Addressing Variability in Hydrologic Systems Using Efficient Uncertainty Quantification

by

Martin C. Dwelle

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Environmental Engineering)
in the University of Michigan
2018

Doctoral Committee:

Associate Professor Valeriy Y. Ivanov, Chair
Professor M. Clara Castro
Professor Nikolaos D. Katopodes
Dr. Khachik V. Sargsyan, Sandia National Laboratories

Martin C. Dwelle

dwellem@umich.edu

ORCID id: 0000-0001-7188-5908

This dissertation is dedicated to my grandfather, the engineer, Donald Arthur Schneider.

Acknowledgements

Many long days were spent staring at words, equations, and code to produce this dissertation. However, none of these pages could have been written without significant support.

I am forever grateful to my advisor, Prof. Valeriy Y. Ivanov, whose guidance, support, and advice were vital to my growth as a researcher. His flexibility in allowing me to perform research outside of Ann Arbor, for both academic and personal reasons, will always be appreciated, as well as his patience during times spent in the research doldrums.

There are many people within the University of Michigan that deserve appreciation. I am grateful to Prof. Nikolaos Katopodes, who reinvigorated my appreciation for math and computing through his numerical methods course. This dissertation would look different if it weren't for the mentorship of Dr. Khachik Sargsyan, a Michigan alumnus. His willingness and patience to mentor a hydrologist in his field of expertise led to a stronger, more complete body of work and enriched my academic life. Faculty within and outside the department have provided challenges which helped me develop as a researcher, including but not limited to: Prof. Veronica Berrocal, Prof. Aline Cotel, Prof. Avery Demond, and Prof. Steven Wright.

I am thankful to the many students that have passed through Prof. Ivanov's lab through the years, including: Jongho Kim, who mentored me in my early years as a grad student; Lingli He, who provided guidance and valuable feedback in numerous funding-seeking efforts; Gregory Ewing, who kept me from getting lost in Brazil and has been a great friend; and Liz Agee, my colleague, office mate, and friend who has been there to share in success and frustration in research and personal life. Others within the university have made the made the good times better and the bad times bearable include: Sara Rimer, Frank Sedlar, Anton

Loukianov, Heather Goetsch, Katya Rahkmatulina, Kevin Fries, Brandon Wong, and Jeseth Delgado Vela.

I cannot adequately express my deepest gratitude to my parents, JoEllen Schneider and Ronald Dwelle. Given my behavior as a teenager it is a surprise to myself, and likely them, that I went to graduate school. Thank you very much for being patient and letting me find the long, circuitous path to a challenging and rewarding life.

Finally, I thank my wife Erica Sivertson. The best choice I made was going to the University of Michigan; not because of this dissertation, but because it was there that I met you. The support you provided during the last four years will take a lifetime to reciprocate, but luckily that is how much time we will have together. You have made life more exciting and rewarding, getting this eremitic engineer to challenge my preconceptions of what is achievable and introducing me to the joys of "seat of your pants" travel. As I close this period of my life as a graduate student, I am excited to grow with you and build a rewarding future together.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

The scale and complexity of environmental and earth systems introduce an array of uncertainties that need to be systematically addressed. In numerical modeling, the ever-increasing complexity of representation of these systems confounds our ability to resolve relevant uncertainties. Specifically, the numerical simulation of the governing processes involve many inputs and parameters that have been traditionally treated as deterministic. Considering them as uncertain with traditional approaches introduces a large computational burden, stemming from the requirement of a prohibitive number of model simulations. Furthermore, within hydrology, most catchments are sparsely monitored, and there are limited, disparate types of data available to confirm the model's behavior. Here I present a blueprint of a general, computationally efficient approach to uncertainty quantification for complex hydrologic models, taking advantage of recent methodological developments.

The framework is used in two basic science problems in hydrology. First, it is applied to the problem of combining heterogeneous data sources representing different physical processes to infer physical parameters for the complex hydrologic model tRIBS-VEGGIE. The inference provides a probabilistic interpretation of bulk soil characteristics and related hydraulic properties for an experimental watershed in central Amazonia. These parameters are then used to propagate uncertainty in hydrologic response to an array of quantities of interest through tRIBS-VEGGIE and determine their sensitivity to uncertain model inputs.

Second, the framework is used to explore landscape controls mediated by subsurface hydrologic dynamics on the distribution of vegetative traits in a mature Amazon rainforest. This study features a large parameter set as uncertain across

three different soil types and three layers of vegetation, explicitly incorporating interactions between subsurface moisture and vegetation biophysical function. Vegetative performance is examined using a hypothesized cost-benefit approach between vegetation carbon uptake and hydraulic effort required to maintain long-term production.

The research enables model-driven inference using a disparate set of observed hydrologic variables including stream discharge, water table depth, evapotranspiration, soil moisture, and gross primary production from the Asu experimental catchment near Manaus, Brazil. Computationally inexpensive model surrogates are constructed and shown to mimic solution of the complex hydrologic model tRIBS-VEGGIE with a high skill. The two applications demonstrate the flexibility of the framework for hydrologic inference in watershed with sparse, irregular observations of varying accuracy. Significant computational savings imply that problems of greater computational complexity and dimension can be addressed. Furthermore, the framework simultaneously yields probabilistic representation of model behavior, robust parameter inference, and sensitivity analysis without the need for greater investment in computational resources.

CHAPTER 1

Introduction

In hydrology, or any large-scale earth system science, models are needed to study the system's processes and outcomes. Unlike disciplines in which experiments can be performed at a lab bench, one cannot generate a physical "model earth" or model watersheds to experiment on. In hydrology, attempts to do such make for valuable science and interesting showcases (e.g., Figure 1.1), but these physical models rely on fixed geometries and allow for minimal carrying out of "what if?" scenarios on the complex system. Here complex system is used in the terms of complexity science, meaning that not only are the elements of a system complicated and hard to model (e.g., biomes, atmosphere, cities, animals, human intervention, climate), but that there are interactions between these different elements of the system which may not (and usually do not) have a mechanistic method of interaction. Because of this complexity, models of earth surface processes need many simplifications to make the solution of the governing equations computationally tractable. Introducing these simplifications in the form of boundary conditions, closure rules, or state variables places uncertainty into models.

Getting best-possible representation of these uncertainties is a major goal in hydrology (Beven 1993; Montanari and Koutsoyiannis 2012; Efstratiadis and Koutsoyiannis 2010; Beven and Westerberg 2011), traditionally around estimation of streamflow for engineering design or watershed management purposes (Beven 2006; Seibert and Beven 2009; Beven 1993). In these assessments, inference is often used to determine parameter or boundary values for a model, but these attempts usually provide a scalar value for a parameter that is then used as an estimate of a "true" value. However, the complex nature of hydrologic systems leads to heterogeneous, anisotropic features and behaviors that cannot be accounted for

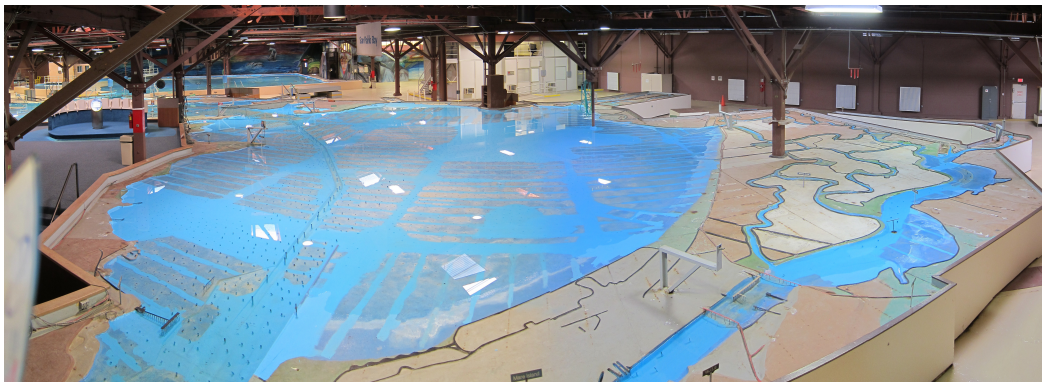

Figure 1.1: Picture of the United States Army Corps of Engineers Bay Model, approximately 11,500 $[m^2]$ in area, provides a hydraulic scale model of the greater San Francisco Bay area. The model was previously used for scientific research and engineering planning between 1958-2000 and is now used as a public education tool. The model is scaled 1:1000 on the horizontal and 1:100 on the vertical axis to enable the shallow water approximation, and operates at a time scale of 1:100. The construction of the model required approximately 1,300 metric tonnes of concrete. Today, the hydraulic processes of this model can be simulated on a computational cluster.

with a scalar value, and attempting to resolve uncertainties in this way neglects variability of parameter values and watershed processes. General, focused efforts are required to address the effects of uncertainty on hydrologic systems and the models that represent them.

1.1 Aim of uncertainty quantification

> Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity. — George Box (1976)

In computational science and engineering, the major goal is to accurately represent physical problems through simulation. The simulation approach offers several advantages such as quick iteration in the design cycle, ability to represent domains for which physical models cannot be constructed, observe alternate histories, or investigate potential future realizations of the problem of interest. In all of these

approaches, the main goal is to map a space of inputs $\mathcal{X}$ to a space of potential outputs $\mathcal{Y}$.

The exact mapping from $\mathcal{X}$ to $\mathcal{Y}$ is not analytic, i.e., we do not know the exact properties of the mapping. To this end we represent the mapping by using conservation laws, relying on approximations to these laws from mathematical closures. The models made from these approximations are often computationally expensive, and makes it difficult to experiment with a large number of unique realizations of $\mathcal{X}$. Even if perfect information about the parameterizations of these model were known, errors could still appear in forecasting or inference. Despite the great advances in physical modeling since George Box's famous quote above, all models are still "wrong," but valuable work has been put forth into making models *less wrong*. A core tenet of uncertainty quantification (UQ) is the investigation of the sources and magnitude of *wrongness* or inaccuracy in numerical simulations. In the domain of UQ, both $\mathcal{X}$ and $\mathcal{Y}$ are treated probabilistically, and the goal becomes to get adequate sampling of $\mathcal{X}$ to get a robust classification of $\mathcal{Y}$. Therefore, the inputs and outputs are treated as random variables.

Most UQ problems are classified as *uncertainty propagation* or *inverse* estimation. In uncertainty propagation, we sample from the input space $\mathcal{X}$ in order to get an output probability density for $\mathcal{Y}$. In inverse estimation, we have a sample of the output $\mathcal{Y}$ and would like to infer what is the most likely selections for $\mathcal{X}$. The goal in either problem is to use the probability spaces for prediction, risk analysis, or estimation of model parameters. For example, in hydrology, there has been a decades-long goal of quantifying uncertainty in rainfall/runoff models to get estimates of streamflow (e.g., Kirchner 2009a; Burns et al. 2001; Milly et al. 2002; Orlandini et al. 1996). With hydrologic systems, the mapping of $\mathcal{X}$ to $\mathcal{Y}$ (or vice-versa) includes macro and micro-level physical processes often described through a large number of model parameters or inputs (e.g., Gutiérrez-Jurado et al. 2006; Weiler and McDonnell 2004; Meerveld and Weiler 2008; Krause et al. 2015). This large input space leads to the desire to have information on sensitivities of the inputs, i.e., how perturbations in the values of $\mathcal{X}$ change the output $\mathcal{Y}$. This information allows for both better understanding of the interactions between processes in the mapping function (model), but also can lead to understanding of the underlying processes being modeled by this mapping (Sobol 2001; Saltelli 2002).

In hydrology, limited observational data is often available because many basins are sparsely gauged or ungauged (Winsemius et al. 2009; Hrachowitz et al. 2013). Furthermore, it is expensive or infeasible to install monitoring equipment in order to obtain data to inform models. Therefore, hydrologic modeling often needs to assess outcomes with limited data availability. A Bayesian approach has been used in hydrology precisely due to its ability to work with limited amounts of data (e.g., Montanari et al. 2009). The strength of the Bayesian approach is that it allows for the use of available prior knowledge, however limited it may be, in probabilistic estimation. If we have two events A and B in a probability space, then Bayes' theorem can be presented as

$$p(B \mid A) = \frac{p(A \mid B)p(B)}{p(A)}. \tag{1.1}$$

In the case of inverse problems, if we have a set of observations $\mathbf{y}_{\text{obs}} = \{y_i\}_i^N \in \mathcal{Y}$, we would like to find the possible inputs $x \in \mathcal{X}$ that could lead to the observed data. We assume that the values in $\mathcal{X}$ follow a probability distribution $p(x)$ called the *prior*. After we observe $\mathbf{y}$, we can update the prior according to Bayes' rule as

$$p(x \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid x)p(x)}{p(\mathbf{y})}, \tag{1.2}$$

which is called the *posterior* distribution. This approach can be favorable to the deterministic approach because it provides probabilistic assessments of quantities of interest, risk assessments, etc. These solutions are updating knowledge about the system in question. The drawbacks to this approach are large computational costs to estimate posterior probabilities, but these costs have been readily decreasing with an increase of computational power.

1.2 Research scope

The call for uncertainty assessments in hydrology has been a constant drumbeat (e.g., Winter 1981; Chapman 1986; Beven 1993; Liu et al. 2009; Montanari and Koutsoyiannis 2012). However, assessments are still urgently needed in hydrology for the following purposes, among others:

- where limited data are available (Seibert and Beven 2009; Borga et al. 2008; Cibin et al. 2014; Hrachowitz et al. 2013);
- making management and engineering decisions in problems of flood forecasting (Borga et al. 2008; Norbiato et al. 2008; Bogner and Pappenberger 2011; Salamon and Feyen 2009; Fontanazza et al. 2012; Villarini et al. 2010; Wong et al. 2015);
- investigating the implications of changing climate on earth and hydrologic systems (Cayan et al. 2010; Najafi et al. 2011; Milly et al. 2002; Kundzewicz et al. 2014; Peel and Blöschl 2011; Blöschl and Montanari 2010; Kim and Ivanov 2015); and
- for a better understanding of the underlying dynamics of these hydrologic systems using complex, process-based models (Beven 2000; Kim and Ivanov 2014; Kim et al. 2016; McCord et al. 1991; Ivanov et al. 2010; Rosenbaum et al. 2012).

In all of these cases, a general uncertainty quantification approach is desired such that a hydrologic model can help to build the probabilistic understanding available through the use of Equation (1.2). In the following chapters, a novel approach to uncertainty quantification in hydrologic models is presented which allows the probabilistic interpretation of model parameterizations, outputs, and sensitivities.

In Chapter 2, the uncertainty quantification framework is presented. The framework relies on the approximation of model behavior over a range of parametric variability using non-intrusive spectral projection through polynomial chaos (PC) expansions. These expansions fit a series of polynomials to the model's input-to-output mapping. This mapping, once trained, allows the approximation of the model's behavior through computation of polynomials instead of complex hydrologic models, and is much cheaper computationally than using Monte Carlo for uncertainty propagation, estimation of model parameters, and global sensitivity analysis. The construction of these PC surrogates is aided by the use of Bayesian compressive sensing—a technique that facilitates dimensionality reduction. Finally, the use of these surrogates in Bayesian inference is presented.

Chapter 3 outlines the novel adaptation of this UQ framework to the complex, three-dimensional ecohydrologic model tRIBS-VEGGIE to a tropical watershed in Amazonia. The work provides a blueprint of a general approach to uncertainty

quantification for complex hydrologic models, taking advantage of recent methodological developments in uncertainty quantification. Heterogeneous data sources in both time and space are used to perform probabilistic, parametric inference of soil properties controlling the flow of water in the watershed and the availability of water for ecologic function. The advantages and limitations of the approach are presented, along with implications of the approach being used to address larger problems in hydrology.

In Chapter 4, the framework presented in Chapter 3 is used in a basic science application to the functional, adaptive strategy of vegetation function in an Amazon rainforest. In a first of its kind work in terms of the scale of the dimensionality and complexity of modeled processes, uncertainty is prescribed to heterogeneous plant and soil types for a computational transect in an Amazonian rainforest. By applying the framework, uncertainty in the complex ecohydrologic system is addressed and competitive vegetation strategy is ascertained through parametric inference and the introduction of a thermodynamics-based performance metric for soil-water controlled vegetative function.

Chapter 5 summarizes this dissertation and addresses ongoing and future studies. The major conclusions and critical assumptions of the conducted research are presented, along with the feasibility of expanding the framework to hydraulic and engineered systems for use in flood forecast and resilience modeling.

CHAPTER 2

Uncertainty quantification

2.1 Introduction

Since the birth of modern computational methods, scientific computing and advancements in engineering have complimented each other and spurned innovation. Some of these innovations have gone on to be widely adopted throughout many scientific disciplines or provided greater capabilities for numerical simulation, e.g., the Monte Carlo method for sampling, Runge-Kutta methods for approximating ordinary differential equations, or the Kalman filter for estimating unknowns through data assimilation (Owen 2013; Flegal et al. 2008; DeChant and Moradkhani 2012). As advances in computational science continue, programming packages are freely available for specific tasks. In addition to this software availability, advances in computational power mean that larger (in computational domain) and more complex systems can be simulated.

The accuracy of these simulations rely on an adequate level of knowledge about the system being simulated. Uncertainty arises in any system when simplifications are made in order to make the expression of the system through numerical modeling feasible. Within hydrology, as computational power has increased, the level of complexity of these models has also increased. In its simplest form, one can model a hydrologic system with a "bucket" model (Manabe 1969; Romano et al. 2011), where water enters the top of the bucket and drains out of the bottom, with parameterizations controlling the drainage from the bucket. The uncertainty associated with this type of model is small as its dynamics are fully represented with a few equations and parameters. As an effort to more accurately model the physical processes controlling hydrology, very complex models have been created which

account for stochastic rainfall, saturated and unsaturated flow in the subsurface, infiltration, runoff and run-on, and energy fluxes, among many other processes (Liu et al. 2009; Maxwell et al. 2014). The introduction of these complexities leads to far greater uncertainties than those arising from a bucket model.

The trade off of greater uncertainties is that the understanding of these smaller scale processes could lead to greater understanding of higher-level characteristics of hydrologic systems. This line of thought is consistent with the idea of a complex system, i.e., a system where the characteristics of the high-level behavior is the result of low-level interactions, which is explicitly argued with respect to hydrologic systems (Allan et al. 1997; Bonan 2008; Kirchner 2009a). Therefore, having a greater understanding of the controls on water and energy balances in a hydrologic model can benefit the understanding of higher level hydrologic behavior. In a hydrologic model, these lower-level processes are often parameterized as a result of a mathematical closure of physical phenomena. These parameterizations lead to simplifications in the way physical processes are expressed through a model, and the selection of values for these parameters is dependent on any available data of properties of the system, parameters describing a property of a medium, or the modeler's expert opinion. The selection and estimation of parameter values has been the subject of numerous studies in hydrology (Yeh 1986; Russo and Bouton 1992; Abbaspour et al. 2004; Kowalsky et al. 2004; Salamon and Feyen 2009; Vrugt et al. 2008; Schoups and Vrugt 2010; Romano et al. 2011; Yu and Coulthard 2015).

In addition to these parameters, there are input fields which describe spatial or temporal states of the domain. These parameters and fields are the assumptions made about the makeup and behavior of our domain, and these will be inherently uncertain since there is no possible way to have perfect information about the system being modeled, and this is especially true for natural systems. The data that we do have is often limited in both quantity and quality by variability, bias in measurements, and by the degrees to which information about the system is able to be collected (Grabe 2014). Attempting to model these systems numerically, we need to account for these limitations as well as take into account the limitations of the numerical techniques and mathematical models used to mimic the system (Beven 2008; Higham 2002). This is the objective of uncertainty quantification.

Due to the ubiquity of these uncertainties, significant investment has been made to address these various sources of uncertainty within hydrology (e.g., Liu et al.

2009; Bulygina and Gupta 2011; Beven 1993; Gupta et al. 2014; Salamon and Feyen 2009; Chen et al. 2011a; Reichle et al. 2008; Grayson et al. 2002; Evensen 2009; Montanari and Koutsoyiannis 2012). However, many of these approaches require a Gaussian assumption on model inputs or behavior, or make other assumptions which result in a loss of interpretability of the model results. For example, using neural networks to emulate model behavior requires calculations of many coefficients of hidden nodes with no physical meaning, making interpreting the process that led to results difficult (Lin et al. 2006; Srivastava et al. 2013; Alemohammad et al. 2016). Although these frameworks can work well, there is still a desire within hydrology to have a general framework for quantifying uncertainty which allows for interpretability of results as well as understanding of the internal dynamics of the model, and therefore the internal dynamics of the hydrologic system being modeled.

More broadly, a completely holistic approach to UQ encapsulates *all* uncertainties of computer simulations within a specific context. This is a valuable pursuit for many computational sciences, and as such, UQ has emerged in the last two decades as very active research field which incorporated applied mathematics, engineering, and physical sciences.

Within engineering and the physical sciences we often want to distinguish between *epistemic* and *aleatory* uncertainties (De Rocquigny 2012; Beven et al. 2016; Beven 2013). Epistemic uncertainty is the lack of knowledge of the mechanism being modeled due to limited data or incomplete representation of the process. Aleatory uncertainty (or variability) emerges from the natural stochasticity of the system being modeled. These uncertainties are amplified by the complexity of problems tackled through simulations in modern engineering and physical sciences. However, in the world created through numerical experiments, these different types of uncertainties often get combined (Der Kiureghian and Ditlevsen 2009), meaning that it often makes sense to apply UQ to approximate both epistemic and aleatory uncertainty.

Within a complete UQ analysis of a model or mechanism, a significant portion of the following analyses may be undertaken (in approximate order):

- Determine output of model to study.
- Identify impactful parameters and inputs to this output and specifying their uncertainty. This requires the knowledge of the underlying framework of

the model, i.e., the governing equations and assumptions, and any associated parameters or forcings.

- Generate a training and validation set from uncertain inputs and propagate the uncertainty through the model. This is the *uncertainty propagation* step of UQ. Within this step, we are looking for the full probability density function (PDF) of specified model outputs — quantity of interest (QoI) — based on the PDF of uncertain inputs. A QoI is any model output or derivative of model output that can be represented through a scalar, e.g., monthly mean streamflow or a basin-averaged water table depth. A discussion on the use of multiple QoIs within the UQ framework is provided in Section 2.3.1.
- Investigate uncertainty in the model's output and the sensitivity of the outputs to changes in the inputs through global sensitivity analysis (Saltelli 2002).
- If interested in estimation of parameter values using observed data, perform inverse estimation of uncertain parameters. This is referred to as *inverse estimation* and its goal is to infer parameter values for the system given the response of the QoI to uncertain states or parameters during a dynamic period of simulation.
- Once obtained, the PDF for inferred parameters can be used to further propagate uncertainty through the model, or be used to investigate behavior of other QoIs that may not have an associated observations used for inference.

In order to perform the process above, many model runs must be performed. This has traditionally been done using Monte Carlo methods (Neal 1993; MacKay 1998), but the computational expense of the model being evaluated in the Monte Carlo framework can lead to a total cost of evaluations that is prohibitive. For an explanation of how Monte Carlo methods scale poorly with dimension, see Appendix A.

The models used in these Monte Carlo methods will impact the feasibility and scale of investigation into uncertainties. In hydrology, the physical, process-based models that have become standard for investigation into surface/subsurface interactions (e.g., ParFlow, HydroGeoSphere, tRIBS+VEGGIE, Maxwell et al. 2014; Kollet and Maxwell 2006; Brunner and Simmons 2012; Ivanov et al. 2008b) are also computationally expensive, taking on order of days of simulation time for a moderately-sized watershed. Therefore, uncertainty quantification driven with Monte-Carlo methods may require thousands of model simulations, quickly mak-

ing computation infeasible. An alternative approach that has been gaining traction in the physical modeling community is to approximate the process-based, physical model with a "surrogate model" that maps the process-based model's input to outputs using methods at a reduced computational cost. Surrogate models can range in complexity, replacing the complex model using regression, radial basis functions, Gaussian processes, neural networks, or polynomial expansions (Jin et al. 2001; Xiu and Karniadakis 2002). Polynomial chaos expansions (PCEs) are one such surrogate modeling technique that provides an approximation to a computer model through the model's spectral representation on a basis of polynomial functions (Najm 2009; Elsheikh et al. 2014a). The reduction of computation time stemming from this surrogate modeling enables the use of UQ in "real-world" applications for risk analysis and to inform decision making.

This chapter will be used to introduce key aspects of the uncertainty quantification framework used throughout this dissertation. It will focus on the probabilistic approach to UQ enabled by polynomial chaos expansions (PCEs) for both uncertainty propagation and Bayesian inference. This will serve as an overview for the hydrologic application of UQ introduced in Chapter 3, including recent advances in UQ which are utilized for this work.

2.2 Propagation

As an introduction to uncertainty quantification, we will look at the probabilistic propagation of uncertainty, whose purpose is to quantify the contribution of the uncertainty of model parameterizations to the outcome of a deterministic model (Arnst and Ponthot 2014). Representing these parameterizations as random variables following a known probability distribution, this uncertainty is propagated through a model and we are interested in the distribution of model outputs.

2.2.1 The model

Consider a *forward* model of the system of interest. In this work, this model will be a hydrologic model describing the movement of water, but it can be generalized as follows. Say we have a model $\mathcal{M}$ with inputs $\mathbf{x} \in \mathcal{P}$ from the parameter domain $\mathcal{P} \in \mathbb{R}^M$ with $M > 0$. This model is used to predict some output QoI $\hat{y} \in \mathbb{R}$, where

we keep $\hat{y}$ as a scalar quantity, but it can (and will) be used for multiple QoIs in this work. Therefore, this model is a simple function that maps uncertain inputs to an uncertain output:

$$\hat{y} = \mathcal{M}(\mathbf{x}). \tag{2.1}$$

We assume that $\mathcal{M}$ is a *deterministic* model that can scale in complexity depending on the problem being solved (e.g., in its simplest form it could be $\mathcal{M}(\mathbf{x}) = \sum x_i$). Where uncertainty quantification becomes interesting is when $\mathcal{M}$ is complex, such as the numerical solutions to the governing equations of fluid flow. It is not uncommon for investigations of fluid flow to make single evaluations of the model represented by Equation 2.1 and treat it as a "black box," i.e., the internal model dynamics are assumed unknown and only a relationship between model inputs and outputs is analyzed. For some models, the internal dynamics may truly not be known; in others, the model may be computing hundreds of different numerical approximations to governing laws at each time step. The only requirement that is imposed on the model is that it must be executable, i.e., provided its inputs $\mathbf{x}$, the model produces its output $\hat{y}$.

2.2.2 Inputs and outputs

If we believe that the model $\mathcal{M}$ adequately captures the dynamics of the system being modeled, then we can also believe that the output $\hat{y} = \mathcal{M}(\mathbf{x})$ we receive from the model gives an accurate prediction, when the values of the input parameters/variables $\mathbf{x}$ are known. However, a typical case for the majority of models of environmental systems is that input values are uncertain, either due to natural variability in the system or inadequate knowledge of parameter values, in general. In this case, it is consistent to represent these input parameters as uncertain, so they follow a random vector

$$\boldsymbol{X} \sim \pi(\mathbf{x}), \tag{2.2}$$

where $\boldsymbol{X} = (X_1, X_2, \ldots, X_M)$ are independent and $\pi(\mathbf{x})$ is a vector of marginal probability density functions (PDF) describing the variability in each of the M uncertain variables in $\boldsymbol{X}$, which are known and defined prior to modeling. Random input variables to the model result in QoIs that can also be treated as random, i.e.:

$$Y = \mathcal{M}(\boldsymbol{X}). \tag{2.3}$$

With uncertainty propagation, we are interested in determining the probability distribution of Y. For example, if Y roughly follows a Gaussian distribution, the mean and variance of Y could be considered as estimates of the location and scale of the distribution. However, with complex interactions occurring within the model $\mathcal{M}$, the distribution of Y could be multi-modal or have large skewness or kurtosis, making conventional methods relying on the assumption of Gaussian behavior a poor approximation. As an illustration, consider that if the model output of interest has a similar variability as the prediction quantity Y in Figure 2.1, one cannot describe the output distribution through its moments, and the full characterization of the density function is required.

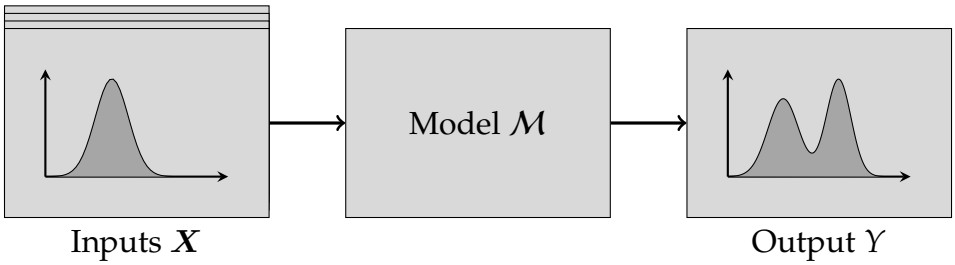

Figure 2.1: Diagram of forward propagation.

2.2.3 Approximating output with Monte Carlo simulation

Perhaps the simplest approach to carrying out the propagation in Figure 2.1 is to use random sampling with Monte Carlo (MC) simulation. The use of random sampling for statistics and probability is very old in mathematics, dating back to at least 1733 when Georges Louis Leclerc (known as Comte de Buffon) used random sampling to answer the question "Given a needle of length a and an infinite grid of parallel lines with common distance d between them, what is the probability that a needle, tossed at the grid randomly, will cross one of the parallel lines?" (Ramaley 1969). There are variations of this problem, sometimes involving a noodle instead of a needle for more complex geometries, or a coin tossed onto a finite floor with congruent squares, but it showed that random sampling was a powerful tool to solve complex integration problems. Buffon's work in using random sampling was relatively unknown and unused until Pierre Simon Laplace used the methodology to estimate the value of π later in the 18th century.

To approximate the output of the model $\mathcal{M}$, take N input samples $\boldsymbol{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ that are independent and identically distributed (iid) from the distribution for the input parameters from Equation 2.2. Then, the computation of the outputs $Y = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N\}$ is done via propagation where $\hat{y}_i = \mathcal{M}(\mathbf{x}_i)$ are realizations of the random variable in Equation 2.3. The first two moments of the output distribution can be approximated as:

$$\begin{aligned} \hat{\mu}_Y &= \mathbb{E}[\mathcal{M}(\boldsymbol{X})] \approx \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i, \\ \hat{\sigma}_Y^2 &= \mathbb{E}\left[\left(\mathcal{M}(\boldsymbol{X}) - \hat{\mu}_Y\right)^2\right] \approx \frac{1}{N-1}\sum_{i=1}^{N} \left(\hat{y}_i - \hat{\mu}_Y\right)^2. \end{aligned} \tag{2.4}$$

When the output is unimodal and close to Gaussian, the mean and variance in Eq. (2.4) may accurately represent the location and scale. If the output distribution is multimodal (as in Fig. 2.1) or more complex, higher-order moments of Y may be required. In the case where $\mathcal{M}$ is computationally expensive, we will want a faster way to calculate the moments in Equation 2.4. In the next section we'll look at surrogate modeling as a technique to get a computationally inexpensive approximation of $\mathcal{M}$.

2.3 Surrogate modeling

A *surrogate model* or *metamodel* is generally used to simulate the behavior of a more computationally complex model. The purpose of using a surrogate model is that it is computationally inexpensive compared to the original model, and it can therefore be rigorously sampled for uncertainty propagation, parameter inference, or sensitivity analysis. This does not come for free, however. If many uncertain inputs are taken, or there are high-order interactions between uncertain inputs in the computationally expensive model $\mathcal{M}$, then significant effort will need to be expended. However, as we will see later, recent advancements have made constructing surrogates more computationally efficient, enabling surrogate construction for models of increasing complexity.

There are multiple frameworks that fall into the class of surrogate models, e.g.: Guassian process models (Rasmussen 2006; Kennedy and O'Hagan 2000),

artificial neural networks (Ripley 1996), support vector machines (Abe 2010), and polynomial chaos expansions (Xiu and Karniadakis 2002; Le Maître and Knio 2010). These different classes of surrogates were often developed in parallel fields over the last two decades, with Gaussian process (GP) models and PCEs mainly being used in statistics and engineering (often computational fluid dynamics), whereas neural networks and support vector machines were more used in data-intensive applications.

This work focuses on using PCEs as the main workhorse to construct surrogate models. The methodology was developed 80 years ago (Wiener 1938), but laid largely untouched until recently as advancement in computational power made the method feasible for addressing real engineering challenges (e.g., Xiu and Karniadakis 2003; Xiu and Tartakovsky 2004; Najm 2009; Marzouk et al. 2007). We'll start by looking at the spectral expansion as it pertains to PCEs, then focus on the polynomial approximations to our complicated model $\mathcal{M}$.

Spectral expansion To construct the spectral expansion for the model, we assume that the model results belong in the square integral space L^2.[1] This means that we have a set in our parameter domain $\mathcal{P}$ and a weighting function w that maps to the positive real-valued domain: $w : \mathcal{P} \mapsto [0, \infty)$. We further assume that the model belongs to the function space $L^2_w(\mathcal{P})$ that has the inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|_w$, which are defined as

$$\begin{aligned} \langle u, v \rangle_w &= \int_{\mathcal{P}} u(\mathbf{x}) v(\mathbf{x}) w(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \\ \|u\|_w &= \langle u, u \rangle_w^{1/2} \end{aligned} \tag{2.5}$$

for the two elements $u, v \in L^2_w(\mathcal{P})$, where $L^2_w(\mathcal{P})$ is a *Hilbert space*.[2] As a Hilbert space, the integral of the weighting function across all real numbers is finite:

$$\int_{\mathcal{P}} u^2(\mathbf{x}) w(x) \, \mathrm{d}\mathbf{x} < \infty.$$

[1] This section is included for completeness. A brief background on measure-theoretic probability theory for this section can be found in many texts. An introduction based on the material of this thesis is included in Appendix A of Le Maître and Knio (2010)

[2] Hilbert space is simply an abstraction of Euclidean space. It allows for the use of vector calculus in any finite or infinite number of dimensions.

To construct the spectral expansion, we introduce the vectors $\Psi_i \in L^2_w(\mathcal{P})$ which are part of the orthonormal set $\{\Psi_i\}$ for all positive integers i. Due to their orthonormality, for all $i, j \in \mathbb{N}^+$, $\langle \Psi_i, \Psi_j \rangle = \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta. This allows to construct the spectral expansion for our model:

$$\mathcal{M} = \sum_{i=1}^{\infty} c_i \Psi_i, \text{ where} \tag{2.6}$$

$$c_i = \langle \mathcal{M}, \Psi_i \rangle_w . \tag{2.7}$$

The infinite series in Equation 2.6 needs to be truncated because evaluation of the model $\mathcal{M}$ is required for each i in the summand. To get a visual representation of this expansion and projection, Figure 2.2 projects a model result $\mathcal{M}$ in three-dimensional space to the truncated model result $\mathcal{M}_2$ in two-dimensional space.

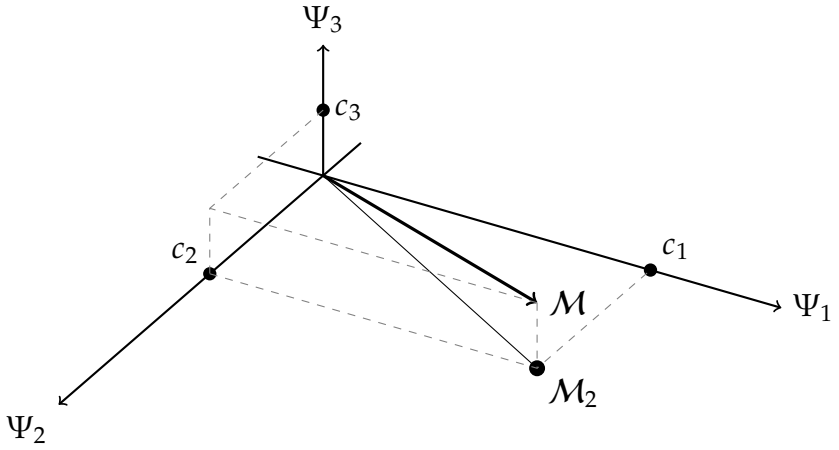

Figure 2.2: Orthogonal projection for $i = 3$. Here, the error from truncation of Equation 2.6 from $\mathcal{M}$ to $\mathcal{M}_2$ is $\mathcal{M} - \mathcal{M}_2 = c_3\Psi_3$ since $\mathcal{M} = c_1\Psi_1 + c_2\Psi_2 + c_3\Psi_3$.

To interpret the spectral expansion probabilistically, we can associate the weighting function with the probability density of the model inputs from Equation 2.2: $w(\mathbf{x}) = \pi(\mathbf{x})$. From this, u, v are random variables $u(\boldsymbol{X})$ and $v(\boldsymbol{X})$, and the expectation of their product is equal to the inner product from Equation 2.5:

$$\langle u, v \rangle_\pi = \mathbb{E}\left[u(\boldsymbol{X})v(\boldsymbol{X})\right].$$

2.3.1 Polynomial chaos expansion

In this section we will construct the polynomial chaos expansions that are the workhorse for the uncertainty analysis used in this dissertation. We concluded the last section mentioning that weighting functions that are probability densities offer a probabilistic interpretation for spectral expansions. Additionally, a subset of probability densities have polynomials orthogonal with their inner products. Four well known distributions are listed with their orthogonal polynomials and support in Table 2.1. In order to use these polynomial approximations we must have our input parameters $\mathbf{x}$ follow a distribution like those in Table 2.1. We will assume that our data are already in such a form, but if that is not the case, parameters can be shifted and scaled, often to a standard uniform or standard Gaussian distribution.

Table 2.1: Potential distributions for $\pi(\mathbf{x})$ and their orthogonal polynomials.

Distribution	Support	Polynomial
Uniform	$[-1, 1]$	Legendre
Gaussian	$(-\infty, \infty)$	Hermite
Beta	$[-1, 1]$	Jacobi
Gamma	$[0, \infty)$	Laguerre

Each variable for our uncertain input x_i will have an associated univariate polynomial $\{\Psi_{i,\alpha_i}\}$, where α_i are members of the multi-index $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_M)$, where M is the dimensionality of our inputs.[3] Then the multivariate polynomials $\{\Psi_{\boldsymbol{\alpha}}(\mathbf{x})\}$ is defined as

$$\Psi_{\boldsymbol{\alpha}}(\mathbf{x}) = \prod_{i=1}^{M} \Psi_{i,\alpha_i}(x_i). \tag{2.8}$$

We can now expand our model through the multivariate polynomial basis—as in Equations (2.6) and (2.7):

$$\mathcal{M}(\mathbf{x}) = \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}, \tag{2.9}$$

$$c_{\boldsymbol{\alpha}} = \frac{\langle \mathcal{M}, \Psi_{\boldsymbol{\alpha}} \rangle_{\pi}}{\|\Psi_{\boldsymbol{\alpha}}\|_{\pi}^2} \tag{2.10}$$

Using these equations, we can take the *polynomial chaos expansion (PCE)* of the

[3] A multi-index is introduced for simplification of notation. Each member α_i of the multi-index is a n-tuple of non-negative integers, which in our case are related to the order of a polynomial.

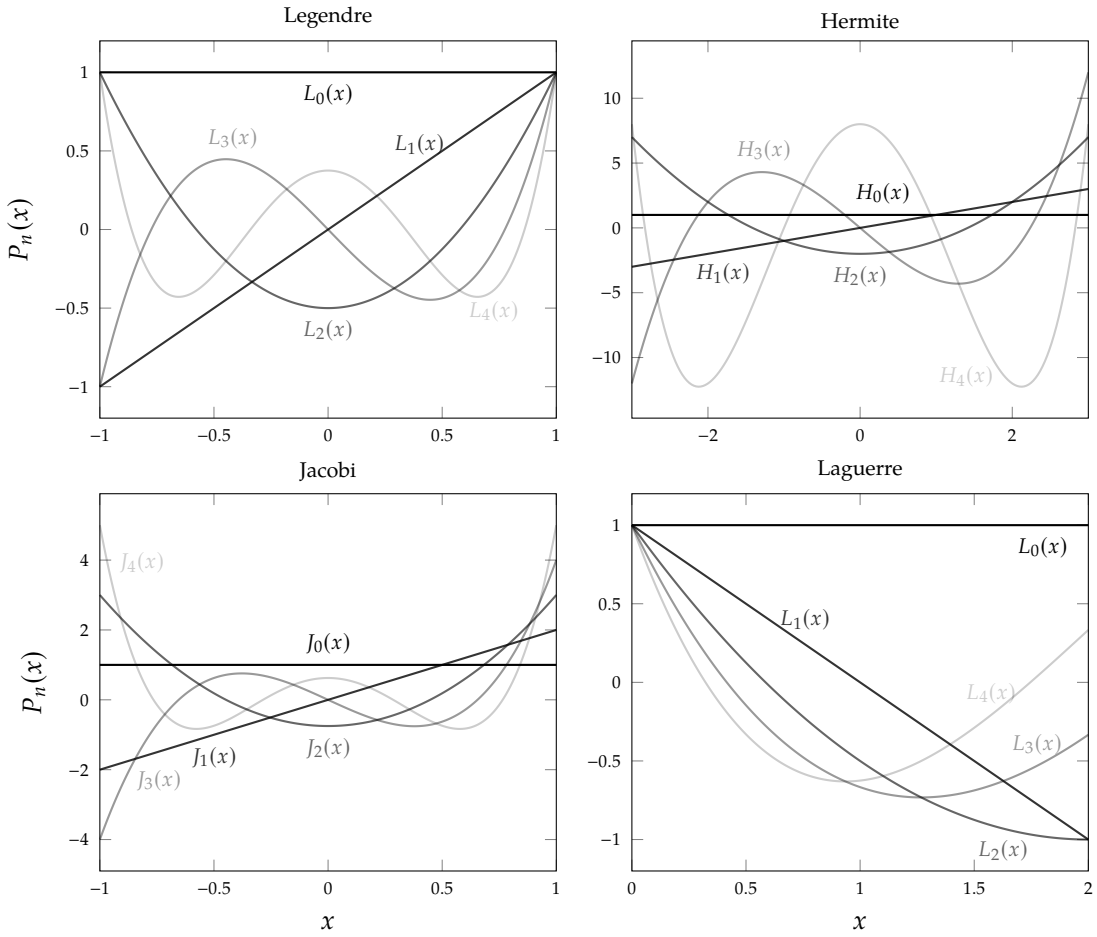

Figure 2.3: Polynomial plots of order 4 as given in Table 2.1. From top left, going clockwise: Legendre, Hermite, Jacobi, and Laguerre.

random input variable from Equation 2.2:

$$Y = \mathcal{M}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \tag{2.11}$$

As mentioned previously, the right hand side of Equation 2.11 needs to be a finite sum through a truncated PCE (Xiu and Karniadakis 2002; Lin and Karniadakis 2009). We select a finite number of terms by only selecting those with a total degree of polynomials (from Equation 2.8) smaller than a certain value p. This redefines our multi-index $\boldsymbol{\alpha}$ as $\mathcal{A}_p = \left\{\alpha : \|\boldsymbol{\alpha}\|_1 \leq p\right\}$.[4] The resulting size of this set depends on the dimensionality M of the input parameters $\boldsymbol{X}$ and the maximal degree p of

[4]The 1-norm of a multi-index is defined as $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^{M} |\alpha_i|$.

the polynomials $\Psi_{\alpha \in \mathcal{A}_p}$. This truncation leads the following equation for the PCE:

$$\mathcal{M}(\boldsymbol{X}) \approx \mathcal{M}^{PC}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_p} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}). \tag{2.12}$$

In practice, the right hand side of Equation 2.12 can be written as:

$$Y = \mathcal{M}(\boldsymbol{X}) \approx \mathcal{M}^{\mathrm{PC}}(\boldsymbol{X}) = \sum_{j=0}^{P} c_j \Psi_j(\boldsymbol{X}), \tag{2.13}$$

where j is a count of the multiindices α with a predefined order; this is discussed further in B. There are a number $P + 1$ polynomial basis functions. A typical truncation rule, a total degree truncation of degree p, i.e. $\sum_{i=1}^{M} \alpha_i \leq p$ leads to (see Xiu and Karniadakis 2002):

$$P + 1 = \frac{(M + p)!}{M! p!}, \tag{2.14}$$

As one might suspect from the factorials in Equation 2.14, the value for P grows quickly. From Figure 2.4, one can see that moderate values of M and p quickly lead to a number of terms that would require a near-prohibitive amount of simulations if the simulation time of $\mathcal{M}$ were more than a couple of hours. For example, with $M = p = 6$, one is looking at 924 basis terms in Eq. (2.11), requiring at least as many simulations of $\mathcal{M}$.

This approximation is convenient when our model $\mathcal{M}$ is physics-based because it offers an easy-to-interpret representation of the model. Each polynomial basis has terms directly related to the input parameter variables $\boldsymbol{X}$, which means that the interactions between these polynomials represent the interactions between the parameters in the models. This means that we can classify the terms in the polynomial based on the degree of the polynomials as well as their multivariate interactions, allowing for differentiation between low- and high-order interactions between variables.

Once the selection of polynomial basis has been made, one can run the polynomial chaos expansion non-intrusively. To do this, we treat the fully deterministic model $\mathcal{M}$ as a heuristic model to inform the surrogate model based on the polynomial chaos expansion. The goal is to obtain the right hand side of Equation

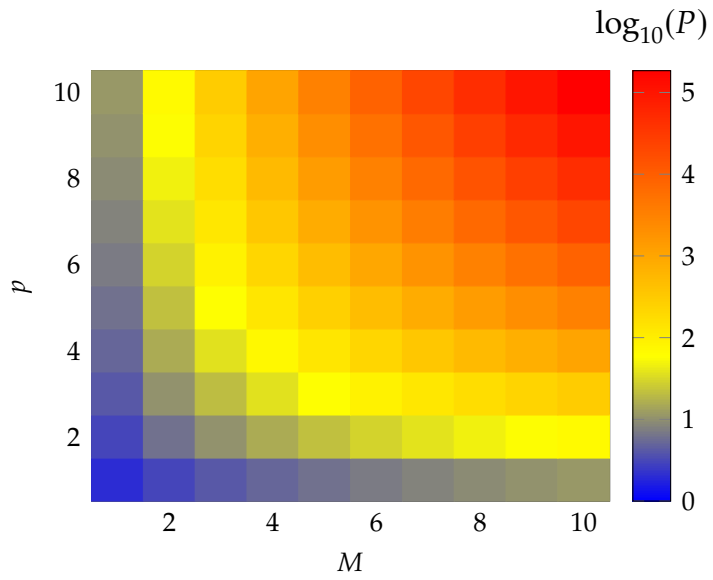

Figure 2.4: Growth of P from Equation 2.14.

(2.13), where c_j are deterministic weighting coefficients, and Ψ_j are the polynomial expansions of the order associated with the index j for realized (sampled) values of $\boldsymbol{X}$. The values for the coefficients $\mathbf{c} = \{c_0, c_1, \ldots, c_P\}$ allow the calculation of the distribution of model output Y as it was induced from the model input $\boldsymbol{X}$.

These coefficients are calculated by solving Equation (2.13) through either Gaussian quadrature (Smolyak 1963), regression (Tibshirani 1996; Blatman and Sudret 2008; Blatman and Sudret 2011; Berveiller et al. 2006), or Bayesian approaches (Sargsyan et al. 2014; Doostan and Owhadi 2011). The next step is to estimate the distribution of model output Y as informed by the uncertain (but pre-defined through marginal PDFs) model input $\boldsymbol{X}$. The approximate response of Y is obtained through sampling, often Markov-Chain Monte Carlo (MCMC) (Haario et al. 2001), because the constructed surrogate model in Equation (2.13) is much more computationally efficient than a complex original deterministic model $\mathcal{M}$. The surrogate model $\mathcal{M}^{\text{PC}}$ is assessed using samples of the random vector $\boldsymbol{X}$. These samples can be taken via, e.g., random uniform, stratified, or Latin hypercube sampling (McKay et al. 1979), where the latter is used in this work.

