A FULLY BAYESIAN APPROACH TO UNCERTAINTY QUANTIFICATION OF GROUNDWATER MODELS

BY

TIANFANG XU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Albert J. Valocchi, Chair
Professor Feng Liang
Professor Paolo Gardoni
Professor Yu-Feng Forest Lin
Professor Praveen Kumar
Professor Ming Ye

Abstract

Effective water resources management typically relies on numerical models to analyze groundwater flow and solute transport processes. Since the important hydrogeological parameters of these models (e.g., hydraulic conductivity) cannot be measured, they are normally estimated by model calibration. In addition, groundwater models are often subject to input data errors, as some of the input forcings (such as recharge and well pumping rates) are unknown or estimated. Furthermore, model structural error is ubiquitous, due to simplification and/or misrepresentation of the real system. The presence of errors in input data and model structure raises questions regarding the suitability of conventional least squares regression-based (LSR) calibration.

We present a Bayesian framework that explicitly recognizes errors in input forcings and model structure and is tailored for groundwater models. The framework implements a marginalizing step to account for input data variability when evaluating the likelihood, and explicitly describes the model structural error statistically in an inductive, data-driven way. We adopt a fully Bayesian approach that integrates Gaussian process error models into the calibration, prediction and uncertainty analysis.

We test the usefulness of the fully Bayesian approach with synthetic case studies of the impact of pumping on surface-ground water interaction and a real-world case study. In the real-world case study, Bayesian inference is facilitated using high performance computing and fast surrogate models (based on machine learning technique) as a substitute for the

computationally expensive groundwater model. We demonstrate in the case studies that when input error or model structural error is present but not explicitly taken into account, the parameters can be overly adjusted to compensate for input data and model structural error, thus leading to biased and overconfident parameter estimates. Parameter compensation is found to have deleterious impact when the model is used to make prediction under new scenarios. In contrast, the presented Bayesian framework effectively alleviates parameter compensation and gives predictions that are more consistent with validation data in all case studies. The results highlight the importance of explicit treatment of errors in input forcings and model structure especially in circumstances where subsequent decision-making and risk analysis require accurate prediction and uncertainty quantification. Follow-up studies will further investigate the feasibility of joint inference of input and model structural errors, particularly for real-world modeling practice.

To my family

Acknowledgments

First and for most, I would like to thank my supervisor, Professor Albert Valocchi, for his invaluable guidance, mentoring and support throughout my graduate study over the past six years. His enthusiasm and devotion to research and teaching inspire me to continue the journey of discovery and pursue a career in academia.

I gratefully acknowledge my Ph.D. committee members, Prof. Paolo Gardoni and Prof. Praveen Kumar, Prof. Feng Liang and Prof. Ming Ye, for their time and support. Specifically, I would like to thank Prof. Feng Liang, my co-advisor, and Prof. Ye, for their critical, in-depth, detailed comments and suggestions on many parts of this dissertation. I also thank Dr. Yufeng Lin for all the constructive discussions on groundwater modeling practices.

This work was supported by the National Science Foundation Hydrologic Science Program under Grant No. 0943627, as well as the University of Illinois Computational Science and Engineering Fellowship.

I would also like to take this chance to thank all my friends at the Ven Te Chow Hydrosystems Laboratory. They made this PhD journey much easier and more enjoyable.

Finally, special thanks to my parents for their never-ending love and support, and to my husband, who is always ready to read my first draft.

Table of Contents

Chapter 1

INTRODUCTION

1.1 Background

Numerical groundwater models are being used to inform decisions and policies with enormous social, economical and political implications, such as water resources management and assessing risks of subsurface contamination. Therefore, it is critically important to ensure accuracy and quantify the intrinsic uncertainties of these models. It has been recognized that systematic structural error is ubiquitous in groundwater models, for example due to simplified or improper interpretation of geological structure and conceptualizations of flow and contaminant transport processes [17, 62, 69]. As a result, the model residual (the difference between observed and simulated quantities) contains both aleatoric and epistemic errors.

The groundwater inverse modeling literature has mainly focused on parameter uncertainty. The least squares regression (LSR) method is commonly used to estimate parameters and associated uncertainty from historical observations. The calibrated model is then used for subsequent prediction and uncertainty analysis. Linear prediction intervals are computed using parameter confidence interval and sensitivity of prediction with respect to parameters [37]. A great deal of effort has been made by the groundwater modeling community to improve the characterization of heterogeneous subsurface systems. Traditional model construction and calibration practice usually defines a handful of parameters by lumping subsurface properties into "zones", such that parameter values can be uniquely determined from available calibration data [37, 62]. In contrast to the parsimonious strategy, there has

been a trend of developing highly parameterized models to represent heterogeneity of subsurface systems. This trend is greatly facilitated by the increasing availability of measurements and computing power, as well as the emergence of sophisticated automatic calibration techniques and software, such as PEST [21] and UCODE [65]. Regularized inversion techniques [24, 81] are proposed to solve the nonuniqueness issue that often arises when calibrating a highly parameterized model. Following the calibration process, predictive uncertainty due to the heterogeneous parameter field can be assessed using either linear prediction intervals or explored using Monte Carlo methods [16, 17, 31, 82].

Underlying these existing approaches is the assumption that the model residual is dominated by aleatoric measurement error. The model residual is commonly assumed to be independent identically distributed (i.i.d.), Gaussian with zero mean and a constant variance. These assumptions are often violated in surface and groundwater modeling applications [37, 63] in addition to other domains of Earth system modeling [27, 93]. For example, Tiedeman and Green [80] discussed a scenario where multiple observations used as calibration targets were computed from overlapping sets of direct measurements. They then computed a full (i.e. with non-zero off-diagonal entries) covariance matrix of observation errors for use in calibration. This led to substantially different parameter estimates compared to using a diagonal covariance matrix that neglects correlation among observations.

Model structural error can also lead to spatial and temporal correlation in model residual [23, 98]. The lack of explicit treatment of model structural error in traditional calibration and prediction of groundwater models based on regression analysis can be problematic in circumstances where model structural error is the dominant contributor to model residual [23, 69]. Unlike in [80], the correlation structure is typically unknown before calibration in this case. Lu et al. [54] proposed an iterative two-stage method in the context of maximum likelihood Bayesian model averaging [61]. The method estimates the error covariance matrix by fitting an autoregressive model to the calibration residual, and then calibration is

repeated using the new covariance matrix.

Error in input data has been recognized as another source of systematic bias in model outputs in rainfall-runoff modeling. For example, Kavetski et al. [44] reasoned that the application of traditional least squares regression ignoring the high spatial and temporal uncertainty of precipitation can lead to biased parameter estimates and compromised prediction.

Handling input uncertainty of groundwater models is of great importance when an indirect method or another model has been used to estimate forcings such as precipitation recharge, percolation from irrigation, evapotranspiration and well pumping rates. Transient boundary conditions, such as flux from adjacent basins and stream flow or stages, can also be considered as model input. For example, irrigation pumping rates are rarely metered, but commonly inferred from well database, irrigation acreage, electricity usage, and other types of records [39, 58]. The amount and timing of precipitation recharge can be estimated by watershed models, by water-budget methods and by measuring unsaturated zone physical properties [34, 39, 51].

Uncertainty due to indirect estimation of these input forcings is sometimes characterized by multipliers. The multipliers are adjusted with other model parameters during the calibration process [58, 91]. In surface hydrology, Kavetski et al. [44] introduced storm event-based multipliers to characterize variability of rainfall, and inversely inferred the multipliers with rainfall-runoff model parameters via Bayesian calibration. This method has been used in later studies [40, 85], and it has been found that explicit treatment of precipitation error significantly altered the shape of the posterior distributions of model parameters and led to increased prediction uncertainty.

However, the multiplier approach typically results in a high dimensional calibration problem with nonuniqueness issues. In cases where input errors are much larger than output errors,

calibrating input multipliers essentially conditions the input on the output, often lead to nearly perfect fit to calibration data (streamflow observations). Since true rainfall input is rarely known in modeling practice, validation of the inferred multiplier, and therefore the model parameters and outputs, is difficult [2, 40]. The nonuniqueness issue tends to be more severe in groundwater modeling problems, as it would take a great number of multipliers to describe inputs that vary temporally and spatially. In the groundwater modeling literature, there exists no systematic approach to investigate the impact of input data uncertainty on calibration and prediction.

Similarly, when calibrating an imperfect model, parameters may be over-adjusted to compensate for errors in model structure[23, 25]. In groundwater literature, Moore and Doherty [59] first introduced the strategic use of "compensatory parameters" within the framework of highly parameterized calibration. Follow-up studies showed that use of many parameters helped to localize parameter compensation [22, 23, 91]. Specifically, "correction parameter" that do not necessarily have strict physical basis can be introduced to absorb model structural error, so that other parameters can be better estimated [22].

Some investigators have shown that the impact of parameter compensation on model predictive ability is prediction dependent [22, 23, 59, 91]. If predictions are very similar (e.g. in location or corresponding forcing scenarios to calibration data, parameter compensation may improve predictive accuracy. Thus they proposed the conjunctive use of multiple models; a model is calibrated multiple times, each time against a different type of observations and would be used for the corresponding type during prediction [22]. In addition, calibration targets that most resemble the prediction quantity can be assigned higher weights to enforce good fit. However, the deleterious implication of "compensatory parameters" can be significant when models are used to predict under a future scenario different from historical conditions reflected by calibration data [91]. In this case, while small calibration residual can be obtained by calibrating a highly parameterized model, parameter compensation leads

to biased prediction and underestimated predictive uncertainty.

Doherty and Christensen [22] proposed a paired complex/simple model method based on linear subspace analysis to quantify calibration-induced predictive bias. This method requires Monte Carlo runs of a complex model and repeated calibration of the simple model, and therefore is very computationally expensive. The linear subspace analysis was later used in a less computationally demanding way to estimate the prediction error covariance matrix [91]. Both studies relied on a complex, highly parameterized model that represents the complexity and variability of the true subsurface systems. The underlying hypothesis is that the contribution of model structural error to predictive error can be described by a linear model of "omitted parameters", i.e. parameters present in the reality but omitted in the approximate numerical model. However, other aspects, such as the interpretation of geological structure and conceptualizations of flow and transport processes, can also contribute to model structural error [62, 69]. Unlike the heterogeneity of subsurface properties, these aspects cannot be parameterized straightforwardly.

Other approaches have been proposed to accommodate correlated and non-Gaussian model residual in the field of surface hydrology, mostly in a Bayesian context. Beven and Freer [8] proposed the generalized likelihood uncertainty estimation methodology (GLUE), which uses a subjective likelihood function to allow for users' judgment of model goodness-of-fit. The idea is related to Approximate Bayesian Computation (ABC), a "likelihood free" method recently introduced to hydrologic model inversion [73]. Progress has also been made to construct formal likelihood functions based on statistical characterization of residual. For example, Schoups and Vrugt [75] constructed a formal generalized likelihood function to handle residual errors that are correlated, heteroscedastic, non-Gaussian and exhibit kurtosis and skewness. The characterization of residual distribution was jointly estimated with parameters of the computation model by a Markov chain Monte Carlo (MCMC) sampler. Applications of the generalized likelihood functions to rainfall-runoff [75] and reactive trans-

port [76] showed that proper representation of the distribution of residual provided improved estimates of parameter and predictive uncertainty over statistical characterization based only on measurement error. Similar approaches were applied to calibrate effective parameters of a layered unsaturated flow column model [25]. As increasing attention to model structure uncertainty has arisen, there have been debates over whether a formal likelihood function should be used instead of an informal or subjective one [10, 73, 79, 86].

The use of formal and subjective likelihood functions in a variety of applications [9, 25, 75, 76] suggests the utility of a Bayesian framework to handle model structural error. While these applications span a variety of fields including rainfall-runoff, unsaturated flow and groundwater uranium reactive transport modeling, they all work with time series data. If a formal likelihood function is to be used, the challenge lies in how to configure the form of the likelihood function to be capable of characterizing the distribution of complicated spatiotemporal residual fields of groundwater models. Using a subjective likelihood or the likelihood free ABC could circumvent this difficulty. However, both GLUE and ABC require users' subjective choice to determine model goodness-of-fit and/or whether a particular parameter set is behavioral or non-behavioral. In situations where knowledge about potential model structural error is limited, improper subjective choice may induce bias in the calibration process.

Fortunately, the statistical characterization of model residual can be approached from an inductive, data-driven modeling prospective. A variety of machine learning techniques, such as artificial neural networks and support vector machines, have been successfully applied to build error models that correct for the systematic residual of rainfall-runoff [1, 32, 63, 78]. These machine learning techniques do not require explicit assumption of residual distribution. Instead, they are able to learn complex relations between the dependent variable (i.e. model residual, in the context of error modeling) and selected predictors from historical data. Therefore, they comprise good candidates to statistically characterize residual distribution.

These techniques have been extended to groundwater hydrology to statistically characterize groundwater model residuals, which are usually spatiotemporal and substantially more complicated than time series data [19, 33, 95, 98]. For example, Xu et al. [98] built complementary data-driven error models to account for the epistemic error of groundwater models. By learning from the historical error of the groundwater model, the machine learning algorithms (clustering, support vector regression and instance based weighting) are capable of correcting its bias when the model is used for forecasting or extrapolation purposes. The method was applied to two regional-scale groundwater models that have different hydrogeologic settings, parameterization and calibration methods. In both case studies, the error models significantly improve the prediction accuracy of groundwater head. Xu and Valocchi [97] extended this method to not only reduce the predictive bias of physically-based groundwater models, but also provide prediction intervals. The prediction uncertainty due to the aleatoric component of groundwater model residuals are estimated using both parametric and non-parametric (quantile regression forest) distribution estimation methods. The new method was tested on a real-world groundwater modeling case study. Compared to using only the physically-based groundwater model, the new method provided more accurate monthly baseflow predictions along with prediction intervals with coverage probability consistent with validation data.

In the above mentioned applications of machine learning techniques, error models are constructed in a postprocessing way that the error model is estimated from the residuals of a single calibrated hydrologic model [63, 78, 90]. As repeated evaluation of the physically-based model is not required to construct error models, postprocessor approaches are computationally efficient. However, postprocessor approaches yield statistical error models that are conditioned on an existing calibrated physically-based model; the calibration has been implemented using conventional methods that do not account for correlated error. In this sense, postprocessor approaches ignore interactions between model structural error and parameters [26].

In contrast, a method that jointly infers the error model and the parameters of the hydrologic model can provide a complete assessment of the contribution to predictive uncertainty from parameter, model structural and measurement uncertainty [46, 70, 75]. Kennedy and O'Hagan [46] proposed a Bayesian formulation that allows for explicit treatment of errors in both input data and model structure. In particular, the framework integrates a Gaussian process error model to characterize predictive uncertainty of numerical simulation models. Gaussian process regression belongs to the family of nonparameteric Bayesian kernel models, which have become popular in the machine learning literature in the last decade [11, 52, 67]. In [46], the Gaussian process error model corrects for model structural error revealed by the model residual, thus preventing parameter compensation during the calibration process. The idea of using an error model to absorb model structural error is related to the strategic use of compensatory parameters in highly parameterized calibration [59, 22], but here the error model is not physically-based. The Bayesian approach [46] has inspired applications and extension in various fields [7, 35], including river water quality [70, 20] and rainfall-runoff modeling [38]. The authors also pointed out the link between the Bayesian formulation and multiobjective calibration.

1.2 Main Contributions and Key Findings

This dissertation proposes a fully Bayesian calibration and uncertainty quantification framework tailored for groundwater models. To account for input data variability, the framework implements a marginalizing step when evaluating the likelihood. The framework incorporates error models to explicitly handle errors in model structure and input data, while previous applications of Bayesian approaches in the groundwater modeling literature concentrated on parameter uncertainty [28, 45, 49, 53]. In particular, by integrating the data-driven error modeling technique [19, 98, 97] with Bayesian calibration [46], the framework is capable of

statistically characterizing complex, spatiotemporal structural error of groundwater models. In addition, existing approaches to handling structural error of environmental models [47, 54, 70, 72, 76, 75] focus on time series data, and mostly rely on relatively simple statistical characterization of model residual distribution. This study integrates statistical learning techniques to correct for groundwater model structural error, which is usually spatiotemporal and substantially more complicated.

An important adaptation from [46, 70] is that the error model inputs can include a variety of information including simulation results of the physically-based groundwater model and other relevant data which are not used directly to construct the groundwater model. Using our new method it is therefore possible to extrapolate the error model to predictions under conditions different from the calibration period. In addition, by fully coupling the groundwater model with data-driven error models in a Bayesian formulation, the presented approach facilitates the joint estimation of physically-based model parameter and structural uncertainties. In this way, it extends beyond our preceding studies [98, 97] that constructed error models for already calibrated groundwater models.

The framework is applied to two synthetic and one real-world case studies. By comparing results obtained using conventional calibration techniques, we investigated the impacts of input data and model structural errors on parameter estimates and predictions made by the calibrated model.

We also illustrate through case studies strategies that we developed to address two challenges faced by Bayesian inference, especially for real-world complicated groundwater models. First, it has been noted in the literature that the interactions among different uncertainty sources could render joint inference methods less robust than postprocessor approaches [26]. Second, the computational cost associated with joint inference is often high and could be infeasible for complex models having long evaluation time. We found in the case studies that cautious

specification of error model priors helps alleviate the identifiability issue due to interaction, delivering reasonable uncertainty analysis performance even with a complicated regional groundwater model. To render the Bayesian inference computationally feasible in the real-world case study, we constructed computationally frugal surrogate models to emulate the behavior of the physically-based groundwater model. The surrogates are used in the Bayesian inference process.

The key findings are summarized as follows.

1. We demonstrated through a synthetic case study of surface-ground water interaction under changing pumping conditions, that calibration using biased input data would undermine the quality of parameter estimates and model predictions. Applying the proposed Bayesian approach with input error model, we showed that explicit treatment of errors in model structure and input data (groundwater pumping rate) has substantial impact on the posterior distribution of groundwater model parameters. Using error models reduces predictive bias caused by parameter compensation. In addition, input variability increases parametric and predictive uncertainty. A manuscript based on the results is in preparation for *Journal of Hydrology*.

2. In the second synthetic case study of surface-ground water interaction under changing pumping conditions, we investigated the role of model structural error in calibration and prediction in groundwater flow modeling practice. We first demonstrated that conventional least squares regression (LSR) yields biased (and often overconfident) predictions under a scenario differing from the calibration period. This finding is consistent with others in the literature reporting the deleterious impact of parameter compensation on prediction performance. We then tested the Bayesian framework on the case study and found that Gaussian process error models can represent the underlying model structural error reasonably well, although not perfectly. Integrating error models into Bayesian calibration reduces the degree

of parameter compensation, leading to parameter posteriors that differ substantially from LSR estimates. We also showed that the Bayesian framework with error model achieves more accurate prediction and more robust prediction intervals compared to both LSR and the classical Bayesian inference without error model. The results are published in Xu and Valocchi [96] in *Water Resources Research*.

3. We presented a new recalibration strategy that circumvents the drawback that error models adjusting the physically-based model simulation results may violate mass balance, because such physical constraints are not enforced on the data-driven error model. The recalibration strategy incorporates model structural error into least squares regression by using a full error covariance matrix. It was found in the second case study that the recalibration strategy yields different parameter estimates and more accurate prediction compared to the conventional LSR calibration and Bayesian calibration without error models. The results are published in Xu and Valocchi [96] in *Water Resources Research*.

4. We further tested the Bayesian framework on a real-world case study to calibrate a regional groundwater flow model. The regional model was developed by a multi-institution team and parameters were calibrated by conventional LSR [39]. We use this as the basis for the real-world case study. Efficient implementation strategies are developed to facilitate Bayesian inference. Similarly as in the second synthetic case study, the integration of Gaussian process error models substantially improves the prediction accuracy of the groundwater model when compared to the classical Bayesian calibration without error models. A manuscript based on this case study is in preparation.

1.3 Thesis Outline

The dissertation is organized as follows. Chapter 2 reviews the two general approaches for model calibration and uncertainty analysis, namely least squares regression and the classical Bayesian inference. Chapter 3 proposes the fully Bayesian approach that is capable of handling errors in model structure and input data. This chapter also presents an innovative recalibration strategy that aims to preserve the physical basis of the groundwater hydrology model (i.e. water balance), while allowing for integration with a data-driven error model. Chapter 4 introduces a synthetic case study of river-aquifer interaction under changing pumping conditions. This case study considers the impact of input data error on calibration, and the results of applying the proposed method is presented and compared with results of existing methods (as described in Chapter 2). Chapter 5 presents a second synthetic case study focusing on model structural error. In Chapter 6, the Bayesian approach is further tested on a real-world regional groundwater modeling case study containing structural error. Finally, Chapter 7 concludes the dissertation.

Chapter 2

METHODS OF MODEL CALIBRATION AND UNCERTAINTY ANALYSIS

In this chapter, we briefly review two calibration techniques that are commonly used in surface and groundwater modeling community. We also briefly describe a Monte Carlo Markov Chain (MCMC) sampler used to carry out Bayesian inference.

2.1 Least Squares Regression

Automatic calibration using the method of least squares has been the standard way to determine parameter values of numerical groundwater models in recent decades [37]. Assume that a groundwater system can be represented as

$$z = M(\mathbf{x}, \theta) + \epsilon, \tag{2.1}$$

where $M(\mathbf{x}, \theta)$ is typically a nonlinear, numerical model with input $\mathbf{x}$ and parameters θ, and ϵ denotes residual error. Both $\mathbf{x}$ and θ can be vectors. Both z and M are vectors that represent the system response at various time and locations. Input $\mathbf{x}$ typically includes boundary conditions and stresses, such as pumping. Parameters θ can be e.g. hydraulic conductivity, storativity, dispersivity and other hydrogeologic properties. Given a set of n observations $\mathbf{z} = \{z_i, i = 1, ..., n\}$, standard least squares calibration seeks $\hat{\theta}$ that minimizes the sum of the squares of residual $\sum_{i=1}^{n} r_i^2$, where $r_i = z_i - M_i(\mathbf{x}, \theta)$. Here, i denotes index that differentiates the observations (or model outputs) at different time and locations. The underlying assumption is that the errors r are uncorrelated, have zero mean and constant variance σ^2.

Generalized least squares method relaxes these assumptions and assumes instead that the errors $r_i, i = 1, ..., n$ have a multivariate Gaussian distribution with zero mean and covariance matrix $\sigma^2\Sigma_r$. The parameter estimate $\hat{\theta}$ is obtained by minimizes the weighted sum of squared residual

$$\min_{\theta} \phi(\theta) = \min_{\theta} \mathbf{r}^{\mathbf{T}}\Sigma_r^{-1}\mathbf{r}/\sigma^2 \tag{2.2}$$

where $\mathbf{r}$ is the residual vector with ith element as $r_i = z_i - M_i(\mathbf{x}, \theta)$.

Based on parameter estimates, parameter confidence intervals and linear prediction intervals can then be derived. First assume that $M(\mathbf{x}, \theta)$ is the correct model for the mean of z, i.e. $E(z) = M(\mathbf{x}, \theta)$. Here θ indicates the real but unknown value of parameters. In addition, the nonlinear model $M(\mathbf{x}, \theta)$ is approximated by first-order Taylor series expansion for θ near $\hat{\theta}$:

$$M_i(\mathbf{x}, \theta) \approx M_i(\mathbf{x}, \hat{\theta}) + \sum_{j=1}^{p} J_{ij} \cdot (\theta_j - \hat{\theta}_j), \tag{2.3}$$

where $J_{ij} = \frac{\partial M_i(\mathbf{x},\theta)}{\partial \theta_j}\big|_{\theta=\hat{\theta}}$ is element ij of the Jacobian matrix J, and p is the number of parameters. The parameter covariance matrix is then given by

$$C = \hat{\sigma}^2[J^T\Sigma_r^{-1}J]^{-1} \tag{2.4}$$

where $\hat{\sigma}^2 = \phi(\hat{\theta})/(n-p)$ is the estimate of the error variance σ^2. The $100(1-\alpha)\%$ linear confidence interval for an individual parameter θ_j is then calculated as

$$\hat{\theta}_j \pm t(n-p, 1.0-\alpha/2)\sqrt{C_{jj}} \tag{2.5}$$

where $t(n-p, 1.0-\alpha/2)$ is the Student t-statistic for $n-p$ degrees of freedom and a significance level of α. Prediction of interest can be computed using the calibrated model, i.e. $\hat{z}_i^* = M(\mathbf{x}_\mathbf{i}^*, \hat{\theta})$. The covariance matrix for multiple predictions $z_i^*, i = 1, ..., k$ is given by

$$C^* = J_*^T C J_* \tag{2.6}$$

where J_* is the prediction Jacobian matrix with element $J_{ij} = \frac{\partial z_i^*}{\partial \theta_j}\big|_{\theta=\hat{\theta}}$. The $100(1-\alpha)\%$ linear prediction interval is

$$\hat{z}_i^* \pm t(n-p, 1.0-\alpha/2)\sqrt{C_{ii}^* + \hat{\sigma}^2}. \tag{2.7}$$

Calibration of groundwater models typically assumes that r_i are uncorrelated observation errors, and omits off-diagonal elements of Σ_r. One exception was mentioned in Section 1.1 where a full error covariance matrix was used to account for correlation in observations computed from a set of direct measurements [80].

While least squares regression theory requires weights assigned as the inverse of observation error, it is often considered appropriated in practice to designate weights to ensure approximately equal goodness of fit of various types of outputs [39, 23]. In addition, the strategy of regularized calibration of highly parameterized models recommends assigning weights for observations according to their resemblance to predictions of interest [23]. As a result, the parameter confidence interval and linear prediction interval derived using these weights lack sound statistical foundation and often depend on subjective weighting decisions.

2.2 Bayesian Calibration and Prediction

In this section we overview the classical Bayesian calibration and prediction using the system defined in Eqn. (2.1). More details can be found in [41, 83]. In a Bayesian framework, the system output $\mathbf{z}$, inputs $\mathbf{x}$ and parameters θ are random variables.

