Title
Inverse Modeling of Geological Heterogeneity for Goal-Oriented Aquifer Characterization

Permalink
https://escholarship.org/uc/item/07n5t8pw

Author
Savoy, Heather Marie

Publication Date
2017

Peer reviewed|Thesis/dissertation
Inverse Modeling of Geological Heterogeneity for Goal-Oriented Aquifer Characterization

by

Heather Marie Savoy

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Engineering - Civil and Environmental Engineering and the Designated Emphasis in Computational and Data Science and Engineering in the Graduate Division of the University of California, Berkeley

Committee in charge:

Professor Yoram Rubin, Chair
Professor Fotini Chow
Professor David Brillinger

Spring 2017
Abstract

Inverse Modeling of Geological Heterogeneity for Goal-Oriented Aquifer Characterization

by

Heather Marie Savoy

Doctor of Philosophy in Engineering - Civil and Environmental Engineering and the Designated Emphasis in Computational and Data Science and Engineering

University of California, Berkeley

Professor Yoram Rubin, Chair

Characterizing the spatial heterogeneity of aquifer properties, particularly hydraulic conductivity, is paramount in groundwater modeling when the transport and fate of contaminants need to be predicted. The field of geostatistics has focused on describing this heterogeneity with spatial random functions. The field of stochastic hydrogeology uses these functions to incorporate uncertainty about the subsurface in groundwater modeling predictions. Bayesian inference can update prior knowledge about the spatial patterns of the subsurface (e.g. plausible ranges of values) with a variety of information (e.g. direct measurements of hydraulic conductivity as well as indirect measurements such as water table drawdown at an observation well) in order to yield posterior knowledge. This dissertation focuses on expanding the tools for Bayesian inference of these spatial random functions.

First, the development of open-source software tools for guiding users through the Bayesian inference process are described. There is an desktop application that implements the Method of Anchored Distributions and is referred to as MAD#. It is built in a modular fashion such that it can be coupled with any geostatistical software and any numerical modeling software. This modularity allows for a wide variety of spatial random functions and subsurface processes to be incorporated in the Bayesian inference process. There is also a R package, called anchoredDistr, that supplements the MAD# software. While the MAD# software handles the communication between the geostatistical software and the numerical modeling software, the anchoredDistr package provides more flexibility in analyzing the results from MAD#. Since R is an open-source statistical computing language, the anchoredDistr package allows users to take advantage of the plethora of statistical tools in R to calculate the posterior knowledge in the Bayesian process. Although MAD# provides a post-processing module to calculate this posterior knowledge, it does not provide all of the options that the R community can provide for modifications.

Second, the expansion of which kinds of data and knowledge can be incorporated into the Bayesian process is explored. Incorporating time series (e.g. the drawdown of a water
table from pumping over time) as indirect data in Bayesian inference poses a computational problem referred to as the ‘curse of dimensionality’. Since each additional measurement in time is correlated with the measurements before and after it, the calculation of probability distributions of these data become multi-dimensional. A synthetic case study incorporating drawdown time series in the Bayesian inference process is explored. A second form of information, conceptual models of geology, is also explored with a synthetic case study. Conceptual models of geology (e.g. a graphical representation of assumed geologic layering) can be described with images. There is a geostatistical technique called Multipoint Statistics that uses images as its input. The synthetic case study provides a proof-of-concept example in which the Bayesian inference process can infer conceptual models of geology using Multipoint statistics.

Third, the issue of devising spatial models with realistic geology while constraining the complexity of the model is explored. An aquifer analog is used as the basis for an example. An aquifer analog is a data set with data of hydraulic properties at high spatial resolution, i.e. much higher than expected for ordinary field measurements. The aquifer analog used in this dissertation has ten soil types distributed in three-dimensional space. The objective posed is to predict the early arrival time of a contaminant traveling through the analog. Given this prediction goal, the task is to simplify the analog into a simplified structure without changing the prediction outcome. The purpose of this exercise is to take a goal-oriented approach to defining a parsimonious spatial model for describing this complex aquifer analog such that a geostatistical model can be inferred for this kind of geology in a computationally efficient manner.

Ultimately, any uncertainty quantification regarding the spatial heterogeneity of subsurface properties has the goal of improving groundwater modeling prediction efforts. With the addition of freely available software tools, the ability to integrate more forms of information, and methodology for translating complex geological structures into parsimonious spatial models, the characterization of our groundwater resources improves.
To my family

who will not read this dissertation,
but would be proud to have a printed copy on a bookshelf somewhere anyways.
Contents

1 Introduction 1
   1.1 State of Stochastic Hydrogeology 2
   1.2 Research Objectives 6

2 Background 9
   2.1 Geostatistics 9
   2.2 Bayesian Inverse Modeling 14

3 Open-source Software Tools for Inversion 22
   3.1 A Software Framework for Inverse Modeling and Uncertainty Characterization 23
   3.2 A Package for the Bayesian Inversion of Geostatistical Parameters with Multi-type and Multi-scale Data 43

4 Integration of Multiple Forms of Information 56
   4.1 Inversion Data: Handling the Curse of Dimensionality 57
   4.2 Prior Knowledge: Conceptual Models of Geological Heterogeneity 73

5 Geological Realism 80
   5.1 Introduction 80
   5.2 Methods 83
   5.3 Results and Discussion 87
   5.4 Conclusion 92

6 Conclusion 96

References 99
I would like to thank the many people in my life that made the past few years unforgettable. First, there is Brad Harken and Karina Cucchi with whom I have shared an office with the tiniest of windows for the past four years - without you, there would have been a lot more despair and a lot less warmth. There is also Ching-Fu Chang and Nigel Chen, who have no window in their office, but they still have been a joy to get to know. Michelle Newcomer and Matthew Over have graduated since I started, but they were both role models in professionalism and kindness. Every week, I look forward to my chat with Falk Heße, who has been both a mentor and a friend across the years that we have moved between California and Germany. There was Jon Sege who always lit up group meetings, Carlos Osorio-Murillo who was always patient with all my questions, and then Changhong Wang and Thanh Thuy Nguyen who are probably the most welcoming people I have ever met, even though they were visiting us. And of course, it was Yoram Rubin who brought us all together and provided to me all of the opportunities that I have had these past few years. Thank you all.

Outside the lab, I would like to thank Tina Chow and David Brillinger for serving on my dissertation committee, and Sally Thompson for sitting on my oral exam committees. I have appreciated your patience and guidance. Thanks to Linda von Hoene for all that she taught me on how to be an educator and Shelley Okimoto for always having time to listen to my questions and problems. To Thomas Kalbacher and Peter Dietrich at the Helmholtz Centre for Environmental Research in Germany, thanks for hosting me for three summers as I found my way through the early stages of my research. I owe gratitude to the Jane Lewis Fellowship Committee and the Trustees of the Roy G. Post Foundation for providing me financial support during those years as well.

To my many sisters at the Convent, where I lived for the majority of my time in Berkeley, thank you for providing me a family away from home. Especially so for Saalem Adera, who has been my housemate, classmate, climbing partner, cooking partner, and dearest friend. For my family in Florida, I would like to thank them for their patience with my need to stay in school so long and for never complaining about having to pick me up at the Tampa airport at 5am on Christmas morning because I could only afford the most inconvenient flight. And my rock throughout it all has been Henry Becker, who kept me sane during my panic before oral exams, made me go outside and enjoy the sunshine during my darkest times, and has been cheering me on all along.
Chapter 1

Introduction

The subsurface is an important water resource. The total groundwater withdrawn in the United States is 80 billion gallons per day [Maupin et al., 2014], which is an average of approximately 250 gallons of groundwater per person everyday. Approximately 130 million people, nearly 45% the population, rely on groundwater as a drinking water source [DeSimone et al., 2014]. In a national assessment of groundwater sources for drinking water, 22% of examined wells tested positive for contamination by at least one chemical regulated by the Environmental Protection Agency [DeSimone et al., 2014]. To protect such a vital resource for drinking water, we need to be able to predict how contaminants are transported in the subsurface.

Numerical groundwater contamination modeling strives to predict how contamination travels through the subsurface such that questions regarding where contaminants are heading, when they will arrive at a certain location such as a drinking water well, and which concentration levels will occur can be answered. The movement of these solutes in the groundwater through the subsurface depends on the physical properties of the subsurface. These properties vary in three-dimensional space such that predictions regarding groundwater flow and transport depend on characterizing the spatial distribution of these properties. For example, hydraulic conductivity (the measure of how easily water flows through a porous medium) can vary over several orders of magnitude on a decimeter scale. By disregarding the spatial distribution of this hydraulic conductivity, solute transport simulations could neglect preferential flow paths created by the higher hydraulic conductivity regions and thus underestimate the transport of the solute.

The main obstacle of incorporating the spatial distribution of subsurface properties in numerical models is that measurements are sparse due to the measurement techniques being either invasive/destructive, at poor spatial resolution compared to the subsurface features, prohibitively expensive, or a combination thereof. The sparse nature of measurements compared to the spatial distribution of subsurface properties results in an ill-posed problem, i.e. there are significantly more unknown values than known values in the subsurface. To alleviate this ill-posedness, the field of geostatistics aims to represent the spatial distribution of properties with statistical models. These models, known as spatial random functions
(SRFs), represent spatial distributions with a set of parameters to describe global patterns as opposed to predict values everywhere. See Section 2.1 for a detailed account of geostatistical models and the algorithms associated with them.

The next section reflects on two series of opinion papers regarding the state of stochastic hydrogeology and the barriers to its use outside academia. Recognizing room for improvements in the field sets the foundation for my research objectives and their contribution to the field of stochastic hydrogeology.

1.1 State of Stochastic Hydrogeology

Introduction

Stochastic hydrogeology has a history over the past few decades for handling uncertainty regarding the spatial configuration of subsurface properties, however, there is a disconnect between the research community and practicing hydrogeologists when it comes to the applicability of stochastic hydrogeology in practice. Previous series of articles have entailed the perspectives of hydrogeology researchers on why this disconnect exists. This section summarizes the points raised on both why there is this disconnect and the suggestions provided on how to remove it.

First, the meaning of ‘stochastic’ in this context should be stressed. Any method dealing with uncertainty in a theoretically sound manner in conditions ranging from no data to abundant data and every condition in between can be considered stochastic. Here, ‘theoretically sound’ implies that there is a justifiable random component that reflects our state of uncertainty about the medium and/or processes studied.

The conversation of how practitioners and scientists consider uncertainty can be found in other contexts. For example, Höllermann and Evers [2017] elicited perspectives from practitioners and scientists in the water management sector on how uncertainty enters their decision-making process. The authors point out how practitioners differ from scientists regarding what sources of uncertainty are of greatest concerns, e.g. the uncertainty in the public response to the length of the project versus the uncertainty in numerical model parameters. Of interest is the translation from uncertainty to risk, as practitioners are prone to do implicitly and as decision-makers can use directly.

To begin the discussion, the comments in two series of articles addressing stochastic hydrogeology in practice are summarized. These series were in the Stochastic Environmental Research and Risk Assessment (SERRA) journal in 2004 and in Water Resources Research (WRR) journal in 2016. The 2004 SERRA series asked nine individual stochastic hydrogeology researchers two specific questions: 1) ‘Why have there not been many real-world applications of stochastic theories and approaches, despite the significant progress in developing such rigorous theories and approaches for studying fluid flow and solute transport in heterogeneous media?’, and 2) ‘In your opinion, what must be done in order to render stochastic theories and approaches as routine tools in hydrogeologic investigation and mod-
The 2016 WRR series is more open-ended, asking teams of stochastic hydrogeology researchers to debate on ‘stochastic subsurface hydrology from theory to practice’. These two series were chosen because they asked similar questions, they combine multiple perspectives from a variety of researchers, and the 12-year timespan between them allows us to consider the evolution of stochastic hydrogeology. However, due to the more open nature of the 2016 WRR series, not all of the discussion points can be compared between the two series.

Summary of perspectives from 2004 and 2016

The 2004 SERRA series

The 2004 SERRA series contains perspectives from nine stochastic hydrogeology researchers. Based on their responses to the question regarding why stochastic hydrogeology is not used in practice, a list of five over-arching topics appearing across the nine papers have been compiled (see Table 1.1). Those five topics are:

The influence of regulations and/or the court system This topic appears with comments regarding how hydrogeology practitioners have the incentive to use status quo, as opposed to state-of-the-art, procedures for reducing liability in adversarial legal or regulatory scrutiny [Freeze, 2004]. Additionally, it is stated that there is the incentive to reduce the possibility of uncertainty analysis being perceived as their lack of training as opposed to a lack of knowledge [Freeze, 2004]. In the realm of environmental risk assessment, the uncertainty in hydrogeological modeling may not be perceived as significant as the uncertainty in the financial or political atmosphere regarding the project [Freeze, 2004]. Also provided is the suggestions that decision-makers are hesitant to interpret stochastic analyses based on the tail of distributions or that regulations mandate point estimates for values as opposed to intervals or probability distributions [Sudicky, 2004]. These points were largely stated without concrete examples and Christakos [2004] takes the opposite opinion that the stochastic hydrogeology community can not place blame on the court system when other fields use models with uncertainty, e.g. biomedical models for DNA analysis, under legal scrutiny.

The role of higher education This topic appeared either in terms of the lack of proper training in stochastic methods for hydrogeology practitioners [Neuman, 2004] or in terms of that the theory is at a level not digestible by students [Winter, 2004].

The lack of appropriate measurement technology and/or data This topic mainly appears as that technology does not provide measurements attributable to the parameters and scales of stochastic models [Molz, 2004]. Additionally stated was that, in practice, too much data is needed to fit those parameters. However, Rubin [2004] points out the misconception regarding data needs since stochastic methods exist to handle the lack of data.
The lack of user-friendly software that applies theory This topic mainly appears as the call for user-friendly software that integrates multiple forms of information \cite{Christakos2004} and that is computationally efficient enough for practitioners to use \cite{Neuman2004}.

The lack of applicability of theory to real-world problems Arguments here were either saying that theoretical research is not applicable in real-world problems (e.g. oversimplification of spatial structure, or minimizing uncertainty in parameters when other uncertainties need to be minimized) \cite{Molz2004}, or at least there is a lack of applications to showcase the applicability of theories in real-world problems \cite{Ginn2004, Neuman2004}.

The second question in this series requests suggestions on what needs to be done to have the field of stochastic hydrogeology adopted by practitioners. The suggestions provided by the nine authors transcended the five topics listed above, so they are presented separately below:

Cross-disciplinary collaborations There should be cross-disciplinary collaborations with either hydrogeology practitioners for translating theory to practice \cite{Dagan2004, Ginn2004}, numerical modelers for embedding stochastic tools in software \cite{Dagan2004}, and decision-makers or health-related researchers for making stochastic analyses relevant \cite{Rubin2004, Christakos2004}. These collaborations would produce literature that is application-oriented and readable outside the stochastic hydrogeology community \cite{Sudicky2004}.

Software There is a need to develop flexible, powerful, and efficient tools for data integration frameworks \cite{Christakos2004}, to embed stochastic methods into pre-existing and popular software \cite{Dagan2004, Ginn2004}, or to make stand-alone stochastic software capable of coupling with flow and transport models \cite{Neuman2004}.

Geological Realism Catalogs of properties and geological structures \cite{Dagan2004} need to be built. Alternatively, there should be advancement in theory for incorporating geological realism into stochastic models \cite{Molz2004}.

Education There should be improvement in education of stochastic concepts \cite{Neuman2004, Winter2004}, its data needs \cite{Sudicky2004, Winter2004}, and its relationship with respect to deterministic methods \cite{Neuman2004}.

For comparison to a more recent series of papers addressing similar questions, the five topics described above were also searched for in the 2016 WRR series to gauge progress since the SERRA 2004 series. Since no new significantly different topics were discerned, the summary of the 2016 WRR series is provided as a function of how the state of the five 2004 topics have updated. How those topics evolved per 2016 series’ perspective can be seen in Table 1.2 The topics most prevalently mentioned in the 2016 WRR series are
Topic 4 (user-friendliness of concepts) and Topic 5 (theory too limited or not showcased in applications). Complaints regarding the availability of useful software tools remain [Cirpka and Valocchi, 2016], however there are concessions on at least some stochastic modules existing in forward models [Fiori et al., 2016]. The extensive discussion on geological realism [Fogg and Zhang, 2016] and transport processes [Cirpka and Valocchi, 2016] in the 2016 WRR series is encapsulated in the issue of simplicity - both concepts relate to the practical concerns of applicability in practice. Of interest, however, is the mention of goal-oriented frameworks by [Fiori et al., 2016], which is akin to the translation of uncertainty into risk at the knowledge-decision interface that Höllermann and Evers [2017] recommends.

Table 1.1: 2004 Authors and Issue Topics

<table>
<thead>
<tr>
<th>Author</th>
<th>Measurement technology and availability of data</th>
<th>User-friendliness of concepts</th>
<th>Theory too limited or not showcased in applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freeze [2004]</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Rubin [2004]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dagan [2004]</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Sudicky [2004]</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Ginn [2004]</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Molz [2004]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Christakos [2004]</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Neuman [2004]</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Winter [2004]</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Returning to our definition of stochastic, i.e. dealing with uncertainty in a theoretically sound manner, we can look back at the claims on why stochastic methods are not used by practitioners. Some clarification can be made regarding the 2004 SERRA and 2016 WRR series’ comments on data needs. First, the notion that stochastic methods require extensive data is misleading. The purpose of stochastic methods is to compensate for uncertainty. To say one needs more information to describe uncertainty is just skipping the part about
CHAPTER 1. INTRODUCTION

Table 1.2: 2016 Status of Issue Topics

<table>
<thead>
<tr>
<th>Issues/Topics</th>
<th>Measurement technology and availability of data</th>
<th>User-friendliness of concepts</th>
<th>Theory too limited or not showcased in applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remains in 2016</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓</td>
</tr>
<tr>
<td>Components resolved by 2016</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
</tr>
<tr>
<td>Components added in 2016</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

actually dealing with uncertainty. For example in the issue of spatial variability, there is a difference in how well a variogram can describe the spatial structure of hydraulic conductivity in a soil and how easy it is to decide on the parameters of that variogram. The variogram is a stochastic method that represents structural patterns with a model that is parsimonious compared to a field of unknown values but allows for the uncertainty in the local variability in un-sampled locations. The lack of direct field data does make fitting a variogram difficult, but that does not imply that the method itself is not applicable. As mentioned by Cirpka and Valocchi [2016], Bayesian methods can use a variety of forms of information including prior knowledge and indirect data to help constrain uncertainty in models such as this.

Overall, from these two series of papers on why stochastic hydrogeology is not commonly used in practice, there are repeated areas in which research into effectively quantifying uncertainty can help. First is the availability of software that can integrate multiple generic forms of sparse information and handle the more esoteric aspects of stochastic hydrogeology in a user-friendly manner. Second is the development of methods to integrate more common forms of information obtained from practitioners. Third is the recognition that stochastic methods ought to be able to model and constrain the uncertainty regarding realistically complex geology.

1.2 Research Objectives

In response for the needs, detailed above, of the field of stochastic hydrogeology, this dissertation aims to contribute with the following research objectives:

**Provide Inverse Modeling Software** The first objective is to expand the availability of software tools for integrating multiple data types for subsurface characterization. The development of open source software is discussed. First, there is a desktop application with a graphical user interface that allows the user to upload direct and indirect measurements, prior knowledge, relevant forward models, and the random field generator of choice. The software is designed to be modular such that any forward model and random field generator can be connected such that any application with spatial un-
CHAPTER 1. INTRODUCTION

certainty can be characterized. Second, a supplemental package for the R statistical language is provided for more flexibility in the distribution calculations if the user is interested.

Expand Data Integration The two forms of information that are focused on are time series for inversion data and prior knowledge of conceptual models of geology. The first is common from the field (e.g. pumping drawdown curves or concentration breakthrough curves), but poses a problem for Bayesian inference since the inversion data is correlated in time and suffers from the ‘curse of dimensionality’, i.e. there is significant computational burden with each additional time step. How these time series can be efficiently incorporated into the Bayesian inference process without sacrificing information is explored. The second data type is common in that experts can have a conceptual idea of what geological patterns look like, but these patterns are either difficult to translate into geostatistical parameters or too complex to be accurately described by classical geostatistics. How training images can be used to incorporate conceptual models of geology into the Bayesian inference framework such that practitioners can incorporate their prior knowledge without translating it to mathematical expressions are also explored.

Assess Geological Realism The final research direction takes a step back from uncertainty quantification of spatial models and considers a three-dimensional data that exhibits realistically complex subsurface heterogeneity. This data set, referred to as an ‘aquifer analog’, is dissected and analyzed in terms of how the spatial patterns can be simplified when a prediction goal is specified. The approach of designing subsurface characterization specific to a pre-determined goal, referred to as ‘goal-oriented’ characterization, is a technique to highlight the usability of theoretical techniques in practice. Here, a goal-oriented approach is taken to consider which aspects of complex patterns need to be retained in parsimonious spatial models for characterization goals to be met while balancing model complexity and geological realism.

