 4.5% at early times (note that the concentration will be overestimated at later time due to the constraint that the approximate MsFV solution is conservative and therefore the mean arrival time of the contaminant must be exact with this type of boundary conditions). DKM leads to curves that are less smooth due to the reduction of statistical space, which deteriorates the estimates of quantile curves; the average maximum fluctuations are of the order of 3%. LEM provides smoother curves than DKM, whereas GEM gives an excellent estimation of the uncertainty (average maximum fluctuations

LEM and GEM are of 1.8% and 1.5% respectively). MsFV bias is effectively corrected and the uncertainty is correctly represented by the MsFV approximate curves.

For BCH, MsFV quantile curves are in good agreement with the reference (maximum difference between the curves is of 5.2%). DKM estimate is less good and the 20 exact medoid responses provide a worst uncertainty estimate than the set of approximate responses (fluctuations of 6.1%). This shows that in some case DKM can lead to a deteriorated prediction of approximate solutions on which it is based. LEM also smooths the DKM estimation for this set of boundary conditions, but P10 and P90 remain underestimate (averaged maximum fluctuations of 4.3%); GEM leads again to an excellent estimate (3.3%).

2.5.4 Quantification of the quality of the estimate and stability

To illustrate the dependence of clustering on the initialization, DKM, LEM, and GEM are applied 500 times time with a different initial guess of the medoid set (seed). The overall quality of the different models is evaluated by considering the l_2-norm of the quantile error

$$l_2: \qquad \sqrt{\sum_t \left(P_T(t) - P_E(t)\right)^2}, \qquad (2.20)$$

where $P_T(t)$ is the reference quantile curve and $P_E(t)$ is the estimated quantile.
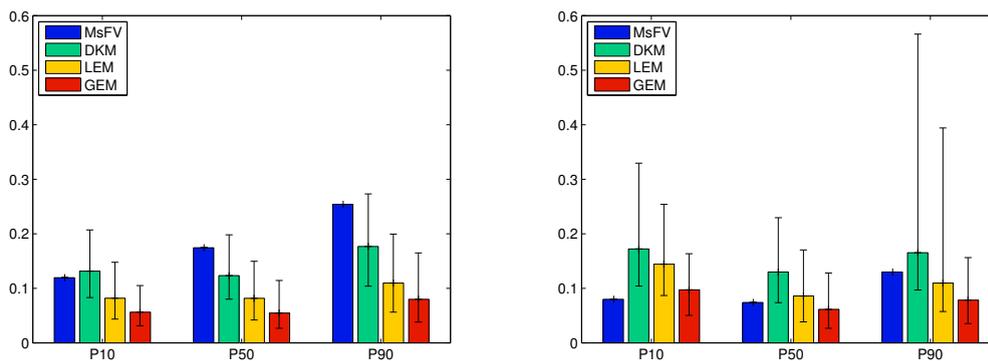
Figure 2.6: Errors on the quantile curves measured by the l_2-norm between the models and the reference curves. The bar plots represent the mean error of each method for each quantile curve. The error bars show the 80% confidence interval obtained for 500 results computed with different seeds. BCF is shown on the left and BCH on the right. Results for the l_∞-norm are shown in Fig. 2.7.

Figs. 2.6 and 2.7 shows the errors for the two set of boundary conditions and for the 500
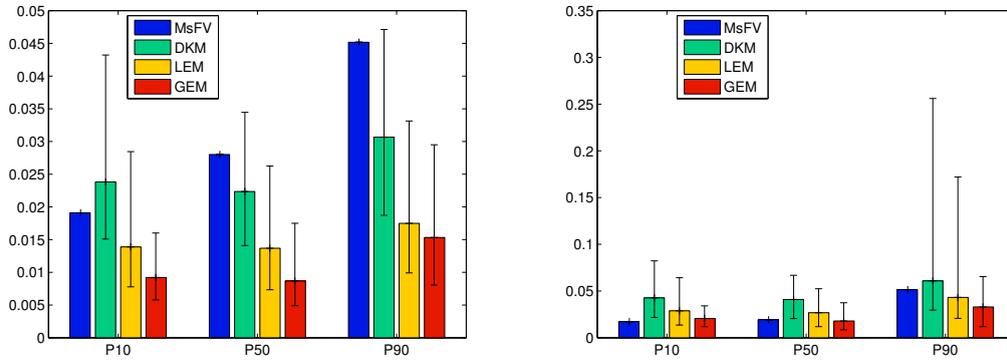
Figure 2.7: Errors on the quantile curves measured by the l_∞-norm between the models and the reference curves, for BCF (left) and BCH (right).

seeds. For each quantiles, the mean error of each method is represented by a bar plot, whereas the error bars represent the 80% confidence interval (i.e., the interval in which one finds 80% of the 500 results obtained with different seeds). These plots clearly show that the DKM error can be much larger than what observed in Figs. 2.8 and 2.9, which correspond to an initialization leading to an error close to the mean of the results from the 500 initializations.
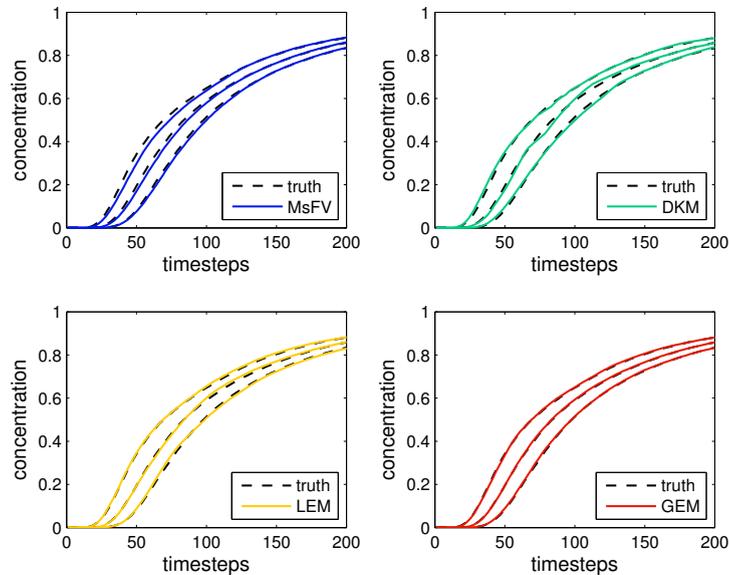
Figure 2.8: Quantiles curves estimated by MsFV (blue), DKM (green), LEM (yellow), and GEM (red) for the BCF.

For BCF, DKM performs, in average, better than MsFV for P50 and P90. MsFV responses
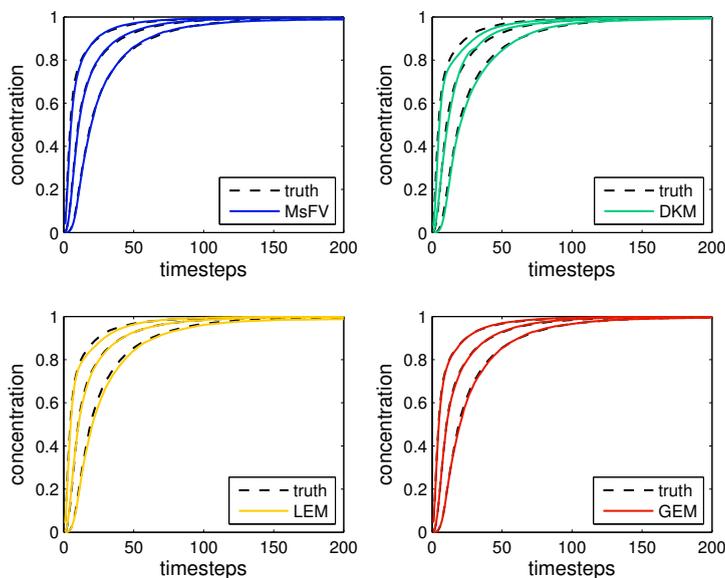
Figure 2.9: Quantiles curves estimated by MsFV (blue), DKM (green), LEM (yellow), and GEM (red) for BCH

provide an accurate selection of representative realizations, but yield a relatively poor estimate of the uncertainty due to the systematic underestimation of the concentration (see Fig. 2.8). However, DKM shows a large variability depending on the initialization of the clustering algorithm and for some seed can lead to larger errors than MsFV (1.7 times higher for P10 in 10% of the cases). LEM and GEM result in a much better estimate and lead to a considerable reduction of the dependency on the initialization of the algorithm. GEM performs better than LEM on both aspects (although for P90 GEM shows a slightly larger seed dependency).

For BCH, the MsFV estimate yields a sensibly lower error than the one obtained by DKM, and this despite the fact that information from 20 exact simulations is used in DKM. This is likely due to the large instability of the clustering algorithm that can lead to very unreliable estimates. This example clearly demonstrate how dangerous could be to rely only on medoid information, thus on an extremely small stochastic space, for estimate P10 and P90. The error models can correct this problem and lead to a better estimate than MsFV for GEM. For P90, one can observe a dramatic reduction of the seed dependency with respect to DKM, whose upper bound of the 80% confidence interval lays at 0.57; LEM reduces this to 0.39 and GEM to 0.16.

In conclusion, MsFV provides a good estimate of the statistical variability but tend to

present some systematic bias. DKM provides a good subset of representative realizations, but is strongly affected by the reduction of statistics. Both error models improve substantially the quantification of uncertainty by combining the whole available information. They both lead to a reduction of dependency on the algorithm seed; and GEM provides an excellent and much more stable estimate in both situations.

2.5.5 Cumulative distribution function at a given time step
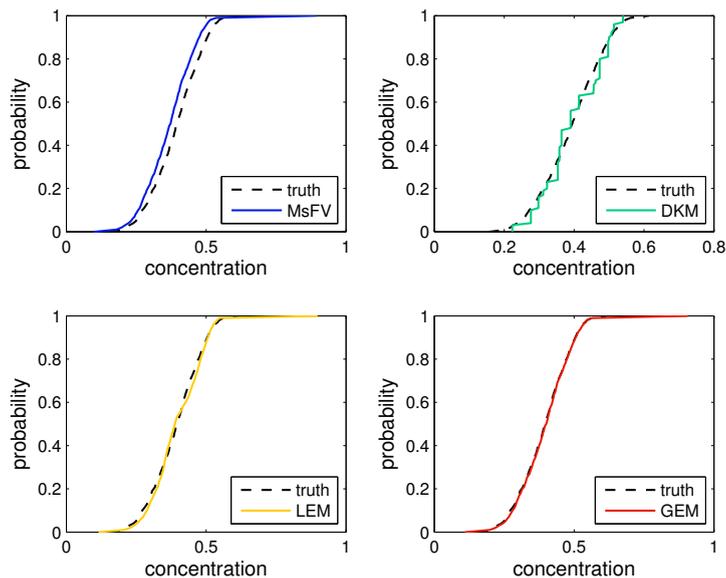
Figure 2.10: CDF of contaminant concentration at time step t = 70 for BCF.

Finally, we consider the estimated Cumulative Distribution Function (CDF) at two time steps: $t = 70$ for BCF (Fig. 2.10), and $t = 14$ (Fig. 2.11) for BCH, respectively. The CDFs in Figs. 2.10 and 2.11 refer to a single initialization (seed) of DKM, which has been chosen to be representative of the average result. Depending on the cluster initialization, however, the quality of the DKM results would be different.

For fixed-flux boundary conditions (BCF), one can observe a systematic shift of the MsFV CDF towards smaller concentrations; whereas for fixed-head boundary conditions (BCH), the MsFV CDF is close to the reference. Depending on the percentile, the error of the DKM estimate could be as high as 5% of concentration for BCF and 12% for BCH. The DKM CDF exhibits a staircase behavior, which is the result of the clustering and the subsequent reduction
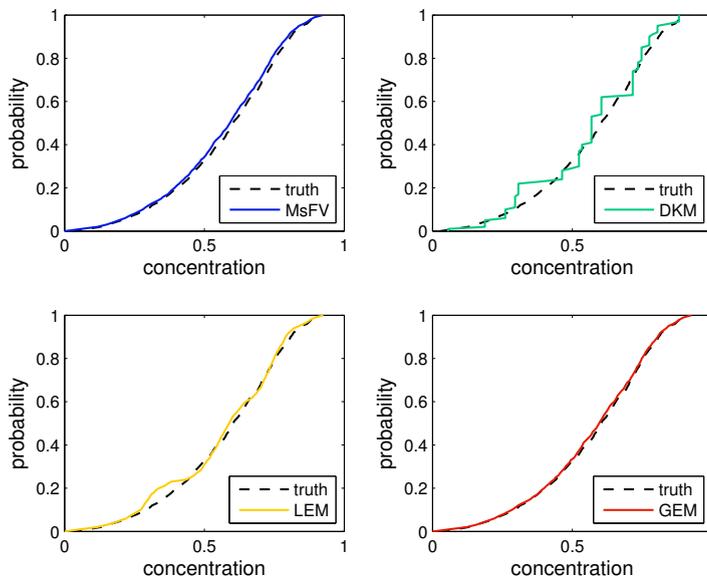
Figure 2.11: CDF of contaminant concentration at time step t = 14 for BCH.

of the number of realizations used to compute the CDF: the DKM estimate employs only the N_{Γ_k} medoid curves and neglects intra-cluster variability. This problem can be overcome by using the LEM or the GEM which construct a sample containing the same number of realizations as the original set, $\{C_i^{lem}(t)\}_{i=1,2,\dots,Nr}$. As a consequence a smooth CDF is obtained and the error is reduced. GEM provides an excellent estimate of the CDF for BCF and does just as well as MsFV for BCH.

2.6   Conclusions

The DKM is applied to estimate uncertainty at lower computational costs than a brute-force Monte Carlo approach. The method relies on an approximate model to select a subset of representative realizations for which the exact model is solved; then, the uncertainty is estimated only on the basis of the exact-response subset with no additional use of the approximate solutions. This approach neglects intra-cluster variability, leads to a dimensional reduction of the statistical space, and provides uncertainty estimates with a lower resolution than allowed by the original set of realizations. For our numerical test case, the DKM is not stable with respect to the initialization (seed) of the clustering algorithm and that this can lead to inaccurate predictions: in most critical cases, the DKM can even deteriorate the uncertainty

estimate provided solely by the approximate solutions. On the other hand, however, using only the approximate responses obtained with the MsFV method can lead to biased estimates of the uncertainty due to the localization assumptions, which reduce the accuracy of the solution in presence of long structures spanning several coarse cells. If this is an issue in a deterministic context (where iterative schemes are usually required to achieve the desired accuracy), in a stochastic framework this is a minor problem, which can be solved by means of an error model.

Two error models are devised that aim at exploiting the whole available information and combine the MsFV approximate responses with the exact responses obtained for medoids selected by the DKM. Both models employs the difference between approximate and exact solutions for the medoid realizations, but differ in the way this discrepancy is interpolated to correct each realization. The LEM applies the same correction to all realizations belonging the same cluster and can be seen as a natural way to model the intra-cluster variability of the responses; the GEM corrects each realization by a linear interpolation of all medoid errors (weighted by a function of the distance in the feature space) regardless to the cluster to which it belongs. Both models improve the DKM estimate and reduce the dependency on the initialization of the clustering. The GEM leads to excellent uncertainty estimates and performs systematically better than the LEM; this is likely due to the fact that a global error model (which does not rely only on intra-cluster information) is more consistent with the data considered in this study, which are not separated in clearly defined clusters.

The framework presented here is not specific to the methods considered (namely MsFV and DKM) but can be applied to other combinations of approximate models and techniques to select a subset of realizations. For instance, it can be used in a multiphysics context where the approximate model employs a simplified physical description and an error model is developed to predict a more complicated physical process (e.g., single phase vs. multiphase flow problems).

Some of the steps can be extended and generalized to ameliorate the reliability of the error model for challenging test cases. In particular four main improvements can be suggested: the selection of the representative realizations can be modified to obtain a larger number of realizations in regions of interest rather than uniformly covering the entire feature space; the subset of representative realizations could be iteratively enlarged until a number of realizations is selected that allows the required level of accuracy (note that this would require an a-posteriori estimate of the accuracy to define the stopping criterion); the weights used in the global error model can be obtained from the solution of an optimization problem, which can be tailored to guarantee that the corrected responses satisfy certain physical constraints (this entails a

more profound re-thinking of all steps to determine the ideal subset); finally, Functional Data Analysis (FDA) can be used to keep an explicit time dependence and work with breakthrough curves in a functional space rather than with points in a feature space.

Acknowledgments

Chapter 3

Functional error modeling for uncertainty quantification in hydrogeology

Laureline Josset, David Ginsbourger and Ivan Lunati

published in *Water Resources Research*[1]

3.1 Abstract

Approximate models (proxies) can be employed to reduce the computational costs of estimating uncertainty. The price to pay is that the approximations introduced by the proxy model can lead to a biased estimation. To avoid this problem and ensure a reliable uncertainty quantification, we propose to combine Functional Data Analysis and Machine Learning to build error models that allow us to obtain an accurate prediction of the exact response without solving the exact model for all realizations. We build the relationship between proxy and exact model on a learning set of geostatistical realizations for which both exact and approximate solvers are run. Functional principal components analysis (FPCA) is used to investigate the variability in the two sets of curves and reduce the dimensionality of the problem while maximizing the retained information. Once obtained, the error model can be used to predict the exact response of any realization on the basis of the sole proxy response. This methodology is purpose-oriented as the error model is constructed directly for the quantity of interest, rather than for the state of the system. Also, the dimensionality reduction performed by FPCA allows a diagnostic of the quality of the error model to assess the informativeness of the learning set and the fidelity of the proxy to the exact model. The possibility of obtaining a prediction of the exact response for any newly generated realization suggests that the methodology can be effectively used beyond the context of uncertainty quantification, in particular for Bayesian inference and optimization.

3.2 Introduction

The major challenge in hydrogeology is to deal with an incomplete knowledge of aquifer properties, which are usually measured only at few, discrete locations. This lack of information makes it impossible to address hydrogeological problems in a deterministic sense. The problem is typically stated in a stochastic framework and Monte Carlo simulations are used to propagate the uncertainty on aquifer properties to the quantities of interest (Dagan, 2002). A typical example is the prediction of the fate of a contaminant, which depends on the heterogeneity structure of the aquifer. The uncertainty on the contaminant breakthrough curve at a given location is estimated by solving the transport problem in a set of realizations, which represent the uncertainty on the permeability of the aquifer. The ensemble of the responses in the different realizations provides a sample of reference of the breakthrough curves.

Despite the appealing conceptual simplicity of this approach, problems arise when many realizations have to be considered and a large number of expensive flow and transport simulations have to be performed: computational cost quickly becomes prohibitive. To avoid this computational bottleneck, the problem is approximated either by coarsening the description of aquifer properties (standard upscaling techniques can be used to this end (Wen and Gómez-Hernández, 1996; Renard and de Marsiliy, 1997; Christie, 1996; Durlofsky, 2005)) or by simplifying the description of the physical processes, thus employing an approximate model or proxy (e.f., Scheidt and Caers (2009a)).

The price to pay for these simplifications is that inference based on the computed responses could lead to a wrong uncertainty quantification. If the approximation is physically motivated, the bias can be safely ignored. Effective computational gains, however, usually require very crude approximations whose effects on the uncertainty quantification is difficult to assess beforehand. To avoid this problem, the proxies are typically employed only to identify a representative subset of realizations for which the exact model is solved. This is the strategy of ranking methods (McLennan and Deutsch, 2005; Ballin *et al.*, 1992), or distance kernel methods (Scheidt and Caers, 2009a). In such case, it is crucial to evaluate to which extent the proxy is informative of the exact model response.

While it is generally acknowledged that an error analysis is necessary (Christie *et al.*, 2005), it is rarely performed. Although approaches that entail a systematic analysis and the construction of error models have been applied to flow in porous media (e.g., to correct fluid-properties approximations (O'Sullivan and Christie, 2006a,b) or approximate numerical solvers (Josset and Lunati, 2013)), in most cases the appraisal of approximate methods is performed for a very limited number of test cases, and it is assumed that they behave similarly for a wider range of applications. This approach is not exempt from problems because the informativeness of the proxy also depends on flow regimes and on the specific quantities of interest.

In this paper, we propose a novel methodology to systematically build statistical error models that describe the discrepancy between exact and approximate responses. Once the error model is constructed, it is used to correct the approximate responses and predict the responses expected from the exact model for all realizations. A characteristic of our approach is that the error model is purpose oriented, that is, it is established directly for the quantities of interest (in our case the breakthrough curve of a contaminant) and not for the state of the system (for instance, the full saturation -or concentration- and pressure fields). This reduces the complexity of the data to be handled (e.g., time-dependent curves rather than

time-dependent fields) while retaining all the relevant information.

Despite some similarities with the error models proposed by Josset and Lunati (2013), two additional key features characterize the present approach: the description of sparse data as continuous variables (time-dependent breakthrough curves), and the reduced dimensionality of the problem that is solved to construct the error model. To this end we employ Functional Principal Component Analysis (FPCA (Henderson, 2006)), which is a functional extension of PCA. The theoretical background is provided by Functional Data Analysis (FDA), a discipline that gathers mathematical tools to construct and treat continuous data. The description of continuous variables from sparse data is a problem faced in many fields of research and not only in environmental applications. While functional analysis is well established, FDA has been integrated as a whole only recently and promoted by Ramsay (2006); Ramsay *et al.* (2009). It has since been applied in various areas such as biomedical science, biomechanics, medicine or linguistic among others. We refer to Ullah *et al.* (2013) for a recent review of the application of FDA over the last 20 years. More specifically to the domain of groundwater protection problem, FPCA has been applied to interpret various contaminant concentrations in river quality (Henderson, 2006).

The paper is organized as follows. After a general problem statement (Sec. 3.3), we introduce the formalism used and describe the methodology in detail (Sec. 3.4). Then, the methodology is evaluated for a synthetic test case that represents a typical groundwater problem (Sec. 3.5). The paper ends with a discussion of the performance and of prospective applications (Sec. 3.6).

3.3   Problem statement

We consider a contamination problem in which a non-aqueous phase liquid (NAPL) is accidentally released and forms a plume that contaminates the fresh water. We are interested in predicting the breakthrough curve of the pollutant at a given location (typically a drinking well or a river that can be contaminated). Examples of NAPL contamination are hydrocarbons spills, or leakage of chlorinated solvents such as TCE. As the NAPL is not miscible with water and forms a separate phase, the evolution of the contamination plume is governed by a set of nonlinear transport equations (Appendix 3.7.1), which complicates both the contaminant behaviour and the numerical resolution of the equations.

Due to sparse measurements, the properties of the aquifer are only partially known. Their

uncertainty is represented by a set of N_r geostatistical realizations of the permeability and porosity fields $\{R_i\}_{i=1,\ldots,N_r}$. In brute force Monte Carlo approaches, this uncertainty is propagated by solving the nonlinear multiphase transport model (hereafter "exact model") and computing the NAPL breakthrough curve in each realization. Here it is assumed that the resulting set of curves, $\{y_i(t)\}_{i=1,\ldots,N_r}$, provides an accurate representation of the uncertainty on the travel time.

Our goal is to find an approximation of the uncertainty without computing the full set of exact curves $\{y_i(t)\}_{i=1,\ldots,N_r}$. To this end we use a simplified model based on the linear single-phase transport equations (hereafter "approximate model" or "proxy"), which allows a relatively inexpensive calculation of the approximate breakthrough curves, $\{x_i(t)\}_{i=1,\ldots,N_r}$. To provide an accurate approximation of the uncertainty, we need to learn the relationship between the proxy and the exact responses, such that an exact response can be predicted from each proxy response.

We formulate this step in a standard machine learning framework: a statistical model relating the exact response curves (treated as outputs of the statistical model) to the proxy response curves (treated as inputs of the statistical model) is postulated. The parameters are estimated based on a learning set (or training set), i.e., a collection of pairs of response curves obtained with the two models for $N_l < N_r$ geostatistical realizations, $\{(x_i(t), y_i(t))\}_{i=1,\ldots,N_l}$.

The statistical model relating the two sets of response curves (exact and proxy) is here restricted to the class of functional linear models (Ramsay, 2006), in which the relationships between the responses is

$$y_i = T(x_i) + \varepsilon_i \qquad i \in [1,\ldots,N_r], \tag{3.1}$$

where T is a bounded linear operator from the Hilbert space L_2 to itself, and the error functions ε_i are centered, independent, and typically assumed to meet further technical conditions (Cuevas *et al.*, 2002).

Since the identification of such statistical model is ill-posed, in practice further restrictions on the form of T are introduced to enable inferring T from the learning set. Two methods are suggested by Ramsay (2006); Ramsay *et al.* (2009): the full functional regression model and the Concurrent model. The full functional regression model allows capturing complex behaviours, but it is costly and requires the fine tuning of several smoothing parameters. The Concurrent model consists of a simpler functional linear regression. This method is fast, but

quite rudimentary because the model uses only concurrent features of the curves (additional details about the two models can be found in Appendix 3.7.2).