The goal of Bayesian calibration is to infer the posterior distribution of parameters θ conditioned on available observations $\mathbf{z} = (z_1, ...z_n)^T$. Let $p(\mathbf{z}|\theta)$ denote the joint density of observations $\mathbf{z}$ conditioned on θ. This is usually referred to as the likelihood function, denoted as $L(\theta|\mathbf{z})$. It is commonly assumed that errors ϵ_i follow multivariate Gaussian distribution

$N(\mathbf{0}, \sigma_\epsilon^2 \Sigma)$. In this case, log likelihood is given by:

$$\log L(\theta|\mathbf{z}) = -\frac{n}{2}\log 2\pi - \log \sigma_\epsilon - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathbf{r}^{\mathbf{T}}\Sigma^{-1}\mathbf{r}, \tag{2.8}$$

where $\mathbf{r}$ is the residual vector with ith element as $r_i = z_i - M_i(\mathbf{x}, \theta)$. The term $\mathbf{r}^{\mathbf{T}}\Sigma^{-1}\mathbf{r}$ corresponds to the weighted sum of squared error. This highlights the relation between Bayesian calibration and least squares regression [53]. Often ϵ_i are considered i.i.d. observation errors, leading to $\Sigma = I_n$. In recent years, cases where ϵ_i are correlated and/or non-Gaussian have been considered. For example, Lu et al. [54] used a full error covariance (i.e. Σ has nonzero off-diagonal elements) in Equation (2.8) to account for error correlation due to model structure uncertainty when simulating column experiments of uranium reactive transport. Schoups and Vrugt [75] proposed a generalized likelihood function to handle non-Gaussian errors in hydrologic models.

Bayesian calibration allows integration of other source of information about the parameters via the prior distribution, $p(\theta)$. Normal, log-normal and uniform priors are widely used [83]. The prior and the likelihood are combined using Bayes' theorem to provide the posterior distribution $p(\theta|\mathbf{z})$:

$$p(\theta|\mathbf{z}) = \frac{p(\theta)p(\mathbf{z}|\theta)}{\int p(\theta)p(\mathbf{z}|\theta)d\theta} \propto L(\theta|\mathbf{z})p(\theta). \tag{2.9}$$

The posterior distribution can be used to derive point and interval estimate of parameters. In particular, the maximum a posteriori probability (MAP) estimate of θ is defined as the mode of the posterior distribution $p(\theta|Z)$. The posterior distribution of parameters is also useful in practical applications where the model is used to predict quantities of interest. The probability density function of prediction z^* conditioned on available observations $\mathbf{z}$ can be computed using

$$p(z^*|\mathbf{z}) = \int p(z^*|\theta, \mathbf{z})p(\theta|\mathbf{z})d\theta. \tag{2.10}$$

2.3 Markov Chain Monte Carlo Sampling

While Bayesian calibration and prediction have simple formulations, analytical solutions of Equations (2.9) and (2.10) are often intractable if nonconjugate prior distributions are used or the integral is high dimensional. The Markov chain Monte Carlo (MCMC) method is the most widely used numerical approximation technique for Bayesian calibration and prediction. The MCMC method is based on the assumption that a Markov chain $\theta^{(n)}$ with states θ can be constructed such that its stationary distribution is equal to the posterior distribution $p(\theta|\mathbf{z})$ of interest. This algorithm starts from arbitrary value of θ, then iteratively generates trial moves from the current position $\theta^{(k)}$ to a new state $\theta^{(k+1)}$ until practical convergence is judged [41].

Various MCMC samplers have been proposed with different strategies of generating new states. Among these samplers, the Differential Evolution Adaptive Metropolis (DREAM) algorithm developed by Vrugt et. al [85, 87] has been shown to be effective and computationally efficient in a variety of environmental modeling applications, and is therefore used in this study. Let $\theta_i, i = 1, ..., N$ denote N (a sufficiently large number) samples from $p(\theta|\mathbf{z})$ by a MCMC sampler. Each θ_i is a vector of p parameters. According to strong law of large numbers, $g(\theta)$, the conditional expectation of an arbitrary function of θ, can be approximated by

$$E[g(\theta)|\mathbf{z}] \cong \frac{1}{N}\sum_{i=1}^{N} g(\theta_i). \tag{2.11}$$

Point and interval estimates of prediction of interest z^* can be computed using the above approximation.

Chapter 3

BAYESIAN APPROACH FOR MODELS WITH STRUCTURAL AND INPUT DATA ERRORS

In Sections 3.1-3.3 we derive the fully Bayesian approach with input data and model structural error models. Section 3.4 describes the surrogate modeling strategy to reduce computational cost associated with Bayesian inference. Finally in Section 3.5 we propose a recalibration strategy that utilizes the Bayesian inference results while preserving physical constraints such as mass balance.

3.1 Statistical Description of Input Data Error

Consider a groundwater system defined in Eqn. (2.1) where z is the quantity of interest that can be observed, M denotes a model with inputs $\mathbf{x}$ and parameter θ, and ϵ is measurement error. Both z and M can be vectors that denote the system output at various time and locations. Input $\mathbf{x}$ typically includes boundary conditions and stresses. In modeling practice, some inputs (e.g. river stage) are measured with relatively high accuracy. However, some input forcings are not measured, but estimated indirectly from relevant information. For example, in the Republican River Compact Administration (RRCA) model [58], the irrigation pumping rates were estimated based on irrigation acreage, farm efficiency, crop water requirement among other information. We compared the estimated total annual pumping rate at county level with metered pumping rate for three counties (Perkins, Chase and Dundy) in Nebraska from 1980 to 2006 (data courtesy of Nicholas Brozovic, written communication). It was found that the estimated annual pumping rate is 1%-26% lower than the metered pumping rate. Besides pumping rate, groundwater recharge is a well-known input forcing

that is hard to estimate accurately. As further discussed in Chapter 4, we will focus on groundwater pumping and recharge rates. In this section, we omit the accurate inputs from Equation (2.1) and use $\mathbf{x}$ to denote input data that may contain error.

As discussed in Section 1.1, uncertainty in input data has been studied in surface water literature within a Bayesian framework. For rainfall-runoff models, rainfall error is typically considered as the primary source of input uncertainty. The existing approach is to introduce a series (e.g. 100) of rainfall "multipliers" to adjust the measured rainfall rate of each storm event; the multipliers are then jointly sampled with model parameters (which are constant over time) using a MCMC sampler based on calibration data [40, 43, 44, 85]. However, assigning multipliers to each rainfall event results in hundreds to thousands of parameters, depending on the duration of simulation. The resulting high dimensionality poses computational challenges to both least squares regression and Bayesian inference. Moreover, nonuniqueness or nonidentifiability issues arises from correlation among model input, parameters and output. When input errors are substantially larger than output errors, calibrating input multipliers essentially conditions the input on the output, often leading to nearly perfect fit to calibration data [40]. In rainfall-runoff modeling, it was argued that inferring the posterior of rainfall event multipliers is of less interest than inferring model parameters. Since true rainfall input is almost never available in modeling practice, validation of the inferred multiplier is unattainable [2, 40].

The nonidentifiability issue tends to be more severe in groundwater modeling problems. For example, it would take a great number of multipliers to describe recharge rate that varies temporally and spatially, or to describe pumping rates and each individual wells. In groundwater modeling practice, recharge is sometimes calibrated while pumping rate is normally fixed during calibration.

In this study, we take a different approach based on the observation that input data can

often be estimated from other sources of information with small to medium degree of uncertainty. For example, recharge rate is sometimes calculated using a surface hydrology model, such as PRMS [55] and SWAT [3]. As the simulation results are usually validated by comparing with streamflow observations, the model simulated recharge can be considered as a reasonable estimate with somewhat confidence. A similar treatment of input data error, albeit with different motivation and implementation, can be found in [2] for rainfall-runoff modeling.

It is assumed *a priori* that analysis of the estimation method can provide a distribution of the true input conditioned on the estimated value, denoted as $p(\mathbf{x}|\hat{\mathbf{x}})$. A reasonable choice would be the normal distribution $\mathbf{x}|\hat{\mathbf{x}} \sim N(\hat{\mathbf{x}}, \sigma_{\mathbf{x}}^2)$. The underlying assumption is that the estimated input $\hat{\mathbf{x}}$ represents modelers' best *a priori* knowledge, and there is no indication that the estimate is biased. If any bias is suspected, the estimate should be adjust to eliminate the bias. The conditional distribution of real input given an estimate comprises the input data model, and is analogous to the measurement error model that $\epsilon \sim N(0, \sigma_\epsilon^2)$.

Next, the probability of observing $\mathbf{y}$ given the true input $\mathbf{x}$ and parameters θ is given by

$$\mathbf{y}|\mathbf{x}, \theta \sim N(M(\mathbf{x}, \theta), \sigma_\epsilon^2 I_n), \tag{3.1}$$

where I_n is a n-by-n identity matrix. In order to account for the uncertainty associated with $\mathbf{x}$, we derive the *marginal* likelihood:

$$L(\theta|\mathbf{y}, \hat{\mathbf{x}}) = p(\mathbf{y}|\hat{\mathbf{x}}, \theta) = \int p(\mathbf{y}|\mathbf{x}, \theta) p(\mathbf{x}|\hat{\mathbf{x}}) d\mathbf{x}. \tag{3.2}$$

Because of the integration step to derive the marginal likelihood in the above equation, the presented method is referred to as the marginalizing method hereafter. Following Equation

(3.2), the posterior distribution of parameters θ can be written as

$$p(\theta|\hat{\mathbf{x}}, \mathbf{y}) \propto \int p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}|\hat{\mathbf{x}})d\mathbf{x} \cdot p(\theta). \tag{3.3}$$

Let $\mathbf{x}^*$ denote uncertain inputs in the prediction scenario, and assume that $\mathbf{x}^*|\hat{\mathbf{x}}^* \propto N(\hat{\mathbf{x}}^*, \sigma^2_{\mathbf{x}^*})$, where $\hat{\mathbf{x}}^*$ denotes the estimated inputs. The distribution of a prediction y^* can then be inferred:

$$p(\mathbf{y}^*|\hat{\mathbf{x}}, \hat{\mathbf{x}}^*, \mathbf{y}) = \int p(\mathbf{y}^*|\theta, \hat{\mathbf{x}}, \mathbf{x}^*)p(\theta|\hat{\mathbf{x}}, \mathbf{y})p(\mathbf{x}, \mathbf{x}^*|\hat{\mathbf{x}}, \hat{\mathbf{x}}^*)d\theta d\mathbf{x} d\mathbf{x}^*. \tag{3.4}$$

The dependency of the prediction $\mathbf{y}^*$ on the input during the calibration period, $\hat{\mathbf{x}}$, is because the model is often used to make forecast in a future scenario, using the simulation results in the calibration period as initial condition.

3.2 Statistical Description of Model Structural Error

Proper treatment of model structural error is critical for calibration and subsequent prediction uncertainty analysis. As discussed in Section 1.1, model structural error has been approached by adopting a generalized likelihood function in a Bayesian framework [75, 73], by constructing an error covariance matrix based on autoregression analysis [54], and by paired complex/simple model methods based on linear subspace analysis [22, 91]. The existing approaches rely on either a simple statistical characterization (e.g. autoregression) of model residual, or the assumption that the model structural error can be described as a linear function of environmental models' parameters.

Kennedy and O'Hagan [46] first proposed a fully Bayesian calibration framework that can handle errors from multiple sources, in particular, from model structural inadequacy. In their formulation, model structural error is expressed as an additive term:

$$z = M(\mathbf{x}, \theta) + b(\mathbf{x}) + \epsilon, \tag{3.5}$$

where $\mathbf{x}$ would be the location and time corresponding to z. Because systematic error is accounted for by the model structural error term, the remaining ϵ can be considered as random measurement error. In their benchmark paper, Kennedy and O'Hagan [46] proposed to place a Gaussian process prior on $b(\mathbf{x})$. This Bayesian framework was successfully implemented in river water quality modeling [20, 70] and rainfall-runoff modeling [38] in a time series context.

In this study, we integrate the data-driven error modeling technique [98] into the Bayesian calibration framework [46] to develop a new framework that is tailored for large-scale geoscience models with structural error and input variability. The new formulation models an observation z as

$$z = M(\mathbf{x}, \theta) + b(\mathbf{y}, \phi) + \epsilon, \tag{3.6}$$

where the model structural error term $b(\mathbf{y}, \phi)$ is represented as a function of its own inputs $\mathbf{y}$ and tuning parameters ϕ. An important adaptation from the original framework in [46] is that, the error model input $\mathbf{y}$ may consist of the physically-based model's output $M(\mathbf{x}, \theta)$ and other relevant information in addition to time and location of quantity of interest. This allows for assimilating data that are not used directly to construct model M, therefore making it possible to extrapolate to conditions different from the calibration period [98, 97].

Non-parametric Bayesian kernel regression methods, such as Gaussian process (GP) as used in [46], are good candidates for the statistical error model. Gaussian process regression [11, 67] has been shown to achieve remarkable performance in a variety of benchmark applications. Because of this, and its compatibility with Bayesian calibration and prediction principles, GP is selected to construct the error model in this study. While the formulation below is illustrated with GP, it should be noted that the proposed framework does not exclude other types of Bayesian kernel methods [52, 64, 77].

A brief introduction to Gaussian process regression is included here. More details can be

found in [11, 67, 92]. In the following overview, we adapt conventional notations to be consistent with Equation (3.6). A Gaussian process refers to a set of random variables $\{b(\mathbf{y})|\mathbf{y} \in R^d\}$ ($\mathbf{y}$ is a $d-$dimensional vector) for which any finite set of $\{\mathbf{b}\}$ has a joint multivariate Gaussian distribution. A GP is fully specified by its mean function $\mu(\mathbf{y}) = E[b(\mathbf{y})]$ and covariance function $k(\mathbf{y}, \mathbf{y}') = E[(b(\mathbf{y}) - \mu(\mathbf{y}))(b(\mathbf{y}') - \mu(\mathbf{y}'))]$. In this study, we consider two simple mean functions: constant zero $\mu(\mathbf{y}) = 0$ and linear $\mu(\mathbf{y}) = \beta^T\mathbf{y}$. We use a popular category of covariance function that takes the *squared exponential* form [67]:

$$k(\mathbf{y}, \mathbf{y}') = \sigma^2 \exp\left[-\sum_{l=1}^{d} \frac{(y_l - y_l')^2}{\lambda_l^2}\right]. \tag{3.7}$$

In Equation (3.7), σ^2 controls the marginal variance of $b(\mathbf{y})$, and $\lambda_1, ...\lambda_d$ control the dependence strength in each of the component directions of $\mathbf{y}$. For the *isotropic* squared exponent covariance function, $\lambda_1 = \lambda_2 = ... = \lambda_d = \lambda$, and λ is usually referred to as the *characteristic length scale*. In the geostatistics literature, σ^2 and $\lambda_1, ...\lambda_d$ are often called *sill* and *range*, respectively. In Equation (3.6), the parameter vector ϕ consists of these tuning parameters in the covariance function and the mean function, i.e. $\phi = \{\beta, \lambda, \sigma^2\}$.

Specifying a Gaussian process prior on $b(\mathbf{y}, \phi)$, the prior distribution of $\mathbf{b}$ would be a multivariate Gaussian distribution $N(\mu(\mathbf{y}, \phi), \Sigma(\phi))$. The covariance matrix Σ is calculated using the specified covariance function, and its ij-th entry is $\Sigma_{i,j} = k(\mathbf{y}_i, \mathbf{y}_j)$. Note that both the mean and the covariance depend on ϕ. For the sake of conciseness, hereafter in this section we will omit the conditioning on ϕ. Equation (3.6) can be re-arraged into $z - M(\mathbf{x}, \theta) = b(\mathbf{y}, \phi) + \epsilon$. Therefore, $z - M(\mathbf{x}, \theta)$ can be considered as noisy observations of $\mathbf{b}$. Let $\{Y, \mathbf{z} - \mathbf{M}\} = \{(\mathbf{y}_1, z_1 - M(\mathbf{x}_1, \theta)), ..., (\mathbf{y}_n, z_n - M(\mathbf{x}_n, \theta))\}$ denote a set of n training data; $\mathbf{z}$ and $\mathbf{M}$ are vectors representing, respectively, observations and physically-based model outputs at different locations and time, and Y is a n by d (input dimension) matrix. The measurement error ϵ can be considered as white noise with variance σ_ϵ^2. It follows that

the log *marginal likelihood* of observations $\mathbf{z}-\mathbf{M}$ is given as [67]:

$$\log p(\mathbf{z}-\mathbf{M}|Y) = -\frac{1}{2}(\mathbf{z}-\mathbf{M}-\mu)^T\left(\Sigma+\sigma_\epsilon^2 I\right)^{-1}(\mathbf{z}-\mathbf{M}-\mu)-\frac{1}{2}\log|\Sigma+\sigma_\epsilon^2 I|-\frac{n}{2}\log 2\pi, \quad (3.8)$$

where μ is the prior mean vector and I is the n by n identity matrix. The first term evaluates the goodness-of-fit, the second term is the complexity penalty, and the last term is a normalization constant.

Based on the training data $\{Y, \mathbf{z}-\mathbf{M}\}$, predictions can be made for new input $Y^* = \{\mathbf{y}_j^*, j = 1, ..., m\}$, i.e. to estimate the probability density of $b(Y^*)$, abbreviated to $\mathbf{b}^*$, conditioned on training data. Similarly as Σ, define Σ^*, Σ^{**} such that $\Sigma_{i,j}^* = k(\mathbf{y}_i, \mathbf{y}_j^*)$ and $\Sigma_{i,j}^{**} = k(\mathbf{y}_i^*, \mathbf{y}_j^*)$. We first write out the *a priori* joint distribution of $\mathbf{z}-\mathbf{M}$ and $\mathbf{b}^*$:

$$\begin{bmatrix} \mathbf{z}-\mathbf{M} \\ \mathbf{b}^* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu^* \end{bmatrix}, \begin{bmatrix} \Sigma+\sigma_\epsilon^2 I & \Sigma^* \\ \Sigma^{*T} & \Sigma^{**} \end{bmatrix}\right), \quad (3.9)$$

The posterior distribution of $\mathbf{b}^*$ conditioned on training data can therefore be derived [67]:

$$\mathbf{b}^*|\mathbf{z}-\mathbf{M}, Y, Y^*, \phi \sim N\left(\bar{\mathbf{b}^*}, cov(\mathbf{b}^*)\right). \quad (3.10)$$

The posterior mean $\bar{\mathbf{b}^*}$ and covariance $cov(\mathbf{b}^*)$ are given below:

$$\bar{\mathbf{b}^*} = E\left[\mathbf{b}^*|\mathbf{z}-\mathbf{M}, Y, Y^*, \phi\right] = \mu^* + \Sigma^{*T}\left(\Sigma+\sigma_\varepsilon^2 I\right)^{-1}(\mathbf{z}-\mathbf{M}-\mu), \quad (3.11)$$

$$cov(\mathbf{b}^*) = \Sigma^{**} - \Sigma^{*T}\left(\Sigma+\sigma_\varepsilon^2 I\right)^{-1}\Sigma^*. \quad (3.12)$$

In Gaussian process regression, assumptions about the target function are imposed via specifying a prior probability distribution over a family of possible functions. The prior is then "sculpted" into a posterior using observation data. This feature and the use of the covariance function give GP more flexibility compared to parametric regression methods that restrict the class of functions. In this study, Gaussian process inference is implemented using GPML

MATLAB toolbox version 3.4 documented in [67].

3.3 Bayesian Calibration and Prediction with Error models

This section briefly reviews the Bayesian framework to handle model structural error, and then discusses numerical implementation details. Complete derivation of Bayesian inference can be found in [46, 70].

In the fully Bayesian framework, the physically-based model parameters θ will be jointly estimated with model structural error $b(\mathbf{y}, \phi)$. This allows for a complete assessment of uncertainty from parameter and model structure. Bayesian calibration starts from specifying the prior distribution of parameters $\{\theta, \phi\}$. In general they are independent unless there is evidence otherwise. According to Bayes' theorem,

$$p(\theta, \phi|\mathbf{z}) \propto p(\mathbf{z}|\theta, \phi)p(\theta)p(\phi), \tag{3.13}$$

where $p(\mathbf{z}|\theta, \phi)$ is the likelihood and can be calculated using Equation (3.8). Calculating the posterior $p(\theta, \phi|\mathbf{z})$ is typically analytically intractable, so sampling techniques such as Markov chain Monte Carlo (MCMC) algorithms are often used to sample from the posterior. In this study, we use DREAM-ZS (DiffeRential Evolution Adaptive Metropolis algorithm) [75].

Once sufficient samples $\{\theta_i, \phi_i\}, i = 1, ..., N$ are generated using MCMC, Bayesian inference of prediction uncertainty can be carried out. For every sample $\{\theta_i, \phi_i\}$:
1. Use θ_i to run the MODFLOW model in prediction mode to obtain $\mathbf{M}_i^*$.
2. Conditioning on ϕ_i, compute the covariance matrices $\Sigma, \Sigma^*, \Sigma^{**}$ using Equations (3.7). The MODFLOW model output $\mathbf{M}_i$ and $\mathbf{M}_i^*$ will be used in this step if they are included in the input of GP error model.

3. Generate one realization $\mathbf{b}_i^*$ from the posterior of the error model using Equations (3.10) – (3.12). Note that $\mathbf{b}_i^*$ is conditioned on calibration residuals, which have been calculated during the calibration phase.

4. Compute $\mathbf{z}_i^* = \mathbf{M}_i^* + \mathbf{b}_i^* + \epsilon_i$, where ϵ_i is a vector comprised of random draws based on inferred measurement error.

Finally, the posterior mean of predictions is given by $\bar{\mathbf{z}^*} = \sum_{i=1}^{N} \mathbf{z}_i^*$. Here, $\bar{\mathbf{z}^*}, \mathbf{z}_i^*$ are vectors of predictions at various locations and time. Predictive quantiles $\mathbf{z}_{\alpha/2}^*, \mathbf{z}_{1-\alpha/2}^*$ corresponding to a specified confidence level α can be derived by sorting $\mathbf{z}_i^*, i = 1, ...N$.

3.4 Numerical Implementation

In Bayesian inference, sampling from the posterior distribution requires tens to hundreds of thousands of model evaluation. For computationally intensive groundwater models, the computational cost of MCMC sampling may be prohibitively high. Under such situations, computationally frugal surrogate models can be used as a substitute for the original model when evaluating the likelihood. Surrogate models can be constructed from the original model by reducing numerical resolution, increasing tolerance and/or omitting processes [5]. However, the parameters of reduced-order models may not be defined exactly the same as in the original model, making the inference of parameter posterior less straightforward. In contrast, response surface methods attempt to statistically mimic the relationship between explanatory variables (i.e. model parameters) and response variable(s). For example, [60] used a radial basis function to approximate the calibration objective, such as the Nash-Sutcliffe index. The resulting response surface was then updated in the optimization approach. Similarly, Gaussian process regression, a machine learning algorithm, was used to emulate the response surface of logarithm of likelihood for the calibration of a rainfall-runoff model [89]. In addition to calibration objective function, response surface methods can also emulate model state variables, or outputs, such as groundwater head. Recently, generalized

polynomial chaos expansion (gPC) [56] and sparse grid methods have been used to construct surrogate models in various hydrology applications including groundwater modeling [49, 99, 100].

As further described in Chapter 6.5, surrogate modeling was implemented on a real-world case study with realistic degree of complexity. The surrogate models take as inputs the parameters to be inferred, and output the original model's simulation results. The surrogate models were constructed based on Support Vector Regression (SVR) [84]. SVR has been applied to many fields including rainfall-runoff modeling [68], radioactive soil contamination [42] and groundwater hydrology [4, 98]. The SVR algorithm has good generalization performance, because it seeks to minimize an upper bound of the generalization error rather than minimize the training error. The solution of SVR is globally optimal, while many other statistical learning tools (e.g. artificial neural network) may converge to local minima.

This section briefly overviews ε-SVR as will be used in Chapter 6.3. Given a set of training data $\{\mathbf{x}_i, y_i\}, i = 1, ..., n$, where $\mathbf{x}_i$ denotes input and y_i denotes output that has been observed, the idea of SVR is to first project input $\mathbf{x}$ to a higher dimensional feature space by the map $\Phi : \mathcal{X} \to \mathcal{F}$, and then carry out a linear regression of y in the feature space $\Phi(\mathbf{x})$:

$$f(\mathbf{x}) = w \cdot \Phi(\mathbf{x}) + b. \tag{3.14}$$

The coefficients $\mathbf{w}$ and b are estimated by solving the following optimization problem

$$\textit{minimize} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3.15a}$$

$$\textit{subject to} \quad (\mathbf{w}^T\Phi(\mathbf{x}_i) + b) - y_i \leq \varepsilon + \xi_i, \tag{3.15b}$$

$$y_i - (\mathbf{w}^T\Phi(\mathbf{x}_i) + b) \leq \varepsilon + \xi_i^*, \tag{3.15c}$$

$$\xi_i, \xi_i^* \geq 0, \ i = 1, ..., n. \tag{3.15d}$$

The first term in Eq. (3.15a) represents the complexity of the regression model and therefore acts as regularization. The second term represents goodness-of-fit to training data; the slack variables ξ_i, ξ_i^* are introduced to cope with otherwise infeasible constraints of the optimization problem. They are derived from the ε-insensitive loss function $|\xi|_\varepsilon = max\{0, |y_i - f(\mathbf{x}_i)| - \varepsilon\}$. The constant C in Eqn. (3.15a) determines the trade-off between the flatness of f and deviations exceeding ε.

In general, the map $\Phi : \mathcal{X} \to \mathcal{F}$ is implemented implicitly via *kernel functions*. This study adopts the commonly used *radial basis function* (RBF) kernel:

$$\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j),$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2). \tag{3.16}$$

The regularization hyperparameter C is chosen according to the training data following the recommendations of Cherkassky and Ma [15]. The loss function hyperparameter ε and kernel width hyperparameter γ are tuned by five-fold cross validation. The LIBSVM toolbox [14] is used to implement ε-SVR.

3.5 Recalibration Strategy

As discussed in Sections 1.1 and 2.1, calibration of groundwater models typically uses a diagonal error covariance matrix because the correlation structure caused by model structural error is usually unknown. We propose a "recalibration" strategy that utilizes, in a data assimilation fashion, the model structural error inferred by the fully Bayesian approach.

