Title

Development of A Bayesian Geostatistical Data Assimilation Method and Application to the Hanford 300 Area

Author

Murakami, Haruko

Publication Date

2010

Development of A Bayesian Geostatistical Data Assimilation Method and Application to the Hanford 300 Area

by

Haruko Murakami

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering-Nuclear Engineering

in the

Graduate Division
of the
University of California at Berkeley

Committee in charge:
Professor Yoram Rubin, Co-Chair
Professor William E. Kastenberg, Co-Chair
Professor Per F. Peterson
Professor David R. Brillinger

Fall 2010

Development of A Bayesian Geostatistical Data Assimilation Method and Application to the Hanford 300 Area

Abstract

Development of A Bayesian Geostatistical Data Assimilation Method and Application to the Hanford 300 Area

by

Haruko Murakami
Doctor of Philosophy in Engineering-Nuclear Engineering

University of California at Berkeley

Professor Yoram Rubin, Co-Chair
Professor William E. Kastenberg, Co-Chair

Probabilistic risk assessment of groundwater contamination requires us to incorporate large and diverse datasets at the site into the stochastic modeling of flow and transport for prediction. In quantifying the uncertainty in our predictions, we must not only combine the best estimates of the parameters based on each dataset, but also integrate the uncertainty associated with each dataset caused by measurement errors and limited number of measurements. This dissertation presents a Bayesian geostatistical data assimilation method that integrates various types of field data for characterizing heterogeneous hydrological properties. It quantifies the parameter uncertainty as a posterior distribution conditioned on all the datasets, which can be directly used in stochastic simulations to compute possible outcomes of flow and transport processes. The goal of this framework is to remove the discontinuity between data analysis and prediction. Such a direct connection between data and prediction also makes it possible to evaluate the worth of each dataset or combined worth of multiple datasets. The synthetic studies described here confirm that the data assimilation method introduced in this dissertation successfully captures the true parameter values and predicted values within the posterior distribution. The shape of the inferred posterior distributions from the method indicates the importance of estimating the entire distribution in fully accounting for parameter uncertainty. The method is then applied to integrate multiple types of datasets at the Hanford 300 Area for characterizing a three-dimensional heterogeneous hydraulic conductivity field. Comparing the results based on the different numbers or combinations of datasets shows that increasing data do not always contribute in a straightforward way to improving the posterior distribution: increasing numbers of the same data type would not necessarily be beneficial above a certain number, and also the combined effect of multiple datasets could be very different from the individual effects. This research thus shows that any site characterization should consider not only the worth of individual datasets, but also the combined effects of multiple datasets and the sequence in which they are assimilated.

To my family
and
in memory of my grandmothers

Contents

List of Figures

List of Tables

Acknowledgements

I still remember my excitement while taking the course taught by my research advisor Professor Yoram Rubin in 2007. Since then, I have been fascinated by the depth and elegance of geostatistics and stochastic hydrology. I was very fortunate that I had a chance to work with Professor Rubin and pursue my career in this area. As an advisor, he has given me invaluable guidance with his expertise and passion. His extraordinary creativity and energy in research always opened my eyes and pushed my research to much higher level. Without his guidance, I could not have come to the acomplishment of finishing my Ph.D. program.

I have also benefitted significantly from interaction with other professors at UC Berkeley. My academic advisor Professor William Kastenberg provided helpful support during my Ph.D. program, giving me many pieces of advice regarding various academic matters. He also introduced me the concept of probabilistic risk assessment and stochastic modeling in engineering systems, which fueled my interest in applying geostatistics and stochastic hydrogeology to engineering problems. Professor David Brillinger has provided me knowledge and insight in statistics, which has become nuts and bolts in my dissertation research. During my first several years in Berkeley, I learned a lot from my master advisor Professor Joonhong Ahn, who taught me valuable analytical skills such as how to make assumptions and how to simplify problems that has become an invaluable asset in my research career. I would also like to thank Professor Per Peterson for his valuable advice and feedback, especially in my qualifying exam.

In addition to Jane Lewis Fellowship, this work was supported by the Integrated Field Research Project at the Hanford 300 Area funded by the U.S. Department of Energy, Office of Biological and Environmental Research through DOE-ERSP grant DE-FG02-06ER06-16. It was such a fortunate that I was able to work on the high-quality data and to have discussions with the world-class researchers in the project, including Dr. John Zachara, Mr. Vince Vermerl, Dr. Glenn Hammond and Dr. Peter Lichiter. Dr. Mark Rockhold was especially helpful by giving me advice and helping me develop papers and chapters in this dissertation.

I have enjoyed working with my group members and officemates; Xingyuan Chen, Felipe De Barros, Gretchen Miller, Melanie Hahn, Yi Liu, Hang Bai, Wolfgang Nowak, Zepu Zhang, Matthew Over, Yarong Yang, Bob Borreli, Denia Djokic and Erwan Bouvier. I would like to thank Xingyuan Chen with whom I worked closely for a year; her deep knowledge in statistics often prevented me from rushing in the wrong direction. I am also lucky to have many friends in Berkeley during my Ph.D. time including Lance Kim, Edward Branford, Sven Chilton, Christina Legget, Kurt Terrani, Koichiro Ito and Mariko Nakamura. Climbing with Melanie Hahn and Stephanie Gindlesperger was very helpful in maintaining my peace of mind in the last year of my Ph.D. I would also like to give special thanks to the student advisor in nuclear engineering Lisa Zemmerman, who helped me with funding and administrative issues. Dr. Daniel Hawkes has helped me to edit papers and dissertation chapters and provided me many tips and advises on English writing.

In the course of my life, I am very grateful for many people who inspired me to purse research career and encouraged me to study in the U.S. including Dr. Yung Liu, Dr. Shiu-Wing Tam, Dr. Nachappa Gopalsami at Argonne National Laboratory, Dr. Tetsujii Yamaguchi and Dr. Shinichi Nakayama at Japanese Atomic Energy Agency, and Professor Hirotake Moriyama, Professor Ikuji Takagi, Professor Takayuki Sasaki, Professor Hajimu Yamana, Professor Tsuyoshi Mishima and Professor Ikuo Kanno at Kyoto University.

Finally, I would like to thank my family for giving me much support and allowing me to be a student until this age.

Chapter 1

Introduction

1.1 Motivation

Groundwater contamination has been one of the environmental concerns central to the nuclear industry since the early 1950s. During the years of the Cold War, effluent from nuclear weapon production contaminated groundwater at various production sites in the United States. Accidental tritium release has also been reported at U.S. nuclear power plants. Although the nuclear industry is currently one of the most highly regulated fields, there still remain many contaminated sites, and the potential contamination from the nuclear waste repositories is a major public concern.

Currently, according to the regulations set by the U.S. Nuclear Regulatory Commission (NRC) (http://www.nrc.gov/), any construction of a nuclear power plant or waste repository requires a rigorous assessment of potential groundwater contamination. Sites contaminated by nuclear material also require assessments to evaluate remediation strategies [USEPA, 2001]. The critical component in these assessments is predicting groundwater flow and contaminant (e.g., radionuclide) transport to determine any future consequences in human health or the regional ecological system. To account for uncertainty, probabilistic risk assessments–originally developed for nuclear power plants–have become a preferred choice for assessing groundwater contamination [USEPA, 2001]. However, in contrast to nuclear power plants, in which all the components are known and a failure rate of the components is determined from repeated experiments, groundwater contamination assessment is subject to much larger uncertainty, since the heterogeneity of the natural subsurface environment makes it impossible to create a complete map of hydrologic properties over the domain.

Stochastic hydrologeology has been developed in the past three decades [Dagan, 1989; Rubin, 2003] to address uncertainty caused by spatially variable properties in flow and transport modeling. Since stochastic modeling yields the consequences of flow and transport in a probabilistic manner (e.g., probability of concentration at a certain location being higher than a certain value given available information), the modeling results can be seamlessly incorporated within the framework of probabilistic risk assessment [Andricevic and Cvetkovic, 1996; Maxwell and Kastenberg,1999; Maxwell et al., 1999]. From data to prediction, the stochastic hydrogeology framework consists of two parts: forward modeling and inverse modeling.

The goal of forward modeling is to describe flow and transport processes based on known physical models and known parameter distributions that account for parameter uncertainty. In stochastic hydrologeology, based on geostatistics originally developed for mining operations [Kitanidis, 1997], a spatially variable hydrogeologic property is modeled as a random variable, or more generally as a random field (i.e., a collection of jointly dependent random variables, indexed by spatial position). The random field is described by a probabilistic model based on various control parameters, called structural parameters, such as trend, variance, and scale of the field, using a two-point correlation structure called a covariance model. Although forward models sometimes have a closed-form analytical solution, often modelers use numerical simulations for complex processes. In stochastic simulations, we first generate a set of parameters from the distributions of the parameters, then generate multiple random fields for a given set of parameters, and finally calculate flow and transport for each field to determine possible consequences [Maxwell et al., 1999; Rubin, 2003]. The computed ensemble of possible consequences yields a probability distribution of the possible consequences or probability of the possible consequence exceeding the regulatory standard, which is used for assessing risks and making decisions.

The goal of inverse modeling is to estimate the parameters or parameter distributions used for the forward simulations, based on field observations. Traditionally, in geostatistics, the variogram approach was used for estimating the structural parameters from point-scale direct measurements. Recently, a Bayesian model-based geostatistics was developed to infer a joint distribution of structural parameters [Diggle and Ribeiro, 2003]. A substantial effort has been invested in using large-scale data–observations affected by a large area of a field such as, for example, pumping-test or tracer-test data. With large-scale data, each observation is a function of the entire field, so that it is difficult to characterize local features of the field from such data. Although inverting large-scale data to map a heterogeneous field is the same concept employed in medical imaging, the data density in hydrogeology is much lower, thus making it impossible to obtain a deterministic picture of the entire field. In hydrogeology, therefore there have been many different methods proposed for inverting large-scale data in the stochastic modeling framework [Kitanidis, 1986,1995; Carrera and Neuman, 1986a, 1986b; Rubin and Dagan, 1987a, 1987b; Rama Rao et al., 1995; Rubin, 2003; Bellin and Rubin, 2004; Hernandez et al., 2006].

In previous work, these two modeling schemes, i.e., forward and inverse modeling, have often been considered and developed separately, which frequently has led to discontinuities between them. For example, most inverse modeling methods yield only a (so called) best estimate and asymptotic confidence bounds of the parameters, without considering connection to the forward simulations. It is known that the best estimate of the field is too smooth and unrealistic to represent a real process [Rubin et al., 2010]. Moreover, the distributions of parameters are often assumed or made up. With such discontinuity, the information contained in the data, especially uncertainty associated with the data, cannot be fully transferred to predictive simulations. This discontinuity becomes even more apparent when we try to integrate (or assimilate) datasets of different types. A typical field characterization project includes two or more different complementary sources of information, such as core analyses, pumping tests, tracer tests, lithology information, or flowmeter tests. They have different accuracy, scale, and relationship to the target variable. Each data type is often analyzed separately and differently using a data-specific inverse modeling method, which makes integrating all the datasets difficult.

Following upon the previous work and current research needs, the research questions addressed in this dissertation are the following:

1. How can we transfer uncertainty associated with the data to prediction in the stochastic hydrology framework?

2. How can we integrate different types of data consistently?
3. Given that we integrate different types of data consistently, can we determine which data type or which combination of data types are more beneficial for characterization?

Answering these questions requires an inverse modeling method that can integrate various types of data in a consistent manner and transfer uncertainty associated with the data into predictions. The method must be tested and validated not only in synthetic studies but also in real subsurface characterizations.

1.2 Scope of the Dissertation

Following the motivation discussed in the previous section, this dissertation presents a Bayesian geostatistical data assimilation method and its application to real data taken from subsurface characterization. The following chapters include an individual introduction, methodology, results, and summary, organized as follows.

Chapter 2 introduces the Bayesian geostatistical data assimilation method, called the method of anchored distributions, which is a general inverse modeling method that integrates various types of data for characterizing a spatially variable hydrological property directly within the stochastic modeling framework. The central element of this method is a new concept called *an anchor*, which is a device for localizing large-scale data. Anchors are used to convert large-scale, indirect data into local distributions of the target variables. The goal of the inversion is to determine a joint distribution of the anchors and structural parameters, conditioned on all of the field data. The structural parameters describe global trends of the field, whereas the anchors capture local heterogeneities. Following the inversion, the joint distribution of anchors and structural parameters can be directly used to generate random fields of hydrogeological properties and compute possible outcomes of flow and transport processes. This method enables us to remove the discontinuity between data analysis and transport predictions. The method is tested through a synthetic study to characterize a two-dimensional transmissivity field based on direct measurements of transmissivity and hydraulic head measurements from natural gradient flow.

Chapter 3 describes an application of this method to aquifer tests at the Integrated Field Research Challenge (IFRC) site at the Hanford 300 Area, Washington. The interdisciplinary Hanford IFRC project explores the complex behavior of subsurface uranium transport, involving hydrogeology, geochemistry, and microbiology. By having several hypotheses to explain the persistence of uranium plume at the site based on geochemical and microbial reactions, the project intends to enhance basic understanding of the subsurface transport mechanism of uranium. Three-dimensional hydrogeological characterization is crucial for understanding uranium transport in a heterogeneous subsurface environment. The main goal in Chapter 3 is to assimilate large-scale depth-averaged data from constant-rate injection tests with small-scale depth-discrete data from electromagnetic borehole flowmeter (EBF) tests for characterizing the three-dimensional hydraulic conductivity field. The method is first applied to determine a joint posterior distribution of two-dimensional geostatistical parameters and local log-transmissivities at multiple locations. After combining the distribution of transmissivities with depth-discrete relative-conductivity profiles from the EBF data, the three-dimensional structural parameters of the log-conductivity field are inferred.

In Chapter 4, the data assimilation is expanded to include two more datasets at the IFRC site: lithology profiles and tracer tests. Sequential data assimilation is explored to integrate different types of datasets in a consistent manner and to reduce the computational cost required in assimilation. After assimilating different combination of the datasets, the inversion results are compared, particularly focusing on the impact of lithology profiles and tracer tests. This comparison utilizes different metrics based on breakthrough curves observed during the tracer test not used in the inversion, such as mean arrival time, peak concentration, and peak time. The purpose of this comparison is to evaluate several possible assimilation strategies, and investigate how and how much different datasets contribute to improving the characterization, which will be useful in guiding further characterization efforts at the site and also in future characterization projects at other sites.

Finally, Chapter 5 summarizes the main findings and conclusions in the previous chapters, and provides the overall conclusion, including answering the research questions raised in this chapter section. Appendices at the end of this dissertation include detailed derivations and discussions of the algorithms used in this dissertation.

Chapter 2

A Bayesian Approach for Inverse Modeling; Application of the Markov-Chain Monte-Carlo Sampling Method

2.1 Introduction

This chapter presents a new approach for inverse modeling called Method of Anchored Distributions (MAD). MAD is a geostatistical inversion method for estimating the joint distribution of parameters that control spatially variable properties. MAD addresses several of the main challenges inherent to inverse modeling. These challenges fall into two broad categories: data assimilation and modularity.

Data assimilation in inverse modeling involves using multiple and complementary types of data as sources of information for the target variable, which is a spatially variable property. In hydrogeological applications, for example, one may be interested in mapping a three-dimensional heterogeneous hydraulic conductivity field [Kitanidis, 1986, 1995, 1997; Carrera and Neuman, 1986a, 1986b; Hernandez et al., 2006], using measurements of hydraulic head, measurements of concentrations and travel times in solute transport experiments [Bellin and Rubin, 2004], and measurements of geophysical attributes from geophysical surveys. Since each type of measurement is related to the target variable, i.e., the conductivity field, through different numerical or analytical models, the challenge is to combine the multiple sources of data into a coherent map of the target variable. Given that a stochastic approach has become more popular for flow and transport simulations, it is especially important to take into account the uncertainties associated with different types of measurements.

Modularity requires an inverse modeling approach that is not tied to any particular model or data type, which results in the needed flexibility to accommodate a wide range of models and

---

[1] This chapter is based on a published article in Water Resources Research, 2010 (Rubin, Y., Chen, X., Murakami, H., and Hahn, M. S.: A Bayesian approach for inverse modeling, data assimilation and conditional simulation of spatial random fields, Water Resour. Res, 46, W10523, doi:10.1029/2009WR008799, 2010.)

data types. Inverse methodologies and numerical simulations of physical processes have over the past few decades become highly intertwined in a way that limits their application. This is largely attributed to the increasing complexity of the processes being analyzed and of the computational techniques needed for their analysis. For example, inverse modeling in hydrogeology evolved from Theis type-curve matching [Theis, 1935] into modern studies that include complex and specialized elements, such as (1) adaptive and parallel computing techniques, (2) linearization or other approximation approaches, and (3) complex optimization, sampling or search algorithms. The range of skills needed for implementing these elements led researchers to build an inversion procedure based on their own preferred numerical schemes. This resulted in limiting the potential for expanding the range of applications and data types beyond the original application. Modularity is a strategy for alleviating this difficulty, by pursuing a model-independent inverse modeling framework.

This chapter, through presentation of the MAD concept, explores these issues using a Bayesian framework. A theoretical approach is developed, and this approach is demonstrated with a synthetic case study.

2.2 Data Classification and Definition of Anchors

This section summarizes and then expands upon a few developments included in an unpublished manuscript by Zhang and Rubin [2008] and in a conference presentation by Zhang and Rubin [2007].

The MAD approach to inverse modeling is built upon two elements: (1) data classification and (2) a strategy for localizing nonlocal data. The localization strategy serves to create a unified approach for dealing with any type of data. These two elements are integrated into a Bayesian formulation for data assimilation. Integration of these two elements is done modularly, accommodating a wide variety of data types and physical processes that connect data to the target variables of interest. The rationale for formulating the inverse problem using a statistical formalism has been sufficiently discussed in the literature [Kitanidis, 1986; Rubin, 2003] and is not repeated here for the sake of brevity.

We consider a target variable as a spatial random process denoted by $Y(x)$, where x represents a single space coordinate. The entire field of Y is denoted by $\tilde{\boldsymbol{Y}}$. A realization of $\tilde{\boldsymbol{Y}}$ is denoted by $\tilde{\boldsymbol{y}}$. The data vector $\boldsymbol{z}$ contains multiple observations of multiple data types. Given data $\boldsymbol{z}$, the goal is to determine a conditional distribution of the field, $p(\tilde{\boldsymbol{y}} \mid \boldsymbol{z})$, from which we can generate random samples of the field $\tilde{\boldsymbol{Y}}$ for stochastic simulations of prediction.

The field $\tilde{\boldsymbol{Y}}$ is defined through a vector of model parameters $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. The $\boldsymbol{\theta}$ part of this vector includes a set of parameters designed to capture the global trends of $\tilde{\boldsymbol{Y}}$, such as the mean of the field and correlation structures. The $\boldsymbol{\vartheta}$ component of this vector consists of the anchored distributions. Anchored distributions, or anchors, in short, are devices used to capture the local effects that cannot be captured using the global parameters represented by $\boldsymbol{\theta}$. The anchors are given in the form of statistical distributions of Y at known or chosen locations.

The overall strategy for determining $p(\tilde{\mathbf{y}} \mid \boldsymbol{z})$ is to determine a joint conditional distribution of the model parameters $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z})$, which in turn allows us to generate multiple realizations of the field $\tilde{\boldsymbol{Y}}$ from $p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$. The way to determine the distribution $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z})$ should be both general and flexible enough to accommodate a wide range of formulations of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$, as well as a wide range of data types included in $\boldsymbol{z}$.

2.2.1 Data Classification

MAD classifies data into two types, i.e. Type-A data and Type-B data, based on the data's relationship to the target variable Y and support volume. We denote all available data by $\boldsymbol{z}$, Type-A data by $\boldsymbol{z}_a$, and Type-B data by $\boldsymbol{z}_b$. Type-A and Type-B data refer to on-site data. Information in the form of expert opinion or from geologically similar sites is treated differently and falls under the category of prior information.

Type-A data is small-scale local data such that the support volume of Type-A data is the same as the one of $\boldsymbol{Y}$. Each of Type-A data is a function of each of the point values $\boldsymbol{Y}$. When we have n_a measurements of Type-A data at locations $\boldsymbol{x}_a$, we can write a equation to connect $\boldsymbol{Y}(\boldsymbol{x}_a)$ and $\boldsymbol{z}_a$ as,

$$\boldsymbol{z}_a = \boldsymbol{y}(\boldsymbol{x}_a) + \boldsymbol{\epsilon}_a, \tag{2.1}$$

where $\boldsymbol{\epsilon}_a$ is a n_a-vector of zero-mean errors. $\boldsymbol{\epsilon}_a$ represents uncertainty associated with measurement errors, errors in regression or petrophysical relationships [Ezzedine et al., 1999]. In Equation (2.1), we assume that the support volume of Type-A measurements is the same as the one of the point-scale target variable Y. When $\boldsymbol{Y}$ is permeability, for example, Type-A data may include measurements of permeability obtained using permeameters, or predicted permeability from soil texture and grain-size distributions using petrophysical models [Ezzedine et al., 1999; Mavko et al., 1998].

Type-B data is large-scale nonlocal data, defined as a function of the field $\tilde{\boldsymbol{Y}}$, which is described by the following equation:

$$\boldsymbol{z}_b = \mathcal{M}(\tilde{\boldsymbol{y}}) + \boldsymbol{\epsilon}_b \tag{2.2}$$

where $\mathcal{M}$ is a known numerical or analytical model of the spatial field, and $\boldsymbol{\epsilon}_b$ is a n_b-vector of zero-mean errors. In hydrogeological applications, $\boldsymbol{z}_b$ may include data obtained from large-scale pumping tests, solute transport experiments, continuous observations of hydraulic heads, and geophysical surveys. The model $\mathcal{M}$ then includes mathematical models that seek to reproduce experiments corresponding to the data. When $\boldsymbol{z}_b$ includes data from multiple sources, $\mathcal{M}$ in Equation (2.2) should be viewed as a collective name for all the models that relate $\boldsymbol{z}_b$ to $\tilde{\boldsymbol{y}}$.

2.2.2 Concept of Anchors

Anchors $\boldsymbol{\vartheta}$ are given in the form of statistical distributions of Y at known or chosen locations, which establishes a connection between the unknown $\tilde{\boldsymbol{Y}}$ field and the data along with the structural parameters $\boldsymbol{\theta}$. Using anchors and structural parameters, we will be able to generate multiple fields $\tilde{\boldsymbol{Y}}$ that are conditioned on Type-A and Type-B data. MAD defines two types of anchors corresponding to its data classification, i.e., Type-A anchors $\boldsymbol{\vartheta}_a$ and Type-B anchors $\boldsymbol{\vartheta}_b$. We denote the entire set of anchors by $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_a, \boldsymbol{\vartheta}_b\}$.

Type-A anchors $\boldsymbol{\vartheta}_a$ are co-located with Type-A data such that we have one anchor per one Type-A measurement at the same location. A set of Type-A anchors is equivalent to $\boldsymbol{y}(\boldsymbol{x}_a)$ in Equation (2.1). Given $\boldsymbol{\vartheta}_a$, $\boldsymbol{z}_a$ follows the distribution of $\boldsymbol{\epsilon}_a$ through regression or petrophysical models.

Type-B anchors $\boldsymbol{\vartheta}_b$ capture information from Type-B data. Since a set of Type-B anchors has influence on Type-B data through the field $\tilde{\boldsymbol{Y}}$ and the physical models denoted by $\mathcal{M}$ in Equation (2.2), there may not be a one-to-one correspondence between Type-B anchors and Type-B data, which is different from Type-A data. We plant Type-B anchors at multiple locations $\boldsymbol{x}_b$ to capture the information contained in the Type-B data relevant to the field $\tilde{\boldsymbol{Y}}$. In other words, we transform the large-scale nonlocal Type-B data into a distribution of the local Y values, i.e. the Type-B anchors. MAD determine the distribution of $\boldsymbol{\vartheta}_b$ though the inversion described in Section 2.3.

Anchor placement is an important element of MAD: we would want to place the anchors such that they capture all the relevant information contained in the data, but at the same time we want to minimize the number of anchors, since the number of anchors is directly related to the computational burden. This is trivial for Type-A anchors, since Type-A anchors are co-located with Type-A data. However, this is a complex issue for Type-B anchors, since there is often a complex relationship between $\boldsymbol{z}_b$ and $\tilde{\boldsymbol{Y}}$.

2.3 Inversion in MAD Framework

As a starting point, let us consider $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z}_a, \boldsymbol{z}_b)$, the joint distribution of the model parameters, including structural parameters and anchors, conditioned on Type-A and Type-B data. Applying Bayes rule, we rewrite the posterior distribution:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z}_a, \boldsymbol{z}_b) &\propto p(\boldsymbol{z}_b \mid \boldsymbol{z}_a, \boldsymbol{\theta}, \boldsymbol{\vartheta}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z}_a), \\ &= p(\boldsymbol{z}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z}_a), \end{aligned} \tag{2.3}$$

where $\propto$ represents that the left-hand side of the equation is proportional to the right-hand side. $p(\boldsymbol{z}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ denotes the likelihood of the Type-B data, which is the key for relating the model parameters and Type-B data. $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z}_a)$ has a dual role of being the posterior probability given Type-A data, and the prior probability with respect to Type-B data. We drop $\boldsymbol{z}_a$ from conditioning whenever it is coupled with Type-A anchors $\boldsymbol{\vartheta}_a$, considering that Type-A anchors carry the information provided by $\boldsymbol{z}_a$. This also assumes that Type-A and Type-B are conditionally independent given $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$.

We can further factorize the second term in Equation (2.3):

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z}_a) &= p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_a, \boldsymbol{\theta}_b \mid \boldsymbol{z}_a), \\ &= p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_a) p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_a \mid \boldsymbol{z}_a), \\ &\propto p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_a) p(\boldsymbol{\vartheta}_a \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\vartheta}_a \mid \boldsymbol{z}_a), \end{aligned} \tag{2.4}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$ before having any data at the site. The distribution $p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_a)$ is the prior distribution of Type-B anchors with respect to Type-B data. The distribution $p(\boldsymbol{\vartheta}_a \mid \boldsymbol{\theta})$ is the likelihood of Type-A anchors given the structural parameters. In Equation (2.3) and (2.4), the critical components for inversions are the following: prior distribution $p(\boldsymbol{\theta})$, Type-A data inversion $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_a \mid \boldsymbol{z}_a)$ and likelihood of Type-B data $p(\boldsymbol{z}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$. In the following subsections, we discuss these elements individually.

2.3.1 Prior Distribution

The prior $p(\boldsymbol{\theta})$ summarizes the information that is available prior to including any measurements at the site. We may rely on literature or previous studies conducted at the geologically similar sites. There are a number of approaches for determining $p(\boldsymbol{\theta})$. Broad statistical perspective is provided by Kass and Wasserman [1996], Woodbury and Ulrych, [1993], Woodbury and Rubin [2000], Hou and Rubin [2005] and Diggle and Ribeiro [2006]. MAD can take any prior distribution, depending on the modelers discretion.

2.3.2 Type-A data Inversion

In this section, we determine the posterior distribution given Type-A data denoted by $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_a \mid \boldsymbol{z}_a)$. This constitutes our entire process when we have only Type-A data $\boldsymbol{z}_a$. The application of MAD in this case includes a sequence of two steps. First, we determine $p(\boldsymbol{\vartheta}_a \mid \boldsymbol{z}_a)$ according to Equation (2.1). Second, we infer the structural parameters as $p(\boldsymbol{\theta}_a \mid \boldsymbol{\vartheta}_a)$ on every possible $\boldsymbol{\vartheta}_a$ from $p(\boldsymbol{\vartheta}_a \mid \boldsymbol{z}_a)$. Since $\boldsymbol{\vartheta}_a$ represents the point Y values, we can use an analytical form for $p(\boldsymbol{\vartheta}_a \mid \boldsymbol{z}_a)$, often assuming that Y and $\boldsymbol{\vartheta}_a$ follow a multivariate normal distribution [Hoeksema and Kitandis, 1984; Kitanidis, 1986, Diggle and Ribeiro, 2006]. Diggle and Ribeiro [2006] used Box-Cox transform or a generalized linear model to deal with non-normal distributions.

2.3.3 Prior Distribution of Type-B Anchors

Prior to the Type-B data inversion, we need to place Type-B anchors and determine the prior distribution of Type-B anchors $p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_a)$. Unlike Type-A anchors, Type-B anchors do not have a one-to-one match to Type-B data. Instead, modelers are required to place anchors at selected locations. We discuss anchor placement in Section 2.3.5. Here we assume that we have already selected Type-B anchor locations $\boldsymbol{x}_b$.

After determining the locations $\boldsymbol{x}_b$, we determine the distribution of Type-B anchors. Since the anchors are a subset of the field $\tilde{\boldsymbol{Y}}$, Type-B anchors follow the same geostatistical model and covariance structure as $\tilde{\boldsymbol{Y}}$. For example, when the field is modeled as a multivariate Gaussian field, for given structural parameters and Type-A anchors $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}_a\}$, Type-B anchors follow the multivariate Gaussian distribution with the covariance structure specified by $\boldsymbol{\theta}$ and $\boldsymbol{x}_b$, and conditioning points of $\boldsymbol{\vartheta}_a$.

2.3.4 Likelihood Estimation

The likelihood $p(\boldsymbol{z}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ in Equation (2.3) connects the various Type-B data to $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. The likelihood can be determined either analytically under certain assumptions [Hoeksema and Kitanidis, 1984; Dagan, 1985; Rubin and Dagan, 1987a, 1987b] or by numerical simulations. MAD can implement both approaches.

An extensive body of literature is devoted to multidimensional density estimation, which is equivalent to likelihood estimation [Scott and Sain, 2005]. These strategies fall into two categories, parametric and nonparametric. The parametric approach assumes a certain distribution, leaving only the need to estimate a few parameters. In many cases, a multivariate normal distribution is adopted for the likelihood [Dagan, 1985; Carrera and Neuman, 1986a, 1986b; Hoeksema and Kitanidis, 1984; Kitanidis, 1986; Rubin and Dagan, 1987a, 1987b].

The nonparametric approach does not assume any distributions. The appeal of nonparametric methods lies in their ability to reveal structure in the data that might be missed by parametric methods. However, this advantage is associated with a heavy price tag: nonparametric methods are much more computationally demanding than their parametric counterparts. The MAD algorithm is flexible in its ability to employ both parametric and nonparametric methods. Scott and Sain [2005] and Newton and Raftery [1994] summarized several alternatives for nonparametric likelihood estimation. Many algorithms are publicly available, such as the algorithms described by Hayfield and Racine [2008], which are part of the R-Package [R Development Core Team, 2007].

For either parametric or nonparametric methods, MAD requires the modelers to select the most appropriate representation of the likelihood depending on the data. For example, when z_b is comprised of pressure-head measurements taken from a uniform-in-the-average flow in an aquifer domain characterized by a small variance in log conductivity, a multivariate normal likelihood function is reasonable, because the head can be expressed as a linear function of the log conductivity [Dagan, 1985].

MAD estimates the likelihood $p(z_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ using numerical simulations, as follows. For any given $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ sampled from the prior distribution $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_a)$, we generate multiple conditional realizations of fields $\tilde{\boldsymbol{Y}}$. With each realization, a forward simulation of the physical process, following Equation (2.2), provides a prediction of z_b in the form of $\tilde{z}_b$. In other words, z_b is viewed as a measured outcome from random process, while $\tilde{z}_b$ is one of many possible realizations given $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. After simulating for all the generated fields, we obtain the ensemble of $\tilde{z}_b$ given $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. MAD uses this ensemble of $\tilde{z}_b$ to estimate the density distribution $p(\tilde{z}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ and determine the density at z_b, which is equivalent to the likelihood $p(z_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$.

2.3.5 Type-B Anchor Placement

Anchor placement, i.e., the number and location of Type-B anchors, depends on various factors. When we generate fields for prediction, we assume that $p(\tilde{y} \mid z) = p(\tilde{y} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$, which means that conditioning on the data z is equivalent to conditioning on the anchors and structural parameters $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. Consider, for example, a local decline observed in water-table elevation. This decline could be controlled by the presence of local features such as high- and/or low-conductivity areas. We would want to capture these features by anchors. At the same time, we need to minimize the number of anchors; since the anchors are model parameters, and increasing the number of anchors leads to an increasing computational burden. We need to have a strategy that guides us to place the anchors such that a small number of anchors capture most of the important information in the data.

We propose a two-step approach for anchor placement. In the first step, the anchors are placed based on geological conditions, characterization goals, and data locations. The second step is a test to check whether the number of anchors is sufficient The first step requires modelers to consider physical principles and/or to rely on experience, while the second step is more mechanistic in nature.

For the first step, we may refer to relevant previous studies. Bellin et al. [1992] investigated solute transport in heterogeneous media numerically, and found that the spatial resolution of numerical models needs to be on the order of one quarter of the integral scale of the log conductivity for capturing the effect of spatial heterogeneity on solute transport. Castagna and Bellin [2009] suggested the same value, i.e., one quarter of the integral scale (i.e. characteristic length of heterogeneity determined by an integral of the two-point correlation over the distance [0, ∞]), in relation to crosshole tomography. Although the true integral scale is usually not known a priori, we may find reliable prior information from field studies conducted in similar formations [Scheibe and Freyberg, 1995; Hubbard et al., 1999; Rubin, 2003, Ch. 2; Sun et al., 2008; Ritzi et al., 2004; Ramanthan et al., 2008; Rubin et al., 2006].

Placing anchors at every quarter of the integral scale, however, can lead to too many anchors in a practical domain. When the number of Type-B data is small, there is not enough information to support such a large number of anchors. It would also be obvious that the anchors far from the observation locations may not contribute to estimation. To localize Type-B information, we need to identify the locales that are the most sensitive to Type-B data.

An effective strategy for identifying such locations is sensitivity analysis. Castagna and Bellin [2009] used sensitivity analysis for such a purpose in the context of hydraulic tomography. Vasco et al. [2000] used sensitivity analysis in the context of tracer tomography. Both studies indicated that certain locations, e.g., near the injection wells and observation wells in pumping and tracer tests, are much more sensitive to Type-B data than others. Such locations are prime targets for placing anchors.

We may note that Type-B anchors are useful only when Type-B data includes information regarding local heterogeneity. Some types of data do not have such local information. For example, the spatial moments of small solute plumes depend very much on local effects, whereas the spatial moments of large solute plumes depend only on the global parameters of spatial variability [Rubin et al., 1999; Rubin et al., 2003]. In another example, Sanchez-Vila et al. [1999] showed that the large-time drawdown response in a pumping test led to the global mean of log transmissivity in the field, regardless of the observation well locations. When Type-B data does not contain local information, the inverse modeling should be limited to identifying structural parameters [Copty et al., 2008].

Locations far from the observation wells would not require many anchors, since there appears little useful information to be gleaned from those locations. In this case, a dense grid of anchors could be placed where prediction accuracy is critical, while a low-density grid could be used for the rest of the domain. The high-density portion of the grid would be effective for capturing local features, whereas the low-density grid would be useful for estimating the global-trend parameters. In another example, anchors could be placed in the locations that are most beneficial for predictions. This is the idea of network design pursued by Janssen et al. [2008] and Tiedeman et al. [2003, 2004].

In the second step, we place additional anchors to test whether the number of anchors is sufficient. We repeat the inversion using this expanded anchor set, including a few additional anchors placed at potentially valuable locations. When the marginal distributions of the original anchors do not change with the additional anchors, we might conclude that the anchors capture all the information contained in z_b, and that the additional anchors are redundant in terms of information. We may also compare the distribution of predicted variables before and after including the additional anchors.

2.3.6 Multiple Type-B Data and Sequential Updating

Since increasing the dimension of z_b increases the computational burden of the likelihood estimation significantly, we might want to divide the vector z_b into L segments as $z_b = \{z_{b(1)}, z_{b(2)}, \ldots, z_{b(L)}\}$, where $z_{b(i)} (i = 1, 2, \ldots, L)$ can be a single data value or a vector of multiple data values. We can decompose the likelihood into each segment as,

$$\begin{aligned} p(z_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}) &= p(z_{b(1)}, z_{b(2)}, \ldots, z_{b(L)} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}), \\ &= \prod_{i=1}^{L} p(z_{b(i)} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}). \end{aligned} \tag{2.5}$$

In Equation (2.5), we assume that the data segments $z_1, z_2, \ldots, z_L$ are conditionally independent for a given $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$, since we assume that $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ contains information equivalent to the data. This equality strictly holds when the data segments are completely independent of each other for example, when the data locations are beyond the zone of influence in the flow or transport process. It approximately holds when the data segments are only weakly correlated, such as with different types of data at the same site. As Hou and Rubin [2005] pointed out, assuming independence leads to higher entropy and makes the estimation less informative.