The above considerations result in the feasibility of calculating a polynomial chaos expansion of the model response by using Monte-Carlo or other sampling techniques (Marzouk and Xiu 2009; Eldred and Burkardt 2009). In addition to

the convenience of numerical approximation, the first two moments of the model output are obtained from the coefficients of the constructed PCE such that the mean μ and variance σ^2 of a scalar model output can be calculated as in (Le Maître and Knio 2010):

$$\mathbb{E}[\mathcal{M}(\boldsymbol{X})] = \mu_Y = c_0 \|\Psi_0\|_\pi = c_0 \quad (2.15)$$

$$\mathrm{Var}[\mathcal{M}(\boldsymbol{X})] = \sigma_Y^2 = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_p} c_{\boldsymbol{\alpha}}^2 \|\Psi_{\boldsymbol{\alpha}}\|_\pi^2 . \quad (2.16)$$

Once the surrogate PCE is constructed, the calculation of these first two moments is simply the evaluation of the polynomials in Equations (2.15) and (2.16).

Curse of dimensionality

The dimensionality of inputs to the model $\mathcal{M}$ can have a large effect on the feasibility of performing uncertainty analysis for models of complex systems. This *curse of dimensionality* (Caflisch 1998; Davis and Rabinowitz 2007) arises when dealing with geometry of high-dimensional spaces with small volumes relative to this high-dimensional space. An illustration of this behavior via the approximation of the volume for a unit sphere within a unit cube is provided in Appendix A. Within UQ applications, the curse of dimensionality arises within PCEs through high dimensional inputs (Eq. (2.14)).

Luckily, this roadblock can be remediated through *sparsity* in high dimensional spaces, meaning that there is a low-dimensional structure which carries much of the information contained within the high-dimensional space. If one is able to discover the structure of these low-dimensional structures, it can ease or even eliminate the curse of dimensionality. Methodologies have been developed to exploit this sparsity (e.g., Lee and Verleysen 2007; Ji et al. 2008; Babacan et al. 2010; Constantine et al. 2014), and we will take advantage of this in Section 2.4.1.

Multivariate output

Up until this point, we have been dealing with a single output from a model $\hat{y} = \mathcal{M}(\mathbf{x})$. We will now consider a model which maps our multivariate input $\mathbf{x}$ to

[5] Because the PDF π integrates to 1, $\|\Psi_0\| = 1$ and the mean estimate is c_0.

multivariate output vector

$$\mathbf{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{\hat{y}_1}(\mathbf{x}) \\ \mathcal{M}_{\hat{y}_2}(\mathbf{x}) \\ \vdots \\ \mathcal{M}_{\hat{y}_N}(\mathbf{x}) \end{bmatrix}, \tag{2.17}$$

where N is the number of output QoIs. Here, each model on the right hand side of Equation 2.17 can be substituted with Equation 2.12. Practically, this just means that a polynomial surrogate model needs to be computed for each output QoI. Assuming that the model that produces the QoI provides many outputs that can be manipulated into the desired QoIs, then no additional model evaluations are required to train the separate surrogate models. Therefore, the methodology naturally expands to multivariate output and the only limitations are one's ability to generate model output and the small amount of time to construct the surrogate for the desired QoIs.[6]

2.3.2 Sensitivity analysis

Once the surrogate model (2.13) has been constructed, the use of Monte Carlo methods allows the computation of Sobol' indices for global sensitivity analysis of the model to its input uncertain variables (Sobol 2001; Saltelli 2002; Sudret 2008). In the context of PCE surrogate models, estimates of Sobol' indices is gained directly from the PC surrogate, offering a convenient and computationally efficient way to determine the relative importance of uncertain inputs to the variability of the quantities of interest. Sobol' indices are split into *main* and *joint* sensitivities, where the former measures the fraction of variance in the output that can be attributed to the uncertain model input variable X_i:

$$S_i = \frac{\mathrm{Var}\left[\mathbb{E}\left[\mathcal{M}_{\mathbf{c}}^{\mathrm{PC}}(\boldsymbol{X}) \mid X_i\right]\right]}{\mathrm{Var}\left[\mathcal{M}_{\mathbf{c}}^{\mathrm{PC}}(\boldsymbol{X})\right]}, \tag{2.18}$$

[6]The time to construct the surrogate is often miniscule compared to the full model evaluations. Within this dissertation, the model simulation time is O(hours to days) whereas the time to construct a surrogate is O(seconds).

where $\mathbb{E}$ and Var are operators for expectation and variance, respectively. Similarly, the joint sensitivity measures the fraction of variance in the output that can be explained by to the joint contribution of variables X_i and X_j, and is defined as

$$S_{ij} = \frac{\text{Var}\left[\mathbb{E}\left[\mathcal{M}_{\mathbf{c}}^{\text{PC}}(\boldsymbol{X}) \mid X_i, X_j\right]\right]}{\text{Var}\left[\mathcal{M}_{\mathbf{c}}^{\text{PC}}(\boldsymbol{X})\right]} - S_i - S_j. \tag{2.19}$$

An additional benefit of using PCE machinery for the surrogate model is that the Sobol' sensitivities in Equations (2.18) and (2.19) can be calculated directly from the coefficients of the PCE using the relations from Eqs. (2.15) and (2.16), one can write the main and joint Sobol' indices in terms of the PCE coefficients. This yields an estimate for the main effect index $\hat{S}_i^{\text{main}}$ as:

$$\hat{S}_i^{\text{main}} = \frac{\sum\limits_{\boldsymbol{\alpha} \in \mathcal{A}_i^{\text{main}}} c_{\boldsymbol{\alpha}}^2 \left\langle \Psi_{\boldsymbol{\alpha}}^2 \right\rangle}{\sum\limits_{\boldsymbol{\alpha} \in \mathcal{A}, \boldsymbol{\alpha} \neq 0} c_{\boldsymbol{\alpha}}^2 \left\langle \Psi_{\boldsymbol{\alpha}}^2 \right\rangle}, \tag{2.20}$$

where $\mathcal{A}_i^{\text{main}} = \left\{\boldsymbol{\alpha} \in \text{A} : \alpha_i > 0, \alpha_{i \neq j} = 0\right\}$. Similarly, one can use the PCE coefficients to account for the total variance contribution of X_i through the estimate of the total effect index $\hat{S}_i^{\text{total}}$:

$$\hat{S}_i^{\text{total}} = \frac{\sum\limits_{\boldsymbol{\alpha} \in \mathcal{A}_i^{\text{total}}} c_{\boldsymbol{\alpha}}^2 \left\langle \Psi_{\boldsymbol{\alpha}}^2 \right\rangle}{\sum\limits_{\boldsymbol{\alpha} \in \mathcal{A}, \boldsymbol{\alpha} \neq 0} c_{\boldsymbol{\alpha}}^2 \left\langle \Psi_{\boldsymbol{\alpha}}^2 \right\rangle}, \tag{2.21}$$

where $\mathcal{A}_i^{\text{total}} = \{\boldsymbol{\alpha} \in \text{A} : \alpha_i > 0\}$. The benefit of Eqs. (2.20) and (2.21) is that, once the PCE surrogate $\mathcal{M}^{\text{PC}}$ is constructed, global sensitivity analysis via Sobol' indices can be conveniently gained by performing simple arithmetic on the coefficients of the PCE surrogate.

For both the main and joint sensitivities, the posterior distribution of the PC coefficients $\mathbf{c}$ are available. It is possible to calculate uncertainty in the sensitivity indices by sampling from the posterior distribution of $\mathbf{c}$ to calculate Equations (2.18–2.21), but this study will use only the mean estimates of the coefficients for sensitivity calculations.

2.4 Dimensionality reduction

When the model $\mathcal{M}$ being used in UQ is computationally complex, as in the case of many process-based hydrologic models, there is a large set of input parameters that can impact the model's output that could be treated as uncertain. As mentioned in 2.3.1, one is somewhat limited in the number of uncertain inputs that can be treated using PCEs due to the growth in basis terms (Eq. (2.14)). Additionally, some uncertain model inputs may not be parameters, but rather stochastic fields (e.g., stream inflow, soil moisture fields, water table fields, distribution of precipitation in both time and space). This section introduces the Bayesian compressive sensing methodology for efficiently finding the coefficients $\boldsymbol{c}$ in Eq. (2.12). The possible treatment of uncertain fields is also introduced by constructing an uncertain field from data using Gaussian process regression, then parameterizing the resulting field using Karhunen-Loève expansions.

2.4.1 Bayesian compressive sensing

The BCS approach provides marginal posterior probability distributions of the vector of coefficients in the PCE model, $\mathbf{c} = \{c_0, c_1, \ldots, c_P\}$. Given available data $\mathcal{D}$, Bayes' formula (Jaynes and Bretthorst 2003) for this situation can be written as

$$q(\mathbf{c}) \propto L_{\mathcal{D}}(\mathbf{c}) p(\mathbf{c}), \tag{2.22}$$

where $q(\mathbf{c})$ is the posterior PDF, $p(\mathbf{c})$ is the PDF representing prior information on the PC coefficient vector $\mathbf{c}$, and $L_{\mathcal{D}}(\mathbf{c})$ is the likelihood function, i.e., a measure of goodness-of-fit for the PCE surrogate model $\mathcal{M}^{\mathrm{PC}}$ from Equation (2.12) to the fully deterministic model $\mathcal{M}$. Assuming a Gaussian noise model with a standard deviation σ representing a tolerance of the discrepancy between $\mathcal{M}^{\mathrm{PC}}$ and $\mathcal{M}$ for the likelihood:

$$L_{\mathcal{D}}(\mathbf{c}) = (2\pi\sigma^2)^{(-N/2)} \exp\left[-\sum_{k=1}^{N} \frac{\left(\mathcal{M}_k - \mathcal{M}_{\mathbf{c}}^{\mathrm{PC}}(\boldsymbol{X}_k)\right)^2}{2\sigma^2} \right], \tag{2.23}$$

where $k = 1, \ldots, N$ correspond to realizations of the random input parameters $\boldsymbol{X}$. Note that Equation (2.23) implies independence of marginal likelihood functions.

The prior PDF $p(\mathbf{c})$ represents prior information on the PC coefficient vector $\mathbf{c}$,

the posterior PDF $q(\mathbf{c})$ is the outcome of the inference given the data set $\mathcal{D}$. In the case of PCEs, the prior information of $\mathbf{c}$ should be flat, so no knowledge is assumed in the calculation of the posterior for $\mathbf{c}$. A flat prior is preferred because the use of BCS motivates that many of the terms in $\mathbf{c}$ will be very close to zero, leading to a lower number of polynomial basis terms that give valuable information for the expansion in Equation (2.12).

Achieving a sparse posterior is strongly supported by sparse priors that give vanishing values for the coefficients unless there is strong evidence to the contrary. As such, this study uses the sparse Laplace prior (Babacan et al. 2010), that assumes coefficient independence:

$$p(\mathbf{c}) = \left(\frac{\beta}{2}\right)^{P+1} \exp\left(-\beta \sum_{j=0}^{P} |c_j|\right), \tag{2.24}$$

where β is a positive shape parameter that also controls the optimization problem in Equation (2.25). The vector $\mathbf{c}$ that maximizes the posterior $q(\mathbf{c})$ is given by the solution to

$$\arg\max_{\mathbf{c}} \left(\log L_{\mathcal{D}}(\mathbf{c}) - \beta||\mathbf{c}||_1\right), \tag{2.25}$$

which is the compressive sensing algorithm used in signal processing (Candès et al. 2006), where the $-\beta \|\mathbf{c}\|_1$ term is due to the l_1 norm-based regularization approach in BCS. The regularization approach is used to reduce overfitting while learning the coefficients $\mathbf{c}$. Different approaches would lead to different regularization terms in Equation (2.25). Details of the implementation of this approach are left to (Sargsyan et al. 2014), but one of the key points is the selection of stopping criterion. Specifically, the algorithm iterates finding the basis terms $\mathbf{c}$ until it reaches a stopping criterion ϵ comparing the relative change in the maximal value of the evidence E — the integrated likelihood. The stopping criterion is defined as $(E^n - E^{n-1})/(E^n - E^1) < \epsilon$, where n is the iteration number. As ϵ decreases, more iterations are required, meaning that fewer basis terms are retained in the final polynomial surrogate. A discussion on the selection of ϵ is included in Section

3.3.1.

For example, in (Sargsyan et al. 2014), the Community Land Model with carbon-nitrogen cycling (Thornton et al. 2007) was modeled with 79 uncertain input parameters, where second, third, and fourth-order polynomials lead to order 10^3, 10^5, and 10^6 basis terms, respectively. Using 10,000 model simulations, the BCS methodology demonstrated an excellent performance skill for a very large uncertain parameter set by avoiding the calculation of all these basis terms and selecting only those relevant to the QoIs. During testing, 17 model simulations failed resulting in 9,983 simulations used for training the surrogate model. This displays another advantage of the BCS method, as failed simulations do not limit the solution of the coefficients as it would in quadrature methods. The methodology also allows additional simulations to be added if an initial simulation set is determined to be insufficient to train the surrogate model.

2.4.2 Uncertain input fields

In hydrology, many inputs to a model can be thought of as a field. These could be fields in space: soil moisture, hydraulic conductivity, water table, soil type, land cover, etc, and could also be thought of in time, e.g., meteorologic time series. In order to represent these fields under a formal uncertainty quantification framework, approaches are needed to represent these fields in a manner which can be represented with polynomial chaos machinery. This approach requires the reduction of these complex fields into a small number of parameters (Salamon and Feyen 2009).

In order to express the uncertainty in these functions (fields), we need to assign some probability measures on the fields. These functions or fields can express uncertainties in external forcing, initial conditions, field parameters, or boundary conditions. Once the uncertainties in these quantities are quantified, they can be propagated using simulation to understand outcomes of the simulations, e.g., flooded extent, cumulative evapotranspiration, inundation risk, etc. A convenient way of quantifying the uncertainties in these fields is to use Gaussian processes (Rasmussen 2006).

Gaussian processes

A Gaussian process, in its simplest form, approximates a function (or field) based on the inputs and outputs to that function. In Eq. (2.26), $f(\mathbf{x})$ is treated as uncertain, and we can express beliefs about $f(\mathbf{x})$ as $f(\mathbf{x}) \sim P(f(\mathbf{x}))$. The GP $f(\mathbf{x})$ requires a mean and covariance function. Having these allows the definition of a probability measure on the space of the function:

$$f(\mathbf{x}) \sim P(f(\mathbf{x})) = \mathcal{GP}\left(f(\mathbf{x}) \mid m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right). \tag{2.26}$$

The mean function describes what the values of $f(\mathbf{x})$ are, and models the expectation of $f(\mathbf{x})$:

$$m(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x})\right].$$

When there are two different inputs, $\mathbf{x}$ and $\mathbf{x}'$, the covariance function models the belief of similarity between $f(\mathbf{x})$ and $f(\mathbf{x}')$, i.e., how close the corresponding outputs are:

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right].$$

In order to express a function as a GP we can write

$$f(\cdot) \sim \mathcal{GP}\left(f(\cdot) \mid m(\cdot), k(\cdot, \cdot)\right) \tag{2.27}$$

Gaussian process regression

If we take the input features $\mathbf{x} \in \mathbb{R}^d$, then we likely have some intuition about a function of interest $f(\mathbf{x})$. For instance, if $f(\mathbf{x})$ is the precipitation intensity for certain times $\mathbf{x}$, then we know that $\min\left[f(\mathbf{x})\right] = 0$. This is prior knowledge or a prior belief for that process. Saying that $f(\mathbf{x})$ is a GP means that it is a random variable and a function, and can be written

$$f(\mathbf{x}) \mid m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}') \sim \mathcal{GP}\left(f(\mathbf{x}) \mid m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right) \tag{2.28}$$

where $f(\mathbf{x})$ is a random function, $m(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is a mean function, and $k(\mathbf{x}, \mathbf{x}') : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a covariance function.

A GP comes from the multivariate Gaussian distribution. Let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$,

be n points in $\mathbb{R}^d$. Let $\mathbf{f} \in \mathbb{R}^n$ be the output of $f(\mathbf{x})$ on each element of $\mathbf{x}$,

$$\mathbf{f} = \begin{pmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{pmatrix}.$$

Since $f(\mathbf{x})$ is a GP with mean and covariance function, we can treat the vector of outputs $\mathbf{f}$ at the inputs $\mathbf{x}$ as multivariate-normal.

$$\mathbf{f} \mid \mathbf{x}, m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}') \sim \mathcal{N}\left(\mathbf{f} \mid m(\mathbf{x}), K\left(\mathbf{x}, \mathbf{x}'\right)\right) \tag{2.29}$$

where $\mathbf{m}$ is the mean vector

$$\mathbf{m}(\mathbf{x}) = \begin{pmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_n) \end{pmatrix},$$

and $\mathbf{K}$ is the covariance matrix

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}.$$

The mean function, $m(\mathbf{x})$, gives the most probable value for $f(\mathbf{x})$, e.g., $m(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x})\right]$.

The covariance function, $k(\mathbf{x}, \mathbf{x}')$, describes the similarity of the input space. Choosing which covariance function to use has been widely discussed (e.g., Rasmussen 2006; Berrocal et al. 2008; Kennedy and O'Hagan 2000). If we take the vector $\mathbf{x}$, then $k(\mathbf{x}, \mathbf{x})$ is the variance of the random variable $f(\mathbf{x})$, i.e., $\mathbb{V}\left[f(\mathbf{x})\right] = \mathbb{E}\left[\left(f(\mathbf{x}) - m(\mathbf{x})\right)^2\right]$. For $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, the covariance function $k(\mathbf{x}, \mathbf{x}')$ explains how the random variables $f(\mathbf{x})$ and $f(\mathbf{x}')$ are correlated, and can be expressed as

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{C}\left[f(\mathbf{x}), f(\mathbf{x}')\right] = \mathbb{E}\left[\left(f(\mathbf{x}) - m(\mathbf{x})\right)\left(f(\mathbf{x}') - m(\mathbf{x}')\right)\right].$$

Properties of the covariance function There are a few important properties of the covariance function: 1) via the definition of $k(\mathbf{x})$ as the variance of the random variable $f(\mathbf{x})$, we can say that for any $\mathbf{x} \in \mathbb{R}^d$, $k(\mathbf{x}, \mathbf{x}) \geq 0$; and 2) the covariance

matrix $K(\mathbf{x}, \mathbf{x}')$ is positive definite for $\mathbf{x} \in \mathbb{R}^{n \times d}$. It is often assumed that the covariance $k(\mathbf{x}, \mathbf{x}')$ is a monotonically decreasing function of the distance between $\mathbf{x}$ and $\mathbf{x}'$, i.e., $k(\mathbf{x}, \mathbf{x}') = \tilde{k}(|\mathbf{x} - \mathbf{x}'|)$, with $\tilde{k}(\cdot)$ being a decreasing function. This is not a necessary condition for covariance functions (Rasmussen 2006), but is often a logical assumption for physical systems. Commonly used covariance functions that follow this assumption are the squared exponential or Matérn class of covariance functions.

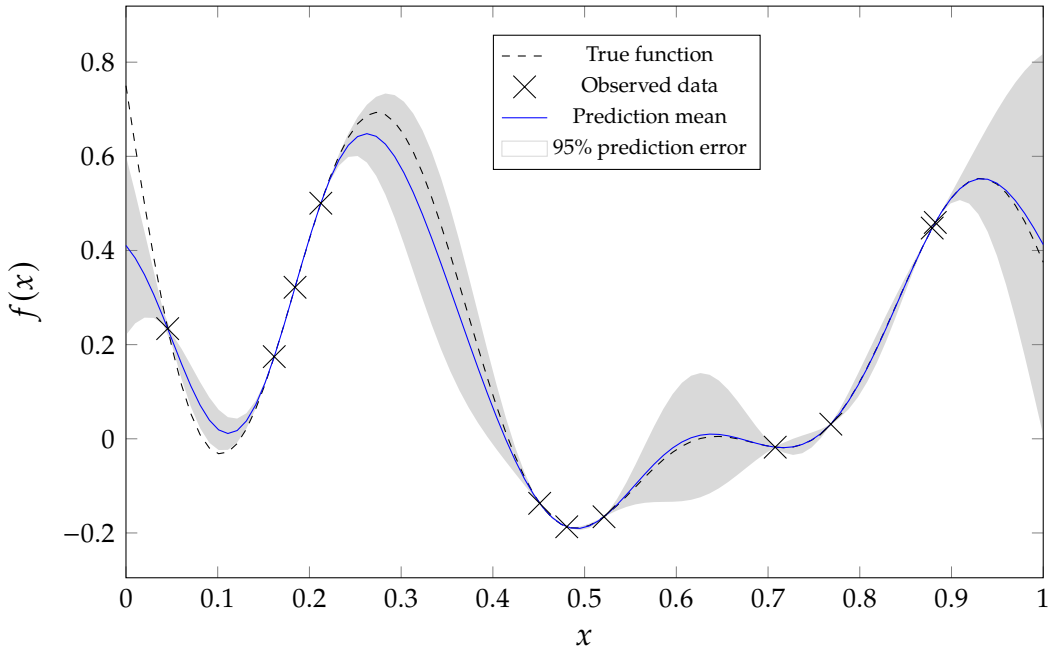

Figure 2.5: Example fit of a Gaussian process regression for the function $f(x) = (1.125x - 3/4)\,[\sin(5\pi x) - \cos(\pi x)]$ using eleven sampling points.

Sampling a Gaussian process

If we represent the Guassian process as

$$P(f(\cdot) \mid I) = \mathcal{GP}\left(f(\cdot) \mid m(\cdot), k(\cdot, \cdot)\right), \tag{2.30}$$

where I is known information about the process (observed input/output data, for example). Once we have this GP, it can be sampled from by selecting test locations, $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. The function values at these locations is then $\mathbf{f} = \left\{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)\right\}$. Then, by definition, the probability of these function values at

the test locations is normally distributed with a mean of $\mathbf{f}$:

$$P(\mathbf{f} \mid \mathbf{x}, I) = \mathcal{N}(\mathbf{f} \mid \mathbf{m}, \mathbf{K}).$$

Given that there are n locations at which the function is evaluated, we need n random variables in order to sample from this. This is often intractable given grid sizes, and is especially so in hydrology. If we have a domain that is 100×100, then we would need 10,000 random variables in order to represent the GP, which bring in the curse of dimensionality. Fortunately, it is possible to describe (2.30) using just a few random variables using Karhunen-Loève expansions.

Karhunen-Loève expansions

A Karhunen-Loève expansion (Karhunen 1946; Zheng and Dai 2017) provides a method of presenting a stochastic process with a small number of random variables. If we take the GP from (2.30), it can be expressed as a Karhunen-Loève expansion

$$f(\mathbf{x}; \mathbf{X}) = m(\mathbf{x}) + \sum_{i=1}^{\infty} X_i \sqrt{\lambda_i} \varphi_i(\mathbf{x}), \tag{2.31}$$

where:

- X_i are standard normal variables, i.e., $\mathbf{X} = \{X_1, X_2, \cdots\} \sim \mathcal{N}(0, 1)$.
- λ_i and $\varphi_i(\mathbf{x})$ are the eigenvalues and eigenfunctions of the covariance function, respectively. Given that the covariance matrix is positive definite, the eigenvalues λ_i will be positive and therefore $\sqrt{\lambda_i}$ is well-defined.

The Karhunen-Loève expansion can be truncated after a number of terms, and still reasonably express the stochastic process, if the eigenvalue term — $\sqrt{\lambda_i}$ — is sufficiently small. The ordering of Eq. (2.31) is done by the magnitude of λ, and truncation depends on how fast value of λ_i decrease as i increases. This truncation offers the reduction of dimensionality and we can then write the truncated Karhunen-Loève expansion as

$$f(\mathbf{x}; \mathbf{X}) = m(\mathbf{x}) + \sum_{i=1}^{d} X_i \sqrt{\lambda_i} \varphi_i(\mathbf{x}). \tag{2.32}$$

The decision for which value of d to use is based on capturing a desired amount

of variance, or energy, of the field

$$\sum_{i=1}^{d} \lambda_i = a \sum_{i=1}^{\infty} \lambda_i,$$

where a is chosen based on the desired amount of captured variance, and is typically chosen to be a value of around $a = 0.95$ (Zheng and Dai 2017).

A graphical representation of the Karhunen-Loève expansion is given in Figure 2.6. This figure shows the construction of a single realization of the stochastic process determined from the GP regression in Fig. 2.5. Here $m(\mathbf{x})$ is the mean from the GP regression, and the remaining terms are the terms in the summation of (2.32), where $X_i \sim \mathcal{N}(0, 1)$.

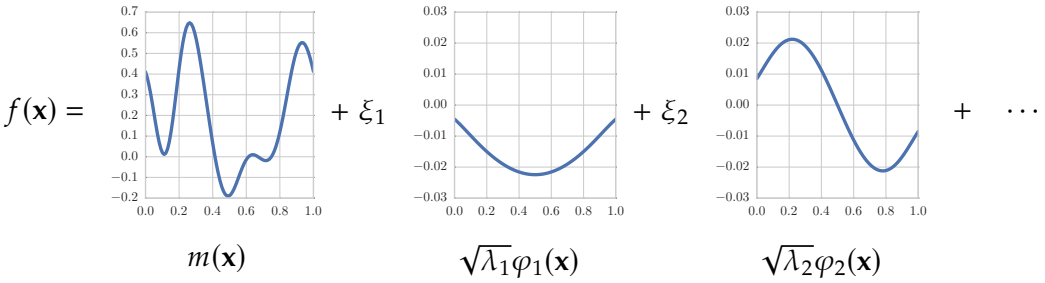

Figure 2.6: Construction of a Karhunen-Loève expansion using the first two terms in the summand.

2.5 Inference

The previous sections displayed a framework which allows propagation of parametric and field uncertainties to model outputs, and possible ways to efficiently use many uncertain inputs to a hydrologic model using approaches to dimensionality reduction. This section deals with the inverse question: *given some observed output data, how can this information be used to learn values of model parameters that would produce the observed data?* The workhorse for this purpose will be Bayesian inference, which provides a probabilistic framework to quantify uncertainty with what is often little data availability in hydrology. Bayesian inference's main feature, with respect to this work, is that it quantifies the degree uncertainty before and

after introduction of observed data through the prior and posterior probability distributions.

Much like PCEs, Bayesian probability has been widely used with increases in computational power. Named after its original formulator, the British Reverend Thomas Bayes, though popularized by the French mathematician Pierre-Simon Laplace, Bayesian probability has been used for over two centuries to answer questions about gambling, cryptography, location of lost nuclear weapons, and the authorship of the Federalist Papers (McGrayne 2011). It has gained popularity due to its ability and flexibility in quantifying uncertainties in a wide range of complex problems.

This section briefly introduces Bayesian inference based on the likelihood and the prior and posterior distributions. Then, it introduces the inverse problem and how it applies to hydrology and the methodology outlined in the previous sections of this chapter.

2.5.1 Introduction of Bayesian inference

Bayesian inference relies on the application of Bayes' theorem in Equation 1.1. To make the notation more consistent with the previous sections, consider M unknown parameters of a statistical model as $\mathbf{x} = (x_1, x_2, \ldots, x_M)$. Furthermore, there are N measurements of observed data which is the target of the statistical model $\boldsymbol{\mathcal{D}} = (y_1, y_2, \ldots, y_N)$. Bayes' theorem can then be rewritten from Eq. 1.1 in terms of unknown parameters and observed data:

$$p(\mathbf{x} \mid \boldsymbol{\mathcal{D}}) = \frac{p(\boldsymbol{\mathcal{D}} \mid \mathbf{x})p(\mathbf{x})}{p(\boldsymbol{\mathcal{D}})}. \tag{2.33}$$

To gain any valuable information about the unknowns $\mathbf{x}$ from the observed quantities $\boldsymbol{\mathcal{D}}$, there needs to be a relationship relating the two:

$$Y \mid \mathbf{x} \sim p(\boldsymbol{\mathcal{D}} \mid \mathbf{x}). \tag{2.34}$$

Here, as in Equation 2.3, Y is a random variable. In this case, it is treated as a random realization of the observed values.[7]

[7]It should be noted that Bayesian inference was used in Section 2.4.1 for surrogate construction through the coefficients $\boldsymbol{c}$. This section focuses on using Bayesian inference of model parameters

Likelihood

The probabilistic model relating the observed to the unknown (Equation 2.34) is known as the *likelihood function*, and it is defined as

$$\mathcal{L}(\mathbf{x}) = p(\boldsymbol{\mathcal{D}} \mid \mathbf{x}). \tag{2.35}$$

The likelihood function incorporates assumptions about the discrepancy of the data $\mathcal{D}$ and the model outputs stemming from input $\boldsymbol{x}$. As such, any data added irrelevant to the parameters $\mathbf{x}$ will not affect the value of the likelihood (Jaynes and Bretthorst 2003).

Furthermore, one can get a point estimate of the unknown parameters, the maximum likelihood estimate, as

$$\hat{\mathbf{x}}_{\text{MLE}} = \arg\max_{\mathbf{x}} \mathcal{L}(\mathbf{x}). \tag{2.36}$$

That is, the values of $\mathbf{x}$ that maximize the likelihood in Equation 2.35. However, these are point estimates and are unable to capture the uncertainty in parameter estimates, which is one motivating factor behind using Bayesian inference.

Prior distribution

In addition to imparting some randomness to the data $\boldsymbol{\mathcal{D}}$ (Equation 2.34), within the Bayesian framework, randomness is also imparted into the unknown parameters $\mathbf{x}$ The uncertainty around the values of the parameters before data is accounted for is taken into account by assigning the random vector

$$\boldsymbol{X} \sim p(\mathbf{x}), \tag{2.37}$$

where $p(\mathbf{x})$ is the *prior density*. This representation means that the epistemic uncertainty around the parameter values are modeled as a probability distribution. As with the observed data, the "true" values of unknown parameters are a realization of the random variable $\boldsymbol{X}$.

The prior distribution represents the knowledge of the modeler of the distribution of the unknown parameters. The selection of the prior has long been a

with hydrologic observations.

controversial aspect of Bayesian inference. In constructing the prior, one can include both qualitative and quantitative information, including expert knowledge as well as physical knowledge (e.g., a prior on a mass of an object cannot be negative). Distilling diverse sources of information into a probability distribution is not straightforward exercise. Additionally, one can use the posterior from previous studies or experiments to inform the prior, a practice called sequential updating (Spiegelhalter and Lauritzen 1990; Kong et al. 1994).

In practice, one often selects from standard families of distributions to construct the prior in order to represent the uncertainty. For example, if an uncertain parameter were to come from recorded sensor data on which the sensor's specification stated that the errors in the observed values are normally distributed, it would make the most sense to use a Gaussian distribution for the prior. Uniform distributions are used for parameters where there is believed to be some minima or maxima creating bounded values, and one can use Guassian distributions for unbounded parameters, or lognormal for parameters which are positive. Deciding on logical probabilistic mappings for a model's parameters is an important part of experimental design and is based on known values of the parameters in question and expert knowledge about logical distributions a parameter may follow.

Posterior distribution

Given the marginal distribution of unknown parameters (Equation 2.37) and the conditional distribution of the observed data (Equation 2.34), one can write the joint distribution of the unknown parameters and data:

$$\boldsymbol{Y}, \boldsymbol{X} \sim p(\boldsymbol{\mathcal{D}}, \mathbf{x}) = p(\boldsymbol{\mathcal{D}} \mid \mathbf{x}) p(\mathbf{x}) \tag{2.38}$$

This is the full probability model, where the true parameter values and observed data are realized. However, the true parameter values $\mathbf{x}$ are still unknown. But, given the likelihood function and the prior density, the *posterior density* can be calculated using Bayes' theorem:

$$p(\mathbf{x} \mid \boldsymbol{\mathcal{D}}) = \frac{\mathcal{L}(\mathbf{x}) p(\mathbf{x})}{\int \mathcal{L}(\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x}} \,, \tag{2.39}$$

where the term in the denominator is a normalizing constant called the *marginal likelihood*, which is used to ensure $p(\mathbf{x} \mid \mathcal{D})$ is a proper probability denity.

This formulation allows for an updating of the posterior density when new information is added (or removed) in Equation 2.39. This updating from the prior to posterior is shown for a single unknown in Figure 2.7. One of the goals of this updating, as seen in the figure, is to reduce the epistemic uncertainty of the unknown. In this figure, one can see that the mass of the parameter is becoming more concentrated, leading to higher probabilities. In complex models, the posterior is unlikely to be a simple probability distribution like that shown in the figure, and could possess higher order moments.

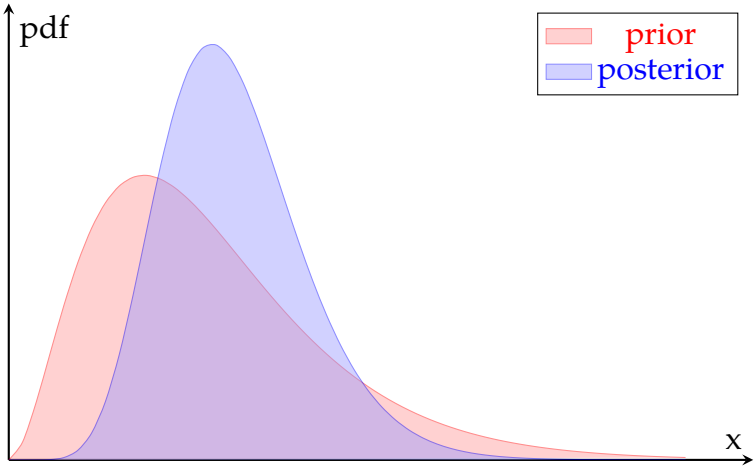

Figure 2.7: Prior and posterior density function. A goal of inference is to potentially decrease uncertainty about input densities, e.g., reducing the variance as shown here.

Given the posterior, one can gain information for a given QoI h, such as its expected value:

$$\mathbb{E}\left[h(\boldsymbol{X}) \mid \mathbf{y}\right] = \int h(\mathbf{x}) p(\mathbf{x} \mid \mathbf{y}) \, \mathrm{d}\mathbf{x} \tag{2.40}$$

In particular, one can calculate the first two moments of the posterior, where the expected value and covariance matrix are given by

$$\mathbb{E}\left[\boldsymbol{X} \mid \mathbf{y}\right] = \int \mathbf{x} p(\mathbf{x} \mid \mathbf{y}) \, \mathrm{d}\mathbf{x}, \tag{2.41}$$

$$\mathrm{Cov}\left[\boldsymbol{X} \mid \mathbf{y}\right] = \int (\mathbf{x} - \mathbb{E}\left[\boldsymbol{X} \mid \mathbf{y}\right]) (\mathbf{x} - \mathbb{E}\left[\boldsymbol{X} \mid \mathbf{y}\right])^{\mathrm{T}} p(\mathbf{x} \mid \mathbf{y}) \, \mathrm{d}\mathbf{x}, \tag{2.42}$$

where the posterior mean can be taken as a point estimate of the parameter value

and the covariance quantifies the uncertainty in the estimate. Another measure for point estimation is the *maximum a posteriori* (MAP) estimate, which is the value that maximizes the posterior density, or more simply, the mode of the posterior distribution

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}) \tag{2.43}$$

In the analog to frequentist statistics, one can also estimate the *credible region* for the posterior mass. To calculate this, one needs to calculate if a certain set contains the parameter value in question, or the posterior probability that $\boldsymbol{X} \in A$:

$$\mathbb{P}_{\boldsymbol{X}|\boldsymbol{Y}}(A \mid \mathbf{y}) = \int_A p(\mathbf{x} \mid \mathbf{y}) I_A \, \mathrm{d}\mathbf{x}, \tag{2.44}$$

where A is a set of potential parameter values being investigated and I_A is an indicator function where

$$I_A = \begin{cases} 1: & \boldsymbol{X} \in A \\ 0: & \text{otherwise} \end{cases}.$$

Using Equation 2.44, one can calculate the probability that the parameter value belongs to a set, allowing for the calculation of credible regions.

2.5.2 Inverse modeling

Inverse modeling is a class of problems where, given some data and a parameterized model, determine the "true" values of parameters of the model given the observed data. Here, "true" is in quotes because they are the parameters for the model, not necessarily the actual values in the studied domain. Consider two hydrologic models, $\mathcal{M}_A$ and $\mathcal{M}_B$. If one performs an inverse problem for a parameter of saturated conductivity using $\mathcal{M}_A$, the posterior value of saturated conductivity will not necessarily fit the observed data the best when used in model $\mathcal{M}_B$. Nevertheless, getting the value for the model $\mathcal{M}_A$ is valuable as it provides more information about the parameter, and therefore the domain, than existed previously, while also better calibrating $\mathcal{M}_A$ for future studies.

Inverse problems have been used in many disciplines, including geosciences, engineering mechanics, imaging, and hydrology (e.g., Butler et al. 2015; Elsheikh et

al. 2014b; Marzouk and Najm 2009; Marzouk et al. 2007; McLaughlin and Townley 1996; Neuman et al. 1980; Yeh 1986; Idier 2013; Kowalsky et al. 2004). In these cases, the numerical model is used to provide a mapping from the observed quantity of interest and the parameter which cannot be observed which controls some aspect of the behavior of that quantity of interest. Running mechanistic, process-based models forward, i.e., defining the model inputs and calculating the physical behavior of the outputs is a "well-posed" problem. In the mathematical sense, a well posed problem is one where the solution exists, is unique, and stable to changes in initial conditions (Hadamard 1902). As long as the model is robust to changes in initial conditions, forward hydrologic modeling is generally well-posed. However, the inverse problem is generally ill-posed, which means that the above conditioned may not be met, i.e., a solution may not exist or not be unique, and could be discontinuous given the data. In such a case, the ill-posed problem must be regularized (Tikhonov and Arsenin 1977), where treating the problem in a Bayesian frame and assigning priors is viewed as a form of regularization (Idier 2013).

Again, consider a model $\mathcal{M}(\boldsymbol{X}, \mathbf{t})$ that maps the variables $\boldsymbol{X} \in \mathbb{R}^M$ and $\mathbf{t} \in \mathbb{R}^d$ to the observables $\hat{\mathbf{u}}$, where $\boldsymbol{X}$ are the unknown model parameters to be found and $\mathbf{t}$ are well-determined experimental conditions such as locations in space or time that are shared between the model and data. Additionally, it is assumed that the model gives an adequate approximation of the observed values, $\mathbf{u} \approx \mathcal{M}(\boldsymbol{X}; \mathbf{t})$, and it is common to that there is some residual error η such that

$$\mathbf{u} = \mathcal{M}(\boldsymbol{X}, \mathbf{t}) + \eta, \tag{2.45}$$

where η combines both prediction errors and noise in the observed data. There are ways to modify this assumption and separate model error from data error (e.g., Sargsyan et al. 2018), however the representation in Equation 2.45 is adequate for the work in this thesis.

In most inversion problems, this residual error is modeled as a random vector $\boldsymbol{\mathcal{H}}$, of the same length as $\mathbf{u}$, often following a normal distribution:

$$\boldsymbol{\mathcal{H}} \sim \mathcal{N}(0, \boldsymbol{\Sigma}). \tag{2.46}$$

From Equation 2.46, it is assumed that the residual error is mean zero with a

covariate matrix $\mathbf{\Sigma}$, where the covariate structure is often dependent upon the experimental conditions or measurements points, such that $\mathbf{\Sigma} = \mathbf{\Sigma}(\mathbf{t})$, i.e., a seasonal or spatial relationship between an observable like streamflow or soil moisture.

The use of Bayes' rule allows for the computation of the posterior parameter values conditioned on the observed data:

$$\Pi\left(\boldsymbol{X} \mid \mathbf{u}\right) \propto L\left(\mathbf{u} \mid \boldsymbol{X}\right) p(\boldsymbol{X}), \tag{2.47}$$

where $p(\boldsymbol{X})$ is the prior distribution, $L(\mathbf{u} \mid \boldsymbol{X})$ is the likelihood function which represents the probability of obtaining the data given the set of parameters, and $\Pi(\boldsymbol{X} \mid \mathbf{u})$ is the posterior distribution for $\boldsymbol{X}$, which represents the probability of having the parameter values given the observed data.

To formulate a likelihood function, one must represent the discrepancy between the model and observations: $\eta = \mathbf{u} - \mathcal{M}$. Assuming that the components of η are independent and identically distributed random variables with some marginal density p_η, the likelihood function can be written as

$$L(\mathbf{u} \mid \boldsymbol{X}) = \prod_{d=1}^{D} p_\eta(u_d - \mathcal{M}_d(\boldsymbol{X})), \tag{2.48}$$

where there are D conditions (e.g., time snapshots of measured streamflow, monthly evapotranspiration, etc.) that are being used for inference.

If one assumes that the errors η_d are independent and normally distributed $\eta_d \sim N(0, \sigma^2)$, Equation (2.48) can be written as:

$$L(\mathbf{u} \mid \boldsymbol{X}) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^D} \prod_{d=1}^{D} \exp\left[-\frac{(u_d - \mathcal{M}_d(\boldsymbol{X}))^2}{2\sigma^2}\right], \tag{2.49}$$

where the logarithm of this likelihood function corresponds to the least-squares form of the objective function often used for deterministic parameter estimation (Sargsyan et al. 2015).

If the variance or error in measurements are not know, it can be valuable to introduce σ^2 as a hyper parameter for the likelihood (Sargsyan et al. 2015) and

rewrite the joint posterior of Equation (2.47) as

$$\Pi\left(\boldsymbol{X}, \sigma^2 \mid \mathbf{u}\right) \propto \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{d=1}^{D} \exp\left[\frac{-(u_d - \mathcal{M}_d(\boldsymbol{X}))^2}{2\sigma^2}\right] p(X_1) \cdots p(X_n) p(\sigma^2). \quad (2.50)$$

The prior for the model parameters $p(X_i)$ are based on their *a priori* knowledge, e.g., that the parameters are uniform within a range or normally distributed with some mean and variance. As the variance of the error noise, σ^2 must be positive, we therefore use a Jeffreys prior (Jaynes and Bretthorst 2003):

$$p(\sigma^2) = \begin{cases} \frac{1}{\sigma^2} & \text{for } \sigma^2 > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2.51)$$

To infer values for uncertain model parameters, the posterior from Equation (2.50) needs to be sampled using methods such as the Metropolis-Hastings MCMC (MacKay 1998). Note that sampling from posterior requires repeated evaluation of the likelihood, implying multiple evaluations of the model $\mathcal{M}$. This tends to be computationally expensive, and it is therefore expeditious to replace the model $\mathcal{M}$ with its PCE surrogate $\mathcal{M}^{\text{PC}}$. In this way, the methodology outlined previously in this section can be combined with those in Sections 2.3 and 2.4 to create a computationally efficient, flexible framework to infer uncertain parameter values for a complex, process-based hydrologic model with multiple inputs applied to a sparsely monitored watershed. The benefit of the methodology is that it is able to answer the question of inverse inference: *what is the distribution of the model's uncertain parameters given observed data*? Importantly, it is also able to address the question of: *what are the possible outcomes for certain quantities of interest given the uncertainty in the model's inputs*?

2.6 Uncertainty quantification workflow

A diagram outlining the uncertainty quantification workflow is provided in Figure 2.8, encompassing the methods outlined in this chapter, with the exception of Section 2.4.2.

2.6.1 Example of workflow applied to a kinematic wave

To demonstrate the efficacy of the workflow in Fig. 2.8, we will consider an example applied to a flow problem with the kinematic wave equation. The kinematic wave equation is a simplification of the one-dimensional Saint-Venant equations assuming no inertial terms and locally uniform flow, and is represented as:

$$\frac{\partial h}{\partial t} + w(h)\frac{\partial h}{\partial x} = 0, \tag{2.52}$$

where h is the flow depth, t is time, and $w(h)$ is the kinematic wave speed:

$$w(h) = \frac{1}{n} m \sqrt{S_0} h^{m-1}, \tag{2.53}$$

where n is Manning's roughness coefficient, $m = 5/3$ is an empirical parameter corresponding to the Manning equation, and S_0 is the surface slope.

For the example problem, consider a 60×60 [m] parking lot that drains into a gutter. The slope in the direction normal to the gutter is $S_0 = 0.005$, with no slope in the direction of the gutter. The surface is characterized by a Manning coefficient of $n = 0.01$, roughly corresponding to debris-free concrete. Consider a rainfall event from times $t = [0, 180]$ [s], with an intensity $R = 50$ [mm hr^{-1}]. With this information, we'd like to calculate the outflow hydrograph per unit width for $t < 1200$ [s].

To construct the hydrograph, we need to consider the time of concentration of the parking lot, i.e., the amount of time required for a drop of water to travel from the furthest hydrologic point to the gutter. The time of concentration is calculated as:

$$t_c = \left[\frac{L}{\alpha R^{m-1}} \right]^{1/m}, \tag{2.54}$$

where $\alpha = \sqrt{S_0}/n$. For this specific example, the time of concentration is greater than the time of duration ($t_d = 180$ [s]) of the rainfall, so the hydrograph has three

sections, calculated as:

$$Q(t) = \begin{cases} \alpha(Rt)^m & \text{for } t < t_d \\ \alpha(Rt_d)^m & \text{for } t_d \leq t \leq t^* \\ \text{solve } \frac{Q(t)}{R} + mQ(t)^{1-1/m}\alpha^{1/m}(t - t_d) - L = 0 & \text{for } t > t^* \end{cases} \tag{2.55}$$

where t^* is the time where the recession curve starts, and is calculated by:

$$t^* = t_d \left[1 + \frac{1}{m} \left(\left(\frac{t_c}{t_d} \right)^m - 1 \right) \right]. \tag{2.56}$$

In the last case of Eq. (2.55), $Q(t)$ needs to be iteratively solved for to satisfy that equation.

The UQ setup for this problem is as follows, the slope and Manning's coefficient are treated as uncertain, such that $S_0 = X_1 \sim U[0.0045, 0.0055]$ and $n = X_2 \sim U[0.006, 0.014]$. Values for S_0 are treated as less uncertain since the slope is an easily-measured quantity, whereas Manning's coefficient (n) is an empirical value and therefore has more associated uncertainty. A simple script to calculate the outflow is used for $\mathcal{M}(\boldsymbol{X})$, and 50 training simulations were performed to construct fourth-order surrogate $M^{\text{PC}}(\boldsymbol{X})$ through Bayesian compressive sensing.[8] An additional eight simulations are used to validate surrogate performance.

The quantities of interest are selected to be the outflow at one-minute intervals up to a time of 20 minutes. A residual error η from Eq. (2.45) is added to the analytic solution such that $\eta \sim N(0, 0.025Q(t))$. This error term is added in order to represent outflow data with a small measurement error. These simulated data and associated quantities of interest are use to perform parametric inference for S_0 and n (box B of Fig. 2.8). The posterior values for these parameters are then propagated through $\mathcal{M}(\boldsymbol{X})$ to compare to the analytic solution of $Q(S_0 = 0.005, n = 0.01)$.

The analytic solution, quantities of interest, and data used for inference are shown in Fig. 2.9a. The median of the training simulations shows good agreement with the analytic solution, but a slight underestimation of the analytic solution between times 180 to 360 seconds.

[8] With a low number of uncertain variables, alternative methods to Bayesian compressive sensing (mentioned in Section 2.3.1) will require fewer training simulations to construct a surrogate as there are only $P + 1 = 15$ basis terms. However, BCS is still used here to demonstrate the workflow shown in the chapter.

Parametric inference is performed to get the posterior values in Fig. 2.9b, which indicates a near-uninformative posterior for S_0. This corresponds to the sensitivity information in Fig. 2.9c, which shows that the total sensitivity (Eq. (2.21)) of $Q(t)$ is dominated by the change in values of n, with S_0 having a very small ($< 5\%$) contribution for all times. Since changes in S_0 have little effect on the variation of $Q(t)$, its posterior changed little from the uniform prior. The opposite can be said for n, which has a large impact on the variability of $Q(t)$, and therefore its posterior is far different from its prior, with a MAP and mean estimate of 0.0101197 and 0.01019415, respectively. The error in the MAP estimate from the true value of n is 1.2%, showing good agreement between the posterior parameter value and the true value used to generate data for inference.