As described in Section 2.2, the Bayesian prediction process will provide a size N ensemble of predictions of interest $\mathbf{z}_i^* = \mathbf{M}_i^* + \mathbf{b}_i^* + \epsilon_i, i = 1, ..., N$. The mean $\bar{\mathbf{z}}^* = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{M}_i^* + \mathbf{b}_i^*)$

will be used as calibration targets in the recalibration process. The calibration targets can be weighted according to measurement error. Another option is to use a full error covariance matrix in Equation (2.2) to account for residual correlation due to model structural error. A matrix Σ_b representing model structural error can be conveniently computed from the Bayesian posterior realizations $\mathbf{b}_i^*, i = 1, ..., N$.

In other words, the groundwater model will be calibrated against predictions made by the Bayesian approach, which is based on calibration data that have been observed and (possibly weak) prior knowledge on model parameters and structural error. Therefore, the recalibration strategy does not require any additional information that is not available in the model construction and calibration phase. The recalibrated model is then used to make forecast along with associated linear prediction interval.

The advantage of the recalibration strategy is twofold. First, the mean prediction given by the Bayesian approach is expected to be more accurate than that given by the initially LSR calibrated model. This is because the Bayesian approach involves the error-correcting Gaussian process error model; the recalibrated model should fit the Bayesian posterior mean of prediction reasonably well, and therefore is expected to yield more accurate prediction while preserving mass balance and other physical constraints inherent in a physically-based model. The recalibration strategy is analogous to smoothing in the context of data assimilation, only that the Bayesian prediction, rather than true observation, is used. Recalibration is also related to the strategic use of compensatory parameters [23]. By recalibrating the model using Bayesian predictions as targets, we are allowing parameters to be over-adjusted in order to compensate for model structural error. As discussed in the Introduction section, for predictions that are under similar conditions as calibration data, parameter compensation may improve predictive accuracy [22]. Second, Tiedeman and Green [80] showed that ignoring error correlation can have substantial effect on parameter estimates, predictions and associated uncertainty. The recalibration strategy utilizes the residual correlation structure

estimated by the Bayesian approach. Therefore, it is expected that more realistic parameter estimates can be achieved via recalibration than using the conventional LSR with a diagonal error covariance matrix.

Chapter 4

SYNTHETIC CASE STUDY WITH INPUT DATA ERROR

In this chapter we describe a synthetic case study used to investigate the impact of input data error on calibration and prediction and test the performance of the proposed Bayesian approach. The hypothetical case study uses a virtual reality to represent realistic hydrogeologic conditions that are common in the field and serves to generate synthetic observations. Meanwhile, we build a working simplified model that represents the limited knowledge modelers would possess about the virtual reality. The working model (hereafter referred to as "model") is calibrated against the synthetic observations generated by the virtual reality and subsequently used to make forecast under changing scenarios.

The synthetic case study simulates the effect of pumping on two-dimensional groundwater flow in an unconfined aquifer that is hydraulically connected to a stream. There have been many documented cases where pumping-induced groundwater piezometric head decline and stream depletion lead to water right conflicts and/or threaten ecosystem services [6, 74]. Numerical models are increasingly being used to support conjunctive regulation of groundwater and surface water resources [50, 58, 88].

4.1 Synthetic Models

Both the virtual reality and the model are transient single-layer MODFLOW2000 models that simulate an unconfined aquifer with impermeable bottom and surrounding no-flow boundaries. Both models have 50×50 grid cells of size $200m \times 200m$, and hence the model

domain spans 10 by 10 km^2 (Figure 4.1). The virtual reality runs for 6 years with monthly stress step and weekly time step. The simple model has the same monthly stress period, but has only one time step for every stress period (monthly time step). The virtual reality has irregular bottom elevation and non-permeable boundaries. The model has linearly inclined (north to south, sides to stream location) bottom elevation and straight surrounding boundaries.

Specific yield is homogeneous in both the virtual reality and the model. For the virtual reality, specific yield equals 0.2; the natural log conductivity field was generated using a sequential Gaussian simulation code SGeMS [71], with a mean of 30 m/day and a sill of 1 (for natural logarithm $\ln K$). An anisotropic spherical variogram was used, with range 6 km in the east-west direction and 4 km in the north-south direction. In the field of groundwater hydrology, it is a common practice to assume that the logarithm of hydraulic conductivity (K) follows a normal distribution and geostatistically represent the natural spatial variability of K via a covariance function or variogram (e.g., [22, 29, 91]). In the simple model, on the other hand, the log conductivity field was interpolated from 12 pilot points (location shown in Figure 4.5), using Ordinary Kriging and a spherical variogram with a range of 4 km and a sill of 2. The variogram used in the model is different from the variogram used in the virtual reality, reflecting inaccurate prior knowledge of the spatial correlation structure of logarithm of K. Similar implementation can be found in other studies, such as [22, 91]. The log conductivity values at pilot points will be calibrated.

The stream is modeled using the MODFLOW SFR1 package [66], and the stream stage is routed by Manning's Formula at each time step. In both models, Manning's n is set to 0.03, and the streambed slope is 0.0005. A rectangular streambed cross section is used, and the channel width is 14 m. The streambed hydraulic conductivity is uniform throughout the whole reach. Seasonally varying inflow is specified at the inlet at the north boundary in the virtual reality. For the model, the inflow is generated by perturbing the inflow in the virtual

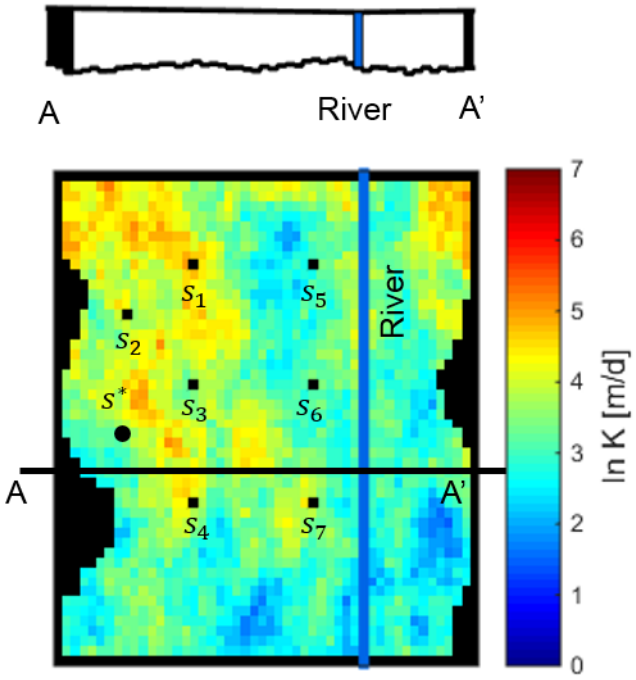

Figure 4.1: Modeling domain and cross section showing the unconfined unit with a stream running from north to south as simulated by the virtual reality. Locations of drawdown calibration targets $s_1, ..., s_7$ and validation data s^* are shown. Color encodes the natural logarithm hydraulic conductivity field of the virtual reality.

reality, assuming that the streamflow measurement has a coefficient of variation (CV) of 0.01.

While necessarily restricted by use of a specific complex numerical model to represent reality, the case study can nevertheless provide insights into the potential of the presented approach to handle errors in common types of forcings.

4.2 Input data

This synthetic case study considers two types of uncertain input data: groundwater pumping and precipitation recharge. As discussed in Section 3.1, pumping rates are often not me-

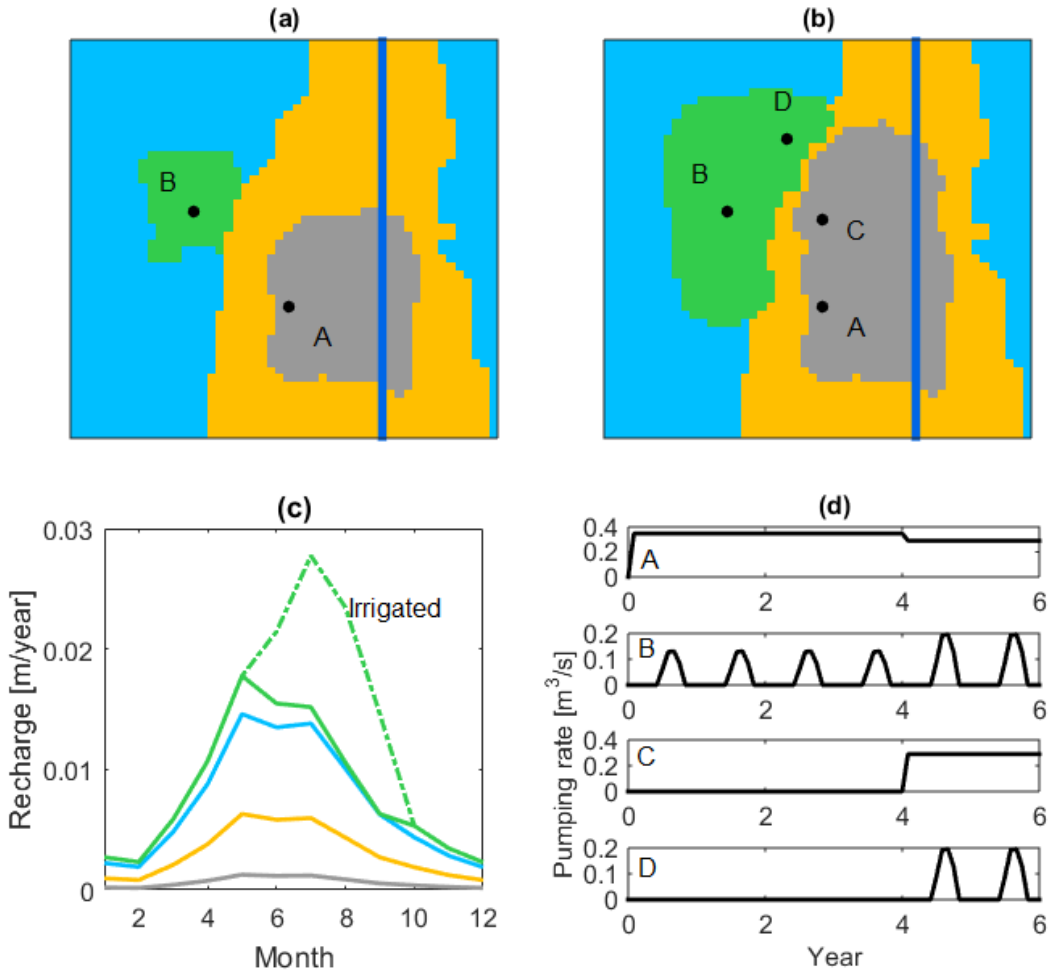

Figure 4.2: Recharge zones and location of pumping wells during the calibration (a) and validation (b) periods. The recharge rates for each zone are shown in (c) and pumping rates at four wells shown in (d).

tered, but indirectly estimated from related information such as power usage and irrigation requirements. Precipitation recharge rate can be calculated using an assumed penetration ratio, using a surface water model, or estimated as model parameters through calibration.

The case study simulates four pumping wells; their locations are marked in Figure 4.2. Among the four wells, B and D are irrigation wells and are turned on during the growth season. On the other hand, A and C are municipal supply wells and are pumped at a constant rate. Wells A and B starts pumping from the first transient stress period. Wells C and D start pumping from the 5th year. The pumping rates at four wells are shown in Figure 4.2.

Recharge is specified with four zones (Figure 4.2). Zones 1, 2 and 3 receive recharge from precipitation only. Zone 4 also receives groundwater irrigation return flow during the growth season, which is assumed to equal 20% of total pumping rates in that zone divided by the area. The return flow rate 20% is chosen according to commonly reported irrigation efficiency [58]. Figure 4.2 shows the monthly varying recharge rates for four zones in the virtual reality. The values are specified based on typical recharge condition in the Nebraska portion of the Republican River basin [58].

4.3 Calibration and validation data

The simulation starts from a steady-state stress period with no groundwater pumping, which mimics natural equilibrium state before development. The virtual reality then runs for 6 years and generates quarterly synthetic drawdown (s) (location shown in Figure 4.1) and stream gain-and-loss (ΔQ) observations, both are the most commonly used types of observation when calibrating a groundwater flow model. Drawdown targets are computed by subtracting the groundwater head at a time step from the head at steady state. The stream gain-and-loss (ΔQ) is computed by summing up the cell-by-cell flow exchange rates between the stream and the aquifer cell across the whole reach. A negative value indicates groundwater discharges to stream, and a positive value means stream loses to groundwater. Stream gain-and-loss targets are important to constrain parameters; if only head observations are used in calibration, hydraulic conductivity parameters would often become highly correlated and cannot be uniquely determined.

The synthetic observations in the first 4 years are contaminated with measurement error and used to calibrate the model. The drawdown measurement error is assumed to be independent and Gaussian distributed with zero mean and a standard deviation of 0.02 m.

The streamflow measurement is also independent and Gaussian distributed with zero mean and a coefficient of variation (CV) of 0.01. The streamflow measurement error variance is computed by summing up the variance of upstream inflow and downstream outflow [37]. A relatively low streamflow measurement CV is assumed because the case study is intended to focus on uncertainties other than measurement error. If a more realistic CV value, e.g. 0.05 [39], is used, CV of ΔQ can exceed 100% because ΔQ is small compared to upstream and downstream flow. Synthetic data of the remaining 6 years are reserved for validation. As can be seen from Section 4.2, the validation period represents an increased groundwater demand scenario that is substantially different from the calibration period.

4.4 Experiments, Results and Discussion

To investigate the impact of inaccurate input data on uncertainty analysis, we carried out three sets of experiments with the synthetic case study as described below.

4.4.1 Experiment A: Benchmark case with true inputs

In experiment A, we use the classical Bayesian method to calibrate the model with "true" recharge and pumping rates. The result of this experiment will serve as the benchmark to which results from experiments B and C will be compared. As can be seen from section 4.1, there exist a few differences between the virtual reality and the model including the modeling domain geometry and the specification of aquifer hydraulic conductivity. Therefore, it is not straightforward to compare the estimated values of model parameter with the "true" value in the virtual reality.

The synthetic data during the first 4 years were used to calibrate 16 parameters, namely the specific yield (S_y), natural logarithm of the hydraulic conductivity of the streambed

($\ln K_{rb}$) and at locations given by the pilot points ($\ln K_1, ..., \ln K_{12}$), the drawdown measurement error standard deviation (σ_s), and the stream gain-and-loss measurement coefficient of variation ($CV_{\Delta Q}$). Relatively vague prior distributions are specified for all parameters as shown in Table 4.1. Specifically, the joint prior distribution of $\ln K_1, ..., \ln K_{12}$ is specified as a multivariate normal distribution with a mean of 4.1 and a covariance matrix Σ_K. The covariance matrix is computed using the variogram used to interpolate logarithm hydraulic conductivity from pilot points (a spherical variogram with a range of 4 km and a sill of 2). As mentioned in Section 4.1, this variogram is different from the true anisotropic variogram used to generate the $\ln K$ field in the virtual reality. In Table 4.1, the mean of prior distributions are chosen to be different from the true value to reflect inaccurate prior knowledge. For $\ln K$, the "true value" refers to the $\ln K$ values in the virtual reality corresponding to locations given by the pilot points and is shown in Figure 4.1.

Table 4.1: Prior distributions of calibrated parameters and assumed distribution of inputs. Please see Section 4.4 for explanation.

Notation	Unit	Distribution
S_y	m	$N(0.18, 0.036^2)$
$\ln K_{rb}$	m/d	$N(0.69, 0.69^2)$
$[\ln K_1, ..., \ln K_{12}]^T$	m/d	$N\left([4.1, ..., 4.1]^T, \Sigma_K\right)$
$CV_{\Delta Q}$	-	Uniform on $[0.0001, 0.5]$
σ_s	m	Uniform on [0.0001,0.5]
Q_A	m^3/d	$N(\hat{Q}_A, (0.2\hat{Q}_A)^2), \hat{Q}_A = 1.2Q_{A,0}$
Q_B	m^3/d	$N(\hat{Q}_B, (0.2\hat{Q}_B)^2), \hat{Q}_B = 1.2Q_{B,0}$
$R_i, i = 1, ..., 4$	mm/y	$N(\hat{R}_i, (0.25\hat{R}_i)^2), \hat{R}_i = 0.75R_{i,0}$
$\lambda_i, i = 1, ..., 4$	-	$N(1, 0.25^2)$

In this study, we used DREAM-ZS (DiffeRential Evolution Adaptive Metropolis algorithm), a Markov chain Monte Carlo sampler developed in [75] to sample from the posterior distributions of parameters. The DREAM-ZS runtime settings were configured following the recommendations in [85]. Three Markov chains were used to generate 15,000 samples from the joint posterior distribution of 16 parameters after convergence was determined based on the $\hat{R}$ statistic of [30], visual inspection of trace plots and other diagnostics [18]; about 40,000

model evaluations were required to converge (burn-in). The marginal posterior distributions of specific yield and natural logarithm streambed hydraulic conductivity are shown in Figure 4.3 and Figure 4.4, respectively. The hydraulic conductivity (K) field interpolated from the MAP estimates at the 12 pilot points is shown in Figure 4.5.

Due to model structural error (i.e. the discrepancy between the model and the virtual reality), the estimated parameters deviate from the real values. The results in experiment A represents the best results that can be obtained given the model structural error as true inputs are used. The results will serve as benchmark, and results from experiments B and C will be compared with the benchmark to assess the performance of different calibration strategies.

In the prediction phase, the model is run repeatedly using the posterior samples for the whole simulation period of 6 years. The posterior mean and 95% credible intervals of drawdown at three locations (Figure 5.1) and stream gain-and-loss are shown in Figures 4.6 - 4.9. Due to model structural error, predictive bias can be observed for drawdown and stream gain-and-loss. The root-mean-square-error (RMSE) of the calibrated model prediction in years 5-6 is listed in Figures 4.6 - 4.9. For all drawdown, the RMSE is significantly greater than the standard deviation of measurement error 0.02 m.

4.4.2 Experiment B: Biased pumping rates

In experiment B, it is assumed that the pumping rates to be used in the model were estimated with bias. For illustration purpose, the pumping rate of wells A and B are overestimated by 20% of the true values (Figure 4.1) through the calibration period. Using the same parameter priors as in section 4.4.1 and Table 4.1, the model is calibrated using the classical Bayesian and the marginalizing methods, respectively, resulting in two calibrated models, $M_{B,1}$ and $M_{B,2}$. For the marginalizing method (Section 3.1), it is assumed that the pumping

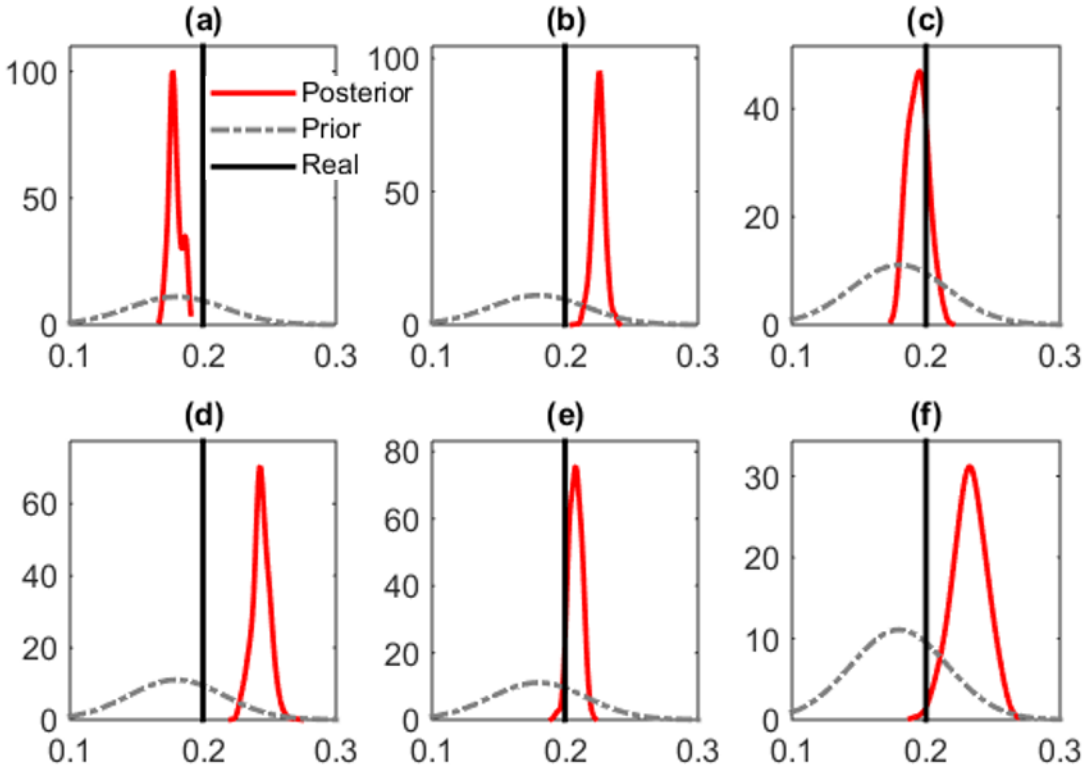

Figure 4.3: Posterior distribution of specific yield S_y of the calibrated models M_0 (a), $M_{B,1}$ (b), $M_{B,2}$ (c), $M_{C,1}$ (d), $M_{C,2}$ (e) and $M_{C,3}$ (f).

rates Q_A and Q_B are normally distributed around the estimated value. More specifically, $Q_A \sim N(\hat{Q_A}, (0.2\hat{Q_A})^2), Q_D \sim N(\hat{Q_B}, (0.2\hat{Q_B})^2)$, where $\hat{Q_A}$ and $\hat{Q_B}$ are the estimated pumping rates and equal to $1.2Q_{A,0}$ and $1.2Q_{B,0}$, respectively (Table 4.1).

Both the classical Bayesian and marginalizing calibration methods were implemented via DREAM-ZS. After convergence, posterior samples of model parameters are collected to run the model in forecast mode. For both methods, true pumping rates (as used in virtual reality) are used during the validation period (years 5 to 6). This is consistent with common modeling practice that uses a groundwater model to make forecast under prescribed future condition. The method can be easily extended to handle prediction under an uncertain future scenario, e.g., climate change and population projection, by marginalizing over the input in the prediction period. During the first 4 years, the classical Bayesian method used

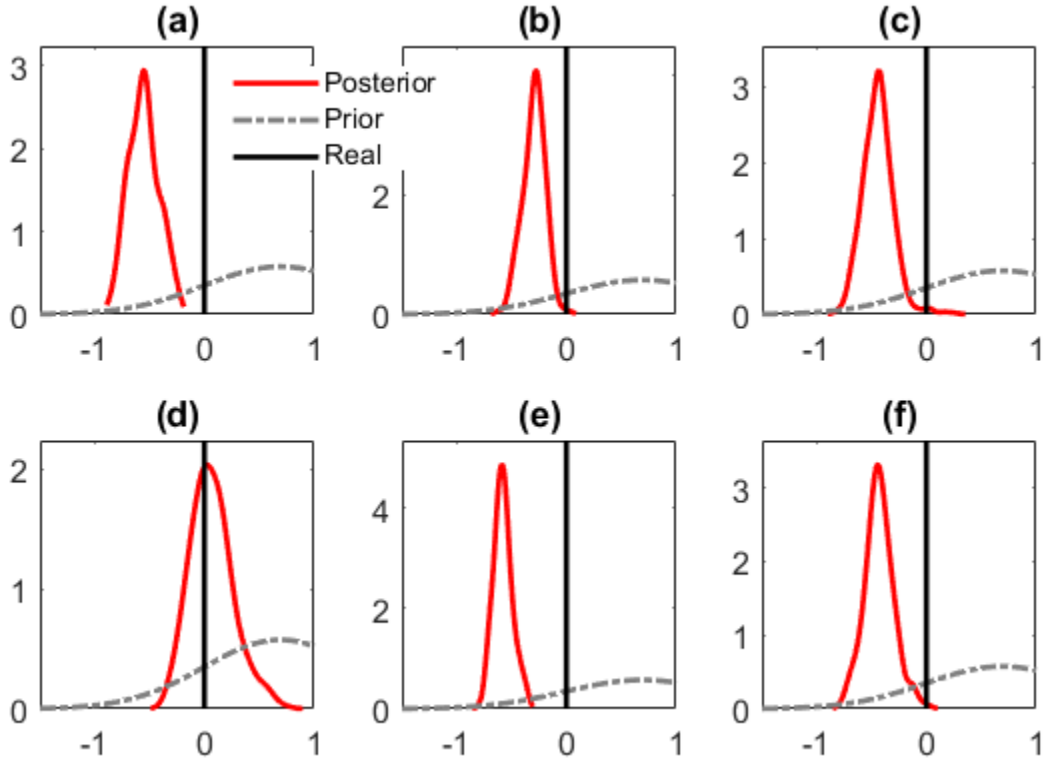

Figure 4.4: Posterior distribution of natural logarithm riverbed hydraulic conductivity $\ln K_{rb}$ of the calibrated models M_0 (a), $M_{B,1}$ (b), $M_{B,2}$ (c), $M_{C,1}$ (d), $M_{C,2}$ (e) and $M_{C,3}$ (f).

the estimated pumping rates $\hat{Q_A}$ and $\hat{Q_B}$, whereas the marginalizing method propagates the uncertainty in calibration pumping rates via marginalizing.