Using the same assumption, i.e. independence among the data segments, we can also sequentially assimilate Type-B data as,

$$p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_a, z_b) \quad \propto \quad p(z_{b(L)} \mid \boldsymbol{\vartheta}, \boldsymbol{\theta}) p(\boldsymbol{\vartheta}, \boldsymbol{\theta} \mid z_a, z_{b(1)}, z_{b(2)}, \ldots, z_{b(L-1)}), \tag{2.6}$$

where $p(\boldsymbol{\vartheta}, \boldsymbol{\theta} \mid z_a, z_{b(1)}, z_{b(2)}, \ldots, z_{b(L-1)})$ is the posterior distribution with respect to 1 to $L-1$ datasets as well as the prior distribution with respect to $z_{b(L)}$, and $p(z_{b(L)} \mid \boldsymbol{\vartheta}, \boldsymbol{\theta})$ is the likelihood of $z_{b(L)}$. This suggests that we can update the posterior distribution every time a new dataset arrives.

These two approaches, Equation (2.5) and (2.6), which yield the same result, since they are based on the same assumptions, give us great flexibility. For example, Equation (2.5), is suitable when all the segments of datasets are collected or simulated at the same time, whereas Equation (2.6) is useful when the characterization projects involve multiple experiments or tests separately over a long time frame.

2.3.7 A Flow Chart for MAD

Figure 2.1 shows a diagram of MAD with the same notation provided in Equations (2.3) and (2.4), particularly emphasizing the modular structure of the MAD approach. There are three blocks in MAD, labeled Blocks I, II and III, respectively, with two auxiliary blocks, labeled Auxiliary Blocks A and B.

Block I is the Type-A data module, which includes the steps described in Section 2.3.2. There is no particular approach to modeling the prior or Type-A data that is hard-wired into MAD. The output of Block I is the conditional distribution $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_a)$, which is the posterior distribution of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ with respect to z_a. This provides the prior distribution for assimilating Type-B data in Block II.

Block II is the Type-B data module. It incorporates Type-B data through the likelihood function $p(z_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$, as described in Section 2.3.4. When combined with Block I product $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_a)$, it yields the posterior $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_a, z_b)$. The model parameters $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ are linked to the observations z_b through the $\tilde{\boldsymbol{Y}}$ field and the forward model $\mathcal{M}$, according to Equation (2.2). MAD may use several forward models simultaneously or sequentially for various types of Type-B data, as indicated by the vertical bars linked to forward simulations.

Block III is the prediction block. It covers the post-inversion analyses needed for predicting a future process of interest. It can connect directly to Block I in the absence of Type-B data.

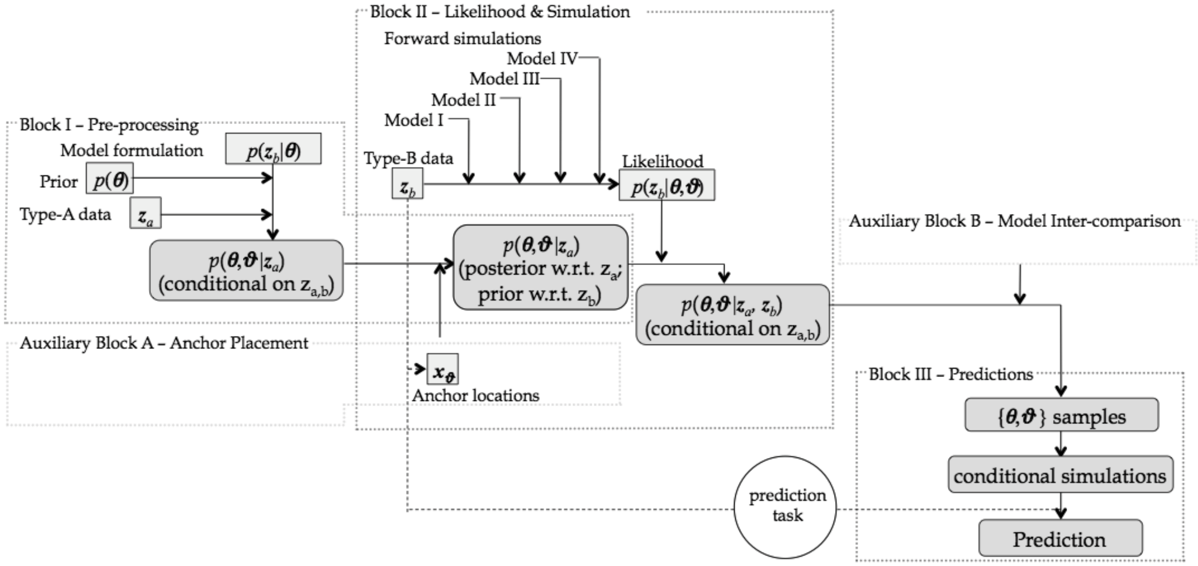


Figure 2.1: Graphical Representation of the MAD approach. The data at the site and prior information are integrated into MAD for prediction (from left to right). Parameter uncertainty is characterized by the posterior distributions at different stages. Anchors here refer to both Type-A and Type-B anchors (with variations depending on the application). Forward simulations in Block II refer to simulations needed for estimating the likelihood.

The forward simulation step in Block III can guide the selection of anchor locations in Block II by evaluating alternative anchor placement schemes. As with the other blocks, Block III could be linked with a wide range of forward simulation codes and computational techniques.

The prediction block includes multiple realizations of the random fields conditioned on every possible set of the parameters $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ sampled from the joint distribution of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. This leads to a complete characterization of parameter uncertainties associated with the model. It creates a stark contrast to commonly used optimization-based approaches such as maximum-likelihood (ML) or maximum a-posteriori (MAP) approaches [Kitanidis, 1995; Ramarao et al., 1995;McLaughlin and Townley, 1996], both of which yield only the best estimates and asymptotic bounds of the parameters under the normality assumption.

The auxiliary blocks include Block A, dedicated to anchor placement analysis, and Block B, dedicated to validating and testing the sufficiency of the number of anchors. They are not considered as core blocks, because they contain elective procedures that are not absolutely necessary for a complete application of MAD.

2.4 Markov-Chain Monte-Carlo Sampling

Practical application of MAD requires implementing a proper sampling algorithm for the parameter vector $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. Since the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z)$ may not be a standard analytical distribution and the parameter dimension is usually quite high, it is impossible to determine the posterior distribution directly. Instead, we need a sampling algorithm to sample parameter vectors based on the prior and likelihood, such that the generated samples mimic samples drawn from the posterior distribution. The posterior distribution is constructed by taking the histogram of the generated samples.

Originally Zhang and Rubin [2009] used the weighted resampling algorithm [Section 3.5.2 in Gamerman and Lopes, 2006]. The algorithm proceeds the following:
(1) Generates independent and identically-distributed samples of the parameter vector from the prior distribution (we call these samples ”prior samples”)
(2) Calculates the likelihood of each prior sample generated in (1), using the ensemble of $\tilde{z}$
(3) Resamples the original samples considering the likelihood to be a probability of each sample (we call these samples ”posterior samples”)
Since the resampled parameter sets follow the posterior distribution, we may create the posterior distribution by estimating density of the posterior samples or make predictions using the posterior samples as inputs.

This approach, however, cannot easily cover the higher-dimensional parameter spaces, as we increase the number of anchors and the dimension of the parameter vector. In practice, we may see the following problems. The first problem is that we have prior samples that do not contribute to the posterior at all. Although MAD intends to capture the lower probability event, some of the prior samples generated from the prior have extremely low likelihood relative to others (e.g., less than 10^{-30}). Since they would never appear in resampling, it is waste of computational resources to have many samples in such a low-likelihood region. The second problem is that we may not have enough samples to resolve the high-likelihood region. As the posterior distribution becomes narrower with more information, only a few samples are left with reasonable probability to be resampled, which makes it impossible to quantify the uncertainty.

In this section, we consider implementing the Markov-Chain Monte-Carlo (MCMC) sampling method as a possible solution for these problems. MCMC is one of the most popular sampling method in Bayesian inference for high-dimensional parameter vectors. In fact, for high-dimensional parameter problems, the MCMC method is the only known general method to provide a solution within a reasonable time and computation [Andrieu et all, 2003].

To mitigate the problems mentioned above, we want to sample less in the low-likelihood region and more in the high-likelihood region. MCMC is iterative and adaptive, so that the chain can spend more time in the most important region. In addition, while the weighted resampling algorithm resamples the prior samples of the entire parameter vector at once, MCMC can update each element or each block of the parameter vector separately at each time, enabling us to investigate the high-likelihood region effectively [Andrieu et all, 2003].

However, MCMC also has some disadvantages compared to the weighted resampling algorithm. First, MCMC is more difficult to implement, especially in a parallel computing environment, because the adaptive nature of MCMC requires us to implement all the procedures (e.g., sampling, numerical simulations, and likelihood estimations) simultaneously on one computer platform. On the other hand, the weighted resampling method allows us to separate the numerical forward simulations completely from the inference part, simplifying parallelization significantly and giving us more flexibility to test different likelihood estimation strategies in post-processing without repeating the numerical simulations. Second, in MCMC, scaling of parallelization is strictly limited to the number of simulations at each iteration step, since the next sample is generated according to the likelihood estimation at the current sample. In the weighted resampling algorithm, we can use as many processors as available on the platform, since we can perform the numerical simulations all at once.

The above discussion suggests that we must select a sampling algorithm based on available computational platforms as well as on progress of the research project. For earlier stages of the project, e.g., while testing different likelihood estimation strategies, the weighted resampling algorithm appears more appropriate. This is also the case when the available computational platform has a large number of processors (e.g., on the order of 10,000). When we have a computer with several to1,000 processors, and a properly fixed likelihood estimation strategy, MCMC is more advantageous. MAD is modular and flexible, so that it can accommodate either method.

2.4.1 Block Metropolis-Hastings Algorithm

Among several variations of the MCMC algorithms, we use the block Metropolis-Hastings (MH) algorithm [Andreiu et al., 2003]. This algorithm, one of the most flexible among many variations of MCMC, allows us to use a nonanalytical form of the likelihood function.

Here we state the general algorithm for data $\boldsymbol{z}$ and a parameter vector $\boldsymbol{\alpha}$, which is partitioned into J blocks as $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_J\}$. The goal is to sample from the posterior distribution $p(\boldsymbol{\alpha} \mid \boldsymbol{z})$. One block may represent one parameter or may contain several parameters. We denote the parameter vector at i-th iteration by $\boldsymbol{\alpha}^{(i)}$. We define the parameter vector without j-th block and with the first $(j-1)$ blocks already updated as $\boldsymbol{\alpha}_{-j}^{(i)} = \{\boldsymbol{\alpha}_1^{(i)}, .., \boldsymbol{\alpha}_{j-1}^{(i)}, \boldsymbol{\alpha}_{j+1}^{(i-1)}, .., \boldsymbol{\alpha}_J^{(i-1)}\}$.

The algorithm for generating N samples $(i = 0, 1, \ldots, N-1)$ is the following:

1. Initialize $\boldsymbol{\alpha}^{(0)}$

2. For $i = 1$ to $N-1$,

For $j = 1$ to J,

In order to draw j-th block in i-th iteration, denoted by $\boldsymbol{\alpha}_j^{(i)}$,

(1) Draw a potential value or potential values for $\boldsymbol{\alpha}_j$, denoted by $\boldsymbol{\alpha}_j^*$ from the proposal distribution, $q(\boldsymbol{\alpha}_j \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z})$: $\boldsymbol{\alpha}_j^* \sim q(\boldsymbol{\alpha}_j \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z})$

(2) Calculate the acceptance ratio r as,

$$r = \frac{p(\boldsymbol{\alpha}_j^{(i)} \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z})}{p(\boldsymbol{\alpha}_j^* \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z})} \frac{q(\boldsymbol{\alpha}_j^* \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z})}{q(\boldsymbol{\alpha}_j^{(i)} \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z})} \tag{2.7}$$

(3) Draw $u \sim Unif[0, 1]$.

(4) Determine $\boldsymbol{\alpha}_j^{(i)}$ such that:

if $u > \min[1, r]$

$$\boldsymbol{\alpha}_j^{(i)} = \boldsymbol{\alpha}_j^{(i)}, \tag{2.8}$$

else

$$\boldsymbol{\alpha}_j^{(i)} = \boldsymbol{\alpha}_j^*. \tag{2.9}$$

In other words, MH algorithm samples each parameter block $\boldsymbol{\alpha}_j^{(i)}$ based on the conditional distribution of $\boldsymbol{\alpha}_j^{(i)}$ given the data and the other parameters in the current iteration, denoted by $p(\boldsymbol{\alpha}_j \mid$

$\boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z})$, and the proposal distribution, denoted by $q(\boldsymbol{\alpha}_j^* \mid \boldsymbol{\alpha}^{(i)}, \boldsymbol{z})$. As a result, we obtain N sets of parameter vectors $\{\boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(N-1)}\}$ that follow the joint posterior distribution $p(\boldsymbol{\alpha} \mid \boldsymbol{z})$.

The convergence, i.e. when to stop sampling or whether N is large enough, is always an issue for MCMC. The common practice is to generate multiple chains and compare them; we stop sampling when the difference among the chains becomes insignificant [Gamerman and Lopes, 2006].

2.4.2 MH Algorithm for MAD

In MAD, the goal is to determine the posterior joint distribution $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z})$. In applying the MH algorithm, we first need to consider how to separate the parameter vector into multiple blocks. According to Andrieu et al. [2003], it is better to group the most strongly correlated parameters together, and smaller blocks are better for faster convergence. In MAD, this suggests that we should group adjacent anchors together, since they are strongly correlated due to the geostatistical covariance structure. To evaluate the acceptance ratio at every iteration, we apply Bayes rule in the same manner as Equation (2.3) to modify the acceptance ratio in Equation (2.7):

$$r = \frac{p(\boldsymbol{z}_b \mid \boldsymbol{\alpha}_j^*, \boldsymbol{\alpha}_{-j}^{(i)}) p(\boldsymbol{\alpha}_j^{(i)} \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \boldsymbol{\alpha}_j^{(i)}, \boldsymbol{\alpha}_{-j}^{(i)}) p(\boldsymbol{\alpha}_j^* \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z}_a)} \frac{q(\boldsymbol{\alpha}_j^* \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z}_a)}{q(\boldsymbol{\alpha}_j^{(i)} \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z}_a)} \tag{2.10}$$

In Equation (2.10), we may have some analytical form for the distribution of structural parameters and anchors conditioned on Type-A data, denoted by $p(\boldsymbol{\alpha}_j \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z}_a)$, as is discussed in Section 2.3.2. We can consider $p(\boldsymbol{\alpha}_j \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z}_a)$ as the prior distribution with respect to Type-B data. To assimilate Type-B data $\boldsymbol{z}_b$, every iteration requires two likelihood estimations, i.e. one for the numerator $p(\boldsymbol{z}_b \mid \boldsymbol{\alpha}_j^*, \boldsymbol{\alpha}_{-j}^{(i)})$ and one for the denominator $p(\boldsymbol{z}_b \mid \boldsymbol{\alpha}_j^{(i)}, \boldsymbol{\alpha}_{-j}^{(i)})$ in Equation (2.10).

There are a variety of approaches for determining the proposal distribution [Andrieu et al.,2003]. The simplest form is independent sampling, which has a proposal distribution independent of the current parameter vector. The prior distribution is commonly used as the proposal distribution. When we sample from the prior distribution for Type-B data $p(\boldsymbol{\alpha}_j \mid \boldsymbol{\alpha}_{-j}^{(i)}, \boldsymbol{z}_a)$, the acceptance ratio can be simplified as:

$$r = \frac{p(\boldsymbol{z}_b \mid \boldsymbol{\alpha}_j^{(i)}, \boldsymbol{\alpha}_{-j}^{(i)})}{p(\boldsymbol{z}_b \mid \boldsymbol{\alpha}_j^*, \boldsymbol{\alpha}_{-j}^{(i)})}. \tag{2.11}$$

This means that the acceptance ratio is the ratio of two likelihoods conditioned either on the current parameter vector $\boldsymbol{\alpha}_j^{(i)}$ or on the parameter vector including the proposed block $\boldsymbol{\alpha}_j^*$.

2.5 Case Study

The goal of this case study was to characterize a two-dimensional transmissivity field using a sparse network of Type-A and Type-B data. We used one set of the randomly generated transmissivity fields and computed the hydraulic head field as our baseline case. We selected transmissivities and hydraulic heads at selected locations as Type-A and Type-B data, respectively.

2.5.1 Methods

The target variable is the natural log-transform of the transmissivity (T) denoted by $Y = \ln T$. Let $Y(x)$ be a natural-log transmissivity at a single location $x = (x_1, x_2)$ in the two-dimensional domain. We assume that a vector $\boldsymbol{Y}$, containing Y at multiple locations $\boldsymbol{x}$, follows a multivariate Gaussian distribution with isotropic exponentially-decaying correlation. We define a structural parameter vector as $\boldsymbol{\theta} = \{\mu, \sigma^2, \phi\}$, including uniform mean μ, variance σ^2, and integral scale ϕ. This means that when we have two points with distance ξ, the two-point correlation follows an exponential function of distance, defined by $\exp(-\xi/\phi)$.

Information available for inversion includes Type-A data in the form of direct Y measurements, and Type-B data in the form of head measurements. The data vector for inversion is $\boldsymbol{z} = \{\boldsymbol{z}_a, \boldsymbol{z}_b\}$, where $\boldsymbol{z}_a(\boldsymbol{x}_a)$ includes n_a measurements of Y taken at the vector of locations $\boldsymbol{x}_a$ of length n_a, whereas $\boldsymbol{z}_b(\boldsymbol{x}_b)$ includes n_b head measurements taken at $\boldsymbol{x}_b$. In this case study, we assume that the Type-A data is error-free, so that Type-A anchors $\boldsymbol{\vartheta}_a$ is identical to $\boldsymbol{z}_a$, and can be given by $p(\boldsymbol{\vartheta}_a \mid \boldsymbol{z}_a) = \delta(\boldsymbol{\vartheta}_a - \boldsymbol{z}_a)$ with δ being Dirac's delta function.

The flow field is assumed at steady state and uniform on average, resulting from the constant head difference across two opposing boundaries and no-flow boundary conditions at the other two boundaries. The flow equation, given in Appendix A, connects the transmissivity field and the head field, which constitutes the forward model $\mathcal{M}$ in Equation (2.2).

We define the vector $\boldsymbol{\vartheta}$ as representing a set of anchors. The vector $\boldsymbol{\vartheta}$ is a vector of order $n_1 + n_2$, defined by $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_a, \boldsymbol{\vartheta}_b\}$, where $\boldsymbol{\vartheta}_a$ is a vector of order $n_1 = n_a$, corresponding to n_a measurements of Y. Since the anchors are a subset of the field, $p(\boldsymbol{\vartheta} \mid \boldsymbol{\theta})$ is a multivariate Gaussian distribution with mean μ and covariance $\sigma^2\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}}, \boldsymbol{x}_{\boldsymbol{\vartheta}})$, where $\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}}, \boldsymbol{x}_{\boldsymbol{\vartheta}})$ is an auto-correlation matrix as a function of ϕ and the locations of $\boldsymbol{\vartheta}, \boldsymbol{x}_{\boldsymbol{\vartheta}}$. The multivariate Gaussian distribution allows us to partition $\boldsymbol{\vartheta}$ into $\boldsymbol{\vartheta}_a$ and $\boldsymbol{\vartheta}_b$. Since $\boldsymbol{\vartheta}_a$ is fixed and known as $\boldsymbol{\vartheta}_a = \boldsymbol{z}_a$, $p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_a)$ is a multivariate Gaussian distribution with conditional mean $\mu_{\boldsymbol{\vartheta}_b|\boldsymbol{\vartheta}_a}$ and conditional covariance $\sigma^2\mathbf{R}_{\boldsymbol{\vartheta}_b|\boldsymbol{\vartheta}_a}$, where the mean and covariance conditioned on the anchors are defined as:

$$
\begin{aligned}
\mu_{\boldsymbol{\vartheta}_b|\boldsymbol{\vartheta}_a} &= \mu + \mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_b}, \boldsymbol{x}_{\boldsymbol{\vartheta}_a})\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_a}, \boldsymbol{x}_{\boldsymbol{\vartheta}_a})^{-1}(\boldsymbol{\vartheta}_a - \mu), \\
\mathbf{R}_{\boldsymbol{\vartheta}_b|\boldsymbol{\vartheta}_a} &= \mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_b}, \boldsymbol{x}_{\boldsymbol{\vartheta}_b}) - \mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_b}, \boldsymbol{x}_{\boldsymbol{\vartheta}_a})\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_a}, \boldsymbol{x}_{\boldsymbol{\vartheta}_a})^{-1}\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_a}, \boldsymbol{x}_{\boldsymbol{\vartheta}_b}),
\end{aligned}
\tag{2.12}
$$

where $\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_a}, \boldsymbol{x}_{\boldsymbol{\vartheta}_a})$ and $\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_b}, \boldsymbol{x}_{\boldsymbol{\vartheta}_b})$ are the auto-correlation matrices for $\boldsymbol{\vartheta}_a$ and $\boldsymbol{\vartheta}_b$, respectively. $\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_a}, \boldsymbol{x}_{\boldsymbol{\vartheta}_b})$ is a cross-correlation matrix between $\boldsymbol{\vartheta}_a$ and $\boldsymbol{\vartheta}_b$. $\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}_a}, \boldsymbol{x}_{\boldsymbol{\vartheta}_a})$ is a positive semidefinite matrix so that inverse of the matrix always exists.

With all the anchors $\boldsymbol{\vartheta}$, the distribution of $\boldsymbol{Y}$ conditioned on the structural parameters and anchors $p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ is a multivariate Gaussian distribution with conditional mean $\mu_{\boldsymbol{Y}|\boldsymbol{\vartheta}}$ and conditional covariance $\sigma^2\mathbf{R}_{\boldsymbol{Y}|\boldsymbol{\vartheta}}$, where the mean and covariance conditioned on the anchors are defined as

$$
\begin{aligned}
\mu_{\boldsymbol{Y}|\boldsymbol{\vartheta}} &= \mu + \mathbf{R}(\boldsymbol{x}, \boldsymbol{x}_{\boldsymbol{\vartheta}})\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}}, \boldsymbol{x}_{\boldsymbol{\vartheta}})^{-1}(\boldsymbol{\vartheta} - \mu), \\
\mathbf{R}_{\boldsymbol{Y}|\boldsymbol{\vartheta}} &= \mathbf{R}(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{R}(\boldsymbol{x}, \boldsymbol{x}_{\boldsymbol{\vartheta}})\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}}, \boldsymbol{x}_{\boldsymbol{\vartheta}})^{-1}\mathbf{R}(\boldsymbol{x}_{\boldsymbol{\vartheta}}, \boldsymbol{x}),
\end{aligned}
\tag{2.13}
$$

where $\mathbf{R}(\boldsymbol{x}, \boldsymbol{x})$ is the auto-correlation matrix for $\boldsymbol{Y}$, and $\mathbf{R}(\boldsymbol{x}, \boldsymbol{x}_{\boldsymbol{\vartheta}})$ is the cross-correlation matrix between $\boldsymbol{Y}$ and $\boldsymbol{\vartheta}$. Note that when we generate random fields $\tilde{\boldsymbol{Y}}$, each field follows the same distribution as $\boldsymbol{Y}$ such that $p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ is a multivariate gaussian distribution with conditional mean and covariance defined at all the grid locations.

For the inversion, we modify Equation (2.3) as:

$$p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_b \mid \boldsymbol{z}_a, \boldsymbol{z}_b) \propto p(\boldsymbol{\theta} \mid \boldsymbol{z}_a) p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\vartheta}_a, \boldsymbol{\theta}) p(\boldsymbol{z}_b \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}), \tag{2.14}$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{\vartheta}_a) = p(\boldsymbol{\theta} \mid \boldsymbol{z}_a)$, since $\boldsymbol{\vartheta}_a = \boldsymbol{z}_a$. We determine $p(\boldsymbol{\theta} \mid \boldsymbol{\vartheta}_a)$, using the model-based Bayesian geostatistics [Diggle and Ribeiro, 2006] (Appendix B). The prior and likelihood are specified in the following subsections.

Prior Distribution

For the prior distribution of the parameter vector, we follow Jeffreys' multi-parameter rule [1946] and Pericchi [1981]:

$$p(\boldsymbol{\theta}) = p(\mu, \sigma^2 \mid \phi) \propto \sigma^{-(d+1)} p(\phi), \tag{2.15}$$

where d is the dimension parameter, which is $d = 1$ for a stationary field in this case study.

The conditional prior for σ^2 and μ, $p(\mu, \sigma^2 \mid \phi)$, is noninformative with regard to σ^2 and μ, i.e., the prior densities of μ and $\log(\sigma^2)$ are both flat on $(-\infty, \infty)$. The unspecified component $p(\phi)$ is flexible. We used a uniform distribution in this case study. Diggle and Ribeiro [2006] adopted the same prior, and noted that although it is an improper prior because its integral over the parameter space is infinite, the substitution of this prior into Equation (2.4) leads to a proper distribution.

Specification of Likelihood

As mentioned in Section 2.3.4, Dagan [1985] showed through mathematical derivation that the hydraulic head values follow a multivariate Gaussian distribution in a uniform-on-average flow field when the variability of T is small. Although nonparametric density estimation is available as is discussed in Section 2.3.4 , the multivariate Gaussian likelihood estimation is computationally advantageous, as the dimension of the data increases. In this synthetic study, we estimated the likelihood assuming a multivariate Gaussian distribution. We checked using numerical simulations that this assumption reasonably holds, which is shown in Section 2.5.2.

For a given parameter set $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$, we generated multiple random fields $\tilde{\boldsymbol{y}}$, on each of which we simulated the hydraulic head field. It yielded an ensemble of possible head observations $\tilde{\boldsymbol{z}}$ for the given $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. Since we employed a parametric approach, we estimated the mean and covariance of this ensemble to estimate the density and compute the likelihood of each $p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$.

2.5.2 Results

Figure 2.2(a) shows the base field, i.e., the (assumed) true aquifer in this synthetic study, along with the locations of the Type-A and Type-B data. The base field was generated using a structural parameters parameter set $\{\mu, \sigma^2, \phi\} = \{8.7, 1.5, 15.0\}$. We obtained maximum likelihood estimates of the true parameters as $\{8.64, 1.49, 15.1\}$ by near-exhaustive sampling (one out of every five points) [GeoR package by Ribeiro and Diggle, 2001]. Figure 2.2(b) is the contour plot of the hydraulic head field simulated on the base field. The mean flow direction is parallel to the y-axis.

Figure 2.3 shows a number of different anchor configurations. We applied MAD to each of these different layouts to compare the different numbers and locations of the anchors. In order to investigate the impact of anchor locations, we created three layouts with eleven anchors; Figure 2.3(b) has a dense network of anchors in part of the domain (dense), Figure 2.3 (c) has a sparse network of anchors over the entire domain (sparse), and Figure 2.3(d) has a dense network of anchors in part of the domain and a few anchors outside (dense+sparse).

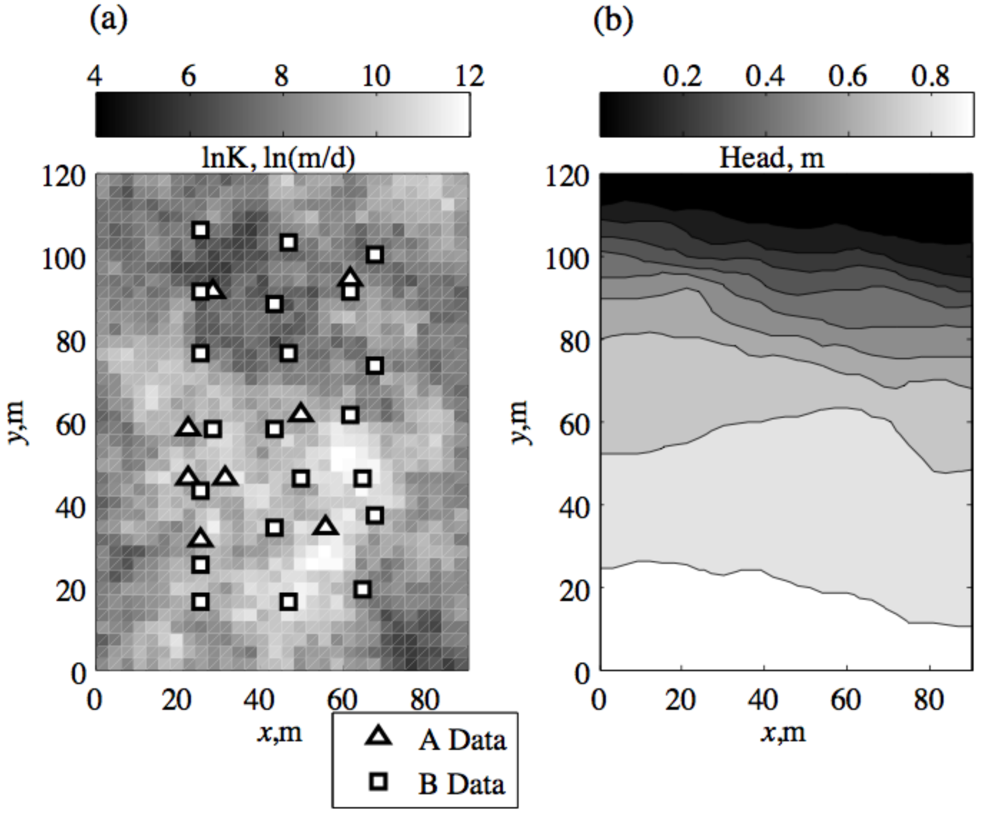


Figure 2.2: (a) The base transmissivity field for the synthetic study, (b) The hydraulic conductivity field.

We first checked our assumption on the likelihood, i.e., a multivariate Gaussian distribution. We generated 300 fields with 36 anchors (Figure 2.3f) and a geostatistical parameter set $\{\mu, \sigma^2, \phi\} = \{9.11, 3.2, 33.3\}$, which has a relatively high variance among the generated parameter sets conditioned on Type-A data. On each field, we simulated the hydraulic head at five of the Type-B data observation wells. We also estimated the density assuming the multivariate Gaussian distribution, using the mean and covariance of the 300 simulated heads at all the Type-B data locations. In Figure 2.4, the histogram of the hydraulic head at each of the five wells was compared to the marginal distribution of the estimated multivariate Gaussian density. Even though some of the wells are close to the domain boundaries, the estimated distributions have a reasonable fit to the simulated ones. The normality of the distribution at each well was also checked by the quantile-

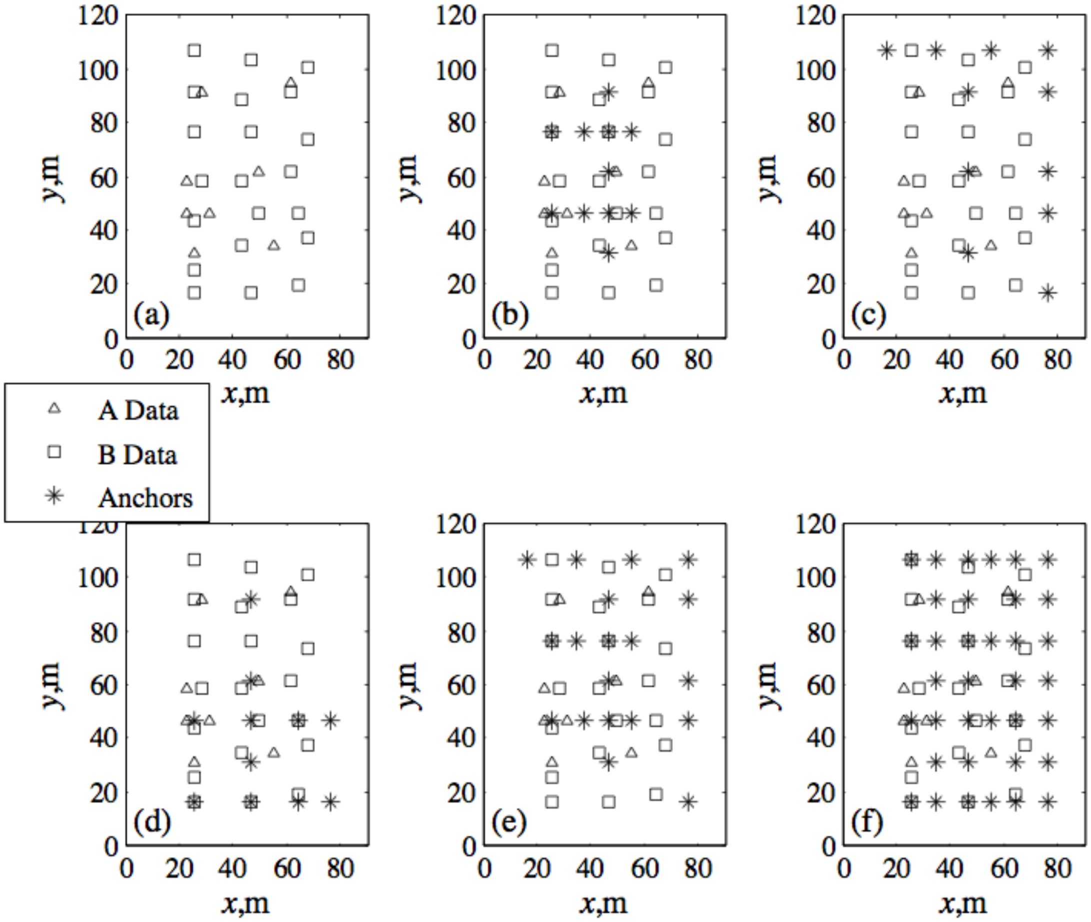


Figure 2.3: Various layouts showing the locations of Type A data, Type B data, and anchors. The number of anchors used in each configuration is as follows: (a) no anchors used; (b) 11 anchors; (c) 11 anchors; (d) 11 anchors; (e) 19 anchors; (f) 36 anchors.

quantile plot shown in Figure 2.5, where we can see that majority of the points are distributed on the straight lines. These results suggest that it is reasonable to use the parametric multivariate Gaussian density estimation for inversion in this case study.

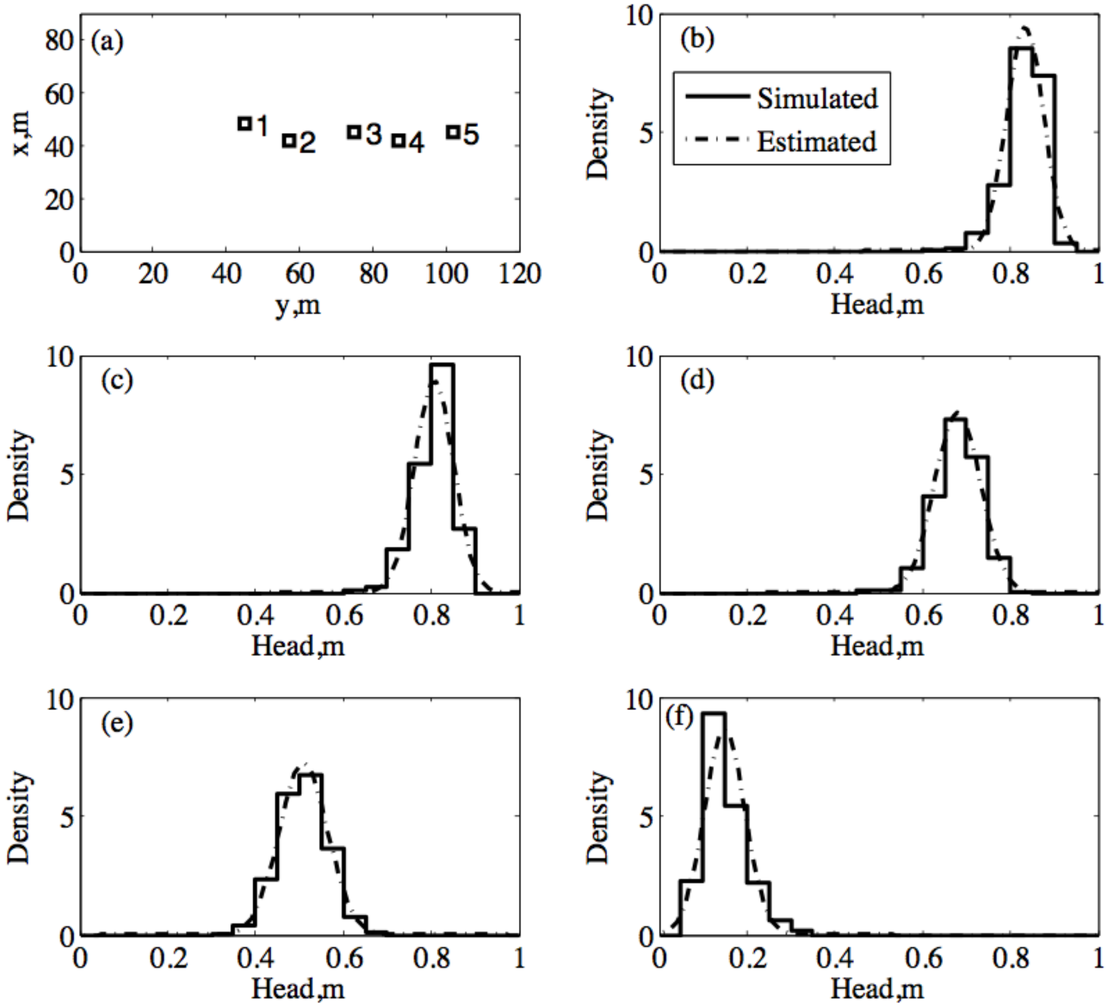


Figure 2.4: Marginal distributions of the head observation at five observation wells. The blue lines are histgrams of simulated heads on the randomly generated fields. The read lines are estimated distribution with the procedure described above:(a) locations of five wells; (b) Well 1; (c) Well 2; (d) Well 3; (e) Well 4; (f) Well 5.