The posterior values for S_0 and n in Fig. 2.9b are then propagated through $M(\boldsymbol{X})$ (box C of Fig. 2.8) in order to obtain $Y = p(\mathcal{M}(\boldsymbol{X}) \mid \mathcal{D})$ for the QoIs, shown in Fig. 2.9d. This plot shows very good agreement between the analytic solution and the median posterior values. To quantify the agreement, consider the relative L_2 norm between the analytic and computed solutions:

$$L_{2,\text{rel}} = \left[\frac{\sum_{i=1}^{N_Y} (Y_i - Q_i)^2}{\sum_{i=1}^{N_Y} Q_i^2} \right]^{1/2},$$

where Q_i is the analytic solution for QoI i, and $N_Y = 20$ is the number of QoIs in this example. The results in Fig. 2.9d provide $L_{2,\text{rel}} = 0.019$, and a maximum error of $\max_{1 \le i \le N_Y} |Y_i - Q_i| = 6.76 \times 10^{-6}$.

The results for this example demonstrate the efficacy of the UQ workflow outlined in this chapter. The construction of PC surrogates provide a computationally inexpensive avenue to perform parametric inference. For this simple example, the computation time for $\mathcal{M}(\boldsymbol{X})$ was approximately 0.07 seconds. Performing parametric inference with this model—assuming a chain length of 50,000—would take approximately an hour. The use of the surrogate in inference decreased the time for inference by a factor of 15, and this efficiency only increases as $M(\boldsymbol{X})$ becomes more computationally complex.

However, the primary benefit is the outcomes of the UQ workflow, e.g., parametric inference (Fig. 2.9b), global sensitivity analysis (Fig. 2.9c), and computation of QoIs using inferred parameter values (Fig. 2.9d). Combined, these outcomes

provide explanatory value to the underlying modeled processes and the impacts of uncertainty on these processes. In Chapters 3 and 4, this UQ workflow will be applied to problems of greater scale and complexity, demonstrating its value to hydrology and ability to address basic science questions.

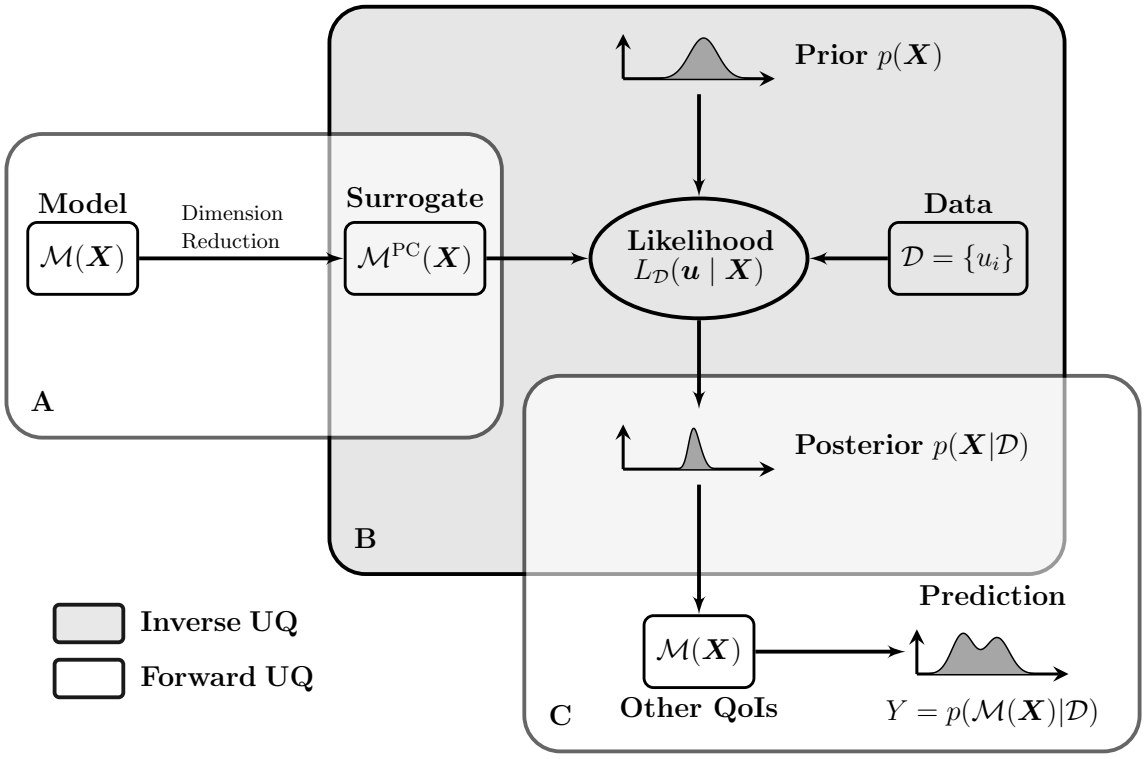

Figure 2.8: An overview of an uncertainty quantification (UQ) workflow. The set of methods in each box can be carried out on its own, or used within the general UQ implementation scope illustrated here. In the "Forward UQ" boxes, uncertain inputs ($\boldsymbol{X}$ or $p(\boldsymbol{X} \mid \mathcal{D})$) are propagated through a model. The processes "Dimension Reduction" and "Likelihood" require a modeling decision to be made on the structure of the surrogate model and likelihood function, respectively. Box (A) is the process of constructing a polynomial surrogate of the hydrologic model. Box (B) describes how hydrologic data $\mathcal{D}$ are used to perform inverse inference on a set of model parameters $\boldsymbol{X}$ to obtain parameter posteriors $p(\boldsymbol{X} \mid \mathcal{D})$. The resultant posteriors can be used in a set of procedures in box (C) that propagate uncertainty in the original model $\mathcal{M}$ (or $\mathcal{M}^{\mathrm{PC}}$) in order to get probabilistic estimates of QoIs.

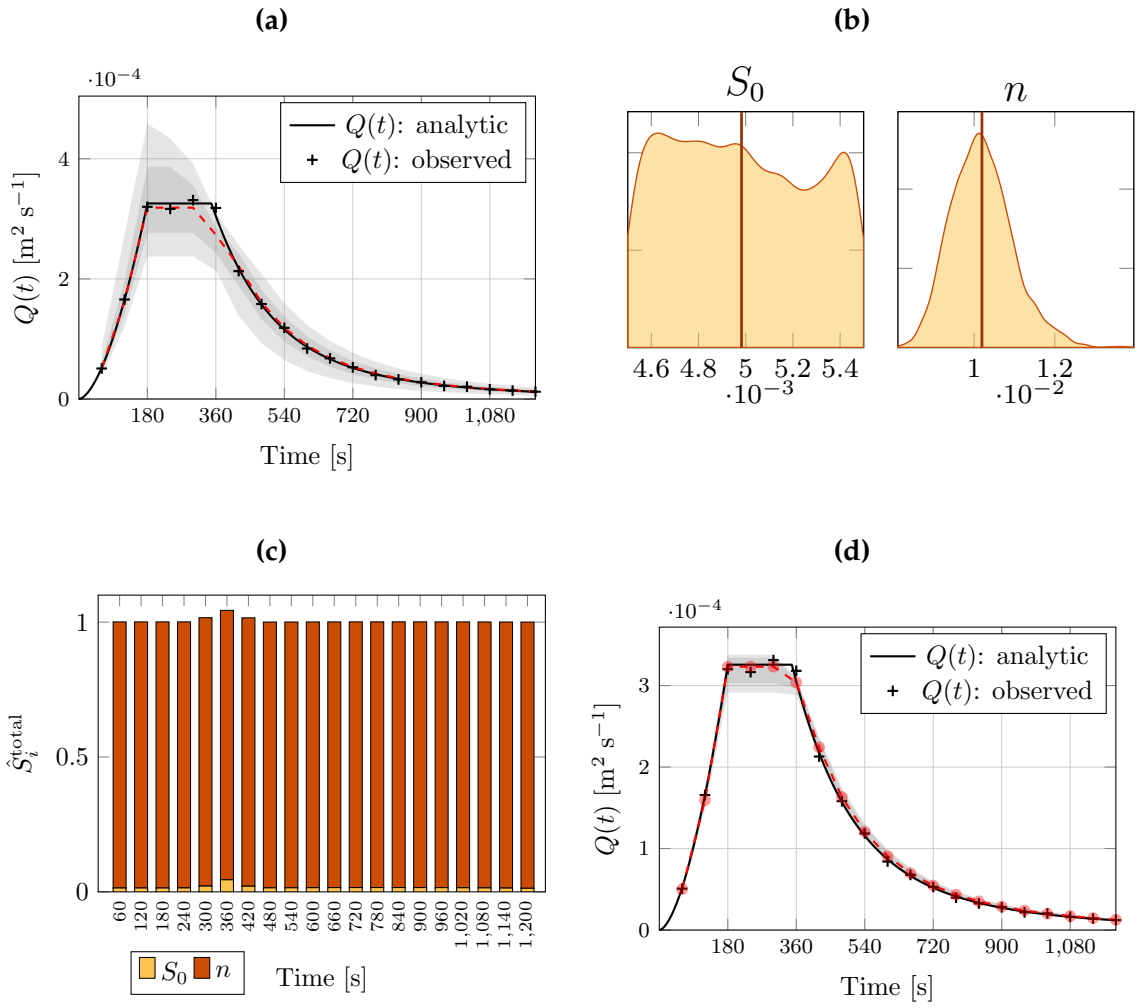

Figure 2.9: Figures for the kinematic wave example. The analytic and observed values for outflow are provided in (a), along with the uncertainty bounds of the training simulations. The gray regions represent the 25/75 quantiles and the min/max of the training simulations, and the dashed red line is the median value of the training simulations. The posterior as a result of parametric inference is provided in (b), where the vertical line in each plot represents the mean value. Total sensitivity information for each quantity of interest $Q(t)$ is given in (c), with the times specified on the x-axis. The results of the posterior propagated through $\mathcal{M}(\boldsymbol{X})$ is given in (d), with the displayed information being the same as in (a). Additionally, the median values of the QoIs are shown with the transparent red circles in order to visually compare their values with the analytic solution.

CHAPTER 3

Parameter inversion with heterogeneous data sources

3.1 Introduction

In research areas of physical hydrology and ecohydrology, computational models are used to improve the understanding and predictions of watershed and ecosystem dynamics. Recent developments towards these objectives include modeling at higher resolutions and investigating sensitivities of hydrologic response to watershed properties and climate forcings (e.g., Ringeval et al. 2014; Rudorff et al. 2014; Getirana et al. 2014; Kim and Ivanov 2015; Krakauer et al. 2014; Guan et al. 2015). Likewise, in climate assessment studies, resolving complex feedback systems requires the representation of relevant dynamics at commensurate spatial and temporal scales (Brown and Lugo 1982; Detwiler and Hall 1988; Cramer et al. 2004; Abril et al. 2014). Tackling this complexity requires models that rely on details of mechanistic interactions and therefore demand large computational resources to provide more robust assessments and predictions (Bisht et al. 2017).

Estimates from computational models are affected by a number of uncertainty sources that can be partitioned into *epistemic* and *aleatory* uncertainties (De Rocquigny 2012; Beven et al. 2016; Beven 2013). The former type refers to lack of knowledge of the mechanism being modeled due to limited data or an incomplete representation of the process; epistemic uncertainty can theoretically be reduced by gathering more data through a better representation of the modeled mechanism. The aleatory type of uncertainty (or variability) emerges from the structural randomness of the system being modeled. As an example, soil properties within a

watershed can be sources of both types of uncertainties. The soil saturated conductivity of a soil can be a source of epistemic uncertainty as soil samples can be gathered and tested in order to obtain relevant magnitudes. However, even with these measurements, there are still uncertainties about the heterogeneity and anisotropy of conductivity throughout a catchment, and this falls under aleatory uncertainty. Indeed, this can be measured, and as such one can "transform" aleatory uncertainty into epistemic uncertainty through measurement (Faber 2005), but given the size and scope of what can be measured for the purposes of watershed hydrologic analysis, accepting aleatory uncertainty is often preferred.

Furthermore, uncertainties can be amplified due to the complexity and nonlinearity of problems addressed through hydrologic simulations. In the world created through numerical experiments, the two types of uncertainties often become combined (Der Kiureghian and Ditlevsen 2009). Therefore it is prudent to apply a formal uncertainty quantification (UQ) machinery to evaluate both epistemic and aleatory sources of uncertainty. More broadly, one needs a holistic approach to UQ to seamlessly encapsulate *all* uncertainties of computer simulations within their specific contexts. This is a valuable pursuit for many computational sciences, not just hydrology, and as such, UQ has emerged in the last two decades as a very active research field, which has incorporated applied mathematics, engineering, and physical sciences (e.g., Xiu and Tartakovsky 2004; Najm 2009; Knio and Le Maître 2006; Ghanem and Doostan 2006; Sargsyan et al. 2014; Gilbert et al. 2016).

The overarching goal of UQ is accurate assessments, improvements in predictions, and understanding of key sources and magnitudes of uncertainty, which can inform decision making and control for management of natural and engineered environmental systems (da Cruz et al. 1999; Morss et al. 2005; Ascough Ii et al. 2008; García et al. 2015). The quantification of uncertainties related to a prediction of a physical system involves two associated problems: (1) the estimation of model input variables (e.g., process parameters, input forcings), addressed by comparing model simulations with available observational data or "data products" (i.e., synthesized data and model estimates), and (2) the forward propagation of uncertainty from input variables to output quantities of interest (QoIs). Both of these approaches have traditionally focused on quantifying epistemic uncertainty in hydrologic modeling, though UQ allows investigation into aleatory uncertainties as well.

Quantifying uncertainties has long been a goal in hydrologic modeling (Beven 1993; Renard et al. 2010; Beven and Westerberg 2011). The first problem of inference of input parameters is common in hydrologic modeling (e.g., Vrugt et al. 2008; Abbaspour et al. 2004; McLaughlin and Townley 1996). However, robust quantification of uncertainties for complex models remains an area of high interest (Krzysztofowicz 2001; Chen et al. 2011b; Hall et al. 2014; Beven et al. 2015). Firstly, traditional UQ methods carry computational burden that makes working with models of higher complexity difficult. Secondly, simpler, lumped models in hydrology cannot provide information on variables that originate from physically rich solutions; they therefore cannot take the full advantage of heterogeneous (in terms of space-time coverage or target variables) observational data sets that are typical of sparsely monitored watersheds. Many UQ studies have used of conceptual rainfall runoff models (e.g., Vrugt et al. 2008; Renard et al. 2010) that permit fast computation and use of variations of Markov Chain Monte Carlo (MCMC) sampling (Hastings 1970; Gilks et al. 1995) for UQ. Complex, integrated models of hydrology (e.g., Maxwell et al. 2014; Kollet et al. 2017), however, require much greater computational resources making the (tens of thousands of) simulations required via MCMC analysis computationally prohibitive. An approach to reduce this computational burden is to construct a *surrogate* or *metamodel* to approximate the behavior of the complex hydrologic model.

Recent advancements in UQ applications have examples of comprehensive, three dimensional fully integrated surface and subsurface flow models (Gilbert et al. 2016; Miller et al. 2018). The methodologies used in these studies still required hundreds of model simulations in order to accomplish rigorous uncertainty assessments. In the case where wall-clock simulation time for a larger-scale, complex watershed is considerable (e.g., days to weeks), more efficient methods are required. This study offers an approach with a UQ framework applied to a quasi-three-dimensional fully-integrated hydrologic model with surface and variably-saturated subsurface flows, as well as vegetation biophysical dynamics. This framework allows the likelihood-based estimation of input parameters to this complex model, allowing the application of a diverse set of observations for the parameter inference. To display the abilities of the framework on a complex terrain, the study domain is a small, sparsely monitored tropical catchment in the Amazon rainforest.

Due to its size and richness, the Amazon rainforest is one of the most important

biomes in the world. Large-scale studies of the Amazon have been undertaken to understand the importance of how hydrology and carbon cycles interact (e.g., Fan and Miguez-Macho 2010; Miguez-Macho and Fan 2012a; Miguez-Macho and Fan 2012b; Pokhrel et al. 2013; Lin et al. 2015). Many of these studies have simulated the entire Amazon region, many with grid cell discretization of $O(10$ [km]) (e.g., Coe et al. 2008; Beighley et al. 2009; Yamazaki et al. 2011), with others having a 2 [km] discretization (Miguez-Macho and Fan 2012a; Miguez-Macho and Fan 2012b). However, as has been previously shown (e.g., Miguez-Macho and Fan 2012b), this discretization does not resolve the basic functional hydrologic units — hillslopes and, as a result, the lateral mass fluxes from higher-elevation areas to the valleys of the drainage network. Ignoring the connection between upstream recharge areas and downstream discharge regions can have important consequences on robustness of studies that depend on understanding space-time variability of the hydrologic regime (Salvucci and Entekhabi 1995; Kim et al. 1999). Important ecohydrologic processes occur in these upstream, lower-order, headwater catchments (Richey et al. 2009; Richey et al. 2011). If one considers the upland regions to be those with a water table depth greater than 5 [m] (as in Miguez-Macho and Fan 2012b), such areas account for approximately 30 to 40% of the Amazon basin, depending on the season (Miguez-Macho and Fan 2012b), and better understanding of the hydrologic processes in these regions is scientifically justified and urgently needed.

Smaller scale studies of these upland areas have been undertaken (e.g., Fleischbein et al. 2006; Nobre et al. 2011; Cuartas et al. 2012; Fang et al. 2017), but they relied on frameworks unable to address the epistemic uncertainty associated with numerical representation of these catchments. Limited sensitivity experiments performed in (Vertessy and Elsenbeer 1999; Fang et al. 2017) were applied to these catchments, but a sufficiently general framework of uncertainty quantification and sensitivity analysis of hydrologic response of these catchments is still absent in the literature.

One source for hydrologic uncertainties in the Amazon is the presence of deep soils which give rise to fluctuating groundwater across climes and seasons (Miguez-Macho and Fan 2012b; Cuartas et al. 2012). Due to difficulties in instrumenting and measuring groundwater, data are sparse: and even experimental catchments have few wells drilled for measuring water table depth. The spatial distribution and initial states of groundwater can impact hydrologic models, such as spin up per-

formance of the model (e.g., Seck et al. 2015; Ajami et al. 2014) or providing better estimates of the impacts on below-surface processes to earth system models (Clark et al. 2015; Riley et al. 2011). Additionally, accurate representation of groundwater processes at smaller scales can provide valuable information to larger scale groundwater processes and their impacts on earth system processes (Fan 2015; Fan et al. 2013; Krakauer et al. 2014; Riley and Shen 2014).

The large soil column depths in the Amazon mean that the parameterizations of soil properties in a hydrologic model may have a large effect on the simulated groundwater and vadose zone dynamics. Some field or lab measurements may exist for the soil properties through core testing or well pump tests, but these are limited to accessible areas and may not necessarily represent bulk soil properties in the catchment (Russo and Bouton 1992; Kowalsky et al. 2004), which in turn are a primary driver of hydrologic variability within the catchment. Therefore, we assume uncertainty in the parameterization of these properties.

We introduce a way of accounting for and propagating these uncertainties by implementing what has been termed non-intrusive spectral projection (Le Maître and Knio 2010) to construct a surrogate model to emulate the behavior of a complex hydrologic model. This methodology allows us to model, at high spatial resolutions, the ecohydrologic interaction between groundwater and surface water in the Asu research catchment (e.g., Tomasella et al. 2008; Cuartas et al. 2012) in the Amazon rain forest. We account for uncertainties in model parameters, which propagate to the initial conditions of the groundwater surface. Additionally, streamflow, water table depth, soil moisture, and evapotranspiration data are available within the catchment. We assume that all the data represent hydrologic processes which are informed by soil characteristics, and therefore these data can inform the parameterization of bulk soil properties. Specifically, this work focuses on quantifying uncertainty in the soil parameterization of a small upland catchment in Amazonia, focusing on the challenge of probabilistic estimates of bulk soil properties in a sparsely-monitored catchment.

In the methodology section, we introduce (i) the mechanistic model for this study: tRIBS-VEGGIE (TIN-based Real-time Integrated Basin Simulator—Vegetation Generator for Interactive Evolution), (ii) construction of a surrogate model for tRIBS-VEGGIE through polynomial chaos expansions (PCEs), (iii) dimensionality reduction methods to more efficiently construct the PCE surrogate model, and (iv)

accelerated inference of tRIBS-VEGGIE model parameterizations using PCE surrogates. The case study of the Asu watershed demonstrates the construction of a surrogate model representation and sensitivity analysis carried out with it. This representation is then used to perform parametric inference, highlighting the flexibility of the framework to identify uncertainties and use diverse observational data for parameter estimation. The parameters obtained from inference are then used to compute hydrologic output from tRIBS-VEGGIE. The benefits and limitations of this framework are addressed in the end, with a focus on issues in hydrologic modeling that benefit from an uncertainty quantification approach.

3.2 Methods and study design

This study provides a framework to derive uncertain model parameters for a sparsely gauged catchment using a physically rich model tRIBS-VEGGIE and its simplified mathematical representation, i.e., a surrogate model. The sparse availability of groundwater, soil properties, and streamflow data in the watershed is recognized and accounted for in the design so that different data types can be used to inform the model's behavior. This section reports an approach to dealing with the lack of observational data available within the modeling and uncertainty quantification framework.

3.2.1 Hydrologic model

The representation of the hydrologic response of a tropical catchment strongly depends on reliable modeling of subsurface flows. tRIBS-VEGGIE (Ivanov et al. 2008a; Ivanov et al. 2010) emulates essential processes of water and energy dynamics over the complex topography of a river basin. Each computational element has a canopy layer that contains two "big-leaves" (sunlit and shaded) representing the canopy. Above-ground processes are coupled to a multi-layer soil model that computes soil moisture, root water uptake, and heat transport using the one-dimensional Richards equation (Hillel 1980) and the heat diffusion equation, in the direction normal to the element's surface. Gravity-driven flow for the unsaturated lateral exchange is assumed and the Dupuit-Forchheimer approximation (Bear 1979) for the saturated lateral exchange is implemented. Subsurface flows are routed using

the D-∞ flow routing algorithm (Tarboton 1997), and the flow directions change dynamically for the saturated zone, leading to spatial dynamics that reproduce the three-dimensional numerical solutions (Hopp et al. 2015). In this study, vegetation dynamics are not simulated. Only the biochemical model of photosynthesis and canopy stomatal behavior (Collatz et al. 1991; Farquhar et al. 1980; Leuning 1990; Leuning 1995) is used to simulate the response of latent heat flux to above- and below-ground conditions. The amount of leaf area as well as other structural characteristics of vegetation are imposed as pre-determined model input (see Section 3.2.3).

3.2.2 Uncertainty quantification framework

tRIBS-VEGGIE has a large set of inputs that could be treated as uncertain; these are closures for certain approximations to physical laws or parameters describing a property of a medium, for example, important for the movement of water in the domain (e.g., saturated conductivity for different soil types present in the domain). In addition to these parameters, there are input fields that describe spatial or temporal states of the simulation watershed. These parameters and fields represent the assumptions made about the makeup and behavior of the domain. They are inherently uncertain since it is not possible to have perfect information about the system being modeled, and this is especially true for natural systems. Therefore, instead of encoding assumptions into single estimates about the system in question, we embrace the uncertainty in our knowledge and attempt to quantify the consequences of this uncertainty on simulation results.

The uncertainty quantification framework from Chapter 2 will be used. Let's consider a model $\mathcal{M}$ (i.e., tRIBS-VEGGIE in this study) with inputs $\mathbf{x}$, where $\mathbf{x} = \{x_1, x_2, \ldots, x_M\}$ (e.g., $M = 7$ for this study). This model is used to predict some output quantity of interest (QoI) $\hat{y} \in \mathbb{R}$. Specifically, polynomial chaos machinery will be used to construct a surrogate model of tRIBS-VEGGIE with Bayesian compressive sensing to perform accelerated inference of input parameters to the model. Repeating from Eq. (2.13), the output QoI can be approximated through a

PCE as

$$Y = \mathcal{M}(\boldsymbol{X}) \approx \mathcal{M}^{\mathrm{PC}}(\boldsymbol{X}) = \sum_{j=0}^{P} c_j \Psi_j(\boldsymbol{X}). \tag{3.1}$$

This study uses Bayesian compressive sensing (BCS) (Section 2.4.1, Ji et al. 2008; Babacan et al. 2010; Sargsyan et al. 2014) to find a sparse set of coefficients to compute the coefficients in Eq. (3.1).

Parameter inference

Given a suite of results from a mechanistic model $\mathcal{M}$ and its constructed polynomial surrogate $\mathcal{M}^{\mathrm{PC}}$, one can infer which values of uncertain input parameters $\boldsymbol{X}$ are *most likely* to provide results that match an observed quantity. An advantage of the approach outlined here is that $\mathcal{M}^{\mathrm{PC}}$ enables very efficient inverse analysis (Marzouk and Xiu 2009). More generally, inverse problems occur when there are related observations but they are not necessarily the ultimate quantity of interest. Within hydrology, and particularly in sparsely monitored basins, there is a long history of parameter identification through some form of inversion (e.g., Neuman et al. 1980; Yeh 1986; McLaughlin and Townley 1996; Kirchner 2009b). Surrogate models with dimension reduction as outlined in Sections 2.3 and 2.4, provides a novel approach, enabling faster computation, inversion, and the ability to solve the inverse problem on a larger set of uncertain model parameters (Section 2.5.2).

Summary of UQ framework

Sections 2.3-2.5 develop a general framework of high flexibility to infer model parameters for a hydrologic model in a computationally efficient manner by using polynomial surrogates. Figure 2.8 provides a diagram outlining this framework. Generally, one uses a hydrologic model to construct a polynomial surrogate model (Box A) that allows for fast computation of output quantities of interest from the hydrologic model. In the case of many uncertain input parameters, dimensionality reduction tools such as BCS are used to alleviate the burden of multi-dimensionality of uncertain inputs for constructing the surrogate models. After a surrogate has been constructed, one can then use it for accelerated, computationally inexpensive inference of the uncertain parameters for the hydrologic model, provided that

there are available data matching a quantity of interest (Box B). Once the posterior distributions of the uncertain parameters have been calculated, they can then be used within a model $g(\boldsymbol{X})$. In theory, $g(\boldsymbol{X})$ can be any model that uses the parameters $\boldsymbol{X}$, but it is prudent to use the posterior values in the same model used for inference. Within this study, $g(\boldsymbol{X})$ is going to be tRIBS-VEGGIE ($\mathcal{M}$) in order to estimate other quantities of interest (Section 3.2.3).

A methodological step that accounts for model error is not shown in Figure 2.8 and therefore is not accounted for in this study. More specifically, the likelihood function in Box B accounts for *data error*, but does not consider the structural error of tRIBS-VEGGIE. Accounting for model error is an active, ongoing area of research within UQ (e.g., Sargsyan et al. 2018), but is beyond the scope of this study.

The strength of this approach in the context of uncertainty quantification is the relationship between QoIs—as discussed in Sec. 2.3—and observational data. The generality of coupling the PCE and inference approaches means that QoIs can be any output from the hydrologic model $\mathcal{M}$ which is then used to construct the surrogate model $\mathcal{M}^{\text{PC}}$. If one can relate observational data to the QoI being addressed through $\mathcal{M}^{\text{PC}}$, then accelerated inference for diverse outputs (e.g., hydrologic, hydraulic, ecologic, biogeochemical, etc.) is possible within complex hydrologic models.

3.2.3 Simulation setup

The watershed domain is located approximately 76 [km] northwest of Manaus, Brazil (Figure 3.1a). The watershed is part of activities carried under the Large-Scale Biosphere-Atmosphere (LBA) Experiment in Amazonia managed by the National Institute of Amazonian Research (INPA). This location was chosen due to the long record of available atmospheric forcing data from a flux tower installed in 1999 as well as the availability of relevant data from streamflow gauges, soil moisture sensors, and groundwater piezometers. This catchment is one of the most instrumented catchments in the Amazon Basin, surrounded by undisturbed rainforest. This region is characterized by a tropical monsoonal climate, with average annual rainfall of approximately 2,400 [mm], average annual temperature of 26°C, and a wet season from November–May and dry season from June–October (Nobre et al. 2011; Cuartas et al. 2012; Cuartas et al. 2007).

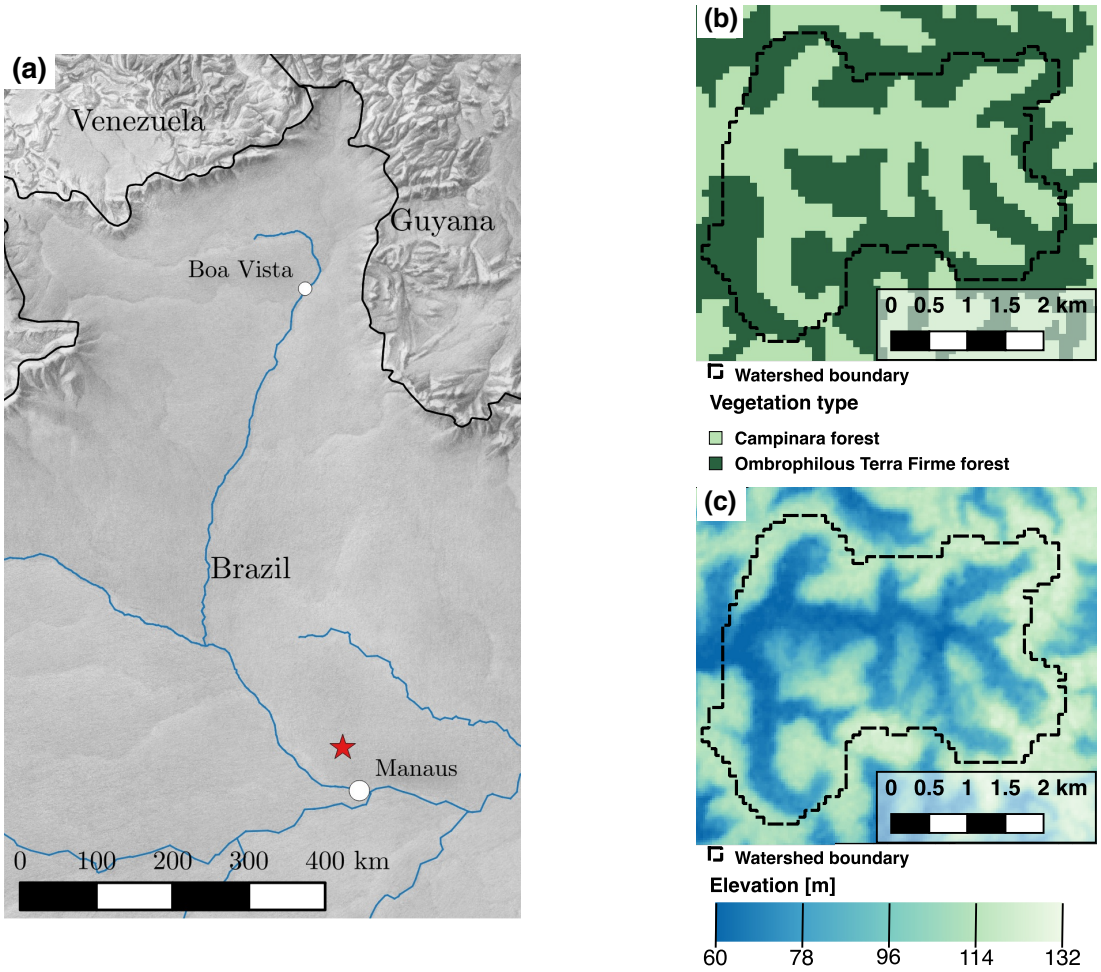

Figure 3.1: The study location, labeled by the red star in (a), is approximately 76 kilometers N-NW of Manaus, Brazil. Manaus lies at the confluence of the Negro and Solimões rivers, and at this confluence the Amazon River begins. The vegetation types for the Asu watershed are shown in (b) as derived in (Cuartas et al. 2012). The spatial distribution of elevation within the watershed is illustrated in (c).

Representation of simulation domain

The simulation domain, the Asu watershed (Figure 3.1b, c), represents a zero- to third-order basin in Amazonia, previously detailed in (Cuartas et al. 2007; Nobre et al. 2011; Cuartas et al. 2012). The total area is 12.4 [km^2], and the watershed has varying soil thickness, with a maximum between 40 and 50 meters (Cuartas et al. 2012). To represent its subsurface domain, the layer thickness is fixed at 40 meters using 35 irregularly resolved mesh layers, increasing from 0.04 [m] for the surface layer, to 2.5 [m] for the layers between 5 and 40 [m]. The thickness of the layers increase following a geometric series such that, as the soil depth increases, each layer is some fraction r thicker than the previous layer: $\Delta z_{i+1} = \Delta z_i(1+r)$, up to the depth $z_i = 5$ [m]. For this domain we chose $r = 0.296$, which allows for smaller soil layers near the surface, and larger layers towards the bottom of the soil domain. This discretization enables the capture of the dynamics of infiltration and lateral water movement in the vadose zone, while maintaining computational efficiency. In the horizontal plane, the domain is represented using 3 arc-second (90×90 [m]) spacing from the SRTM digital elevation model (Jarvis et al. 2008), resulting in 1,554 square Voronoi cells. Overall, this gives $1,554 \times 35 = 54,390$ computational nodes in the domain.

Soil type and land cover

Previous classification of soils for this site have been undertaken in (Cuartas et al. 2012; Fang et al. 2017; Tomasella et al. 2008), but focused on the near-surface soil properties at few locations that are hard to interpret in terms of their changes with depth (see also Fang et al. 2017). The detailed soil classification such as the one given in (Cuartas et al. 2012) can be useful, but understanding effective, watershed-scale properties (i.e., that represent the catchment as a whole) is frequently of more relevance, since the vast majority of basins are ungauged. Consequently, this study assumes that there is a *single* soil type in the watershed and there is a need to have its hydraulic, drainage, and retention characteristics understood.

Specifically, the uncertain parameterizations used for the retention characteristics of the soil are calculated using the pedotransfer function for Brazilian soils from (Tomasella et al. 2000). This study used multivariate linear regression relying on texture (percentages of sand, silt, and clay), organic carbon, moisture equivalent,

and bulk density to fit a second-order polynomial for the dependent variables of α, n, θ_r and θ_s of the van Genuchten soil water retention and model (Genuchten 1980). Furthermore, the study of (Broedel et al. 2017) provides data on texture and bulk density for the Asu catchment up to a depth of 14.3 [m], and the estimate for moisture equivalent is given in (Tomasella et al. 2000). Values for the saturated conductivity and the anisotropy ratio were not a part of this pedotransfer function, and reasonable value ranges were estimated using the studies of (Cuartas et al. 2012; Fang et al. 2017). The soil properties make up the uncertain model parameters $\boldsymbol{X}$, which are treated as random variables with distributions provided in Table 3.1.

The two classes of vegetation present in the catchment are (a) *terra firme* forest on the plateau and sloped areas, and (b) *Campinarana* forest (Ranzani 1980) in the valleys and ecotone areas of the watershed (Figure 3.1b). The Type (a) forest has dense evergreen tropical vegetation with heights of approximately 30 [m]. The Type (b) forest is less dense with tree heights typically around 20-25 [m] (Cuartas et al. 2012). The only difference in the parameterization between the two types is the vegetation height, as the uncertainty in vegetation were intentionally de-emphasized in order to focus on the impact of uncertainty in soil characteristics.

Table 3.1: Uncertain soil parameters $\boldsymbol{X}$ used in the workflow of Figure 2.8. The fraction of coarse and fine sand (F_{CS}, F_{FS}), required for the pedotransfer function in (Tomasella et al. 2000), is determined based on the sampled values of F_C and F_S, such that $F_{CS} = \alpha_{cs}(1 - F_C - F_S)$ and $F_{FS} = (1 - \alpha_{cs})(1 - F_C - F_S)$. $U[A, B]$ denotes the uniform distribution with support $[A, B]$.

Parameter	Description	Distribution
F_C	Fraction of clay [%]	$U[45, 65]$
F_S	Fraction of silt [%]	$U[15, 25]$
α_{cs}	Fraction of sand that is coarse [%]	$U[45, 55]$
M_e	Moisture equivalent [g g^{-1}]	$U[0.1, 0.25]$
ρ_b	Bulk density [g cm^{-1}]	$U[1.1, 1.2]$
k_s	Saturated hydraulic conductivity [mm hr^{-1}]	$U[2.0, 30]$
a_r	Horizontal:vertical anisotropy ratio [-]	$U[1, 2]$

Climate forcing

Climate forcing data are available at hourly intervals for 26,300 hours from January 1, 2003 to January 1, 2006. Aggregated time series of data used for forcing the model are shown in Figure 3.2 (Restrepo-Coupe et al. 2013).

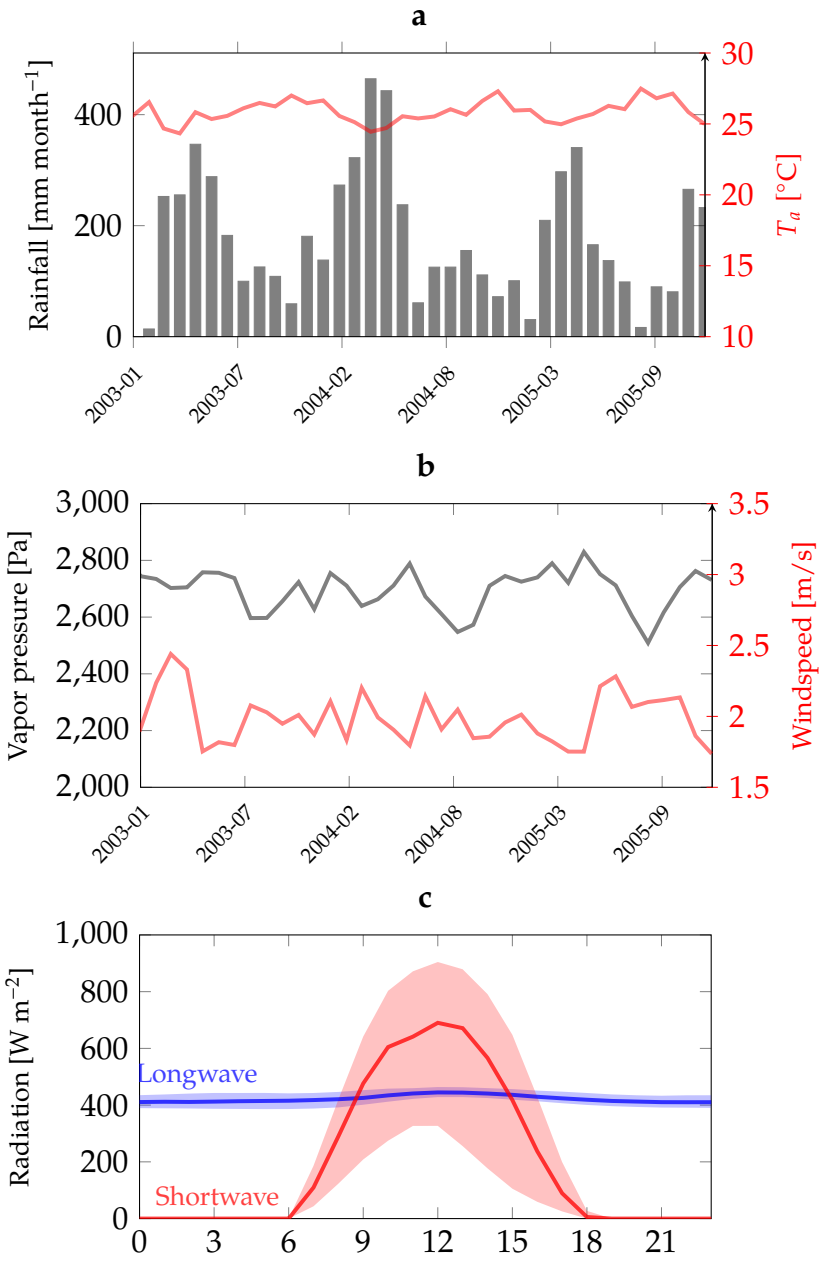

Figure 3.2: Atmospheric forcings used in simulations. (**a**): Monthly aggregated rainfall (gray bars) and air temperature (red line). (**b**) Monthly averaged vapor pressure and wind speed. (**c**): The diurnal cycle of longwave and shortwave radiation estimated for the entire simulation period. The line of each represents the median, and the shading is between the 10- and 90-percent quantiles.

Initial and boundary conditions

For solving subsurface flow dynamics, the flux (Neumann) boundary condition was specified at the surface (net rainfall) and bottom (zero flux) of the domain, allowing for infiltration, runoff, and exfiltration fluxes. For surface flow, an open boundary (in the form of free outfall) was assumed at the downstream end (Kim et al. 2016). The watershed was delineated from the downstream end, therefore for all other boundaries of the watershed, the no-flux (solid slip wall) boundary condition was specified.

Water table initializations given soil parameter uncertainty

Given that the soil parameterizations are treated as uncertain, water table initializations that adequately represent the initial state of the water table within the watershed are required. For example, a shallow water table will result under a simulation with a soil type with a low hydraulic conductivity, and a deeper water table will result with a higher hydraulic conductivity. Setting the water table to some fixed value for the entire basin, then allowing steady flow conditions to develop for a given soil type through forcing and draining requires computational resources (e.g., Seck et al. 2015), which would be a detriment to the desired efficiency from the UQ approach in this paper. Groundwater well data for the watershed (e.g., Cuartas et al. 2012; Fang et al. 2017) are available only for a few locations along a transect in the watershed. This limitation, in both spatial and topographic spaces (e.g., elevation, height above nearest drainage, slope, etc.) implies that the available data are inadequate to create meaningful realizations of initial depth to water table to be used in simulations. However, groundwater plays an important role in seasonal flooding and ET dynamics of the Amazon (Miguez-Macho and Fan 2012b; Miguez-Macho and Fan 2012a). It must be accounted for in any comprehensive hydrologic modeling and therefore some uncertainty should be associated with the water table initialization.

In this work, uncertainty in initial depth to water table is propagated through parametric uncertainty using a subset of the uncertain parameters in Table 3.1. Specifically, the initial water table depth is estimated using an adapted mapping function from (Sivapalan et al. 1987). This method supposes that there is a steady state groundwater profile throughout the basin and that the streamflow at initial-

ization time $Q(t = 0)$ is derived from a constant, spatially uniform recharge to the groundwater. Furthermore, each location in the basin has a contribution q_i defined by its position in the watershed:

$$q_i = \frac{a_c Q}{A}, \tag{3.2}$$

where q_i is the saturated lateral flow, a_c is the surface contributing area of a computational cell, Q is the semi-steady state discharge at the basin outlet, and A is the total basin area. This approach also assumes that the water table is parallel to the soil surface, so the subsurface drainage aligns with the topographical gradient. The water table depth for a single computational cell is given by:

$$N_{\text{iwt}} = -\frac{1}{f} \ln \left[\frac{f a_c Q}{A K_0 a_r W \tan S_0} \right], \tag{3.3}$$

where N_{iwt} is the initial local water table depth, K_0 is the surface saturated conductivity, f is the exponential decay parameter of the saturated hydraulic conductivity with depth (Beven 1982), a_r is soil anisotropy ratio (Ivanov et al. 2004), W is the width of the saturated flow between cells, and S_0 is the surface slope.

The decay parameter f is not used in the representation of soil hydraulic properties in this work. However, the rest of the parameters of Equation (3.3) are known and therefore a value of f can be solved for a given set of soil parameters and local watershed characteristics at location i. Specifically, the water table depths N_{iwt} are derived by taking the temporal average of water table depths for each observation location; the values of K_0 and a_r are treated as uncertain and obtained from sampling (see Table 3.1); and the rest of the variables in Equation (3.3) are derived from the basin topography. This leaves f as the only unknown of the equation and therefore, f can be derived for each soil type in the UQ framework through least squares optimization to fit the data to the model in Equation (3.3). The estimated value of f for each soil type enables the generation of an initial spatial distribution of water table, specific to the soil parameterization used in the simulation.

Selections for quantities of interest

Quantities of interest (QoIs) are outputs from the tRIBS-VEGGIE model, which are used on the left hand side of Equation (2.13) to construct the surrogate model. Each QoI needs to be a scalar, and the number of QoIs is only limited by the memory available to store data during computations outputting for QoIs, as well as the time required to construct surrogate models. In practice thousands of QoIs can be defined, such as the time series (i.e., multiple time instants), or a single QoI may be selected, if it is believed to carry a lot of information about the phenomenon being studied (e.g., the mean water table depth). For this study, a mix of targeted quantities of interest is selected. They represent both domain-aggregated quantities of interest as well as time series of specific QoIs, which are provided in Table 3.2. Those QoIs that coincide with observations are used for parametric inference. Common metrics of interest in hydrology are selected for QoIs not used in inference, highlighting the use of the UQ methodology in a partially gauged basin.

The approach taken here does not follow the common model confirmation paradigm where one separates available data for calibration and validation, typically used when there is a single variable representing the behavior of the dynamic system, such as streamflow. Rather, the methodology utilizes the total amount of data that might exist for a hydrologic system, such that if one obtains a reasonable skill of the model for the suite of available data, then the model is trained given the observations. In situations with limited data availability, preference is given to include all data in inference, as exclusion reduces the chance that the resulting posterior is accurate. In cases where included data increases uncertainty of inferred parameter values, e.g., an increase in the standard deviation in the posterior, this still provides valuable information about the representation of uncertainty within the target domain.

The number of surrogates constructed for each QoI is given in the fourth column of the Table 3.2, with a total of 499 surrogates constructed for this study. QoIs that were used in surrogates were constrained to periods where data were available, e.g., the soil moisture data in the study area only exists for January-October of 2005 (see Figure 3.3). Outlet streamflow was collected starting December 2004 and running through December 2005, however these data were only sampled once daily, often with two to four days, or sometimes periods of weeks between sampling. Due to the absence of continuous observed streamflow, monthly aggregated mean streamflow

was used as a QoI to construct the surrogate model.

Water table data at this site were sampled with temporal resolution similar to streamflow between 2012-2015, outside the time period for this study, and were sampled similarly to streamflow. While twelve well locations were sampled, not each location was sampled during both the wet and dry season, and several wells were in close proximity to each other. Due to these issues, wells that had at least ten recordings in both the wet and dry seasons were kept for analysis. After exclusion, the remaining ten wells were aggregated into groups based on their location within computational cells in tRIBS-VEGGIE, i.e., if two or more well locations were in the same computational cell, the data from these locations were combined into a group for analysis. After this aggregation, six well groups remained with water table depths between 0.5 and 5 [m] (Figure 3.3).

Table 3.2: Quantities of interest selected for this study. Those denoted $\mathcal{TS}(\cdot)$ are time series of a specific QoI. The "Inference" column denotes whether that QoI was used in inference of model parameters in Section 3.3.2, and N is the number of surrogates constructed for each QoI.

QoI	Description	Inference	N
$\mathcal{TS}(Q)$	Daily time series of streamflow [m^3 s^{-1}]		365
$\mathcal{TS}(Q_{\text{month}})$	Monthly aggregated streamflow [m^3 s^{-1}]	Y	13
$Q_{0.95}$	Daily streamflows in the 95th percentile [m^3 s^{-1}]		1
$\mathcal{TS}(\theta_{1\text{m}})$	Soil water content in top 1m	Y	10
$\mathcal{TS}(ET_{\text{dry}})$	Evapotranspiration in dry periods [mm day^{-1}]		92
$\mathcal{TS}(ET_{\text{month}})$	Mean monthly evapotranspiration [mm day^{-1}]	Y	12
WT	Depth to water table [m]	Y	6

In Table 3.2, the distinction for the "dry" period for $\mathcal{TS}(ET_{\text{dry}})$ refers to a month exhibiting a cumulative water deficit (CWD):

$$CWD_i = \begin{cases} \sum_{j=1}^{D} P_j - ET_j & \text{if } \sum_{j=1}^{D} P_j - ET_j < 0 \\ 0 & \text{if } \sum_{j=1}^{D} P_j - ET_j > 0 \end{cases}, \tag{3.4}$$

where P_j and ET_j are the daily accumulated daily precipitation and evapotranspiration, respectively, where j denotes the day in month i. The time period for all reported QoIs are the year 2005, and during this period there was a negative CWD in August, September, and October of 2005 of -49.1, -80.6, and -87.5 [mm], respectively.