The marginal posterior distributions of specific yield and natural logarithm streambed hydraulic conductivity are shown in Figure 4.3 and Figure 4.4, respectively. Comparing Figure 4.3 (b) and (c), and Figure 4.4 (b) and (c), it can be seen that the marginalizing method yielded flatter, wider posteriors for both S_y and $\ln K_{rb}$, because of the propagation of input uncertainty to posterior parametric uncertainty. For S_y, the posterior given by the marginalizing method is closer to the benchmark results (Figure 4.3a). The hydraulic conductivity (K) field interpolated from the MAP estimates at the 12 pilot points is shown in Figure 4.5. Compared with the benchmark results in (a), both (b) and (c) capture the overall pattern of higher K in the northwestern part and lower K on the south. The inference of spatial

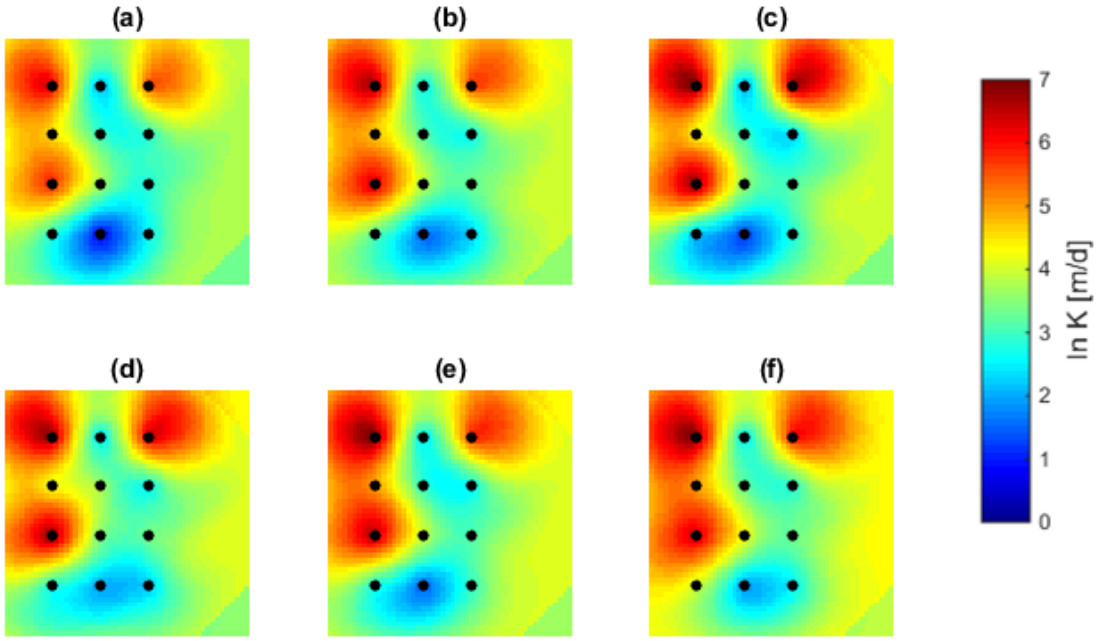

Figure 4.5: Natural log hydraulic conductivity field of the calibrated models M_0 (a), $M_{B,1}$ (b), $M_{B,2}$ (c), $M_{C,1}$ (d), $M_{C,2}$ (e) and $M_{C,3}$ (f). Locations of the 12 pilot points are shown. The K fields are interpolated according to the maximum a posterior (MAP) estimate of pilot points.

variation of K is likely to be primarily controlled by drawdown observations and is insensitive to bias in pumping and recharge rates.

The posterior mean and 95% credible intervals of drawdown at three locations (Figure 4.1) and stream gain-and-loss are shown in Figures 4.6 - 4.9. It can be seen from Figures 4.6 and 4.7 that the calibration error of M_0 and $M_{B,1}$ are both small, while slight bias can be observed for $M_{B,2}$ in year 1-4 at s_1 and s_6. This is because the marginalizing step (Equation (3.2)) leads to a likelihood with inflated variance term and is therefore more tolerant to systematic bias. Despite small calibration error, the classical Bayesian calibrated model $M_{B,1}$ shows significant prediction bias in years 5-6 at s_1 and s_6 (and also other locations that are not shown here), as well as at s^*, a monitoring well not used for calibration (Figure 4.8). In contrast, the marginalizing method yields significantly less biased head predictions.

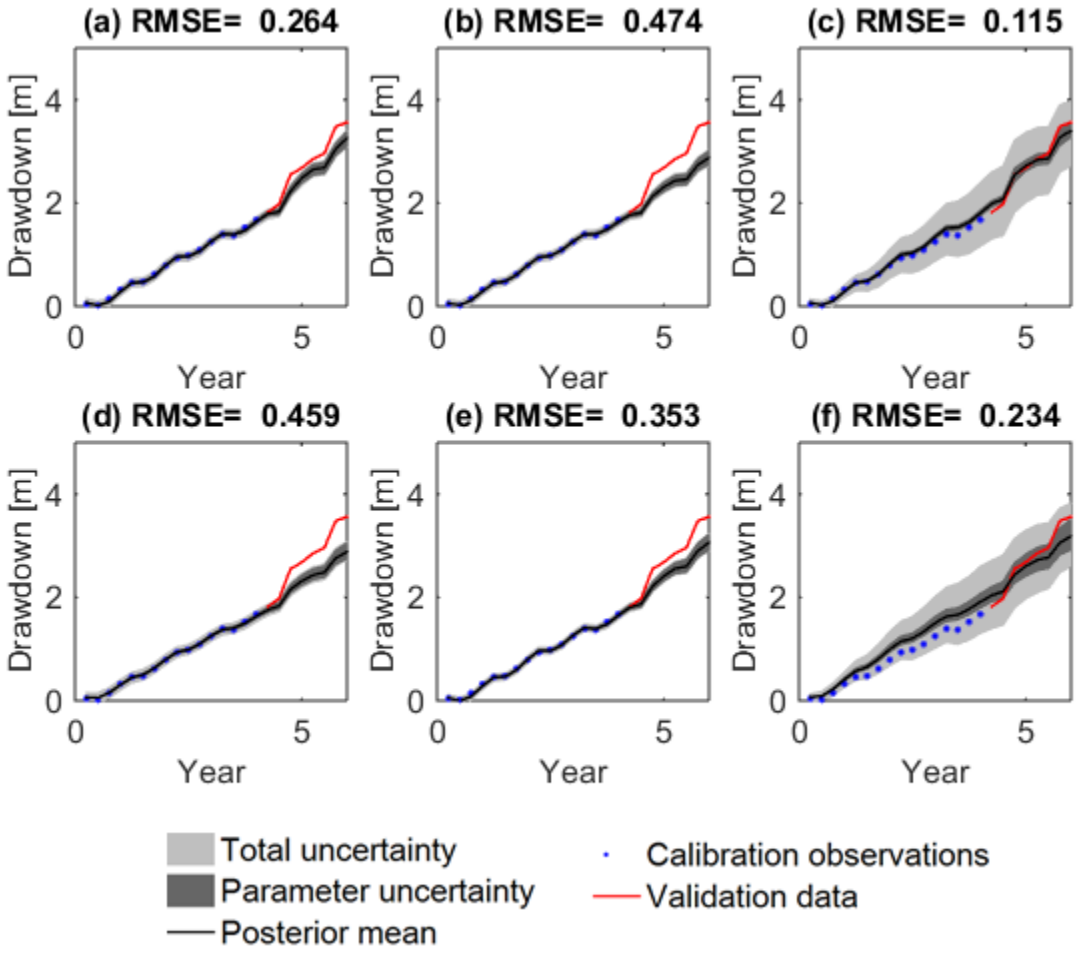

Figure 4.6: Simulation results of drawdown s_1 using the calibrated models M_0 (a), $M_{B,1}$ (b), $M_{B,2}$ (c), $M_{C,1}$ (d), $M_{C,2}$ (e) and $M_{C,3}$ (f).

As for stream gain-and-loss predictions, both methods deliver similar performance with the classical Bayesian approach yielding slightly lower RMSE. As will be further discussed in Section 4.4.3, the marginalizing method assumes that the true pumping rate follows a normal distribution centered around the biased estimated value. Stream gain-and-loss is an important component in the water budget of this case study, and is sensitive to bias in pumping rates.

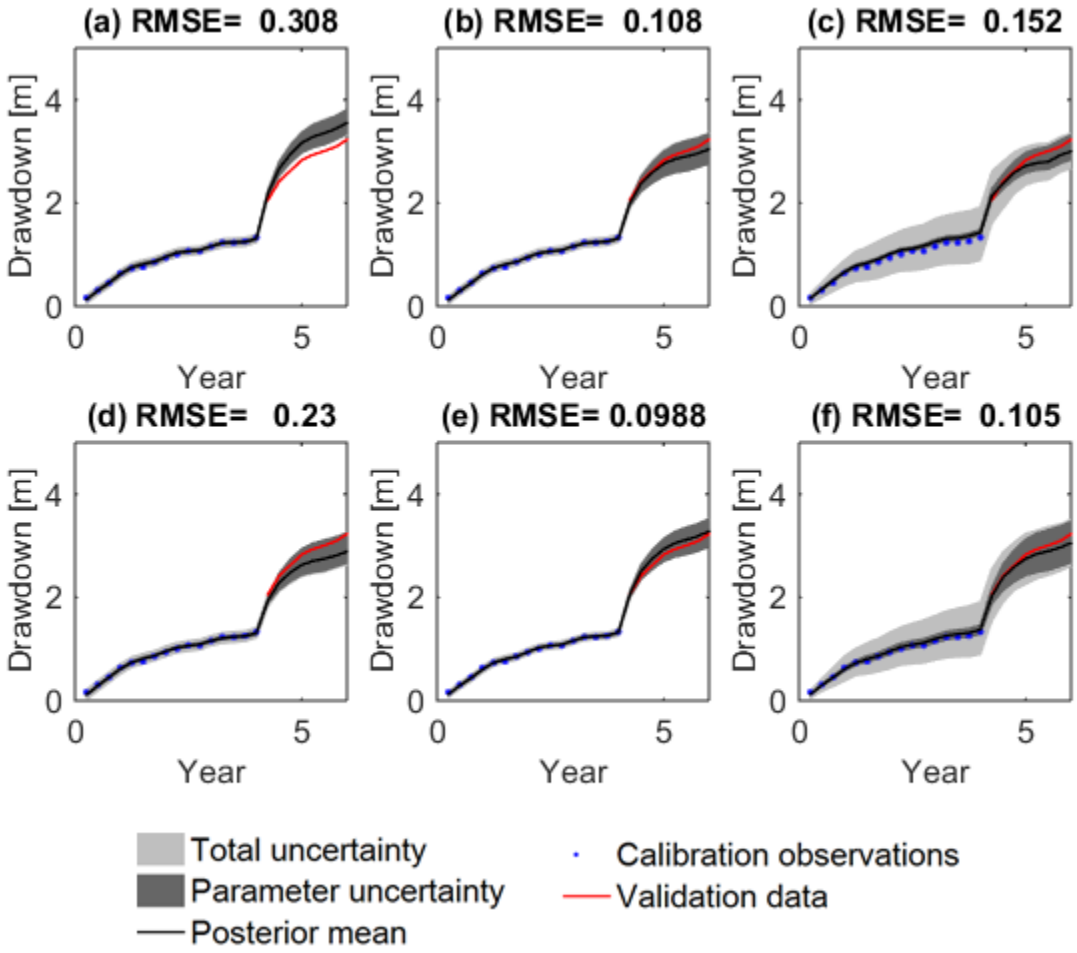

Figure 4.7: Simulation results of drawdown s_6 using the calibrated models M_0 (a), $M_{B,1}$ (b), $M_{B,2}$ (c), $M_{C,1}$ (d), $M_{C,2}$ (e) and $M_{C,3}$ (f).

4.4.3 Experiment C: Biased pumping and recharge rates

Experiment C considers the case that during the calibration period the pumping rates to be used in the model are overestimated, and the recharge rates are underestimated. More specifically, $\hat{Q_A} = 1.2Q_{A,0}, \hat{Q_B} = 1.2Q_{B,0}, \hat{R_i} = 0.75R_{i,0}, i = 1, ..., 4$. In this situation, the bias from overestimation of pumping rate and the bias from underestimation of recharge rate cannot cancel off, thus is expected to induce parameter compensation. The estimated values of pumping and recharge rates were used as input in classical Bayesian calibration, leading to calibrated model $M_{C,1}$. A second calibration strategy was implemented for experiment C that introduces four recharge multipliers to be calibrated along with the specific

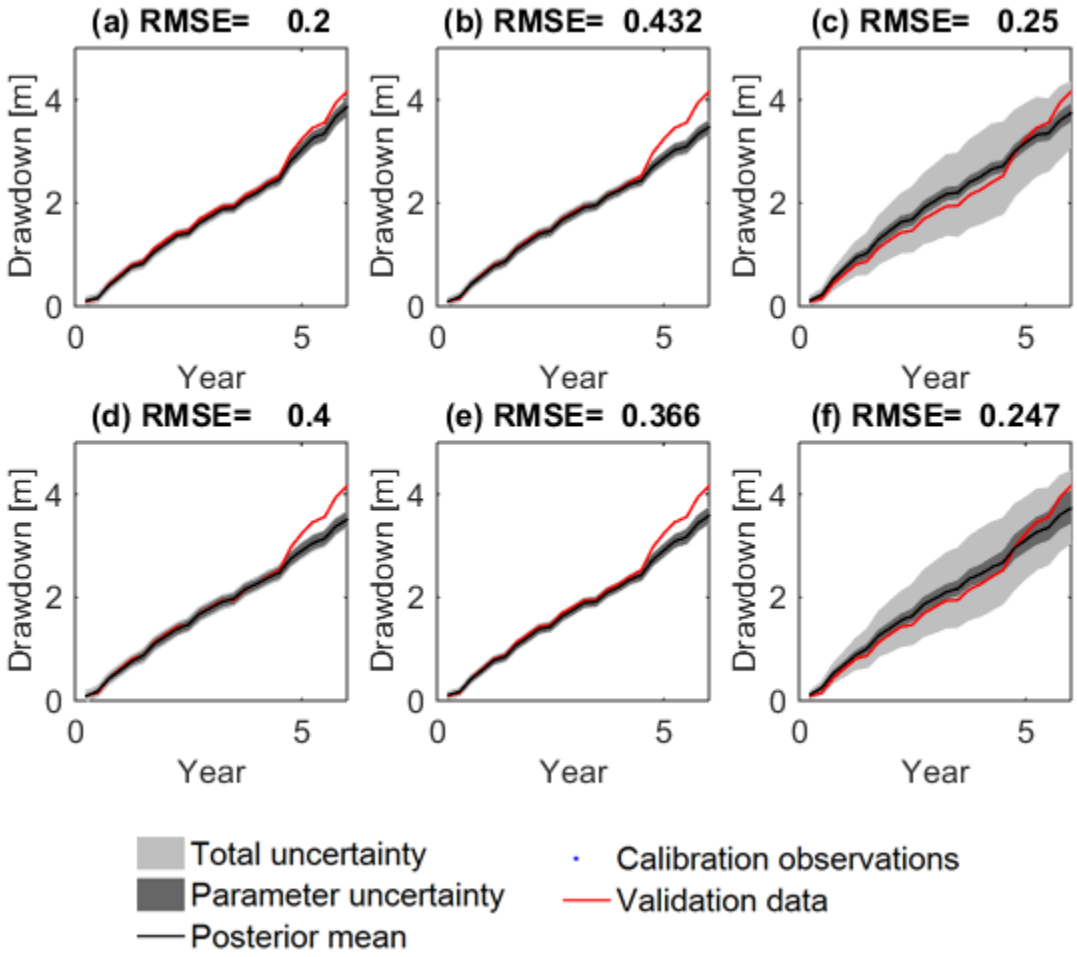

Figure 4.8: Prediction of drawdown s^* using the calibrated models M_0 (a), $M_{B,1}$ (b), $M_{B,2}$ (c), $M_{C,1}$ (d), $M_{C,2}$ (e) and $M_{C,3}$ (f).

yield, hydraulic conductivities and measurement error parameters. The recharge multipliers $\lambda_i, i = 1, ..., 4$ are defined as the ratio of the recharge in a zone over the estimated recharge rate at that zone, i.e. $R_i = \lambda_i \hat{R}_i, i = 1, ..., 4$. The prior marginal distributions of $\lambda_i, i = 1, ..., 4$ are listed in Table 4.1. In total 20 parameters will be calibrated. This strategy will be referred to as "augmentation" method in the remaining part of this chapter, and the resulting model is denoted by $M_{C,2}$. The augmentation strategy is included in Experiment C because sometimes recharge rates are adjusted during calibration in groundwater modeling practice. Finally, as the third calibration strategy, the marginalizing method assumes that the pumping and recharge rates follow normal distributions as listed in Table 4.1 and gives

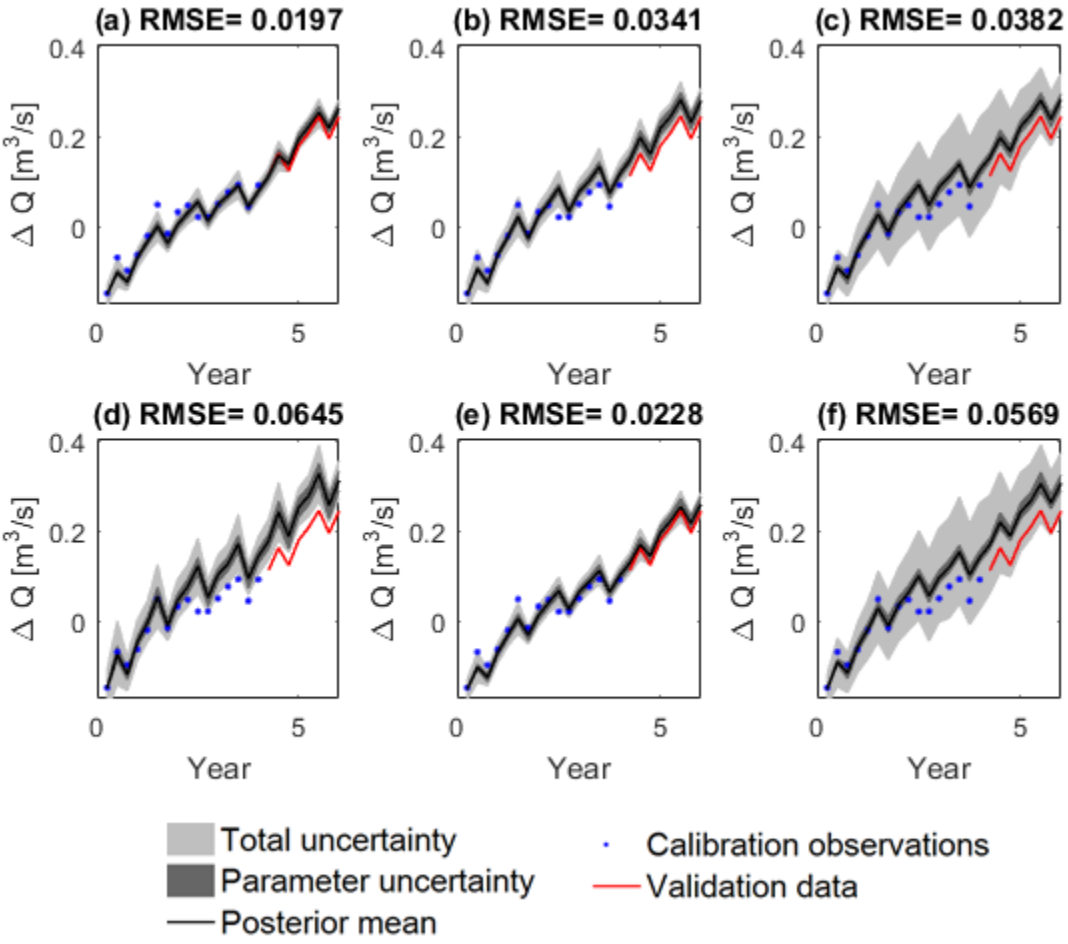

Figure 4.9: Simulation results of stream gain-and-loss ΔQ using the calibrated models M_0 (a), $M_{B,1}$ (b), $M_{B,2}$ (c), $M_{C,1}$ (d), $M_{C,2}$ (e) and $M_{C,3}$ (f).

model $M_{C,3}$.

In the prediction phase, the calibrated models $M_{C,1}, M_{C,2}, M_{C,3}$ were run and produce forecast in the validation period (years 5-6) under specified pumping rates (Figure 4.2). While biased estimated pumping rates $\hat{Q}_A, \hat{Q}_B$ were used for years 1-4 (the calibration period), true pumping rates $\hat{Q}_{A,0}, ...\hat{Q}_{D,0}$ were used in years 5-6 (the validation period). On the other hand, biased estimated recharge rates $\hat{R}_i, i = 1, ..., 4$ were used throughout the 6 years of simulation period for the classical Bayesian and marginalizing methods. The augmentation method used the posterior distributions of recharge multipliers $\lambda_i, i = 1, ..., 4$. The

marginalizing method propagates the uncertainty in calibration pumping rates via marginalizing. The reason of such implementations is that, in groundwater modeling practice, the calibrated model is often used to make forecast with a prescribed future demand such as pumping rate. Precipitation recharge, however, is often estimated via a set of infiltration rates based on land use or soil type; once determined, these infiltration rates will also be used in the prediction period. Note, however, that these implementation details do no affect the major conclusions drawn in the chapter.

The marginal posterior distributions of specific yield and natural logarithm streambed hydraulic conductivity are shown in Figure 4.3 and Figure 4.4, respectively. Comparing Figure 4.3(b) with (d), it can be seen that the posterior of S_y given by the classical Bayesian method becomes more biased when recharge rates are biased in addition to the pumping rates. This is anticipated with the overestimation of pumping rate and underestimation of recharge rate. With biased pumping and recharge rates, the calibration process tries to "fill in" the missing water, while matching the observed drawdown generated by the virtual reality under true pumping and recharge rates. Therefore, the specific yield Sy is overestimated. Similarly, classical Bayesian calibration gives streambed conductivity $\ln K_{rb}$ that is higher than the benchmark result (Figure 4.4(a)); higher k_{rb} produces more inflow from stream to the aquifer to make up for the missing water. The marginalizing method and in particular the augmentation strategy yielded S_y and $\ln K_{rb}$ posteriors that are closer to the benchmark results. However, it can be seen from Figure 4.4(e) that the posterior distribution of $\ln K_{rb}$ given by augmentation method is slightly overconfident compared to (a).

The hydraulic conductivity (K) field interpolated from the MAP estimates at the 12 pilot points is shown in Figure 4.5. The K fields estimated by the three calibration strategies are similar with each other. Similarly as in Section 4.4.2, the inference of spatial variation of K is likely to be primarily controlled by drawdown observations and is insensitive to bias in pumping and recharge rates.

The posterior mean and 95% credible intervals of drawdown at three locations (Figure 4.1) and stream gain-and-loss are shown in Figures 4.6 - 4.9. Similarly as in section 4.4.2, it can be seen that despite relatively small calibration error of $M_{C,1}$ and $M_{C,2}$, the prediction made by the two models has significant bias. In contrast, the marginalizing method yields significantly less biased predictions during the validation period at calibration target wells s_1, s_6 and another monitoring well s^*. As for stream gain-and-loss ΔQ, Figure 4.9 shows that the classical Bayesian calibrated model $M_{C,1}$ yields biased and overconfident prediction; the marginalizing method yields smaller bias, while the augmentation method achieves the most accurate prediction. The main reason is that the augmentation method resulting in reasonable calibrated recharge rates for the two zones to which the stream gain-and-loss are most sensitive. However, when model structural error is present in addition to input data error, it is likely that the augmentation method may overly adjust inputs to compensate for model structural error. For example, the augmentation method produces biased and overconfident drawdown prediction at s_1 and s^*, possibly due to the compensation effect.

It is worth mentioning that for the marginalizing method, the recharge rates are assumed to follow a normal distribution centered around the biased estimates. Stream gain-and-loss is an important component in the water budget of this case study, and is sensitive to bias in pumping and recharge rates. The marginalizing method does not correct for the input data bias, but accounts for the possibility that the true input deviates from the estimated value. We recognize this as the limitation of the marginalizing method. As can be seen in Figure 4.9(f) the Bayesian posterior prediction is biased, while the 90% prediction interval still encompasses the validation data. We also note that the prediction interval given by the marginalizing method depends on the quality of the statistical model specified *a priori* to characterize input data error. Low confidence in the estimated input data will result in prediction intervals that are too wide to be informative. The marginalization method is suitable when input data are estimated with low to medium level of uncertainty.

When prior estimation of input data is highly uncertain and significant bias is likely to exist, an alternative is to jointly infer the hyperparameters of the input data error model during the calibration process. For example, we can assume that $Q_B \sim N\left(\mu\hat{Q}_B, (\beta\hat{Q}_B)^2\right)$ and jointly sample μ, β with other parameters. In this way, the alternative method can correct for possible bias in input data. This approach will be further investigated in followup studies.

4.5 Summary

We demonstrated the Bayesian approach through a synthetic case study of surface-ground water interaction under changing pumping and land use conditions. It is found that explicit treatment of errors in input data (groundwater pumping and recharge rates) has substantial impact on the posterior distribution of groundwater model parameters. Using statistical models to explicitly account for input error reduces predictive bias caused by parameter compensation.

Compared to classical Bayesian results, the marginalizing approach yields more accurate predictions. However, one limitation of the marginalizing approach is that it only recognizes the uncertainty associated with input data, but does not correct for potential input bias. As a result, for quantities that are highly sensitive to biased input data, such as streamflow gain-and-loss in this case study, the prediction made by the marginalizing approach may still be biased, although the prediction interval still encompasses validation data.

The marginalizing method results indicate that input variability increases parametric and predictive uncertainty, in contrast with the augmentation result. This is because the marginalizing approach does not update the assumed input distribution, while the augmentation method calculates a posterior input distribution by combining the prior with likelihood. In

this sense, we recommend the marginalizing approach to be used for situations in which (1) substantial knowledge is available to specify a reasonable input distribution, and (2) augmentation method may not work due to identifiability issues.

Finally, in this case study the input uncertainty is dominant among various sources of uncertainty. In general, the Bayesian approach allows for a comparison among the contributions from various error sources, which could inform future model improvement and data collection efforts on how to best direct resources towards reducing predictive uncertainty.