In this paper, we follow a slightly different strategy: we appeal to a spectral approach and decompose the elements of the learning set on two *ad hoc* bases, one for the proxy and one for the exact responses. The response curves are then described in two spaces of dimensions $D_{ex} < N_l$ for the exact responses and $D_{app} < N_l$ for the proxy responses. A statistical model is constructed to relate the coefficients of the elements of one space, $y_i(t)$, to the coefficients of the elements of the other space, $x_i(t)$, as illustrated in Fig. 3.1.

Once the approximation $\hat{T}$ of T is obtained from the learning-set, it is used to predict the exact responses of all realizations from of the approximate responses, i.e.,

$$\{\hat{y}_i = \hat{T}(x_i)\}_{i=1,\ldots,N_r}, \tag{3.2}$$

and the uncertainty is quantified from the ensemble of predicted curves.
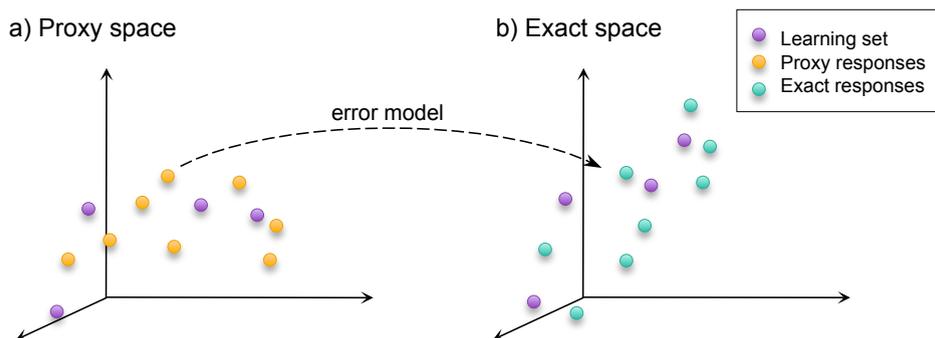
Figure 3.1: A statistical model is built on the learning set to relate the coefficients of the elements $x_i(t)$ in the proxy-response space to the coefficients of the elements $y_i(t)$ in the exact-response space. It is used as error model to predict the exact response from the proxy response.

3.4 Methodology

The construction of the error model consists of four steps: first, functional objects are built from the data in the learning set; second, the dimensionality of the problem is reduced by decreasing the dimensions of the two functional spaces; third, the relationship between the approximate and exact responses is constructed; fourth, the error model is used to predict the exact responses from the proxy responses. These steps are illustrated in the flowchart in Fig.

3.2.
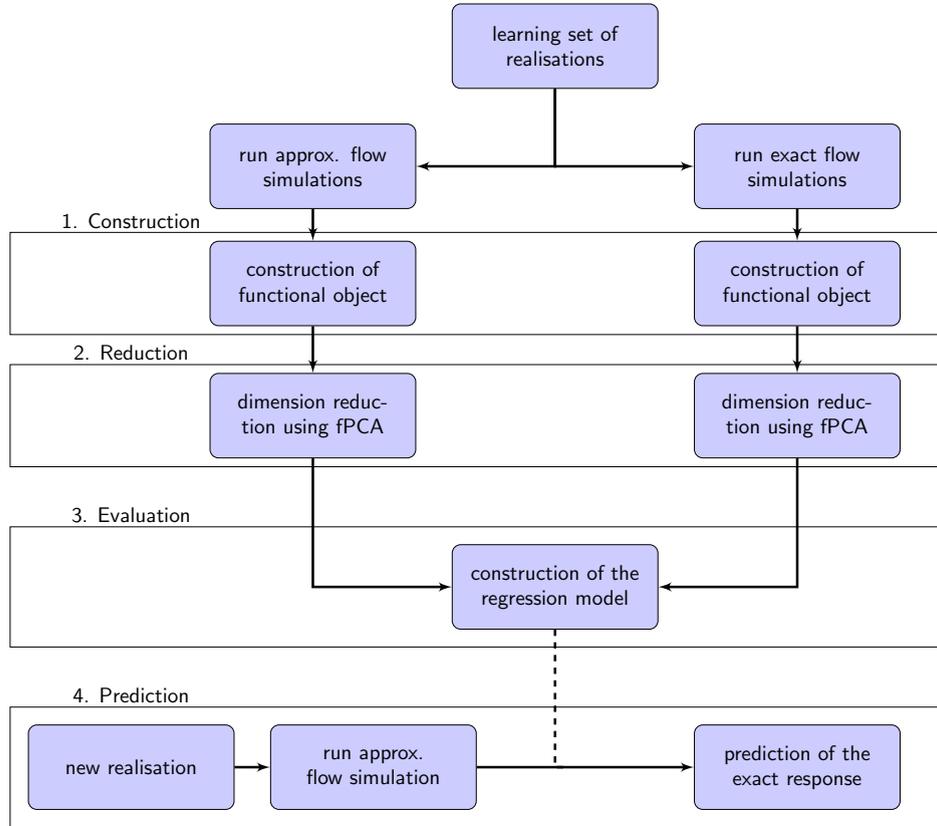
Figure 3.2: Flowchart of the methodology. After a learning set of realization has been constructed by selecting a subset of realizations and calculating pairs of proxy- and exact-response curves, the exact responses for the realizations that are not in the learning set can be predicted in four steps: 1. first, the functional objects are constructed by spline interpolation, 2. then, the dimensions of the subspaces of exact and proxy responses are reduced by means of FPCA, 3. next, a regression model is constructed between the proxy and the exact scores; 4. finally, the regression model is used to predict the exact responses of the realizations that are not in the learning set.

3.4.1 Recasting discretized curves as functional data

Both exact and proxy responses are obtained from numerical simulations and are represented by contaminant breakthourgh curves defined at discrete times. Therefore, we recast the time-discrete curves into time-continuous functions. This has two practical advantages: first, it allows us to use the formalism of functional data analysis and the tools that have been developed in this context; second, it permits to work with asynchronous information about the curves, i.e., curves that have been sampled at different times. Note that this step is essential

in applications in which analytic solutions are used as proxies or if the exact responses are provided by field measurements, which are typically acquired with different temporal resolution.

Many functional bases are available to recast discretized curves into functional data. Here, we use a K-dimensional B-spline basis denoted by $\{\varphi_k(t)\}_{k \in [1,K]}$. To determine the coefficients, a linear combination of the elements of this basis is fitted to the data, which are represented as time dependent functions of the form

$$f(t) = \sum_{k=1}^{K} c_k \varphi_k(t) \tag{3.3}$$

Ramsay (2006) suggests two strategies to choose the basis and fit the coefficients to data: either a low-dimension basis is used and the data are plainly projected (e.g., by ordinary least squares), or a high-dimension basis is used with a roughness penalty to mitigate overfitting. Both strategies allow not only to distinguish noise from information but also to impose various constraints on the functional objects, e.g. positivity and/or monotonicity. As our data (contaminant breakthrough curves) are typically fairly smooth, a standard B-spline basis of small dimension can be used. We refer the readers to (Ramsay, 2006; Ramsay et $al.$, 2009) for more details about the notions of roughness penalty and incorporation of constraints.

3.4.2   Functional reduction of the dimensionality

The previous step allows representing each exact response and each proxy response as a continuous function, i.e., $y_i(t)$ and $x_i(t)$, respectively. To decrease the dimension of the response spaces and the size of the regression problem, we employ Functional Principal Component Analysis, which is a functional extension of standard PCA and allows highlighting the main modes of variability in a sample of functions. Beside a small computational advantage, using spaces of lower dimension reduces the risk of over-fitting and allows us to visualize the data to assess the informativeness of the proxy response with respect to the exact response.

We apply FPCA to the exact and proxy responses in the learning set. Given the sample of proxy functions in the learning set, $\{x_i(t)\}_{i=1,...,N_l}$, with average $\bar{x}(t) = \frac{1}{N_l} \sum_{i=1}^{N_l} x_i(t)$ and estimated covariance function

$$\nu(t', t) = \frac{1}{N_l - 1} \sum_{i=1}^{N_l} [x_i(t') - \bar{x}_i(t')][x_i(t) - \bar{x}_i(t)], \tag{3.4}$$

FPCA constructs a non increasing sequence of eigenvalues of the estimated covariance function, $\mu_1^\circ \geq \mu_2^\circ \geq \cdots \geq \mu_{N_l-1}^\circ$, by solving the functional eigenequation

$$\int \nu(t',t)\zeta_i^\circ(t)dt = \mu_i^\circ\zeta_i^\circ(t'). \tag{3.5}$$

The sequence of eigenfunctions (or harmonics) of the covariance function, $\{\zeta_1^\circ, \ldots, \zeta_{N_l-1}^\circ\}$, satisfies the condition

$$\int \zeta_i^\circ(t)\zeta_j^\circ(t)dt = \delta_{ij}, \tag{3.6}$$

(where δ_{ij} is the Kronecker delta), and, together with the average $\bar{x}(t)$, form an orthonormal basis for the space of the sampled approximate responses. The eigenvalue μ_i is also denoted as the *probe score variance* and the eigenfunction $\zeta_i^\circ(t)$ as *harmonic* (Ramsay *et al.*, 2009). The dimensionality of the response space can be optimally reduced considering only the first D_{ex} and D_{app} for the exact response space and the proxy response space, respectively. The fact that the sequence of eigenvalues is non increasing guarantees that no other basis of size D_{app} can describe better the data; the total squared error introduced by discarding the eigenfuncions $(\zeta_i^\circ(t))_{i>D_{app}}$ is $\sum_{i=D_{app}+1}^{N_l-1} \mu_i^\circ$.

The basis allows us to approximate each proxy response as

$$x_i(t) \approx \tilde{x}_i(t) = \bar{x}(t) + \sum_{j=1}^{D_{app}} b_{ij}^\circ\zeta_j^\circ(t) \tag{3.7}$$

where

$$b_{ij}^\circ = \int [\bar{x}(t) - x_i(t)]\zeta_j^\circ(t)dt \tag{3.8}$$

is the projection of the deviation from the mean of the ith approximate curve on the jth harmonic ($\tilde{x}_i(t)$ denotes the approximation of $x_i(t)$ in terms of the first D_{app} harmonics). As in standard PCA, these coefficients are typically referred to as *scores*.

Although it offers an optimal dimensionality reduction with respect to the total mean squared error, the orthonormal basis might not be ideal to represent the information. The *varimax* algorithm (Kaiser, 1958) can be applied to find a suitable rotation that improve data interpretation while preserving the optimality of the result in terms of explained variance (Richman, 1986; Ramsay *et al.*, 2009). Therefore, without any further loss of information, the

approximate curves can be written as

$$\tilde{x}_i(t) = \bar{x}(t) + \sum_{j}^{D_{app}} b_{ij}\zeta_j(t), \qquad (3.9)$$

where

$$b_{ij} = \int [\bar{x}(t) - x_i(t)]\zeta_j(t)dt \qquad (3.10)$$

is the projection of the deviation from the mean of the ith curves on the rotated harmonic $\zeta_j(t)$.

An analogous procedure is applied to the sample of exact responses in the learning set, $\{y_i(t)\}_{i=1,\ldots,N_l}$, which is approximated as

$$\tilde{y}_i(t) = \bar{y}(t) + \sum_{j}^{D_{ex}} c_{ij}\eta_j(t), \qquad (3.11)$$

where $\bar{y}(t)$ is the average, $\eta_j(t)$ the jth harmonic of the (varimax) rotated orthonormal basis $\{\eta_i(t)\}_{i=1,\ldots,D_{ex}}$, and

$$c_{ij} = \int [y_i(t) - \bar{y}(t)]\eta_j(t)dt \qquad (3.12)$$

the score with respect to $\eta_j(t)$. (As for the proxy curve, the *tilde* denotes the restriction to the first D_{ex} harmonics).

3.4.3 Regression and error model

Once the problem dimensionality has been reduced by FPCA, we investigate the relationships between the two sets of curves in the learning set approximated by considering the first D_{app} and D_{ex} harmonics, $\{\tilde{x}_i(t), \tilde{y}_i(t)\}_{i=1,\ldots,N_l}$. The goal is to find a transformation between the spaces of exact and proxy responses. (Notice that the *varimax* rotation does not affect the representation of the curves, but might affect the quality of the transformation).

Here, we restrict ourselves to functional linear regression models of the form given in Eq. 4.6. Training such a functional linear model in full generality is not straightforward. A simple choice to restrict the class of linear regression models is to postulate that, at any time t, $\tilde{y}_i(t)$ depends on $\tilde{x}_i(t)$ solely through its value at that time t. This assumption leads to the Concurrent model

$$\tilde{y}_i(t) = \beta_0(t) + \tilde{x}_i(t)\beta_i(t) + \varepsilon_i(t), \qquad (3.13)$$

which is a particular case of the functional linear model in Eq. 4.6 and corresponds to $T(x_i)(t) = \beta_0(t) + x_i(t)\beta_i(t)$. The Concurrent model will be used as baseline in our numerical application, and compared to our FPCA-based prediction approach.

To simplify the exposition, in the following we assume that the same number of harmonics is retained for the two spaces, i.e., $D = D_{ex} = D_{app}$. However, the number of harmonics depends on the inherent variability of the learning set, which can be different for the exact and proxy responses. Ultimately, the number of harmonics to be employed depends on how rapidly the eigenvalues of the FPCA decomposition decrease for the specific problem. It has to be chosen large enough to guarantee an exhaustive representation of the variability of the response curves, but small enough with respect to the number of elements in the learning set to avoid over-fitting when the regression model is constructed.

Given $N_l \leq N_r$ pairs of accurate and proxy responses, $\{(\tilde{x}_i(t), \tilde{y}_i(t))\}_{i=1,\dots,N_l}$, we postulate that there exists a $(D+1) \times D$ matrix of real-valued coefficients $\boldsymbol{\beta}$ (with line index starting at 0, by convention) and a $N_l \times D$ error matrix $\mathbf{E}$, such that for any $(i,j) \in [1, N_l] \times [1, D]$,

$$c_{ij} = \beta_{0j} + \sum_{\ell=1}^{D} b_{i\ell}\beta_{\ell j} + e_{ij}, \tag{3.14}$$

where β_{ij} and e_{ij} are the components of $\boldsymbol{\beta}$ and $\mathbf{E}$, respectively. The errors, e_{ij}, are implicitly assumed to be Gaussian with zero mean and variance σ_j^2, which depends only on j. In matrix notation, the statistical model reads

$$\mathbf{C} = \mathbf{B}\boldsymbol{\beta} + \mathbf{E}, \tag{3.15}$$

where $\mathbf{C}$ is the $N_l \times D$ matrix containing the scores of the exact responses, c_{ij}, and $\mathbf{B}$ is the $N_l \times (D+1)$ with elements of the first column $b_{i0} = 1$ by convention, and containing the scores of the proxy responses $b_{i(j-1)}$.

In the statistics literature, solving Eq. 3.15 for the coefficient matrix $\boldsymbol{\beta}$ is referred to as a *multivariate multiple regression* problem (Hastie *et al.*, 2009; Fox and Weisberg, 2011). A simpler regression problem can be obtained by separating the regression models for the D responses, hence solving D independent regression problems

$$\mathbf{C}_{(j)} = \mathbf{B}\boldsymbol{\beta}'_{(j)} + \mathbf{E}'_{(j)} \quad (1 \leq j \leq D), \tag{3.16}$$

where $\mathbf{C}_{(j)}$ is the jth column of the score matrix $\mathbf{C}$. A very convenient fact is that the columns of the Ordinary Least Squares (OLS) estimator of $\boldsymbol{\beta}$ coincides with the concatenated OLS estimators of $\boldsymbol{\beta}'_{(j)}$ (Hastie *et al.*, 2009), that is

$$\widehat{\boldsymbol{\beta}}_{(j)} = \widehat{\boldsymbol{\beta}}'_{(j)} \quad (1 \leq j \leq D), \tag{3.17}$$

where $\widehat{\boldsymbol{\beta}}_{(j)}$ are the columns of the OLS estimator $\widehat{\boldsymbol{\beta}}$ (hereafter, the hat denotes the OLS estimator of the quantity). However, test statistics and confidence bands of the multivariate regression model cannot be directly derived from those obtained for the multiple linear regressions in Eq. 3.16 and have to be computed for the general regression model in Eq. 3.15. The formula of the simultaneous confidence bands is given in appendix 3.7.3, together with a brief outline of the derivation.

3.4.4 Prediction of the exact response from the proxy response

Once the OLS estimator $\hat{\boldsymbol{\beta}}$ has been obtained, the regression model is used to predict the exact response for all N_r geostatistical realizations on the basis of the corresponding proxy responses $\tilde{x}_i(t)$. The predicted exact response for the ith realization is

$$\hat{y}_i(t) = \bar{y}(t) + \sum_{j=1}^{D} \hat{c}_{ij} \eta_j(t). \tag{3.18}$$

where

$$\hat{c}_{ij} = \hat{\beta}_{0j} + \sum_{\ell=1}^{D} \hat{\beta}_{j\ell} b_{i\ell}, \tag{3.19}$$

are the estimates of the scores with respect to the rotated harmonics.

The estimator of the linear regression model allows us to predict the $\hat{c}_{ij}$ scores solely from the scores b_{ij} of the proxy responses, hence predicting $\tilde{y}_i(t)$ without solving the exact model. We emphasize the difference between the proxy response $x_i(t)$ (or $\tilde{x}_i(t)$, which is the projection onto the lower dimensional space defined by the first D harmonics, $\{\zeta_j\}_{j=1,\dots,D}$), and the predicted exact response $\hat{y}_i(t)$: they both approximate the "true" response $y_i(t)$, but, while $x_i(t)$ is simply the result of the proxy model and lives in the space defined by the basis of the proxy curves, $\hat{y}_i(t)$ results from applying the error models to the proxy response and lives in the space of the exact responses (more precisely: in the subspace defined by the orthonormal basis formed by the first D harmonics, $\{\eta_j\}_{j=1,\dots,D}$).

Surrogating $y(t)$ by $\hat{y}(t)$ is prone to errors: first, $\{\eta_i(t)\}_{i=1,\ldots,N_l}$ depends on the quality of the learning set; second, the subspace of the prediction is further reduced by considering only the first D harmonics; third, the coefficients $\hat{c}_{ij}$ are predicted through the OLS estimator of a linear regression model, and thus entails statistical uncertainties and possibly systematic errors due to the choice of a simple linear model.

3.5 Numerical test case: An idealized NAPL pollution problem

We consider an idealized groundwater pollution problem in which the fate of a NAPL plume has to be predicted. We model a portion of aquifer as a vertical 2D domain of length $10.8m$ and depth $5.1m$ discretized into cells of size $10cm \times 10cm$. Gravity effects are neglected, which implies that the density of the NAPL phase is equal to the water density. No-flow boundary conditions are imposed at the upper and lower boundaries, whereas the pressure is fixed at the right boundary. The contaminant is released at the left boundary (a constant influx is assigned) and displaces the water initially present in aquifer. We are interested in the time evolution of NAPL saturation at the right boundary. Two cases are investigated; first, we estimate the uncertainty on the contaminant breakthrough curve computed by averaging the saturation along the right boundary; then, we consider a single-point breakthrough curve obtained by sampling the saturation in a single cell (Sec. 3.5.5). As the NAPL is immiscible with water, the exact model solves the multiphase flow and transport equations, which require solving a pressure equation and a highly nonlinear phase-transport equation (see, e.g., Marle, 1981; Helmig *et al.*, 1997). The two equations are highly coupled and characterized by fluxes that exhibit a non-linear dependence on NAPL saturation. (The full system of equations is described in Appendix 3.7.1.)

The uncertainty on the transport properties of the aquifer (permeability and porosity) is represented by a set of $N_r = 1000$ geostatistical realizations that are generated by a multipoint geostatistical method (DeeSse) (Mariethoz *et al.*, 2010) with a training image obtained from data of facies-distribution collected at the Herten site (Germany) (Bayer *et al.*, 2011). As an example, three realizations are shown in Fig. 3.3.

d)	Hydraulic conductivity in m/s	Porosity
I	$2.2 \cdot 10^{-5}$	0.21
II	$1.2 \cdot 10^{-3}$	0.32
III	$6.1 \cdot 10^{-5}$	0.13
IV	$2.4 \cdot 10^{-4}$	0.16
V	$8.4 \cdot 10^{-2}$	0.25
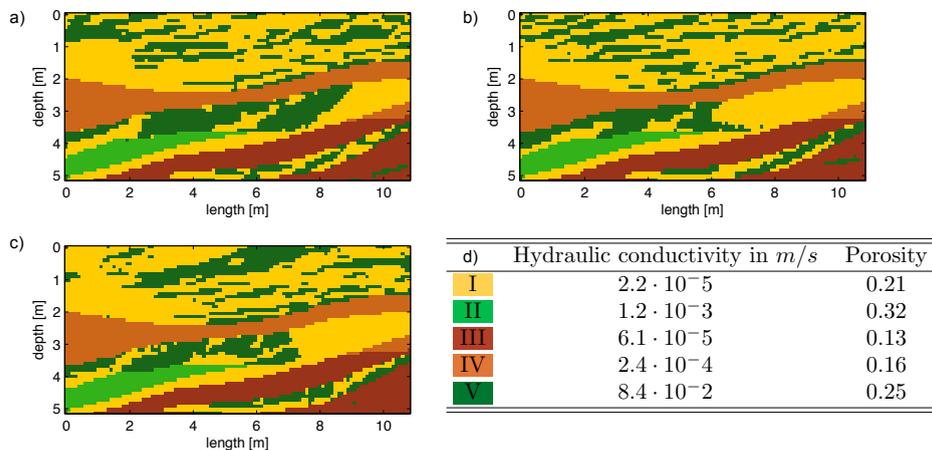
Figure 3.3: (a), (b) and (c): three examples of geostatistical realizations generated by a multipoint methods (DeeSSe, (Mariethoz *et al.*, 2010)) with training image from the Herten site (Germany) (Bayer *et al.*, 2011). The different colors correspond to 5 different facies, whose properties are reported in (d). The three realizations belong to the set of realizations used to construct the learning set; the corresponding NAPL breakthrough curves obtained with the exact and with the approximate models are highlighted in Fig. 3.4.

3.5.1 The proxy model

The proxy model simplifies the physics of the problem by treating the NAPL as an ideal tracer, thus solving a linear transport problem. Although it is possible to further improve the computational efficiency by simplifying the description of the heterogeneity (e.g., by some upscaling or multiscale methods (see, e.g., Josset and Lunati, 2013)), here we do not approximate the aquifer properties.

In practical situations, replacing a multiphase flow problem by a single-phase (tracer-transport) problem considerably reduces the computational costs. Indeed, a large part of the cost of solving the flow and transport system stems from the solution of the elliptic (or parabolic) equation that governs the pressure. Due to the effects of the saturation on the fluxes, this equation has to be solved at every time step in multiphase problems. In contrast, if the pollutant is considered as an ideal tracer, the saturation does not impact the velocity, and the pressure equation has to be solved only once. The NAPL transport equation becomes linear and can be solved very efficiently by streamline methods (here, we use a Finite-Volume upwind scheme that can be seen, in some sense, as a very rudimentary streamline method without sub-grid interpolation of the velocity field).

3.5.2 The learning dataset

After the proxy responses have been obtained by solving the ideal transport problem and computing the contaminant breakthrough curves for the whole sample of 1000 realizations, we construct the learning set by identifying a subset of $N_l = 20$ realizations. The realizations can be selected in several ways, including a simple random choice. Here, we use a clustering technique to group the proxy responses based on their l_2-distance, and we choose the k-medoid curves as representative of the clusters (Distance Kernel Method (Scheidt and Caers, 2009a)). The medoids define the subset of realizations, $\{R_i\}_{i=1,...,N_l=20}$, for which the exact responses are computed by solving the multiphase transport problem. Additional tests (not reported here) with learning sets consisting of $N_l = 50$ and $N_l = 100$ realizations did not show a significant improvement of the quality of the learning set. This suggests that only 20 realizations are sufficient to obtain a satisfactory error model for the present test case. Cross validation tests can be performed to identify the optimal size of the learning set.

As the numerical NAPL breakthrough curves are discrete in time, a spline basis is defined to interpolate the discrete data and construct the functional objects. In the present test case, data points are fairly smooth and a rather small number of basis functions is necessary for an accurate representation of the data (here, only 50 splines are used as basis functions). The 20 pairs of spline-interpolated proxy and exact curves in the learning set, $\{(x_i(t), y_i(t))\}_{i=1,...,N_l=20}$, are shown in Fig. 3.4.

3.5.3 Understanding the data using FPCA

To extract the relevant information from the data and to reduce the problem dimensionality, we apply FPCA independently to both sets of approximate and exact curves in the learning set. As in standard PCA, if all the components (harmonics) are considered, no approximation is made and the data are represented exactly. However, the eigenvalues of higher order harmonics decrease so fast that the first three components describe more than 97% and 99% of the variability of proxy and exact curves, respectively. In the subspaces defined by the first three harmonics, each curve is described by the corresponding three scores and by the sample means. To improve the interpretability of the data, a rotation is sought with the varimax algorithm (Ramsay *et al.*, 2012). The rotated harmonics for both sets of curves are shown in Fig. 3.5.