For sampling the parameter vector $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$, we used the MCMC sampling method. The algorithm is shown in Appendix C. We found that 2,000 samples are enough to attain the convergence after discarding the first 200 samples as a common practice of MCMC, since increasing the number of samples further did not change any of the results shown in this chapter. We found that 300 realizations of random fields conditional to anchors were sufficient to estimate the likelihood for each parameter set, since the likelihood values did not change any further with increasing numbers of realizations. Each case consumed around 47 hours on a single core of an Intel Core 2 Quad

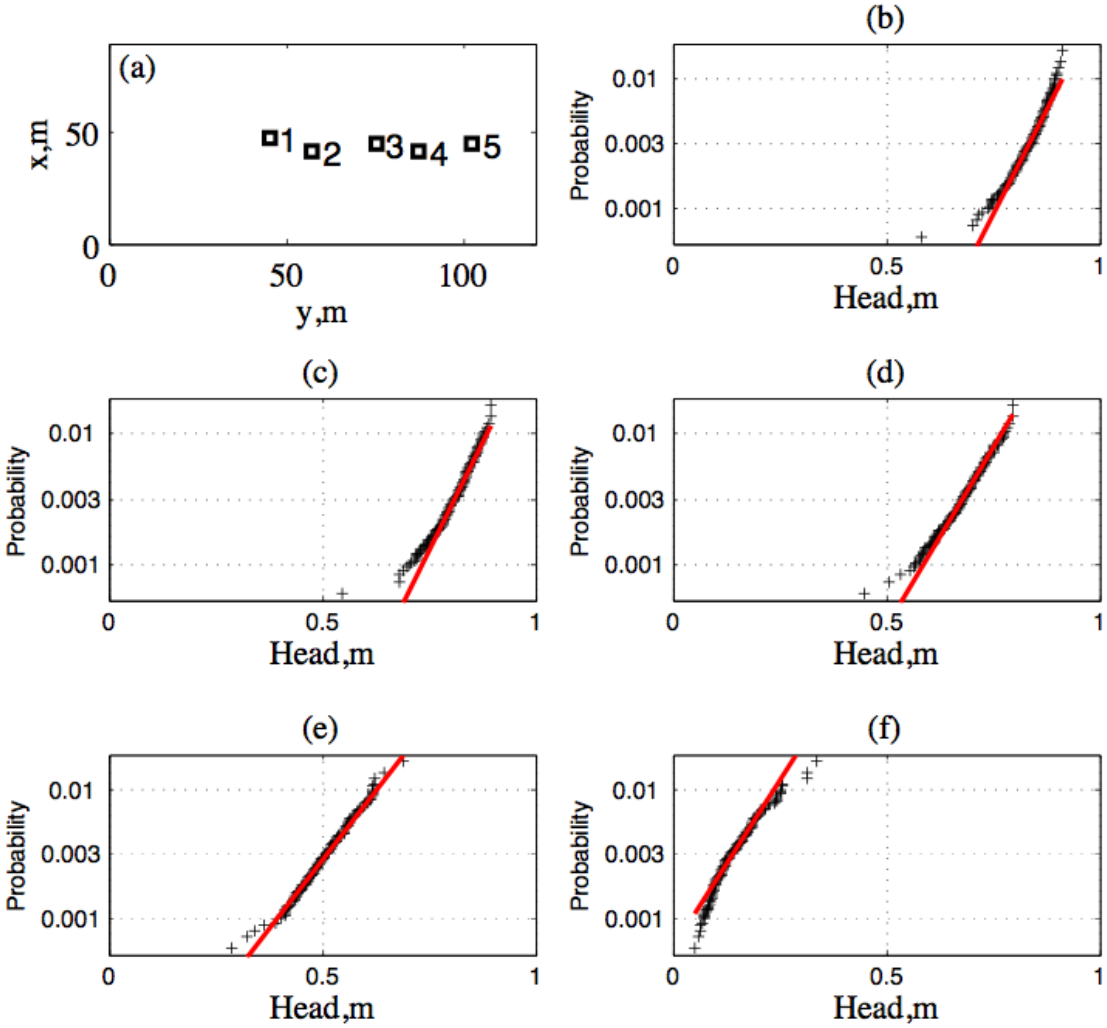


Figure 2.5: Quantile-quantile plots of the head observation at five observation wells: (a) locations of five wells; (b) Well 1; (c) Well 2; (d) Well 3; (e) Well 4; (f) Well 5.

Q6600 3.0 GHz processor.

Figure 2.6 shows the marginal posterior distributions of the scale parameter ϕ obtained using the various layouts of data anchors shown in Figure 2.3. The boundaries of each plot correspond to the minimum and maximum values of the prior distribution. Although the result became worse when we included Type-A data without anchors, the posterior distribution improves as we added more anchors, in the sense that the posterior distribution becomes distributed around the true value. Figure 2.6(f) (36 anchors) depicts the mode of the distribution close to the actual value of the scale. Among the layout with 11 anchors, the layout in Figure 2.3(d) (sparse+dense) has the best result, as shown in Figure 2.6(d). It suggests that the anchor placement for the structural parameters might be related to the optimal design of direct (Type-A) measurements, since previous studies [Zimmerman, 2006; Diggle and Ribeiro, 2007] found that a dense measurement network with a few sparse points outside is the best configuration to estimate the scale. Figure 2.7 shows the marginal distributions for the variance σ^2 . Here, we notice that the contribution of the Type-B data appears more significant compared to the scale in Figure 2.6, even when no use is made of anchors.

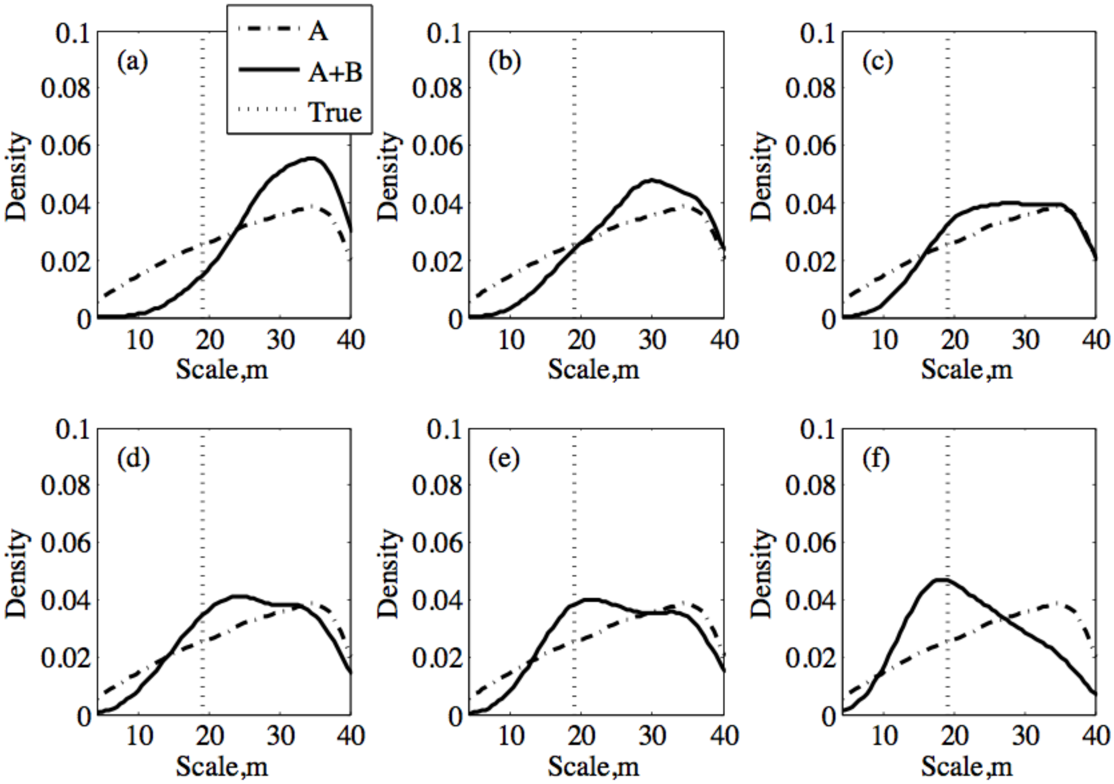


Figure 2.6: Marginal posterior distributions of the scale obtained for the corresponding anchor layouts shown in Figure 2.3. The distributions were estimated by kernel-smoothing of the posterior samples.

Figure 2.8 evaluates the combined impact of measurements and anchors on two transects along the aquifer. The transect on the left is surrounded by measurements and anchors. The transect on the right is outside of the measurement network, and relies mostly on anchors as sources of

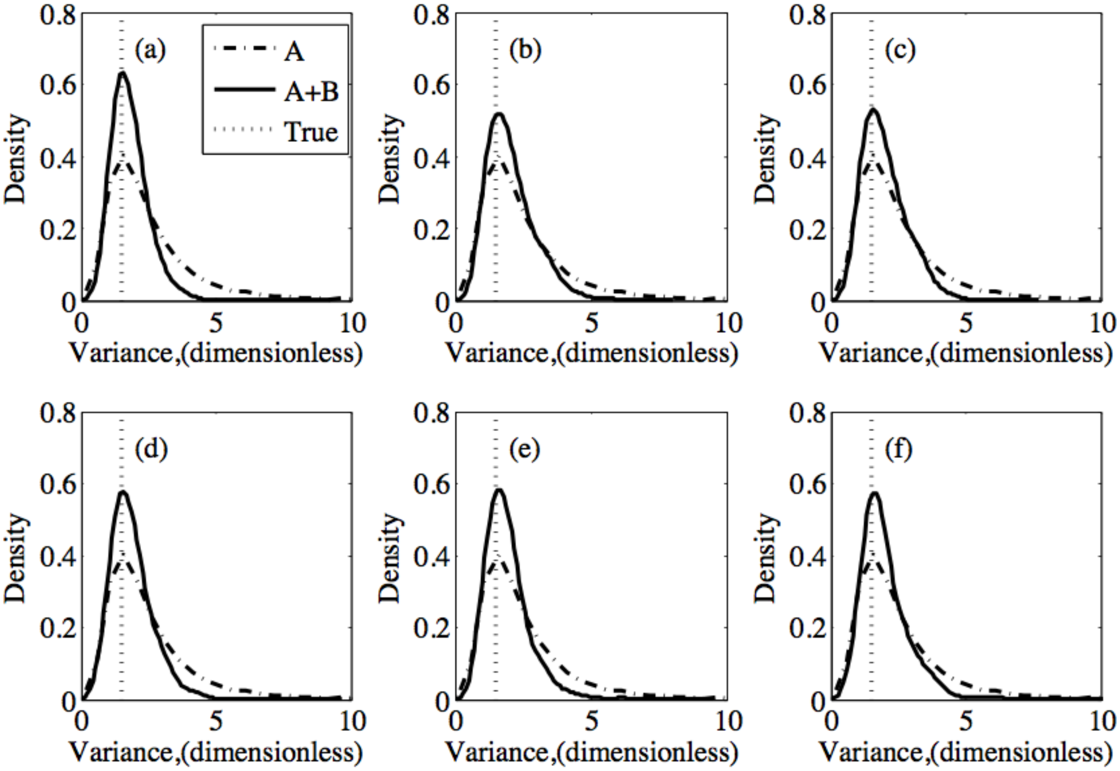


Figure 2.7: Marginal posterior distributions of the variance obtained for the corresponding anchor layouts shown in Figure 2.3. The distributions were estimated by kernel-smoothing of the posterior samples.

local information. Both transects show tighter bounds as the database is augmented with Type-B data (Figure 2.8b) and with the addition of anchors (Figure 2.8c). Besides the observation that the 36-anchor case has tighter bounds than the 0-anchor case, the 36-anchor case also captures the true field within the 90% confidence interval (CI) (Figure 2.8c), while the true field is out of the CI in the 0-anchor case (Figure 2.8b). We may note that none of the CIs in the transect on the right captured the local low-T zone near y=0, since there is no measurement around that zone.

Another way to check the quality of the inversion results is to evaluate their predictive capability. This was done through application of Block III (see Figure 2.1). We selected several test points throughout the domain and compared the baseline hydraulic head with distributions of the predicted head. Figure 2.9 shows the head distribution at the test point shown in Figure 2.9 (i). Comparing between Figures 2.9 (a) and 2.9 (b), we find that Type-B data is effective in improving the predictive capability, since the distribution in (b) is more tightly distributed around the true value than (a). This improvement is much more significant compared to the parameters inference (Figures 2.6 and 2.7). Anchors also play a significant role in improving predictions such that, with more anchors, the distributions are more tightly distributed around the baseline (i.e., true) head value. Since the improvement from 19 to 36 anchors is minor, we may consider that 19 anchors are sufficient for this case, when we are interested in the prediction.

Comparing the three layouts with 11 anchors, we found that the sparse network over the domain (Figure 2.3c) gave the best result, even though there are fewer anchors in the vicinity of that test point. As a rationale for this, we consider that anchors covering a large area are advantageous for predicting hydraulic head, because the hydraulic head is nonlocal, and (as such) it has a influence over much larger distances compared to Y.

2.6 Summary

We presented a Bayesian approach for inverse modeling of a spatially variable field, called the Method of Anchored Distributions (MAD). MAD is general and flexible, in the sense that it is not limited to particular data types or to any particular models. MAD is built around two new elements: (1) the concept of anchored distributions and (2) a data classification scheme. These two elements enable us to assimilate various types of data systematically into an improved characterization of the field.

MAD is based on stochastic modeling of a spatially variable field, employing the geostatistical concept that enables the field to be modeled as a random field, controlled by a combination of structural parameters and anchors. The structural parameters are meant to capture the global trends and statistical characteristics of the fields, whereas the anchors capture the local features. The anchors are statistical devices that attempt to extract information from data that is directly or indirectly (and locally or nonlocally) related to the field. The information is captured as a joint distribution of the structural parameters and anchors.

MAD classifies the data depending on its relationship to the field. Each of Type-A data is small-scale local data given as a function of a point value of the field, whereas each of Type-B data is large-scale nonlocal data given as a function of the field. We place Type-A and Type-B anchors, corresponding to these two data types. Type-A anchors are co-located with Type-A data, the distribution of which is obtained through direct measurements or regression. We place Type-B anchors at chosen locations to capture the information from Type-B data. Bayes rule combined with a likeli-

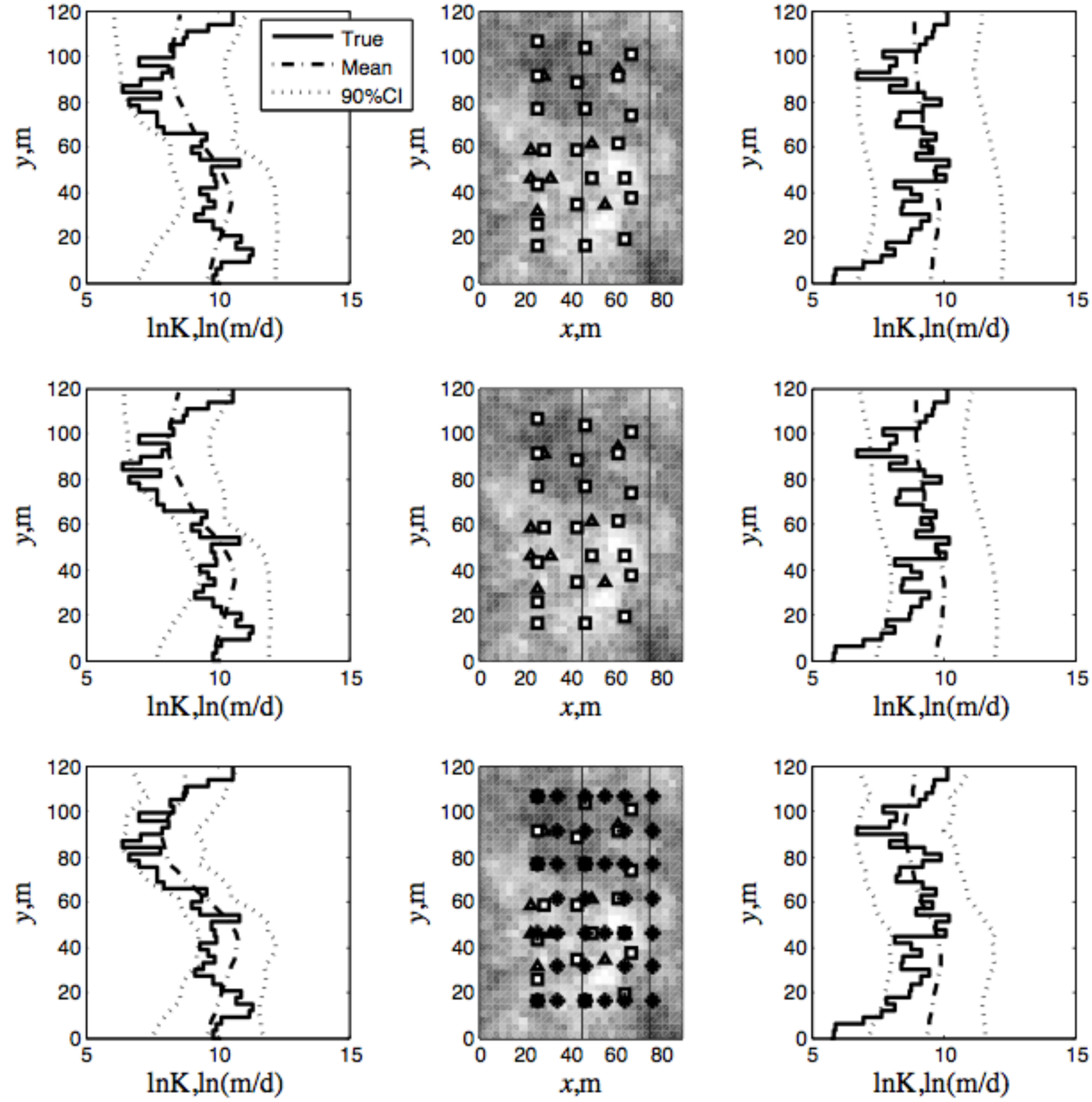


Figure 2.8: Profiles of the log-transmissivity along the two vertical lines shown in the central figure for three cases. The figures on the right show the mean actual values, the expected values and the confidence intervals for right transect whereas the figures on the left show the actual values, the expected values and the confidence intervals for left transect. The transects are shown for 3 cases: (a) only Type-A data used; (b) Type-A and Type-B data used and no anchors used; (c) Type-A data and Type-B data used, anchors used as shown in Figure 2.3 (f) (36 anchors)

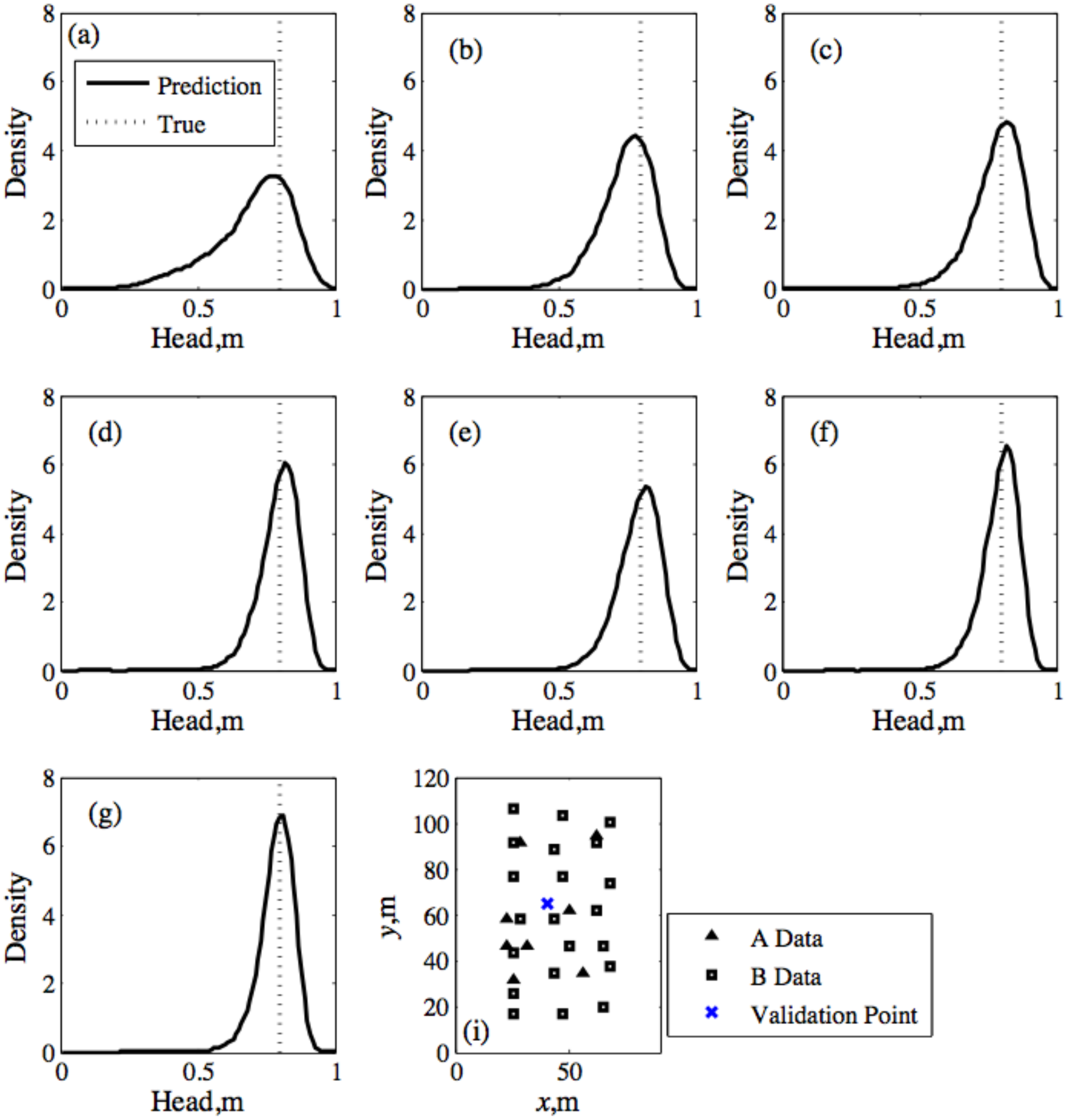


Figure 2.9: The probability density (distribution) of the pressure at the location marked by the X in Figure 2.9(i). The various figures refer to different configurations of data points and anchors as follows: (a) Type-A data only with no anchors; (b) Type-A and Type-B data with no anchors; (c) Type-A, Type-B data and 11 anchors (Figure 2.3b); (d) Type-A, Type-B data and 11 anchors(Figure 2.3c); (e) Type-A, Type-B data and 11 anchors (Figure 2.3d); (f) Type-A, Type-B data and 19 anchors (Figure 2.3e); (g) Type-A, Type-B data and 36 anchors (Figure 2.3f).

hood estimation transfers the information from Type-B data into the distribution of Type-B anchors. This approach provides three benefits: (1) assimilating the different types of data (with different support volumes) in a consistent manner, (2) connecting the data directly to stochastic simulations of the field and flow/transport processes and (3) fully characterizing the parameter uncertainties under certain assumptions.

MAD has a modular structure, which consists of three distinct blocks: a block for analyzing Type-A data, a block for analyzing Type-B data and a block for stochastic simulations to predict the future event of interest. The blocks are defined by their tasks in the total inversion scheme and not by the computational tools employed. The modular structure is advantageous, because it is not limited to any specific type of data or model, and opens the door for applications with a wide variety of data types and modeling tools.

For implementing MAD, we discussed the choice of sampling algorithms, depending on available computational platforms and maturity of the inverse modeling development. A weighted resampling algorithm appears suitable for large-scale parallel computing environments and/or the early stages of a development in which we need to test all the components in MAD. The Markov-chain Monte-Carlo approach seems more advantageous for single core or small-scale parallel computing environments and/or the later stage of the development in which all the components are tested and properly implemented.

We tested MAD with MCMC on a simple two-dimensional flow problem for characterizing the transmissivity field. As synthetic data, we used direct transmissivity measurements (small-scale Type-A data) and hydraulic head measurements (large-scale Type-B data). The results showed that the method successfully captured the true parameter values and predicted values within the posterior distribution. It also showed the power of anchors, in sense that introducing anchors led to a significant improvement by reducing uncertainty and capturing true values in the field prediction and flow measurements. We investigated different anchor location settings and found that a dense network with sparse locations is more advantageous for inferring the parameters, but the sparse network is for predicting the hydraulic heads, and also found that more anchors are necessary for inferring the scale parameter than predicting the hydraulic heads.

Chapter 3

Bayesian approach for three-dimensional aquifer characterization at the Hanford 300 area

3.1 Introduction

Hydrogeological characterization plays a key role in various projects involving groundwater flow and contaminant transport. A detailed three-dimensional (3-D) description of spatial variability in subsurface hydraulic properties is imperative for predicting water and solute movement in the subsurface [Rubin, 2003]. Recent focus on geochemical and microbiological reactions in field studies, for example, requires flow parameters to be fully characterized a priori for testing their research hypotheses [Scheibe et al., 2001; Scheibe and Chien, 2003; Fienen et al., 2004].

One of the main challenges in hydrogeological characterization is to integrate datasets of different types and scales. Typical field studies usually include two or more different complementary sources of information, which may include depth-discrete small-scale data such as core analysis, slug tests and electromagnetic borehole flowmeter (EBF) tests and large-scale data such as pumping tests and tracer tests. With stochastic modeling of flow and transport becoming increasingly common, it is important not only to combine best-fitted values from each dataset, but also to correctly quantify and weigh errors and uncertainty associated with different datasets, and to transfer the uncertainty to the final prediction [Maxwell et al., 1999; Hou and Rubin, 2005; De Barros et al., 2009].

To tackle this challenge, various researchers have applied Bayesian approaches to the problem of subsurface characterization [Copty et al., 1993; Woodbury and Rubin, 2000; Chen et al., 2001]. Within a Bayesian framework, the probability density function of a parameter vector can be updated sequentially to include more datasets in a consistent manner. In addition, the resulting predictive distribution can properly account for the parameter uncertainty inherent in estimating the parameter values from the data [Diggle and Ribeiro, 2002]. Two recent developments in particular

---

[2] This chapter is based on a published article in Hydrology and Earth System Sciences, 2010 (Murakami, H., Chen, X., Hahn, M. S., Liu, Y., Rockhold, M. L., Vermeul, V. R., Zachara, J. M., and Rubin, Y.: Bayesian approach for three-dimensional aquifer characterization at the Hanford 300 Area, Hydrol. Earth Syst. Sci., 14, 1989-2001, doi:10.5194/hess-14-1989-2010, 2010.)

have increased the potential of the Bayesian approach for subsurface characterization: (1) Bayesian model-based geostatistics, and (2) the method of anchored distributions (MAD).

Bayesian model-based geostatistics, introduced by Kitanidis [1986] and Handcock and Stein [1993], assumes a parametric model for a spatial stochastic process and infers geostatistical structural parameters based on small-scale datasets or point measurements [Diggle and Ribeiro, 2006]. While the traditional variogram approach determines best-fitted estimates of geostatistical structural parameters and their asymptotic confidence interval, the Bayesian model-based approach yields a posterior distribution of the parameters. Diggle and Ribeiro [2006] showed that correlation parameters such as variance and scale follow non-Gaussian and skewed distributions, which suggests that the first two moments are not enough to characterize the distribution.

The method of anchored distributions (MAD) is a general Bayesian method for inverse modeling of spatial random fields that addresses complex patterns of spatial variability, multiple sources and scales of data available for characterizing the fields, and the complex relationships between observed and target variables [Zhang and Rubin, 2008a, b; Rubin et al., 2010]. The central element of MAD is a new concept called "anchors." Anchors are devices for localizing large-scale data: they are used to convert large-scale, indirect data into local distributions of the target variables. The goal of the inversion is to determine the joint distribution of the anchors and structural parameters, conditioned on all of the measurements. The structural parameters describe large-scale trends of the target variable fields, whereas the anchors capture local heterogeneities. Following the inversion, the joint distribution of anchors and structural parameters can be directly used to generate random fields of the target variable(s). Different from most of the other inversion methods that determine a single best estimate of the field and asymptotic uncertainty bounds [Kitanidis, 1995; Ramarao et al., 1995; McLaughlin and Townley, 1996], MAD yields a posterior distribution of the parameters.

In this chapter, we assimilate EBF tests and constant-rate injection tests for characterizing a 3-D hydraulic conductivity K field at the Integrated Field Research Challenge (IFRC) site in the Hanford 300 Area (http://ifchanford.pnl.gov). Since the EBF tests yield only relative K values along each of the EBF test wells, we need a local transmissivity T value at each of the EBF test wells to convert the relative values to absolute K values [Javandel and Witherspoon, 1969; Molz et al., 1994; Young et al., 1998; Fienen et al., 2004]. The local T values can be determined by inverting the large-scale constant-rate injection tests. This assimilation requires us to quantify the uncertainty in T values based on the injection tests and to combine that uncertainty with the one from the EBF data.

The particular difficulty in inverting injection-test or pumping-test data is the computational effort associated with a long time series. Li et al. [2005] and Zhu and Yeh [2006] have applied temporal moments of drawdowns to estimate T and storativity S fields. The sandbox experiment by Liu et al. [2007] has shown that the moment approach can successfully characterize a T field. The advantage of employing temporal moments is that we can compute them using steady-state flow equations, which can reduce the computational burden significantly. In addition, when the interest is limited to T, the zeroth-order temporal moment can eliminate the effects of the uncertainty in S, since it does not depend on S [Zhu and Yeh, 2006].

In the following sections, we first describe the site and the experimental procedure. We then present our approach, including the geostatistical inversion method and the inference of the 3-D geostatistical parameters. After presenting the inversion results in a synthetic study to demonstrate and verify our approach, we discuss the results using the actual data at the site.

3.2 Site and Experiment Description

The Hanford 300 Area is located at the southern part of the Hanford Nuclear Reservation one mile north of Richland, Washington, USA. The IFRC site is located within the footprint of a former disposal facility for uranium-bearing liquid wastes known as the South Process Pond, approximately 250 m west from the Columbia River. As is shown in Figure 3.1, the triangular well field consists of 25 wells fully screened through the saturated portion of the Hanford formation, ten wells partially screened at different depths, and one deep characterization well [Bjornstad et al., 2009].

In this study, we focus on the saturated portion of the highly permeable and coarse-grained Hanford formation, which is a shallow unconfined aquifer. The main lithology is a poorly sorted mixture, dominated by gravel up to boulder size, with lesser amounts of sand and silt [Bjornstad et al., 2009]. It overlies the Ringold formation, the upper portion of which is a continuous low-permeability layer consisting of cohesive and compact, well-sorted fine sand to silty sand. The saturated thickness is variable over the site, ranging from about 5 m to 8 m, with daily and seasonal fluctuations of the water table in response to changes in the river stage. The prior estimates of hydraulic conductivity are 1000–100 000 m/day for the Hanford formation and 0.01–3.00 m/day for the Ringold formation [Meyer et al., 2007].

The aquifer in the Hanford formation has been very difficult to characterize, since typical methods, such as permeameter tests and slug tests, do not provide reliable results due to the coarse-grained and highly permeable nature of the aquifer [Meyer et al., 2007; Newcomer, 2008; Vermeul et al., 2009]. In addition, traditional analysis of pumping and injection tests yields only the averaged property over a large domain, since the zone-of-influence expands very rapidly. Combining EBF tests and injection tests through inverse modeling is one of the few feasible alternatives to characterize the 3-D heterogeneous structure of the aquifer.

Fourteen constant-rate injection tests were conducted, each of which had one injection well and 7 to 10 observation wells. All the wells used in the tests are fully screened over the saturated portion of the Hanford formation. The distance between the injection and observation wells ranges between 8 and 60 m. The injection duration and rate are approximately 20 minutes and 315–318 gpm (1.19–1.20 $\mathrm{m}^3\,\mathrm{min}^{-1}$), respectively. The preliminary analysis of the late-time curve data, using the Cooper-Jacob straight-line method, has shown that most of the observation wells yield similar estimates for the T values in each test, which is considered to be the geometric mean of T, T_G, over the entire well field, as is mathematically proved by Sanchez-Villa et al. [1999]. It suggests that the zone-of-influence expands very rapidly and the conventional pumping test analysis yields only an effective property, smoothing out the local heterogeneity at the well field.

The EBF test data were obtained at 19 fully screened wells, which yielded 283 depth-discrete relative hydraulic conductivities with 0.3–0.6 m depth intervals. The pumping rate was 1.04–1.55 gpm (3.94×10^{-3}–3.94×10^{-3} $\mathrm{m}^3\,\mathrm{min}^{-1}$), and kept constant during the test at each well. The vertical profiles indicated that the hydraulic conductivity over the central third of the Hanford formation was lower than its top and bottom thirds at many of the wells. Although the thickness and contact depths for this lower permeability material vary across the site, this general pattern was observed to some extent at most of the monitoring well locations.

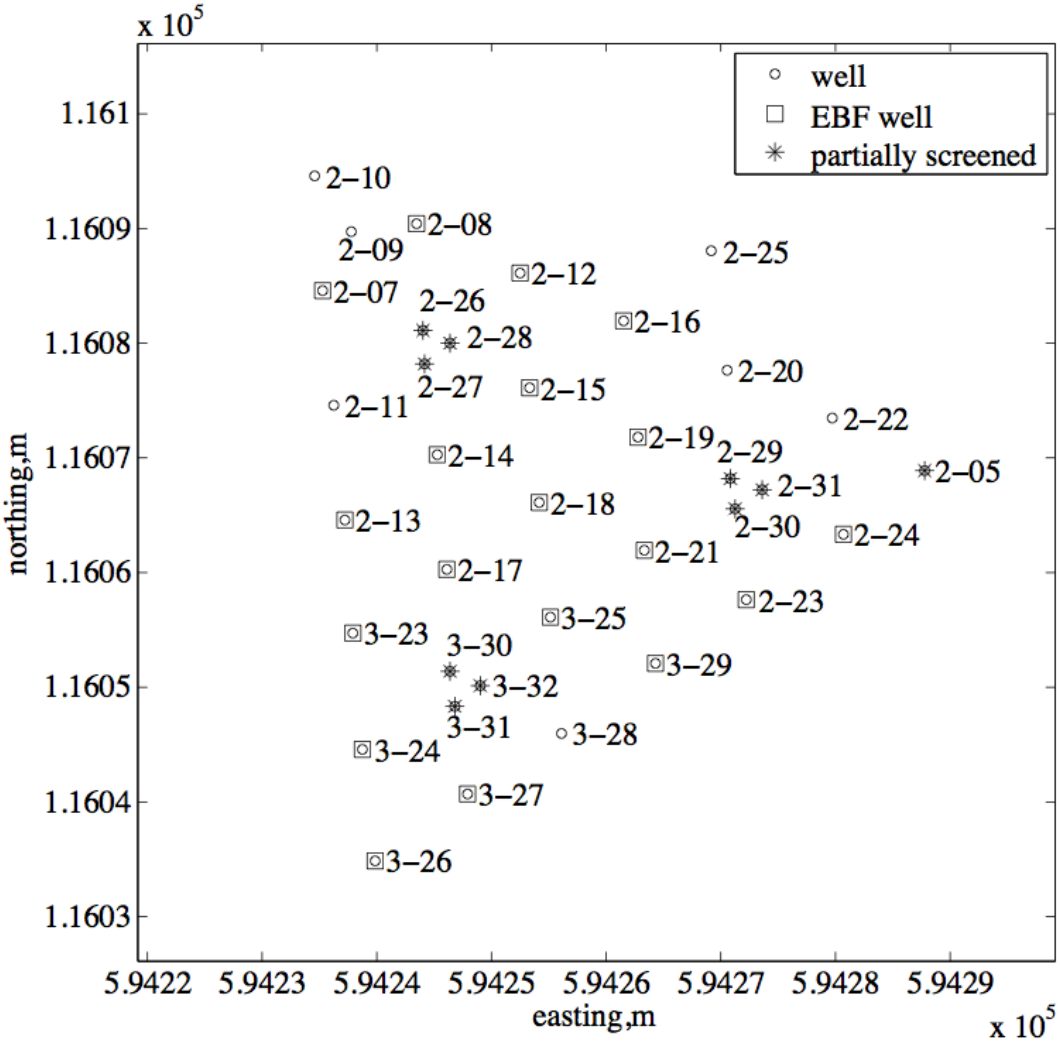


Figure 3.1: Site map of the IFRC site (The coordinate system follows the convention used at the Hanford site).

The more detailed description of the site and data is available in Bjornstad et al. [2009] and Rockhold et al. [2010].

3.3 Integration Strategy of EBF and Injection Tests

For the 3-D characterization, we employ a two-step approach. First, we invert the large-scale injection tests to characterize the 2-D T field. We apply MAD to invert the zeroth-order moments of pressure-buildup curves at multiple observation wells. As a result, we obtain a joint posterior distribution of T at the EBF test wells. Second, we combine this distribution with the EBF data for determining the absolute K values. Instead of a single K value at each of the EBF data point, we obtain the distribution of K at each point. Based on the distribution of the absolute K, we infer the 3-D geostatistical parameters, using the Bayesian model-based geostatistics.