The contributions from these research objectives for the field of stochastic hydrogeology include the following:

- The desktop application provides a user-friendly tool to help integrate multiple forms of information about a field site to quantify the uncertainty about the spatial heterogeneity of hydraulic properties. Before its release, there was no software as widely applicable for combining geostatistical packages and numerical models for such a goal. With this ability, a wide range of applications with spatial heterogeneity can have their uncertainties quantified.

- The R package supplements the desktop application for users who are more comfortable with programming and statistics. For example, if the user wants to modify the
algorithm for the kernel density estimation step, then the R package provides a platform for adjusting the likelihood calculation while keeping the rest of the Bayesian inference process intact.

- Exploring an example for incorporating time series into the Bayesian inference process provides a methodology for future studies in incorporating data series in general. Other time series (e.g. breakthrough curves of a tracer) or other forms of data series (e.g. depth profiles of hydraulic head) can be utilized. Although these time series have been used before in aquifer characterization studies (e.g. pumping tests), they have not been explored in the Bayesian inference process for inferring geostatistical parameters for general applications.

- Providing a proof-of-concept study of inferring conceptual models with the Bayesian inference process opens the door for an even wider array of applications. By removing the distributional assumptions of the geostatistical parameters used in previous examples, Bayesian inference of spatial heterogeneity is shown to be applicable to realistic geology, such as channels of soil types. It is shown that training images to describe complex spatial patterns can be inferred from candidate training images with Bayesian inference and the desktop application described above.

- The synthetic exercise of simplifying an aquifer analog provides an example on how realistically complex geologies can be simplified into spatial models. Tailoring the simplicity of spatial patterns, which is a necessity when we only have sparse information from the field, to the prediction goal can ensure that the ultimate goal of improving groundwater modeling efforts is achieved.

In Chapter 2, background material is provided on the field of geostatistics and the concepts behind applying Bayesian inversion for geostatistical parameters. The remainder of this dissertation addresses the above three research objectives. Chapter 3 addresses both the desktop application (Section 3.1) as well as the R package (Section 3.2). Chapter 4 addresses integrating both the time series (Section 4.1) as well as the conceptual models of geological heterogeneity (Section 4.2). Chapter 5 addresses the geological heterogeneity. Chapter 6 summarizes the contributions of this research to the field of stochastic hydrogeology.
Chapter 2

Background

This chapter serves to provide a background on the geostatistical and Bayesian inference theory and algorithms used in the following chapters. Geostatistics allows for creating models to describe global spatial patterns of the subsurface (hence ‘geo-’) in a stochastic manner in order to incorporate uncertainty regarding the local variability (hence ‘-statistics’). Two approaches for creating these models are explained along with the respective algorithms used to generate random realizations with those two approaches. Bayesian inference comes into play when we want to quantify and constrain our uncertainty regarding the parameters of these geostatistical models. The inference framework can integrate multiple forms of information that we have about a site to provide a description of the plausible values of the geostatistical parameters.

2.1 Geostatistics

Spatial random functions

Spatial random functions (SRFs) are random variables that are a function of space. A random variable comes from a distribution which can be described by statistics such as mean and variance. For a SRF, the mean and variance are functions of space. SRFs are used to describe spatially-variable fields (also known as random fields) and each point in the field can be thought of as a random variable. Collections of points are conditional on one another. Each point has a conditional cumulative distribution function (CCDF) that describes the probability of that point having certain values based on the values of the neighboring points. For example, the field $Z$ has spatial coordinates $x$ and its CCDF is

$$F_Z(z, x, d_n) = P(Z(x) \leq z \mid d_n) \quad \text{CCDF} \quad (2.1)$$

$$d_n = \{Z(x_1), Z(x_2), \ldots, Z(x_n)\} \quad \text{Data events at the other } n \text{ points in } Z \quad (2.2)$$

This CCDF describes the cumulative probability that a specific point in space will have a certain value given the values at neighboring points. For example, if a point in the subsurface...
is surrounded by hydraulic conductivity values on the order of $10^{-2} \text{m/s}$, then the probability of that point also having hydraulic conductivity on the order of $10^{-2} \text{m/s}$ is high. How to handle the CCDF of $Z$ depends on the application. The following two sections describe two approaches, two-point and multi-point statistics.

**Two-point statistics**

Two-point statistical methods describe how the covariance between pairs of points change over increasing distance between pairs. Covariance is the measure of how two random variables change with one another, so it is a convenient tool to describe the relationship between different points in a random field. For example, if two locations in the subsurface are 1 mm apart, then their hydraulic conductivity values are probably correlated. If that distance increases to 1 cm, then the two values are still probably correlated but maybe not as much as when the distance was 1 mm. Then, if the distance becomes 1 m or 10 m, the two locations may not be correlated at all. The correlation, and thus covariance, is decreasing with distance between the two locations. For two locations, $a$ and $b$, in the random field $Z$, the two-point covariance function in general is:

$$C_Z(a, b) = \text{Cov}(Z(a), Z(b))$$

(2.3)

$$= E[(Z(a) - E[Z(a)]) - (Z(b) - E[Z(b)])]$$

(2.4)

$$= E[Z(a)Z(b)] - E[Z(a)]E[Z(b)]$$

(2.5)

Where $E[\cdot]$ denotes the expected value operator. Assuming stationarity, where the mean $\mu$ and covariance do not change over the field, the covariance can be described by the relative position between the two points, which is described by the distance $h$ between them:

$$C_Z(a, b) = C_Z(|b - a|) = C_Z(h) = E[Z(x_i)Z(x_i + h)] - \mu^2$$

(2.6)

where $\mu^2$ is the global mean of the field, e.g. the average hydraulic conductivity across a field site. This covariance function $C(h)$ is commonly expressed as one of several covariance models such as exponential, Gaussian, or spherical. Although, the variogram is also often used to express the covariance structure of a field:

$$\gamma_Z(h) = \frac{1}{2}E[(Z(x_i) - Z(x_i + h))^2]$$

(2.7)

$$\gamma_Z(h) = \sigma_Z^2 - C_Z(h)$$

(2.8)

where $\sigma_Z^2 = C_Z(0)$ and represents the global variance of the field. Table 2.1 gives examples of common variogram models used in practice, were $l$ represents the correlation length scale.

The empirical (or raw) variogram can be fitted to data to estimate the parameters using weighted ordinary least squares, but a sufficient amount of data is not commonly available. For example, field sites may have adequate measurements across vertical distances from drilling boreholes into the subsurface, but these boreholes are often spread out such that
there is not enough information across horizontal distances. See Section 3.1 for how to handle parameter uncertainty due to sparse direct data. Given the geostatistical parameters, the mean and variance of individual locations can be calculated with the kriging method. Kriging entails solving a system of equations relating the measurement values and their locations along with a chosen variogram model. One could apply kriging to estimate the average field via geostatistical interpolation, but this does not incorporate the uncertainty in local variability. Next, we look into how these two-point geostatistical models are implemented to generate random fields for both continuous and discrete variables.

**Gaussian random fields (continuous variables)**

Gaussian fields are modeled as multivariate Gaussian distributions where the locations in the field are the dimensions. This SRF is appropriate when the field is a continuous variable and the values are normally distributed when spatial coordinates are ignored (e.g. the natural logarithm of hydraulic conductivity).

\[ Z \sim \text{MVN}(\mu, \Sigma) \tag{2.9} \]

where \( \mu \) is a vector of the expected values of each location and \( \Sigma \) is the covariance matrix for all the locations. The covariance matrix is filled by the covariance model \( C(h) \) chosen for the field. The algorithm for simulating a Gaussian random field is described in Algorithm 1. Essentially, every point in the field is assigned a value from sampling its CCDF based on previously filled points and all the points are filled by a random order.

**Indicator random fields (categorical variables)**

Indicator random fields describe categories, such as soil types or geologic layers, instead of continuous variables. Simulating an indicator field is very similar to simulating a Gaussian field but with just a different CCDF formulation. The probability distribution for a point is a Bernoulli random variable described by parameter \( p \). The algorithm for simulating an indicator random field is described in Algorithm 2.

In summary, both continuous and categorical random fields can be described and simulated with the use of two-point statistical methods. However, there is a limitation to the spatial patterns that these methods can describe. The following section describes multipoint methods that can describe more complex spatial patterns.
CHAPTER 2. BACKGROUND

Algorithm 1: Sequential Gaussian Simulation

Data: measurements of the field, if available; Covariance model $C(h)$

Result: random field adhering to covariance model and conditioned on measurements (if available)

1 Create grid of desired size and resolution
2 if available then assign measurements to grid;
3 while unassigned elements remain do
4 Randomly pick a remaining unassigned element to be current element;
5 Determine which assigned elements are in neighborhood surrounding current element;
6 Apply kriging to obtain estimate $\mu_{ok}$ and variance $\sigma_{ok}^2$ for current element using assigned elements in neighborhood and $C(h)$;
7 Pull one sample from the univariate normal distribution $N(\mu_{ok}, \sigma_{ok}^2)$;
8 Assign sampled value to current element;
9 end

Algorithm 2: Sequential Indicator Simulation

Data: measurements of the field, if available; Covariance model $C(h)$

Result: random field adhering to covariance model and conditioned on measurements (if available)

1 Create grid of desired size and resolution
2 if available then assign measurements to grid;
3 while unassigned elements remain do
4 Randomly pick a remaining unassigned element to be current element;
5 Determine which assigned elements are in neighborhood surrounding current element;
6 Apply kriging to obtain estimate $p_{ok}$ for current element using assigned elements in neighborhood and $C(h)$;
7 Pull one sample $u$ from the univariate uniform distribution $U(0, 1)$;
8 if $u < p_{ok}$ then
9 assign the value 1 to current element
10 else
11 assign the value 0 to current element
12 end
13 Assign sampled value to current element;
14 end
CHAPTER 2. BACKGROUND

Multipoint statistics

Multipoint statistical (MPS) methods differ from two-point methods by not only considering pairs of points to build the CCDF, but rather multiple points. Instead of covariance models, MPS uses a training image (TI) to describe spatial patterns (see Figure 2.1). There have been multiple algorithms proposed to construct the CCDF for MPS simulations, but the initial ones were limited by memory constraints of computers when all spatial patterns in the TI were recorded. A relatively new algorithm, Multipoint Direct Sampling (MPDS) [Mariethoz et al., 2010], is based on the idea that the TI does not have to be initially scanned for a survey of the patterns it holds. In MPDS, the TI is randomly sampled in that a random subset of the TI is defined as a search neighborhood and that neighborhood is scanned for a certain pattern as needed for the current element for assignment. The same sequential approach as in Sequential Gaussian Simulation and Sequential Indicator Simulation is used in that a random path is taken to sequentially assign values to unassigned elements while taking into consideration the previously assigned elements nearby. The key differences are that the neighborhood size for incorporating conditional data events changes over time in MPDS and that the TI is sampled instead of an explicit distribution. The algorithm for simulating a multipoint field is described in Algorithm 3.

![Figure 2.1: An example training image for multipoint statistics.](image)

Since configurations of more than two points in the field are searched during each iteration, the spatial patterns simulated can have more complicated patterns than if only two-point configurations were used. Depending on how the search neighborhood is set, or depending on the complexity of the training image, MPS can simulate the same spatial patterns as two-point methods. However, if the spatial pattern can be explained by two-point methods, the computational cost is less than the multipoint methods.
Algorithm 3: Multipoint Direct Sampling

Data: measurements of the field, if available; Training Image (TI)
Result: random field adhering to patterns in TI and conditioned on measurements (if available)

1 Create grid of desired size and resolution
2 if available then assign measurements to grid;
3 while unassigned elements remain do
4     Randomly pick a remaining unassigned element to be current element;
5     Determine neighborhood size needed to encompass the right number of assigned elements;
6     Randomly pick a subset of the TI the size of the neighborhood;
7     Scan subset for configuration of values matching assigned elements from original neighborhood until found;
8     Assign found value to current element;
9 end

2.2 Bayesian Inverse Modeling

Method of Anchored Distributions

The Method of Anchored Distributions (MAD) is a Bayesian framework for applying inverse modeling to geostatistics. The results are posterior distributions such that the uncertainty in the geostatistical parameters is quantified and can be propagated into future stochastic modeling. Multiple types and scales of field information can be incorporated and the anchored distributions convert indirect data into direct data.
CHAPTER 2. BACKGROUND

Definition of terms:

\[ Y(\mathbf{x}) : \text{Spatial random function} \quad (2.10) \]
\[ y(\mathbf{x}) : \text{Realization (random field) of } Y \quad (2.11) \]
\[ \mathbf{x} : \text{Spatial coordinates} \quad (2.12) \]
\[ z_a(\mathbf{x}) = f(y(\mathbf{x})) + \epsilon_a : \text{Direct (Type-A) data} \quad (2.13) \]
\[ z_b(\mathbf{x}) = M(y) + \epsilon_b : \text{Indirect (Type-B) data} \quad (2.14) \]
\[ M : \text{Numerical model} \quad (2.15) \]
\[ \theta : \text{Parameters of } Y \quad (2.16) \]
\[ \vartheta : \text{Anchored distributions ("anchors")} \quad (2.17) \]
\[ P(\theta) : \text{Prior Distribution} \quad (2.18) \]
\[ P(z|\theta) : \text{Likelihood Distribution} \quad (2.19) \]
\[ P(\theta|z) : \text{Posterior Distribution} \quad (2.20) \]

MAD derivation

Bayesian inference is the method of updating prior knowledge (e.g. a range of hydraulic conductivity values at similar sites) with evidence (e.g. sparse measurements of hydraulic conductivity at the current site) in order to obtain posterior knowledge and is derived from Bayes’ Theorem. MAD is an application of Bayesian inference where the knowledge we wish to update is about both global (\(\theta\), e.g. the mean hydraulic conductivity across the field site) and local (\(\vartheta\), e.g. the hydraulic conductivity at a specific location at the field site) spatial distributions and the evidence comes in the form of direct (\(z_a\), e.g. measurements of hydraulic conductivity) and indirect (\(z_b\), e.g. measurements of hydraulic head) measurements. Below is the derivation from conditional probability to Bayesian inference to MAD as implemented.
\[ P(AB) = P(A|B)P(B) = P(B|A)P(A) \] Joint and conditional probability (2.22)

\[ P(A|B) = \frac{P(A)P(B|A)}{P(B)} \] Bayes’ Theorem (2.23)

\[ P(\theta|z) \propto P(\theta)P(z|\theta) \] Bayesian Inference (2.24)

\[ P(\theta, \vartheta|z_a, z_b) \propto P(\theta, \vartheta)P(z_a, z_b|\theta, \vartheta) \] Extend parameters and data (2.25)

\[ \propto P(\theta, \vartheta)P(z_b|\theta, \vartheta, z_a)P(z_a|\theta, \vartheta) \] Separate the data types (2.26)

\[ \propto P(\theta, \vartheta)P(z_b|\theta, \vartheta, z_a)P(\theta, \vartheta|z_a) \] Use Bayes’ to replace last term (2.27)

\[ \propto P(\theta, \vartheta|z_b|\theta, \vartheta, z_a)P(\theta|\vartheta, z_a) \] Cancel terms (2.28)

\[ \propto P(z_b|\theta, \vartheta, z_a)P(\vartheta|\theta, z_a)P(\theta) \] Pull out \( \theta \) prior (2.29)

\[ P(\theta, \vartheta|z_a, z_b) \propto P(\theta)P(\vartheta|\theta, z_a)P(z_b|\theta, \vartheta, z_a) \] MAD as implemented (2.30)

The reason why the initial equality in Bayes’ Theorem is subsequently expressed as a proportionality is that the missing denominator \( P(z_b) \) is not easily obtained and ultimately not necessary. Since we know the posterior distribution on the LHS must integrate to unity with respect to the parameters for it is be a proper PDF and that \( P(z_b) \) is a constant with respect to the parameters, we can scale the righthand side of the proportion after it is calculated. How each term on the righthand side is handled in applications is described in the next subsection.

**MAD implementation**

The MAD terms are implemented as such:

\( P(\theta) \) : This is the prior distribution of the structural parameters (i.e. the parameters of the geostatistical model that describes the global heterogeneity pattern) that are to be inferred. For each parameter, previous literature and expert opinion are consulted to define a distribution of plausible values. This can be a PDF ranging over values that previous studies have found at similar sites (using a uniform distribution is common). It can also be a discrete distribution if the parameter is discrete, such as with training images. If there is more than one parameter being inferred, the distributions can be formed separately and then sampled such that the parameter space is fully represented (e.g. with latin hypercube). This assumes that the parameters are independent.

\( P(\vartheta|\theta, z_a) \) : These are the anchors’ prior distributions, which are conditioned on the structural parameters and the Type-A measurements. The construction of these prior distributions is dependent on the geostatistical model formulation being used. For two-point methods, ordinary kriging is used to determine the relevant parameters (i.e. \( \mu_{\text{ok}} \) and \( \sigma_{\text{ok}}^2 \) for continuous, or \( p_{\text{ok}} \) for categorical) of the estimate for the location of the anchor and relevant distribution (i.e. Gaussian or Bernoulli) with those parameters is the prior distribution for
the anchor. If multipoint statistics formulations are used, training images can be scanned to estimate the probability of values at the anchor locations based on the configurations of the direct data.

\[ P(z_b|\theta, \vartheta, z_a) : \] This is the likelihood term and it is estimated via executing the numerical model \( M \) on each field \( y \) generated from samples of \( P(\theta)P(\vartheta|\theta, z_a) \) (hence the conditions) and comparing the simulated values at the Type-B measurement locations with the measured values. Kernel density estimation is used to non-parametrically estimate the likelihood value for each sample. See the next section for a discussion on kernel density estimation limitations.

\[ P(\theta, \vartheta|z_a, z_b) : \] After finding the likelihood term \( P(z_b|\theta, \vartheta, z_a) \), it is used to weight the prior distribution in order to update the prior knowledge into posterior knowledge. After the weighting, the values are scaled such that they, and thus the posterior distribution, integrate to unity.

A desktop software MAD# [Osorio-Murillo et al., 2015] (see Section 3.1) has been developed for users to supply prior distributions, measurement values, and numerical models (\( M \)) which are taken and run in batch jobs in order to estimate the likelihood and, ultimately, the posterior distributions. The software is modular in that it is designed to be coupled with a variety of random field generators and numerical models. The source code is open and available at mad.codeplex.com along with extensive documentation. See Section 3.1 for a thorough discussion on this software.

**Density estimation**

Density estimation in general is the process of converting a histogram into a continuous function that integrates to unity, known as a probability density function (PDF). There are two main classes of density estimation: parametric and non-parametric. Parametric is when the distribution type has a known analytical formula and only its parameters need to be fitted. For example, if a selection of samples appear to be normally distributed, then parametric density estimation would be estimating the mean and variance from the samples and generating a normal distribution with those parameters. In MAD, the density distribution of the likelihood distribution needs to be estimated. For example, the samples of simulated hydraulic conductivity at a specific point can be shown with a histogram. If that histogram appears normally-distributed, then parametric density estimation can be used. However, hydraulic head is often not normally-distributed in this context. If a parametric distribution can not be assumed, such as in this case for hydraulic head, then non-parametric techniques are needed. A common technique is kernel density estimation (KDE), which essentially assigns a kernel to each sample and vertically sums the kernels across the sample space and scales in order to integrate to unity. A kernel in this case is a non-negative function that integrates to unity, has a zero mean, and has a parameter called a ‘bandwidth’ that controls its width across the sample space.
CHAPTER 2. BACKGROUND

\[ \hat{f}_h = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) \]  

(2.31)

where \( K(\cdot) \) is the kernel, \( h \) is the bandwidth, and \( x_i \) is the \( i \)th sample. A common kernel is a zero-mean normal distribution with the variance as the bandwidth parameter. Figure 2.2 is an example of kernel density estimation for a few samples.

The selection of the bandwidth effects the shape of the resulting density function and depicts uncertainty. There is a trade-off between incorporating the evidence of the samples for which we want the bandwidth to be as small as possible and reflecting the true underlying distribution. As the number of samples increases, the bandwidth can be reduced. There are various approaches for choosing the correct bandwidths for each kernel in an estimation and KDE functions in R generally come with those approaches as default options such that the user does not have to specify any. In MAD#, the bandwidth is left to the default approach, but is increased when Type-B measurement error is provided by the user. Figure 2.3 shows the effect of the bandwidth of KDE for the same samples as in Figure 2.2.

The disadvantage of non-parametric techniques is that they generally require more samples in order to resolve the density function over the sample space because the general shape is unknown (or at least not assumed). This disadvantage is amplified if the sample space is multi-dimensional as per the curse of dimensionality, which is described in the next section.

The curse of dimensionality

The ‘curse of dimensionality’ is the increasing need for more samples or computational effort to characterize a system as more dimensions or variables are introduced. In MAD, this manifests via each inversion data measurement that we have that is correlated with other inversion
data measurements. Each additional correlated measurement includes another dimension in the probability distribution \( p(z_b | \hat{z}_b) \) (where \( z_b \) is the vector of observed measurements and \( \hat{z}_b \) is the matrix of co-located simulated values from the realizations of a given sample). This distribution is used to estimate the likelihood of a sample. Correlated measurements can be from a series of measurements in time (e.g. a concentration breakthrough curve) or in space (e.g. a vertical depth profile from a borehole). Correlated measurements cannot have each of their likelihood estimated in a univariate PDF because they are dependent variables.