In the subset of the exact responses, the first rotated component explains the deviation from the mean behavior measured at late time. The second rotated component describes
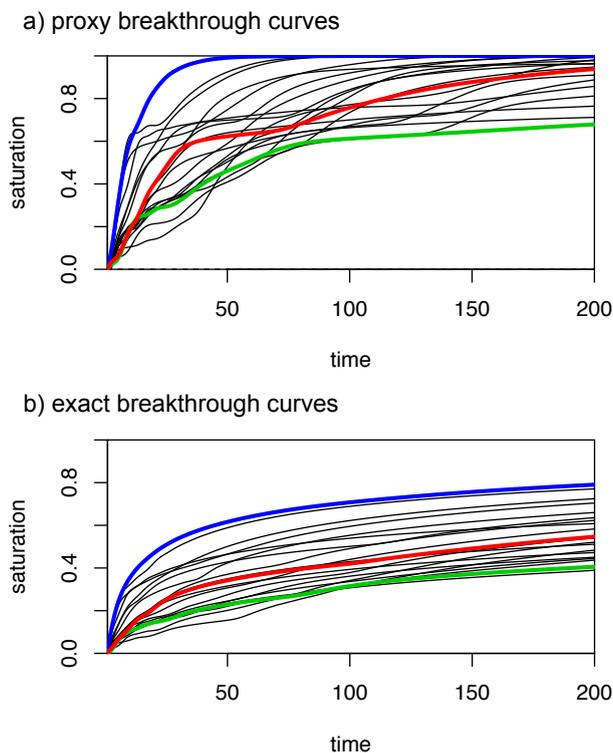
Figure 3.4: The learning set: (a) proxy curves and (b) exact curves recast as functional objects for the $N_l = 20$ realizations in the learning set. The thicker blue curves correspond to the realization in Fig. 3.3a), the red curves to 3.3b), and the green curves to 3.3c).

the variation at the beginning of the breakthrough curve, thus enlightening high-connectivity paths. The third component explains the variation observed at intermediate time. In the proxy subset, the first rotated component describes the initial variability; the second component highlights the variation at high saturation; and the third component explains the variation observed at intermediate time. By analyzing the projection of the curves on these components, it is possible to gain information about the data, for instance about the link between the early-time responses and the late-time variations. We refer to Henderson (2006) for an example in hydrology.

3.5.4 Regression model and evaluation of the proxy

The linear regression model is built between the scores of proxy and exact curves, which represent their coordinates with respect to the two orthonormal bases formed by the first three harmonics. Three linear regression problems (one for each exact-response score, $j = 1, 2, 3$) are solved to establish a relationship with the three proxy-response scores. The resulting
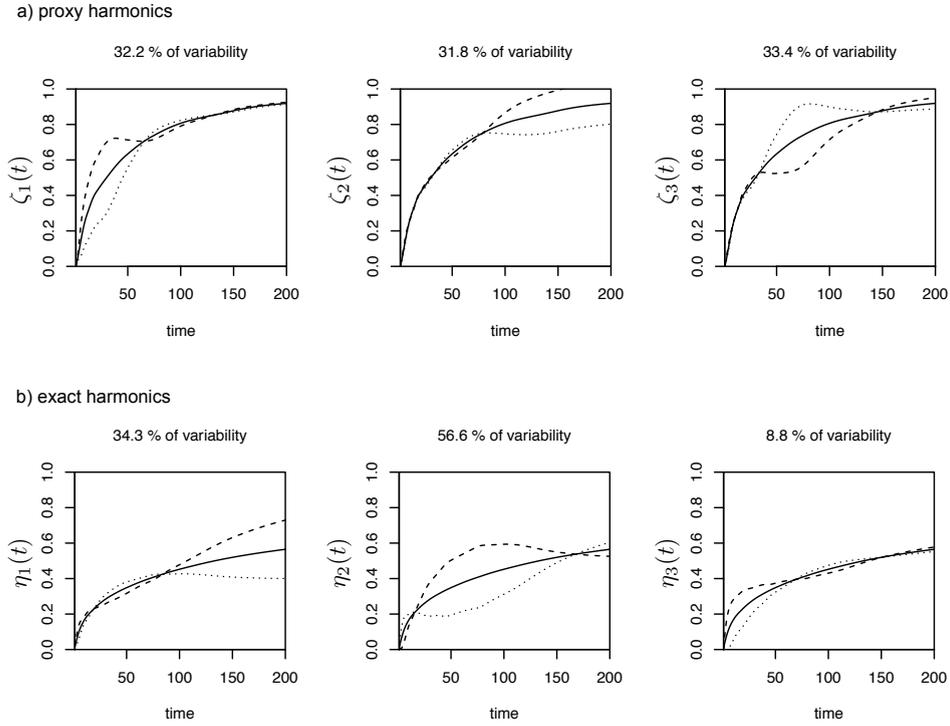
Figure 3.5: The three first rotated functional principal components (harmonics) extracted from the learning set are plotted for the proxy curves (top) and for the exact curves (bottom). The solid line is the mean curve and the dotted lines represent the variability around the mean described by the corresponding harmonic.

coefficients of the three regression models are

	β_{0j}	β_{1j}	β_{2j}	β_{3j}	R^2	p-value
$j = 1$	$-2.3 \cdot 10^{-16}$	0.42	0.18	-0.37	0.99	$< 2 \cdot 10^{-16}$
$j = 2$	$4.4 \cdot 10^{-17}$	0.82	-0.02	0.37	0.99	$< 2 \cdot 10^{-16}$
$j = 3$	$1.6 \cdot 10^{-16}$	0.51	0.03	0.08	0.97	$1.3 \cdot 10^{-12}$

$$(3.20)$$

Notice that the R^2 values are quite high and that $\beta_{0j} \approx 0$, which suggests that the linear regression model preserves the mean. The dependency among scores is illustrated in Fig. 3.6. The relationships between the scores of the three harmonics of the exact curves and the scores of the first harmonic of the proxy curves are rather well approximated by the linear regression. The scores of the second harmonic of the proxy curves are less important as it is indicated by the low values of β_{22} and β_{23}. This might be due to the fact that the proxy second harmonic explains the variability of the curves for saturations close to one, a situation that is not observed in the two-phase responses.
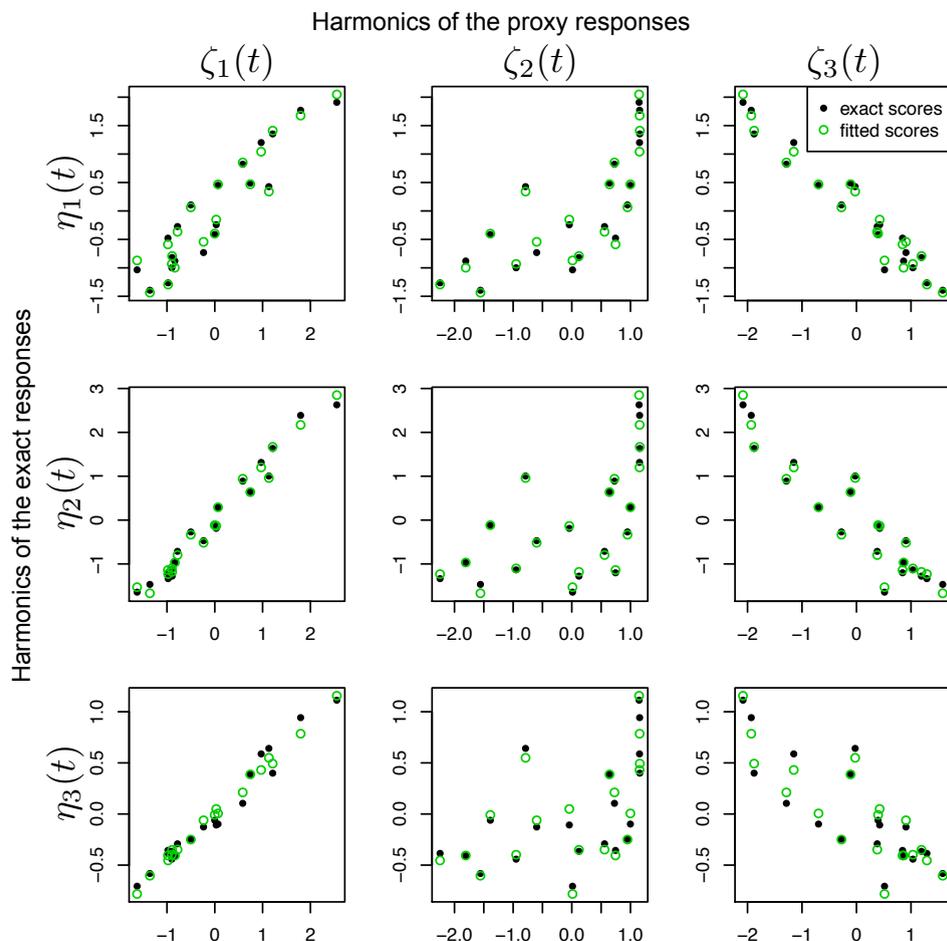
Figure 3.6: The scores, with respect to the first three harmonics $\{\eta_i(t)\}_{i=1,2,3}$, of the exact curves are plotted as functions of the scores for the approximate curves with respect to the harmonics $\{\zeta_i(t)\}_{i=1,2,3}$. The filled (black) circles correspond to the exact score, the empty circles (green) to the prediction of the scores by the OSL linear regression. The visualization is helpful to assess whether the linear regression model describes the relationship between proxy and exact curves in the learning set.

3.5.5 Performance of the regression model as error model

In general, the proxy-curve scores are informative of the exact-curve scores, at least for the curves pairs in the learning set. This suggests that, despite the rather primitive physical model employed, the regression model can be effectively used to predict the exact responses of the realizations for which only the proxy solution is available. The exact response is predicted on Eqs. 4.9 and 4.10.

Prediction of the average breakthrough curve at the outlet

We start by considering the prediction of the breakthrough curve calculated by averaging the saturation at the right-hand boundary. Examples of two predicted curves are shown in Fig. 3.7a and b. Despite the fact that the curves are very different for the two realizations, both predictions are in good agreement with the exact responses. In general, the behaviour of the exact response is well predicted, with the exception of some fluctuations at early times. The error model greatly improves the proxy solution and provides a much better prediction than the Concurrent model, which is unable to significantly modify the shape of the curves due to the use of only concurrent information.

The differences between predicted and exact curves are illustrated in Fig. 3.7c for all $N_r = 1000$ realizations, together with the mean error. The maximum differences in the saturation are observed at early time and are about 10%; later, the saturation discrepancy remains below $\pm 1.8\%$ for 68% of the realizations and below $\pm 4\%$ in the worst cases. The mean error is very close to zero, which shows that the predicted curves conserve the mean behaviour of the exact curves, and that the subset of 20 realizations selected in the learning set is representative of the whole sample to describe the mean behaviour.

Fig. 3.8a shows the histograms for the l_2-norm and the l_∞-norm of the errors. We compare the performance of the error model based on FPCA with the Concurrent functional linear regression model. The histogram of the l_∞-norm shows that on average the maximum deviation is 4.5% for FPCA, and about 8% for the Concurrent model. The l_2-error is on average more than three times lower for the FPCA-based model.

In many applications, the uncertainty is quantified in terms of the quantiles of the responses. Fig. 3.7d displays the quantile curves obtained using the different models. The Concurrent model fails to reproduce the 90th percentiles, because it is unable to modify the plateau of the proxy curves close to saturation one; it performs better for the other quantiles. The quantiles curves computed using only the learning set of exact responses (as suggested by (Scheidt and Caers, 2009a,b)), are slightly biased estimates of the exact quantiles. An excellent estimate is obtained with the functional error model, which is able to correct the approximate responses and predicts quantiles close to the exact ones.

55
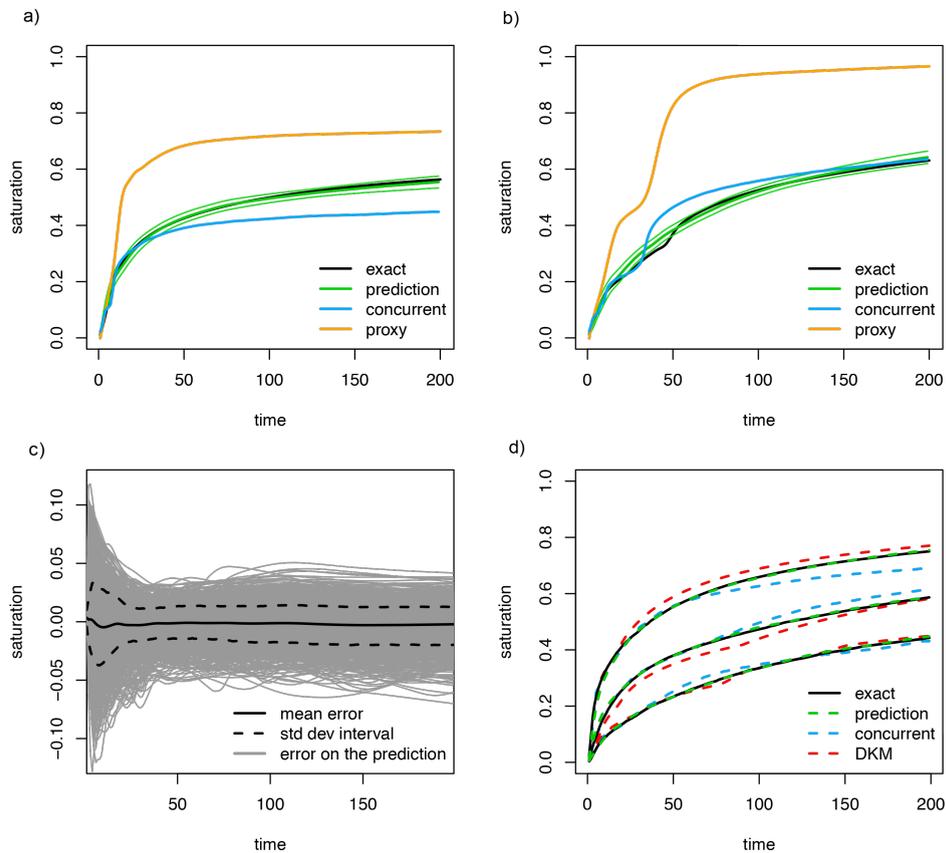
Figure 3.7: (a and b) the predicted responses (with 2σ-confidence intervals) of two realizations that are not in the learning set. (c) Prediction error of all $N_r = 1000$ realizations (gray curves), the mean error (continuous line), and the mean $\pm$ one standard deviation (dotted lines) are represented. (d) P10, P50 and P90 quantiles curves obtained with the different models and compared to the reference quantile curves computed using the whole set of exact responses (solid black line).

Prediction of single-point breakthrough curve

In this second test case, we are interested in predicting the breakthrough curve of the contaminant at a precise location, defined by a single cell of the numerical grid, which is located at mid-depth at the outlet. In contrast to the breakthrough curves averaged over the whole outlet, in which the effects of extreme permeability structures (flow barriers or preferential pathways) are smoothed, the single-point breakthrough curves display a variety of shapes. The large contrast in permeability and in connectivity at the sampling location leads to important differences, particularly in the first arrival time.

In this case, it is useful to apply a translation in time to redefine the origin, which is chosen to be the first arrival time. This procedure is referred to as registration in the FDA literature Ramsay (2006); Ramsay *et al.* (2009). For the translated responses in the learning set FPCA

a) FPCA error model
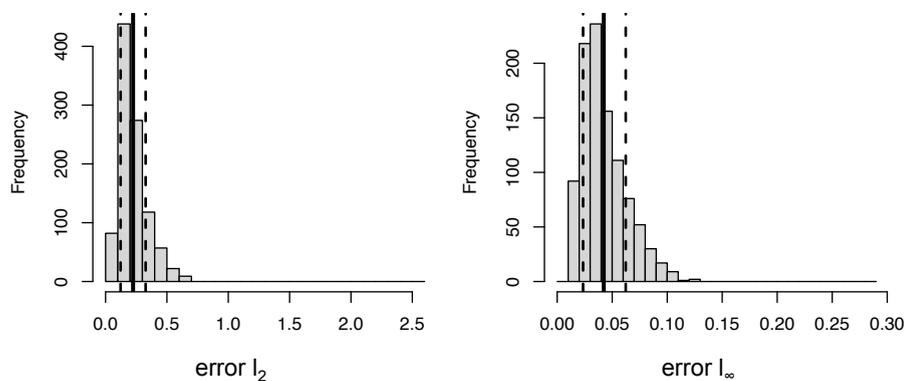
b) Concurrent model
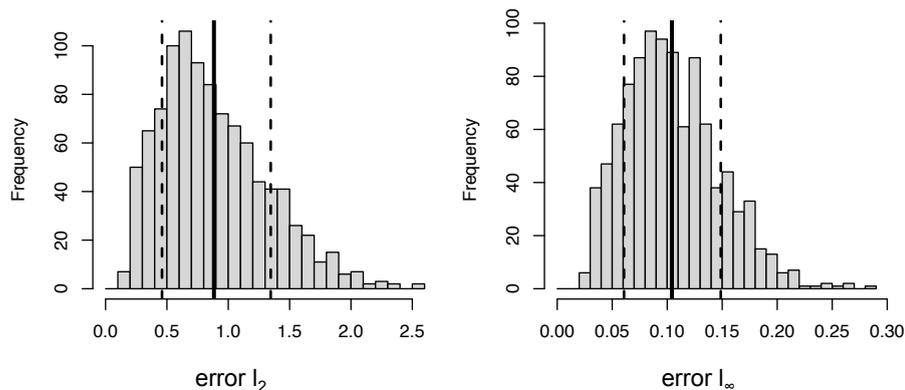
Figure 3.8: Histograms of the distribution of the l_2 error (left) and l_∞ error (right), (a) for the predictions of the FPCA model and (b) for the predictions of the concurrent model. The mean (continuous line) together with the mean $\pm$ one standard deviation (dotted lines) are represented.

is then applied and the dimensionality is reduced as described above. Again, we use the first three harmonics, which describe more than 98% of the variability of the shape of the curves after the registration. An example of proxy, predicted and exact curves after registration is shown in Fig. 3.9a for a realization that does not belong to the learning set.

Beside the prediction of the shape, it is now necessary to predict the first arrival time and translate back the predicted curves. The first arrival time is predicted jointly to the scores of the harmonics by solving a 4×4 regression model, where the 4th dimension is the first arrival times of the proxy responses, which have been used for the registration. Fig. 3.9b compares the proxy and exact curves with the predicted curve after translation by the predicted arrival time (these curves correspond to the registered curves in Fig. 3.9a). For the whole sample of realizations, the mean saturation error is close to zero and with a standard deviation that

remains below 0.04 (Fig. 3.9c).

The predicted quantile curves (shown in Fig. 3.9d) are in good agreement with the exact quantile curves for P50 and P90, but P10 is biased. As the concurrent model would perform very poorly in this case because it is unable to deal with curves characterized by different arrival times, we compare our methodology with the quantile curves obtained directly from the exact response in the learning set (this procedure corresponds to the classical DKM). As both the functional error model and the DKM estimates depend on the clustering, we have applied both methodologies 200 times. The example shown in Fig. 3.9 is representative of the typical behaviors of the methods (i.e., the quantiles are close to the average quantiles obtained form the 200 applications of the methods shown in Fig. 3.9e and d. In average, the functional error model is more robust than DKM and provides a better prediction of the P10 quantile curve.

Effects of the number of harmonics

Here, we investigate the effects of the number of harmonics on the prediction of single-point breakthrough curves. In order to increase the difficulty of the problem, we do not apply the registration as in the previous section (i.e., the breakthrough curves are not translated by their first arrival times). On one hand this requires more harmonics to describe the variability of the curves; on the other hand it allows us to demonstrate that the functional error model is able to correct for different arrival times also without registration.

We consider 200 different learning sets, which are selected by DKM clustering with different initialization. For each learning set we apply FPCA and then construct the functional error models by employing a different number of harmonics. The quality of the prediction is measured by the l_2 distance between the predicted and exact responses for all 1000 realizations.

The performance of the method (expressed as median error and confidence interval of the responses of the 200 learning sets) is presented in Fig. 3.10 as a function of the number of harmonics. The error exhibits a minimum around 5-7 harmonics. Indeed, when the number of harmonics is increased from 2 to 5, the variability of the learning set represented increases from 92% to 99%, leading to an improved error model. If the number of harmonics is increased further, the error increases quite rapidly. For 12 harmonics errors are very large and fluctuate greatly depending on the choice of the learning set. This behavior is a clear signature of over-fitting, as the large number of harmonics is not balanced by the size of the learning set
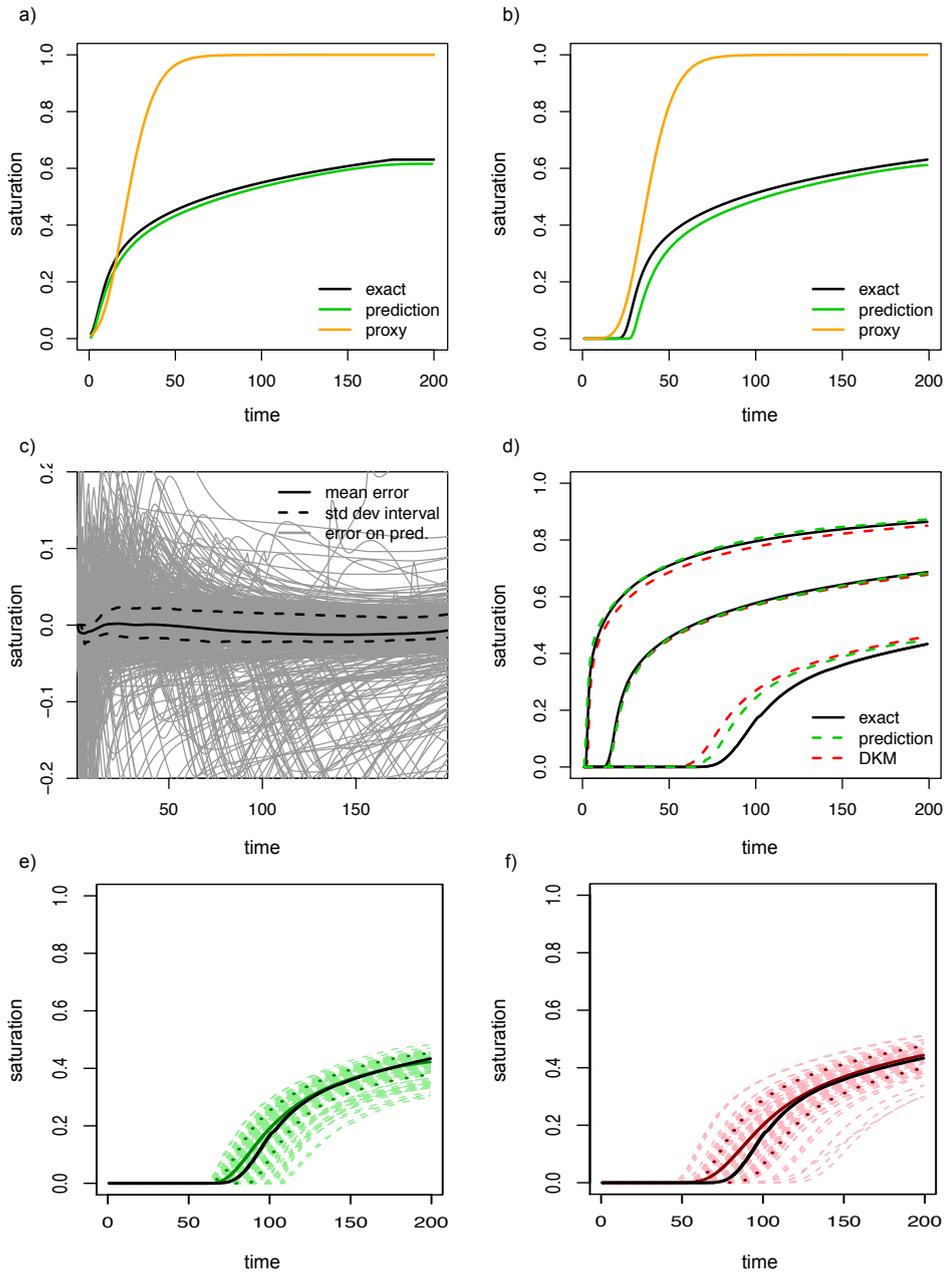
Figure 3.9: (a and b) predicted responses (before and after translation) of a realization that is not in the learning set. (c) Prediction error of all $N_r = 1000$ realizations (grey curves), the mean error (continuous line), and the mean $\pm$ one standard deviation (dotted lines) are represented. (d) P10, P50 and P90 quantiles curves obtained with the different models and compared to the reference quantile curves computed using the whole set of exact responses (solid black line). (e), respectively (f), shows the P10 FPCA, respectively DKM, predictions of the P10 quantile for the 200 clusterings.