Compared to direct coupling of the EBF and pumping tests used in Li et al. [2008], this two-step approach has a significant computational advantage. This approach is possible because we can model the flow process during the injection tests as 2-D planar flow in the horizontal plane, due to the particular site conditions as the following. The coarse-grained and highly permeable nature of the aquifer caused the elastic response and drainage effect to occur very rapidly (less than 30 s after the injection started), so that the radial flow regime dominated the pressure buildup responses [Neuman, 1975]. In addition, despite the large injection rate, the maximum pressure buildup at the nearest observation wells was less than several centimeters, which is much smaller than the aquifer thickness (5–8 m). Although the EBF tests suggested vertical heterogeneity in the saturated zone, Dagan et al. [2009] recently showed that Dupuit's assumption is still valid – when the aquifer thickness is not large compared to the vertical integral scale, and the ratio between the vertical and horizontal integral scale is large, which is the case at this site.

3.4 Geostatistical Inversion for Transmissivity Field

3.4.1 MAD for large-scale data

In this section, we summarize the Bayesian geostatistical inversion method, MAD, which was introduced in Chapter 2. Since we use it to invert the injection-test data without direct measurements of transmissivity, we modify the explanation focusing only on the large-scale data.

We denote a spatial random process by $Y(x)$, where x is the space coordinate. We further denote an entire field of Y by a random vector $\tilde{\boldsymbol{Y}}$, and denote a realization of the field by $\tilde{\boldsymbol{y}}$. The dimension of $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{y}}$ is equal to the number of elements in the discretized field. The field $\tilde{\boldsymbol{Y}}$ is defined through the vector of model parameters $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. The $\boldsymbol{\theta}$ part of this vector, called the structural parameter vector, includes a set of parameters designed to capture the global features of $\tilde{\boldsymbol{Y}}$, such as the mean of the field and correlation structures. The $\boldsymbol{\vartheta}$ component of this vector consists of the anchored distributions, or anchors in short. Anchors are devices used to capture local features of $\tilde{\boldsymbol{Y}}$ that cannot be captured by the global parameters $\boldsymbol{\theta}$. In their simplest form, an anchor would be error free measurements of Y. Other forms of anchors include measurements coupled with error distributions and/or anchors that are obtained by inversion.

The data $\boldsymbol{z}$ is a vector of multiple observations of a physical process. The data can be described by the following equation:

$$\boldsymbol{z} = \mathcal{M}\left(\tilde{\boldsymbol{y}}\right) + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathcal{M}$ is a known function, or a set of functions, numerical or analytical, of the spatial field, and $\boldsymbol{\epsilon}$ is a vector of zero-mean errors. The goal of the inversion is, first, to derive a posterior distribution of the model parameters conditioned on the data $\boldsymbol{z}$, $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\boldsymbol{z})$. This distribution then allows us to generate multiple realizations of the field $\tilde{\boldsymbol{Y}}$ for prediction.

Using Bayes' rule, we can define the posterior distribution of parameters as:

$$p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z}) \propto p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{3.2}$$

where $p(\boldsymbol{\theta})$ is the prior distribution, $p(\boldsymbol{\vartheta}|\boldsymbol{\theta})$ is the anchor distribution given a structural parameter vector $\boldsymbol{\theta}$, and $p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is the likelihood of data $\boldsymbol{z}$ given a parameter set $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$.

We estimate the likelihood $p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\vartheta})$ using the Monte Carlo simulations. Since the model parameters $\{\boldsymbol{\theta},\boldsymbol{\vartheta}\}$ and the data $\boldsymbol{z}$ are connected through the field, we generate multiple conditional realizations of the field $\tilde{\boldsymbol{Y}}$ for any given $\{\boldsymbol{\theta},\boldsymbol{\vartheta}\}$; with each realization, the forward model provides a prediction of $\boldsymbol{z}$ in the form of $\tilde{\boldsymbol{z}}$, according to Equation (3.1). In other words, $\boldsymbol{z}$ is viewed as a measured outcome from a random process, whereas $\tilde{\boldsymbol{z}}$ is one of many possible realizations, given a particular parameter set of $\{\boldsymbol{\theta},\boldsymbol{\vartheta}\}$. By generating random fields for a given parameter set $p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\vartheta})$ and simulating the physical process on each field, we obtain multiple realizations of $\tilde{\boldsymbol{z}}$, i.e., an ensemble of $\tilde{\boldsymbol{z}}$. The ensemble of $\tilde{\boldsymbol{z}}$ is then used for estimating the probability density function (pdf) of $\tilde{\boldsymbol{z}}$. After determining the pdf, it is straightforward to calculate the density at a point $\boldsymbol{z}$, $p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\vartheta})$, which is the likelihood.

As the dimension of $\boldsymbol{z}$ increases, a larger number of realizations of $\tilde{\boldsymbol{z}}$ are necessary to estimate the pdf accurately, which increases the computational burden. To accommodate the large-dimensional data, we may divide the vector $\boldsymbol{z}$ into L segments as $\boldsymbol{z}$=$\{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_L\}$. We can decompose the likelihood into each segment as,

$$\begin{aligned} p(\boldsymbol{z}\mid\boldsymbol{\theta},\boldsymbol{\vartheta}) &= p(\boldsymbol{z}_1,...,\boldsymbol{z}_L\mid\boldsymbol{\theta},\boldsymbol{\vartheta}) \\ &= p(\boldsymbol{z}_L\mid\boldsymbol{z}_1,...,\boldsymbol{z}_{L-1},\boldsymbol{\theta},\boldsymbol{\vartheta})p(\boldsymbol{z}_{L-1}\mid\boldsymbol{z}_1,...,\boldsymbol{z}_{L-2},\boldsymbol{\theta},\boldsymbol{\vartheta}),...,p(\boldsymbol{z}_2\mid\boldsymbol{z}_1,\boldsymbol{\theta},\boldsymbol{\vartheta})p(\boldsymbol{z}_1\mid\boldsymbol{\theta},\boldsymbol{\vartheta}) \\ &\approx \prod_{l=1}^{L} p(\boldsymbol{z}_l\mid\boldsymbol{\theta},\boldsymbol{\vartheta}). \end{aligned} \tag{3.3}$$

In Equation (3.3), we assume that the data segments $\boldsymbol{z}_1$, $\boldsymbol{z}_2$,..., $\boldsymbol{z}_L$ are conditionally independent for a given $\{\boldsymbol{\theta},\boldsymbol{\vartheta}\}$, since we consider that $\{\boldsymbol{\theta},\boldsymbol{\vartheta}\}$ contains information equivalent to the data. This equality strictly holds when the data segments are independent of each other – for example, when the data locations are beyond the zone-of-influence or zone-of-correlation. It approximately holds when the data segments are only weakly correlated, such as with different types of data at the same site. As Hou and Rubin [2005] pointed out, assuming independence leads to higher entropy and makes the estimation less informative.

3.4.2 Specification of a 2-D Geostatistical Model

Here we specify the geostatistical model for the 2-D T field. Let $Y(x)$ be natural-log transmissivity, $\ln T(x)$, at the location $x=(x_1,x_2)$ in the 2-D domain. We assume that a vector $\boldsymbol{Y}$, containing Y at multiple locations $\boldsymbol{x}$, follows a multivariate Gaussian distribution with exponential covariance. We define a structural parameter vector as $\boldsymbol{\theta}=\{\mu,\sigma^2,\phi\}$, including uniform mean μ, variance σ^2, and integral scale ϕ, which are used at a geologically similar site (i.e. unconsolidated glacial materials) [Rubin, 2003; McLaughlin et al., 1993].

We define a vector $\boldsymbol{\vartheta}(\boldsymbol{x}_\vartheta)$ to represent a set of anchors. Since the anchors are a subset of the field, $p(\boldsymbol{\vartheta}|\boldsymbol{\theta})$ is a multivariate Gaussian distribution with mean μ and covariance $\sigma^2\mathbf{R}^{(2-D)}(\boldsymbol{x}_\vartheta,\boldsymbol{x}_\vartheta)$, where $\mathbf{R}^{(2-D)}(\boldsymbol{x}_\vartheta,\boldsymbol{x}_\vartheta)$ is an auto-correlation matrix as a function of ϕ and the locations of $\boldsymbol{\vartheta}$, $\boldsymbol{x}_\vartheta$. The distribution of $\boldsymbol{Y}$ conditioned on the structural parameters and anchors $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\vartheta})$ is a multivariate Gaussian distribution with conditional mean $\mu_{Y|\vartheta}$ and conditional covariance $\sigma^2\mathbf{R}^{(2-D)}_{Y|\vartheta}$, where the mean and covariance conditioned on the anchors are defined as

$$\begin{aligned} \mu_{Y|\vartheta} &= \mu+\mathbf{R}^{(2-D)}(\boldsymbol{x},\boldsymbol{x}_\vartheta)\mathbf{R}^{(2-D)}(\boldsymbol{x}_\vartheta,\boldsymbol{x}_\vartheta)^{-1}(\boldsymbol{\vartheta}-\mu), \\ \mathbf{R}^{(2-D)}_{Y|\vartheta} &= \mathbf{R}^{(2-D)}(\boldsymbol{x},\boldsymbol{x})-\mathbf{R}^{(2-D)}(\boldsymbol{x},\boldsymbol{x}_\vartheta)\mathbf{R}^{(2-D)}(\boldsymbol{x}_\vartheta,\boldsymbol{x}_\vartheta)^{-1}\mathbf{R}^{(2-D)}(\boldsymbol{x}_\vartheta,\boldsymbol{x}) \end{aligned} \tag{3.4}$$

where $\mathbf{R}^{(2-D)}(\boldsymbol{x},\boldsymbol{x})$ is the auto-correlation matrix for $\boldsymbol{Y}$, and $\mathbf{R}^{(2-D)}(\boldsymbol{x},\boldsymbol{x}_\vartheta)$ is the cross-correlation matrix between $\boldsymbol{Y}$ and $\boldsymbol{\vartheta}$.

3.4.3 Specification of Likelihood

We consider the data z consisting of L injection tests $(l = 1, 2, \ldots, L)$. We divide z into L segments as $z = \{z_1, z_2, \ldots, z_L\}$, where z_l is the vector containing the zeroth-order moments at multiple observation wells in the l-th injection test. The governing equation and the temporal moment formulation are shown in Appendix D.

In order to determine the likelihood $p(z|\boldsymbol{\theta}, \boldsymbol{\vartheta})$, we first compute the likelihood in each injection test $p(z_l|\boldsymbol{\theta}, \boldsymbol{\vartheta})$. Since, during the forward modeling analysis, we have found that the zeroth-order moments are approximately Gaussian, we use a multivariate Gaussian distribution for the likelihood estimation. Although nonparametric density estimation is available, the Gaussian-likelihood estimation is computationally advantageous as the dimension of the data increases.

Using the ensemble of $\tilde{z}_l$ simulated on the multiple fields conditioned on each parameter set $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$, we calculate the mean and covariance to determine $p(z_l|\boldsymbol{\theta}, \boldsymbol{\vartheta})$ (Robert and Casella, 2005). When we include the multiple injection tests in the inversion, we multiply the likelihoods of multiple tests, according to Equation (3.3), to obtain the likelihood for the entire data $p(z|\boldsymbol{\theta}, \boldsymbol{\vartheta})$. We have observed that the zeroth-order moments from the different injection tests are not strongly correlated, so that we may use Equation (3.3).

3.4.4 Placement of Anchors

The success of MAD depends on placement of the anchors from two reasons. First, careful placement of anchors will maximize their ability to extract information from observations. Second, the computational burden is linked to the number of anchors, and a smaller number would improve computational efficiency. Hence, we need to place anchors (1) at sensitive locations to the data, (2) to capture local features of the field, and (3) according to the goal of the inversion. A detailed discussion is available in Rubin et al. [2010]. Here we discuss the issues relevant to our inversion.

First, to find sensitive locations, we refer to the sensitivity analysis. Li et al. [2005] has formulated the sensitivity of zeroth-order moments to a $\ln T$ value at a specific location, using the adjoint-state method [Sun, 1994]. For the current estimate of T field required in the sensitivity analysis, we may use a uniform T field, since our priors are only for the global parameters (i.e., mean, variance, and scale) and we do not have any local information before the injection test. In this case, sensitivity is high around observation well locations, which is consistent with findings by Castagna and Bellin [2009] and Vasco et al. [2000].

Second, to capture heterogeneity, we would ideally have more than one anchor per integral scale. Although the real integral scale is not known in advance, we may consider the minimum possible integral scale at the site. Anchors outside the well plot, far from any of the observation wells, are not effective in resolving spatial heterogeneity, so that we need fewer anchors outside the well plot.

Third, to achieve our goal, which is to obtain the $\ln T$ values at the EBF well locations, we need anchors at all EBF well locations. All the EBF wells are used as observation wells during the injection tests, so that we do not need additional anchors for this purpose.

3.5 3-D Geostatistical Model for Hydraulic Conductivity Field

The 2-D inversion of the injection tests yielded a joint distribution of $\ln T$ at the EBF well locations. Since we placed anchors at all those locations, we can use the anchor distribution directly. Let us denote the $\ln T$ values at the EBF well location by a vector $\boldsymbol{\vartheta}_{\mathrm{EBF}}$, which is a subset of $\boldsymbol{\vartheta}$. Marginalizing the other parameters leads to the posterior distribution of $\boldsymbol{\vartheta}_{\mathrm{EBF}}$ conditioned on the injection test data $\boldsymbol{z}$ as $p(\boldsymbol{\vartheta}_{\mathrm{EBF}}|\boldsymbol{z})$.

Let $K(x_1, x_2, x_3)$ and $k(x_1, x_2, x_3)$ be the absolute and relative K values at the location $x = (x_1, x_2, x_3)$ in the 3-D domain, respectively. Based on Javandel and Witherspoon [1969], we have the correlation between the absolute and relative K values [Moltz et al., 1994; Fienen et al., 2004] as:

$$K(x_1, x_2, x_3) = T(x_1, x_2)k(x_1, x_2, x_3)/b(x_1, x_2), \tag{3.5}$$

where $b(x_1, x_2)$ is the aquifer thickness at the horizontal location (x_1, x_2). We can determine the natural log-conductivity $u = \ln K$ at (x_1, x_2, x_3) by

$$u(x_1, x_2, x_3) = \vartheta_{\mathrm{EBF}}(x_1, x_2) - \ln b(x_1, x_2) + \ln k(x_1, x_2, x_3). \tag{3.6}$$

We use a N-vector $\boldsymbol{k}$ containing all the relative conductivity values from the EBF data at N locations $\boldsymbol{x}$, and a N-vector $\boldsymbol{u}$ containing all the $\ln K$ values at the same locations as $\boldsymbol{k}$. Equation (3.6) allows us to combine $\boldsymbol{k}$ and $p(\boldsymbol{\vartheta}_{\mathrm{EBF}}|\boldsymbol{z})$ into $p(\boldsymbol{u}|\boldsymbol{k}, \boldsymbol{z})$, which is the distribution of $\boldsymbol{u}$ conditioned on both the injection test data and the EBF data.

We construct a 3-D geostatistical model, assuming that $\boldsymbol{u}(\boldsymbol{x})$ follows a multivariate Gaussian distribution with mean β and covariance $(\eta^2\mathbf{R}^{(3-D)}(\boldsymbol{x}, \boldsymbol{x}) + \nu^2\mathbf{I})$, where η^2 is the variance of variability in $\ln K$, $\mathbf{R}^{(3-D)}(\boldsymbol{x}, \boldsymbol{x})$ is the auto-correlation matrix for $\boldsymbol{u}(\boldsymbol{x})$ as a function of the locations $\boldsymbol{x}$, the horizontal integral scale λ_h and the vertical integral scale λ_v, $\mathbf{I}$ is the identity matrix of order N, and ν^2 is the nugget, which represents the EBF measurement errors. The structural parameter vector of the 3-D geostatistical model is $\{\beta, \eta^2, \lambda_h, \lambda_v, \nu^2\}$. Our goal here is to obtain a joint posterior distribution of the parameters conditioned on both data $\{\boldsymbol{k}, \boldsymbol{z}\}$ through $\boldsymbol{u}$:

$$p\left(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2 | \boldsymbol{k}, \boldsymbol{z}\right) = \int p\left(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2 | \boldsymbol{u}\right) p\left(\boldsymbol{u}|\boldsymbol{k}, \boldsymbol{z}\right) d\boldsymbol{u}. \tag{3.7}$$

For the prior distribution of the parameters, we assume the Jeffreys prior for the mean and variance [Jeffreys, 1946], which is the least informative prior for those two parameters. The prior distribution of all the structural parameters is defined as

$$p\left(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2\right) \propto \frac{1}{\eta^2}\pi\left(\lambda_h, \lambda_v, \nu^2\right). \tag{3.8}$$

For the rest of the prior distribution $\pi(\lambda_h, \lambda_v, \nu^2)$, we use an independent uniform distribution for each of $\{\lambda_h, \lambda_v, \nu^2\}$ bounded by each set of the minimum and maximum possible values. Following Diggle and Ribeiro [Chapter 6, 2006], we obtain an analytical expression for $p(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2|\boldsymbol{u})$ (Appendix E). In addition, based on the joint distribution of the structural parameters and $\boldsymbol{u}$, determined by $p(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2, \boldsymbol{u}| \boldsymbol{k}, \boldsymbol{z}) = p(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2| \boldsymbol{u})p(\boldsymbol{u}|\boldsymbol{k}, \boldsymbol{z})$ we can sample multiple parameter sets $\{\beta, \eta^2, \lambda_h, \lambda_v, \nu^2, \boldsymbol{u}\}$, and generate multiple random fields conditioned on each of the parameter sets.

3.6 Implementation

3.6.1 Organization of Constant-rate Injection Test Data

To demonstrate our approach, we used seven out of the fourteen constant-rate injection tests at the site for the synthetic study, and four for the real data analysis (injection at Well 2-09, 2-18, 2-24, and 3-24), the locations of which are well balanced within the IFRC site. Figure 3.2 shows the configuration of the injection and observation wells for each of the four tests.

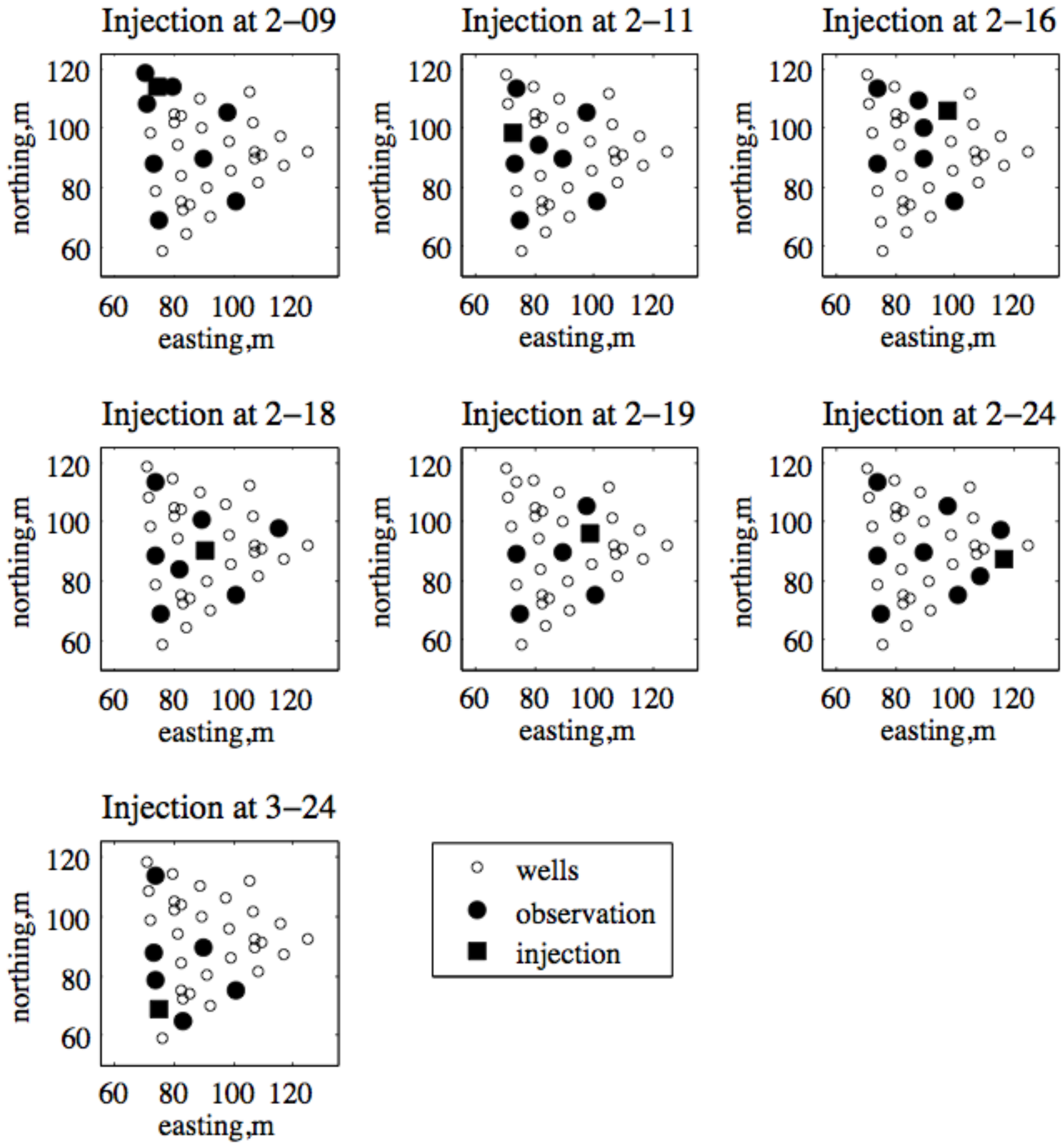


Figure 3.2: Configuration of injection and observation wells in each test used in this chapter. The reference point of local coordinates is at (594 164 m, 115 976 m) in the Hanford coordinates.

Table 3.1: The lower and upper bounds of the prior distribution for the structural parameters of the 2-D transmissivity field.

	Minimum	Maximum	Reference
Mean μ, m^2 s-1 (ln$T$)	-4.82*	2.17*	Meyer et al. (2007) (Inverse model estimates)
Variance, σ^2	0.5	3.0	Rubin (2003), Table 2.1 and Table 2.2
Scale, ϕ, m	8	30	Rubin (2003), Table 2.1 and Table 2.2

* The upper bound and lower bounds of K multiplied by the average aquifer thickness 7.62 m.

For each test, we calculated the zeroth-order moments at multiple observation wells by integrating the pressure-buildup curves. Since the well field is located near the Columbia River, the water table changes according to the river stage fluctuation. Since the change was mostly linear within 20 min after the injection started, we removed the ambient head contribution by linear interpolation. For quantifying the measurement errors, we followed Nowak et al. [2006] and Li et al. [2008], who determined the errors based on fluctuation or noise in the pressure measurements. Since the noise in our case is contained within the range of the instrument resolution, we determined the standard deviation of measurement error based on the resolution of the instrument, 0.003 ft (0.09 cm) by integrating it over the injection duration.

3.6.2 Prior Distribution for MAD Inversion

For the prior distribution of the 2-D structural parameters $\boldsymbol{\theta}$, we used three independent uniform distributions bounded by the minimum and maximum values, as are shown in Table 3.1. The prior distributions of each parameter sufficiently cover possible values from the historical data at the site [Meyer et al., 2007] or literature values for similar geological formations [Rubin, 2003]. The uniform distributions are considered to be less informative than Gaussian distributions, which have been commonly used in the Bayesian geostatistical inversion [Li et al., 2005]. Three thousand sets of $\boldsymbol{\theta}$ are generated from $p(\boldsymbol{\theta})$ using a quasi Monte-Carlo sampling method [Krommer and Ueberhuber, 1998].

As is shown in Figure 3.3, we placed 44 anchors at all the well locations inside the well plot and at sparse locations outside the well plot, following the discussion in Section 3.4.4. For each set of $\boldsymbol{\theta}$, we generated 12 sets of anchors $\boldsymbol{\vartheta}$ from $p(\boldsymbol{\vartheta}|\boldsymbol{\theta})$, so that the number of prior parameter sets $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ is 36 000. We gradually added more parameter sets until we observed the posterior distribution remaining the same – not changing along with increasing numbers of sets.

3.6.3 Forward Simulation in MAD

Figure 3.3 shows the 2-D computational domain used for the forward simulations. For simulating the zeroth-order temporal moments on multiple random fields, we followed the approach by Firmani et al. [2006], since the mathematical expression for the zeroth-order moments is the same as the one for steady-state flow toward a well.

Firmani et al. [2006] determined the grid and domain sizes according to the integral scale.

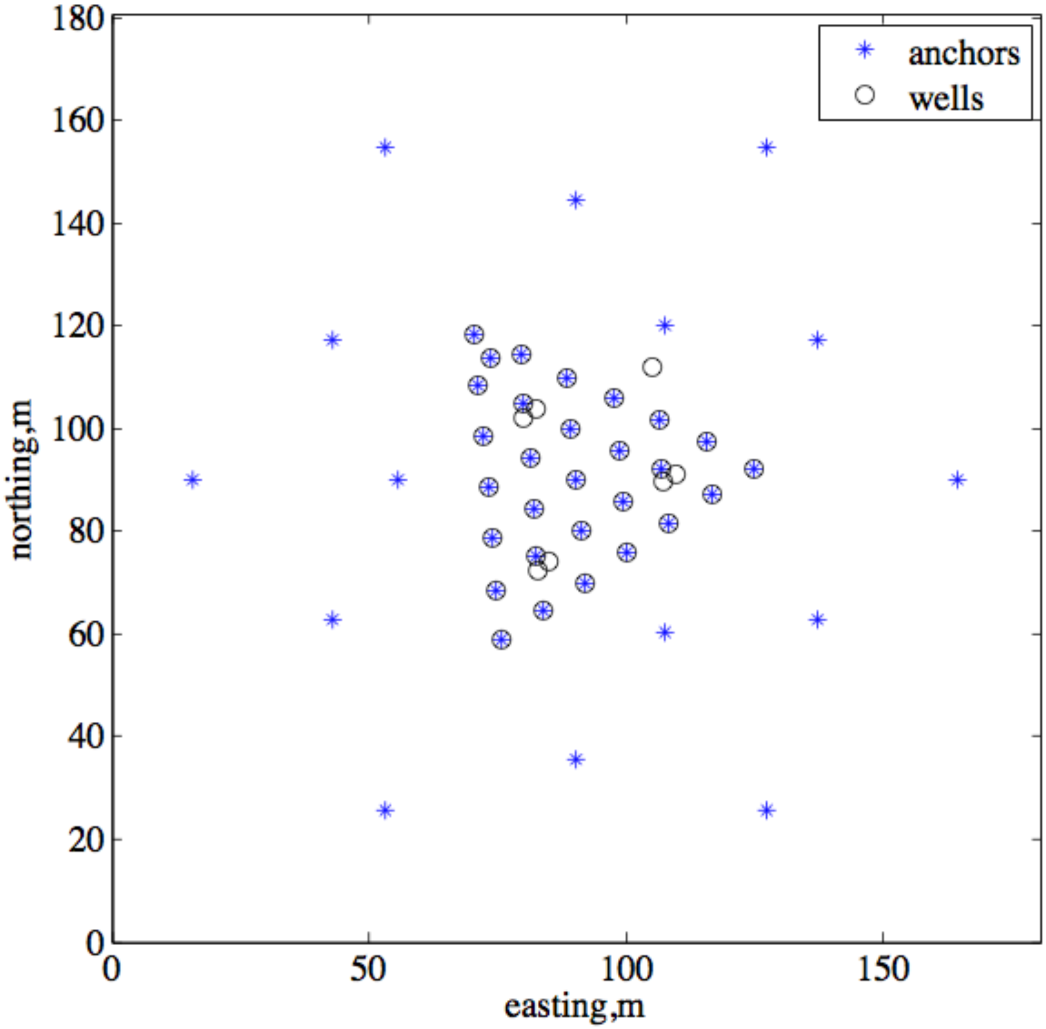


Figure 3.3: Anchor locations in the domain for the constant-rate injection test inversion. The reference point of local coordinates is at (594 164 m, 115 976 m) in the Hanford coordinates.

Table 3.2: The lower and upper bounds of the prior distribution for the horizontal scale, vertical scale and nugget variance of the 3-D hydraulic conductivity field.

	Minimum	Maximum	Reference/Justification
Horizontal scale, λ_h, m	8	50	Rubin (2003), Table 2.1 and Table 2.2
Vertical scale, λ_v, m	0.5	10	Saturated thickness <10 m
Nugget variance, ν^2	10^{-3}	0.25	Less than 50 % of standard deviation

Although the integral scale is unknown in our framework, we have the minimum or maximum possible integral scale ($\phi_{\min}$ and $\phi_{\max}$), which are the upper and lower bounds of the uniform prior distribution. We used these two values so that any possible integral scale can satisfy the requirement for the domain and grid sizes.

The computational grid size is uniform equal to $0.2\phi_{\min}$ in both x_1 and x_2 directions during the field generation. During the flow simulations, the grid blocks near the injection well are divided into non-uniform grid from $0.04\phi_{\min}$ at the injection well location to $0.2\phi_{\min}$ at a distance of $0.8\phi_{\min}$, satisfying the condition that the ratio between the neighboring block size should not exceed 1.5 [Firmani et al., 2006]. We determined the domain size such that the observation wells were $2\phi_{\max}$ away from the boundaries to reduce the boundary effect. During the flow simulation, we added another buffer zone with width $\phi_{\max}$ and uniform T equal to T_G between the field and the boundaries for further reducing the boundary effect. We intended to satisfy $3\phi_{\max}$ between any observation wells and the boundaries, based on the theory developed by Rubin and Dagan [1988].

We used the SGSIM program from GSLIB [Deutsch and Journel, 1998] to generate spatially correlated Gaussian random fields conditioned on each set of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. We then simulated zeroth-order moments on each field, using a finite-element method with linear elements. We used 250 realizations of random fields and moments for each $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ to calculate the likelihood $p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\vartheta})$. We gradually added more realizations, and found that the likelihood values converged with 250 realizations. The 9 000 000 forward simulations took 60 000 computational hours in total. It was divided into several batches, and used up to 9000 cores on the Franklin supercomputer at the National Energy Research Scientific Computing Center (Berkeley, USA), each core of which is a 2.3 GHz single AMD Opteron processor.

3.6.4 3-D Geostatistical Model

After calculating $p(\boldsymbol{u}|\boldsymbol{k}, \boldsymbol{z})$ from the relative K values and $\ln T$ values at the EBF well locations, we generated a thousand sets of $\boldsymbol{u}$ from $p(\boldsymbol{u}|\boldsymbol{k}, \boldsymbol{z})$. For each set of $\boldsymbol{u}$, we computed a posterior distribution $p(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2|\boldsymbol{u})$, based on the uniform prior distribution of λ_h, λ_v and ν^2 bounded by the values shown in Table 3.2. We then integrate the distribution numerically to determine $p(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2|\boldsymbol{k}, \boldsymbol{z})$.

3.7 Results and Discussion

We first tested the MAD and numerical setting in a synthetic study for inverting the injection test data, and then we applied it to the actual data at the Hanford site.

3.7.1 Synthetic Study for 2-D Transmissivity Field

We generated a synthetic reference 2-D $\ln T$ field with a 2-D geostatistical parameter set $\{\mu, \sigma^2, \phi\}$=$\{-1.8, 1.5, 20\}$, shown in Figure 3.4. We obtained maximum likelihood estimates of the true parameters as $\{-1.76, 1.46, 20.0\}$, with near-exhaustive sampling (one out of every five points) [GeoR package by Ribeiro and Diggle, 2001]. We then calculated the zeroth-order moments on the reference field and superposed a zero-mean independent Gaussian measurement error, which has the same variance as the actual data from the study site.

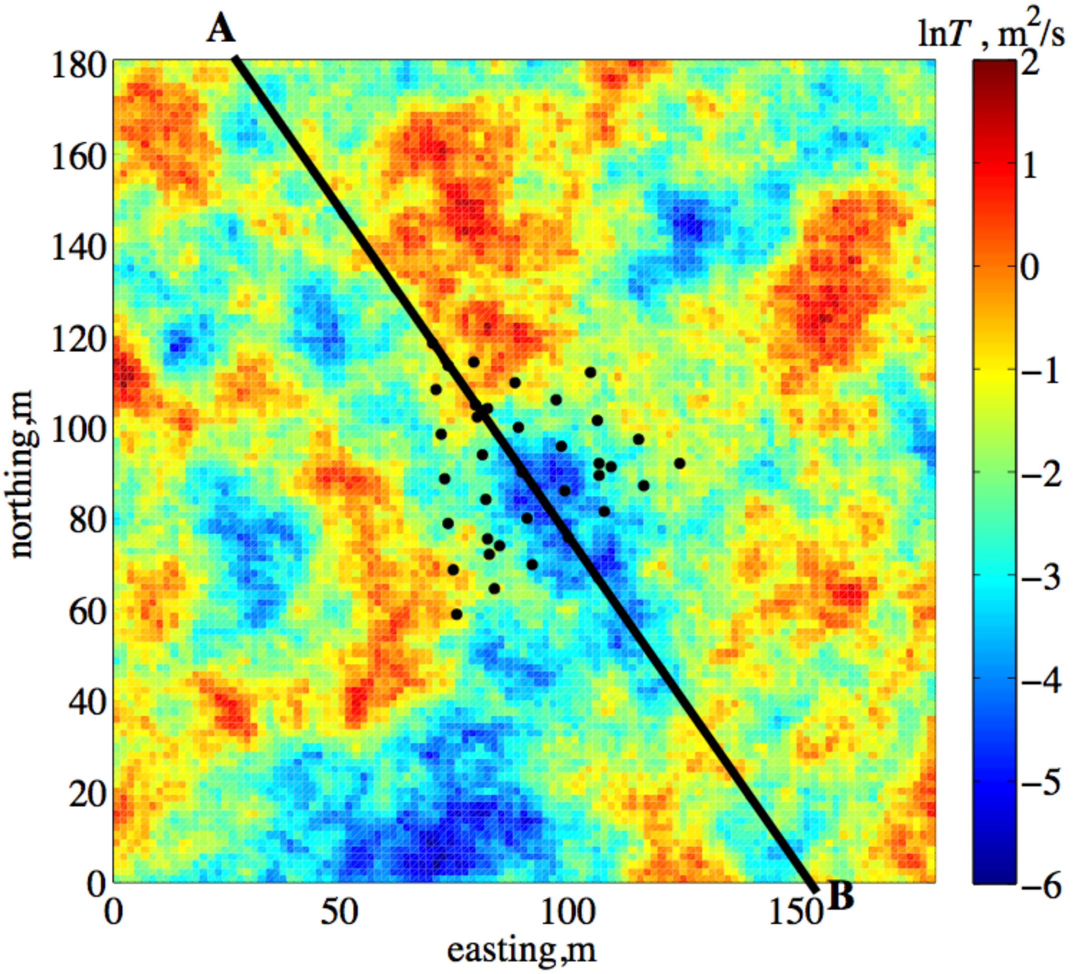


Figure 3.4: Reference field for the synthetic study. Line A-B is used for the transect.

Our inversion process is based on the same sets of injection and observation wells as the actual experiments conducted at the IFRC site (Figure 3.2). We also combined the different number of injection tests in the inversion: one injection test (injection at Well 2-18), two tests (injection at Well 2-09 and 2-24), three tests (injection at Well 2-09, 2-24, and 3-24), four tests (injection at Well 2-09, 2-18, 2-24, and 3-24), five tests(Well 2-09, 2-11, 2-18, 2-24 and 3-24), six tests (Well 2-09, 2-11, 2-16, 2-18, 2-24 and 3-24) and seven tests (Well 2-09, 2-11, 2-16, 2-18, 2-19, 2-24 and 3-24). They are compared to show the effect of additional information from the multiple tests.

Figure 3.5 shows the marginal posterior distributions of the 2-D geostatistical structural parameters $\{\mu, \sigma^2, \phi\}$ based on the various number of tests, with their corresponding true values. While the mean has a symmetric Gaussian-like distribution, the variance and scale has broad and skewed distributions. The results are improved with increasing number of tests up to three tests, i.e. the posterior distributions become narrower and biases are reduced. The improvement by additional tests is more significant for the variance and scale, which suggests that the estimation of variance and scale requires more observations. The improvement, however, is not significant for more than three tests, and the distributions based on four to seven tests are very close to each other, which would suggest that the effect of increasing number of tests could be saturated due to the measurement errors and redundancy of information in the data. Although we may expect tighter distributions with more information, some of the uncertainty cannot be eliminated due to measurement errors and the limited number of observation wells. In addition, the same observation wells were used repeatedly for several tests. Yeh and Li [2000] and Zhu and Yeh [2005] also reported that increasing the number of pumping tests did not improve the estimation above a certain number (three to four tests in their cases).

To examine the effect of anchors and evaluate the random fields generated based on the posterior distributions, we generated $200\,000$ fields from the posterior distribution of parameters (5000 posterior sample parameter sets with 40 fields per parameter set), and compared the ensemble with the true field. Two cases are studied: one based on a single test (injection at Well 2-18) and the other based on three tests (injection at Well 2-09, 2-24, 3-24).