So why is it a ‘curse’ to need so many dimensions? The first reason is that the number of samples needed to resolve density distributions increases exponentially with each additional dimension. For example, to estimate a univariate density distribution for a given variable, \( N \) samples would suffice given the method (non-parametric methods would need more than parametric methods). If there was a second variable that needed its univariate density distribution estimated, then another \( N \) samples of the variable would be needed. However, if these two variables are dependent, you would need enough samples to cover the whole parameter space which is two-dimensional now and no longer one-dimensional. The samples would need to be of pairs of the two variables and be distributed over both axes such that \( N^2 \) samples are needed. If there are \( M \) correlated measurements, then \( N^M \) samples are needed to resolve the distribution. Figure 2.4 shows densities estimated by KDE in one and two dimensions from samples pulled from a normal distributions. We can expect a bell curve for the 1D case and concentric circles for the 2D case. The rows are for 50, 500, 5000, and 50000 samples.

Beyond just needing more and more samples, there is also a numerical reason why high dimensionality, i.e. too many correlated measurements, in MAD is a curse. The area under the curve of the estimated PDF must integrate to unity given the definition of probability. This holds true no matter the dimensions. However, more dimensions means a greater probability space over which probability must integrate to unity. This results in the likelihood
Figure 2.4: The curse of dimensionality.
values becoming more extreme (either approaching infinity if variables are defined over ranges of values less than one, or approaching zero if variables range over values greater the one) as dimensions increase. As the magnitudes of the likelihoods become of greater and greater order of magnitude, there is greater variety between samples’ likelihood values. The second most likely sample can be two orders of magnitude less in likelihood compared to the most likely sample and the resulting posterior density is $100\times$ higher, which is highly improbable. The result is a posterior distribution that is solely based on one sample and is effectively only the kernel of that samples scaled to sum to unity over the parameter space.
Chapter 3

Open-source Software Tools for Inversion

This chapter discusses the software development for implementing the Bayesian inference framework into user friendly applications and packages. Below are abstracts for the following sections on a desktop application with a graphical user interface and an R package for extending the flexibility of the likelihood calculations. For an application of the software, see Heße et al. [2015].

Section 3.1 Estimation of spatially random fields (SRFs) is required for predicting groundwater flow, subsurface contaminant movement, and other areas of environmental and earth sciences modeling. This paper presents an inverse modeling framework called MAD# for characterizing SRFs, which is an implementation of the Bayesian inverse modeling technique Method of Anchored Distributions (MAD). MAD# allows modelers to “wrap” simulation models using an extensible driver architecture that exposes model parameters to the inversion engine. MAD# is implemented in an open source software package with the goal of lowering the barrier to using inverse modeling in education, research, and resource management. MAD# includes an intentionally simple user interface for simulation configuration, external software integration, spatial domain and model output visualization, and evaluation of model convergence. Two test cases are presented demonstrating the novel functionality of this framework to apply inversion to calibrate the model parameters characterizing a groundwater aquifer.

Section 3.2 The Method of Anchored Distributions (MAD) is a method for Bayesian inversion designed for inferring both local (e.g. point values) and global properties (e.g. mean and variogram parameters) of spatially heterogenous fields using multi-type and multi-scale data. Software implementations of MAD exist in C++ and C# to import data, execute an ensemble of forward model simulations, and perform basic post-processing of calculating likelihood and posterior distributions for a given application. This article describes the R package anchoredDistr that has been built to provide an R-based environment for this method. In particular, anchoredDistr provides a range
3.1 A Software Framework for Inverse Modeling and Uncertainty Characterization

This section is copied, with permission, from the publication Osorio-Murillo et al. [2015].

Introduction

Overview

Spatial phenomena variability is typically evaluated through analytical and numerical models that describe the general properties of spatial random fields (SRFs). These models employ parameters and observations to define spatial variability. The characteristics - and hence variability - of an SRF can be discerned by the relationship between model parameters, direct, and indirect information. A number of hydrogeological studies have been conducted using SRF analysis [Delhomme, 1979; Carrera and Neuman, 1986; Dagan, 1987; Bates and Townley, 1988; Bellin and Rubin, 1996; Yeh, 2002; Kanso et al., 2003; Gallagher and Doherty, 2007; Farmani et al., 2008]. This paper introduces an open source inverse modeling framework, called MAD# (pronounced “mad sharp”), focused on the characterization of SRFs using the Method of Anchored Distributions (MAD), a Bayesian inverse modeling technique [Rubin et al., 2010].

The process of estimating model parameters from the inversion of governing equation(s) and observations is called inverse modeling. For over fifteen years, researchers have advocated for the development of flexible and easy-to-use inverse modeling tools, with the understanding that the shortage of such tools hinders the development of comprehensive and credible uncertainty quantification tools [Poeter and Hill, 1997, 1999; Rubin, 2004; Dagan, 2011]. Carrera et al. [2005] identified five features that are needed for broad adoption of inverse modeling tools in hydrogeology: 1) incorporating geological data, 2) improving the flexibility of the code and procedures to handle any and all relevant data types, 3) a complete quantification of uncertainty, 4) reducing the difficulty of code operation, and 5) coupling inverse modeling techniques with a geographic information system (GIS) platform.

A number of existing simulation model software tools include model parameter estimation and uncertainty characterization as embedded functions within the program. For example, WEAP [Yates et al., 2005] and PMWIN [Chiang and Kinzelbach, 2001] both are applications that use forward models (FMs) and model parameter estimation software applications like...
PEST [Doherty, 1994]. These and related software tools have aided adoption of uncertainty characterization and inverse modeling to some degree. However, we recognize a need for additional tools that provide a more general set of capabilities and that address the issues raised by Carrera et al. [2005].

MAD has been shown by Rubin et al. [2010], Murakami et al. [2010], and Chen et al. [2012] to be a flexible stochastic inverse modeling technique that addresses the first three challenges posed by Carrera et al. [2005]. Specifically, MAD can account for geology (Challenge #1) via the representation of geological features through SRFs modeled using structural parameters; handles multiple relevant data types (Challenge #2) through use of direct measurements and measurements that are indirectly related to the variable modeled; and accommodates uncertainty (Challenge #3) by explicitly incorporating observation uncertainties and quantifying uncertainty of geostatistical structural parameters and a new concept called “anchors”.

Research goals

We have endeavored to address Carrera’s Challenges #4 and #5 by implementing and testing MAD in an extensible, user-friendly software framework. Specific goals for the developed framework include:

1. It should be capable of generically accommodating FMs that relate target variables with observations.
2. It also should be flexible in supporting the use of other user-specified software packages for random field generators (RFGs).
3. It should be able to characterize the uncertainty associated with SRFs.
4. It should be well documented and transparent with independently verifiable results.

The remainder of this paper presents our approach to meeting the research goals noted above in the form of an open source inverse modeling software framework called MAD#. This new inverse modeling application builds upon a prototype architecture [Osorio et al., 2012], in which MAD was implemented as a HydroDesktop [Ames et al., 2012] plugin using an embedded steady-state head solver written in R statistical software. MAD# is a standalone desktop application and includes an architecture for adding custom random field generator drivers (RFGDs) and forward model drivers (FMDs) for incorporating new models. We present an architectural overview of MAD# and descriptions of drivers currently implemented. We also present a demonstration of MAD# in two synthetic pumping experiments using a MODFLOW [Harbaugh and McDonald, 1996] project created in the PMWIN MODFLOW interface [Chiang and Kinzelbach, 2001].

The work presented here is related to an active area of research and development within the broader context of integrated environmental modeling in that our software framework is indeed a method for “integrating” different modeling software packages into a single cohesive environment. This approach is related to the approach supported by as OpenMI [Castronova...
et al., 2013, Knapen et al., 2013. Ridler et al. 2014 follow a strikingly similar approach to developing a data assimilation framework using OpenMI and an open data assimilation library, using the C# programming language. Another model integration framework that is rapidly growing in adoption is the Community Surface Dynamics Modeling System (CSDMS) which uses a “wrapper-style” common modeling interface approach which is similar in nature and purpose to the forward model driver approach we present herein Overeem et al., 2013, Castronova and Goodall, 2010.

It is worth noting that the OpenMI and CSDMS approaches both presume the existence of software packages that implement particular hydrologic or environmental numerical models. In other words, these models exist as software packages that require specific input and output files - not simply as conceptual mathematical models. A somewhat different approach has been taken by integrated modeling efforts such as The Object Modeling System (OMS) which is suited to integrating small functions or codes that represent individual physical processes rather integrating large software packages David et al., 2013.

In comparison to the OpenMI and CSDMS integrated modeling systems, our MAD# approach is relatively simple. Rather than facilitating the transmission of inputs and outputs between various numerical model packages, our approach focuses on a tightly managed system of a single model package connected through a single wrapper (or “driver”) directly to our inversion software. This is much more manageable than the alternative of linking multiple models to multiple models, and helps us avoid the challenges and issues of the so-called “Integronsters” Voinov and Shugart, 2013.

Methods

MAD theoretical background

Although a complete description of MAD is outside the scope of this paper, a brief introduction to the method is presented here. MAD is a Bayesian inverse modeling technique focused on characterizing SRFs by using Bayes’ theorem and the following concepts intended to address the challenges stated in the previous section:

- Geostatistical models are used to capture large-scale trends and reproduce patterns of spatial variability in terms of SRFs.

- Data classification - MAD classifies data (measurements) in a general format that is not limited (or specific to) any particular discipline or application. MAD categorizes data as:
  - Type A data, $z_i = y(x_i) + \epsilon_i$, $i = 1, ..., N$, which could include direct measurements (including measurement error $\epsilon$) of the target variables (e.g. hydraulic conductivity) at location $x_i$, $i=1,..,N$, or other types of measurements (e.g. transmissivity) at $x_i$ that could be directly related to the target variable at $x_i$,
Type B data, $z_b = M(x_i) + \epsilon_b, j = 1, \ldots, M$, which include all measurements (including measurement error $\epsilon$) that do not conform with Type-A, but are related to the target variable via a forward model, $M$ (e.g. pressure head).

- Localization through anchored distributions (or “anchors”). An anchor is a statistical distribution of a target variable at a given location. Anchors can be employed for multiple target variables and/or locations. Anchors intend to capture local effects in the field of the target variables by conditioning realizations on fields.

MAD defines a target variable as a SRF, which is represented by a vector of geostatistical structural parameters ($\theta$) capturing the global tendency, and anchors ($\vartheta$) for quantifying local variations of the parameter field. MAD relies on the following proportionality [Rubin et al., 2010].

$$p(\theta, \vartheta | z_a, z_b) \propto p(\theta, \vartheta | z_a) p(z_b | \theta, \vartheta, z_a)$$  \hspace{1cm} (3.1)

where $p$ indicates a probability density function (pdf) and $p(\theta, \vartheta | z_a)$ is the joint prior distribution of the structural parameters and anchors conditional on Type-A data vector $z_a$, and $p(z_b | \theta, \vartheta, z_a)$ is the likelihood of observing the Type-B data vector $z_b$ given the structural parameters, anchors and Type-A data. Finally, $p(\theta, \vartheta | z_a, z_b)$ is the joint posterior distribution of the structural parameters and anchors conditional on both Type-A and Type-B data.

**MAD methodological approach**

MAD is applied in three stages: 1) Strategy, 2) Implementation, and 3) Assessment. These three stages are described in Figure 3.1 and are discussed in depth in the following three subsections.

![Figure 3.1: Structure of the MAD process](image)

**Strategy**

The first stage of applying MAD to a case study is the formulation of a strategy. This strategy entails the following six elements: (1) identifying target variables, (2) selecting
appropriate priors for the SRF parameters, (3) identifying types of data available, (4) selecting numerical modeling strategy, (5) selecting locations for anchors, and (6) planning post-calibration model testing. A target variable is typically a heterogeneous variable that needs to be characterized. Measurements of target variables or directly related variables are classified as Type-A data. A SRF model type to describe a target variable’s spatial variability in the field needs to be chosen, and this choice can be made from analyzing the measurements and previous literature. The second consideration is choosing appropriate priors for the parameters of the SRF model types chosen for the target variables. After these first two steps, the target variable is set up for MAD.

The third step focuses on identifying the data that could be used for the inversion process. Since the inversion requires indirect data, measurements need to be taken of a variable that the target variable influences via a mathematical model. These measurements are classified as Type-B data. This inversion data type generally describes larger-scale phenomena than the target variable and thus the combination of the two better informs the inversion process.

The fourth step is creating a numerical model. The numerical model needs to take the target variable, or a variable directly related, as an input and produce the inversion data type as an output. Any relevant environmental influences from the site (e.g. wells or streams) need to be numerically represented. Other considerations that are necessary for building a well-posed model are also required, including dimensionality, boundary conditions, and time dependence. The Type-A and Type-B data need to be collected from within the domain selected and adequately distant from the boundaries to prevent interference.

The fifth step is placing the anchors in locations that are influential in the environmental process being modeled. Optimal selection of anchor locations was discussed by [Yang et al. 2012]. Each anchor also requires a prior distribution defined by prior knowledge. Finally, one must choose an evaluation method for assessing the success of the inversion approach and testing calibration. Multiple inversions can be performed and cross-validation can be used to determine which approach is most successful. At this point, the strategy is developed and MAD can be implemented.

**Implementation**

The four steps of the implementation stage are (1) sampling from prior distributions, (2) creating realizations, (3) executing numerical model simulations, and (4) extracting results. The first two steps cover the sampling strategy. In the first step, each SRF model parameter and anchor will need to have its prior distribution sampled, creating \( \theta_i \) and \( \vartheta_i \), \( i = 1, \ldots, N \) where \( N \) is the number of samples. The number of samples needs to be high enough to cover the parameter space, but choosing the number is not an exact science. Evaluating if there are not enough samples is covered in the assessment stage. In the second step, the samples from the previous step are used to create realizations of the target variable field. Each \( \theta_i \) will define a SRF model to create realizations conditional on \( \vartheta_i \) and the Type-A data. The third step in the implementation stage is to run the simulations. For each realization created, the numerical FM is applied and a simulation is created. In the fourth step, the relevant Type-B
simulated values are extracted and used to calculate likelihoods and posterior distributions. The extracted Type-B values are the $z_b$ vector for which likelihoods are calculated. The method for calculating likelihood is not specific to MAD such that any applicable to the $z_b$ vector and Type-B measurements may be used. The likelihood method is applied to each sample, yielding a likelihood distribution across the sample space. This is multiplied by the prior distributions resulting in the posterior distributions and concluding the implementation stage.

**Assessment**

The assessment stage of the MAD approach focuses on assessing: (1) convergence and (2) general strategy. Convergence is assessed in two ways: if there were enough realizations and if there were enough samples. Generally, enough realizations are needed to estimate accurate likelihood values and enough samples are needed to resolve accurate posterior distributions. Graphical examples are given in the next section on how to assess convergence.

The second kind of assessment, assessing general strategy, is more open-ended. The success evaluating technique chosen in the strategy stage is applied here. If the chosen success criteria are not met, then the strategy can be modified in several ways, including: increasing the total number of measurements if possible, changing the SRF model type or which parameters are random, or modifying the FM. If the success criteria are indeed met, it is still advisable to compare different parameters of the SRF model (e.g. covariance functions), or FMs to address model uncertainty. With this evaluation and acceptance of success, MAD has been thoroughly applied.

**MAD# software framework**

This section describes the design and development of an open source software framework that implements the MAD methodological approach described above. The software, MAD#, is designed as an extensible architecture that uses generic functions for sharing information, executing processes and extracting data of FMs and RFGs, which can be stand-alone software applications, packages of a statistical frameworks, or libraries. To support these diverse applications, MAD# framework uses a driver approach to connect the FMs and RFGs through forward model drivers (FMDs) and random field generator drivers (RFGDs) respectively with the framework. These drivers are libraries that implement a set of interfaces of the MadInterfaces library (Figure 3.2).

This driver-based approach is expected to facilitate adoption and extension of the system by 3rd parties who can create new drivers for supporting new FMs or RFGs. MAD# is programmed using the .NET framework and the open source DotSpatial GIS programming library [www.dotspatial.org](http://www.dotspatial.org) following a similar approach as used in [Ames et al., 2012]. An open source approach was chosen to support transparency and adaptability of the software [Alexandrov et al., 2011]. The DotSpatial library provides the geographic and display functionalities of MAD#, including data management, control, projection, symbology, and
extension management. Statistical libraries including Math.NET \cite{team2014} and ALGLIB \cite{Bochkanov2014} are also used in the framework. A customized version of MapWindow 6 \cite{Dunsford2011} serves as the final user interface of MAD#.

Data structure

The MAD# data structure is based on three entity-relationship models (Figure 3.3) implemented using open source database SQLite. Three SQLite file databases “XMAD”, “XRESULT” and “XDATA” are created in different stages of the MAD# process. The XMAD file stores the information generated in pre-processing module. The main tables of the XMAD file structural are explained. The variable table contains the list of variables provided by the FMD. These variables are used as key words during whole MAD# process. The Domain table stores the geographic information of the FM domain. The Measure table manages the Type-A and Type-B data and anchors. The prior information of the structural parameters and anchor are stored in the PriorData table. The $z_b$ vector is stored in the SelectionValues table. This table is used to link the information generated by the simulation of the FM. In the processing module is generated a XDATA file database for each sample. This structure was selected due to size of the files generated by large simulation projects. These files store the simulated Type-B per realization in the ResultSelection table, which is related to the SelectionValues table matching each output with the $z_b$ vector. Also, the processing module creates a XRESULT file database that contains a copy of all tables of the XMAD file and the parameter used by the user to execute the simulation in the ConfigurationResult table. The XRESULT file is used in the post-processing module to
store the likelihood calculation in the LikelihoodGroupValue table.

![XMAP database diagram](image)

**Figure 3.3: MAD# database structure.**

**Drivers**

The FMDs are simple model wrappers that can be developed by 3rd parties to enable use of specific models with MAD#. The FMDs also expose the list of inversion target variables, domain and temporal types. MAD# supports, by default, two forward models: PMWIN - MODFLOW 96 [Chiang and Kinzelbach, 2001], and HYDRUS 1D [Simunek et al., 1998].

A second type of MAD# driver is intended to support external programs capable of generating random spatial data fields (e.g. using geostatistics). RFGDs are used to define structural parameters (e.g. for a geostatistical model); establish random and deterministic structural parameters; and generate conditional fields. The RFGDs implemented in MAD# are: GSTAT - Based on GSTAT R package and stand-alone [Pebesma and Wesseling, 1998], R_Base_Package - Based on Mvtnorm [Genz and Bretz, 2009], Msm [Jackson, 2011], and Tmvtnorm [Wilhelm and G, 2015] R packages.
Although the likelihood calculation is not a driver, in future releases, the calculation of the likelihood will be managed as a driver. This approach will allow to couple different methods to calculate the likelihood can be added in the framework. In the current version, the likelihood is calculated using a nonparametric kernel method through the R statistical package \texttt{np} [Hayfield and Racine, 2008a].

**Implementation of MAD stages**

The strategy stage is addressed in the pre-processing module (Figure 3.4). This is an input module where the user selects a FMD and RFGD. The selected FMD obtains information about the geographical domain of the FM and the temporal nature of project (steady-state or transient). Geographic domain information is used to reproduce the same domain of the FM. A list of available variables classified in Type-A and Type-B are through FMD. These variables allow users to identify the target variable and type of data available. The target variable is considered a SRF, and defined by SRF parameters through RFGD. The selected SRF parameters are managed as inversion parameters. The module contains tools for introducing the location of observations, and anchors which are also inversion parameters. These parameters should be associated to prior distributions, which can be generated by the MAD# framework or imported by the user.

![Figure 3.4: MAD# pre-processing module flow chart.](image-url)
The implementation stage is handled by the processing module (Figure 3.5) which is the core of the MAD# framework. The user defines the number of samples and realizations per samples to be executed. Using each sample, MAD# requests from the RFGD a number of realizations (defined by the user). The RFGD returns the realizations of the sample. MAD# processes the realizations of each sample in the FM using the FMD. The FMD extracts the simulated data at the location of each Type-B data measurement. Finally, the processing module generates a result file, which is a database file with all parameters used in the process, and output files with the information extracted from multiple executions of the FM.

![Figure 3.5: MAD# processing module flow chart.](image)

The post-processing module complain the assessment stage (Figure 3.6). This module begins by assembling the simulated Type-B of the processing module by sample. The MAD# user can define different subsets of the Type-B data vector to evaluate the likelihood. The simulated Type-B and the Type-B subset are compared to calculate the likelihood per sample.