To construct surrogates for each QoI, one generates a set of training and validation samples from the uncertain parameters $\boldsymbol{X}$ in Table 3.1. Each parameter X_i is scaled to a standard uniform variable for computational input, $\xi_i \in [-1, 1]$. These are then run a set of training simulations through $\mathcal{M}$ to construct the surrogate in (2.13), and the performance of the surrogate is evaluated using the set of validation simulations.

3.3 Results

This section provides an overview of the results of the construction of the surrogate model from Section 3.2.2, uses the performance of the surrogate model against observations (Table 3.2, "Inference" column) to infer parameter values, and uses the latter to investigate the response of hydrologic variables of interest within the catchment.

The data used to construct the surrogate and perform inference is summarized in Figure 3.3. One would like the training simulations to overlap the observed values used for inference, but this is not the case for all QoIs, especially those in the discharge and water table groups. However, Section 3.3.2 shows how this can be overcome using the inference techniques to better confirm tRIBS-VEGGIE with the observed data.

3.3.1 Surrogate construction

The polynomial chaos surrogates $\mathcal{M}^{\text{PC}}$ were constructed as in Equation (2.13) for the QoIs in Table 3.2. To have a well-performing surrogate, $\mathcal{M}^{\text{PC}}$ should match the simulations of the mechanistic model tRIBS-VEGGIE $\mathcal{M}$. In Figure 3.4, the absolute errors between the constructed surrogate ($\mathcal{M}^{\text{PC}}$) and simulations of tRIBS-VEGGIE ($\mathcal{M}$) are shown as illustration. Simulation results used for training purposes of constructing the surrogate are shown. Also shown are the results corresponding to validation of the surrogate, i.e., a comparison of the forward tRIBS-VEGGIE model simulations and outputs of the trained surrogate for the same QoIs.

A quantitative error measure of the surrogate accuracy is the relative L_2-norm

as a representation of error, defined as:

$$L_{2,\text{rel}} = \left[\frac{\sum_{i=1}^{N_s} \left(\mathcal{M}_i^{\text{PC}} - \mathcal{M}_i \right)^2}{\sum_{i=1}^{N_s} \mathcal{M}_i^2} \right]^{1/2}, \tag{3.5}$$

where N_s is the number of training (or validation) simulations performed. The value used for N_s depends on the computational expense of the model. A useful heuristic is to have at least 3-4 samples in each uncertain parameter dimension, which in this study would result in at least $3^7 = 2,187$ runs of tRIBS-VEGGIE. These samples are taken at random, with the heuristic providing generally good coverage of the sampling domain. To summarize surrogate performance, the QoIs are aggregated into streamflow, water table depth, evapotranspiration, and soil moisture groups. These groups are used to illustrate the relative surrogate error as a function of PC order in Figure 3.5, which shows that for training simulations, the decrease in $L_{2,\text{rel}}$ is muted after a PC order of 5 or 6.

Qualitatively, Figures 3.4 and 3.5 both show that the surrogates performs slightly better (i.e., lower values of $L_{2,\text{rel}}$ in Fig. 3.5) using training rather than validation results. This is the desired behavior as it means that the stopping criteria ϵ from the BCS method (Section 2.4, 2.4.1) are chosen correctly. In the case where reducing ϵ further improves the performance at the training set, at the expense of performance at validation set, indicates that overfitting has occurred in surrogate training. This means that the surrogate is being trained to only capture the behavior near the parameter samples at the training locations, and will not accurately capture the behavior at a significant distance away from these training samples. Evidence of overfitting would be that $L_{2,\text{rel}}$ decreases with training data but remains the same or increases in the validation data set. This is what occurs if higher-order terms in the expansion are retained for most groups of QoIs.

Thus, we have selected a surrogate order $p = 6$ for $\mathcal{M}^{\text{PC}}$ as the optimal order that leads to a sufficiently accurate surrogate without overfitting. Given the seven uncertain input parameters (Table 3.1), Equation (2.14) gives a required $P + 1 = 1,716$ basis terms in $\mathcal{M}^{\text{PC}}$. Due to the use of the BCS method outlined in Section 2.4, there were 100 simulations of tRIBS-VEGGIE used to train the surrogate. An additional 10 simulations were used to validate that the constructed surrogate accurately rep-

resented the QoIs from tRIBS-VEGGIE. In this case, the BCS methodology enabled the reduction of the number of required simulations by approximately 93% without considerable accuracy loss.

Sensitivity analysis

The goal of variance-based sensitivity analysis is to relate the proportion of variance in the model's output to the uncertain input parameters. The constructed PCE model allows for efficient computation of the main and joint effect sensitivity indices from Equations (2.18) and (2.19). One can assess at the sensitivity contributions for a single QoI, or for uncertain parameters across multiple QoIs.

For the first instance, the main and joint sensitivities of the uncertain parameters for the QoIs of water table group 5 and evapotranspiration in April 2004 are shown in Figure 3.5c and 3.5d, respectively. The main effect sensitivities — the fraction of variance described by changing each uncertain parameter in isolation, averaged over the input distribution of the other parameters — are given on the diagonal of these figures. For the water table and evapotranspiration groups, one sees that k_{sat} is the dominant contributor in the variation of the QoIs shown. This displays that the model is qualitatively consistent, as a higher value of saturated conductivity allows faster drainage from the soil and impacts water available for vegetation transpiration in the case of Figure 3.5d. The lower diagonal on Figures 3.5c, 3.5d represents the joint effect sensitivities, e.g., in Fig. 3.5c, terms in the expansion containing both k_s and a_r account for approximately 9.5% of the variability seen in the fifth water table group, while the equivalent contribution from terms containing both a_r and F_C is approximately 0.45%. Recalling from Eqs. (2.20) and (2.21), these sensitivities are computed directly from the PCE coefficients, so sparsity or very low values in the lower diagonal (e.g., Figure 3.5c) represents sparsity of the coefficients multiplying terms containing the uncertain parameters. The implication of this sparsity is that interaction between those parameters in the model have an insignificant impact on the resultant value of the QoI.

These indices are computed for each QoI, and summarized sensitivities across all QoIs are given in Figure 3.5b. Based on the contributions to the variances of the QoIs, it is clear that k_{sat} is the uncertain parameter with the largest impact on model sensitivity for the identified QoIs. This confirms intuition since values of k_{sat} control both infiltration excess runoff as well as impact of lateral flows in the

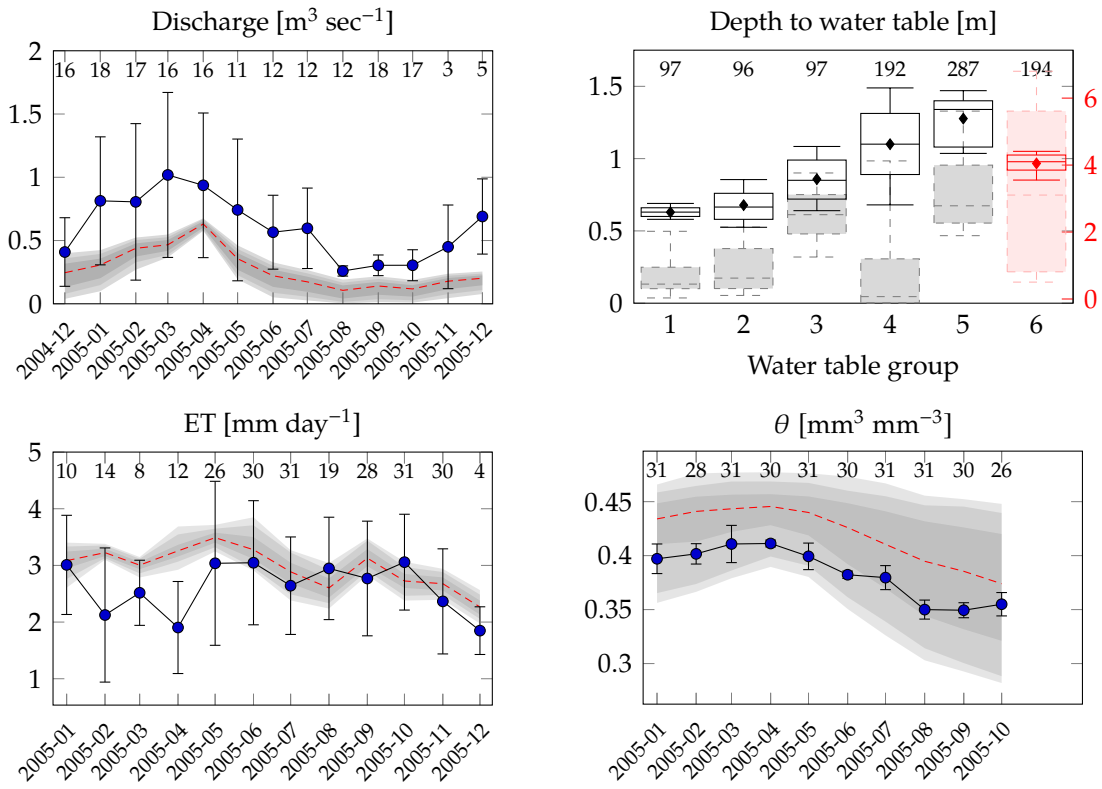

Figure 3.3: Plots for observed discharge, depth to water table, evapotranspiration (ET), and soil moisture content (θ). The discharge, ET, and θ figures plot the mean of the time series of available data with the error bars being the standard deviation of the recorded data. The discharge plot also shows the number of records aggregated into each month. Water table depth is displayed in box and whisker plots. The whiskers represent the 10- and 90-percent quantiles, and the box limits represent the 25- and 75-percent quantiles, with the median lying within the box. The diamond within each box is the mean water table value that is used for inference. The shaded regions in the plot represent the training simulations from tRIBS-VEGGIE that are used to train the surrogate model. The different shading levels represent the 25/75, 10/90, and 5/95 percentiles of the 100 training simulations. Additionally, the red line used in the time series plots represents the median of the training simulations. The colors within the water table plot are divided into two y-axes to provide better readability and clarity values in groups (1-5) and 6.

hydrologic system, which is not included in the pedotransfer function of (Tomasella et al. 2000).

3.3.2 Inference of soil properties

After the surrogate model is obtained, it is possible to calculate the posterior distribution of model parameters through MCMC sampling,as described in Section 3.2.2 above, by replacing $\mathcal{M}$ with $\mathcal{M}^{\mathrm{PC}}$. This enables faster computation and benefits hydrologic models that take more than a few minutes to perform a single simulation. The marginal and pairwise joint marginal posteriors of the hydrologic parameters in Table 3.1 are shown in Figure 3.6. These posteriors are summarized in Table 3.3 by their moments, *maximum a posteriori* (MAP) estimate, i.e., the mode of the posterior distribution, and the coefficient of variation (ratio of standard deviation to the mean value).

Table 3.3: First two moments and MAP estimate of parameter posteriors.

X_i	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mathrm{MAP}}$	$\hat{\sigma}/\hat{\mu}$ [%]
F_c	56.0	1.871	56.3	3.34
F_S	24.2	0.677	24.5	2.79
α_{CS}	0.525	1.64E-3	0.525	0.31
M_e	0.184	5.75E-3	0.184	3.13
ρ_b	1.2	1.19E-3	1.2	9.9E-2
k_s	15.8	0.645	15.9	4.09
a_r	1.45	6.11E-2	1.44	4.22

These posterior distributions provide information about the bulk soil properties of the watershed, e.g., that the soil is mostly clay and silt, with a saturated conductivity around 20 [mm/hr], and an accompanying anisotropy ratio of approximately 1.25. These values are in agreement with previous studies of the catchment (e.g., Tomasella et al. 2008; Cuartas et al. 2012), which found four different soil types within the watershed. The clay contents for these four soil types (below 1 [m]) ranged between 5–90% clay, where the soil type with 80–90% clay accounting for 45% of the catchment, and the soil type with 5% clay accounting for 30% of the catchment. Using a single soil type in the catchment represents an aggregation of the physical properties, whereby the posterior mean for clay content of 56% is reasonable based on the observed properties. The other parameters of the pedo-

transfer function fall within ranges given in (Tomasella et al. 2000), with additional qualitative agreement between the textural classification shown for Manaus. The hydraulic properties of k_s and a_r in (Cuartas et al. 2012) have sixteen classifications based on landscape classes and depth, with the inferred values in this study indicating a soil type similar to the "plateau" or "slope" landscape which are at least 15 [m] above the nearest stream node, accounting for approximately 45% of the catchment area. Similarly, (Fang et al. 2017) contains hydraulic properties with a single landscape classification at four depths, where k_s is approximately between 9 and 15 [mm/hr] with $a_r = 1$, which qualitatively agrees to the inferred value of 15.8 [mm/hr]. Additionally, one can assess the pairwise correlations between parameters using Figure 3.6, e.g., that the anisotropy ratio (a_r) is negatively correlated with saturated conductivity (k_s).

Posterior predictions of the QoIs are demonstrated in Figure 3.7. Here, the joint posterior from Figure 3.6 is sampled in order to calculate the QoI values using the surrogates constructed in Section 3.3.1. A comparison can then be made between the posterior QoI values and the observed data values. One sees that the posterior QoIs for ET and soil moisture match the data better than for discharge and depth to water table. There are two main factors contributing to this: (1) the simulated values (using $\mathcal{M}$) for ET and soil moisture match the data better than the simulated discharges and water table depths (see Figure 3.3), and (2) data noise for these QoIs are much lower than those for discharge. If one were interested in getting a better fit to the data for a set of QoIs, one could: (i) select a larger training set to have more chances for the simulation results of $\mathcal{M}$ to match observations; (ii) perform inference using a subset of the observed data and QoIs in an attempt exclusively fit for that subset of data; or (iii) attempt to collect more data to constrain the data noise. In the cases of options (ii) and (iii), the PCE framework provides a benefit of not needing to rerun simulations of the computationally-expensive $\mathcal{M}$, but also allows fine-tuning of model parameters for investigations into specific QoIs.

The posterior QoIs as a result of performing inference separately on each group of QoIs (i.e., only data of a given group are used to infer the entire set of parameters in Table 3.1) is shown in Figure 3.8. In this figure, the data noise—σ^2 in Equation (2.50)—was set as a hyper-parameter and was also inferred because it provides a better fit of the mean prediction of $\mathcal{M}^{\text{PC}}$ to the data. This was done to illustrate an approach that can be undertaken when data noise is large to the point of it being

uninformative, e.g., in the case of discharge and ET data. In this case a Jefferys prior (Eq. 2.51) is used for σ^2. Performing inference on each group of QoIs is done at the expense that the parameter posteriors are constrained for use on each separate group of QoIs. The posteriors in Figure 3.8 for the variables X_i contain other important details about the information that can be gained with inference. For example, the posteriors for the bulk density (ρ_b) are very close to the uniform prior for most groups of QoIs, since this parameter is not informed by the available data (except for soil moisture).

Whether one takes the approach shown in Figure 3.7 or 3.8 will depend on the questions being investigated. Generally, the Bayesian framework prefers including all data to perform inference, as long as the data are informative. In the case where data come from disparate sources, e.g., sporadic sampling of discharge vs. hourly evapotranspiration measurements, one may want to separate the inference using these data, or incorporate more detailed structure in the data noise representation among groups of QoIs to better inform their impact on inference.

3.3.3 Computing QoIs from posteriors

Frequently, in hydrologic modeling there are no direct observations of the studied quantity of interest. With the framework presented in this study, one can construct a surrogate model for quantities that have been observed (Section 3.3.1) and use this surrogate to confirm the model (Section 3.3.2). This confirmed model can then be used to assess the performance to unobserved QoIs or the same observed variables using the parameter posterior distributions. This allows one investigate a wide variety of model behavior, such as higher-frequency or aggregated quantities at coarser or finer temporal/spatial intervals. In this work, the parameter posteriors from Figure 3.8 resulted in a chain of 19,000 posterior values for each parameter in Table 3.1, which can be directly sampled and used as input to a constructed surrogate (box (C) in Figure 2.8).

From Table 3.2, the QoIs held out from inference were: Q_{95} — the 95th percentile of *daily-averaged* discharge during 2005 (representing the probability that 18 days during the year have a mean daily discharge larger than Q_{95}), $\mathcal{TS}(Q)$ — the daily time series of streamflow in 2005, and $\mathcal{TS}(ET_{\text{dry}})$ — daily evapotranspiration in months with cumulative water deficit as defined in Equation (3.4). Illustration Plot

and sensitivity information for these QoIs are given in Figure 3.9. Note that the surrogates for these higher-frequency and aggregated QoIs were not constructed in Section 3.3.1, but doing so is straightforward and follows the methodology of Section 2.3.1.

These results are provided to illustrate the flexibility of the UQ framework. The benefits of the approach include that one can investigate hydrologic response at higher temporal and spatial resolutions. For example, one can see the pattern of discharge in Figure 3.9 follows that of Figure 3.8, but unlike the original data sourced in Figure 3.3, this provides daily estimates of discharge.

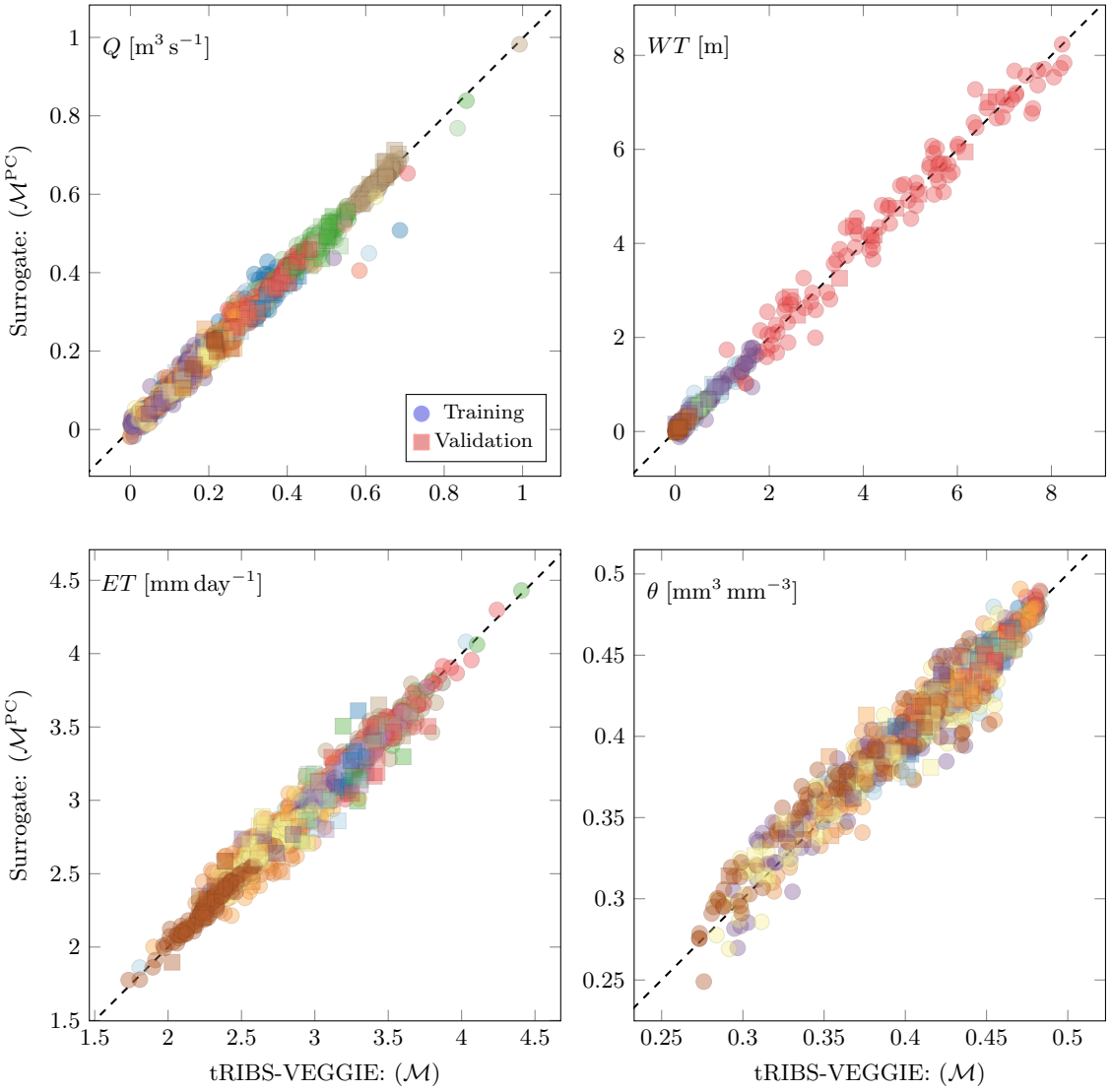

Figure 3.4: Comparison of tRIBS-VEGGIE and constructed surrogate for PC order $p = 6$. The surrogates are separated into four groups depending on the QoI. Colors within groups signify a different QoI, e.g., there are 12 colors in the bottom-left plot for ET since there is a QoI for ET in each month of 2005. Circular and square marks represent the 100 training and 10 validation simulations, respectively. A $y = x$ line is added to each plot, such that points on the black line represent agreement between tRIBS-VEGGIE and the constructed surrogate. The simulations are colored by the QoI, e.g., the different colors in the plots for Q, ET, and SM represent a classifier to distinguish different months for the QoI. The relative errors of the surrogates for each group are: $L_{2,\text{rel}}(Q) = 0.081$, $L_{2,\text{rel}}(WT) = 0.10$, $L_{2,\text{rel}}(ET) = 0.025$, and $L_{2,\text{rel}}(\theta) = 0.029$.

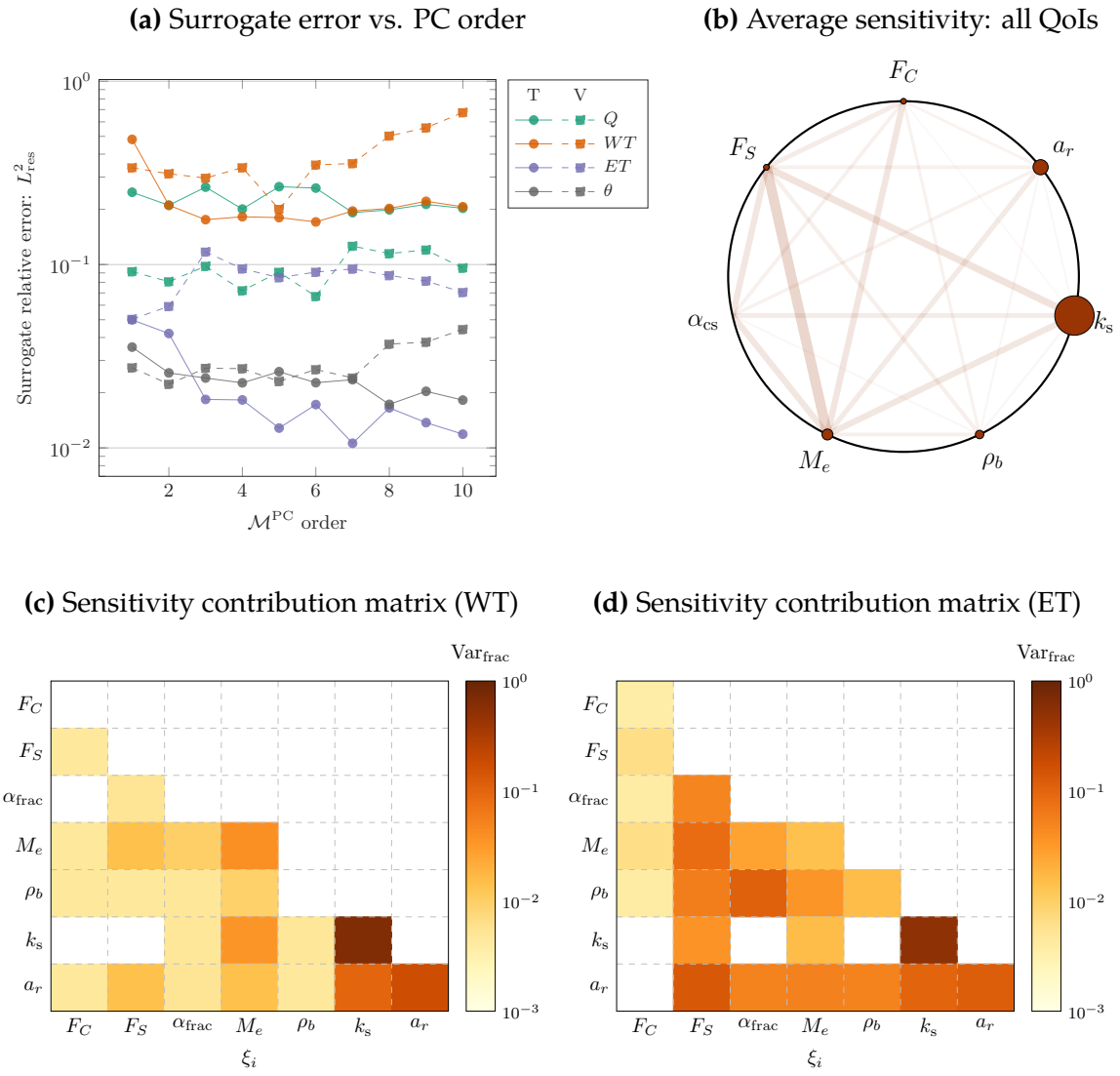

Figure 3.5: Plots of surrogate errors and sensitivities. The training and validation surrogate error is given in (a). Each color represents a different group of QoIs, and the marker shape differentiates between training and validation samples. A qualitative representation of sensitivities is given in (b). Here, the diameter of the nodes around the circle are proportional to the main effect sensitivities, and the width and opacity of the lines connecting the nodes around the circle are proportional to the joint sensitivities, where the main and joint sensitivities are calculated for each QoI, and the plot shows the average over all QoIs. The lower triangular matrices in (c) and (d) show the main and joint sensitivities for water table group 5 (c) evapotranspiration in April, 2005 (d). The main effect sensitivities are on the diagonal, and joint sensitivities between parameters are lower triangular, with a minimum sensitivity threshold of 10^{-5}.

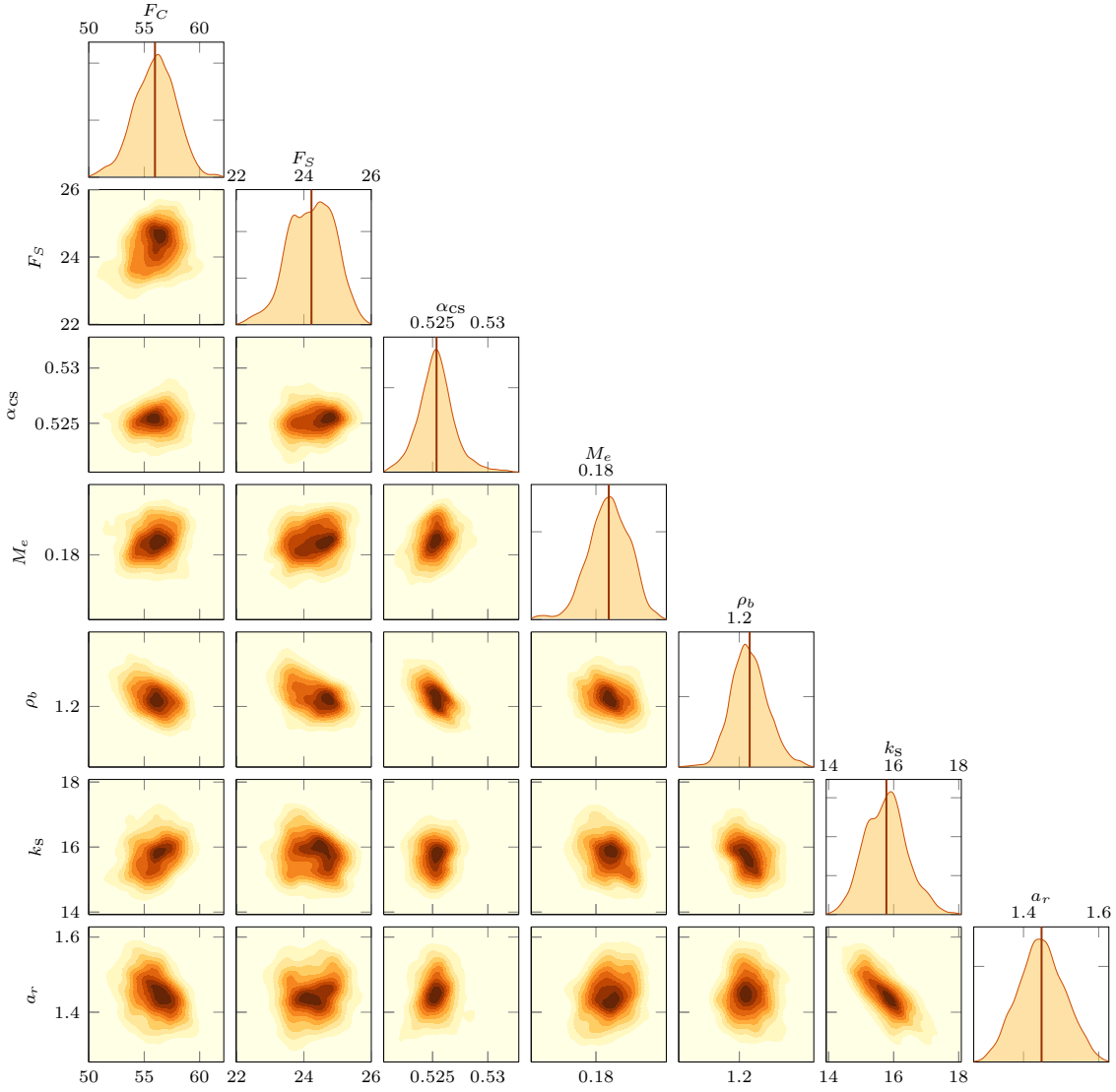

Figure 3.6: Posterior distributions of parameters used in inference. The vertical line in each of the diagonal is the estimated mean of the posterior. Summary statistics for the marginal distributions are given in Table 3.3.

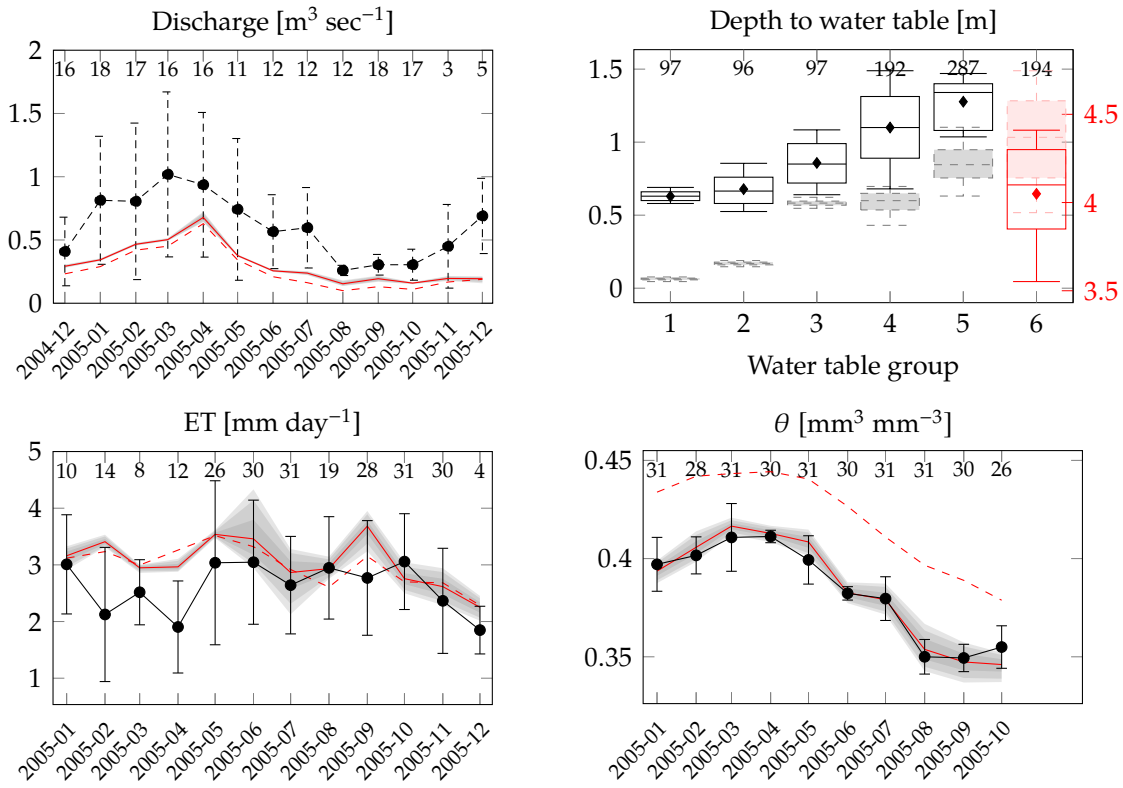

Figure 3.7: Plots for observed discharge, depth to water table, evapotranspiration (ET), and soil moisture content (θ) compared to surrogate calculations using the samples of posteriors from Figure 3.6, which are summarized in Table 3.3. The dashed red line corresponds to the median training simulation values as in Figure 3.3, while the solid red line represents the median values of the surrogate simulations with the posterior parameter samples. Please refer to Figure 3.3 for a description of other presentation details.

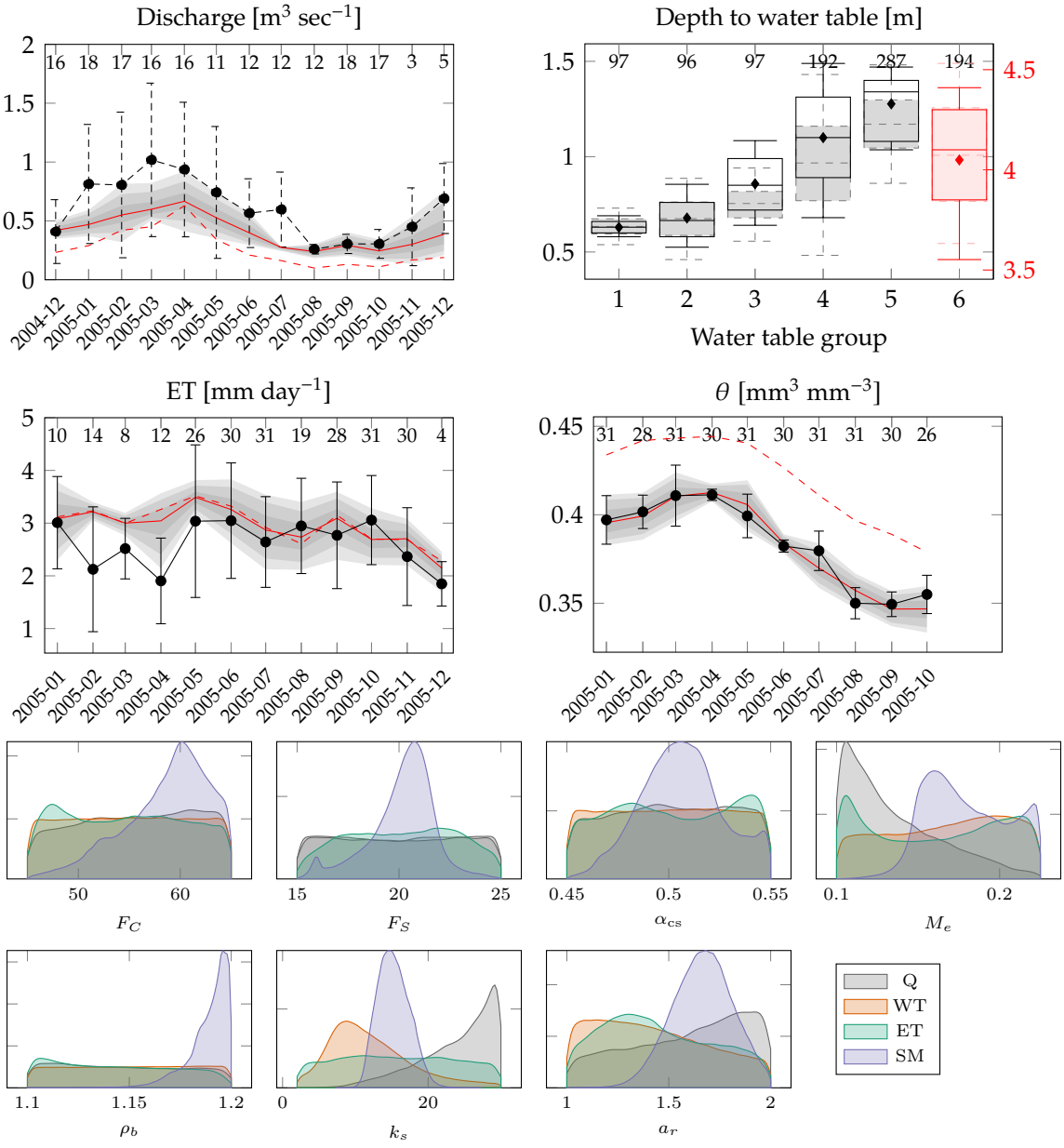

Figure 3.8: Plots for the four output groups of observed discharge, depth to water table, evapotranspiration (ET), and soil moisture content (θ) compared to surrogate calculations using posterior parameter values that were inferred for each individual output group. This path of inference generates four separate posterior PDFs for $\boldsymbol{X}$, which are shown at the bottom of the figure. The dashed red line corresponds to the median training simulation values as in Figure 3.3, while the solid red line is the median value of the surrogate simulations with the posterior parameter samples. For each output group, the gray regions represent the 95% uncertainty bound from propagating the parameter posteriors for that output group (Q, WT, ET, SM) through the respective surrogates.

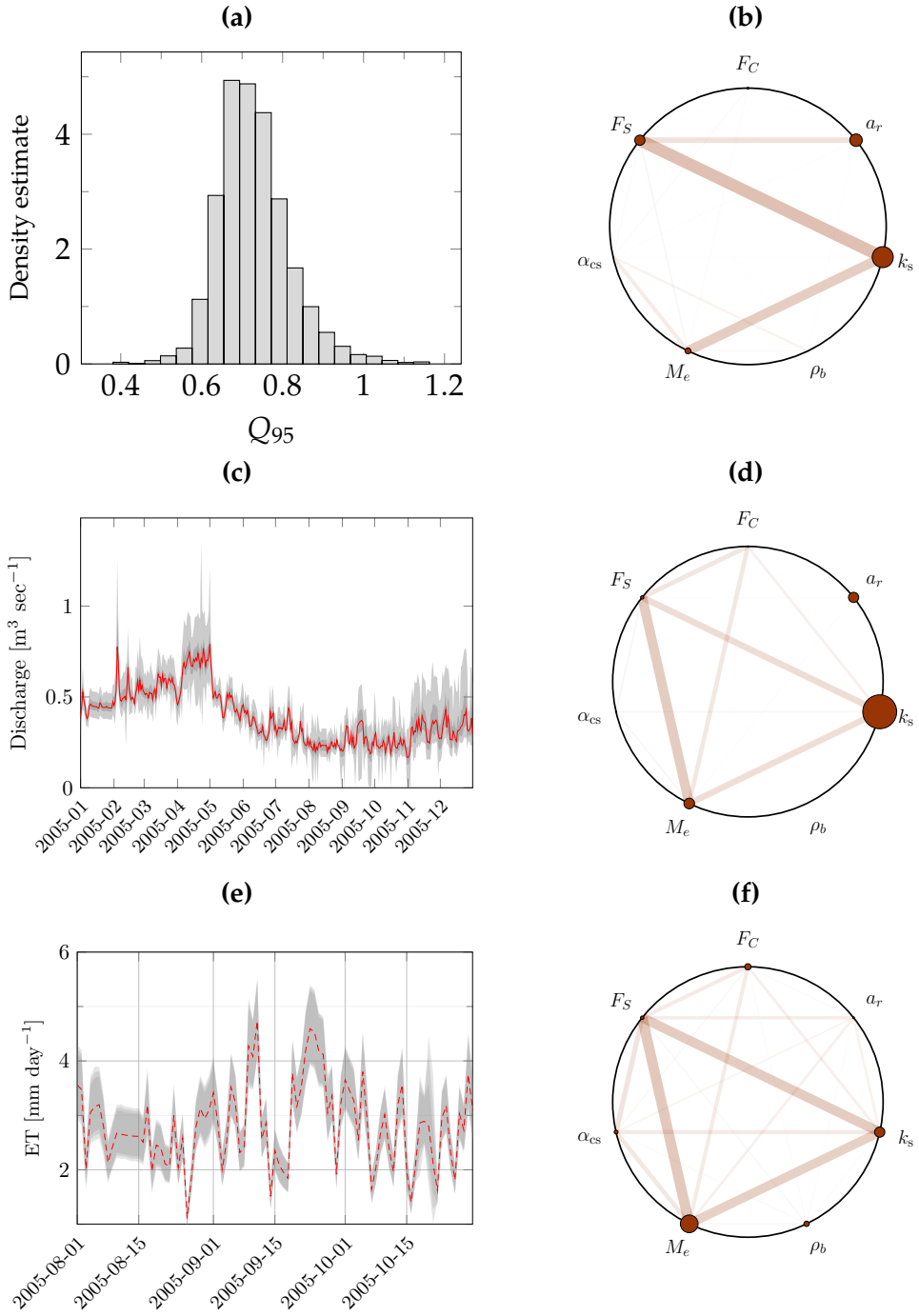

Figure 3.9: Posterior QoIs (left) and their sensitivities (right) corresponding the the QoIs not included in inference in Table 3.2, due to insufficiency/lack of observational data. The surrogates for these QoIs were constructed using the same 100 training and 10 validation simulations as in Figure 3.3. In plots (c) and (e), the grayed region represents the 5% and 95% quantiles of surrogate simulations evaluated using the posterior parameter distributions in Figure 3.8. See Figure 3.5b for details on the representation of sensitivities in (b), (d), and (f).

3.4 Discussion

Despite recent advances in computational power, simulation times for complex hydrologic models, even for low-order catchments can be significant. When a large number of uncertain parameters or, generally, uncertain model inputs are used, thousands of model simulations may be required to perform robust uncertainty, parameter inference, and sensitivity assessment. Even reduced-order simulation approaches recently introduced to environment and earth-system modeling would be prone to computational issues. This study addresses this challenge and adopts an efficient methodology to enable uncertainty quantification and stochastic simulation with deterministic, process-based hydrologic models of higher complexity. This work applies recent developments that combine reduced-order modeling based on polynomial chaos expansions with Bayesian compressive sensing to construct computationally inexpensive surrogate formulations of the complex hydrologic model, while reducing the number of required simulations. The surrogate representation enables Bayesian inversion and calibration of uncertain model variables for any model output that can be compared to observable data, even when these are sporadic and have time-varying accuracy. Furthermore, surrogate formulation can be used to propagate uncertainty through a hydrologic model for any of the model's outputs, enabling one to estimate uncertainties of QoIs that are difficult or too costly to measure.

This approach is flexible, but relies on *a prior* identification of input variables of high impact before modeling. This means that although one can change the QoIs being investigated, one cannot change the parameters that are treated as uncertain without having to rerun training and validation simulations. Given the reduction in computational due to the use of the Bayesian compressive sensing methodology, it is possible to treat dozens of model parameters as uncertain (Sargsyan et al. 2014; Ricciuto et al. 2018), essentially assuming no *a priori* knowledge of impactful variables. However, in most cases in hydrology, it is beneficial to constrain the number of uncertain parameters using expert knowledge about the governing processes controlling QoIs and the respective variables impacting these processes. Alternatively, one can do an initial screening of model parameters with Bayesian compressive sensing. This will not produce an accurate surrogate, but will yield accurate sensitivities, so the parameter space can be reduced and the procedure

repeated to gain a more accurate surrogate.

Given the scale and complexity of environmental systems, one must include uncertainty if one hopes to capture or discover complex hydrologic behavior. Uncertainties are abound within the field of hydrology, e.g., a measurement of a basin-aggregated metric such as streamflow (Figure 3.3) is subject to aleatory uncertainty within the catchment. Or there may be interactions with model input variables, meaning that to best capture epistemic uncertainty about variables, description through probability densities (Figure 3.6) is preferred over scalar, deterministic values. In either case, the presented approach offers a novel, sufficiently general way forward to address uncertainties in hydrology, reducing or defining the uncertainties of model inputs.

Generally, the framework outlined in this study provides flexibility to address computationally expensive problems in hydrology (e.g. high-resolution modeling of soil moisture (Krasnosel'skii and Pokrovskii 1989; Ivanov et al. 2010), representation of macroporosity (Beven and Germann 1982, e.g.,), etc.) by enabling high-fidelity simulations under uncertainty. In the cases where these simulations are highly sensitive to input parameters *and* forcings, the presented approach can be expanded to any process that can be represented through a set of parameters $\boldsymbol{X}$. For example, spatial model input can be viewed as a stochastic field that can be represented with a finite number of parameters using a Karhunen-Loéve expansion (Karhunen 1946; Zheng and Dai 2017). Following this philosophy, uncertain soil moisture, precipitation, or water table fields could be propagated through a hydrologic model.

Relevant software tools have been developed in recent years, so that the hydrologist need not also be an expert in uncertainty quantification to apply these methodological approaches. Those that are freely available for research use and actively developed include the Uncertainty Quantification Toolkit (UQTk) (Debusschere et al. 2017, version 3.0.4 used in this study), UQLab (Marelli and Sudret 2014), and the MIT Uncertainty Quantification Library (Parno et al. 2014).

CHAPTER 4

Inference of landscape controls on distribution of vegetation traits

4.1 Introduction

The Amazon rainforest exhibits large effects on global water, carbon and energy cycles, as shown in numerous previous studies (e.g., Salati and Vose 1984; Pan et al. 2011; Davidson et al. 2012). In turn, the production and function of global forests are related to the spatial and temporal variability in rainfall (Fatichi et al. 2012), and approximately half of the Amazon basin experiences a drought season (Marengo and Espinoza 2016). This inter-annual variability is also associated with prolonged droughts, often coinciding with the El Niño Southern Oscillation events and warm tropical Atlantic surface temperatures (Trenberth 2011). Furthermore, the frequency and severity of drought may not be stationary, with some studies showing reduced precipitation in the basin in the 21st century (Marengo and Espinoza 2016; Phillips et al. 2009).

These periods of limited water availability also coincide with increased water and carbon fluxes, and an increase in photosynthetic activity (Saleska et al. 2003; Huete et al. 2006; Santos et al. 2018). Multiple possible explanations for this behavior exist, one being that in periods of high rainfall the forest is light-limited from cloud cover, and therefore the reduction of cover during dry periods allow for an increase in transpiration flux (Hutyra et al. 2007). However, previous modeling efforts did not discover this pattern, with decreased transpiration fluxes during the dry season. Other hypotheses to explain drought tolerance include: (i) water availability from high water tables (Miguez-Macho and Fan 2012a), (ii) hydraulic redistribution within the root zone, where the root system of the tree transports

water vertically in the root zone to respond to water potential gradients (Baker et al. 2008), and (iii) very deep root systems capable of root water uptake during dry seasons (Baker et al. 2008).

There is partial and often inconclusive support for each of these hypothesis. In the first instance, some 60-70% of the Amazon basin may have a water table depth of less than 5 [m] (Miguez-Macho and Fan 2012a), providing water at shallow rooting depths. However, this still does not account for the remainder of the upland areas of the Amazon basin, which have much deeper water table depths. Secondly, there is mixed support for the hypothesis of hydraulic redistribution. Though evidence exists for hydraulic distribution in three tree species in the Amazon rainforest (Tapajós National Forest in Brazil, Oliveira et al. 2005), this study also showed a reversal of sap flow, which appears orthogonal to the evolutionary strategy to increase photosynthetic function (Eagleson 2005). Other studies have postulated that hydraulic redistribution is not significant enough to adequately explain soil water dynamics (Romero-Saltos et al. 2005; Markewitz et al. 2010). Lastly, there is clear evidence of "deep" roots, where deep is defined as a rooting depth of greater than three (3) [m]. However, the functionality of these roots remain unclear, as there are questions around how the hydraulic functionality of the root and xylem[1] system is able to transport water from deep roots to the top of the canopy. Further studies have estimated that a small amount of the rooting mass lay at depths greater than 4 [m], but the ability to infer global characteristics of rooting behavior from these studies is highly uncertain (Fang et al. 2017; Fan et al. 2017; Brum et al. 2018; Ivanov et al. 2012).

