**Title**
Hydrogeologic Modeling at Unsampled Locations: A Bayesian View of Uncertainty Quantication and Reduction



**Author**
Chang, Ching-Fu

**Publication Date**
2019

Peer reviewed|Thesis/dissertation

**Hydrogeologic Modeling at Unsampled Locations: A Bayesian View of Uncertainty Quantification and Reduction**

by

Ching-Fu Chang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yoram Rubin, Chair
Professor David Brillinger
Associate Professor Sally Thompson

Spring 2019

**Hydrogeologic Modeling at Unsampled Locations: A Bayesian View of Uncertainty Quantification and Reduction**

## Abstract

Hydrogeologic Modeling at Unsampled Locations: A Bayesian View of Uncertainty Quantification and Reduction

by

Ching-Fu Chang

Doctor of Philosophy in Engineering - Civil and Environmental Engineering

University of California, Berkeley

Professor Yoram Rubin, Chair

Groundwater plays a crucial role in our water resources. To counteract the growing demand and depletion of groundwater resources and devise sustainable management plans, a wide range of models have been applied to make estimates/predictions of various hydrogeologic responses. However, uncertainty arises in all modeling applications, and the quantification/reduction of modeling uncertainty has been a challenge for all hydrogeologists, especially at unsampled locations. The main challenge posed at unsampled locations is the lack of in-situ data, forcing us to search for alternative sources of information, and systematically assimilate the said information in order to obtain conditioned estimates of hydrologic responses. To that end, the primary objective of this dissertation is the advancement of stochastic modeling approaches targeting at unsampled locations. Under the context of the primary objective, we propose three different stochastic modeling approaches that are designed to assimilate three alternative sources of information, respectively.

First, we propose the Rapid Impact Modeling (RIM) approach to efficiently assimilate in-situ soft data (i.e., in-situ data that are related to the target response via transfer functions) for obtaining conditioned estimates. RIM improves upon the existing approximate Bayesian computation approaches by (1) bypassing the estimation of posterior distributions of model parameters, thus reducing the computation burden, and (2) relaxing the need to reduce data into summary statistics, thus avoiding losing information. To demonstrate the power of RIM, we address the challenge of data scarcity against the backdrop of a 7 $km$ long and hundreds of meters deep underground tunnel in China, a typical example of heavy-impact yet poorly sampled site. Through the demonstration, we also recognize that goal-oriented site characterization is in many cases more useful in applications compared to parameter-oriented characterization.

Second, we turn our attention to the assimilation of ex-situ data (i.e., data from locations other than the location of interest) via regionalization, transferring information obtained at sampled locations to unsampled ones. The reliability of regionalization depends on (1) the underlying system of hydrologic similarity, as well as (2) the approach by which information

is transferred. We propose a nested structure to couple classification tree with Bayesian additive regression tree, named the nested tree-based modeling approach. The nested tree-based modeling approach is designed as an advanced regionalization technique that features the capability of modeling non-linear predictor-response relationship, as well as Bayesian representation of uncertainties of the model parameters and the model structure. In addition, we integrate the approach with a hypothesis of two-leveled hierarchical hydrologic similarity to investigate the dynamic behavior of hydrologic similarity. In a case study of groundwater recharge estimation, we show how the nested tree-based modeling approach and the hierarchical similarity hypothesis can reveal the variation of the controls of hydrologic similarity under different conditions. The proposal of the nested tree-based modeling approach and our hypothesis of hierarchical similarity contribute to the understanding of the physical principles governing robust information transfer.

Third, we look at situations with extreme data scarcity where in-situ data are unavailable, and the ex-situ data takes the form of bounds of plausible value rather than point observations. We propose a nuanced two-level Bayesian hierarchical model to assimilate ex-situ bounds, where ex-situ bounds are assimilated via truncation of distributions rather than data imputation, thus avoiding artificial biases. Furthermore, our approach features the capability of modeling ex-situ bounds as random variables to account for the potential uncertainties of ex-situ bounds. Our proposed approach not only contributes to the Bayesian regionalization using ex-situ bounds but also provides guidance for future applications in the establishment of ex-situ bounds.

The three approaches are all based on the concepts of Bayes' rule and can all be considered as applications of Bayesian inference. They represent sophisticated assimilation of various alternative forms of information, and are designed to tackle the ultimate challenge of large modeling uncertainty in the face of data scarcity. We expect the approaches proposed in this dissertation to contribute to the advancement of Bayesian uncertainty quantification and reduction at unsampled locations.

To my lovely wife, Yu-Chun Hsu

without whom I would not have come to Berkeley.

# Contents

# List of Figures

# List of Tables

# Acknowledgments

Back in the Winter of 2014, a young man thought his application for the doctoral program at UC Berkeley CEE was a fun but futile attempt. Years later, he is writing the acknowledgment for this very dissertation. What a magical journey, full of wonderful people along the way.

I would first like to thank my Ph.D. advisor, Yoram Rubin, for his guidance since the very first Skype call we had. I have learned a lot from him, and it has truly been my honor and my pleasure to join Rubin's group. Speaking of the group, I would like to thank Jon Sege, Changhong Wang, and Carlos Osorio-Murillo, who warmly welcomed me on my first day on campus and sacrificed their space for my desk. I would also like to thank Nigel Chen, Thanh Thuy Nguyen, Benjamin Tan, Jianqin Chen, Bo Tan, and Hu Ting, for the wonderful time we have had on and off campus. I am also grateful for the advice, in both research and life, from my academic big brother/sisters, Michelle Newcomer, Bradley Harken, Heather Savoy, and Karina Cucchi.

Outside the group, I would like to thank Sally Thompson for inspiring and supporting me throughout these years, thank Tina Chow and Marc Stacey for their advice and support, thank Dennis Baldocchi for serving on my qualifying exam committee, and thank David Brillinger for serving on both my qualifying exam and dissertation committees. Thanks to Falk Heße and Chris Paciorek for their help in my research. I am also thankful to Mark Hansen and Nicholas Sitar for the fruitful experiences I had working with them. I also want to thank the Jane Lewis Fellowship Committee for providing me financial support, thank the UC Berkeley International Office for the special help for international students, and thank Shelley Okimoto for always kindly helping me with all sorts of problems.

As an international student, even the everyday life is an adventure, full of excitement, but not without difficulties. I would like to thank my friends in the Berkeley Association of Taiwanese Students, for making Berkeley more like home. To my family in Taiwan, thank you for being proud of me. Thanks to my aunt Iris Cheng and my uncle Eric Chen, who picked me up at SFO when I first got to the U.S. helped me settle down. Lastly, I would like to thank my wife, Yu-Chun Hsu, for the support no words can ever describe.

# Chapter 1

# Introduction

Our heavy reliance on groundwater resources has been well documented. Groundwater provides almost half of all drinking water worldwide and more than 40 % all consumptive use of water for irrigation (Smith et al., 2016). A global inventory of water consumption for irrigation published in 2010 suggested that the global consumptive groundwater use for irrigation alone surpassed 500 $km^3yr^{-1}$; furthermore, the inventory showed growing trends of the percentage of irrigated area that was irrigated with groundwater in the U.S.A. and in India over the $20^{th}$ century (Siebert et al., 2010). In response to the growing demand, the groundwater has been increasingly depleted since the $20^{th}$ century, as the abstraction of groundwater at least tripled over the last 50 years (Smith et al., 2016; Wada et al., 2010), which highlights the importance of sustainable groundwater resources management. To that end, our ability to estimate and manage groundwater storage, groundwater table, and/or groundwater recharge is of great importance. In this regard, numerical models with different natures (e.g., physically based or data-driven), at different levels of complexity, and at different spatiotemporal scales have been applied in a myriad of studies to model hydrologic/hydrogeologic responses (e.g., Gilbert and Maxwell, 2017; Kim et al., 2008; Mogaji et al., 2015; Rahmati et al., 2016).

In all modeling applications, however, various forms of uncertainty exist, both within and outside of the hydrologic/hydrogeologic modeling process. Starting from the outside, uncertainty resides in the interface between hydrology/hydrogeology and related disciplines such as climate science, sociology, and political science. Rubin et al. (2018) identified this interface as the source of "unknown unknowns", i.e., things we do not know that we do not know. They further described a blind spot of stochastic hydrogeology where the modelers focus on known unknowns, while some other variables are intentionally or unintentionally ignored. The inclusion of the interface between hydrogeology and the related fields into the modeling process is daunting. Fortunately, there is a growing recognition that real-world situations are ripe with unknown unknowns (Rubin et al., 2018), and Harken (2017) and Harken et al. (2019) have attempted a promising start by establishing a framework which facilitates conversations on uncertainty among hydrogeologists, regulators, managers/decision makers, and the general public.

Moving from the outside to the inside, we now turn our attention to uncertainties that appear along the entire modeling process. Uncertainty resides in the forcing and response data, in the model structure, and in the model parameters and/or the physical properties that characterize the underlying physical processes (Beven, 2016; Refsgaard et al., 2007).

- Data uncertainty is primarily dependent on the processes through which observations or measurements of a target property are made, as well as the interpretation of the data. Despite the importance of data uncertainty, its assessment has been pointed out as complex and full of pitfalls (Refsgaard et al., 2007), and the identification of an error in the data may be computationally challenging depending on the model structures and parameters tested (Beven, 2016).

- Model structure uncertainty is primarily caused by either the lack of knowledge or the misinterpretation of the physical system that is being modeled, and may be observed through equifinality, where there are multiple plausible models capable of explaining the observations (Beven, 2006; Beven and Freer, 2001). The *a priori* presumption of a known and fixed model structure, without justification, could obscure the modeling results as the uncertainty in model structure gets compensated by other types of uncertainty (Ajami et al., 2007; Beven et al., 2008; Nowak et al., 2010), especially by model parameter uncertainty: a conceptually wrong model could possibly output results that match the observations quite well, provided a forcible calibration process that makes the model do so.

- Compared to the other two, model parameter uncertainty has been widely recognized, quantified, and even reduced in many modeling applications by considering the parameters as random variables (e.g., Li et al., 2018; Rubin and Dagan, 1987a; Rubin et al., 2010; Woodbury and Rubin, 2000).

In the following Sections, we start by providing an overview of how stochastic approaches have been applied to tackle the uncertainties in hydrologic/hydrogeologic modeling, followed by a discussion on a challenge in the application of stochastic approaches.

## 1.1 Stochastic modeling: Overview

In a typical modeling application, after data collection and the establishment of the model structure, the next step would be to calibrate and validate the model (or alternatively, model fitting) (Refsgaard et al., 2007). The conventional idea of calibration is to find the "best" set of model parameters that reduces the expected predictive error to the best degree. This concept can be applied to physically based models (e.g., Rodhe and Bockgård, 2006; Xie et al., 2017) as well as empirical or data-driven models (e.g., Chen et al., 2002; Gemitzi et al., 2017; Mogaji et al., 2015). Hybrid models have also been built; for example, Xu and Valocchi (2015) integrated statistical learning techniques with a calibrated groundwater model to obtain predictive intervals that agree with available data.

In the past few decades, the concept of model fitting has evolved from searching for an optimal set of parameters that yields a minimum error, to estimating conditional metrics and summary statistics (e.g., statistical moments, range, quantiles, etc.) or even conditional distributions of the parameters and/or the modeled responses. The evolution is observed together with the growing application of stochastic approaches, here used to refer to any approach that use radon variables to account for the uncertainty associated with at least one component of the model (parameters in the model, or the model structure, etc.). Various stochastic modeling approaches have been developed and applied in that regard, which has been exemplified by the blooming adoption of Bayesian inference and its variants in hydrologic/hydrogeologic research (e.g., Beven and Binley, 2014; Nott et al., 2012; Nowak et al., 2010; Over et al., 2015; Rubin et al., 2010; Smith et al., 2014; Xu et al., 2017; Yang et al., 2015), and also exemplified by various ready-to-use software solutions to stochastic modeling approaches (Table A.1). Regardless of the role stochastic approaches play in the entire modeling process, ranging along the spectrum from the overarching framework at one end to just an add-on supplemental tool at the other, stochastic approaches offer a way to quantify and/or reduce uncertainty systematically in probabilistic terms.

Nonetheless, the application of stochastic approaches is not without challenges. Besides the various factors outside the realm of hydrogeologic modeling (e.g., the law, regulations, higher education, and profitability) that affect the application of stochastic hydrogeologic models (Rubin et al., 2018), of particular interest in this dissertation is the challenge posed by the lack of in-situ data, which will be discussed next.

## 1.2 Challenges in Stochastic Modeling: Lack of In-situ Data

In all modeling applications, stochastic or not, of significant importance is the availability of **in-situ data**, here defined as observations/measurements taken within the spatial domain of the site of interest, and within a reasonable temporal domain that does not make the observations obsolete. Unfortunately, most watersheds in the world still remain ungauged (Hrachowitz et al., 2013; Ibrahim and Cordery, 1995). The term "ungauged watershed" originates from the effort in the field of surface hydrology to predict runoff or its derivatives (e.g., hydrograph, low flow, flood) at watersheds which are either (1) genuinely ungauged, (2) poorly gauged, or (3) previously gauged, i.e., sampling was terminated (Blöschl et al., 2013; Hrachowitz et al., 2013; Loukas and Vasiliades, 2014). Here the term "unsampled" is used interchangeably with "ungauged", for "unsampled" has a somewhat more general image signifying the lack of in-situ observations of hydrological responses in any form (including runoff gauge, piezometer reading, weir overflow, observation well, infiltrometer reading, etc.). The lack of in-situ data poses a great challenge at both the intra-model level and the inter-model level. There would be no data to condition model parameters on, and furthermore, there would be no data to support any specific model structure. Consequently, one would

not be surprised if large uncertainty is associated with any stochastic modeling application that is without in-situ data.

The aforementioned large uncertainty might be misconceived as the seeming "failure" of stochastic approaches, which might fool one into thinking that a successful stochastic model application requires a lot of data so that the uncertainty can be significantly reduced (e.g., obtaining a well-fitted variogram with many data points for a spatial random variable). On one hand, a lack of in-situ data does affect the performance of a stochastic model (e.g., it would be difficult to model the space random function, or that the plausible value of a model parameter would have a wide range, etc.). On the other hand, the lack of data does not make uncertainty disappear, and in fact, it only accentuates the need for quantifying uncertainty, where stochastic methods shine the brightest (Rubin et al., 2018).

Now with the misconception sorted out, we are still faced with large uncertainty due to the lack of in-situ data. Is this the best one can do at unsampled locations? Fortunately, the answer is no. Bayesian stochastic approaches, which are based on Bayes's rule and Bayesian inference, offer a solution to the lack of in-situ data by having the capability of using a variety of forms of information (Cirpka and Valocchi, 2016; Li et al., 2018; Rubin et al., 2018). In the following subsections, we provide brief introductions to several types of information that Bayesian approaches can use, along with some associated potential challenges that still need to be addressed.

## In-situ Soft Data

The concept of soft data originates from the concept of hard data, which is generally used to described direct and collocated measurements of the target response variable. Broadly speaking, soft data means data that are not hard data. The definition of soft data is somewhat vague and has been discussed in several previous studies. Rubin (2003) used the term "soft information" to describe imprecise measurements in general, such as measurements that are subject to sampling error, or range of values, or even probability density function (pdf) of the target variable. Winsemius et al. (2009) used the same term to describe the situation where the conditioning on the information is less effective and/or where the uncertainty of the measurements cannot be objectively quantified. Stemming from the high cost of obtaining direct measurements of soil hydraulic properties, Segal et al. (2008) used "soft data" to describe estimates of the target variable from indirect approaches (e.g., estimates of soil texture from electrical conductivity), which are of lower quality and confidence. Hou and Rubin (2005) described soft information as measurements of variables that are related to the target variable via transfer functions. Rubin et al. (2010) used the term "Type-B data" to describe the same concept: data that are the outputs from a transfer function of the target variable. This definition is more general, as it does not specify the form of the transfer function; in this context, all the other definitions can be viewed as different cases with different transfer functions. Therefore, in this dissertation, the term "soft data" is defined as observations/measurements of other responses that can be related to the target response via transfer functions.

Soft data has been applied as a source of information in addition to hard data, and various studies have shown that soft data can reduce the uncertainty in **inverse modeling** (i.e., obtaining estimates of model parameters that are conditioned on observations, Rubin et al., 2010; Segal et al., 2008; Winsemius et al., 2009). However, few studies have directly linked soft data with **forward modeling** (i.e., obtaining estimates of the target response using a given set of parameters). Using soft data for forward modeling may seem natural: information is first transferred from soft data to estimates of model parameters via inverse modeling, and then transferred to estimates of hydrologic response via forward modeling. Nonetheless, at unsampled sites with complicated settings, this transition of information may be computationally demanding due to (1) large uncertainties in the parameters because of the lack of in-situ hard data, which necessitates a large number of simulations, (2) the combined computational demand for both inverse and forward modeling, and (3) the additional computational demand for the transfer function between hard and soft data. Although computational demand is not absolutely insurmountable, it is still of great practical importance in many modeling applications. This poses **the need for efficient conditioning of hydrologic response on soft data**.

## Ex-situ Data and Regionalization

Studying unsampled watersheds has been a popular research topic for more than a decade, especially since The Prediction in Ungauged Basins (PUB) initiative by the International Association of Hydrological Sciences (IAHS) (Sivapalan et al., 2003). A way to tackle the lack of in-situ data is to extract information from **ex-situ data** (namely, observations of the target response that are not taken at the site of interest), and transfer the information to the unsampled locations of interest. This information transfer is also termed "regionalization", which could be applied to constrain the model in the form of (1) relationships between model parameters and site characteristics, (2) subsets of the parameter space, or (3) plausible parameter values from models at hydrologically similar watersheds (Blöschl et al., 2013; Kuczera, 1982; Razavi and Coulibaly, 2017; Singh et al., 2014; Wagener and Montanari, 2011).

However, the application of regionalization is not without challenges. One of the key factors of predictive uncertainty identified by the PUB initiative is the unsuitability of information transfer techniques, due to a lack of comparative studies across watersheds and a lack of understanding of the physical principles governing robust regionalization (Hrachowitz et al., 2013). Different regionalization techniques have been applied in different cases with different assumptions. For example, Li et al. (2018) attempted a simple form of regionalization, where kernel density estimation was applied on recharge values obtained from various hydrologically similar sites, in order to build an ex-situ prior distribution (i.e., a prior distribution conditioned on ex-situ data). However, one limitation in Li et al. (2018) was that hydrologic similarity was treated as a Boolean variable, and therefore, there was no way to systematically distinguish a highly similar site from a slightly similar site.

To pursue this further, we must ask the following question: How can we tell that two watersheds are hydrologically similar? Sawicz et al. (2011) applied Bayesian mixture clustering to watersheds across the eastern U.S. They found that spatial proximity was a valuable first indicator of hydrological similarity because it reflected strong climatic control in their study area. Oudin et al. (2008) reported similar findings based on 913 French watersheds, despite acknowledging the lack of some key physical descriptors in their data set. However, Smith et al. (2014) attempted regionalization of hydrologic model parameters in eastern Australia, and suggested that spatial proximity was an unreliable metric of hydrological similarity. For their part, Tague et al. (2013) presented successful regionalization of hydrologic parameters based on geologic similarity at watersheds in the U.S. Oregon Cascades, a mountain range that features geological heterogeneity. Although not directly shown, their findings also went against the use of applying spatial proximity, for they discussed the sharp contrasts in hydrology at proximal watersheds based primarily on geological differences. The indication from these findings is that, although spatial proximity is of practical importance due to its common use, its simplicity, and its demonstrated effectiveness in specific areas (Smith et al., 2014), it is not the true controlling factor, but rather a confounding factor.

One can resort to other physical characteristics of watersheds for the determination of hydrologic similarity. However, what those characteristics are may be a complicated question. Razavi and Coulibaly (2017) tested the effect of combinations of neural-network-based classification techniques and regionalization techniques in Canada, and found that classifying watersheds before regionalization improves regionalization for streamflow, baseflow, and peak flow predictions, but also discovered that the best combination of techniques varied from one watershed to another. Singh et al. (2014) applied classification and regression tree to determine the relationship between catchment similarity and regionalization in the U.S., finding that the dominant controls of successful regionalization vary significantly with the spatial scale, with the region of interest, and with the objective function used. Similarly, Kuentz et al. (2017) found that different physiographic variables controlled various flow characteristics across Europe, showing how different descriptors could account for different dominant hydrologic processes and flow characteristics. Loritz et al. (2018) suggested an interesting perspective describing a dynamic hydrologic similarity system, where similarity and uniqueness are not mutually exclusive; rather, they suggested that hydrologic systems operate by gradually changing to different levels of organization in which their behaviors are partly unique and partly similar. These studies indicate an important challenge, that the factors determining hydrologic similarity may vary under different conditions, and a universal system of hydrologic similarity still remains unavailable.

The need to understand the dynamic behavior of hydrologic similarity has been emphasized above, but how have researchers attempted to investigate it? Gibbs et al. (2012) provided a generic framework of regression regionalization, which involves a multi-objective optimization for calibration, a sensitivity analysis to determine the most important model parameters, and a final step relating watershed characteristics with model parameters. The framework is capable of assimilating information from exogenous variables to obtain an informed inference of hydrologic similarity, while also accounting for the interaction between

parameters. However, the framework does not include a straightforward quantification of uncertainties in calibration and in regionalization. In comparison, Bayesian approaches offer a solution to the quantification of uncertainty by outputting conditional distributions. Despite the lack of in-situ data, one can still apply Bayesian approaches to establish prior distributions that are informed by data from previous studies or well-established databases (Hou and Rubin, 2005; Woodbury, 2011; Woodbury and Rubin, 2000). More advanced pooling of information from multiple sampled sites has also been demonstrated with the application of Bayesian hierarchical models (Cucchi et al., 2019; Smith et al., 2014), which can account for both intra- and inter-site uncertainty of the parameters. However, the aforementioned Bayesian approaches have several disadvantages, including: (1) requiring a system of hydrologic similarity *a priori* that helps us decide which sampled sites or databases are suitable as "information donor", (2) requiring known or assumed distributional forms of the parameters, and (3) difficulties in accounting for complicated and highly non-linear dependence on exogenous variables. Adding onto the challenge is that the aforementioned approaches, regression-based or Bayesian, do not have a built-in component for the consideration of model structure uncertainty. Model structure uncertainty does not only apply to the model for the hydrologic response, but also to the model with which the regionalization is carried out (e.g., the regression model between watershed characteristics and hydrologic model parameters). In summary, from the literature review above we have identified **the need for a regionalization approach that can simultaneously account for the dynamic hydrologic similarity and the uncertainties in the model parameters as well as the model structure**.

## Ex-situ Soft data

In some cases, the scarcity of data could reach a high level that even the in-situ soft data and ex-situ data are in shortage. This extreme lack of data forces the modeler to consider anything else that could possibly offer an additional piece of information. The importance of ex-situ soft data becomes noticeable in such cases. Following the definitions introduced in the previous subsections, ex-situ soft data are observations that are (1) not taken at the site of interest, and (2) are some other responses that can be related to the target response through transfer functions. Ex-situ soft data can be conceptualized as the outcome of two sequential transfer functions: the transfer function between hard and soft data, and the transfer function embedded in the regionalization process that transfers information from one site to another. This naturally makes the ex-situ soft data only weakly correlated with the target response, when compared to in-situ soft data or ex-situ hard data.

There are many potential sources of ex-situ soft data; among all, a promising one is expert elicitation, which has been applied in various fields in previous studies. Garthwaite et al. (2005) pointed out that the growing sophistication of computational approaches has led to a dramatic increase in the breadth and complexity of Bayesian applications, which relates to the increasing interest in expert elicitation in Bayesian context. Sebok et al. (2016) elicited information from 35 experts to estimate watershed-scale water balance and

to analyze the corresponding uncertainty. Warmink et al. (2010) analyzed the uncertainties in river model and identified the most important components of uncertainty based on expert opinion. Ye et al. (2008) integrated expert elicitation into a framework of Bayesian model averaging to establish prior distribution of different groundwater recharge models. As they pointed out, for complicated hydrologic systems, expert judgment is the basis of conceptual model development, and may even be more informative than limited observations, which is particularly true for hydrogeology.

Depending on the transfer functions, soft data can appear in various forms, including statistical moments, quantiles/percentiles, point estimates with or without error estimates, and bounds of plausible values (bounds, in short), the last of which is particularly common in hydrogeology. For example, plausible bounds of hydraulic properties corresponding to different soil or rock types can be found in almost every groundwater textbook (e.g., Fetter, 2001; Singhal and Gupta, 2010; Todd and Mays, 2004). Of course, it would be better to have site-specific information, but when data scarcity is extremely limiting, these bounds provide valuable information for the establishment of a weakly informative distribution of the variable of interest.

Unfortunately, the assimilation of ex-situ soft data is still in its infant stage in hydrogeology. Cucchi et al. (2019) provided a great start, by proposing a practical approach that uses data imputation to convert various forms of ex-situ soft data into ex-situ hard data. While a versatile method that can cater to many forms of soft data is ideal and desirable, the involvement of imputation makes the approach sensitive to the number of imputed data, and thus make the approach prone to artificial biases. To that end, we see the effort of Cucchi et al. (2019) as a basis upon which we can improve, by identifying **the need for an approach for the assimilation of ex-situ bounds that is free of imputation-induced biases**.

## 1.3 Research Objectives

While Bayesian approaches are capable of assimilating both soft data and ex-situ data, some challenges have been identified in Section 1.2. In response, the overarching goal of this dissertation is the development of advanced Bayesian stochastic modeling approaches that overcome the aforementioned challenges. Under this context, there are three specific research objectives, each of which involves the proposal of an innovative approach in order to tackle a specific challenge, explained in details as follows.

### Efficient Assimilation In-situ Soft Data for Forward Modeling

The first objective is to propose an efficient and goal-oriented Bayesian computation approach that features significantly reduced computation demand for the conditioning of estimates of hydrologic responses on in-situ soft data.

### Systematic Determination of Hydrologic Similarity and Full Bayesian Regionalization

The second objective is to propose a general approach with which regionalization can be carried out with (1) systematically determined hydrologic similarity, (2) full Bayesian representation of both the model parameter uncertainty and the predictive uncertainty, and (3) consideration of the model structure uncertainty. We also wish to be able to explain how modeling uncertainty and the assimilation of ex-situ data are affected by the dynamic behavior of hydrologic similarity with the proposed framework.

### Imputation-free Assimilation of Bounds

The third objective is to formally and analytically derive a Bayesian model where one specific type of ex-situ soft data —bounds —can be assimilated. By sacrificing versatility and focus only on bounds, we propose this approach to solve the issue of imputation-induced artificial biases in the approach proposed by Cucchi et al. (2019).

## 1.4 The Structure of This Dissertation

The remainder of the dissertation is organized as follows. Chapter 2 provides the mathematical background of Bayesian inference, Bayesian computation, and some of the derivatives of Bayesian inference, providing the basis of all the innovations in our proposed Bayesian stochastic approaches. In Chapter 3 through 5, we describe the innovations in the proposed approaches corresponding to the three research objectives, respectively, each of which will be demonstrated in a case study. Finally, Chapter 6 summarizes the findings and the contributions of this dissertation.

# Chapter 2

# Background

This Chapter provides the mathematical background of Bayesian inference and Markov Chain Monte Carlo simulations. Then, several derivatives of Bayesian inference are introduced to provide background knowledge for the innovations in the rest of the dissertation.

## 2.1 Bayesian Inference

The core of Bayesian inference is Bayes' rule. Let $Y$ be the target variable of interest (e.g., a hydrological response), which we modeled as a random variable. Let $y$ be the data (i.e., observations/measurements of the target variable), and let $\theta$ be parameter(s) of the model we use to estimate/predict $Y$. We use the term **forward modeling** to refer to the process where we obtain estimates of $Y$ with a given $\theta$ and some other model inputs. We use the term **inverse modeling** to refer to the process where we obtain estimates of $\theta$ conditioned on $y$. Bayesian statistical conclusions about $\theta$ are made in terms of probability statements (Gelman et al., 2014). By the rule of conditional probability, the joint distribution of $y$ and $\theta$ can be expressed as follows:

$$p(\theta, y) = p(\theta)p(y|\theta) = p(y)p(\theta|y). \tag{2.1}$$

The term $p(\theta)$ is referred to as the prior distribution (prior, in short) of $\theta$. Rearranging terms, one would obtain the general form of Bayes' rule:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{2.2}$$

where the left hand side is the posterior distribution (posterior, in short) of $\theta$. The term $p(y|\theta)$ is by definition the conditional distribution of $y$ if considered as a function of $y$, but when it is considered as a function of $\theta$ for a given $y$, it is referred to as the likelihood function (of $\theta$). Bayesian inference refers to the statistical inference one makes by applying Bayes' rule to update the prior and obtain the posterior.

The term $p(y)$ could be obtained by the total probability rule as follows:

$$p(y) = \begin{cases} \sum_\theta p(\theta)p(y|\theta) & \text{if } \theta \text{ is discrete} \\ \int_\theta p(\theta)p(y|\theta)d\theta & \text{if } \theta \text{ is continuous} \end{cases}. \tag{2.3}$$

As it only depends on $y$, $p(y)$ is often omitted in many applications of Bayesian inference, yielding the unnormalized posterior distribution, which is the right hand side in the following equation:

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{2.4}$$

If a stochastic approach results in obtaining the posteriors of the variables of interest, it can be described as "full Bayesian" to differentiate it from other approaches that only results in obtaining posterior statistics (e.g., posterior mean, posterior 95% confidence intervals, etc.).

## 2.2 Markov Chain Monte Carlo (MCMC) Simulations

An analytical solution to Equation 2.2 can be obtained when both the prior and the likelihood can be expressed analytically. An even more analytically convenient situation is when the prior and the likelihood form a conjugate pair (e.g., two Gaussian distributions, Beta and binomial distributions, Gamma and Poisson distributions, etc.), so that the posterior takes the same analytical form as the prior. However, for complicated or unusual models, or in high dimensions, $p(y)$ and $p(y|\theta)$ can be difficult to obtain, and thus more elaborate algorithms are required to numerically approximate the posterior distribution.

In that regard, many clever methods have been devised for simulation-based approximations of the posterior. Among all the methods, Markov Chain Monte Carlo (MCMC) simulation is adopted in this dissertation for it has been well developed and integrated with many stochastic software packages (e.g., Chipman et al., 2010; Valpine et al., 2017). MCMC is an indirect iterative method based on drawing values of $\theta$ from approximate distributions and then correcting those drawn values to better approximate the target posterior; it is used when direct sampling from $p(\theta|y)$ is not feasible (due to model complexity, high dimensionality, etc.) (Gelman et al., 2014). The Monte Carlo sampling is done sequentially where the distribution of the current draw depends on the last drawn value, which fits the definition of a Markov Chain where the probability of each event in a sequence of events depends on the state attained in the previous event, thus the name Markov Chain Monte Carlo. The critical factor of a successful application of MCMC is that the distribution of the draws eventually converges towards a stationary distribution, which is the target posterior. The application of MCMC usually starts with an initial condition, $\theta^{(0)}$, and then follows a transition distribution for the $l^{th}$ draw, $p_T\left(\theta^{(l)}|\theta^{(l-1)}\right)$.

Brief introductions to two of the commonly used MCMC algorithms that are adopted in this dissertation are provided as follows.

## Gibbs Sampler

The Gibbs sampler is particularly useful when the dimensionality of the parameter space is high (i.e., $\theta$ consists of many parameters) (Gelman et al., 2014). Suppose the parameter vector, $\theta$, can be partitioned into $D$ subvectors, $\theta = \{\theta_1, ..., \theta_D\}$, such that sampling of each subvector given all the other subvectors and the data is feasible. If so, the Gibbs sampler proceeds by applying the following transition distribution on the $d^{th}$ subvector for the $l^{th}$ draw:

$$p\left(\theta_d^{(l)} | \theta_{-d}^{(l-1)}, y\right), \tag{2.5}$$

where $\theta_{-d}^{(l-1)} = \left\{\theta_1^{(l)}, ..., \theta_{d-1}^{(l)}, \theta_{d+1}^{(l-1)}, ..., \theta_D^{(l-1)}\right\}$. Thus, it avoids the challenging sampling from $p(\theta|y)$ by sampling from Equation 2.5. Within each MCMC draw, the Gibbs sampler sample each subvector sequentially, and then proceeds to the next draw.

## Metropolis and Metropolis-Hastings Algorithms

The Metropolis algorithm is an adaptation of a random walk with an acceptance/rejection rule to converge to the target distribution (Gelman et al., 2014).

It requires an initial value, $\theta^{(0)}$, as well as an initial distribution, $p(\theta^{(0)})$. A jumping distribution (also termed proposal distribution) is defined for the $l^{th}$ draw, denoted as $J_l(\theta^*|\theta^{(l-1)})$, where $\theta^*$ is a proposal drawn from this jumping distribution. The jumping distribution is symmetric, i.e., $J_l(a|b) = J_l(b|a)$ for all $a$, $b$, and $l$.

After drawing the proposal, an acceptance ratio of the unnormalized posteriors is calculated as follows:

$$r_J = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(l-1)})p(\theta^{(l-1)})}. \tag{2.6}$$

Finally, the $l^{th}$ draw is drawn from the following distribution:

$$\theta^{(l-1)} = \begin{cases} \theta^* & \text{with probability } \min(r_J, 1) \\ \theta^{(l-1)} & \text{otherwise} \end{cases} \tag{2.7}$$

The Metropolis-Hastings algorithm generalizes the Metropolis algorithm by relaxing the assumption that the jumping distribution must be symmetric, so the acceptance ratio is calculated as follows:

$$r_J = \frac{p(y|\theta^*)p(\theta^*)/J_l(\theta^*|\theta^{(l-1)})}{p(y|\theta^{(l-1)})p(\theta^{(l-1)})/J_l(\theta^{(l-1)}|\theta^*)}. \tag{2.8}$$

Other than the acceptance ratio, the Metropolis-Hastings algorithm proceeds in the same way as the Metropolis algorithm does, where the transition distribution of the MCMC is a mixture of a point probability mass centered at the previous draw and a weighted version of the jumping distribution that adjusts for the acceptance ratio.

## 2.3 Derivatives of Bayesian Inference

In this Section, we will introduce several well-documented Bayesian approaches, each of which augments the simple Bayes' rule in a certain way.

### Approximate Bayesian Computation

Revisiting Equation 2.4, to obtain the posteriors of $\theta$ requires going through the calculation of the likelihood function, which is often computationally-intensive. This effort becomes extremely unwieldy when a likelihood function cannot be analytically specified, requiring alternatives such as non-parametric likelihood functions or some surrogate indicators of likelihood that can be obtained rather easily. The computational challenge has been addressed in previous studies, but it is still formidable, especially in situations where $\theta$ is a vector of a large number of parameters. Among all, an appealing alternative is offered by Approximate Bayesian Computation (also known as ABC, Csilléry et al., 2012; Turner and Van Zandt, 2012). In ABC, the computation of the likelihood is bypassed altogether. Instead, a set of summary statistics is computed from the simulated parameters and compared to the summary statistics obtained from the in-situ data (Csilléry et al., 2012), leading to approximate posterior distributions of parameters. A similar concept, although in a different context, was offered by Beven and Binley (1992) in the form of Generalized Likelihood Uncertainty Estimation (GLUE). Although not a Bayesian approach, GLUE intends to bypass expensive optimization searches in complex parameter spaces by ranking parameter sets based on a comparison between simulated and observed data.