59

(consisting of 20 pairs of curves) and the parameters of the regression model are not constrained enough by the data.
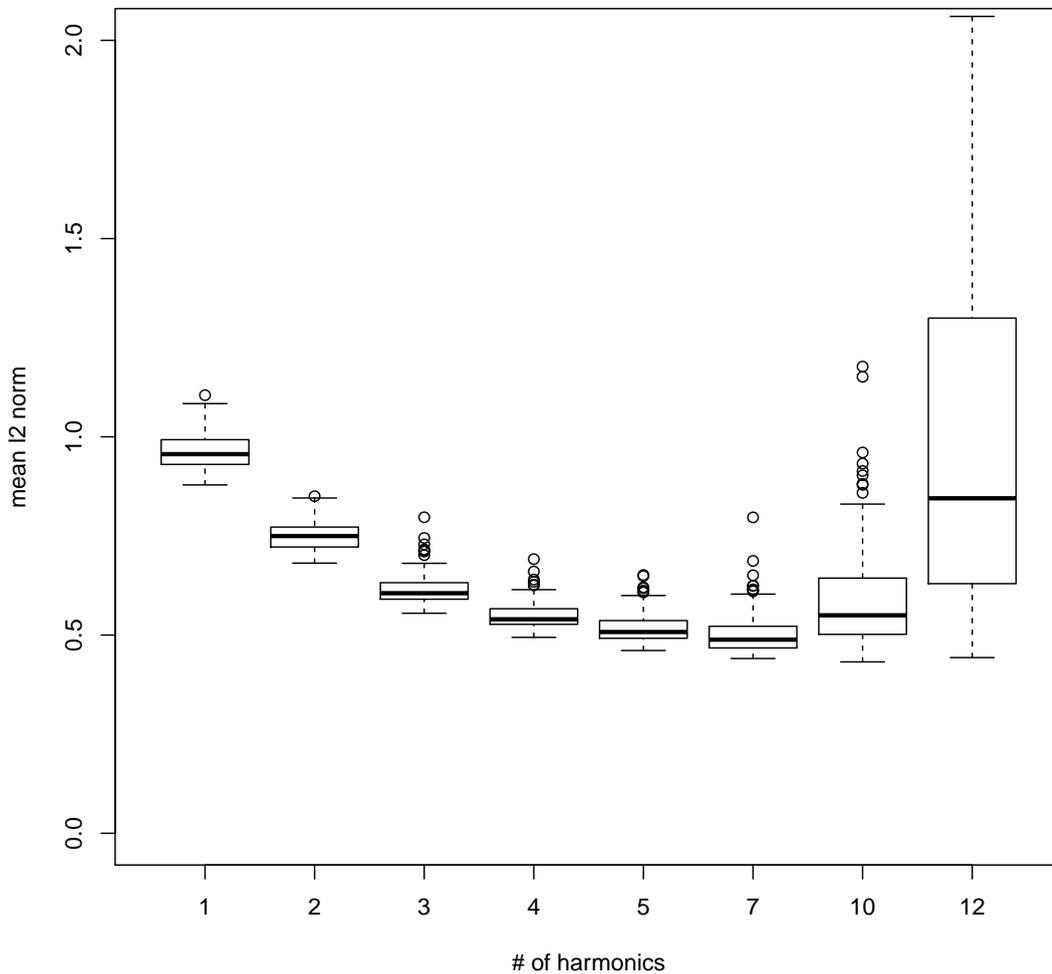
Figure 3.10: Boxplots of the prediction error (calculated as mean l_2 norm of the error of the predicted curves) as a function of the number of harmonics used to describe the proxy and exact curves in the learning set. The boxplots represent the statistics of the prediction errors over 200 clusterings in function of the number of harmonics. The thick line indicates the median error; the box the 1σ interval; the bars the 2σ interval; and the circles are the outliers (for 12 harmonics they are out of scale).

3.6 Conclusions

We have presented a novel methodology that combines elements of Functional Data Analysis and Machine Learning to construct error models that improve uncertainty quantification. The approach is purpose-oriented as it is formulated directly on the quantity of interest (in the case considered here, the contaminant breakthrough curve) rather than on the state of the system (e.g., the entire saturation and pressure fields).

The core idea of the method is to construct an error model from a learning set containing pairs of proxy and exact responses of a subset of realizations, and to predict the exact responses of the entire sample without solving the exact model for all realizations. FPCA is employed to separately reduce the dimensionality of the spaces of exact and proxy responses in the learning set. The advantage is twofold: on one hand, the small dimension allows a diagnostic of the regression model on scores to assess the informativeness of the proxy for the application at hand; on the other hand, using spaces of lower dimension reduces the risk of over-fitting when the regression model is constructed.

The method has been tested for a synthetic contamination problem, in which the breakthrough curve of a NAPL contaminant is predicted with the help of a tracer transport simulation (as proxy model). We have obtained excellent results with a learning set consisting of 20 pairs of curves (corresponding to 20 realizations out of a sample of 1000) and considering only the first three harmonics, which describe more than 97% of the variability. Visual inspection of the score scatter plots shows that the proxy is indeed potentially very informative of the exact response (this is confirmed by a linear determination coefficient $R^2 = 0.97$). Notice that this is not necessarily an indication of the quality of the predictions as the size of the learning set and the number of harmonics also influence the accuracy of the prediction. For both test case, the error model allows us to solve a two-phase problem only for the 20 realizations, whereas a simple tracer transport problem is solved for all realizations in the sample. The gain in computational efficiency is evident as multiphase transport requires solving the pressure problem at every time step, in contrast to ideal tracer transport, which requires solving the pressure equation only once.

In comparison to the Concurrent model (an existing methodology used to correct proxy responses), we have demonstrated an error reduction by a factor 3 when the functional error model is employed. Also, the error model improves the uncertainty quantification with respect to the estimate obtained solely on the basis of the 20 exact responses in the learning set

(this approach corresponds to the DKM, which uses the proxy responses only to cluster the realizations). Beside an increase in accuracy, the methodology presents two advantages over the DKM. First, the error model allows us to use the proxy response to predict the exact response for any new geostatistical realization that might be successively generated; this clearly opens new possibilities to use the model beyond the context of uncertainty quantification, and in particular for Bayesian inference, model calibration and optimization. Second, simultaneous confidence bands of the predicted curves can be defined by propagating the errors of the multivariate regression model. Notice that the residual uncertainty due to the size of the learning set and to the truncation of the basis should be taken into account.

Combining FPCA and machine learning can be seen as a general framework in which each component can be modified and improved, if it is required to improve accuracy. For instance, the rather crude linear regression model between the three-dimensional spaces of exact and proxy responses can be made more complex by increasing the dimensions (possibly with different truncations for the proxy and the exact model) or by refining the mathematical form of the statistical model to predict the scores. Possible enhancements include linear regression models with more complex basis functions (polynomials or others), but may also entail kernel methods like co-kriging. Almost any multivariate prediction may be adapted to this problem once the dimensionality reduction is performed. Another potential improvement is to perform the dimensionality reduction jointly for the proxy and the exact spaces, in order to optimize the informativeness of the proxy rather than the description of the variability of each response space independently. Indeed, in very complex test cases, it might occur that some small-eigenvalue harmonics of the proxy response might explain large-eigenvalue characteristics of the exact curves. This can be done by replacing FPCA by Functional Canonical Correlation Analysis (Ramsay, 2006) or by Functional Partial Least Squares (Cuevas, 2014).

Finally, we observe that the proposed framework can be applied far beyond the contamination example that we have presented. It can be useful in virtually any situation in which the most reliable technique has to be surrogated by an approximate method. Applications are not limited to the case in which evaluating exact response involves the solution of a complex numerical model, but also to situations in which the proxy or the exact responses consist of experimental data. The FDA framework would be then important to compare information with different temporal resolutions. Also, the error model can potentially be very useful in the context of Bayesian inference, when the number of responses that have to be evaluated (e.g., in Metropolis-Hastings algorithms and alike) is typically of the order of 10^5. In this case, a

functional error model capable to predict the exact responses only on the basis of the proxy responses can substantially speed up MCMC algorithms, as it reduces the cost of likelihood estimation. This would improve the efficiency of the calibration and optimization algorithms, which are often used in hydrogeological applications.

3.7 Appendix

3.7.1 Multiphase and single-phase transport equations

Assuming that both phases are incompressible and neglecting gravity and capillary effects, the saturation of the NAPL, S, is governed by the following system of equations:

$$\nabla \cdot \left[\left(\frac{k_n(S)}{\mu_n} + \frac{k_w(1-S)}{\mu_w} \right) k \nabla p \right] = 0, \tag{3.21}$$

$$\frac{\partial}{\partial t}(\phi S) - \nabla \cdot \left(\frac{k_n(S)}{\mu_n} k \nabla p \right) = 0, \tag{3.22}$$

where the absolute permeability, k, and the porosity, ϕ, are aquifer properties; p is the pressure; μ_n and μ_w are the viscosities of NAPL and water, respectively; and k_n and k_w are the relative permeabilities of NAPL and water, respectively, which are nonlinear functions of the saturation. Together with the constitutive relationships for the permeabilities (here, they are assumed quadratic i.e., $k_n(S) = S^2$ and $k_w(S) = (1-S)^2$), the two equations above form a complete system of equations that can be solved for p and S to calculate the NAPL breakthrough curves. These curves are the responses of the exact (multiphase) model.

Due to the nonlinearity of the relative permeability, the system above is computationally expensive because the two equations are coupled and the pressure equation has to be solved at any time step. This problem can be avoided by neglecting the nonlinearity of the permeabilities, hence approximating the system above as

$$\nabla \cdot \left(\frac{k}{\mu_w} \nabla p \right) = 0, \tag{3.23}$$

$$\frac{\partial}{\partial t}(\phi S) - \nabla \cdot \left(S \frac{k}{\mu_w} \nabla p \right) = 0, \tag{3.24}$$

which corresponds to a simple tracer transport problem without mechanical dispersion.

3.7.2 Linear models for functional responses with functional predictors

A simple class of linear models is the Concurrent model (Ramsay, 2006). The value of the response variable $y(t)$ is predicted solely by the value of the functional covariate at the same time t

$$y_i(t) = \alpha(t) + x_i(t)\beta(t) + \varepsilon_i(t), \tag{3.25}$$

where $\varepsilon_i(t)$ are the functional errors and the functions $\alpha(t)$ and $\beta(t)$ are estimated by

minimizing the sum of squares under some penalty on the roughness of the functions to avoid overfitting and loose predictability power. Despite the rather arbitrary choice of the degree of smoothness of the functional parameters, this method is quite fast but also rudimentary because there is a priori no reason to assume that only concurrent features of the curves are relevant (this is well illustrated by the synthetic test to predict the single-point breakthrough curve in Sec. 4.5.2).

A generalized formulation is when the functional variable contributes to the prediction for all possible time values

$$y_i(t) = \alpha(t) + \int x_i(s)\beta(s,t)ds + \varepsilon_i(t) \qquad (3.26)$$

which allows the predicted response to depend on the functional covariate at all times, but $\beta(s,t)$ is now bivariate. The application of this model is known to be particularly challenging as the smoothing constraints to be imposed is of paramount importance.

3.7.3 Simultaneous confidence bands for multiple multivariate linear regression

To take into account the uncertainty stemming from the linear regression, we derive simultaneous confidence bands for the predicted curve $\hat{y} = \mathbf{b}'\hat{\beta}\boldsymbol{\eta}(t)$, where $1 - \alpha$ is the level of confidence that the exact curve $\tilde{y}(t) = \mathbf{b}'\beta\boldsymbol{\eta}(t)$ is within the confidence bands for all t, that is

$$Pr\Big(\tilde{y}(t) \in [\hat{y}(t) - w_\alpha(t), \hat{y}(t) + w_\alpha(t)] \text{ for all } t\Big) = 1 - \alpha \qquad (3.27)$$

and, following the sketch of proof below, where $D_{ex} + D_{app} < N_l$ is assumed,

$$\begin{aligned} w_\alpha(t) \quad &= \sqrt{\tfrac{D_{ex}(N_l - D_{app} - 1)}{N_l - D_{ex} - D_{app}} F_{D_{ex}, N_l - D_{ex} - D_{app}}(\alpha)} \\ &\times \sqrt{(1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}) \tfrac{N_l}{N_l - D_{app} - 1} \boldsymbol{\eta}'(t)\hat{\boldsymbol{\Sigma}}\boldsymbol{\eta}(t)}, \end{aligned} \qquad (3.28)$$

where $\boldsymbol{\eta}(t)$ the values of the exact harmonics; $F(\alpha)$ Fisher's α-quantile; and $\hat{\boldsymbol{\Sigma}}$ the covariance matrix of the errors estimated on the learning set.

The key step of the derivation is the use of Scheffe's Lemma that states that, for a symmetric and positive definite matrix $\Gamma \in \mathbb{R}^{p \times p}$, the following statements are equivalent for any vector

$\mathbf{v} \in \mathbb{R}^p$ and constant $c > 0$

$$\left(\mathbf{v}'\Gamma\mathbf{v} \leqslant c^2 \right) \iff \left(|\boldsymbol{\psi}'\mathbf{v}| \leqslant c\sqrt{\boldsymbol{\psi}'\Gamma^{-1}\boldsymbol{\psi}} \qquad \forall \boldsymbol{\psi} \in \mathbb{R}^p \right) \qquad (3.29)$$

Sketch of proof

The residuals $\hat{\mathbf{E}} = \hat{\mathbf{C}} - \mathbf{C}$ are centred and with covariance $\mathbb{E}[\hat{\mathbf{E}}'\hat{\mathbf{E}}] = (N_l - D_{app} - 1)\boldsymbol{\Sigma}$, where $(\boldsymbol{\Sigma})_{jk} = \sigma_{jk}$. Assuming that $\mathbf{E}$ is Gaussian entails that $\hat{\boldsymbol{\beta}}$ is Gaussian, whereof $\mathbf{c} \sim \mathcal{N}_{D_{ex}}\left(\mathbf{b}'\beta, (1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}')\boldsymbol{\Sigma} \right)$. Then $\left(\frac{\mathbf{b}'\hat{\beta} - \mathbf{b}'\beta}{\sqrt{1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}}} \right)'\left(\frac{1}{N_l - D_{app} - 1}\boldsymbol{\Sigma} \right)^{-1}\left(\frac{\mathbf{b}'\hat{\beta} - \mathbf{b}'\beta}{\sqrt{1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}}} \right)'$ follows a Chi-squared distribution $\chi^2_{D_{ex}}$. On the other hand, the usual estimator $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ follows a Wishart distribution independently from $\hat{\beta}$. We then obtain the following

$$t^2 = \left(\frac{\mathbf{b}'\hat{\beta} - \mathbf{b}'\beta}{\sqrt{1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}}} \right)'\left(\frac{N_l}{N_l - D_{app} - 1}\hat{\boldsymbol{\Sigma}} \right)^{-1}\left(\frac{\mathbf{b}'\hat{\beta} - \mathbf{b}'\beta}{\sqrt{1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}}} \right) \sim T^2_{D_{ex}, N_l - D_{app} - 1}.$$
$$(3.30)$$

As the Hotelling T^2-distribution can be expressed in term of the F-distribution, we can write that, with probability $1 - \alpha$,

$$t^2 \leqslant \frac{D_{ex}(N_l - D_{app} - 1)}{N_l - D_{ex} - D_{app}} F_{D_{ex}, N_l - D_{ex} - D_{app}}(\alpha), \qquad (3.31)$$

where $F_{p,q}(\alpha)$ stands for the α-quantile of the Fisher-Snedecor distribution with parameters p and q.

Using Scheffe's Lemma (eq. 3.29) for $\mathbf{v} = \mathbf{b}'\hat{\boldsymbol{\beta}}$ the vector of predicted scores and $\boldsymbol{\psi}$ the vector of the exact harmonics values $\boldsymbol{\eta}(t)$, the second statement gives us the simultaneous confidence bands on the prediction.

Acknowledgments

Chapter 4

Accelerating Monte Carlo Markov chains with proxy and error models

Laureline Josset, Vasily Demyanov, Ahmed H. Elsheikh and Ivan Lunati

accepted in *Computers & Geosciences*[1]

---

[1]L. Josset, V. Demyanov, A. H. Elsheikh and I. Lunati. Accelerating Monte Carlo Markov chains with proxy and error models. *Computers & Geosciences*.

4.1 Abstract

In groundwater modeling, Monte Carlo Markov Chain (MCMC) simulations are often used to calibrate aquifer parameters and propagate the uncertainty to the quantity of interest (e.g., pollutant concentration). However, this approach requires a large number of flow simulations and incurs high computational cost, which prevents a systematic evaluation of the uncertainty in presence of complex physical processes. To avoid this computational bottleneck, we propose to use an approximate model (proxy) to predict the response of the exact model. Here, we use a proxy that entails a very simplified description of the physics with respect to the detailed physics described by the "exact" model. The error model accounts for the simplification of the physical process; and it is trained on a learning set of realizations, for which both the proxy and exact responses are computed. First, the key features of the set of curves are extracted using functional principal component analysis; then, a regression model is built to characterize the relationship between the curves. The performance of the proposed approach is evaluated on the Imperial College Fault model. We show that the joint use of the proxy and the error model to infer the model parameters in a two-stage MCMC set-up allows longer chains at a comparable computational cost. Unnecessary evaluations of the exact responses are avoided through a preliminary evaluation of the proposal made on the basis of the corrected proxy response. The error model trained on the learning set is crucial to provide a sufficiently accurate prediction of the exact response and guide the chains to the low misfit regions. The proposed methodology can be extended to multiple-chain algorithms or other Bayesian inference methods. Moreover, FPCA is not limited to the specific presented application and offers a general framework to build error models.

4.2 Introduction

Simulations of subsurface flow is important in many applications, such as groundwater protection and remediation, water prospection, exploration of hydrocarbon resources, and nuclear waste disposal. One of the main challenges is to estimate a continuous distribution of the underground model parameters from a sparse set of observational sites. This lack of information on model input propagates to the quantities of interest (for instance, the concentration of a pollutant in a drinking well), whose exact values remain uncertain. Model calibration using historical integrated data (for example, time series of concentration or pressure at observation

wells) is often used to reduce the uncertainty on model parameters by relying on Bayes theorem. A widespread approach for numerical application of Bayes rule is to use Monte-Carlo Markov-Chain (MCMC) simulations (Robert and Casella, 2004) to sample the posterior probability density function. While MCMC is theoretically robust and ensures convergence to the true posterior distribution under mild constraints, in practice it is subject to several limitations due to the cost of the large number of required flow simulations, which can become prohibited in presence of limited computational resources. Indeed, the finite length chains should be able to explore all areas of the prior space in order to provide samples from the posterior distribution. To achieve this goal, it is tempting to increase the step length of the chains, but this would result in a drastic reduction of the acceptance rate (which should ideally remain around 20-50% in multidimensional space) and subsequently in a high number of wasted simulations (Roberts *et al.*, 1997).

To avoid these issues, Efendiev *et al.* (2005, 2006) and Christen and Fox (2005) have introduced a two-stage MCMC, which employs a less computationally expensive solver to obtain a first evaluation of the proposal and decide whether it is useful to run the exact solver. This allows them to reduce the number of exact simulations that will be rejected and thus increase the acceptance rate. This methodology has been first explored by Christen and Fox (2005) to recover resistor values of an electrical network from measurements performed at the network boundary. They have obtained an increase in acceptance rate (the number of exact simulations accepted over the number of exact simulations run; first-stage simulations are not taken into account as their cost is assumed to be negligible). Both Efendiev *et al.* (2006) and Christen and Fox (2005) have shown that, under certain hypotheses, the solution converges to the posterior distribution. Efendiev *et al.* (2005, 2006); Dostert *et al.* (2008) have applied this methodology in the context of flow in porous media. As first-stage solver they have used a multiscale method, which combines a global coarse solution with a number of local fine solutions. If the coarse solution is accepted, local solutions are employed to reconstruct a finer solution on the original grid, based on which the second-stage evaluation is performed. While this allows for the necessary convergence assumptions to be satisfied (namely, smoothness and strong correlation), the computational gain of the two-stage set-up is limited. Indeed, the reconstruction step (necessary for the second-stage evaluation) is cheap with respect to the cost of constructing and solving the coarse problem used at the first-stage. Other applications of two-stage MCMC have used polynomial chaos response surfaces (Zeng *et al.*, 2012; Elsheikh

71

et al., 2014; Laloy *et al.*, 2013) as first-stage model. The computational gain is much higher, despite some additional cost required to set up the polynomial chaos model.

The use of inexact solvers requires designing error models to account for the discrepancy between approximate and exact responses. In the context of multiscale approaches, Kennedy and O'Hagan (2001) used a Gaussian-process method to represent model inadequacy. O'Sullivan *et al.* (2005); O'Sullivan and Christie (2006a) employed error modeling to reduce the bias in history matching resulting from the use of upscaled reservoir models. Efendiev *et al.* (2009) proposed non-linear error models in the context of ensemble-level upscaling. Scheidt *et al.* (2010), for instance, used a distance metric to account for upscaling errors in ensemble history matching. More specifically to two-stage MCMC, Cui *et al.* (2011) proposed to adapt the error model at each iteration: they used information on the discrepancy between the exact and approximate models at the previous iteration to correct the result of the successive iteration. However, this approach works and provides a good correction only for problems that are smooth enough.

Here, we propose a different strategy that combines a two-stage MCMC set-up with a methodology recently presented by Josset *et al.* (2015). We use an approximate model (proxy) that assumes a very simplified physics with respect to the problem under consideration, and we construct an error model to account for the approximation errors. The error model is purpose oriented as it is tailored directly for the quantities of interest following an approach typical of machine learning. For a subset of realizations, the responses of both the proxy and the exact models are evaluated and the mapping between the two is learned by means of tools from functional data analysis (Ramsay, 2006; Ramsay *et al.*, 2009). Josset *et al.* (2015) applied this methodology to propagate the uncertainty on the permeability field to the concentration of a pollutant in the observational well. Here, the methodology is tested on a complex problem of Bayesian inference, the Imperial College Fault (ICF) test case, which is a benchmark problem first published by Tavassoli *et al.* (2004) and repeatedly explored in many studies (e.g., Demyanov *et al.* 2010; Mohamed *et al.* 2011, 2012).

The paper is structured as follows: we first describe the ICF test case and review the literature about the calibration of this model (Section 2). Next, we present the novel methodology, which uses a purpose-oriented error model within a two-stage MCMC set-up (Section

3). Then, we specifically construct and evaluate the error-model approach for the ICF problem (Section 4.1). Finally, we compare and discuss the results of the two-stage MCMC with the classic Metropolis-Hastings algorithm (Section 4.2).

4.3 The Imperial College Fault (ICF) test case

The ICF test case was first published by Tavassoli *et al.* (2004, 2005) as a simple yet challenging example of history matching in petroleum engineering applications. Since then, ICF has proved a difficult test for optimization techniques due to numerous local minima. The ICF model consists of a layered reservoir disrupted by a fault (figure 4.1), in which water is injected at the left-hand boundary while the displaced fluids are recovered at the right-hand boundary. The layer-cake model of the reservoir permeability is described by three parameters: the conductivity of the high permeability facies, K_{high}, the conductivity of the low permeability facies, K_{low}, and the fault throw, h. The true parameters are $K_{high} = 131.6$ md, $K_{low} = 1.3$ md and $h = 10.4$ ft. A uniform distribution $\mathcal{U}_{[a,b]}$ (where a and b are the bounds of the distribution) is attributed to each parameter as prior.
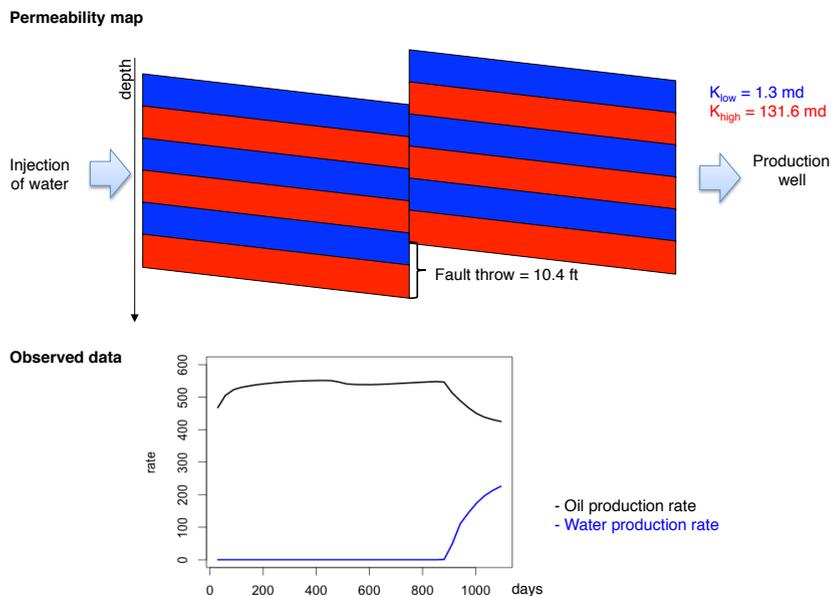
Figure 4.1: The permeability map of the ICF test case and the observed data used for the history matching. As prior, a uniform distribution is attributed to each parameter, i.e., $P(h) = \mathcal{U}_{[0,60]}$ for the fault throw h, $P(K_{high}) = \mathcal{U}_{[100,200]}$ for the permeability of the most permeable facies K_{high}, and $P(K_{low}) = \mathcal{U}_{[0,50]}$ for the permeability of the least permeable facies K_{low}.