Related to rooting depth, (Ivanov et al. 2012; Brum et al. 2018) investigated the role of possible rooting depth niches between canopy layers in explaining the photosynthetic behavior during dry periods and found that such a separation across rooting depth may provide an adaptive strategy for changing water demands in dry seasons. Further evidence leading to the plausibility of varying root depths come from the rainfall exclusion experiments in (Nepstad et al. 2002), which showed asymmetric response from overstory and understory trees. The overstory trees were found to be the most vulnerable, potentially due to their higher exposure to radiation and reliance on deeper soil water reservoir that was depleted during the

[1]Xylem is the conductive tissue in plants that is the pathway for water from the root to the leaf of the plant.

exclusion experiment. In addition to rooting depth, there has been linkages to the role of other plant traits to drought resilience in Amazonia, including root hydraulic redistribution (Oliveira et al. 2005), xylem embolism resistance (Anderegg et al. 2016; Meinzer et al. 2009), and community symbiosis (Silvertown et al. 2015). Any, or likely some mixture of these traits may have explanatory value on the resilience of Amazonian forests to drought.

Examining the combination of these traits in the field is a costly endeavor in both time and money. To further experimental investigation, numerical modeling has emerged as a way to investigate these traits in concert with field experiments (e.g., Ivanov et al. 2012; Fan et al. 2017; Fang et al. 2017). The simulation-based studies to investigate plant behavior rely on mechanistic representation of hydrologic, coupled water/energy, and vegetation processes. The computational expense of these complex, coupled hydrology-vegetation models limits the amount of uncertainty one can evaluate related to the multi-dimensional space of the model's parameters and inputs.

This study makes a further effort to investigate the ecohydrology of the Amazon rainforest, expanding on the work of root niche separation in (Ivanov et al. 2012) and emerging evidence of vegetation trait adjustment in relation to a position within a canopy Brum et al. 2018. The study further hypothesizes that the frequency of water-logging conditions should also play a role in landscape distribution of traits. By addressing landscape effects on a broad set of plant traits defining above-ground carbon uptake and below-ground acquisition of soil water, this research explores further how the hydrologic regime shapes ecosystem structure and function.

By employing data inference and uncertainty quantification (UQ), the performance of vegetation is investigated. Specifically, we hypothesize that plants in these environments evolved to *maximize reproductive performance at the least possible hydraulic effort*. The reasoning behind this is that plants evolved targeting to invest the least amount of resources possible towards constructing mechanical and biochemical strategies to cope with periodic droughts and water-logging conditions. Given constraints on carbon uptake when the root zone is in anoxic (i.e., saturated) and drought conditions, this research seeks to identify traits and performance metrics that can demonstrate whether vegetation performs well in both dry and wet conditions. To do this, we need to investigate the below-ground hydrologic controls on root water uptake, and assess what types of constraints apply to either

vegetation traits or soil hydrologic behavior.

The process to perform this investigation is as follows: (1) construct surrogates for the computationally expensive tRIBS-VEGGIE for observed quantities of interest to infer parameter values that accurately represent hydrologic and vegetative function in the case study domain; (2) Use these inferred parameter values to propagate uncertainty through both QoIs used for inference as well as unobservable QoIs. (3) Construct a performance metric that balances the maximization of vegetation productivity and the cost of maintaining its performance in drought conditions. (4) Investigate plant functional characteristics using sensitivity analysis *from the construction of surrogates* and relationships with other QoIs across both time and transect location.

In the following sections we introduce the study domain and the framework for uncertainty quantification, providing reasoning behind the parameters treated as uncertain in the study. A performance metric for vegetative function is introduced in Section 4.2.6. Construction of surrogate models and the results of parametric inference are reported, and the performance metric is used to interpret the role of plant functional traits at differing locations in the study domain.

4.2 Methods and study design

4.2.1 Site description

The data are from the Asu experimental watershed near Manaus, Brazil in the Cuieiras Biological Reserve, home to the ZF2 K34 flux tower (-2.61, -60.21) (Araújo et al. 2002), which was installed in 1999. This location will be referred to as "ZF2" in this study. The site is part of the Large-Scale Biosphere-Atmosphere (LBA) Experiment in Amazonia, managed by the National Institute of Amazonian Research (INPA), and is also one of the sites chosen for the Department of Energy's Next-Generation Ecosystem Experiments—Tropics (NGEE-Tropics) program and has been used in a number of previous studies; across these two projects the site has been featured in a number of previous studies (e.g., Cuartas et al. 2007; Tomasella et al. 2008; Cuartas et al. 2012; Broedel et al. 2017; Fang et al. 2017; Dwelle et al. 2018). The site has an average annual temperature and rainfall of 26°C and approximately 2,400 [mm], respectively, with forcing for the study period given in

Figure 4.1. The site exhibits a wet season from November–May and dry season from June–October, (Cuartas et al. 2007; Cuartas et al. 2012; Nobre et al. 2011). Some months in the simulation period contain a cumulative water deficit (CWD), defined as a month with cumulative evapotranspiration greater than the accumulated daily precipitation (Eq. (3.4)). These months are highlighted in Figure 4.1.

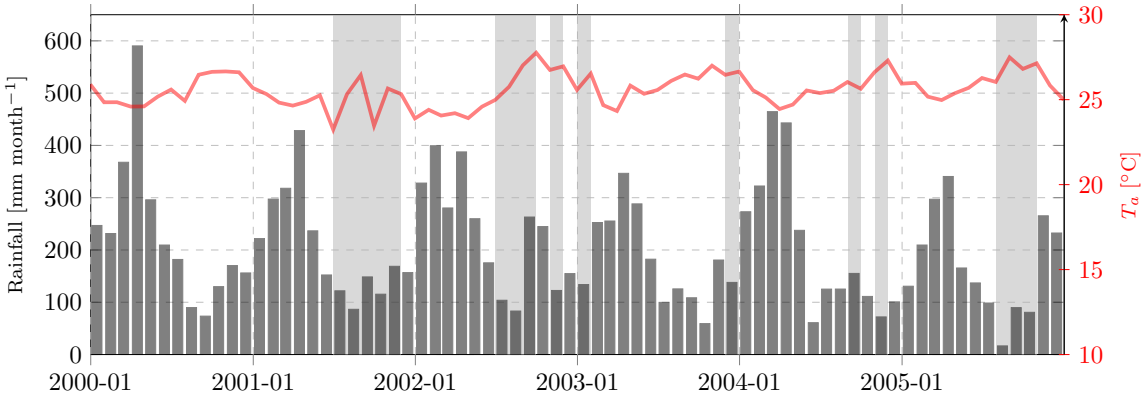

Figure 4.1: Plot of forcing climate during simulation. Accumulated monthly precipitation and average monthly temperature. The shaded months indicate months with a cumulative water deficit.

Selection of computational transect

The domain is represented using the 3 arc-second spacing from the SRTM digital elevation model (Jarvis et al. 2008). To study the effect of rooting and vegetation function to the spatial and temporal availability of water in the watershed, a transect was selected that connects the upper and lower sections of the watershed. The rationale for such a selection of the study domain is that one expects similar hydrologic behavior for many areas of the watershed. Choosing a transect allows us to avoid the unnecessary computational burden for areas that have similar hydrologic dynamics. This transect was chosen to start at the ZF2 flux tower, and follow a flow path using the D8 flow routing algorithm (O'Callaghan and Mark 1984).

The resultant transect terminates at the main stream channel in the watershed and contains three different soil types, with a total of ten square Voronoi cells in the horizontal dimension, each measuring 8,100 [m^2]. The elevations, soil types, and initial water table position are provided in Figure 4.2. The subsurface domain of each of these elements is represented with a 40-meter soil layer with 201 regularly-

resolved mesh layers, providing 200 subsurface elements of depth 200 [mm]. The horizontal and vertical discretization provides a total of 2,000 computational nodes in the domain, resulting in two-dimensional flow in the depth dimension and along the length of the transect.

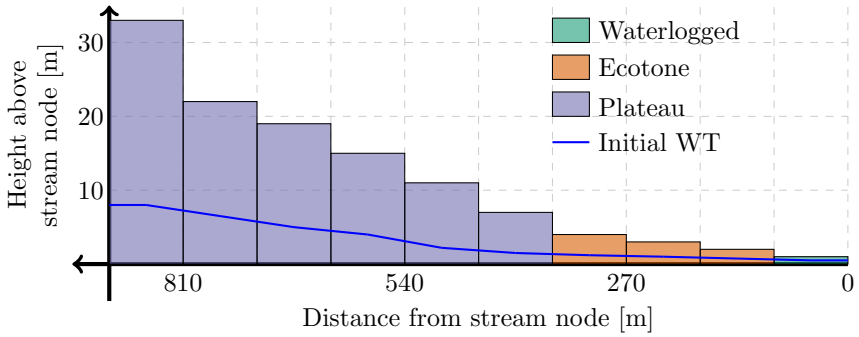

Figure 4.2: Transect elevation, distance from stream, and soil type from Cuartas et al. (2012).

4.2.2 Soil hydraulic properties

A description of the soil and vegetation of the site is provided in (Cuartas et al. 2012), and shows four different soil types in the watershed, with clay-rich soil in the upland areas of the watershed and a clay-poor soil in the lowland regions. These soils are classified by the landscape and terrain classes in (Cuartas et al. 2012; Nobre et al. 2011), and are called (from lowest to highest elevation): waterlogged (abbreviated in the following as W), ecotone (E), slope, and plateau (P). As a result of the transect creation, three soil types: waterlogged, ecotone, and plateau, are present in the transect. Additional on-site soil measurements are available for the catchment up to a soil depth of 14.3 [m] (Broedel et al. 2017). The soil properties as a result of model calibration (Cuartas et al. 2012) or on-site sampling (Broedel et al. 2017) provide a large range of possibly physical soil characteristics. These on-site measurements are limited to areas around the flux tower, and uncertainties related to the calibration of the soil hydraulic properties in (Cuartas et al. 2012) lead to these properties being treated as uncertain in this study.

As in (Dwelle et al. 2018), the van Genuchten-Mualem (vG) soil hydraulic model (Genuchten 1980) is used. In order to have correct dependence of the soil hydraulic model parameters within the uncertainty quantification framework, a pedotransfer

function (PTF) for Brazilian soils (Tomasella et al. 2000) is used. This PTF relies on the soil texture (percentages of sand, silt, and clay), organic carbon, moisture equivalent, and bulk density to fit a second-order polynomial for the vG model variables of α, n, θ_r, θ_s. At the catchment level, with one soil type representing the entire domain, hydrologic outputs are generally most sensitive to the saturated conductivity and anisotropy ratio of the soil (Dwelle et al. 2018). Combined with the physical description of the soil in (Cuartas et al. 2012), only a few variables of the PTF will be treated as uncertain. The soil hydraulic properties are provided in Table 4.1. Given that at least one PTF variable is treated as uncertain for each soil type in the domain, the resulting soil water retention (SWR) characteristics will also be uncertain. As shown in Figure 4.3, each soil type has an envelope of possible curves in the SWR plot. This is valuable as the relationship between soil water content (θ) and soil matric potential (Ψ) plays a role in the energy required for water uptake by plant roots (Hildebrandt et al. 2016).

Table 4.1: Soil hydraulic properties and variables for the pedotransfer function. Values enclosed in brackets ($[a, b]$) represent the lower and upper bounds of the uncertain parameterization.

Parameter	Waterlogged	Ecotone	Plateau
	PTF variables		
F_C	5	25	85
F_S	[5, 50]	[5, 50]	10
α_{cs}	0.5	0.5	0.5
M_e	[0.1, 0.225]	[0.1, 0.225]	[0.1, 0.225]
ρ_b	1.2	1.2	1.2
	Soil hydraulic parameters		
k_{sat}	[35, 85]	[22, 52]	[3, 7]
a_r	[1E-3, 1E-2]	[0.1, 1.5]	[10, 50]

4.2.3 Selection of vegetation traits

A three "big-leaf" representation of the forest canopy is used to represent vertical structure as in Ivanov et al. (2012). The three vegetation layers present are the "top-canopy", "mid-canopy", and "bottom-canopy" trees. The different trees in each layer are assumed to be sufficiently different in their physical and chemical properties that it is prudent to treat them separately, as characterizations are expressed

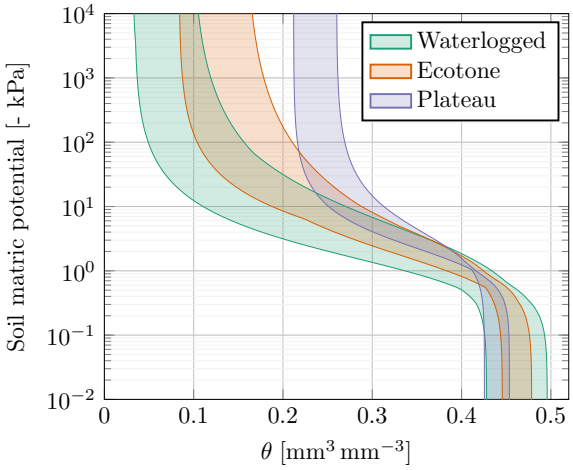

Figure 4.3: Bounds of soil water retention (SWR) curves for uncertain parameterizations.

differently between plants exposed to full sun and those in shaded regions (Taiz and Zeiger 2006). This delineation results in three different plant functional types. For a thorough approach to uncertainty quantification, parameterizations for each function type needs to be taken, and these are given in Table 4.2, where details of how each parameter is used in tRIBS-VEGGIE is provided in (Ivanov et al. 2008a).

Many of the parameters are held constant, but those that are denoted as $[a, b]$ in Table 4.2 are treated as uncertain as uniform random variables $X_i \sim U[a, b]$. These parameters are defined as:

- d_{leaf} [m] is the mean leaf size of the tree;
- χ_L is a parameter controlling the departure of leaf angles from random distribution where values of -1, 0, 1 represent vertical, random and horizontal leaves, respectively;
- $\alpha_{\Lambda}^{\text{leaf}}$ and $\tau_{\Lambda}^{\text{leaf}}$ [-] are the leaf reflectance and transmittance, respectively;
- $\alpha_{\Lambda}^{\text{stem}}$ and $\tau_{\Lambda}^{\text{stem}}$ [-] are the stem reflectance and transmittance, respectively, and the "VIS" and "NIR" identifiers represent the visible and near-infrared spectral bands;
- $V_{\text{max},25}$ [μmol CO_2 m^{-2} s^{-1}] is the maximum catalytic capacity of Rubisco at 25°C;
- $\overline{K}$ [-] is the time-mean PAR extinction coefficient parameterizing the decay of nitrogen content in the canopy;
- m [-] is an empirical slope parameter for a linear model relating the net assimilation rate and stomatal conductance;

- b [μmol m^{-2} s^{-1}] is the minimum stomatal conductance;
- $\epsilon_{3,4}$ [μmol CO_2 μmol^{-1} photons] is the intrinsic quantum efficiency of CO_2 uptake;
- $D_{s,0}$ [Pa] is the Leuning coefficient of stomatal sensitivity to humidity deficit;
- K_c [mm hour^{-1}] is the canopy water drainage rate coefficient;
- g_c [mm^{-1}] is the exponential decay parameter of the canopy water drainage rate;
- Ψ^* [MPa] is the soil water potential at which stomatal closure begins;
- Ψ^w [MPa] is the soil water potential at which plant wilting begins;
- $\Psi^*_{H_2O}$ [MPa] is the soil water potential at which stomatal closure begins due to waterlogged conditions;
- $\Psi^w_{H_2O}$ [MPa] is the soil water potential at which plant wilting begins due to being waterlogged; and
- R_d [-] is the decay parameter for the exponential function of root fraction as a function of depth.

The rooting depth is defined by an exponential root distribution with respect to depth, controlled by the coefficient of decay R_d and the fraction of root biomass R_f that defines the rooting depth. The rooting depth is then defined as

$$Z_r \text{ [m]} = \frac{-\ln(1 - R_f)}{R_d} \tag{4.1}$$

where $R_f = 0.95$ and the ranges for R_d for each plant functional type are in Table 4.2

Stomatal conductance as a function of soil water content is regulated through the use of a soil moisture availability factor $\beta_T \in [0, 1]$, following the relationship of (Feddes et al. 2001):

$$\beta_T = \sum_{i}^{N_{\text{roots}}} \beta_{T,i}(z_i) r_i(z_i), \tag{4.2}$$

$$\beta_{T,i} = \max\left[0, \min\left(1, \frac{\theta_i(z_i) - \theta^w}{\theta^* - \theta^w}, \frac{\theta_i(z_i) - \theta^w_{H_2O}}{\theta^*_{H_2O} - \theta^w_{H_2O}}\right)\right], \tag{4.3}$$

where $\theta_i(z_i)$ [m^3 m^{-3}] is the soil moisture content at the mesh node depth z_i, and the soil moisture contents θ^w, θ^*, $\theta*_{H_2O}$, $\theta^w_{H_2O}$ [m^3 m^{-3}] correspond to the wa-

ter potentials $\Psi^*, \Psi^w, \Psi^*_{H_2O}, \Psi^w_{H_2O}$ [MPa], respectively. Stomatal closure begins at the water potentials $\Psi^*, \Psi^*_{H_2O}$ due to being stressed and waterlogged, respectively. Similarly, plant wilting begins at the water potentials $\Psi^w, \Psi^w_{H_2O}$ due to being stressed and waterlogged, respectively. These potentials are treated as uncertain, and the range of their values is provided in Table 4.2 and Figure 4.4a. Note that the relationship is linear with respect to water potentials, but the relationship to soil moisture contents depends on realized soil water retention parameters for each soil type (Figure 4.3) following the van Genuchten model (Genuchten 1980). Similarly, the values of $\beta_{T,i}(z_i)$ will depend on the root profile for each plant functional type at depth z_i.

This factor accounts for variability in the root profile using weighting by the fraction of root biomass $r_i(z_i)$. A decrease in β_T indicates a control of soil on the transpiration flux. Within the model, $\beta_{T,i}(z_i)$ is computed at hourly intervals for each plant functional type, which are then used to constrain the maximum catalytic capacity of Rubisco (Ivanov et al. 2008a). This approach is a simplified representation of root water uptake, where studies have shown that plants may be able to change their areas of active root water uptake when other parts of the root system are stressed (Sharp and Davies 1985; Garrigues et al. 2006). This compensation behavior cannot be effectively modeled by the Feddes approach of Eq. (4.3).

The leaf area index [m^2 m^{-2}] (LAI) for each layer was determined by averaging LAI derived from LiDAR measurements in a similar forest in the Adolfo Ducke Reserve just north of Manaus, Brazil (Stark et al. 2012), approximately 50 [km] away from the ZF2 study area. These values for LAI were then used to generate seasonal variability of LAI, which was applied to each layer giving the pattern for LAI in Figure 4.4b. The pattern of LAI for each layer are the same and remain constant over the simulation period, but qualitatively agree with the seasonal variation of LAI in other studies (Stark et al. 2012; Poulter et al. 2009; Wu et al. 2016). Importantly, this does not allow for leaf growth/death responses to excess moisture or drought conditions, however the bulk measure of LAI is well represented using this approach (Stark et al. 2012).

Table 4.2: Vegetation biophysical, photosynthesis, interception, and water uptake parameters. Values enclosed in brackets ($[a, b]$) represent the lower and upper bounds of the uncertain parameterization.

Parameter	Top-Canopy	Mid-Canopy	Bottom-Canopy
Biophysical parameters			
d_{leaf} [m]	0.05	0.05	0.05
χ_L [—]	[-0.4, 0.6]	[-0.4, 0.6]	[-0.4, 0.6]
$\alpha_{\Lambda}^{\text{leaf}}$-VIS [—]	0.11	0.05	0.02
$\alpha_{\Lambda}^{\text{leaf}}$-NIR [—]	0.5	0.5	0.5
$\alpha_{\Lambda}^{\text{stem}}$-VIS [—]	0.2	0.2	0.2
$\alpha_{\Lambda}^{\text{stem}}$-NIR [—]	0.45	0.45	0.45
$\tau_{\Lambda}^{\text{leaf}}$-VIS [—]	0.07	0.02	0.01
$\tau_{\Lambda}^{\text{leaf}}$-NIR [—]	0.33	0.32	0.32
$\tau_{\Lambda}^{\text{stem}}$-VIS [—]	1E-4	1E-4	1E-4
$\tau_{\Lambda}^{\text{stem}}$-NIR [—]	1E-4	1E-4	1E-4
Photosynthesis parameters			
$V_{\text{max},25}$ [μmol CO_2 m^{-2} s^{-1}]	[7.25, 76.25]	[7.25, 76.25]	[7.25, 76.25]
$\overline{K}$ [—]	0.35	0.35	0.35
m [—]	[6, 12]	[6, 12]	[6, 12]
b [μmol m^{-2} s^{-1}]	1E4	1E4	1E4
$\epsilon_{3,4}$ [μmol CO_2 μmol^{-1} photons]	0.055	0.06	0.065
$D_{s,0}$ [Pa]	[1E3, 4E3]	[1E3, 4E3]	[1E3, 4E3]
Interception parameters			
K_c [mm hour^{-1}]	0.15	0.15	0.15
g_c [mm^{-1}]	3.7	3.7	3.7
Water uptake parameters			
Ψ^* [MPa]	[-0.6, -0.4]	[-0.6, -0.4]	[-0.6, -0.4]
Ψ^w [MPa]	[-3, -2]	[-3, -2]	[-3, -2]
$\Psi^*_{\text{H}_2\text{O}}$ [MPa]	[-6E-5, -4E-5]	[-6E-5, -4E-5]	[-6E-5, -4E-5]
$\Psi^w_{\text{H}_2\text{O}}$ [MPa]	[-1.2E-5, -8E-6]	[-1.2E-5, -8E-6]	[-1.2E-5, -8E-6]
R_d [—]	[0.08, 0.4]	[0.4, 2]	[2, 14]
Z_r [m]	[7.49, 37.4]	[1.5, 7.49]	[0.21, 1.5]

4.2.4 Uncertainty quantification framework

This study employs an uncertainty quantification framework to allow input parameters to the hydrologic-vegetation model to be uncertain, i.e., uncertain variables in Tables 4.1 and 4.2. These uncertainties are accounted from using polynomial chaos machinery (Le Maître and Knio 2010) combined with Bayesian compressive sensing (Sargsyan et al. 2014). Details of the framework are left to (Dwelle et al. 2018),

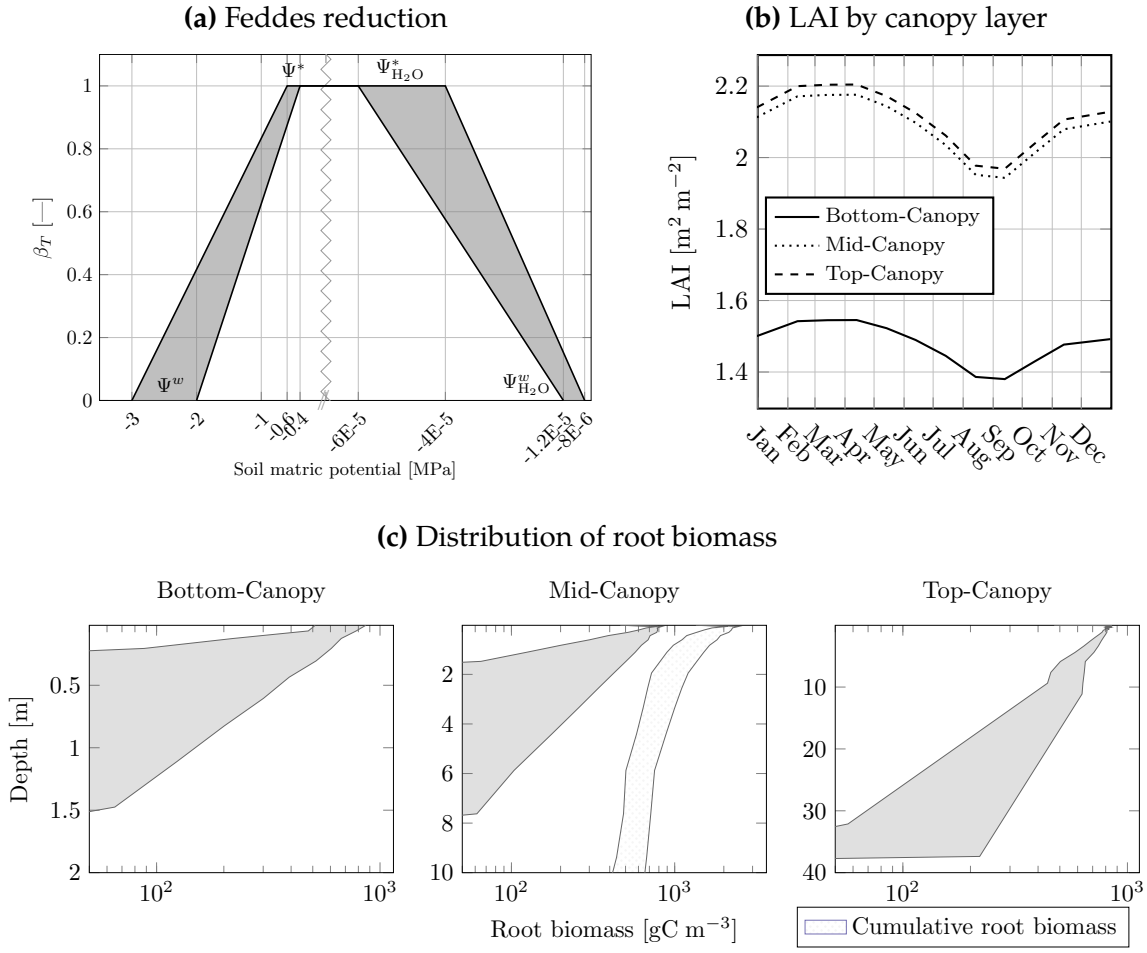

Figure 4.4: Figures for describing vegetation function. The behavior of the soil moisture availability factor regulating stomatal conductance is given in (a), showing the window of possible realizations for all vegetation types given the uncertain priors. The annual-varying LAI for each canopy layer is given in (b), and remains constant for each year in model simulations, and the uncertain bounds of root biomass as a function of depth for each canopy layer is given in (c, shaded areas in gray), with the cumulative root biomass also provided in the center panel (dotted area).

Chapter 2, and references therein; material introduced in this section is included for notation and interpretation of results.

Consider the model $\mathcal{M}$ used in this study, tRIBS-VEGGIE, has inputs $\mathbf{x}$ and an output quantity of interest (QoI) $\hat{y}$. With respect to the framework, the model is a "black box" function that maps uncertain inputs to uncertain output:

$$\hat{y} = \mathcal{M}(\mathbf{x}). \tag{4.4}$$

In the case where the inputs $\mathbf{x}$ are uncertain, they can be represented by a random vector $\boldsymbol{X} \sim \pi(\mathbf{x})$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_M)$ and $\pi(\mathbf{x})$ is a vector of marginal probability density functions describing the variability in each of the M uncertain variables in $\boldsymbol{X}$, which are known and defined prior to modeling. By propagating the uncertain inputs through the model, one can obtain the output QoI as a random variable $Y = \mathcal{M}(\boldsymbol{X})$.

Multiple model evaluations need to be performed in order to get a density for Y. This study employs the Bayesian compressive sensing (BCS) approach to calculating the PCE coefficients $\boldsymbol{c}$ (Sargsyan et al. 2014). The benefit of using the BCS approach is to minimize the "curse of dimensionality" associated with propagating uncertainties from a large number of variables (Caflisch 1998; Davis and Rabinowitz 2007). For example, with $M = 38$ uncertain model parameters and a $p = 4$ order polynomial, from Equation (2.14) there are 111,930 basis terms in Equation (2.13). To solve analytically (e.g., with quadrature methods), one would need to evaluate $\mathcal{M}$ 111,930 times. Consider simulation times of M to be 1 second, 1 minute, and 1 hour; these would require total wall clock times of 1.3 days, 78 days, and 12.7 years, respectively. Even the simplified domain used in this study requires simulation times on the order of hours, and without the use of BCS the investigation of a large number of uncertain model parameters would not be possible.

Once constructed, the surrogate model in Equation (2.13) can be used for model calibration through inference from observational data at the study location. The parameter priors in this study are uniform and given in Tables 4.1 and 4.2, and the posteriors can be calculated conditional on observational data as in Eq. (2.50). where σ^2 is also introduced as an additional hyper parameter. To infer values for uncertain model parameters, the posterior from Eq. (2.50) needs to be sampled. Such methods as the Metropolis-Hastings MCMC or another similar sampling method (MacKay 1998) can be used. One can sample the posterior directly from simulation result of $\mathcal{M}$, but it is computationally expensive, and it is therefore expeditious to replace the deterministic $\mathcal{M}$ in the posterior with its PCE surrogate $\mathcal{M}^{\text{PC}}$.

4.2.5 Selection of quantities of interest

Quantities of interest were selected such that they coincided with observable data from ZF2 flux tower at the highest elevation of the transect. Data that were available from the ZF2 flux tower included gross primary productivity (GPP) [gC m^{-2} day^{-1}], evapotranspiration (ET) [mm day^{-1}], and soil water content in the top five meters (θ_{5m}) [mm^3 mm^{-3}]. Each of these observables are aggregated to the daily time scale, and the daily value is aggregated to a monthly average if there are at least ten daily values available to aggregate. Soil water content is observed at the flux tower in a five-meter deep soil pit, with measurements reported at one meter increments, and therefore θ_{5m} is calculated by taking the mean of these values. Given that the footprint of the K34 flux tower is estimated to be approximately 2-3 [km^2] (Araújo et al. 2002), values for the GPP and ET quantities of interest are averaged over all 10 computational nodes.

The time resolution of data is not regular across the simulation period of 2000-2005, with a significant number of months not having a record. The resulting QoIs are shown in Figure 4.6. These are the QoIs used for parameter inference in this study, representing plant function (GPP and ET) as well as hydrologic processes in the vadose zone (θ_{5m}).

4.2.6 Construction of performance QoI

In order to investigate hydrologic controls on plant function, a new performance metric was established. The metric was selected from the perspective of a cost-benefit analysis, where the allocation of carbon to the root system must result in a positive carbon growth over the life of the tree, e.g., the ability of a tree to respond positively in drought conditions, increasing carbon uptake (Huete et al. 2006; Saleska et al. 2003; Ivanov et al. 2012). Such a metric will follow the form of

$$\mathcal{F} = \text{Benefit} - \text{Cost}. \tag{4.5}$$

The "Benefit" represents plant 'biochemical performance'—a variable that measures the plant's evolutionary objective to increase its reproductive capacity. Gross Primary Production (GPP) is used in this study, as it is a well-studied and understood metric for forest photosynthetic response with many studies performed

in the Amazon rainforest (Saleska et al. 2003; Graham et al. 2003; Poulter et al. 2009; Brando et al. 2010), consistent with the Darwinian approach in ecohydrology developed by Eagleson (2005), while "Cost" is to be determined. We only consider "Cost" to be a function of hydrologic conditions, aiming to quantify the cost related to protection against drought such as design of water transport tissues in xylem that can sustain low water potentials without embolism. The goal of this metric is to allow a minimum 'hydraulic effort' on the part of the plant, given hydrologic controls of environment in which this plant is located. Therefore the metric should contain information about both plant function and hydrologic states.

However, there is not a well-defined metric to represent hydrologic control on plant function. As a proxy for this type of behavior, we propose the use of exported energy from root water uptake (Hildebrandt et al. 2016), which is defined as the product of the potential and flow across the root collar of the plant:

$$J_{\mathrm{E,exp}} = g\rho_w\psi_X J_{\mathrm{wu}}, \tag{4.6}$$

ψ_X [m] is the xylem water potential, J_{wu} [$\mathrm{m}^3\ \mathrm{s}^{-1}$] is the total root water uptake, and ρ_w, g are the density of water and gravitational acceleration, respectively. For simplicity, it is assumed that the total root water uptake is equal to the transpiration flux for each plant functional type, which is obtainable as model output from tRIBS-VEGGIE. Similarly, the xylem potential needs to be calculated for each plant functional type, and is given by:

$$\psi_X = \frac{J_{\mathrm{wu}} + \sum_{i=1}^{N}\left(\psi_{M,i}K_{r,i}\right)}{\sum_{i=1}^{N}K_{r,i}}, \tag{4.7}$$

where $K_{r,i}$ [$\mathrm{m}^2\ \mathrm{s}^{-1}$] is the effective radial conductivity of roots in soil compartment i, and $\psi_{M,i}$ [m] is the water potential in each soil compartment, where N is the number of soil compartments considered. The soil compartments are the computational elements within the root zone for each plant functional type, which is a function of the uncertain rooting depth exponential parameter R_d. Therefore, the summation carried out in Equation (4.7) will carry with it propagated uncertainty.

The effective radial conductivity, $K_{r,i}$ is not explicitly computed in tRIBS-VEGGIE, so it is estimated using available information. $K_{r,i}$ can be expressed as the ease of flow of soil water into the root xylem, and is traditionally normalized by root sur-

face area, i.e., $K_{r,i} = L_r A_r$. Here, L_r has units of [m^3 s^{-1} m^{-2} m^{-1} head], and A_r is the root area. L_r is an intrinsic ecophysiological value of a root or root systems, with values $O(10^{-6})$ [m^3 s^{-1} m^{-2} m^{-1}head]. The extrinsic value of root area still needs to be determined. Given a dry root biomass density [g m^{-3}], one can determine the root length density (RLD) [m m^{-3}] by using the measure of length of root grown per gram carbon invested, called specific root length [m g^{-1}]. One can then get root area from the product of root length density [m m^{-3}], root diameter (d_{root}) [m], and volume of soil (V_s) [m^3]. Therefore, one can write the effective radial conductivity as

$$K_{r,i} = L_r \text{RLD}_i d_{\text{root}} V_s, \tag{4.8}$$

and Equation (4.7) can be rewritten as

$$\psi_X = \frac{J_{\text{wu}} + L_r d_{\text{root}} V_s \sum_{i=1}^{n} \left(\psi_{M,i} \text{RLD}_i\right)}{L_r d_{\text{root}} V_s \sum_{i=1}^{N} \text{RLD}_i}, \tag{4.9}$$

where V_s is a constant due to the subsurface mesh having a regular discretization, otherwise it would need to be included in the summation. The assumptions for Equation (4.9) include: there is a single root type throughout the entire soil domain; this root type has a constant diameter and specific root length, and that the root biomass density is calculated from the fraction of roots in the subsurface as

$$R_b(z) = \frac{R_{\text{frac}}(z)}{\Delta z} \overline{R_{b,\text{total}}}, \tag{4.10}$$

where $R_b(z)$ [-] is the fraction of roots in the system present in the soil node at depth z, and $\overline{R_{b,\text{total}}}$ [g m^{-2}] is the root biomass averaged over depth. Given Equations (4.9) and (4.10), the exported energy from Equation (4.6) can be calculated from available information from tRIBS-VEGGIE.

To aid in comparison of performance across the transect and time, values of GPP and $J_{\text{E,exp}}$ are scaled from $[0, 1]$, such that for a variable γ the scaling is $(\gamma - \gamma_{\min})/(\gamma_{\max} - \gamma_{\min})$, where γ can be either GPP or $J_{\text{E,exp}}$. The minimum and maximum are computed over all simulations for each month of simulation, pixel, and PFT. This allows for a comparison between parameterizations, leading to 2,160 quantities of interest (72 months $\times$ 10 pixels $\times$ 3 PFTs). To compare the performance of plant function and a hydrologic control on this function, the difference between

the two metrics are taken and the performance quantity of interest becomes:

$$\mathcal{F} = \widetilde{GPP} - \widetilde{J_{\mathrm{E,exp}}}, \tag{4.11}$$

where the tilde represents the scaled version of the quantity in Eq. (4.5). Generally, one could classify four regions of performance based on the saturation of the soil (Figure 4.5); when GPP is relatively high, one could have performance in wetter ($\mathcal{F}_{\mathrm{HW}}$) or drier ($\mathcal{F}_{\mathrm{HD}}$) conditions. Similarly, one can have lower performance in wetter or drier conditions, $\mathcal{F}_{\mathrm{LW}}$ and $\mathcal{F}_{\mathrm{LD}}$, respectively. It is reasonable to assume that the best performance will occur in the region of $\mathcal{F}_{\mathrm{HW}}$, due to no limitation of water for plant function. However, in the case with a reduction in stomatal conductance in waterlogged conditions (Eq. (4.3), Fig. 4.4a), the upper right quadrant of Fig. 4.5 may not be the optimal strategy. For example, this could be the case where one has a deep root system in an area of shallow water table, where much of the root biomass will be in a waterlogged area and therefore experience a large overall reduction in stomatal conductance, thereby pushing the performance into the $\mathcal{F}_{\mathcal{LW}}$ quadrant.

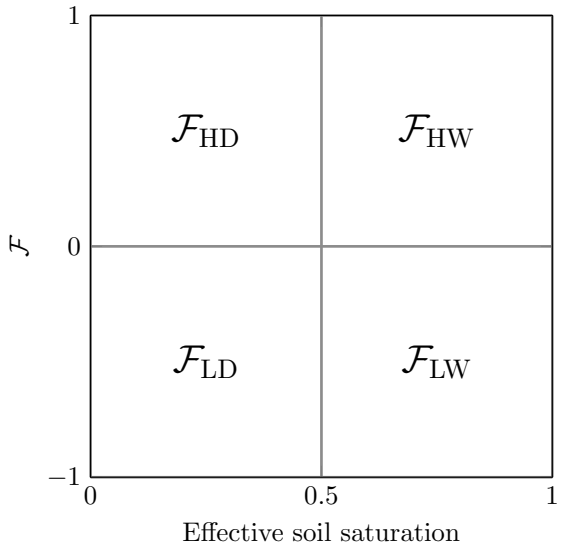

Figure 4.5: Qualitative classification of the performance metric from Equation (4.11) for different effective soil saturation.

A valuable plant functional strategy would be the ability to operate in the $\mathcal{F}_{\mathrm{HD}}$ quadrant, where vegetation production can be maintained in dry conditions with relatively high xylem potentials. This would indicate a plants ability to succeed

in drought-like conditions, like those experienced in Amazonia in 2005, 2010, and 2015 (Phillips et al. 2009; Lewis et al. 2011; Santos et al. 2018; Saatchi et al. 2013; Brum et al. 2018).

4.3 Results

This section provides an overview of the construction of surrogates and inference of model parameters using these surrogates. The data used to construct the surrogate and perform inference is given in Figure 4.6. For inference, one wants training data that overlaps observations, which is the case for the vast majority of the QoIs, with a few exceptions in the ET time series.

4.3.1 Surrogate construction

Polynomial chaos surrogates $\mathcal{M}^{\text{PC}}$ were constructed as in Equation (2.13) for the observed monthly quantities in Figure 4.6. A total of 1,000 training, and 150 validation simulations were performed using a 96-core computational cluster. A limit for wall clock time was set for each simulation, which resulted in a total of 983 completed training and 150 completed validation simulations. To test the accuracy of the surrogate, a qualitative comparison can be made by comparing simulation and surrogate values in Figure 4.7, with the relative error from Eq. (3.5) given above each subplot. Due to the number of data points in these plots (983 training simulations $\times\{61, 45, 46\}$ QoIs = $\{59963, 44235, 45218\}$ training data records for {GPP, ET, SM} QoI groups, respectively), data are hexagonally binned to display the distribution of fit. The plots show that the surrogate slightly overestimates tRIBS-VEGGIE for GPP and ET, but has very good agreement ($L_{2,rel}$ < 1E-2) for θ_{5m}.

These surrogates were constructed at a total polynomial order of $p = 4$, which was chosen as it provided error agreement between training and validation simulations. Other polynomial orders between 2-5 were tested, but these other orders either increased the errors $L_{2,\text{rel}}$ in Figure 4.7, or decreased the training errors at the expense of the validation errors, where the latter behavior indicates that overfitting is occurring around the parameter space of the training samples.

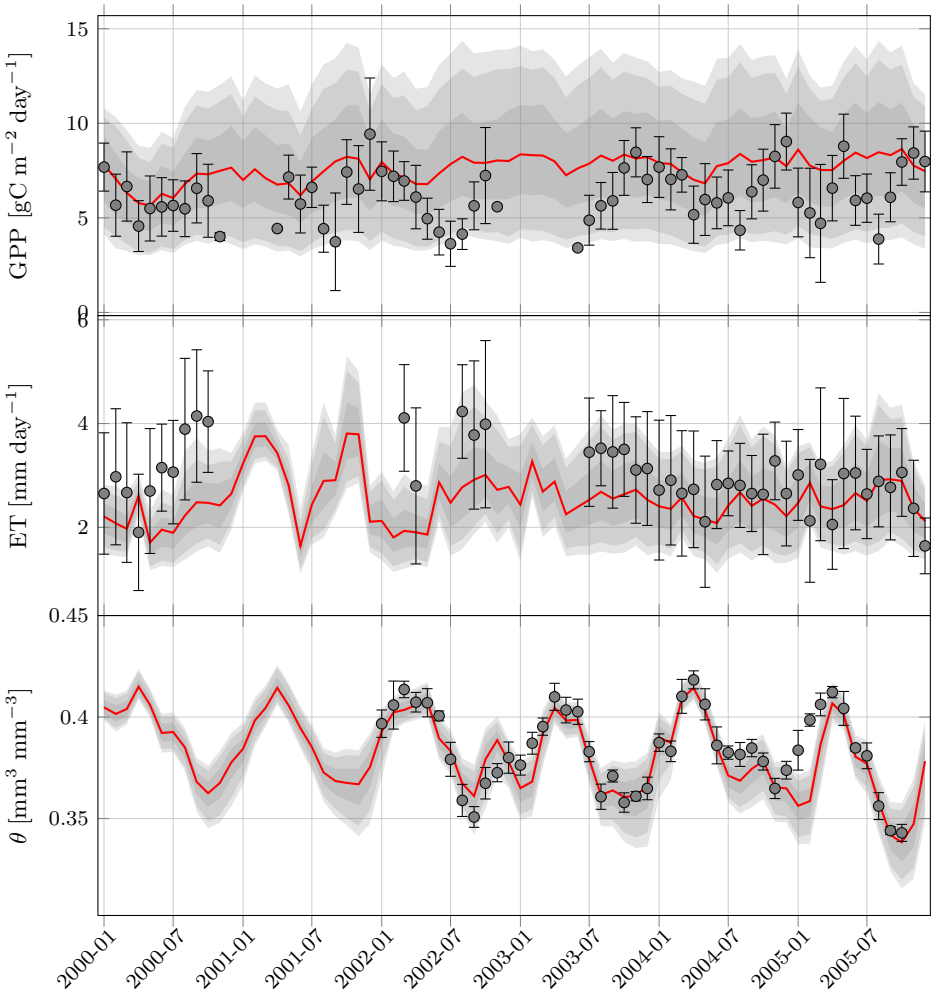

Figure 4.6: Observed quantities of interest and training simulations of tRIBS-VEGGIE. Observed quantities are displayed as circles with the error bars representing the standard deviation of the daily recorded data used to aggregate into a monthly mean. The shaded regions represent the training simulations from tRIBS-VEGGIE that are used for construction of the surrogate and inference. The shading levels represent the 5/95, 10/90, and 25/75 percentiles of training simulations, and the red line in each plot is the median value from training simulations.

Sensitivity of observed quantities

One of the benefits of using polynomial chaos based methods for surrogate modeling is that one obtains global sensitivity information "for free" from the calculation of the polynomial coefficients in Equation (2.13). The main and joint sensitivities of the 38 uncertain parameters, averaged over all QoIs, is provided in Figure 4.8. The

sparsity of the lower-diagonal indicates sparsity in the polynomial basis, allowing for deployment of the Bayesian compressive sensing approach to calculating PC coefficients. Although sparse, 17 parameters still have a averaged main effect contribution $> 10^{-3}$. One can see that a small subset of the parameters represent contributions to the variation in realized values for QoIs. More specifically, the largest sensitivity contributions comes from soil hydraulic parameters, rooting depths, and $V_{\text{max},25}$ for the top-canopy plant functional type.

4.3.2 Parametric inference

Given acceptable performance of the surrogates constructed in Section 4.3.1, one can perform inference using the surrogate. Inference was performed using the errors associated with the standard deviations shown in Figure 4.6. Moment estimates of the posterior distributions are provided in Tables 4.3 and 4.4 for the soil and vegetation parameters, respectively. Given the limitations of plotting pairwise joint distributions of the large number of parameters, plots of the marginal distributions for the ten most impactful parameters, as measured by main effect sensitivity, are provided in Figure 4.9.

These posteriors provide information about the bulk soil properties along the transect, and provide insight into the differences between plant functional types along a transect. The inferred hydraulic parameters coincide with the parameters in Cuartas et al. (2012) where different soil types are present at differing depths. The geometric means over the first three soil layers (to a depth of 4.8 [m]) in Cuartas et al. (2012) for k_s are 62.4, 19.2, 4.78 [mm hr^{-1}] and for a_r are 8.3E-4, 0.62, and 36.1 [-] for the soil types of plateau, ecotone, and waterlogged, respectively. The inferred MAP values given one soil layer in this work for k_s are 41.6, 22.1, and 4.28 [mm hr^{-1}], and the inferred MAP values for a_r are 1.24E-3, 0.87, and 44.0 [-] for the plateau, ecotone, and waterlogged soil types, respectively. Given the differences across models and different simulations assumptions, e.g., multiple soil layers in (Cuartas et al. 2012) vs. one soil layer in this study, there is good agreement between soil parameters, with the added benefit that given the probabilistic nature of this study, the full posterior can be used to propagate remaining uncertainty in these parameter values through future simulations.

Less data are available in literature to compare the vegetation parameters to the

ones inferred in this study. Indeed, this is often the case in ecohydrologic studies as *in situ* measurements are difficult to obtain for many vegetation parameters, and therefore many modeling assumptions are made in order to carry out modeling efforts. However, the values most impactful vegetation parameters — $V_{\text{max},25}$ — are consistent with observations in the K34 watershed (Santos et al. 2018).

Inference provided a posterior chain that was sampled 3,200 times to compute monthly average GPP, ET, and SM in Figure 4.10. Comparing the posterior simulations in this figure with the training simulations in Figure 4.6, one sees that a much better fit of the data is using the calibrated model. This is aided to in a large part by the range of training simulations available in Fig. 4.6, e.g., in the plot for GPP training simulations overlapped with nearly every available observation. The dashed red line in Fig. 4.10 is the *median* of the training simulations, and the solid red line is the median of the posterior simulations. One important thing to note is that the median of each plot is *not* the values for the median simulation, but rather the median across all simulations. Where there is a large amount of uncertainty in simulations, as in the training GPP simulations, the pattern of the median value may not follow physical expectations. However, when uncertainty in the simulations is constrained, one will begin to see the more physical behavior of the model, (as in, e.g., SM Fig. 4.6 and all plots in Fig. 4.10).

4.3.3 Balance between productivity and hydraulic effort

Performance as a function of landscape position

As the performance metric $\mathcal{F}$ is not observed along the transect—GPP is measured, but root zone soil moisture is not—forward UQ was performed on the metric. For this to take place, $\mathcal{F}$ was calculated as an aggregated output of tRIBS-VEGGIE from the 983 training and 150 validation simulations described in 4.3.1, which was then used to construct surrogates for $\mathcal{F}$ at each pixel, in each month, and for each of the three PFTs. Once the surrogates were constructed, the posterior parameter values from 4.3.2 were propagated through the surrogate model. The change in $\mathcal{F}$ along the length of the transect is given in Figure 4.11, with the top three contributors to fraction of variance of $\mathcal{F}$ plotted in the right column.

Given that the performance metric $\mathcal{F}$ is a balance between productivity and hydraulic effort (Eq. (4.11)), the presence of $V_{\text{max},25}$ as the main contributor to vari-

Table 4.3: Posterior values of soil hydraulic properties and variables for the pedotransfer function. The estimates for *maximum a posteriori* (MAP), mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) are provided for the waterlogged, ecotone, and plateau soil types present in the transect (Figure 4.2).