The simplest form of ABC is the ABC rejection sampling algorithm (Turner and Van Zandt, 2012). First, a plausible parameter vector, $\theta^*$, is sampled from the prior, $p(\theta)$. Then, we use $\theta^*$ in the forward model to obtain an estimate of the response, denoted by $y^*$. Here we invoke a distance function, denoted by $\rho(y, y^*)$, to quantify the deviation of $y^*$ from $y$. If $\rho(y, y^*)$ is smaller than a user-defined threshold, then $\theta^*$ is kept because it is considered to have a nonzero probability of being in the approximate posterior. Otherwise, $\theta^*$ is discarded (or "rejected") and we move on to sample another plausible parameter vector (hence the name rejection sampling). One may notice that the premise behind ABC is that $\rho(y, y^*)$ should be defined by the way of sufficient statistics (namely, the statistics that provides the same information as the sample itself; e.g., sample mean is a sufficient statistics for the mean of a Gaussian distribution with known variance, because no more information can be extracted from the sample once the sample mean is known). The choice of $\rho(y, y^*)$ can be tricky and would affect the results of ABC. For the details of fairly robust choices of $\rho(y, y^*)$ with respect to some particular summary statistics, as well as some other variants of ABC that feature more efficient sampling of $\theta^*$, please refer to Nott et al. (2012) and Turner and Van Zandt (2012).

## Bayesian Hierarchical Model

In many applications of statistical models, multiple model parameters can be regarded as related or connected in some way by the structure of the problem, implying that a joint probability model for these parameters should reflect their dependence (Gelman et al., 2014). A key feature of such applications is that the observed data are somewhat grouped. For example, a watershed can be divided into sub-watersheds, where a physical property of interest (e.g., slope) can be modeled as a random variable of a fixed distribution form, but the parameters defining that distribution vary from one sub-watershed to another. The available dataset consists of observations grouped by sub-watersheds, where observations from different groups are inter-correlated. Nonhierarchical models are usually inappropriate for such data because in order to model the inter-correlation, the model often ends up using too many parameters, which leads to overfitting. In contrast, hierarchical models can use a population distribution to structure dependence into the parameters, thereby avoiding problems of overfitting (Gelman et al., 2014).

One might be curious about why Bayesian hierarchical models are relevant to unsampled sites. Recall that in Section 1.2 we introduced ex-situ information as a source of information at unsampled locations. Ex-situ data may be collected at multiple sampled locations, and may be inter-correlated from one location to another. The assimilation of such data, as well as the transferring of information from sampled locations to unsampled ones, requires a systematic way to structure the dependence of parameters among locations. Here, we provide a concise mathematical introduction a two-leveled hierarchical model, while interested readers are referred to Gelman et al. (2014).

We start by having the data that are divided into $N_i$ groups, each of which includes $J_i$ observations:

$$\mathbf{y} = \{y_{i,j}; j = 1, ..., J_i, i = 1, ..., N_i\}, \tag{2.9}$$

where the subscripts in $y_{i,j}$ indicate the $j^{th}$ observation in the $i^{th}$ group. Each point observation in each group follows a group-specific distribution, which is parameterized by the group-specific parameter vector, $\Theta_i$:

$$y_{i,j} \sim p(y|\Theta_i), \tag{2.10}$$

Uncertainties within each site, such as sampling uncertainty or the natural variability of the target response, are all parameterized by $\Theta_i$ and modeled with this group-specific distribution.

Second, adding onto Equation 2.10 is the second level: the inter-group level, where uncertainties among groups are taken into account. To that end, the group-specific parameters are modeled by an inter-group distribution that depends on the hyperparameter vector ($\Phi$):

$$\Theta_i \sim p(\Theta_i|\Phi). \tag{2.11}$$

The conditional distribution of $\Phi$ can then be obtained by integrating the joint conditional distribution of all parameters over the group-specific parameters:

$$p\left(\Phi|\mathbf{y}\right) = \int_{\vec{\Theta}} p\left(\Phi|\mathbf{y},\vec{\Theta}\right) p\left(\vec{\Theta}|\mathbf{y}\right) d\vec{\Theta}, \tag{2.12}$$

where $\vec{\Theta} = \{\Theta_1, ..., \Theta_{N_i}\}$. The conditional distribution of $\Phi$ is also an informative distribution indicating an underlying mechanism that is shared among all groups. Such a hierarchical model can account for (1) intra-group uncertainty, (2) inter-group uncertainty, and (3) inter-group correlation/dependence of the group-specific parameters.

## Bayesian Additive Regression Tree

The forward model can be physically based or data-driven/statistical. A typical type of data-driven model is regression model, which models the relationship between some predictor variables (or predictors) and the response variable of interest. A general form of data-driven models can be expressed as follows:

$$Y = \hat{Y} + \epsilon = f\left(\theta, \xi\right) + \epsilon, \tag{2.13}$$

where $f\left(\cdot\right)$ denotes the statistical forward model that estimates of the response variable ($\hat{Y}$), with parameters ($\theta$) and predictors ($\xi$), and $\epsilon$ is an error term characterized by Gaussian white noise with finite variance, i.e., $\epsilon \sim N(0, \sigma^2)$. Among data-driven models, classification and regression tree (CART, Breiman, 1984) is particularly suitable for studying various hydrologic responses because:

1. it can account for non-linear and interaction effects of the predictors at low computational costs, which is desirable for complicated and non-linear hydrologic processes, and

2. it can handle quantitative (e.g., annual rainfall), binary (e.g., snow existence), and multinomial predictors (e.g., land cover type).

A schematic diagram of a CART model is shown in Figure 2.1**(a)**, which resembles an upside-down tree (root on top and leaves at the bottom). The root node of the tree represents the space spanned by the predictor(s). As one moves downward from root to leaves, the said space is recursively partitioned by a sequence of binary partitioning rules. This partitioning and the corresponding partitioning rules define the tree structure, and can be represented by the tree structure variable, denoted by $T$. After partitioning, output response values are assigned to each and every leaf, where each leaf represents a partitioned subspace. These output values can be collectively denoted by $\mathbf{M}$. A CART model can be fully defined by knowing its $T$ and $\mathbf{M}$. To further improve the predictive performance on an individual CART, one can establish an additive ensemble tree model by summing $J$ individual trees

(Figure 2.1**(b)**), each of which has its tree structure ($T_j, j = 1, ..., J$) and its set of leaf values ($\mathbf{M}_j, j = 1, ..., J$), shown as follows:

$$Y = \hat{Y} + \epsilon = \sum_{j=1}^{J} g\left(T_j, \mathbf{M}_j, \xi\right) + \epsilon, \tag{2.14}$$

where $g(\cdot)$ denotes an individual tree. The output of an additive ensemble tree model is the sum of the outputs from the $J$ trees. If in-situ data are available, one can fit the ensemble model with data by searching for the best $T_j$ and $\mathbf{M}_j$ for every $j$, where the definition of "best" can be different under different contexts (e.g., least classification error, lowest predictor error, etc.). However, this model fitting process does not allow for a straightforward quantification of the uncertainty in the parameters.

Fortunately, Bayesian additive regression tree (BART) offers a solution. The details of BART, including the establishment of prior, the calculation of likelihoods, and the posterior inference statistics are well documented in Chipman et al. (2010) and Kapelner and Bleich (2016). Here, we provide a brief conceptual introduction. BART replace model fitting with Bayesian inferencing by first defining the following joint prior of all the tree structures, all the sets of leaf values, and the white noise variance:

$$p\left(T_1, \mathbf{M}_1, ..., T_j, \mathbf{M}_j, \sigma^2\right) = p(\sigma^2) \prod_{j=1}^{J} p(T_j) P(\mathbf{M}_j | T_j) \tag{2.15}$$

BART then applies a tailored version of backfitting MCMC simulation to condition the prior on $r$, where backfitting means the $j^{th}$ tree is iteratively updated with its partial residual. The stationary distribution toward which the MCMC simulations converge is then used to approximate the true posterior:

$$p\left(T_1, \mathbf{M}_1, ..., T_j, \mathbf{M}_j, \sigma^2 | r\right). \tag{2.16}$$

A schematic diagram of the MCMC simulation iteration procedure is shown in Figure 2.1**(c)**. Within each MCMC simulation, the $T_j$ are sampled with Metropolis samplers, while $\mathbf{M}_j$ are sampled with Gibbs samplers conditioned on the previously sampled $T_j$; this process is done iteratively for all $j$, and is conceptualized as the loop in the blue circle in Figure 2.1**(c)**. After simulating all the trees, the error variance ($\sigma^2$) is simulated with a Gibbs sampler conditioned on previously sampled $T_j$ and $\mathbf{M}_j$ for all $j$. The sampling of $\sigma^2$ marks the end of one MCMC simulation. We can then see by the loop in the red square in Figure 2.1**(c)**, the MCMC simulation is continuous, until the simulated values converge to a stationary distribution. These post-convergence simulated values approximate realizations from Equation 2.16, and thus we approximate the true posterior in Equation 2.16 by the stationary distribution obtained by MCMC simulation. At this point, we have reached a BART model that is conditioned on the response data, because all the BART parameters (tree structures, leaf node values, and the white noise variance) have been conditioned on the response data.

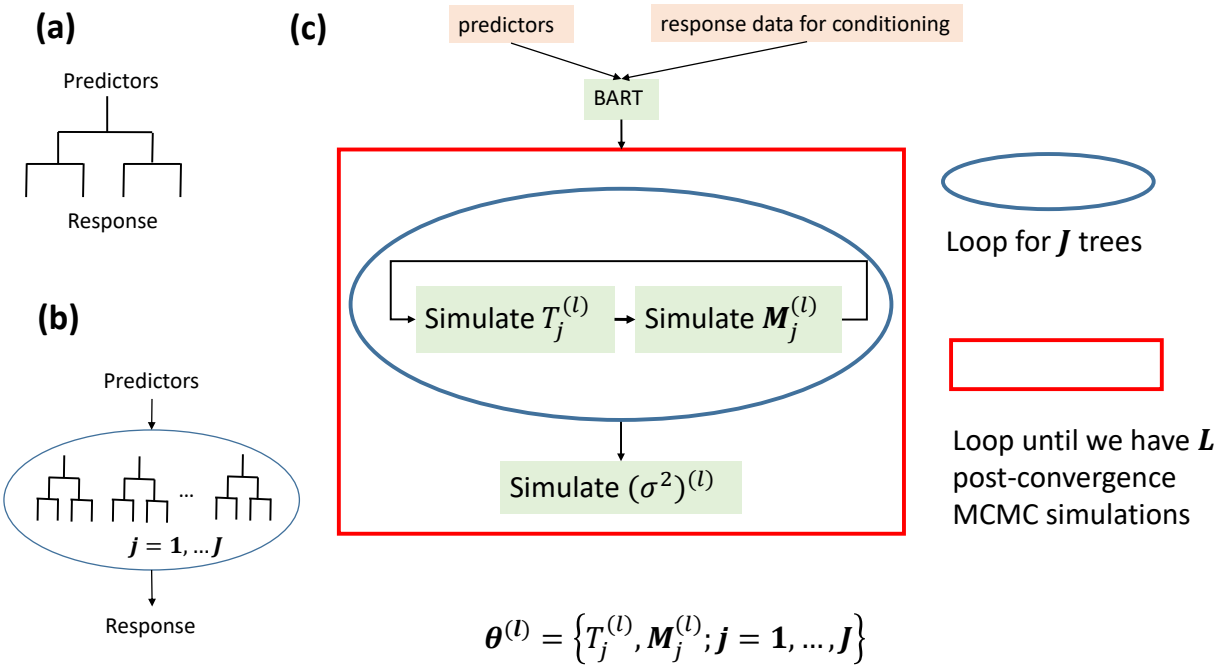


Figure 2.1: Schematic diagrams of **(a)** a regression tree model, **(b)** an ensemble tree model which consists of $J$ additive regression tree models, and **(c)** the loops structure that BART uses to draw MCMC simulations (indexed by $l$), consisting of an inner loop for $J$ additive regression tree models and an outer loop for a total of $L$ MCMC simulations.

In summary, BART can be understood as Bayesian inference applied on many individual CART models, so that we obtain posteriors of the parameters, which could lead to conditional predictive distributions of the target response variable.

Now let us tern our attention to making predictors given a new vector of predictors, denoted by $\tilde{\xi}$. Plugging $\tilde{\xi}$. Firstly, Equation 2.13 can be rewritten as:

$$Y \sim N\left(\hat{Y}, \sigma^2\right). \tag{2.17}$$

Both the mean and the variance in Equation 2.17 are uncertain, and have their respective posteriors. By combining Equations 2.14 and 2.17, and after plugging in the post-convergence MCMC simulated values and $\tilde{\xi}$, we obtain a plausible realization (indexed by the superscript $l$, $l = 1, ..., L$) of predictive distribution as follows:

$$N\left(\hat{Y}^{(l)}, (\sigma^2)^{(l)}\right) = N\left(\sum_{j=1}^{J} g\left(T_j^{(l)}, \mathbf{M}_j^{(l)}, \tilde{\xi}\right), (\sigma^2)^{(l)}\right). \tag{2.18}$$

The collection of many plausible realizations yields an approximated posterior of predictive

distributions. Thus, for response of interest, we have now obtained a fully Bayesian Gaussian predictive model, where the mean and the variance have their respective posteriors.

BART fits itself in a nice niche in between various data-driven models (such as regression models, neural network models, etc.) and various Bayesian models. Compared to other data-driven models, the advantage of BART is that it allows for the determination of a full Bayesian posterior of parameters and thus full Bayesian predictive distributions, rather than just a few estimates/predictions or posterior statistics. Compared to Bayesian models, the advantage of BART is its built-in capability of taking multiple exogenous variables (i.e., the predictors) in various forms (continuous, discrete, multinomial, etc.) into account to model complicated non-linear predictor-response relationship.

# Chapter 3

# Rapid Impact Modeling: Efficient Computation for The Assimilation In-situ Soft Data

## 3.1 Introduction

In this Chapter, we approach the challenges of efficient assimilation of in-situ soft data at unsampled sites. Although a general description has been provided in Section 1.2, to illustrate the challenges more vividly, let us turn our attention to an interesting example.

The Mingtang deep underground tunneling project is located in a mountainous region in China, of 7.5 $km$ in length and at depths exceeding 450 $m$ in a mixed granite/gneiss rock formation. From a few boreholes along the tunnel, prior to the tunnel construction the local groundwater table was found to be only a few meters to twenty meters below the ground surface, which is significantly smaller than the depth to the tunnel. During the tunnel construction, significant groundwater drawdown was observed, and local farmers reported that the local paddy fields were drying up (Figure 3.1), potentially as a consequence of groundwater inflow into the tunnel. The main environmental impacts of concern relate to ecosystem health and sustainable agriculture. The drawdown is expected to have a direct impact on local residents in terms of domestic and agriculture water supply. First, local groundwater is a relevant source for water uptake, especially during drought periods when the soil moisture reserve is low (Domec et al., 2010; Peñuelas and Filella, 2003; Vincke and Thiry, 2008). Severe drawdown could deprive the ecosystem of this drought resilience. Second, rice paddies may be a valuable source of groundwater recharge (Anan et al., 2007; Greppi, 2004; Imaizumi et al., 2006; Liu et al., 2005), but recharge is not the primary purpose of actively farmed paddies, and drying paddy fields is against the local farmers' interest because rice paddies should be submerged in a shallow layer of water during the growing season. Unfortunately, although the Mingtang tunnel connects two major urban areas, it is located in a largely uninhabited area in a mountainous region with limited economic activity.

As a result: (1) there is no information on the spatial variability any hydraulic parameter, and there is no information on groundwater recharge; (2) there is no information available to indicate the spatial extent of the groundwater basin and the boundary conditions; (3) there is no information on water table elevation except at a few boreholes along the tunnel. The only immediately available data are the measurements of groundwater infiltration into the tunnel during the tunnel construction, which is an example of in-situ soft data.

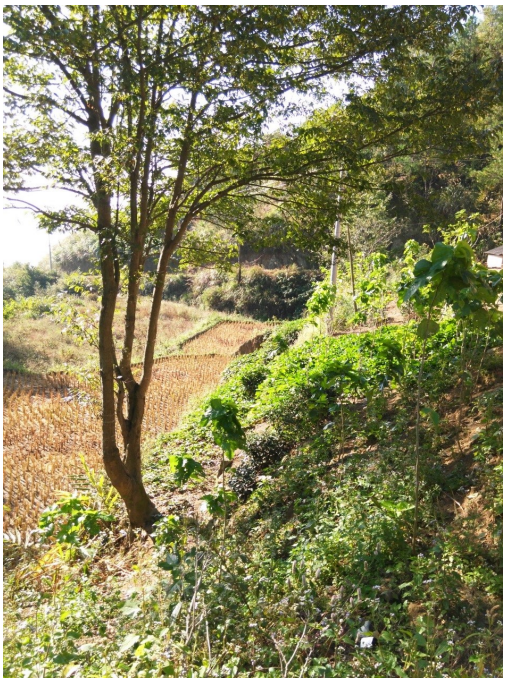

Figure 3.1: Picture taken near the Mingtang site; following tunnel construction, the paddy fields were dried up and ceased to be agriculturally productive.

This is a typical example of a heavy-impact yet poorly sampled project, a situation quite common in large-scale projects in general, and deep underground structures in particular (e.g., Molinero et al., 2002; Yang et al., 2009). The Mingtang Tunnel project quite accurately reflects the challenges we want to address in this Chapter: the complexity of the site, on one hand, and the scarcity of in-situ hard data on the other. The lack of in-situ hard data forces us to look for alternative information sources (such as soft data), but the large spatial scale and the heterogeneity in the area lead to: (1) high computational demand in inverse modeling due to the difficulty in the computation of the likelihood function, (2) high computational demand in forward modeling due to the large spatial scale and the complexity of the site. Consequently, it is necessary to come up with an efficient approach to assimilate the in-situ soft data, in order to provide informed estimates of the hydrologic response of interest, assessment of the environmental impact of drawdown.

## 3.2 Rapid Impact Modeling: Theoretical Background

To achieve our goal, we have to tackle the computational demand in the inverse modeling and in the forward modeling. The approach of ABC introduced in Section 2.3 provides a great starting point, but that is not enough. Wagener and Montanari (2011) pointed out that the ultimate goal of predictions at ungauged watersheds is not to define parameters of a model, but rather, to understand the expected behavior at the ungauged watersheds of interest. Following the same thought, we recognize that in many situations the ultimate targets of obtaining posteriors of parameters are not the model parameters *per se*, but the resulting hydrologic/hydrogeologic responses (or their environmental impacts) from forward models.

With that in mind, we propose a new approach, named Rapid Impact Modeling (RIM, Li et al., 2018), which follows a somewhat similar line of thinking as ABC but takes it in a different direction. By realizing that posterior distributions of parameters are but an interim step along the way, we can avoid altogether the need to estimate posteriors of the model parameters, focusing instead on stochastic modeling of hydrologic responses and the associated environmental impacts (e.g.,, impacts of groundwater drawdown). With RIM, one can address uncertainty in the presence of limited data availability, in a practical way and without sacrificing rigor, and to evaluate the environmental impact of concern. The mathematical details are as follows.

Following the same notation as in Section 2, let $Y$ denote the random variable we use to model the target response, let $y$ denote the observations of the target response (which is of course unavailable at unsampled sites), and let $\theta$ denote the vector of parameters for the forward model.

1. Since $y$ is unavailable, the first step in RIM is to determine a total of $C_i$ intervals of $y$ (e.g., 0-1, 1-2, etc.), denoted as $y_i', i = 1, ..., C_i$.

2. We can then characterize $Y$ by its probability mass function:

$$P(Y \in y_i'), \tag{3.1}$$

   which denotes the probability the $Y$ is within $y_i'$. Equation 3.1 provides the unconditional distribution of $Y$.

3. Equation 3.1 can be solved with the aid of the total probability theorem, as follows:

$$P(Y \in y_i') = \int_\theta \{P(Y \in y_i')|\theta\}\, p(\theta) d\theta, \tag{3.2}$$

   where the integration is taken over the parameter space spanned by $\theta$. Equation 3.2 can be solved using Monte Carlo integration. This entails generating multiple realizations of $\theta$ from $p(\theta)$ (just like the first step in ABC), followed by running the forward model. Here we let $\theta^*$ denote a plausible realization of $\theta$, and let $y^*$ denote the estimated response using $\theta^*$.

4. Further simplification of Equation 3.2 is enabled by dividing the parameter space into hypercubes, with the aid of the Total Probability Rule, as follows:

$$P(Y \in y_i') = \sum_{j=1}^{C_j} \{P(Y \in y_i')|A_j\}\, P(A_j), \tag{3.3}$$

where $A_j$ denotes one of the $C_j$ hypercubes defining subsets of the parameter space. The probabilities $P(A_j), j = 1, ..., C_j$ are obtained from priors, subject to the constraint $\sum_{j=1}^{C_j} P(A_j) = 1$. Equation 3.3 provides us with the means to assign probabilities to $y^*$ based on prior information.

5. The next step is to condition the probabilities in Equation 3.2 on in-situ soft data. Let $z$ denote the in-situ soft data, and let $z^*$ denote the output from the transfer function using $y^*$. The conditioning is done by determining intervals of the absolute difference between $z$ and $z^*$, expressed as $|z - z^*| \in e_k, k = 1, ..., C_k$ for a total of $C_k$ intervals. The conditional probabilities can be expressed as follows:

$$P(Y) \in y_i'|e_k) = \sum_{j=1}^{C_j} \{P(Y \in y_i')|e_k, A_j\}\, P(A_j|e_k). \tag{3.4}$$

6. Because of the correspondence between $z^*$ and $\theta^*$, by classifying $|z-z^*|$ based on $e_k$, we are simultaneously classifying $\theta^*$ based on $e_k$. This will allow us to obtain the statistics of $Y$ for any given error interval. This is the essence of RIM as it provides the means to analyze $Y$ (and its impact) directly, without going through a computationally-intensive inverse modeling.

Because RIM directly links $e_k$ with $\theta^*$, it bypassed not only the interim step parameter estimation, but also the step that converts the posterior of parameters into the posterior of the hydrologic response. Thus, the computation demand for obtaining estimates that are conditioned on in-situ soft data has been significantly reduced. Besides efficiency, a noteworthy advantage of RIM is that the conditioning information is used in its original form, without being reduced to summary statistics, and hence RIM does not lose any information compared to alternative methodologies such as ABC. In addition, although the highlight of RIM is on efficient assimilation of in-situ soft data, ex-situ information can also be accounted for in terms of an ex-situ prior (i.e., a prior that is informed by ex-situ information), leading to realizations $\theta^*$ that are informed by ex-situ information.

One should note that the selection of the maximum error bracket to consider in Equation 3.4 is subjective. The range of values considered for the errors should reflect a wide range of scenarios. Large errors may be associated with low probabilities, but they could represent extreme scenarios, and need to be considered. This not only allows the modelers to cover a wide range of scenarios, but also offers a simple mechanism by which the modelers can adjust the degree of conditioning based on case-specific considerations.

## 3.3 Case Study: The Mingtang Tunnel Project

In this Section, we will demonstrate the power of RIM in a case study. The case study is built around the Mingtang tunnel project introduced in Section 3.1. As mentioned, the Mingtang project is a typical example of a heavy-impact yet poorly sampled project. The impact of groundwater drawdown due to tunnel construction is of concern, but only very limited in-situ information is available, provided as follows.

### Location

The Mingtang Site is located in Yuexi County, Anhui Province, mid-eastern China (Figure 3.2)). It connects Hubei province and Anhui province. The tunnel is 7.548 $km$ long and runs from east to west. The maximum depth of the tunnel is about 548 m below ground surface. Our model domain consists of a study area of approximately 180 $km^2$ around the Mingtang tunnel.

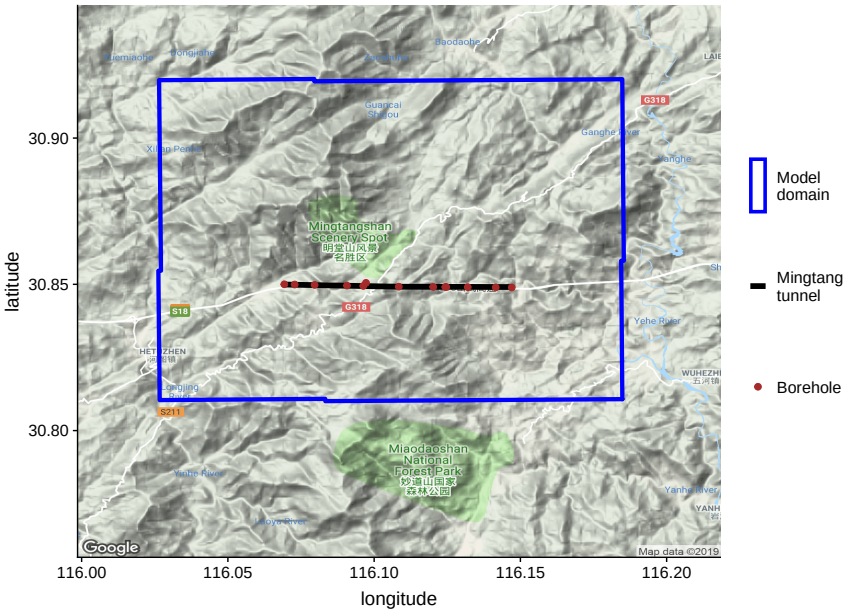


Figure 3.2: Location of Mingtang tunnel, the boreholes, and the model domain.

## In-situ Information

Despite the data scarcity, limited information about the Mingtang site is still available. The topography is mountainous with moderate slopes. The average temperature is 14.5°C. The rainfall in spring and summer account for more than 70% of the roughly 1400 $mm$ average annual precipitation. The land cover is primarily pine forests, with sparsely distributed farmlands. At the boreholes, the pre-construction groundwater table was found to be about several meters to twenty meters below the ground surface.

Information about the local geology is available along the tunnel and at the boreholes. The rock mass along the tunnel axis consists primarily of moderately weathered granite and gneiss (Figure 3.3). There are three major faults along the tunnel axis, marked as F7, F6, and F14, respectively from west to east, but only two (F6 and F7) were encountered during excavation and thus considered in the model. Additional details are provided in (Chen et al., 2017; Li et al., 2016).

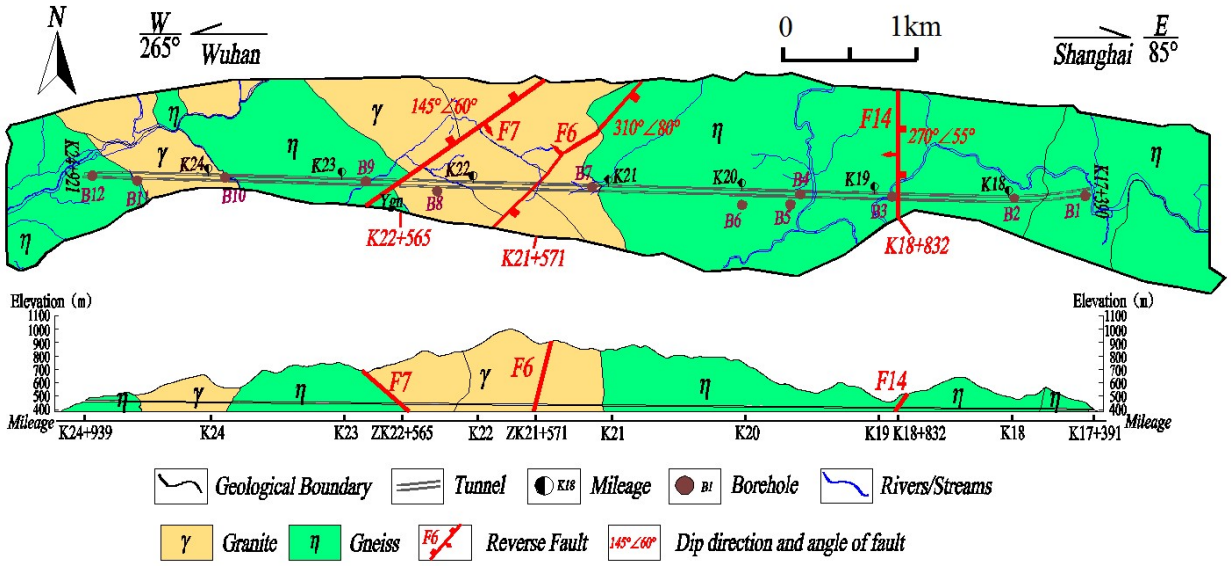


Figure 3.3: Geological cross-sections of the study area. The upper and lower diagrams provide a planar view and vertical cross-section, respectively.

The only in-situ measurements are the groundwater infiltration into the tunnel, monitored during tunnel construction. The infiltration was first collected by circumferential drains installed outside of the tunnel's waterproof later, and then gathered into a central drainage pipe buried under the pavement of the tunnel (Figure 3.4**(a)**). The infiltration flow rate was estimated as the flow cross-section (calculated from measured flow depth, Figure 3.4**(b)**) multiplied by the average flow velocity (Figure 3.4**(c)**) in the drainage pipe. Note that the measured flow is the total infiltration of groundwater along the tunnel. Normalized by the tunnel length, the average infiltration is $0.65m^2/day$, and the measurement error is estimated to be $0.03m^2/day$.

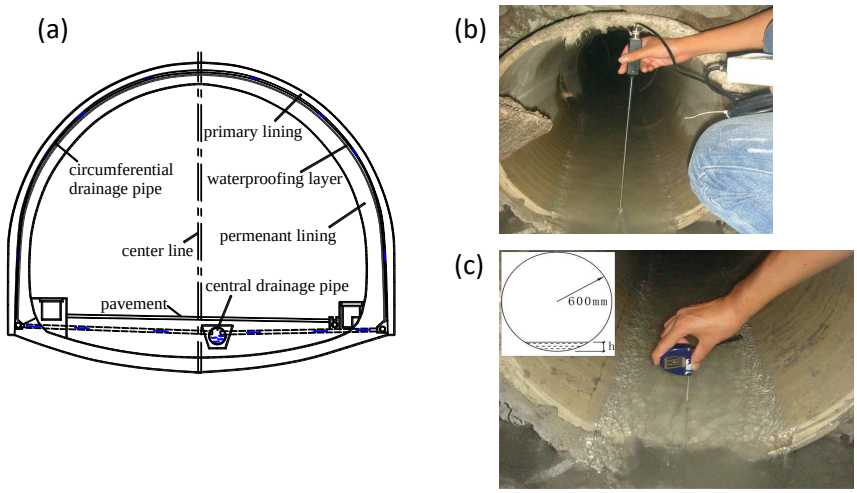


Figure 3.4: **(a)** Schematic diagram of tunnel cross-section and drainage system. **(b)** Measurement of flow depth in the drainage pipe using a pocket tape. **(c)** Measurement of flow velocity in the drainage pipe using a portable flowmeter.

## 3.4 Stochastic Modeling of Groundwater Drawdown with RIM

We follow the same theoretical setup from Equation 3.1 through Equation 3.4 apply RIM at the Mingtang site. Ideally, the hydraulic conductivity as well as the recharge to the aquifer should be considered as spatial (or spatiotemporal) random variables, and we should model the associated heterogeneity and uncertainty. However, due to the lack of in-situ data, it would be difficult to do so as the uncertainty is likely to be prohibitively large. In designing a solution strategy, we should note the multiple length scales that define a groundwater flow and transport problem. This includes the scale of the flow domain, the scales of heterogeneity, the scales of the target variables, measurement scales and problem-specific scales (e.g., the scale of the solute plume). The relationship between the scales offers opportunities for simplification (e.g., De Barros and Rubin, 2011; Rubin et al., 1999). The strategy adopted for this study calls for populating the cells in the forward groundwater model with effective hydraulic conductivities, which is modeled as a random variable in this case study. The effective conductivity represents homogenization over large spatial extents. With this simplification we trade the modeling local heterogeneity off against the stochastic modeling of homogenized responses. Likewise, we apply the same simplification to recharge, and consider the average recharge (here referring to recharge homogenized over a year and over the whole model domain) as a random variable in this case study. Consequently, we end up with a forward groundwater model with two random variables: the effective hydraulic conductivity and the average recharge. To further reduce the computation demand, we made a special design of model domain discretization, which is is detailed in Appendix B.

Since groundwater drawdown is a spatial function, we let $x$ denote the location and rewrite Equation 3.4 as

$$P(Y(x) \in y_i' | e_k) = \sum_{j=1}^{C_j} \{P(Y(x) \in y_i') | e_k, A_j\} P(A_j | e_k). \tag{3.5}$$

In this case study, $Y(x)$ is the random variable we use to model the drawdown, $y_i'$ denotes the $i^{th}$ interval of drawdown value, and $e_k$ is the $k^{th}$ interval of the error, obtained by comparing the modeled infiltration into the tunnel with the measured infiltration.

Up to this point, the only thing left undefined is the prior. We need to define the prior for the effective conductivities of granite and gneiss, as well as the prior of the average recharge. At the Mingtang Site we have limited in-situ information, primarily of a qualitative nature, not enough for establishing priors. However, qualitative in-situ information is sufficient for providing some indications of which sampled sites could be used for borrowing information. Compared to totally uninformative priors, even a weakly informative prior may greatly improve computational efficiency in the application of RIM, because we may rule out unlikely values of the parameters that do not need to be considered for sampling. Therefore, we will be relying on ex-situ data for the establishment of priors; hence we refer to the priors as ex-situ priors. The establishment of ex-situ priors, in a nutshell, involves a Boolean variable that determines whether a candidate site is similar to the Mingtang site or not, and a set of criteria we use to determine the value of the said Boolean variable. The concept of using ex-situ priors is important in efficient assimilation of in-situ soft data because the sampling in RIM is thus made more efficient. On the other hand, the details of how we established ex-situ priors are not the innovation of RIM we want to highlight here, and thus are provided in Appendix C.

## 3.5 Case Study Results

This Section explores the unconditional and conditional statistics of the drawdowns at multiple locations, and evaluates: (1) the predictive power of the ex-situ informative priors (unconditional predictions), and (2) the effect of conditioning the predictions on the tunnel infiltration fluxes.

### Exploring The Parameter Space

A scan of the parameter space is provided in Figure 3.5, showing, for each combination of parameter values, the corresponding estimation error of the tunnel infiltration flux. The scan is performed over the full range of parameter values as defined by the priors in Appendix C. This scan provides the means to distinguish between subsets of parameters based on their predictive capabilities. For example, the region dominated by blue is defined by parameter sets that most accurately predict the tunnel infiltration fluxes, whereas the

yellow-dominated regions are populated by parameter combinations that provide a poorer match with observations of tunnel infiltration flux.

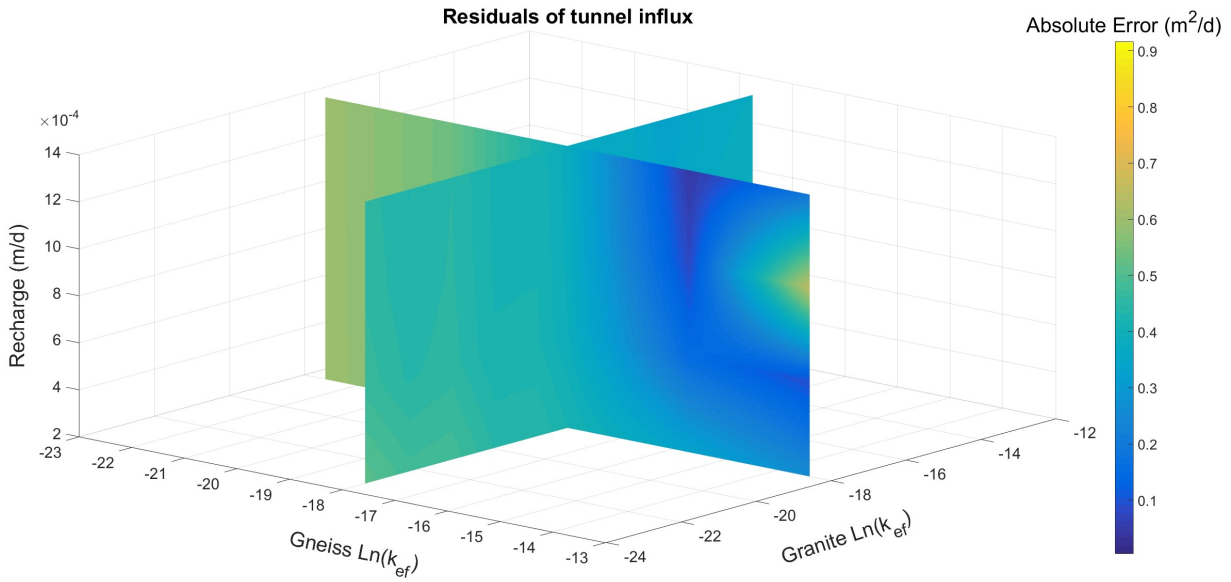


Figure 3.5: Vertical cross-sections of the 3-D parameter space. The color code represents the absolute error (in $m^2/day$) in predicting the tunnel infiltration flux (defined as the absolute difference between computed and measured fluxes). Note: Natural log scale used for the conductivities.