Figure 3.6 shows the mean and 98% confidence interval of the predicted $\ln T$ fields along the centerline of the well field as shown in Figure 3.4 (Line A–B). The centerline also corresponds to the line passing through large variability, from high near the top to low in the middle of the well plot. The figure also includes five realizations that depict the level of heterogeneity in the randomly-generated fields to be used for stochastic simulations. The mean field and random fields along the line follows a general trend of the true field, especially so with more tests assimilated. If there were no anchors, the mean field would be a flat line at the global mean, and the random fields would be distributed around the flat line. Therefore the deviation from such a straight line is attributed to anchors that capture local heterogeneity. The uncertainty bounds are found to be tighter near the center, where there are more observation wells and more anchors. Increasing the number of tests not only reduces the uncertainty, but also reduces the bias by moving the mean field closer to the true field.

3.7.2 IFRC Data Analysis

After we gained confidence from the synthetic study, we applied the same scheme to the data from the IFRC site. Since the synthetic study indicated that four tests would be enough, we used up to four tests (the same sets): one injection test (injection at Well 2-18), two tests (injection at Well 2-09 and 2-24), three tests (injection at Well 2-09, 2-24, and 3-24) and four tests (injection at Well 2-09, 2-18, 2-24, and 3-24). Since the true values are unknown in this case, we validated the posterior distribution by predicting a subset of the data and comparing it to the actual data. The data used for validation was not used for inversion.

Figure 3.7 shows the marginal posterior distributions of the three 2-D structural parame-

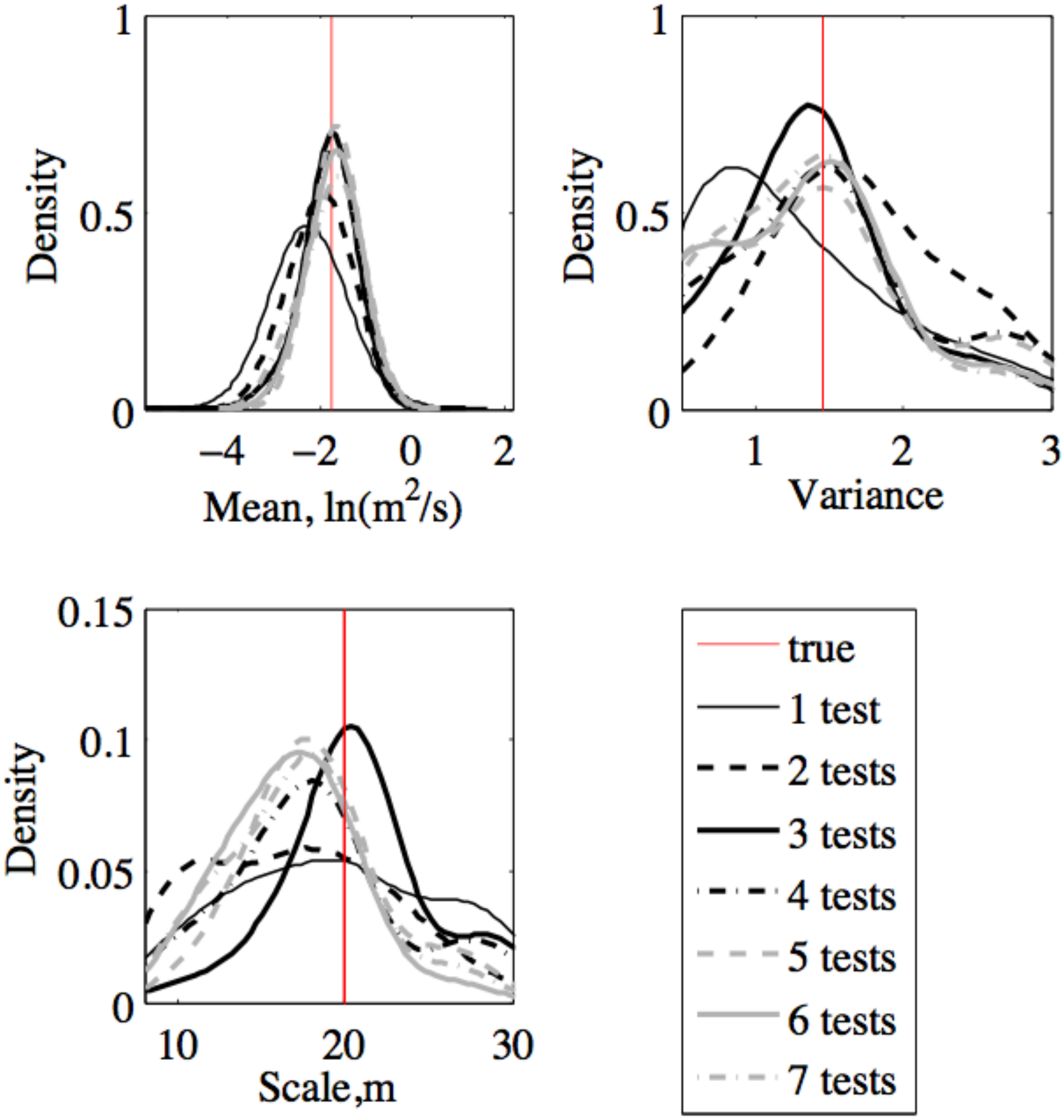


Figure 3.5: Marginal posterior distributions of the structural parameters (mean, variance and scale) in the synthetic study, with their corresponding true values. The ones based on the different number of tests are compared.

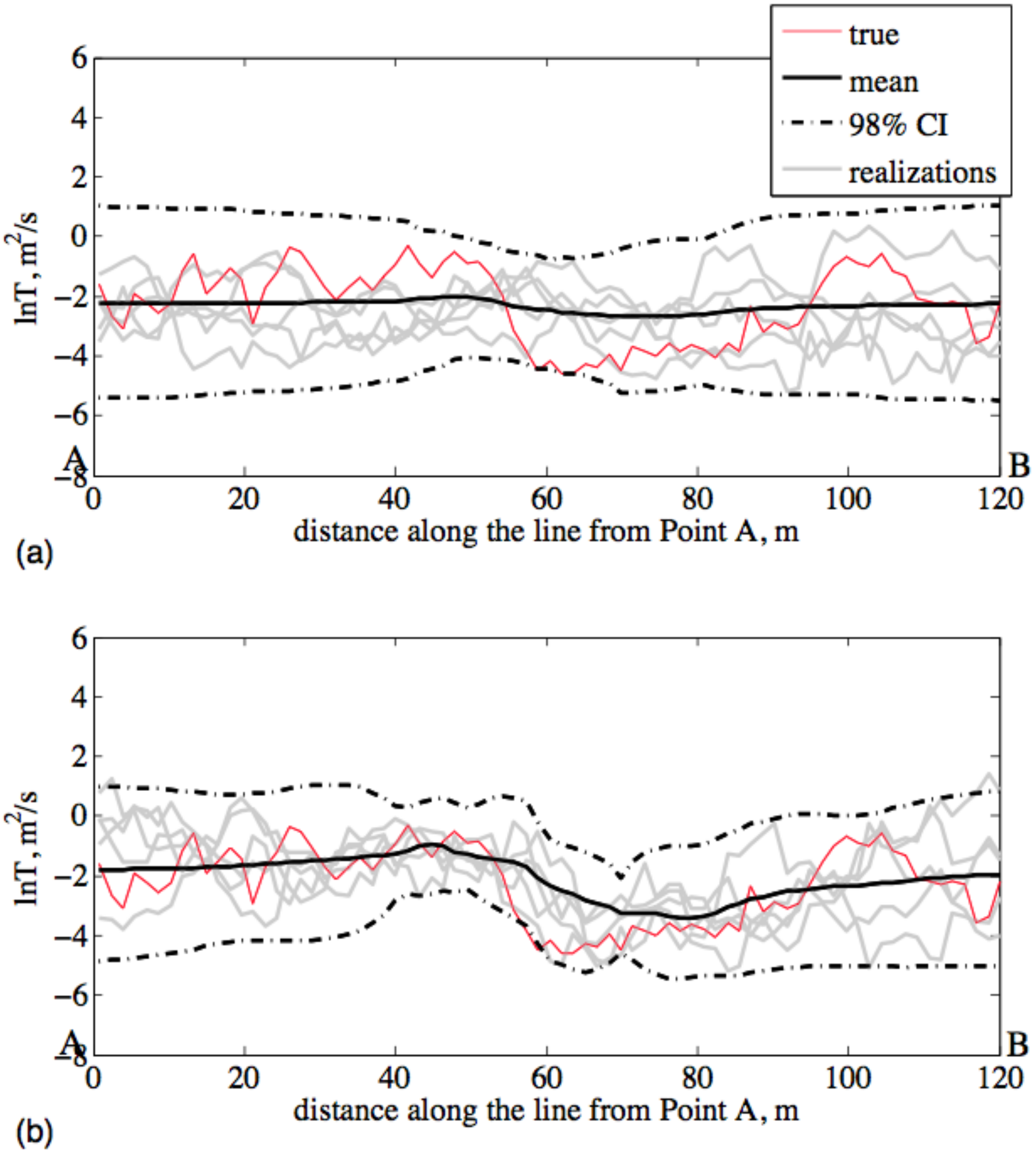


Figure 3.6: Comparison among the reference field, the mean field and the 98% confidence interval of the generated fields, along the center line of the IFRC well field (Line A-B in Figure 3.4), for the inversion based on **(a)** one test (injection at Well 2-18) and **(b)** three tests (injection at Wells 2-09, 2-24 and 3-24).

ters for the 2-D $\ln T$ field at the IFRC site. These three plots show a similar feature to the synthetic study in Figure 3.5 such as a Gaussian-like distribution for the mean, broad and skewed distributions for the variance and scale, and the effect of increasing the number of injection tests in the inversion.

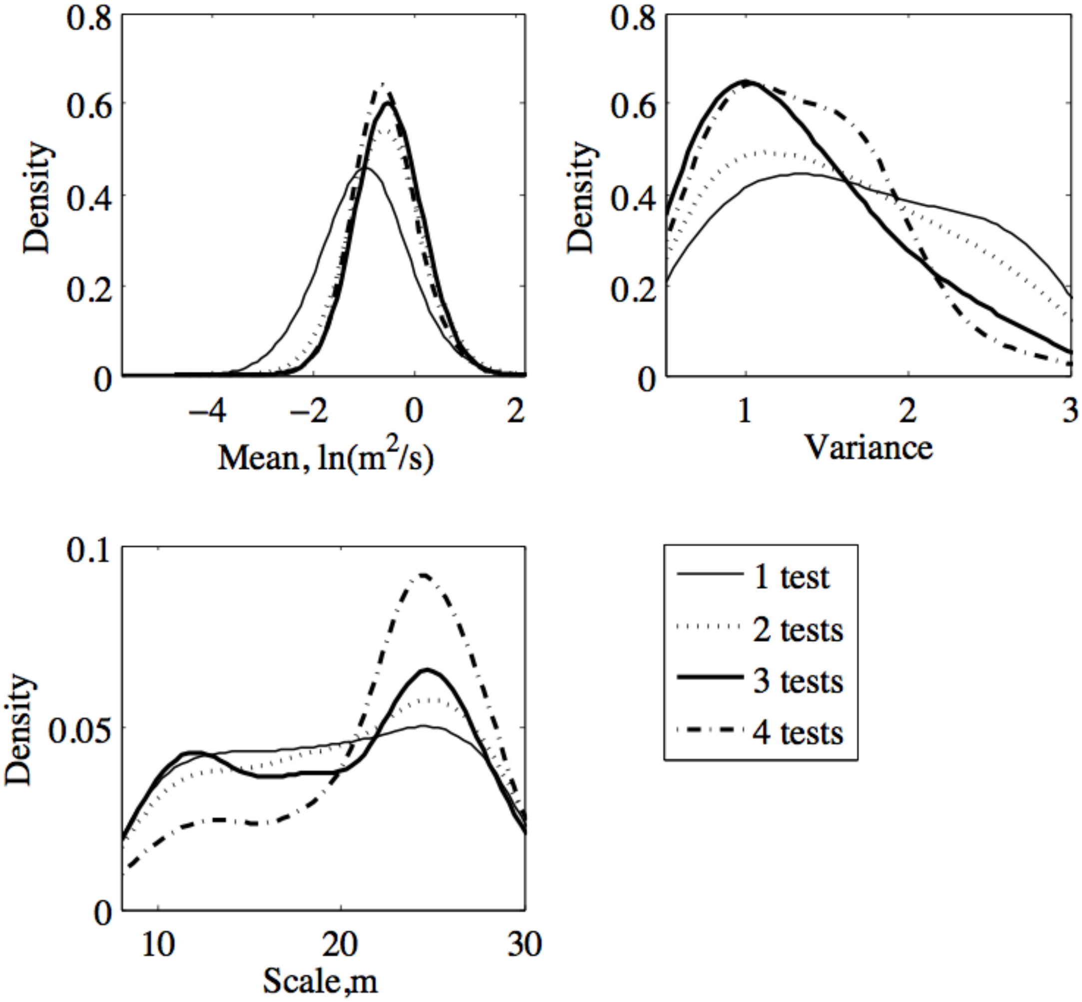


Figure 3.7: Marginal posterior distributions of the structural parameters (mean, variance and scale) for the Hanford IFRC site data.

For the first validation of the 2-D $\ln T$ inversion, we generated 200 000 fields (5000 posterior sample parameter sets with 40 fields per parameter set) based on one test (injection at Well 2-09), two tests (injection at Well 2-09 and 3-24) and three tests (injection at Well 2-09, 2-24 and 3-24). We then predicted the zeroth-order temporal moments in the injection test at Well 2-18, which was not the part of the estimation. Figure 3.8 shows the marginal predictive distributions of the observed moments at two observation wells, based on one, two and three tests, compared with the actual data. The true value is contained within the range of values defining the predictive distributions, and the combination of the three tests improves the prediction by narrowing the distributions.

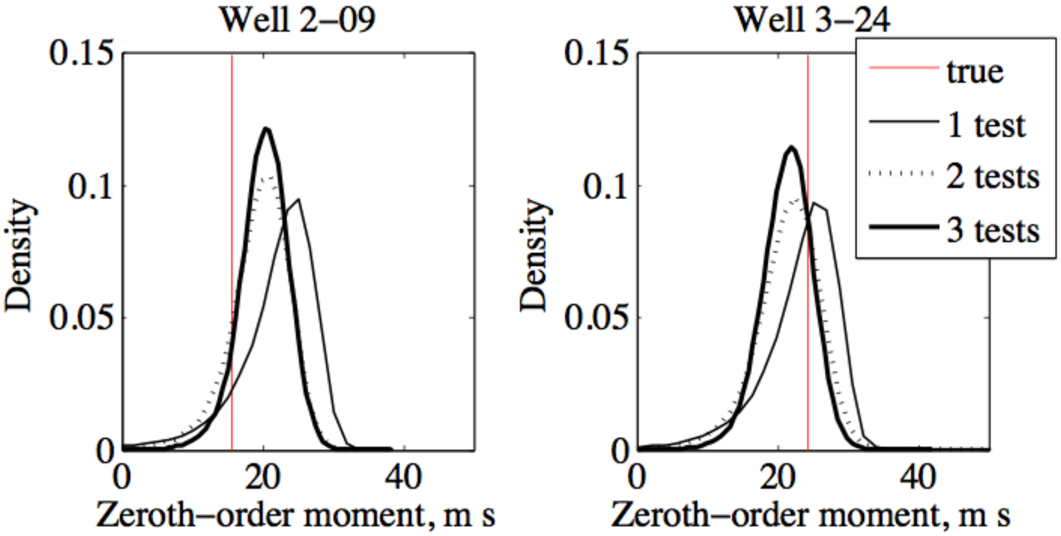


Figure 3.8: Comparison between the zeroth-order moments observed at Well 2-09 and 3-24 in the injection test at Well 2-18, and predictive posterior distributions from the inversion, including the different number of injection test.

As another validation, we obtained the maximum a posteriori (MAP) estimate of $T_G = \exp(\mu)$ in Figure 3.7, which is $0.52\,\mathrm{m^2\,s^{-1}}$. We compared this value with T_G estimated from the Cooper-Jacob analysis (Sanchez-Villa et al., 1999), in which we fitted the late-time pressure-buildup curves at multiple observation wells in the injection test at Well 2-18. The 95% confidence bound of T_G was 0.38–0.57 $\mathrm{m^2\,s^{-1}}$. As we expected, our estimate of T_G corresponded to the estimates based on conventional analysis for a large-scale injection test.

Figure 3.9 shows the marginal distribution for the 3-D geostatistical structural parameters conditioned on the EBF data and injection test data. For the horizontal scale, vertical scale, and nugget, the upper and lower bounds of the x-axis correspond to the bounds of the prior distributions. We can see that the marginal posterior distributions of the structural parameters are skewed except for the mean, which suggests that the entire distribution is necessary to quantify the parameter uncertainty.

The horizontal scale has a particularly broad distribution, which is not zero at the bounds. This is because available data is insufficient for narrowing down the distribution [Diggle and Ribeiro, 2002; Hou and Rubin, 2005]. For the horizontal scale, spacing between the wells restricts variation in horizontal lags, although we have many different lags for the vertical scale. It suggests the importance of setting reasonable bounds for the prior distribution based on the information from geologically similar sites.

We also compared the distributions based on the different number of tests included in the injection test inversion. As it turned out, increasing the number of injection tests did not reduce the parameter uncertainty in the 3-D structural parameters as significantly as in the 2-D parameters.

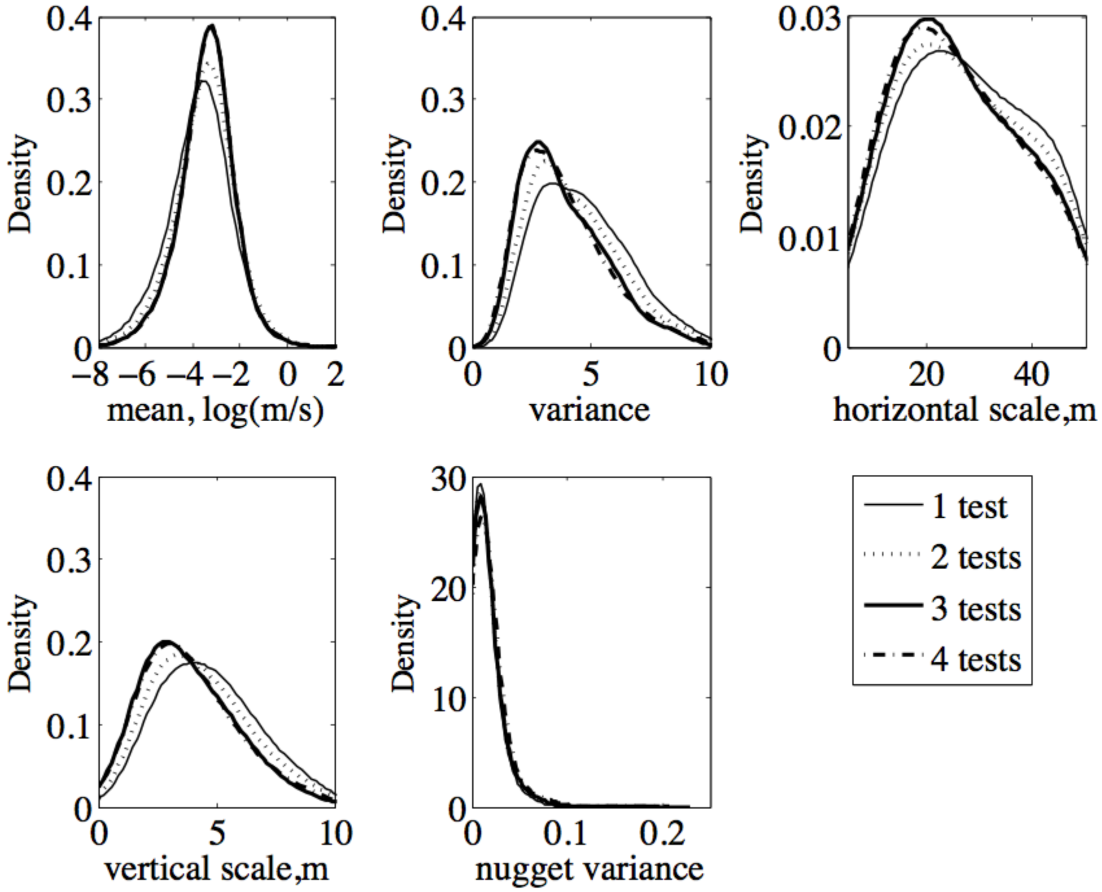


Figure 3.9: Marginal posterior distributions of 3-D geostatistical structural parameters of ln$K$ values at the Hanford IFRC site, based on the different number of injection test.

This is because the uncertainty and sparseness of the EBF data obscures additional information of the increasing number of injection tests in the 3-D spatial inference. These findings are consistent with Li et al. [2008], who also found that the 3-D characterization of the aquifer was limited by the EBF data density.

Figure 3.10 shows the 3-D mean field, based on the 5 000 parameter sets generated from the distribution $p(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2 | \boldsymbol{k}, \boldsymbol{z})$. We can see the high-low-high layers in ln$K$ along the centerline of the field, which is consistent with the observations in the tracer tests later conducted at the site [Rockhold et.al., 2010; Zachara, 2010].

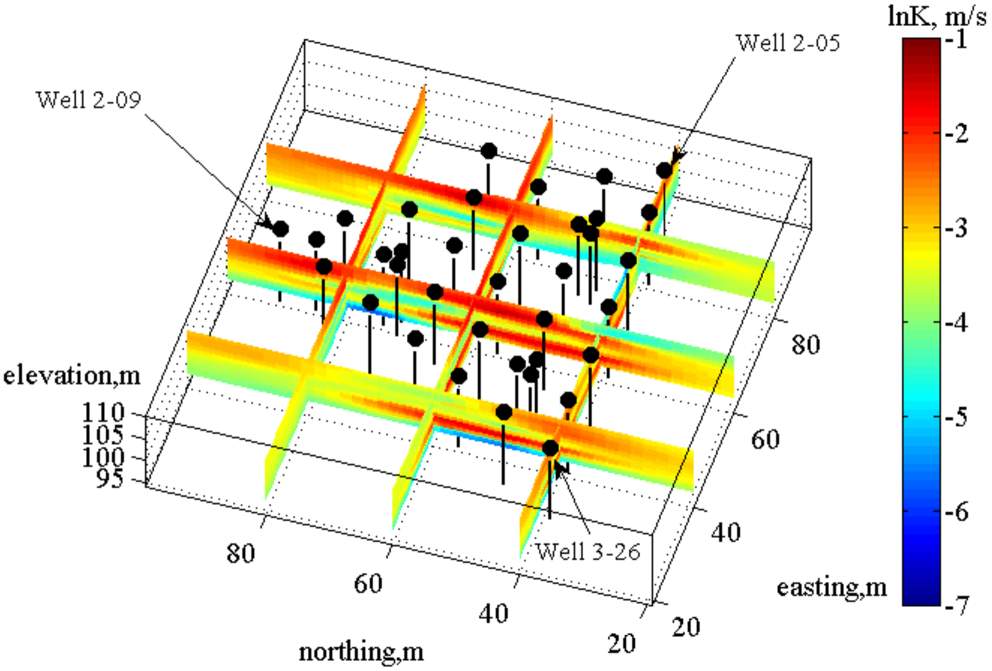


Figure 3.10: 3-D mean ln$K$ field in the saturated portion of the Hanford formation. The black dots and lines represent the well locations. The reference point of local coordinates is at (594 164 m, 115 976 m) in the Hanford coordinates.

3.8 Summary

In this chapter, we presented a Bayesian approach for characterizing a 3-D K field by assimilating the EBF and constant-rate injection tests. We employed a two-step approach – first inverting the constant-rate injection test data for obtaining the joint distribution of local T values at the EBF well locations, and then converting the EBF data to local K values for the 3-D characterization. For the injection test inversion, we used MAD, which is a newly developed Bayesian geostatistical inversion framework. We inverted zeroth-order moments of pressure buildups at multiple observation wells, which can eliminate uncertainty in a storage coefficient, as well as significantly reduce computational cost.

In a synthetic study, we first showed that MAD could successfully infer the geostatistical parameters and predict the 2-D ln$T$ field. As we included more tests, we could further reduce the uncertainty, and better capture the local heterogeneity. The improvement, however, was saturated at three to four tests due to the measurement errors and redundancy of information in the data, which is consistent with the previous studies [Yeh and Li, 2000; Zhu and Yeh, 2005].

By applying the method to the actual data, we obtained the posterior distribution of geostatistical structural parameters and the anchor values of the 2-D ln$T$ field for the Hanford 300 Area IFRC site. We validated the result using the predictive distribution of the zeroth-moments in the injection test that were not part of the inversion. In addition, the MAP estimate of the mean ln$T$ coincided with the lnT_G value obtained from the Cooper-Jacob analysis, which confirmed our methods consistency with conventional pumping test analysis.

We then combined the relative K values from the EBF data with the distribution of ln$T$ at EBF wells so that we obtain the distribution of the depth-discrete absolute K in the 3-D domain. The uncertainty in T (2-D) is consistently carried on into the ln$K$ values (3-D) as a probability distribution. We thus constructed a 3-D geostatistical model for the ln$K$ field using the model-based Bayesian geostatistical approach.

We demonstrated the advantages of MAD such that MAD was directly connected to the stochastic forward simulations, and it directly inferred the joint distribution of the parameters to be used as an input of the simulations. Compared to the other inverse modeling methods that yield only best-estimates and confidence bounds, MAD fully quantifies the parametric uncertainty as statistical distributions, which is necessary to capture skewed distributions found for the variance and scale parameters.

For the field application, we showed that combining EBF and injection tests is promising for characterizing a heterogeneous ln$K$ field in a coarse-grained and highly permeable aquifer, where the other conventional techniques fail to provide information of local heterogeneity. We found that the 3-D characterization is restricted by the EBF measurement density, in the sense that increasing the amount of depth-averaged information from the injection tests did not contribute significantly to narrowing down the posterior distribution of the 3-D geostatistical parameters.

Chapter 4

Sequential Bayesian Geostatistical Inversion and Evaluation of Combined Data Worth for Aquifer Characterization at the Hanford 300 Area

4.1 Introduction

Subsurface characterization plays a central role in various projects involving groundwater flow and contaminant transport. A detailed three-dimensional (3-D) map of subsurface hydrogeological properties is critical for predicting water and solute movement in the subsurface [Rubin, 2003]. In addition, an accurate map of hydrogeological properties is the basis for understanding more complex physical processes, such as geochemical and microbiological reactions. [Scheibe et al., 2001].

Such characterization projects typically include various types of data and experiments as complementary sources of information. To determine a hydraulic conductivity field, for example, we typically have small-scale datasets such as permeameter tests, grain-size analysis and flowmeter tests, and large-scale datasets such as pumping tests and tracer tests. For characterizing a heterogeneous field of target properties, we need to integrate all the types of data in a consistent manner with uncertainty properly quantified. In addition, in order to distribute limited resources to multiple types of data, we need to know how effective each dataset is in reducing uncertainty and mismatch in predictions.

As Rubin et al. [2010] noted, it is a challenge to integrate data and evaluate its data worth, due to the variety of data and complexity of groundwater modeling. In addition to different accuracy and scales, each type of data is connected to the target properties by a different physical-process model, which sometimes requires complex numerical simulations. In addition, some data types become usable only in the presence of other type of data. For example, lithology data is useful only when the hydraulic conductivity values are known in each sedimental class using the site-specific correlation.

The gap between forward and inverse modeling of flow and transport within a stochastic framework presents another challenge. In the stochastic forward simulations, we generate multiple possible realizations of a 3-D heterogeneous field of target properties, and simulate flow and transport for each field. Such stochastic simulations start from a known forward model and probability distribution that quantifies the uncertainty of the model parameters [Maxwell et al., 1999; Rubin, 2003]. Most of the inverse modeling methods, however, focus primarily on determining the best estimates and asymptotic uncertainty bounds of the parameters, and do not provide the distribution [Kitanidis, 1995; Ramarao et al., 1995]. This gap hinders not only taking uncertainty into account properly in predictions, but also evaluating the data worth, since the information cannot be fully transferred from data to prediction.

Recently, the method of anchored distribution (MAD) has been developed for integrating different types of data in a consistent manner, including uncertainty associated with each dataset [Zhang and Rubin, 2008a, 2008b; Rubin et. al., 2010]. MAD is a general Bayesian method for inverse modeling of spatial random fields that addresses multiple sources and scales of datasets available in characterization, as well as the complex relationships between the data and target properties. In MAD, in addition to typical structural parameters, the parameter vector includes point values of the heterogeneous field, called anchors, which capture local trends and reduce uncertainty in prediction. Different from most of the other inversion methods, MAD yields a posterior distribution of the parameters, which can be directly used in the forward simulations.

Much work within the framework of stochastic hydrogeology has been devoted to evaluating the worth of data. Following Rubin and Dagan [1992], many investigators have examined the impact of conditioning data on groundwater model predictions [McKenna and Poeter, 1995; Maxwell et al., 1999; Scheibe and Chien, 2003; Fu and Gomez-Hernandez, 2007]. Several studies also pursued an optimal design of measurements [Freeze et al., 1992; Nowak et al., 2010]. Most of these studies, however, considered only one or two types of data, and many of them dealt with only synthetic data. Although Scheibe and Chien [2003] evaluated the different types of data at the Oyster Site, it did not include large-scale data, which required inverse modeling.

In this chapter, we investigate the data assimilation strategy with respect to the hydraulic conductivity field at the Integrated Field Research Challenge (IFRC) site in DOEs Hanford 300 Area (http://ifchanford.pnl.gov). The Hanford IFRC project, initiated in 2007, is an interdisciplinary projectinvolving the fields of hydrogeology, geochemistry, and microbiologythat seeks to understand the complex behavior of subsurface uranium transport. Characterizing the hydraulic conductivity field prior to uranium transport experiments is critical for understanding the chemical behavior of uranium in the natural environments. To characterize the three-dimensional hydraulic conductivity field, we conducted a nonreactive tracer experiment in addition to aquifer tests such as electric borehole flowmeter (EBF) and constant-injection tests. All the wells have the lithology data available, which is less informative and often ignored in characterizing a conductivity field [Scheibe and Chien, 2003; Li et al., 2008]. In this analysis, we are particularly interested in examining the impact of lithology data and tracer-test data.

In evaluating the worth of real field data, it is common to compare the predicted and observed breakthrough curves during a tracer test at observation wells not used for inversion [Rubin, 2003; Scheibe and Chien, 2003]. Based on the breakthrough curves, we can determine several performance metricssuch as early arrival time, mean arrival time, and peak concentrationand several performance measures to represent mismatch and uncertainty in prediction, such as bias and variance. Previous work comparing the performance of inverse modeling or worth of data found that inverse-modeling performance depends on the performance metrics and measures [Zimmerman et al., 1998; Scheibe and Chien, 2003; Franssen et al., 2009; Nowak et al., 2010]. It would be most beneficial to evaluate to which metric and measure each type of data can contribute, rather than just determining which data is the best.

This chapter proceeds as follows: first, in Section 4.2, we present the general data assimilation method. After we describe the site and data collection in Section 4.3, we present the assimilation procedure for each dataset in Section 4.4. In Section 4.6, we organize the comparison metrics and comparison measures to evaluate multiple data combinations. In the last section, after we present all the results, we discuss the results and their implications for field characterization strategy.

4.2 A Bayesian sequential data assimilation method

In this section, we briefly describe our Bayesian data assimilation method, the method of anchored distributions (MAD), introduced in Chapter 2. We further generalize the method to accommodate sequential assimilation.

In MAD, we consider a spatial random process denoted by $Y(x)$, where Y is a random variable and x represents a single space coordinate. We further denote an entire field of Y by $\tilde{\boldsymbol{Y}}$, and denote a realization of the field by $\tilde{\boldsymbol{y}}$. The field is defined through a vector of model parameters $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. The $\boldsymbol{\theta}$ part of this vector, called the structural parameter vector, includes a set of parameters designed to capture the global features of $\tilde{\boldsymbol{Y}}$, such as the mean of the field and correlation structures. The $\boldsymbol{\vartheta}$ component of this vector consists of the anchored distributions, or (in short) anchors. Anchors are devices used to capture the local features of the field that cannot be captured using the global parameters represented by $\boldsymbol{\theta}$. The anchors are given in the form of statistical distributions of Y at known or chosen locations.

MAD classifies data into two data types, Type-A and Type-B data. Type-A data have a small support volume, each of which is a function of a point value of Y, while Type-B data are a function of the entire field. We denote all available data by $\boldsymbol{z}$–Type-A data by $\boldsymbol{z}_a$, and Type-B data by $\boldsymbol{z}_b$. MAD defines two types of anchors corresponding to its data classification. Type-A anchors $\boldsymbol{\vartheta}_a$ are co-located with Type-A data, and correlated with Type-A data though a point-scale model such as a petrophysical model or small-scale flow model. Type-B anchors $\boldsymbol{\vartheta}_b$ are placed at selected locations to capture information from Type-B data through inversion. There may not be a one-to-one correspondence between Type-B anchors and Type-B data, since Type-B data is a function of the field. As is discussed in Rubin et al. [2010], we need to place Type-B anchors at the locations sensitive to Type-B data and locations that can capture a local trend of the field. We denote the entire anchor set by $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_a, \boldsymbol{\vartheta}_b\}$. In the MAD method, the goal of the inversion is to determine a posterior distribution of model parameters conditioned on the data, i.e., $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid \boldsymbol{z})$. Once this distribution is determined, we can generate random fields given any random sample of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ from this distribution.

For multiple types of datasets, we consider all the available data divided into subsets as $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N\}$, each of which can be either Type-A or Type-B data. Within the data assimilation scheme, it is more efficient to include each subset, i.e., each dataset, sequentially so that we do not have to re-do the entire assimilation from scratch when new data arrive. We expand the original scheme [Rubin et al., 2010] such that we can assimilate either type of data in any order as long as a certain condition is satisfied.

For the first data z_1, we determine the posterior distribution using Bayes rule:

$$p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_1) \propto p(z_1 \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta}), \tag{4.1}$$

where $p(z_1 \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ is the likelihood of z_1 and $p(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is the prior distribution for the parameter vector before taking any data into account. When the prior pertains only to the structural parameters, we determine the distribution of parameter sets by $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = p(\boldsymbol{\vartheta} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$. To include i-th dataset z_i, $(i = 2, \ldots, N)$, we update the posterior distribution as follows:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_1, z_2, \ldots, z_i) &\propto p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}, z_1, z_2, \ldots, z_{i-1}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_1, z_2, \ldots, z_{i-1}), \\ &= p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_1, z_2, \ldots, z_{i-1}), \end{aligned} \tag{4.2}$$

where $p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ is the likelihood of z_i and $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_1, z_2, \ldots, z_{i-1})$ is the prior distribution with respect to z_i as well as the posterior distribution with the data $\{z_1, z_2, \ldots, z_{i-1}\}$ assimilated. For equality in the second line, we assume the conditional independence between z_i and $\{z_1, z_2, \ldots, z_{i-1}\}$ given parameter set $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$, since the parameter vector $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ contains the spatial information relevant to $\{z_1, z_2, \ldots, z_{i-1}\}$. At any stage of the sequential assimilation, we can generate random fields for prediction. In the stochastic simulations, we sample multiple parameter sets of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ from the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z)$, and generate multiple random fields given set of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$.

The likelihood $p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ $(i = 1, 2, , N)$ is the key to update the posterior distribution for every dataset to be assimilated. For Type-A data, we typically have an analytical form for the likelihood $p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$, using a model-based geostatistical approach [Diggle and Ribeiro 2006]. For Type-B data, we usually need to estimate the likelihood numerically, using numerical simulations of a flow/ transport model, except for some special cases where an analytical form is available [Rubin, 2003].

When the analytical form is not available, we estimate the likelihood of $p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$ using Monte Carlo simulations. Since the model parameters $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ and data z are connected through the field $\tilde{\boldsymbol{Y}}$, we generate multiple conditional realizations of the field $\tilde{y}$ conditioned on a set of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$ sampled from the prior distribution; with each realization, a forward model provides a prediction of z in the form of $\tilde{z}$. As a result of the forward simulations, we obtain the ensemble $\tilde{z}$ for each sample of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}\}$. After estimating the distribution of $\tilde{z}$ parametrically or nonparametrically, we can determine the density at z, which is equivalent to the likelihood $p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta})$.