X_i	Waterlogged			Ecotone			Plateau		
	MAP	$\hat{\mu}$	$\hat{\sigma}$	*MAP*	$\hat{\mu}$	$\hat{\sigma}$	*MAP*	$\hat{\mu}$	$\hat{\sigma}$
				PTF variables					
F_S	37.9	37.5	6.0	41.4	31.1	4.16	44.0	42.1	4.46
M_e	0.188	0.195	1.32E-2	0.172	0.16	7.39E-3	1.73E-1	2.15E-1	5.87E-3
				Soil hydraulic parameters					
k_{sat}	41.6	44.6	4.68	22.1	22.7	5.29E-1	4.28	4.47	1.77E-1
a_r	1.24E-3	2.02E-3	1.06E-3	1.4	8.69E-1	4.12E-2	44.0	42.1	4.46

Table 4.4: Parameter posterior values for vegetation biophysical, photosynthesis, interception, and water uptake parameters.

Parameter	Top-Canopy			Mid-Canopy			Bottom-Canopy		
	MAP	$\hat{\mu}$	$\hat{\sigma}$	MAP	$\hat{\mu}$	$\hat{\sigma}$	MAP	$\hat{\mu}$	$\hat{\sigma}$
Biophysical parameters									
χ_L	5.71E-1	3.55E-1	9.84E-2	-1.79E-1	9.34E-2	7.27E-2	-1.07E-1	-2.87E-2	1.76E-1
Photosynthesis parameters									
$V_{\text{max},25}$	21.4	22.8	1.41	30.0	34.2	4.9	50.1	36.5	8.36
m	11.4	9.41	4.11E-1	11.9	11.7	2.66E-1	8.06	9.0	7.92E-1
$D_{s,0}$	2760	2580	334	2880	2630	279	2630	2850	580
Water uptake parameters									
Ψ^*	-4.97E-1	-5.19E-1	1.89E-2	-5.03E-1	-5.16E-1	1.75E-2	-5.23E-1	-4.77E-1	1.71E-2
Ψ^w	-2.71	-2.43	8.07E-2	-2.37	-2.29	9.03E-2	-2.44	-2.68	1.52E-1
$\Psi^*_{\text{H}_2\text{O}}$	-5.23E-5	-4.77E-5	3.61E-6	-5.01E-5	-4.99E-5	3.21E-6	-4.30E-5	-4.9E-5	4.04E-6
$\Psi^w_{\text{H}_2\text{O}}$	-9.92E-6	-9.59E-6	5.61E-7	-9.09E-6	-1.01E-5	5.34E-7	-1.002E-5	-9.56E-6	2.69E-7
R_d	1.94E-1	2.53E-1	0.962	1.3	1.71	1.8E-2	4.79	4.92	8.04
Z_r [m]	15.4	11.8	3.11	2.3	1.75	0.166	0.625	0.608	0.37

ation in $\mathcal{F}$ is expected for each plant functional type. Values for $\mathcal{F}$ are nearly always in the "benefit" region ($\mathcal{F} > 0$), which is indicative of the calibrated parameter values and model corresponding to a "wet" domain. However, the trend of $\mathcal{F}$ across the transect depends on the PFT. For the top- and mid-canopy trees, $\mathcal{F}$ decreases approaching the stream node, indicating that the deeper roots in the profile are becoming waterlogged and are limiting stomatal function and decreasing GPP. The decrease in variability of $\mathcal{F}$ in the top-canopy (Fig. 4.11) are related to a correlation between GPP and $J_{\mathrm{E,exp}}$, as the latter is a product of the tree's transpiration and the xylem potential. Therefore, when the deep root profile of the top canopy is very saturated, both its GPP and water uptake will be very low, leading to a decrease in the variability in $\mathcal{F}$, which can be seen at distances less than 200 [m] from the stream node. For the bottom-canopy trees, the inverse relationship is seen, performance increases with proximity to the stream node. If one considers the shallow root zone, the bottom-canopy trees will be much more susceptible to variations in near-surface soil moisture, and the variability in near-surface soil moisture decreases closer to the stream node, where a shallower water table will provide capillary rise and a more consistent water supply for photosynthetic function.

The change of the performance metric with respect to mean water table depth is given in Figure 4.12. The figure shows the relative contribution of $\widetilde{\mathrm{GPP}}$ to $\mathcal{F}$, meaning the difference between the two is the measure for hydraulic effort, $J_{\mathrm{E,exp}}$. This hydraulic effort stays similar in magnitude across the mid- and bottom-canopy layers, but appears inversely correlated to water table depth for the top-canopy. Two possible explanations are: (1) as absolute GPP decreases, so does the hydraulic effort. This coincides with the appearance of transpiration in the calculation of hydraulic effort (J_{wu} in Eq. (4.6)). where the absolute values of GPP in the top canopy are larger than those in the mid- and bottom-canopy. (2) The much deeper roots in the top-canopy (see Table 4.4) mean that a much higher proportion of the root zone is saturated in the lowland areas of the transect, which is therefore reducing the hydraulic effort.

Sensitivity to soil and vegetation parameters

The interaction between the rooting depth and water table position exhibits a control on the relative performance of the top-canopy vegetation as it is located in regions of the transect with shallower water table (Fig. 4.11), as judged by sensitivity

of the rooting depth parameter on the performance metric $\mathcal{F}$. As estimated in Section 4.3.2, MAP rooting depths are 15.4, 2.3, and 0.63 [m] for the top, mid, and bottom-canopy plant functional types, respectively. This hydraulic control is also seen in the bottom-canopy tree type, but with an opposite interaction; changes in the rooting depth of the bottom canopy are more sensitive in the lowland regions of the transect. In this region, closer to the stream node, the rooting depth is much closer to the water table and therefore is able to access this groundwater through capillary rise or may become waterlogged due to root biomass being below the water table. This sensitivity is not present in the mid-canopy tree type, with a possible explanation that the distribution of root biomass (Fig. 4.4c) results in the majority of roots being in non-waterlogged regions throughout the majority of the transect and therefore does not have a large impact on performance.

Given the UQ framework, it is possible to investigate other possible controls on vegetative function, as long as those controls can be described by one of the parameters treated as uncertain. From Fig. 4.11, one sees that soil parameters also impact the performance metric $\mathcal{F}$. Of specific note is the anisotropy ratio in the waterlogged region (labeled "W: a_r" in the figure). This parameter controls the flow of water out of the domain through the seepage face, and therefore the water table depth at the last node in the transect as well as having a large impact on the water table depths in its upslope neighbors, as witnessed by the sensitivity experienced in distances < 250 [m] away from the stream node. For the mid-canopy plant functional type M_e, the moisture equivalent of the soil in the plateau region impacts performance. This is likely due to M_e being the only parameter augmenting the soil water retention curve for the plateau soil type (Tab. 4.1, Fig. 4.3).

To separate the contribution of vegetation and soil parameterizations, Figure 4.13 provides the three vegetative and soil parameters with the highest sensitivity contribution to $\mathcal{F}$. As in Fig. 4.11, one sees that $V_{\text{max},25}$ is the dominant parameter with respect to sensitivity for each plant functional type. The most sensitive parameters within each PFT generally come from that PFT (e.g., top-canopy traits are impactful only to top-canopy $\mathcal{F}$), with the exception of the mid-canopy PFT having a sensitivity contribution coming from χ_L from the top-canopy PFT. This indicates that χ_L, the parameter controlling leaf orientation, has a significant effect on the performance of vegetation in the mid-canopy layer. This can be explained by χ_L of

the top-canopy controlling the amount of light available to the mid-canopy, showing that there are interactions between functional types important for vegetative performance. Additionally, Fig. 4.13 displays a transition from canopy-dominated parameters ($V_{\max,25}$, m, χ_L) to root-dominated parameters (R_d, Ψ^*) going from top- to bottom-canopy functional types.

One can partition the soil parameters present in Fig. 4.12 into two classes, those related to water infiltration and drought (M_e, k_s) and those related to drainage (a_r). In particular, the top-canopy PFT is sensitive to drought and infiltration in the upland areas of the transect, with no real soil sensitivities present in the lower areas of the transect. This can be explained by the fact that the rooting depths (Table 4.4), when compared with the water table depths (Fig. 4.12), indicates that a large portion of the root zone will be saturated in the ecotone and waterlogged regions. With the mid- and bottom-canopy functional types, one see a dominance of drainage-related sensitivities to $\mathcal{F}$. In particular, the anisotropy ratio for the waterlogged region controls the drainage from the transect, and therefore water table depths, and impacts the waterlogged and ecotone regions of the transect. Additionally, both of these PFTs show a preference towards the importance of infiltration in the upland areas of the transect, which would then control the soil moisture content impacting ψ_X and therefore $\mathcal{F}$.

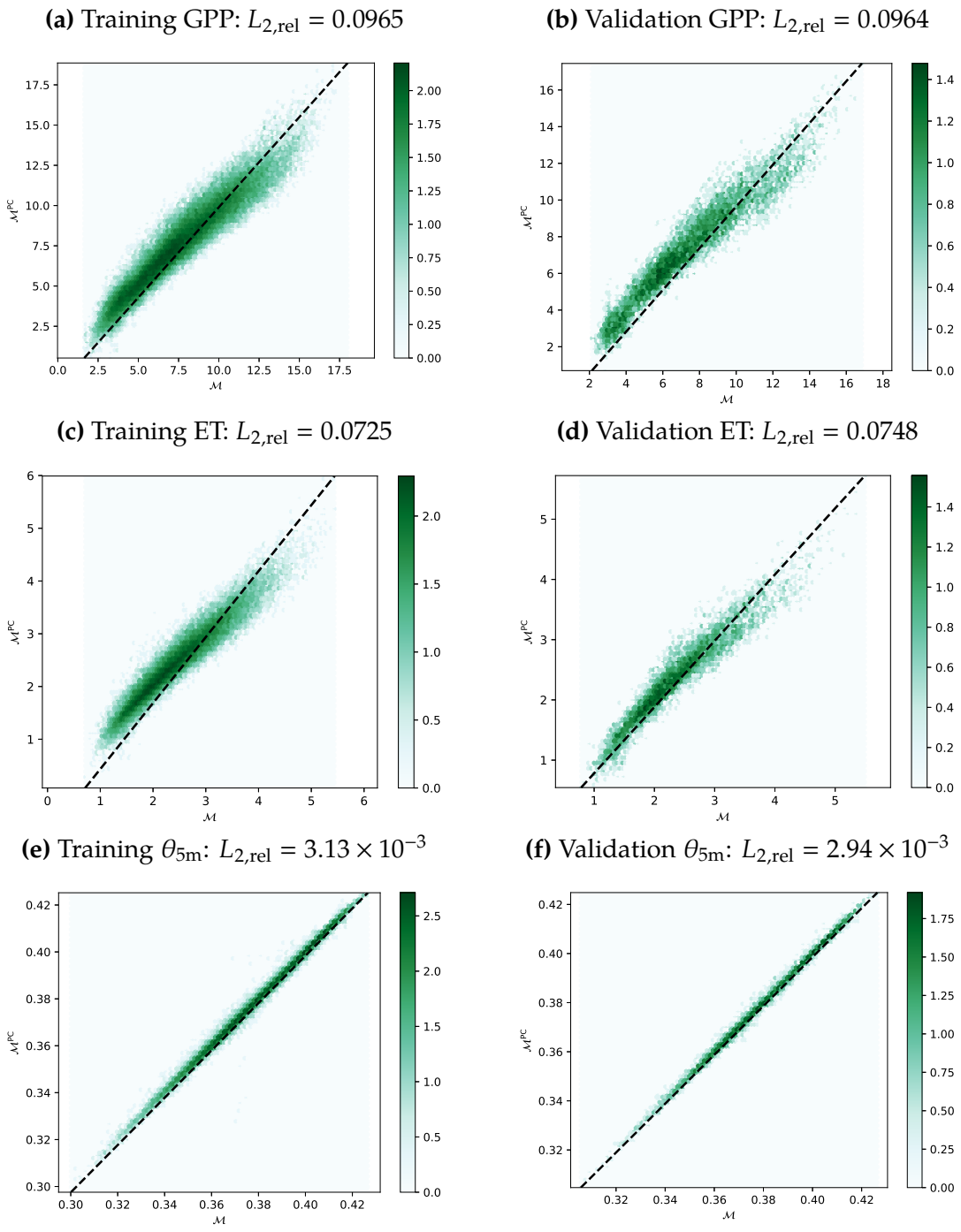

Figure 4.7: Comparison of tRIBS-VEGGIE and constructed surrogate for PC order $p = 4$. The surrogates are separated into six groups based on the QoI and simulation type (training vs. validation). Due to the large number of simulations performed, values are hexagonally binned, with the color scale representing the $\log_{10}$ count of simulations in each bin. A $y = x$ line is added to each plot, such that points on this black line represent agreement between tRIBS-VEGGIE and the constructed surrogate. The mean relative error ($L_{2,\text{rel}}$) of the surrogates for each group are given in the plot captions.

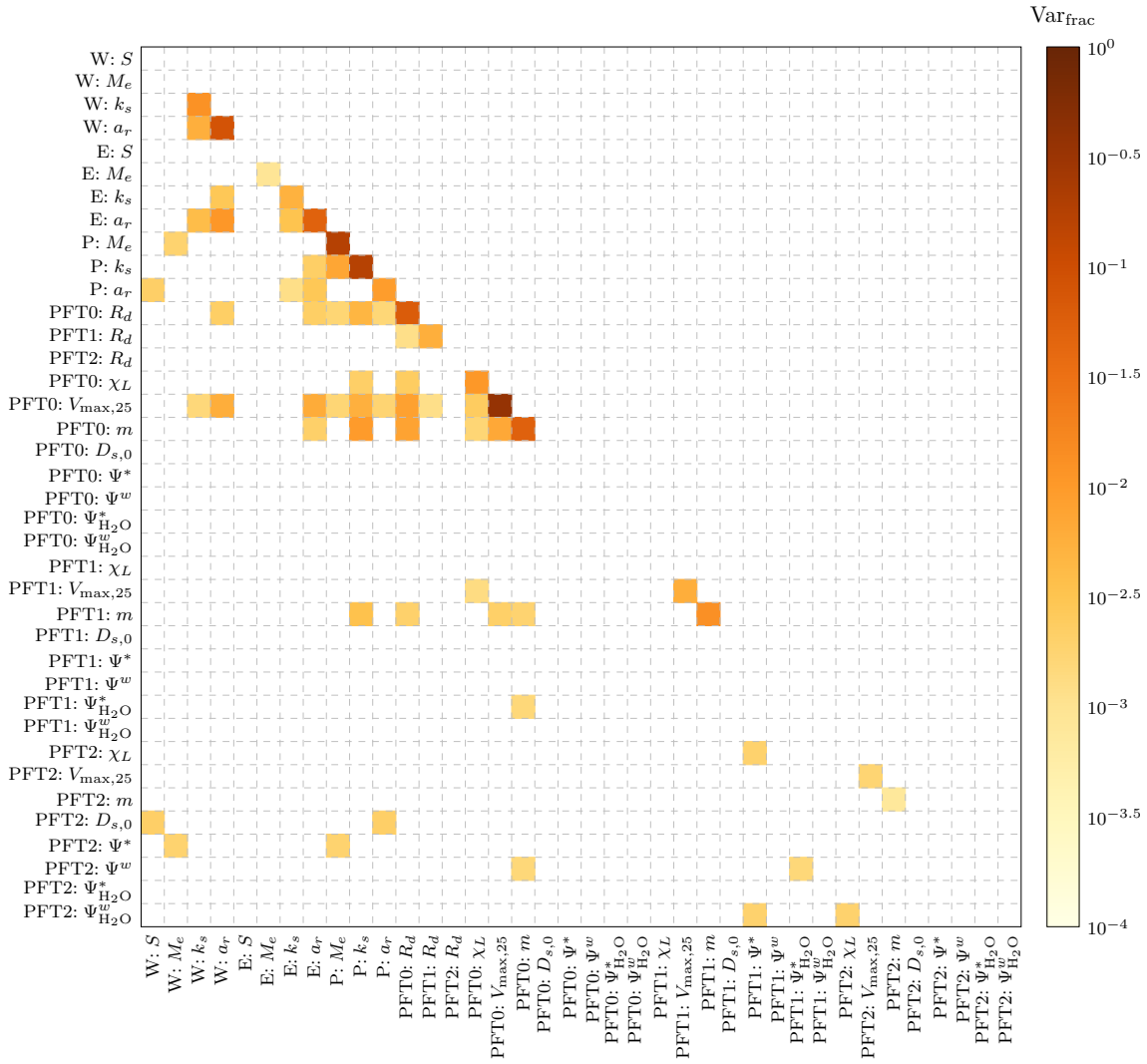

Figure 4.8: Averaged sensitivity contributions to all QoIs. The diagonal of the plot contains the main effect sensitivity contribution, and the lower diagonal plot is the joint sensitivity contribution from interaction between parameters. Sensitivity values were averaged for all sensitivity contributions greater than 10^{-5}, with inclusion in the plot requiring an average contribution of at least 10^{-3}, with color scaling on the plot selected for visibility.

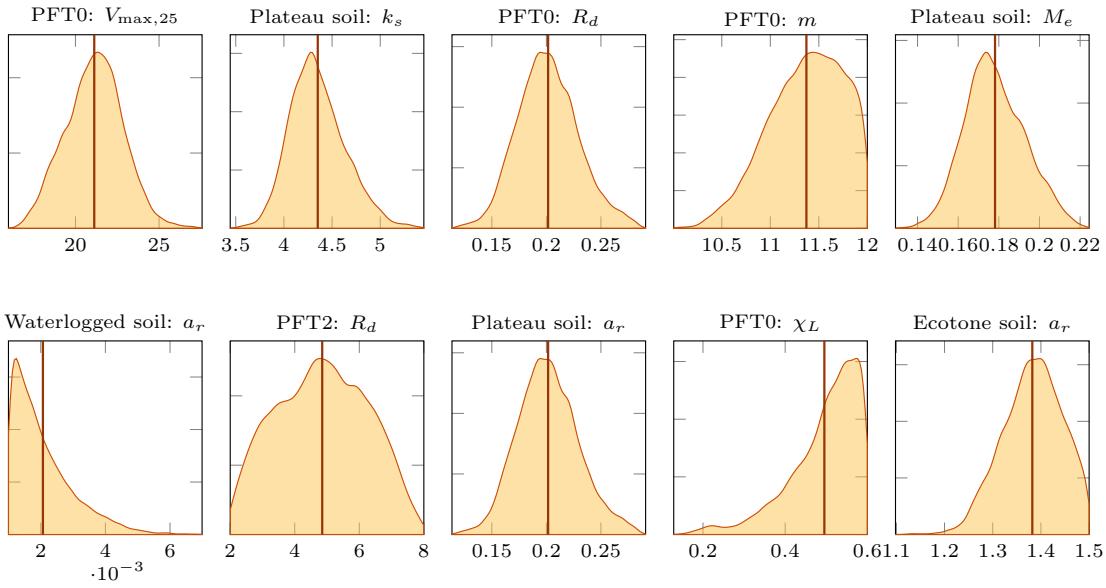

Figure 4.9: Inferred parameter values for parameters with highest main contribution to variance, column then row-wise. Therefore, the top-left plot is the parameter with the largest main sensitivity, then the top right, bottom left, and bottom right have the 5th, 6th, and 10th largest main sensitivity, respectively.

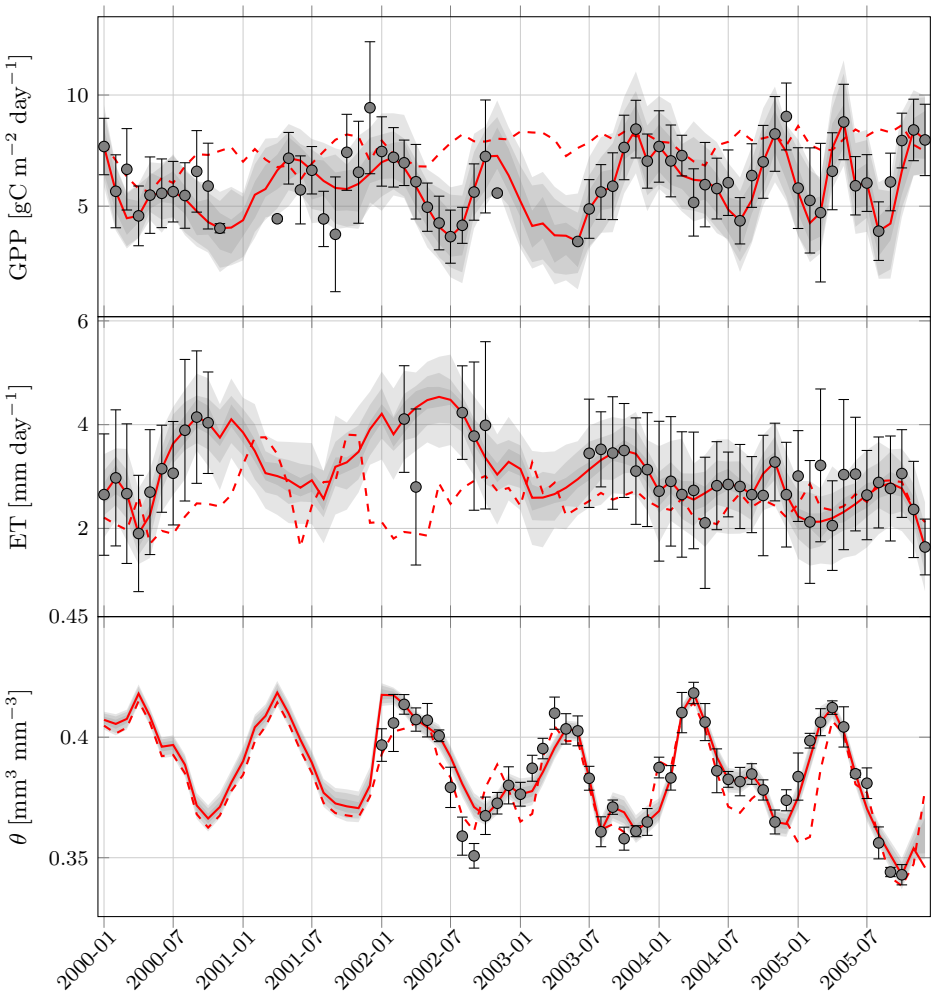

Figure 4.10: Observed quantities of interest generated with constructed surrogates and the posterior parameters. Presentation of quantities is the same as in Figure 4.6, except the solid red line is the median of the simulations generated through posteriors, and the dashed red line is the median of the training simulations from Figure 4.6.

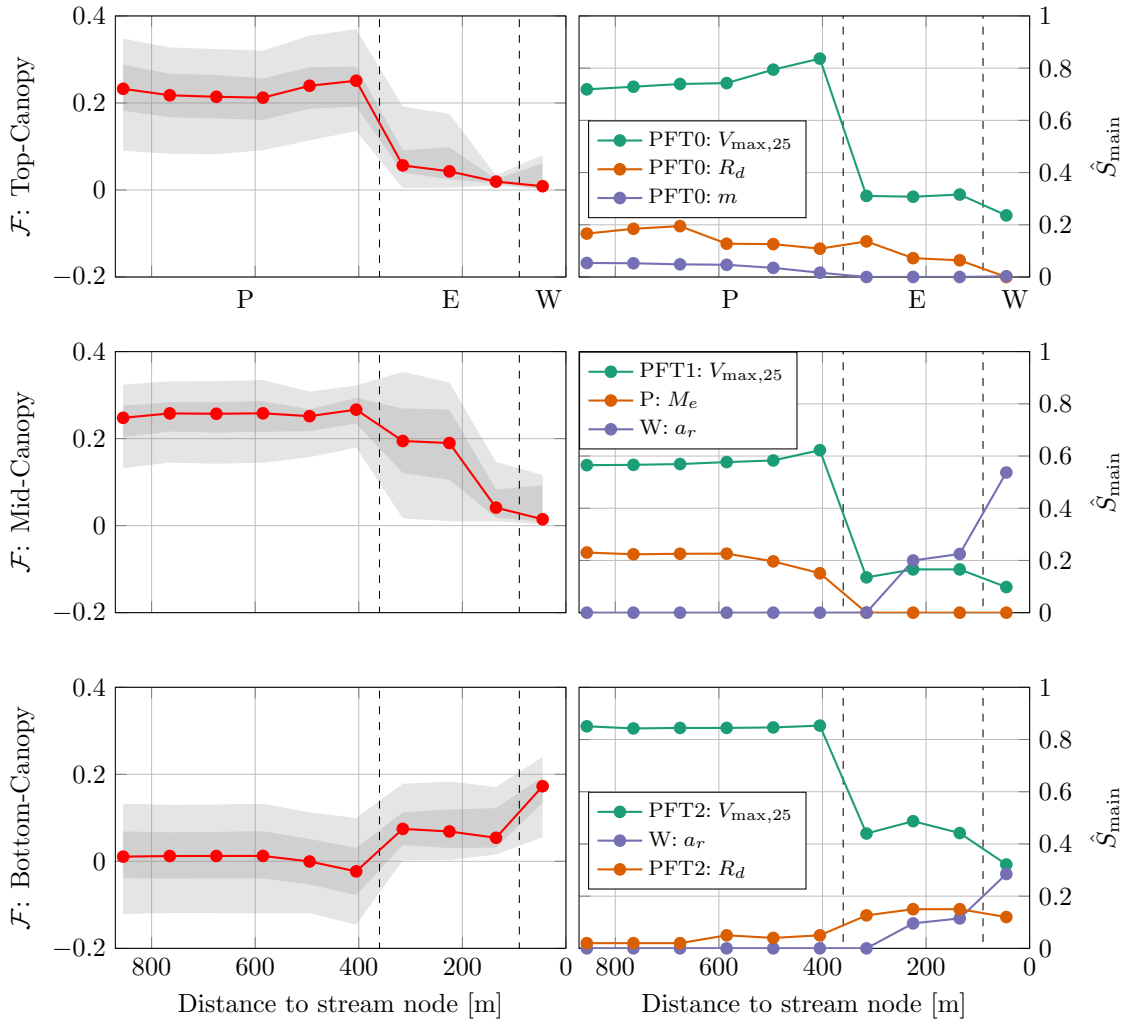

Figure 4.11: Left: Uncertain values of the performance metric $\mathcal{F}$ for each plant functional type given parametric uncertainty resulting from inference from Section 4.3.2. The gray regions represent the 5/95 and 25/75 percentile uncertainty regions, and the red line is the median. Each circle represents the center of a computational pixel. Right: Plots of main sensitivity estimates along the transect for each plant functional type. The parameters with the three largest maximum values of $\hat{S}_{\text{main}}$ are provided in each plot, showing the transition in the sensitivity of $\mathcal{F}$ to each of the three parameters along the transect. In each plot the plateau, ecotone, and waterlogged regions of the transect are delineated by the vertical dashed lines, and denoted as "P", "E", and "W", respectively.

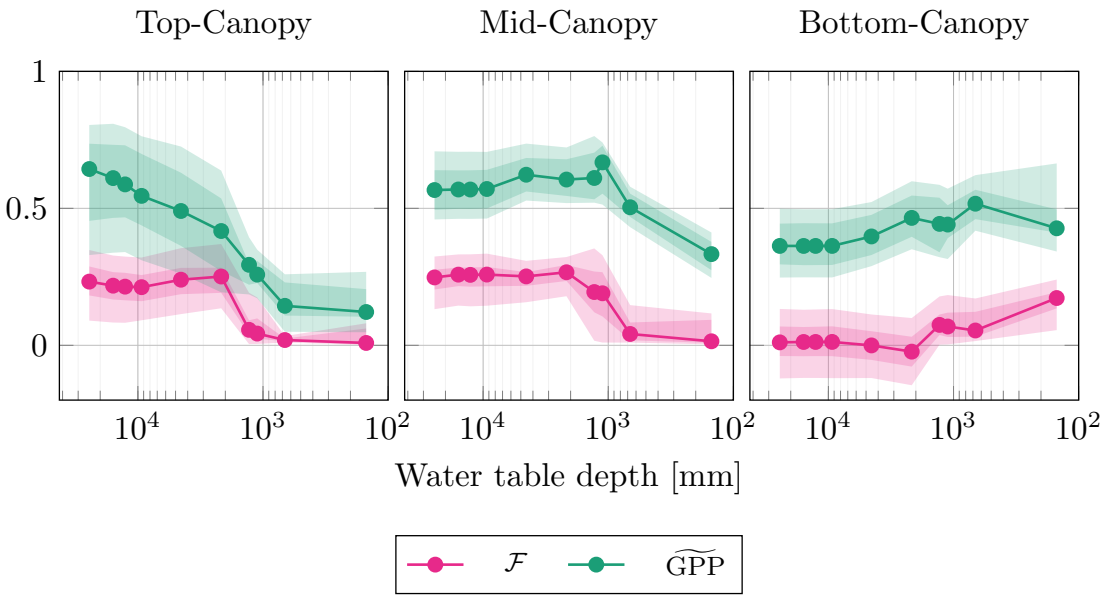

Figure 4.12: Uncertain values of the performance metric $\mathcal{F}$ and scaled GPP ($\widetilde{\text{GPP}}$) for each plant functional type with respect to the mean water table depth at each computational pixel. The representation of the performance metric is the same as in the left plots of Fig. 4.11. The water table depth decreases monotonically along the transect, so the ordering of marks is the same as those in Fig. 4.11.

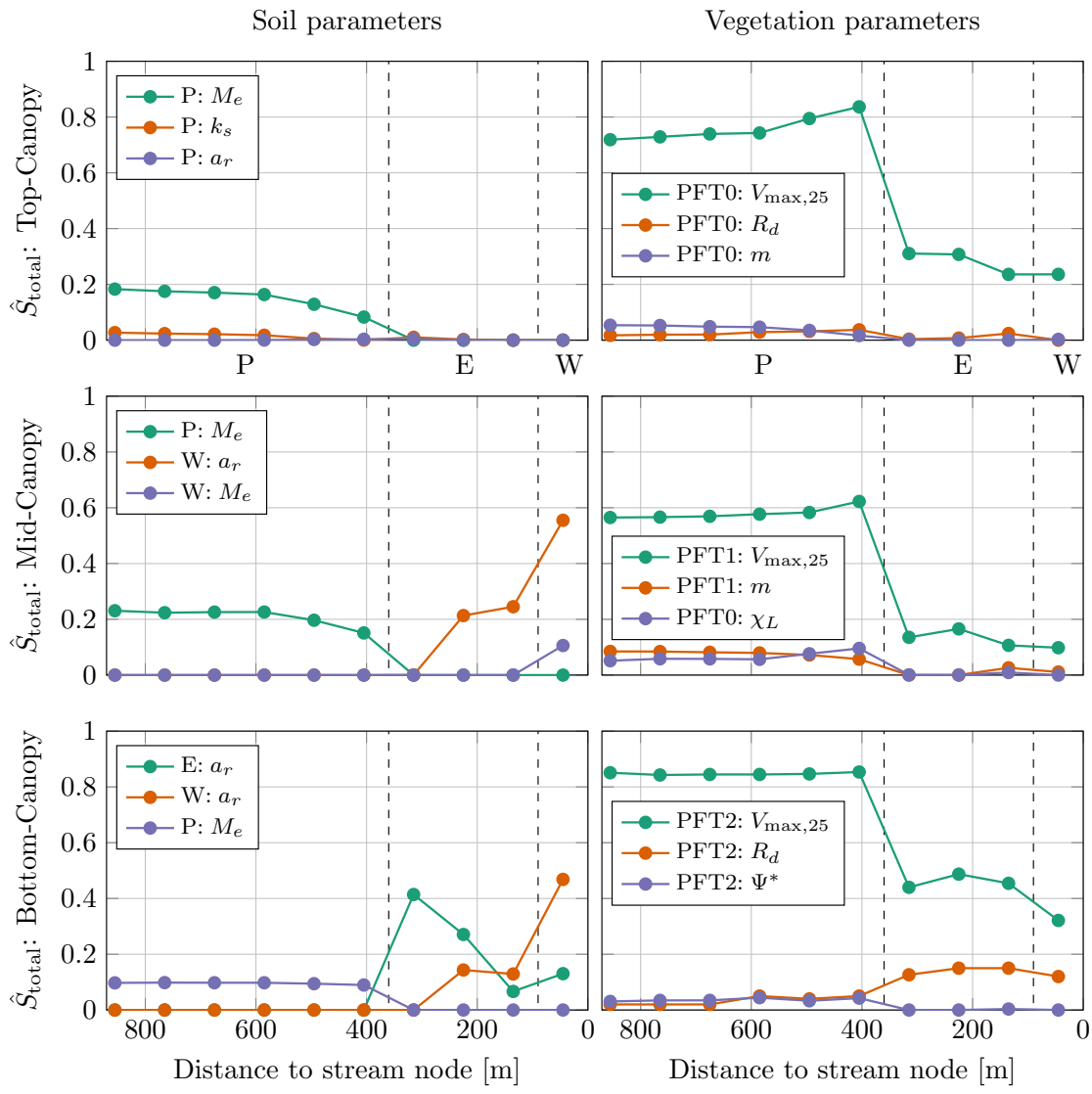

Figure 4.13: Plots of total sensitivity for each plant functional type, separated into soil (left) and vegetation (right) parameters. The plateau, ecotone, and waterlogged regions of the transect are delineated by vertical dashed lines, and denotes as "P", "E", and "W", respectively.

4.4 Discussion

4.4.1 The principle of optimality

Throughout time, one expects a complex, dynamic system like a tropical rainforest to evolve towards efficient performance. In terms of vegetation, this is related to the biochemical cost of carbon investment into the canopy, roots, and water transport tissues to aid in photosynthesis. Previous efforts to characterize this performance focused on the maximization of a plant's reproductive potential, and therefore its investment into annual biomass through GPP (Eagleson 2005). However, temporally dynamic hydrologic controls across a watershed or catchment can impact water available to plants, and there needs to be a trade off between a plant acting to increase its reproductive potential (i.e., increase of biochemical productivity through investment into the canopy) vs. carbon investment to increase its drought resilience (i.e., costs associated with water transport tissues).

Given the deep water table present at the catchment of the studied transect, and the relatively high amount of evapotranspiration occurring in the dry season (Fig. 4.9), there is sufficient evidence for deep roots within the transect (as shown in other areas of the Amazon rainforest, e.g., Davidson et al. 2011; Ivanov et al. 2012). The separation of root niches as a strategy for drought resilience in Amazonia was first implemented into modeling in Ivanov et al. (2012), and in an attempt to mimic this design, each plant functional type was assigned uncertain rooting depths of depending on their canopy location (Table 4.2). Unlike the domain in (Davidson et al. 2011; Ivanov et al. 2012), the K34 catchment transitions from a deep ($O(25)$ [m]) to shallow ($O(1)$ [m]) water table from the upslope to valley regions, exhibiting a larger variation on water table depth. One might expect to see this variation translated to hydraulic controls on vegetative traits, which precipitated the introduction of waterlogged photosynthetic restrictions through $\Psi^{w}_{H_2O}$ and $\Psi^{*}_{H_2O}$ into tRIBS-VEGGIE. Having these additional parameters enables the representation of a decrease of vegetative function in both wet and dry conditions, where both are possible within the K34 watershed and the study domain.

The investigation into the cost-benefit tradeoff displayed variation in performance of plant function with its position in the landscape, related to the water table depth and saturation in the root zone. This indicates that there is a transition in the contributions from vegetative to abiotic sources to plant performance,

controlled by landscape position. Furthermore, certain landscape niches exhibit the importance of interaction between vegetation types. These findings indicate complex spatial relationships in vegetative performance controlled by the landscape, which also influences the degree of importance of interaction between plant functional types. Additionally, the introduction of a cost-benefit analysis between biochemical production and hydraulic resilience is required to further investigate the tradeoffs of spatial and temporal variability of vegetative function in tropical landscapes.

Given the complex nature of these tradeoffs, the resulting models representing them also exhibit a high degree of complexity (Ivanov et al. 2012), which makes accounting for uncertainties in model inputs difficult. This work shows that by using a general uncertainty quantification approach, one can account for uncertainty in a large number of model parameters, allowing for parametric inference using heterogeneous data sources. As a result of this inference, one is able to train the complex hydrologic model on the ecohydrologic dynamics of the system in question, and then use the inferred parameters to investigate behavior derived from model outputs. Given sufficient process-based representation available in tRIBS-VEGGIE, this work was able to investigate the performance of three different plant functional types along a tropical transect in response to their intrinsic vegetative traits as well as the controls exerted on vegetation through interaction with hydrology.

4.4.2 Impacts of model representation and assumptions

The limits of the investigation into vegetative traits and function is limited to what can be expressed through tRIBS-VEGGIE. For example, it has been proposed that a "compensation" effect exists for root systems, where a plant will shift its water uptake to deeper roots during periods of drought (Doussan et al. 2006; Lobet et al. 2014). This possible control cannot be addressed in this study as the factor β_T controlling root water uptake cannot account for such behavior, and therefore the model lacks a mechanism or parameter to model this compensation effect. Other limitations include the inability to look at potential effects from changes in anything other than specified parameters. For instance, the response of the Amazon rainforest to drought is gaining increasing interest with numerous studies devoted

to the possible responses of the forest to prolonged periods of drought (Lewis et al. 2011; Davidson et al. 2011; Brando et al. 2010; Phillips et al. 2009; Nepstad et al. 1994; Santos et al. 2018; Brum et al. 2018; Chitra-Tarak et al. 2018; Ivanov et al. 2012). Using the UQ methodology presented here and in Dwelle et al. (2018), one could possibly investigate the ecohydrologic response to uncertain precipitation forcing by representing atmospheric forcing as a stochastic process. These types of investigation, incorporating uncertain forcings in addition to uncertain parameterizations for ecohydrology, is an area yet to be presented in the research literature.

This study has addressed interactions between plant production, subsurface moisture dynamics, and vegetative functional traits for a transect in an Amazonian rainforest. Simplifying assumptions were made for modeling, but unlike previous efforts in the research literature, great effort was undertaken to allow uncertainty in model parameterizations. A total of 38 model parameters were allowed to be uncertain, covering bulk soil properties, soil hydraulic properties, and photosynthetic and water uptake parameters for three different plant functional types across three different soil types within the transect. These uncertainties were propagated through the model to create an inexpensive computational surrogate that was then used to infer parameter values based on available observational data covering vegetation function an subsurface soil moisture dynamics in the vadose zone.

A metric for vegetation performance, striking a cost-benefit balance between vegetation production and biologic cost of water uptake was investigated using the posterior parameter values. It was found that a large number (17) of the uncertain parameters have a non-negligible impact on the variation of the performance metric. The large, impactful parameter set indicates that the trees can have a number of strategies in order to meet the observed production. This study confirms that rooting depth plays a large role in performance of vegetation for plant functional types that have very deep or very shallow roots (Fig. 4.11), as well as the soil exhibiting a control on the performance through availability of water via the water table depth, or the value of water at the xylem potential through the shape of the soil water retention curve (Fig. 4.3). These soil controls indicate that care needs to be taken when choosing soil properties as they may have an impact on vegetative function.

CHAPTER 5

Research summary and future applications

5.1 Summary of research

Hydrologic systems are complex and interactions between different processes lead to emergent phenomena. These phenomena can be small in scale, e.g., how a specific tree utilizes water near the saturated/unsaturated surface in the subsurface, or they could be very large, e.g., how an entire forest finds symbiosis between different tree types to maintain resilience during prolonged droughts. Regardless of the phenomena being studied, understanding the variability of the phenomena is informative for how the system may change in the future. This is especially important for tropical forests and the Amazon as it has experienced prolonged droughts this century and could experience many more (Pokhrel et al. 2014; Lewis et al. 2011).

This dissertation presents my work to develop the experimental approach to account for numerous uncertainties in hydrologic modeling in order to enable scientific investigation. The uncertainty quantification framework provides a holistic, robust approach to accounting for natural stochasticity in environmental systems. It provides the benefit of an interpretable, probabilistic framework on which to make inferences about the drivers of model behavior, as well as the sensitivities of the model's output to the uncertain inputs. Many approaches in hydrologic uncertainty assessment focus on either the problem of uncertainty propagation or inversion, and in the latter cases nearly all work is distributed around inferring model inputs based on a single metric, streamflow (Q) (e.g., Beven 2006; Seibert and Beven 2009). Previous efforts were often limited to this one metric due to availability of data or computational resources required for inversion. However the UQ

framework employed in this dissertation vastly reduces the computational burden on the hydrologists and allows inference based on disparate, heterogeneous data sources.

The strengths of the framework presented in Chapter 2 are as follows:

1. Provide a non-intrusive methodology for UQ. Do not need to make changes to the underlying structure of the computational model; the UQ framework treats model as a "black box."
2. Preserves much of the physical structure of model outputs. Provided that the surrogate is accurate (see, e.g., Section 3.3.1), it will conform to model outputs, and not introduce statistical noise as a way to account for variability in model outputs.
3. Provides global sensitivity information "for free" from computing the expansion coefficients. This allows easy investigation into the sensitivity of model outputs to inputs. Understanding changes with respect to changes in inputs provides valuable insight into the physics behind a process and how well the model is capturing those physics.
4. Is computationally inexpensive. The effective dimensionality reduction from using Bayesian compressive sensing to compute the surrogate coefficients allows exploration of larger uncertain parameter sets.
5. Enables large-scale inference. Due to the framework being computationally efficient, it allows inference using large observational sets.

Using the benefits above, I applied the UQ framework to a sparsely gauged Amazonian watershed with disparate data sources to perform parametric inference for uncertain soil properties in Chapter 3. The results demonstrate the flexibility of the framework for hydrologic inference in watersheds with sparse, irregular observations of varying accuracy. Significant computational savings imply that problems of greater computational complexity and dimension can be addressed using accurate, computationally inexpensive surrogates for complex hydrologic models. This will ultimately yield probabilistic representation of model behavior, robust parameter inference, and sensitivity analysis without the need for greater investment in computational resources.

Chapter 4 displayed the strength of this framework applied to the assessing the possible distribution of vegetation traits for efficient and evolutionary-sound water uptake strategies. A set of 38 parameters were treated as uncertain across

multiple soil types and plant functional types and their values were inferred using observational data representing vegetation production and below-surface moisture dynamics. It was determined that parametric inference was able to get disperse training simulations to follow the behavior of the ecohydrologic processes in the study domain. Furthermore, these trained parameters were then used to investigate the cost-benefit relationship between plant production and carbon investment, finding a difference of preferred strategies as dependent upon functional type and location in the study domain.

5.1.1 Assumptions and limitations of the research

Any modeling work has its limitations or assumptions associated with it. Critical assumptions affecting the framework and findings are given here.

- Extrapolation; this was not done in the presented work, but is still worth noting. As part of the UQ framework, distributions for uncertain parameters are defined and these uncertainties are propagated through the model. Inference can then be performed on these parameters. One should constrain the posterior distributions from inference to be within the prior if one wishes to then propagate the posterior back through the model, as was done in Chapters 3 and 4. Because the uncertainty propagation is performed using the surrogate model that was trained using parameter samples from the prior distribution, attempting to propagate values outside of the prior distribution could likely lead to erroneous or non-physical results.
- Expression of parameter distributions. Any distribution that follows the Askey scheme (Xiu and Karniadakis 2002) could be chosen as the prior distribution for an uncertain parameter, but it is much more common to choose more "standard" distributions, like those in Table 2.1. Furthermore, many software packages that can construct PC surrogates only account for a few polynomial schemes, and may not allow mixing of polynomial types as input, so one needs to be aware of these limitations during experimental design.
- Independence of uncertain parameters. The framework assumes that parameters are independent, and this will not be the case when large number of uncertain parameters are being chosen.
  - If one has parameters that are not independent, and in fact have some

relationship or constraint, methods exist to transform the two or more variables into uncertain variables for use in the UQ framework (Sargsyan et al. 2015; Torre et al. 2017). This approach seems valuable to the hydrologic community.

- Process-based model limitations. Every hydrologic model will have its own sets of assumptions and limitations. For discussions of the limitations of the model used in this dissertation, tRIBS-VEGGIE, see (Ivanov 2006; Ivanov et al. 2008a; Ivanov et al. 2008b; Kim et al. 2016). It is important to keep these in mind when performing UQ because the adage "garbage in, garbage out" applies to surrogate construction framework as well. Just as one needs verify the performance of the surrogate against the model (e.g., Fig. 4.7), one needs to be confident in the behavior of the model generating the training set in order to perform UQ. The quantification of model structural error is an active area of research in UQ (Sargsyan et al. 2015; Sargsyan et al. 2018) and its application to hydrologic models is a future area of research.

5.2 Future studies

5.2.1 Forest assessment studies with uncertainty under climate change

A goal of future studies is investigation into the uncertainty of forest and hydrologic response to uncertain climatalogic signals for an Amazonian watershed system. In addition to the parametric uncertainties addressed in this research, sources of uncertainties for such an assessment include:

- Uncertainty in climate projections. Climate projections are often derived from an ensemble of global circulation models (GCMs), which exhibit a large degree of uncertainty, both spatially and temporally, for values for precipitation. Ways to account for the stochastic nature of forcing are needed, i.e., ways to systematically treat the stochastic input in a formal UQ framework.
- Extreme precipitation events. GCMs do not capture extreme events well, such as prolonged periods of drought or excess precipitation. These events have the capacity to affect forest structure.
- Dynamic vegetation in response to extremes. During these extreme events,

the possibility of plant die-off needs to be accounted for to capture the feedback system, i.e., long-term changes of water, carbon, and energy balances after an extreme event.

5.2.2 Accounting for input dependencies in UQ framework

Dependence among parameterizations is common in hydrologic models, but UQ frameworks treat inputs as independent. Therefore, the ability to model the dependence structure of multivariate inputs is needed for a proper accounting of both prior and posterior dependencies of input parameters. Recent work modeling dependencies through copulas (Torre et al. 2017) provides one potential avenue to addressing this issue. Accounting for this dependence can help address questions of: how can correlations between uncertain climate forcings be accounted for in a UQ framework (Section 5.2.1)? Does accounting for the dependence structure reveal emergent relationships across scales of hydrologic processes? Are there significant impacts in modeled outcomes when accounting for this structure vs. assuming independence?

5.2.3 Uncertainty quantification for urban flooding

This section briefly introduces the problem of uncertainty quantification for flooding and presents a small case study that has been used as a proof-of-concept work for applying the model tRIBS-OFM (Kim et al. 2012) to flooding in urban areas.

Despite substantial advances in civil engineering in the modern era, urbanized areas are vulnerable to extreme flooding, since urban drainage and stormwater management infrastructure are generally designed for flows of 10-20 year return periods (Hoang and Fenner 2016). Flooding in densely populated areas has remained the costliest natural hazard of all weather-related events in terms of fatalities and material costs. For example, in the U.S., while there has been a gradual negative trend in annual flood-related fatalities, their 30-year (1985–2014) average remains high: 81-85 people per year. Past events are clear harbingers of what is yet to come: urban areas already occupy 2% of the Earth's land surface concentrating over 45% of the world's population (Cohen 2003) and recent estimates indicate that the number of people residing in the flow path of high-risk floods will double — from one to two billion, within two generations (De Groeve et al. 2014). Furthermore, global

extreme floods are clearly on the rise (Hirabayashi et al. 2013) and continued climate change is likely to promote the intensification of the land-surface hydrologic cycle and flooding. Consequently, problems associated with flooding have the potential to rapidly proliferate in the very near future, affecting populated environments of all types, from small rural dwellings, to megacities. The emerging need is both to understand how heterogeneity of urban environments impact extreme floods and engineer a comprehensive modeling capacity relevant to decision making in the critical times of flooding.