The calibration of the parameters to the observational data (oil and water production rates) appeared to be a challenging history matching problem. Due to the nature of the permeability field, several parameter combinations, corresponding to narrow regions of the parameter space, can reproduce the observational data with satisfactory accuracy. Between these regions of good quality, the misfit is very high due to the very irregular response surface that results from the strong fluctuations of the connectivity across the fault when h is varied. We refer to figure 4.9 for a 1D cross-section cut of the complex misfit surface that characterizes this problem.

Many optimizations and inference techniques have been applied to the ICF problem over the years. The first studies of this test case (Tavassoli *et al.*, 2004, 2005; Carter *et al.*, 2006) have employed a pure Monte Carlo approach, which required nearly 160'000 samples of the parameter space. Christie *et al.* (2006) demonstrated that a good representation of the uncertainty can be inferred from a few thousand samples using Genetic Algorithm Important Sampling with artificial neural network proxy. More recently, Demyanov *et al.* (2010) have used Support Vector Machines (SVM) with a small number of flow simulations (about 700); and Mohamed *et al.* (2011) have employed Particle Swarm Optimization (PSO) using 2050 flow simulations. A Bayesian inference approach close to two-stage MCMC has been presented by Mohamed *et al.* (2012), who used a population MCMC method with 45'000 simulations. We refer to Mohamed *et al.* (2011) for a more detailed review of the literature on the ICF problem.

4.4   Methodology

Our objective is to sample the geostatistical parameter space conditioned on some flow observations. Using Bayes theorem, this can be written as

$$P(k|d) \propto P(d|k)P(k) \tag{4.1}$$

where $P(k|d)$ the is probability of the realization with the parameters, k, conditioned on the data, d, and $P(d|k)$ the likelihood distribution. The most common technique to tackle this problem uses the Metropolis-Hasting (MH) algorithm (Robert and Casella, 2004), which is very demanding in terms of CPU time. We propose to employ a two-stage MCMC algorithm in

which the first stage allows us to reject samples from low likelihood regions of the parameter space based only on the responses of an approximate model. The latter is constructed by combining a proxy model with an error model that permits the reduction of the proxy bias. This approach is illustrated in figure 4.2.

4.4.1 Error modeling based on Functional Principal Component Analysis (FPCA)

The number of flow simulations required for MCMC or two-stage MCMC can become prohibitive in case of very complex physical processes that requires performing computationally expensive simulations. An inexpensive proxy that relies on a very simplified physical description can be used to reduce the computation cost. However, direct inference from the proxy response is extremely dangerous, because the proxy model neglects important physical couplings inherent to the system, which likely bias the predictions. However, if we are able to devise an effective model of the errors arising from the use of the proxy, we can account for the neglected complexity and correct the bias of the prediction.

A purpose-oriented error model can be constructed directly on the quantity of interests by training a regression model on a subset of response pairs obtained by evaluating the proxy and the exact model for a selected subset of realizations (Josset *et al.*, 2015). The flowchart of the regression-model construction is detailed hereafter and illustrated in figure 4.2.

Construction of the learning set of curves

The first step consists in constructing the learning set from pairs of proxy and exact response curves corresponding to the same realizations. To obtain a learning sample of N realizations, which is assumed representative of most plausible solutions, we use the Latin Hypercube Sampling (Carnell, 2009). Other sampling methods (e.g., basic random sampling of the prior or stratified sampling) could be successfully employed as long as the various regions of the prior are sampled.

Once the learning realizations are identified, the proxy and the exact solutions are computed to get the time-dependent response curves. The functional proxy curves, $\{x_i(t)\}_{i=1,\dots,N}$, and functional exact curves, $\{y_i(t)\}_{i=1,\dots,N}$, are obtained by interpolating the responses produced by the numerical models, which are discrete in time, by means of a basis of spline
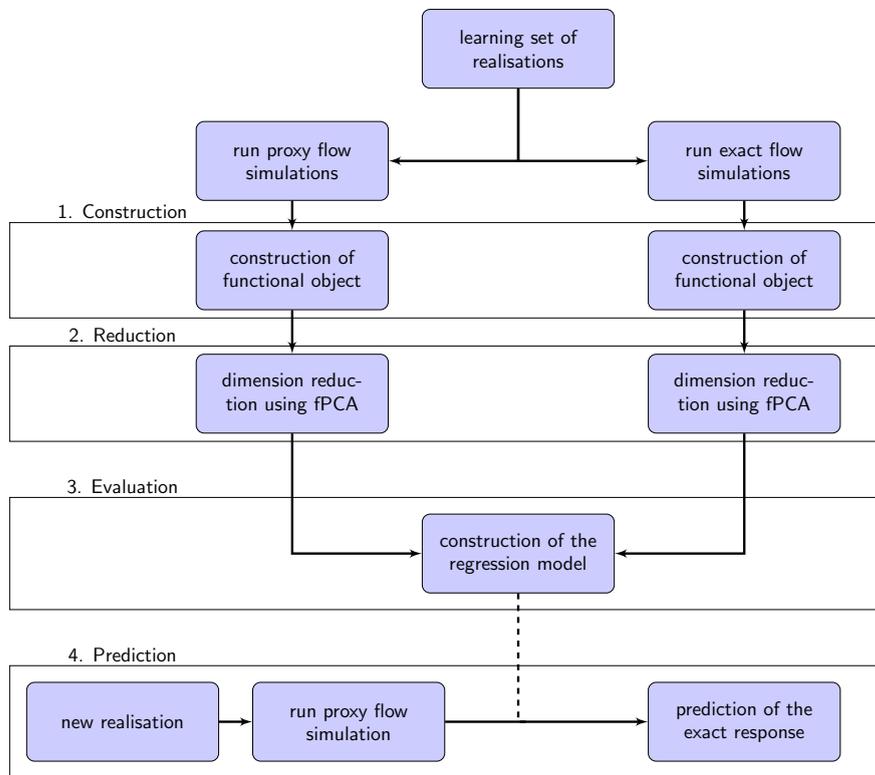
Figure 4.2: Flowchart of the construction of the error model as proposed in Josset *et al.* (2015). Numbering refers to the sub-sections in section 5.2
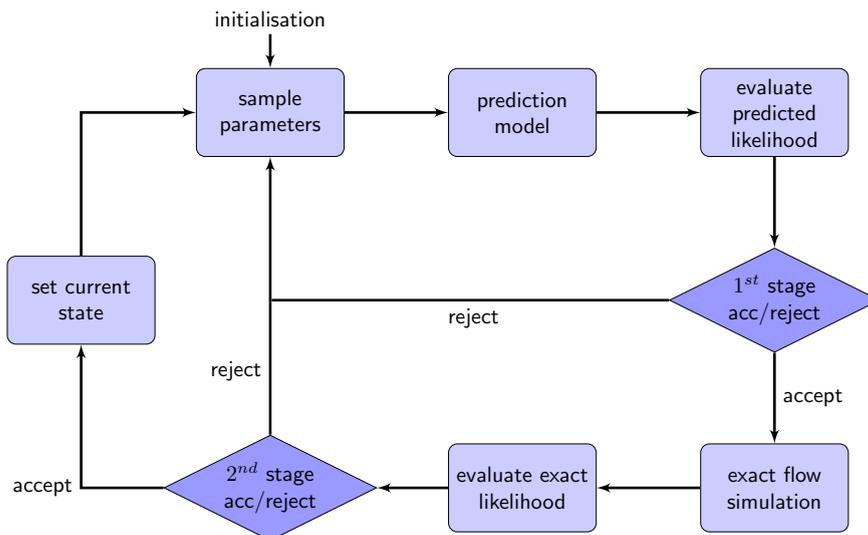
Figure 4.3: Flowchart of the two-stage MCMC algorithm.

functions.

Notice that a functional representation of the curves is necessary to deal with data acquired with different time resolution, as it is always the case when the numerical solvers employ adaptive time stepping techniques. The drawback is that a functional full-regression model between continuous curves is difficult to implement and requires introducing and fine-tuning additional parameters. To avoid these problems we proceed to a functional reduction of the problem dimensionality.

Functional reduction of the dimensionality

We reduce the dimension of the response spaces by means of Functional Principal Component Analysis (FPCA, Henderson 2006), which is a rather straightforward functional extension of standard PCA. Beside the indubitable computational advantages, low-dimensional spaces allow us to visualize the most relevant modes that describe data variability and help us to evaluate the suitability of the proxy model for the quantities of interest. FPCA is applied separately to the two sets of exact and proxy responses. The dimensionality of the response spaces is reduced considering only the first D harmonics, where D is chosen to achieve the desired degree of accuracy.

Although FPCA offers an optimal dimensionality reduction with respect to the total mean squared error, any rotation of the basis preserves the accuracy. The choice of a proper rotation of the basis might allow a better interpretation of the data (Richman, 1986; Ramsay *et al.*, 2009). Therefore, we use the *varimax* algorithm (Kaiser, 1958) to find an appropriate rotation. As a results, each proxy response is approximated by projection on the rotated FPCA basis as

$$x_i(t) \approx \tilde{x}_i(t) = \bar{x}(t) + \sum_j^D b_{ij}\zeta_j(t), \tag{4.2}$$

where $\bar{x}(t)$ is the mean curve, and

$$b_{ij} = \int [\bar{x}(t) - x_i(t)]\zeta_j(t)dt \tag{4.3}$$

is the projection of the deviation from the mean of the ith proxy curve on the jth rotated harmonic $\zeta_j(t)$. Following the same procedure, the N exact responses in the learning set are

approximated as

$$y_i(t) \approx \tilde{y}_i(t) = \bar{y}(t) + \sum_j^D c_{ij}\eta_j(t), \tag{4.4}$$

where $\bar{y}(t)$ is the mean exact response, $\eta_j(t)$ the jth harmonic of the (varimax) rotated orthonormal basis $\{\eta_j(t)\}_{j=1,\ldots,D}$, and

$$c_{ij} = \int [y_i(t) - \bar{y}(t)]\eta_j(t)dt \tag{4.5}$$

the score with respect to $\eta_j(t)$.

Regression and error model

The relationships between the two sets of curves in the learning set approximated is investigated by considering the first D harmonics, $\{\tilde{x}_i(t), \tilde{y}_i(t)\}_{i=1,\ldots,N}$. As sketched in figure 4.4, the goal is to find a mapping from the space of proxy responses onto the space of exact responses that allows us to predict the exact responses for the realizations that do not belong to the learning set (hence, without actually solving the exact model). This is commonly referred to the model's predictive ability.
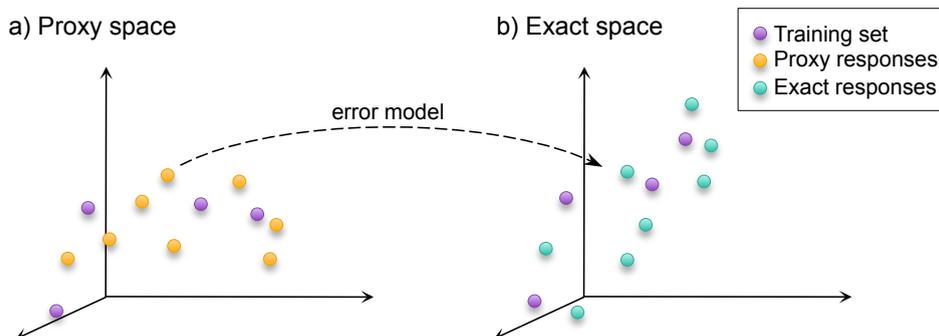
Figure 4.4: A statistical model is built on the learning set to relate the coefficients of the elements $x_i(t)$ in the proxy space to the coefficients of the elements $y_i(t)$ in the exact-model space. It is used as an error model to predict the exact response from the proxy response.

Here, we restrict ourselves to functional linear regression models that minimize the l_2-norm of the residuals

$$\varepsilon_i = y_i - \hat{T}(x_i) \qquad i \in [1, \ldots, N], \tag{4.6}$$

where $\hat{T}$ is the estimator on the learning set. Training such a functional linear model in full

generality is not straightforward, but we can take advantage of the FPCA basis to define a multivariate multiple regression problem of the form (Hastie *et al.*, 2009; Fox and Weisberg, 2011; Weisberg, 2014)

$$c_{ij} = \beta_{0j} + \sum_{l=1}^{D} b_{il}\beta_{lj} + e_{ij} \quad (i,j) \in [1,N] \times [1,D], \tag{4.7}$$

where β_{lj} are the coefficients of the regression, and e_{ij} are the errors, which we assume to be Gaussian with variance σ_j^2.

A further simplification is obtained by splitting the regression model into D independent problems of the form

$$c_i^{(j)} = \beta_0^{(j)} + \sum_{l=1}^{D} b_{il}\beta_l^{(j)} + e_i^{(j)}. \tag{4.8}$$

This simplification does not affect the operator estimators, which are identical for the problems in Eqs. 4.7 and 4.8, i.e., $\hat{\beta}_{jl} = \hat{\beta}_l^{(j)}$. However, confidence bands of the multivariate regression model cannot be directly derived from those obtained for the regressions in equation 4.8, which complicates their derivation (Josset *et al.*, 2015).

Prediction of exact response from the proxy response

The regression model can be used to predict the exact response of any new realization r for which the proxy response $\tilde{x}_r(t)$ is known. Indeed, the estimator of the linear regression model allows us to predict the scores of exact response curve, $\hat{c}_{rj}$, without solving the exact model. Therefore, solely on the basis of the scores of the proxy responses, b_{rl}, we can estimate the exact response as

$$\hat{y}_r(t) = \bar{y}(t) + \sum_{j=1}^{D} \hat{c}_{rj}\eta_j(t), \tag{4.9}$$

where

$$\hat{c}_{rj} = \hat{\beta}_{0j} + \sum_{\ell=1}^{D} \hat{\beta}_{j\ell}b_{r\ell}, \tag{4.10}$$

are the estimates of the exact scores predicted by the error model.

4.4.2 Two-stage MCMC

Two-stage MCMC has been introduced by (Christen and Fox, 2005; Efendiev *et al.*, 2005, 2006) to improve the acceptance rate of the Metropolis-Hastings algorithm (MH). For optimal convergence conditions of standard MCMC algorithms it is necessary to tune the random-walk step of the chain in order to obtain an acceptance rate between 20% and 50%. As flow simulations are performed at each step to compute the likelihood $\mathcal{L}$ of the proposed sample ϕ, the low acceptance rate implies that 50% to 80% of the flow simulations are performed on rejected samples and do not contribute to the posterior distribution.

Moreover, in order to satisfactorily explore the prior space under the constraint of limited computer resources, the length of the random-walk step is often increased with the result that the acceptance rate is drastically reduced (for instance, an acceptance rate around 10^{-5} is reported by Efendiev *et al.* (2005)).

The goal of two-stage MCMC is to decrease the computational cost by reducing the number of full-physics flow simulations that are performed on rejected samples. This is achieved by employing an approximate model to identify samples in low likelihood regions that might be rejected and avoid running the exact simulator on these samples and at the same time to identify the samples that are more likely to be accepted by the exact model. Proposing samples that are more likely to be accepted at the second stage will eventually boost the acceptance rate. In other words, the approximate likelihood $\tilde{\mathcal{L}}$ of the proposed sample ϕ is estimated by using the approximate model response, $\hat{y}_\phi(t)$, from which the first-stage acceptance,

$$\tilde{\alpha} = min\left\{1, \frac{\tilde{\mathcal{L}}(\hat{y}_\phi(t))}{\tilde{\mathcal{L}}(\hat{y}_\theta(t))}\right\}, \tag{4.11}$$

is computed. If the sample is accepted, the response of the exact model, $y_\phi(t)$, is calculated to compute the exact likelihood $\mathcal{L}(y_\phi(t))$ and the proposal is tested again using a modified acceptance/rejection condition

$$\tilde{\alpha} = min\left\{1, \frac{\mathcal{L}(y_\phi(t))}{\mathcal{L}(y_\theta(t))}\frac{\tilde{\mathcal{L}}(\hat{y}_\theta(t))}{\tilde{\mathcal{L}}(\hat{y}_\phi(t))}\right\}. \tag{4.12}$$

A schematic diagram of the two-stage MCMC algorithm is depicted in Figure 4.3.

Efendiev *et al.* (2006) demonstrated that the two-stage MCMC converges to the true posterior distribution under two mild assumptions: first, the proposal distribution has to satisfy $q(\phi, \psi) > 0$ for any (ϕ, ψ) in the posterior distribution; second, the support of the exact pos-

terior distribution belongs to the support of the approximate distribution (see theorem 3.2 in (Efendiev *et al.*, 2006)).

The first condition is easily satisfied when a Gaussian random walk is used as proposal distribution: a step size sampled from a normal distribution guaranties that $q(\phi, \psi) > 0$ for any (ϕ, ψ). The second condition is met assuming a Gaussian error model for the likelihoods for both proxy, $\hat{y}_\phi$, and exact, y_ϕ, solutions, i.e.,

$$\tilde{\mathcal{L}} \propto \exp(-\frac{||y_{obs} - \hat{y}_\phi||^2}{\sigma_{app}^2}) \qquad \text{and} \qquad \mathcal{L} \propto \exp(-\frac{||y_{obs} - y_\phi||^2}{\sigma_{ex}^2}), \qquad (4.13)$$

respectively. The likelihoods distributions are non-compact, and thus the supports of both posterior distributions are identical to the one of the prior distribution.

Numerically, it is probable that the likelihood values are very close to zero, which prevents the chain to reach all regions of the parameter space. However, under the condition that the exact and approximate misfits are correlated, Efendiev *et al.* (2006) have shown that it is possible to choose σ_{app} such that the second assumption is verified and that the optimal acceptance rate can be obtained by setting σ_{app}^2 to σ_{ex}^2/α_o, if the correlation can be described by a linear relationship

$$||y_{obs} - y_\phi||^2 \approx \alpha_0 \cdot ||y_{obs} - \hat{y}_\phi||^2 + \alpha_1. \qquad (4.14)$$

4.5   Application to the IC Fault test case

In this section, we first assess the performance of the functional error model to satisfactorily describe the misfit between the proxy and the exact models for the ICF test case. Then, the proxy (corrected by the error model) is used as first-stage solver in two-stage MCMC, and the results are compared with a pure Metropolis-Hastings approach in order to illustrate the potential of error modeling in the context of Bayesian inference.

4.5.1   Error model

The objective of functional error modeling is to correct the proxy response to estimate an unbiased exact response. The first step is to choose an appropriate proxy that is sufficiently informative of the behavior of the exact model but considerably cheaper in terms of computational cost.

Choice of proxy model

Here, we are interested in sampling the space of the parameters that describe the permeability field, while the properties of the fluids and the physical processes are known. We consider the simultaneous flow of two immiscible liquids that form two separate phases (oil and water) and we are interested in the production rates of both fluids. Under these conditions, the fluid transport is governed by a set of coupled nonlinear equations, which complicates the numerical solution of the equations. The high degree of coupling between the pressure and the saturation equations renders the transport problem computationally expensive.

A natural choice of proxy is to neglect the nonlinearity of the permeabilities and the two-way coupling between the equations by solving a simple tracer transport problem. This means using a single phase solver as a proxy for a two-phase solver. Further simplifications are introduced by neglecting capillarity and gravity, so that the pressure equation has to be solved only once per proxy simulation.

Construction of the learning set

The construction of the learning set requires making choices on the method of selection and on the size of the set. Here, we train the error model on a subset of 100 realizations selected by performing a Latin hypercube sampling in the 3D parameter space. The learning set consists of two pairs of curves par realization: water and oil production rates obtained with the proxy and the exact models. Comparison with other sampling techniques and learning-set sizes has indicated that the effects of these variables on the error model is limited. Additional tests (not reported here) have suggested that 20 realizations might be sufficient to obtain a satisfactory error model, but with such few realizations the performances would vary greatly from one learning set to another. The choice of a subset of 100 realizations has been made for the sake of robustness. The proxy and exact curves in the learning set are plotted in figure 4.5.a.

Dimensionality reduction and interpretation of the information

For each realization in the learning set we have four subspaces of response curves: the spaces of the proxy and exact production rates of water and oil. For each subspace, we subtract the average response from each response curve and then apply FPCA to obtain a basis of the subspace. To reduce the dimensionality of the problem we truncate the basis by considering only the first three functional principal components, which capture more than 96% of the
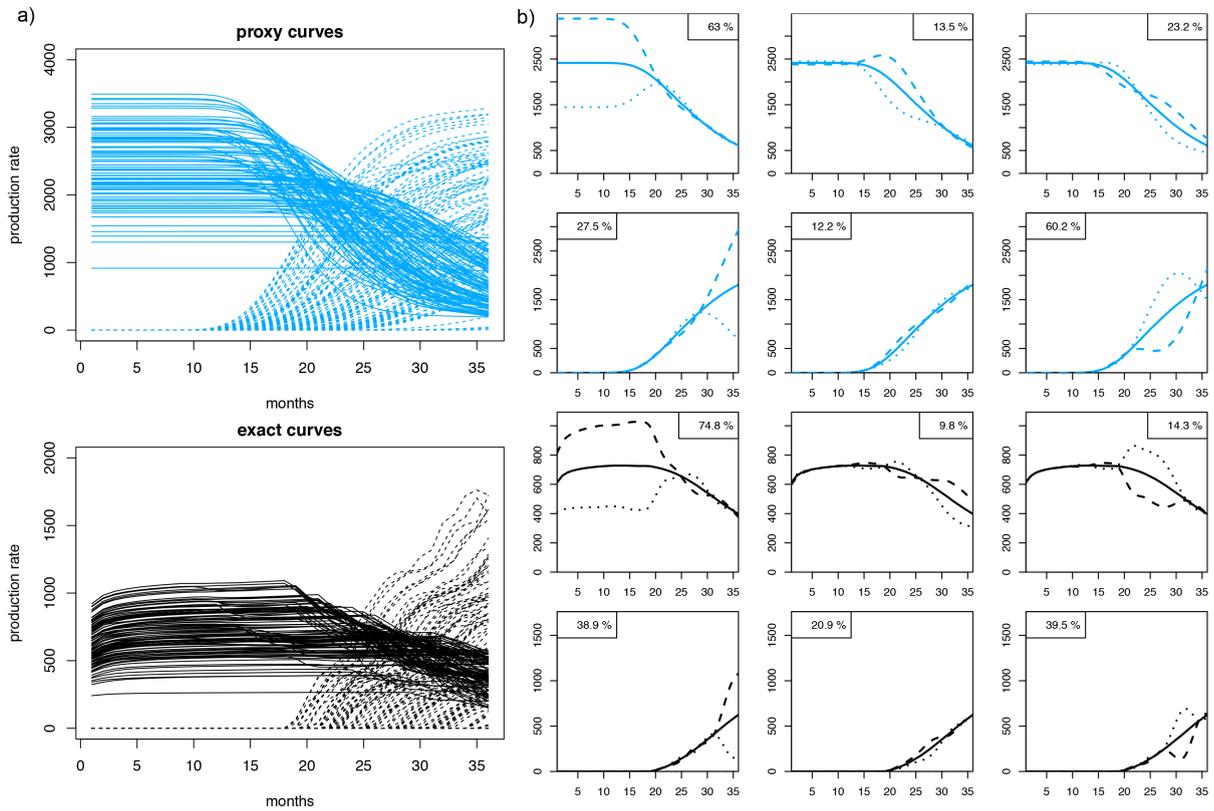
Figure 4.5: a) The learning set of curves is constructed by running both proxy (top) and exact (bottom) models on the sampled geostatistical realizations. The production rates of oil (full lines) and water (dashed lines) are plotted in bbl/day in function of time. b) The three first rotated functional principal components (harmonics) extracted from the learning set are represented here for the two sets of pairs of production rate curves. The solid lines are the mean curves and the dotted lines represent the variability around the mean described by the corresponding harmonic. The legends report the percentage of the total variability, which is explained by each harmonic.

variability within the learning set.