The parameter scan is used to construct the conditional distributions of the model parameters, following Equation 3.4. Figure 3.6 shows the unconditional and conditional distributions of the log effective conductivity of Granite. As mentioned in Section 3.3, the average measured flux is about 0.65 $m^2/day$, with measurement error approximately 0.03 $m^2/day$. Thus, we expect the total error to be in the "blue" parameter range in Figure 3.6, but we allowed for a wider range of errors, corresponding to the parameter space scan, to be on the conservative side and to allow project managers to decide on acceptable error ranges based on the projected impacts.

One can observe from Figure 3.6 that as the error decreases, the distribution gradually converges toward a unimodal and approximately Gaussian distribution, with a concentration of probability around a well-defined mode. It shows a marked reduction in spread and modality compared to unconditional or less conditional distributions. It is, of course, to be expected that conditioning on field-data would have a favorable effect, but what is significant here is the conditioning of the effective conductivity on soft information (the infiltration flux). Furthermore, recalling that the effective conductivity represents homogenization over large spatial extents, this result is important in that it shows the conditioning on the tunnel infiltration flux carries an impact on a variable defined over large support volumes. The second point above deserves further discussion. The effective conductivity is not a spatial

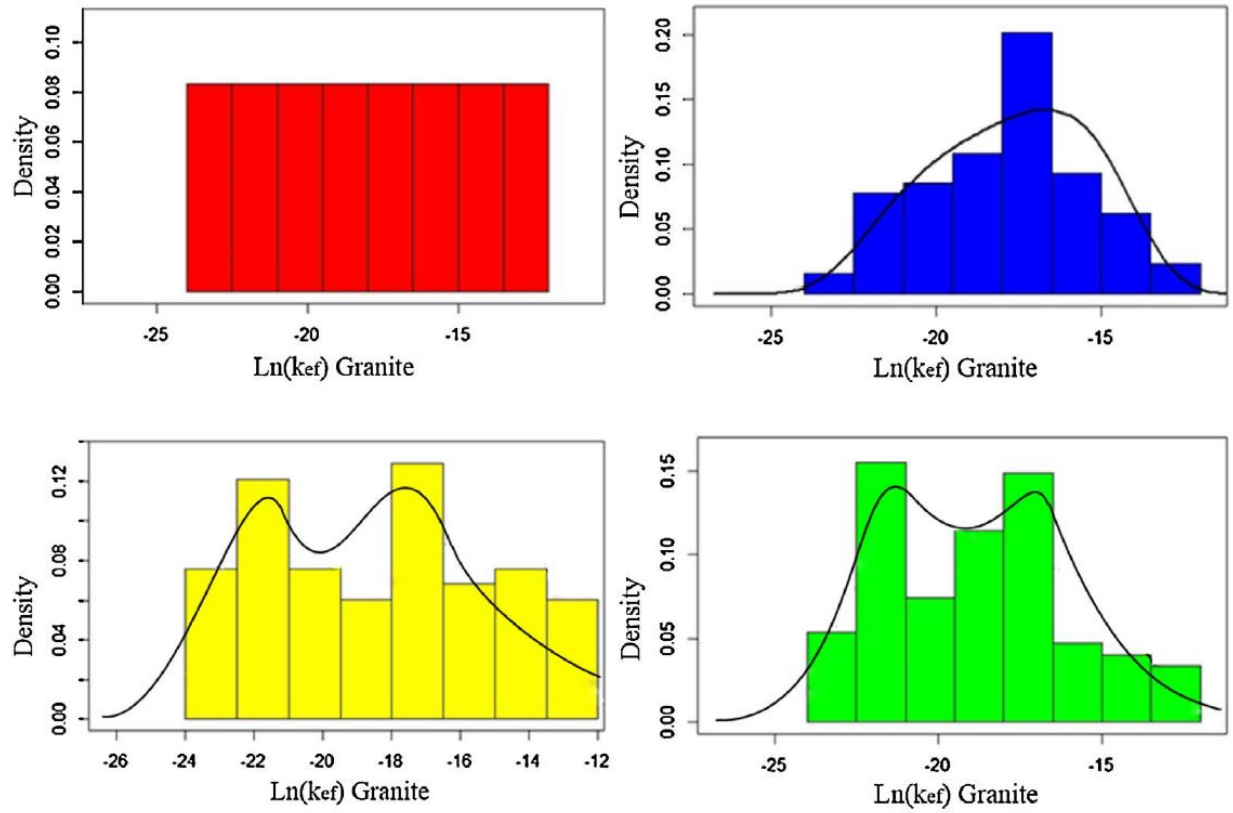


Figure 3.6: The prior (unconditional, shown in red) and posterior (conditional) distributions of the effective hydraulic conductivity of granite, $\ln(k_{ef})$. The colors represent error intervals: (1) Yellow: 0.45 - 0.65 $m^2/day$, (2) Green: 0.25 - 0.45 $m^2/day$, and (3) Blue: below 0.25 $m^2/d$, as defined in Figure 3.5. The plot in red is the unconditional ex-situ prior.

random function and it is not, in general, defined by a statistical distribution because it is defined over the ergodic limit of a stationary variable. In the present study, the effective conductivity is defined as a uniformly distributed random variable based on the prior, varying between its lower and upper bounds, which opens the door for conditioning.

## Statistics of Drawdown

This section analyzes the drawdowns at 4 different locations, shown in Figure 3.7, representing each a different combination of distances from the tunnel and from the fault planes. Locations 1 and 2 are quite close to the tunnel and to faults. The cumulative distribution function (CDF) of drawdown at these two locations are very similar, and are shown in Figure 3.8. The CDFs of drawdown at location 3 and 4 are shown in Figures 3.9 and 3.10, respectively. These results suggest the following: (1) conditioning on the average flux has

stronger impact at locations 3 and 4, whereas locations 1 and 2 show limited sensitivity to conditioning; (2) the parameter set that leads to the smallest error (the "blue" in Figure 3.6) provides the narrower CDFs of simulated drawdown; (3) drawdowns increase as we get closer to the tunnel or to the faults, but they are still significant at distances of several kilometers from the tunnel. The drawdowns in locations 3 and 4 are quite similar, although location 4 is much closer to the tunnel. In both cases the drawdowns are much smaller than those observed for locations 1 and 2. This shows the strong impact of the faults on the drawdowns, with the effects of the faults stretching out over large distances away from the tunnel. Also, locations 3 and 4 show much more sensitivity to conditioning, and that have a much stronger impact on drawdowns compared to the tunnel. To summarize: the drawdowns in locations 1 and 2 are dominated by the faults and are insensitive to conditioning. Proximity to the tunnel (not near faults) does not lead to extreme drawdowns.

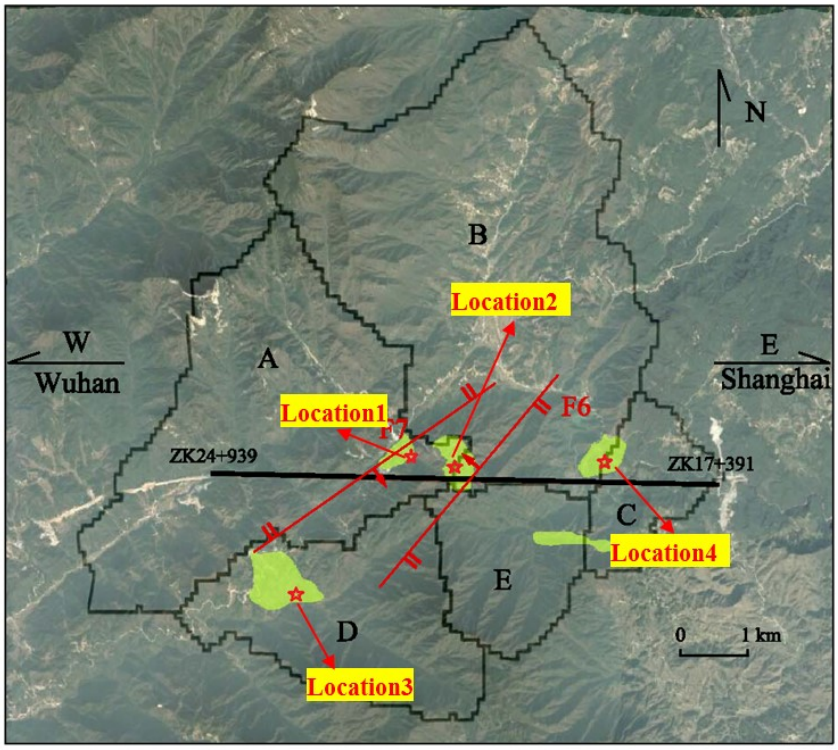


Figure 3.7: Locations analyzed for drawdowns. The Mingtang tunnel is indicated by the thick black line, with the boreholes at both ends marked. The green spots represent agricultural areas identified from satellite images.

To understand the effect of conditioning, we should note that the tunnel infiltration flux, i.e., the in-situ soft data used for conditioning, represents an average taken over a very large stretch of the tunnel, and as such is not representative of local conditions, except in an average sense. The infiltration flux is controlled by the effective conductivity and

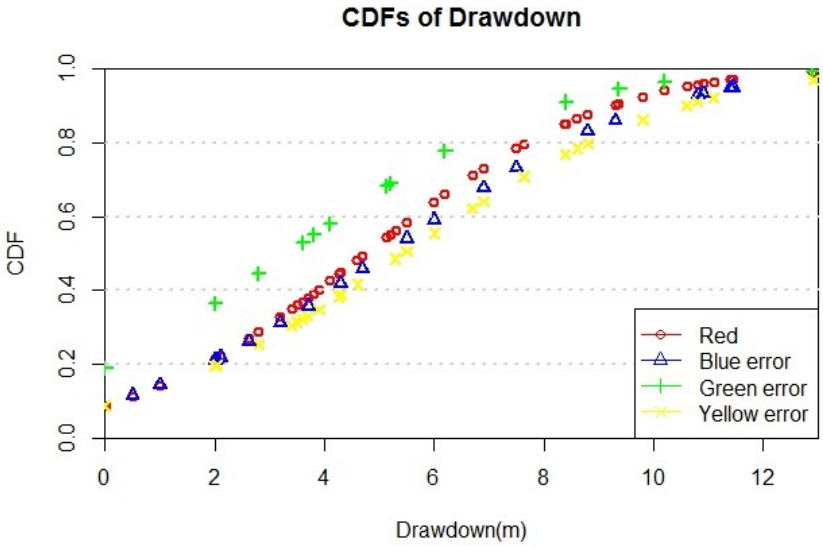


Figure 3.8: CDFs of the drawdowns for Location 1 and 2. The various colors represent the error level used to condition the simulations, as defined in Figure 3.6.

not by local conditions, and as such, estimating the effective conductivity benefits from conditioning on the fluxes. The drawdowns at locations 1 and 2 are dominated by local conditions (proximity to the faults) and hence they are not informed by conditioning on the average flux. At locations 3 and 4, on the other hand, being far removed from the faults, the factors affecting the drawdown are less local, and hence the drawdown statistics are informed by, and are sensitive to, the information on the average flux.

The takeaway messages are as follows.

1. The tunnel engineering perspective: tunnel structures experience enormous pressures. The pressure can be reduced by allowing water to drain through the tunnel's walls. This obviously has an impact on construction costs, but it comes with an environmental price. The larger the savings on construction costs, the larger would be the environmental impacts.

2. The environmental impact perspective: despite the mitigating effects of the tunnel's construction details, drawdowns are significant, even at distances of a few kilometers from the tunnel. For the agricultural practices common in the study area, even a small drawdown of the order of 1–5 m is detrimental.

3. The stochastic approach perspective: the new concept that emerges from this discussion is that of selective conditioning through the Total Probability Theorem. It was formally summarized in Equation 3.4. What that means is that rather than deriving the

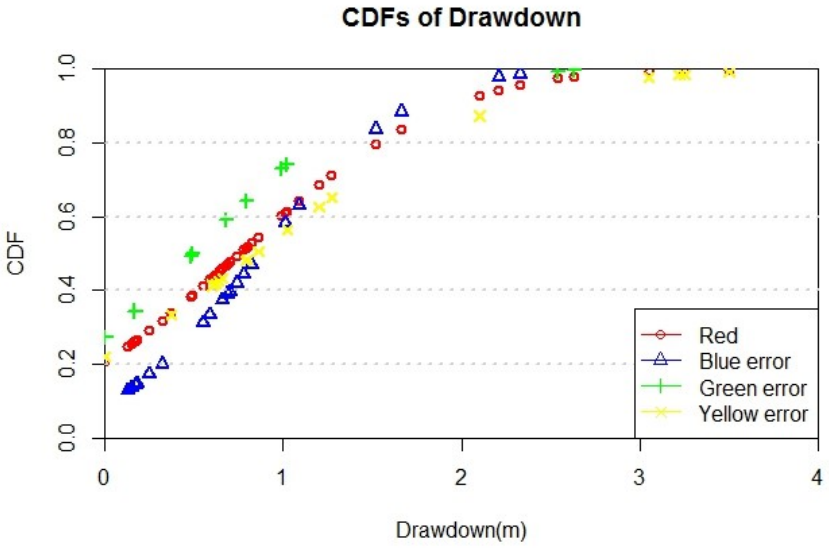


Figure 3.9: CDFs of the drawdowns for Location 3. The various colors represent the error level used to condition the simulations, as defined in Figure 3.6.

posterior distributions of the effective conductivity, conditional to the tunnel infiltration flux, we derive the conditional distribution of the drawdowns, limiting predictions to only those parameter sets that are in the best agreement with the observations. This offers significant savings in computations, because it eliminates the need to derive the likelihood function (which is extremely demanding at the tails).

## The Environmental Impact of The Drawdown

The pine forest at the Mingtang site is the dominant land cover. According to the criteria listed by Eamus et al. (2006), it is likely a part of a groundwater-dependent ecosystem. To assess the potential for impacts, a comparison is needed between the rooting depth of the vegetation and the CDFs of drawdown. Canadell et al. (1996) showed that the maximum rooting depth of pine trees on weathered granite could range from 4 $m$ to 7.5 $m$, while Peñuelas and Filella (2003) found evidence of hydraulic lift by pine tree at 8 $m$ in depth. This depth 8 $m$ is deeper than the drawdown shown in Figures 3.9 and 3.10 , but corresponds only to roughly the median values of the CDF value in Figures 3.8 . In other words, the drawdowns at locations 1 and 2 are likely to exceed the maximum rooting depth. This leads to considering how the pine trees would respond to drawdown in a drier environment.

The effects of drier environments on pine rooting depth have been found to be positive. Laiho and Finér (1996) studied pine mire in southern Finland and showed that following drawdown of tens of centimeters, the living root biomass increased considerably. Bakker

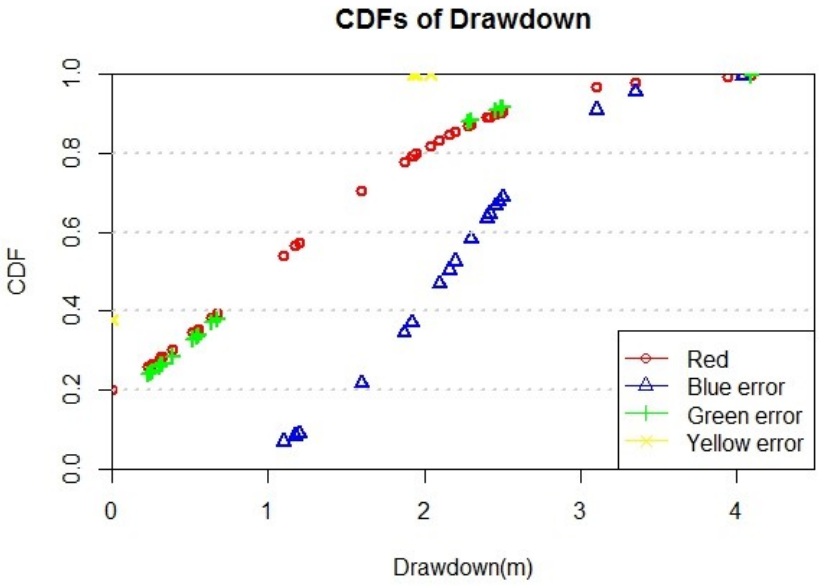


Figure 3.10: CDFs of the drawdowns for Location 4. The various colors represent the error level used to condition the simulations, as defined in Figure 3.6.

et al. (2006) studied the fine roots of *Pinus pinaster* and showed that when the water table depth was 3 $m$ instead of 1.3 $m$, the rooting depth was 3 $m$ instead of 0.9 $m$. In other words, it is reasonable to expect pine vegetation to be capable of growing deeper roots to fetch deeper water. The depth to the water table at the Mingtang site ranges from 0.1 to 10 m. Given the depth range at which pine trees can access water, a drawdown of the order of 1 m is considered surmountable by either growing deeper roots (e.g., Bakker et al., 2006; Laiho and Finér, 1996) or hydraulic lifts (e.g., Domec et al., 2010; Peñuelas and Filella, 2003). According to Figures 3.9 and 3.10, the drawdowns at locations 3 and 4 are very likely to be smaller than 3 m, which is likely to be overcome. However, the drawdown at locations 1 and 2 (as shown in Figure 3.8) could be as extreme as 10 $m$. Therefore, an environmental impact would be expected at locations 1 and 2.

Drawdown is significantly higher when close to both tunnel and faults. If we move away from the tunnel and the faults (e.g., the northern part of Basins A and B, and the southern part of Basin D, in Figure 3.7), the modeled drawdown is generally smaller than 1 m, regardless of values of hydraulic parameters. Thus at these areas, the environmental impact is expected to be negligible. One should note that drawdown does not necessarily lead to wilting of the pine trees, as the available water content in the soil is still accessible. The drawdowns are expected to (and in fact, shown, by local farmers) to affect the rice paddies. During the growing season, rice paddies should be submerged in a shallow layer of water. Schmidt et al. (2011) scanned the soil layers in a rice paddy and showed that the

rooting depth of rice is roughly 20 cm; Morita et al. (1988) showed that the root density of rice plunged significantly at depths larger than 25 cm. Both studies showed that the rooting depth of rice is of orders of magnitude much smaller than the modeled drawdowns. Thus, the effect of drawdown on rice paddies is obvious and significant and can happen at all locations shown in Figure 3.7.

## 3.6 Conclusions

In this Chapter, we propose the approach of Rapid Impact Modeling, RIM, for efficient conditioning of estimates of hydrologic responses on in-situ soft data. RIM's primary innovations are (1) its capability of using all the information available, including in-situ soft information and ex-situ information, to condition the estimates of hydrologic responses, and (2) the efficiency in conditioning the estimates of hydrologic responses, and rapid assessment of impacts, achieved by while bypassing the need for deriving statistical distributions of model parameters. RIM provides a huge reduction in computational effort because it replaces the computation of a likelihood function with integration over the parameter space defined by priors. Thus, RIM offers a systematic way to assimilate both in-situ soft data and ex-situ data for stochastic impact evaluation. By focusing on the total probability theorem instead of posterior parameter distributions, RIM recognizes that in many applications the identification of parameters is just an interim step towards the end goal, which is the estimation of hydrologic responses and the associated impact assessment, and that this interim step incurs disproportionally large computational effort and (in many cases) requires unnecessary data acquisition efforts. With this, RIM complements and augments previous efforts on goal-oriented site characterization (Barros and Rubin, 2008; Barros et al., 2012; Maxwell et al., 1999; Nowak et al., 2010) and on approximate Bayesian computation.

We demonstrated the power of RIM in a case study at the Mingtang tunnel project, a typical example of heavy-impact yet poorly sampled project. The issues of primary concerns include the potential impacts of drawdown in the water table on the natural vegetation of the overlying ecosystem and on agricultural activities. Given that, estimating the extent of the groundwater drawdown is identified as a variable of major concern. We showed how it can be efficiently done to assess the impact of drawdown conditioned on the average groundwater infiltration into the tunnel, a source of soft data, while also systematically accounting for the associated uncertainty. From a broader perspective, this case study also addressed the misconception mentioned in Section 1.2 that a successful application of stochastic modeling requires lots of data.

# Chapter 4

# Nested Tree-based Modeling: Hierarchical Similarity and Bayesian Regionalization

## 4.1 Introduction

In Chapter 3 and Appendix C, we have introduced the establishment of ex-situ priors. Borrowing information from other sampled sites has been actively studied, especially in the form of regionalization (see Section 1.2). However, treating similarity as a Boolean variable may not be the best approach, because we then lose the ability to distinguish a highly similar site from a slightly similar site. It is important and advantageous to consider similarity as a continuous function: from previous studies referenced in Section 1.2, we have learned that the factors determining hydrologic similarity may vary under different conditions, and a universal system of hydrologic similarity still remains unavailable. From there, we recognized the need for an approach to simultaneously (1) account for the dynamic hydrologic similarity system and (2) quantify uncertainties in the parameters, the model structures, and the estimates. To that end, in this Chapter the objectives are twofold. First, we would like to propose a general approach of regionalization that serves the aforementioned need. Second, we would like to integrate the approach with our hypothesized theory to explain how modeling uncertainty and the assimilation of ex-situ data are affected by the dynamic behavior of hydrologic similarity.

## 4.2 Nested Tree-based Modeling

We propose a general approach of regionalization, named the nested tree-based modeling approach, explained in details in this Section.

## Bayesian Additive Regression Tree and Its Advantages

First, let us recall the Bayesian Additive Regression Tree (BART) method introduced in Section 2.3, which is, in a nutshell, Bayesian inference done on an ensemble of classification and regression tree (CART) models. There are two advantages of BART that makes it a good starting point for our need. First of all, BART combines the non-linear regression for the predictor-response relationship with Bayesian inference, allowing for full Bayesian quantification of model parameter uncertainty, which means that we get the conditional distributions of the parameters, rather than a few conditional estimates or conditional statistics. Second, it is a pure data-driven machine learning technique, so it is applicable as long as there are data, even when the data are ex-situ. Thus, regionalization can be done by conditioning the BART parameters on ex-situ data, and one may apply the conditioned BART model at any unsampled site of interest.

The second advantage above deserves a more detailed discussion, as one may argue where does physical knowledge come into play. Prior knowledge of physics is only minimally accounted for in BART, primarily in terms of the composition of the predictor set. Namely, if one believes that a variable $X$ can be used to estimate the response $Y$, then one should include $X$ in the predictor set of the BART model for $Y$. The underlying physical predictor-response relationship is inferred from the ex-situ data via obtaining conditional distributions of the BART parameters, and thus is implicitly embedded, rather than explicitly defined as in physically based models. This of course is a common limitation of many data-driven approaches: the extent to which physics could be inferred is restricted by the conditioning data —in our context, the ex-situ data. However, in compensation, we avoid a disadvantage of the application of physically based models in the case of unsampled sites. The available information at unsampled sites is limited, and it is unrealistic to expect that a certain piece of information should be known. Information unavailability could hinder the implementation of powerful hydrologic models (Razavi and Coulibaly, 2017) because some of the required model inputs may be unavailable (Gemitzi et al., 2017; Xie et al., 2017). It is possible to treat missing inputs as parameters, and run simulations to impute them or apply stochastic methods to estimate them. Nonetheless, the corresponding computational demand grows in power law with the number and the plausible range of the missing inputs, which is of great practical importance when evaluating the pros and cons of an approach. Note that we have no intention to show the superiority of either the data-driven or the physically based models. As Wagener and Montanari (2011) pointed out, the ultimate goal of predictions at unsampled locations is not to define parameters of a model, but rather, to understand what behavior we should expect. We have simply shown why BART is suitable for modeling at unsampled sites.

With its two advantages, BART offers an elegant way to account for model parameter uncertainty while at the same time performing regionalization. Nonetheless, there is still a limitation behind the user-defined composition of the predictor set. The predictor set is defined by answering the question "What variables should be considered for the estimation of the target response?", and thus implies a presumed model structure. As emphasized in

Chapter 1, uncertainty resides not only in model parameters but also in model structures, and thus model structure uncertainty should be considered. An intuitive solution to that is the proposal of multiple plausible BART models, each of which features a unique predictor set. However, that introduces the following problem: defining the weights of these plausible models, or in probabilistic terms, the establishment of a probability mass function (PMF) of plausible models. One can always define an uninformative PMF, i.e., a uniform distribution, but if priors of parameters can be informed by ex-situ data, why cannot the PMF? To that end, we invoke a nesting framework of BART, which will be explained in the next subsection.

## Nesting BART under CART

We approach the challenge of defining a PMF of plausible BART models by using a proposal-comparison procedure, which we termed the nested tree-based modeling approach.

We start by proposing $K$ plausible BART models, denoted as $B_k, k = 1, ..., K$. The model structure uncertainty is accounted for by obtaining a probability mass function of the $K$ plausible BART models, denoted by $p(B_k)$. The determination of $p(B_k)$ can be informed by the data (namely, in an empirical Bayes way, where the prior is informed by the data). At each available data point, we evaluate the performance of the plausible BART models by a performance metric; a typical example is the mean squared error, but it can be any metric that can be obtained from the conditioned predictive distribution. Then, a label is given to each data point, indicating which plausible model has the highest performance measured by the metric. Finally, we build a CART model to classify the data points based on their labels. We refer to this as nesting multiple BART models in one CART model. Like introduced in Section 2.3, this CART model is capable of partitioning the space spanned by all predictors, resulting in partitioned subspaces represented by the leaf nodes. At each leaf node, we obtain an empirical multinomial distribution of the $K$ plausible BART models, indicating $p(B_k)$ in that predictor subspace. Thus, by investigating all the leaf nodes together, one can study the variation of $p(B_K)$ with various predictors, thus exploring the variation of the dominant controls of hydrologic similarity under different conditions.

A very simple example is illustrated in Figure 4.1, where we compare the performances of two BART models ($K = 2$) using one predictor and a simple two-leveled classification tree. The predictor space is partitioned into the positive subspace and the negative subspace by the partitioning rule indicated in the diamond box. Thus, for any new data point with positive predictor value, we would use $p(B_1) = 0.76$ and $p(B_2) = 0.24$ as the probability mass function of plausible models. In real applications, of course, one can use an arbitrary number of predictors to compare an arbitrary number of plausible BART models.

Up to this point, we have introduced the nested tree-based modeling approach. The essence of it is the combination of the proposal-comparison-based consideration of model structure uncertainty, and the various advantages of BART that were discussed before. For estimation purpose, one would be interested in accounting for model structure uncertainty by averaging the estimates over $p(B_k)$, which can be done by invoking Bayesian model averaging. However, the capability of the nested tree-based modeling approach does not stop here, as

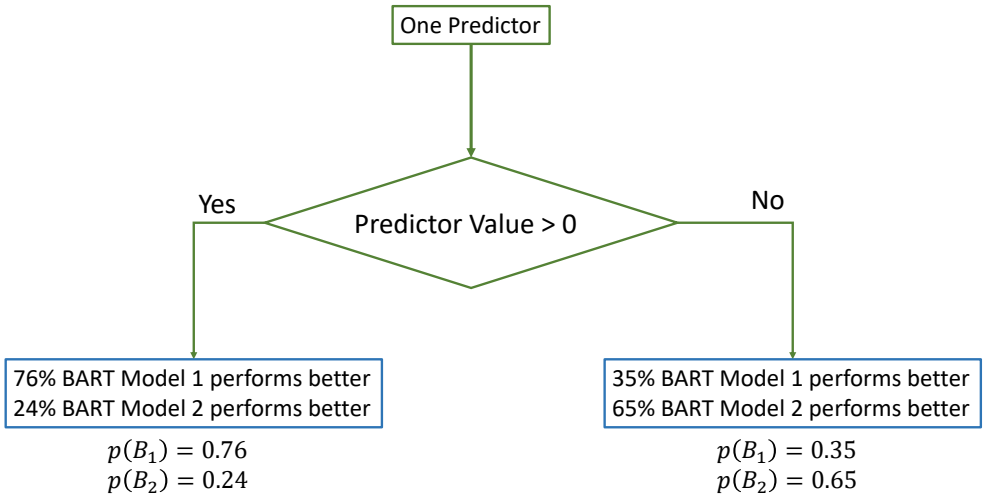


Figure 4.1: Schematic diagrams of an example of nesting two BART models under a simple two-leveled CART model, using only one predictor. The partitioning rule is expressed in the diamond box, and the leaves are represented in blue boxes.

the approach also outputs the variation of $p(B_k)$ under various conditions. This could be an indication of the behavior of a dynamic hydrologic similarity system, and will be explained in details next.

## Hypothesis of hierarchical similarity

The variation of $p(B_k)$ indicates the behavior of a dynamic hydrologic similarity system, but the underlying theory for physical interpretation is still missing. Here, we propose a hypothesis of hierarchical similarity to do so. We hypothesize the hydrologic similarity follows a hierarchy that has two levels:

1. The lower level is termed the **predictor similarity**, meaning that if two sets of predictors are similar in some parts, their corresponding response will be similar. In hydrology context, if two watersheds have some similar characteristics, then their hydrologic responses will be similar. This lower level corresponds to the BART models in the nested tree-based modeling approach.

2. The higher level is the **regionalization similarity**, meaning that if two sets of predictors are similar in some parts, their corresponding predictor-response relationships will be similarly controlled. In hydrology context, if two watersheds have some similar characteristics, then their hydrologic responses will be governed by similar functions/mechanisms. This higher level corresponds to the classification tree in the nested tree-based modeling approach.

Put simply, regionalization similarity determines the predictor-predictor relationship and tells us which predictors to extract information from, while predictor similarity determines the predictor-response relationship that actually estimates the response using the said extracted information. Note that the two sets of predictors respectively determining the two levels of similarity are not mutually exclusive: they may or may not overlap. To elaborate on the difference between the two levels of similarity, we present the following two example statements under the example context of groundwater recharge estimation.

1. **Systematic trends in recharge rates are often associated with climatic trends (Healy, 2010).** This is a statement of predictor similarity, indicating a predictor-response relationship. One would be informed to association recharge rates with climatic variables.

2. **In arid regions, focused recharge from ephemeral streams is often the dominant form of recharge (Healy, 2010).** This is a statement of regionalization similarity, indicating a predictor-predictor relationship. One would be informed to pay more attention to the dominant factors of ephemeral streams, if the study area of interest is in arid regions.

The two levels in the hypothesis correspond to the two levels in the nested tree-based modeling approach. We use the BART models to explore predictor similarity with different predictor sets, and use the classification tree to explore regionalization similarity by investigating the variation of $p(B_k)$ under various conditions. Note that as the condition changes, the best performing BART model may change and so does the set of dominant predictors in the predictor-response relationship. This offers a potential theory of why under different conditions, the hydrologic similarity may be controlled by different watershed characteristics.

## 4.3 Case Study: The Eastern U.S.

In this Section, we will demonstrate how the nested tree-based modeling approach can be used to condition estimates of hydrologic responses on ex-situ data while also quantifying the associated uncertainty. More importantly, we will demonstrate how the approach can inform us about the different controls of hydrologic similarity under different conditions. The case study involves the estimation of mean annual groundwater recharge, a spatiotemporally homogenized hydrologic response, in the Eastern U.S.

To estimate recharge, one should first identify a number of plausible predictors. Although few studies have directly identified the controlling factors, some insights can be learned from previous studies. For example, the effective recharge (i.e., the net source term in the groundwater flow equation) in a steady, depth-integrated, and unbounded groundwater flow was found to be correlated with the spatial distributions of transmissivity and hydraulic head (Rubin and Dagan, 1987a,b). From a recharge-mechanism-based perspective, previous studies have also found a list of plausible controlling factors of recharge via recharge potential

mapping (Naghibi et al., 2015; Rahmati et al., 2016; Yeh et al., 2009, 2016). These variables include watershed topography, land cover, soil properties, and geology. At the regional scale, climate variables have been found to be among the primary controlling factors of groundwater table depth (Fan et al., 2013), mean annual groundwater recharge (Nolan et al., 2007), and mean annual baseflow (Rumsey et al., 2015), the latter of which is often used as a surrogate of recharge under the steady state assumption. Other examples include Xie et al. (2017), who showed that evapotranspiration data provided more conditioning power and more uncertainty reduction than soil moisture data in long-term mean recharge estimation, and Hartmann et al. (2017), who reported variations of the sensitivity of annual groundwater recharge to annual precipitation with aridity. Although these studies did not apply regionalization explicitly and did not target ungauged watersheds directly, their findings provide guidance for us to identify some watershed characteristics—especially climate variables—that might play an important role in the regionalization process for recharge estimation.

We will try to follow the indication from previous studies for the collection of predictors. It is important to note, however, that this case study is not aimed at a thorough investigation of the recharge mechanism, nor is the goal obtaining the most accurate recharge estimates. Rather, the primary goals are the demonstration of the power of our approach, and showing how the approach helps us understand the dynamic behavior of hydrologic similarity in the study area.

## Watersheds and Recharge Estimates

The conterminous United States can be divided into eight major river basins (MRBs, see Figure 4.2), each of which consists of thousands of watersheds (Brakebill and Terziotti, 2011; USGS, 2005). At each and every watershed, watershed-average annual recharge estimate and watershed characteristics data are retrieved from publicly available databases, and will be described in the following subsections. In our work, the recharge estimates are used as the target response while the characteristics are used as predictors in the regionalization process. as explained in Section 4.2.

In 2002, annual groundwater recharge at each watershed was estimated via baseflow analyses by the U.S. Geological Survey (USGS) (Wieczorek and LaMotte, 2010d; Wolock, 2003, also shown in Figure 4.2). Streamflow-based estimation of recharge, such as baseflow analysis, is commonly used in humid regions. As put forward by Healy (2010), there are three key questions that should be carefully checked before applying baseflow analysis: (1) Is all recharging water eventually discharged into the stream where the baseflow is measured? (2) Do low flows consist entirely of groundwater discharge? (3) Does the contributing area of the aquifer differ significantly from that of the watershed? Without a rigorous proof, we make a working assumption about the reliability of baseflow analysis. Fortunately, from a post hoc check, the recharge estimates fall within the typical scales at which baseflow analysis is more suitable: a recharge scale from hundreds to thousands $mm$ per year, a spatial scale of hundreds of $m^2$ to hundreds of $km^2$, and temporal scales from months to decades (Scanlon et al., 2002).

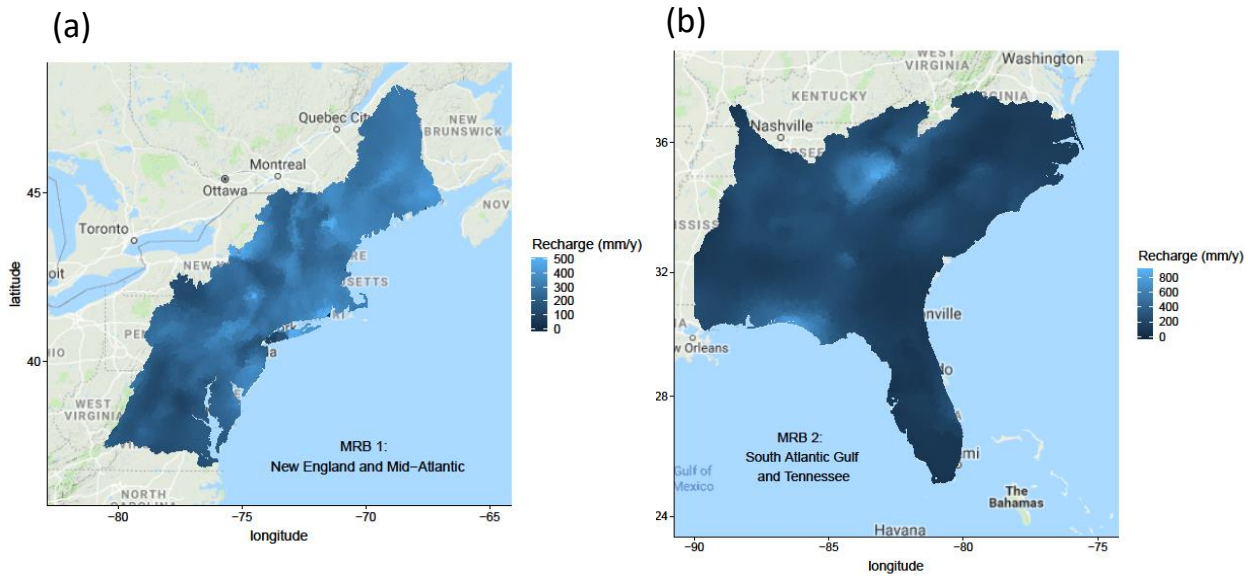


Figure 4.2: The study area includes **(a)** MRB 1 and **(b)** MRB 2 in the eastern U.S., colored by the estimated annual groundwater recharge in the year of 2002 (Wolock, 2003). For the details of the delineation of MRBs please refer to USGS (2005).