The convergence of the likelihood of each sample is evaluated using a graphical tool, which uses the likelihood of a sample with different amount of realizations. The evaluation of convergence of number of samples is done comparing the likelihood of all samples with different number of realizations. When the number of realizations or the number of samples is insufficient, it is necessary to add more realizations or samples. This process is executed again until the likelihood converges adequately. Using the likelihood of all samples is used to generate a posterior of the structural parameters and anchors.
Example test cases

A base synthetic case with lateral confined groundwater flow through a heterogeneous aquifer was used for creating four test cases. These heterogeneous aquifers are of interest because in reality aquifers are heterogeneous and this heterogeneity can affect the travel time of contaminants to as sensitive areas like drinking water wells. A synthetic $\ln(\text{Transmissivity})$ field was created using GSTAT [Pebesma and Wessel09 1998] with isotropic exponential covariance and no trend, which is common on synthetic projects (Li et. al, 2005). The SRF uses the following structural parameters: mean = -2, variance = 0.15, range (length scale) = 28 meters, and nugget = zero meters. The field is 400x400 meters discretized into a 40x40 uniform rectangular grid. From this baseline field, the Type-A data were collected and the anchor values are known. Also, the specific storage was constant of 0.001. The aquifer was evaluated as steady and transient state. Three periods of 3, 10, and 15 days were defined in the transient state. A well was place at center of the domain, pumping only in the second period. The boundary conditions in all test cases were: constant heads 105m and 100m at south and north respectively, and no lateral flow, generating a hydraulic gradient of 1.25%. 

Figure 3.6: MAD# post-processing module flow chart.
Case studies: strategy

The general strategy for this example is as outlined above and starts with choosing the target variables in each test case (Table 3.1). The structural parameters are not assumed and will be random variables in the inversion process. The prior distributions of all structural parameter were chosen to be uniform distributions in order to be conservative. The Type-A measurements in all test cases were taken from the three locations that form a triangle placed at the center of the domain (Figure 3.7). The inversion data type for these test cases is hydraulic head. Head provides ideal Type-B data since head gradients are a function of the hydraulic conductivity field, being this field numerically equals that transmissivity field managed in the test cases, where the thickness is one. The Type-B measurement locations are along the predominant flow path (south to north) for test case I. The test case II, the Type-B measurements are along of the flow direction generated by the pumping. The test cases III and IV use the same Type-B measurement configuration.

Table 3.1: Description of test cases

| Test case | Type  | Pumping | Target variable      | θ    | Priors bounds | Anchors | |z_b|
|-----------|-------|---------|----------------------|------|---------------|---------|---|
| I         | Steady| -       | Transmissivity       | Mean | [-5, -1]      | 8       | 4 |
|           |       |         |                      | Partial Sill | [0.1, 0.7] |         |   |
|           |       |         |                      | Range | [10, 120]     |         |   |
| II        | Transient | 5m³/d | Transmissivity       | Mean | [-5, -1]      | 8       | 12|
|           |       |         |                      | Partial Sill | [0.1, 0.7] |         |   |
|           |       |         |                      | Range | [10, 120]     |         |   |
| III       | Steady| -       | Transmissivity       | Mean | [-5, -1]      | 0       | 11|
|           |       |         |                      | Partial Sill | [0.1, 0.7] |         |   |
|           |       |         |                      | Range | [10, 80]      |         |   |
| IV        | Transient | 5m³/d | Transmissivity       | Mean | [-5, -1]      | 0       | 11|
|           |       |         | Specific Storage     | Mean | [0.0005, 0.002] |         |   |

The FM used was MODFLOW-96 [Chiang and Kinzelbach, 2001] since that FM software is well-established in the groundwater community, is easily available which appeals to the MAD# community resource objective, and already has a FMD written for it. The MODFLOW-96 project used for the FM simulations in MAD# is based off of that used for creating the synthetic Type-B data in all test cases. This eliminates the possibility of model uncertainty associated with the FM, a condition which does not reflect reality, but this is an elementary example to show the basic application of MAD#. The one disparate aspect of the FM used in MAD# compared to the baseline FM is that the synthetic transmissivity field is not utilized in the former. The same domain extent and discretization along with boundary and initial conditions were used.

The placement of anchors is in a diamond configuration settled between the Type-B data locations. There are eight anchors, and the distances between them and a Type-B data location varies. In the assessment stage, the influence of the distance between the anchors
Figure 3.7: MAD# software showing the study area, Type-B and Type-A measurements, and anchors.  

a) Test case I. Steady state, 3 structural parameters of transmissivity range, partial sill and mean.  
b) Test case II. Transient scenario with 3 structural parameters of transmissivity range, sill and mean.  
c) Test case III. Steady state without anchors and 3 structural parameters of transmissivity range, sill and mean.  
d) Test case IV. Transient state without anchors, and 2 structural parameters, transmissivity mean and specific storage mean.
and Type-B data will be analyzed. Strategic placement of anchors is a subject of current research, and is addressed in Yang et al. [2012].

The goal of this example is to compare the difference between the posterior distributions to the respective prior distributions, for both over the parameter space and in relation to the true values of the synthetic baseline case. Three structural parameters and anchors will be assessed this way. The test case I and II will be compared in order to evaluate the effect of more Type-B measurements and transient state in the inversion. The test cases III and IV are linked with the objective to determine the posterior distribution of the structural parameter mean of transmissivity and specific storage. The maximum likelihood value of the structural parameters range and variance of the test case III were used in the test case IV. All of the information provided in this subsection is entered into the pre-processing module of MAD#, as discussed in previous sections.

Case studies: implementation

The implementation of the strategy is carried out in the MAD# processing module; the first step of sampling prior distributions was conducted in each test case using MAD#. To obtain samples for the structural parameters, the prior distributions were independently randomly sampled using Math.NET team [2014] library included in MAD#. The number of unique samples for each structural parameter was 100. To create samples for the anchor location, first a unique sample from the structural parameter set was used to generate conditioned fields of the Type-A data. The conditioned fields at the anchor locations were then used to obtain a vector of values that describe normal distribution. In total, with the 100 unique structural parameter samples each having 10 anchor samples, there were 1000 prior samples.

Each of the 1000 prior samples had 300 realizations generated with the GSTAT as the RFG. Each of these realizations were passed to MODFLOW-96 and the FM was executed a total of 300,000. Since the forward simulations are independent, MAD fits into the ‘embarrassingly parallel’ category of parallel algorithms. The high-throughput computing resource HTCondor was utilized to distribute the forward simulations a computer lab. The total processing time for the simulations is presented in the Table 3.2. HTCondor controls the number of instances in each simulation process. The computer lab has computers with different CPUs from Intel I7 - 8 cores, 8 GBytes of memory to Intel I5 - 4 GBytes of memory.

<table>
<thead>
<tr>
<th>Test case</th>
<th>Physical hours</th>
<th>CPU hours</th>
<th>Number of computers</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>3.83</td>
<td>230.1</td>
<td>51</td>
</tr>
<tr>
<td>II</td>
<td>6.37</td>
<td>282.9</td>
<td>46</td>
</tr>
<tr>
<td>III</td>
<td>1.67</td>
<td>88.4</td>
<td>22</td>
</tr>
<tr>
<td>IV</td>
<td>1.40</td>
<td>124.8</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 3.2: Description of test cases
After the simulations were run and MAD\# extracted the simulated Type-B results from the Type-B measurement locations, the likelihood distributions were calculated. The likelihood calculation method utilized was non-parametric kernel density estimation [Hayfield and Racine, 2008a]. The advantage of the non-parametric method is that there is no assumption on the shape of the distribution, but the disadvantage is that more realizations are needed to resolve the shape compared to parametric methods.

Case studies: assessment

Within the MAD assessment stage, if a sample has consistent likelihood over increasing numbers of realizations, and has the same relative likelihood when compared to other samples, then its likelihood is converged. Figure 3.8 shows one convergence plot for each of the four test cases. Seven random samples are chosen, and their likelihoods over a range of realizations are calculated. After 260 realizations, all but one sample out of all test cases holds a consistent likelihood value. This suggests that 300 realizations are enough to proceed to calculating posterior distributions.

By multiplying the converged likelihood distributions by the prior distributions, the posterior distributions are calculated. Figures 9-12 compare posterior distributions to their respective prior distributions and the true values. There are two ways to determine if a posterior is an improvement over a prior. Primarily, the posterior should cover a narrower range of values in the parameter space which concentrates the probability density. Second, the posterior can increase the probability of the true value, however this is only applicable in synthetic case studies.

Comparing the four test cases, some insight can be gained on the influence of Type-B data on posteriors. Test case II - which has more Type-B measurements in both space and time compared to test case I - had posteriors with narrower ranges and higher probabilities for the true values for all but one anchor (Figure 3.9). The same success can be seen for the partial sill (Figure 3.10). The mean’s posterior is narrower in test case II, but the true value’s probability is approximately the same as in the prior, while test case I had a high probability for the true value. For the range structural parameter, the posterior had a narrower range, but the true value does not fall within it. However, in test case III, which has more Type-B variables in space than test case I, the range structural parameter has a posterior with the probability density concentrated closer to the true value. Estimation of the range structural parameter shows a large uncertainty with respect the true value; similar results were found by [Firmani et al., 2006]. The maximum likelihood (ML) value of the structural parameters range and partial sill of the test case III (Figure 3.11) were used in the test case IV. The ML value of the range 14.33 m and partial sill 0.17 were considered constant values in the test case IV. The posterior pdf in test case IV (Figure 3.12), where there are transient Type-B measurements but one less than test case II, the mean has the best posterior out of all the test cases and the storage coefficient’s mean also had a successful posterior.
Discussion and conclusions

The MAD approach and the MAD# software differ significantly from other common inverse modeling methods and tools. PEST [Doherty 1994], UCODE [Poeter and Hill 1999] and ITOUGH2 [Finsterle and Zhang 2011] have as a primary purpose, deterministic parameter estimation through minimizing linear or nonlinear objective functions and determining a single value per parameter. Conversely, MAD# uses a Bayesian method for transferring information from observations to anchors and structural parameters to obtain posterior distributions for each parameter.

Inverse modeling applications generally require configuring and changing input and output files, executing a forward model, and evaluating results. While methods such as Joint Universal Parameter Identification and Evaluation of Reliability (JUPITER) [Banta et al. 2008] have been developed to reduce the complexity of parameter estimation, users are required to alter input files of the FM to fit within the specific parameter estimation framework. MAD# implements a new approach, in which users are not required to create template files. Using a plug and play approach, the user just specifies the FMD for executing the inverse
Figure 3.9: Prior and posterior pdf comparison (red line prior, blue line posterior, vertical line true value). a) Anchors test case I. b) Anchors test case II.
Figure 3.10: Prior and posterior pdf comparison of structural parameters between test case I and II. a) Structural parameters test case I. b) Structural parameters test case II.
Figure 3.11: Prior and posterior pdf of structural parameters of test case III. a) Mean b) Partial Sill c) Range

Figure 3.12: Prior and posterior pdf of structural parameters of test case IV. a) Structural parameter mean of the transmissivity. b) Structural parameter mean of the primary storage coefficient.
process. The expectation is that this approach will increase the ease of using new FMs and simplify the application of inverse modeling techniques.

The generic configuration of the MAD# framework allows accommodation of different FMs via drivers. This paper demonstrated how a FM can be linked with the MAD# framework. The generation of realizations in MAD# is also managed via a driver. The MAD# user can select the appropriate generator for each project. The generated information in the inversion process is stored in simple SQLite databases, which can be accessed by generic SQLite manager applications and integrated with the GIS environment. The inversion process required a long processing time, which suggests that MAD# team should work in mechanism to execute in High Performance Computing frameworks.

In the case studies presented here, the PMWIN-MODFLOW driver manages the MODFLOW project generated by PMWIN without changing information in the original model control files. The SRF Ln(Transmissivity) is characterized using the indirect measurements of pressure head via solving head equation implemented in MODFLOW. The anchors defined in the SRF domain show the posterior distribution of Ln(Transmissivity) and the uncertainty at the anchor locations. The posterior pdfs of structural parameters and anchors characterized the global and Ln(Transmissivity) field, respectively. The posterior pdfs obtained with more Type-B measurements produced values closer to true values. MAD# provides a user interface that allows comparison of multiple scenarios. The range structural parameter

In summary, the MAD# software framework was designed, developed, and described here to aid scientists, modelers, and students in the application of inverse modeling and SRF characterization. MAD# specifically addresses the five criteria proposed by Carrera et al. [2005] and represents a potentially valuable step forward for inverse modeling in general and the MAD method specifically.

Software availability

MAD# is made available through collaboration with the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Data Center. MAD# source code and documentation can be accessed at the MAD code repository website http://mad.codeplex.com. MAD# and its source code are released under the New Berkeley Software Distribution (BSD) License which allows for liberal reuse of the software and code.

Acknowledgements

This work was supported by the National Science Foundation under grant EAR-1011336, “The Method of Anchored Distributions (MAD): Principles and Implementation as a Community Resource.” Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We also acknowledge the extensive programming, testing, and documentation work of Daniel Gunnel, and others.
3.2 A Package for the Bayesian Inversion of Geostatistical Parameters with Multi-type and Multi-scale Data

Introduction

The field of Geostatistics originated in the 1950s with the pioneering work of [Krige 1951] and [Matheron 1962] who tried to estimate the characteristics of subsurface properties with the limited measurements typically available in this field. This scarcity, caused by the high explorations costs, is exacerbated by the strong heterogeneity that many such subsurface properties exhibit. Both these factors combined make it impossible to describe any subsurface process with certainty, therefore necessitating the application of statistical tools. Today, Geostatistics is used in many fields of Earth Science such as Geology [Hohn, 1962], Hydrogeology [Kitanidis, 2008], plus Hydrology and Soil science [Goovaerts, 1999]. To meet this demand, many software packages have been developed that provide practitioners and scientists alike with the much needed tools to apply Geostatistics. In R, the best collection of such tools is arguably found in the \texttt{gstat} package [Pebesma, 2004] developed and maintained by Pebesma and colleagues. With \texttt{gstat}, it is possible to estimate (Kriging) and simulate (Gaussian process generation) heterogeneous fields in one, two or three dimensions, therefore providing necessary tools for geostatistical analysis.

Any such statistical analysis should draw on all available data that are connected to the variable of interest to infer, i.e. to learn about, its spatial distribution as much as possible. Examples for such spatially distributed variables in Earth Sciences would be, e.g. the hydraulic conductivity of an aquifer, evapotranspiration rates of different land surface areas, and soil moisture. In classical statistics, such information may consist of measurements of the variable itself or so-called local variables. Here, local means that a point-by-point relationship between both variables exists. However, many data are non-local, which means they are connected to the variable of interest via a complicated forward model. For instance, hydraulic conductivity may be connected by a solute transport model to break-through curves of said solutes and soil moisture may be connected by a hydraulic catchment model to river discharge. To learn about the input from the output of such forward models means to invert them, hence the name inversion for such techniques.

The Method of Anchored Distributions (MAD) provides a Bayesian framework for the geostatistical inversion of spatially heterogeneous variables. MAD solves the aforementioned problem by converting non-local data into equivalent local data using the tools of Bayesian inference. The result of such a conversion is the consistent representation of all data (local and non-local) as local data only, which is then amendable to further geostatistical analysis [Rubin et al., 2010]. So far, applications of MAD have been focused on Hydrogeology [Murakami et al., 2010, Chen et al., 2012, Heße et al., 2015] as well as Soil Science [Over et al., 2015]. However, given the explanations above, MAD is in no way limited to these
fields and can be employed wherever non-local data need to be incorporated into a geostatistical framework. This generality also extends to the spatial model being inferred. While there are R packages utilizing Bayesian inference for spatial models such as spBayes [Finley et al., 2015], spTimer [Bakar and Sahu, 2015], and R-INLA [Lindgren and avard Rue, 2015], these packages have several constraints compared to anchoredDistr. First, each method assumes a Gaussian process for the spatial variability. MAD has no inherent distributional assumptions, which allows its application to a wide variety of scenarios where, for example, Gaussian fields are not justified. In addition, these packages are either geared toward large data sets (spBayes and spTimer) or applied to only local data (spBayes, spTimer, and R-INLA) while MAD focuses on addressing uncertainty due to sparse data sets by incorporating non-local data. Finally, MAD employs a non-parametric likelihood estimation, which allows for great flexibility, in particular for non-linear forward models. The presented R package anchoredDistr provides an interface to the C# implementation of MAD. It allows post-processing of calculating likelihood and posterior distributions as well as visualization of the data.

The Method of Anchored Distributions

Equation 3.2 displays the general procedure of Bayesian inference where $\theta$ represents the parameters of the variable being inferred (e.g., hydraulic conductivity) and $z$ represents the data informing the inference:

$$p(\theta|z) \propto p(\theta) p(z|\theta).$$  

(3.2)

An important element of MAD is a strict classification of all data into local $z_a$ and non-local data $z_b$, with the latter being the target of inversion. MAD employs Bayesian inference in the realm of Geostatistics by expanding the supported parameters into $\theta$ for global parameters (describing overall trend and spatial correlation) and $\vartheta$ for local parameters. Since MAD is a Bayesian scheme, these $\theta$ and $\vartheta$ both have probability distributions. As mentioned above, MAD turns non-local data into equivalent local data $\vartheta$ by inverting the forward model that connects both. The non-local data therefore become anchored in space, hence the name Method of Anchored Distributions. Equation 3.3 displays the general form of MAD:

$$p(\theta, \vartheta|z_a, z_b) \propto p(\theta) p(\vartheta|\theta, z_a) p(z_b|\theta, \vartheta, z_a).$$

(3.3)

Open-source software implementations for applying the entirety of MAD are available both with a graphical interface and a command-line interface to guide users through connecting their forward models and random field generators and to execute the ensemble of forward simulations [Osorio-Murillo et al., 2015]. This software (available at http://mad.codeplex.com) was inspired by the claim that inverse modeling will be widely applied in Hydrogeology only if user-friendly software tools are available [Carrera et al., 2005].

The package anchoredDistr described here focuses on extending the post-processing capabilities of MAD software, particularly the calculation of the likelihood distribution $p(z_b|\theta, \vartheta, z_a)$ and the posterior distribution $p(\theta, \vartheta|z_b, z_a)$ after the ensemble of forward model
simulations is already complete. The MAD# software has basic post-processing capabilities, but does not offer the degree of flexibility as R for the post-processing analysis. For example, when handling $z_b$ in the form of time series, dimension reduction techniques are necessary for calculating the likelihood values. By having the R package `anchoredDistr`, users have the support to attach whichever applicable technique for their data.

**General workflow**

In the current version of `anchoredDistr`, which only handles the post-processing of a MAD application, it is assumed that prior distributions of local and global parameters, $p(\vartheta|\theta, z_a)$ and $p(\theta)$ respectively, have already been defined and sampled and that forward model simulations based on those samples have been executed within the MAD# software. This data is stored by MAD# in databases (extensions .xresult for project metadata and .xdata for each sample). The package `anchoredDistr` primarily consists of methods for the S4 class "MADproject" that extract and analyze data from these databases, i.e. handling information regarding the samples from the prior distributions and the resulting ensemble of simulated $z_b$ data. The usage of `anchoredDistr` will generally follow the workflow below (also see Figure 5.3):

1. Read data from databases generated from MAD software into a "MADproject" object via the `readMAD` function
2. View the observations and realizations with `plotMAD`
3. Apply any necessary dimension reduction techniques to $z_b$ with `reduceData`
4. Test the convergence of the likelihood distribution with respect to the number of realizations with `testConvergence` (return to MAD software to run additional realizations if unsatisfactory)
5. Calculate the likelihood and posterior distributions with `calcLikelihood` and `calcPosterior`, respectively
6. View the posterior distribution with `plotMAD`.