It is possible to place additional anchors during the sequential assimilation if necessary, depending on the locations or types of newly arrived data. When we assimilate i-th data z_i, we may want to add anchors $\boldsymbol{\vartheta}_i$ to the previous anchors $\boldsymbol{\vartheta}_{i-1}$ for $\{z_1, z_2, \ldots, z_{i-1}\}$, as $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_{i-1}, \boldsymbol{\vartheta}_i\}$. The assimilation of z_i requires modifying Equation (4.2) as follows:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_1, z_2, \ldots, z_i) &\propto p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}, z_1, z_2, \ldots, z_{i-1}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_{i-1} \mid z_1, z_2, \ldots, z_{i-1}), \\ &= p(z_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{i-1}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_{i-1} \mid z_1, z_2, \ldots, z_{i-1}). \end{aligned} \tag{4.3}$$

The joint distribution $p(\boldsymbol{\vartheta}_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{i-1})$ is the prior distribution of $\boldsymbol{\vartheta}_i$. The equality in the second line of Equation (4.3) assumes that both z_i and $\boldsymbol{\vartheta}_i$ are conditionally independent of $\{z_1, z_2, \ldots, z_{i-1}\}$ given $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}_{i-1}\}$. When z_i is Type-A data, $\boldsymbol{\vartheta}_i$ represents Type-A anchors, which have the same locations as z_i. From the definition, z_i is a function of $\boldsymbol{\vartheta}_i$ so that:

$$p(\boldsymbol{\theta}, \boldsymbol{\vartheta} \mid z_1, z_2, \ldots, z_i) \propto p(z_i \mid \boldsymbol{\vartheta}_i) p(\boldsymbol{\vartheta}_i \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{i-1}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_{i-1} \mid z_1, z_2, \ldots, z_{i-1}). \tag{4.4}$$

The assumptions in Equations (4.2) and (4.3) restrict the sequence of adding more datasets and anchors. The newly introduced dataset z_i and anchors $\boldsymbol{\vartheta}_i$ should not be strongly correlated to the datasets previously assimilated $\{z_1, z_2, \ldots, z_{i-1}\}$ given $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}_{i-1}\}$. As Rubin et al. [2010] formulated, the independence strictly holds when we add Type-B data and Type-B anchors after all Type-A anchors, since Type-A data depends on the point values and the Type-A anchors $\boldsymbol{\vartheta}_{i-1}$ carry all the information from Type-A data. When we have enough anchors, it is also valid, since $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}_{i-1}\}$ can capture all the features from $\{z_1, z_2, \ldots, z_{i-1}\}$. In the sequential assimilation, we may need to evaluate this condition every time we introduce additional data and anchors.

4.3 Site and Data Description

4.3.1 Hanford 300Area IFRC Site

The Hanford 300 Area is located at the southern part of the Hanford Nuclear Reservation, one mile north of Richland, Washington, USA. The IFRC site is located within the footprint of a former disposal facility for uranium-bearing liquid waste known as the South Process Pond, 250 m west from the Columbia River. As is shown in Figure 4.1, the triangular well field consists of 25 wells fully screened through the saturated portion of the Hanford formation, and ten wells partially screened at a different depth [Bjornstad et al., 2009].

In this study, we limit ourselves to characterizing the saturated portion of the highly permeable and coarse-grained Hanford formation, which is a shallow unconfined aquifer. The main lithology is a poorly sorted mixture, dominated by gravel up to boulder size, with lesser amounts of sand and silt [Bjornstad et al., 2009]. It overlies the Ringold formation, the upper portion of which is a continuous low-permeability layer consisting of cohesive and compact, well-sorted fine to silty sand. The saturated thickness is variable over the site, ranging from about 5 m to 8 m, with daily and seasonal fluctuations of the water table in response to changes in river stage. The prior estimates of hydraulic conductivity are 1,000–100,000 m/day for the Hanford formation and 0.01–3.00 m/day for the Ringold formation [Meyer et al., 2007].

4.3.2 Hydrogeological Data at the IFRC Site

Constant-rate Injection Test

Fourteen constant-rate injection tests were conducted, each of which used one injection well and 7 to 10 observation wells, as is shown in Figure 4.2(a). All the wells used in the tests were fully screened over the saturated portion of the Hanford formation. The distance between the injection and observation wells ranged between 8 and 60 m, and the injection duration and rate are 20 minutes and 315–318 gpm (1.19–1.20 m^3/min), respectively. A preliminary analysis of the late

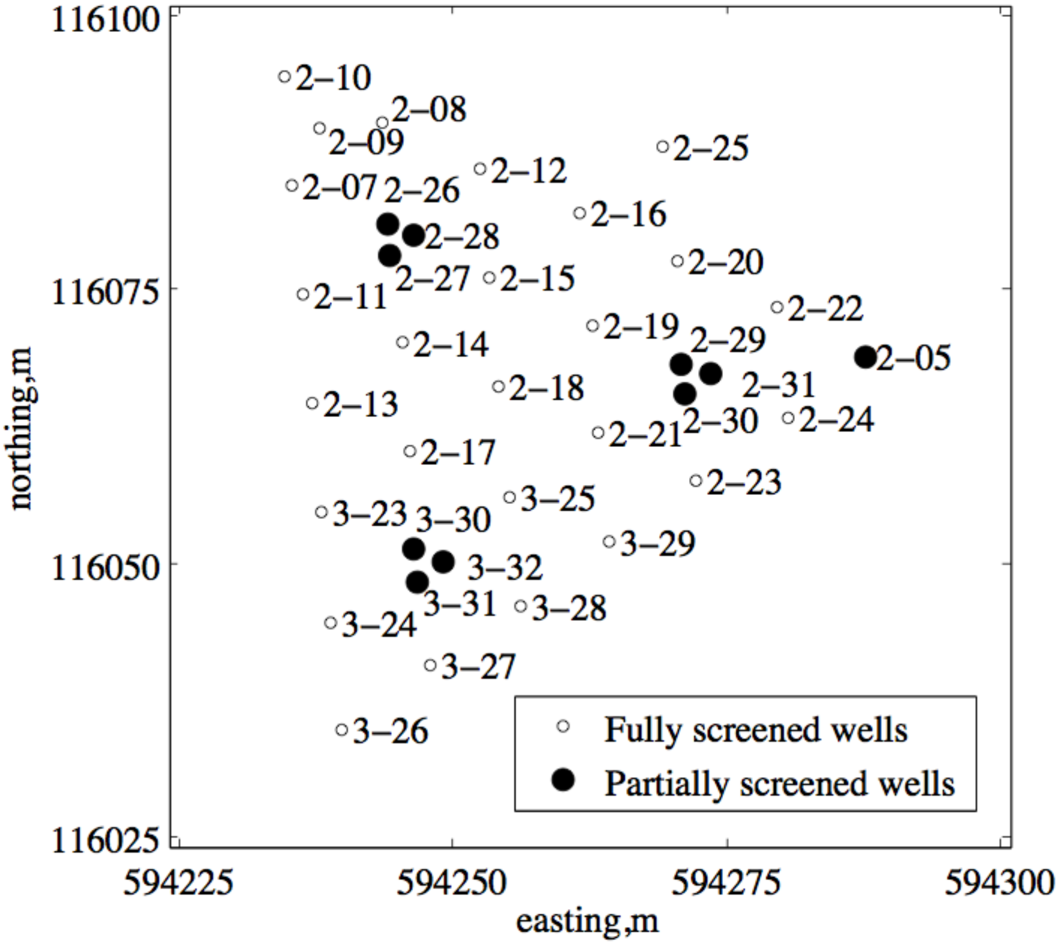
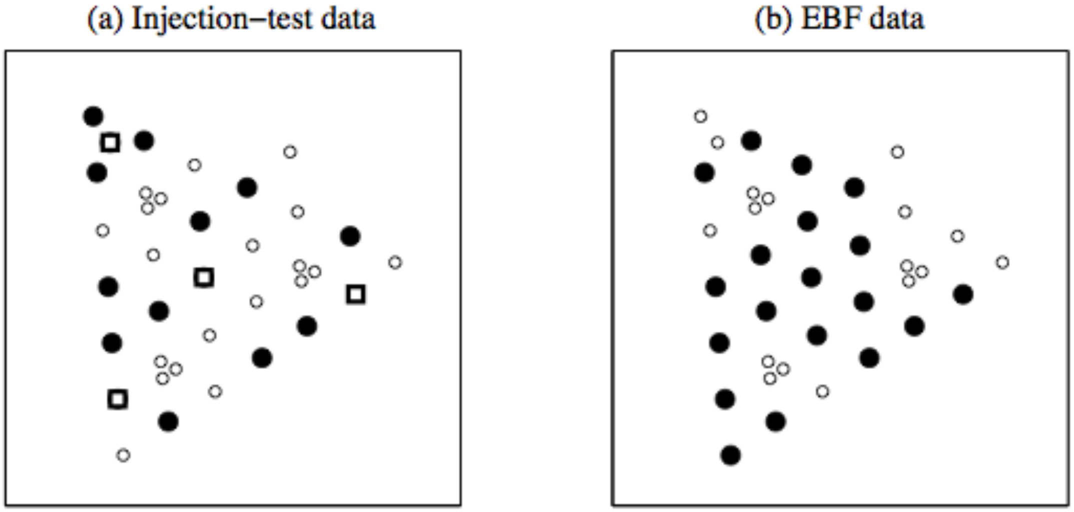


Figure 4.1: Site map of the IFRC site (The coordinate system follows the convention used at the Hanford site).

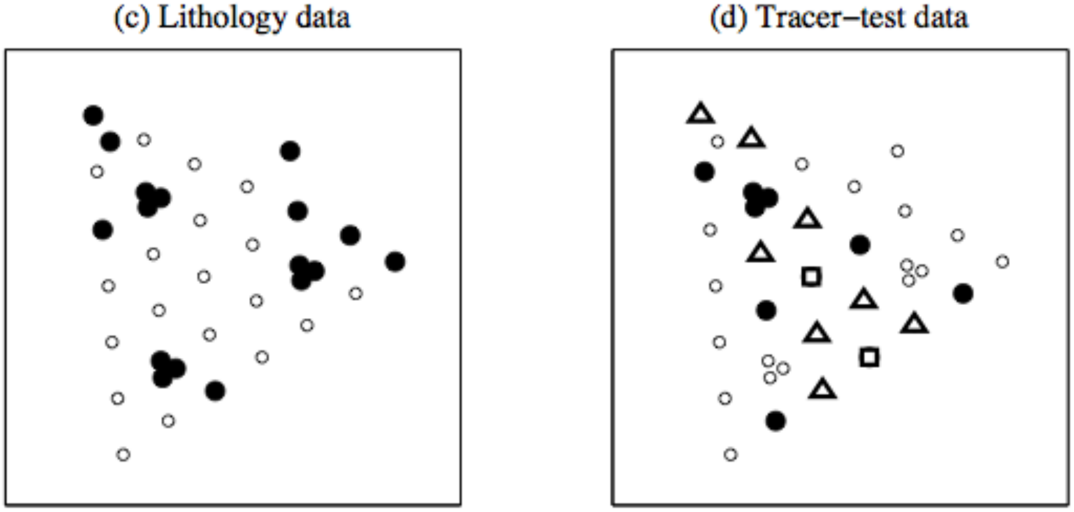
time-curve data, using the Cooper-Jacob straight-line method, showed that most of the observation wells yielded similar estimates for the transmissivity values in each test. As is shown by Sánchez-Villa et al. [1999], this estimate is the geometric mean transmissivity over the entire well field, which suggests that the zone-of-influence expands very rapidly and the conventional pumping test analysis yields only an effective property, smoothing out the local heterogeneity at the well field.

EBF

The EBF tests at the site were conducted at nineteen fully screened wells, which yielded 283 depth-discrete relative hydraulic conductivities with 0.3–0.6 m depth intervals (Figure 4.2b). The pumping rate, 1.04–1.55 gpm ($3.94 \times 10 - 3 - 3.94 \times 10^{-3}$ m^3/min), was kept constant during the test at each well. Vertical profiles indicated that the hydraulic conductivity over the central third of the Hanford formation was lower than at the top and bottom thirds of many of the wells.


(a) Injection−test data

(b) EBF data


(c) Lithology data

(d) Tracer−test data

Figure 4.2: Data locations in the well plot: (a) injection-test data, (b) EBF data, (c) lithology data and (d) tracer-test data. In (a), the black circles are the wells used for observation, and the white squares are the wells used for injection and observation. The subset of the observation wells are used for each injection test. In (d), the black circles are the training-set wells at which M_1 was used, the white squares are the training-set wells at which both M_1 and M_1/M_0 were used, and the white triangles are the testing-set wells.

Although the thickness and contact depths for this lower permeability material varied across the site, this general distribution was observed to some extent at many of the monitoring well locations.

Lithology Data

The lithology profile along each well, the so-called geologist log, was recorded when the well was constructed. Historically, at the Hanford site, the lithology profile has been recorded at nearly every well construction. There has been a significant effort to evaluate and standardize the lithology information by cross-calibrating it to the grain-size analysis [Last et al., 2007].

The dominant lithology at the site is gravelly sand and sandy gravel, in which large variations in hydraulic conductivity (K) were observed during the EBF tests. The EBF tests also indicated higher K in the gravel zone typically found just above the Ringold formation, and lower K in or around the silt zone, which is part of the rip-up clasts of the semiconsolidated, fine-grained Ringold Formation [Bjornstad et al., 2009].

Tracer Test

A nonreactive bromide tracer (average concentration 95 mg/L) was injected into Well 2-09 at a constant rate (71.7 gpm) for 10 hours, which resulted in a total injected volume of 40,600 gallons. The plume was left under natural gradient conditions after the injection stopped. A network of observation wells, including nine depth-discrete wells, monitored tracer concentrations for the next 12 days. Downhole Ion Selective Electrodes (ISE) determined the concentration in situ, calibrated by measuring concentrations of several collected samples with ion chromatography.

Although the March 2009 test was conducted during a relatively stable Columbia River stage, the calculated head gradient and flow direction still showed variations during the monitoring period. The selected wells within the IFRC well field were monitored continuously to measure water levels, with hourly or subhourly measurement frequency, and these measurements were used for setting the boundary conditions in the flow and transport simulations.

4.4 Data Assimilation at the Hanford IFRC Site

In this section, we describe the procedure for assimilating each type of datasets. Although except for the lithology data, all the datasets have been analyzed separately in previous studies, [Murakami et al. 2010; Chen et al., 2010], we briefly repeat the assimilation procedure here for completeness.

4.4.1 Constant-Rate Injection Test

As described in Murakami et al. [2010], we modeled the flow process during the injection tests as a two-dimensional (2-D) planar flow in the horizontal plane to characterize the two-dimensional (2-D) transmissivity (T) field. Such modeling was possible since we observed that the radial flow regime dominated the pressure-buildup responses [Neuman, 1975], and that the maximum pressure buildup at the nearest observation wells was on the order of several centimeters, much smaller than the aquifer thickness (58 m). We transformed the pressure-buildup curves to the zeroth-order moment at each of the multiple observation wells, which allowed us to remove the alias effect from the storage coefficient, as well as reduce computational cost [Li et al., 2004; Zhu and Yeh, 2005].

We constructed a 2-D geostatistical model, assuming that the natural logarithm of T (ln$T$) was multivariate Gaussian, and applied MAD to invert the zeroth-order moments. Since the pumping tests are regarded as large-scale Type-B data, we planted Type-B anchors at selected locations, which represented the local ln$T$ values. We selected four tests to use for inversion, since we observed in the synthetic study that increasing numbers of tests more than four did not change the posterior distribution. As a result, we obtained a joint posterior distribution of 2-D structural parameters and anchors. The distribution of ln$T$ at the EBF-test wells were used to determine the distribution of the absolute K values at the EBF data locations, as described in the next subsection. In addition, we determined the depth-averaged log-conductivity at the wells that did not have EBF data but had the saturation thickness exactly known during the injection tests. The depth-averaged loc-conductivity values, denoted by $\boldsymbol{\vartheta}_{DA}$, were jointly distributed with ln$T$. We used $\boldsymbol{\vartheta}_{DA}$ as Type-A anchors using the point-to-volume covariance [Beherens et al.,1998;Chen et al., 2010].

4.4.2 EBF

As described in Murakami et al. [2010], we assimilated the EBF data after the injection-test data. In the MAD method, the EBF data is Type-A data, owing to its small support volume [Scheibe and Chien, 2003]. We placed Type-A anchors at all the EBF data locations (283 data points), denoted by $\boldsymbol{\vartheta}_{EBF}$. The anchors $\boldsymbol{\vartheta}_{EBF}$ represented the natural logarithm of hydraulic conductivity (ln$K$) at the EBF data locations.

Since the EBF tests yield only relative K values along each of the EBF test wells, local transmissivity T at each of the EBF wells is necessary to convert the relative values to absolute K (Javandel and Witherspoon, 1969; Molz et al., 1994; Young et al., 1998; Fienen et al., 2004). We used the joint distribution of ln$T$ values from the constant-rate injection tests to determine the joint distribution of the absolute K values, i.e., the distribution of Type-A anchors at the EBF data locations [Murakami et al., 2010]. We denote the distribution of $\boldsymbol{\vartheta}_{EBF}$ given the injection-test and EBF data, jointly distributed with $\boldsymbol{\vartheta}_{DA}$, by $p(\boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA} \mid \mathbf{z}_{INJ}, \mathbf{z}_{EBF})$.

As is shown in Figure 4.3(a), we observed that the overall distribution of ln$K$ from the injection and EBF tests is approximately Gaussian, so that they we inferred the distribution of a three-dimensional (3-D) structural parameter vector $\boldsymbol{\theta}$, assuming that ln$K$ follows a multivariate Gaussian distribution. The 3-D structural parameter vector $\boldsymbol{\theta}$ includes mean, variance, horizontal scale, vertical scale, and nugget variance with an exponential covariance model. We determined the joint distribution by $p(\boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA} \mid \mathbf{z}_{INJ}, \mathbf{z}_{EBF}) = p(\boldsymbol{\theta} \mid \boldsymbol{\vartheta}_{EBF}) p(\boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA} \mid \mathbf{z}_{INJ}, \mathbf{z}_{EBF})$, following Equation (4.2). The distribution for the structural parameters of the anchors $p(\boldsymbol{\theta} \mid \boldsymbol{\vartheta}_{EBF})$ was determined analytically using model-based geostatistics [Chapter 6, Diggle and Ribeiro, 2006]. We did not use $\boldsymbol{\vartheta}_{DA}$ for inferring $\boldsymbol{\theta}$, since the depth-average conductivity does not have significant information about the structural parameters.

4.4.3 Lithology Data

We generated the ln$K$ values from the distribution at the EBF locations $p(\boldsymbol{\vartheta}_{EBF} \mid \mathbf{z}_{INJ}, \mathbf{z}_{EBF})$ and analyzed the values in each facies. As is shown in Figure 4.3(b)-(d), the gravel and silt zones have narrow high-ln$K$ and low-ln$K$ distributions, respectively. Others in Figure 4.3(d) includes other sediment classes such as gravelly sand, sandy gravel, and sand, among which we did not observe any significant difference. Last et al. [2007] noted that the geologists distinction among those

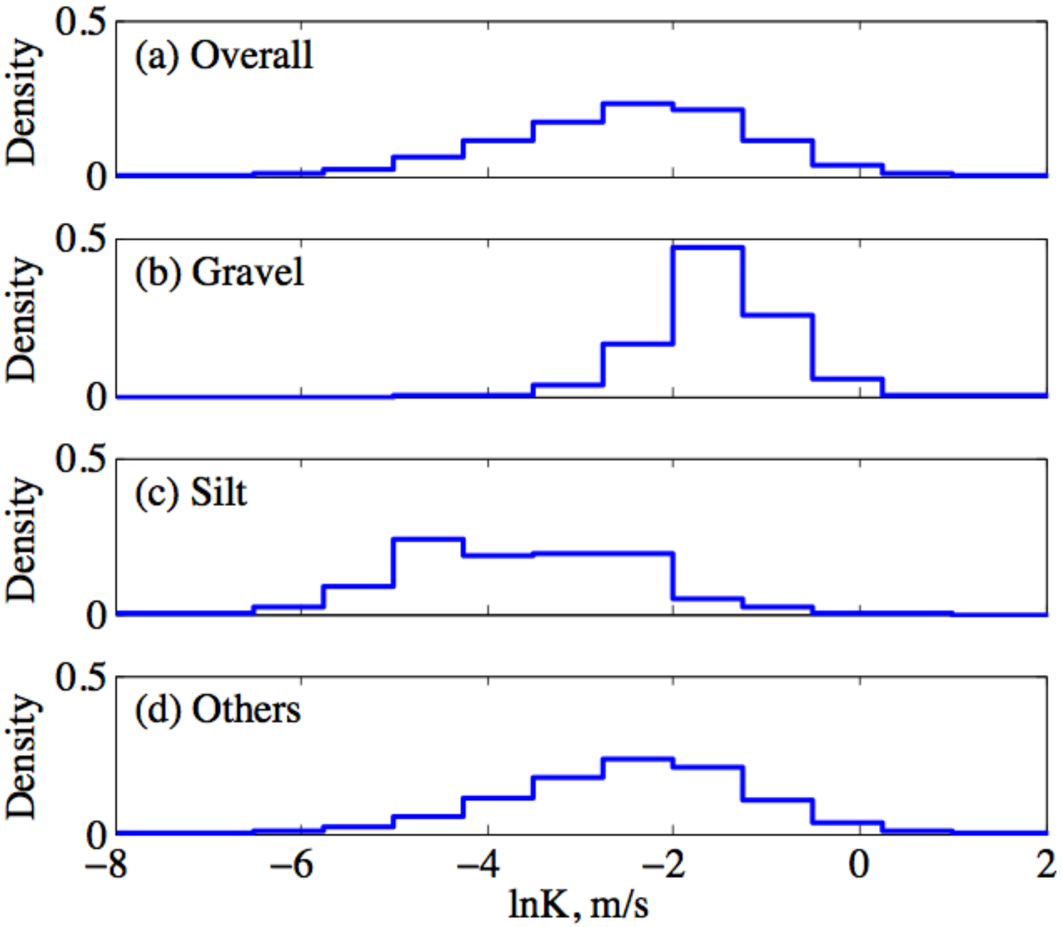


Figure 4.3: Overall distribution of ln$K$ and distributions of ln$K$ in each lithology conditioned on the injection-test and EBF data. Others include sandy gravel, gravelly sand and sand.

classes is sometimes not accurate, which would be another reason to leave those classes together as Others.

Although the lithology information is categorical and contains substantial uncertainty, this information can be useful for the wells that do not have EBF data, especially near the injection well of the tracer test, since Nowak et al. [2010] have shown the importance of characterizing the hydraulic conductivity field near the source. In addition, the lithology information would be helpful in capturing the fast flow path along the centerline of the well plot near the Hanford-Ringold interface. As Chen et al. [2010] noted, this fast flow path was not captured by the EBF and injection tests. Since the sediment class at the bottom of Well 2-26 is gravel, which is the high-K zone, we expect that the lithology data would be helpful in capturing this fast flow path.

To use the lithology data, modelers typically apply an indicator simulation or a transition probability approach, which generates facies over the domain [Carle and Fogg, 1996; Deutsch and

Journel, 1998]. In this study, however, we assumed the Gaussian field for ln$K$, as stated above, since the overall distribution of the ln$K$ values is unimodal and approximately Gaussian, as shown in Figure 4.3(a). In addition, several laboratory-scale experiments have confirmed that the Gaussian ln$K$ field approach still works even for the domain with distinct facies [Nowak et al. 2006; Liu et al., 2007].

MAD allows us to assimilate the lithology data in a consistent manner. Following its data classification scheme, the lithology data (z_L) is Type-A data, which requires additional Type-A anchors ($\boldsymbol{\vartheta}_L$) at the same locations. From the continuous lithology profile, we took data points at every 2 ft in the saturated region along seventeen wells (173 data points), except for several intervals missing in the geologists log. Based on observations in Figure 4.3(b)-(d), we defined three classes {Gravel, Silt, Others} for z_L. To determine the posterior distribution of $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_L\}$, conditioned on the three data sets $\{z_{INJ}, z_{EBF}, z_L\}$, we followed Equation (4.4):

$$p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_L \mid z_{INJ}, z_{EBF}, z_L)$$
$$\propto p(z_L \mid \boldsymbol{\vartheta}_L) p(\boldsymbol{\vartheta}_L \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA} \mid z_{INJ}, z_{EBF}). \tag{4.5}$$

Note that in the posterior distribution, we removed the depth-averaged log-conductivity $\boldsymbol{\vartheta}_D A$, because the lithology data locations overlap the depth-averaged log-conductivity locations. In addition, following a common practice [Maxwell et al., 1999], we updated only $\boldsymbol{\vartheta}_L$ without updating the structural parameters $\boldsymbol{\theta}$, since the lithology data contains substantial uncertainty, which would have an adverse effect on inferring the structural parameters $\boldsymbol{\theta}$.

For determining $p(z_L \mid \boldsymbol{\vartheta}_L)$, we require a model or function to describe the lithology given ln$K$. In our case, this function is site-specific and determined by the ln$K$ distribution in each sedimental class based on the EBF and injection test. In other words, each value of Type-A data z_L is a function of Type-A anchor at the same location, whereas this function, i.e., the shape of the distribution, is determined by the co-located lithology data and ln$K$ values at the EBF locations. We determine this distribution $p(z_{L,j} \mid \boldsymbol{\vartheta}_{L,j})$, i.e., the probability of lithology class given ln$K$ value at j-th lithology data point, by using the generated ln$K$ values based on the EBF and injection tests and applying the nonparameteric kernel density classification algorithm [Section 6.6.2, Hastie et al., 2001].

4.4.4 Tracer Test

As reported in Chen et al. [2010], we placed additional Type-B anchors $\boldsymbol{\vartheta}_{TRS}$ to capture the plume movement for the tracer-test assimilation, as shown in Figure 4.4(a). As the tracer-test data z_{TRC}, we used the first-order temporal moments of the breakthrough curves (M_1) at ten observation wells, and the first-order moments divided by the zeroth-order temporal moments (M_1/M_0) at two wells, as indicated in Figure 4.2(d), following Cirpka and Kitanidis [2003], who analyzed the physical meaning of temporal moments of breakthrough curves. The tracer-test data z_{TRC} is assimilated according to Equation (4.3) as:

$$p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA}, \boldsymbol{\vartheta}_{TRC} \mid z_{INJ}, z_{EBF}, z_{TRC})$$
$$\propto p(z_{TRC} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA}, \boldsymbol{\vartheta}_{TRC}) p(\boldsymbol{\vartheta}_{TRC} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA}) p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA} \mid z_{INJ}, z_{EBF}). \tag{4.6}$$

We estimated the likelihood $p(\boldsymbol{z}_{TRC} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_{DA}, \boldsymbol{\vartheta}_{TRS})$ using nonparametric density estimation, based on the ensemble of $\tilde{\boldsymbol{z}}_{TRC}$ from the forward simulations of flow and transport. A parallel 3-D flow-and-transport code, PFLOTRAN, facilitated the large number of forward simulations on multicores, including the complex site conditions such as water-table fluctuation included in the boundary conditions [Hammond et al., 2010].

As a modification of the original tracer-test inversion reported by Chen et al. [2010], in this chapter, we assimilated the tracer-test data again after the lithology data. Since the ln$K$ values at the lithology data locations, represented by $\boldsymbol{\vartheta}_L$, affect the outcome of the tracer test, the lithology data needs to be assimilated before the tracer-test data in order to satisfy the assumptions in Equations (4.2). Following Equation (4.2), we determine the posterior distribution after the four datasets:

$$\begin{aligned} & p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_L, \boldsymbol{\vartheta}_{TRC} \mid \boldsymbol{z}_{INJ}, \boldsymbol{z}_{EBF}, \boldsymbol{z}_L, \boldsymbol{z}_{TRC}) \\ & \propto p(\boldsymbol{z}_{TRC} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_L) p(\boldsymbol{\vartheta}_{TRC} \mid \boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_L) p(\boldsymbol{\theta}, \boldsymbol{\vartheta}_{EBF}, \boldsymbol{\vartheta}_L \mid \boldsymbol{z}_{INJ}, \boldsymbol{z}_{EBF}, \boldsymbol{z}_L). \end{aligned} \tag{4.7}$$

For implementation, we use the same procedure and the same observation wells as in the original case reported by Chen et al. [2010] (to be consistent). The anchor location is indicated in Figure 4.4(b). This modification can be considered such that adding the lithology data improved the prior distribution with respect to the tracer test.

4.5 Comparison Metrics and Measures

In this chapter, we evaluate the improvement in characterization of the ln$K$ field as we include more datasets–particularly focusing on the impact of lithology data and tracer-test data. Since the real ln$K$ field is unknown, we evaluate the performance of the inversion results based on breakthrough curves (BTCs) during a tracer test not used in the inversion [Scheibe and Chien, 2003]. Using BTCs matches the IFRC project goal as well, since the project will test hypotheses regarding the uranium transport behavior, using the uranium injection experiments with the same observation wells. We may note that although the improvement might depend also on different conceptual models for geostatistical and transport models or different inverse modeling methods, it is beyond the scope of the current study.

Following a common practice in statistics, we divide the tracer data into training-set wells (i.e., observation wells used in inversion) and testing-set wells (i.e., wells used for testing or validating the inversion result). We chose eight wells that have BTCs with significant magnitude as the testing-set wells (Well 2-08, 2-10, 2-14, 2-15, 2-21, 3-25, 2-23, and 3-28), which are indicated in Figure 4.2(d). Among them, we can identify three pairs that have the same distance from the source and are located symmetric to the centerline of the well field (Wells 2-14 and 2-15, Wells 2-21 and 3-25, and Wells 2-23 and 2-28). Since the centerline of the well field is aligned to the mean flow direction averaged over time, an analytical solution of transport based on the structural parameters [Gelhar and Collins, 1971; Datta-Gupta et al., 2002; Bellin and Rubin, 2004] yields identical BTCs for each of those pairs. The difference within each pair is therefore attributed to the local features within the well plot. Comparing within each pair therefore would indicate whether our assimilation method is capturing the local features.

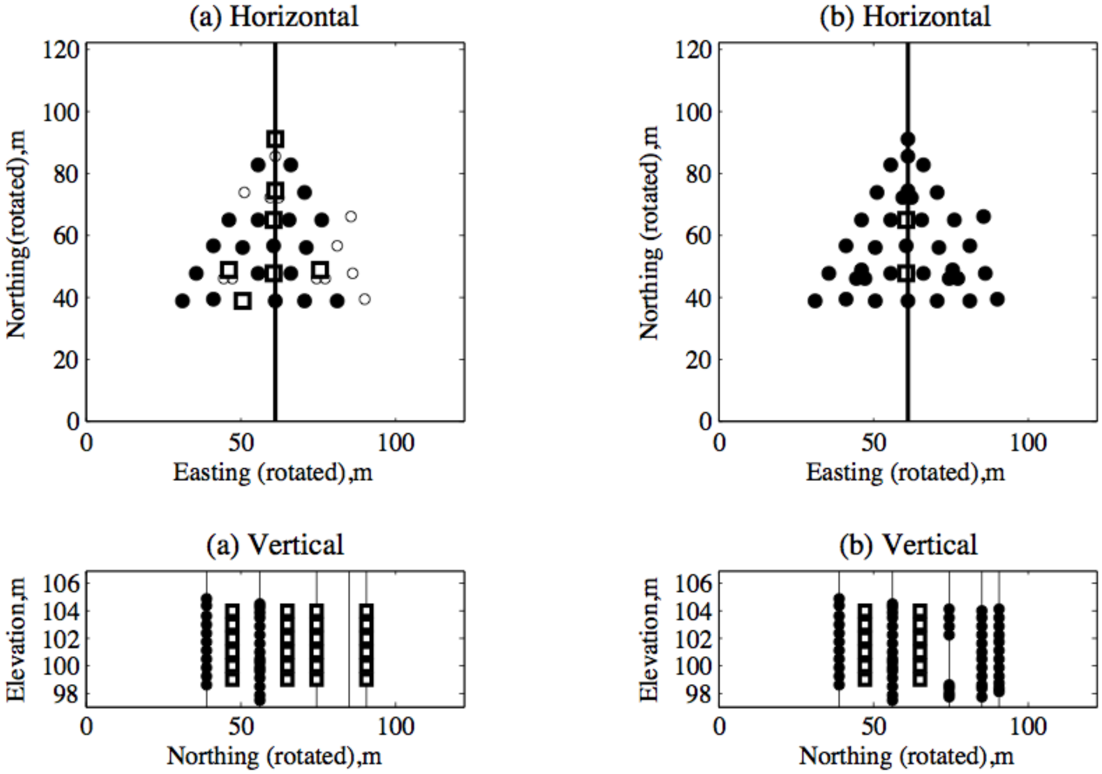


Figure 4.4: Type-A anchor locations (black circles) and Type-B anchor locations (squares) for (a) tracer-test inversion without lithology and (b) tracer-test inversion with lithology. The vertical views are along the centerlines in the horizontal views.

4.5.1 Comparison Cases

We consider the following cases for comparison, with particular interest in the impact of lithology data and tracer-test data.

- Case 1 (ie) : Injection tests and EBF
- Case 2 (ieL): Injection tests, EBF and lithology
- Case 3 (ieT): Injection tests, EBF and tracer test
- Case 4 (ieLT): Injection tests, EBF, lithology and tracer test

In addition to case number, we added the first letters of the data names to make it easier to identify the datasets in each case. We may note that adding the lithology data includes not only the lithology data itself, but also the effect from the EBF and injection test, since these two datasets are used to calibrate ln$K$ in each facies. Adding the lithology data therefore is not equal to the worth of the lithology data itself, but rather represents a combined effect.

4.5.2 Comparison Metrics

In addition to evaluating the entire predicted BTCs, we calculate the single-valued metrics from each BTC to evaluate the improvement. We chose four metrics highly weighted in Scheibe and Chien [2003].
(1) Area under the BTC curve (zeroth-order moment of BTC)
(2) Time to arrival of center of mass (mean arrival time)
(3) Peak arrival time
(4) Peak concentration
We calculate these four metrics for each of the testing-set wells. Among the metrics, area under the BTC is the zeroth-order temporal moment of BTC, denoted by M_0. The time to arrival for center of mass is equivalent to mean arrival time in our case, according to Cirpka and Kitanidis [2000], since the tracer injection time is much shorter than the observation time frame. The mean arrival time is calculated by dividing the first-order temporal moment of BTC by the zeroth-order moment.

4.5.3 Comparison Measures

The stochastic forward simulations yield the ensemble of possible outcomes $\tilde{z}_{test}$ at the testing-set wells conditioned on all the data integrated in each case, denoted by $p(\tilde{z}_{test} \mid z_{test})$. We compare this distribution with the measured outcome z_{test} at the same wells. For comparing the four cases, we calculate the following measures to combine the errors at all the testing-set wells:
(1) Sum of mean squared error (SMSE):

$$\sum_{j=1}^{n} \mathbb{E}[(\tilde{z}_{test,j} - z_{test,j})^2] \tag{4.8}$$

where n is the number of testing-set wells (n = 8), and $\mathbb{E}[\bullet]$ represents the expected value of the quantity inside the bracket. SMSE corresponds to the trace of the error covariance matrix defined by $Cov[\tilde{z}_{test,j} - z_{test,j}]$. It is one of the most common single-valued statistics for representing the overall error in testing multivariate data [Davision, 2003]. Since MSE is the sum of bias square and variance, it can honor both mismatch and uncertainty. We may note that although MSE is one of the common error quantifications in hydrogeological applications [Scheibe and Chien, 2003; Franssen et al., 2009], the definition is often different from Equation (4.8). Here, we define the MSE according to the statistical principle [Davision, 2003].
(2) Sum of absolute errors (SAE):

$$\sum_{j=1}^{n} \mid \mathbb{E}[\tilde{z}_{test,j}] - z_{test,j} \mid \tag{4.9}$$

This is the sum of bias, which measures the mismatch between the mean of the predictive distribution and the true value. SAE is equivalent to mean absolute error used in Franssen et al. [2009], since n does not change in each case.
(3) Sum of variance (SV):

$$\sum_{j=1}^{n} \mathbb{E}[(\tilde{z}_{test,j} - \mathbb{E}[\tilde{z}_{test,j}])^2] \tag{4.10}$$

The sum of variance (SV) represents the magnitude of uncertainty in the prediction. As is noted in Franssen et al. [2009], the smaller SV does not mean good performance, since the distribution may not capture the true value. SV is still useful, however, in evaluating the magnitude of uncertainty in the predictive distributions.

4.6 Results

Based on our four cases, we generated 210,000 realizations of ln$K$ fields and tracer-test predictions (100 realizations for 2,100 parameter sets from the posterior distribution). We have confirmed that the number of realizations and parameter sets is enough to obtain convergence, by gradually adding parameter sets until the distribution stops changing along with the increase in simulations.

4.6.1 Breakthrough curves

Each plot in Figure 4.5 shows the mean and 95-percentile confidence interval (CI) of the predicted BTCs along with the true BTC at each of the four testing-set wells in each of the four cases. We may evaluate the performance based on tightness of the CI around the true BTC, and closeness of the mean BTC to the true one. The mean BTC itself is the averaged value of all the predicted BTCs, so that it does not represent one best-predicted BTC. However, closeness between the mean and the true BTC is a good indication that the predicted BTCs are distributed around the true BTC.

As indicated in Figure 4.5, at Well 2-08, Case 1 (ie) does not capture the true BTC in the CI. Although adding the lithology data in Case 2 (ieL) makes the CI tighter, it has an adverse effect in terms of CI capturing the true BTC. When only the tracer test is added, Case 3 (ieT) has a mean BTC close to the true one, but the CI still does not include the true one. When both datasets are included, however, the CI captures the true BTC.