The more arid U.S. Midwest may have more pronounced localized recharge (Vries and Simmers, 2002), which cannot be effectively captured by baseflow analysis (Scanlon et al., 2002). This, then, does not fit well with our working assumption. Therefore, following the suggestion of Nolan et al. (2007), our study area includes only the relatively humid eastern parts of the U.S., namely MRB 1 and 2 (Figure 4.2). After excluding watersheds with less desirable data coverage, we consider a total of 3609 watersheds in MRB 1 and 7413 watersheds in MRB 2. The distributions of the recharge data from all the watersheds in the study area are shown in Figure 4.3**(a)**.

## Climate

At each watershed included in the study, the following data are retrieved from publicly available databases: the long-term average annual precipitation ($\bar{P}$) averaged from 1970 to 2000 (Wieczorek and LaMotte, 2010a), the annual precipitation in the year 2002 ($P$) (Wieczorek and LaMotte, 2010h), and the long-term average annual potential evapotranspiration ($E_p$) averaged from 1960 to 1990 (Title and Bemmels, 2017). Note that limited by data availability, the average periods of $\bar{P}$ and $E_p$ are different. Thus, we also make a working assumption that at the decadal scale the averaged climate variables remain steady, with

which we ignore the potential effect of climate change on the difference between the average from 1960 to 1990 and that from 1970 to 2000. Given the precipitation and evapotranspiration, we obtained two additional climate variables: the long-term aridity index, estimated as $\bar{\phi} = E_p/\bar{P}$, and the 2002 aridity index, estimated as $\phi = E_p/P$. Given that the recharge data are based on baseflow analysis for the year 2002, $P$ and $\phi$ represent the climate controls of that same year, while $\bar{P}$, $E_p$, and $\bar{\phi}$ represent climate controls over the long-term. The distributions of $P$, $\bar{P}$, and $E_p$ are shown in Figures 4.3**(b)**, **(c)**, and **(d)**, respectively.

### Normalization and Transformation of Recharge

The annual recharge data (in volume of water per unit watershed area) can be normalized by $P$ (also in volume of water per unit watershed area), as in Figure 4.3**(e)**. This stems from the concept of water budgets and has been commonly used in hydrological studies worldwide (e.g., Heppner et al., 2007; Magruder et al., 2009; Obuobie et al., 2012; Rangarajan and Athavale, 2000; Takagi, 2013; Yang et al., 2009). Here, we apply logit transformation, which is common for proportions or probabilities (Gelman et al., 2014), to that normalized recharge, relaxing the physical bounds (0 and 1) of the values of the target variable (Figure 4.3**(f)**). This step is advantageous as it opens the opportunity to estimate recharge with classical parametric statistical approaches without special accommodations for the bounds. Therefore, in this case study the logit normalized recharge (LNR) is used as the target response.

## Non-climate Predictors

We also consider various non-climate watershed characteristics in this study, including topography, land cover, soil properties, and geology. The land cover is based on data published in 2001, which we feel is close enough to 2002 to provide the appropriate information. The other characteristics are based on raw data obtained in different years before 2002; it is assumed that they remain steady at sub-century time scales.

The topographic predictors are taken from publicly available databases (Wieczorek and LaMotte, 2010b); they are summarized in Table 4.1. The land cover variables are the percentages of watershed area corresponding to each land cover class (Wieczorek and LaMotte, 2010e); these are summarized in Table 4.2. The land cover classes are based on the 2001 National Land Cover Database (NLCD2001), the categories of which include water, developed land, barren land, forest, shrubland, herbaceous land, cultivated land, and wetland, with each having its own sub-classes. The details of NLCD2001 can be found in Homer et al. (2007).

The soil property predictors include watershed scale statistics (e.g., average, upper bound, and lower bound) of soil properties (Wieczorek and LaMotte, 2010f); these are summarized in Table 4.3. The spatial statistics of the soil properties within each watershed were obtained over gridded source data values from the State Soil Geographic database (STATSGO)

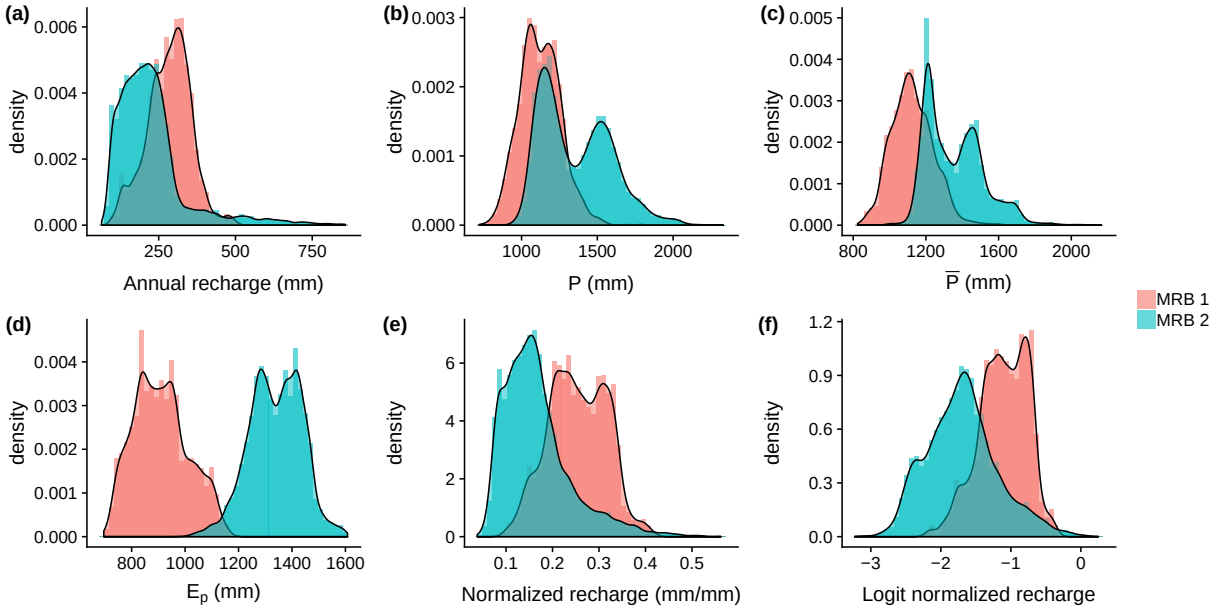


Figure 4.3: Histograms of **(a)** annual recharge in 2002, **(b)** annual precipitation in 2002, **(c)** long term average annual precipitation, **(d)** long term average annual potential evapotranspiration, **(e)** normalized recharge, and **(f)** logit normalized recharge (LNR) at all the watersheds in MRB 1 and 2. The black curves are estimates of the distributions based on kernel density estimation.

(Schwarz and Alexander, 1995), which were depth-averaged overall soil layers (Wolock, 1997).

The geology predictors used in this study are also retrieved from publicly available databases (Wieczorek and LaMotte, 2010c,g) and they can be classified into two subcategories: surficial geology (surface sediment) and bedrock geology. As the predictors, we used fractions of the watershed area corresponding to each of the 45 surficial geology types (Clawges and Price, 1999) and each of the 162 bedrock geology types (Schruben et al., 1994). Details regarding each geology type can be found in Wieczorek and LaMotte (2010c,g). Note that in geological terminology, rock type or rock composition data are referred to as lithology data. Compared to lithology, structural geology data might be more informative for groundwater studies (e.g., orientation, fracture properties, discontinuity, etc.). However, structural geology information usually requires in-situ investigation, which cannot be expected at unsampled watersheds. Therefore, we consider only lithology data in this study.

## Data partitioning

Because we cannot evaluate the predictive accuracy at real unsampled watersheds due to the lack of in-situ recharge observations, we follow the holdout method to partition the

Table 4.1: Watershed topography predictors.

| **Variable** | **Explanation** | **Unit** |
|---|---|---|
| Basin index | Watershed area divided by squared watershed perimeter | dimensionless |
| Stream density | Reach length divided by watershed area | $m^{-1}$ |
| Sinuosity | Reach length divided by the distance between the beginning and the ending of the reach | dimensionless |
| Slope | Mean watershed slope calculated from digital elevation data | degree |

watersheds into two mutually exclusive sets: the watersheds in MRB 1 are the **testing** watersheds, and the watersheds in MRB 2 are the **training** watersheds. The testing watersheds will be treated as if they were unsampled, and we only condition the plausible BART models on data from the training watersheds (which are the ex-situ data with respect to the testing watersheds).

There are two reasons for this MRB-based data partitioning:

1. For reasons touched on in Section 1.2, we do not consider spatial proximity as a predictor in this study. Separating the two MRBs partly ensures the exclusion of the confounding effect of spatial proximity, and thus the regionalization is solely based on the watershed characteristics.

2. Considering the distributions of LNR (Figure 4.3**(f)**), the range of values in MRB 2 fully covers the range of values in MRB 1. However, the reverse is not true. It is thus advantageous to train the models with MRB 2 to avoid poor model fitting due to lack of data coverage.

After partitioning the watersheds, we now turn our attention to the partitioning of predictors. As mentioned in Section 4.1, climate variables are among the most important factors in hydrology at the regional scale, but there might be other controlling factors to consider as well, and the dominance of climate variables may not be always present. To investigate the various effects of different predictors, we conceptually divide the predictors into four sets: (1) climate controls that determine the input amount of water into the system, (2) surface controls that determine the distribution of water at the surface, (3) soil controls that determine the infiltration of water, and (4) lithology controls that indicates the properties of the aquifer. We further break of the first set into three subsets to investigate the effect

Table 4.2: Land cover classification by NLCD2001.

| **Class** | **Subclass** |
|---|---|
| Water | Open water<br>Perennial ice |
| Developed | Open space<br>Low intensity<br>Medium intensity<br>High intensity |
| Barren | Barren land |
| Forest | Deciduous<br>Evergreen<br>Mixed |
| Shrubland | Dwarf shrub<br>Shrub/scrub |
| Herbaceous | Grassland<br>Sedge<br>Lichens<br>Moss |
| Cultivated | Pasture/hay<br>Crops |
| Wetlands | Woody wetland<br>Emergent herbaceous wetland |

of dimensionless predictors. Therefore, we define a total of six different predictor sets to build six unique BART models, which are indexed by $k$, $k = 1, 2, ...6$ (Table 4.4). Note that the determination of the six predictor sets is guided by a conceptual division of predictors and the idea of testing the relative importance of different categories of predictors under different conditions, instead of aiming for high accuracy and precision. Therefore, by no means is Table 4.4 an exhaustive list of all possible sets, nor does it necessarily include the "best" set that leads to the "best" predictive performance. The design of the six predictor sets simply facilitates the investigation of the effects of various categories of predictors on predictive accuracy and uncertainty.

In addition to the six BART models, we also build a simple model by using the estimated distribution of LNR at the training watersheds via kernel density estimation (R Core Team, 2018; Sheather and Jones, 1991), without considering any predictor. In other words, this is simply using the distribution of LNR at all the training watersheds as the predictive distribution. This is a model that ignores hydrologic similarity altogether, and it can be

Table 4.3: Soil property predictors.

| Soil property | Unit | Statistics* |
|---|---|---|
| Calcium carbonate equivalent | % | Lower/higher bounds |
| Cation exchange capacity | cmolc $kg^{-1}$ | Lower/higher bounds |
| Depth to the seasonally high water table | m | Average and Lower/higher bounds |
| Soil thickness | m | Lower/higher bounds |
| Hydrologic soil group classification | % | Average |
| Soil erodibility factor | dimensionless | Average |
| Permeability | m $h^{-1}$ | Average and Lower/higher bounds |
| Available water content | fraction | Average and Lower/higher bounds |
| Bulk density | g $cm^{-3}$ | Average and Lower/higher bounds |
| Organic matter content | % | Average and Lower/higher bounds |
| Clay soil content | % | Average and Lower/higher bounds |
| Silt soil content | % | Average |
| Sand soil content | % | Average |
| Percent finer Than nos.4, 10, and 200 sieve | % | Average and Lower/higher bounds |

* Spatial statistics calculated across the watershed.

Table 4.4: Table of the six different predictor sets.

| k | predictors included | Number of predictors |
|---|---|---|
| 1 | $\bar{\phi}$ and $\phi$ | 2 |
| 2 | $\bar{P}$ , $P$, and $E_p$ | 3 |
| 3 | All climate predictors: $\bar{P}$ , $P$, $E_p$, $\bar{\phi}$ and $\phi$ | 5 |
| 4 | Topography and land cover predictors | 20 |
| 5 | Soil predictors | 48 |
| 6 | Geology predictors | 206 |

considered as an extreme case of the ex-situ prior in Li et al. (2018), with a lot more watersheds and much less stringent criteria of similarity. From this point forward, we refer to this model as the benchmark model, for it is used as a benchmark against which the BART models are compared.

## Evaluation of predictive distributions

To apply the nested tree-based modeling approach, we should label each testing watershed by the best performing model. Thus, the metric with which we evaluate predictive distributions matters. In this study, two different accuracy metrics are adopted. The first is

the root mean squared error (RMSE), defined as

$$E_{i,k} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\hat{R}_{i,k}^{(l)} - \tilde{r}_i\right)^2} \tag{4.1}$$

where $\tilde{r}_i$ is the recharge data at the $i^{th}$ testing watershed, and $E_{i,k}$ is the RMSE of the $k^{th}$ model at the $i^{th}$ testing watershed. Note that $\hat{R}_{i,k}^{(l)}$ is obtained using BART (Equation 2.17), but we change the notation to specifically represent the target response, LNR, and now subscripts are added to indicate that we plug in the predictors from the $i^{th}$ testing watershed to the $k^{th}$ model. This metric evaluates the predictive performance in an estimation problem, where we wish to obtain a "best estimate" of recharge with minimal expected error.

The second metric is the median log predictive probability density (LPD) at the value of recharge observation, defined as

$$L_{i,k} = \text{median}_{l=1,\ldots,L}\left\{\ln\left[p\left(R = \tilde{r}_i|\hat{R}_{i,k}^{(l)}, (\sigma^2)_k^{(l)}\right)\right]\right\}, \tag{4.2}$$

where $L_{i,k}$ is the LPD of the $k$th model at the $i$th testing watershed. The subscript of $(\sigma^2)_k^{(l)}$ indicates the $k$th model. This metric evaluates the predictive performance in a simulation problem, where we wish the realizations from the predictive distributions are likely to be the same as the observation.

In addition to accuracy, we also quantify the predictive uncertainty. This is done by first recognizing the two components of uncertainty for the $k$th model at the $i$th testing watershed:

1. $\sigma_k^2$, which we refer to as the **predictive variance**, and is approximated as the sample median of $(\sigma^2)_k^{(l)}$ over $l = 1, ..., L$, and

2. the posterior variance of $\hat{R}_{i,k}$, which we refer to as the **estimate variance**, and is approximated as the sample variance of $\hat{R}_{i,k}^{(l)}$ over $l = 1, ..., L$.

The predictive variance indicates how informative the inferred predictor-response relationship is, while the estimate variance indicates how uncertain the said relationship is. In this case study we weigh the two components equally, as we wish to obtain an informative relationship with certainty. To that end, we define the **total predictive variance** as the summation of the two components, and use it as the metric of predictive uncertainty in this study.

## 4.4 Case Study Results: Predictor Similarity

As discussed above, we built six BART models (Table 4.4) with ex-situ data. In-situ predictors were then fed into the models to yield posterior realizations of predictive distributions (Eq. 2.18). With the metrics of accuracy and uncertainty defined, we are then able

to quantify the predictive performance of the BART models, and classify them based on either the RMSE-based labels or the LPD-based labels with the nested tree-based modeling approach. This allows for the investigation of the effects of various predictors under different conditions, which will be presented in this Section.

## Predictive Uncertainty

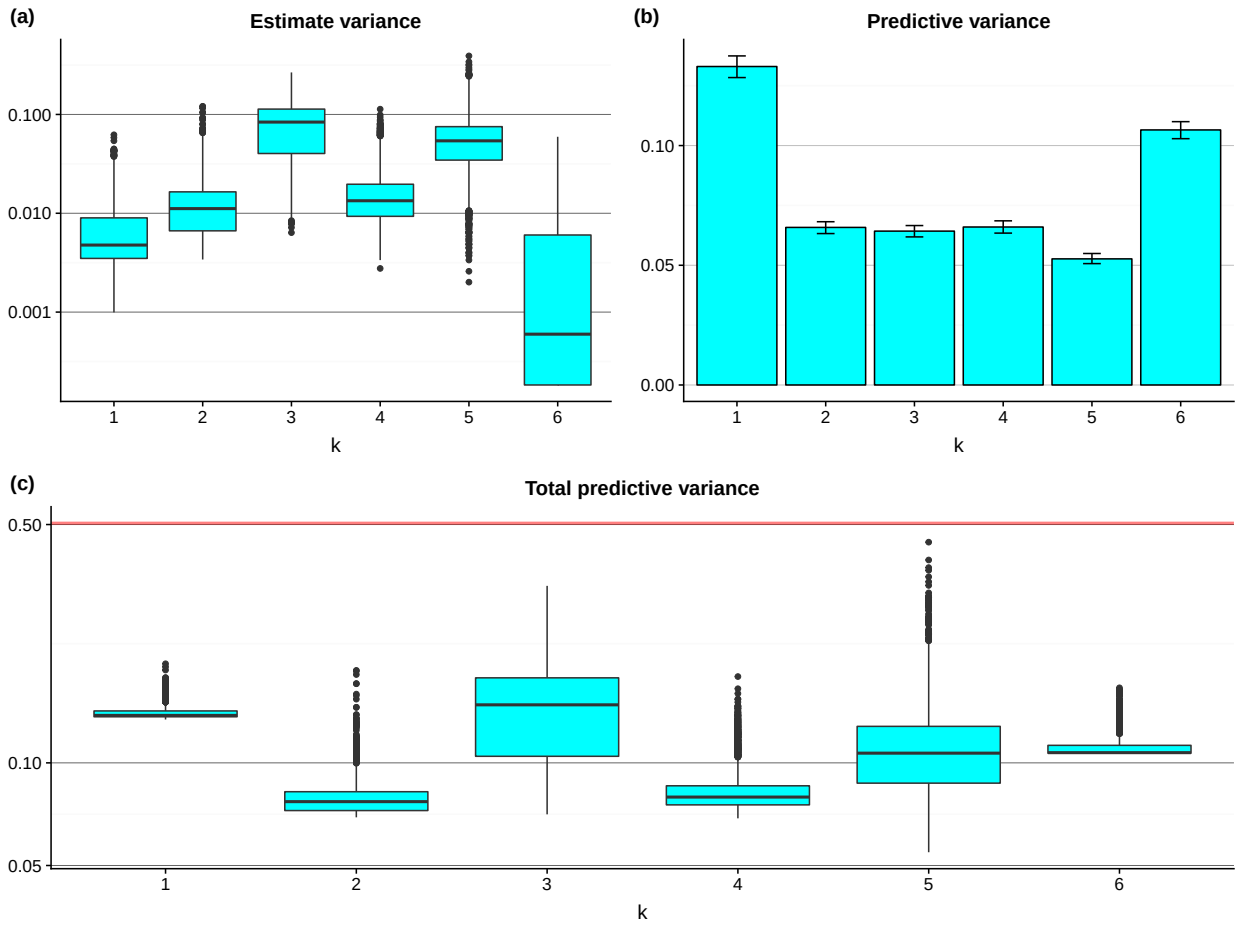


Figure 4.4: **(a)** The box plots of the estimate variances at the testing watersheds, **(b)** the bar plot of the predictive variances with 95% intervals shown by the error bars , and **(c)** the box plots of the total predictive variances at the testing watersheds. The red line indicates the variance of the benchmark model, used for comparison.

The effect of regionalization with the different predictor sets on predictive uncertainty is shown in Figure 4.4. The estimate variance (Figure 4.4 **(a)**) represents how well the BART models capture the predictor-response relationships. We see that the geology predictors

lead to the lowest estimate variance, probably because of the significantly larger number of predictors used (see Table 4.4).

Yet, there is a surprise in Figure 4.4**(a)**. First, at $k = 1$ and $k = 2$ the estimate variances are generally quite low, despite the low number of predictors. However, at $k = 3$, the estimate variances increase significantly. Intuitively, since aridity is the ratio of evapotranspiration to precipitation, one would expect that the variances at $k = 3$ would be similar to, if not lower than, those at $k = 1$ and $k = 2$. One plausible explanation here is that although aridity indices and precipitation/evapotranspiration carry ample information to be extracted and conditioned upon, the respective predictor-response relationships we get might be significantly different. When used together, the BART models were not able to formulate a universal relationship. This will be revisited later in Section 4.6.

The predictive variance (Figure 4.4**(b)**) represents how informative the predictor-response relationships are, which is a different aspect of uncertainty compared to the estimate variance. One could obtain a predictor-response relationship fairly confidently (low estimate variance), but the relationship is less informative (high predictive variance), like that found at $k = 6$. The opposite case is that one could not confidently obtain a predictor-response relationship, but once that relationship is obtained it is quite informative, like that found at $k = 5$.

The total predictive variance (Figure 4.4**(c)**) provides an overall metric that considers the above two sources of uncertainties. While the medians are rather similar, the spread of the box plots does vary significantly with $k$. The condensed box plots (e.g., $k = 1$ and $k = 6$) indicate that the total predictive variances are essentially constant throughout all testing watersheds, while the spread-out box plots (e.g., $k = 5$) indicate that the effect of the predictors may vary significantly from one testing watershed to another. This indicates that there might not be one single predictor set that always leads to the lowest uncertainty, and thus the effects of predictors on predictive uncertainty may vary from one condition to another. That said, regardless of the testing watersheds and predictor sets, the total predictive variance is always lower than the variance of the benchmark model, which clearly shows that regionalization using watershed characteristics definitely improves predictive precision.

### Predictive Accuracy

The effect of regionalization with the different predictor sets on RMSE is shown in Figure 4.5. The RMSE of the benchmark model (Figure 4.5 **(a)**) at each testing watershed is simply the difference between the sample mean of the ex-situ recharge data and the in-situ recharge observation. For the BART models (Figure 4.5 **(b)**), it is calculated by the root of the average squared errors over post-convergence MCMC simulations. Regardless of $k$, we see that, compared with the benchmark model, RMSE is reduced at least at half of the testing watersheds. Surprisingly, the largest overall RMSE reduction is observed when only the aridity indices are used for regionalization, indicating that at most of the watersheds tested in this study, aridity similarity implies LNR similarity at regional and annual scales to a high degree. On the other hand, we observe some outliers that have high RMSE reduction at $k = 4$ through $k = 6$, indicating that topography, land cover, soil properties, and geology

may not have an overall effect that is as strong, but under certain circumstances, they could still be important factors.

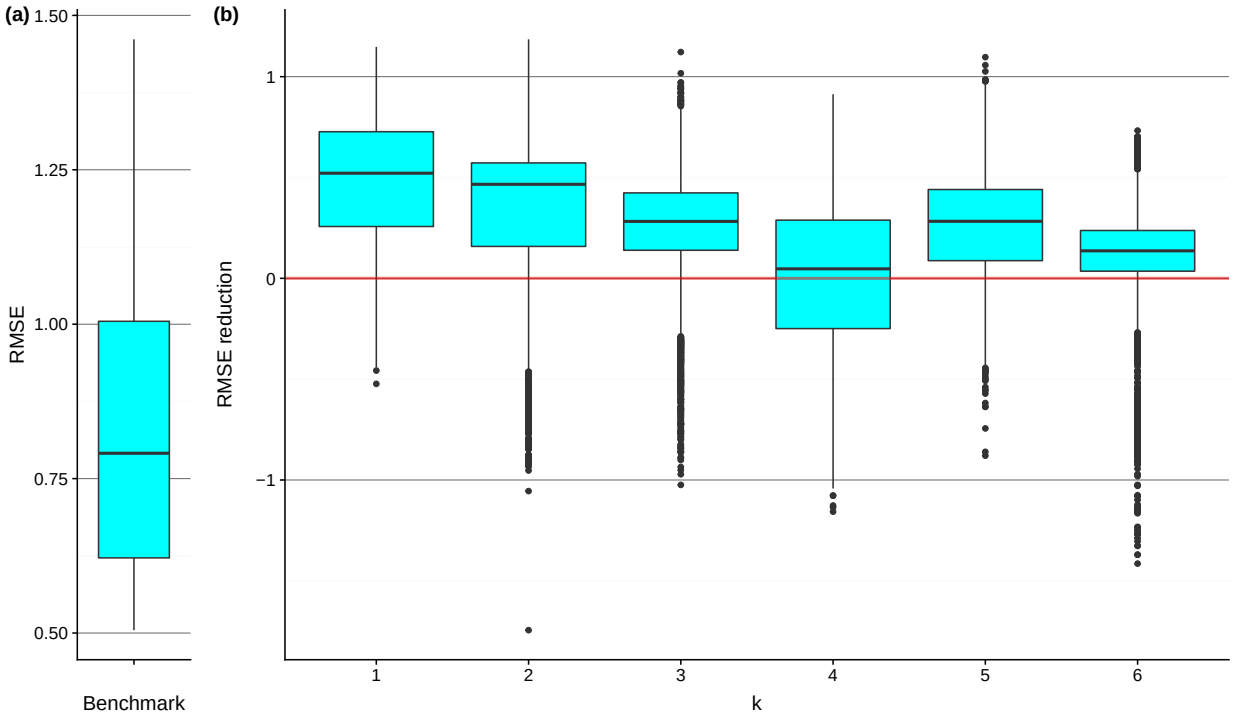


Figure 4.5: **(a)** The box plot of the RMSE of the benchmark model at the testing watersheds, and **(b)** the box plots of the RMSE reduction introduced by applying the BART models at the testing watersheds. The red line indicates zero RMSE reduction, used for comparison.

The effect of regionalization with different predictor sets on LPD is shown in Figure 4.6. It is immediately clear that the accuracy improvement is not as prominent as that seen in Figure 4.5. Only when $k = 1$ is LPD increased at most of the watersheds . We also find that all of the distributions of LPD are heavily negatively skewed with a lot of outliers.

Looking at Figures 4.4 through 4.6 together, one can observe the different effects of the predictor sets on predictive accuracy, stemming from the different natures of an estimation and a simulation problem. From the point of view of the overall effect, for $k = 2$ through $k = 5$ (i.e., the predictors other than aridity indices) RMSE is reduced at more than half of the testing watersheds, but LPD does not increase to the same extent. This suggests that the predictive distributions are centered closer to the in-situ observations due to regionalization, but that the conditioning also significantly reduces the predictive variances, causing the predictive distribution to be too narrow. Therefore, compared to a relatively flat, spread-out, and uninformative or weakly informative distribution, the predictive density decays too quickly when deviating from the predictive mean, resulting in low LPD. This might be a sign of over-conditioning, or the disproportional reduction of predictive uncertainty, as

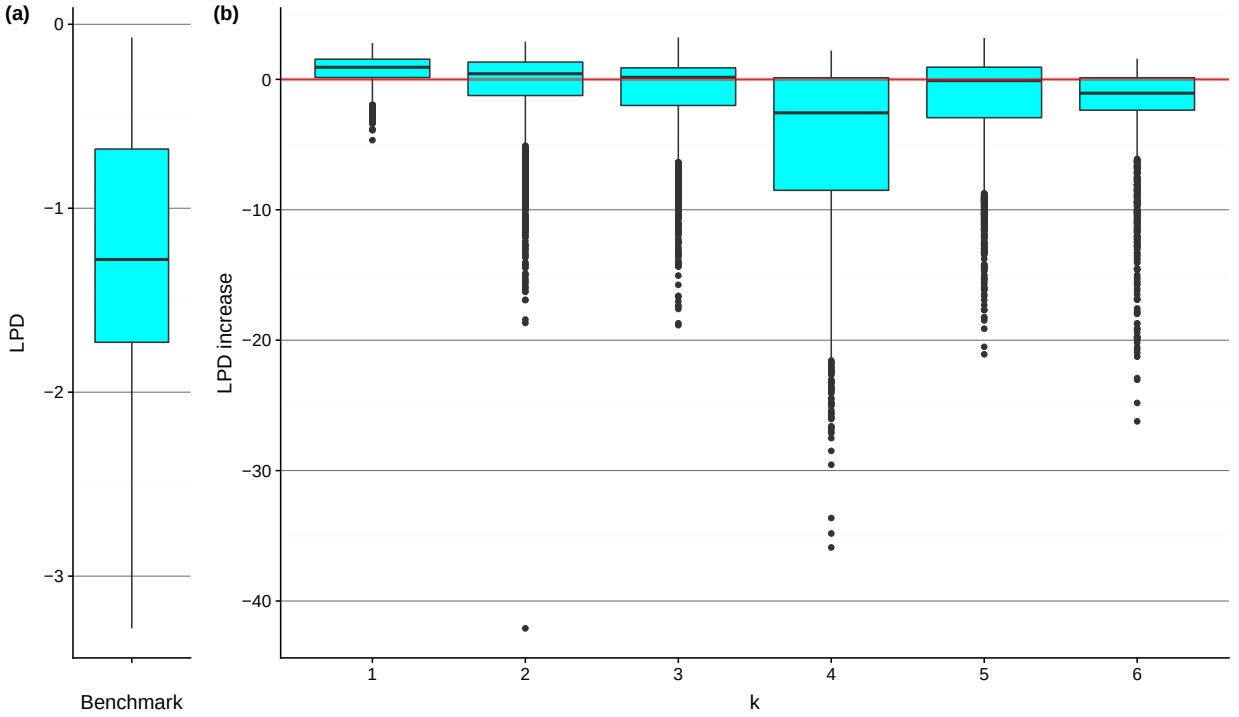


Figure 4.6: **(a)** The box plot of the LPD of the benchmark model at the testing watersheds, and **(b)** the box plots of the LPD increase introduced by applying the BART models at the testing watersheds. The red line indicates zero LPD increase, used for comparison.

exemplified in Figure 4.7. The cyan curve is an example of an over-conditioned distribution. Although its mean is somewhat close to the true value, the small variance causes rapid decay of probability density; therefore, at the true value (red vertical line) the predictive density is no better than that of the weakly informative or uninformative distributions. How could this ever happen? Take $k = 5$ in Figure 4.4 as an example: the predictive variance is small, meaning that the predictive distribution should be rather peaked (just like the cyan curve in Figure 4.7). The only way one can get a high predictive density is then to make the predictive mean close to the true value. Nonetheless, this would be very difficult at some of the watersheds where the estimate variance is large. The only predictor set that improves both RMSE and LPD at most of the testing watersheds is $k = 1$, the aridity indices, and one could expect the corresponding predictive distributions to be somewhat similar to the case of the ideal dark blue curve in Figure 4.7.

Over-conditioning can occur when model fitting or model calibration leads to well-constrained parameters that are, in fact, subject to different forms of model uncertainty (Beven et al., 2008; Hutton et al., 2014), which is an indication of why the determination of $p(B_k)$ is important. In this case study, we focused more on the variation of $p(B_k)$ under various conditions (to be shown shortly), and less on improving the estimates. However, in

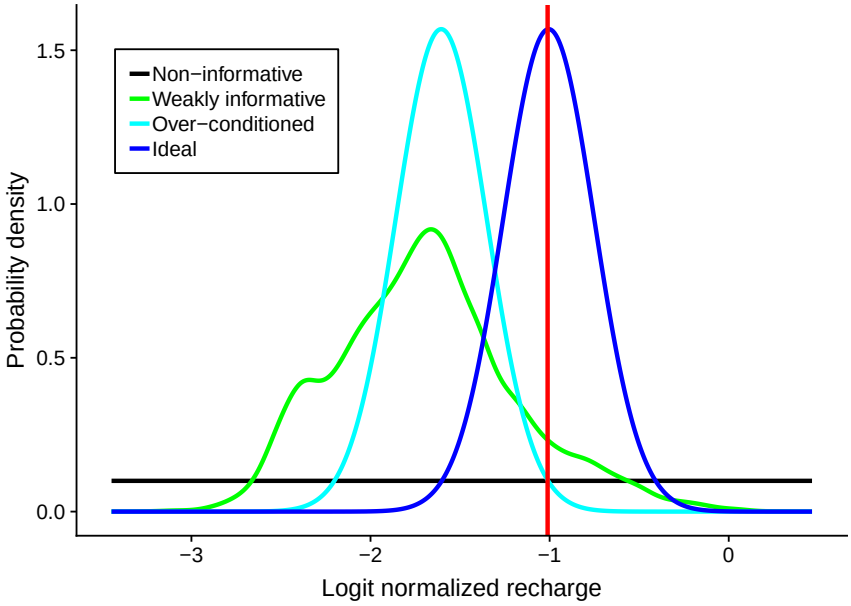


Figure 4.7: An example of over-conditioning: the probability density at the true value (indicated by the red vertical line) of the over-conditioned distribution is not higher than that of the non-informative distribution or that of the weakly informative distribution, not because the conditioning does not work, but because of the disproportional reduction of the variance of the distribution.

another application where the estimates are to be improved, model structure uncertainty should be and can be considered in order to refine the estimates (e.g., via Bayesian model averaging).

## 4.5 Case Study Results: Regionalization Similarity

The box plots in Figures 4.4 through 4.6 show different distributions of the predictive performance metrics for the different predictor sets. An interesting follow-up question here would be how model performance varies with watershed characteristics. It was shown that, consistent with previous studies, aridity is indeed the most important controlling factor at regional and annual scales on average, but there are few cases where this aridity dominance is replaced. In other words, how might we identify the conditions under which a specific predictor set could be more informative than others?

To investigate this further, we give each testing watershed two labels: the model with the lowest RMSE, and the model with the highest LPD; we refer to these labels as the RMSE labels and the LPD labels, respectively. The possible values of each label include $k = 1$ through $k = 6$ and *benchmark*, representing the six BART models and the benchmark

model, respectively. Then, using all the available predictors, we built two CART models to classify watersheds based on the RMSE labels (Figure 4.8), and the LPD labels (Figure 4.9).