To install the `anchoredDistr` package, the `devtools` package [Wickham and Chang, 2016] is required in order to download the necessary files from GitHub:

```r
library(devtools)
install_github("hsavoy/anchoredDistr")
library(anchoredDistr)
```

Other packages used by `anchoredDistr` include `RSQLite` [Wickham et al., 2014] for reading from MAD databases, `np` [Hayfield and Racine, 2008b] for estimating non-parametric
density distributions, plyr [Wickham, 2011] and dplyr [Wickham and Francois, 2016] for efficient data manipulation, and ggplot2 [Wickham, 2009] for plotting. The methods included in anchoredDistr are listed in Table 3.3 and two examples utilizing these methods are provided next.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>readMAD()</td>
<td>Reads data from databases generated by MAD software</td>
</tr>
<tr>
<td>reduceData()</td>
<td>Applies dimension reduction to $z_b$ time series</td>
</tr>
<tr>
<td>testConvergence()</td>
<td>Tests for convergence of likelihood values for increasing number of realizations</td>
</tr>
<tr>
<td>calcLikelihood()</td>
<td>Calculates the likelihood values for the samples</td>
</tr>
<tr>
<td>calcPosterior()</td>
<td>Calculates the posterior values for the samples</td>
</tr>
<tr>
<td>plotMAD()</td>
<td>Plots the observations, realizations, reduced data, and/or posteriors</td>
</tr>
</tbody>
</table>

Table 3.3: The methods for the "MADproject" S4 class provided by anchoredDistr.
Example 1: aquifer characterization with steady-state hydraulic head from multiple wells

Scenario setup

In this first example, we will use the tutorial example available from the MAD website [http://mad.codeplex.com](http://mad.codeplex.com). For the sake of easier reproducibility, a "MADproject" object for this example is included in `anchoredDistr` as accessed by `data(tutorial)`. The variable of interest is transmissivity $T$, an aquifer property that represents how much water can be transmitted horizontally through an aquifer. We will use the one-dimensional heterogeneous field of the decimal log transform of $T$ (see Figure 3.14) as our baseline field from which we can generate virtual measurements and validate our resulting posterior distributions. The field was generated as a Gaussian process by the `gstat` package in R with a mean $\mu_{\log_{10}T} = -2$ and an exponential covariance function with a variance $\sigma^2_{\log_{10}T} = 0.4$ and length scale $l_{\log_{10}T} = 3$ m. Within the scope of this example, we assume these global parameter values to be known. Furthermore, we assume that we have local data in the form of measurements of $T$ at three different locations. In addition, non-local data are available in the form of head measurements (indication of water pressure) at the same locations. The forward model used to solve the groundwater flow equation and relate $T$ to head is the software MODFLOW-96 [Harbaugh and Mcdonald, 1996], part of the open source MODFLOW series that is the industry standard for groundwater modeling. To convert the non-local data into equivalent local data of $T$, we will place four anchors at selected unmeasured locations. The number of anchors needs to be justified by the data content of the measurements such that the complexity of the model does not become disproportionate to the information available. The locations of these anchors reflect locations where there is no other local data available but there is non-local data nearby for conversion (see [Yang et al. 2012] for more discussion on anchor placement). The locations of the measurements and anchors are depicted in Figure 3.14. The prior distributions for these anchors are based on simple kriging with the local data $z_a$ for conditioning and the known Gaussian process for the covariance function:

$$p(\hat{\theta} | \theta, z_a) = \mathcal{N} \left( \mu = \hat{Z}(y_i), \sigma^2 = Var \left( Z(y_i) - \hat{Z}(y_i) \right) \right),$$

(3.4)

where $Z$ generally represents $\log_{10}T$, $y_i$ is the $y$-coordinate of the $i^{th}$ anchor, $\hat{Z}(y_i)$ is the kriging estimate at the $i^{th}$ anchor, and $Var \left( Z(y_i) - \hat{Z}(y_i) \right)$ is the kriging variance at the $i^{th}$ anchor. The goal of the example is to compare the posterior distributions of the four anchors resulting from the inversion to their prior distributions which will indicate the information gain from the inclusion of the non-local data $z_b$.

Reading and viewing data

In the first step, a "MADproject" object is created with the `new()` function. Three arguments must be provided to read the MAD databases: `madname` (the name of the MAD project, e.g.
Figure 3.14: The one-dimensional baseline field of $\log_{10} T$ used in Example 1 with locations of measurements (co-located $z_a$ and $z_b$) marked along with the anchors to be inferred.

the filename for the .xmad database), `resultname` (the name of the result from MAD, e.g. the result folder name), and `xpath` (the path to where the .xresult database and result folder are located). These three arguments ensure the MAD databases can be read by the method `readMAD`, which will read in the prior distribution samples for the global and local parameters plus the observations and forward model predictions for the $z_b$. Note that `anchoredDistr` could be used independently of the MAD software, if desired, as long as the slots filled in by `readMAD` (see Table 3.4) are provided manually. The object creation and database reading steps have been completed for this example already to provide the internal data tutorial.

<table>
<thead>
<tr>
<th>Slot</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>madname</code></td>
<td>MAD project name</td>
<td>user provided</td>
</tr>
<tr>
<td><code>resultname</code></td>
<td>MAD result name</td>
<td>user provided</td>
</tr>
<tr>
<td><code>xpath</code></td>
<td>Path to .xresult database</td>
<td>user provided</td>
</tr>
<tr>
<td><code>numLocations</code></td>
<td>Number of $z_b$ locations</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>numTimesteps</code></td>
<td>Number of time steps measured at each $z_b$ locations</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>numSamples</code></td>
<td>Number of samples drawn from prior distributions</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>numAnchors</code></td>
<td>Number of local parameters / anchors placed in field</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>numTheta</code></td>
<td>Number of random global parameters to infer</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>truevalues</code></td>
<td>True values for the parameters to infer, if known</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>observations</code></td>
<td>Observed values of the $z_b$ locations and time steps</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>realizations</code></td>
<td>Simulated values of the $z_b$ locations and time steps</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>priors</code></td>
<td>Samples from the prior distributions of each parameter</td>
<td>readMAD</td>
</tr>
<tr>
<td><code>likelihoods</code></td>
<td>Likelihood values for each sample</td>
<td>calcLikelihoods</td>
</tr>
<tr>
<td><code>posteriors</code></td>
<td>Posterior values for each sample of each parameter</td>
<td>calcPosteriors</td>
</tr>
</tbody>
</table>

Table 3.4: The slots for the "MADproject" S4 class provided by `anchoredDistr`.
The prior distributions can be viewed by calling the `plotMAD` function with the "MADproject" object and the string "priors" (see below). Figure 3.15 shows the prior distributions for the four anchors in Example 1. The distributions roughly follow a Gaussian distribution due to the baseline field being a Gaussian field and the prior distributions based on the kriging mean and variance at these four locations from the \( z_a \) data and the known spatial random function. The x-axis labels are pulled from the "MADproject" object’s `priors` slot, which contains the random parameter names as provided in the MAD software.

```
data(tutorial)
plotMAD(tutorial, "priors")
```

Figure 3.15: The relative frequency (gray bars) and estimated density (red line) of the prior distributions for the four anchor locations based on samples supplied in Example 1.

**Calculating likelihoods and posteriors**

After the information contained in the MAD databases has been read into the "MADproject" object, the likelihood and posterior distributions can be calculated by `calcLikelihood` and `calcPosterior`, respectively. The method `calcLikelihood` uses non-parametric kernel density estimation (from the package `np`) to estimate the probability density of measured inversion data from the probability density function of inversion data simulated from the realizations per sample. The method `calcPosterior` multiplies the resulting likelihood distribution across the samples and the provided prior distribution to calculate the posterior.

First, we can call the `testConvergence()` method to visually inspect if we have enough realizations for the likelihood values of samples to converge (this method calls the `calcLikelihood` internally to perform this test). Figure 3.16 depicts this qualitative convergence test for Example 1 by plotting the likelihood values of a sample with increasing number of realizations.
In order to prevent cluttering, the default number of samples to display is set to seven samples randomly selected from those available in the project. Convergence is achieved when the likelihood stabilizes with increasing realizations. For this example, it appears that the log likelihood of the samples have started to stabilize by 50 realizations, but more realizations may be warranted.

The posterior distributions for each random parameter can be seen by calling `plotMAD` with the "MADproject" object and the string "posteriors". Figure 3.17 shows the posteriors for Example 1 along with the prior distribution and the true values for each of the four anchors. The posterior distributions for Anchors 2 and 3, which were surrounded by $z_b$ measurements, show an increase in probability near the true value, indicating a successful information transfer from the non-local $z_b$ into equivalent local data.

```r
  testConvergence(tutorial)
tutorial <- calcLikelihood(tutorial)
tutorial <- calcPosterior(tutorial)
plotMAD(tutorial, "posteriors")
```

![Figure 3.16: Convergence testing for Example 1 by plotting the decimal log of likelihood of a collection of randomly selected samples with increasing number of realizations.](image)

**Example 2: aquifer characterization with one pumping drawdown curve**

**Scenario setup**

The second example depicts a different aquifer characterization scenario for a two-dimensional field where the natural log transform of hydraulic conductivity ($K$) is assumed to be an
isotropic Gaussian field with variance $\sigma_{\ln K}^2 = 1$ and length scale $l_{\ln K} = 10\text{m}$ but unknown mean $\mu_{\ln K}$ (Figure 3.18). There are no anchors placed in this example, leaving the mean as the only parameter to infer. Unlike Example 1, Example 2 is therefore a demonstration of how MAD can be employed as a regular Bayesian inversion scheme, too. The prior distribution for global parameters ideally come from previous knowledge of similar sites, e.g. the distribution of mean $\ln K$ observed at other aquifers with the same geological setting. For this example, we will compare three equally spaced samples for $\ln K$ to represent a uniform prior distribution for the mean. The data include four local data $z_a (K)$ at four different locations and one non-local data series $z_b$ (hydraulic head drawdown) at a single location (see Figure 3.18). The $z_b$ location provides 100 time steps, i.e. data points, of drawdown measurements (Figure 3.19). The forward model used to solve the groundwater flow equation and relate $K$ to drawdown is OpenGeoSys [Kolditz et al., 2012], an open source software that simulates a variety of subsurface processes. This second example uses a different forward model than the first example to showcase the MAD software’s modular design that does not assume or rely on specific forward models. The data for this example is provided within the package as data(pumping).

When the pumping dataset is initially loaded, we can view the observation of $z_b$, i.e. drawdown time series (Figure 3.19), the prior distribution of the three samples (Figure 3.20), and the interquartile range of the time series simulated by the forward model for the samples (Figure 3.21):

```r
data(pumping)
plotMAD(pumping, "observations")
plotMAD(pumping, "priors")
plotMAD(pumping, "realizations")
```
Applying dimension reduction to time series

Even though we have the time series of drawdown, we cannot use these 100 individual values to calculate the likelihood because they are correlated and the multivariate likelihood distribution would be 100-dimensional. Such dimensionality would require an unrealistic number of realizations to resolve, known as "the curse of dimensionality." To overcome this obstacle, dimension reduction is needed and the method to use depends on the type of non-local data $z_b$. For this example, we will simply use the \texttt{min} function to collect the minimum head value in the time series since the observed head reduces and converges to a stable head value with time (Figure 3.19). The \texttt{anchoredDistr} package can handle any non-parameterized function, such as \texttt{min}, or a parameterized function if initial values for each parameter are given and the \texttt{nls} function \cite{R Core Team, 2016} can perform the fitting (see the package vignette for an example). The \texttt{reduceData} function is used to perform the dimension reduction on the time series:

```r
pumping.min <- reduceData(pumping, min)
plotMAD(pumping.min, "realizations")
```

The \texttt{reduceData} function returns a "MADproject" object with a \texttt{realizations} slot with reduced dimensions. The reduced data can be viewed by calling \texttt{plotMAD} with the string
Figure 3.19: The observed time series of hydraulic head drawdown to be used as non-local data $z_b$ in Example 2.

Figure 3.20: The histogram (gray bars) and estimated density (red line) of the prior distributions for the mean $\ln K$ Example 2.
Figure 3.21: The observed time series of drawdown at the $z_b$ location along with the inter-quartile range of simulated values for each time step for the three samples. "realizations". The plot shows the distributions of each parameter for each sample. In this case, Figure 3.22 shows the minimum head value distribution for the three samples, which will be used to calculate the three likelihood samples.

Figure 3.22: The reduced $z_b$ data (minimum of drawdown curve) for Example 2. Distributions are estimated from the realizations’ reduced data per sample.

With this new "MADproject" object, calcLikelihoods and calcPosteriors can be called. In Figure 3.23 the posterior distributions are shown for the three samples along with the true value of $-10$. The posterior distribution assigns greater probability toward the true value.
Figure 3.23: The prior (red) and posterior (blue) distributions with the true value (black) for the mean \( \ln K \) locations in Example 2.

```r
pumping.min <- calcLikelihood(pumping.min)
pumping.min <- calcPosterior(pumping.min)
plotMAD(pumping.min, "posteriors")
```

**Summary**

The examples given above show how the `anchoredDistr` package allows flexible post-processing of results by virtue of the MAD software such that users can apply their own post-processing analyses, such as dimension-reduction techniques. The first example shown here is available for download as MAD databases from [http://mad.codeplex.com/releases/](http://mad.codeplex.com/releases/) or as an included dataset in the main `anchoredDistr` package as `data(tutorial)`. The second example is also included in `anchoredDistr` as `data(pumping)` and is further detailed in the package vignette. The `anchoredDistr` package is hosted on GitHub and can be installed via calling `devtools::install_github("hsavoy/anchoredDistr")` or by downloading from [http://hsavoy.github.io/anchoredDistr](http://hsavoy.github.io/anchoredDistr).
Chapter 4

Integration of Multiple Forms of Information

Given the software tools to implement the Bayesian inversion framework, we can move onto how to integrate multiple forms of information. In this Chapter, the focus is on data series, which pose the curse of dimensionality problem discussed in Section 2.2 and conceptual models of geology, which are useful in the application of multipoint statistics as described in Section 2.1.

Section 4.1 There are many forms of field measurements that can be taken from the subsurface when trying to characterize the subsurface properties, such as boreholes to measure the properties directly, or wells for providing indirect information concerning the water pressure and chemical concentrations in the subsurface. All of these cases can be in the form of data series that are correlated in space or time. This section focuses on how time series can be effectively incorporated in the Bayesian framework for inferring geostatistical parameters. The synthetic case examined uses the change in drawdown at observation wells with increasing distance from a pumping well.

Section 4.2 The use of multipoint statistics for random field generation has great potential for subsurface applications. The geostatistical approach uses training images, as opposed to explicit statistical parameters and models, to describe spatial patterns. When working in a multidisciplinary team characterizing a field site where practitioners may not be conversant in geostatistics, conceptual models, e.g. what geologists image the spatial pattern to look like given the site location, are still informative. However, there may be competing alternative conceptual models. A Bayesian framework is used here to infer the appropriate use of training images given sparse direct and indirect measurements of a field.
CHAPTER 4. INTEGRATION OF MULTIPLE FORMS OF INFORMATION

4.1 Inversion Data: Handling the Curse of Dimensionality

Introduction

An accurate description of the subsurface is vital for a wide range of applications such as the modeling of contaminant transport, water flow, and the reaction of a groundwater reservoir to management policies. Characterizing the hydraulic properties of the subsurface is very challenging due to their pronounced spatial heterogeneity. Representing this heterogeneity by virtue of modern computer systems would theoretically be possible, but is hampered by the sparse nature of direct measurements typically available. Here, ‘direct’ refers to measurements of the variable itself (e.g., the hydraulic conductivity) or covariates for which simple relationships can be provided. Such a situation, where the number of unknowns is potentially much larger than the number of data, is known as an ill-posed problem. To alleviate this situation, practitioners need to use all data available to inform the characterization as much as possible. For this reason, time series, e.g., from pumping test and tracer experiments, have attracted much attention since they provide a large amount of data for a comparably modest effort. However, while the amount of raw data might be large, the actual information content can be very low, since data points from drawdown or break-through curves are highly correlated. In addition, such data are indirect, which means they are connected to the variable of interest via a complex physical process that needs to be inverted before becoming amenable for site characterization.

One of the most prominent responses to the problem of ill-posedness was the development of the field of Stochastic Hydrogeology [Gelhar, 1993, Rubin et al., 2003]. Here, the highly-variable subsurface quantities are modeled as spatially-distributed stochastic processes, typically known as a spatial random field or spatial random function (SRF). Conceptualizing the subsurface as SRFs solves the above problem, since such models are fully parametrized by only a handful of parameters describing the global behavior of the subsurface. For example, the Gaussian process, which is arguably the most prominent of these models, is fully parametrized by specifying its mean, covariance model, variance and integral scale. As a result, the problem becomes well posed and the parameters can be estimated by using direct or indirect measurements, with the latter being done by fitting analytical solutions to the observations [Rubin and Dagan, 1987, Sánchez-Vila et al., 1999, Firmani et al., 2006, Riva et al., 2009, Pechstein et al., 2016, Zech et al., 2016]. Many more models of subsurface heterogeneity have been proposed over the years to account for some inherent shortcomings of the Gaussian process model [Gómez-Hernández and Wen, 1998, Zinn and Harvey, 2003, Linde et al., 2015]. However, the overall idea remains the same, i.e., a few global parameters are used to describe the complex subsurface [Berkowitz et al., 2002, Newman and Tartakovsky, 2009]. A second approach to account for the ill-posedness inherent in subsurface characterization is the method of regularization [Tikhonov, 1963a,b]. Here, additional constraints are imposed on the spatial variability of the subsurface that typically
CHAPTER 4. INTEGRATION OF MULTIPLE FORMS OF INFORMATION

reflect some physically-based reasoning.

However, these methods are not free from problems. For example, regularized inversion may be theoretically well posed, but inverting, e.g., the conductivity values even for moderate numerical grid sizes becomes quickly impracticable. Likewise, stochastic models are very parsimonious due to describing the global behavior only, but often local data are also available that should not be discarded. To address the former problem, the Pilot-Point method (PPM) has been developed, where only a limited number of locations (i.e., the eponymous pilot points) are used for the inversion [RamaRao et al., 1995, Doherty, 2003, Alcolea et al., 2006]. However, PPM has been criticized since it cannot provide a measure for the uncertainty of the inference [Rubin et al., 2010]. In addition, the pilot points are a tool to guide the optimization but it does not provide a method for estimating the amount of information in the data. Complexity should be a function of the data [Schoeniger et al., 2015].

The Method of Anchored Distributions (MAD), first presented in Rubin et al. [2010], is a fully probabilistic framework for subsurface site characterization and can address these problems. First, similar to PPM, MAD is able to represent both global and local information on the subsurface behavior. Unlike PPM, MAD is a probabilistic framework, i.e., all quantities are represented by virtue of a multivariate probability distribution. As a result, MAD does not suffer from problems such as over-paramterization, equifinality, or ill-posedness. MAD has been successfully applied to a number of applications (see, e.g., Murakami et al., 2010, Chen et al., 2012). However, so far the assimilation of time series in MAD has been hampered by natural limits of the likelihood estimation, which is a crucial part of any Bayesian inference [Scott, 1992]. For example, Chen et al. [2012] used only the temporal moments of the series and Heße et al. [2015] used only a fixed number of data points to reduce the dimensions of the estimation procedure. If data series such as concentration breakthrough curves are to be used within MAD, the so-called ‘curse of dimensionality’ needs to be addressed and incorporated into the generic workflow of the Bayesian framework. The term curse of dimensionality goes back to Bellman [1961] who demonstrated how the computational effort in combinatorial optimization grows exponentially with the dimensionality of the problems.

For the question of density estimation, the curse of dimensionality refers to the observation that bins of a histogram in higher dimensions have rapidly diminishing number of samples from which to estimate the density. To keep the number of samples per bin the same, the total sample size has to grow rapidly as well.

To address this problem, the high-dimensional sample space is typically projected into lower-dimensional subspaces, while keeping the information content of the data as intact as possible. Several reasons may facilitate such an information-conserving projection. For instance, different data points of a time series are often highly correlated. This means their combined information content may be much lower than the sheer number of data points (i.e., their dimensionality) suggests. In such cases, information-conserving projection into a lower-dimensional subspace can be achieved by virtue of fitting analytical curves with only a few number of degrees of freedom. The better the fit the more information is preserved during the projection. Other data series, like depths profiles, may not show such a high correlation between different data points. However, dimension-reduction methods like
principal component analysis may still be able to glean the bigger part of the information from the data with only a few degrees of freedom. In this study, we will use a number of different high-dimensional data series like break-through curves, drawdown series and, depths profiles of head along with a series of dimension-reduction methods. The measure of success will be the amount of information that the dimension-reduced data sets still contain. This will be measured by virtue of their impact on the inference, i.e., the reduction of uncertainty. To measure this reduction, we use the Kullback-Leibler divergence. The results will provide an insight to the relative information gain from the alternative derivative forms of the data series for the geostatistical parameters.

This Section aims to explore the use of data series in the Bayesian inversion of geostatistical parameters. First, there is an overview of the theory of MAD and how the curse of dimensionality comes into play regarding the estimation of the likelihood distribution. Additionally, the criteria by which dimension reduction techniques are deemed successful are also discussed. Next, a synthetic test case in which dimension reduction successfully integrates data series into the MAD framework is addressed.

Methods

Before the discussion on dimension reduction, we first detail the theory behind MAD. With the formulation of the likelihood distribution under review, the effect of data series on the dimensionality of the likelihood distribution becomes apparent. To choose applicable dimension reduction techniques for a specific application, we consider a set of criteria that the methods must meet.

Method of Anchored Distributions

The fundamentals of MAD are summarized here while a thorough derivation can be found in Rubin et al. [2010]. MAD evolves from Bayesian inference which states that the probability density function (PDF) of parameters $\theta$ given data $z$, can be found from the PDF of $\theta$, the PDF of $z$ given $\theta$, and the PDF of $z$:

$$p(\theta|z) = \frac{p(\theta)p(z|\theta)}{p(z)} \quad (4.1)$$

However, in practice, the denominator consisting of $p(z)$ is often not estimated due to the entire right-hand side of Equation (4.1) having to be a proper PDF with respect to $\theta$ and $p(z)$ being a constant with respect to $\theta$. By estimating and multiplying $p(\theta)$ and $p(z|\theta)$ and then re-scaling the product to integrate to one, the left-hand side can be obtained. Hereafter, our Bayesian framework will be constructed from the resulting proportion between $p(\theta|z)$, known as the ‘posterior’ distribution, and $p(\theta)$ and $p(z|\theta)$, known as the ‘prior’ and ‘likelihood’ distributions, respectively:

$$p(\theta|z) \propto p(\theta)p(z|\theta) \quad (4.2)$$
MAD is an application of Bayesian inference where $\theta$ represents both global and local geostatistical parameters. For global parameters, e.g. the mean of a SRF, we will continue to use the variable $\theta$, where bold font indicates hereafter a vector of parameters or data points. For local parameters, e.g. the value of the field at an area of interest, we will use $\vartheta$. Similarly, the data that can be utilized to infer global and local geostatistical parameters can also consist of multiple scales. We denote data locally and directly related to the target variable as $z_a$, while data that exist on a larger scale and are indirectly related to the target variable will be denoted as $z_b$. Introducing these new variables into Equation (4.2) and manipulating terms for future ease yields:

$$p(\theta, \vartheta | z_a, z_b) \propto p(\theta) p(\vartheta | \theta, z_a) p(z_b | \theta, \vartheta, z_a)$$ (4.3)

In general, Equation (4.3) represents the inference of global and local geostatistical parameters from multi-scale and multi-type data by integrating the prior information of global parameters, the subsequent information of local parameters based on global parameters and direct data alone, and the likelihood of observing the indirect data given the parameters and direct data. The estimation of the likelihood term $p(z_b | \theta, \vartheta, z_a)$ is both the most computationally burdensome component of applying MAD and the source of dimensionality concerns.