By close inspection of the rotated harmonics (figure 4.5.b), we notice that the first principal component captures the variability of the initial plateau of oil production rate (i.e. prior to the water breakthrough, figure 4.5.a bottom). The second harmonic of the proxy and the third harmonic of the exact model describe the production drop after water breakthrough. The third harmonic of the proxy and the second harmonic of the exact model capture the remaining late-time variability. A similar analysis can be done for the harmonics of the water production rate curves. The first harmonics (both of the exact and proxy models) explain the variability at the end of the simulation time, the second harmonics capture small variabilities

83

at the water breakthrough time, and the third harmonics describe most of the variabilities occurring at intermediate time between the water breakthrough and the end of the simulation.

Evaluation of the informativeness of the proxy and self-consistency of the error model

After the dimensionality reduction, each functional space has a six-dimensional basis (three harmonics for the water production and three harmonics for the oil production). In addition to decreasing the computational cost of constructing the error model, the reduction to six dimensions facilitates a visual inspection of the relationships between proxy and exact curves, providing insight into whether the proxy response is informative of the full-physics response.

Figure 4.6.a) plots the one-to-one relationship between the scores (i.e., the projections on the harmonics) in the proxy space versus the scores in the exact space. A clear linear relationship can be observed in the upper-left plot, which illustrates the relationship between the first harmonics of the oil production. This indicates that the height of the plateau of the exact oil-production curves is well explained by the proxy plateau. On the other hand, the second harmonic of the proxy oil curves (plots in the second column) does not display a simple relationship with any harmonic of the exact curves. Also, the second and third harmonics of the exact oil-production curves do not display a simple relationship with any of the proxy harmonics (second and third rows). This indicates that the proxy is not very informative of the features described by the second and third harmonics of the exact oil curves and one can expect that the error model will be less accurate in predicting those harmonics.

The error model maps the space of the proxy responses onto the space of the exact responses and it is constructed by solving six independent linear regression models as explained in section 4.4.1. Figure 4.6.b) shows the correlation between exact scores and the scores predicted by the error model (in the space of the exact curves) for all the 100 realizations of the learning set. As expected, the projection on the first oil-production harmonic, which describes the plateau at early time, is well predicted with an R^2 value of 0.91. The projections on the second and third harmonics are predicted with lower accuracy ($R^2 = 0.77$ and 0.79, respectively). The water-production scores are rather well predicted with R^2 values around 0.9. The underestimation of the largest score values for the first and the second harmonics of the water production rates (figure 4.6.b) demonstrates the limitation of the linear model. Indeed, as the proxy curves are always positive, not all scores values are possible. In particular, for the second water

harmonic (figure 4.6.a), a clear lower bound in the exact scores is displayed and biases the linear regression.
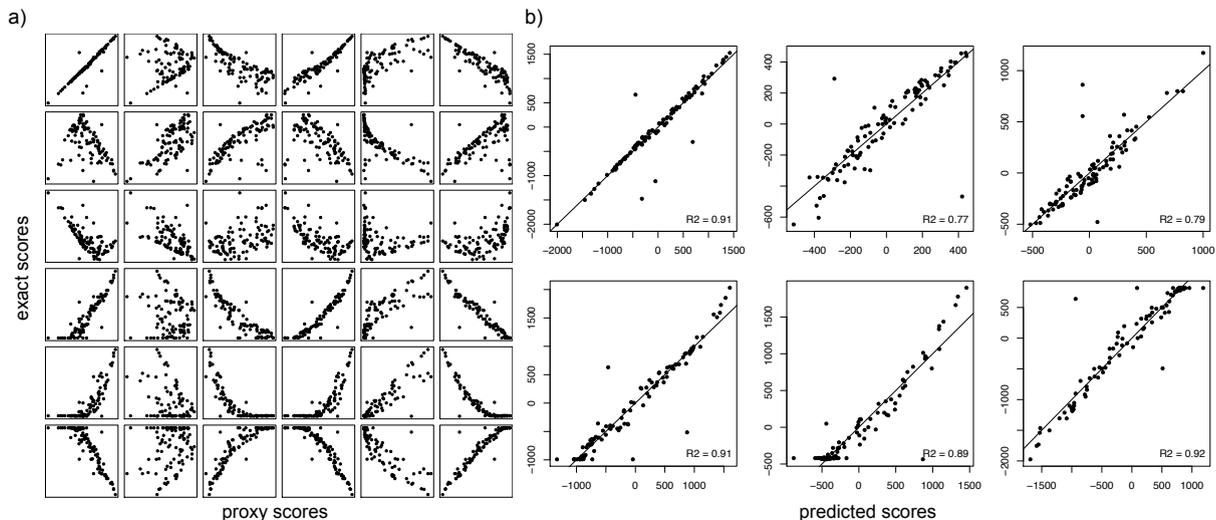
Figure 4.6: a) Dependency between exact and proxy scores. The scores of the first three harmonics of the exact oil production rate $\{\eta_i^o(t)\}_{i=1,2,3}$ and water production rate $\{\eta_i^w(t)\}_{i=1,2,3}$, are plotted as function of the scores of the proxy curves with respect to the harmonics $\{\zeta_i^o(t)\}_{i=1,2,3}$ and $\{\zeta_i^w(t)\}_{i=1,2,3}$. b) Results of the linear model: the exact scores are plotted as function of the predicted scores; also shown is the identity line. Both plots are helpful to assess whether the linear regression model is appropriate to describe the relationship between proxy and exact scores, thus the level of informativeness of the learning set.

Evaluation of predictive power of the error model

For a new point in the parameter space, the corresponding realization is built and the proxy model is run. Then, from the output of the proxy model (i.e., the time-discrete recovery rates resulting from the numerical simulations), continuous oil and water production rates are reconstructed and projected on the harmonics. The proxy scores are used as input of the error model, which allows prediction of the corresponding exact scores that are used to reconstruct the two-phase response curves.

In order to evaluate the performance of the error model, proxy and exact simulations were run for a test set of 1000 realizations sampled in the entire parameter space by means of Latin Hyper Cube sampling. Figure 4.7 compares the exact responses with the predicted responses for four points sampled in the parameter space. Figure 4.8.a) plots the error of the prediction as a function of time. The error of the mean of the predicted curves is very close to zero for
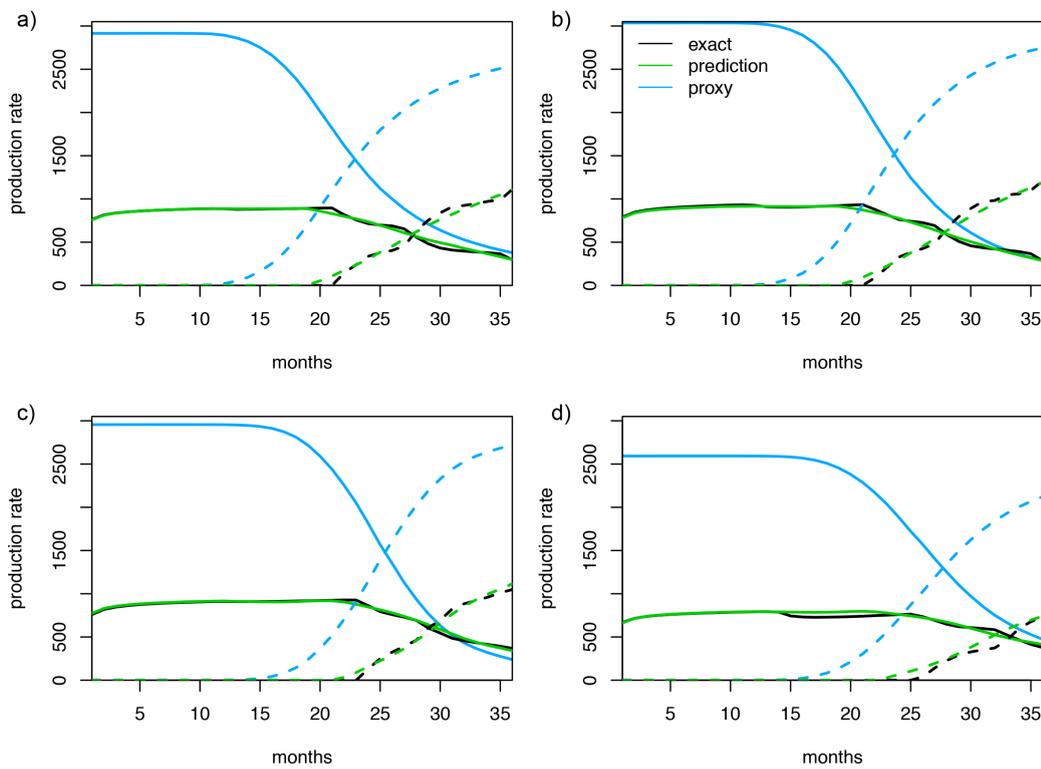
Figure 4.7: Four predictions that are representative in term of l_2 error norms: a) and b) have errors close to the median, c) to the 25% percentile, and d) to the 75% percentile. The continuous lines are the oil production rates, the dashed lines the water production rate. The proxy curves (blue) are effectively corrected by the error model and the predicted curves (green) match well the exact curves (black).

both the oil and the water production rates, which indicates that the predicted mean is not biased. The histograms in figure 4.8.b) show the distribution of the l_2 and l_∞ error norms. On average, the maximum error made is around 80bbl/day for oil and 180bbl/day for water, respectively.

In the context of Bayesian inference, a correct prediction of the misfit to the observed data is crucial. Figure 4.8.c) illustrates the correlation between the misfit computed from the predicted curves and the misfit computed from the exact curves for the observational data shown in figure 4.1. The overall correlation between the exact and predicted misfits is good as indicated by the high correlation coefficients in R^2. Therefore, the prediction model is expected to be efficient at rejecting realizations. However, for small misfits (i.e., for realizations whose responses deviate less from data) the error model is less accurate and tends to overestimate the misfit. This explains the lower Kendall correlation coefficient (a measure of rank correlation)

with respect to the Pearson coefficient (a measure of the degree of linear dependence).
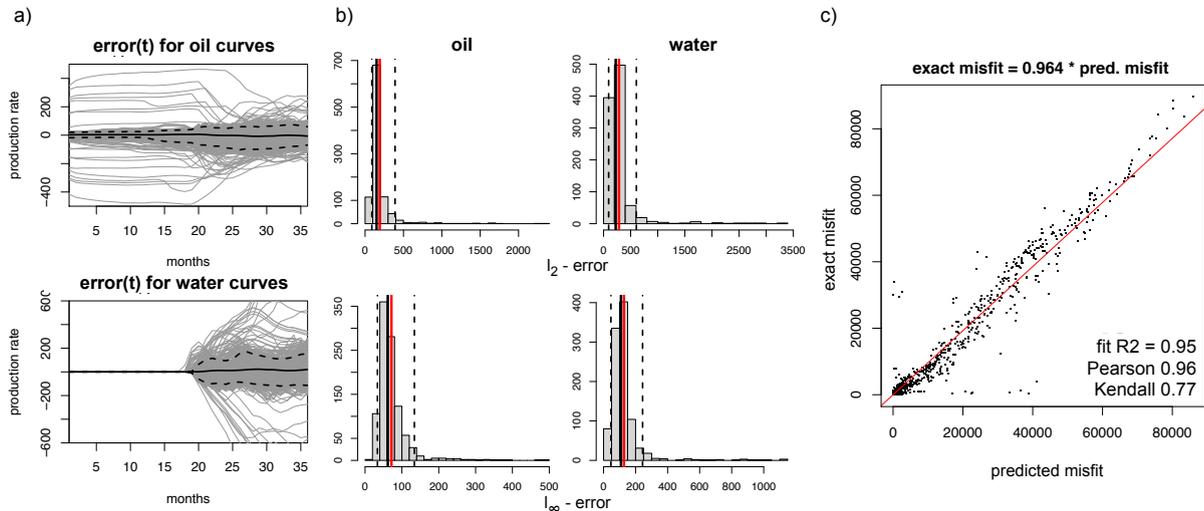
Figure 4.8: The quality of the error model is evaluated on a test set of 1000 new realizations. a) Difference between production rates predicted with the error model and the exact production rates (grey curves) for the oil (top) and the water (bottom). b) Histograms of the l_2 and l_∞ error. c) Exact misfit versus predicted misfit with respect to the observations (the identity line is plotted in red); the R^2, Pearson and Kendall correlation coefficients are reported to indicate the quality of the prediction.

4.5.2 Two-stage MCMC

In this section, we first introduce the definition of the misfit necessary to compute the like-lihoods in Eq. 4.13; then we investigate the fidelity of the response surface predicted by the error model; and finally we show that a two-stage MCMC set-up is able to explore a larger portion of the parameter space than MH at the same computational cost, which can be a substantial advantage for challenging problems as the ICF test case.

Definition of the misfit and response surfaces

Here we employ the definition of the misfit that is commonly used to investigate the ICF test case, i.e.,

$$\mathcal{M}_j = \sum_{i=1}^{36} \frac{(C_o^j(t_i) - C_o^{ref}(t_i))^2}{\sigma_o^2(i)} + \sum_{i=27}^{36} \frac{(C_w^j(t) - C_w^{ref}(t))^2}{\sigma_w^2(i)} \tag{4.15}$$

where $\sigma_o(i) = 0.03 \cdot C_o^{ref}(t_i)$ and $\sigma_w(i) = 0.03 \cdot C_w^{ref}(t_i)$. The likelihood is then obtained from the misfit as $\mathcal{L} = \exp(\mathcal{M}_j)$. Notice that only the water production rate at later time ($i \geq 27$)

contributes to the misfit.

The three first original papers on ICF (Tavassoli *et al.*, 2004, 2005; Carter *et al.*, 2006) have employed a slightly different definition of the misfit, which considers the contribution of the water production rate at any time (i.e., with $i = 1$ instead of $i = 27$ in the second summation in Eq. 4.15). However, this choice leads to a very discontinuous response surface, for which hardly any method beside classical Monte Carlo would be able to provide a reasonable solution. The modified misfit function defined in equation 4.15 has been introduced to make the problem more tractable and is commonly used in all investigations of the ICF test case.

Comparison of the response surfaces

To further assess the performance of the error model, figure 4.9 compares the 1D response surface of the misfit of both the exact model and the prediction given by the error model, as a function of the fault-throw value. The response surface of the exact model exhibits several local minima separated by large misfit regions. This situation is particularly challenging for any MCMC approach because many realizations are required to cross large misfit regions with small random-walk steps.
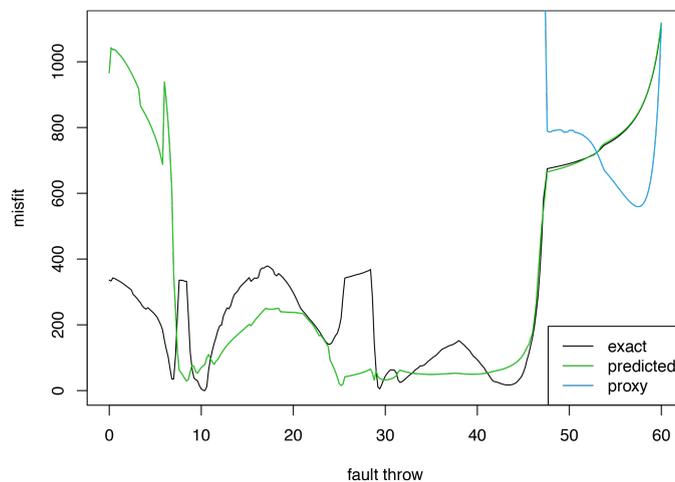
Figure 4.9: The 1D response surface of the ICF problem for the misfit definition given in equation 4.15. K_{high} and K_{low} are set to the reference values, while the fault throw varies between 0 and 60 feet. Shown are the response surfaces obtained from the exact model (black), from the responses predicted by the error model (green), and from the proxy curves alone (blue).

The predicted response surface (which provides the basis of the first-stage rejection deci-

sion) is in excellent agreement with the exact response surface for $h > 48$ft. For a fault throw between 8 and 48ft, the discrepancies between the two curves are more important, but the main features of the curves are reproduced. We can expect that the low misfit values of the predicted response curve will be able to guide the chain into this region. For values between 0 and 8 feet, the misfit is greatly overestimated but the shape of the curve is reproduced. If inference is made only based on the prediction model, the minimum around 7ft would not be identified. However, in a two-stage set-up the relative values of the misfit are more relevant than the absolute values.

An error model that predicts a response surface that roughly preserves the shape of the exact surface may be sufficient to drive the chain to minimum misfit regions at a lower computational cost than it would be possible with the exact model alone. Sharp misfit contrasts, as the one observed around 8ft, might impair the mobility of the chain, preventing the exploration of the entire parameter space. Note, however, that in multidimensional spaces (e.g., in the full 3D parameter space of the ICF test case) sharp contrast might be less problematic than in 1D, because the higher dimension might allow the chain to bypass the misfit peak.

MCMC results

In a MCMC set-up the choice of proposal distribution is crucial. To obtain optimal convergence of the chain, the acceptance rate should be in the range between 20% and 50% (see Sec. 4.4.2). This is achieved by tuning the standard deviation of the random walk, which is defined as

$$
\begin{aligned}
h^{(i+1)} &= h^{(i)} + s_h \cdot \delta_h{}^{(i)}, & \delta_h &\sim \mathcal{N}(0, \sigma^2) \\
K_h^{(i+1)} &= K_h^{(i)} + s_{K_h} \cdot \delta_{K_h}{}^{(i)}, & \delta_{K_h} &\sim \mathcal{N}(0, \sigma^2) \\
K_l^{(i+1)} &= K_l^{(i)} + s_{K_l} \cdot \delta_{K_l}{}^{(i)}, & \delta_{K_l} &\sim \mathcal{N}(0, \sigma^2)
\end{aligned}
\tag{4.16}
$$

where σ is the standard deviation of the random walk; and s_h, s_{K_h}, and s_{K_l} are the scaling factors ensuring that each prior is visited at the same rate. To determine the standard deviation that corresponds to the optimal acceptance rate for Metropolis-Hastings algorithm we have launched several chains of 1'000 iterations with different standard deviations, and found an optimal value $\sigma = 5 \cdot 10^{-3}$.

First, we compare three MH chains with three two-stage MCMC chains. All chains are launched with the optimal value $\sigma = 5 \cdot 10^{-3}$ and have a length of 10'000 iterations. The statistics of the chains are reported in table 4.1. A representative example of chain is plotted

for each of the two methods in figure 4.10 (first and fourth columns). The acceptance rate of MH is approximately in the optimal interval, ranging from 14% to 36%, whereas for two-stage MCMC we obtain a slightly suboptimal acceptance rate, which ranges from 8% to 23%. In all cases the chains have been able to explore only a limited portion of the parameter space, despite a length of 10'000 iterations.

In order to enlarge the portion of the parameter space that is explored, we multiply the standard deviation of the random walk by a factor 5 ($\sigma = 1 \cdot 10^{-2}$) and 10 ($\sigma = 5 \cdot 10^{-2}$), we launch again three chains for both values of σ. The length of the MH chains remains fixed to 10'000 iterations, whereas the length of the two-stage MCMC chains is chosen to approximately match the computational cost of the MH chains. (This is done assuming that the computational gain of the proxy with respect to the exact model is equal to the number of time steps per simulation, which is about 43). The statistics of MH and two-stage MCMC chains with the modified parameters are reported in table 4.1, and two examples of chains are shown in figure 4.10. The MH chains acceptance rate drops from an average 23% for $\sigma = 5 \cdot 10^{-3}$ to 11% and 1% for $\sigma = 1 \cdot 10^{-2}$ and $\sigma = 5 \cdot 10^{-2}$, respectively.

In addition to the fact that these values are not optimal for convergence, the low acceptance rate implies that many of the full-physics simulations are run without providing any information gain, thus wasting computational resources. One of the main results of the work is that, at approximately the same computational cost, the two-stage MCMC set-up allows us to increase the acceptance rate by a factor 1.5 to 4 (the average acceptance is 16% and 4.5% for $\sigma = 1 \cdot 10^{-2}$ and $\sigma = 5 \cdot 10^{-2}$, respectively, see table 4.1). Moreover, as the proxy model is much cheaper than the exact model, two-stage MCMC chains reach lenghts of about 15'000 and 30'000 iterations (which corresponds to an increase in length of a factor 1.5 to 3) and allows a larger portion of the parameter space to be sampled.

While those results are very promising, none of the two-stage MCMC chains visited the reference point. The reference point was visited only by one of the MH chains, which was randomly initialized very close. Overall, this test case remains very challenging for single chain MCMC set-up and multiple chains solutions (Mohamed *et al.*, 2012) should be considered.
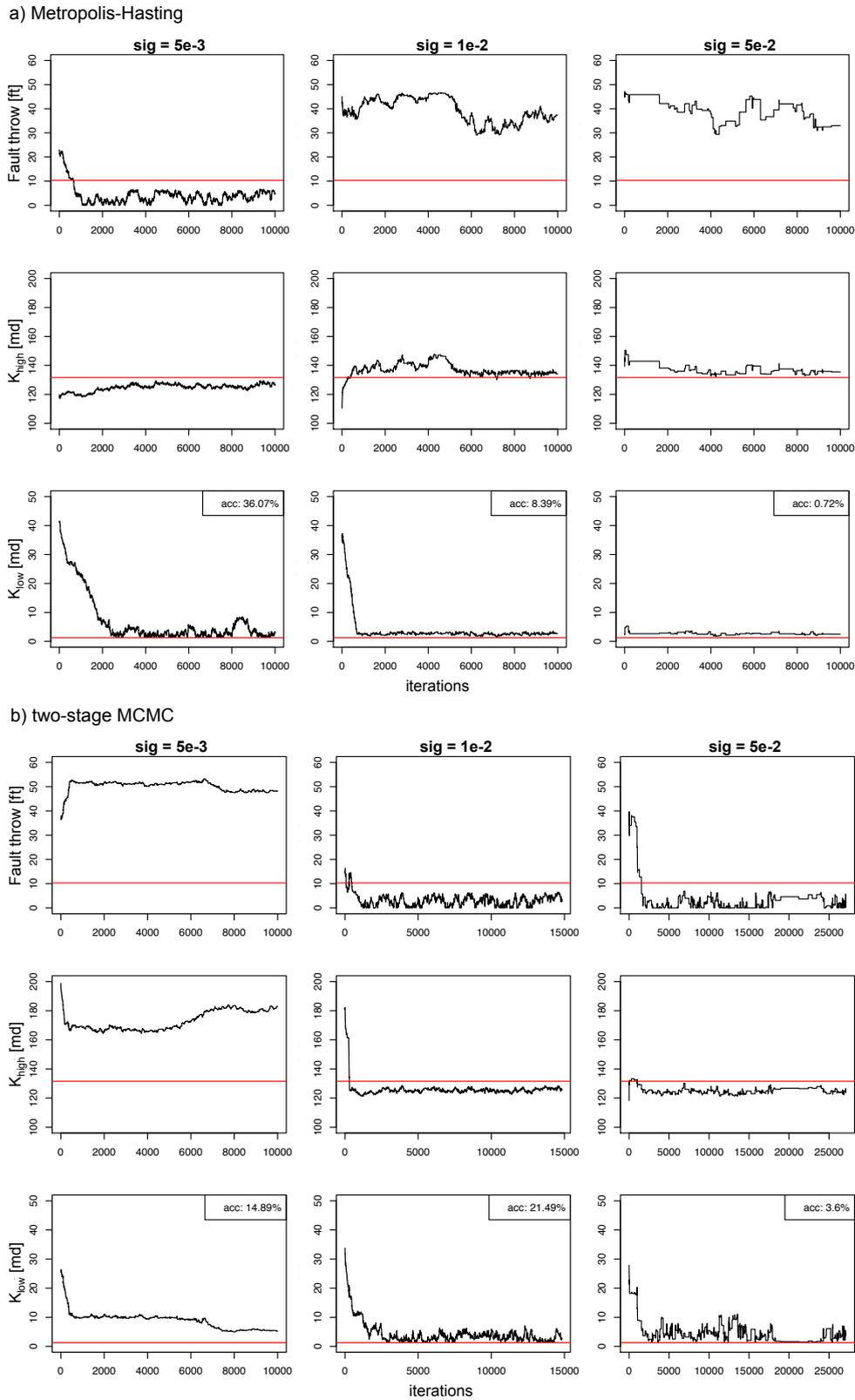
Figure 4.10: The chains are represented by their movements in the parameter space (vertically h, K_{high} and K_{low}) in function of iterations. For each of the three values of the random walk step length σ, one Metropolis-Hasting chain and one two-stage MCMC are plotted. The acceptance rates indicated in the legends are improved for the two-stage MCMC chains when σ is increased, allowing for much longer chains at the same computational cost.

random walk	number of iterations			number of accepted simulations 1st stage			2nd stage			acceptance rate			
σ	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	mean
					Metropolis-Hasting								
$5 \cdot 10^{-3}$	10'000	10'000	10'000				1'631	3'247	1'291	18.1%	36.1%	14.3%	22.8%
$1 \cdot 10^{-2}$	10'000	10'000	10'000				1'683	755	628	18.7%	8.4%	7.0%	11.4%
$5 \cdot 10^{-2}$	10'000	10'000	10'000				179	65	48	2.0%	0.7%	0.5%	1.1%
					Two-stage MCMC								
$5 \cdot 10^{-3}$	10'000	10'000	10'000	4'760	5'299	176	367	789	41	7.7%	14.9%	23.3%	15.3%
$1 \cdot 10^{-2}$	14'372	14'815	31'738	9'666	9'656	7'820	2'060	2'075	331	23.3%	21.5%	4.2%	16.3%
$5 \cdot 10^{-2}$	28'337	31'777	27'108	9'341	9'261	9'370	393	518	337	4.2%	5.6%	3.6%	4.5%

Table 4.1: Results of Metropolis-Hasting and two-stage MCMC algorithms for three chains (C1, C2, and C3): the standard deviation of the random walk, σ; number of iterations (i.e. total length of the chain); the number of accepted simulations at the first-stage; the number of accepted simulations at the second-stage; and the acceptance rate (i.e., the ratio of accepted exact simulations to the number of exact simulations that have been performed).