### Nesting by RMSE

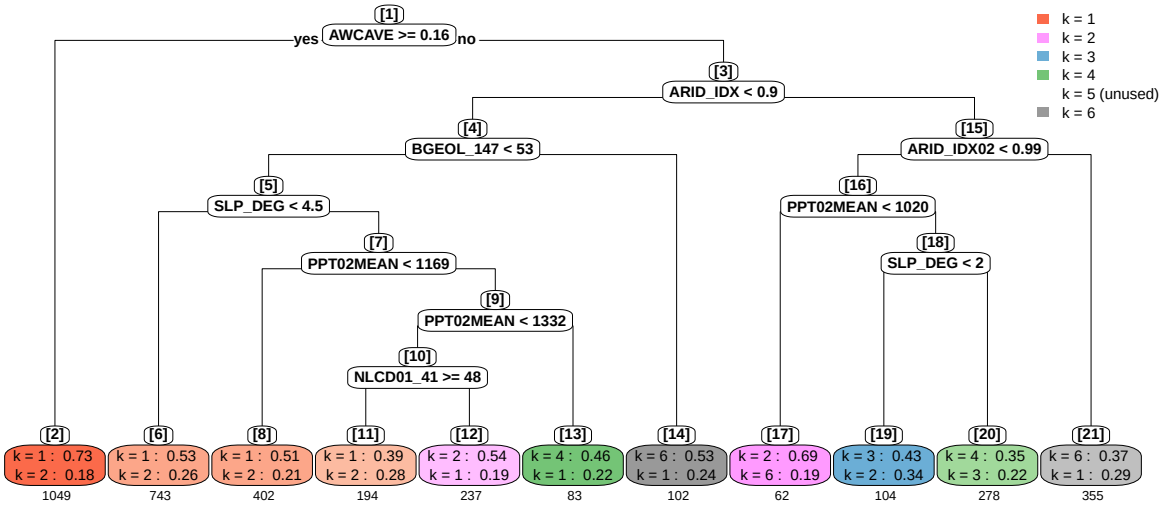


Figure 4.8: CART model classifying the RMSE labels of the testing watersheds. Splitting rules are shown in white nodes, while leaf nodes are colored based on the classification results. For each leaf node, the brightness of the coded color indicates the node impurity (the brighter the more impure), where impurity is defined as the probability that two randomly chosen watersheds within the node have different labels. On top of every node, in brackets, is the node number, provided for convenient referencing. The predictors in the splitting rules are expressed in code names for convenience; a reference list is found in Table 4.5. For each leaf node, the model of the highest multinomial probability of having the best performance is shown first, which also determines the classification result, followed by the model of the second highest probability, also to indicate the impurity. Underneath each leaf node box is the number of watersheds belonging to the leaf. Note that the legend does not include benchmark because the benchmark model is never the best-performing model at any testing watershed. $k = 5$ is marked as "unused" in the legend because there is no leaf node where $p(B_5)$ is the highest.

Figure 4.8 shows the variation of the top two best performing BART models and the corresponding $p(B_k)$ values under various conditions, where the performance of each BART model is defined by the RMSE. This variation indicates the regionalization similarity in the study area. At first glance, the available water content (AWC) stand out to be the first indicator of regionalization similarity (Figure 4.8 node 1): at watersheds with high AWC, aridity stands out as the dominant factor, which is consistent with the previous studies

Table 4.5: Reference list of the splitting variables in Figures 4.8 and 4.9.

| Code name | Description |
|---|---|
| AWCAVE | Average available water content |
| ARID_IDX | Long term average aridity index |
| ARID_IDX02 | Aridity index in 2002 |
| PPT02MEAN | Annual precipitation in 2002 |
| SLP_DEG | Average slope in degree |
| NLCD01_41 | % area of Deciduous forest |
| BGEOL_147 | % area of Paragneiss and Schist bedrock |

cited in Sect. 4.1. However, there is a potential risk if one uses aridity as the primary indicator of hydrologic similarity regardless of AWC. In previous studies, AWC was found to be an important predictor correlated with surface runoff, baseflow, and groundwater recharge (Arnold et al., 2000), and it was among the most important parameters to which water balance models are sensitive Finch1998. In the current study, we are not claiming that AWC cannot be a predictor, but rather, we are suggesting a hierarchical structure in which AWC is placed —together with other predictors —to help estimate LNR at ungauged watersheds. Since AWC is governed by field capacity and wilting point, it is an indicator of the storage capacity of the soil for usable/consumable water: the larger the storage capacity, the higher the degree to which the system is supply-limited, thus pointing to aridity. If the storage capacity is low, on the other hand, the more complicated interplay among various predictors needs to be considered, and one cannot simply assume that aridity is the primary indicator of hydrologic similarity. We also found the soil organic matter content a quite competitive surrogate for AWC, meaning that if organic matter content was used here instead of AWC, we would end up with a slightly less accurate but overall similar classification. We conjecture that this is because of the high positive correlations between organic matter content and AWC (Hudson, 1994).

Further down the classification tree, watersheds with lower AWC are classified roughly as arid or humid watersheds by the long-term aridity index. For the more humid watersheds (Figure 4.8, nodes 4 through 14), regionalization similarity is controlled by different predictors, but the dominant predictors for LNR estimation are almost always the climate variables (nodes 6, 8, 11 and 12, which contain 1576 watersheds in total). Only at a handful of watersheds (nodes 13 and 14, which contain only 185 watersheds in total) are aridity indices not dominant. However, some interesting conjectures can be made by taking a closer look at these two nodes.

Node 14 is a small but unique cluster, featuring watersheds that have low AWC, are humid, and have relatively homogeneous paragneiss and/or schist bedrock. Both of these bedrock types belong to the category of crystalline rock, and often feature layering in a particular orientation. The groundwater movement in such rock formation often depends on foliation, i.e., rock breaks along approximately parallel surfaces, which affect the direction

of the regional groundwater flow (Singhal and Gupta, 2010). Hence we observe a condition where the ample water supply cannot be substantially held by the soil due to low AWC, and the regional groundwater movement might be controlled by bedrock layering and foliation. Low AWC is an indication of less clayey soils, and implies that infiltration/percolation through the soil layer might be facilitated by relatively higher permeability. Water could thus easily enter the bedrock layer, which is rather horizontally homogeneous. To that end, those predictor sets other than $k = 6$ become less informative, while the predictor set $k = 6$ becomes relatively more informative. In fact, these watersheds are mostly the positive outliers at $k = 6$ in Figure 4.5**(b)**, where the predictive power of the geology predictors is at its best.

Node 13 features watersheds that have low AWC, are humid, are not dominated by homogeneous paragneiss and/or schist, have a relatively steep average slope, and have a large amount of annual precipitation. The low aridity is primarily driven by precipitation rather than evapotranspiration. In fact, these watersheds are mostly outliers featuring extremely low aridity index (below 0.65) due to ample precipitation. Under such condition, evapotranspiration is expected to operate to its full potential, i.e., it is shifting from water-limited state to energy-limited and canopy-controlled state. In addition, as evapotranspiration is near its full potential, the drainage of the excess precipitation would be controlled by the topography of the watershed (e.g., the slope and the sinuosity of the stream). Fast drainage leaves less water available for infiltration and recharge, and vice versa. To that end, the land cover type and topography now start to play a dominant role in hydrologic similarity. It is noteworthy to point out node 20 here. Node 20 features watersheds that are relatively humid among the arid watersheds ($\bar{\phi}$ in the range from 0.9 to 0.99) and have ample precipitation. The similarity of node 20 with node 13 supports our conjecture that the dominance of land cover and topography predictors is due to the precipitation-driven humid environment that is relatively more capable of catering to the evapotranspiration water demand and features excess precipitation.

On the other side of the tree (Figure 4.8, node 15 through 21), the resulting classification is quite diverse, and the impurity of each node is relatively high. Aridity no longer plays the dominant role, and the hierarchical similarity structure becomes complicated that it is difficult to make straightforward physical interpretations. The most important message we get is the significant risk one would face if one considers aridity, or any climate variable in general, as the primary indicator of hydrologic similarity when AWC is low and aridity index is high. In summary, although climate predictors are still the most important ones on average, within the context of the hierarchical similarity we have identified certain conditions under which either non-climate predictors become dominant or no dominant predictor set can be straightforwardly identified, all of which contribute to the understanding of the dynamic hydrologic similarity.

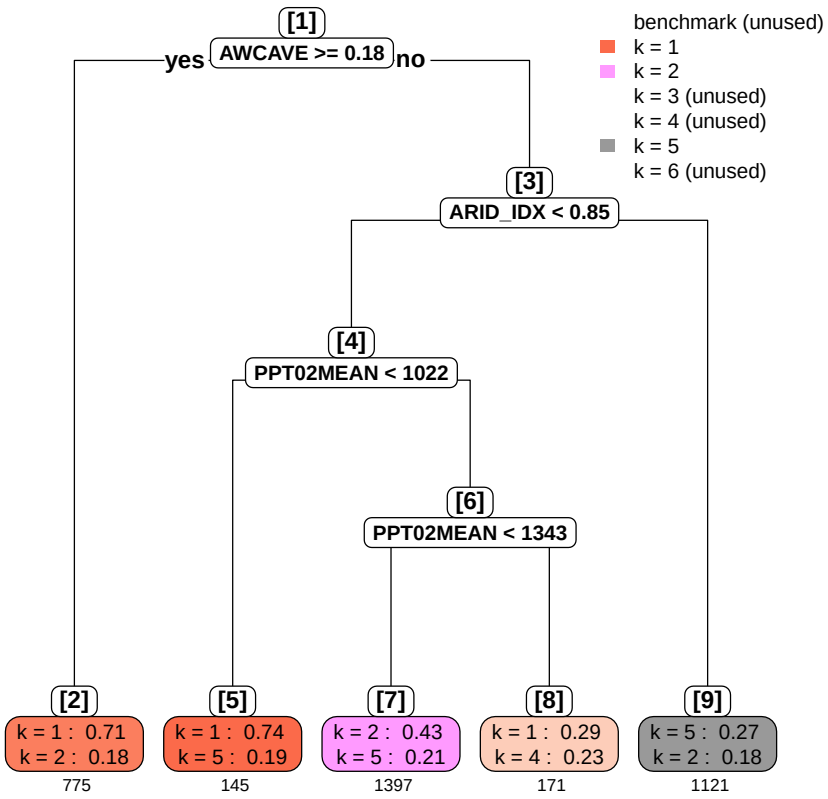


Figure 4.9: Same as Figure 4.8, except here the classification is done using LPD labels. The predictors in the splitting rules are expressed in code names for convenience; a reference list is found in Table 4.5.

### Nesting by LPD

The classification of the LPD labels is shown in Figure 4.9. In general, the root part of the classification tree (node 1 through 3) is quite similar to that found in Figure 4.8, where AWC and long-term aridity define two sequential overarching separations of watersheds. However, further down the tree the leaf part is significantly different. The classification essentially leads to only three big clusters (Figure 4.9, nodes 2, 7, and 9), and the other leaf nodes only contain a few watersheds. Node 9 features arid watersheds with low AWC, where we end up with a highly impure leaf node, and even the highest multinomial probability is only 0.27. No further splitting rule could significantly reduce classification error. This is supportive towards our previous argument that when aridity index is high and AWC is low, it is risky to resort to climate variables for hydrologic similarity, as shown here that it is difficult to even identify a dominant predictor set. As mentioned earlier, underestimation of the predictive variance ($\sigma_k^2$) leads to low LPD, and thus it is difficult to make physical interpretation our

of the results in Figure 4.9, except for node 1 through 3, which are quite similar to their counterparts in Figure 4.8. Therefore, with the LPD labels we are only able to identify the overarching regionalization similarity controlled by AWC and long-term aridity.

RMSE and LPD represent views of predictive accuracy in an estimation problem and a simulation problem, respectively. Intuitively, if one only considers unimodal predictive distribution with limited skewness, a high predictive density at a value directly implies a closeness of the distribution central tendency to that value. However, the reverse is not necessarily true: either over- or underestimation of variance might possibly lead to low predictive density, even if the mean is close to the target value (e.g., Figure 4.7). Based on whether RMSE or LPD is used as the accuracy metric —which implies the scope of LNR estimation —we can observe some common features as well as some distinctions of the structure of the hypothesized hierarchical similarity.

Fortunately, regardless of the metric of predictive accuracy, in both Figures 4.8 and 4.9 the first three nodes are remarkably consistent, and the effect of the metric of predictive accuracy is only manifested at watersheds with low AWC. This supports the suggestion that AWC plays a pivotal role in hydrologic similarity for mean annual LNR estimation.

## 4.6 Discussion

In this Section, we discuss the key features of the approach, the key findings from the case study, as well as the limitations of the case study.

### The Nested Tree-based Modeling Approach

The nested tree-based modeling approach proposed in this Chapter is essentially a coupling of BART and CART. ABoth BART and CART are independent of the physical background, and are pure data-driven machine learning techniques. Therefore, in principle as long as there are data, the nested tree-based modeling approach is applicable like any other data-driven approach. However, one may argue that (1) the in-principle applicability does not set the nested tree-based modeling approach apart from other data-driven machine-learning approaches, and that (2) it would be counter-intuitive to advocate a data-driven approach with a seemingly data-rich case study (here "data-rich" refers to the fact that each MRB consists of thousands of watersheds) when the whole dissertation actually emphasizes ungauged watersheds.

Our explanation starts with explaining two significant advantages of the nested tree-based modeling approach. First of all, one great advantage of BART is that it outputs the posteriors of the model parameters, which could lead to posteriors of the target response. The advantage of having the posteriors is that the users/modelers can then derive the desired information at will, such as percentiles, moments, information gain, or the posterior mean and variances like what was demonstrated in the case study. Conditional simulation is also made easy when the posteriors are available, opening the door for Monte-Carlo analyses.

Second, following the statement that one can obtain the statistics or representative metric of interest, the nesting of BART models under CART can be done with the said metric, resulting in the corresponding probability mass function of the plausible BART models. For example, the classification shown in Figure 4.8 is based on RMSE, which is then based on the posterior mean values. This is essentially a proposal-comparison-based consideration of model structure uncertainty.

How do the aforementioned two advantages of the nested tree-based modeling approach justify the use at ungauged watersheds? First, of course the performance of the model depends on the quality and the quantity of training data. In this sense all modeling approaches are the same, and applying BART does not disproportionally enhance the predictive accuracy when the data are limited. However, what sets BART apart is the Bayesian feature that accounts for model parameter uncertainty properly in the form of conditional distribution, which cannot be done as easily with only a few point estimates or a few posterior statistics. Second, uncertainty exists not only for the model parameters but also for the models themselves. The nested tree-based modeling approach can help us obtain an informed empirical probability mass function, $p(B_k)$, of the plausible BART models (which was also exemplified in the case study). The fact that at ungauged watersheds in-situ hard data are absent and ex-situ data can be limited in quantity and/or quality accentuates the importance of uncertainty quantification, and the nested tree-based modeling approach offers a Bayesian solution to that, making itself not only applicable but also advantageous at ungauged watersheds.

One may then argue that how would a modeler make an informed proposal of plausible BART models in the first place? This is where physical knowledge come into play, and the proposal is indeed case specific. This is why we proposed the hypothesis of hierarchical similarity, which can be integrated with the nested tree-based modeling approach to study the behavior of a dynamic hydrologic similarity system, like what was demonstrated in the case study. Unlike the generality and the merits of the nested tree-based modeling approach, our findings regarding the variation of $p(B_k)$ and the shifts in dominant controlling factors of recharge are indeed specific to the context of the case study, which will be discussed next.

## The Hierarchical Similarity Hypothesis and The Shift in Dominant Physical Processes

With BART's ability to simultaneously model non-linear and/or interaction effects and present uncertainty in a fully Bayesian fashion, we are able to show how the controlling factors of hydrologic similarity vary among different watersheds, among different conditions, and among different accuracy metrics. These are all manifested in the case study under the context of the hierarchical similarity hypothesis.

Climate variables have been identified as the dominant factors in previous studies (see Section 4.1), and they are indeed on average the most dominant factors in our case study. However, the hierarchical similarity shows potential risks if one resorts to climate variables to define hydrologic similarity without considering other physical watershed characteristics,

especially the soil available water content.

The details of the hierarchical similarity are inferred from the data in the fashion of supervised machine learning, using six BART models and one benchmark model nested under one classification tree. It is of great importance to have two levels in such a system, as it allows for identification of the shifts of dominant factors under different conditions. These shifts indicate shifts in dominant physical processes, as exemplified by node 13 and 20 in Figure 4.8 where we observed the shift from water-limited evapotranspiration to energy-limited evapotranspiration. Therefore, we conjecture that it is the shift in dominant physical processes that is driving, and thus is reflecting, the shift in the controlling factors of hydrologic similarity under different conditions.

## Limitations of The Case Study

While proudly presenting and demonstrating the nested tree-based modeling approach, we do recognize several limitations of the case study from the aspects of the data set, the target response, and the partitioning of data.

### Scale of The Target Response

A major limitation of the case study is that the target hydrologic response is the logit normalized watershed-averaged annual groundwater recharge. This is a large-scale spatiotemporally homogenized response, and in this study, the data were based on baseflow analyses. To that end, a working assumption about the reliability of the baseflow analysis was made without rigorous proof. The findings of the case study are all under the context of this working assumption, and thus, they should not be applied to recharge/LNR at other spatiotemporal scales or to other hydrologic responses without careful considerations.

### The MRB-based Partitioning of Watersheds

Although we tried to justify the MRB-based partitioning by the reasons listed in Section 4.3, we acknowledge that this may not be the best partitioning method for demonstrating the full potential of the estimating power of BART. An associated limitation stems from the data not covering a desirable range of values. An example was already presented in Section 4.3 and Figure 4.3. The limitations in the data accentuate the advantage of our approach regarding the consideration of uncertainty, but it is also recognized that it could be challenging to discover the same findings if MRB 1 provided the training data for MRB 2, which is part of the reason why we kept the MRB-based partitioning. Another case of lack of data coverage can be found in our climate predictors data. Since aridity index is the ratio of potential evapotranspiration to precipitation ($\phi = E_p/P$), one might be surprised by the differences among the cases of $k = 1$, $k = 2$, and $k = 3$ in the results. The main reason is revealed in Figure 4.10. The $E_p$ values at the training and testing watersheds are so distinct that, essentially, all the testing watersheds are outliers from the point of view of a BART

model trained at the training watersheds. On the other hand, the $\phi$ values at the training and testing watersheds share the range from about 0.6 to 1.2, and only differ at the two extreme ends. In other words, the predictor-response relationships inferred by using $\phi$ can be transferred due to the overlapping range (Figure 4.10**(c)**), but the relationships inferred using $E_p > 1000mm$ cannot be effectively transferred to watersheds with $E_p < 1000mm$ (Figure 4.10**(b)**). Although it is not shown, a similar case can be found by comparing $\bar{\phi}$ with $E_p$. Although this might have been avoidable by using a more sophisticated design of

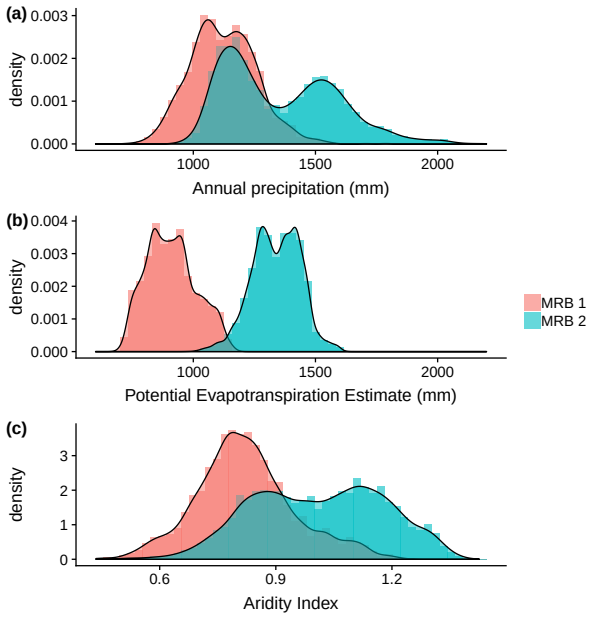


Figure 4.10: Distributions of **(a)** $P$, **(b)** $E_p$, and **(c)** $\phi$, at watersheds in MRB 1 (the testing watersheds) and MRB 2 (the training watersheds).

cross-validation, we kept the MRB-based holdout method on purpose. In addition to the reasons that were explained in Section 4.3, another motivation is that, in reality, the data at hand come in as is. This means there is no guarantee that the measurements will cover a particular range or that the watershed characteristics of the ungauged watersheds of interest are within a desirable range. The prevailing superiority of $\phi$ and $\bar{\phi}$ over $P$, $\bar{P}$, and $E_p$ found in our results shows an important advantage of dimensionless predictors, that they tend to be more transferable from one site to another, and hence, they may be more suitable for studies targeting ungauged watersheds.

### Limited Temporal Data Coverage

Another limitation is the lack of temporal coverage. Given limited data coverage along the time axis, in the case study we only studied the LNR in the year of 2002, and we considered two types of climate predictors: those from the same year and those from the

long term average. However, being the recharge process highly non-linear, it is not impossible that some predictors representing the antecedent conditions, such as precipitation from years prior to the year of 2002, could affect the LNR in the year of 2002. Not having multiple years of climate data prevents us from testing the effects of antecedent conditions or the effects that take place at various multi-year scales, and thus it is clearly a limitation of the case study. Because of this limitation, we made a steady state working assumption, with which we assume that the effect of climate predictors from the previous years are captured by the long term average predictors, and also assume the effect of climate change to be negligible. While acknowledging the inclusion of multiple years of climate data could have made an impact, note that the highly consistent roots of the trees in Figures 4.8 and 4.9 are based on soil AWC and the long term average aridity index, both of which are expected to be relatively insensitive to the inter-annual variation of climate predictors. Therefore, we expect the findings corresponding to the roots of the trees in Figures 4.8 and 4.9 to be relatively less affected by the limitation of not having multiple years of climate data.

**Non-comprehensive list of plausible models**

The proposal of plausible BART models was guided by a conceptual understanding and grouping of the available predictors. Like mentioned in Section 4.3, our proposal does not cover a comprehensive list of plausible models, nor does it necessarily include the "best" or the "true" model. The effect of different proposals of plausible BART models, which represents different perspectives of the conceptual understanding of the underlying physics, was not investigated in the case study, and remains as an interesting follow-up that could be pursued in future studies.

## 4.7 Conclusions

In this work, we proposed a nested tree-based modeling approach with three key features: (1) full Bayesian quantification of parameter uncertainty, (2) non-linear regression in order to model the predictor-response relationship, and (3) proposal-comparison-based consideration of model structure uncertainty. We applied the nested tree-based modeling approach to obtain logit normalized recharge estimates conditioned on ex-situ data at ungauged watersheds in a case study in the eastern U.S. We hypothesized a hierarchical similarity to explain the variation of the probability mass function of plausible models, and thus to investigate the behavior of a dynamic hydrologic similarity system. In addition to the demonstration of the nested tree-based modeling approach for assimilation of ex-situ data and estimating hydrologic responses at unsampled sites, we also showed how the approach can be used to investigate the variation of the controls of hydrologic similarity under different conditions. It should be pointed out that the nested tree-based modeling approach is independent of the target response and the predictors of interest, so it could be integrated with future stud-

ies within or beyond the field of hydrology in search of a hierarchical predictor-response relationship.

Besides the proposal of the nested tree-based modeling approach, this Chapter also provides some findings that are specific to the case study, contributing to the understanding of the physical principles governing robust regionalization among watersheds. Firstly, consistent with previous studies, we found that the climate variables are on average the most important controlling factors of hydrologic similarity at regional and annual scales, which means a climate-based regionalization technique is on average more likely to result in better estimates. However, with our hierarchical similarity hypothesis we revealed certain conditions under which non-climate variables become more dominant than climate variables. In particular, we demonstrated how soil available water content stood out to be the pivotal indicator of the variable importance of aridity in hydrologic similarity. Moreover, we showed that with hierarchical similarity one could identify shifts in dominant physical processes that are reflecting shifts in the controlling factors of hydrologic similarity under different conditions, such as water-limited evapotranspiration versus energy-limited evapotranspiration, or homogeneous and foliated bedrock versus heterogeneous bedrock. As the controlling factors change from one condition to another, the suitable regionalization technique also changes. We demonstrated how the hierarchical similarity hypothesis could indicate mechanisms by which available water content, aridity, and other watershed characteristics dynamically affect hydrologic similarity. The nested tree-based modeling approach can be applied to identify plausible sets of watershed characteristics to be considered in the regionalization process.

The specific contributions in groundwater recharge estimation at unsampled watersheds may be viewed differently depending on individual cases. In a situation where groundwater recharge is the ultimate target variable at ungauged watersheds, the nested tree-based modeling approach offers a systematic way to obtain informative predictive distributions that are conditioned on ex-situ data. In a difference case, where recharge estimation at ungauged watersheds is but one component of a greater project, the aforementioned informative predictive distributions can be treated as informative ex-situ priors, which could be further updated and/or integrated into simulation-based stochastic analyses where recharge is an input/component of other models/functions. At unsampled watersheds that will become gauged in the foreseeable future (as described in Section 2), the informative predictive distributions again serve as informative ex-situ priors that could guide the design of the sampling campaign, as different recharge flux magnitudes require different quantifying techniques (Healy, 2010; Scanlon et al., 2002). Lastly, the hierarchical similarity hypothesis offers one plausible explanation of the dynamic nature of hydrologic similarity, which affects the application of regionalization.

# Chapter 5

# Imputation-free Assimilation of Ex-situ Soft Data: Bounds

## 5.1 Introduction

In Section 1.2, we have discussed how ex-situ soft data could be a potentially valuable source of information in hydrogeology, and we specifically looked at one type of ex-situ soft information: bounds. We recognize the great work by Cucchi et al. (2019), who proposed a practical approach that uses data imputation to convert ex-situ soft data into ex-situ hard data, but also recognized that the involvement of imputation introduces artificial biases. In this Chapter, we aim to improve upon Cucchi et al. (2019) by proposing an approach for the assimilation of ex-situ bounds that is free of imputation-induced biases. In addition, we allow bounds to be imprecise to account for the uncertainty in bounds due to, for example, uncertainties from expert elicitation. The results of this study provide guidelines for the consideration of bounds as a form of ex-situ soft data in situations of extreme data scarcity.

## 5.2 Bayesian Hierarchical Modeling:Assimilation of Ex-situ Bounds

The Bayesian hierarchical model introduced in Section 2.3 (also in Cucchi et al. (2019)) provides a good starting point for our goal in this Chapter. We follow the same intra-site model as in Equation 2.10. Note that when data scarcity is extreme, there might be no ex-situ point observation available. In that case, instead of imputing point observations, we propose to skip the intra-site model altogether and proceed to the inter-site model, where we propose to introduce bounds.

At inter-site level, we rewrite Equation 2.11 as a truncated distribution as follows:

$$\Theta_i \sim p\left(\Theta_i | \Phi, L_i \leq \Theta_i \leq U_i\right), \tag{5.1}$$

where we constrain $\Theta_i$ by not only the hyperparameters but also the site-specific upper bound, $U_i$, and the site-specific lower bound, $L_i$. By truncating the distribution of $\Theta_i$, we force the distribution to agree with the bounds, but still keep the flexibility of the distribution to concentrate at any point within the bounds. To model the uncertainty of the bounds, we model the upper and the lower bounds as random variables:

$$\begin{aligned} L_i &\sim p_L\left(L_i|a_i\right) \\ U_i &\sim p_U\left(U_i|b_i\right), \end{aligned} \tag{5.2}$$

where $a_i$ and $b_i$ parameterize the distribution of $L_i$ and $U_i$, respectively. Thus, we obtain a nuanced version of Equation 5.1 as follows:

$$\Theta_i \sim p\left(\Theta_i|\Phi, a_i, b_i\right). \tag{5.3}$$

Note that Equations 5.1 and 5.3 represent different degrees of uncertainty. In the former, the bounds are certain and given, while in the latter, the bounds are uncertain and we are only given the parameters that determine the distributions of bounds. Bayesian inference is applied to obtain the distribution of the hyperparameters conditioned on all the data:

$$f\left(\Phi|\mathbf{y}, A, B\right), \tag{5.4}$$

where $A = \{a_i; i = 1, ..., N_i\}$, and $B = \{b_i; i = 1, ..., N_i\}$. Note that we kept $\mathbf{y}$ in Equation 5.4 to show that our approach does not deprive Bayesian hierarchical models of the capability of assimilating ex-situ hard data.

Equations 5.3 and 5.4 are the essence of our approach. By incorporating the constraints of bounds without data imputation, we avoid introducing artificial biases.

## 5.3 Synthetic Case Study: Ex-situ Prior of Effective Hydraulic Conductivity

In this Section, we demonstrate our approach of the assimilation of ex-situ bounds in a synthetic case study. In this case study we set the effective hydraulic conductivity of a site as the target variable. Unlike in Chapter 3 where some in-situ soft information was available, in this synthetic case study we focus on an extreme situation of data scarcity where both in-situ hard and soft data are unavailable, and there are only three hydrologically similar sites at hand. The goal of the case study is to first demonstrate the capability of our proposed approach, and then to investigate how informative the ex-situ bounds could be in different situations.

### Data

Log-transformed hydraulic conductivity ($\log_{10} K$) observations from 3 coal-dominant sites are retrieved from the World Wide Hydrogeological Parameters Database (WWHYPDA) (an

open-source database of hydrogeologic parameters, Comunian and Renard, 2008). The three selected sites are the only three coal-dominant sites in WWHYPDA, and can be looked up in WWHYPDA with site ID 19, 21, and 25. The database is publicly available on an as-is basis, with no warranties regarding the correctness of the data. For demonstration purpose in this case study, a working assumption is made that the observations at those 3 coal-dominant sites are observations of the site-homogenized effective conductivity, i.e., the target variable of interest. The histograms and the Gaussian Quantile-Quantile plots (also known as QQ-plot, to show whether the data are in Gaussian distributions) are provided in Figure 5.1. From the QQ-plots, since the data are not falling on the line of Gaussian quantiles, the data are not in Gaussian distributions, which can also be observed from the histograms. The data values range from $-9$ to $-3$, are more or less concentrated between $-5$ and $-6$, and are skewed towards the negative side.

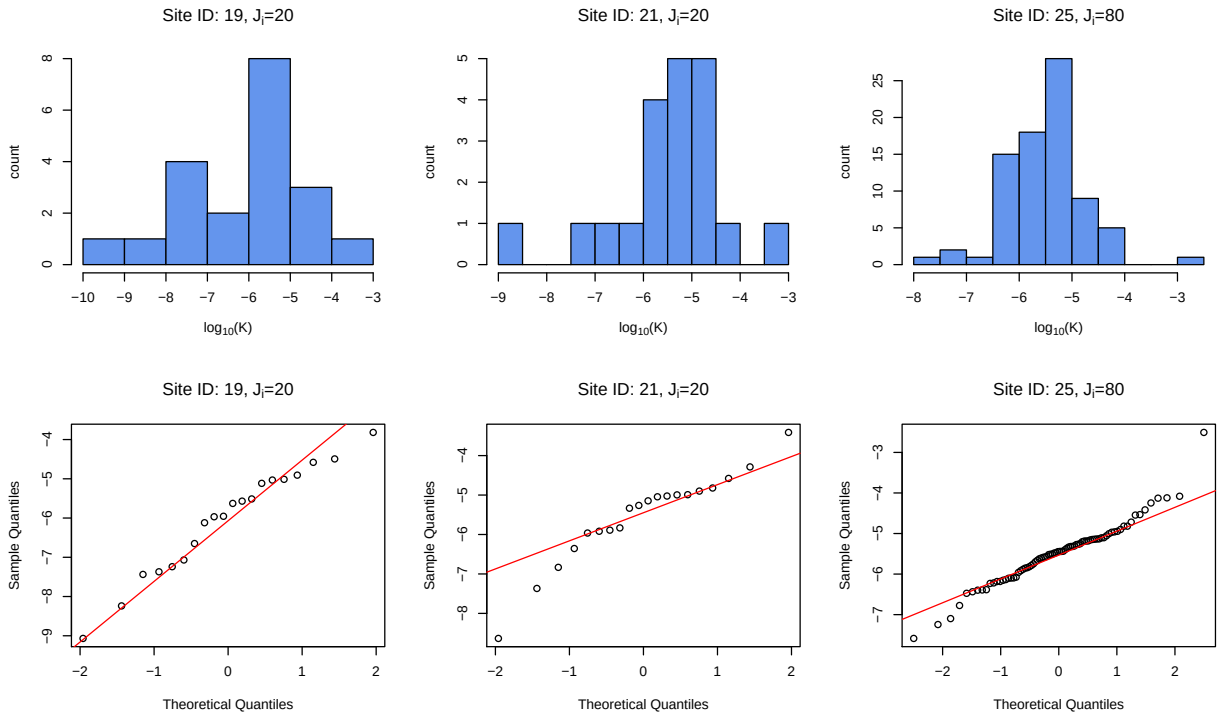


Figure 5.1: Histograms of log-transformed hydraulic conductivity data from the 3 coal-dominated sites (upper row), and the corresponding Gaussian Quantile-Quantile plots (lower row). Each plot is labeled by the its ID number from the WWHYPDA database, and $J_i$ is the total number of observation at the $i^{th}$ site.

## Application of Hierarchical Bayesian Model

In this study, we model the log-transformed conductivity by Gaussian distribution, resulting in the intra-site level Gaussian distribution of the point observations:

$$y_{i,j} \sim N\left(\theta_i, \sigma_i^2\right), \tag{5.5}$$

where the site-specific parameter set includes the Gaussian mean and variance, i.e., $\Theta_i = \{\theta_i, \sigma_i^2\}$. If a sampled site does not have point observations, this intra-site level is skipped.

At the inter-site level, we applied three different distributions depending on the availability of hard and/or soft data at the site from which we are borrowing information (which we refer to as "donor site").

1. If only hard data are available at a donor site, then the site-specific mean follows an inter-site Gaussian distribution:

$$\theta_i \sim N\left(\mu, \tau^2\right), \tag{5.6}$$

and the site-specific variance follows an inter-site inverse-Gamma distribution:

$$\frac{1}{\sigma_i^2} \sim Gamma\left(\alpha, \beta\right), \tag{5.7}$$

where the hyperparameters include the inter-site Gaussian mean and variance, and the inter-site Gamma shape and rate parameters, i.e., $\Phi = \{\mu, \tau^2, \alpha, \beta\}$.

2. If both hard and soft data, i.e., the point observations and the bounds, are available at a donor site, then we first model the bounds from uniform distributions:

$$\begin{aligned} L_i &\sim Uni(a_{1,i}, a_{2,i}) \\ U_i &\sim Uni(b_{1,i}, b_{2,i}). \end{aligned} \tag{5.8}$$

where $a_i = \{a_{1,i}, a_{2,i}\}$ defines the lower of upper bounds of $L_i$, and similarly does $b_i = \{b_{1,i}, b_{2,i}\}$ define the lower and upper bounds of $U_i$. Next, we constrain the inter-site Gaussian distribution by truncating it:

$$\theta_i \sim c \cdot N\left(\mu, \tau^2 | L_i \leq \theta_i \leq U_i\right), \tag{5.9}$$

where $c$ is a scaling constant that makes the truncated distribution integrate to one. The site-specific variance is modeled in the same way as in the previous item, not affected by bounds.

3. If only bounds are available at a donor site, then the site-specific mean is modeled with the truncated Gaussian as in Equation 5.9, but the modeling of site-specific variance is skipped.

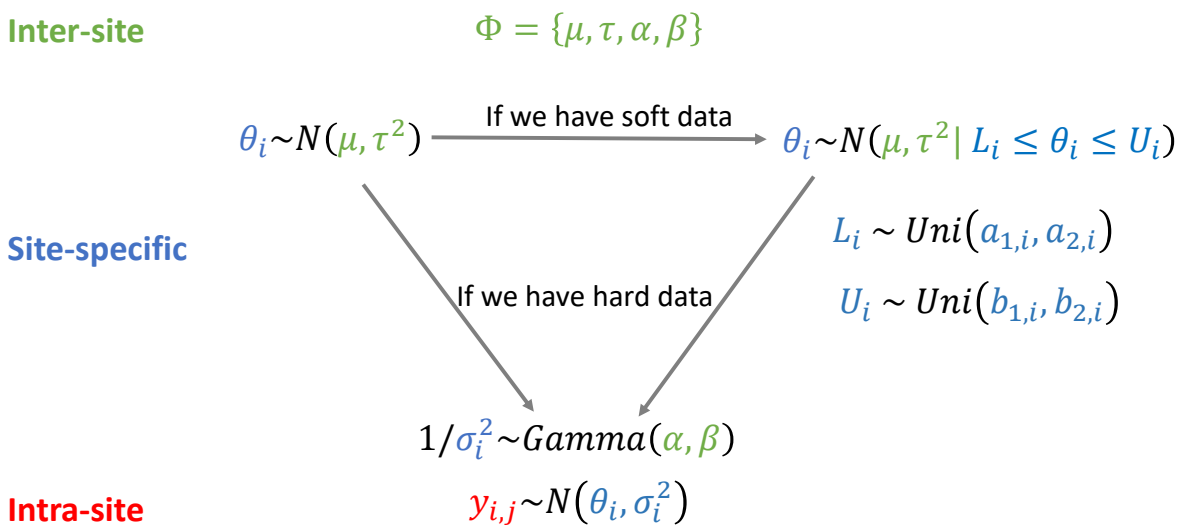


Figure 5.2: Schematic diagram of the application of the hierarchical Bayesian model in this case study.