Chapter 5

SYNTHETIC CASE STUDY WITH MODEL STRUCTURAL ERROR

In this Chapter we present a second synthetic case study in which we examine the impact of model structural error on uncertainty analysis and test the performance of the Bayesian approach with error model and the recalibration strategy. The synthetic case study is based on the one described in Chapter 4 after modifications so that it represents a situation where model structural error is non-negligible. The model is designed to suffer from common types of model inadequacy (or imperfection) in groundwater modeling practice. The presented framework can be applied to other types of model inadequacy beyond those represented in the case study. Calibration and prediction are implemented using standard least squares, standard (classical) Bayesian, the proposed Bayesian approach and the recalibration strategy. Performance of these methods is evaluated and compared. The materials presented in this chapter are based on Xu and Valocchi [96].

5.1 Synthetic Models

In the second synthetic case study, we modified the virtual reality and the working model used in the first case study in Chapter 4 to introduce model structural errors. For conciseness this section focuses on the changes made.

For both the virtual reality and the model, the west and east boundaries are assigned as impermeable, and the north and south boundaries are general head boundaries (GHB) for which the flux is proportional to the difference between cell head and specified boundary

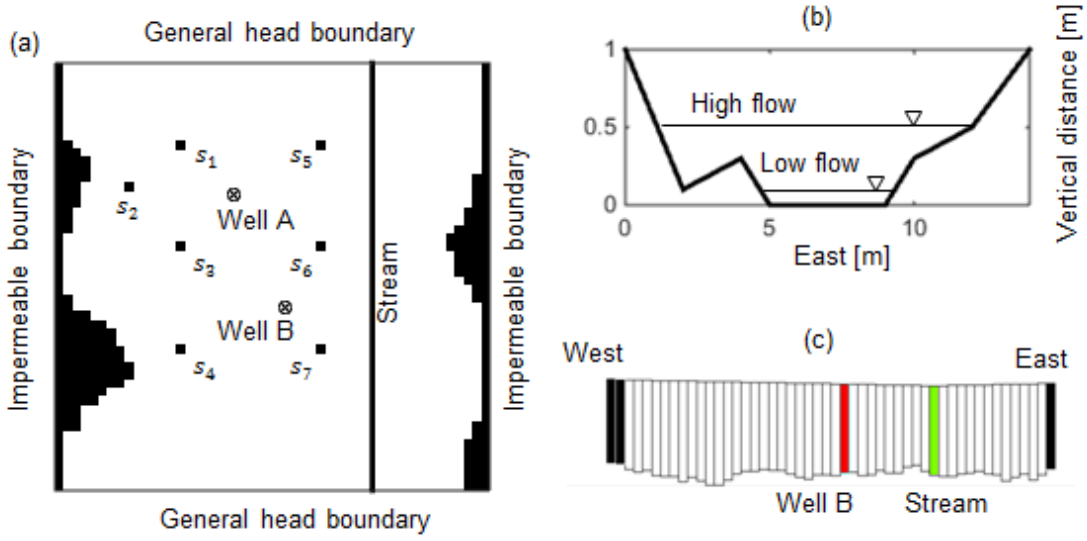

Figure 5.1: (a) Modeling domain showing the unconfined unit with a stream running from north to south. Squares show locations of drawdown calibration targets. Circles with inside cross show the two pumping wells. (b) Stream cross section of the complex model. (c) Aquifer cross section showing irregular bottom elevation.

head (Figure 5.1a). Figure 5.2 shows the difference between the two models in the geometry of the west and east non-permeable boundaries. The discrepancy reflects imperfect knowledge about the subsurface distribution of bedrock surrounding the simulated aquifer. In addition, the head conductance of the north and south GHBs in the simple model is based on prior knowledge that differs from the true value used in the complex model. Similarly as in Chapter 4, the virtual reality has irregular bottom elevation (Figure 5.1c), while the simple model has linearly inclined (north to south, sides to stream location) bottom elevation.

Specific yield is homogeneous in both models and equals 0.25 in the virtual reality. The natural log conductivity($\ln K$) field was generated using a sequential Gaussian simulation code SGeMS [71], with a mean of 22.5 m/day and a sill of 1.1 (for natural logarithm $\ln K$). An anisotropic spherical variogram was used, with range 1 km in the east-west direction and 2 km in the north-south direction. The $\ln K$ field of the simple model was interpolated from 12 pilot points (location shown in Figure 5.2), using Ordinary Kriging and the true variogram used in the virtual reality. The log conductivity values at pilot points will be

calibrated. Using only 12 pilot points cannot characterize the heterogeneity details of the real K field (Figure 5.2a) even though the true variogram is used (Figure 5.2b). The case study is designed to test the capability of Gaussian process error model to compensate for bias resulting from the loss of heterogeneity.

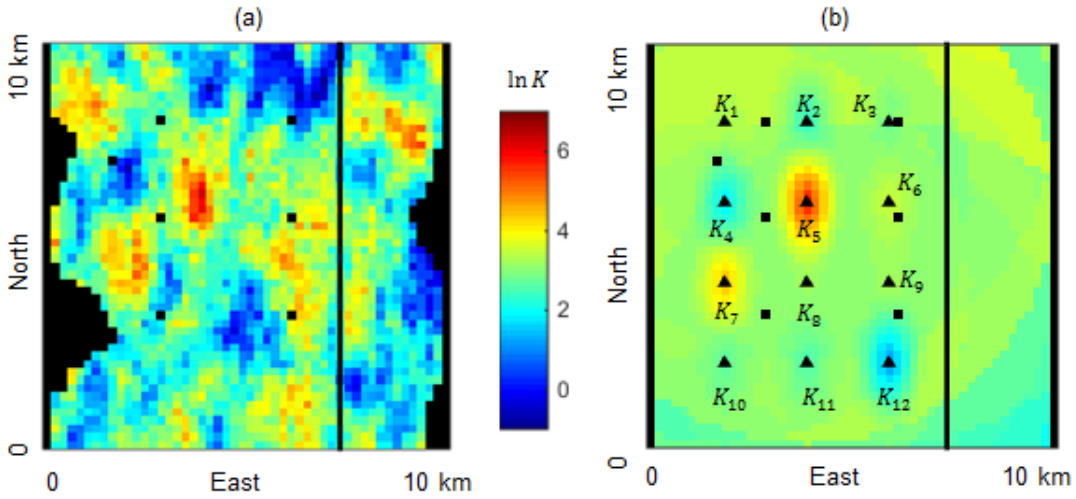

Figure 5.2: Natural log hydraulic conductivity field of (a) the complex model and (b) the simple model. Squares show locations of drawdown calibration targets, same as in Figure 5.1; triangles in (b) indicate pilot points.

In both models recharge and evapotranspiration (EVT) rates vary in space and are higher during summer months. For the complex model, the recharge and EVT rates are generated by multiplying a first order autoregressive time series with time varying spatial factors generated using SGeMS. First, annual recharge rates for 20 years are generated from a first order autoregressive AR(1) model, with a long-term mean of 0.1 m/year. Second, the annual recharge rates were distributed to every month using a fixed set of monthly multiplier. For the recharge rate used in the virtual reality, a spatially varying factor field was simulated for each month, using SGeMS, with a mean of 1, a variance of 1 and an isotropic spherical variogram with range equal to 1 km. The spatial varying recharge field for each month was calculated by multiplying the factor field with the recharge rate of that month. For the recharge rate in the simple model, on the other hand, the spatial factor fields were first

contaminated with noise, sampled at five virtual climate stations, and then extrapolated throughout the whole domain to obtain "smoothed" factor fields. This introduces input error induced by, e.g. limited sampling locations for precipitation, coarse resolution for soil type map. The new factor fields were then multiplied by the monthly recharge rates to calculate the spatiotemporally varying recharge in the simplified model. The EVT fields were generated in a similar way.

On average, the mean annual recharge rates in the virtual reality and simple model are 93.5 mm and 105.1 mm, respectively. The maximum annual EVT is 36.9 mm for the virtual reality and 37.1mm for the simple model, and the extinction depth is uniformly 2 meters beneath the land surface in both models.

Likewise in the first synthetic case study described in Chapter 4, the stream is modeled using the MODFLOW SFR1 package [66]. In this case study, however, for the virtual reality streambed hydraulic conductivity is generated using SGeMS and varies longitudinally. The stream runs from north to south; the inflow fluctuates seasonally (in phase with recharge and EVT rates) and has a mean of $0.7m^3/s$. An eight-point cross section profile is assigned to the whole reach (Figure 5.1b). The cross section represents a main channel and a side channel which is dry except during wet periods. Meanwhile, in the simplified model, the streambed hydraulic conductivity and channel width are assumed to be constant throughout the whole reach. The streambed hydraulic conductivity is to be calibrated, and the channel width is set to 14 m, which is the same as the maximum width in the virtual reality (Figure 5.1b). The inflow in the simplified model is generated by perturbing the inflow in the virtual reality, assuming that the streamflow measurement has a coefficient of variation (CV) of 0.01.

In summary, both models simulate a river-aquifer system, in which the groundwater receives inflow from the northern GHB boundary and rainfall recharge, and discharges to the stream and the southern GHB boundary. As pumping rate increases, the river would

become losing along most segments. The simple model differs from the complex model in the following aspects: simplified geometry (no-flow boundaries and bottom elevation), heterogeneity of hydraulic conductivity specified by 12 pilot point and Kriging interpolation, uniform riverbed conductance and idealized cross section, inaccurate stream inflow and spatiotemporal recharge and ET rates, and coarser time step. Therefore, the simple model has significant model structural error. A highly parametrized strategy could partly resolve model structural error, for example by using more pilot points (reaches) to better represent heterogeneity of aquifer (streambed), by parameterizing the spatially and temporally varying recharge and ET rates, and by parametrization the GHB head and conductance [91]. However, model structural error that arises from model geometry including the stream cross section is less straightforward to be avoided via parameterization. While necessarily restricted by use of a specific complex numerical model to represent reality, the case study can nevertheless provide insights into the potential of presented approach to handle various types of commonly encountered model structural error.

5.2 Calibration and validation data

The virtual reality is used to generate quarterly synthetic drawdown (s) and stream gain-and-loss (ΔQ) observations for calibration. The locations of the drawdown observations are shown in Figure 5.2b. The stream gain-and-loss (ΔQ) is computed by summing up the cell-by-cell flow exchange rates between the stream and underlying aquifer cell across the whole reach. A negative value indicates groundwater discharges to stream, and a positive value means stream loses to groundwater.

To mimic measurement error, noises were added to the virtual reality's simulation results. The drawdown measurement error is assumed to be independent and Gaussian distributed with zero mean and a constant standard deviation of 0.02 m. The streamflow measurement

is assumed to be independent and Gaussian distributed with zero mean and a coefficient of variation (CV) of 0.01. A low streamflow measurement CV is used because the focus of this study is to investigate the role of model structural error. Following [37], the variance of the stream gain-and-loss, ΔQ, is computed by summing up the variance of upstream inflow and downstream outflow. Groundwater discharges to the river during the calibration period, yielding high streamflow at the outlet. If a more realistic CV value, e.g. 0.05, is used, CV of ΔQ can exceed 100%.

Both models start from a steady-state stress period with no groundwater pumping, which mimics natural equilibrium state before development. Pumping starts from the second stress period at well A, at a constant rate of 35,000 m^3/day. The first eight years of data (excluding the steady-state period) are used as calibration targets, which consist of 32 stream gain-and-loss observations and 224 drawdown measurements at seven locations (marked in Figure 5.1a). Drawdown targets are computed by subtracting the groundwater head at a time step from the head at steady state. At the beginning of the ninth year (immediately after the calibration period), well B is turned on and pumped at the same constant rate as well A. Data in the remaining 12 years are reserved for validation. The doubled total pumping rate in the validation period represent an increased water demand scenario to test the extrapolation capacity of the proposed framework.

5.3 Least Squares Regression

We first calibrate the simple model using standard weighted least squares regression. Calibrated parameters include the hydraulic conductivity values at the 12 pilot points, uniform streambed hydraulic conductivity and the uniform specific yield. All parameters except the specific yield are log-transformed. Synthetic observations during the first eight years described in Section 5.1 were used as calibration targets. In this synthetic case study, the

measurement error of drawdown is much smaller than that of streamflow. Preliminary experiments assigned weights as the inverse of corresponding measurement error variance. As a result, the contribution to the objective function from stream gain-and-loss is much smaller than that from drawdown. Therefore, the weights of stream gain-and-loss targets are increased to achieve comparable goodness-of-fit of the two types of observations.

Identifiability issues arise for the hydraulic conductivity values at K_{10}, K_{11}, K_{12} during initial calibration attempts, mainly due to the fact that the three pilot points are located beyond the range of monitoring wells. Therefore prior information was introduced based on initial estimates of the parameters, and equivalent prior distribution will be used in Bayesian calibration. The weights of prior information represent our confidence in the initial estimates and were designated according to the coefficient of variation (CV), which equals 0.2, 1, 0.6 for $Sy, \ln K_{rb}$ and pilot points $\ln K_1, ..., \ln K_{12}$, respectively. The hydraulic conductivity parameters have higher CV because hydraulic conductivity can vary over orders of magnitude and uncertainty of measurements or estimates is typically high. A prior value of Sy, 0.30, was obtained by perturbing the corresponding parameter in the virtual reality with the specified CV value. The prior of $\ln K_{rb}$, 1.61, was generated by perturbing (according to CV) the mean value averaged over the whole stream reach in the virtual reality. We specify the same prior value for the log hydraulic conductivity at all 12 pilot points. The prior value is calculated as the arithmetic mean of the log hydraulic conductivity values in the real K field at the locations of 12 pilot points. The new objective function is

$$\Phi(\theta) = \sum_{i=1}^{n} w_i \left[z_i - M(\mathbf{x}_i, \theta)\right]^2 + \sum_{j=1}^{p} w_j \left[\theta_j - \theta_j^{pr}\right]^2, \tag{5.1}$$

where w_i, w_j denote weights of calibration targets z_i and prior information θ_j^{pr}, respectively, and p is the dimension of parameter vector θ. Since we specified same prior for all pilot points, $\theta_{K_1} = ... = \theta_{K_{12}}$ and $w_{K_1} = ... = w_{K_{12}}$. Optimization started from several sets of initial value in order to prevent from converging to sub-optimal local minima. The least

squares calibration is implemented using PEST [21].

The mean calibration error for stream gain-and-loss is $-116 m^3/day$, and the root-mean-square-error (RMSE) is $1.6 \times 10^3 m^3/day$. The streamflow at the inlet on the north boundary varies in the range of 3.0×10^4 to $1.5 \times 10^5 m^3/day$, and the coefficient of variation of streamflow measurement is 0.01. Therefore, the ΔQ RMSE is of similar magnitude with measurement error. However, autocorrelation was detected for the ΔQ residual (Figure 5.3 and Figure 5.4). For drawdown, the mean calibration error is 0.005 m; the RMSE is 0.073 m, significantly higher than the drawdown measurement error standard deviation (0.02 m). Further residual analysis reveals that the drawdown residuals are heteroscedastic and highly temporally correlated, indicating potential model structural error (Figures S1 and S3, Supporting Information). In addition, correlation among calibration error at different drawdown locations suggests spatial correlation (Figure S4, Supporting Information).

Next, the calibrated model was used to make forecasts for the validation period, in which the second well, well B, pumps at a rate of 35,000 m^3/day. Following the procedures outlined in Section 2.1, linear 95% prediction intervals were then calculated by propagating the parameter covariance matrix through the linearized model.

5.4 Classical Bayesian Method

We then calibrate the simple model using the classical Bayesian method (Section 2.2). As in the least squares calibration (Section 5.3), calibrated parameters include the hydraulic conductivity values at the 12 pilot points, uniform streambed hydraulic conductivity and the uniform specific yield. All parameters except the specific yield are log-transformed. The prior distributions of MODFLOW model parameters are consistent with the prior information employed in least squares calibration. Specifically, the prior of parameter θ_j is a normal

distribution $N(\theta_j^{pr}, (CV\ \theta_j^{pr})^2)$. The CV value for Sy, lnK_{rb} and pilot points $\ln K_1, ..., \ln K_{12}$ are 0.2, 1, 0.6, respectively, the same as for the conventional LSR calibration.

Besides MODFLOW model parameters, two additional likelihood parameters are calibrated, including the coefficient of variation of stream flow measurements ($CV_{\epsilon,\Delta Q}$) and the standard deviation of drawdown measurement error ($\sigma_{\epsilon,s}$). We choose a uniform distribution on $[0.0001, 0.5]$ as the prior of $\sigma_{\epsilon,\Delta Q}$ and $\sigma_{\epsilon,s}$. The priors are loose since in practice knowledge about measurement accuracy is usually available.

For linear and quasi-linear problems, the least squares regression and Bayesian calibration should result in almost equivalent parameter estimates and predictions, provided that equivalent priors are used [53]. In this case study, the model is found to be moderately nonlinear using the modified Beale's measure [37]. The modified Beale's measure tests model linearity with respect to the calibration targets, focusing on the parameter confidence region. Therefore, the LSR calibration results could be different from results obtained by classical Bayesian due to the linearity assumption of LSR. We carried out classical Bayesian calibration to obtain benchmark results, in addition to LSR, to be compared with the calibration results using the proposed Bayesian approach with error model.

5.5 Bayesian Method With Error Model

A premise of the Bayesian calibration framework is that the discrepancy between the model and the reality is reflected by the mismatch between model simulation and observed data [46]. As stated in Section 5.3, residual analysis suggested the presence of model structural error: the calibration error of the conventional LSR method contains bias and temporal and spatial correlation structure. Still, researchers have reported possible confounding (or identifiability) issues between the physically-based and error models [13, 70]. Because of overfitting, the

physical model parameters may over-compensate for structural error, while the error model may over-compensate for physical model parameters. This issue will be further discussed in Section 5.6.2. The Bayesian formulation provides a natural solution to alleviate these issues via specifying priors that incorporate soft expert knowledge [7]. For example, as described later this section, the prior of the drawdown error model "encourages" the model structural error to be zero. In this way, the error model takes the compensation role only when supported by the data. In addition, while the case study uses vague priors for the hydraulic conductivity parameters, it is often possible in practice to specify fairly informative priors for parameters with clear physical meaning based on preliminary measurements or estimates.

Bayesian calibration was conducted using data during the first eight years. In total 21 parameters were calibrated, which consists of 14 MODFLOW model parameters (the same as LSR) and 7 likelihood hyperparameters (Table 5.1). The prior distribution of MODFLOW model parameters are consistent with the prior information deployed in least squares calibration. Specifically, the prior of parameter θ_j is a normal distribution $N(\theta_j^{pr}, (CV\ \theta_j^{pr})^2)$. The CV value for Sy, lnK_{rb} and pilot points $\ln K_1, ..., \ln K_{12}$ are 0.2, 1, 0.6, respectively, same as for the conventional LSR calibration.

The seven likelihood hyperparameters include the coefficient of variation of stream flow measurements ($CV_{\epsilon,\Delta Q}$), the standard deviation of drawdown measurement error ($\sigma_{\epsilon,s}$), and tuning parameters of GP error models. We choose a uniform distribution on $[0.0001, 0.5]$ as the prior of $\sigma_{\epsilon,\Delta Q}$ and $\sigma_{\epsilon,s}$. One GP error model was constructed for drawdown observations at seven locations (Figure 5.1a), and another was constructed for stream gain-and-loss. Hereby they are denoted as b_s and $b_{\Delta Q}$, respectively. Input of b_s included time (t), spatial location of the monitoring well ($\mathbf{u} = (u_x, u_y)$) and MODFLOW model simulated drawdowns (M_s); input of $b_{\Delta Q}$ consisted of time (t) and the MODFLOW model simulated stream gain-and-loss ($M_{\Delta Q}$). Because they are of different magnitudes and units, all input data were linearly scaled.

For each of the GP models, we specify an isotropic squared exponential covariance function to enforce smoothness and reduce confounding. The isotropic squared exponential covariance function is a special case of Equation (3.7) when $\lambda_1 = ... = \lambda_d = \lambda$. For example, follow the notation in Equations (3.6) and (3.7) and let $\mathbf{y}_i = [t_i, \mathbf{u}_i, M_{s,i}], \mathbf{y}_j = [t_j, \mathbf{u}_j, M_{s,j}]$ denote two input data points corresponding respectively to $b_{s,i}, b_{s,j}$. Using Equation (3.7), the prior covariance between $b_s(\mathbf{y}_i)$ and $b_s(\mathbf{y}_j)$ can be computed as $\sigma_s^2 \exp\{-[(t_i - t_j)^2 + (u_{x,i} - u_{x,j})^2 + (u_{y,i} - u_{y,j})^2 + (M_{s,i} - M_{s,j})^2]/\lambda_s^2\}$. Here, σ_s, λ_s are two parameters of error model b_s. The prior covariance function for $b_{\Delta Q}$ is defined similarly with parameters $\sigma_{\Delta Q}, \lambda_{\Delta Q}$.

The characteristic scale length hyperparameters $\lambda_s, \lambda_{\Delta Q}$, can be different for each dimension of input data but are kept the same in this case study to lower the dimension of parameters to be sampled. Anisotropy can be handled by scaling the elements of input data differently. The time (t) and MODFLOW model outputs $M_{\Delta Q}, M_s$ were scaled to $[0, 1]$; $M_{\Delta Q}, M_s$ may sometimes slightly exceed the $[0, 1]$ range depending on specific parameter values used to run the model. For the drawdown GP model, the spatial location $\mathbf{u} = (u_x, u_y)$ is scaled to range $[0, 2]$, which is twice of the range of $t, M_{\Delta Q}, M_s$. This reflects the prior belief that historical model structural error at one monitoring location contains more information for inferring predictive structural error at the same location than other locations. Residual analysis on the LSR calibration error suggests complex correlation structure including strong negative correlation among errors at different drawdown locations (Figure 5.5). Such correlation cannot be fully captured by the simple squared exponential covariance function. In this case study, scaling $\mathbf{u}$ to different ranges, such as $[0, 1], [0, 3]$ did not significantly alter parameter estimates and predictions. While the true model structural error is unknown, residual analysis on the LSR calibration error can be used to guide scaling and choice for the range priors of hyperparameter. For example, autocorrelation and variogram plots could detect the correlation range in time and space when sufficient calibration data is available. In this case study, drawdown observations are only available at seven locations, making variogram

analysis challenging.

For the GP error model $b_{\Delta Q}$, a linear mean function $\mu(\mathbf{y}) = \beta_{\Delta Q}(t + M_{\Delta Q})$ is used based on the prior conjecture that the the model structural error tends to exacerbate as pumping continues. In addition, residual analysis revealed a linear trend in the LSR calibration error of stream gain-and-loss (Figure A2, Appendix). Similar to $\lambda_{\Delta Q}$, $\beta_{\Delta Q}$ are the same for t and $M_{\Delta Q}$, although they can be different if supported by prior knowledge. A constant zero mean prior was assigned to the drawdown GP error model. Residual analysis of drawdown LSR calibration error did not reveal clear pattern that can be cast into a prior mean of simple form.

In summary, the GP error models have five hyperparameters: characteristic scale lengths λ_s and $\lambda_{\Delta Q}$, standard deviations σ_s and $\sigma_{\Delta Q}$, and linear prior mean coefficient $\beta_{\Delta Q}$. The prior distributions are summarized in Table 5.1. Since the model structural error is expected to be smooth, we use a Gamma prior distribution with mean 1 and variance 0.2 for λ_s and $\lambda_{\Delta Q}$, following similar practice in [13]. We truncate the prior distribution of λ_s at 0.6 to enforce a longer characteristic scale length. This is because the GP error models will be extrapolated during the validation period. With a small λ_s, the GP error model predicted model structural error will essentially equal to the prior after one or two λ_s. With a larger λ_s, however, the information that the GP error model learned during the calibration period can be carried over to the validation period. This is not necessary for $\lambda_{\Delta Q}$ because a linear prior mean is used. Both $\lambda_{\Delta Q}, \lambda_s$ are truncated at 3 to ensure numerical stability.

The standard deviation hyperparameters, $\sigma_s, \sigma_{\Delta Q}$, reflect the prior knowledge about the magnitude of model structural error. For example, $\sigma_s = 1$ m suggests that *a priori*, the bias is unlikely to exceed 1.96 m, which is the $0.975-$th quantile of a normal distribution $N(0, 1)$. Compared to measurement error hyperparameters, less is known about $\sigma_s, \sigma_{\Delta Q}$. The exponential distribution with mean μ, $f_X(x) = \exp(-x/\mu)/\mu$, is specified as the prior of $\sigma_s, \sigma_{\Delta Q}$. This distribution favors smaller values unless data support otherwise, thus expressing the

modeler's preference of smaller model structural error [70]. Other types of prior distribution can also be used as long as they encourage smaller model structural error. In this way, the risk of overfitting the Gaussian process error model is reduced.

The GP error model $b_{\Delta Q}$ assumes a linear mean function as prior. A normal distribution with mean 0 and standard deviation 0.5 is specified for $\beta_{\Delta Q}$, so that lower bias is encouraged.

Table 5.1: Prior marginals of the parameters.