The lithology data has a greater impact at Well 2-10, such that the CIs become tighter around the true value with the lithology data added. With both datasets, Case 4 (ieLT) has a tight CI and a mean BTC close to the true one. Co-location with the lithology data would be one of the reasons for such a large impact, since the vertical profile of ln$K$ at the well influences the flux-averaged concentration. The lithology data also leads to the improvement at Well 2-23 when combined with the tracer-test data such that with both datasets included, the mean BTC is very close to the true one, even though Well 2-23 is far from the source and not co-located with the lithology data.

At Well 2-15, the CI misses the true curve, even after including the tracer-test data; although the breakthrough time is roughly captured, the predicted magnitude of BTC is lower than the true one. Including the tracer-test data raises the CI slightly, but not enough to capture the true BTC. This might be caused by a very local preferential flow path that was not captured by any of the point-scale datasets. It is known that large-scale data has difficulty capturing highly local features due to the averaging effect in the data. As Scheibe and Chien [2003] pointed out, this also suggests the difficulty of predicting BTCs in the natural environment.

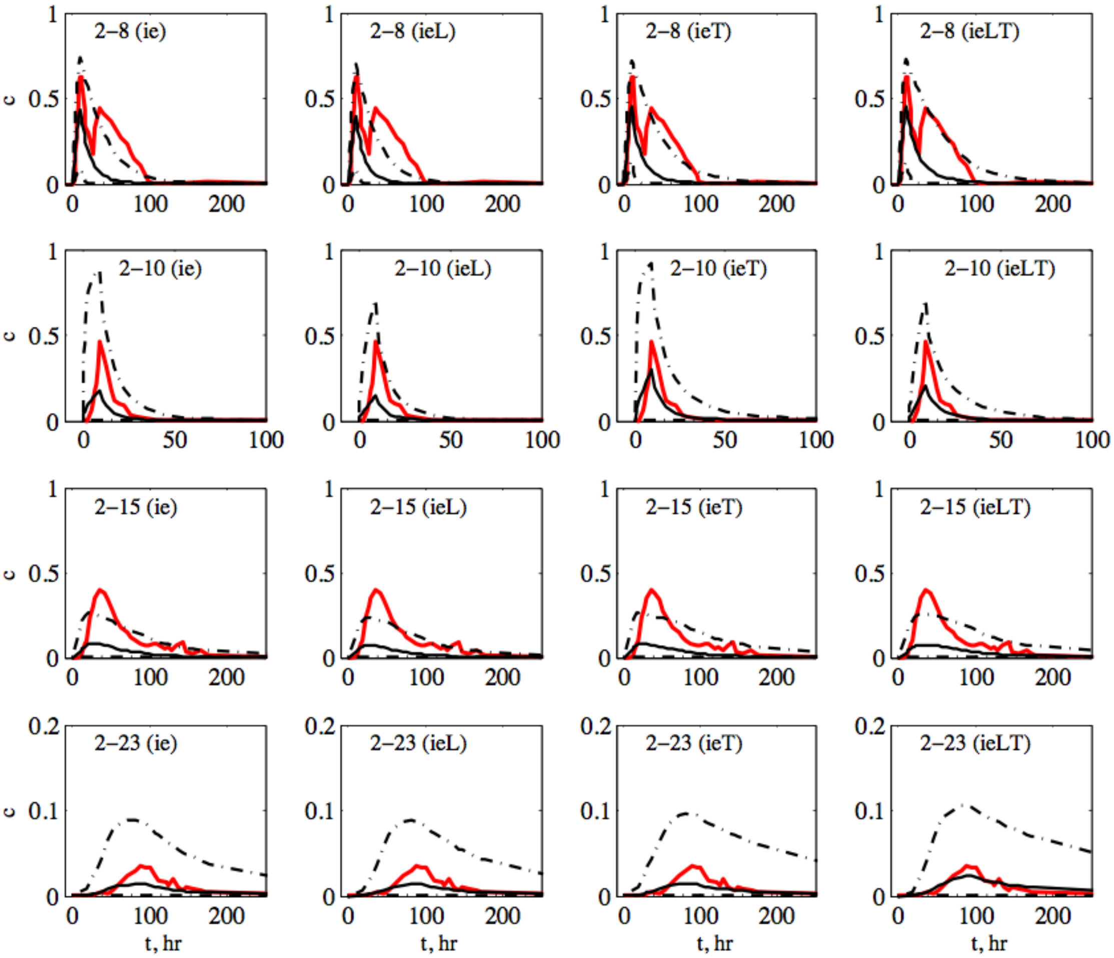


Figure 4.5: Breakthrough curves (BTCs) at Well 2-08, Well 2-10, Well 2-15 and Well 2-23, based on four different combinations of datasets; (ie) injection tests and EBF, (ieL) injection tests, EBF and lithology, (ieT) injection tests, EBF and tracer tests and (ieLT) injection tests, EBF, lithology and tracer tests. In each plot, the black line is the mean BTC, the dash lines are 95% CI and the read line is the observed BTC.

4.6.2 Distribution of Comparison Metrics

We converted each BTC to four comparison metrics, as is described in the previous section. As a result, we obtained the ensemble and predictive distribution for four metrics in four cases. Since we have only eight testing-set wells, we can visualize the CI and mean of each metric at each of the testing-set wells.

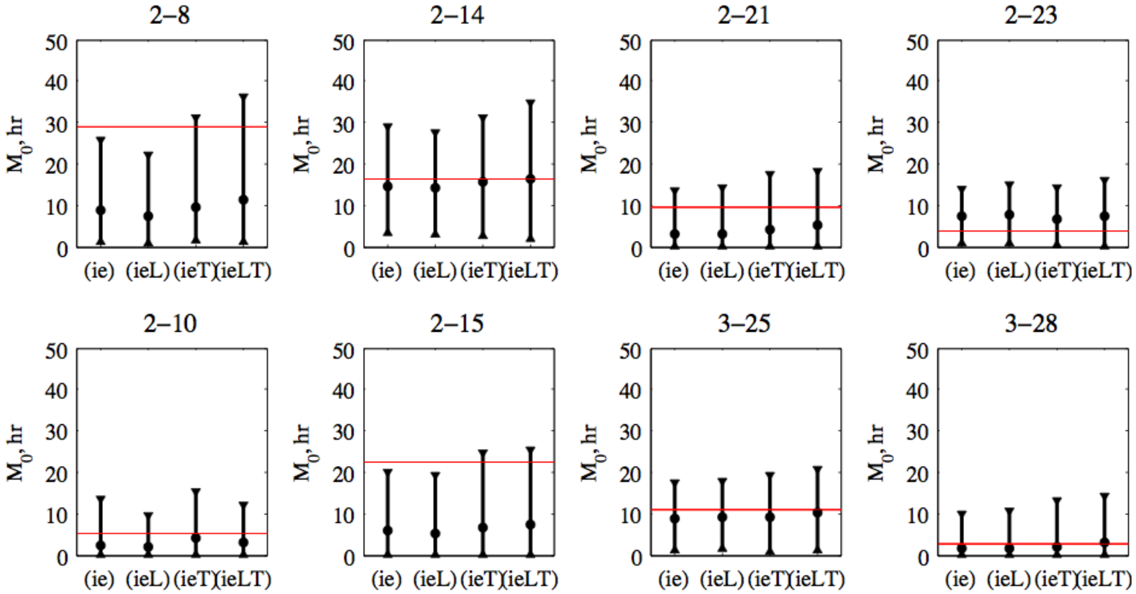


Figure 4.6: 95% CI and mean of M0 at the eight testing-set wells, based on four different combinations of datasets; (ie) injection tests and EBF, (ieL) injection tests, EBF and lithology, (ieT) injection tests, EBF and tracer tests and (ieLT) injection tests, EBF, lithology and tracer tests. In each plot, the black lines are 95% CI, the black circles are the mean, and the red line is the observed value.

Area under the BTC curve

Figure 4.6 shows the change in the mean and CI of the predicted M_0 compared to the true M_0 at eight wells. Each plot includes four cases with different combinations of datasets. We can see that including the tracer-test data, i.e., Case 3 (ieT) and Case 4 (ieTL), increases the width of the CIs at all the wells, which is the direct opposite of our original expectation. However, we may consider this as an improvement, since with tracer-test data added, all the CIs include the true M_0, which are missed in Case 1 (ie) and Case 2 (ieL) at Well 2-08 and Well 2-15. In addition, including the tracer-test data leads to the mean values approaching the true values.

With only the lithology data added in Case 2 (ieL), the CIs become narrower than Case 1 (ie) near the source (Well 2-08, 2-10, 2-14 and 2-15), which implies that the lithology data reduces uncertainty. At some wells, however, the mean values and CIs move away from the true values, compared to the Case 1 (ie), which implies that the lithology data increases mismatch or bias. There is little difference between with-lithology and without-lithology cases at the wells far from the source. With both the lithology and tracer-test data in Case 4 (ieLT), while the CIs are the largest among the four cases at some wells (Well 2-08 and Well 2-14), the mean values are the closest to the true values at most of the wells.

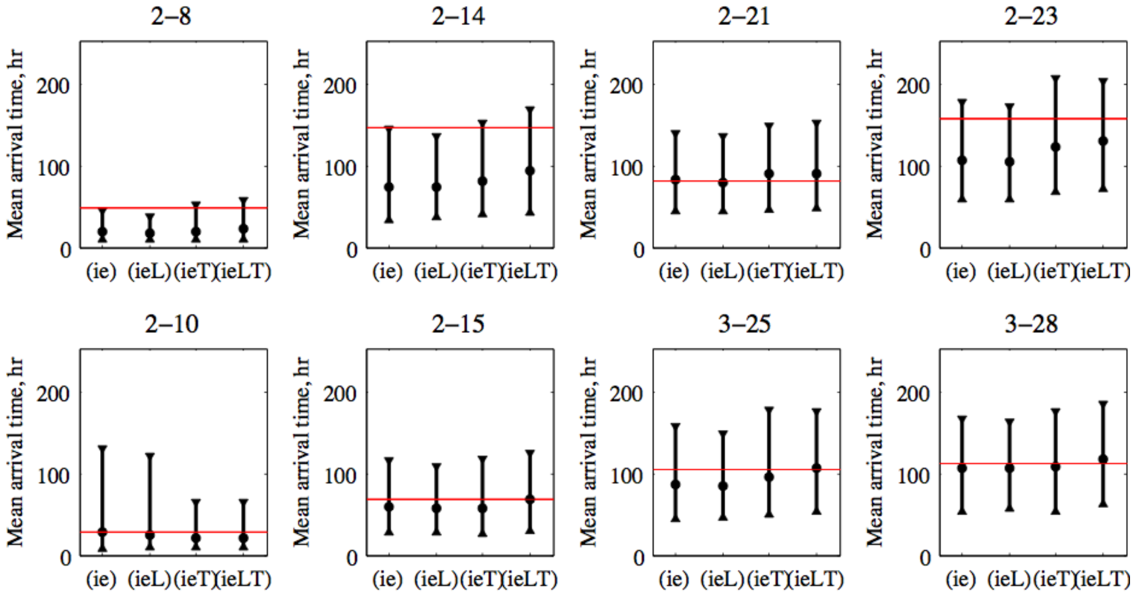


Figure 4.7: 95% CI and mean of mean arrival time at the eight testing-set wells, based on four different combinations of datasets; (ie) injection tests and EBF, (ieL) injection tests, EBF and lithology, (ieT) injection tests, EBF and tracer tests and (ieLT) injection tests, EBF, lithology and tracer tests. In each plot, the black lines are 95% CI, the black circles are the mean, and the red line is the observed value.

Mean Arrival Time

In Figure 4.7, we can see similar features to M_0, such that the tracer-test data moves the mean closer to the true values, and the lithology data reduces uncertainty but increases bias at some wells. Compared to M_0, the CIs do not become wider when the tracer-test data is included except for Well 2-14, at which the without-lithology cases (ie, ieL) do not capture the true values. At Well 2-14, although Case 4 (ieLT) has the largest CI, the mean value is significantly closer to the true one, and the CI contains the true value better than Case 3 (ieT) without the lithology data. We can also see that the CIs capture the difference within each of the equal-distance pairs. For example, Well 2-14 has a much larger observed mean arrival time than Well 2-15; the predicted CIs at Well 2-14 are also located higher than the ones at Well 2-15, even without including the tracer-test data. The other two pairs indicate the same feature. This implies that our assimilation scheme is able to capture the local features of the field to some extent, since it captured the difference in a short distance (i.e., in the 10 m interval between Well 2-14 and 2-15).

Overall, the performance in the mean arrival time is better than M_0, in the sense that the CIs did not become wider when the tracer-test data was added, and the CIs and mean moved towards the true values with increasing data. It would be because mean arrival time is not only less susceptible to errors in other properties, but is also (according to Cirpka and Kitanidis, [2000]) more closely related to ln$K$.

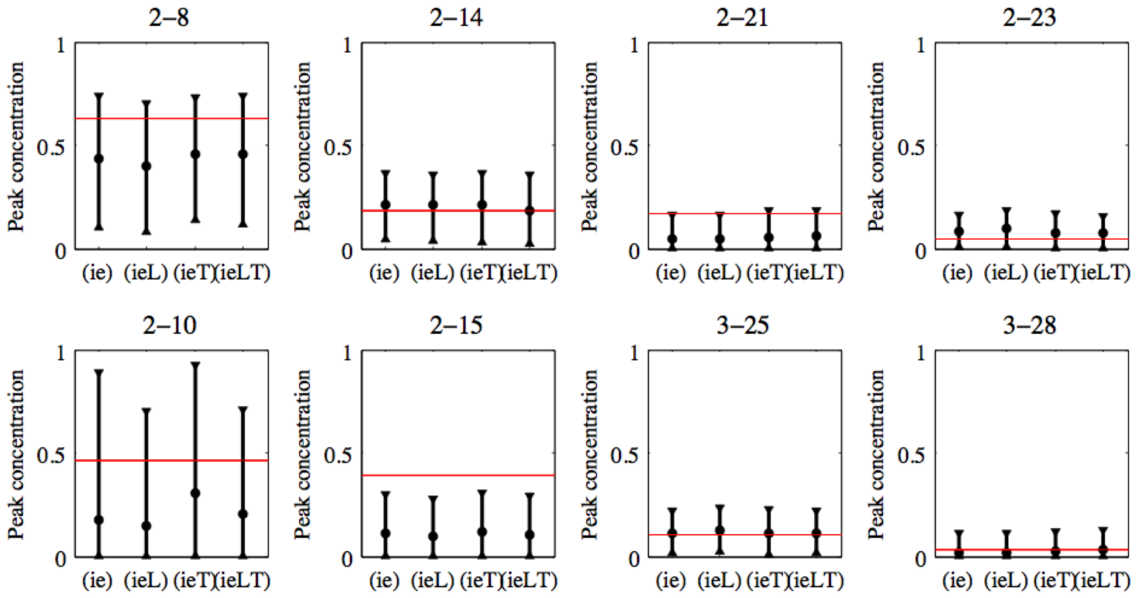


Figure 4.8: 95% CI and average of peak arrival time at the eight testing-set wells, based on four different combinations of datasets; (ie) injection tests and EBF, (ieL) injection tests, EBF and lithology, (ieT) injection tests, EBF and tracer tests and (ieLT) injection tests, EBF, lithology and tracer tests. In each plot, the black lines are 95% CI, the black circles are the mean, and the red line is the observed value.

Peak Arrival Time

In Figure 4.8, the general features are similar to M_0 and the mean arrival time, although the CIs become much larger when the tracer-test test is included, compared to the mean arrival time. Compared to the mean arrival time, the CIs of peak arrival time are smaller at the wells near the source and larger at the wells far from the source. One of the reasons is that the BTCs were truncated at 250 hrs in the simulations, and the peak did not appear in some of the realizations at the wells far from the source. In addition, the mean arrival time includes the effect of BTC over the entire observation time, whereas the peak arrival time does not account for later-time BTCs. The peak time therefore has less uncertainty near the source, whereas the mean arrival time has the similar magnitude of uncertainty at all the wells.

At Well 2-14, the true value is far from the mean values without the tracer-test data in Case 1 (ie) and Case 2 (ieL). When all the datasets are included in Case 4 (ieLT), the CI and mean values move closer to the true values, although there is no difference between Case 1 (ie) and Case 3 (ieT), where the tracer test is included without lithology data. This suggests that the lithology data was not beneficial by itself, but contributed in conjunction with the tracer test.

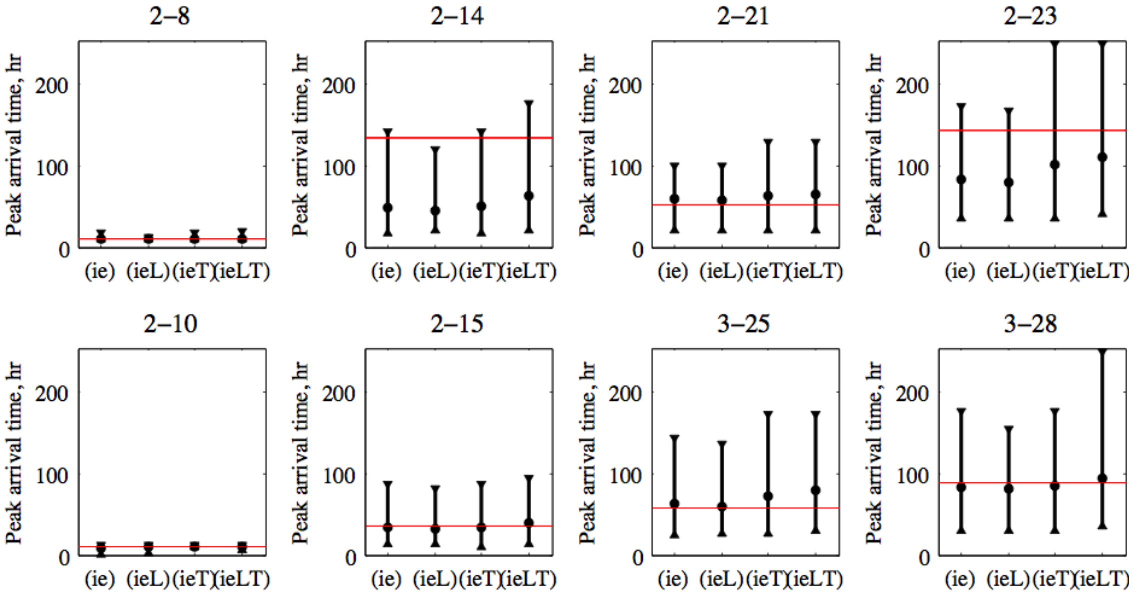


Figure 4.9: 95% CI and average of peak concentration at the eight testing-set wells, based on four different combinations of datasets; (ie) injection tests and EBF, (ieL) injection tests, EBF and lithology, (ieT) injection tests, EBF and tracer tests and (ieLT) injection tests, EBF, lithology and tracer tests. In each plot, the black lines are 95% CI, the black circles are the mean, and the red line is the observed value.

Peak Concentration

In Figure 4.9, the CIs and mean show little difference among all the cases at most of the wells. Including the lithology and tracer-test data does not improve the CIs and the mean values, except that the lithology data slightly reduces the CIs at Well 2-10. At Well 2-21, there is a slight improvement with addition of the tracer-test data, such that the CIs include the true values in Case 3 (ieT) and Case 4 (ieLT). At Well 2-15, the CIs do not capture the true values even with all the datasets included, and the CIs do not change with additional datasets.

Bellin and Rubin [2004], noting that peak concentration was informative for a ln$K$ field, used peak concentration at multiple wells for inferring the structural parameters of the ln$K$ field. However, their framework was limited to the situation where the analytical solution was applicable. In our case, as is mentioned in Chen et al. [2010], local features likely governed peak concentrations. We may see such local features by comparing the peak concentration at Well 2-14 and 2-15. Although they are only 10 m away from each other and at the same distance from the source, Well 2-15 has a peak twice as high as the one at Well 2-14.

4.6.3 Comparison Measures

To compare our four cases with all the testing-set wells combined, Figure 4.10 shows the three comparison measures for each metric that we introduced in the previous section. To evaluate

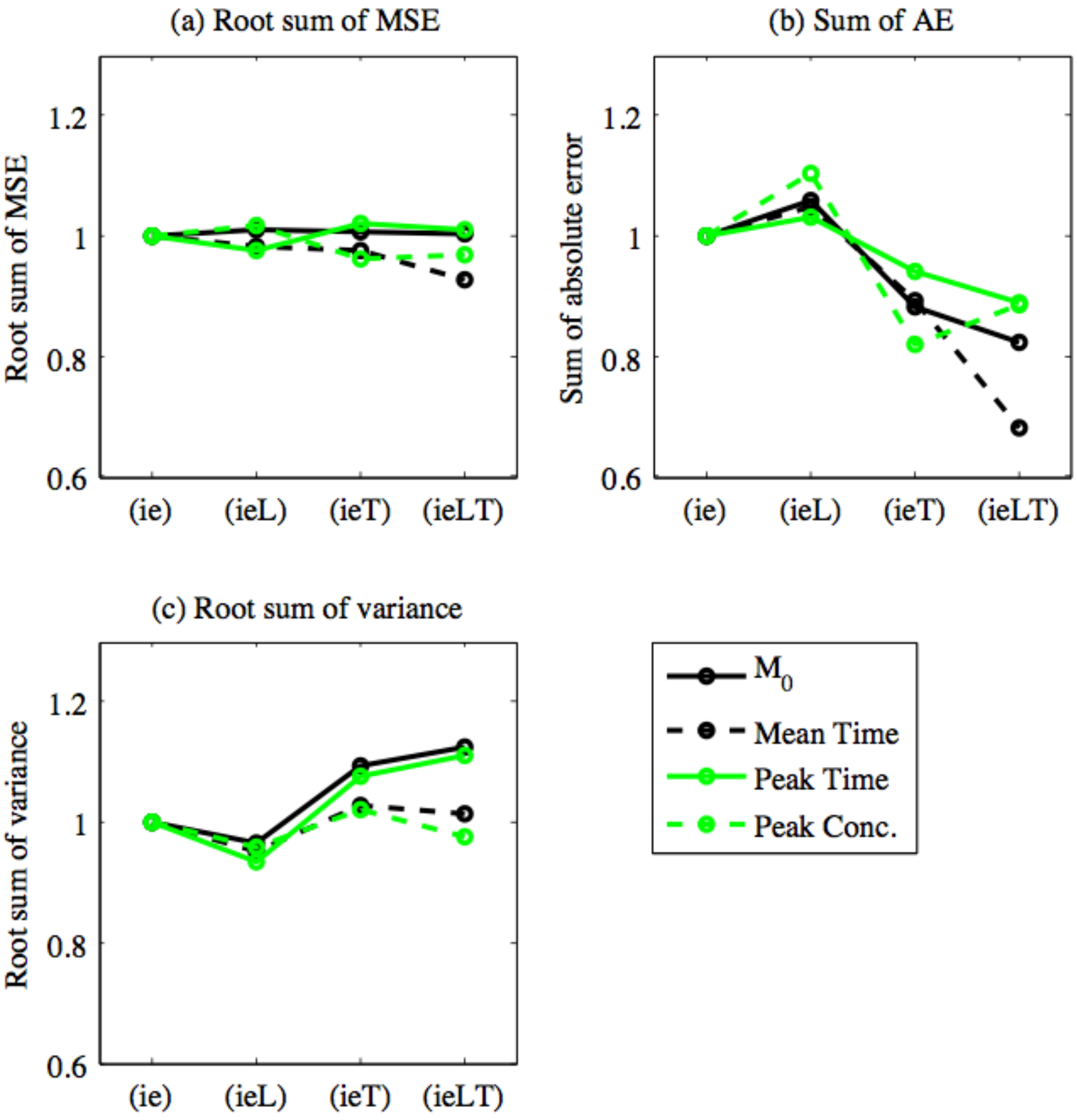


Figure 4.10: Change in the comparison measures with lithology and tracer-test data added: (a) Root sum of MSE, (b) Root sum of AE and (c) Root sum of variance. Each plot has the different combinations of datasets in the same manner as Figure6-9: (ie) injection tests and EBF, (ieL) injection tests, EBF and lithology, (ieT) injection tests, EBF and tracer tests and (ieLT) injection tests, EBF, lithology and tracer tests. The values are normalized to Case 1(ie) in each metric.

the effect of lithology data and tracer-test data, all the values are normalized to Case 1 (ie).

In Figure 4.9(a), the SMSE stays almost the same for all four cases. We can see a slight improvement in mean arrival time, which is consistent with Scheibe and Chien [2003]. The SAE in Figure 4.9(b) shows improvement with the tracer-test data and increased mismatch with the lithology data in all the metrics. With all the data combined, Case 4 (ieLT) has the biggest improvement, except for the peak concentration. The improvement is the most significant for mean arrival time. In Figure 10(c), the SV increases when the tracer-test data are included, and decreases when the lithology data are included. The opposite trend between the SAE and SV leads to a near-constant SMSE, since MSE represents the combined impact of mismatch and uncertainty.

These three measures identified features similar to what we found in the mean values and CIs, since the absolute error is the distance between the mean and true values, and the variance can be interpreted as the width of interval. Since CIs are more informative but cumbersome to analyze, this would imply that these measures are effective for comparing the performance of assimilation.

4.7 Discussion and Conclusion

In this study, we assimilated the constant-rate injection tests, EBF, lithology data, and tracer test at the Hanford IFRC site within a Bayesian assimilation method, the method of anchored distributions (MAD), a recently developed modular-based inversion method. MAD allowed us to directly infer the joint distribution of parameters required for stochastic simulations, and to assimilate multiple types of datasets sequentially by updating the posterior distribution. After we assimilated all the datasets, we evaluated the combined data worth, particularly focusing on the lithology data and tracer-test data. Different from other studies, which focus on the data worth of each data point or each data type, our study focused on the combination of the datasets, since each dataset cannot be separately assimilated. This study provided useful insights into subsurface characterizations, specifically regarding (1) the impact of including lithology data, (2) the impact of including tracer-test data, (3) the differences among performance metrics and (4) the importance of the sequence of assimilating different data types. Each point is discussed further below.

In terms of the lithology data, we observed that including the lithology data made the CIs and the distributions tighter, both in the performance metrics and BTCs. Since the lithology data is point-scale data, adding the lithology data increased the special coverage of point-scale ln$K$ values, which increased constraints on the randomly generated fields and reduced uncertainty. The lithology data by itself, however, sometimes had an adverse effect on prediction, such that the mean values or CIs diverged from the true values.

Including both the lithology and tracer-test data yielded significant improvement. Perhaps paradoxically, lithology data had an adverse effect on prediction without tracer-test data, but a favorable effect when combined with tracer-test data. This is attributed to our anchor concept: each Type-A anchor is not a single value, but rather a distribution. When we have only lithology data, the prediction honors anchor values close to the mean of the distribution, which might be distant from the true values. When we assimilate tracer-test data, the likelihood can honor the values closer to the true value from the underrepresented region of the distribution. In inverse modeling, having more point-scale information, i.e., Type-A data and anchors, not only reduces uncertainty but also avoids or reduces nonuniqueness.

As we discussed in the introduction, the lithology data is often ignored in characterization due to its large uncertainty [Scheibe and Chien, 2003]. We showed that it is useful as long as it is integrated in a consistent manner, with the uncertainty property quantified. In our case, the lithology data is especially effective, since we have sedimental classes with distinct differences in ln$K$. This finding is consistent with Fu and Gomez-Hernandez [2007], who found that the point-scale data reduced uncertainty and led to further improvement when combined with other types of data.

Regarding the tracer-test data, we observed that adding the tracer-test data moved predictive distributions and mean values closer to true values, except for the peak concentration. Although we had expected that the tracer-test data would reduce uncertainty, i.e., reduce the width of CIs and SV, we did not see such an effect at most of the wells. One of the possible reasons for this is that the depth-averaged information at most of the training-set wells did not contribute to reducing the uncertainty in depth-discrete ln$K$ values. Another reason would be that EBF data and lithology data could not capture some local features, which created bias in the prior distribution. The tracer-test data contributed to fixing that bias, but the uncertainty stayed the same or larger (this is known as bias-variance tradeoff). Hou and Rubin (2005) discussed the impact of the prior distribution as well. From the perspective of stochastic modeling, however, we may consider that the tracer-test data contributed significantly, since the 95% CIs captured all the true values, which were missed without the tracer-test data.

Among the four metrics, the mean arrival time showed the best performance, in the sense that additional datasets led to the consistent improvement such as reducing both the SAE and SMSE. This would be because the mean arrival time is related more closely to the ln$K$ field, and is less affected by the errors in the other properties [Cirpka and Kitanidis, 2000]. The peak concentration had the worst performance: the CIs did not change with lithology and tracer-test data, and failed to capture the true values at some wells even with the additional datasets. This finding contrasts with Bellin and Rubin [2004], who used the peak values for inference. In our case, as we can see from the differences between nearby wells, strong local heterogeneity might have affected the peak concentration. In addition, the peak concentration might be more prone to errors in other properties.

Within a sequential data assimilation method, these results would suggest that we must consider the data assimilation procedure with respect not to the worth of each separate set of data, but to their combined effect. The lithology data had an adverse effect by itself, but it contributed significantly when the tracer-test data was included afterwards. In addition, we learned that bias is hard to overcome once it is created, since the posterior distribution from one data type becomes the prior distribution for the next data. In a characterization campaign, therefore, it is important to plan the sequence of assimilation carefully and avoid potential bias. In our particular case, as is discussed in Chen et al. [2010], the problem seems to be that the local preferential flow path was not captured sufficiently by the wells. Introducing geophysical measurements would be beneficial in such a situation, since they could possibly capture local features between the wells before large-scale data are introduced [Scheibe and Chien, 2003; Rubin and Hubbard, 2005].

Chapter 5

Summary

This dissertation describes the development of a Bayesian data assimilation method to integrate multiple data types for probabilistic prediction of groundwater flow and contaminant transport, and application of the method to the real data at the Hanford 300 Area. The method particularly focuses on characterizing a heterogeneous field of hydrogeological properties such as a hydraulic conductivity field, which significantly affects flow and transport processes. The previous three chapters investigated different aspects of the development and application, and presented a number of associated findings.

Chapter 2 introduced a newly developed Bayesian geostatistical data assimilation method, the method of anchor distribution (MAD), for characterizing a heterogeneous field of hydrogeological properties based on various types of datasets. MAD starts by classifying the available data depending on its scale, using two types of anchors (corresponding to two data types), with Type-A data as the small-scale local data defined by a function of a point value of the field, and Type-B data as the large-scale nonlocal data defined by a function of the field. Geostatistical structural parameters describe global features of the field, whereas the anchors capture local trends of the field. The goal for inversion of these data is to obtain a joint posterior distribution of anchor values and structural parameters. The major advantages of MAD are that (1) it is directly connected to stochastic modeling of groundwater flow and transport, since it can infer the joint distribution of the parameters to be directly used as an input for stochastic simulations; (2) it fully quantifies the parametric uncertainty as a statistical distribution; (3) it does not incur overfitting problems, since it uses neither fitting nor optimization procedures, and (4) it has a modular structure, such that it is not limited to any type of data or model, which enables it to be applied to a wide variety of data types and modeling tools.

As indicated in Chapter 2, implementing MAD involves a choice of parameter sampling algorithms–between the weighted resampling method and the Markov-chain Monte-Carlo (MCMC) method. The method itself is sufficiently flexible to accommodate either method, depending on available computational platforms and maturity of each component such as setting a likelihood estimation strategy. The weighted resampling method is suitable for a large-scale parallel computing platform and/or the early stages of developing each component, since we can run as many forward simulations as possible and we can test various likelihood strategies without redoing forward simulations. Methods based on MCMC are more advantageous for a single-core or a small-scale parallel computing platform and the later stages of component development, in which all components are properly fixed.

A synthetic study was conducted to test MAD with MCMC on a simple two-dimensional flow problem. The results showed that the method successfully captured the true parameter values and predicted values within the posterior distribution. It also showed the power of anchors, in that introducing anchors led to a significant improvement by reducing uncertainty and capturing true values in predicting the field and hydraulic head. Comparing the different anchor location settings suggested that the number or location of anchors would depend not only on the physical process but also on the goal of the inverse modeling such as inference of parameters and predictions.

In Chapter 3, the method is applied to a three-dimensional characterization of a heterogeneous hydraulic conductivity field within the Hanford 300 Area, by assimilating large-scale depth-averaged data from constant-rate injection tests with small-scale depth-discrete data from electromagnetic borehole flowmeter (EBF) tests. The method was used first to invert zeroth-order temporal moments of pressure-buildup curves during the injection tests for determining a joint posterior distribution of two-dimensional geostatistical parameters and local log-transmissivities at multiple locations. After combining the distribution of transmissivities with a depth-discrete relative-conductivity profile from the EBF data, the three-dimensional geostatistical parameters of the log-conductivity field were inferred.

The results in Chapter 3 showed that the combination of EBF and injection tests is promising for characterizing a heterogeneous conductivity field in a coarse-grained and highly permeable aquifer, which has been very difficult to characterize by other, conventional techniques. The results showed broader and skewed distributions of geostatistical parameters except for the mean, which suggested the importance of inferring the entire distribution to quantify the parameter uncertainty. One particularly interesting finding was that increasing the number of injection tests led to an improvement up to three or four tests, but improvement beyond that was limited because of measurement errors, and the repeated use and limited number of observation wells. The 3-D characterization was restricted also by the EBF measurement density.

Chapter 4 described expanding data assimilation to include two more data sets: lithology profile tests and tracer tests. The original method was further generalized for assimilating multiple datasets sequentially in any order as long as the new data were conditionally independent of the previously integrated datasets given parameters. The four datasets were assimilated in four different combinations, and then the inversion results were compared among four cases, particularly focusing on the impact of lithology and tracer-test data. In inversion, integrating the lithology data involved the kernel density classification method, and the tracer-test data assimilation used the breakthrough-curve temporal moments for reducing the data dimension. For comparison, four metrics were computed based on the tracer-test breakthrough curves that were not used in the inversion, including zeroth-order temporal moments of breakthrough curves, mean arrival time, peak concentration, and peak time.

In the results described in Chapter 4, the two added datasets each had a different impact: the lithology data reduced uncertainty but sometimes increased bias in the prediction, and the tracer test data did not reduce uncertainty but was beneficial in reducing bias. In addition, the results suggested that it would be necessary to consider the data assimilation procedure with respect not to the worth of each set of data, but to the combined effect, since the combined effect of both data types led to significant improvement in the prediction. Among the comparison metrics, the best and most consistent improvement was found in mean arrival time, and the least improvement was found in peak concentration. One persistent difficulty was that bias is hard to overcome once it is created, since the posterior distribution from one dataset becomes the prior distribution for the next dataset. In any characterization campaign, therefore it is important to avoid potential bias and carefully plan the order of assimilation.

Overall, the method removes the gap between data integration and prediction with respect to parameter uncertainty; more specifically, it allows for directly inferring a joint distribution of parameters to be used as input in stochastic simulations. It can integrate multiple types of datasets in a consistent manner, such that the predictive distribution can reflect the uncertainty associated with each type of dataset, including measurement errors and limited number of measurements. Such a direct connection between data and prediction makes it possible to investigate the worth of each dataset or the combined effect of multiple datasets.

This research revealed some interesting insights for subsurface characterization. First, it is important to characterize the entire distribution of parameters for taking full account of parameter uncertainty in stochastic simulations. Second, we should keep in mind that additional datasets may not always yield what we would usually expect, i.e., smaller uncertainty with the true values captured. Instead, we should consider that (1) repeated experiments of the same kind using subsets of the same observation wells might cause redundancy in information, so that improvement would saturate (severely diminish) above a certain number of tests; (2) configuration of data points or observation wells restricts the available lags and spatial coverage for inferring geostatistical parameters; (3) not only different data types have different effects individually, but also the combined effects could be very different from those individual effects, and (4) in the sequential assimilation, we need to plan the order or sequence carefully, since if one dataset causes a significant bias, it is difficult for additional datasets to correct the bias.

Finally, although the method is applicable for any site and data, the data assimilation procedure is site-specific, in the sense that available data are always different, which requires us to carefully examine the different assumptions and conditions in the flow and transport model. The modular structure of the method, however, very effectively and consistently integrates multiple datasets.

References

Alcolea, A., Carrera, J. and Medina, A.: Pilot points method incorporating prior information for solving the groundwater flow inverse problem, Adv. Water Resour. 29, 1678-1689, 2006.

Andreu C., de Freitas, N., Doucet, A. and Jordan, M. I.: An Introduction to MCMC for Machine Learning, Machine Learning, 50, 5-43, 2003.

Andricevic, R., and Cvetkovic, V.:, Evaluation of Risk from Contaminants Migrating by Groundwater, Water Resour. Res., 32(3), 611621, 1996.

Behrens, R.A., MacLeod, M. K., Tran, T. T. and Alimi, A. O.: Incorporating seismic attribute maps in 3D reservoir models, SPE Reserv. Evalu. Eng., 122Y126, 1998.

Bellin, A., Salandin, P. and Rinaldo, A.: Simulation of dispersion in heterogeneous porous formations: Statistics, first-order theories, convergence of computations, Water Resour. Res., 28(9), 2211-2227, 1992.