Dimensionality and likelihood estimation

The likelihood distribution is a PDF with respect to $z_b$ such that its dimensionality is related to the dimension of $z_b$, i.e. the number of inversion data points. The PDF of $z_b$ can be simplified in terms of dimensionality when the values of inversion data are not dependent on each other. For example, Murakami et al. [2010] investigated the spatial distribution of hydraulic conductivity and inferred the Gaussian SRF parameters with measurements of hydraulic head as inversion data. The hydraulic conductivity was assumed to have a variance low enough that the hydraulic head field could also be described as a Gaussian SRF. The authors argued that since the borehole locations where hydraulic head measurements were taken were spaced at distances greater than multiple integral scales of the head field, that the measured head values were not dependent on each other. Therefore, the PDF of $z_b$ is equivalent to the products of PDFs of each $z_b$. In implementation, this translates to estimating a collection of univariate PDFs independently and then taking their product. This independence avoids the dimensionality issue.

When using inversion data such as a time series of hydraulic head at a single well, the dependence cannot be so easily avoided and the $n$-dimensional PDF for the $n$ inversion data points needs to be estimated. However, as $n$ increases, the number of samples needed to estimate the PDF increases non-linearly. For example, if $s$ samples are roughly needed to estimate a univariate p.d.f, then $s^n$ samples would be needed to estimate a $n$-dimensional PDF. However, the process of generating samples within the MAD framework occurs in a double-nested Monte Carlo loop. The outer loop contains samples $s$ from $p(\theta)$ and $p(\vartheta | \theta, z_a)$ while the inner loop consists of an ensemble of realizations $r$ adhering to the SRF of each
sample. The likelihood estimation occurs for each sample in MAD and is constructed from the ensemble of realizations for the sample along with the inversion data. This double loop creates the need for on the order of $sr^n$ estimated $z_b$ from executions of the forward model that simulates the relationship between $z_a$ and $z_b$. By limiting the number of $z_b$, we prevent the necessity of processing an insurmountable number of numerical models.

**Dimension reduction methods and criteria of success**

Dimension reduction methods in the context of MAD mean techniques to replace a data series with a smaller number of representative values such that the likelihood calculation occurs over fewer dimensions than if the data series was used directly. Possible candidates for dimension reduction methods include picking subsets of uncorrelated data points from a data series, extracting statistical moments or other statistics of the data series such as the maximum value, or fitting a function or model to the data series. Of course, describing a data series by any of these lower-dimensional representations has the disadvantage of omitting unique characteristics or possibly introducing numerical artifacts into the inference. Therefore, we have compiled a collection of criteria by which we assess dimension reduction methods.

First, the observed data series should be adequately reconstructed and the ensembles of realizations should be equally reconstructed. To define “adequate” is subjective and can be quantified by metrics such as root mean square error, but more important is the notion that the data series generated within the ensemble of realizations is also able to be characterized by the dimension reduction method. This concern is more geared toward the case of conceptual model error or other cases where the model output could be significantly different in shape than the observation. In our synthetic investigation, we minimize this concern but still confirm the similarity in shape between our observations and realizations.

Second, the representative values generated by the dimension reduction should be varied enough across the realizations to inform the inference. For example, if the observation and realization time series all begin at a value of zero, then including this value is not informative because the likelihood value can not be calculated. After a dimension reduction method has been applied to an ensemble of realizations, histograms of the representative values can be assessed for how informative each value is.

Third, convergence of the likelihood values for each of the samples needs to be reached with increasing number of realizations. This convergence depends on the total number of dimensions used in calculating the likelihood with more realizations required to reach convergence for more dimensions. This criteria can be compared against the computational expense of each forward model and the computational resources available.

Lastly, the resulting posterior distribution should suggest the underlying true values for the variables being inferred and the posterior distribution should contain more information than the prior distribution. Being able to compare the posterior to true values is only possible in synthetic studies such as this one, but information gain can be quantified in any study by the use of Kullback-Leibler divergence, which is an asymmetric distance metric to compare two probability distributions. The asymmetry is useful here in a Bayesian framework to
directly quantify the information content gained by updating from a prior distribution to a posterior distribution:

$$D_{KL}(p(\theta|z) \parallel p(\theta)) = \int_{-\infty}^{\infty} p(\theta = t|z) \log_2 \frac{p(\theta = t|z)}{p(\theta = t)} dt$$  \hspace{1cm} (4.4)$$

where $D_{KL}(p(\theta|z) \parallel p(\theta))$ is in units of bits.

Scenario: water table drawdown from pumping

The first scenario considered is a site in which there is a pumping well that is causing a cone of depression in spatially variable soil. The soil is characterized by its hydraulic conductivity ($K$) and there are four direct measurements of $K$. The field is discretized into a 100-by-50m grid with one pumping well. The groundwater flow process is simulated with the open-source software OpenGeoSys [Kolditz et al., 2012] with 5000 quadrilateral elements and 5151 nodes comprising the finite element mesh. The hydraulic head is initially 0 m across the grid with all boundary conditions of 0 m head. The pumping well starts pumping at the beginning of the simulation such that drawdown is equivalent to the hydraulic head. The time discretization of the simulation is 100 time steps. The simulation outputs the time series of drawdown at three observation wells. See Figure 4.1 for a visual of the site configuration and Figure 4.2 for the drawdown curves from the three observation wells used as indirect data.

The synthetic spatially variable $K$ field is represented by an isotropic Gaussian SRF of natural log transformed $K$ with mean $\mu_{lnK} = -10$, variance $\sigma_{lnK}^2 = 1$, and correlation length scale $l_{lnK} = 10$. A baseline field is created for validation purposes and exhibits the following parameters: $\mu_{lnK} = -10.09$, $\sigma_{lnK}^2 = 1.07$, and $l_{lnK} = 9.51$ (see Figure 4.3). For this example, $\sigma_{lnK}^2$ and $l_{lnK}$ are assumed known and the prior distribution for $\mu_{lnK}$ is

$$\mu_{lnK} \sim U(-11, -9)$$  \hspace{1cm} (4.5)$$

and can be seen in reference to the baseline values in Figure 4.4. No anchors, i.e. probability distributions representing local values of hydraulic conductivity, were used.

The prior distribution was sampled 50 times in latin-hypercube fashion, resulting in 50 equally spaced values in between the prior distribution bounds due to the uniform nature of the prior distributions. From each of these samples, 2000 realizations of random fields, conditioned on the four direct measurements, were generated and treated as input in the forward model executed by OpenGeoSys.

After the simulations were complete, the 2000 sets of simulated drawdown curves from the three indirect measurement locations from each of the 50 samples can be compared to the three observed drawdown curves to calculate likelihood values for each sample. Two directions are taken regarding how the simulated drawdown curves were compared to the observations: combinations of individual time steps or fitting a model. Using individual time steps serves as a naive baseline since it is the most straightforward method, but neglects the majority of the information available in the data series. We consider sets of one, two, and
Figure 4.1: The configuration of the site for the pumping drawdown example with four direct measurement locations, three indirect measurement locations, and one pumping well.

three times steps to compare the inference to a model with three parameters. The model is a modified Matérn function:

\[
D_\kappa(t) = \sigma^2 - \sigma^2 \frac{2^{1-\kappa}}{\Gamma(\kappa)} \left(\sqrt{2\kappa \frac{t}{\lambda}}\right)^\kappa K_\kappa\left(\sqrt{2\kappa \frac{t}{\lambda}}\right)
\] (4.6)

where \(t\) is time, \(\kappa\) is the shape parameter, \(D_\kappa(t)\) is the drawdown as a function of time, \(\sigma^2\) is the eventual steady-state drawdown that \(D_\kappa(t)\) converges to with time, \(\Gamma\) is the gamma function, \(K_\kappa\) is the modified Bessel function of the second kind, and \(\lambda\) is the correlation time scale. The Matérn function is primarily used in geostatistics as a covariance model that allows for flexibility of near-origin behavior, e.g. it can simplify to a Gaussian covariance model or an exponential covariance model depending on the value of the shape parameter \(\kappa\). Here, we use \(D_\kappa(t)\) to describe drawdown curves whose near-origin behavior also varies. The main difference with the modified Matérn function in Eq. 4.6 compared to the traditional
Figure 4.2: The drawdown curves from the three observation wells used as indirect data.

Figure 4.3: The baseline lnK field of the site for the pumping drawdown example.
use as a geostatistical covariance function is that the former is subtracted from $\sigma^2$ because $D_\kappa(t)$ is negative and that the physical dimension is time instead of space. The flexibility in shape at the near-origin behavior allows for Eq. 4.6 to fit to the drawdown curves from the observations wells at increasing distance and thus delayed drawdown. See Figure 4.5 to see how well the Matérn function can be fitted to the three observed drawdown time series.

For each of the simulated drawdown curves, the three Matérn function parameters, $\sigma^2$, $\kappa$, and $\lambda$, are fitted by non-linear least squares. These three parameters are used as substitutes for the 100 individual time steps in the inversion, which reduced the dimensionality of the likelihood calculation from 100 dimensions down to just three, in line with general recommendations of only using up to five dimensions. To compare the relative information content of the three parameters and to investigate potential further reductions in dimensionality, we tried a variety of inversion data configurations. Table 4.1 lists the seven configuration used. The configurations vary between one and three dimensions as one to three parameters are used as inversion data.

When only evaluating one individual times step, all 100 time steps were used independently. For two and three time steps, we used 24 random combinations of each. This led
Figure 4.5: Modified Matérn fitted to the three observed drawdown curves.

Table 4.1: Inversion data configurations of Matérn function parameters. All the parameters were fitted with non-linear least squares, but the configurations differ in which fitted parameters were used in the inversion. Those marked with $z_b$ were used as inversion data and those withheld are marked by ‘-’.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>$\sigma^2$</th>
<th>$\lambda$</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$z_b$</td>
<td>$z_b$</td>
<td>$z_b$</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>$z_b$</td>
<td>$z_b$</td>
</tr>
<tr>
<td>3</td>
<td>$z_b$</td>
<td>-</td>
<td>$z_b$</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>$z_b$</td>
</tr>
<tr>
<td>5</td>
<td>$z_b$</td>
<td>$z_b$</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>$z_b$</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>$z_b$</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
to 148 posterior distributions for each observation well for each parameter. To show the posterior results in a succinct manner, we recorded the posterior distribution assigned to the true value along with the Kullback-Leibler Divergence (KLD). Together, these two values describe how well the time steps inferred the true value and how much the posterior is different from the prior. The ideal result is a high posterior density paired with a higher KLD. Figures 4.6 shows the distribution of posterior densities estimated at the true values for mean by how many times steps were used in the inference and which observation well the data came from. For $\mu_{\ln K}$, increasing the number of time steps improved the posterior of the true value for all three observation wells, with the middle well having the highest posterior density for any number of time steps. The KLD also increased with increased number of time steps, so the inference generally improves with more time steps in general.

Gaussian noise ($\epsilon \sim N(0, 1)$) was added to the drawdown curve observations to highlight the disadvantages of using individual time steps. Figure 4.7 shows that the one, two, and three times steps either maintained the baseline value of $\mu_{\ln K}$ from prior to posterior, or reduced the probability density from prior to posterior for most combinations of time steps. The KLDs significantly increased, indicating that the posteriors from noisy data diminishes the ability of inferring the baseline $\mu_{\ln K}$ from individual time steps.

From the seven configurations of Matérn parameters, the relative success of inferring $\mu_{\ln K}$ from the individual parameters can be gleaned. Configuration 1 uses all three parameters and increases the density at the true value from the prior to the posterior. The shape of the posteriors in Configuration 1 are similar to that of Configurations 2, 5, and 6, which are the other four configurations that use $\lambda$ as inversion data. Configuration 6 only uses $\lambda$ for inversion data, but it does not increase the probability density of the baseline values of $\mu_{\ln K}$ for the two farther observation wells. The inclusion of using $\sigma^2$ in the inversion with Configuration 5 does improve the inference from all three observation wells. The inclusion of $\kappa$ as inversion data does not seem to improve the inference, as seen in the flat nature of Configuration 4 that solely used $\kappa$. From these seven configurations, it suggests that the dimensionality of inversion can successfully be decreased from 100 down to two.

One benefit of fitting an analytical function as opposed to using measured values at specific time steps is the reduced susceptibility to noisy data (Figure 4.9) compared to individual time steps. The posterior distribution were re-calculated (Figure 4.10). There are no significant changes in the inference with the addition of this noise, so the dimensionality decrease from 100 to two holds for even noisy data, unlike with individual time steps.

In summary, the inference of geostatistical parameters from pumping drawdown curves performs better with Matérn parameters than individual time steps. The improvement is not that noticeable without noise, but remarkable with noise. Since real data is often noisy, the Matérn parameters are deemed the better method for reducing the dimensionality of the drawdown curves in this scenario. The Matérn model has three parameters, but similar results were achieved when the shape parameter was not used in the inference compared to when it was. In the end, we were able to reduce the dimensionality from 100, i.e. the number of time steps that drawdown was simulated over, down to two, i.e. the $\sigma^2$ and $\lambda$ parameters of the Matérn model.
Figure 4.6: The distribution of posterior densities at the baseline value of the mean (top) and the Kullback-Leibler Divergence values (bottom) for one, two, and three time steps used in the inversion of $\mu_{lnK}$. The horizontal line represents the prior density.
Figure 4.7: The distribution of posterior densities at the baseline value of the mean (top) and the Kullback-Leibler Divergence values (bottom) for one, two, and three time steps used in the inversion of $\mu_{lnK}$ with noisy data. The horizontal line represents the prior density.
Figure 4.8: Posterior distributions for $\mu_{\ln K}$ from Configurations 1-7 at measurement locations 1 (closest), 2 (middle), and 3 (farthest). Vertical lines marking the baseline parameter values as requested (solid) and estimated (dashed).
Conclusion

An example of indirect data series - pumping drawdown time series - is investigated regarding how the data series can be effectively and efficiently integrated in a Bayesian framework for inferring geostatistical parameters. A modified Matérn function was used to model the drawdown curves at three observation wells, which were at three different distances from the pumping well. The flexible near-origin behavior of the Matérn function compensates for the delayed drawdown at farther observation locations. By using the three Matérn parameters to represent the drawdown curve from 100 time steps, the dimensionality of the inversion reduces from 100 down to three. When compared to using the measured drawdown from three time steps, which has equivalent dimensionality as the three Matérn parameters, the posterior distributions for the Matérn parameters perform better.

The examples shown are for non-periodic time series. The work can be expanded to periodic time series, such as the diurnal variation of water temperature in the vadose zone, by using the coefficients in Fourier transforms as inversion data. In general, fitting non-linear analytical expressions requires user input, such as initial values of parameters. To apply our methodology to data series from applications other than the ones provided here as examples,
CHAPTER 4. INTEGRATION OF MULTIPLE FORMS OF INFORMATION

Figure 4.10: Posterior distributions for $\mu_{\ln K}$ from Configurations 1-7 at measurement locations 1 (closest), 2 (middle), and 3 (farthest) with noisy data. Vertical lines mark the baseline parameter values as requested (solid) and estimated (dashed).
application-specific knowledge is required and similar synthetic cases as conducted here can provide insight on how to effectively reduce the dimensions of the inference.

4.2 Prior Knowledge: Conceptual Models of Geological Heterogeneity

Up until this Section, the prior distributions in the Bayesian framework have been describing the geostatistical parameters of Gaussian fields. Here, we move to multipoint statistics and its use of training images instead of quantitative parameters for describing the spatial patterns of the subsurface.

Introduction

Heterogeneity exists in nature and is often difficult to measure, so great uncertainty stems from sparse measurements. Conceptual models of geological heterogeneity are graphical representations of how we expect the geological units to be spatially distributed. Both geostatistical approaches described in Section 2.1 (two-point and multipoint) aim to describe global spatial patterns. The multipoint approach relies on cataloging patterns of values in more than two locations while the variogram-based approach only uses two locations. With the inclusion of more locations in the spatial variability pattern description, multipoint methods can describe more complex spatial patterns than variogram-based methods [Guardiano and Srivastava, 1993]. To model complex patterns with multipoint statistics, a training image depicting that complex pattern needs to be provided. On the other hand, variogram-based methods only need a mean and variogram to describe patterns. Although training images contain more information than variograms and thus may be seen as more work intensive to create, the data is visual and can be collected from previous studies and a variety of information sources. Variograms, although described with a handful of parameters only, are more abstract descriptions of spatial patterns.

Geostatistics can aid in modeling subsurface heterogeneity, and much work has been done for two-point methods. For geological heterogeneity, i.e. the spatial distribution of relatively homogeneous geological units, variograms have been applied over a variety of scales (e.g. Chen and Rubin [2003], White et al. [2003], D’Or et al. [2009]). However, even with heavy conditioning, i.e. the incorporation of extensive direct data, variogram-based methods are not capable of simulating complex patterns such as connected channels [Klise et al. 2009]. Multipoint methods are advantageous for complex and repetitive patterns which training images can describe, but there has been less work in uncertainty characterization compared to variogram-based methods. Examples of where multipoint methods have been used include filling in missing satellite data [Mariethoz et al. 2012], describing and sharpening geological delineations in geophysical tomography [Lochbühler et al. 2014], and extrapolating gravel pit cross-sections into three-dimensional space [Comunian et al. 2011].
The Method of Anchored Distributions (MAD) is a general Bayesian inverse modeling framework designed for geostatistics [Rubin et al., 2010]. It has been applied in cases using variogram-based methods [Murakami et al., 2010], transformed variogram-based methods [Heße et al., 2015], and homogeneous layering methods [Over et al., 2015] so far. This Section is a proof-of-concept for using MAD to quantify the uncertainty in the training images in MPDS for multipoint geostatistical simulation of geological heterogeneity, as opposed to the uncertainty in quantitative parameters of Gaussian fields.

Statistical concepts

The conceptual differences between multipoint geostatistics and variogram-based geostatistics as it pertains to how the statistical model is parameterized are described here. For a more detailed mathematical description of the statistical theories and algorithms for both approaches, see Section 2.1. I will focus on the multipoint statistics Direct Sampling (MPDS) method as it is the most efficient algorithm in terms of computation time and memory costs compared to the older alternative multipoint statistics algorithms [Mariethoz, 2010]. I will also assume the SRF in question is a binary categorical variable for this discussion, although both approaches also support continuous variables. The variogram-based algorithm for simulating fields of binary categorical variables is called Sequential Indicator Simulation [Journel, 1983]. Both approaches can generate random fields of a binary categorical variable, but the algorithms differ in how the cumulative distribution function for a certain grid cell is calculated and sampled.

The Direct Sampling method

MPS is first introduced by [Guardiano and Srivastava, 1993], and expanded by Strebelle, 2002, but it was not until the Direct Sampling algorithm in [Mariethoz et al., 2010] that MPS was computationally efficient enough for practical applications. The Direct Sampling algorithm randomly samples the training images directly, as opposed to sampling a CCDF constructed in a pre-simulation stage. It is argued that randomly sampling the training image to find particular configurations of conditional points is equivalent to sampling the CCDF constructed by cataloging all available configurations. Direct sampling is also more efficient in terms of time and memory since the cataloging does not have to be executed or stored. Since multiple points are considered when searching conditional configurations, more complex patterns can be simulated, but there is a computational cost compared to variogram-based methods. If the spatial pattern does not exhibit complex structures, then it may be advantageous to use variogram-based methods instead of MPDS.

The Method of Anchored Distributions

The basics of MAD are discussed in Section 2.2 and originally in [Rubin et al., 2010], but this subsection is dedicated to how MAD’s generality makes it ideal for quantifying uncertainty
in training images. Since MAD is not restricted to any specific geostatistical models, distribution types, or physical applications, it is compatible with the flexible MPDS algorithm. Figure 4.11 depicts the process of inferring training images in the MAD framework. First, the prior distribution for training images is set as a probability distribution with the probability mass assigned to each training image proportional to the relative certainty regarding the training images. Then, each training image is used to create an ensemble of realizations conditioned on any available direct data and/or anchors. Those fields are inputed into the relevant forward model to yield outputs associated with the available indirect data. The comparison of the simulated indirect data to the observed indirect data yields the likelihood value of each training image. The prior distribution is then weighted by these likelihood values to yield the posterior distribution of the training images. Not explicitly shown here is the prior distributions of anchors. To estimate $p(\vartheta|\theta, z_a)$, where $\theta$ represents the training image here, one would record the configuration of conditional data $z_a$ surrounding the anchor(s), then scan the entire training image for the same configuration of values, and record all instances of values corresponding to the anchor location(s) in the original configuration. The distribution of recorded corresponding values would be the distribution for the anchor(s).