A summarizing schematic diagram is also provided in Figure 5.2.

The Bayesian inference computation is done within R environment using the package *nimble* (Valpine et al., 2017), with which we run MCMC simulations (see Section 2.2) to approximate the conditional distribution of $\Phi$. Although there are four hyperparameters, for this case study we focus only on $\mu$, which is of greater importance in terms of estimating the effective conductivity at an unsampled site of interest.

## The Benchmark Ex-situ Prior

The ideal case in this synthetic case study is that the three donor sites are all sampled, so one can obtain an ex-situ prior conditioned on ex-situ hard data from the 3 sites described in Figure 5.1. This ex-situ prior is denoted by $p(\mu|\mathbf{y})$, and is the most informative ex-situ prior one can get if ex-situ bounds are not considered (Figure 5.3). From this point forward, we refer to it as the benchmark prior. To investigate the effect of ex-situ bounds, we will proceed in two different directions by conducting two different numerical experiments, both of which involve building different ex-situ priors and comparing them with the benchmark prior.

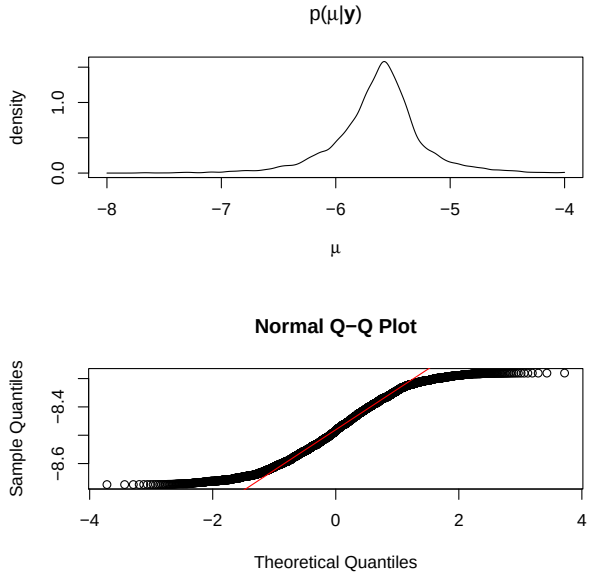


Figure 5.3: The benchmark ex-situ prior, and its Gaussian QQ-plot.

## Numerical Experiment 1: Removal, Replacement, and Augmentation of Ex-situ Hard Data by Bounds

The first numerical experiment is meant to test the information gain from the bounds. Looping over the three donor site ($i = 1, ..., 3$), we build three additional ex-situ priors by applying the following three designs:

1. Remove the $i^{th}$ donor site from the ex-situ data set altogether, thus ending up with an ex-situ prior of $\mu$ conditioned only on data from the other two donor sites, denoted by $p(\mu|\mathbf{y}_{-i})$ where $\mathbf{y}_{-i}$ includes all point observations but those from the $i^{th}$ site.

2. Remove the point observations from the $i^{th}$ site, but add the bounds of the $i^{th}$ site into the ex-situ database. The bounds are defined according to the empirical confidence intervals of the point observations, as follows: the 90% confidence intervals are used to define $a_{2,i}$ and $b_{1,i}$, and the 95% confidence intervals are used to define $a_{1,i}$ and $b_{2,i}$. The resulting ex-situ prior of $\mu$ is conditioned on hard data from two sites and soft data from one site, denoted by $p(\mu|\mathbf{y}_{-i}, a_i, b_i)$.

3. Do not remove any point observation, and add the bounds at the $i^{th}$ donor site; namely we augment the hard data by bounds at the $i^{th}$ donor site. The bounds are defined in the same way as defined above. The resulting ex-situ prior of $\mu$ is conditioned on hard data from all three sites, as well as soft data from the $i^{th}$ donor site, denoted by $p(\mu|\mathbf{y}, a_i, b_i)$.

The three measures represent removal, replacement, and augmentation of ex-situ data at the $i^{th}$ site. By comparing the ex-situ priors resulting from these measures with the benchmark prior, we investigate the effect of bounds from different aspects: as an additional source of information, or as an alternative source of information.

To quantify the difference between two ex-situ priors, we adopt the Kullback-Leibler Divergence (KLD) as a quantitative metric. Corresponding to the aforementioned three designs, three KLD values can be calculated:

$$\Lambda_{remove,i} = \int_{\infty}^{\infty} p(\mu|\mathbf{y}) \ln \frac{p(\mu|\mathbf{y})}{p(\mu|\mathbf{y}_{-i})} d\mu, \tag{5.10}$$

$$\Lambda_{replace,i} = \int_{\infty}^{\infty} p(\mu|\mathbf{y}) \ln \frac{p(\mu|\mathbf{y})}{p(\mu|\mathbf{y}_{-i}, a_i, b_i)} d\mu, \tag{5.11}$$

$$\Lambda_{augment,i} = \int_{\infty}^{\infty} p(\mu|\mathbf{y}, a_i, b_i) \ln \frac{p(\mu|\mathbf{y}, a_i, b_i)}{p(\mu|\mathbf{y})} d\mu, \tag{5.12}$$

where $\Lambda_{remove,i}$ and $\Lambda_{replace,i}$ quantify the information loss due to removing and replacing the hard data at the $i^{th}$ donor site, respectively, and $\Lambda_{augment,i}$ quantifies the information gain due to augmenting the hard data at the $i^{th}$ donor site.

## Numerical Experiment 2: Tightness and Precision of Bounds

The second numerical experiment is meant to test how our approach can account for the uncertainty in the ex-situ bounds if the bounds are imprecise, and test the effect of bound precision and tightness on the ex-situ prior of $\mu$. In this numerical experiment, we introduce the fourth donor site ($i = 4$), a hypothetical site, to the ex-situ data set. The fourth site comes without point observations and only with bounds, but the bounds can be of different levels of tightness or precision. The goal is to see the effect of bound precision and tightness at this fourth site on the resulting ex-situ prior.

To begin with, we calculate the inter-quartile range of $\mathbf{y}$,

$$Q = IQR(\mathbf{y}) \approx 0.52, \tag{5.13}$$

which will be used as a unit of variability in this numerical experiment. We define three levels of bound tightness as follows:

1. tight: $b_{1,4} - a_{2,4} = 0.5Q$;
2. less tight: $b_{1,4} - a_{2,4} = Q$;
3. loose: $b_{1,4} - a_{2,4} = 1.5Q$.

Similarly, we define three levels of bound precision:

1. precise: $b_{2,4} - b_{1,4} = a_{2,4} - a_{1,4} = 0.5Q$;

2. less precise: $b_{2,4} - b_{1,4} = a_{2,4} - a_{1,4} = Q$;

3. imprecise: $b_{2,4} - b_{1,4} = a_{2,4} - a_{1,4} = 1.5Q$.

As a consequence, we can obtain nine different ex-situ priors, corresponding to nine different bounds from the fourth site. These ex-situ priors are compared against the benchmark ex-situ prior to investigate the effect of bound tightness and precision.

## 5.4 Case Study Results: Removal, Replacement, and Augmentation of Ex-situ Hard Data by Bounds

The ex-situ priors obtained by removing, replacing, and augmenting the ex-situ hard data at the three donor sites are shown in Figure 5.4. First of all, augmentation has a limited effect, if any, on the ex-situ prior of $\mu$. This shows that if the bounds are closely agreeing with the point observations (e.g., the bounds are equal to the confidence intervals of the point observations, as defined here), then having bounds as an additional source of information does not have a significant effect.

However, if point observations are not available, then having bounds as an alternative source of information does have a significant effect, as shown by the noticeable differences between the red curves and the green curves in Figure 5.4. When bounds are available, the mean value of the ex-situ prior is more concentrated towards that of the benchmark, and the peak probability density is also higher. The same finding can be obtained from a different aspect: the KLD values (Figure 5.5). At all three sites, we observe $\Lambda_{remove,i} > \Lambda_{replace,i}$, showing that the information loss is reduced if we at least have bounds. Information gain due to augmentation is only noticeable at the site with ID 19, and even so, the magnitude of $\Lambda_{augment,i}$ is at least on order smaller than that of $\Lambda_{remove,i}$ or $\Lambda_{replace,i}$.

It might sound tautological that having ex-situ bounds is better than having nothing, and previous studies have already pointed out that soft data could be informative (e.g., from expert elicitation; Garthwaite et al., 2005). However, what is significant here is the quantification of the effect of ex-situ bounds, in terms of the conditional distributions of $\mu$ or KLD, under different condition.

### Case Study Results: The Effect of Tightness and Precision of Bounds

The effect of bound tightness is shown in Figure 5.6. Regardless of the level of precision, the tightest bounds consistently lead to the most informative distribution with the highest peaked probability density, which is expected. Surprisingly, however, the less tight bounds are not always more informative than the loose bounds, indicating that some interplay between tightness and precision might exist. Equally surprising is that the most precise bounds do not necessarily lead to the most informative distribution at each level of tightness

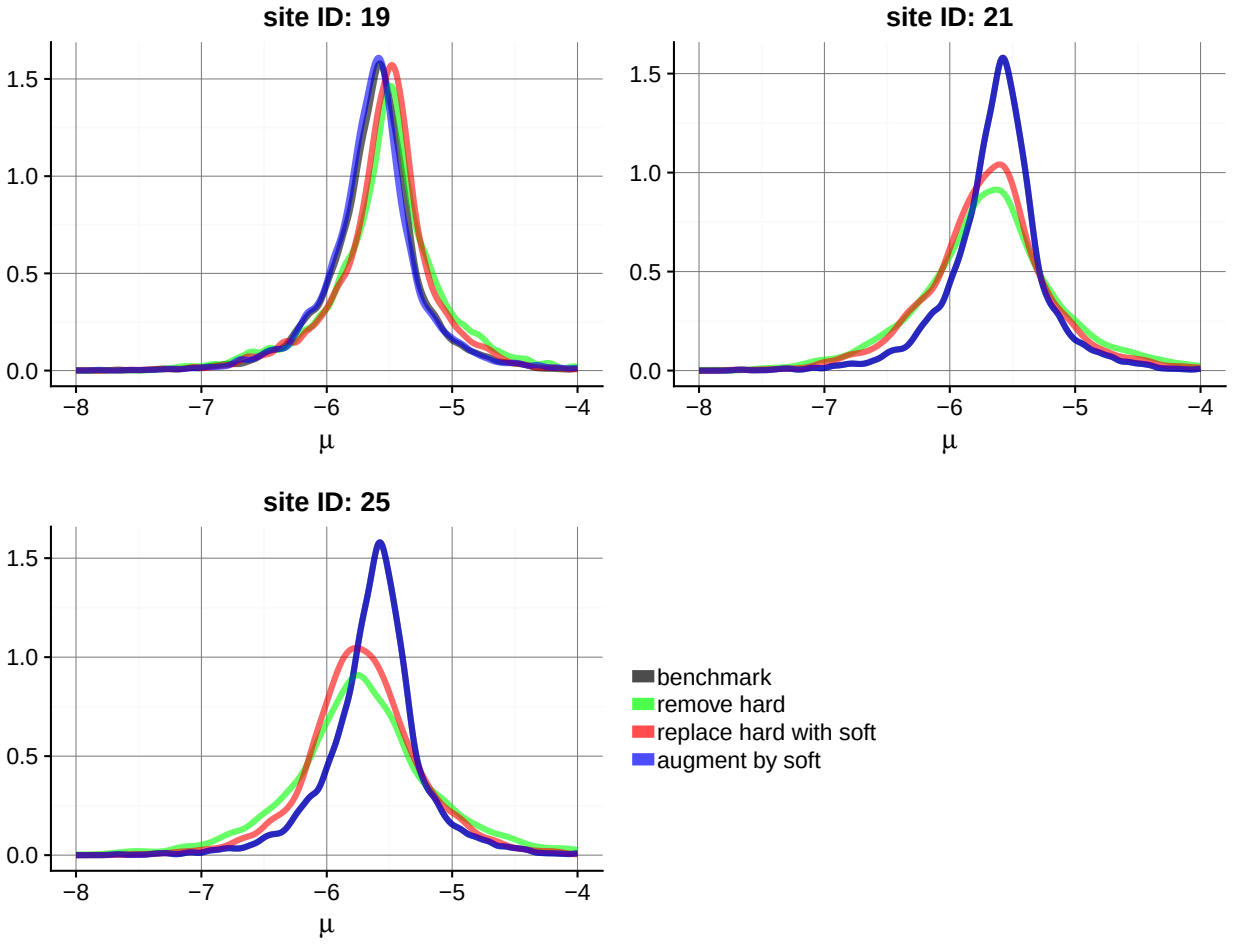


Figure 5.4: The ex-situ priors obtained by removing, replacing, and augmenting the ex-situ hard data at the three sampled sites, compared against the benchmark ex-situ prior.

(Figure 5.7). What is expected, though, is that the most imprecise bounds consistently lead to the least informative distribution, at each level of tightness.

The findings consist of some expected results as well as some surprising results. For the expected part, of course when we have bounds, we would like them to be tight and precise, thus potentially carrying more information for conditioning. So it is expected that the most imprecise bounds lead to the least informative distribution at all tightness levels, and the tightest bounds lead to the most informative distribution at all precision levels. What is unexpected is that, if we are to improve a set of loose and imprecise bounds, improvement in precision might have a slightly greater effect than an improvement in tightness.

That said, also note that Figures 5.6 and 5.7 are zoomed-in view of the ex-situ priors at their peaks. If we zoom out and take the holistic view, the differences become less significant. This is shown in Figure 5.8, where we compare the ex-situ prior obtained with imprecise and

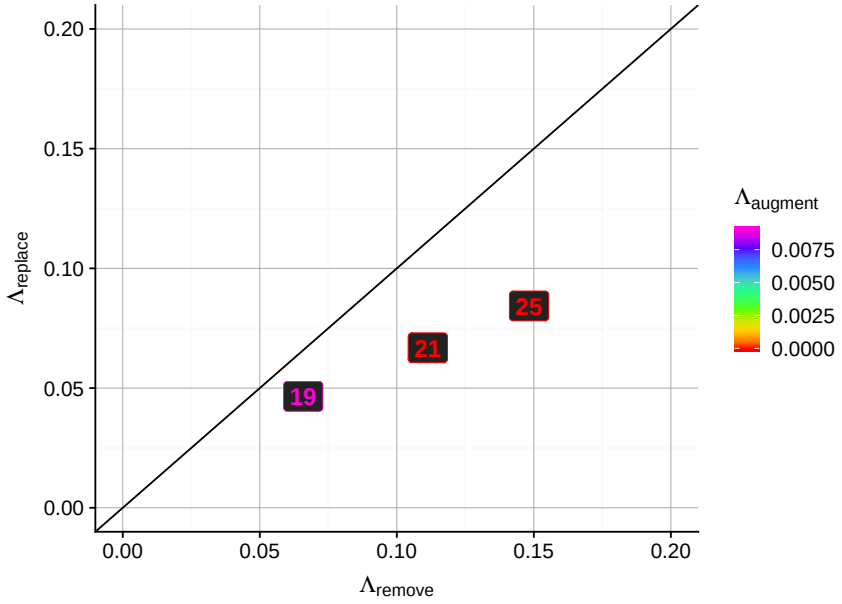


Figure 5.5: Scatter plot of the KLD values calculated from Equation 5.10 through 5.12 at the three sampled sites. The black line is the 45-degree line, indicating the case of $\Lambda_{remove,i} = \Lambda_{replace,i}$. The sites are labeled by their WWHYPDA site ID. The color, as shown by the color bar, indicates the value of $\Lambda_{augment,i}$.

loose bounds at the fourth site with a "wrong example" of ex-situ prior, obtained with bounds at the fourth site that are tight and precise but two orders of magnitude ($-2$ on log scale) smaller. It is obvious that even a set of imprecise and loose bounds can provide additional information to improve the benchmark ex-situ prior, but if the bounds are bounding a "wrong" range of values that are is contradicting the data from the other donor sites, we end up with a much less informative distribution, even when the bounds are quite precise and tight. Note that in the case of imprecise/loose bounds, we allowed a room of adjustment for precision and tightness to be roughly 1.5 orders of magnitude ($1.5Q$ on log scale, with $Q \approx 0.52$), which is comparable to the $-2$ for the "wrong example". Thus, the result is not due to disproportional adjustment the tightness, precision, and center of the bounds.

The takeaway message is that, when it comes to determining the bounds (e.g., via expert elicitation), it is better to take a conservative view where we try to account for uncertainties and determine relatively wide/imprecise bounds that have a higher chance of including the true value, than to take a progressive view where we determine relatively narrow/precise bounds that might fail to include the true value. The effects of tightness and precision are indeed minor, when compared to the effect of accuracy.

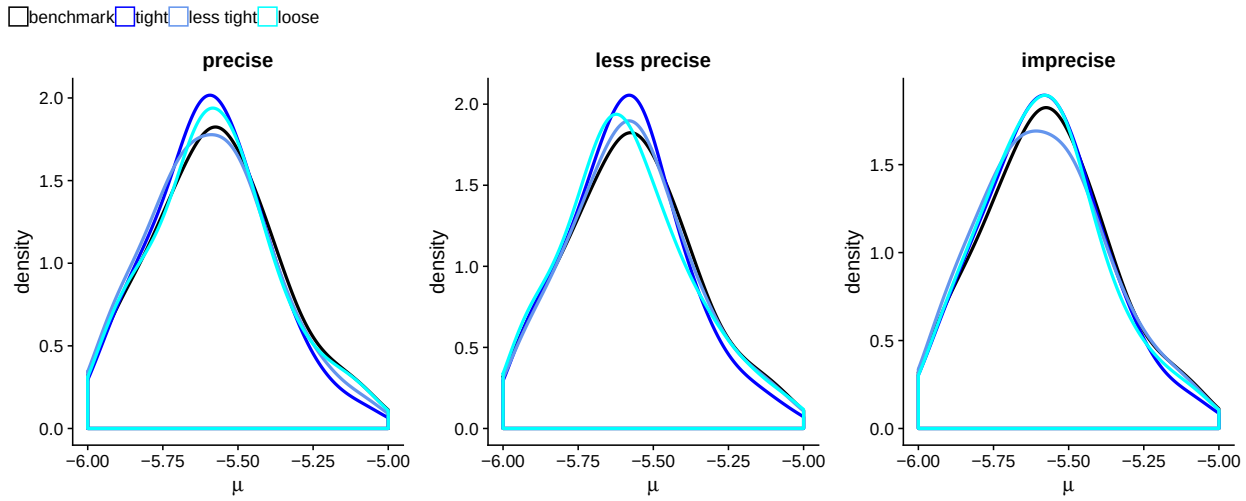


Figure 5.6: Ex-situ priors at different levels of tightness and precision, highlighting the effect of tightness. Note that this is the zoomed-in view on the $x$ axis, in order to make noticeable the differences between ex-situ priors.

## 5.5 Conclusions

In this Chapter, we analytically established a Bayesian hierarchical model. In terms of assimilation of ex-situ bounds, we improved upon Cucchi et al. (2019) by replacing data imputation by truncation of distributions and allowing the truncating values to be imprecise, thus (1) avoiding imputation-induced artificial biases and (2) accounting for potential uncertainties in the bounds. Compared to Cucchi et al. (2019), our improvement is made by sacrificing the versatility of dealing with various forms of ex-situ soft data simultaneously. We consider the search for an equally versatile and imputation-free model an interesting follow-up study that could be pursued in the future.

We demonstrated the approach in a synthetic case study, and obtained a few specific findings. While subject to limitations and working assumptions (e.g., the distribution forms assumed in the case study, the assumption that the hard data are for the effective hydraulic conductivity, and the choice of data in the case study), our specific findings are listed as follows.

1. We investigated the information gain from bounds either as an additional or as an alternative source of information, and quantitatively proved that when facing extreme data scarcity, ex-situ bounds could be a significant source of ex-situ information.

2. When determining bounds, is it recommended to account for various sources of uncertainty (e.g., in expert elicitation there could be multiple experts providing different opinions, etc.) and be more conservative to obtain bounds that are more likely to include the true value. The effect of tightness and precision of the bounds are relative

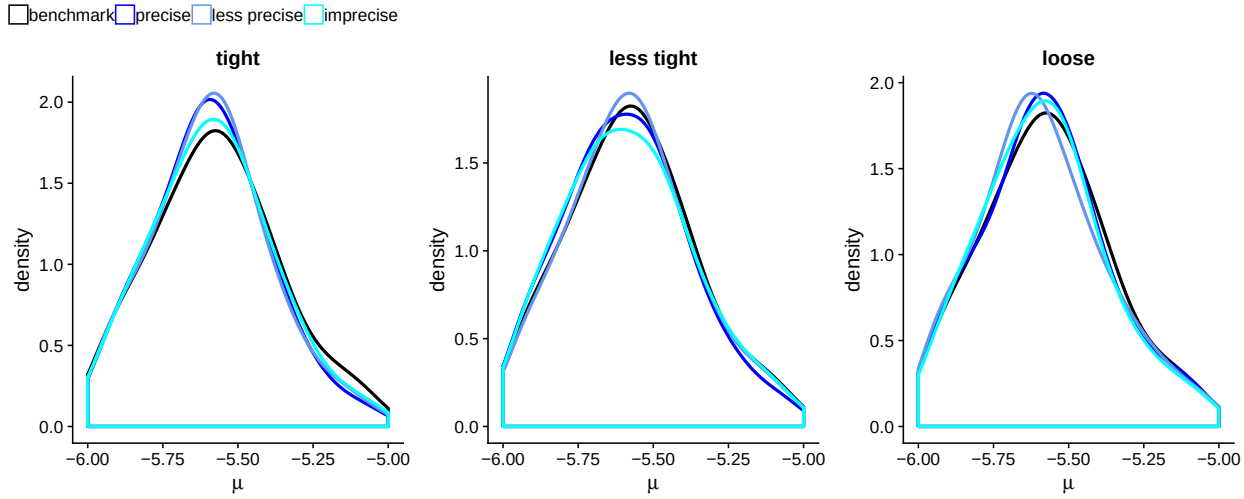


Figure 5.7: Ex-situ priors at different levels of tightness and precision, highlighting the effect of precision. Like Figure 5.6, this is the zoomed-in view on the $x$ axis, in order to make noticeable the differences between ex-situ priors.

minor, compared to the potentially immense price we pay when we falsely adopt precise and tight bounds.

These suggestions may contribute to future applications of expert elicitation by highlighting what the modelers need to consider and converse with experts about, to obtain a potentially more informative ex-situ prior of the target variable of interest.

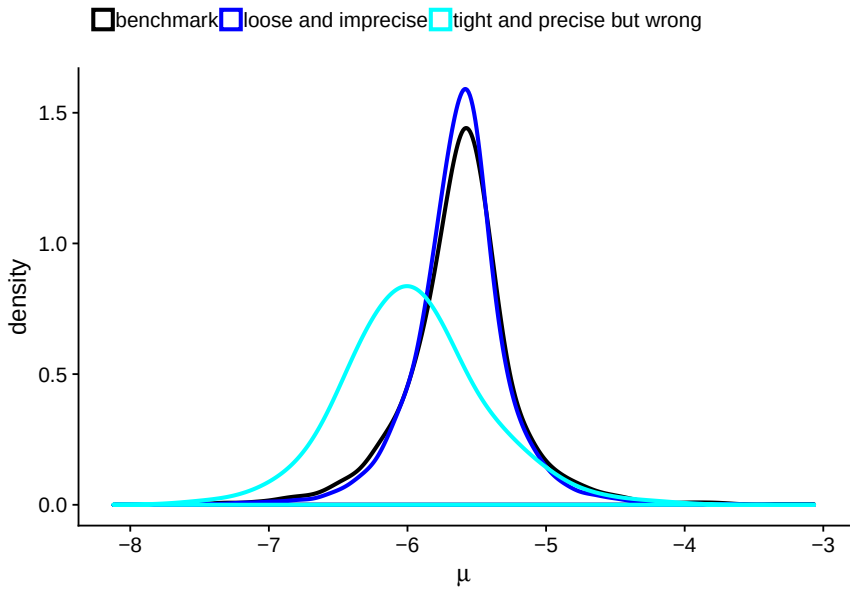


Figure 5.8: Comparing the case where imprecise, loose, but accurate bounds are used with the case where tight, precise, but inaccurate bounds are used to obtain the ex-situ prior of $\mu$.

# Chapter 6

# Conclusions

This dissertation focuses on quantification and reduction of uncertainty in hydrogeological modeling at unsampled locations. A Bayesian view is adopted throughout the dissertation, where the objective is to obtain informative distributions of the target variable, e.g., hydrogeological parameters or responses, that are conditioned upon limited information from various sources. In Chapter 1 we identified and discussed the challenges in stochastic hydrogeologic modeling under in-situ data scarcity. The overarching goal of this dissertation is to develop advanced Bayesian stochastic modeling approaches that overcome the challenge of the lack of in-situ data at unsampled locations. With that goal in mind, the research objectives involve the proposal of three new approaches that improve upon the theoretical basis introduced in Chapter 2.

In Chapter 3, we proposed a rapid, approximate Bayesian computational technique, named Rapid Impact Modeling (RIM), that features efficient assimilation of in-situ soft data in stochastic hydrogeologic modeling. The innovation of RIM behind its efficiency is replacing the daunting and demanding tasks of likelihood computation and inverse modeling with direct conditioning of the estimates of hydrogeologic responses on the available information. Uncertainty in the estimates is accounted for in the form of conditional predictive distribution. The advantages that set RIM apart from other data assimilation approaches are its capability of assimilating both in-situ and ex-situ data, its efficiency, and its suitability for large-scale hydrogeologic modeling at complicated, heavy-impact, yet poorly sampled sites. We demonstrated RIM at the Mingtang tunnel project to investigate the environmental impact of groundwater drawdown. With RIM, we conditioned the estimates of groundwater drawdown on measured groundwater infiltration into the Mingtang tunnel, and successfully assessed the environmental impact in the face of data scarcity.

In Chapter 4, we further investigated the assimilation of ex-situ data by proposing the nested tree-based modeling approach, which is essentially a nested coupling of Bayesian Additive Regression Tree (BART) and Classification/Regression Tree (CART). Although both BART and CART have been published for decades, the innovation of the approach is its nested structure, which makes our approach simultaneously feature (1) full Bayesian quantification of parameter uncertainty, (2) non-linear regression for modeling the hydrologic

response of interest, and (3) proposal-comparison-based consideration of model structure uncertainty. In addition to the approach, we also proposed a hypothesis of hierarchical similarity. The integration of the nested tree-based modeling approach and the hypothesis of hierarchical similarity allows us to investigate the dynamic behavior of hydrologic similarity, and to explain the variation of the controls of hydrologic similarity under different conditions. In the case study of groundwater recharge estimation in the Eastern U.S., we identified several conditions where hydrologic similarity is dominated by non-climate variables, such as soil available water content, and revealed the potential risk one may face if one resorts to climate variables for hydrologic similarity regardless. The proposed nested tree-based stochastic modeling approach can be applied in any field to study a dynamic and hierarchical predictor-response relationship, and the hierarchical similarity hypothesis contributes to the understanding of the physical principles governing robust regionalization among watersheds.

In Chapter 5, we improved upon Cucchi et al. (2019) by analytically deriving a Bayesian hierarchical model for the imputation-free assimilation of one type of ex-situ soft data: ex-situ bounds. We replace data imputation with distribution truncation to avoid artificial biases, and allow the truncating values to be imprecise to account for the uncertainty in the ex-situ bounds. The approach was demonstrated in a synthetic case study that was designed to represent extreme data scarcity, where in-situ information is unavailable and ex-situ information is limited both in quantity and in quality. We quantitatively proved the usefulness of ex-situ bounds under extreme data scarcity. Furthermore, we recommended conservative ex-situ bounds that are more likely to cover the true value, over progressive bounds that might underestimate uncertainty and lease to false conditioning.

All the stochastic approaches proposed in this dissertation are improvements based on existing open-source and ready-to-use approaches. The proposed approaches are initially designed for hydrogeologic modeling at unsampled locations, with the goal of representation and reduction of uncertainty under Bayesian context, conditioning on in-situ soft data and/or ex-situ data. While the proposed approaches were all demonstrated with hydrogeologic case studies, the approaches themselves are general and independent of the field of study. Thus, the new approaches proposed in this dissertation are expected to contribute to future applications of stochastic modeling, within or beyond the field of hydrogeology.

# Appendix A

# Compilation of Ready-to-use Software Solutions to Stochastic Hydrgeologic Modeling

Table A.1: Table of software solutions for stochastic hydrogeology; modified from Rubin et al. (2018)

| **Software** | **Main focus area** | | | **User friendliness** | **Cost** | |
|---|---|---|---|---|---|---|
| | Random field simulation | Forward modeling | Inverse modeling | GUI | Free | Commercial |
| DeeSee (Straubhaar, 2015) | ✓ | | | ✓ | | ✓ |
| FIELDGEN (Doherty and Hunt, 2010) | ✓ | | | | ✓ | |
| GMS (Aquaveo, 2013) | ✓ | ✓ | ✓ | ✓ | ✓ | |
| GSLIB (Deutsch and Journel, 1998) | ✓ | | | ✓ | | ✓ |
| GSTAT (Pebesma, 2004) | ✓ | | | ✓ | ✓ | |
| Isatis (Bleines et al., 2004) | ✓ | | | ✓ | | ✓ |
| PEST (Doherty, 2005) | | | ✓ | ✓ | ✓ | |
| iTOUGH2 (Finsterle, 2011) | | | ✓ | ✓ | | ✓ |
| MAD (Osorio-Murillo et al., 2015) | | | ✓ | ✓ | ✓ | |
| SGeMS (Remy et al., 2009) | | ✓ | ✓ | ✓ | | ✓ |
| TPROGS (Carle, 1999) | ✓ | | | ✓ | | ✓ |
| spMC (Sartore, 2013) | | | ✓ | | | ✓ |
| HYDRO_GEN (Bellin and Rubin, 1996) | ✓ | | | | ✓ | |
| Isim3D (Gómez-Hernández and Srivastava, 1990) | ✓ | | | ✓ | ✓ | |

# Appendix B

# The Groundwater Forward Model for The Mingtang Tunnel Case Study

The geographical scales of Mingtang tunnel and the impacted domain surrounding it are challenging: the horizontal extent of the domain is in the tens of kilometers, with depths to the tunnel reaching 500 m and more. Adding to the complexity is the need to employ variable grid density, using small grid blocks (of the order of a few meters) to model the tunnel and faults, while coarsening the grid (to tens and hundreds of meters) farther away from these elements. To address the scale challenges, we adopted a nested-modeling strategy, which we describe next.

To accurately simulate the groundwater flow system, a structured grid refinement method referred to as nested modeling, also known as telescopic mesh refinement (TMR), was employed. Regional inter-basin flow effects were incorporated into the modeling analysis of smaller model domains, and critical areas encompassing the local study site were simulated with high grid resolution. Nested modeling with variable density grid has the advantage of providing significant savings in computational cost (Dunning et al., 2004).

Figure B.1 presents schematically the nested modeling concept, as implemented in this study. The nested modeling follows a progression of steps, from the multi-catchment to regional to the local scales. This progression intends to (1) systematically communicate boundary conditions across modeling scales; (2) preserve continuity and consistency across scales, and (3) allow the use of fine grid near the tunnel and coarsening of the grid with distance from the tunnel. As a starting point in this approach, the hydraulic heads at the edges of the coarse-grid, multi-catchment scale model were determined using a regression model which correlates water table elevations with topography, obtained from borehole data. While similarly used in previous studies (Yang et al., 2009) and reasonably accurate, the regression may not be upheld near major flow conduits and barriers such as rivers and faults. Therefore, to ensure the accuracy of results near the tunnel, the regression model is not directly implemented at the regional model boundaries where these features may be close by. Next, the multi-catchment scale model is simulated to assign hydraulic head boundaries to the regional model. The regional model is then processed with coarse grids to simulate

regional flow effects outside of the local model area. Values of the hydraulic heads along the domain of the local scale model were then used as prescribed boundary conditions around the local flow model of the Mingtang tunnel site. The local model employs a fine grid, with block size of the order of 0.1 m (following Zaidel et al., 2010). MODFLOW (Harbaugh, 2005) drain package was applied around the tunnel face to simulate zero pressure conditions along the tunnel interior and the infiltration.

The nested modeling approach intends to transfer information across scales without affecting the simulated flow effects in the local model. Specifically, the boundaries of the local model should capture regional trends without affecting the local flow conditions in the vicinity of the tunnel. Towards this end, the effects of boundary positions on simulated drawdown and head results in the local model were iteratively analyzed to establish a safe distance. We concluded that positioning the nested model boundaries at a close distance of 100 m away from the tunnel resulted in drawdown simulations being greatly subdued. In contrast, expanding the local model beyond a threshold distance of 500m away from the tunnel displayed no observable changes in the gradient and values of simulated drawdowns and heads. Hence, nested model boundary conditions were applied at a minimum of 500 m away from the length and ends of the modeled Mingtang tunnel. This finding is in line with Raposo et al. (2010).

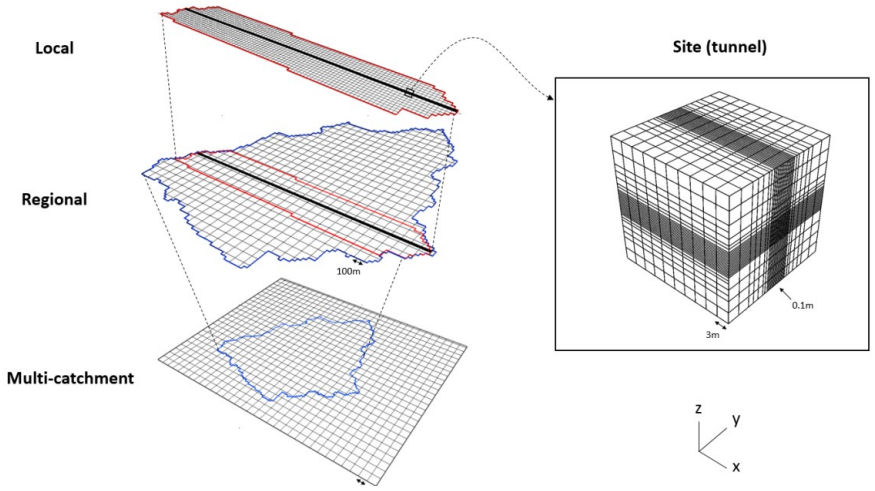


Figure B.1: Schematic representation of the concept of nested modeling of the Mingtang site. Prescribed hydraulic heads were transferred between model scales along the displayed colored head boundaries.

# Appendix C

# Defining the Ex-situ Informative Priors for The Mingtang Tunnel Case Study

**The Ex-situ Prior of The Effective Conductivity**

The objective here is to enrich the conditioning power by assimilating ex-situ data and establish the priors for the effective conductivities. The approach we pursue here combines the following elements: (1) establishing priors based on published data from similar sites (namely, the ex-situ priors); (2) in the absence of ex-situ priors, exploring the bounds of physical plausibility; (3) reducing data needs by identifying potential for homogenization.

While not directly informative on hydraulic conductivity, the rock type information (granite and gneiss) is useful for identifying hydrogeologically similar and better sampled sites around the world, from which priors could be defined. Hydrogeologically similar sites are obtained by screening candidate sites using a list of similarity criteria. Thus, similarity is modeled as a Boolean variance in this study: a sampled site is either similar to the Mingtang site if the criteria are met, or dissimilar otherwise.