Notation	Unit	Distribution
S_y	m	$N(0.30, 0.06^2)$
$\ln K_{rb}$	m/d	$N(1.61, 1.61^2)$
$\ln K_i, i = 1, ..., 12$	m/d	$N\left(3.06, 1.84^2\right)$
$CV_{\epsilon,\Delta Q}$	-	Uniform on $[0.0001, 0.5]$
$\sigma_{\epsilon,s}$	m	Uniform on [0.0001,0.5]
$\beta_{\Delta Q}$	$10^4\ m^3/d$	Normal, $N(0, 0.5^2)$
$\lambda_{\Delta Q}$	-	Gamma, $k = 5, \theta = 0.2$, truncated at 3
λ_s	-	Truncated Gamma, $k = 5, \theta = 0.2$, on $[0.6, 3]$
$\sigma_{\Delta Q}$	$10^4\ m^3/d$	Exponential, $\mu = 0.25$
σ_s	m	Exponential, $\mu = 0.25$

The DREAM-ZS runtime settings were configured following the recommendations in [85]. The Bayesian calibration and prediction processes were carried out following the procedures outlined in Section 2.2. Three Markov chains were used to generate 15,000 samples from the joint posterior distribution of θ and ϕ after convergence was determined based on the $\hat{R}$ statistic of [30], visual inspection of trace plots and other diagnostics [18]. In the prediction phase, the Gaussian process error model uses as input $\mathbf{y}^* = [t, \mathbf{u}, M_s^*]$ for drawdown, and $\mathbf{y}^* = [t, M_{\Delta Q}^*]$ for stream gain-and-loss. The GP error models can predict the model structural error throughout the model domain, including locations not included in the calibration data. The final outcome of the Bayesian framework is an ensemble (of 15,000 samples) of predictions comprised of stream gain-and-loss $\mathbf{\Delta Q}_i^* = \mathbf{M}_{\Delta Q}^*(\theta_i) + \mathbf{b}_{\Delta Q,i}^* + \epsilon_{\Delta Q,i}$ and drawdown $\mathbf{s}_i^* = \mathbf{M}_s^*(\theta_i) + \mathbf{b}_{s,i}^* + \epsilon_{s,i}, i = 1, ..., 15,000$. Here, $\mathbf{\Delta Q}^*$ is a time series representing the stream gain-and-loss rate at different times during the prediction period; $\mathbf{s}_i^*$ denotes draw-

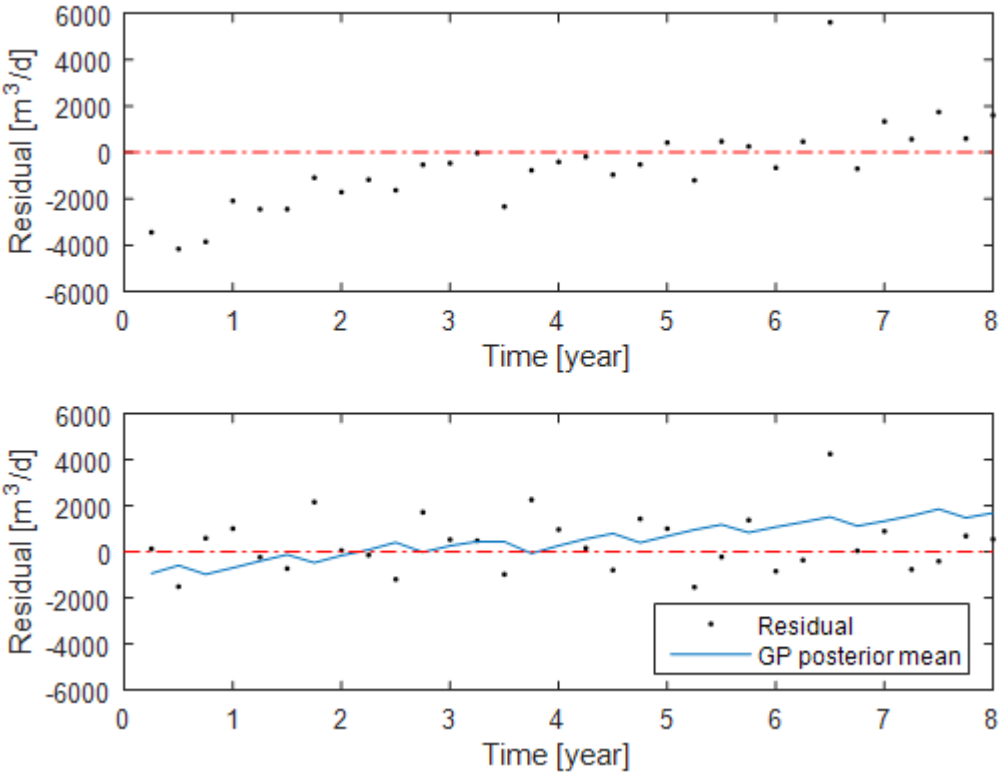

Figure 5.3: Stream gain-and-loss (Q) calibration error time series of the standard LSR method (upper) and the Bayesian approach (bottom). An increasing trend can be observed from the LSR calibration error.

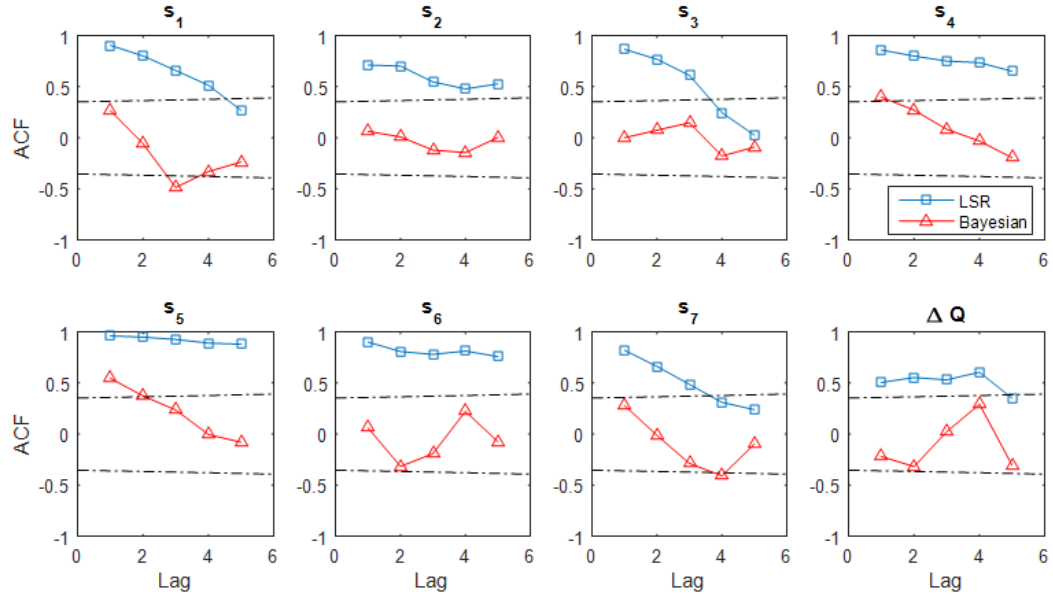

Figure 5.4: Autocorrelation function (ACF) of calibration error of the conventional LSR (blue) and the Bayesian approach (red). The dash-dotted lines enclose 95% confidence interval that the true correlations were 0. One lag equals three month as drawdowns and stream gain-and-loss are observed quarterly.

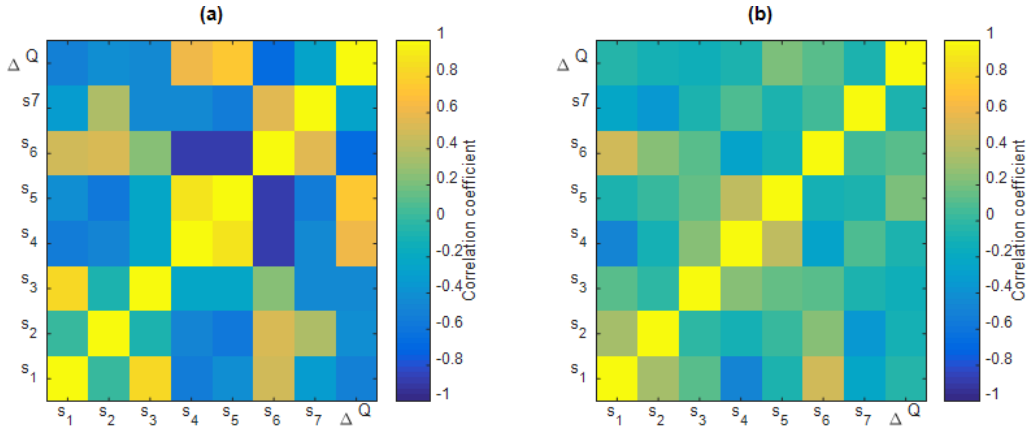

Figure 5.5: Correlation coefficients among the calibration error for the seven drawdown locations and stream gain-and-loss, resulted from the conventional LSR (a) and the Bayesian approach (b).

down varying in both space and time. The Bayesian posterior of prediction can be estimated by collecting the realizations in the ensemble. The Bayesian posterior mean is given by $\boldsymbol{\Delta}\bar{\mathbf{Q}}^*$ and $\bar{\mathbf{s}}^*$.

5.6 Results and Discussion

5.6.1 Parameter Estimates

The 95% confidence interval estimated by conventional LSR calibration and the posterior marginal distribution given by the classical Bayesian method are shown in Figure 5.6. It can be seen that for most of the parameters, the classical Bayesian maximum a posterior (MAP) estimates agree with the LSR estimates. Exceptions including $\ln Krb, \ln K_3, \ln K_7$ are likely caused by the moderate degree of non-linearity of the MODFLOW model. Non-linearity also explains the observation that for some parameters, such as $\ln K_4$ and $\ln K_7$, the Bayesian posterior distribution is much narrower than the 95% linear confidence interval given by LSR.

We then examine the parameter estimation performance of the proposed fully Bayesian framework by comparing parameter posterior distributions with the true values used in the

virtual reality and the posterior yielded by classical Bayesian calibration, as shown in Figure 5.7. We will discuss the recalibration results later in Section 5.7.

As described in Section 5.4, classical Bayesian and Bayesian calibration with GP error model used equivalent priors for MODFLOW model parameters and measurement error hyperparameters. For the specific yield S_y, it can be seen from Figure 5.7 that the classical Bayesian estimate seems overconfident. The posterior (blue) does not encompass the true value indicated by the vertical line, although the bias is small. The posterior of S_y given by the error model approach encompasses the true value, and the mode is less biased than the classical Bayesian posterior. This indicates that the Gaussian process error model can indeed reduce the degree of S_y compensating for model structural error. The uncertainty associated with S_y clearly reduces after calibration: both posterior pdfs are significantly narrower than the prior.

Validation of hydraulic conductivity parameters is less straightforward because in the virtual reality both the streambed and the aquifer have heterogeneous conductivity. For streambed k_{rb}, the true value indicated by the vertical black line is calculated by taking the arithmetic mean of streambed conductivity values over all stream cells. For hydraulic conductivity values at pilot points (locations shown in Figure 5.2b), $K_1, ..., K_{12}$, the "true" value equals the hydraulic conductivity of the cell in the virtual reality at which the pilot point is located. As mentioned in Section 5.3, using only 12 pilot points cannot characterize the heterogeneity details of the real K field (Figure 5.2a) even if true values are specified at pilot points (Figure 5.2b). Ideally, Gaussian process error models would compensate for the bias from multiple model deficiencies including the simplification of hydraulic conductivity field. Therefore, the fully Bayesian calibration approach should yield marginal posterior distributions that overlap with the "true" value.

Figure 5.7 shows that both posteriors of $\ln K_{rb}$ are underestimating the "true" value. Un-

derestimation is not surprising, because the simple model assumes a rectangular channel with a constant width of $14m$. In the virtual reality, the streambed conductance varies with stream width, which in turn varies with flow rate. The maximum width in the virtual reality is 14 m, and the effective width will in general be smaller than that. To maintain the same streambed conductance, the calibration process will arrive at a smaller streambed hydraulic conductivity, if all other conditions are the same between the virtual reality and the simple model. One way to deal with this underestimation issue is to calibrate streambed conductance instead of hydraulic conductivity. This approach is not adopted here because streambed conductance varies in time and along the stream in the virtual reality. More importantly, calibrating streambed conductance cannot avoid the deleterious effect of parameter compensation on prediction, as will be shown in Section 5.6.2. The GP error model only slightly reduce parameter compensation of K_{rb}, although it improved model predictive performance as will be shown later. The reason is that the noisy calibration data and GP prior are not adequate to identify model structural error and K_{rb} [13].

Similarly with S_y and K_{rb}, the classical Bayesian posterior of $K_3, K_4, K_5, K_6, K_8, K_9, K_{12}$ are biased and overconfident, suggesting parameter compensation. The Bayesian posteriors of these parameters are clearly less biased and mostly encompassing the "true" values. Interestingly, K_9 is an important parameter for predicting the effect of well B on drawdown and stream depletion during the validation period. This pilot point is close to well B, thus a very low K_9 may lead the model to a false forecast that well B would turn dry. The LSR estimate of K_9 is as low as 1.2 m/day, significantly lower than the Bayesian estimate and the true value. For $\ln K_7$, both methods gives biased posterior, while the with error model approach is less biased. The classical Bayesian method gives better estimates for K_1 and K_2 than the Bayesian approach. A possible reason is that the calibration data are not sufficient for the GP error model to infer the effect of incorrect eastern and northern boundary conditions.

For $K_1, K_2, K_4, K_5, K_6, K_7, K_8, K_9, K_{12}$ as well as S_y, the Bayesian posteriors show higher

parameter variability in comparison with classical Bayesian results. This indicates the confounding effect between model structural error and model parameters [13, 70]. For example, Figure 5.8 suggests correlation between the posterior samples of S_y and GP error model b_{s_1} at the end of year 8. The underlying mechanism is that, for any value of S_y within a feasible range, the MCMC sampling procedure will be able to find a corresponding b_{s_1} such that the MODFLOW model and the GP error model combined can fit calibration data reasonably well (reflected by a high likelihood value), due to the flexibility of GP regression. In some cases the confounding effect raises identifiability issues that may not necessarily be solved by increasing the amount of calibration data [13]. The parameter uncertainty arising from this confounding is inherent in the calibration problem with given data and prior; neglecting model structural error could lead to overconfident parameter estimates.

Finally, Figure 5.7 indicates lack of identifiability for K_{10}, K_{11}, K_{12}. This is anticipated when calibration data provide limited information to alter the prior because of low sensitivity of calibration data to K_{10}, K_{11}, K_{12}. As can be seen from Figure 5.2, the three pilot points are located on the south of all drawdown observations.

5.6.2 Prediction

Next we investigate the prediction capability of the fully Bayesian approach. In practice, it is usually more important to achieve accurate prediction than parameter estimates when the numerical groundwater models are used to support water resources management decision making. Comparison of predictive capability between LSR and Bayesian approaches is shown in Figures 5.9 -5.11.

In Figure 5.9, the Bayesian calibration error is calculated as the difference between calibration targets and Bayesian posterior mean. It can be seen that the Bayesian approach resulted in errors of smaller magnitude and more evenly spreading around 0.

Figure 5.10 shows the simulation results at two monitoring wells s_3, s_6 and stream gain-and-loss ΔQ for both the calibration and prediction periods. It can be seen in Figure 5.10a-c that in the calibration period, the LSR calibrated model outputs fit the measurements reasonably well, although with some bias for drawdown (Section 5.3). Despite moderate calibration error, the LSR calibrated model yields notable biased forecast during the prediction period. Figure 5.10a shows that the LSR calibrated model systematically over-predicts drawdown s_3 by 2.5 m by the end of the prediction period (20th year), and the 95% confidence interval does not encompass the validation data. For drawdown s_6, prediction is accurate for later stage of the prediction period. However, the LSR calibrated model underestimates the drawdown before year 14. A possible reason is that LSR gives lower biased estimates of K_5 and K_6, the hydraulic conductivity values at two pilot points close to the pumping well A and location of s_6. Figure 5.11a shows the LSR calibrated model drawdown prediction error throughout the model domain at the end of year 12. The model significantly overestimates drawdown in the central area, mainly because the S_y estimate is lower than the true value. The model underestimates drawdown in a small area close to the stream because LSR underestimates the hydraulic conductivity value at two pilot points close to the stream (K_6, K_9). In addition, drawdown near K_4 is underestimated, possibly because LSR estimated K_4 is lower than the "true" value.

The classical Bayesian method yields similar results with LSR calibration. Comparing Figure 5.10a and d, it can be seen that the classical Bayesian method gives even narrower prediction interval that does not encompass the validation data. For streamflow gain-and-loss, however, the classical Bayesian method did not fit the calibration data well, and produces very wide error bars. This is because ΔQ data have higher measurement error. Note that we have manually increased the weights assigned to ΔQ targets in LSR calibration.

The drawdown prediction performance of the fully Bayesian approach is demonstrated in

Figure 5.10d, e and Figure 5.11b. Because of the confounding between model structural error and MODFLOW model parameters, only total uncertainty is presented. Figure 5.10d, e shows that the proposed Bayesian method 95% credible intervals of drawdown are wider than the LSR 95% prediction interval because of higher posterior parameter uncertainty due to model structural error. The proposed Bayesian approach perfectly reproduces the calibration error, which can be anticipated based on the adaptive nature of Gaussian process regression (Section 3.2). The Bayesian with error model prediction of s_3 is substantially closer to the true validation data (Figure 5.10d) compared to the LSR and classical Bayesian calibrated models prediction. Slight underestimation after year 15 is because the GP posterior approaches the prior, which is zero, when extrapolating to a later stage of pumping. The performance can be further improved by imposing a more informative prior, e.g., a linear mean function and a higher value of characteristic length scale. For s_6, the Bayesian with error model posterior mean slightly overpredicts the drawdown after year 11. A probable reason is that, in the forecast scenario, the prediction made by the GP error model at s_6 is affected by information at other locations such as s_5. The model structural error at s_5 is significantly larger than at s_6 due to the influence of inaccurate boundary conditions, thus leading to overestimation of b_s^* at s_6. This issue may be addressed by using a more complicated covariance function and/or including additional information in GP input to reflect the spatially varying pattern of model structural error.

Comparing Figure 5.11 a and b, it can be seen that using a GP error model effectively reduces the overall predictive bias for drawdown. The improved predictive capability of the Bayesian approach is particularly evident in the central area, due to better estimate of S_y and that the GP error model compensates for model structural error. However, the Bayesian posterior mean shows slightly higher predictive bias than the LSR calibrated model near the west and north boundaries outside of the range of calibration data. In these regions, the extrapolated GP cannot fully compensate for model structural error.

As shown in Figure 5.10c, the LSR calibrated model substantially underestimates the stream gain-and-loss ΔQ. A positive value of ΔQ indicates stream leakage to the aquifer, while a negative value means that the aquifer discharges to the stream. Therefore, higher positive value of ΔQ indicates higher degree of stream depletion, and based on Figure 5.10c the LSR calibrated model under-predicts stream depletion. This is a combined result from several factors: (1) the error in north and south boundary conditions of the simple model leads to over-predicted boundary inflow, (2) the simple model, assuming spatially uniform streambed conductance, does not describe a high conductance stream segment close to well B, and (3) the LSR calibrated model underestimates drawdown in the near-stream area (Figure 5.11a). In comparison, the fully Bayesian framework yields substantially less biased ΔQ prediction because of GP error correcting and better estimates of S_y and most of the pilot points.

Figure 5.10c, f show that neither LSR nor the Bayesian approach is able to fully reproduce the magnitude of seasonal fluctuation of ΔQ. As described in Section 5.1, in the virtual reality the streambed conductance is higher during high flow seasons due to the increase in channel width. During the calibration period, the stream is primarily gaining water from the aquifer. When the streamflow rate is high, stream stage is high, and therefore the groundwater discharge to stream is low, although streambed conductance is high. In the prediction period, however, pumping at well B leads to a primarily losing stream. Stream leakage is in general high during high flow seasons due to high stream stage and large conductance. Therefore, the seasonal fluctuation magnitude tends to be larger during the prediction period as the stream seasonally switches between high and low flow regimes. The change of fluctuation magnitude cannot be simulated by the simple model because it uses a simplified rectangular channel cross section with constant width. In this case study, the GP error model has difficulty in inferring the increase of fluctuation magnitude from calibration data, because such increase is not visible during the calibration period.

Figure 5.10c shows that the LSR estimated prediction variability due to parameter uncer-

tainty (darker shades) is rather small, while the 95% linear prediction interval is wide. The wide 95% prediction interval is a combined product of streamflow measurement error and the fact that the LSR calibration was not able to fit drawdown observations to the degree of measurement error (Equation (2.7)). However, the prediction interval does not encompass the validation data. On the other hand, Figure 5.10i shows that the fully Bayesian framework yields credible intervals of ΔQ^* that encompasses most of the validation data points. As a result of the confounding effect, the relatively high prediction uncertainty is inherent of the calibration problem with given data and prior, and may be reduced if a more informative prior is available for $b_{\Delta Q}$.

Finally, it was found that the residual of the posterior mean obtained by the Bayesian with error model approach has weaker temporal and spatial correlation compared to results obtained by LSR and the classical Bayesian. Figure 5.3 shows that the GP posterior mean captured the linear trend in LSR residual, therefore the remaining residual is more evenly distributed around 0. In Figure 5.4, strong temporal correlation can be observed for LSR calibration error within the time span of one year, leading the use of a diagonal error covariance matrix dubious. On the other hand, the Bayesian residual has significantly weaker temporal correlation, mostly falling within the 95% confidence bound. This is because the GP error model captures the correlation structure in model structural error. In Figure 5.5, strong correlation can be observed among drawdown locations and between drawdowns and stream gain-and-loss. Similarly with temporal correlation, the presence of such correlation makes the use of a diagonal error covariance matrix dubious. For the Bayesian calibration error, the correlation among calibration targets is significantly smaller, within the range of $(-0.5, 0.5)$, indicating that the GP error model indeed captures the correlation structure in model structural error and renders nearly white-noise remnant error.

5.7 Recalibration

As discussed in Section 3.5, least squares recalibration is a strategy introduced to utilize the prediction given by the Bayesian approach while preserving mass balance of the MODFLOW model and other physical constraints such as the relation between groundwater discharge rate to stream and groundwater head. In the recalibration phase, the simple model is recalibrated with calibration targets given by the Bayesian posterior mean throughout the whole 20-year simulation period. The Bayesian posterior mean is used instead of observation data in the first eight years (the calibration period) because the posterior mean is expected to represent the underlying noise-free system response. Theoretically, one can include drawdown at any location/time and stream gain-and-loss at any time as recalibration targets. For illustration purposes, we assume that the quantity of interest in the case study is the impact of pumping on drawdown at the seven observation wells and stream gain-and-loss through the 20-year simulation period. Hence, we use the following as recalibration targets: $\bar{\boldsymbol{\Delta}\mathbf{Q}}, \bar{\boldsymbol{\Delta}\mathbf{Q}}^*$ and $\bar{\mathbf{s}_i}, \bar{\mathbf{s}_i}^*, i = 1, ..., 7$. Meanwhile, a full error covariance matrix Σ_b associated with the recalibration targets is computed from the error model realizations $\mathbf{b}^*_{\Delta Q,i}$ and $\mathbf{b}^*_{s,i}, i = 1, ..., 15,000$ given by the Bayesian approach (Section 3.5). Preliminary experiments indicate that the recalibration targets suffice for identifying parameter values, hence the prior information as used in the initial LSR calibration is not directly used in the recalibration process. However, the prior information is encapsulated in the recalibration targets, which depend on the prior information used in Bayesian calibration.

Finally, it is worth noting that numerical instability might occur if the error covariance matrix approaches singularity [23]. Therefore, recalibration targets should be chosen such that they are not highly correlated with each other. In the case study, it was found that model structural error has spatial and temporal correlation. Therefore, it would be inappropriate to use drawdown at every grid cell and at every stress period, as this would lead to Σ_b with a high condition number. However, using quarterly drawdown at the seven observation wells

and stream gain-and-loss as recalibration targets still leads to some numerical instability due to strong temporal correlation. To reduce the condition number of Σ_b, a small amount equivalent to 5% of measurement error was added to the diagonal entries.

Based on Figures 5.6 and 5.7, for some parameters the recalibration yields estimates different from the classical LSR calibration results. This is not surprising since recalibration uses different targets and error covariance matrix. The recalibration 95% CI is narrower for most of parameters, because (1) more calibration "data" are used, and (2) the full error covariance matrix imposes less penalty to systematic bias when computing the objective function in Equation (2.2) [54].

Comparing the 20-year simulation results of the recalibrated model with the Bayesian posterior mean used as recalibration targets, the mean error is 0.16 m (drawdown) and $2.2 \times 10^3 m^3/day$ (stream gain-and-loss), and the RMSE is 0.42 m (drawdown) and $2.8 \times 10^3 m^3/day$ (ΔQ). The Bayesian posterior realizations have been corrected by the GP error model. Therefore, it is not surprising that the simple model cannot match the posterior mean perfectly, given the model structural error. For drawdown, comparing Figure 5.11a and c, it can be seen that the recalibration strategy improved the predictive accuracy over the standard LSR calibrated model in the center part of domain. In addition, the recalibration strategy reduces the underestimation of LSR calibrated model in a small area close to the stream, which is consistent with the observation from Figure 5.7 that recalibration yields less biased estimates of K_6 and K_9. Comparing Figure 5.11c with a and b, the recalibrated model has higher bias near the west boundary, which can be considered as a negative impact of parameter compensation. For stream gain-and-loss, the recalibrated model prediction has smaller bias compared to the original LSR calibrated model (Figure 5.10l); this is related to more accurate prediction of drawdown near stream.

Figure 5.10j-l provide 95% prediction intervals associated with the recalibrated model. The

PIs are not computed using the methods outlined in Section 2.1. The recalibration strategy uses an error covariance matrix computed from the realizations of the error model given by the Bayesian approach. It is then straightforward to estimate prediction variability by drawing samples from the error covariance matrix. The parameter uncertainty in Figure 5.10j-l is computed based on 10,000 independent samples drawn from $N(0, \Sigma)$; Σ is computed from the posterior samples $\mathbf{b}_i^* + \epsilon_i, i = 1, ..., 15,000$. It is not surprising that the resulting 95% PIs in Figure 5.10j-l are similar in width with the Bayesian with error model credible intervals in Figure 5.10g-i.

In summary, the results suggest that recalibration is a promising method to achieve more robust prediction of quantities of interest by assimilating the GP error model updated predictions and using a full error covariance matrix inferred by the Bayesian approach. The recalibration strategy represents a tradeoff between the aims of obtaining realistic parameter estimates and accurate predictions. Using a full error covariance matrix could potentially alleviate the degree of parameter compensation. This is because when the errors are positively correlated, using a full error covariance matrix indicates higher tolerance for systematic bias.

The implementation configuration of the recalibration strategy is problem specific. For example, if obtaining accurate prediction is of central importance, the measurement error covariance matrix can be used instead of the full error covariance in order to force a good fit to the Bayesian posterior mean, i.e., overfitting on purpose. This scenario does not require the full error covariance matrix given by Bayesian inference with the error model. Postprocessor approaches that construct statistical error models conditioned on an existing calibrated physically-based model [26, 98, 97] can provide corrected prediction to be used as recalibration targets. The postprocessor approaches are computationally efficient because they typically do not require repeated evaluation of the physically-based model. The error model corrected prediction is overall more accurate than the prediction given by the initial calibrated model using conventional LSR. Therefore, if the recalibration is able to fit the

corrected predictions reasonably well, the recalibrated model can be expected to yield more accurate prediction than the initial calibrated model. When recalibration error is large, multiple models can be recalibrated separately, each using one subset of recalibration targets, as suggested in [22] in a different context.