Figure 4.11: The process of inferring training images in MAD.

This inversion process is implemented by a driver program within the MAD# software [Osorio-Murillo et al., 2015] (see Section 3.1) and the DeeSse software developed at the University of Neuchâtel that implements the MPDS algorithm. Once a posterior distribution is estimated, the uncertainty in the training images can be propagated into subsequent predictions. The posterior mass assigned to each training image is the proportion of total simulations that will receive conditional realizations from each of the training images.
CHAPTER 4. INTEGRATION OF MULTIPLE FORMS OF INFORMATION

Synthetic case study

Scenario setup

The synthetic case study is based on a two-dimensional baseline field with an indicator variable that can take two values. These two values represent two soil types, signified by their hydraulic conductivity $K$ of $10^{-1}$ and $10^{-3}$ m/s in the physical model, but take on the values 0 and 1 in the geostatistical algorithms. The baseline represents the channeling geological patterns relevant to braided river deposits with the assumption that the $K$ is relatively homogenous within one soil type relative to the difference of $K$ between soil types.

The physical model is constructed in OpenGeoSys [Kolditz et al., 2012] to simulate groundwater flow and the time evolution of the concentration field for a tracer released on the $x = 0$ m boundary. The hydraulic head field is not transient and has an average gradient of 1% in the $x$-direction. The concentration field is initially zero and the source releases continuously. The direct data includes four measurements of $K$ and only one location for indirect data. The location measures the hydraulic head as well as the concentration time series over 100 time steps. See Figure 4.12 for the locations of the direct and indirect data.

Creating the training images

Four candidate training images are considered (Figure 4.13). The first training image, denoted hereafter as TI1, is a discretized image from a braided-river geology. The other three training images, TI2-4, are derivatives of TI1 and represent three alternative geostatistical methods. TI2 is indicator variogram, TI3 is truncated Gaussian, and TI4 is object-based. All four training images have the same mean and variogram, but only MPS is able to create the channeling structure in TI1. The choice of training images is inspired by the interest in being able to infer connected channels, or the lack thereof, while controlling for the first two spatial moments.

The baseline field was generated from TI1 with a size of 150-by-50 grid. From this baseline, observations can be extracted for the synthetic exercise. The soil type is sampled directly from the baseline from four locations, and the indirect data are extracted from the same location after the baseline case is used as input for the forward model. This exercise neglects measurement and physical model error.

Quantifying uncertainty

From the four training images, 1000 realizations of the geology were generated. The flow and transport processes were simulated on each of the realizations. The likelihood values were calculated non-parametrically by the anchoredDistr package described in Section 3.2. The prediction variable of interest considered is the time at which a concentration threshold is passed, denoted $\tau$. The same location as the indirect data placement is considered, and the resulting posterior distribution is sampled to create an ensemble of 100 predictions of $\tau$. 
CHAPTER 4. INTEGRATION OF MULTIPLE FORMS OF INFORMATION

Figure 4.12: The configuration of direct and indirect data locations.

Figure 4.13: The four training images.
CHAPTER 4. INTEGRATION OF MULTIPLE FORMS OF INFORMATION

Results

The posteriors resulting from using the hydraulic head, the final concentration, and the peak concentration are shown in Figure 4.14. The progression of these indirect data lead to improved posterior distributions with success defined as increased posterior mass assigned to the ‘true’ training image, TI1. The hydraulic head field is less affected by the channel network in TI1 due to the smooth nature of hydraulic head fields, so it is expected that provides less information than concentration for the inference. The concentration field is significantly more affected by the channeling network than the hydraulic head field. The final concentration at a location may or may not be indicative of the channel network depending on the length of the simulation, but the peak concentration should have high inference worth given that conductive channels should attract a high proportion of the tracer mass through a small channel, creating higher concentrations. The posterior distributions reflect these intuitions.

![Figure 4.14: The posterior distributions for the four training images based on the measurement of hydraulic head (left), final concentration (center), and peak concentration (right). The horizontal lines mark the prior probability mass.](image)

Using the posterior distribution from the peak concentration indirect data, an ensemble of 100 realizations of $\tau$ is generated. Similarly, the prior distribution is also used to generate $\tau$ values. Figure 4.15 compares the distribution of $\tau$ values from the prior and posterior along with the value calculated from the baseline field. The posterior’s $\tau$ distribution is more concentrated near the baseline value compared to the prior’s. Thus, the uncertainty in the environmental performance metric $\tau$ has been constrained, as well as quantified.

Conclusion

MAD was able to infer the correct training image given single values of indirect data. The posterior mass assigned to the correct training image increased with the progression of the indirect data coming from hydraulic head, final concentration, to peak concentration. The
relative success of concentration data compared to hydraulic head in the ability to infer connectivity in the field is in line with the findings of Heße et al. [2015]. The shape of the posterior distributions highlight MPS’s superiority over other geostatistical methods in modeling the connectivity of fields. When comparing the relative success of the prior and posterior distributions in predicting the time at which a concentration threshold was exceeded, the posterior distribution improved the prediction. By sampling a posterior distribution for future predictions, the quantified uncertainty regarding training images can be propagated to future predictions. Next steps include incorporating the dimension reduction methodology described in Section 4.1 using each training image as a baseline source for repeated baselines to estimate the rate of success in inferring the correct training image, and the incorporation of anchors in the inference.
Chapter 5

Geological Realism

Until this chapter, the focus of the dissertation has been on the Bayesian inversion of geostatistical models, including those from multipoint statistics which help improve the geological realism. Here in the penultimate chapter, I change direction to focus on geological realism and consider how detailed datasets called aquifer analogs can help inform the preliminary analysis of spatial heterogeneity and its effect on goal-oriented predictions before spatial models are formed.

Geological heterogeneity, i.e. the spatial variability of discrete hydrogeological units, is investigated in an aquifer analog of glacio-fluvial sediments to determine how such a geological setting can be simplified for characterization needs. The aquifer analog consists of ten hydrofacies whereas the scarcity of measurements in typical field studies precludes such detailed spatial models of hydraulic properties. Transport through three realizations of the aquifer analog is modeled with numerical particle tracking to ascertain the fast flow channel through which early arriving particles travel. Three simplification schemes of two-facies models are considered to represent the aquifer analogs and the velocity within the fast flow channel is used to estimate the apparent hydraulic conductivity of the new facies. The facies models in which the discontinuous patches of high hydraulic conductivity are separated from the rest of the domain yield the closest match in early arrival times compared to the aquifer analog, but assuming a continuous high hydraulic conductivity channel connecting these patches yields underestimated early arrivals times within the range of variability between the realizations, which implies that the three simplification schemes could be advised but pose different implications for field measurement campaigns.

5.1 Introduction

To predict the transport of contaminants in groundwater, the spatial heterogeneity of aquifer properties needs to be characterized. This characterization is often executed within the stochastic formulation of geostatistics such that the uncertainty of the subsurface due to sparse measurements can be incorporated into the modeling framework. In geostatistics, a
parsimonious spatial model represents the spatial structure of a target variable, e.g. hydraulic conductivity ($K$). One form of spatial structure to address in the subsurface is geological heterogeneity \cite{marsily2005, eaton2006}, which describes the distribution of hydrofacies, i.e. disparate soil types with relatively homogeneous hydraulic properties \cite{fogg1986, anderson1989}. This geological heterogeneity is important to consider in settings such as glacial and/or fluvial sediments that exhibit $K$ values that can vary over several orders of magnitude at the sub-meter scale \cite{labolle2001, alexander2011}.

In recent years, the collection of exhaustive, i.e. completely describing a field site, data into datasets referred to as ‘aquifer analogs’ has been pursued to convey more realistic complexity in geological heterogeneity \cite{bayer2011a, bayer2015, bayer2011b, hoyng2014, lochbuhler2014}. Here, we use ‘complexity’ to suggest spatial patterns neither easily described nor commonly assumed in classical geostatistics (i.e. two-point covariance models), such as curvilinear structures that need more than two points to describe \cite{guardiano1993}. The creation of these analogs does not intend to reflect the feasible amount of data resulting from standard measurement campaigns. In fact, the analogs are created for numerical studies to investigate how complex geological heterogeneity affects the physical processes that occur in the subsurface at locations similar to the aquifer analog sources to better inform future measurement campaigns or modeling efforts. However, the translation from insights gained by studying aquifer analogs to guidelines for field campaign design is needed. To accomplish that translation, the complexity of the analogs needs to be simplified into a spatial model that is parsimonious enough to be well-posed in the face of sparse information \cite{schoeniger2015} and structured enough to reflect the geology and physical processes being simulated \cite{linde2015}.

There are two obstacles in characterizing geological heterogeneity to overcome: facies delineation and hydraulic property assignment. The first obstacle manifests in the simplification process by needing to select the number of facies or zones to incorporate in the spatial model and how they are oriented in space. In field studies, the methods of choosing this number include interpreting the multimodality in geophysical results such as cross-hole tomography \cite{hyndman1994} or borehole results for $K$ or lithology \cite{labolle2001, burns2010, bianchi2011}, using cluster analysis with $K$ samples \cite{riva2008}, expert opinion on local geology, or a combination thereof. The delineation of these facies is often left to geostatistical methods or geologic interpretation that interpolate between sparse measurements. However, with the aid of aquifer analogs, the opportunity to test the method of facies delineation on process outcomes presents itself. To conduct this test, the second obstacle needs to be included since simulation of any process, such as transport, requires the facies delineation but also the hydraulic properties of each facies. There has been extensive research into estimating effective or equivalent $K$ to upscale aquifer properties by homogenization while trying to reflect complex spatial patterns and processes on different spatial scales \cite{knudby2006, fleckenstein2008, oriani2014, tyukhova2016}, but the goal of simplifying complex heterogeneity is to maintain the heterogeneity features that are relevant for a
certain prediction goal while reducing less relevant complexity, i.e. the homogenization is not absolute and multiple facies remain. In the simplification process, an apparent $K$, a $K$ value to encapsulate the mechanistic behavior of the remaining facies in the system given simplification is of the same philosophy as homogenization, but with a different goal of a mechanistic evaluation of the geological heterogeneity, similar to Ronayne et al. [2010], that results in multiple facies. Ultimately, the $K$ values assigned to the simplified facies need to be in concert with the structure of the new facies to estimate the quantity pertaining to the goal of the investigation. However, as concluded by Feyen and Caers [2006], the facies delineation is more influential on transport than the assignment of hydraulic conductivity, thus the priority of this study will be on comparing facies models, along the line of Klise et al. [2009] and Bianchi et al. [2011] but with a three-dimensional aquifer analog.

The benefit of converting aquifer analogs into simplified facies models for specific prediction goals is that future measurement campaigns for site characterization can be tailored to finding relevant features. For example, Parra et al. [2006] studied the effectiveness of cross-well seismic reflection data to delineate flow units, but commented that the method is only advisable when lateral continuity can be assumed. If predicting early arrival times of a solute is dependent on incorporating discontinuity of high $K$ facies, then a measurement technique is needed at an adequate spatial resolution in at least two spatial dimensions, i.e. geophysical methods, and likely a combination of them. Tronicke et al. [2002]. In Sassen et al. [2012], geological prior knowledge hinted at spatial continuity of sediments with relevant biogeochemical properties, thus a combination of high-resolution geophysical methods sensitive to the chemistry were used to delineate ‘reactive facies’. Specific to glacial-fluvial deposits, Alexander et al. [2011] calls for more advanced site characterization methods for strongly heterogeneous geologic media after being unable to systemically improve drawdown estimates in their field site with continuous multichannel tubing, slug tests, and permeameters. Resources such as Binley et al. [2015] can be a vital resource for identifying the applicable characterization techniques for a given site after the investigation of aquifer analogs or genetic- or process-based models are investigated to ascertain which features are relevant for a specific prediction goal.

This study aims to provide guidelines for future site characterization efforts from insights gained by simplifying aquifer analogs in a goal-oriented approach. To accomplish this, we use an aquifer analog of glacio-fluvial sediments in southwestern Germany as provided by Bayer et al. [2011b] and focus on the goal of predicting early arrival times for a tracer transported through the analog. Section 5.2 provides details regarding the aquifer analog, the numerical transport modeling, and the simplification process. Section 5.3 describes the transport modeling results in the aquifer analog, the resulting apparent $K$ values from three alternative simplification schemes, and the verification of early arrival predictions of those schemes. Section 5.4 concludes with an overview of the investigation and implications for future field campaigns that can be extracted from the study.
5.2 Methods

The aquifer analog

The aquifer analog used in this study is from a gravel pit in the Rhine valley in southwestern Germany which exhibits fluvioglacial sedimentology modeled as ten hydrofacies over the extent of $16 \times 10$ m horizontally and 7 m vertically. The field site is in the Rhine basin, where glacial and fluvial deposits dominate the geology. This form of depositional environment causes discontinuous layers of significantly varying hydraulic properties, such as $K$ ranging over seven orders of magnitude in this analog ($6 \cdot 10^{-7} - 1.3 \cdot 10^{-1}$ m/s). The measurement campaign and soil classification of the six measured outcrops used to create the analog are discussed by Bayer et al. [2011a]. The expansion of the six outcrops into the three-dimensional aquifer analog using a combination of two-point and multi-point geostatistics is detailed by Bayer et al. [2011b]. Because the aquifer analog was created in a stochastic framework, there is an ensemble of realizations available for investigation. In this study, we will use the three realizations provided as supplemental data by Bayer et al. [2011b]. Although Comunian et al. began to model transport with the analog, the influence of the hydrofacies on preferential flow was not examined beyond the relative and qualitative similarity of two-dimensional plumes between simulated and measured cross-sections to suggest plausibility for their realizations; hence, the quantification of arrival times through the analog remains unaddressed.

The original analog resolution was coarsened from 5cm to 25cm both horizontally and vertically by selecting every fifth element to reduce the computational burden of the numerical simulations. Fig. 5.1 shows the intersection of three cross sections of the coarsened analog (Realization 1) used in this study. Each hydrofacies is assigned distinct $K$ and porosity values (Table 5.1) and their prevalence in the analog is similar for the three realizations (Fig. 5.2). Hereafter, the hydrofacies are numbered by increasing $K$ from 1 to 10 and referred to as ‘H1’ and so forth. The higher $K$ hydrofacies, H8-H10 ($2.6 \cdot 10^{-2} - 1.310^{-1}$ m/s), constitute 2.8-3.1% of the realizations, but are multiple orders of magnitude greater in $K$ than the other hydrofacies ($6 \cdot 10^{-7} - 2.3 \cdot 10^{-3}$ m/s). In all cross-sections, the high-$K$ hydrofacies H8-H10 appear clustered in the middle of the analog. From the cross-sections of Fig. 5.1, it can be seen that the geological heterogeneity is still complex at 25cm resolution, particularly for the horizontal cross-section (Fig. 5.1).

<table>
<thead>
<tr>
<th>Hydrofacies</th>
<th>$K$ (m/s)</th>
<th>$n$ (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>$6.0 \cdot 10^{-7}$</td>
<td>0.20</td>
</tr>
<tr>
<td>H2</td>
<td>$4.3 \cdot 10^{-5}$</td>
<td>0.22</td>
</tr>
<tr>
<td>H3</td>
<td>$6.1 \cdot 10^{-5}$</td>
<td>0.13</td>
</tr>
<tr>
<td>H4</td>
<td>$1.4 \cdot 10^{-4}$</td>
<td>0.36</td>
</tr>
<tr>
<td>H5</td>
<td>$2.3 \cdot 10^{-4}$</td>
<td>0.15</td>
</tr>
<tr>
<td>H6</td>
<td>$2.5 \cdot 10^{-4}$</td>
<td>0.17</td>
</tr>
<tr>
<td>H7</td>
<td>$2.3 \cdot 10^{-3}$</td>
<td>0.27</td>
</tr>
<tr>
<td>H8</td>
<td>$2.6 \cdot 10^{-2}$</td>
<td>0.26</td>
</tr>
<tr>
<td>H9</td>
<td>$9.5 \cdot 10^{-2}$</td>
<td>0.23</td>
</tr>
<tr>
<td>H10</td>
<td>$1.3 \cdot 10^{-1}$</td>
<td>0.26</td>
</tr>
</tbody>
</table>
The numerical transport model

The numerical model employed to simulate groundwater flow and particle tracking on the three-dimensional aquifer analog was executed in OpenGeoSys [Kolditz et al., 2012], a finite element software for simulating various subsurface processes. The groundwater flow configuration in the domain is designed to be uniform in the average with the predominant flow direction along the x-axis. At \( x = 0 \) m and \( x = 16 \) m, the boundaries of the x-axis, there are constant head boundary conditions constructing a head gradient of approximately 1% with flow traveling in the positive x-direction. The remaining boundaries on the y- and z-axes are no flow boundaries. The flow configuration is designed to be steady-state. Using the domain’s \( K \) values of the elements and the head and flow boundary conditions, OpenGeoSys solves the groundwater flow equation with the finite element method to yield the head field over the domain.

With the solved head field and the porosity values from the analog, the velocity field can also be obtained. The global node-based method for obtaining the continuous velocity field via the finite element method [Park et al., 2008b] is used here to prevent discontinuities in velocities across element boundaries for more accurate particle tracking. Only advection is considered in this study since the goal is early arrival time prediction, and the early arrivals result from a fast flow channel created by high-\( K \) zones where advection is dominant and mechanical dispersion is accounted for in the \( K \) variability between hydrofacies, similarly
argued as Zheng and Gorelick [2003] and Bianchi et al. [2011].

To analyze transport paths taken through the analog, the particle tracking feature in OpenGeoSys was used [Park et al., 2008a]. A total of 40,000 particles were uniformly distributed across the x=1m plane with a buffer of 0.25m along the edges, resulting in approximately 42 particles/element. The number of particles and the time interval used to advance them were chosen by assessing the qualitative convergence of the arrival time distribution as the number of particles increased and the time intervals decreased. The appropriate time interval to resolve the particles’ paths was determined to be every 400 seconds out of a total simulated time of 23.15 days.

The simplification process

Based on the particle tracking results, the fast flow channel through the three aquifer analog realizations can be determined to help the simplification process for predicting early arrival times. Fig. 5.3 depicts the conceptual workflow of considering simplification schemes and apparent $K$ values. The following list describes the steps taken for each aquifer analog realization to define the fast flow channel, select simplification schemes, estimate apparent $K$ values per facies within a simplification, and verify early arrival times.

1. **Identify the particles representing the prediction goal:** Since the prediction goal of this study is early arrival times, the first 100 particles to arrive at the control plane of x=16m are identified as representing the prediction goal. Alternatively, the early arrival particles could be defined by a cutoff arrival time, particularly in cases in which a pre-defined critical arrival time is of interest, such as in the case of being imposed by regulations.

2. **Identify the goal-relevant portion of the domain that is associated with the prediction goal particles:** The goal-relevant portion of this analog was considered
the finite element nodes and elements through which the early arriving particles traveled, which we will use to define the fast flow channel. The nodes that the particles were closest to for each time step until their arrival at the control plane were identified with the nearest neighbor algorithm [Elseberg et al. 2012] by comparing the particle path output to the mesh definition, both files from OpenGeoSys. This is the nodal definition of the fast flow channel that allows the comparison of velocity. Additionally, the elements attributed to the fast flow channel were identified as every element that is defined by at least one fast flow channel node. This is the element definition of the fast flow channel that allows the comparison of hydrofacies and $K$.

3. **Compare the hydrofacies proportions in the domain and the goal-relevant portion to define simplification schemes:** Histograms of hydrofacies presence in the analog and the fast flow channel are compared to identify which hydrofacies are at higher proportions in the fast flow channel than the analog as a whole. Multiple simplifications may be made subjectively for comparison to ascertain if a hydrofacies elevated presence in the fast flow channel is due to that hydrofacies driving early arrival time or due to coincidence such as juxtaposition to the hydrofacies that drive early arrival time. Two-facies models are the simplest simplification besides complete homogenization, so we begin with only considering simplifications with two facies. If all the two-facies simplifications fail to reproduce the prediction goal (see Step 5 below), then increasing the numbers of facies may be warranted.