A general description of the criteria is provided in Table C.1. We consider two different perspective of site similarity: (1) the intrinsic perspective considers intrinsic properties of the aquifer, such as lithology and tectonic activity, etc., and (2) the epistemic perspective relates to our knowing of the aquifer, such as how and where were the sample collected. Although an extended list of similarity criteria is always desirable as it increases the potential of finding sites that are hydrogeologically similar, it is difficult to obtain the information needed to evaluate all criteria, and we are often limited to employing a subset of criteria. A subset of criteria consisting of the field testing method employed, the rock type, the fracture condition, and the tectonic condition is used to establish the ex-situ priors in this case study. The selection of the criteria is based on data availability at the Mingtang site and at the candidate sites.

Using the criteria listed in Table C.1, we collect an ensemble of sampled sites, from which

Table C.1: Site similarity criteria for hydraulic conductivity with emphasis on rock formations. Light-grey-colored rows indicate criteria used in this study.

| Group | Criteria | General description | Mingtang site description |
|---|---|---|---|
| Intrinsic perspective | Lithification | A process of porosity reduction through compaction and cementation | Granite (igneous) and gneiss (metamorphic) |
| | Mineral ingredient | The mineral(s) of which the geological structure consist | Rock mass consisting of quartz, feldspar, and mica |
| | Geological structure | The outcome of the geological formation process | Fine-grained structure, granoblastic texture, and gneissic structure |
| | Weathering degree | A qualitative engineering perspective of rock mass | Overlay strata thickness of strongly weathered rock reaching 20m below ground surface |
| | Fracture condition | Separation in a geologic formation, such as a joint or a fault that divides the rock into two or more pieces; fluid movement may be enhanced along open fractures | Highly developed joint fissures, water bearing and conductive in the shallow rock formations |
| | Tectonic condition | Common tectonic activities include earthquakes, volcanoes, and orogeny, where the edges of two or more plates are in contact along huge linear faults | Two or three major fault zones almost normally intersecting with tunnel longitudinal axis |
| | Ground stress | Stresses could deform or crash rocks, which may induce space for groundwater storage and flow. | Geostress consisting of gravitational stress and vertical tectonic stress |
| Epistemic perspective | Sampling depth | The depth at which the sample is taken | The maximal depth of the tunnel reaches 400m below ground surface |
| | Sampling method | Including lab test, and in-situ pumping test, injection test and packer test etc., with each test having its own assumptions and limitations and support scale | Support volume much smaller than heterogeneity scale |

the information can be borrowed to establish the ex-situ priors. The sites for granite are listed in Table C.2, while those for gneiss are listed in Table C.3.

Statistics of the conductivity values from the sites listed in Tables C.2 and C.3 are used to establish the priors of the effective conductivity. Our general approach for modeling the conductivity of the granites is to homogenize the conductivity field and to represent it using effective conductivities. Homogenization of the conductivity field is a viable strategy for modeling the effects of spatial variability when dealing with average fluxes. In this study we are interested in modeling large-scale average fluxes, and such fluxes can be modeled using effective conductivities as has been shown in Dagan (1989) and Rubin (2003).

There are several models for the effective conductivity reported in the literature, and in general they are associated with assumptions regarding the flow regime (2D, 3D, uniform in the average, radial flow, etc., e.g., Binley et al., 1989; Dykaar and Kitanidis, 1992; Indelman and Dagan, 1993) and the statistical distribution of the conductivity (e.g., normal, log-normal, bimodal, etc.). As it is difficult to associate the Mingtang site flow regime with any

Table C.2: List of similar sites for the ex-situ prior of the hydraulic conductivity of granite.

| **Study** | **Site** | **Measurement depth (m)** | **Measurement Value (m/s)** | **Sampling method and support volume dimension** |
|---|---|---|---|---|
| Martinez-Landa and Carrera (2005) | Grimsel, Switzerland | 450<br>450 | 5E-7<br>5E-7 | Borehole pulse tests, 1 - 10 m |
| Zhao (1998) | Singapore | 40 - 70<br>60<br>120 | 5E-7<br>5E-11<br>1R-9 | Borehole packer tests, 3 m |
| Heath (1985) | Carnmenellis, England | 525<br>580 | 1E-7<br>1E-9 | Borehole packer tests, 1 - 7 m |
| Magnusson and Duran (1984) | Krakemala, Sweden | 330<br>140 | 1E-5<br>5E-9 | Borehole packer tests, 2 m |
| Lee and Kim (2015) | Daejeon, Korea | 1 - 4<br>4 - 7 | 3.1E-11<br>6E-7 | Borehole packer tests, 3 m |
| Mejías et al. (2009) | Cadalso de losVidrios, Spain | 190<br>161 | 2E-8<br>8E-11 | Borehole packer tests, 1 - 6 m |
| Stober (1996) | Black Forest, Germany | 100<br>100 | 7E-6<br>2E-9 | Borehole pumping tests |
| Farvolden et al. (1988) | Atikokan, Canada | 100<br>300 | 1E-7<br>1E-12 | Borehole lab tests, <1 m |
| Kun (2012) | Shandong, China | 200 | 1.1E-8 | Borehole injection tests |
| Yin et al. (2005) | Liaoning, China | 317 | 5.2E-8 | Borehole injection tests |

Table C.3: List of similar sites for the ex-situ prior of the hydraulic conductivity of gneiss.

| **Study** | **Site** | **Measurement depth (m)** | **Measurement Value (m/s)** | **Sampling method and support volume dimension** |
|---|---|---|---|---|
| Magnusson and Duran (1984) | Krakemala, Sweden | 330<br>140 | 1E-5<br>5E-9 | Borehole packer tests, 2 m |
| Stober (1996) | Black Forest, Germany | 100<br>100 | 7E-6<br>2E-9 | Borehole pumping tests |
| Äikäs et al. (2000) | Olkiluoto, Sweden | 210<br>300 | 4E-6<br>1.5E-8 | Borehole lab tests, 2 - 5 m |
| Farvolden et al. (1988) | Atikokan, Canada | 100<br>300 | 1E-7<br>1E-7 | Borehole lab tests, <1 m |
| Ma (2010) | Shandong, China | 200 | 3.4E-8 | Borehole packer tests, 2 - 5 m |
| Guo (2005) | Hebai, China | 110 | 1E-7 | Borehole pumping tests |
| Cui (2010) | Liaoning, China | 25 | 1.3E-6 | Borehole packer tests, 5 - 7 m |
| Chen (2013) | Shanxi, China | 18 | 8.3E-8 | Borehole pumping tests |

set of assumptions that would allow us to select a tightly-defined model, we must resort to the least restrictive (and hence the most general) model (Rubin, 2003), which states that the effective conductivity is bounded between the arithmetic and harmonic means of the conductivity, depending neither on the flow regime nor on the shape of the distribution of the conductivity. For this model to be valid, the following conditions must be met. First, the flow domain needs to be stationary. As shown in Tables C.2 and C.3, the conductivity decreases with depth, and so we limited the assumption of stationarity for fixed depth intervals, with the bounds computed based on conductivity measurements of equal depth. Second, the domain must be larger than the correlation length. Review of multiple studies (Erhel et al., 2009; Follin, 1990; Jaquet et al., 2013; Poteri et al., 2002; Sawada et al., 2015; Walker et al., 2005; Wen and Kung, 1993) indicates that the horizontal correlation lengths in granite rocks vary from 1 to 300 m, with most domains showing correlation lengths between 20 to 50 m. These length scales are smaller than the dimensions of the Mingtang flow domain, and at the same time are larger than the measurement scale, which is typically of the order of 1 to 7m (see Tables C.2 and C.3).

To summarize, our approach for selecting values of the effective conductivities follows these steps:

1. Define an ensemble of similar sites (following intrinsic and epistemic screening criteria).

2. Generate a scatterplot of the conductivities from the comparable sites while avoiding bias due to disparities in the number of samples between sites.

3. Compute the upper and lower bounds on the effective conductivity at each depth interval. With $n$ values of the conductivity within a depth interval around a nodal depth, $K_i, i = 1, ..., n$, the arithmetic and harmonic mean values ($\bar{K}_A$, $\bar{K}_H$) of that nodal depth interval are given by the following definitions (Dagan, 1989; Dykaar and Kitanidis, 1992; Indelman and Dagan, 1993; Rubin, 2003):

$$\bar{K}_A = \frac{1}{n} \sum_{i-1}^{n} K_i, \tag{C.1}$$

$$\bar{K}_H = \left( \frac{1}{n} \sum_{i-1}^{n} K_i^{-1} \right). \tag{C.2}$$

4. Perform this analysis sequentially for different depth intervals.

5. Model the effective conductivity as a random variable, distributed uniformly between the bounds.

6. Generate values for the vertical profiles of the effective conductivity to be used in the numerical model. Each profile defines a realization of the conductivity field. The ensemble of realizations is then used to generate the statistical distributions of the dependent variables.

The scatterplot of the log-conductivity of granite as a function of depth below ground surface with data drawn from multiple, similar sites is shown in Figure C.1**(a)**. The arithmetic, geometric, and harmonic mean values of data at each nodal depth are shown in Figure C.1**(b)**, with the arithmetic and harmonic means constituting the upper and lower bounds on the hydraulic conductivity, respectively (Rubin, 2003). The bounds are found to decrease exponentially with depth, which is consistent with previous studies (Bense et al., 2013). The bounds, as represented by the regression lines, form the upper and lower bounds on the effective conductivity. As shown in Figure C.1, the two Mingtang conductivity values lie next to the computed upper bounds (as defined by the regression curves). These conductivity values were measured near the faults' damage zones, which are known to be characterized by relatively high conductivity values. Indeed these values are in better agreement with the prior information on damage zone conductivity as seen later in Figure C.1. Similarly,

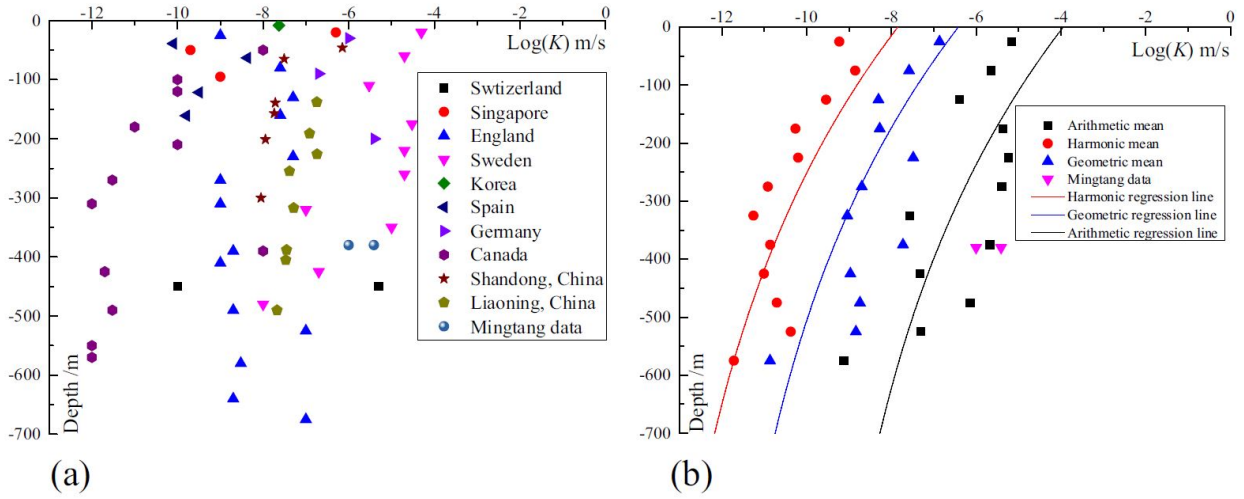


Figure C.1: Prior of hydraulic conductivity for granite. (a) Scatterplot of hydraulic conductivities from the sites listed in Table C.2; (b) Effective conductivity bounds. In both cases, decimal log scale used.

the scatterplot of log-conductivity and the regression curves for the bounds on the effective conductivity as well as the geometric mean are shown in Figure C.2**(a)** and **(b)**, respectively. The Mingtang measurements are again closer to the upper bound, and scattered around the regression curve.

**The Ex-situ Prior of Fault Zone Hydraulic Conductivity**

The hydraulic conductivities of the fault zones could be very distinct from those of the fractured rocks. In general, average hydraulic conductivities are often assigned to fault zones, which could be based on scaling factors from neighboring rock stratum (Yang et al., 2009), reproduction of model parameters that fit observed water tables (Raposo et al., 2010), derivation from an average of nearby borehole measurements (Molinero et al., 2002),

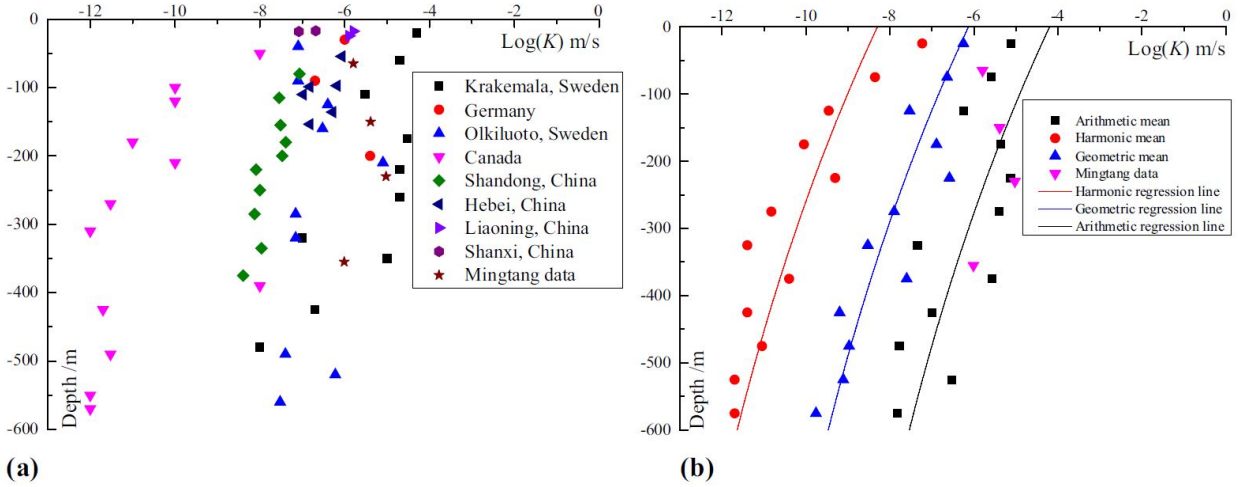


Figure C.2: Prior of hydraulic conductivity for gneiss. (a) Scatterplot of hydraulic conductivities from the sites listed in Table C.3; (b) Effective conductivity bounds. In both cases, decimal log scale used.

or statistical simulations assuming a representative average fault conductivity (Kitterød et al., 2000). However, Bense et al. (2013) showed that this is not the case, instead depicting the fault zone of crystalline rocks as a combination of highly conductive outer damage zones enveloping a barely conductive inner fault core. These fault zones are thus highly anisotropic, in that the hydraulic conductivities in the damage zone have been reported in many studies to be several orders of magnitude higher than those near to the fault core. In this manner, the damage zones act as conduits for flow parallel to the fault plane, whereas the fault core acts as a barrier to flow normal to the fault plane. While the aforementioned core-damage-zone framework may assume that several fault zone processes reducing fault permeability are negligible (e.g., dissolution and cementation in fracture networks), we consider this to be reasonable in light of the large water influx quantities observed at fault intersections during the Mingtang tunnel excavation. At the Mingtang site, two faults were recognized to be intersecting the tunnel, as described in Section 3.3. The outer damage zone has an approximate width (normal to the fault plane) of 5–6m on either side of the 30 cm fault core. A similar approach as the one applied for the conductivity of fractured rock is also applied here for the fault zones, but two adjustments are made. First, due to less available information about fault zones at the Mingtang site, among all criteria listed in Table C.1 only the geological structure similarity criterion is adopted. Second, instead of 50 m, the depth interval at which mean values are computed is set as 100 m, accounting for the reduction in available observations at similar sites. This is expected, since boreholes only intersect fault-affected areas at certain depths depending on a variety of factors including the width of the fault architecture, fault inclination, etc.

Geologically similar sites for the ex-situ prior of the hydraulic conductivity of fault zones

Table C.4: List of similar sites for the ex-situ prior of the hydraulic conductivity of fault zones.

| Study | Site | Rock type |
|---|---|---|
| Evans et al. (1997) | Wyoming, U.S.A. | Granite/gneiss |
| Gleeson and Novakowski (2009) | Ontario, Canada | Granite/gneiss |
| Morrow and Byerlee (1992) | California, U.S.A. | Granite/gneiss |
| Farvolden and Fritz (1988) | Ontario, Canada | Granite/gneiss |
| Stober, 1996 | Black Forest, Germany | Granite/gneiss |
| Rutqvist (2015) | Gidea, Sweden | Crystalline rock (assumed Granite/gneiss) |
| Géraud et al. (2010) | Soultz-sous-Forets, France | Granite |
| Lockner et al. (2009) | Nojima, Japan | Granite |
| Mejías et al. (2009) | Cadalso de losVidrios, Spain | Granite |
| Morrow and Lockner (1997) | Illinois, U.S.A. | Granite |

are listed in Table C.4, including information on whether the fault zones are embedded in granite or gneiss, or a combination of both. Note that there is no information exclusively for faults embedded in gneiss, with all compiled sites having either granite or a combination of granite and gneiss as the dominant rock type. In reality, granite and gneiss rocks can either be fused together in formation, or separated in alternating bands. However, it is unclear if most conductivity measurements arise from granite or gneiss rocks exclusively. To that end, an assumption is made that the ex-situ prior for faults embedded in gneiss can be established by conditioning on the measurements from sites characterized by a combination of granite and gneiss. The ex-situ prior for faults embedded in granite, on the other hand, can be established by considering the measurements from all the sites listed in Table C.4.

Following the same steps, the prior uncertainty bounds of can be established, as shown in Figure C.3. The difference between the hydraulic conductivities of damage zones and of fault cores is about 3 orders of magnitude, which is actually quite consistent with the range reported in the literature of ex-situ priors. The hydraulic conductivities of fault zones, depending on whether they are damage zones or fault cores or are embedded in granite or gneiss, are thus bounded by the upper and lower bounds shown in Figure C.3.

**The Ex-situ Prior of Average Groundwater Recharge**

In this study the average annual groundwater recharge ($R$) is modeled by $R = C_r P$, whereP is the average annual precipitation and $C_r$ is the recharge-precipitation ratio. The model stems from the concept of water budget under the assumption that hydrologic processes are distributing the total precipitation into the remaining hydrologic responses, which include groundwater recharge. It is used quite commonly (e.g., Heppner et al., 2007; Magruder et al., 2009; Obuobie et al., 2012; Rangarajan and Athavale, 2000; Vries and Simmers, 2002). We shall adopt the model as a working hypothesis, without claiming optimality. The estimation of recharge now becomes the estimation of Cr. The same approach is applied here as the one applied for establishing ex-situ priors for the conductivities, but with a different set of similarity criteria. It would be ideal if we could consider watershed topography,

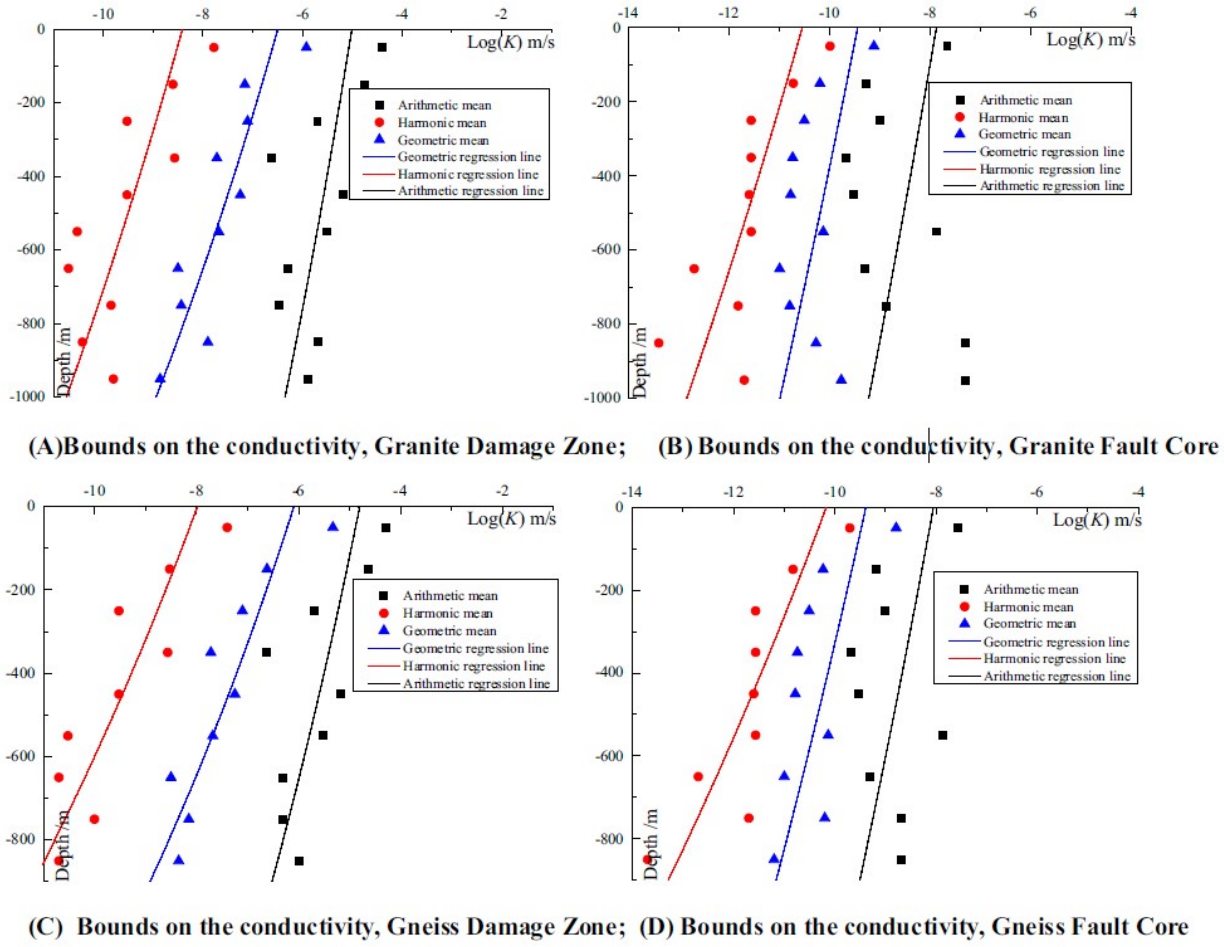


Figure C.3: Prior of hydraulic conductivity for (A) damage zone embedded in granite, (B) fault core embedded in granite, (C) damage zone embedded in gneiss, and (D) fault core embedded in gneiss. In all cases, decimal log scale used.

climate, land cover, geological structure, and the existence of preferential flow path for the intrinsic similarity, and consider the sampling method as well as the spatiotemporal scale for epistemic similarity. Due to limited documentation at the Mingtang site as well as the potentially similar sites collected in this study, only a subset of three intrinsic similarity criteria is considered in this study: rock type, precipitation characteristic, and the dominant land cover, and they represent similarities in geology, climate, and land cover, respectively.

The three selected criteria are applied qualitatively by direct comparison with the Mingtang site. The dominant rock types at the Mingtang site are granite and gneiss. Satellite images and land use maps show that the Mingtang site is highly forested. The annual precipitation pattern over the Mingtang site is characterized by prominent wet-dry seasons; roughly 45% of rainfall occurs from June to August and roughly 16% occurs from November

to February; furthermore, snow/ice melt contribution is negligible. A number of studies satisfying at least one of the aforementioned criteria are listed in Table C.5, as well as their corresponding $C_r$ values. The histograms and the estimated ex-situ priors (by kernel density estimation) of $C_r$ are shown in Figure C.4.

Table C.5: List of similar sites for the ex-situ prior of mean annual groundwater recharge. Grey-colored cell indicates that it is considered a similar condition as that at the Mingtang site.

| Study | Site | Rock type | Dominant land cover | Precipitation characteristic | Average $C_r$ (mm/mm) |
|---|---|---|---|---|---|
| Rangarajan et al. (2009) | Tamil Nadu, India | Half gneiss half alluvium | Cultivation, shrub, and grass | Monsoon from Jun to Sep | 0.106 |
| Chand et al. (2004) | Bairasagara, India | Granite and schist | Cultivation | Monsoon | 0.07 |
| Chand et al. (2005) | Kongal river, India | Pink granite, weathered granite and gneiss | | Monsoon | 0.044 |
| Rangarajan and Athavale (2000) | Upper Hatni, India | Archaean granites, quartzite and phyllite, Cretaceous sandstone, and recent alluvium. | | Monsoon | 0.1 |
| Ruiz et al. (2010) | Mule Hole, India | Precambrian peninsular gneiss intermingled with mafic and ultramafic rocks | Dry deciduous forest | Bimodal due to S-W monsoon and N-E monsoon | 0.176 |
| Ofterdinger et al. (2014) | Western Gotthard Massif of the Swiss Alps | Late Hercynian Rotondo granite and orthogneisses of early Palaeozoic age | | Significant glacier melt contribution | 0.19 |
| Raposo et al. (2010) | Ferrol, Spain | Hercynian post-tectonic granites | City | Concentrated precipitation in Fall and Winter | 0.136 |
| Massuel et al. (2007) | Musi, India | Archaean granites with overlaying Deccan Traps | 70 % cultivation | Monsoon Jun to Oct | 0.094 |
| Obuobie et al. (2012) | White Volta River, Ghana | Precambrian granite-gneiss greenstone rocks, deformed metamorphic rocks and anorogenic intrusions, shale and sandstone | Savannah | Wet season May to Oct | 0.1 |
| Yang et al. (2009) | Tseng-Wen Reservoir basin, Taiwan | Miocene sedimentary rock with layers of sandstone or shale or their alternations | Forest | Wet season May to Sep | 0.24 |
| Mondal et al. (2012) | Near Dingdigul town, India | Achaean granites and gneisses, intruded by dykes | | S-W monsoon and N-E monsoon | 0.064 |
| Mogaji et al. (2015) | Perak and Selango , Peninsular Malaysia | Recent alluvium, sedimentary rocks, acidic and undifferentiated granitoid | | Equatorial maritime climate | 0.2 |
| Katsuyama et al. (2010) | Kiryu watershed, Japan | Cretaceous biotite granite (also called Tanakami granite) | Japanese cypress | Wet season Jun to Oct | 0.17 |
| Takagi (2013) | Miyazaki University Forest, Japan | Shale | Planted Japanese cypress and natural broad-leaf trees | Wet season Jun to Oct | 0.25 |
| Marécchal et al. (2006) | Maheshwaram pilot watershed, India | Archean granite | Agriculture | Monsoon Jun to Oct | 0.145 |

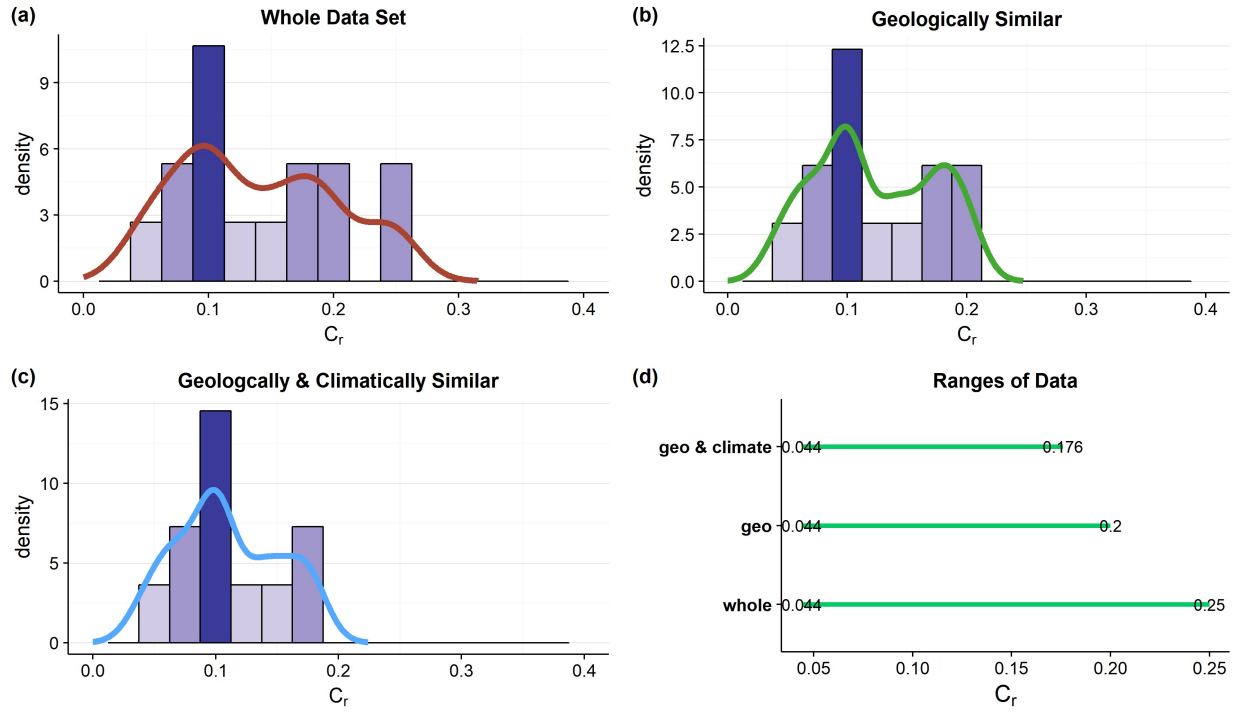


Figure C.4: Histograms and priors from the sites listed in Table 5 based on (a) the whole data set, (b) the subset of geologically similar sites, (c) the subset of geologically and climatically similar sites, and (d) the ranges of each subset (the green lines spans the interval between maximum and minimum values). The color of the histogram bars is scaled by value: the darker the larger.

# Bibliography

Äikäs, Kari et al. (2000). "Engineering rock mass classification of the Olkiluoto investigation site". In: *Posiva Oy, report POSIVA* 8.

Ajami, Newsha K., Qingyun Duan, and Soroosh Sorooshian (2007). "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction". In: *Water Resources Research* 43.1. DOI: doi:10.1029/2005WR004745.

Anan, Mitsumasa et al. (2007). "Quantification of the effect of rice paddy area changes on recharging groundwater". In: *Paddy and Water Environment* 5.1, pp. 41–47. DOI: 10.1007/s10333-006-0059-1.

Aquaveo (May 2013). "GMS User Manual (v9.1) The Groundwater Modeling System". In:

Arnold, J. G. et al. (2000). "Regional estimation of base flow and groundwater recharge in the Upper Mississippi river basin". In: *Journal of Hydrology* 227.1–4, pp. 21–40. DOI: http://dx.doi.org/10.1016/S0022-1694(99)00139-0.

Bakker, M. R., L. Augusto, and D. L. Achat (2006). "Fine root distribution of trees and understory in mature stands of maritime pine (Pinus pinaster) on dry and humid sites". In: *Plant and Soil* 286.1, pp. 37–51. DOI: 10.1007/s11104-006-9024-4.

Barros, F. P. J. de and Y. Rubin (2008). "A risk-driven approach for subsurface site characterization". In: *Water Resources Research* 44.1, n/a–n/a. DOI: 10.1029/2007WR006081.

Barros, Felipe P. J. de, Souheil Ezzedine, and Yoram Rubin (2012). "Impact of hydrogeological data on measures of uncertainty, site characterization and environmental performance metrics". In: *Advances in Water Resources* 36, pp. 51–63. DOI: http://dx.doi.org/10.1016/j.advwatres.2011.05.004.

Bellin, Alberto and Yoram Rubin (1996). "HYDRO_GEN: A spatially distributed random field generator for correlated properties". In: *Stochastic Hydrology and Hydraulics* 10.4, pp. 253–278. DOI: 10.1007/BF01581869.

Bense, V. F. et al. (2013). "Fault zone hydrogeology". In: *Earth-Science Reviews* 127, pp. 171–192. DOI: http://dx.doi.org/10.1016/j.earscirev.2013.09.008.

Beven, Keith (2006). "A manifesto for the equifinality thesis". In: *Journal of Hydrology* 320.1, pp. 18–36. DOI: https://doi.org/10.1016/j.jhydrol.2005.07.007.

— (2016). "Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication". In: *Hydrological Sciences Journal* 61.9, pp. 1652–1665. DOI: 10.1080/02626667.2015.1031761.

Beven, Keith and Andrew Binley (1992). "The future of distributed models: model calibration and uncertainty prediction". In: *Hydrological processes* 6.3, pp. 279–298.

— (2014). "GLUE: 20 years on". In: *Hydrological Processes* 28.24, pp. 5897–5918. DOI: doi:10.1002/hyp.10082.

Beven, Keith and Jim Freer (2001). "Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology". In: *Journal of Hydrology* 249.1-4, pp. 11–29. DOI: 10.1016/s0022-1694(01)00421-8.

Beven, Keith J., Paul J. Smith, and Jim E. Freer (2008). "So just why would a modeller choose to be incoherent?" In: *Journal of Hydrology* 354.1, pp. 15–32. DOI: https://doi.org/10.1016/j.jhydrol.2008.02.007.

Binley, Andrew, Keith Beven, and John Elgy (1989). "A physically based model of heterogeneous hillslopes: 2. Effective hydraulic conductivities". In: *Water Resources Research* 25.6, pp. 1227–1233. DOI: 10.1029/WR025i006p01227.

Bleines, C et al. (2004). *ISATIS software manual*.

Blöschl, Günter et al. (2013). "Runoff prediction in ungauged basins: synthesis across processes, places and scales". In:

Brakebill, J. W. and S. E. Terziotti (2011). "A Digital Hydrologic Network Supporting NAWQA MRB SPARROW Modeling - MRB_E2RF1WS". In:

Breiman, Leo (1984). *Classification and regression trees*. ISBN: 1351460498.

Canadell, J. et al. (1996). "Maximum rooting depth of vegetation types at the global scale". In: *Oecologia* 108.4, pp. 583–595. DOI: 10.1007/bf00329030.

Carle, Steven F (1999). "T-PROGS: Transition probability geostatistical software". In: *University of California, Davis, CA* 84.

Chand, R. et al. (2004). "Estimation of natural recharge and its dependency on sub-surface geoelectric parameters". In: *Journal of Hydrology* 299.1–2, pp. 67–83. DOI: http://dx.doi.org/10.1016/j.jhydrol.2004.04.001.

Chand, Ramesh et al. (2005). "Reliable natural recharge estimates in granitic terrain". In: *Current science* 88.5, pp. 821–824.

Chen, Changqing (2013). "Geological Evaluation of Wenxi-Jiyuan ZhingTiaoShan Highway Tunnel Project". In: *Engineering Geology* Master.

Chen, Jianqin et al. (2017). "Geostatistical method for inferring RMR ahead of tunnel face excavation using dynamically exposed geological information". In: *Engineering Geology* 228, pp. 214–223. DOI: 10.1016/j.enggeo.2017.08.004.

Chen, Zhuoheng, Stephen E. Grasby, and Kirk G. Osadetz (2002). "Predicting average annual groundwater levels from climatic variables: an empirical model". In: *Journal of Hydrology* 260.1–4, pp. 102–117. DOI: http://dx.doi.org/10.1016/S0022-1694(01)00606-0.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch (2010). "BART: Bayesian additive regression trees". In: *Ann. Appl. Stat.* 4.1, pp. 266–298. DOI: 10.1214/09-AOAS285.

Cirpka, Olaf A. and Albert J. Valocchi (2016). "Debates: Stochastic subsurface hydrology from theory to practice—Does stochastic subsurface hydrology help solving practical problems of contaminant hydrogeology?" In: *Water Resources Research*, n/a–n/a. DOI: 10.1002/2016WR019087.