Understanding and predicting floods across a range of space-time scales at the relevant level of detail and with uncertainty assessment remains a poorly addressed challenge. Large-scale operational modeling efforts, such as NASA's Global Land Data Assimilation System, are capable of yielding modeling results on runoff in near-real time (Rodell et al. 2004) at 10^4-10^5 [m] spatial resolution. The NOAA National Weather Service recently launched the National Water Model to provide streamflow for 2.7 million river reaches in the U.S. and other hydrologic variables on $10^3$3 [m] and 250 [m] grids. These systems however inherently rely on over-simplified physics of hydrologic and hydrodynamic processes and their coarse resolutions entirely neglect intricate drainage routes and lump the structure of urban landscapes, preventing their direct use for predicting flooding dynamics at local scales (10^0 -10^2 [m]), which are most relevant to the end-users: communities, water utilities, and humans who are in the pathways of flood waves.

Furthermore, uncertainties remain unaddressed in a variety of types of input information, such as precipitation, soil moisture, hydraulic characteristics of water pathways below and above surface. Academic community-enabled capabilities to simulate relevant processes across a range of temporal and spatial scales have been steadily increasing and hold a lot of promise (Kollet and Maxwell 2006; Ivanov et al. 2008b; Kim and Ivanov 2015; Kim et al. 2012; Mirus and Loague 2013; Maxwell and Condon 2016; Fatichi et al. 2016; Ivanov et al. 2004; Bates 2012). The development of the coupled hydrologic-hydraulic models targeted the physical consistency of transferring information from larger-scale drivers to local-scale dynamics with a smaller number of simplifying approximations, lately reaching the scales that are relevant for flood-forecasting.

The science of flood modeling has been steadily changing from a data-scarce to data-rich environment because of the growing availability of geospatial, remote

sensing datasets, as well as real-time sensor data for hydrologic systems (Lincoln 2007; Baldassarre and Montanari 2009; Hill et al. 2014). For example, due to increasingly available LiDAR data on urban geometry, hydrodynamic modeling is now possible at sub-meter resolutions and demonstrates excellent skill (Schubert and Sanders 2012; Bates 2012). Finer spatial resolutions of relevant data sets on topography, building layout, and civil infrastructure describe heterogeneity of the real world and have the potential to advance knowledge in urban hydrology and hydraulics, but also come at an extreme computational cost. Specifically, while the temporal scales of flood events are fairly small (10^0 – 10^1 [day]), the level of spatial detail that can take full advantage of geospatial data (10^0 – 10^1 [m]) and is most useful to end-users creates extreme challenges of feasibility for any real-time applications, especially when input uncertainty is explicitly accounted for. For example, a general circulation model (GCM) used for global-scale climate modeling at a 0.5-degree resolution has approximately $5{\times}10^6$ grid cells in three dimensions; even on a high-end computer at a national center, a heavily parallelized simulation takes approximately 3.5 hrs for a 1-year run. A relatively small, city-scale flooding problem can have the same order of computational nodes (e.g., variable mesh spacing for a flooding domain for the city of Nashville, TN, drainage area < 1,000 [km^2]). The physical equations describing hydrologic and hydrodynamic processes have approximately the same degree of non-linearity, as compared to a GCM.

However, the use of the UQ framework can help to reduce the computational burden of this model, but further advancements will be required. Consider the problem presented in Chapter 4, where 38 parameters were treated as uncertain and 1,000 training simulations of tRIBS-VEGGIE were performed. The domain size and complexity of resolving flow pathways in urban environments would render simulation times of tRIBS-OFM to be orders of magnitude greater than those witnessed running tRIBS-VEGGIE in the work of this dissertation. However, recent advancement in UQ such as multi-fidelity modeling (Ng and Eldred 2012; West and Gumbert 2017), where one uses variable resolutions of the simulation domain to approximate and resolve discrepancies of the model outputs at various levels of complexity, may provide an avenue to tackle uncertainty quantification for the urban flooding problem.

5.2.4 Small-scale case study in Nashville, TN

Here I briefly present a case study for the 2010 "1000-year flood" event in Nashville, TN (Moore et al. 2011) using the framework presented in Chapter 2, using solely uncertainty propagation techniques and constructing surrogates using Bayesian compressive sensing. The storm inundated large portions of downtown Nashville, including the industrial and downtown district, causing major disruptions to industry and tourism. Unlike natural hydrologic systems, urban settings require the modeler to also resolve uncertainties in the built environment, with the largest uncertainty being the placement and "porosity" of structures (Guinot 2012; Kim et al. 2015).

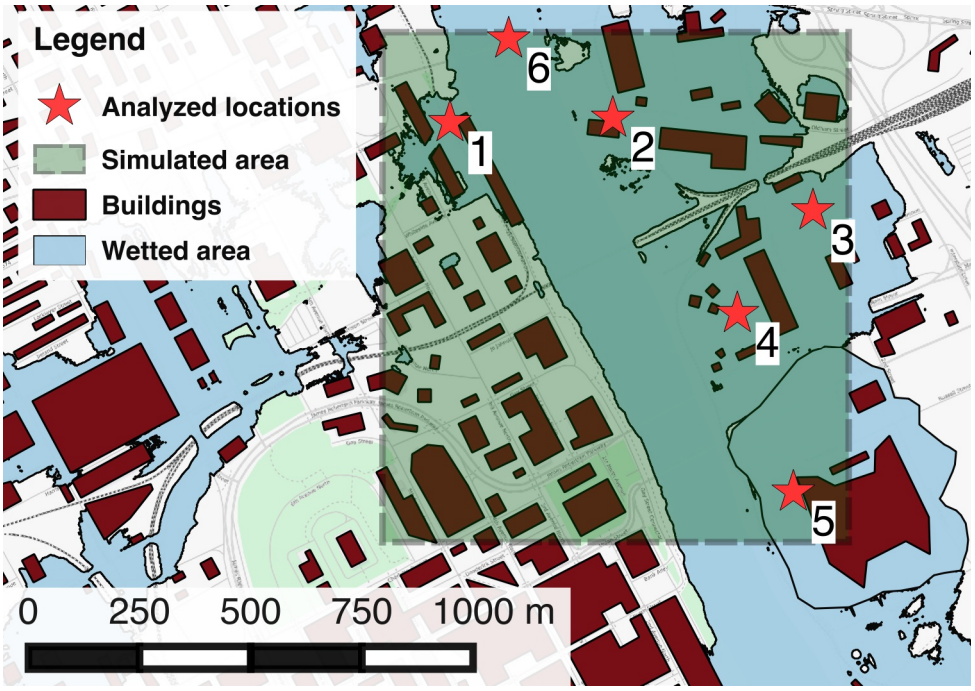

Figure 5.1: Small study area (1.2 [km^2]) used for Nashville case study. This location is located at the heart of downtown next to the main tourist avenue and the Tennessee Titans football stadium. The river flows from the bottom to the top of the figure along the channel of the Cumberland River. Six locations were selected for analysis, representing different levels of inundation throughout the storm. The blue outline represents the inundation map generated by the Federal Emergency Management Agency.

The parameters treated as uncertain in the domain are given in Table 5.1. Manning's roughness coefficient parameters n were treated as uncertain for both the land and river channel. Four different land use types: prairie, grass, undeveloped,

and developed were present within the simulation domain, and the saturated conductivity of the soil (K_{sat})in each were treated as uncertain to account for potential drainage before the flood wave entered the simulation domain. The saturated soil moisture content θ_{sat} was also treated as uncertain for the land use types, with the exception of prairie. Finally, the inflow hydrograph was obtained from a USGS gauge just upstream of the simulation domain, containing measurements of streamflow [m^3s^{-1}] during the storm. These data were taken and used to generate a Gaussian process in order to treat the inflow data as a stochastic time series of inflow (Figure 5.2) following Section 2.4.2. The resulting process was used to generate six uncertain parameters $\xi_{1,\dots,6}$ using Karhunen-Loève decomposition of the stochastic process (Section 2.4.2), the truncation at six parameters was performed to achieve a 95% explanation of the variability in the stochastic process (Eq. (2.32)).

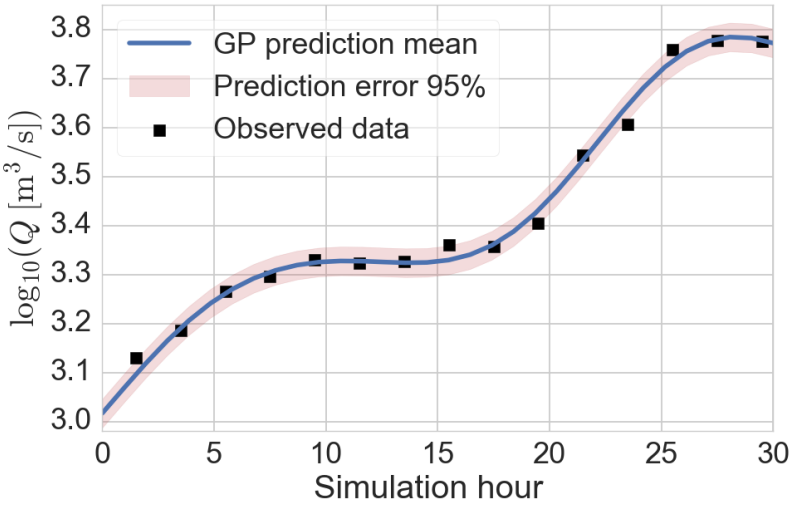

Figure 5.2: Gaussian process of $\log_{10}(Q)$ used for uncertain inflow boundary condition. The logarithm of streamflow was used instead of the non-transformed value as it provided a smoother function to approximate.

A total of 750 training simulations were performed for the 15 uncertain parameters, and QoIs were given as the time series of water depth at each of the analyzed locations in Fig. 5.1 as well as an aggregated value of inundated area (Figure 5.3).

Sensitivity analysis of the constructed surrogates showed that parameters relating to the inflow boundary condition were the most impactful, followed by the Manning's roughness coefficient for the channel (n_{channel}). The preliminary results display the ability of used stochastic fields as uncertain model inputs, and this is a

Table 5.1: Uncertain parameterizations for the Nashville case study.

Parameter	Distribution
n_{land}	$U[0.02, 0.07]$
n_{channel}	$U[0.01, 0.03]$
$k_{\text{sat, prairie}}$	$U[30, 45]$
$k_{\text{sat, grass}}$	$U[15, 30]$
$\theta_{\text{sat, grass}}$	$U[0.45, 0.5]$
$k_{\text{sat, undeveloped}}$	$U[3, 15]$
$\theta_{\text{sat, undeveloped}}$	$U[0.4, 0.55]$
$k_{\text{sat, developed}}$	$U[0, 3]$
$\theta_{\text{sat, developed}}$	$U[0.4, 0.55]$
$\xi_{1,\ldots,6}$	$N(0, 1)$

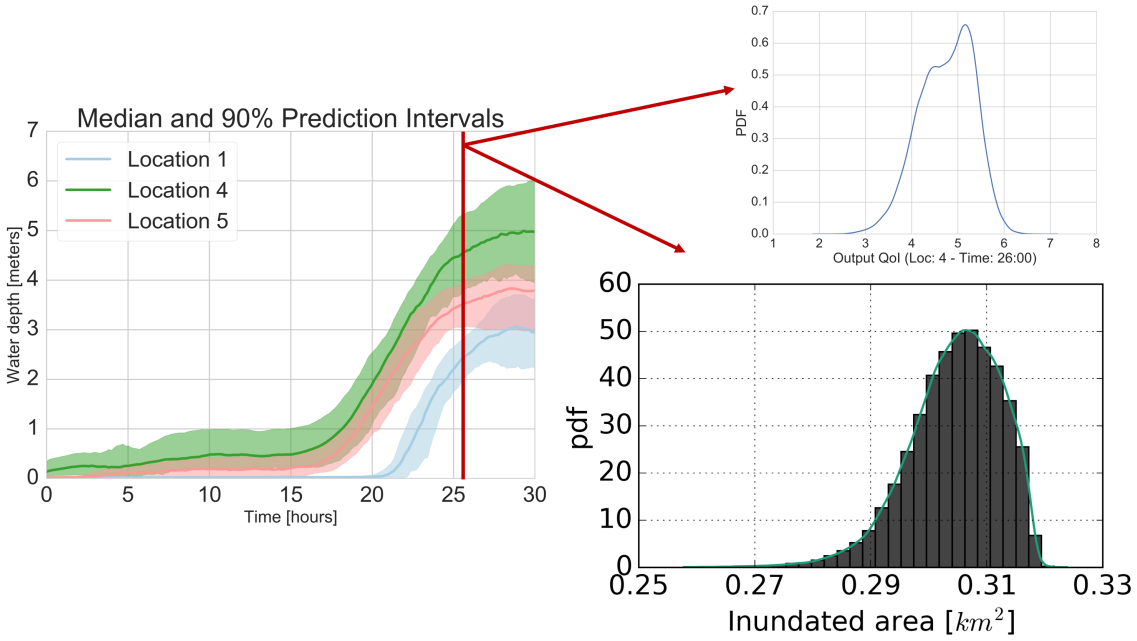

Figure 5.3: Time series (15 min.) water depth over simulation time at three locations from Figure 5.1

further area of research to be applied to not just inflow hydrographs, but also soil moisture and water table fields, as well as atmospheric forcing conditions.

Appendices

APPENDIX A

Monte Carlo: why it fails for complex problems

Monte Carlo analysis is very convenient for numerical integration or to generate samples from a distribution. The accuracy of the method is independent of the dimensionality of the problem, but it also scales poorly for complex problems. Imagine using Monte Carlo to estimate the area of a high-dimensional function by taking random points and tabulating how many fall above or below the curve. The error in the integration estimate decreases proportional to $1/\sqrt{n}$, where n is the number of samples used for approximation. This error is therefore independent of dimension and a robust methodology which has been used in a large range of problems (e.g., Gilks et al. 1995; Rajabi and Ataie-Ashtiani 2014; Vrugt et al. 2008; Mishra et al. 2012; Aronica et al. 2012).

However as the dimensionality d of the problem grows, it becomes increasingly difficult to get informative samples. As an illustrative example, consider trying to estimate the volume of a unit sphere in d dimensions of radius 1 by randomly sampling within the d-dimensional unit cube in $[-1, 1]$. Assume that d is even so $d = 2a$, then the volume $V(\cdot)$ of the box and ball will be:

$$\begin{aligned} V(\text{cube}) &= 2^{2a} = 4^a \\ V(\text{sphere}) &= \frac{\pi^{2a/2}}{(2a/2)!} = \frac{\pi^a}{a!}, \end{aligned} \tag{A.1}$$

so the ratio of the volume of the sphere in the cube is $\pi^a/(4^a a!)$. How this scales with dimension can be seen in A.1. Sampling the volume of the sphere inside the cube, one would estimate that the volume of the sphere is zero unless a large number

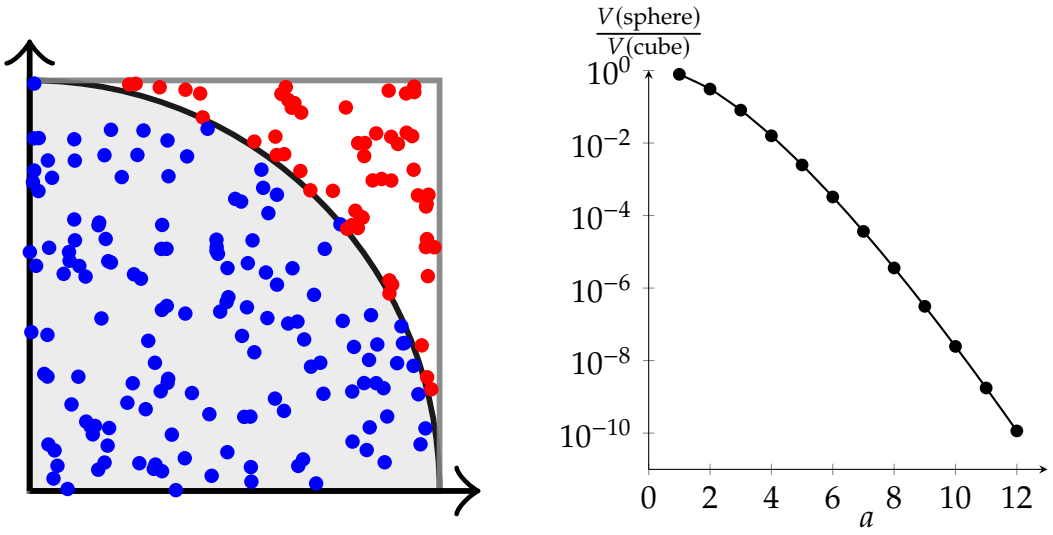

Figure A.1: Left: Figure of Monte Carlo sampling ($n = 200$) of volume of a unit sphere in a unit cube with dimension $d = 2$ (only $1/4$ showed due to symmetry), coinciding to $a = 1$ in (A.1). The dots are colored blue if they fall within the sphere and red otherwise. Right: Scaling of the ratio of the volume of a unit sphere to the volume of the unit cube in $d = 2a$ dimensions.

of samples are taken. Whether or not this is an issue depends on the question being asked and the problem being solved. An estimate of volume zero provides an *absolute* error that is very small, since the volume of the sphere is relatively small compared to the sampling space.

When integrating with Monte Carlo methods, a small *relative* error is usually desired, and one therefore wants an a large proportion of sampling points to be in the region of interest. The desire is to have an *efficient* method, one where a large proportion of samples are useful. This is an area of continuous development, and includes importance sampling and transport maps (Siegmund 1976; Glynn and Iglehart 1989; Parno and Marzouk 2018; Spantini et al. 2017). The samples in this research are being generated from numerical simulations of a complex hydrologic model, and therefore each rejected could be considered as "wasted" computational expense, as more simulations are required to get a desired number useful samples. For this reason, surrogate models (Section 2.3) are deployed, meaning that the computational cost of getting a useful sample is significantly reduced.

APPENDIX B

Construction of PC surrogate using multi-index notation.

In order to construct the polynomial surrogate it is often helpful to use simplified multi-index notation in place of the explicit terms being added to the index. To aid interpretation of the notation, this section explicitly constructs a polynomial chaos surrogate using the multi index notation, and then shows how terms are gathered for sensitivity analysis as mentioned in Section 2.3.2.

Surrogate construction

As before, consider the multi-index $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$, a tuple of integers up to m, where m is the number of uncertain variables considered: $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)$. These are used to construct the multivariate polynomial used in the expansion in Equation (2.13), constructed as in Equation (2.8) and repeated here:

$$\Psi_{\boldsymbol{\alpha}} = \psi_{\alpha_1}(\xi_1)\psi_{\alpha_2}(\xi_2)\cdots\psi_{\alpha_m}(\xi_m).$$

If one assumes that the random variables in $\boldsymbol{\xi}$ are Gaussian, then Hermite polynomials would apply (see Table 2.1). The recurrence relation for Hermite polynomials is:[1]

$$\psi_{n+1}(\xi) = \xi\psi_n(\xi) - n\psi_{n-1}(\xi),$$

[1]This is the *probabilists* polynomial, or Hermite polynomial of the *first kind*. There also exists a *physicists* version, or Hermite polynomial of the *second kind*.

which gives the first five terms as:

$$\begin{aligned}
\psi_0(\xi) &= 1 \\
\psi_1(\xi) &= \xi \\
\psi_2(\xi) &= \xi^2 - 1 \\
\psi_3(\xi) &= \xi^3 - 3\xi \\
\psi_4(\xi) &= \xi^4 - 6\xi^2 + 3
\end{aligned}$$

A plot of Hermite and Legendre polynomials are shown in Figure 2.3.

To construct the polynomial surrogate in Equation (2.13), one first selects the graded lexicographic ordering of the multi-index $\boldsymbol{\alpha}$ to get the unique combinations of polynomials and variables in $\boldsymbol{\xi}$ under a certain total degree. For this ordering, one first sorts by the total degree (sum of polynomial order), then sorts lexicographically, i.e., $\xi_1 \prec \xi_2 \prec \xi_3$, where $a \prec b$ is read as: *a precedes b in order*. As an example, consider a polynomial that is indexed by its arguments with three variables such that $P_{\alpha_1,\alpha_2,\alpha_3}(\xi_1, \xi_2, \xi_3) = \xi_1^{\alpha_1}\xi_2^{\alpha_2}\xi_3^{\alpha_3}$, then its graded lexicographic ordering would be:

$$1 \prec \xi_1 \prec \xi_2 \prec \xi_3 \prec \xi_1^2 \prec \xi_1\xi_2 \prec \xi_1\xi_3 \prec \xi_2^2 \prec \xi_2\xi_3 \prec \xi_3^2 \prec \xi_1^3 \prec \xi_2^3 \prec \dots.$$

The values of the members of the multi-index $(\alpha_1, \alpha_2, \alpha_3)$ are the order of the polynomial that is a function of (ξ_1, ξ_2, ξ_3) Writing in terms of the Hermite polynomials above, this becomes the following, where the underbraces are the values of the multi-index tuple $(\alpha_1, \alpha_2, \alpha_3)$.

$$\begin{aligned}
&\underbrace{1}_{(0,0,0)} \prec \underbrace{\psi_1(\xi_1)}_{(1,0,0)} \prec \underbrace{\psi_1(\xi_2)}_{(0,1,0)} \prec \underbrace{\psi_1(\xi_3)}_{(0,0,1)} \prec \underbrace{\psi_2(\xi_1)}_{(2,0,0)} \prec \underbrace{\psi_1(\xi_1)\psi_1(\xi_2)}_{(1,1,0)} \prec \underbrace{\psi_1(\xi_1)\psi_1(\xi_3)}_{(1,0,1)} \\
&\quad \prec \underbrace{\psi_2(\xi_2)}_{(0,2,0)} \prec \underbrace{\psi_1(\xi_2)\psi_1(\xi_3)}_{(0,1,1)} \prec \underbrace{\psi_2(\xi_3)}_{(0,0,2)} \prec \underbrace{\psi_3(\xi_1)}_{(3,0,0)} \prec \underbrace{\psi_3(\xi_2)}_{(0,3,0)} \prec \dots
\end{aligned} \tag{B.1}$$

When truncating the expansion in Equation (2.13), polynomials that have at most total order p are kept.[2] The total order is calculated from the sum of the

[2]This is the *total degree* truncation scheme. Other truncation schemes (see, e.g., Sargsyan et al. 2014) can be chosen as a preprocessing step to select a different polynomial basis.

polynomial orders, which can be easily calculated by the members of the multi-index, and is denoted

$$p = \|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^{M} |\alpha_i|.$$

This means that both the terms $\psi_3(\xi_1)$ and $\psi_2(\xi_1)\psi_1(\xi_2)$ have total polynomial order 3 as their multi-indices are $(3, 0, 0)$ and $(2, 1, 0)$, respectively.

To construct a PC surrogate, take an example with three uncertain variables and polynomial order $p = 2$. One can gather the terms and use the ordering in Equation (B.1), adding the respective coefficients. The truncated surrogate of these collected terms would be:

$$\begin{aligned}\mathcal{M}^{\mathrm{PC}}(\xi_1, \xi_2, \xi_3) =& c_0 + c_1\psi_1(\xi_1) + c_2\psi_1(\xi_2) + c_3\psi_1(\xi_3) + c_4\psi_2(\xi_1) + \\ & c_5\psi_1(\xi_1)\psi_1(\xi_2) + c_6\psi_1(\xi_1)\psi_1(\xi_3) + c_7\psi_2(\xi_2) + \\ & c_8\psi_1(\xi_2)\psi_1(\xi_3) + c_9\psi_2(\xi_3)\end{aligned} \tag{B.2}$$

From Equation (2.14), the expected (3+2)!/3!/2! = 10 terms appear in Equation (B.2). Using Equation (B.2) and Figure 2.3, one gets a clear visual sense that, once the PCE coefficients are determined, $\mathcal{M}^{\mathrm{PC}}$ is simply a linear combination of polynomials in order to determine the value for the QoI. The graded lexicographic ordering mentioned previously is a way to assign and manage accounting of the coefficients in Equation (B.2), and is the standard used in most software that constructs PCE surrogates, though any ordering would work as long as it is used consistently in calculations.

Given the PCE surrogate in (B.2), one can calculate the first two moments as in Equations (2.15) and (2.16) as:

$$\mu^{PC} = c_0 \tag{B.3}$$

$$(\sigma^2)^{PC} = \sum_{j=1}^{9} c_j^2 \left\langle \Psi_j^2 \right\rangle ., \tag{B.4}$$

where $\left\langle \Psi_j^2 \right\rangle$ is the inner product or squared norm of the basis function in Equation (2.8), with the subscript j on the polynomial basis represents the multi-index associated with the coefficient term c_j. The inner product is calculated through

numerical quadrature, and is given by:

$$\langle u, v \rangle = \int_{\mathcal{D}_x} u(x)v(x)\pi(x)\,\mathrm{d}x, \tag{B.5}$$

where $\mathcal{D}_x$ is the domain of quadrature which is defined by the chosen polynomial family, and $\pi(x)$ is the weighting function. This can also be called the squared norm since the norm is defined by:

$$\|u\| = \langle u, u \rangle^{1/2}. \tag{B.6}$$

Sensitivity analysis with calculated coefficients

As mentioned in Section 2.3.2, using PCE surrogates provides global sensitivity analysis "for free" from the values of the coefficients $\boldsymbol{c} = (c_1, c_2, \ldots, c_{P+1})$ (recall that c_0 gives the estimated mean of the surrogate). Continuing the example from the previous section, the main and total sensitivities will be calculated from $\mathcal{M}^{\mathrm{PC}}$. Using Equations (B.2) and (2.16), the variance relation can be written as:

$$\begin{aligned}\mathrm{Var}\left(\mathcal{M}^{\mathrm{PC}}\right) =& c_1^2\left\langle\psi_1^2\right\rangle + c_2^2\left\langle\psi_1^2\right\rangle + c_3^2\left\langle\psi_1^2\right\rangle + c_4^2\left\langle\psi_2^2\right\rangle + c_5^2\left\langle\psi_1^2\right\rangle\left\langle\psi_1^2\right\rangle + \\ & c_6^2\left\langle\psi_1^2\right\rangle\left\langle\psi_1^2\right\rangle + c_7^2\left\langle\psi_2^2\right\rangle + c_8^2\left\langle\psi_1^2\right\rangle\left\langle\psi_1^2\right\rangle + c_9^2\left\langle\psi_2^2\right\rangle\end{aligned} \tag{B.7}$$

The Sobol sensitivity indices can be calculated using Equation (B.7), which is also the denominator in Equations (2.20) and (2.21). The numerators of each equation are given below, with the main effect contribution to the indices (Equation (2.20)) from each uncertain parameter being:

$$\hat{S}_i^{\mathrm{main}} = \begin{cases} \xi_1: & c_1^2\left\langle\psi_1^2\right\rangle + c_4^2\left\langle\psi_2^2\right\rangle \\ \xi_2: & c_2^2\left\langle\psi_1^2\right\rangle + c_7^2\left\langle\psi_2^2\right\rangle \\ \xi_3: & c_3^2\left\langle\psi_1^2\right\rangle + c_9^2\left\langle\psi_2^2\right\rangle \end{cases}, \tag{B.8}$$

and the corresponding contributions from each variable towards the total effect

indices (Equation (2.21)) are:

$$\hat{S}_i^{\text{total}} = \begin{cases} \xi_1: & c_1^2 \left\langle \psi_1^2 \right\rangle + c_4^2 \left\langle \psi_2^2 \right\rangle + c_5^2 \left\langle \psi_1^2 \right\rangle \left\langle \psi_1^2 \right\rangle + c_6^2 \left\langle \psi_1^2 \right\rangle \left\langle \psi_1^2 \right\rangle \\ \xi_2: & c_2^2 \left\langle \psi_1^2 \right\rangle + c_7^2 \left\langle \psi_2^2 \right\rangle + c_5^2 \left\langle \psi_1^2 \right\rangle \left\langle \psi_1^2 \right\rangle + c_8^2 \left\langle \psi_1^2 \right\rangle \left\langle \psi_1^2 \right\rangle \\ \xi_3: & c_3^2 \left\langle \psi_1^2 \right\rangle + c_9^2 \left\langle \psi_2^2 \right\rangle + c_6^2 \left\langle \psi_1^2 \right\rangle \left\langle \psi_1^2 \right\rangle + c_8^2 \left\langle \psi_1^2 \right\rangle \left\langle \psi_1^2 \right\rangle \end{cases} . \tag{B.9}$$

Bibliography

Abbaspour, Karim C., C. A. Johnson, and M. Th Van Genuchten (2004). "Estimating Uncertain Flow and Transport Parameters Using a Sequential Uncertainty Fitting Procedure". In: *Vadose Zone Journal* 3.4, pp. 1340–1352.

Abe, Shigeo (2010). *Support Vector Machines for Pattern Classification*. Advances in Pattern Recognition. Springer, London. ISBN: 978-1-84996-097-7 978-1-84996-098-4. DOI: 10.1007/978-1-84996-098-4_1.

Abril, Gwenaël et al. (Jan. 16, 2014). "Amazon River Carbon Dioxide Outgassing Fuelled by Wetlands". In: *Nature* 505.7483, pp. 395–398. DOI: 10.1038/nature12797.

Ajami, Hoori et al. (Mar. 1, 2014). "Assessing the Impact of Model Spin-up on Surface Water-Groundwater Interactions Using an Integrated Hydrologic Model". In: *Water Resources Research* 50.3, pp. 2636–2656. DOI: 10.1002/2013WR014258.

Alemohammad, S. H. et al. (Nov. 18, 2016). "Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A Statistically-Based Estimate of Global Surface Turbulent Fluxes Using Solar-Induced Fluorescence". In: *Biogeosciences Discuss.* 2016, pp. 1–36. DOI: 10.5194/bg-2016-495.

Allan, David, Donna Erickson, and John Fay (Feb. 1, 1997). "The Influence of Catchment Land Use on Stream Integrity across Multiple Spatial Scales". In: *Freshwater Biology* 37.1, pp. 149–161. DOI: 10.1046/j.1365-2427.1997.d01-546.x.

Anderegg, William R. L. et al. (May 3, 2016). "Meta-Analysis Reveals That Hydraulic Traits Explain Cross-Species Patterns of Drought-Induced Tree Mortality across the Globe". In: *Proceedings of the National Academy of Sciences* 113.18, pp. 5024–5029. DOI: 10.1073/pnas.1525678113. pmid: 27091965.

Araújo, A. C. et al. (2002). "Comparative Measurements of Carbon Dioxide Fluxes from Two Nearby Towers in a Central Amazonian Rainforest: The Manaus LBA Site". In: *Journal of Geophysical Research: Atmospheres* 107.D20, LBA–58.

Arnst, Maarten and Jean-Philippe Ponthot (2014). "An Overview of Nonintrusive Characterization, Propagation, and Sensitivity Analysis of Uncertainties in Computational Mechanics". In: *International Journal for Uncertainty Quantification* 4.5.

Aronica, G. T. et al. (Dec. 30, 2012). "Probabilistic Evaluation of Flood Hazard in Urban Areas Using Monte Carlo Simulation". In: *Hydrological Processes* 26.26, pp. 3962–3972. DOI: 10.1002/hyp.8370.

Ascough Ii, J. C. et al. (2008). "Future Research Challenges for Incorporation of Uncertainty in Environmental and Ecological Decision-Making". In: *Ecological modelling* 219.3-4, pp. 383–399.

Babacan, S.D., R. Molina, and A.K. Katsaggelos (Jan. 2010). "Bayesian Compressive Sensing Using Laplace Priors". In: *IEEE Transactions on Image Processing* 19.1, pp. 53–63. DOI: 10.1109/TIP.2009.2032894.

Baker, I. T. et al. (Mar. 2008). "Seasonal Drought Stress in the Amazon: Reconciling Models and Observations: TAPAJOS Km 83 NEE ANNUAL CYCLE". In: *Journal of Geophysical Research: Biogeosciences* 113.G1, n/a–n/a. DOI: 10.1029/2007JG000644.

Baldassarre, GDi and A. Montanari (2009). "Uncertainty in River Discharge Observations: A Quantitative Analysis". In: *Hydrology and Earth System Sciences* 13.6, pp. 913–921.

Bates, Paul D. (July 30, 2012). "Integrating Remote Sensing Data with Flood Inundation Models: How Far Have We Got?" In: *Hydrological Processes* 26.16, pp. 2515–2521. DOI: 10.1002/hyp.9374.

Bear, J. (1979). *Hydraulics of Groundwater*. McGraw-Hill International Book Co.

Beighley, R. E. et al. (2009). "Simulating Hydrologic and Hydraulic Processes throughout the Amazon River Basin". In: *Hydrological Processes* 23.8, pp. 1221–1235.

Berrocal, Veronica J., Adrian E. Raftery, and Tilmann Gneiting (Dec. 1, 2008). "Probabilistic Quantitative Precipitation Field Forecasting Using a Two-Stage Spatial Model". In: *The Annals of Applied Statistics* 2.4, pp. 1170–1193. JSTOR: 30245130.

Berveiller, Marc, Bruno Sudret, and Maurice Lemaire (Jan. 1, 2006). "Stochastic Finite Element: A Non Intrusive Approach by Regression". In: *European Journal of Computational Mechanics* 15.1-3, pp. 81–92. DOI: 10.3166/remn.15.81-92.

Beven, K. J. (2006). "Searching for the Holy Grail of Scientific Hydrology: Q(t)=H(S,R, Dt)A as Closure". In: *Hydrology and Earth System Science* 10, pp. 609–618.

Beven, K. J. et al. (2015). "Epistemic Uncertainties and Natural Hazard Risk Assessment–Part 1: A Review of the Issues". In: *Nat Hazards Earth Syst Sci Discuss* 3.12, pp. 7333–7377.

Beven, K. J. et al. (2016). "Epistemic Uncertainties and Natural Hazard Risk Assessment–Part 2: Different Natural Hazard Areas". In: *Natural Hazards and Earth System Sciences* 2016.1, pp. 1–1.

Beven, Keith (1982). "On Subsurface Stormflow: An Analysis of Response Times". In: *Hydrological sciences journal* 27.4, pp. 505–521.

— (1993). "Prophecy, Reality and Uncertainty in Distributed Hydrological Modelling". In: *Advances in Water Resources*. Research Perspectives in Hydrology 16.1, pp. 41–51. DOI: 10.1016/0309-1708(93)90028-E.

— (Aug. 15, 2008). "On Doing Better Hydrological Science". In: *Hydrological Processes* 22.17, pp. 3549–3553. DOI: 10.1002/hyp.7108.

Beven, Keith (2013). "So How Much of Your Error Is Epistemic? Lessons from Japan and Italy". In: *Hydrological Processes* 27.11, pp. 1677–1680. DOI: 10.1002/hyp.9648.

Beven, Keith and Peter Germann (1982). "Macropores and Water Flow in Soils". In: *Water resources research* 18.5, pp. 1311–1325.

Beven, Keith and Ida Westerberg (2011). "On Red Herrings and Real Herrings: Disinformation and Information in Hydrological Inference". In: *Hydrological Processes* 25.10, pp. 1676–1680. DOI: 10.1002/hyp.7963.

Beven, Keith J. (2000). "Uniqueness of Place and Process Representations in Hydrological Modelling". In: *Hydrology and Earth System Sciences Discussions* 4.2, pp. 203–213.

Bisht, Gautam et al. (Dec. 12, 2017). "Coupling a Three-Dimensional Subsurface Flow and Transport Model with a Land Surface Model to Simulate Stream – Aquifer – Land Interactions (CP v1.0)". In: *Geoscientific Model Development* 10.12, pp. 4539–4562. DOI: 10.5194/gmd-10-4539-2017.

Blatman, Géraud and Bruno Sudret (2008). "Sparse Polynomial Chaos Expansions and Adaptive Stochastic Finite Elements Using a Regression Approach". In: *Comptes Rendus Mécanique* 336.6, pp. 518–523.

— (2011). "Adaptive Sparse Polynomial Chaos Expansion Based on Least Angle Regression". In: *Journal of Computational Physics* 230.6, pp. 2345–2367.

Blöschl, Günter and Alberto Montanari (Jan. 30, 2010). "Climate Change Impacts — Throwing the Dice?" In: *Hydrological Processes* 24.3, pp. 374–381. DOI: 10.1002/hyp.7574.

Bogner, K. and F. Pappenberger (July 1, 2011). "Multiscale Error Analysis, Correction, and Predictive Uncertainty Estimation in a Flood Forecasting System". In: *Water Resources Research* 47.7, W07524. DOI: 10.1029/2010WR009137.

Bonan, Gordon B. (June 13, 2008). "Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests". In: *Science* 320.5882, pp. 1444–1449. DOI: 10.1126/science.1155121. pmid: 18556546.

Borga, Marco et al. (Aug. 30, 2008). "Surveying Flash Floods: Gauging the Ungauged Extremes". In: *Hydrological Processes* 22.18, pp. 3883–3885. DOI: 10.1002/hyp.7111.

Brando, Paulo M. et al. (Aug. 17, 2010). "Seasonal and Interannual Variability of Climate and Vegetation Indices across the Amazon". In: *Proceedings of the National Academy of Sciences* 107.33, pp. 14685–14690. DOI: 10.1073/pnas.0908741107. pmid: 20679201.

Broedel, Elisângela et al. (Apr. 30, 2017). "Deep Soil Water Dynamics in an Undisturbed Primary Forest in Central Amazonia: Differences between Normal Years and the 2005 Drought". In: *Hydrological Processes* 31.9, pp. 1749–1759. DOI: 10.1002/hyp.11143.

Brown, Sandra and Ariel E. Lugo (1982). "The Storage and Production of Organic Matter in Tropical Forests and Their Role in the Global Carbon Cycle". In: *Biotropica*, pp. 161–187.

Brum, Mauro et al. (July 16, 2018). "Hydrological Niche Segregation Defines Forest Structure and Drought Tolerance Strategies in a Seasonal Amazon Forest". In: *Journal of Ecology*. Ed. by Deepak Barua. DOI: 10.1111/1365-2745.13022.

Brunner, Philip and Craig T. Simmons (Mar. 1, 2012). "HydroGeoSphere: A Fully Integrated, Physically Based Hydrological Model". In: *Groundwater* 50.2, pp. 170–176. DOI: 10.1111/j.1745-6584.2011.00882.x.

Bulygina, Nataliya and Hoshin Gupta (May 1, 2011). "Correcting the Mathematical Structure of a Hydrological Model via Bayesian Data Assimilation". In: *Water Resources Research* 47.5, W05514. DOI: 10.1029/2010WR009614.

Burns, Douglas A. et al. (July 1, 2001). "Quantifying Contributions to Storm Runoff through End-Member Mixing Analysis and Hydrologic Measurements at the Panola Mountain Research Watershed (Georgia, USA)". In: *Hydrological Processes* 15.10, pp. 1903–1924. DOI: 10.1002/hyp.246.

Butler, T. et al. (Apr. 2015). "Definition and Solution of a Stochastic Inverse Problem for the Manning's n Parameter Field in Hydrodynamic Models". In: *Advances in Water Resources* 78, pp. 60–79. DOI: 10.1016/j.advwatres.2015.01.011.

Caflisch, Russel E. (1998). "Monte Carlo and Quasi-Monte Carlo Methods". In: *Acta numerica* 7, pp. 1–49.

Candès, Emmanuel J., Justin Romberg, and Terence Tao (2006). "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information". In: *IEEE Transactions on information theory* 52.2, pp. 489–509.

Cayan, Daniel R. et al. (Dec. 14, 2010). "Future Dryness in the Southwest US and the Hydrology of the Early 21st Century Drought". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.50, pp. 21271–21276.

Chapman, Tom G. (June 15, 1986). "Entropy as a Measure of Hydrologic Data Uncertainty and Model Performance". In: *Journal of Hydrology* 85.1, pp. 111–126. DOI: 10.1016/0022-1694(86)90079-X.

Chen, Fan et al. (Apr. 2011a). "Improving Hydrologic Predictions of a Catchment Model via Assimilation of Surface Soil Moisture". In: *Advances in Water Resources* 34.4, pp. 526–536. DOI: 10.1016/j.advwatres.2011.01.011.

Chen, Jie et al. (2011b). "Overall Uncertainty Study of the Hydrological Impacts of Climate Change for a Canadian Watershed". In: *Water Resources Research* 47.12.

Chitra-Tarak, Rutuja et al. (July 2018). "The Roots of the Drought: Hydrology and Water Uptake Strategies Mediate Forest-Wide Demographic Response to Precipitation". In: *Journal of Ecology* 106.4. Ed. by Gerhard Zotz, pp. 1495–1507. DOI: 10.1111/1365-2745.12925.

Cibin, R. et al. (Feb. 15, 2014). "Application of Distributed Hydrological Models for Predictions in Ungauged Basins: A Method to Quantify Predictive Uncertainty". In: *Hydrological Processes* 28.4, pp. 2033–2045. DOI: 10.1002/hyp.9721.

Clark, Martyn P. et al. (Aug. 1, 2015). "Improving the Representation of Hydrologic Processes in Earth System Models". In: *Water Resources Research* 51.8, pp. 5929–5956. DOI: 10.1002/2015WR017096.

Coe, Michael T., Marcos H. Costa, and Erica A. Howard (2008). "Simulating the Surface Waters of the Amazon River Basin: Impacts of New River Geomorphic and Flow Parameterizations". In: *Hydrological processes* 22.14, pp. 2542–2553.

Cohen, Joel E. (Nov. 14, 2003). "Human Population: The Next Half Century". In: *Science* 302.5648, pp. 1172–1175. DOI: 10.1126/science.1088665. pmid: 14615528.

Collatz, G. J. et al. (1991). "Physiological and Environmental Regulation of Stomatal Conductance, Photosynthesis and Transpiration - a Model That Includes a Laminar Boundary-Layer". In: *Agricultural and Forest Meteorology* 52.2-4, pp. 107–136.

Constantine, Paul G., Eric Dow, and Qiqi Wang (2014). "Active Subspace Methods in Theory and Practice: Applications to Kriging Surfaces". In: *SIAM Journal on Scientific Computing* 36.4, A1500–A1524.

Cramer, Wolfgang et al. (2004). "Tropical Forests and the Global Carbon Cycle: Impacts of Atmospheric Carbon Dioxide, Climate Change and Rate of Deforestation". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 359.1443, pp. 331–343.

Cuartas, Luz Adriana et al. (July 9, 2007). "Interception Water-Partitioning Dynamics for a Pristine Rainforest in Central Amazonia: Marked Differences between Normal and Dry Years". In: *Agricultural and Forest Meteorology* 145.1–2, pp. 69–83. DOI: 10.1016/j.agrformet.2007.04.008.

Cuartas, Luz Adriana et al. (Sept. 2012). "Distributed Hydrological Modeling of a Micro-Scale Rainforest Watershed in Amazonia: Model Evaluation and Advances in Calibration Using the New HAND Terrain Model". In: *Journal of Hydrology* 462-463, pp. 15–27. DOI: 10.1016/j.jhydrol.2011.12.047.

Da Cruz, Paulo S., Roland N. Horne, and Clayton V. Deutsch (1999). "The Quality Map: A Tool for Reservoir Uncertainty Quantification and Decision Making". In: *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Davidson, Eric et al. (Feb. 1, 2011). "Carbon Inputs and Water Uptake in Deep Soils of an Eastern Amazon Forest". In: *Forest Science* 57.1, pp. 51–58. DOI: 10.1093/forestscience/57.1.51.

Davidson, Eric A. et al. (2012). "The Amazon Basin in Transition". In: *Nature* 481.7381, p. 321.

Davis, Philip J. and Philip Rabinowitz (2007). *Methods of Numerical Integration*. Courier Corporation.

De Groeve, T. et al. (Dec. 30, 2014). "Joining Forces in a Global Flood Partnership". In: *Bulletin of the American Meteorological Society* 96.5, ES97–ES100. DOI: 10.1175/BAMS-D-14-00147.1.

De Rocquigny, Etienne (2012). *Modelling under Risk and Uncertainty: An Introduction to Statistical, Phenomenological and Computational Methods*. John Wiley & Sons.

Debusschere, Bert J. et al. (2017). "The Uncertainty Quantification Toolkit (UQTk)". In: *Handbook of Uncertainty Quantification*.

DeChant, Caleb M. and Hamid Moradkhani (Apr. 1, 2012). "Examining the Effectiveness and Robustness of Sequential Data Assimilation Methods for Quantification of Uncertainty in Hydrologic Forecasting". In: *Water Resources Research* 48.4, W04518. DOI: 10.1029/2011WR011011.

Der Kiureghian, Armen and Ove Ditlevsen (2009). "Aleatory or Epistemic? Does It Matter?" In: *Structural Safety* 31.2, pp. 105–112.

Detwiler, Ralph Paul and Charles AS Hall (1988). "Tropical Forests and the Global Carbon Cycle". In: *Science* 239.4835, pp. 42–47.

Doostan, Alireza and Houman Owhadi (2011). "A Non-Adapted Sparse Approximation of PDEs with Stochastic Inputs". In: *Journal of Computational Physics* 230.8, pp. 3015–3034.

Doussan, Claude et al. (2006). "Water Uptake by Plant Roots: II–Modelling of Water Transfer in the Soil Root-System with Explicit Account of Flow within the Root System–Comparison with Experiments". In: *Plant and soil* 283.1-2, pp. 99–117.

Dwelle, M. Chase et al. (2018). "Streamflow, Stomata, and Soil Pits: Sources of Inference for Complex Models with Fast, Robust Uncertainty Quantification". In:

Eagleson, Peter S. (Feb. 17, 2005). *Ecohydrology: Darwinian Expression of Vegetation Form and Function*. Cambridge University Press. 485 pp. ISBN: 978-1-139-43340-2.

Efstratiadis, Andreas and Demetris Koutsoyiannis (Mar. 10, 2010). "One Decade of Multi-Objective Calibration Approaches in Hydrological Modelling: A Review". In: *Hydrological Sciences Journal* 55.1, pp. 58–78. DOI: 10.1080/02626660903526292.

Eldred, Michael and John Burkardt (2009). "Comparison of Non-Intrusive Polynomial Chaos and Stochastic Collocation Methods for Uncertainty Quantification". In: *47th AIAA Aerospace Sciences Meeting Including The New Horizons Forum and Aerospace Exposition*, p. 976.

Elsheikh, Ahmed H., Ibrahim Hoteit, and Mary F. Wheeler (Feb. 1, 2014a). "Efficient Bayesian Inference of Subsurface Flow Models Using Nested Sampling and Sparse Polynomial Chaos Surrogates". In: *Computer Methods in Applied Mechanics and Engineering* 269, pp. 515–537. DOI: 10.1016/j.cma.2013.11.001.

Elsheikh, Ahmed H., Mary F. Wheeler, and Ibrahim Hoteit (Feb. 1, 2014b). "Hybrid Nested Sampling Algorithm for Bayesian Model Selection Applied to Inverse Subsurface Flow Problems". In: *Journal of Computational Physics* 258, pp. 319–337. DOI: 10.1016/j.jcp.2013.10.001.

Evensen, Geir (2009). *Data Assimilation*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-03710-8 978-3-642-03711-5.

Faber, Michael Havbro (2005). "On the Treatment of Uncertainties and Probabilities in Engineering Decision Analysis". In: *Journal of Offshore Mechanics and Arctic Engineering* 127.3, pp. 243–248.

Fan, Y. and G. Miguez-Macho (Oct. 25, 2010). "Potential Groundwater Contribution to Amazon Evapotranspiration". In: *Hydrol. Earth Syst. Sci.* 14.10, pp. 2039–2056. DOI: 10.5194/hess-14-2039-2010.

Fan, Y., H. Li, and G. Miguez-Macho (Feb. 22, 2013). "Global Patterns of Groundwater Table Depth". In: *Science* 339.6122, pp. 940–943. DOI: 10.1126/science.1229881. pmid: 23430651.

Fan, Ying (2015). "Groundwater in the Earth's Critical Zone: Relevance to Large-Scale Patterns and Processes". In: *Water Resources Research* 51.5, pp. 3052–3069.

Fan, Ying et al. (Oct. 3, 2017). "Hydrologic Regulation of Plant Rooting Depth". In: *Proceedings of the National Academy of Sciences* 114.40, pp. 10572–10577. DOI: 10.1073/pnas.1712381114.

Fang, Yilin et al. (Aug. 27, 2017). "Influence of Landscape Heterogeneity on Water Available to Tropical Forests in an Amazonian Catchment and Implications for Modeling Drought Response: Water Available to Tropical Forest". In: *Journal of Geophysical Research: Atmospheres* 122.16, pp. 8410–8426. DOI: 10.1002/2017JD027066.