5.8 Summary and Discussions

We investigated the role of model structural error in calibration and prediction in groundwater flow modeling practice using a synthetic case study of surface-ground water interaction under changing pumping conditions. We first demonstrated that conventional least squares regression and the classical Bayesian method yield biased (and often overconfident) predictions under a scenario differing from the calibration period. This finding is consistent with others in the literature reporting the deleterious impact of parameter compensation on prediction performance.

In order to properly treat model structural error, we present a Bayesian framework that incorporates data-driven error models. The proposed approach allows for a complete assessment of uncertainty by jointly inferring parameter variability and model structural error. We found in the case study that Gaussian process error models can represent the underlying model structural error reasonably well, although not perfectly. Integrating error models into Bayesian calibration reduces the degree of parameter compensation, leading to parameter posteriors that differ substantially from LSR and classical Bayesian estimates. We also showed that the Bayesian framework with error model achieves more accurate prediction and more robust prediction intervals compared to the LSR and classical Bayesian calibrated models.

Using an external error model to correct for model structural error leads to the violation

of mass conservation and inconsistency between parameter estimates and predictions. We argue that profound model structural error could be an indication of error in water budget terms, such as recharge. In circumstances where preserving water budget and other physical constraints are important, we present a recalibration strategy that incorporates model structural error into least squares regression by using a full error covariance matrix. We showed in the case study that the recalibration strategy yields more realistic parameter estimates and more accurate prediction compared to the conventional LSR and classical Bayesian calibration methods.

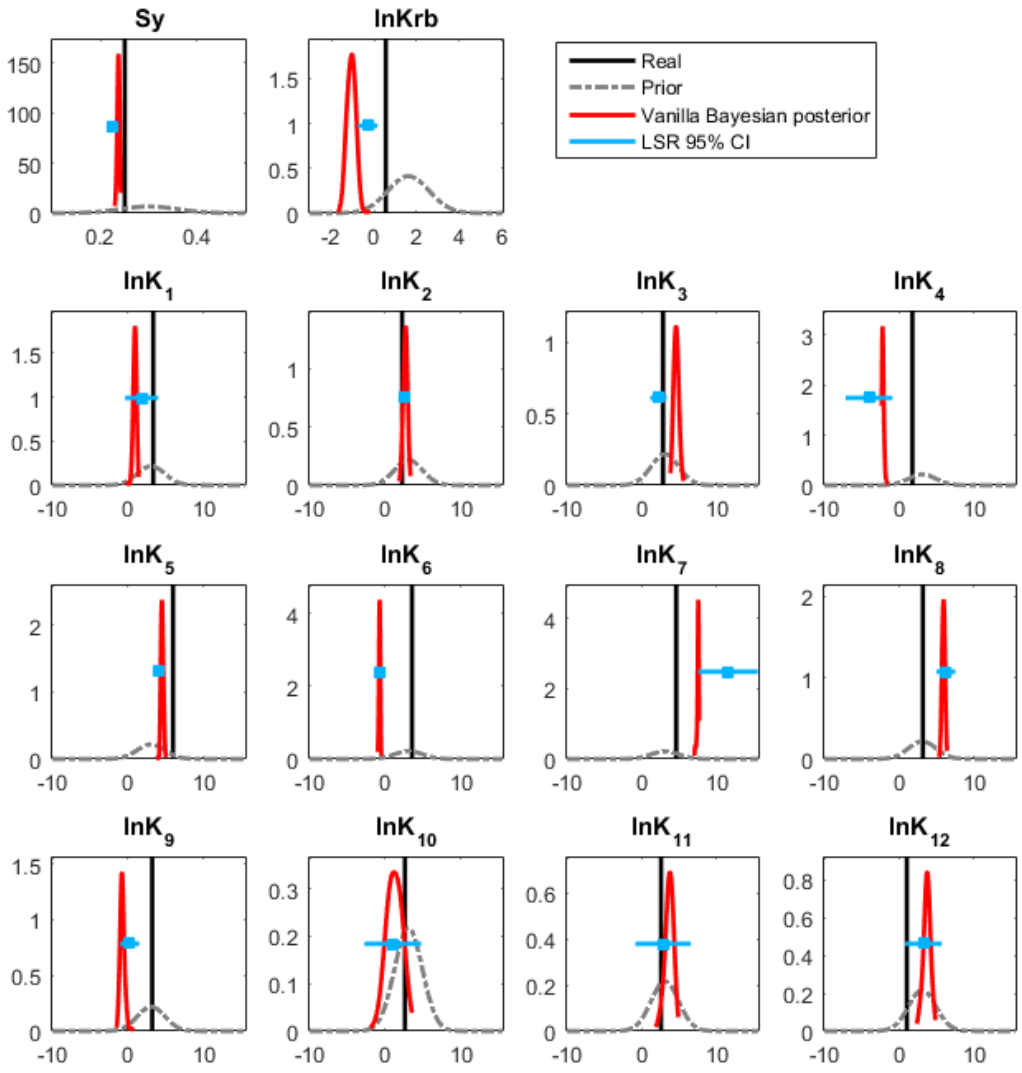

Figure 5.6: Prior distributions (grey, dashed), 95% confidence intervals given by conventional least squares calibration (blue), and marginal posterior distributions given by classical Bayesian calibration (red). Black vertical lines show the "true" values. The specific yield S_y is dimensionless, and the hydraulic conductivities $K_{rb}, K_1, ..., K_{12}$ with unit [m/day] are natural log transformed.

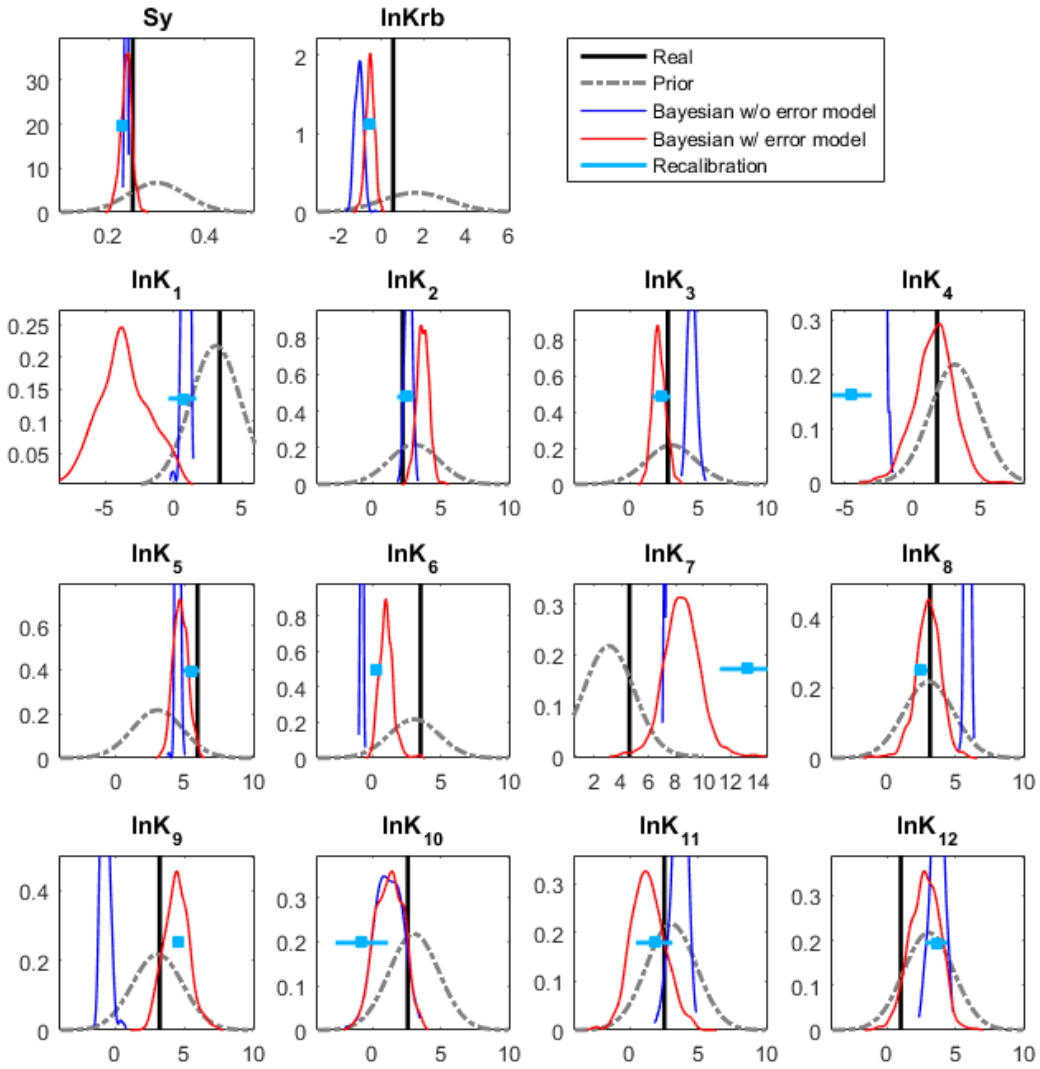

Figure 5.7: Prior distributions (grey, dashed), marginal posterior distributions given by classical Bayesian calibration (blue), marginal posterior distributions given by Bayesian calibration with GP error model (red) and 95% confidence intervals given by the recalibration strategy (light blue bar). Black vertical lines show the "true" values. The specific yield S_y is dimensionless, and the hydraulic conductivities $K_{rb}, K_1, ..., K_{12}$ with unit [m/day] are natural log transformed.

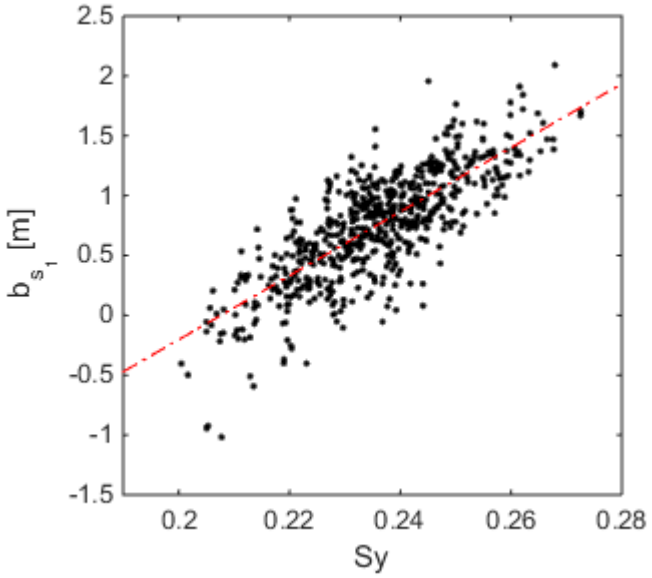

Figure 5.8: Correlation between specific yield S_y and Gaussian process error model posterior b_{s_1} at the end of calibration period (year 8).

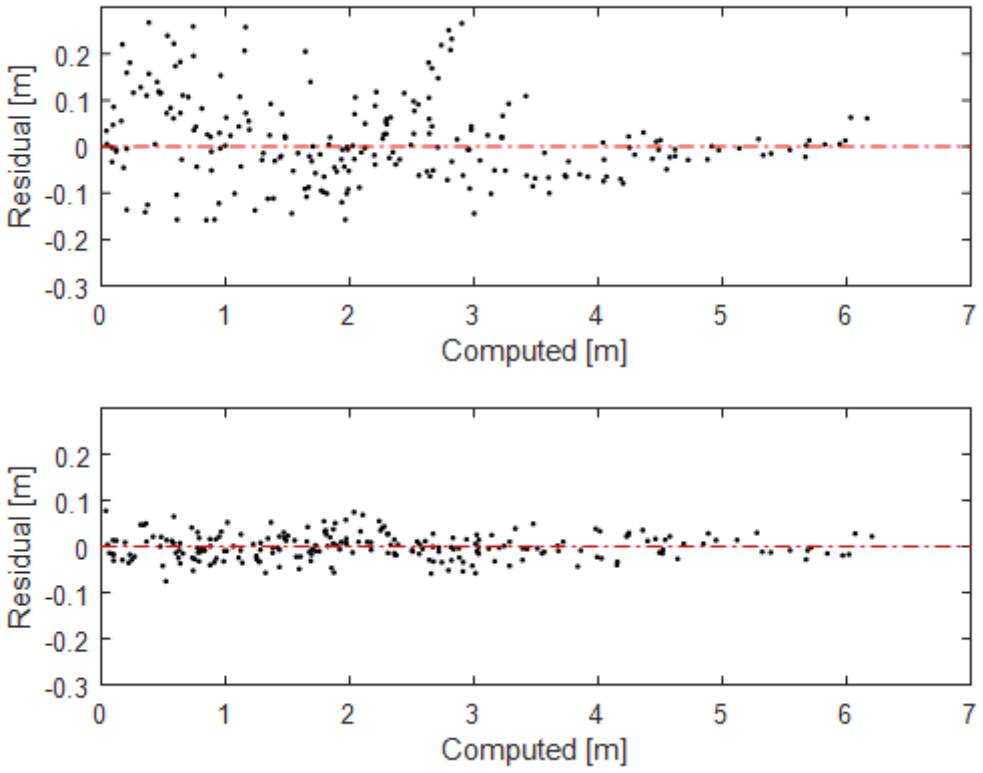

Figure 5.9: Up: Drawdown calibration error of the standard LSR method plotted versus drawdown computed by the LSR calibrated model. Bottom: Drawdown calibration error of the Bayesian approach (with error model) plotted versus posterior mean of the Bayesian approach.

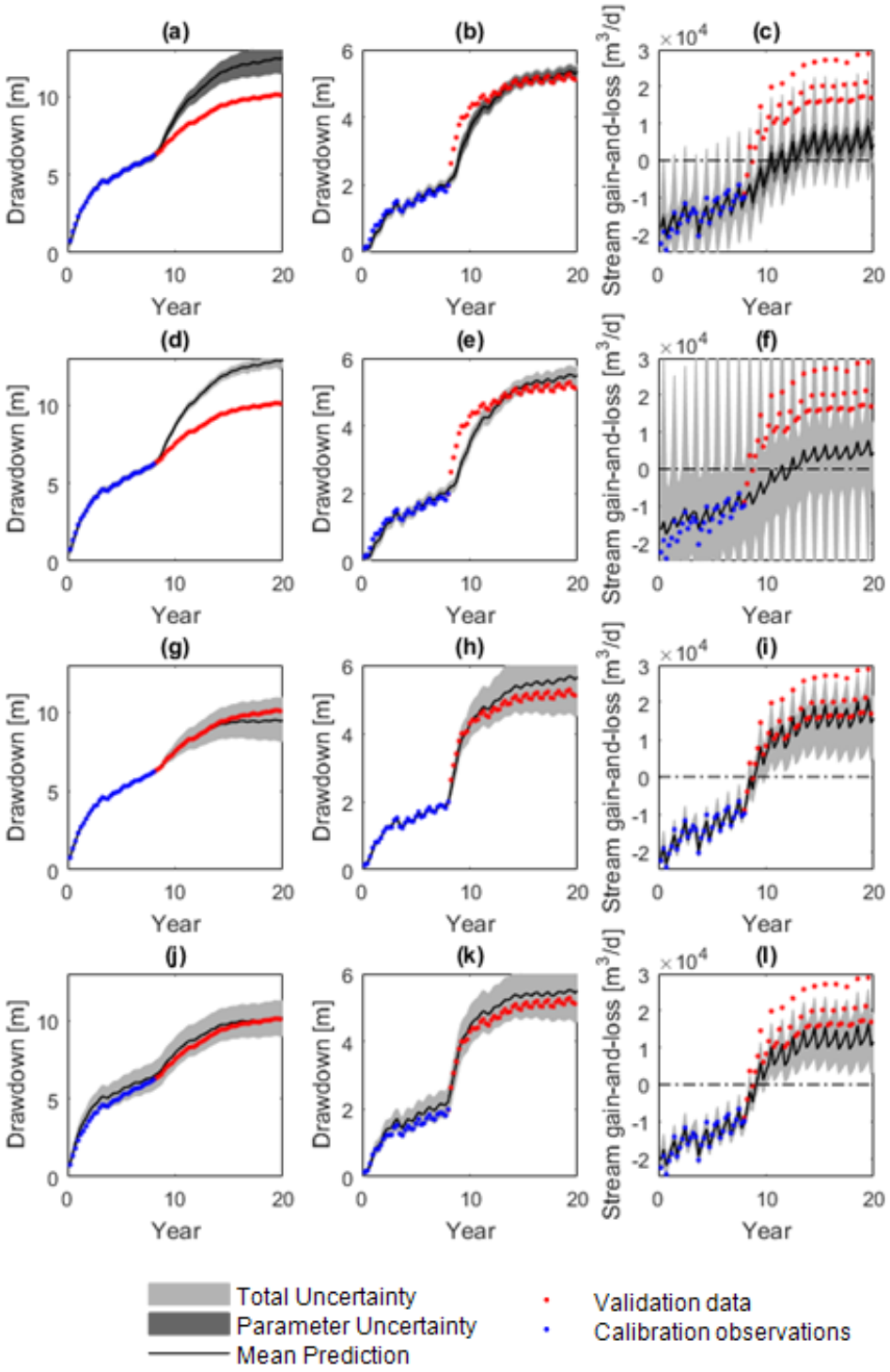

Figure 5.10: Prediction and associated uncertainty of drawdown s_3 (left), s_6 (center) and stream gain-and-loss ΔQ (right) given by least squares regression (a-c), classical Bayesian (d-f), proposed fully Bayesian approach (g-i) and the recalibration strategy (j-l). Dark shades in (a-c) indicate 95% LSR confidence intervals due to parameter uncertainty, and light shades indicate 95% prediction (credible) intervals of total uncertainty. Blue dots show calibration measurements; red dots correspond to noise-free data reserved for validation.

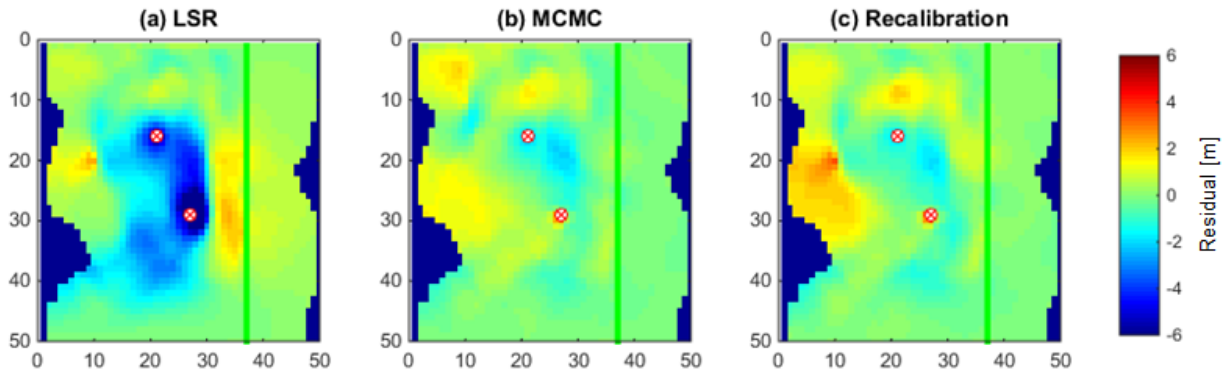

Figure 5.11: Drawdown prediction error at the end of year 12, yielded by (a) conventional least squares regression, (b) proposed fully Bayesian approach and (c) the recalibration strategy. Locations of pumping wells are shown.

Chapter 6

A REGIONAL-SCALE GROUNDWATER MODELING CASE STUDY

In this chapter, we apply the Bayesian with error model approach to calibrate a real-world regional-scale groundwater flow model. This case study is motivated by the findings in our previous work [98] that systematic bias exists in groundwater head simulated by the least squares calibrated model, indicating presence of model structural error. In this chapter, we use a Gaussian process to describe the model structural error, and jointly infer the error model with groundwater model parameters.

We discuss a surrogate modeling strategy we employed to reduce the computational cost associated with Bayesian calibration of a complicated groundwater model. Calibration and prediction are implemented using classical Bayesian and the proposed Bayesian approach with error model. Performance of the two methods is compared. Based on the results presented in this chapter, a manuscript is in preparation.

6.1 The Spokane Valley-Rathdrum Prairie Model

This study is based on a regional-scale groundwater flow model, namely the Spokane Valley-Rathdrum Prairie (SVRP) model. The SVRP aquifer covers approximately 326 square miles across the states of Idaho and Washington, and supplies drinking water to more than 500,000 residents. A MODFLOW-2000 model was jointly developed by the USGS, Idaho Department of Water Resources, the University of Idaho, and Washington State University [39]. We have used the SVRP model as a case study in our previous paper [98] in which data-driven

error models based on machine learning techniques were used as postprocessors to improve the model's head predictive accuracy.

Figure 6.1 shows the SVRP model domain. The model has a uniform cell size of 1,320 by 1,320 ft (402.34 m), and stress period of 1 month from September 1990 through September 2005. The SVRP aquifer is conceptualized as one active layer except in Hillyard Though and the Little Spokane Arm. In those areas, the aquifer was divided by a clay layer (layer 2) into an upper, unconfined unit (layer 1) and a lower, confined unit (layer 3), as shown in Figure 6.1.

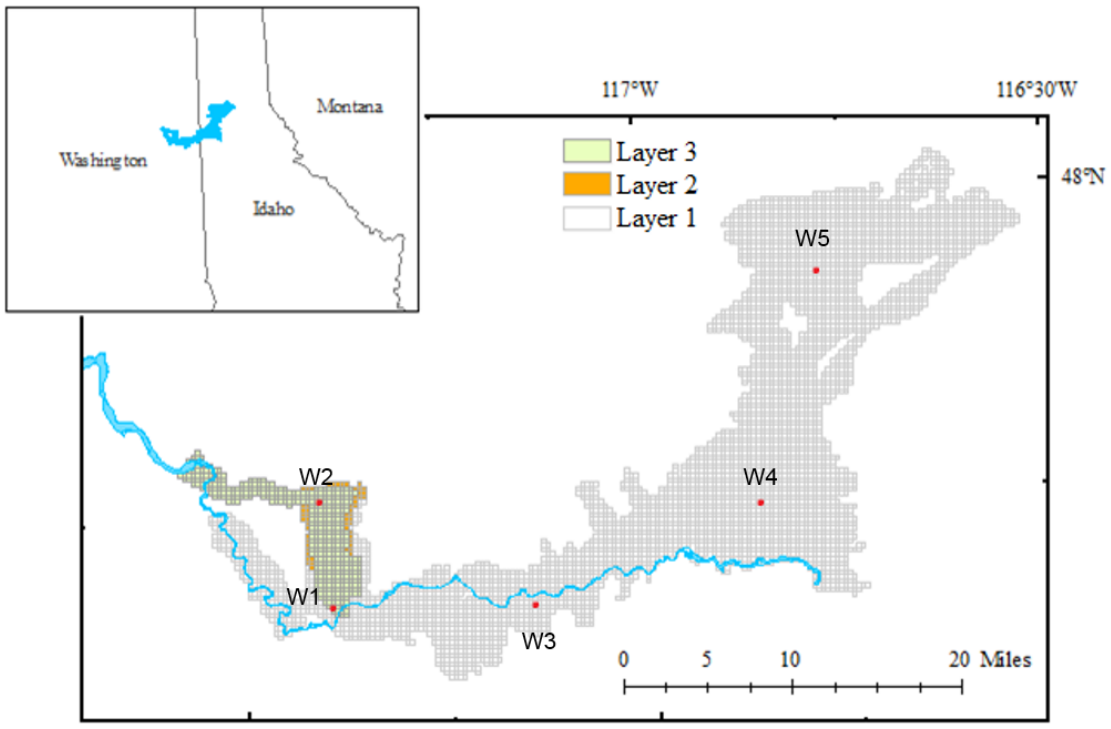

Figure 6.1: The Spokane Valley-Rathdrum Prairie aquifer on the border of Washington and Idaho. The Spokane River is shown in blue. The three layers are shown in different colors. The grids represent the spatial discretization of the MODFLOW model. Also shown are the locations of representative monitoring wells as discussed in Section 6.7. Adapted from "Use of machine learning methods to reduce predictive error of groundwater models," by T. Xu et al., 2014, Groundwater, 52(3):448-460, 2014.

The parameterization of the model is summarized in Table 6.1 and explained in the following

paragraphs. The horizontal hydraulic conductivity (K_h) field in layer 1 was grouped into 22 zones. The value of K_h is uniform within each zone and denoted by HK1-1 through HK1-22. The vertical hydraulic conductivity, K_v, is uniform in all active cells of layer 1. The specific yield S_y is represented using three zones, SY-1, SY-2 and SY-3. For layer 2, K_h and K_v are represented with two zones. For layer 3, K_h is represented using two zones and denoted by HK3-1 and HK3-2. The vertical hydraulic conductivity in layer 3 is uniform. Storativity of layers 2 and 3 are negligible.

The main aquifer (layer 1) receives inflow from adjacent tributary basins, lakes, precipitation recharge, irrigation and septic systems. The aquifer loses water mainly through pumping and exchanges water with the Spokane River and Little Spokane River. The Little Spokane River, Lake Pend Oreille and Coeur d'Alene Lake are simulated using MODFLOW River Package (RIV), and a single conductance is assigned to the river and each lake, denoted as C-LSR, C-PO and C-CDA, respectively. The Spokane River is simulated using Streamflow-routing package (SFR) [66]. The Spokane River within the model domain is divided into 11 sections (the stream sections are shown in Figure 35 in [39]), and one streambed conductance is assigned to each section (denoted by KVSR-1 through KVSR-11).