4. **Search potential combinations of apparent $K$ values for each simplification from the range of $K$ values present in the hydrofacies being combined:** For this study, pairs of 50 $K$ values at an even interval on a logarithmic scale for each facies were searched, i.e. 2500 pairs of $K$ values were considered. For each pair of $K$ values considered for the apparent $K$ values, the velocity field $V_{K,i}$ was solved and compared to the analog velocity field $V_{o,i}$ by the following two metrics concerning the nodes $i = 1 \rightarrow N$, where $N$ is the number of nodes in the fast flow channel. The apparent $K$ values were estimated by selecting the pair of $K$ values that minimizes both metrics. Here, porosity is assumed constant (0.22) in light of the greater variability in $K$. Mean Error (ME): The average difference between across nodes of the decimal log velocity fields:

$$ME = \frac{1}{N} \sum_{i=1}^{N} (V_{K,i} - V_{o,i})$$

Root Mean Squared Error (RMSE): The square root of the mean of the node-based squared error:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (V_{K,i} - V_{o,i})^2}$$

5. **Confirm prediction goal for the simplifications with their estimated apparent $K$ values:** Equipped with apparent $K$ values for each facies in each simplification,
particle tracking can be applied to each simplification to predict the early arrival times. The early arrival time predictions can be used to compare the relative success of the simplifications in estimating the same early arrival times as the analog and judging if the discrepancy between the best performing simplification and the analog is sufficient for the application. If not sufficient, then additional simplification with possibly more facies may be warranted.

Figure 5.3: The conceptual process of testing simplification schemes for a specific realization: Starting with an aquifer analog (left, where color denotes hydrofacies), multiple simplification schemes which represent facies structure are identified (middle, where color denotes facies), and then for each simplification a variety of hydraulic conductivity ($K$) combinations are attempted to reconstruct the transport of early arriving particles (right, where color denotes $K$).

5.3 Results and Discussion

The early arrival times and fast flow channel

The results from Steps 1-3 of Section 5.2 are detailed in this section. Fig. 5.4 depicts the arrival times of particles for the three aquifer analog realizations. Fig. 5.4 zooms in to the first 100 particles, which have arrival times occurring within the range of 24-48 hours after release while the full arrival time distribution occurs over 23.15 days, by which approximately 96.5% of the particles arrive at the control plane. The arrival time curves for the three
realizations appear similar and indicate that the realizations have comparable transport characteristics. The structure of the fast flow channels through which these 100 particles travel is also similar in that the channels meander in three-dimensional space between the release and control planes while being able to have multiple branches (Fig. 5.5).

To devise the simplification schemes to reconstruct these early arrival times, the proportion of hydrofacies present in the fast flow channels versus the analog realizations is compared (Fig. 5.6). For each realization, H4 and H8-10 are present at higher proportions in the fast flow channel than in the analog. For example, H8-10 comprise approximately 3% of the analog in each realization, but increase to 23-28% of the fast flow channels. Similarly, H4 makes up 7-8% of the analog realizations, but 10-16% of the fast flow channels. Due to the several orders of magnitude difference in $K$ between H4 and H8-10 and the relatively weaker gain in abundance for H4, we consider one simplification scheme with H8-10 combined (Simplification 1) and another with H8-10 combined with H4 (Simplification 2). Since these four hydrofacies do not constitute the entire fast flow channel, i.e. collectively only 33-44%, an alternative simplification scheme is considered in which the fast flow channel elements are considered a new facies irrespective of hydrofacies (Simplification 3) for comparison.

The spatial configuration of the simplifications within the domain is depicted in Fig. 5.7. Simplification 1 represents the case of rare (2.8-3.1% of the analog) discontinuous patches of high $K$ embedded in a homogeneous facies of lower $K$. This construction of a two-facies models based on separating the high $K$ values from the rest of the domain is similar to previous studies, such as Bianchi et al. [2011], who segregate $K$ values above and below the 90th percentile value of the natural logarithm of $K$ from measurements. Simplification 2 is similar but incorporates another hydrofacies, H4, that is often juxtaposed next to H8-10, thus acting as a bridge between the high-$K$ patches, that brings the higher-$K$ facies up to 10-11% of the analog, thereby improving the chances of encountering this facies in measurement campaigns while still maintaining similar discontinuity compared to Simplification 1. Simplification 3 represents the case of ignoring discontinuity and assuming fully connected high-$K$ channels.

The apparent hydraulic conductivities

The results from Step 4 of Section 5.2 are detailed in this section. For each of the three simplification schemes for each of the aquifer analog realizations, a search of potential apparent $K$ value pairs resulted in a neighborhood of minimal velocity error metrics as seen in Fig. 5.8, where the axes’ ranges reflect the scope of potential apparent $K$ searched. For example, Simplification 1 combined H8-10 into one facies and H1-7 into another such that the higher apparent $K$ was searched between the $K$ values assigned to H8-10 (i.e. $2.6 \cdot 10^{-3} - 1.3 \cdot 10^{-1}$ m/s) and the lower apparent $K$ was searched between the $K$ values assigned to H1-7 (i.e. $6.0 \cdot 10^{-7} - 2.3 \cdot 10^{-3}$ m/s). Among the three realizations for a specific simplification, the location of the minimal error metrics in terms of apparent $K$ pairs is similar, but the intersection of minimal ME and RMSE occurs at different pairs for the three simplification schemes. To estimate the apparent $K$ pair for each simplification, a location
Figure 5.4: The arrival times in days (top), cumulative arrival times in days (middle), and a close up of cumulative arrival time of the first 100 particles in hours (bottom) for the three aquifer analog realizations before simplification. The first 100 particles will be used to create simplification schemes, and the arrival times of these particles will be used to gauge the success of those simplification schemes in predicting early arrival times.

Figure 5.5: The elements in the three aquifer analog realizations (top, middle, and bottom) attributed to the fast flow channel as determined by the paths of the first 100 arriving particles. This is the goal-relevant portion of the analogs with which simplifications will be defined and their apparent hydraulic conductivity values will be estimated.
CHAPTER 5. GEOLOGICAL REALISM

Figure 5.6: The proportion of hydrofacies in the fast flow channel (red) versus in the entire analog (blue) for each of the three aquifer analog realizations (top, middle, and bottom). The hydrofacies that are at elevated proportions in the fast flow channel compared to the entire analog for all realizations, H4 and H8-10, will be used to construct simplification schemes. Although H7 is at a higher proportion in the fast flow channel than the domain in Realization 1, that is not the case for Realizations 2 and 3, therefore it will not be used for simplification scheme definitions.

Figure 5.7: The spatial configuration of the three simplifications for Realization 1. Simplification 1 (top) is defined by H8-10 and the three colors distinguish between the three hydrofacies. Simplification 2 (middle) is the same as Simplification 1 but with the addition of H4, which is marked by a fourth color and can be seen to act as a bridge between discontinuous patches of H8-10. Simplification 3 (bottom) is equivalent to the fast flow channel and is not based on hydrofacies, so there are no colors assigned.
was handpicked where this intersection appeared for all three of the realizations. For Simplifications 1-3, the apparent $K$ pair is estimated at $5 \cdot 10^{-4}$ and $6 \cdot 10^{-2}$ m/s, $6 \cdot 10^{-4}$ and $2.1 \cdot 10^{-2}$ m/s, and $3.5 \cdot 10^{-5}$ and $3.7 \cdot 10^{-2}$ m/s for the two facies, respectively.

Figure 5.8: The 10 pairs of potential apparent K with the least ME and RMSE for the three aquifer analog realizations for each of the three simplification schemes (Simplification 1-3 from left to right). The areas of concentrated minimal error metrics denote the pairs of $K$ values that minimize the error between the velocity values in the fast flow channel of the analog and the simplifications, i.e. apparent $K$ values for each simplification. These estimated values (shown by the intersection of the black lines) allow for the transport modeling in the simplifications to assess early arrival time prediction success.

Compared to the constitutive hydrofacies’ $K$ values, the estimated apparent $K$ for the lower-$K$ facies of Simplifications 1 and 2 ($5 \cdot 10^{-4}$ and $6 \cdot 10^{-4}$ m/s) are on the same order of magnitude as the value of the geometric mean $K$ in the analogs ($3.21 \cdot 10^{-4} - 3.23 \cdot 10^{-4}$ m/s) while the estimated apparent $K$ for the higher-$K$ facies in all three simplifications ($6.0 \cdot 10^{-2}$, $2.1 \cdot 10^{-2}$, and $3.7 \cdot 10^{-2}$ m/s) are similar to that of the arithmetic mean of $K$ in the fast flow channels ($1.8 \cdot 10^{-2} - 2.3 \cdot 10^{-2}$ m/s). For Simplification 3, the estimated apparent $K$ for the lower-$K$ facies ($3.5 \cdot 10^{-5}$ m/s) is less than the harmonic mean of the analogs ($4.96 \cdot 10^{-5} - 1.0 \cdot 10^{-5}$ m/s).

In terms of error metrics, Simplification 1 had the least error in reproducing the velocity in the fast flow channel. All three simplifications had both positive and negative ME values within the apparent $K$ search and the minimal ME values were at most $3.47 \cdot 10^{-7}$, $1.05 \cdot 10^{-6}$, and $7.83 \cdot 10^{-7}$ m/s for Simplifications 1-3, respectively. The minimum RMSE for Simplifications 1-3 were $1.47 \cdot 10^{-5}$, $2.59 \cdot 10^{-5}$, and $3.15 \cdot 10^{-5}$ m/s, respectively. However, due to the varying resolutions of the apparent $K$ searches and the ultimate goal is predicting
early arrival times, it is more reliable to gauge the success of the simplifications by the transport modeling results.

The comparison of transport in the simplification schemes

The results from Step 5 of Section 5.2 are detailed in this section. Particle tracking was applied with the same specifications as discussed in Section 5.2 but with updated $K$ values reflecting the pairs of estimated apparent $K$ identified for each simplification scheme in the section above. Fig. 5.9 shows the estimated early arrival times from the three simplification schemes for each of the aquifer analog realizations. For Realization 1 and 2, Simplification 1 performed the best with arrival times within 2.3 hours of the aquifer analog’s arrival times while Simplification 2 was as much as 21 hours early. However, for Realization 3, Simplifications 1 and 2 have nearly the same error of 3-6 hours either late or early, respectively. Simplification 3 is consistently early by 8-20 hours early.

One potential source of discrepancy of simplification schemes’ success among the realizations is the difference in fast flow channel structure. In Fig. 5.5, Realization 3 consists of a single channel while Realizations 1 and 2 have branching due to early arriving particles circumventing low-$K$ areas. The influence of this branching or lack thereof may affect the velocity field such that there is less agreement between the realizations when a single pair of apparent $K$ is used. However, in practice, the apparent $K$ will be likely inferred within an ensemble of more than three realizations of the simplified analogs and the uncertainty in the early arrivals can be characterized stochastically, i.e. an ensemble of arrival times will be produced to relieve the influence of individual realizations.

5.4 Conclusion

In this study, we have translated a ten-hydrofacies aquifer analog into three simplified two-facies models for the explicit purpose of predicting early arrival times of particles in transport modeling with a more parsimonious spatial model of $K$. Three realizations of the stochastically generated aquifer analogs were used to assess generality. The three simplification designs for these realizations of the analog were constructed based on the hydrofacies histograms in the fast flow channel, i.e. the collective path of the early arriving particles. Simplification 1 consisted of the top three highest $K$ hydrofacies, H8-10, as one facies with the remaining hydrofacies, H1-7, as the second facies. Simplification 2 was similar but contained H4 along with H8-10 due to its proximity by H8-10 throughout the analog and its elevated presence in the fast flow channel compared to the entire analog. Simplification 3 assigned the elements into facies by their presence within the fast flow channel or lack thereof for comparison to hydrofacies-based simplifications. Estimates of apparent $K$ pairs, i.e. one apparent $K$ for each facies in the two-facies models, were estimated by comparing velocity fields from the simplifications and the analog realizations. For each simplification applied to each realization, the arrival times of early particles were compared to that of the analog to
Figure 5.9: The arrival times of the first 100 particles for the three aquifer analog realizations (top, middle, and bottom) based on three simplification schemes as compared to the analog arrival times. Simplification 1 performs the best across the three realizations, Simplification 2 performs as well as Simplification 1 for Realization 3 only, and Simplification 3 performs the worst across the realizations in terms of accurately predicting the analog arrival times.
compare success of each simplification. Simplification 1 performed the best of the three simplifications with arrival times within 0.23, 1.25, and 5 hours (on average for the realizations) of the analog in which arrival times ranged from approximately 1-23 days. Simplification 2 and 3 estimated early arrivals within 19.6, 13.6, and 4.5 hours and 19.6, 11.3, and 13.76 hours of the analogs, respectively.

Depending on the required accuracy of the early arrival time predictions, either Simplification 1 or 2 could be advised for early arrival time prediction in glacio-fluvial sediments. Both incorporate discontinuity in the higher-K facies representing open framework gravel (H8-10), but Simplification 2 includes the juxtaposed sandy H4 with this gravel. If prior geological knowledge pertaining to this juxtaposition is available, efforts to locate this facies with measurements are aided by the increased presence from 3% to 10% of the higher-K facies in Simplification 1 to Simplification 2. Simplification 3 is not advised since it regularly under-estimated the arrival time and does not directly reflect the distribution of hydrofacies, i.e. does not aid future field campaign designs as clearly as instructions on relevant hydrofacies. The combination of discontinuity in the higher-K facies with the success of Simplification 1 agrees with earlier findings that the connectivity in transport processes, i.e. the manifestation of a fast flow channel, is possible with the lack of connectivity in aquifer properties [Knudby and Carrera, 2006; Oriani and Renard, 2014].

Although one specific aquifer analog was used, sites of similar depositional environments could have similar geological heterogeneity and contributing to the common knowledge of facies heterogeneity is paramount for future characterization efforts, as urged by de Marsily et al. [2005]. For dissimilar depositional environments, the same task of goal-oriented analog simplification can be applied to other aquifer analogs or even process- or genetic-based models to inform future field characterization for their own site-specific heterogeneity and prediction needs. Future field characterization at sites with similar geological settings as the aquifer analog used in this study will not have a site-specific aquifer analog, but can benefit from this study and other studies of the analog that detail the importance of geological heterogeneity in processes such as transport. The analogs circumvent the uncertainty in the subsurface by posing a generic but data-inspired aquifer such that a synthetic study can be used to investigate the role between realistic complexity and processes of interest. Field studies, which must deal with that uncertainty, benefit from the insights gained about the facies structural patterns. Inverse modeling can be used in concert with geostatistics to constrain either the uncertainty in local placement of facies, the hydraulic properties of the facies, or both.

Future expansions of this study could incorporate additional transport processes such as dispersion and/or reactive chemistry if warranted by the prediction goal. For goals such as remediation modeling, which depends on accurate prediction of late arriving particles as opposed to early arriving particles in this study, the facies simplification schemes and resulting apparent $K$ values will certainly be different than the results found in this study. There is an inherent trade-off when simplifying a spatial model, i.e. reducing a ten-facies analog to a two-facies model in which there is less predictive power of the model in terms of the complexity in the response, i.e. only early arrival times are compared for the two-facies
models as opposed to the entire arrival time curve of particles, which is the motivation for the goal-oriented approach to facies structure simplification.
Chapter 6

Conclusion

In Chapter 1, the state of stochastic hydrogeology was discussed. According to published debates on the lack of use of stochastic hydrogeology methods in practice, multiple reasons were suggested as to why stochastic hydrogeology methods may not be adopted by practitioners. Those reasons included the lack of (1) user-friendly and freely available software for applying general stochastic methods, (2) data integration frameworks for all available information, and (3) applicable geological realism in geostatistical models. After discussing background material in Chapter 2, the subsequent chapters addressed contributions to alleviate each of these reasons.

Chapter 3 addressed the first reason. It discussed software that can integrate multiple forms of information from the field to constrain uncertainty about geostatistical parameters. The main advantages of MAD, the technique, include a fully probabilistic treatment of uncertainty that can be propagated into future modeling, the characterization of both global and local parameters (geostatistical parameters and anchored distributions), and the lack of assumptions regarding statistical distributions or physical processes. The main advantage of the software itself is its modular design that upholds the lack of assumptions in MAD - any random field generator and any forward model can be connected with MAD so users are not constrained to any specific geostatistical model or physical process. Both a graphical user interface and an R package were discussed. These provide the software tools for practitioners and researchers to attach their numerical models to geostatistical software to quantify and constrain the uncertainty in spatially-variable fields without having to implement the Bayesian framework themselves. Previously, numerical models have come equipped with parameter estimation modules or external programs have been coupled with numerical models for parameter estimation, but MAD is the first to provide a fully probabilistic framework for spatially-variable fields that can be coupled with numerical models.

Chapter 4 utilized the software tools from Chapter 3 to incorporate two new forms of information into MAD applications: data series and conceptual models of geological heterogeneity. Data series, such as the change in hydraulic head or tracer concentration over time, are not new forms of information but were not until now thoroughly investigated for integration into the MAD framework. It was shown that drawdown time series can be effi-
Ciently incorporated into the Bayesian inference process of MAD by replacing the 100 time step series with two parameters of a modified Mat’ern function. This allows practitioners to utilize a common form of field measurements, i.e. drawdown measurements after pumping, in the MAD framework that allows them to quantify and constrain their uncertainty in the spatially-variable hydraulic conductivity field. The same methodology can be followed to explore the use of other time series or depth profiles. MAD was also shown to be applicable for the other form of new information, i.e. conceptual models of geological heterogeneity. This relied on the use of multipoint statistics, a relatively new geostatistical approach that can describe a wider range of spatial patterns than classical two-point statistics. This approach is praised for being able to represent realistically complex geology, but has the disadvantage of needing a training image supplied. Since MAD was shown to be able to infer the training image from a collection of candidate training images, MAD can now be applied to a wider range of application areas, such as field sites with a channel network of soils.

Chapter 5 reflected on the discussion of conceptual models of geological heterogeneity with respect to how conceptual models can be tailored in ‘goal-oriented’ characterization efforts. By analyzing a three-dimensional ‘aquifer analog’ data set with 10 hydrogeological units, simplification schemes were considered and evaluated on how well the prediction goal was sustained after the analog was simplified. By tailoring the simplicity of the spatial patterns, which is a necessity when we only have sparse information from the field, to the prediction goal, then we can ensure that the ultimate goal of improving groundwater modeling efforts is achieved. It was shown that the 10 hydrogeological units could be simplified down to two hydrogeological units while keeping the predicted early arrival time of a tracer within acceptable limits. Simplifying the spatial structure of the geology while keeping the prediction goal maintained paves the way for building goal-oriented spatial models that are parsimonious enough for uncertainty quantification, such as with MAD. This helps the future characterization and predictions at field sites with complex geology.

Ultimately, any uncertainty quantification regarding the spatial heterogeneity of subsurface properties has the goal of improving predictions of groundwater modeling efforts. With the addition of freely available software tools, the ability to integrate more forms of information that include conceptual models of geology, and methodology for translating complex geological structures into parsimonious spatial models, the characterization of our groundwater resources improves. However, there is always more room for improvement. The work described thus far could be expanding in multiple directions.

First, the lines of research started here could be explored further and even combined. The aquifer analog simplified in Chapter 5 can be translated into geostatistical models, either two-point or multipoint depending on the simplification, and MAD can be used to infer the spatial models with a variety of inversion data. One form of inversion data that is of interest is depth profiles of hydraulic head. From simulating the hydraulic head field over the aquifer analog, it can be seen that there are vertical variations in hydraulic head. However, in practice, hydraulic head is commonly measured in a depth-averaged manner such that vertical variations are not seen. If an application of MAD to the aquifer analog suggests that depth profiles of hydraulic head are more informative than depth-averaged
hydraulic head, then field measurement techniques that can measure these depth profiles, such as direct-push injection logging, can be suggested for field studies and prediction goals in similar geologies. The combination of simplification, spatial model, and inversion data can change with a survey of other prediction goals. Altogether, the investigation into the application of MAD to the aquifer analog would address the call for both geological realism, as well as the suggestion for goal-oriented frameworks, found in Chapter 1.

Second, the MAD software can be applied to more field applications. Over et al. [2015] showed one application of the software and MAD to a field study, but it would be beneficial to provide publications that are geared toward practitioners and that showcase the advantages of the software to applications in hydrogeology with field data. This would be aligned with the suggestions described in Chapter 1 that call for more application-oriented literature for practitioners to encourage the adoption of stochastic hydrogeology methods in practice. Of particular interest would be incorporating data from Geographic Information Systems databases, which are common stores of geospatial data in practice, but have thus far not been shown to be compatible with MAD#. One specific application that has promise is modeling the structure of regolith, i.e. the material from bedrock to the ground surface, on the regional scale. This structure is important to characterize since the water storage capacity of regions, such as the Sierra Nevada which provides drinking water to California, depends on the porosity of the regolith material. There is a multitude of field data related to the regolith material provided by the Critical Zone Observatory network, but the information is still sparse compared to the extent of space that needs to be characterized. This application could be published with practitioners as the target audience, would incorporate new forms of geophysical data and petrophysical models for MAD, and would showcase an application in hydrogeology with a different prediction goal than used so far with MAD.

Third, the use of multipoint statistics with MAD opens many doors for uncertainty quantification outside of hydrogeology. With the ability to model any repetitive pattern with training images offered by multipoint statistics, and given the generality of MAD, uncertainty quantification could potentially be applied in modeling the spatial distribution of vegetation patches on hill slopes in ecohydrology or even imperfections in 3D-printed material in material science.
References


REFERENCES


C. Zheng and S. M. Gorelick. Analysis of solute transport in flow fields influenced by preferential flowpaths at the decimeter scale, 2003. ISSN 0017467X.