4.6 Conclusions

We have investigated the potential of using error models in the context of Bayesian inference. The error model is used to map a proxy model response into the response of the exact model, which can be predicted without actually solving the exact model, thus reducing the computational costs. This methodology was applied to the ICF benchmark test case, which is geometrically simple yet very challenging. The ICF problem is particularly arduous for MCMC methods, because the very intricate surface response, characterized by sharp misfit contrasts, makes it very difficult, if not impossible, to explore the whole space by a single chain at tractable computational costs.

We have compared the performance of classic Metropolis-Hasting chains with a method that couples our error model with a two-stage MCMC algorithm. The use of the error model has increased the acceptance rate of the realizations for which the exact model was run (from 11% to 16% and 1% to 4% for $\sigma = 1 \cdot 10^{-2}$ and $\sigma = 5 \cdot 10^{-2}$, respectively). This has allowed the chain length to be increased up to a factor three with respect to MH at comparable computational costs, potentially permitting us to explore a larger portion of the response space. Based on the results of the few chains reported, it remains unclear whether the decreased computational costs might be sufficient to guide the chain out of areas of local minima, in which MCMC chains remain systematically trapped regardless of the random walk standard deviation σ that is employed. Most likely, this problem will not be solved for irregular response surfaces as the one of the ICF test case. However, the use of an error model can be greatly

beneficial also for multiple-chain algorithms that can be set up to overcome this issue.

We have demonstrated that the relationship trained on the learning set is quite effective in predicting the exact responses, as it is indicated by the correlation indices and by the linear relationships between the exact and predicted misfits. The error model has been very successful to reject bad samples, but slightly less informative to predict the response of the best samples (i.e., for realizations in regions of low misfit).

Notice that the use of the proxy without error model would be very inefficient as first-stage selection criterion. This is evident from simple inspection of the proxy misfit in figure 4.9: the regions of good-quality parameters cannot be identified on the basis of the proxy misfit alone. The error model is thus critical to guide the simulations in the correct regions of the parameter space, avoiding that the two-stage MCMC approach results in a counter-productive increase of simulations in poor quality regions, thus heavily increasing the computational effort.

The question that arises naturally is whether the quality of the proxy is relevant in presence of such an effective error model. To investigate this, we used the input parameters of the model (i.e., the permeabilities of the two facies and the fault throw) as proxy, that is, we directly constructed a regression model between the input parameters and the scores of the exact responses on a learning set. In this case, we have observed a total absence of relationship. This demonstrates that, despite its simplicity, the single-phase proxy provides important information on the connectivity that results from the combined effect of the parameters.

Several improvements can be devised within the framework proposed here. In particular, more complex (nonlinear) regression models could be considered (e.g., by using of kernels) and appropriate data transformations could be employed to avoid unphysical results after correction of the proxy responses, as proposed in Josset and Lunati (2013). In terms of computational cost, a major improvement could be achieved by taking advantage of all the simulations performed along the MCMC chains and iteratively updating the error model as soon as new samples are evaluated (Cui *et al.*, 2011). This option, however, would require overcoming the problem that the likelihood is modified and convergence is not guaranteed. Several alternative approaches to MCMC could also be considered jointly with the error model, e.g., the Nested Sampling (Skilling *et al.*, 2006; Elsheikh *et al.*, 2014) in which resampling is performed at the prior level. In such approaches, the error model would be useful to reject sampled points and the Nested Sampling would avoid entrapments in the inherent structure

of the ICF while allowing an iterative update of the regression model.

Acknowledgments

Chapter 5

Iterative construction of error models to correct proxy responses

Laureline Josset, Camille Kerboas and Ivan Lunati

5.1 Introduction

Stochastic approaches are common practice to represent the uncertainty in the many subsurface properties (Kennedy and O'Hagan, 2001; Wu and Zeng, 2013). Once a set of property realizations is defined, the uncertainty on the quantity of interest is estimated by performing flow simulations, and summary statistics (e.g., point-wise quantiles such as P10, P50 and P90) are used to facilitate interpretation. This process is called uncertainty propagation. While simple in concept, the approach is heavily limited by the associated computational cost (Dagan, 2002; Rubin, 2004; Renard, 2007; Wu and Zeng, 2013). An accurate description of the subsurface uncertainties requires a large number of realizations, and thus many flow simulations have to be performed.

This computational issue was first addressed by resorting to ranking methods (Ballin *et al.*, 1992), where an estimate of the uncertainty is provided by solving the flow for a few realizations that are close to the quantiles that need to be calculated. The performance of ranking methods depends greatly on their ability to identify the relevant realizations. This led to the development of dynamic ranking methods (McLennan and Deutsch, 2005), where a simplified flow solver (or "proxy") is first used to sort the realizations and thus help selecting of the realizations.

More recently, Scheidt and Caers (2009a) have introduced the Distance Kernel Method, where a simplified solver is used to identify a subset of realizations, representative of the ensemble, by using a k-medoids clustering algorithm (Hastie *et al.*, 2009). The point-wise quantile curves are computed based on the exact responses weighted by the size of the clusters (see section 2.4.2 for a detailed description).

The quality of the estimates obtained by dynamic ranking methods and DKM is determined by the choice of the simplified solver, which should balance reduction of computer costs and accuracy of the estimate of the exact simulation. The methodologies proposed in chapters 2 and 3 follow the same approach as DKM but use the subset of exact responses to construct an error model and correct the proxy responses. The predicted curves are then used to compute the corresponding quantiles. Beside an increase in robustness, the error model described in chapter 3 offers an additional advantage : the informativeness of the proxy is evaluated during the construction of the error model.

In the present paper, we propose to combine the concept of error model with an iterative scheme to improve the estimation of the quantile curves. At each iteration, the current error

model is constructed and the predicted responses are used to identify the realizations that are of interest to compute the summary statistics. The realization that is the closest to the computed quantile is added to the learning set of realizations, then the exact simulation is performed for this realization and is used to update the error model. This allows the iterative refinement of the error model around the quantile curves. The proposed methodology is useful beyond the context of uncertainty propagation. In a model calibration procedure, for instance, one is interested in sampling realizations that match the observation data. In this context, the error model is used to identify the realizations that are the closest to the observed curve. As realizations are added, the error model becomes more accurate around the data.

The paper is structured as follows. We start by presenting the proposed methodology (section 5.2). Two applications are then explored: first, we consider the problem of the prediction of an hydrocarbon plume potentially polluting a drinking source as in chapter 3 (section 5.3), then, we investigate the challenge of predicting saline intrusion in a coastal aquifer (section 5.4). For both applications, two problems are considered: uncertainty propagation and model calibration.

5.2   Methodology

In the present paper, we propose a rather naive methodology illustrated in figure 5.1 and detailed hereafter. The initialisation phases starts with the evaluation of the proxy on all N_{tot} realizations to obtain the proxy responses $\{x_i(t)\}_{i=1,...,N_{tot}}$. FPCA is then used to define a basis of limited dimension, and a Partitioning Around Medoids algorithm (Maechler *et al.*, 2015) identifies N_{LS} realizations that cover quite uniformly the FPCA proxy space. The exact flow simulations are then performed for the N_{LS} realizations and the resulting N_{LS} pairs of proxy and exact curves form the learning set.

Follows the iteration phase. A Leave-One-Out Cross Validation (LOOCV) procedure is used to find the appropriate number of harmonics: for each sub-learning set of size $N_{LS}^{it} - 1$, the FPCA decomposition and the regression are performed, and the left out curve is predicted. Indeed, a high number of harmonics allows a finer reproduction of the curves, but also increases the dimensionality of the regression problem. When the number of realizations in the learning set is limited, an overfitting of the error model may then occur if n_{harm} is too large. The appropriate number of harmonics is determined by choosing n_{harm} such that the median of the prediction errors is minimized. The full error model is constructed using the N_{LS}^{it} curves

and for n_{harm}, then the N_{tot} exact responses are predicted. Based on these predictions, the next realization, R_{new}, to be used to enlarge the learning set is chosen with a criterion tailored to the quantity of interest. In the case of model calibration, for instance, one is interested in refining the error model for curves close to the observation. The iteration loop starts over at the LOOCV procedure and is repeated until the maximum number of exact flow simulations is reached ($N_{LS}^{it} = N_{LS}^{max}$).
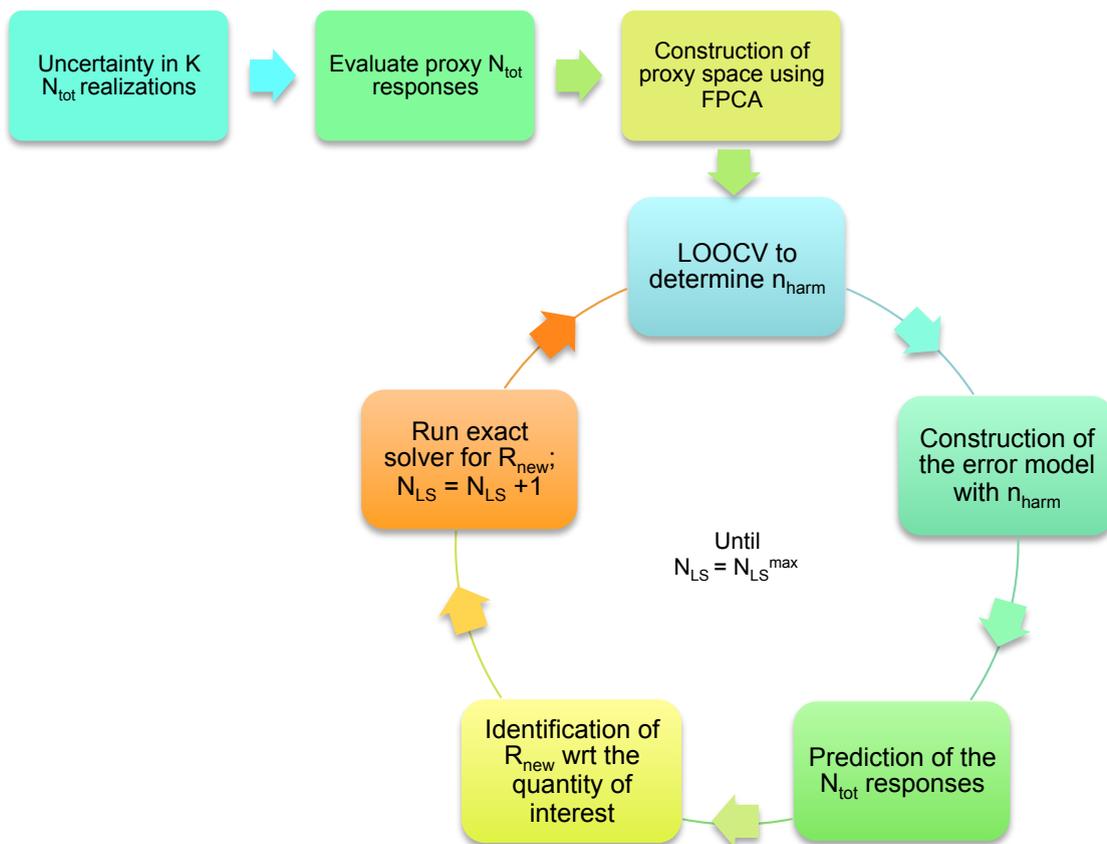
Figure 5.1: Flowchart of the methodology

5.3 Application to a hydrocarbon pollution problem

In the first application we consider the problem of an aquifer contamination by a non-aqueous phase liquid (NAPL). This test case was first explored in section 3.5. A set of 1000 realizations inspired from the Herten test case describes the uncertainty in subsurface permeability and porosity. The contaminant is accidentally released on the left-hand side of the aquifer and

the quantity of interest is the breakthrough curve of the pollutant at a point location (i.e. described by a single cell of the discretization grid) representing, for instance, a drinking well (section 3.5.5). The exact model requires solving the two-phase transport equations as the NAPL is not miscible with water and forms a separate phase, whereas the proxy considers a single-phase transport problem. We refer to section 1.2 and appendix 3.7.1 for more details on flow equations and solvers.

5.3.1   Uncertainty propagation

In the context of uncertainty propagation, a set of N_{tot} geostatistical realizations of the permeability fields $\{R_i\}_{i=1,\dots,N_{tot}}$ is identified to represent the uncertainty in the subsurface properties. The objective is to compute the uncertainty associated with this set for the quantity of interest (here, the saturation in NAPL). To represent the uncertainty on the travel time, it is customary to use as summary statistic the point-wise quantiles curves. The exact solver is computed for each realization $\{y_i(t)\}_{i=1,\dots,N_{tot}}$ and the resulting quantiles are computed $Q_{ref}^{10,50,90}$, to be used as references.
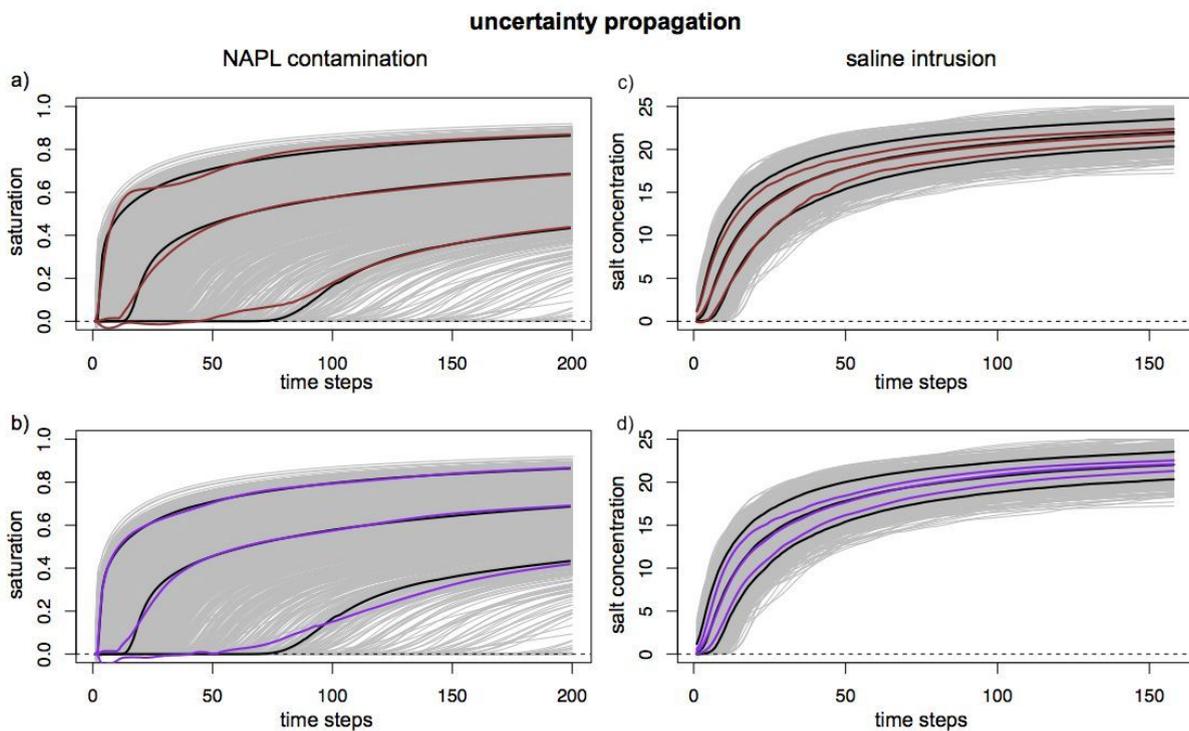
Figure 5.2: Comparison of the quantile curves prediction obtained without iteration (brown curves) and with an iterative scheme (purple curves). The reference quantile curves are indicated in black.

99

We iteratively construct the error model starting with 10 realizations in the learning set and adding a new realization around the predicted quantiles until $N_{LS} = 20$. The results of the iterative scheme are compared with the predicted quantiles obtained with the error model constructed directly with $N_{LS} = 20$, which are identified by a PAM algorithm (Maechler *et al.*, 2015). In figure 5.2.a), the prediction of the quantiles matches quite well the reference for the second half of the measurement time. On the other hand, the first part of the curves exhibits oscillations. The problem is most likely due to the difficulty of finding a basis that represents this set of curves and its large variability at early time. The iterative scheme, in which one curve close to $P10$ is added in alternation with one curve close to $P90$, improves greatly the prediction of P90 quantile curves. The new curve pairs add to the learning set improve the representation with the FPCA basis and allow a finer representation. $P10$ is only slightly improved between the 50 and 100th timesteps and deteriorates at later times. The lower density of curves around $P10$, together with the difficulty of representing late first breakthrough times by a low dimensional space, renders the description of $P10$ more challenging.

In figure 5.3, the error of the predicted quantiles is reported as a function of the iterations for the test case of a learning set of size increasing up to $N_{LS} = 60$. The $P90$ prediction error remains rather constant with some oscillatory behavior as the number of iterations grows. Bootstrap techniques might be used to quantify the stability of the prediction. Indeed, while the prediction of $P90$ with 10 curves in the learning set is just as good as with 50 curves, the confidence bounds around the prediction are much smaller.

In comparison with $P90$, $P10$ error is much larger at the beginning but decreases as the size of the learning set increases. This illustrates that the prediction of the quantiles can be improved by the iterative resampling strategy. The fact that the error seems to converge suggests that a stability (or stagnation) measure, together with a quantification of the uncertainty on the prediction (e.g., using bootstrap techniques), could be used to define a stopping criterion.

5.3.2 Model calibration

Next we consider the challenge of model calibration. One geostatistical realization is randomly chosen as the reference from the ensemble and taken out. The period of observation is defined as the first 60 time steps. The goal is to predict the breakthrough curves at the following time steps. As mean of comparison, model calibration is first done over the ensemble of exact responses by computing the misfit with respect to the reference curve over the period
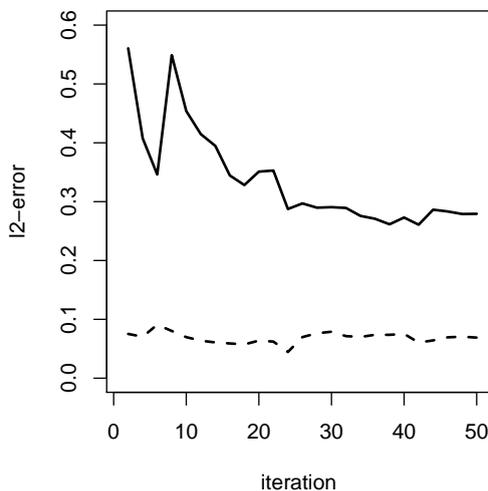
Figure 5.3: Evolution of the l2-error between the predicted and reference quantile with the iterations. The error for $P10$ is indicated by a continuous line, $P90$ by a dotted line.

of observation and by selecting the ones that are under a certain threshold. The "exact" prediction of the curves are shown in figure 5.4.a). The error model performance without any iterative scheme and built directly with 20 realizations in the learning set is shown in figure 5.4.b). The threshold on the l_2-error has to be doubled so that a predicted curve could meet the misfit threshold. While the one identified realization provides a close to perfect prediction, the error model is not capable of reproducing the water breakthrough time. On the other hand, the iterative scheme allows us to fix this issue and improve the prediction of the early time of the curves (figure 5.4.c).

5.4 Application to a saline intrusion in a coastal aquifer problem

We consider a two-dimensional confined aquifer of $10m$ depth and $40m$ length. It is assumed that the aquifer is saturated with freshwater with a density of 1000 kg/m^3 and that saline intrusion occurs on the right-hand side of the aquifer. A hydrostatic gradient is assigned on both vertical sides, assuming fresh water density (1000 kg/m^3) on the left-hand, and saline water density of 1025 kg/m^3 on the right-hand side. The heterogeneity is described by 1000 realizations, with the logarithm of the permeability field that follows a multi-Gaussian distribution and a constant porosity of 20%. The breakthrough curves are computed by vertically averaging the concentration $10m$ away from the sea.

When simulating saline intrusion in an aquifer, the density difference between saline and fresh water requires the explicit description of gravity effects in Darcy's law (see section 1.2). The "exact model" yields a heavy computational cost as the elliptic pressure equation as to be solved at each time step. As proxy model, we choose to perform single-phase transport simulation with the same boundary conditions, which leads to drastic computational savings because the pressure is solved only once. The responses of exact and proxy solvers vary greatly: the absence of coupling between the pressure and the transport equations (which neglects gravity effects) leads to a fully saturated aquifer (with concentration equal to the saline water) and to monotone breakthrough curves. Using the exact model, instead, the saline intrusion is driven to the bottom of the aquifer and the streamlines change dynamically; nothing enforces the monotonicity of the breakthrough curves.

5.4.1 Uncertainty propagation

The reference quantile curves are computed from the 1000 exact responses. In figure 5.2, we compare the prediction of the point-wise quantiles by an error model without (c) and with an iterative scheme (d). No appreciable improvement is provided by the iterative scheme. Prediction of the quantile curves may even deteriorate. Both error models fail to reproduce the $P90$ quantile by 3 to 4 kg/m^3 and $P10$ is overestimated by 1 to 2 kg/m^3.

Additional tests (not reported here) have been run for different learning set size, and showed no improvement of the description of the quantile curves. This is due to the poor informativeness of the proxy for later times (the R^2 coefficient of the linear regression is as low as 0.3). More complex regression models (polynomial and kernel regression models (Hayfield and Racine, 2008)), have been also tested with no success. If the proxy is not informative, the choice of regression model is not relevant and the individual predictions of the curves are of poor quality.

5.4.2 Model calibration

We have arbitrarily chosen to define the period of observation over the 60 first time steps. The model calibration is first performed on the set of exact responses for a given threshold on the l_2-error. The identified curves are shown in figure 5.4.d) and depict quite a wide range of variability, with a mean curve overestimating the salt concentration by 1 or 2 kg/m^3. Using the error model constructed using 20 curves at once, the predictions are in agreement with

the truth but provide a much more limited range of variability (figure 5.4.e). The iterative scheme (figure 5.4.f), in which 10 curves are first used to construct an error model and the next ten are then chosen iteratively by finding the realization that predicts the curve closest to the observations, provides a wider estimate of the uncertainty, but it still underestimate the uncertainty in comparison with the reference.
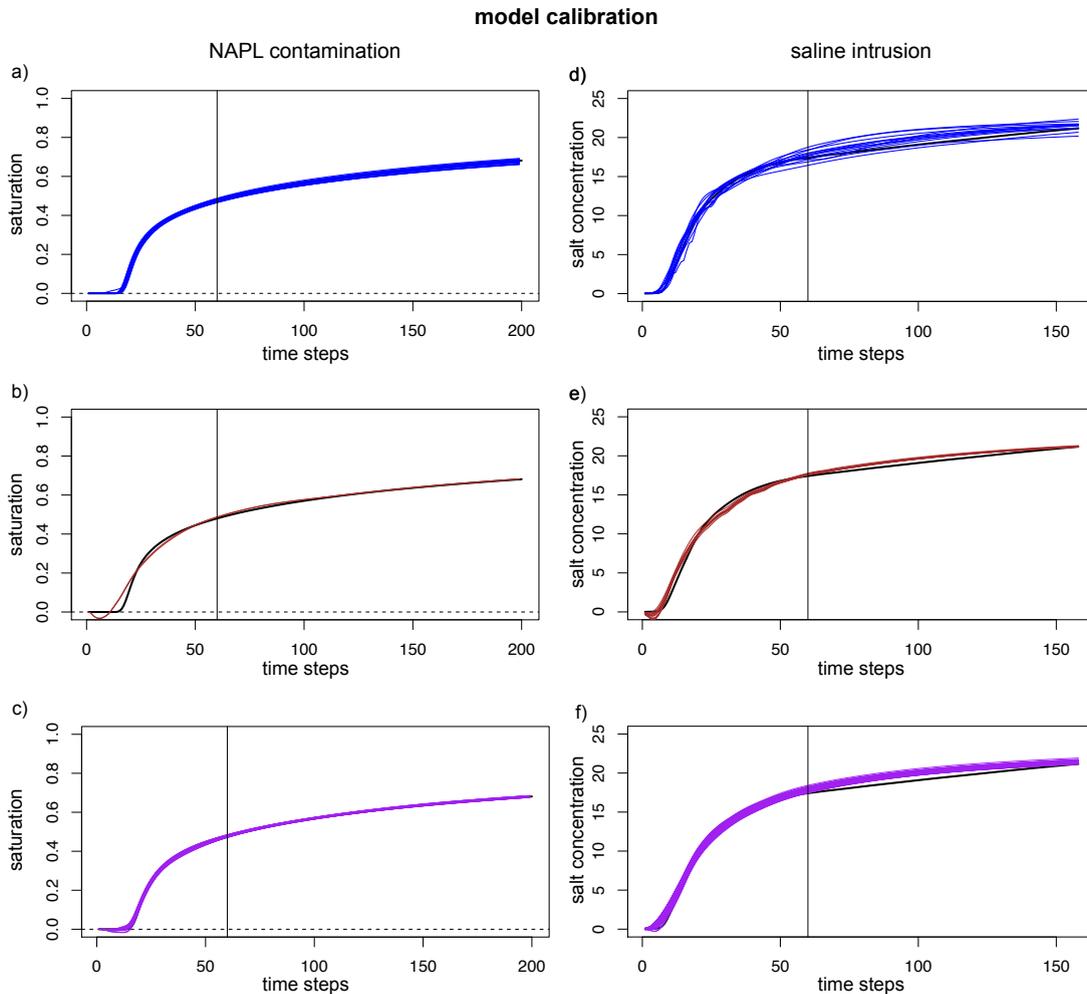
Figure 5.4: Comparison of the model calibration results obtained using the complete set of exact responses (i.e., the reference, in blue), using an error model without iteration (brown curves), and with an iterative scheme (purple curves). The reference calibration curve is indicated in black.