Farquhar, G. D., S. von Caemmerer, and J. A. Berry (1980). "A Biochemical Model of Photosynthetic CO_2 Assimilation in Leaves of C-3 Species". In: *Planta* 149.1, pp. 78–90.

Fatichi, S., V. Y. Ivanov, and E. Caporali (Feb. 1, 2012). "A Mechanistic Ecohydrological Model to Investigate Complex Interactions in Cold and Warm Water-Controlled Environments: 1. Theoretical Framework and Plot-Scale Analysis". In: *Journal of Advances in Modeling Earth Systems* 4.2, p. M05002. DOI: 10.1029/2011MS000086.

Fatichi, Simone et al. (June 1, 2016). "An Overview of Current Applications, Challenges, and Future Trends in Distributed Process-Based Models in Hydrology". In: *Journal of Hydrology* 537, pp. 45–60. DOI: 10.1016/j.jhydrol.2016.03.026.

Feddes, Reinder A. et al. (Dec. 2001). "Modeling Root Water Uptake in Hydrological and Climate Models". In: *Bulletin of the American Meteorological Society* 82.12, pp. 2797–2809. DOI: 10.1175/1520-0477(2001)082<2797:MRWUIH>2.3.CO;2.

Flegal, James M., Murali Haran, and Galin L. Jones (May 1, 2008). "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?" In: *Statistical Science* 23.2, pp. 250–260. JSTOR: 27645897.

Fleischbein, Katrin et al. (2006). "Water Budgets of Three Small Catchments under Montane Forest in Ecuador: Experimental and Modelling Approach". In: *Hydrological Processes* 20.12, pp. 2491–2507.

Fontanazza, C. M., G. Freni, and V. Notaro (Aug. 2012). "Bayesian Inference Analysis of the Uncertainty Linked to the Evaluation of Potential Flood Damage in Urban Areas". In: *Water Science & Technology* 66.8, p. 1669. DOI: 10.2166/wst.2012.359.

García, Luís et al. (2015). "Modeling and Real-Time Control of Urban Drainage Systems: A Review". In: *Advances in Water Resources* 85, pp. 120–132.

Garrigues, Emmanuelle, Claude Doussan, and Alain Pierret (2006). "Water Uptake by Plant Roots: I–Formation and Propagation of a Water Extraction Front in Mature Root Systems as Evidenced by 2D Light Transmission Imaging". In: *Plant and soil* 283.1-2, p. 83.

Genuchten, M. Th van (1980). "A Closed-Form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils". In: *Soil Science Society of America Journal* 44, pp. 892–898.

Getirana, Augusto C. V. et al. (Sept. 26, 2014). "Water Balance in the Amazon Basin from a Land Surface Model Ensemble". In: *Journal of Hydrometeorology* 15.6, pp. 2586–2614. DOI: 10.1175/JHM-D-14-0068.1.

Ghanem, Roger G. and Alireza Doostan (Sept. 1, 2006). "On the Construction and Analysis of Stochastic Models: Characterization and Propagation of the Errors Associated with Limited Data". In: *Journal of Computational Physics*. Uncertainty Quantification in Simulation Science 217.1, pp. 63–81. DOI: 10.1016/j.jcp.2006.01.037.

Gilbert, James M. et al. (June 2016). "Global Spatial Sensitivity of Runoff to Subsurface Permeability Using the Active Subspace Method". In: *Advances in Water Resources* 92, pp. 30–42. DOI: 10.1016/j.advwatres.2016.03.020.

Gilks, Walter R., Sylvia Richardson, and David Spiegelhalter (1995). *Markov Chain Monte Carlo in Practice*. CRC press.

Glynn, Peter W. and Donald L. Iglehart (1989). "Importance Sampling for Stochastic Simulations". In: *Management Science* 35.11, pp. 1367–1392.

Grabe, Michael (2014). *Measurement Uncertainties in Science and Technology*. Springer.

Graham, Eric A. et al. (Jan. 21, 2003). "Cloud Cover Limits Net CO2 Uptake and Growth of a Rainforest Tree during Tropical Rainy Seasons". In: *Proceedings of the National Academy of Sciences* 100.2, pp. 572–576. DOI: 10.1073/pnas.0133045100. pmid: 12518044.

Grayson, Rodger B et al. (Aug. 2002). "Advances in the Use of Observed Spatial Patterns of Catchment Hydrological Response". In: *Advances in Water Resources* 25.8–12, pp. 1313–1334. DOI: 10.1016/S0309-1708(02)00060-X.

Guan, Kaiyu et al. (Apr. 2015). "Photosynthetic Seasonality of Global Tropical Forests Constrained by Hydroclimate". In: *Nature Geoscience* 8.4, pp. 284–289. DOI: 10.1038/ngeo2382.

Guinot, Vincent (Mar. 2012). "Multiple Porosity Shallow Water Models for Macroscopic Modelling of Urban Floods". In: *Advances in Water Resources* 37, pp. 40–72. DOI: 10.1016/j.advwatres.2011.11.002.

Gupta, H. V. et al. (2014). "Large-Sample Hydrology: A Need to Balance Depth with Breadth". In: *Hydrology and Earth System Sciences* 18.2, pp. 463–477. DOI: http://dx.doi.org.proxy.lib.umich.edu/10.5194/hess-18-463-2014.

Gutiérrez-Jurado, H. A. et al. (2006). "Ecohydrology of Root Zone Water Fluxes and Soil Development in Complex Semiarid Rangelands". In: *Hydrological Processes* 20, pp. 3289–3316.

Haario, Heikki, Eero Saksman, and Johanna Tamminen (2001). "An Adaptive Metropolis Algorithm". In: *Bernoulli* 7.2, pp. 223–242.
Hadamard, J. (1902). "Sur Les Problemes Aux Derivees Partielles et Leur Signification Physique". In: *Princeton University Bulletin* 13, pp. 49–52.
Hall, Jonathan et al. (2014). "Understanding Flood Regime Changes in Europe: A State of the Art Assessment". In: *Hydrology and Earth System Sciences* 18.7, pp. 2735–2772.
Hastings, W. Keith (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1, pp. 97–109.
Higham, Nicholas J. (2002). *Accuracy and Stability of Numerical Algorithms*. Vol. 80. SIAM.
Hildebrandt, A., A. Kleidon, and M. Bechmann (Aug. 29, 2016). "A Thermodynamic Formulation of Root Water Uptake". In: *Hydrol. Earth Syst. Sci.* 20.8, pp. 3441–3454. DOI: 10.5194/hess-20-3441-2016.
Hill, David et al. (2014). "Sensing and Cyberinfrastructure for Smarter Water Management: The Promise and Challenge of Ubiquity". In: *Journal of Water Resources Planning and Management* 140.7, p. 01814002. DOI: 10.1061/(ASCE)WR.1943-5452.0000449.
Hillel, D. (1980). *Fundamentals of Soil Physics*. New York, NY, USA: Academic Press.
Hirabayashi, Yukiko et al. (Sept. 2013). "Global Flood Risk under Climate Change". In: *Nature Climate Change* 3.9, pp. 816–821. DOI: 10.1038/nclimate1911.
Hoang, L. and R. A. Fenner (2016). "System Interactions of Stormwater Management Using Sustainable Urban Drainage Systems and Green Infrastructure". In: *Urban Water Journal*, pp. 1–20.
Hopp, Luisa, Simone Fatichi, and Valeriy Y. Ivanov (Nov. 13, 2015). "Simulating Water Flow in Variably Saturated Soils: A Comparison of a 3-D Model with Approximation-Based Formulations". In: *Hydrology Research*, nh2015126. DOI: 10.2166/nh.2015.126.
Hrachowitz, M. et al. (Aug. 1, 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.
Huete, Alfredo R. et al. (2006). "Amazon Rainforests Green-up with Sunlight in Dry Season". In: *Geophysical Research Letters* 33.6. DOI: 10.1029/2005GL025583.
Hutyra, Lucy R. et al. (Sept. 2007). "Seasonal Controls on the Exchange of Carbon and Water in an Amazonian Rain Forest: SEASONALITY OF AMAZONIAN CARBON FLUXES". In: *Journal of Geophysical Research: Biogeosciences* 112.G3. DOI: 10.1029/2006JG000365.
Idier, Jérôme (2013). *Bayesian Approach to Inverse Problems*. John Wiley & Sons.
Ivanov, Valeriy Y. et al. (Oct. 2004). "Preserving High-Resolution Surface and Rainfall Data in Operational-Scale Basin Hydrology: A Fully-Distributed Physically-Based Approach". In: *Journal of Hydrology* 298.1-4, pp. 80–111. DOI: 10.1016/j.jhydrol.2004.03.041.

Ivanov, Valeriy Y., Rafael L. Bras, and Enrique R. Vivoni (Mar. 2008a). "Vegetation-Hydrology Dynamics in Complex Terrain of Semiarid Areas: 1. A Mechanistic Approach to Modeling Dynamic Feedbacks". In: *Water Resources Research* 44.3. DOI: 10.1029/2006WR005588.

— (2008b). "Vegetation-Hydrology Dynamics in Complex Terrain of Semiarid Areas: 2. Energy-Water Controls of Vegetation Spatiotemporal Dynamics and Topographic Niches of Favorability". In: *Water Resources Research* 44.3, W03430. DOI: 10.1029/2006WR005595.

Ivanov, Valeriy Y. et al. (Sept. 1, 2010). "Hysteresis of Soil Moisture Spatial Heterogeneity and the "Homogenizing" Effect of Vegetation". In: *Water Resources Research* 46.9, W09521. DOI: 10.1029/2009WR008611.

Ivanov, Valeriy Y. et al. (Dec. 1, 2012). "Root Niche Separation Can Explain Avoidance of Seasonal Drought Stress and Vulnerability of Overstory Trees to Extended Drought in a Mature Amazonian Forest". In: *Water Resources Research* 48.12. DOI: 10.1029/2012WR011972.

Ivanov, Valeriy Yuryevich (2006). "Effects of Dynamic Vegetation and Topography on Hydrological Processes in Semi-Arid Areas". Thesis. Massachusetts Institute of Technology.

Jarvis, Andy et al. (2008). "Hole-Filled SRTM for the Globe Version 4". In:

Jaynes, E. T and G. Larry Bretthorst (2003). *Probability Theory the Logic of Science*. Cambridge, UK; New York, NY: Cambridge University Press. ISBN: 978-0-511-06589-7 978-0-521-59271-0 978-0-511-06802-7 978-0-511-79042-3 978-1-139-63632-2 978-1-280-41722-1.

Ji, Shihao, Ya Xue, and Lawrence Carin (June 2008). "Bayesian Compressive Sensing". In: *IEEE Transactions on Signal Processing* 56.6, pp. 2346–2356. DOI: 10.1109/TSP.2007.914345.

Jin, R., W. Chen, and T. W. Simpson (Dec. 1, 2001). "Comparative Studies of Metamodelling Techniques under Multiple Modelling Criteria". In: *Structural and Multidisciplinary Optimization* 23.1, pp. 1–13. DOI: 10.1007/s00158-001-0160-4.

Karhunen, Kari (1946). "Zur Spektraltheorie Stochastischer Prozesse". In: *Ann. Acad. Sci. Fennicae, AI* 34.

Kennedy, M. C. and A. O'Hagan (2000). "Predicting the Output from a Complex Computer Code When Fast Approximations Are Available". In: *Biometrika* 87.1, pp. 1–13. JSTOR: 2673557.

Kim, Byunghyun et al. (Apr. 2015). "Urban Flood Modeling with Porous Shallow-Water Equations: A Case Study of Model Errors in the Presence of Anisotropic Porosity". In: *Journal of Hydrology* 523, pp. 680–692. DOI: 10.1016/j.jhydrol.2015.01.059.

Kim, Cornelis P., Guido D. Salvucci, and Dara Entekhabi (1999). "Groundwater-Surface Water Interaction and the Climatic Spatial Patterns of Hillslope Hy-

drological Response". In: *Hydrology and Earth System Sciences Discussions* 3.3, pp. 375–384.
Kim, Jongho and Valeriy Y. Ivanov (Feb. 1, 2014). "On the Nonuniqueness of Sediment Yield at the Catchment Scale: The Effects of Soil Antecedent Conditions and Surface Shield". In: *Water Resources Research* 50.2, pp. 1025–1045. DOI: 10.1002/2013WR014580.
— (Mar. 2015). "A Holistic, Multi-Scale Dynamic Downscaling Framework for Climate Impact Assessments and Challenges of Addressing Finer-Scale Watershed Dynamics". In: *Journal of Hydrology* 522, pp. 645–660. DOI: 10.1016/j.jhydrol.2015.01.025.
Kim, Jongho et al. (Mar. 2012). "Coupled Modeling of Hydrologic and Hydrodynamic Processes Including Overland and Channel Flow". In: *Advances in Water Resources* 37, pp. 104–126. DOI: 10.1016/j.advwatres.2011.11.009.
Kim, Jongho et al. (June 2016). "On the Non-Uniqueness of the Hydro-Geomorphic Responses in a Zero-Order Catchment with Respect to Soil Moisture". In: *Advances in Water Resources* 92, pp. 73–89. DOI: 10.1016/j.advwatres.2016.03.019.
Kirchner, James W. (Feb. 1, 2009a). "Catchments as Simple Dynamical Systems: Catchment Characterization, Rainfall-Runoff Modeling, and Doing Hydrology Backward". In: *Water Resources Research* 45.2, W02429. DOI: 10.1029/2008WR006912.
— (2009b). "Catchments as Simple Dynamical Systems: Catchment Characterization, Rainfall-Runoff Modeling, and Doing Hydrology Backward". In: *Water Resources Research* 45.2.
Knio, O. M. and O. P. Le Maître (Sept. 2006). "Uncertainty Propagation in CFD Using Polynomial Chaos Decomposition". In: *Fluid Dynamics Research*. Recent Topics in Computational Fluid DynamicsRecent Topics in Computational Fluid Dynamics 38.9, pp. 616–640. DOI: 10.1016/j.fluiddyn.2005.12.003.
Kollet, Stefan et al. (Jan. 2017). "The Integrated Hydrologic Model Intercomparison Project, IH-MIP2: A Second Set of Benchmark Results to Diagnose Integrated Hydrology and Feedbacks: Integrated Hydrologic Model Intercomparison, IH-MIP2". In: *Water Resources Research* 53.1, pp. 867–890. DOI: 10.1002/2016WR019191.
Kollet, Stefan J. and Reed M. Maxwell (July 2006). "Integrated Surface–Groundwater Flow Modeling: A Free-Surface Overland Flow Boundary Condition in a Parallel Groundwater Flow Model". In: *Advances in Water Resources* 29.7, pp. 945–958. DOI: 10.1016/j.advwatres.2005.08.006.
Kong, Augustine, Jun S. Liu, and Wing Hung Wong (Mar. 1, 1994). "Sequential Imputations and Bayesian Missing Data Problems". In: *Journal of the American Statistical Association* 89.425, pp. 278–288. DOI: 10.1080/01621459.1994.10476469.
Kowalsky, Michael B., Stefan Finsterle, and Yoram Rubin (June 1, 2004). "Estimating Flow Parameter Distributions Using Ground-Penetrating Radar and Hydrolog-

ical Measurements during Transient Flow in the Vadose Zone". In: *Advances in Water Resources* 27.6, pp. 583–599. DOI: 10.1016/j.advwatres.2004.03.003.
Krakauer, Nir Y., Haibin Li, and Ying Fan (2014). "Groundwater Flow across Spatial Scales: Importance for Climate Modeling". In: *Environmental Research Letters* 9.3, p. 034003. DOI: 10.1088/1748-9326/9/3/034003.
Krasnosel'skii, M. A. and A. V. Pokrovskii (1989). *Systems with Hysteresis*. New York, USA: Springer-Verlag.
Krause, Stefan et al. (June 1, 2015). "Frontiers in Real-Time Ecohydrology – a Paradigm Shift in Understanding Complex Environmental Systems". In: *Ecohydrology* 8.4, pp. 529–537. DOI: 10.1002/eco.1646.
Krzysztofowicz, Roman (2001). "The Case for Probabilistic Forecasting in Hydrology". In: *Journal of hydrology* 249.1-4, pp. 2–9.
Kundzewicz, Zbigniew W. et al. (Jan. 2, 2014). "Flood Risk and Climate Change: Global and Regional Perspectives". In: *Hydrological Sciences Journal* 59.1, pp. 1–28. DOI: 10.1080/02626667.2013.857411.
Le Maître, O. P. and Omar M. Knio (2010). *Spectral Methods for Uncertainty Quantification*. Scientific Computation. Dordrecht: Springer Netherlands. ISBN: 978-90-481-3519-6 978-90-481-3520-2.
Lee, John A. and Michel Verleysen (2007). *Nonlinear Dimensionality Reduction*. Springer Science & Business Media.
Leuning, R (Jan. 1, 1990). "Modelling Stomatal Behaviour and and Photosynthesis of Eucalyptus Grandis". In: *Functional Plant Biology* 17.2, pp. 159–175.
Leuning, R. (Apr. 1, 1995). "A Critical Appraisal of a Combined Stomatal -Photosynthesis Model for C3 Plants". In: *Plant, Cell & Environment* 18.4, pp. 339–355. DOI: 10.1111/j.1365-3040.1995.tb00370.x.
Lewis, Simon L. et al. (Feb. 4, 2011). "The 2010 Amazon Drought". In: *Science* 331.6017, pp. 554–554. DOI: 10.1126/science.1200807. pmid: 21292971.
Lin, G. and G. E. Karniadakis (Nov. 5, 2009). "Sensitivity Analysis and Stochastic Simulations of Non-Equilibrium Plasma Flow". In: *International Journal for Numerical Methods in Engineering* 80.6-7, pp. 738–766. DOI: 10.1002/nme.2582.
Lin, JIAN-YI, CHUN-TIAN CHENG, and KWOK-WING CHAU (Aug. 1, 2006). "Using Support Vector Machines for Long-Term Discharge Prediction". In: *Hydrological Sciences Journal* 51.4, pp. 599–612. DOI: 10.1623/hysj.51.4.599.
Lin, Yen-Heng, Min-Hui Lo, and Chia Chou (May 5, 2015). "Potential Negative Effects of Groundwater Dynamics on Dry Season Convection in the Amazon River Basin". In: *Climate Dynamics* 46.3-4, pp. 1001–1013. DOI: 10.1007/s00382-015-2628-8.
Lincoln, Tim (May 24, 2007). "Hydrology: Flood of Data". In: *Nature* 447.7143, pp. 393–393. DOI: 10.1038/447393b.
Liu, Yanli et al. (Mar. 30, 2009). "Towards a Limits of Acceptability Approach to the Calibration of Hydrological Models: Extending Observation Error". In: *Journal of Hydrology* 367.1, pp. 93–103. DOI: 10.1016/j.jhydrol.2009.01.016.

Lobet, Guillaume et al. (Apr. 1, 2014). "Plant Water Uptake in Drying Soils". In: *Plant Physiology* 164.4, pp. 1619–1627. DOI: 10.1104/pp.113.233486. pmid: 24515834.

MacKay, David JC (1998). "Introduction to Monte Carlo Methods". In: *Learning in Graphical Models*. Springer, pp. 175–204.

Manabe, Syukuro (Nov. 1, 1969). "Climate and the Ocean Circulation. II. The Atmospheric Circulation and the Effect of Heat Transfer by Ocean Currents". In: *Monthly Weather Review* 97, pp. 775–805. DOI: 10.1175/1520-0493(1969)097<0739:CATOC>2.3.CO;2.

Marelli, Stefano and Bruno Sudret (2014). *UQLab: A Framework for Uncertainty Quantification in MATLAB*. ETH-Zürich.

Marengo, J. A. and J. C. Espinoza (Mar. 2016). "Extreme Seasonal Droughts and Floods in Amazonia: Causes, Trends and Impacts: EXTREMES IN AMAZONIA". In: *International Journal of Climatology* 36.3, pp. 1033–1050. DOI: 10.1002/joc.4420.

Markewitz, Daniel et al. (Aug. 2010). "Soil Moisture Depletion under Simulated Drought in the Amazon: Impacts on Deep Root Uptake". In: *New Phytologist* 187.3, pp. 592–607. DOI: 10.1111/j.1469-8137.2010.03391.x.

Marzouk, Youssef and Dongbin Xiu (2009). "A Stochastic Collocation Approach to Bayesian Inference in Inverse Problems". In: *Communications in Computational Physics* 6.4, pp. 826–847. DOI: 10.4208/cicp.2009.v6.p826.

Marzouk, Youssef M. and Habib N. Najm (Apr. 2009). "Dimensionality Reduction and Polynomial Chaos Acceleration of Bayesian Inference in Inverse Problems". In: *Journal of Computational Physics* 228.6, pp. 1862–1902. DOI: 10.1016/j.jcp.2008.11.024.

Marzouk, Youssef M., Habib N. Najm, and Larry A. Rahn (June 10, 2007). "Stochastic Spectral Methods for Efficient Bayesian Solution of Inverse Problems". In: *Journal of Computational Physics* 224.2, pp. 560–586. DOI: 10.1016/j.jcp.2006.10.010.

Maxwell, Reed M. and Laura E. Condon (July 22, 2016). "Connections between Groundwater Flow and Transpiration Partitioning". In: *Science* 353.6297, pp. 377–380. DOI: 10.1126/science.aaf7891. pmid: 27463671.

Maxwell, Reed M. et al. (Feb. 1, 2014). "Surface-Subsurface Model Intercomparison: A First Set of Benchmark Results to Diagnose Integrated Hydrology and Feedbacks". In: *Water Resources Research* 50.2, pp. 1531–1549. DOI: 10.1002/2013WR013725.

McCord, J. T., D. B. Stephens, and J. L. Wilson (1991). "Hysteresis and State-Dependent Anisotropy in Modeling Unsaturated Hillslope Hydrologic Processes". In: *Water Resources Research* 27.7, pp. 1501–1518.

McGrayne, Sharon Bertsch (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.

McKay, Michael D., Richard J. Beckman, and William J. Conover (1979). "Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". In: *Technometrics* 21.2, pp. 239–245.

McLaughlin, Dennis and Lloyd R. Townley (1996). "A Reassessment of the Groundwater Inverse Problem". In: *Water Resources Research* 32.5, pp. 1131–1161.

Meerveld, H. J. Tromp-van and M. Weiler (2008). "Hillslope Dynamics Modeled with Increasing Complexity". In: *Journal of Hydrology* 361, pp. 24–40. DOI: 10.1016/j.jhydrol.2008.07.019.

Meinzer, Frederick C. et al. (Oct. 2009). "Xylem Hydraulic Safety Margins in Woody Plants: Coordination of Stomatal Control of Xylem Tension with Hydraulic Capacitance". In: *Functional Ecology* 23.5, pp. 922–930. DOI: 10.1111/j.1365-2435.2009.01577.x.

Miguez-Macho, Gonzalo and Ying Fan (Aug. 16, 2012a). "The Role of Groundwater in the Amazon Water Cycle: 1. Influence on Seasonal Streamflow, Flooding and Wetlands". In: *Journal of Geophysical Research: Atmospheres* 117.D15, p. D15113. DOI: 10.1029/2012JD017539.

— (Aug. 16, 2012b). "The Role of Groundwater in the Amazon Water Cycle: 2. Influence on Seasonal Soil Moisture and Evapotranspiration". In: *Journal of Geophysical Research: Atmospheres* 117.D15, p. D15114. DOI: 10.1029/2012JD017540.

Miller, K. L. et al. (Jan. 1, 2018). "Efficient Uncertainty Quantification in Fully-Integrated Surface and Subsurface Hydrologic Simulations". In: *Advances in Water Resources* 111, pp. 381–394. DOI: 10.1016/j.advwatres.2017.10.023.

Milly, P. C. D. et al. (Jan. 31, 2002). "Increasing Risk of Great Floods in a Changing Climate". In: *Nature* 415.6871, pp. 514–517. DOI: 10.1038/415514a.

Mirus, Benjamin B. and Keith Loague (May 1, 2013). "How Runoff Begins (and Ends): Characterizing Hydrologic Response at the Catchment Scale". In: *Water Resources Research* 49.5, pp. 2987–3006. DOI: 10.1002/wrcr.20218.

Mishra, S., C. Schwab, and J. Šukys (Jan. 1, 2012). "Multilevel Monte Carlo Finite Volume Methods for Shallow Water Equations with Uncertain Topography in Multi-Dimensions". In: *SIAM Journal on Scientific Computing* 34.6, B761–B784. DOI: 10.1137/110857295.

Montanari, Alberto and Demetris Koutsoyiannis (2012). "A Blueprint for Process-Based Modeling of Uncertain Hydrological Systems". In: *Water Resources Research* 48.9, W09555. DOI: 10.1029/2011WR011412.

Montanari, Alberto, Christine A. Shoemaker, and Nick van de Giesen (Dec. 1, 2009). "Introduction to Special Section on Uncertainty Assessment in Surface and Subsurface Hydrology: An Overview of Issues and Challenges". In: *Water Resources Research* 45.12, W00B00. DOI: 10.1029/2009WR008471.

Moore, Benjamin J. et al. (Sept. 9, 2011). "Physical Processes Associated with Heavy Flooding Rainfall in Nashville, Tennessee, and Vicinity during 1–2 May 2010: The Role of an Atmospheric River and Mesoscale Convective Systems". In: *Monthly Weather Review* 140.2, pp. 358–378. DOI: 10.1175/MWR-D-11-00126.1.

Morss, Rebecca E. et al. (2005). "Flood Risk, Uncertainty, and Scientific Information for Decision Making: Lessons from an Interdisciplinary Project". In: *Bulletin of the American Meteorological Society* 86.11, pp. 1593–1601.

Najafi, M. R., H. Moradkhani, and I. W. Jung (Aug. 30, 2011). "Assessing the Uncertainties of Hydrologic Model Selection in Climate Change Impact Studies". In: *Hydrological Processes* 25.18, pp. 2814–2826. DOI: 10.1002/hyp.8043.

Najm, Habib N. (2009). "Uncertainty Quantification and Polynomial Chaos Techniques in Computational Fluid Dynamics". In: *Annual Review of Fluid Mechanics* 41.1, pp. 35–52. DOI: 10.1146/annurev.fluid.010908.165248.

Neal, Radford M. (1993). "Probabilistic Inference Using Markov Chain Monte Carlo Methods". In:

Nepstad, D. C. et al. (2002). "The Effects of Partial Throughfall Exclusion on Canopy Processes, Aboveground Production, and Biogeochemistry of an Amazon Forest". In: *Journal of Geophysical Research: Atmospheres* 107.D20, LBA 53–1–LBA 53–18. DOI: 10.1029/2001JD000360.

Nepstad, Daniel C. et al. (Dec. 1994). "The Role of Deep Roots in the Hydrological and Carbon Cycles of Amazonian Forests and Pastures". In: *Nature* 372.6507, pp. 666–669. DOI: 10.1038/372666a0.

Neuman, Shlomo P., Graham E. Fogg, and Elizabeth A. Jacobson (1980). "A Statistical Approach to the Inverse Problem of Aquifer Hydrology: 2. Case Study". In: *Water Resources Research* 16.1, pp. 33–58.

Ng, Leo Wai-Tsun and Michael Eldred (Apr. 23, 2012). "Multifidelity Uncertainty Quantification Using Non-Intrusive Polynomial Chaos and Stochastic Collocation". In: *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference
20th AIAA/ASME/AHS Adaptive Structures Conference
14th AIAA*. 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference
20th AIAA/ASME/AHS Adaptive Structures Conference
14th AIAA. Honolulu, Hawaii: American Institute of Aeronautics and Astronautics. ISBN: 978-1-60086-937-2. DOI: 10.2514/6.2012-1852.

Nobre, A. D. et al. (June 29, 2011). "Height Above the Nearest Drainage – a Hydrologically Relevant New Terrain Model". In: *Journal of Hydrology* 404.1–2, pp. 13–29. DOI: 10.1016/j.jhydrol.2011.03.051.

Norbiato, Daniele et al. (Dec. 5, 2008). "Flash Flood Warning Based on Rainfall Thresholds and Soil Moisture Conditions: An Assessment for Gauged and Ungauged Basins". In: *Journal of Hydrology* 362.3–4, pp. 274–290. DOI: 10.1016/j.jhydrol.2008.08.023.

O'Callaghan, J.H and D.M Mark (Dec. 1, 1984). "The Extraction of Drainage Networks from Digital Elevation Data". In: *Computer Vision, Graphics, and Image Processing* 28.3, pp. 323–344. DOI: 10.1016/S0734-189X(84)80011-0.

Oliveira, Rafael S. et al. (Sept. 1, 2005). "Hydraulic Redistribution in Three Amazonian Trees". In: *Oecologia* 145.3, pp. 354–363. DOI: 10.1007/s00442-005-0108-2.

Orlandini, S. et al. (1996). "Local Contributions to Infiltration Excess Runoff for a Conceptual Catchment Scale Model". In: *Water Resources Research* 32.7, pp. 2003–2012. DOI: 10.1029/96WR00897.

Owen, Art B. (2013). *Monte Carlo Theory, Methods and Examples*.

Pan, Yude et al. (Aug. 19, 2011). "A Large and Persistent Carbon Sink in the World's Forests". In: *Science* 333.6045, pp. 988–993. DOI: 10.1126/science.1201609. pmid: 21764754.

Parno, M. and Y. Marzouk (Jan. 1, 2018). "Transport Map Accelerated Markov Chain Monte Carlo". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.2, pp. 645–682. DOI: 10.1137/17M1134640.

Parno, Matthew, Andrew Davis, and Patrick Conrad (2014). "MIT Uncertainty Quantification (MUQ) Library". In:

Peel, Murray C. and Günter Blöschl (Jan. 4, 2011). "Hydrological Modelling in a Changing World". In: *Progress in Physical Geography* 35.2, pp. 249–261. DOI: 10.1177/0309133311402550.

Phillips, Oliver L. et al. (Mar. 6, 2009). "Drought Sensitivity of the Amazon Rainforest". In: *Science* 323.5919, pp. 1344–1347. DOI: 10.1126/science.1164033. pmid: 19265020.

Pokhrel, Yadu N. et al. (Apr. 27, 2013). "The Role of Groundwater in the Amazon Water Cycle: 3. Influence on Terrestrial Water Storage Computations and Comparison with GRACE". In: *Journal of Geophysical Research: Atmospheres* 118.8, pp. 3233–3244. DOI: 10.1002/jgrd.50335.

Pokhrel, Yadu N., Ying Fan, and Gonzalo Miguez-Macho (2014). "Potential Hydrologic Changes in the Amazon by the End of the 21st Century and the Groundwater Buffer". In: *Environmental Research Letters* 9.8, p. 084004. DOI: 10.1088/1748-9326/9/8/084004.

Poulter, Benjamin, Ursula Heyder, and Wolfgang Cramer (2009). "Modeling the Sensitivity of the Seasonal Cycle of GPP to Dynamic LAI and Soil Depths in Tropical Rainforests". In: *Ecosystems* 12.4, pp. 517–533. JSTOR: 40603614.

Rajabi, Mohammad Mahdi and Behzad Ataie-Ashtiani (May 1, 2014). "Sampling Efficiency in Monte Carlo Based Uncertainty Propagation Strategies: Application in Seawater Intrusion Simulations". In: *Advances in Water Resources* 67, pp. 46–64. DOI: 10.1016/j.advwatres.2014.02.004.

Ramaley, J. F. (Oct. 1969). "Buffon's Noodle Problem". In: *The American Mathematical Monthly* 76.8, p. 916. DOI: 10.2307/2317945. JSTOR: 2317945?origin=crossref.

Ranzani, G. (1980). "Identificação e Caracterização de Alguns Solos Da Estação Experimental de Silvicultura Tropical Do INPA". In: *Acta Amazonica* 10.1, pp. 7–41.

Rasmussen, Carl Edward (2006). "Gaussian Processes for Machine Learning". In:

Reichle, Rolf H., Wade T. Crow, and Christian L. Keppenne (Mar. 1, 2008). "An Adaptive Ensemble Kalman Filter for Soil Moisture Data Assimilation". In: *Water Resources Research* 44.3, W03423. DOI: 10.1029/2007WR006357.

Renard, Benjamin et al. (May 1, 2010). "Understanding Predictive Uncertainty in Hydrologic Modeling: The Challenge of Identifying Input and Structural Errors". In: *Water Resources Research* 46.5, W05521. DOI: 10.1029/2009WR008328.

Restrepo-Coupe, Natalia et al. (Dec. 2013). "What Drives the Seasonality of Photosynthesis across the Amazon Basin? A Cross-Site Analysis of Eddy Flux Tower Measurements from the Brasil Flux Network". In: *Agricultural and Forest Meteorology* 182-183, pp. 128–144. DOI: 10.1016/j.agrformet.2013.04.031.

Ricciuto, Daniel, Khachik Sargsyan, and Peter Thornton (Feb. 2018). "The Impact of Parametric Uncertainties on Biogeochemistry in the E3SM Land Model: ELM Biogeochemical Parameter Sensitivity". In: *Journal of Advances in Modeling Earth Systems* 10.2, pp. 297–319. DOI: 10.1002/2017MS000962.

Richey, Jeffrey E. et al. (2009). "The Role of Rivers in the Regional Carbon Balance". In: *Amazonia and Global Change*, pp. 489–504.

Richey, Jeffrey E. et al. (2011). "Land–Water Interactions in the Amazon". In: *Biogeochemistry* 105.1-3, p. 1.

Riley, W. J. and C. Shen (2014). "Characterizing Coarse-Resolution Watershed Soil Moisture Heterogeneity Using Fine-Scale Simulations". In: *Hydrology and Earth System Sciences; Katlenburg-Lindau* 18.7, p. 2463. DOI: http://dx.doi.org.proxy.lib.umich.edu/10.5194/hess-18-2463-2014.

Riley, W. J. et al. (July 20, 2011). "Barriers to Predicting Changes in Global Terrestrial Methane Fluxes: Analyses Using CLM4Me, a Methane Biogeochemistry Model Integrated in CESM". In: *Biogeosciences* 8.7, pp. 1925–1953. DOI: 10.5194/bg-8-1925-2011.

Ringeval, B. et al. (Mar. 21, 2014). "Methane Emissions from Floodplains in the Amazon Basin: Challenges in Developing a Process-Based Model for Global Applications". In: *Biogeosciences* 11.6, pp. 1519–1558. DOI: 10.5194/bg-11-1519-2014.

Ripley, Brian D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press. DOI: 10.1017/CBO9780511812651.

Rodell, M. et al. (Mar. 1, 2004). "The Global Land Data Assimilation System". In: *Bulletin of the American Meteorological Society* 85.3, pp. 381–394. DOI: 10.1175/BAMS-85-3-381.

Romano, N., M. Palladino, and G. B. Chirico (Dec. 21, 2011). "Parameterization of a Bucket Model for Soil-Vegetation-Atmosphere Modeling under Seasonal Climatic Regimes". In: *Hydrology and Earth System Sciences* 15.12, pp. 3877–3893. DOI: 10.5194/hess-15-3877-2011.

Romero-Saltos, Hugo et al. (Mar. 1, 2005). "Rainfall Exclusion in an Eastern Amazonian Forest Alters Soil Water Movement and Depth of Water Uptake". In: *American Journal of Botany* 92.3, pp. 443–455. DOI: 10.3732/ajb.92.3.443.

Rosenbaum, U. et al. (Oct. 1, 2012). "Seasonal and Event Dynamics of Spatial Soil Moisture Patterns at the Small Catchment Scale". In: *Water Resources Research* 48.10, W10544. DOI: 10.1029/2011WR011518.

Rudorff, Conrado M., John M. Melack, and Paul D. Bates (Jan. 1, 2014). "Flooding Dynamics on the Lower Amazon Floodplain: 2. Seasonal and Interannual Hydrological Variability". In: *Water Resources Research* 50.1, pp. 635–649. DOI: 10.1002/2013WR014714.

Russo, David and Moshe Bouton (July 1992). "Statistical Analysis of Spatial Variability in Unsaturated Flow Parameters". In: *Water Resources Research* 28.7, pp. 1911–1925. DOI: 10.1029/92WR00669.

Saatchi, Sassan et al. (Jan. 8, 2013). "Persistent Effects of a Severe Drought on Amazonian Forest Canopy". In: *Proceedings of the National Academy of Sciences* 110.2, pp. 565–570. DOI: 10.1073/pnas.1204651110. pmid: 23267086.

Salamon, Peter and Luc Feyen (Oct. 15, 2009). "Assessing Parameter, Precipitation, and Predictive Uncertainty in a Distributed Hydrological Model Using Sequential Data Assimilation with the Particle Filter". In: *Journal of Hydrology* 376.3–4, pp. 428–442. DOI: 10.1016/j.jhydrol.2009.07.051.

Salati, Eneas and Peter B. Vose (1984). "Amazon Basin: A System in Equilibrium". In: *Science* 225.4658, pp. 129–138. JSTOR: 1693078.

Saleska, Scott R. et al. (Nov. 28, 2003). "Carbon in Amazon Forests: Unexpected Seasonal Fluxes and Disturbance-Induced Losses". In: *Science* 302.5650, pp. 1554–1557. DOI: 10.1126/science.1091165. pmid: 14645845.

Saltelli, Andrea (2002). "Making Best Use of Model Evaluations to Compute Sensitivity Indices". In: *Computer physics communications* 145.2, pp. 280–297.

Salvucci, Guido Daniel and Dara Entekhabi (1995). "Hillslope and Climatic Controls on Hydrologic Fluxes". In: *Water Resources Research* 31.7, pp. 1725–1739.

Santos, Victor Alexandre Hardt Ferreira dos et al. (May 30, 2018). "Causes of Reduced Leaf-Level Photosynthesis during Strong El Niño Drought in a Central Amazon Forest". In: *Global Change Biology*. DOI: 10.1111/gcb.14293.

Sargsyan, K., H. N. Najm, and R. Ghanem (Apr. 1, 2015). "On the Statistical Calibration of Physical Models". In: *International Journal of Chemical Kinetics* 47.4, pp. 246–276. DOI: 10.1002/kin.20906.

Sargsyan, Khachik et al. (2014). "Dimensionality Reduction for Complex Models via Bayesian Compressive Sensing". In: *International Journal for Uncertainty Quantification* 4.1, pp. 63–93. DOI: 10.1615/Int.J.UncertaintyQuantification.2013006821.

Sargsyan, Khachik, Xun Huan, and Habib N. Najm (Jan. 20, 2018). "Embedded Model Error Representation for Bayesian Model Calibration". In: arXiv: 1801.06768 [physics, stat].

Schoups, Gerrit and Jasper A. Vrugt (Oct. 1, 2010). "A Formal Likelihood Function for Parameter and Predictive Inference of Hydrologic Models with Correlated, Heteroscedastic, and Non-Gaussian Errors". In: *Water Resources Research* 46.10, W10531. DOI: 10.1029/2009WR008933.

Schubert, Jochen E. and Brett F. Sanders (June 2012). "Building Treatments for Urban Flood Inundation Models and Implications for Predictive Skill and Modeling

Efficiency". In: *Advances in Water Resources* 41, pp. 49–64. DOI: 10.1016/j.advwatres.2012.02.012.

Seck, Alimatou, Claire Welty, and Reed M. Maxwell (Apr. 1, 2015). "Spin-up Behavior and Effects of Initial Conditions for an Integrated Hydrologic Model". In: *Water Resources Research* 51.4, pp. 2188–2210. DOI: 10.1002/2014WR016371.

Seibert, J. and K. J. Beven (2009). "Gauging the Ungauged Basin: How Many Discharge Measurements Are Needed?" In: *Hydrology and Earth System Sciences* 13.6, pp. 883–892.

Sharp, R. E. and W. J. Davies (1985). "Root Growth and Water Uptake by Maize Plants in Drying Soil". In: *Journal of Experimental Botany* 36.9, pp. 1441–1456.

Siegmund, David (1976). "Importance Sampling in the Monte Carlo Study of Sequential Tests". In: *The Annals of Statistics*, pp. 673–684.

Silvertown, Jonathan, Yoseph Araya, and David Gowing (Jan. 2015). "Hydrological Niches in Terrestrial Plant Communities: A Review". In: *Journal of Ecology* 103.1. Ed. by Will Cornwell, pp. 93–108. DOI: 10.1111/1365-2745.12332.

Sivapalan, Murugesu, Keith Beven, and Eric F. Wood (1987). "On Hydrologic Similarity: 2. A Scaled Model of Storm Runoff Production". In: *Water Resources Research* 23.12, pp. 2266–2278.

Smolyak, Sergey A. (1963). "Quadrature and Interpolation Formulas for Tensor Products of Certain Classes of Functions". In: *Dokl. Akad. Nauk SSSR*. Vol. 4, p. 123.

Sobol, I. M (Feb. 15, 2001). "Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates". In: *Mathematics and Computers in Simulation*. The Second IMACS Seminar on Monte Carlo Methods 55.1–3, pp. 271–280. DOI: 10.1016/S0378-4754(00)00270-6.

Spantini, Alessio, Daniele Bigoni, and Youssef Marzouk (Mar. 17, 2017). "Inference via Low-Dimensional Couplings". In: arXiv: 1703.06131 [stat].

Spiegelhalter, David J. and Steffen L. Lauritzen (Aug. 1, 1990). "Sequential Updating of Conditional Probabilities on Directed Graphical Structures". In: *Networks* 20.5, pp. 579–605. DOI: 10.1002/net.3230200507.

Srivastava, Prashant K. et al. (Apr. 18, 2013). "Machine Learning Techniques for Downscaling SMOS Satellite Soil Moisture Using MODIS Land Surface Temperature for Hydrological Application". In: *Water Resources Management* 27.8, pp. 3127–3144. DOI: 10.1007/s11269-013-0337-9.

Stark, Scott C. et al. (2012). "Amazon Forest Carbon Dynamics Predicted by Profiles of Canopy Leaf Area and Light Environment". In: *Ecology Letters* 15.12, pp. 1406–1414. DOI: 10.1111/j.1461-0248.2012.01864.x.

Sudret, Bruno (2008). "Global Sensitivity Analysis Using Polynomial Chaos Expansions". In: *Reliability Engineering & System Safety* 93.7, pp. 964–979.

Taiz, Lincoln and Eduardo Zeiger (2006). "Plant Physiology. 4th". In: *Sinauer Associate, Sunderland, Mass., EUA*.

Tarboton, D. (1997). "A New Method for the Determination of Flow Directions and Upslope Areas in Grid Digital Elevation Models". In: *Water Resources Research* 33.2, pp. 309–319.

Thornton, Peter E. et al. (2007). "Influence of Carbon-Nitrogen Cycle Coupling on Land Model Response to CO2 Fertilization and Climate Variability". In: *Global biogeochemical cycles* 21.4.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Tikhonov, A. N. and V. Ya Arsenin (1977). *Solutions of Ill-Posed Problems*. Washington, D.C: Winston & Sons.

Tomasella, Javier, Martin G. Hodnett, and Luciana Rossato (Jan. 1, 2000). "Pedotransfer Functions for the Estimation of Soil Water Retention in Brazilian Soils". In: *Soil Science Society of America Journal* 64.1, pp. 327–338. DOI: 10.2136/sssaj2000.641327x.

Tomasella, Javier et al. (June 30, 2008). "The Water Balance of an Amazonian Micro-Catchment: The Effect of Interannual Variability of Rainfall on Hydrological Behaviour". In: *Hydrological Processes* 22.13, pp. 2133–2147. DOI: 10.1002/hyp.6813.

Torre, E. et al. (Sept. 25, 2017). "A General Framework for Data-Driven Uncertainty Quantification under Complex Input Dependencies Using Vine Copulas". In: arXiv: 1709.08626 [stat].

Trenberth, Ke (Mar. 31, 2011). "Changes in Precipitation with Climate Change". In: *Climate Research* 47.1, pp. 123–138. DOI: 10.3354/cr00953.

Vertessy, Robert A. and Helmut Elsenbeer (1999). "Distributed Modeling of Storm Flow Generation in an Amazonian Rain Forest Catchment: Effects of Model Parameterization". In: *Water Resources Research* 35.7, pp. 2173–2187.

Villarini, Gabriele et al. (Nov. 17, 2010). "Towards Probabilistic Forecasting of Flash Floods: The Combined Effects of Uncertainty in Radar-Rainfall and Flash Flood Guidance". In: *Journal of Hydrology*. Flash Floods: Observations and Analysis of Hydrometeorological Controls 394.1–2, pp. 275–284. DOI: 10.1016/j.jhydrol.2010.02.014.

Vrugt, Jasper A. et al. (2008). "Treatment of Input Uncertainty in Hydrologic Modeling: Doing Hydrology Backward with Markov Chain Monte Carlo Simulation". In: *Water Resources Research* 44.12, W00B09. DOI: 10.1029/2007WR006720.

Weiler, M. and J. J. McDonnell (2004). "Virtual Experiments: A New Approach for Improving Process Conceptualization in Hillslope Hydrology". In: *Journal of Hydrology* 285, pp. 3–18. DOI: 10.1016/S0022-1694(03)00271-3.

West, Thomas K. and Clyde Gumbert (Jan. 9, 2017). "Multifidelity, Multidisciplinary Uncertainty Quantification with Non-Intrusive Polynomial Chaos". In: *58th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*. 58th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Ma-

terials Conference. Grapevine, Texas: American Institute of Aeronautics and Astronautics. ISBN: 978-1-62410-453-4. DOI: 10.2514/6.2017-1936.
Wiener, Norbert (1938). "The Homogeneous Chaos". In: *American Journal of Mathematics*, pp. 897–936. JSTOR: 2371268.
Winsemius, H. C. et al. (Dec. 1, 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, W12422. DOI: 10.1029/2009WR007706.
Winter, Thomas C. (Feb. 1, 1981). "Uncertainties in Estimating the Water Balance of Lakes1". In: *JAWRA Journal of the American Water Resources Association* 17.1, pp. 82–115. DOI: 10.1111/j.1752-1688.1981.tb02593.x.
Wong, Jefferson S. et al. (Jan. 15, 2015). "Sensitivity of a Hydraulic Model to Channel Erosion Uncertainty during Extreme Flooding". In: *Hydrological Processes* 29.2, pp. 261–279. DOI: 10.1002/hyp.10148.
Wu, J. et al. (Feb. 26, 2016). "Leaf Development and Demography Explain Photosynthetic Seasonality in Amazon Evergreen Forests". In: *Science* 351.6276, pp. 972–976. DOI: 10.1126/science.aad5068.
Xiu, D. and D. M. Tartakovsky (2004). "Uncertainty Quantification for Flow in Highly Heterogeneous Porous Media". In: *Developments in Water Science*. Ed. by Matthew W. Farthing Cass T. Miller William G. Gray and George F. Pinder. Vol. 55, Part 1. Computational Methods in Water Resources: Volume 1. Elsevier, pp. 695–703.
Xiu, Dongbin and George Em Karniadakis (Jan. 2002). "The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations". In: *SIAM Journal on Scientific Computing* 24.2, pp. 619–644. DOI: 10.1137/S1064827501387826.
— (2003). "Modeling Uncertainty in Flow Simulations via Generalized Polynomial Chaos". In: *Journal of computational physics* 187.1, pp. 137–167.
Yamazaki, Dai et al. (2011). "A Physically Based Description of Floodplain Inundation Dynamics in a Global River Routing Model". In: *Water Resources Research* 47.4.
Yeh, William W.-G. (1986). "Review of Parameter Identification Procedures in Groundwater Hydrology: The Inverse Problem". In: *Water Resources Research* 22.2, pp. 95–108.
Yu, Dapeng and Tom J. Coulthard (May 2015). "Evaluating the Importance of Catchment Hydrological Parameters for Urban Surface Water Flood Modelling Using a Simple Hydro-Inundation Model". In: *Journal of Hydrology* 524, pp. 385–400. DOI: 10.1016/j.jhydrol.2015.02.040.
Zheng, Zhibao and Hongzhe Dai (Sept. 1, 2017). "Simulation of Multi-Dimensional Random Fields by Karhunen–Loève Expansion". In: *Computer Methods in Applied Mechanics and Engineering* 324, pp. 221–247. DOI: 10.1016/j.cma.2017.05.022.