Bellin, A., and Rubin, Y.: On the use of peak concentration arrival times for the inference of hydrogeologic parameters, Water Resour. Res., 40, W07401, doi:10.1029/2003WR002179, 2004.

Box, G. E. P., and Tiao, G. C.:, Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading, Mass, USA, 1973.

Bjornstad, B. N., Horner, J. A., Vermuel, V. R., Lanigan, D. C., and Thorne, P. D.: Borehole completion and conceptual hydrolgeologic model for the IFRC Well Field, 300 Area, Hanford Site PNNL-18340, Pacific Northwest National Laboratory, Richland, Washington, 2009.

Carle, S. F. and Fogg, G. E.: Transition probability-based indicator geostatistics, Math. Geology, 28(4), 453-477, 1996.

Carrera, J., and Neuman, S. P.: Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, Water Resour. Res., 22(2), 199-210, 1986a.

Carrera, J., and Neuman, S. P.: Estimation of aquifer parameters under transient and steady state conditions: 2. Uniqueness, stability, and solution algorithms, Water Resour. Res., 22(2), 211-227, 1986b.

Castagna, M. and Bellin, A.: A Bayesian approach for inversion of hydraulic tomographic data, Water Resour. Res., 45, W04410, doi:10.1029/2008WR007078, 2009.

Chapra, S. C., Canale, R. P.: Numerical Methods for Engineers, McGraw-Hill, 4th edition, 2002.

Chen, J., Hubbard, S., and Rubin, Y.: Estimating the hydraulic conductivity at the South Oyester Site from geophysical tomographic data using Bayesian techniques on the normal linear regression model, Water Resour. Res. 37(6), 103-1613, 2001.

Chen, X., Murakami, H., Hahn, M. S., Hammond, G., Rockhold, M. L. and Rubin. Y.: Bayesian geostatistical aquifer characterization at the Hanford 300 Area using tracer test data, in preparation.

Cirpka, O. A. and Kitanidis, P. K.: Characterization of mixing and dilution in heterogeneous aquifers by means of local temporal moments, Water Resour. Res., 36(5), 1221-1236, doi:10.1029/1999WR900354, 2000.

Copty, N., Rubin, Y., and Mavko, G.: Geophysical-Hydrological Identification of Field Permeabilities Through Bayesian Updating, Water Resour. Res., 29(8), 2813-2825, 1993.

Copty, N. K., Trinchero, P., Sanchez-Vila, X., Sarioglu, M. S. and Findikakis, A. N.: Influence of heterogeneity on the interpretation of pumping test data in leaky aquifers, Water Resour. Res., 44, W11419, doi:10.1029/2008WR007120, 2008.

Cooley, R. L., An analysis of the pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields, Water Resour. Res. 36(4), 1159-1163, 2000.

Cooley, R. L., and Hill, M.C.: Comment on Ramarao et al. [1995] and LaVenue et al. [1995], Water Resour. Res., 36(9), 2795-2797, 2000.

Dagan, G.: Flow and Transport in Porous Formations, Springer Verlag, Berlin, Germany, 1989.

Dagan, G.: Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem, Water Resour. Res., 21(1), 65-72, 1985.

Dagan, G., Lessoff, S. C., and Fiori, A.: Is transmissivity a meaningful property of natural formations? Conceptual issues and model development, Water Resour. Res., 45, W03425, doi:10.1029/2008WR007410, 2009.

Dai, Z., Ritzi, R. W., Huang, C., Rubin, Y. and Dominic, D. F.: Transport in heterogeneous sediments with multimodal conductivity and hierarchical organization across scales, J. Hydrol., 294(1-3), 68-86, 2004.

Datta-Gupta, A., Yoon, S., Vasco, D. W. and Pope, G. A.: Inverse modeling of partitioning interwell tracer tests: A streamline approach, Water Resour. Res., 38(6), 1079, doi:10.1029/2001WR000597, 2002.

De Barros, F. P. J., Rubin, Y., and Maxwell, R. M.: The concept of comparative information yield curves and its application to risk-based site characterization, Water Resour. Res., 45, W06401, doi:10.1029/2008WR007324, 2009.

Deutsch, C. V. and Journel, A. G.: GSLIB: Geostatistical Software Library and User's Guide (2nd ed.), Oxford University Press, New York, 1998.

Diggle, P. J. and Ribeiro, P. J.: Bayesian inference in Gaussian model-based geostatistics, Geographical and Environmental Modeling, 6(2), 129-146, doi:10.1080/1361593022000029467, 2002.

Diggle, P. J. and Ribeiro, P. J.: Model-based geostatistics, Springer, New York, USA, 2006.

Ezzedine, S., Rubin, Y. and Chen, J.: Hydrogeological geophysical Bayesian method for subsurface site characterization: Theory and application to the LLNL Superfund Site, Water Resour. Res, 35(9), 2671-2683, 1999.

Fienen, M. N., Kitanidis, P. K., Walton, D., and Jardine, P.: An application of Bayesian inverse methods to vertical deconvolution of hydraulic conductivity in a heterogeneous aquifer at Oak Ridge National Laboratory, Math. Geol., 36(1), 101-126, 2004.

Firmani, G., Fiori, A., and Bellin, A.: Three-dimensional numerical analysis of steady state pumping tests in heterogeneous confined aquifers, Water Resour. Res., 42, W03422, doi:10.1029/2005WR004382, 2006.

Franssen, H. J. H., Alcolea, A., Riva, M., Bakr, M., van der Wiel, N., Stauffer, F. and Guadagnini, A.: A comparison of seven methods for the inverse modeling of groundwater flow. Application to the characterization of well catchments, Adv. Water Resour., 32 (6), 851-872, doi: 10.1016/j.advwatres.2009.02.011, 2009.

Freeze, R. A., James, B., Massmann, J., Sperling, T. and Smith, L.: Hydrogeological Decision Analysis: 4. The Concept of Data Worth and Its Use in the Development of Site Investigation Strategies. Ground Water, 30: 574-588. doi: 10.1111/j.1745-6584.1992.tb01534.x, 1992.

Fu, J. and Gmez-Hernndez, J. J.: Uncertainty assessment and data worth in groundwater flow and mass transport modeling using a blocking Markov chain Monte Carlo method, J. of Hydrol., 364, 328-341, doi: 10.1016/j.jhydrol.2008.11.014, 2009.

Gamerman, D. and Lopes, H. F.: Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference. Chapman&Hall/CRC, 2nd edition, Boca Raton, USA, 2006.

Gelhar, L. W., and Collins, M.: General analysis of longitudinal dispersion in nonuniform flow, Water Resour. Res., 7, 1511-1521, 1971.

Hammond, G.E., Lichtner, P.C. and Rockhold, M.L.: Stochastic simulation of uranium migration at the Hanford 300 Area, J. of Contam. Hydrol., doi:10.1016/j.jconhyd.2010.04.005, in press, 2010.

Handcock, M. S. and Stein, M. L.: A Bayesian analysis of kriging, Technometrics, 35(4), 403-410, 1993.

Hastie, T., Tibshirani, R. and Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction, Springer, New York, USA, 2001.

Hayfield, T. and Racine, J. S.: R: Nonparametric econometrics: the np package, Journal of Statistical Software, 27(5), 1-32, 2008.

Hernandez, A. F., Neuman, S. P., Guadagnini, A. and Carrera, J.: Inverse stochastic moment analysis of steady state flow in randomly heterogeneous media, Water Resour. Res., 42, W05425, doi:10.1029/2005WR004449, 2006.

Hoeksema, R. J., and Kitanidis, P. K.: An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling, Water Resour. Res., 20(7), 1003-1020, 1984.

Hou, Z. and Rubin, Y.: On minimum relative entropy concepts and prior compatibility issues in vadose zone inverse and forward modeling, Water Resour. Res., 41, W12425, doi:10.1029/2005WR004082, 2005.

Hou, Z., Rubin, Y., Hoversten, G.M., Vasco, D. and Chen, J.: Reservoir Parameter Identification Using Minimum Relative Entropy-Based Bayesian Inversion of Seismic AVA and Marine CSEM Data, Geophysics, 71(6), P.O77-O88, Nov-Dec, 2006.

Hoversten, G.M., Cassassuce, F., Gasperikova, E., Newman, G. A., Rubin, Y., Hou, Z. and Vasco, D.: Direct reservoir parameter estimation using joint inversion of marine seismic AVA & CSEM data, Geophysics, Volume 71, Issue 3, pp. C1-C13, May-June, 2006.

Hubbard S., Rubin, Y. and Majer, E.: Spatial correlation structure estimation using geophysical and hydrogeological data, Water Resour. Res., 35(6), 1809-1825, 1999.

Javandel, I. and Witherspoon, P. A.: A method of analyzing transient fluid flow in multilayered aquifers, Water Resour. Res., 5, 856-869, 1969.

Janssen, G. M. C. M., Valstar, J. R. and van der Zee, S. E. A. T. M.: Measurement network design including travel time determinations to minimize model prediction uncertainty, Water Resour. Res., 44, W02405, doi:10.1029/2006WR005462, 2008.

Jeffreys, H.: An invariant form for the prior probability in estimation problems, Proc. R. Soc. Lon. Ser.-A, 186, 453-461, 1946.

Kass, R.E., and Wasserman, L.: The selection of prior distributions by formal rules, J. Am. Stat. Assoc., 91(435), 1343-1370, 1996.

Kitanidis, P. K., and Vomvoris, E. G.: A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, Water Resour. Res., 19(3), 677-690, 1983.

Kitanidis, P. K.: Parameter uncertainty in estimation of spatial functions: Bayesian analysis, Water Resour. Res., 22(4), 499-507, 1986.

Kitanidis, P. K.: Quasi-linear geostatistical theory for inversing, Water Resour. Res., 31(10), 2411-2419, 1995.

Kitanidis, P. K.: Introduction to Geostatistics: Applications in Hydrogeology, Cambridge University Press., New York, USA, 1997.

Kowalsky, M. B., Finsterle, S. and Rubin, Y.: Estimating flow parameter distributions using ground-penetrating radar and hydrological measurements during transient flow in the vadose zone, Adv. Water Resour., 27, 583-599, 2004.

Krommer, A. R. and Ueberhuber, C. W.: Computational integration, Society for Industrial and Applied Mathematics, Philadelphia, 1998.

Last, G. V., Murray, C. J., Bush, D. A., Sullivan, E. C., Rockhold, M. L., Mackley, R. D. and Bjornstad, B. N.: Standardization of Borehole Data to Support Vadose Zone Flow and Transport Modeling, Vadose Zone Journal, 6(4), 906-912, doi:10.2136/vzj2006.0175, 2007.

Li, W., Nowak, W., and Cirpka, O. A.: Geostatistical inverse modeling of transient pumping tests using temporal moments of drawdown, Water Resour. Res., 41, W08403, doi:10.1029/2004WR003874, 2005.

Li, W., Englert, A., Cirpka, O. A., and Vereecken, H.: Three-dimensional geostatistical inversion of flowmeter and pumping test data, Ground Water, 46(2), 193-201, 2008.

Liu, X., Illman, W. A., Craig, A. J., Zhu, J., and Yeh, T.-C. J.: Laboratory sandbox validation of transient hydraulic tomography, Water Resour. Res., 43, W05404, doi:10.1029/2006WR005144, 2007.

McLaughlin, D., and Townley, L. R.: , A Reassessment of the Groundwater Inverse Problem, Water Resour. Res., 32(5), 11311161, doi:10.1029/96WR00160,1996.

Mavko, G., Mukerji, T. and Dvorkin, J.: The Rock Physics Handbook, Cambridge University Press, New York, USA, 1998.

Maxwell, R. M. and Kastenberg, W. E.: Stochastic environmental risk analysis: An integrated methodology for predicting cancer risk from contaminated groundwater, Stoch. Environ. Res. Risk Assess., 13(1-2), 27-47, 1999.

Maxwell, R., Kastenberg, W. E., and Rubin, Y.: A methodology to integrate site characterization information into groundwater-driven health risk assessment, Water Resour. Res., 35(9), 2841-2885, 1999.

McKenna, S.A., and Poeter, E. P.: Field example of data fusion in site characterization, Water

Resour. Res., 31(12), 3229-3240, 1995.

McLaughlin, D., Reid, L. B., Li, S. G., and Hyman, J.: A stochastic method for characterizing groundwater contamination, Ground Water, 31(2), 237-249, 1993.

Meyer, P. D., Ye, M., Rockhold, M. L., Neuman, S. P., and Cantrell, K. J.: Combined estimation of hydrogeologic conceptual model, parameter, and scenario uncertainty with application to uranium transport at the Hanford Site 300 Area, PNNL-16396, Pacific Northwest National Laboratory, Richland, WA, 2007.

Michalak, A.M., and Kitanidis, P. K.: A method for enforcing parameter nonnegativity in Bayesian inverse problems with an application to contaminant source identication, Water Resour. Res., 39(2), doi:10.1029/2002WR001480, 2003.

Molz, F. J., Boman, G. K., Young, S. C., and Waldrop, W. R.: Borehole flowmeters: Field application and data analysis, J. Hydrol., 163(4), 347-371, 1994.

Murakami, H., Chen, X., Hahn, M. S., Liu, Y., Rockhold, M. L., Vermeul, V. R., Zachara, J. M., and Rubin, Y.: Bayesian approach for three-dimensional aquifer characterization at the Hanford 300 Area, Hydrol. Earth Syst. Sci., 14, 1989-2001, doi:10.5194/hess-14-1989-2010, 2010.

Mugunthan, P., and Shoemaker, C. A.: Assessing the impacts of parameter uncertainty for computationally expensive groundwater models, Water Resour. Res., 42, W10428, doi:10.1029/2005WR004640, 2006.

Neuman, S. P.: Analysis of pumping test data from anisotropic unconfined aquifers considering delayed gravity response, Water Resour. Res., 11(2), 329-342, 1975.

Newcomer, D. R.: 300 Area VOC Program Slug Test Characterization Results for Selected Test/Depth Intervals for Wells 399-2-5, 399-3-22, and 399-4-14, PNNL-17439, Pacific Northwest National Laboratory, Richland, Washington, 2008.

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

Nowak, W. and Cirpka, O. A: Geostatistical inference of hydraulic conductivity and dispersivities from hydraulic heads and tracer data, Water Resour. Res., 42, W08416, doi:10.1029/2005WR004832, 2006.

Nowak, W., de Barros, F. P. J. and Rubin, Y.: Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain, Water Resour. Res., 46, W03535, doi:10.1029/2009WR008312, 2010.

Poeter, E.P., and Hill, M.C.: Inverse models: A necessary next step in ground-water models, Ground Water, Volume 35(2), 1997.

Ramanathan, R., Ritzi, R. W. and Huang, C.C.: Linking hierarchical stratal architecture to plume spreading in a Lagrangian-based transport model, Water Resour. Res., 44, W04503, doi:10.1029/2007WR006282, 2008.

R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2007.

Ribeiro, P. J. and Diggle P. J.: geoR: a package for geostatistical analysis R-NEWS, 1(2), 15-18, 2001.

Ritzi, R. W., Dai, Z., Dominic, D. F. and Rubin, Y.: Spatial correlation of permeability in cross-stratified sediment with hierarchical architecture, Water Resour. Res., 40, W03513, doi:10.1029/2003WR002420, 2004.

Robert, C. P. and Casella, G: Monte Carlo statistical methods (second edition), Springer, Berlin, Germany, 2005.

Rockhold, M. L., Vermeul, V. R., Mackley, R. D., Fritz, B. G., Mendoza, D. P., Newcomer, E. M., Newcomer, D. R., Murray, C. J., and Zachara, J. M.: Hydrogeologic characterization of the Hanford 300 Area Integrated Field Research Challenge site and numerical modeling of the first aquifer tracer Test, Ground Water, submitted, 2010.

Rubin, Y. and Dagan. G.: Stochastic identification of transmissivity and effective recharge in steady groundwater flow: 1. Theory, Water Resour. Res., 23(7), 1185-1192, 1987a.

Rubin, Y. and Dagan, G. : Stochastic identification of transmissivity and effective recharge in steady groundwater flow: 2. Case study, Water Resour. Res., 23(7), 1193-1200, 1987b.

Rubin, Y. and Dagan, G.: Stochastic analysis of the effects of boundaries on spatial variability in groundwater flows: 1. Constant head boundary, Water Resour. Res., 24(10), 1689-1697, 1988.

Rubin, Y, and Dagan, G.: Conditional estimation of solute travel time in heterogeneous formations: Impact of transmissivity measurements, Water Resour. Res., 28(4), 1033-1040, 1992.

Rubin, Y., and Seong, K.: Investigation of flow and transport in certain cases of nonstationary conductivity fields, Water Resour. Res., 30(11), 2901-2912, 1994.

Rubin, Y.: Flow and transport in bimodal heterogeneous formation, Water Resour. Res., 31(10), 24612468, 1995.

Rubin, Y. and Ezzedine S.: The travel times of solutes at the Cape Cod tracer experiment: data analysis, modeling, and structural parameters inference, Water Resour. Res. 33(7), 1537-1547, 1997.

Rubin, Y., Sun, A., Maxwell, R. and Bellin, A.: The concept of block effective macrodispersion, J. Fluid Mech, 395, 161-180, 1999.

Rubin, Y., Bellin, A. and Lawrence, A.: On the use of block-effective macrodispersion for numerical simulation of transport in heterogeneous formations, Water Resour. Res., 39(9), 1242, doi:10.1029/2002WR001727,2003.

Rubin, Y: Applied Stochastic Hydrogeology, Oxford Univ. Press, Oxford, UK, 2003.

Rubin Y. and Hubbard, S.: Hydrogeophysics, Springer, Netherlands, Water and Science Technology Library 50, 523 p., 2005.

Rubin, Y., Lunt, I. A. and Bridge, J. S.: Spatial variability in river sediments and its link with river channel geometry, Water Resour. Res., 42, W06D16, doi:10.1029/2005WR004853, 2006.

Rubin, Y., Chen, X., Murakami, H., and Hahn, M. S.: A Bayesian approach for inverse modeling, data assimilation and conditional simulation of spatial random fields, Water Resour. Res, 46, W10523, doi:10.1029/2009WR008799, 2010.

Rubin, Y., Chen, X., Murakami, H., and Hahn, M. S.: A Bayesian approach for inverse modeling, data assimilation and conditional simulation of spatial random fields, Water Resour. Res, 46, W10523, doi:10.1029/2009WR008799, 2010.

Sanchez-Vila, X., Meier, P. M., and Carrera, J.: Pumping tests in heterogeneous aquifers: An analytical study of what can be obtained from their interpretation using Jacob's method, Water Resour. Res., 35(4), 943-952, 1999.

Scheibe, T. D., and Freyberg, D. L.: Use of sedimentological information for geometric simulation of natural porous media structure, Water Resour. Res., 31(12), 3259-3270, 1995.

Scheibe, T. D., Chien, Y. J., and Radtke J.: Use of quantitative models to design microbial transport experiments in a sandy aquifer, Ground Water, 39(2), 210-222, 2001.

Scheibe, T. D. and Chien, Y. J.: An evaluation of conditioning data for solute transport prediction, Ground Water, 41(2), 128-141, 2003.

Scott, D. W., and Sain, S. R.: Multidimensional density estimation, in Data Mining and Data Visualization, Handbook of Statistics, vol. 24, edited by C. R. Rao, E. J.Wegman, and J. L. Solka, pp. 229-262, Elsevier, Amsterdam, The Netherlands, 2005.

Seong, K. and Rubin, Y.: Field investigation of the WIPP Site (New Mexico) using a non stationary stochastic model with a trending hydraulic conductivity, Water Resour. Res., 35(4), 1011-1018, 1999.

Sun, A. Y., Ritzi, R. W. and Sims, D. W.: Characterization and modeling of spatial variability in a complex alluvial aquifer: Implications on solute transport, Water Resour. Res., 44, W04402, doi:10.1029/2007WR006119, 2008.

Sun, N.-Z.: Inverse Problems in Groundwater Modeling, Theor. Appl. Transp. Porous Media, vol.

6, Kluwer Acad., Dordrecht, Netherlands, 1994.

Theis, C. V.: The relation between the lowering of the piezometric surface and the rate and duration of discharge of a well using ground-water storage. Transactions Am. Geophys. Union, 16th Annual Meeting. Pt. 2, 1935.

Tiedeman, C. R., Hill, M. C., DAgnese, F. A. and Faunt, C. C.: Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the Death Valley regional groundwater flow system, Water Resour. Res., 39(1), 1010, doi:10.1029/2001WR001255, 2003.

Tiedeman, C. R., Ely, D. M., Hill, M. C. and O'Brien, G. M.: A method for evaluating the importance of system state observations to model predictions, with application to the Death Valley regional groundwater flow system, Water Resour. Res., 40, W12411, doi:10.1029/2004WR003313, 2004.

USEPA:Risk Assessment Guidance for Superfund: Volume III - Part A, Process for Con- ducting Probabilistic Risk Assessment, Tech. Rep. Rep.EPA 540/R-02/002,2001.

Vasco, D. W., Keers, H., and Karasaki, K.: Estimation of reservoir properties using transient pressure data: An asymptotic approach, Water Resour. Res., 36(12), 3447-3465, 2000.

Vermeul, V. R., Bjornstad, B. N., Fritz, B. G., Fruchter, J. S., Mackley, R. D., Newcomer, D. R., Mendoza, D. P., Rockhold, M. L., Wellman, D. M., and Williams, M. D.: 300 Area uranium stabilization through polyphosphate injection: Final Report, PNNL-18529, Pacific Northwest National Laboratory, Richland, WA, 2009.

Woodbury, A. D. and Rubin, Y.: A Full-Bayesian Approach to parameter inference from tracer travel time moments and investigation of scale effects at the Cape Cod Experimental Site, Water Resour. Res., 36(1), 159-171, doi:10.1029/1999WR900273, 2000.

Woodbury, A. D., and Ulrych, T. J.: Minimum relative entropy: Forward probabilistic modeling, Water Resour. Res., 29, 2847-2860, 1993.

Yeh, T.-C. J. and Liu, S.: Hydraulic tomography: Development of a new aquifer test method, Water Resour. Res., 36(8), 2095-2105, doi:10.1029/2000WR900114, 2000.

Young, S. C., Julian, H. E., Pearson, H. S., Molz, F. J., and Boman, G. K.: Application of the electromagnetic borehole flowmeter, US Environmental Protection Agency Research Report EPA/600/R-98/058, Ada, Oklahoma, 1998.

Zachara, J. M.: Annual Report to the DOE Office of Science, Climate and Environmental Sciences Division (January 2009 to January 2010), Multi-Scale Mass Transfer Processes Controlling Natural Attenuation and Engineered Remediation: An IFRC Focused on Hanford's 300 Area Uranium Plume, PNNL-19209, Pacific Northwest National Laboratory, Richland Washington, 2010.

Zhang, Z. and Rubin, Y.: MAD: a new method for inverse modeling of spatial random fields with applications in hydrogeology, Eos Trans. AGU, 89(53), Fall Meet. Suppl., Abstract H44C-07, 2008a.

Zhang, Z. and Rubin, Y.: Inverse modeling of spatial random fields, unpublished manuscript, 2008b.

Zhu, J. and Yeh, T.-C. J.: Characterization of aquifer heterogeneity using transient hydraulic tomography, Water Resour. Res., 41, W07028, doi:10.1029/2004WR003790, 2005.

Zhu, J. and Yeh, T.-C. J.: Analysis of hydraulic tomography using temporal moments of drawdown recovery data, Water Resour. Res., 42, W02403, doi:10.1029/2005WR004309, 2006.

Zimmerman, D.A., De Marsily, G., Gotway, C. A., Marietta, M. G., Axness, C. L., Beauheim, R. L., Bras, R. L., Carrera, J., Dagan, G., Davies, P. B., Gallegos, D. P., Galli, A., Gomez-Hernandez, J., Grindror, P., Gutjahr, A. L., Kitanidis, P. K., Lavenue, A. M., McLaughlin, D., Neuman, S. P., RamaRao, B. S., Ravenne, C. and Rubin, Y.: A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow, Water Resour. Res., 34(6), 13731413, 1998.

Zimmerman, D. L.: Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics, 17: 635-652. doi: 10.1002/env.769, 2006.

Appendix A

Forward Model in Chapter 2

The forward model $\mathcal{M}$ used in the case study is the flow equation [Castagna and Bellin, 2009]:

$$\nabla \cdot (T(x)\nabla h(x)) = 0, \tag{A.1}$$

where ∇ is the vector differential operator (the first ∇ represents divergence and the second ∇ represents gradient). T is the heterogeneous transmissivity and h is the hydraulic head. This equation is augmented by Dirichlet-type boundary conditions of constant (but different) heads along the boundaries at $y = 0$ and $y = 120$ with the head difference of 1m/120m (see Figure 2.2), and a Neuman-type no flow boundary conditions at $x = 0$ and $x = 90$, leading to a uniform-in-the-average pressure gradients parallel to the y-axis. The numerical solution is based on the finite-element method reported in Chapra and Canale (2002).

Appendix B

Bayesian Model-based Geostatistics for 2-D Structural Parameters in Chapter 2

According to Diggle and Ribeiro [Chapter 6, 2006], we calculate the posterior distributions of the 2-D geostatistical structural parameters conditioned on $\boldsymbol{z}_a(\boldsymbol{x}_a)$. First, the scale ϕ depends only on $\boldsymbol{z}_a(\boldsymbol{x}_a)$ and the prior distribution $p(\phi)$as

$$p(\phi \mid \boldsymbol{z}_a(\boldsymbol{x}_a)) \propto p(\phi) \mid V_{\hat{\mu}} \mid^{\frac{1}{2}} \mid \mathbf{R} \mid^{-\frac{1}{2}} \left(S^2\right)^{-\frac{N-1}{2}} , \tag{B.1}$$

where each term is defined as the follows:

$$\begin{aligned} V_{\hat{\mu}} &= \left[\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}\right]^{-1}, \\ \hat{\mu} &= V_{\hat{\mu}} \mathbf{1}^T \mathbf{R}^{-1} \boldsymbol{z}_a, \\ S^2 &= \frac{1}{n_a - 1} (\boldsymbol{z}_a^T \mathbf{R}^{-1} \boldsymbol{z}_a - \hat{\mu}^T V_{\hat{\mu}}^{-1} \hat{\mu}), \end{aligned} \tag{B.2}$$

where $\mathbf{R} = \mathbf{R}(\boldsymbol{x}_a, \boldsymbol{x}_a)$ is the auto-correlation matrix for $\boldsymbol{z}_a(\boldsymbol{x}_a)$, and $\mathbf{1}$ is the n_a-vector with all the elements equal to one. The variance σ^2 follows an inverse-scaled χ^2 distribution χ^2_{Scl} with $(n_a - 1)$ degrees of freedom and a scale parameter equal to S^2:

$$p(\sigma^2 \mid \phi, \boldsymbol{z}_a(\boldsymbol{x}_a)) \sim \chi^2_{Scl}(n_a - 1, S^2). \tag{B.3}$$

The mean follows a normal distribution with mean $\hat{\mu}$ and variance $\sigma^2 V_{\hat{\mu}}$:

$$p(\mu \mid \sigma^2, \phi, \boldsymbol{z}_a(\boldsymbol{x}_a)) \sim N(\hat{\mu}, \sigma^2 V_{\hat{\mu}}). \tag{B.4}$$

We multiply Equation (B.1), (B.3), (B.4), and the prior distribution to determine $p(\mu, \sigma^2, \phi \mid \boldsymbol{z}_a)$.

Appendix C

MCMC Implementation in MAD in Chapter 2

We partitions the parameter vector into blocks as $\{\mu, \sigma^2, \phi, \boldsymbol{\vartheta}_1, ..., \boldsymbol{\vartheta}_{J-3}\}$. We divide n_2 anchors into $J-3$ blocks. We sample each block as is described in the following sections.

C.1 Sampling of Structural Parameters

C.1.1 Scale ϕ

Bayes's rule is applied to the conditional distribution for the scale as,

$$p(\phi \mid \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}) = \frac{p(\boldsymbol{z}_b \mid \phi, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a) p(\phi \mid \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}, \tag{C.1}$$

where $p(\phi \mid \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}) = p(\phi \mid \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)$ from Equation (B.1). The acceptance ratio is obtained as:

$$\begin{aligned} r &= \frac{p(\phi^* \mid \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z})}{p(\phi^{(i)} \mid \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z})} \frac{q(\phi^{(i)} \mid \phi^*, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z})}{q(\phi^* \mid \phi^{(i)}, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z})} \qquad \text{(C.2)} \\ &= \frac{p(\boldsymbol{z}_b \mid \phi^*, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \phi^{(i)}, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)} \frac{p(\phi^* \mid \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\phi^{(i)} \mid \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)} \frac{q(\phi^{(i)} \mid \phi^*, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z})}{q(\phi^* \mid \phi^{(i)}, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z})}. \end{aligned}$$

For the proposal distribution, we apply the independent sampling [page 17 in Andreiu et. al. 2003] and sample from the prior distribution, which is assumed to be a uniform distribution as $\phi^* \sim Unif[\phi_{\min}, \phi_{\max}]$.

Then, we have the acceptance ratio as:

$$r = \frac{p(\boldsymbol{z}_b \mid \phi^*, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \phi^{(i)}, \mu^{(i)}, \sigma^{2(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)} \frac{p(\phi^* \mid \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\phi^{(i)} \mid \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}. \tag{C.3}$$

$p(\phi^* \mid \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)$ and $p(\phi^{(i)} \mid \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)$ are analytically determined according to Equation (B.1), since the anchors and Type-A data $\{\boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a\}$ are the point-scale conditioning points, i.e., error-free measurements. The likelihood of $\boldsymbol{z}_b$ is determined through numerical simulations.

C.1.2 Variance σ^2

In the same manner as the scale, Bayes' rule is applied to the invariant distribution as,

$$p(\sigma^2 \mid \phi^{(i+1)}, \mu^{(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}) = \frac{p(\boldsymbol{z}_b \mid \sigma^2, \phi^{(i+1)}, \mu^{(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a) p(\sigma^2 \mid \phi^{(i+1)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \phi^{(i+1)}, \mu^{(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}. \tag{C.4}$$

We use the independent sampling to draw the proposal value as,

$$\sigma^{2*} \sim q(\sigma^2 \mid \phi^{(i+1)}, \mu^{(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}). \tag{C.5}$$

Since we had an analytical form in Equation (B.2), we sampled the variance from the conditional pdf on the direct measurements and anchors $p(\sigma^2 \mid \phi^{(i+1)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)$ The acceptance ratio is

$$r = \frac{p(\boldsymbol{z}_b \mid \sigma^{2*}, \phi^{(i+1)}, \mu^{(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \sigma^{2(i)}, \phi^{(i+1)}, \mu^{(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}. \tag{C.6}$$

It means that we draw the proposed variance from the pdf conditioned on all but the non-local data, and decide to jump depending on the non-local data $\boldsymbol{z}_b$.

C.1.3 Mean μ

In the same manner as the variance and scale, we apply Bayes's rule as:

$$p(\mu \mid \phi^{(i+1)}, \sigma^{2(i+1)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}) = \frac{p(\boldsymbol{z}_b \mid \mu, \phi^{(i+1)}, \sigma^{2(i+1)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a) p(\mu \mid \phi^{(i+1)}, \sigma^{2(i+1)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \phi^{(i+1)}, \sigma^{2(i+1)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}. \tag{C.7}$$

We use the independent sampling again to draw the proposal value from the analytical form, in the same manner as the variance as $\mu^* \sim p(\mu \mid \phi^{(i+1)}, \sigma^{2(i+1)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)$.

The acceptance ratio is

$$r = \frac{p(\boldsymbol{z}_b \mid \mu^*, \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \mu^{(i)}, \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i)}, \boldsymbol{\vartheta}^{(i)}, \boldsymbol{z}_a)}. \tag{C.8}$$

C.2 Anchor block $\boldsymbol{\vartheta}_j$

In the same manner above, the Bayes rule is applied.

$$\begin{aligned} & p(\boldsymbol{\vartheta}_j \mid \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i+1)}, \boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}) \\ = & \frac{p(\boldsymbol{z}_b \mid \boldsymbol{\vartheta}_j, \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i+1)}, \boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}_a) p(\boldsymbol{\vartheta}_j \mid \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i+1)}, \boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i+1)}, \boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}_a)}. \end{aligned} \tag{C.9}$$

We draw the proposal value in the same manner as the previous parameters from the proposal distribution $p(\boldsymbol{\vartheta}_j \mid \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i+1)}, \boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}_a)$, where the vector $\{\boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}_a\}$ becomes the conditioning points. This proposal distribution is the multivariate Gaussian distribution with structural parameters $\{\phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i+1)}\}$ and conditioning points $\{\boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}_a\}$. Then, the acceptance ratio is

$$r = \frac{p(\boldsymbol{z}_b \mid \boldsymbol{\vartheta}_j^*, \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i+1)}, \boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}_a)}{p(\boldsymbol{z}_b \mid \boldsymbol{\vartheta}_j^{(i)}, \phi^{(i+1)}, \sigma^{2(i+1)}, \mu^{(i+1)}, \boldsymbol{\vartheta}_{-j}^{(i)}, \boldsymbol{z}_a)}. \tag{C.10}$$

Appendix D

Temporal Moment Formulation for Injection Tests in Chapter 3

According to Li et al. [2005] and Zhu and Yeh [2006], the k-th-order temporal moment $m_k(x)$ for a pressure-buildup curve $s(x,t)$ is defined by

$$m_k\left(x\right) = \int_0^\infty t^k s\left(x,t\right) dt. \tag{D.1}$$

In this study, we use only the zero-order moment $m_0(x)$ for the inversion to characterize the T field, which can exclude uncertainty in the storage coefficient and avoid an alias effect of the storage-coefficient uncertainty to the T field. Under the constant injection condition, we obtain $m_0(x)$ from the equation:

$$\nabla \cdot \left(T \nabla m_0\right) + \tau Q \delta\left(x - x_p\right) = 0, \tag{D.2}$$

with the boundary condition at the Drichlet boundary Γ_{dri} as,

$$m_0 = 0, \quad \text{on} \quad \Gamma_{\text{Dri}}, \tag{D.3}$$

where $T(x)$ is the depth-integrated T value, τ is the injection duration, Q is the constant injection rate and x_p is the injection well location. The Drichlet boundary condition was imposed at the nearest observation well location, where the m_0 value is known, in the same manner that Firmani et al. [2006] imposed a boundary condition at the injection well location. Note that Equation (D.2) is the same as the one for determining hydraulic head under steady-state flow with a constant injection rate τQ.

Appendix E

Bayesian Model-based Geostatistics for 3-D Structural Parameters in Chapter 3

According to Diggle and Ribeiro [Chapter 6, 2006], we calculate the posterior distributions of the 3-D geostatistical structural parameters conditioned on $\boldsymbol{u}(\boldsymbol{x})$. First, the scales λ_h and λ_v and nugget variance ν^2 depend only on $\boldsymbol{u}$ as

$$p(\lambda_h, \lambda_v, \nu^2|\boldsymbol{u}) \propto \pi(\lambda_h, \lambda_v, \nu^2)|V_{\hat{\beta}}|^{\frac{1}{2}}|(\mathbf{R}^{(3-D)} + \nu^2\mathbf{I})|^{-\frac{1}{2}}(S^2)^{-\frac{N-1}{2}}, \tag{E.1}$$

where each term is defined as the follows:

$$\begin{aligned} V_{\hat{\beta}} &= [\mathbf{1}^T(\mathbf{R}^{(3-D)} + \nu^2\mathbf{I})^{-1}\mathbf{1}]^{-1}, \\ \hat{\beta} &= V_{\hat{\beta}}\mathbf{1}^T(\mathbf{R}^{(3-D)} + \nu^2\mathbf{I})^{-1}\boldsymbol{u}, \\ S^2 &= \frac{\boldsymbol{u}^T(\mathbf{R}^{(3-D)} + \nu^2\mathbf{I})^{-1}\boldsymbol{u} - \hat{\beta}^T V_{\hat{\beta}}^{-1}\hat{\beta}}{N-1}, \end{aligned} \tag{E.2}$$

where N is the dimension of $\boldsymbol{u}$, $\mathbf{R}^{(3-D)} = \mathbf{R}^{(3-D)}(\boldsymbol{x}, \boldsymbol{x})$ is the auto-correlation matrix for $\boldsymbol{u}$, $\mathbf{1}$ is the N-vector with all the elements equal to one, and $\mathbf{I}$ is the identity matrix. The variance η^2 follows an inverse-scaled χ^2 distribution χ^2_{Scl} with $(N-1)$ degrees of freedom and a scale parameter equal to S^2:

$$p(\eta^2|\lambda_h, \lambda_v, \nu^2, \boldsymbol{u}) \sim \chi^2_{Scl}(N-1, S^2). \tag{E.3}$$

The mean follows a normal distribution with mean $\hat{\beta}$ and variance $\eta^2 V_{\hat{\beta}}$:

$$p(\beta|\eta^2, \lambda_h, \lambda_v, \nu^2, \boldsymbol{u}) \sim N(\hat{\beta}, \eta^2 V_{\hat{\beta}}). \tag{E.4}$$

We multiply Equation (E.1), (E.3), (E.4), and the prior distribution to determine $p(\beta, \eta^2, \lambda_h, \lambda_v, \nu^2|\boldsymbol{u})$.