Clawges, Rick M and Curtis V Price (1999). "Digital data set describing surficial geology in the conterminous US". In: *US Geological Survey Open-File Report* 99, p. 77.

Comunian, Alessandro and Philippe Renard (2008). "Introducing wwhypda: a world-wide collaborative hydrogeological parameters database". In: *Hydrogeology Journal* 17.2, pp. 481–489. DOI: 10.1007/s10040-008-0387-x.

Csilléry, Katalin, Olivier François, and Michael G. B. Blum (2012). "abc: an R package for approximate Bayesian computation (ABC)". In: *Methods in Ecology and Evolution* 3.3, pp. 475–479. DOI: 10.1111/j.2041-210X.2011.00179.x.

Cucchi, Karina et al. (2019). "Ex-situ priors: A Bayesian hierarchical framework for defining informative prior distributions in hydrogeology". In: *Advances in Water Resources* 126, pp. 65–78. DOI: 10.1016/j.advwatres.2019.02.003.

Cui, Yang (2010). "Plan to Tunnel Construction Project to Complicated Geologic Tectonic Zone and Measuring Evaluation of Linear Wall". In: *Road and Railway Engineering* Master of Engineering.

Dagan, Gedeon (1989). *Flow and transport in porous formations*. ISBN: 364275015X.

De Barros, Felipe P. J. and Yoram Rubin (2011). "Modelling of block-scale macrodispersion as a random function". In: *Journal of Fluid Mechanics* 676, pp. 514–545. DOI: 10.1017/jfm.2011.65.

Deutsch, Clayton V. and Andre G. Journel (1998). *GSLIB : Geostatistical Software Library and User's Guide (Applied Geostatistics Series)*, pp. 1–369. ISBN: 0-19-510015-8.

Doherty, Je and Rj Hunt (2010). "Approaches to highly parameterized inversion: a guide to using PEST for groundwater-model calibration". In: *U. S. Geological Survey Scientific Investigations Report 2010-5169*, pp. 70–70. DOI: 2010-5211.

Doherty, John (2005). "PEST User Manual, 5th Edition". In:

Domec, Jean-Christophe et al. (2010). "Hydraulic redistribution of soil water by roots affects whole-stand evapotranspiration and net ecosystem carbon exchange". In: *New Phytologist* 187.1, pp. 171–183. DOI: 10.1111/j.1469-8137.2010.03245.x.

Dunning, CP et al. (2004). "Simulation of ground-water flow, surface-eater flow, and a deep sewer tunnel system in the Menomonee Valley, Milwaukee, Wisconsin". In: *US Geol Surv Sci Invest Rep* 5215, pp. 2004–5031.

Dykaar, Bruce B. and Peter K. Kitanidis (1992). "Determination of the effective hydraulic conductivity for heterogeneous porous media using a numerical spectral approach: 2. Results". In: *Water Resources Research* 28.4, pp. 1167–1178. DOI: 10.1029/91WR03083.

Eamus, Derek. et al. (2006). "A functional methodology for determining the groundwater regime needed to maintain the health of groundwater-dependent vegetation". In: *Australian Journal of Botany* 54.2, pp. 97–114. DOI: http://dx.doi.org/10.1071/BT05031.

Erhel, J. et al. (2009). "A parallel scientific software for heterogeneous hydrogeoloy". In: pp. 39–48. ISBN: 978-3-540-92744-0. DOI: 10.1007/978-3-540-92744-0_5.

Evans, James P., Craig B. Forster, and James V. Goddard (1997). "Permeability of fault-related rocks, and implications for hydraulic structure of fault zones". In: *Journal of Structural Geology* 19.11, pp. 1393–1404. DOI: http://dx.doi.org/10.1016/S0191-8141(97)00057-6.

Fan, Y., H. Li, and G. Miguez-Macho (2013). "Global Patterns of Groundwater Table Depth". In: *Science* 339.6122, pp. 940–943. DOI: 10.1126/science.1229881.

Farvolden, R. N. et al. (1988). "Region 12, Precambrian Shield". In: *Hydrogeology*. Geological Society of America, pp. 101–114. DOI: 10.1130/dnag-gna-o2.101.

Farvolden, RN and P Fritz (1988). "Region 12, Precambrian Shiela". In:

Fetter, C.W. (2001). *Applied Hydrogeology*. ISBN: 978-0130882394.

Finsterle, Stefan (2011). "iTOUGH2 User's Guide". In: *LBNL* 123.6, pp. 407–8. DOI: 10.1111/j.1600-0447.2011.01711.x.

Follin, S (1990). "Effective Hydraulic Conductivity for Fractured Media". In: *IN: MODELCARE 90: Calibration and Reliability in Groundwater Modelling. IAHS Publication* 195.

Garthwaite, Paul H, Joseph B Kadane, and Anthony O'Hagan (2005). "Statistical Methods for Eliciting Probability Distributions". In: *Journal of the American Statistical Association* 100.470, pp. 680–701. DOI: 10.1198/016214505000000105.

Gelman, Andrew et al. (2014). *Bayesian data analysis*. Vol. 2.

Gemitzi, Alexandra, Hoori Ajami, and Hans-Hermann Richnow (2017). "Developing empirical monthly groundwater recharge equations based on modeling and remote sensing data – Modeling future groundwater recharge to predict potential climate change impacts". In: *Journal of Hydrology* 546, pp. 1–13. DOI: https://doi.org/10.1016/j.jhydrol.2017.01.005.

Géraud, Yves et al. (2010). "Physical properties of fault zones within a granite body: Example of the Soultz-sous-Forêts geothermal site". In: *Comptes Rendus Geoscience* 342.7–8, pp. 566–574. DOI: http://dx.doi.org/10.1016/j.crte.2010.02.002.

Gibbs, Matt. S., Holger. Robert. Maier, and Graeme. Clyde. Dandy (2012). "A generic framework for regression regionalization in ungauged catchments". In: *Environmental Modelling & Software* 27–28, pp. 1–14. DOI: http://dx.doi.org/10.1016/j.envsoft.2011.10.006.

Gilbert, J. M. and R. M. Maxwell (2017). "Examining regional groundwater–surface water dynamics using an integrated hydrologic model of the San Joaquin River basin". In: *Hydrol. Earth Syst. Sci.* 21.2. HESS http://www.hydrol-earth-syst-sci.net/21/923/2017/hess-21-923-2017.pdf, pp. 923–947. DOI: 10.5194/hess-21-923-2017.

Gleeson, Tom and Kent Novakowski (2009). "Identifying watershed-scale barriers to groundwater flow: Lineaments in the Canadian Shield". In: *Geological Society of America Bulletin* 121.3-4, pp. 333–347. DOI: 10.1130/b26241.1.

Gómez-Hernández, J. Jaime and R. M. Srivastava (1990). "Isim3D - an Ansi-C 3-Dimensional Multiple Indicator Conditional Simulation Program". In: *Computers & Geosciences* 16.4, pp. 395–440.

Greppi, Mauro (2004). "Infiltration process and groundwater table raising in a paddy field area". In: *Paddy and Water Environment* 2.3, pp. 171–179. DOI: 10.1007/s10333-004-0056-1.

Guo, Wenya (2005). "The Study on The Forecast of Water Inflow during Turning Open Pit into Underground Mining in ShiRengou Iron Mine". In: *Resource and Environmental School* Master of Engineering.

Harbaugh, Arlen W. (2005). *MODFLOW-2005 : the U.S. Geological Survey modular groundwater model–the ground-water flow process*. DOI: 10.3133/tm6a16.

Harken, Bradley (2017). "Hydrogeological Modeling and Water Resources Management: Improving the Link Between Site Characterization, Prediction, and Decision Making". PhD thesis. University of California, Berkeley.

Harken, Bradley et al. (2019). "Hydrogeological Modeling and Water Resources Management: Improving the Link Between Data, Prediction, and Decision-Making". In: *Manuscript submitted to Water Resources Research*.

Hartmann, Andreas et al. (2017). "Enhanced groundwater recharge rates and altered recharge sensitivity to climate variability through subsurface heterogeneity". In: *Proceedings of the National Academy of Sciences* 114.11, pp. 2842–2847.

Healy, Richard W (2010). *Estimating groundwater recharge*. ISBN: 1139491393.

Heath, MJ (1985). "Geological control of fracture permeability in the Carnmenellis granite, Cornwall: implications for radionuclide migration". In: *Mineralogical Magazine* 49.2, pp. 233–244.

Heppner, Christopher S. et al. (2007). "Multiple-methods investigation of recharge at a humid-region fractured rock site, Pennsylvania, USA". In: *Hydrogeology Journal* 15.5, pp. 915–927. DOI: 10.1007/s10040-006-0149-6.

Homer, Collin et al. (2007). "Completion of the 2001 national land cover database for the counterminous United States". In: *Photogrammetric engineering and remote sensing* 73.4, p. 337.

Hou, Zhangshuan and Yoram Rubin (2005). "On minimum relative entropy concepts and prior compatibility issues in vadose zone inverse and forward modeling". In: *Water Resources Research* 41.12, n/a–n/a. DOI: 10.1029/2005WR004082.

Hrachowitz, M. et al. (2013). "A decade of Predictions in Ungauged Basins (PUB)—a review". In: *Hydrological Sciences Journal* 58.6, pp. 1198–1255. DOI: 10.1080/02626667.2013.803183.

Hudson, Berman D. (1994). "Soil organic matter and available water capacity". In: *Journal of Soil and Water Conservation* 49.2, pp. 189–194.

Hutton, C. J. et al. (2014). "Application of Formal and Informal Bayesian Methods for Water Distribution Hydraulic Model Calibration". In: *Journal of Water Resources Planning and Management* 140.11, p. 04014030. DOI: 10.1061/(ASCE)WR.1943-5452.0000412.

Ibrahim, Agung B. and I. A. N. Cordery (1995). "Estimation of recharge and runoff volumes from ungauged catchments in eastern Australia". In: *Hydrological Sciences Journal* 40.4, pp. 499–515. DOI: 10.1080/02626669509491435.

Imaizumi, Masayuki, Satoshi Ishida, and Takeo Tuchihara (2006). "Long-term evaluation of the groundwater recharge function of paddy fields accompanying urbanization in the Nobi Plain, Japan". In: *Paddy and Water Environment* 4.4, pp. 251–263. DOI: 10.1007/s10333-006-0056-4.

Indelman, P. and G. Dagan (1993). "Upscaling of permeability of anisotropic heterogeneous formations: 1. The general framework". In: *Water Resources Research* 29.4, pp. 917–923. DOI: 10.1029/92WR02446.

Jaquet, Oliver et al. (2013). *Groundwater flow modelling under ice sheet conditions in Greenland (phase II).*

Kapelner, Adam and Justin Bleich (2016). "bartMachine: Machine Learning with Bayesian Additive Regression Trees". In: *2016* 70.4, p. 40. DOI: 10.18637/jss.v070.i04.

Katsuyama, M., M. Tani, and S. Nishimoto (2010). "Connection between streamwater mean residence time and bedrock groundwater recharge/discharge dynamics in weathered granite catchments". In: *Hydrological Processes* 24.16, pp. 2287–2299. DOI: 10.1002/hyp.7741.

Kim, Nam Won et al. (2008). "Development and application of the integrated SWAT–MODFLOW model". In: *Journal of Hydrology* 356.1–2, pp. 1–16. DOI: http://dx.doi.org/10.1016/j.jhydrol.2008.02.024.

Kitterød, N.-O. et al. (2000). "Simulation of groundwater drainage into a tunnel in fractured rock and numerical analysis of leakage remediation, Romeriksporten tunnel, Norway". In: *Hydrogeology Journal* 8.5, pp. 480–493. DOI: 10.1007/s100400000089.

Kuczera, George (1982). "Combining site-specific and regional information: An empirical Bayes Approach". In: *Water Resources Research* 18.2, pp. 306–314. DOI: 10.1029/WR018i002p00306.

Kuentz, A. et al. (2017). "Understanding hydrologic variability across Europe through catchment classification". In: *Hydrol. Earth Syst. Sci.* 21.6. HESS https://www.hydrol-earth-syst-sci.net/21/2863/2017/hess-21-2863-2017.pdf, pp. 2863–2879. DOI: 10.5194/hess-21-2863-2017.

Kun, Song (2012). "The permeability of granitic gneiss and the stability of water sealed tank well rock". In: *Geotechnical engineering* Ph.D.

Laiho, Raija and Leena Finér (1996). "Changes in root biomass after water-level drawdown on pine mires in southern Finland". In: *Scandinavian Journal of Forest Research* 11.1-4, pp. 251–260. DOI: 10.1080/02827589609382934.

Lee, Hang Bok and Byung-Woo Kim (2015). "Characterisation of hydraulically-active fractures in a fractured granite aquifer". In: *Water SA* 41, pp. 139–148.

Li, Xiaojun, Jianqin Chen, and Hehua Zhu (2016). "A new method for automated discontinuity trace mapping on rock mass 3D surface model". In: *Computers & Geosciences* 89, pp. 118–131. DOI: http://dx.doi.org/10.1016/j.cageo.2015.12.010.

Li, Xiaojun et al. (2018). "Stochastic, goal-oriented rapid impact modeling of uncertainty and environmental impacts in poorly-sampled sites using ex-situ priors". In: *Advances in Water Resources* 111, pp. 174–191. DOI: https://doi.org/10.1016/j.advwatres.2017.11.008.

Liu, Chen-Wuing, Chih-Huang Tan, and Cheng-Chang Huang (2005). "Determination of the magnitudes and values for groundwater recharge from Taiwan's paddy field". In: *Paddy and Water Environment* 3.2, pp. 121–126. DOI: 10.1007/s10333-005-0005-7.

Lockner, David A. et al. (2009). "Geometry of the Nojima Fault at Nojima-Hirabayashi, Japan – I. A Simple Damage Structure Inferred from Borehole Core Permeability". In: *Pure and Applied Geophysics* 166.10, pp. 1649–1667. DOI: 10.1007/s00024-009-0515-0.

Loritz, Ralf et al. (2018). "On the dynamic nature of hydrological similarity". In: *Hydrology and Earth System Sciences Discussions* 2018, pp. 1–38.

Loukas, A. and L. Vasiliades (2014). "Streamflow simulation methods for ungauged and poorly gauged watersheds". In: *Nat. Hazards Earth Syst. Sci.* 14.7. NHESS http://www.nat-hazards-earth-syst-sci.net/14/1641/2014/nhess-14-1641-2014.pdf, pp. 1641–1661. DOI: 10.5194/nhess-14-1641-2014.

Ma, Feng (2010). "The Study of Fractured Rock Permeability in Huangdao Water Sealed Underground Oil Tank". In: *Groundwater Science and Engineering* Master.

Magnusson, Kurt-Åke and Oscar Duran (1984). *Comparative study of geological, hydrological and geophysical borehole investigations.*

Magruder, Ian A., William W. Woessner, and Steve W. Running (2009). "Ecohydrologic process modeling of mountain block groundwater recharge". In: *Groundwater* 47.6. DOI: 10.1111/j.1745-6584.2009.00615.x.

Marécchal, J. C. et al. (2006). "Combined estimation of specific yield and natural recharge in a semi-arid groundwater basin with irrigated agriculture". In: *Journal of Hydrology* 329.1–2, pp. 281–293. DOI: http://dx.doi.org/10.1016/j.jhydrol.2006.02.022.

Martinez-Landa, Lurdes and Jesús Carrera (2005). "An analysis of hydraulic conductivity scale effects in granite (Full-scale Engineered Barrier Experiment (FEBEX), Grimsel, Switzerland)". In: *Water Resources Research* 41.3, n/a–n/a. DOI: 10.1029/2004WR003458.

Massuel, Sylvain et al. (2007). "Groundwater Modeling for Sustainable Resource Management in the Musi Catchment, India". In: *Proceedings, International Congress on Modelling and Simulation, Christchurch, New Zealand* 10, pp. 1429–1435.

Maxwell, Reed M., William E. Kastenberg, and Yoram Rubin (1999). "A methodology to integrate site characterization information into groundwater-driven health risk assessment". In: *Water Resources Research* 35.9, pp. 2841–2855. DOI: 10.1029/1999WR900103.

Mejías, Miguel, Philippe Renard, and Damian Glenz (2009). "Hydraulic testing of low-permeability formations: A case study in the granite of Cadalso de los Vidrios, Spain". In: *Engineering Geology* 107.3–4, pp. 88–97. DOI: http://dx.doi.org/10.1016/j.enggeo.2009.05.010.

Mogaji, Kehinde Anthony, Hwee San Lim, and Khiruddin Abdullah (2015). "Modeling of groundwater recharge using a multiple linear regression (MLR) recharge model developed from geophysical parameters: a case of groundwater resources management". In: *Environmental Earth Sciences* 73.3, pp. 1217–1230. DOI: 10.1007/s12665-014-3476-2.

Molinero, Jorge, Javier Samper, and Rubén Juanes (2002). "Numerical modeling of the transient hydrogeological response produced by tunnel construction in fractured bedrocks".

In: *Engineering Geology* 64.4, pp. 369–386. DOI: http://dx.doi.org/10.1016/S0013-7952(01)00099-0.

Mondal, N. C., V. P. Singh, and S. Ahmed (2012). "Entropy-Based Approach for Assessing Natural Recharge in Unconfined Aquifers from Southern India". In: *Water Resources Management* 26.9, pp. 2715–2732. DOI: 10.1007/s11269-012-0042-0.

Morita, Shigenori, Tetsuya Suga, and Koou Yamazaki (1988). "The Relationship between Root length Density and Yield in Rice Plants". In: *Japanese journal of crop science* 57.3, pp. 438–443. DOI: 10.1626/jcs.57.438.

Morrow, C. A. and J. D. Byerlee (1992). "Permeability of core samples from Cajon Pass Scientific Drill Hole: Results from 2100 to 3500 m depth". In: *Journal of Geophysical Research: Solid Earth* 97.B4, pp. 5145–5151. DOI: 10.1029/90JB00423.

Morrow, C. A. and D. A. Lockner (1997). "Permeability and porosity of the Illinois UPH 3 drillhole granite and a comparison with other deep drillhole rocks". In: *Journal of Geophysical Research: Solid Earth* 102.B2, pp. 3067–3075. DOI: 10.1029/96JB03178.

Naghibi, Seyed Amir, Hamid Reza Pourghasemi, and Barnali Dixon (2015). "GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran". In: *Environmental Monitoring and Assessment* 188.1, p. 44. DOI: 10.1007/s10661-015-5049-6.

Nolan, Bernard T. et al. (2007). "Factors influencing groundwater recharge in the eastern United States". In: *Journal of Hydrology* 332.1, pp. 187–205. DOI: http://dx.doi.org/10.1016/j.jhydrol.2006.06.029.

Nott, David J., Lucy Marshall, and Jason Brown (2012). "Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection?" In: *Water Resources Research* 48.12. DOI: 10.1029/2011wr011128.

Nowak, W., F. P. J. de Barros, and Y. Rubin (2010). "Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain". In: *Water Resources Research* 46.3, n/a–n/a. DOI: 10.1029/2009WR008312.

Obuobie, Emmanuel et al. (2012). "Groundwater level monitoring and recharge estimation in the White Volta River basin of Ghana". In: *Journal of African Earth Sciences* 71–72, pp. 80–86. DOI: http://dx.doi.org/10.1016/j.jafrearsci.2012.06.005.

Ofterdinger, U. S., Ph Renard, and S. Loew (2014). "Hydraulic subsurface measurements and hydrodynamic modelling as indicators for groundwater flow systems in the Rotondo granite, Central Alps (Switzerland)". In: *Hydrological Processes* 28.2, pp. 255–278. DOI: 10.1002/hyp.9568.

Osorio-Murillo, Carlos A. et al. (2015). "Software framework for inverse modeling and uncertainty characterization". In: *Environmental Modelling and Software* 66, pp. 98–109. DOI: 10.1016/j.envsoft.2015.01.002.

Oudin, Ludovic et al. (2008). "Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments". In: *Water Resources Research* 44.3, n/a–n/a. DOI: 10.1029/2007WR006240.

Over, Matthew W. et al. (2015). "Bayesian inversion of Mualem-van Genuchten parameters in a multilayer soil profile: A data-driven, assumption-free likelihood function". In: *Water Resources Research* 51.2, pp. 861–884. DOI: 10.1002/2014WR015252.

Pebesma, Edzer J. (2004). "Multivariable geostatistics in S: The gstat package". In: *Computers and Geosciences* 30.7, pp. 683–691. DOI: 10.1016/j.cageo.2004.03.012.

Peñuelas, Josep and Iolanda Filella (2003). "Deuterium labelling of roots provides evidence of deep water access and hydraulic lift by Pinus nigra in a Mediterranean forest of NE Spain". In: *Environmental and Experimental Botany* 49.3, pp. 201–208. DOI: http://dx.doi.org/10.1016/S0098-8472(02)00070-9.

Poteri, Antti et al. (2002). "Final report of the TRUE Block Scale project. 3. Modelling of flow and transport". In:

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rahmati, Omid, Hamid Reza Pourghasemi, and Assefa M. Melesse (2016). "Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran". In: *CATENA* 137, pp. 360–372. DOI: http://dx.doi.org/10.1016/j.catena.2015.10.010.

Rangarajan, R. and R. N. Athavale (2000). "Annual replenishable ground water potential of India - an estimate based on injected tritium studies". In: *Journal of Hydrology* 234.1–2, pp. 38–53. DOI: http://dx.doi.org/10.1016/S0022-1694(00)00239-0.

Rangarajan, R et al. (2009). "Estimation of natural recharge and its relation with aquifer parameters in and around Tuticorin town, Tamil Nadu, India". In: *Current Science* 97.2, pp. 217–226.

Raposo, Juan Ramon, Jorge Molinero, and Jorge Dafonte (2010). "Quantitative evaluation of hydrogeological impact produced by tunnel construction using water balance models". In: *Engineering Geology* 116.3, pp. 323–332.

Razavi, Tara and Paulin Coulibaly (2017). "An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds". In: *Canadian Water Resources Journal / Revue canadienne des ressources hydriques* 42.1, pp. 2–20. DOI: 10.1080/07011784.2016.1184590.

Refsgaard, Jens Christian et al. (2007). "Uncertainty in the environmental modelling process – A framework and guidance". In: *Environmental Modelling & Software* 22.11, pp. 1543–1556. DOI: https://doi.org/10.1016/j.envsoft.2007.02.004.

Remy, Nicolas, Alexandre Boucher, and Jianbing Wu (2009). *Applied Geostatistics with SGeMS*. DOI: 10.1017/CBO9781139150019.

Rodhe, Allan and Niclas Bockgård (2006). "Groundwater recharge in a hard rock aquifer: A conceptual model including surface-loading effects". In: *Journal of Hydrology* 330.3, pp. 389 –401. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2006.03.032.

Rubin, Y. et al. (1999). "The concept of block-effective macrodispersivity and a unified approach for grid-scale- and plume-scale-dependent transport". In: *Journal of Fluid Mechanics* 395, pp. 161–180. DOI: 10.1017/S0022112099005868.

Rubin, Y. et al. (2018). "Stochastic Hydrogeology's Biggest Hurdles Analyzed and Its Big Blind Spot". In: *Hydrol. Earth Syst. Sci. Discuss.* 2018. HESSD, pp. 1–36. DOI: 10.5194/hess-2018-290.

Rubin, Yoram (2003). *Applied stochastic hydrogeology*. Oxford University Press. ISBN: 0-19-513804-X.

Rubin, Yoram and Gedeon Dagan (1987a). "Stochastic identification of transmissivity and effective recharge in steady groundwater flow: 1. Theory". In: *Water Resources Research* 23.7, pp. 1185–1192. DOI: 10.1029/WR023i007p01185.

— (1987b). "Stochastic identification of transmissivity and effective recharge in steady groundwater flow: 2. Case study". In: *Water Resources Research* 23.7, pp. 1193–1200. DOI: 10.1029/WR023i007p01193.

Rubin, Yoram et al. (2010). "A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields". In: *Water Resources Research* 46.10, pp. 1–23. DOI: 10.1029/2009WR008799.

Ruiz, Laurent et al. (2010). "Water balance modelling in a tropical watershed under deciduous forest (Mule Hole, India): Regolith matric storage buffers the groundwater recharge process". In: *Journal of Hydrology* 380.3–4, pp. 460–472. DOI: http://dx.doi.org/10.1016/j.jhydrol.2009.11.020.

Rumsey, Christine A. et al. (2015). "Regional scale estimates of baseflow and factors influencing baseflow in the Upper Colorado River Basin". In: *Journal of Hydrology: Regional Studies* 4, pp. 91–107. DOI: http://dx.doi.org/10.1016/j.ejrh.2015.04.008.

Rutqvist, J. (2015). "Fractured rock stress-permeability relationships from in situ data and effects of temperature and chemical-mechanical couplings". In: *Geofluids* 15.1-2, pp. 48–66. DOI: 10.1111/gfl.12089.

Sartore, Luca (2013). "spMC: Modelling Spatial Random Fields with Continuous Lag Markov Chains". In: *The R Journal* 5.2, pp. 16–28.

Sawada, Atsushi et al. (2015). "– Groundwater flow and transport modelling of fracture system at regional, block, and single-fracture scale flow and transport, Olkiluoto". In:

Sawicz, K. et al. (2011). "Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA". In: *Hydrol. Earth Syst. Sci.* 15.9. HESS http://www.hydrol-earth-syst-sci.net/15/2895/2011/hess-15-2895-2011.pdf, pp. 2895–2911. DOI: 10.5194/hess-15-2895-2011.

Scanlon, Bridget R., Richard W. Healy, and Peter G. Cook (2002). "Choosing appropriate techniques for quantifying groundwater recharge". In: *Hydrogeology Journal* 10.1. DOI: 10.1007/s10040-001-0176-2.

Schmidt, Hannes, Thilo Eickhorst, and Rolf Tippkötter (2011). "Monitoring of root growth and redox conditions in paddy soil rhizotrons by redox electrodes and image analysis". In: *Plant and Soil* 341.1, pp. 221–232. DOI: 10.1007/s11104-010-0637-2.

Schruben, Paul G Arndt et al. (1994). "Geology of the Conterminous United States at 1: 2,500,000 Scale–A Digital Representation of the 1974 PB King and HM Beikman Map". In:

Schwarz, Gregory E and RB Alexander (1995). "State soil geographic (STATSGO) data base for the conterminous United States". In:

Sebok, E. et al. (2016). "Using expert elicitation to quantify catchment water balances and their uncertainties". In: *Water Resources Research* 52.7, pp. 5111–5131. DOI: 10.1002/2015wr018461.

Segal, Eran et al. (2008). "Integration of Hard and Soft Data to Characterize Field-Scale Hydraulic Properties for Flow and Transport Studies". In: *Vadose Zone Journal* 7.3, pp. 878–889. DOI: 10.2136/vzj2007.0090.

Sheather, S. J. and M. C. Jones (1991). "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 53.3, pp. 683–690. ISSN: 00359246.

Siebert, S. et al. (2010). "Groundwater use for irrigation – a global inventory". In: *Hydrology and Earth System Sciences* 14.10, pp. 1863–1880. DOI: 10.5194/hess-14-1863-2010.

Singh, R., S. A. Archfield, and T. Wagener (2014). "Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments – A comparative hydrology approach". In: *Journal of Hydrology* 517, pp. 985–996. DOI: http://dx.doi.org/10.1016/j.jhydrol.2014.06.030.

Singhal, Brij Bhusan Saran and Ravi P Gupta (2010). *Applied hydrogeology of fractured rocks*. ISBN: 9048187990.

Sivapalan, M. et al. (2003). "IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences". In: *Hydrological Sciences Journal* 48.6, pp. 857–880. ISSN: 0262-6667. DOI: 10.1623/hysj.48.6.857.51421.

Smith, Mark et al., eds. (2016). *Spring : managing groundwater sustainably*. International Union for Conservation of Nature and Natural Resources. DOI: 10.2305/iucn.ch.2016.wani.8.en.

Smith, Tyler, Lucy Marshall, and Ashish Sharma (2014). "Predicting hydrologic response through a hierarchical catchment knowledgebase: A Bayes empirical Bayes approach". In: *Water Resources Research* 50.2, pp. 1189–1204. DOI: 10.1002/2013WR015079.

Stober, I. (1996). "Hydrogeological investigations in crystalline rocks of the Black Forest, Germany". In: *Terra Nova* 8.3, pp. 255–258. DOI: 10.1111/j.1365-3121.1996.tb00754.x.

Straubhaar, J. (2015). *DeeSse User's Guide*.

Tague, C. L., J. S. Choate, and G. Grant (2013). "Parameterizing sub-surface drainage with geology to improve modeling streamflow responses to climate in data limited environments". In: *Hydrol. Earth Syst. Sci.* 17.1. HESS http://www.hydrol-earth-syst-sci.net/17/341/2013/hess-17-341-2013.pdf, pp. 341–354. DOI: 10.5194/hess-17-341-2013.

Takagi, Masahiro (2013). "Evapotranspiration and deep percolation of a small catchment with a mature Japanese cypress plantation". In: *Journal of Forest Research* 18.1, pp. 73–81. DOI: 10.1007/s10310-011-0321-2.

Title, Pascal O. and Jordan B. Bemmels (2017). "ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling". In: *Ecography*, n/a–n/a. DOI: 10.1111/ecog.02880.

Todd, D.K. and L.W. Mays (2004). *Groundwater Hydrology*. ISBN: 9780471059370.

Turner, Brandon M. and Trisha Van Zandt (2012). "A tutorial on approximate Bayesian computation". In: *Journal of Mathematical Psychology* 56.2, pp. 69–85. DOI: https://doi.org/10.1016/j.jmp.2012.02.005.

USGS (2005). "Locations of Regional Assessments of Streams and Rivers". In:

Valpine, Perry de et al. (2017). "Programming with models: writing statistical algorithms for general model structures with NIMBLE". In: *Journal of Computational and Graphical Statistics*. DOI: 10.1080/10618600.2016.1172487.

Vincke, Caroline and Yves Thiry (2008). "Water table is a relevant source for water uptake by a Scots pine (Pinus sylvestris L.) stand: Evidences from continuous evapotranspiration and water table monitoring". In: *Agricultural and Forest Meteorology* 148.10, pp. 1419–1432. DOI: 10.1016/j.agrformet.2008.04.009.

Vries, Jacobus J. de and Ian Simmers (2002). "Groundwater recharge: an overview of processes and challenges". In: *Hydrogeology Journal* 10.1, pp. 5–17. DOI: 10.1007/s10040-001-0171-7.

Wada, Yoshihide et al. (2010). "Global depletion of groundwater resources". In: *Geophysical Research Letters* 37.20. DOI: 10.1029/2010GL044571.

Wagener, Thorsten and Alberto Montanari (2011). "Convergence of approaches toward reducing uncertainty in predictions in ungauged basins". In: *Water Resources Research* 47.6, n/a–n/a. DOI: 10.1029/2010WR009469.

Walker, Douglas D., Björn Gylling, and Jan-Olof Selroos (2005). "Upscaling of Hydraulic Conductivity and Telescopic Mesh Refinement". In: *Ground Water* 43.1, pp. 40–51. DOI: 10.1111/j.1745-6584.2005.tb02284.x.

Warmink, Jord J. et al. (2010). "Identification and Quantification of Uncertainties in a Hydrodynamic River Model Using Expert Opinions". In: *Water Resources Management* 25.2, pp. 601–622. DOI: 10.1007/s11269-010-9716-7.

Wen, Xian-Huan and Chen-Shan Kung (1993). "Stochastic Simulation of Solute Transport in Heterogeneous Formations: A Comparison of Parametric and Nonparametric Geostatistical Approaches". In: *Ground Water* 31.6, pp. 953–965. DOI: 10.1111/j.1745-6584.1993.tb00869.x.

Wieczorek, Michael E. and Andrew E. LaMotte (2010a). "Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: 30-Year Average Annual Precipitation, 1971-2000". In:

— (2010b). "Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: Basin Characteristics, 2002". In:

— (2010c). "Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: Bedrock Geology". In:

Wieczorek, Michael E. and Andrew E. LaMotte (2010d). "Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: Estimated Mean Annual Natural Groundwater Recharge, 2002". In:

— (2010e). "Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: NLCD 2001 Land Use and Land Cover". In:

— (2010f). "Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: STATSGO Soil Characteristics". In:

— (2010g). "Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: Surficial Geology". In:

— (2010h). "Attributes for MRB_E2RF1 Catchments by Major Rivers Basins in the Conterminous United States: Total Precipitation, 2002". In:

Winsemius, H. C. et al. (2009). "On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information". In: *Water Resources Research* 45.12, n/a–n/a. DOI: 10.1029/2009WR007706.

Wolock, David M (1997). "STATSGO soil characteristics for the conterminous United States". In:

Wolock, David M. (2003). "Estimated Mean Annual Natural Ground-Water Recharge in the Conterminous United States". In: *Open-File Report*.

Woodbury, Allan D. (2011). "Minimum relative entropy, Bayes and Kapur". In: *Geophysical Journal International* 185.1, pp. 181–189. DOI: 10.1111/j.1365-246x.2011.04932.x.

Woodbury, Allan D. and Yoram Rubin (2000). "A Full-Bayesian Approach to parameter inference from tracer travel time moments and investigation of scale effects at the Cape Cod Experimental Site". In: *Water Resources Research* 36.1, pp. 159–171. DOI: 10.1029/1999WR900273.

Xie, Yueqing et al. (2017). "Uncertainty of groundwater recharge estimated from a water and energy balance model". In: *Journal of Hydrology*. DOI: https://doi.org/10.1016/j.jhydrol.2017.08.010.

Xu, Tianfang and Albert J. Valocchi (2015). "Data-driven methods to improve baseflow prediction of a regional groundwater model". In: *Computers & Geosciences* 85.Part B, pp. 124–136. DOI: https://doi.org/10.1016/j.cageo.2015.05.016.

Xu, Tianfang et al. (2017). "Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model". In: *Water Resources Research* 53.5, pp. 4084–4105. DOI: 10.1002/2016WR019831.

Yang, Feng-Rong et al. (2009). "The impact of tunneling construction on the hydrogeological environment of "Tseng-Wen Reservoir Transbasin Diversion Project" in Taiwan". In: *Engineering Geology* 103.1–2, pp. 39–58. DOI: http://dx.doi.org/10.1016/j.enggeo.2008.07.012.

Yang, Wen-Hsi, David Clifford, and Budiman Minasny (2015). "Mapping soil water retention curves via spatial Bayesian hierarchical models". In: *Journal of Hydrology* 524, pp. 768–779. DOI: http://dx.doi.org/10.1016/j.jhydrol.2015.03.029.

Ye, Ming, Karl F. Pohlmann, and Jenny B. Chapman (2008). "Expert elicitation of recharge model probabilities for the Death Valley regional flow system". In: *Journal of Hydrology* 354.1-4, pp. 102–115. DOI: 10.1016/j.jhydrol.2008.03.001.

Yeh, Hsin-Fu et al. (2009). "GIS for the assessment of the groundwater recharge potential zone". In: *Environmental Geology* 58.1, pp. 185–195. DOI: 10.1007/s00254-008-1504-9.

Yeh, Hsin-Fu et al. (2016). "Mapping groundwater recharge potential zone using a GIS approach in Hualian River, Taiwan". In: *Sustainable Environment Research* 26.1, pp. 33–43. DOI: http://dx.doi.org/10.1016/j.serj.2015.09.005.

Yin, Li ming et al. (2005). "Application of high water-pressure test to deep borehole". In: *Rock and Soil Mechanics* 26.10, pp. 1692–1694.

Zaidel, Jacob, Bradley Markham, and David Bleiker (2010). "Simulating Seepage into Mine Shafts and Tunnels with MODFLOW". In: *Ground Water* 48.3, pp. 390–400. DOI: 10.1111/j.1745-6584.2009.00659.x.

Zhao, J. (1998). "Rock mass hydraulic conductivity of the Bukit Timah granite, Singapore". In: *Engineering Geology* 50.1–2, pp. 211–216. DOI: http://dx.doi.org/10.1016/S0013-7952(98)00021-0.