5.5   Conclusions

An iterative approach has been tested to identify the next realization and improve the quality of the error model. Two applications have been investigated. The first example concerns the

case of predicting the point-wise curves of hydrocarbon plume. As shown in section 3.5.5, this test case is particularly challenging due to the great variability of the response curves. The iterative construction of the error model allowed a successful model calibration, but in the context of uncertainty propagation the improvement was modest for the case and the iterative scheme considered. An alternative would be to modify the selection scheme. Instead of focusing equitably on both quantiles, the procedure should preferably sample the region around $P10$ where the density of curve is lower and the variability harder to describe. A solution could be to define a criterion to resample, for instance the Expected Improvement criterion proposed in Ginsbourger *et al.* (2013). However, the adaptation of this criterion in a functional context is not straightforward. Another missing element is the definition of a stopping rule in the iterative construction of the learning set. Figure 5.3 indicates that the error model converges when a certain number iterations are performed. A measure of stability could be used to stop the iterative procedure.

Arguably, the application to the saline intrusion problem, does not allow for a conclusive illustration of the relevance of this approach. The difficulties resides in the use of a rather crude proxy. More refined proxy models could be used instead the single-phase transport solver. One could imagine, for instance, to compute the effect of gravity for a limited number of time steps during the simulation.

Despite the limitations of the strategy explored here, we stress that an iterative scheme is crucial to increase the applicability of the approach. The combination of a proxy model with an error model is useful beyond uncertainty propagation or the rather simplistic calibration problem explored here, such as in a Bayesian inference context. In chapter 4, an error model devised a priori is used within a two-stage MCMC to reduce the number of wasted exact simulations. The construction of the error model requires that some computational resources are devoted to it, in this particular case 100 exact simulations were performed for the 100 identified points in the parameter space. We advocate that an iterative scheme would have been helpful not only to determine the appropriate size of the learning set, but also to identify the relevant realizations to be used to optimize the error model.

Chapter 6

Conclusions and Outlook

In the present work, we have introduced a novel methodology to devise error models and to correct the results of proxy simulations. Following a machine learning approach, a learning set is used to construct a mapping between pairs of proxy and exact responses. The mapping relies on a dimensionality reduction of the problem using FPCA. A linear regression model is then built, so that for any new realization a prediction of the exact response can be obtained based solely on its proxy response. The proposed methodology was applied for several subsurface flow problems (i.e. single-phase transport, multi-phase simulation and density driven flow) and in various contexts (i.e. uncertainty propagation, model calibration, history matching). Other applications could be explored such as reactive transport or reactive two-phase simulation, for instance.

Based on the results of our work, several areas of research appear as promising. To improve the quality of the predictions of the error model, methods should be developed that satisfy the physical constraints of the exact responses. In the examples in chapters 2 and 3, for instance, the exact curves are monotonous and constrained between 0 and 1. In chapter 2, a sigmoid transformation was used to constrain the predicted curves in the $(0, 1)$ interval. The procedure works for the open interval, however, the transformation is not defined when the saturation is 0 or 1. Also, this procedure is not sufficient to enforce monotonicity. The FDA framework allows enforcing monotonicity by using smoothness functions (Ramsay, 2006), but do not guarantee the monotonicity of the predictions, as the curves are reconstructed using the FPCA basis. A simple solution would be to smooth the predicted curves using Bernstein Polynomials as proposed in Menafoglio *et al.* (2014) hence enforcing both monotonicity and

interval constraints. An alternative would be to enforce the constraints at the level of the regression model, which suggests to go beyond linear regression models.

In chapters 3 to 5, a linear regression model between the scores is used. There are several reasons for that, the most evident being to keep the approach as simple as possible. A linear model requires fewer data points, owing to the low number of parameters, and allows the derivation of simultaneous confidence bands around the prediction (see appendix 3.7.3). There is no reason, however, to assume a-priori a linear relationship between approximate and exact scores. Any type of regression model is conceivable, ranging from a polynomial regression model to more sophisticated kernel techniques. Also, other types of regression models might enable us to meet the physical constraints on the predicted curves.

Chapter 3 raises the question of how to use FPCA. The complexity of the responses obtained in several tests has led us to consider increasing the number of harmonics to control the loss of information when projecting on the basis (7 instead of 3 harmonics are considered there), and performing a registration of the curves before applying FPCA. More sophisticated techniques, such as a warping of the curves (Sangalli *et al.*, 2009), could be explored in place of a simple translation in time. However, we should keep in mind that any additional parameter required by the transformation will need to be predicted, hence adding a dimension to the regression model.

Another possibility would be to look for alternatives to FPCA. A natural solution would be to consider a Canonical Correlation Analysis (Ramsay, 2006; Satija and Caers, 2015), to find the shared modes of variation. Despite some advantages, CCA necessitates the definition of additional parameters (e.g. smoothness parameters) that may greatly impact the quality of the prediction.

The aforementioned potential improvements are fairly technical and their listing could be prolonged endlessly in function of the considered test cases. Indeed, the success of the proposed procedure ultimately depends on the quality of the proxy model and on the degree of information that it carries about the exact model. A potential extension of the work would be to devise a multi-level approach, in which several proxies of different qualities could be used in series, each having its own error model. Investigating if and how "intermediate" proxies

might help is an intriguing field of research, and it would be particularly relevant within a Multi-Level Monte Carlo algorithm (or any Bayseian inference procedure).

Our objective has been to develop a framework in broad generality and not problem specific. We believe that this philosophy has indeed been explored throughout the thesis and that the proposed methodology may be helpful in any context where a mapping between two sets of functional responses is required. Independently of the specific choice, all techniques introduce parameters that have to be set. Ideally, those parameters should be "user-defined" depending on the specific needs, for instance in terms of the desired level of accuracy. In chapter 5, a Leave-One-Out Cross-Validation procedure is used to identify the number of harmonics htat have to be used. Quite similarly, bootstrap techniques could be used to quantify the accuracy of the prediction and define a stopping criteria for the iterative procedure. Future work in this area must explore the possibility of an iterative and fully automated procedure for the construction of the error model.

Bibliography

Aarnes, J. and Efendiev, Y. (2008). Mixed multiscale finite element methods for stochastic porous media flows. *SIAM Journal on Scientific Computing*, **30**(5), 2319–2339.

Aarnes, J., Kippe, V., and Lie, K. (2005). Mixed multiscale finite elements and streamline methods for reservoir simulation of large geomodel. *Adv. Water Res.*, **28**, 257–271.

Arbogast, T. (2002). Implementation of a locally conservative numerical subgrid upscaling scheme for two phase darcy flow. *Comput. Geosci.*, **6**, 453–481.

Ballin, P., Journel, A., and Aziz, K. (1992). Prediction of uncertainty in reservoir performance forecast. *Journal of Canadian Petroleum Technology*, **31**(4).

Ballio, F. and Guadagnini, A. (2004). Convergence assessment of numerical Monte Carlo simulations in groundwater hydrology. *Water resources research*, **40**(4).

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

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of hydrology*, **320**(1), 18–36.

Borg, I. and Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. Springer.

Carnell, R. (2009). lhs: Latin hypercube samples. *R package version 0.5*.

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., and Slooten, L. J. (2005). Inverse problem in hydrogeology. *Hydrogeology Journal*, **13**(1), 206–222.

Carter, J. N., Ballester, P. J., Tavassoli, Z., and King, P. R. (2006). Our calibrated model has

poor predictive value: An example from the petroleum industry. *Reliability Engineering & System Safety*, **91**(10), 1373–1381.

Chen, Y. and Durlofsky, L. (2008). Ensemble-level upscaling for efficient estimation of fine-scale production statistics. *SPE Journal*, **13**(4), 400–411.

Chen, Y., Park, K., and Durlofsky, L. J. (2011). Statistical assignment of upscaled flow functions for an ensemble of geological models. *Computational Geosciences*, **15**(1), 35–51.

Christen, J. A. and Fox, C. (2005). MCMC using an approximation. *Journal of Computational and Graphical statistics*, **14**(4), 795–810.

Christie, M. (1996). Upscaling for reservoir simulation. *Journal of Petroleum Technology*, **48**(11), 1004–1010.

Christie, M., Demyanov, V., and Erbas, D. (2006). Uncertainty quantification for porous media flows. *Journal of Computational Physics*, **217**(1), 143–158.

Christie, M. A., Glimm, J., Grove, J. W., Higdon, D. M., Sharp, D. H., and Wood-Schultz, M. M. (2005). Error analysis and simulations of complex phenomena. *Los Alamos Science*, **29**(6).

Cox, M. and Cox, T. (2008). Multidimensional scaling. *Handbook of data visualization*, pages 315–347.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, **147**, 1–23.

Cuevas, A., Febrero, M., and Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics*, **30**(2), 285–300.

Cui, T., Fox, C., and O'Sullivan, M. (2011). Adaptive error modelling in mcmc sampling for large scale inverse problems. *Report, Univeristy of Auckland, Faculty of Engineering*.

Dagan, G. (2002). An overview of stochastic modeling of groundwater flow and transport: From theory to applications. *Eos, Transactions American Geophysical Union*, **83**(53), 621.

Demyanov, V., Pozdnoukhov, A., Christie, M., and Kanevski, M. (2010). Detection of optimal models in parameter space with support vector machines. In *geoENV VII–Geostatistics for Environmental Applications*, pages 345–358. Springer.

Deviese, E. (2010). *Modeling Fluvial Reservoir Architecture using Flumy Process*. Ph.D. thesis, TU Delft, Delft University of Technology.

Dostert, P., Efendiev, Y., and Hou, T. (2008). Multiscale finite element methods for stochastic porous media flow equations and application to uncertainty quantification. *Computer Methods in Applied Mechanics and Engineering*, **197**(43), 3445–3455.

Durlofsky, L. (2005). Upscaling and gridding of fine scale geological models for flow simulation. In *8th International Forum on Reservoir Simulation Iles Borromees, Stresa, Italy*, pages 20–24.

Efendiev, Y., Datta-Gupta, A., Ginting, V., Ma, X., and Mallick, B. (2005). An efficient two-stage Markov chain Monte Carlo method for dynamic data integration. *Water Resources Research*, **41**(12).

Efendiev, Y., Hou, T., and Luo, W. (2006). Preconditioning Markov chain Monte Carlo simulations using coarse-scale models. *SIAM Journal on Scientific Computing*, **28**(2), 776–803.

Efendiev, Y., Datta-Gupta, A., Ma, X., and Mallick, B. (2009). Efficient sampling techniques for uncertainty quantification in history matching using nonlinear error models and ensemble level upscaling techniques. *Water Resources Research*, **45**(11).

Elsheikh, A. H., Hoteit, I., and Wheeler, M. F. (2014). Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates. *Computer Methods in Applied Mechanics and Engineering*, **269**, 515–537.

FOEN (2009). *Sustainable protection of groundwater*. the Federal Office for the Environment.

Foster, S. S., Lawrence, A., and Morris, B. (1998). *Groundwater in urban development: assessing management needs and formulating policy strategies*, volume 390. World Bank Publications.

Fox, J. and Weisberg, H. S. (2011). *An R companion to applied regression*, chapter Multivariate Linear Models in R. SAGE Publications, 2nd edition edition.

Freni, G. and Mannina, G. (2010). Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution. *Journal of Hydrology*, **392**(1), 31–39.

Gaganis, P. and Smith, L. (2001). A Bayesian approach to the quantification of the effect of model error on the predictions of groundwater models. *Water Resources Research*, **37**(9), 2309–2322.

Ginsbourger, D., Rosspopoff, B., Pirot, G., Durrande, N., and Renard, P. (2013). Distance-based kriging relying on proxy simulations for inverse conditioning. *Advances in Water Resources*, **52**, 275–291.

Hajibeygi, H. and Jenny, P. (2009). Multiscale finite-volume method for parabolic problems arising from compressible flow in porous media. *J. Comp. Phys*, **228**, 5129–5147.

Hajibeygi, H., Bonfigli, G., Hesse, M., and Jenny, P. (2008). Iterative multiscale finite-volume method. *J. Comp. Phys*, **227**(19), 8604–8621.

Hajibeygi, H., Lee, S. H., Lunati, I., *et al.* (2012). Accurate and efficient simulation of multiphase flow in a heterogeneous reservoir with error estimate and control in the multiscale finite-volume framework. *SPE Journal*, **17**(04), 1–071.

Hassan, A. E., Bekhit, H. M., and Chapman, J. B. (2009). Using Markov chain Monte Carlo to quantify parameter uncertainty and its effect on predictions of a groundwater flow model. *Environmental Modelling & Software*, **24**(6), 749–763.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, **27**(5).

Helmig, R. *et al.* (1997). *Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems*. Springer-Verlag.

Henderson, B. (2006). Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics*, **17**(1), 65–80.

Hou, T. Y. and Wu, X. H. (1997). A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comp. Phys*, **134**(1), 169–189.

Huber, E., Caers, J., and Huggenberger, P. (2015). Geophysical stereology: Quantifying the uncertainty on 3D subsurface structure from 2D GPR data. application to the tagliamento river. *in preparation*.

Jenny, P. and Lunati, I. (2009). Modeling complex wells with the multi-scale finite-volume method. *J. Comp. Phys*, **228**(3), 687–702.

Jenny, P., Lee, S. H., and Tchelepi, H. (2003). Multi-scale finite-volume method for elliptic problems in subsurface flow simulation. *J. Comp. Phys*, **187**(1), 47–67.

Jenny, P., Lee, S. H., and Tchelepi, H. (2006). Adaptive fully implicit multi-scale finite-volume method for multi-phase flow and transport in heterogeneous porous media. *J. Comp. Phys*, **217**, 627–641.

Josset, L. and Lunati, I. (2013). Local and global error models to improve uncertainty quantification. *Mathematical Geosciences*, pages 45:601–620.

Josset, L., Ginsbourger, D., and Lunati, I. (2015). Functional error modeling for uncertainty quantification in hydrogeology. *Water Resources Research*, **51**(2), 1050–1068.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**(3), 187–200.

Kavetski, D., Kuczera, G., and Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 2. application. *Water Resources Research*, **42**(3).

Keating, E. H., Doherty, J., Vrugt, J. A., and Kang, Q. (2010). Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. *Water Resources Research*, **46**(10).

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(3), 425–464.

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

Köpke, C., Irving, J., and Roubinet, D. (2014). Assessing and accounting for the effects of model error in Bayesian solutions to hydrogeophysical inverse problems. *American Geophysics Union Fall Conference*.

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

Künze, R. and Lunati, I. (2012b). *MaFloT - Matlab Flow and Transport*. Published under the GNU license agreement.

Laloy, E., Rogiers, B., Vrugt, J. A., Mallants, D., and Jacques, D. (2013). Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. *Water Resources Research*, **49**(5), 2664–2682.

Lantuéjoul, C. (2002). *Geostatistical simulation: models and algorithms*. Number 1139. Springer Science & Business Media.

Lee, S., Zhou, H., and Tchelepi, H. (2009). Adaptive multiscale finite-volume method for non-linear multiphase transport in heterogeneous formations. *Journal of Computational Physics*, **228**(24), 9036–9058.

Lunati, I. and Jenny, P. (2004). Multi-scale finite-volume method for flow in highly hetero-geneous porous media with shale layers. In *European Conference of Mathematics of Oil Recovery IX, Cannes, France*.

Lunati, I. and Jenny, P. (2006). Multi-scale finite-volume method for compressible flow in porous media. *J. Comp. Phys*, **216**(2), 616–636.

Lunati, I. and Jenny, P. (2007). Treating highly anisotropic subsurface flow with the multiscale finite-volume method. *Multiscale Model. Simul.*, **6**(1), 308–318.

Lunati, I. and Jenny, P. (2008). Multiscale finite-volume method for density-driven flow in porous media. *Comput. Geosci.*, **12**(3), 337–350.

Lunati, I. and Kinzelbach, W. (2004). Water-soluble gases as partitioning tracers to investigate the pore volume-transmissivity correlation in a fracture. *J. Cont. Hydr.*, **75**(1-2), 31–54.

Lunati, I. and Lee, S. (2009). An operator formulation of the multiscale finite-volume method with correction function. *Multiscale Model. Simul.*, **8**(1), 96–109.

Lunati, I., Tyagi, M., and Lee, S. (2011). An iterative multiscale finite-volume algorithm converging to the exact solution. *J. Comp. Phys*, **230**(5), 1849–1864.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2015). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.1).

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

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

Marle, C. (1981). *Multiphase flow in porous media*. Éditions technip.

McLennan, J. and Deutsch, C. (2005). Ranking geostatistical realizations by measures of connectivity. In *SPE International Thermal Operations and Heavy Oil Symposium, 98168, Alberta, Canada*.

Menafoglio, A., Guadagnini, A., and Secchi, P. (2014). A kriging approach based on aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment*, **28**(7), 1835–1851.

Mohamed, L., Christie, M. A., Demyanov, V., *et al.* (2011). History matching and uncertainty quantification: multiobjective particle swarm optimisation approach. In *SPE EUROPEC/EAGE Annual Conference and Exhibition*. Society of Petroleum Engineers.

Mohamed, L., Calderhead, B., Filippone, M., Christie, M., and Girolami, M. (2012). Population MCMC methods for history matching and uncertainty quantification. *Computational Geosciences*, **16**(2), 423–436.

O'Sullivan, A. and Christie, M. (2006a). Error models for reducing history match bias. *Computational Geosciences*, **10**(4), 405–405.

O'Sullivan, A. and Christie, M. (2006b). Simulation error models for improved reservoir prediction. *Reliability Engineering & System Safety*, **91**(10), 1382–1389.

O'Sullivan, A., Christie, M., *et al.* (2005). Solution error models: a new approach for coarse grid history matching. *Paper SPE*.

Phien-Wej, N., Giao, P., and Nutalaya, P. (2006). Land subsidence in Bangkok, Thailand. *Engineering Geology*, **82**(4), 187–201.

Pirot, G., Straubhaar, J., and Renard, P. (2014). Simulation of braided river elevation model time series with multiple-point statistics. *Geomorphology*, **214**, 148–156.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramsay, J. J. O., Hooker, G., and Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer.

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2012). *fda: Functional Data Analysis*. R package version 2.3.2.

Razavi, S., Tolson, B. A., and Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, **48**(7).

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

Renard, P. (2007). Stochastic hydrogeology: what professionals really need? *Groundwater*, **45**(5), 531–541.

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

Renard, P. and de Marsiliy, G. (1997). Calculating equivalent permeability: a review. *Adv. Water Res.*, **20**(5-6), 253–278.

Richman, M. B. (1986). Rotation of principal components. *Journal of climatology*, **6**(3), 293–335.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*, volume 319. Citeseer.

Roberts, G. O., Gelman, A., Gilks, W. R., *et al.* (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.

Rubin, Y. (2004). Stochastic hydrogeology–challenges and misconceptions. *Stochastic Environmental Research and Risk Assessment*, **18**(4), 280–281.

Sangalli, L. M., Secchi, P., Vantini, S., and Veneziani, A. (2009). A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, **104**(485).

Satija, A. and Caers, J. (2015). Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space. *Advances in Water Resources*.

Scheidt, C. and Caers, J. (2009a). Representing spatial uncertainty using distances and kernels. *Mathematical Geosciences*, **41**(4), 397–419.

Scheidt, C. and Caers, J. (2009b). Uncertainty quantification in reservoir performance using distances and kernel methods–application to a west africa deepwater turbidite reservoir. *SPE Journal*, **14**(4), 680–692.

Scheidt, C. and Caers, J. (2010). Bootstrap confidence intervals for reservoir model selection techniques. *Computational Geosciences*, **14**, 369–382.

Scheidt, C., Caers, J., Chen, Y., and Durlofsky, L. (2010). Rapid construction of ensembles of high-resolution reservoir models constrained to production data. In *12th European Conference on the Mathematics of Oil Recovery*.

Scheidt, C., Caers, J., Chen, Y., and Durlofsky, L. J. (2011). A multi-resolution workflow to generate high-resolution models constrained to dynamic data. *Computational Geosciences*, **15**(3), 545–563.

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

Strebelle, S. B., Journel, A. G., *et al.* (2001). Reservoir modeling using multiple-point statistics. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Tavassoli, Z., Carter, J. N., King, P. R., *et al.* (2004). Errors in history matching. *SPE Journal*, **9**(03), 352–361.

Tavassoli, Z., Carter, J. N., and King, P. R. (2005). An analysis of history matching errors. *Computational Geosciences*, **9**(2-3), 99–123.

Ullah, S., Finch, C. F., *et al.* (2013). Applications of functional data analysis: A systematic review. *BMC medical research methodology*, **13**(1), 43.

UNESCO (2012). *Water and Energy*, volume The United Nations World Water Development Report 4. UNESCO and World Water Assessment Programme.

UNESCO (2014). *Managing Water under Uncertainty and Risk*, volume The United Nations World Water Development Report 2014. UNESCO and World Water Assessment Programme.

Weisberg, S. (2014). *Applied linear regression*. John Wiley & Sons.

Wen, X. and Gómez-Hernández, J. (1996). Upscaling hydraulic conductivities in heterogeneous media: An overview. *J. Hydr.*, **183**(1-2), R9–R32.

Wu, J. and Zeng, X. (2013). Review of the uncertainty analysis of groundwater numerical simulation. *Chinese Science Bulletin*, **58**(25), 3044–3052.

Xiu, D. and Karniadakis, G. E. (2002). The wiener–askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, **24**(2), 619–644.

Yustres, Á., Asensio, L., Alonso, J., and Navarro, V. (2012). A review of Markov chain Monte Carlo and information theory tools for inverse problems in subsurface flow. *Computational Geosciences*, **16**(1), 1–20.

Zeng, L., Shi, L., Zhang, D., and Wu, L. (2012). A sparse grid based Bayesian method for contaminant source identification. *Advances in Water Resources*, **37**, 1–9.

Zhou, H. and Tchelepi, H. (2012). Two-stage algebraic multiscale linear solver for highly heterogeneous reservoir models. *SPE Journal*, **17**(2), 523–539.

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

Acknowledgements

With the conclusion of this work comes the time to look back on the formidable adventure that have been the last four years. I have truly enjoyed it, and for that I have many people to thank for.

Foremost, I wish to thank my PhD supervisor Ivan Lunati, with whom I have learned continuously, discovered new subjects, and developed much needed writing skills. I have particularly enjoyed our constant debates on mathematical concepts and his patient teachings on flow modeling. Most importantly, Ivan's constant attention to details taught me to aim for academic excellency while cultivating a strong notion of ethic in science. I could not have hoped for a better teacher and I am happy to say that our collaboration will not stop here.

Many thanks are also due to Philippe Renard. Eleven years ago, Philippe took the time to sit down and explain the fundamentals of hydrogeology to the young student that I was then, just freshly out of high-school. Four years ago, he introduced me to Ivan Lunati to study within the ENSEMBLE project. As initiator and principal investigator, he ensured a constructive and always stimulating atmosphere during our many project meetings. It is with pride that I refer to him as mentor.

I also wish to thank the remaining members of the Jury: Eric Verecchia who brought a joyful atmosphere to the private defense, Klaus Holliger for his external eye on the subject and Mikhail Kanevski for bringing a new perspective on the matter and introducing me to the machine learning community.

Another crucial contributor to this work is David Ginsbourger, who taught me "morally correct" statistics, as he puts it so elegantly. David also initiated and led the Monte CAarlo