The model was calibrated using PEST [21] by the model developers [39]. Calibration parameters include horizontal hydraulic conductivity in layers 1 and 3, specific yield and conductance defined in the RIV and SFR packages. It was found that the calibration data were not sensitive to HK1-21 and KVSR-11. In addition, the estimated value of HK3-2 was considered as unreasonable. Therefore, these three parameters were not adjusted in the calibration process but fixed. In total, there are 38 calibrated parameters [39] as listed in Table 6.1.

The PEST calibration data was comprised of over 1,500 groundwater level (or head) measurements and 313 measurements of streamflow gain-and-loss along segments of the Spokane

Table 6.1: Calibrated model parameters, lower and upper bounds, PEST estimated values and 95% confidence intervals. Adapted from Table 8, "Model parameters, acceptable intervals, estimated values, and 95-percent confidence intervals." in Hsieh et al. [39].

		Enforced bounds			95% confidence interval	
Parameter	Units	Lower	Upper	PEST estimated value	Lower	Upper
HK1-1	ft/d	100	50,000	13,000	6,440	26,400
HK1-2	ft/d	100	50,000	6,170	4,300	8,860
HK1-3	ft/d	100	50,000	17,100	13,500	21,700
HK1-4	ft/d	100	50,000	12,100	10,800	13,500
HK1-5	ft/d	100	50,000	22,100	20,200	24,300
HK1-6	ft/d	100	50,000	19,100	17,800	20,400
HK1-7	ft/d	100	50,000	7,470	6,820	8,170
HK1-8	ft/d	100	50,000	9,500	8,040	11,200
HK1-9	ft/d	1	5,000	2,630	2,400	2,870
HK1-10	ft/d	1	5,000	2,180	2,020	2,360
HK1-11	ft/d	1	5,000	1,980	1,710	2,300
HK1-12	ft/d	1	5,000	608	485	362
HK1-13	ft/d	1	5,000	3,110	2,470	3,920
HK1-14	ft/d	1	5,000	90	82	98
HK1-15	ft/d	1	5,000	1,290	755	2,190
HK1-16	ft/d	1	5,000	55	53	56
HK1-17	ft/d	1	5,000	5	4	7
HK1-18	ft/d	1	5,000	78	74	82
HK1-19	ft/d	1	5,000	95	93	97
HK1-20	ft/d	1	5,000	64	55	76
HK1-22	ft/d	1	5,000	140	131	150
HK3-1	ft/d	1	5,000	207	155	276
C-PO	ft^2/d	10^{-10}	10^{10}	241,000	102,000	572,000
C-LSR	ft^2/d	10^{-10}	10^{10}	40,600	36,100	45,700
C-CDA	ft^2/d	10^{-10}	10^{10}	77,800	40,000	151,000
SY-1	–	.1	.3	.1	.08	.13
SY-2	–	.1	.3	.19	.16	.21
SY-3	–	.1	.3	.21	.18	.23
KVSR-1	ft/d	.01	10	.054	.047	.062
KVSR-2	ft/d	.01	10	.25	.23	.27
KVSR-3	ft/d	.01	10	.054	.047	.062
KVSR-4	ft/d	.01	10	.14	.10	.20
KVSR-5	ft/d	.01	10	9.4	7.3	12.2
KVSR-6	ft/d	.01	10	.01	.005	1.7
KVSR-7	ft/d	.01	10	10	5.6	18
KVSR-8	ft/d	.01	10	.3	.20	.45
KVSR-9	ft/d	.01	10	10	1.70	50
KVSR-10	ft/d	.01	10	10	.63	159

River and Little Spokane River from October 1995 to September 2005. The five years before October 1995 were considered as the warm-up period, thus observations from September 1990 to September 1995 were excluded from the calibration data. More details about the

model can be found in the documentation [39] that is available on the project website (`http://wa.water.usgs.gov/projects/svrp/summary.htm`).

Overall the calibrated model fits the calibration data to a reasonable degree, given the complexity of the model. There is visible mismatch between measured and model simulated streamflow gain-and-loss. Nevertheless, the simulated gains-and-loss are mostly within the error bounds of the measured quantities, mainly because of the relatively large measurement error of streamflow. However, residual analysis revealed that some bias existed in the head residuals of the PEST calibrated model. The mean error is 3.37 ft (1.03 m) and RMSE is 15.50 ft (3.20 m), which is larger than a reasonable estimate of the waterlevel observation error. In addition, our preceding work [98] found that the calibration error is correlated temporally and spatially, indicating presence of model structural error.

6.2 Calibration and Validation Data

Post-audit of the SVRP model is not possible due to the lack of input data beyond the simulation period (from September 1990 to September 2005). Generating new inputs (e.g. recharge and pumping rates) requires a variety of information, such as land use map, irrigation amount of both agricultural and recreational lands, and domestic and public supply pumping records. Not all of the required information is public available. Therefore, it is not possible to run the model in forecast mode beyond the simulation period.

In this case study, we follow the model developers' practice of using the first five years as warm-up period. Groundwater piezometric head and stream gain-and-loss measurements from October 1995 to September 2004 are used as calibration data, while measurements from October 2004 to September 2005 are reserved for validation. October 2004 to September 2005 corresponds to a dry period in that the precipitation recharge is lower than in preceding

years (Figure 9 in [39]). In other words, the validation period exhibits somewhat different hydrogeologic condition compared to the conditions reflected by the calibration dataset.

The whole dataset is comprised of groundwater head and stream gain-and-loss measurements on the Spokane River and Little Spokane River that have been used in the PEST calibration (Section 6.1), as well as additional head observations that became available via the USGS Water Data for the Nation online database (`http://waterdata.usgs.gov/nwis/gw`) after model construction and PEST calibration in 2006. In total, calibration dataset includes 1,552 head data points at 342 wells from October 1995 to September 2004, 177 stream gain-and-loss measurements on segments of the Spokane River, and 87 stream gain-and-loss measurements on segments of the Little Spokane River; the validation dataset is comprised of 554 head measurements at 55 wells and 41 stream gain-and-loss measurements on the Spokane River and 18 on the Little Spokane River, from October 2004 to September 2005.

6.3 Surrogate Models

One forward run for the simulation period (from 1990 to 2005) of the SVRP MODFLOW model takes approximately 2 minutes and 20 seconds (depending on parameter values) on a single 2.0GHz CPU core. Bayesian calibration often requires over tens of thousands model evaluations to sample from the posterior, and therefore can be infeasible for a complicated groundwater flow model like the SVRP model. In order to reduce the computational cost, we construct computationally frugal surrogate models to mimic the model outputs that vary with parameter values (Section 3.4).

In order to generate the training data of the surrogate models, the SVRP model was run repeatedly using 3,200 sets of the 38 calibration parameters drawn using Latin Hypercube sampling. The samples cover a relatively wide range of parameter values that are *a priori*

believed to be reasonable based on the hydrogeologic conditions of the region. The span of parameter values were selected based on the lower and upper bounds (Table 6.1) and starting value of the PEST calibration as documented in [39]. As the 3,200 model evaluations are independent from one another, they were run in parallel using the high performance computing (HPC) resources provided by the Illinois Campus Cluster.

Support Vector Regression algorithm is used to construct surrogate models for each head measurement at a certain time and location and for each stream gain-and-loss observation along one segment of river at a certain time, resulting in a total of 1552 surrogate models for head and 264 for stream gain-and-loss. These surrogate models take as inputs the model parameters. Sensitivity analysis reveals that for many observations, the change in the values of some parameters does not significantly alter the model outputs; insensitive parameters vary among observations. For each surrogate model, we select the subset of parameters to which the model output is the most sensitive according to the predictor importance measure calculated with Random Forest regression [12].

Split-sample validation is carried out to examine the emulation accuracy of SVR surrogates. The 3,200 sets of model simulation results are randomly divided into a training dataset and a testing dataset. The training dataset is comprised of 80% of all data, i.e. 2,560 sets of parameters and corresponding MODFLOW model outputs. Using the training dataset, we then tune the SVR hyperparameters via five-fold cross validation (Section 3.4). We then retrain the SVR surrogates using the whole training dataset; the trained SVR surrogates are then tested on the testing dataset which consists of 640 data points for every drawdown and stream gain-and-loss output.

Figure 6.2 compares the surrogate model prediction results with the MODFLOW model outputs for the testing dataset. Overall, the head emulation coefficient of determination R^2 is 0.999, and the RMSE is 1.3 ft. The RMSE of surrogate models is smaller than model

structural error, as the RMSE of LSR calibrated model is 15.50 ft (Section 6.1). The 3,200 parameter sets cover a wide range, and the corresponding SVRP model simulated head can vary significantly (by over 55 ft for half of calibration targets). For streamflow gain-and-loss on the Spokane River and the Little Spokane River, the coefficient of determination R^2 is 0.996, and the RMSE is 29.5 cfs. For 83% of stream gain-and-loss calibration targets, the SVR surrogate RMSE is smaller than the measurement error standard deviation, which ranges from 9 to 1.2×10^3 cfs. Overall, the SVR emulation accuracy can be considered as acceptable.

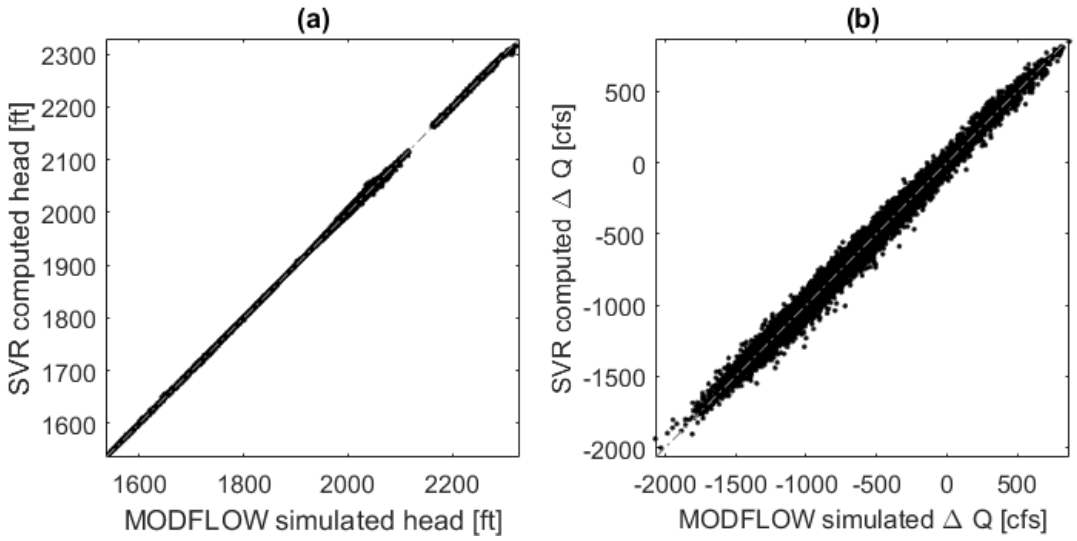

Figure 6.2: (a) Groundwater head simulated by the SVR surrogate models plotted versus the head simulated by the MODFLOW model. (b) Streamflow gain-and-loss simulated by the SVR surrogate models plotted versus the gain-and-loss simulated by the MODFLOW model.

When evaluating the likelihood during MCMC sampling, we will use SVR as surrogates of the MODFLOW model. This introduces additional error to the inference process. More specifically, letting $f_{i,j}$ denote the SVR prediction corresponding to the j-th LHS sample of parameters for the $i-$th MODFLOW output M_i, we have $M_i = f_{i,j} + e_{i,j}$, where $e_{i,j}$ is the surrogate error. As described earlier, we test the trained SVR models on a testing dataset consisting of 640 data points for every drawdown and stream gain-and-loss outputs. Comparing the SVR predicted values with the MODFLOW simulation results, we calculate

the mean squared error (MSE) for each model outputs as $s_i^2 = \frac{1}{n}\sum_{j=1}^{n} e_{i,j}^2$, where n is the number of testing data points and equals 640. The MSE s_i^2 is a good estimate of the surrogate error variance. Incorporating the surrogate error into Equation (3.8), we have

$$\begin{aligned}\log p(\mathbf{z}-\mathbf{M}|Y) = &-\frac{1}{2}(\mathbf{z}-\mathbf{M}-\mu)^T\left(\Sigma+\sigma_\epsilon^2 I+\Sigma_{SVR}\right)^{-1}(\mathbf{z}-\mathbf{M}-\mu)\\ &-\frac{1}{2}\log|\Sigma+\sigma_\epsilon^2 I+\Sigma_{SVR}|-\frac{n}{2}\log 2\pi,\end{aligned} \tag{6.1}$$

where Σ_{SVR} is a diagnonal matrix with elements $s_i^2, i = 1, 2, ...$ Here, we assume that the surrogate errors $e_{i,j}$ are independent because residual analysis did not show any significant correlation.

To further examine the quality of surrogate models, we carried out Bayesian calibration with the same head and streamflow gain-and-loss data and associated weights as used in the PEST calibration [39]. The Bayesian and PEST calibration used the same prior and lower/upper bounds for parameters. This is not a rigorous comparison because PEST calibration entails the quasi-linear assumption of the model with respect to model parameters. Ideally, the resulting joint posterior distribution of parameters should be compared to the results from Bayesian calibration using the MODFLOW model, which would take one month or even more, depending on the MCMC convergence rate. Therefore, such an experiment is usually not possible for real-world applications. Figure 6.3 shows the comparison between the PEST parameter estimates (with 95% confidence intervals) and posterior marginal distributions obtained by MCMC sampling with surrogate models. In general, the posteriors of hydraulic conductivity and specific yield parameters agree with PEST estimates. Exceptions include the hydraulic conductivity for two zones, HK1-1 and HK1-14. These zones are close to the model domain boundary. Near the boundary, the SVRP aquifer has low saturation thickness. Accordingly, the head response surface with respect to parameters could be nonlinear and less smooth, which would be challenging for the surrogate models to emulate. The Bayesian calibration yields streambed conductance posteriors that are different from

PEST results. A possible reason could be that adding the surrogate error variance changes the goodness-of-fit tradeoff between head and stream gain-and-loss targets. Nevertheless, given the overall good agreement between the PEST estimates and Bayesian posterior, we consider the SVR surrogates eligible to be used in later experiments.

6.4 Classical Bayesian Calibration

As benchmark, we run classical Bayesian calibration first to estimate the 38 SVRP model parameters, using the surrogate models as substitute for the SVRP model. The priors of parameters are specified as uniform distributions over a relatively wide range that is considered as physically reasonable given the hydrogeologic condition, as listed in the third and fourth columns in Table 6.1.

In the PEST calibration by SVRP model developers, it was found that the sum of squares of weighted residuals is dominated by head measurements if calibration weights are calculated as the inverse of measurement error standard deviation. This is similar with the findings in our second case study (Section 5.3), and results from two facts. First, the streamflow measurement error standard deviation is 5% of the measured streamflow; the variance of stream gain-and-loss ΔQ is calculated as the sum of the variance of upstream and downstream measurement error. As the magnitude of gain-and-loss is small compared to upstream and downstream flow, the relative error in ΔQ measurements is much higher than that in head observations. Second, the number of head measurements is about six times the number of streamflow gain-and-loss measurements. To ensure approximately equal goodness-of-fit to head and flow measurements, the model developers added 5 ft. to the standard deviation of head measurement errors to reduce the weights of head targets [39].

Accordingly, in the classical Bayesian calibration experiment, we follow this practice and

added 5 ft. to the standard deviation of head measurement errors when evaluating the likelihood during MCMC sampling. In addition, as described in Section 6.3, we added a surrogate error covariance matrix Σ_{SVR} to account for the error resulting from using the SVR surrogates. When deriving the prediction intervals, we do not make these adjustments because we will run the SVRP model during the prediction phase.

The DREAM-ZS runtime settings were configured following the recommendations in [85]. Ten Markov chains were used to generate 1,600 samples from the joint posterior distribution of model parameters after convergence was determined based on the $\hat{R}$ statistic, visual inspection of trace plots and other diagnostics [18]; about 60,000 model evaluations were required to converge (burn-in). The marginal posterior distributions of 38 parameters are shown in Figure 6.4. The parameter estimation and prediction results will be discussed in Sections 6.6 and 6.7.

6.5 Fully Bayesian Calibration with Error Model

Based on the residual analysis results reported in Section 6.1, we constructed an error model to describe model structural error in head. We did not construct error models for stream gain-and-loss because residual analysis did not reveal significant bias and correlation structure in streamflow residuals. The inputs of GP error model are chosen similarly as in Section 5.5. As variogram analysis revealed spatial correlation in head residual [97], the input of the error model should include spatial locations of head measurements, $\mathbf{u} = (u_x, u_y)$. Since in this case study the model needs to make forecasts beyond the calibration time span, using time as one of the inputs would require the error model to be extrapolated in time. Our earlier results in [98] indicate that for temporal prediction it is better not to include time as an input for the data-driven error model. Therefore we used as input the (surrogate) model simulated head (M_h) rather than time. In summary, the input of the error model is

$\mathbf{y} = \{\mathbf{u}, M_h\}$. Because $\mathbf{u}$ and M_h are of different magnitudes and units, they were linearly scaled to a range of $[0, 1]$. It is worth mentioning that other relevant information, such as depth to groundwater and precipitation, can also be incorporated into the inputs. This might lead to an even more robust error model for making forecast under changing conditions and will be explored in future studies.

As discussed in Section 5.5, the confounding (interaction) between the physical model parameters and error model could potentially lead to identifiability issues [13, 70]. To constrain the error model so that it is not overfitting, we specify the prior of the drawdown error model such that it "encourages" the model structural error to be zero. In this way, the error model takes the compensation role only when supported by the data. More specifically, we specify the prior of head error model as a Gaussian Process with constant zero mean. Similarly as in Section 5.5, an isotropic squared exponential covariance function is used to enforce smoothness and reduce confounding between model structural error and parameters. The GP error model has two hyperparameters: characteristic scale length λ_h and standard deviation σ_h. The scale length hyperparameter represents the degree of correlation in the space of GP input. We specify a uniform distribution on $(0, 1]$ for λ_h. Given that the inputs were scaled to the range of $[0, 1]$, the prior has a loose upper bound. The standard deviation represents the amount of model structural error we would accept. A larger σ_h allows the error model to take on more compensation role of the model structural error. We specify a uniform distribution on $(0, 20]$ for σ_h. The upper limit 20 ft. suggests that *a priori*, the bias is unlikely to exceed $20 \times 1.96 = 39.2ft$, which is the 0.975-th quantile of a normal distribution $N(0, 20^2)$. In this case study, the posterior distribution is not sensitive to the choice of prior distribution, as long as the prior covers a fairly wide range. The main reason is because the calibration data provide much more information to constrain the posterior.

In addition to the 38 MODFLOW model parameters (θ) and 2 GP hyperparameters, a likelihood parameter σ_ϵ is jointly inferred. Variogram analysis on head residuals revealed a

nugget effect that is larger than the magnitude of head measurement error. For drawdown observations, the nugget σ_ϵ, SVR surrogate error, and measurement error combined represent the aleatoric error that the error model cannot capture. For stream gain-and-loss, on the other hand, the aleatoric error consists of measurement error and the SVR surrogate error; both terms are calculated before calibration and fixed. We specify a uniform distribution on $(0, 10]$ as the prior of σ_ϵ. For the 38 MODFLOW parameters, the same prior distributions as in Section 6.4 are used.

In total, Bayesian calibration was carried out to infer the joint posterior distribution of 41 parameters, using head and streamflow gain-and-loss during the calibration period (October 1995 to September 2004). The process is similar as described in Section 5.5. The DREAM-ZS runtime settings were configured following the recommendations in [85]. Ten Markov chains were used to generate 1,600 samples from the joint posterior distribution of 41 parameters after convergence was determined based on the $\hat{R}$ statistic [30], visual inspection of trace plots and other diagnostics [18]. As burn-in, 80,000 samples were discarded.

In the prediction phase, the Gaussian process error model uses as input $\mathbf{y}^* = [\mathbf{u}, M_h^*]$. The Bayesian framework yields an ensemble of head predictions $\mathbf{h}_i^* = \mathbf{M}_h^*(\theta_i) + \mathbf{b}_{h,i}^* + \epsilon_{h,i}, i = 1, ..., 1,600$. Here, $\mathbf{h}_i^*$ denotes groundwater head varying in both space and time; $\mathbf{b}_{h,i}^*$ is a vector drawn from the GP error model posterior; $\epsilon_{h,i}$ is randomly drawn from a normal distribution $N(0, \sigma_{\epsilon,i}^2 I + \Sigma_h)$, where Σ_h is a diagonal matrix with head measurement error as diagonal entries. It is worth mentioning that the head measurement error could vary in space and time depending on the accuracy of site land surface altitude, accuracy of depth to groundwater measurement and site status (e.g. if pumped recently) [39]. The Bayesian posterior of prediction can then be estimated by collecting the realizations in the ensemble, and the posterior mean is given by $\bar{\mathbf{h}}^*$.

The groundwater head forecasts $\mathbf{h}_i^*$ could be evaluated using the surrogate models if surro-

gate models for prediction quantities have been trained following the procedures described in Section 6.3. When using surrogate models to make a forecast, more parameter samples can be used to improve the accuracy of Monte Carlo posterior mean. Here we take a more straightforward approach and run the SVRP MODFLOW model to compute predictions. This is feasible because the model runs using different sets of parameters can be executed in parallel. The MODFLOW model is evaluated using $1,600$ samples drawn with DREAM-ZS. With $1,600$ samples, the calculated posterior mean has an error rate of order $O(1/\sqrt{1600})$ (according to the Central Limit Theorem), which is acceptable in this case. For modeling problems with higher accuracy requirement, more posterior samples may be needed at the expense of increased computational burden.

6.6 Results: Parameter Estimates

Figure 6.4 shows the marginal posterior distributions estimated by the classical Bayesian method without error model and the proposed Bayesian approach with error model for head. The names of the 38 SVR parameters were defined in Section 6.1. The priors of the parameters were specified as uniform distributions over a wide range (Sections 6.4 and 6.5). The priors are not displayed in Figure 6.4 because they would be a horizontal line close to the horizontal axes.

In this real-world case study, the "true" value of model parameters is unknown. Therefore, it is not possible to validate the correctness of parameter posteriors, as was possible for the synthetic case study in Chapter 5. For some parameters (KVSR-5, KVSR-7, KVSR-9, KVSR-10), the posterior is on the upper bound. The same phenomena occurred in PEST calibration, and the resulting parameter estimates are considered reasonable [39].

Given the wide priors, the parameter posteriors yielded by both the classical Bayesian and

proposed method can be considered as fairly constrained. Comparing the posteriors given by the two methods, it can be seen that the Bayesian with GP error model approach yields posteriors that are visibly different from the classical Bayesian method for many parameters, such as HK1-1, KVSR-1, SY-1, SY-2, SY-3. This reaffirms our finding in Section 5.6 that accounting for model structural error leads to substantially different parameter estimates.

6.7 Results: Prediction Performance

This section evaluates the performance of the classical Bayesian method and the proposed fully Bayesian approach in terms of predictive capability. Figure 6.5 and Table 6.2 assess how the simulated head compares with observation data. Figure 6.5 plots the difference between observations and posterior mean given by the classical and proposed Bayesian methods. For both the calibration and the validation periods, the proposed Bayesian method simulation error is smaller compared to classical Bayesian results. Table 6.2 summarizes the mean error, mean absolute percentage error, and root-mean-square-error (RMSE) statistics. The mean absolute percentage error is defined as the ratio of absolute error to observed value, averaged over all observations. It can be seen that the integration of an error model into Bayesian calibration effectively improved the accuracy of head prediction of the MODFLOW model, reducing the RMSE by over 50% for the validation period. The error model also removed most of the global bias, reducing the mean error from -2.08 ft. to 0.483 ft., and the mean absolute percentage error from -0.11% to 0.026%.

Figures 6.6-6.10 show head simulation results at four representative wells; the locations of wells are plotted in Figure 6.1. Figure 6.6 and Figure 6.8 plot the head prediction at two wells located in the Spokane Valley. For both wells, the classical Bayesian approach simulation results are biased and overconfident, in that the posterior mean deviates from observation data, and the prediction intervals do not encompass observations. With a GP

Table 6.2: Head simulation error of the classical Bayesian and proposed Bayesian with error model methods. Performance measures are calculated for the calibration period (October 1995 to September 2004) and the validation period (October 2004 to September 2005), respectively.

	Calibration		Validation	
	w/o error model	w/ error model	w/o error model	w/ error model
Mean error (ft)	-1.10	-0.0308	-2.08	0.483
Mean absolute percentage error	0.0595%	-0.0018%	-0.11%	0.0258%
RMSE (ft)	11.4	4.48	7.84	3.55

error model, the proposed Bayesian approach magnified the seasonal fluctuation, yielding head prediction that better matches the validation data. Well W3 is close to the Spokane River near Greenacres. The GP error model is not able to fully recover the observed head rise in autumn, which is caused by the rise in Spokane River stage as the Post Falls Dam opens its gates [39]. The performance could potentially be improved by incorporating relevant information, such as river stage, into the GP error model inputs.

Figure 6.7 shows the hydrograph at a well screened in layer 3. The drawdown in August is due to pumping at a nearby well [39]. The classical Bayesian method calibrated model predicted much higher drawdown than the observation, indicating that layer 3 may not be represented accurately in the SVRP model. The GP error model partially corrected this model structural error and yielded less biased head prediction.

Figure 6.9 shows that for a well in the southern Rathdrum Prairie, the classical Bayesian method yielded head prediction with fluctuation character that does not match measurements. A similar finding was observed for the PEST calibrated model, and a possible reason is that the temporal distribution of recharge for this region used in the model in 2004-2005 is not accurate [39]. Using a GP error model, we were able to improve the prediction accuracy, albeit with slightly overdampened head fluctuation.

For a representative well in northern Rathdrum Prairie (Figure 6.10), the classical Bayesian

method results in good fit to calibration data. However, in the prediction period the simulated fluctuations of the classical Bayesian calibrated model are somewhat larger than measurements. This is likely due to the linear assumption between the time for precipitation infiltration to reach groundwater table and depth of groundwater [39]. In Figure 6.10 bottom panel, it can be seen that the Bayesian approach with error model corrected this issue and yielded more accurate prediction.

6.8 Summary

In this chapter, the Bayesian with error model approach was further tested on a real-world regional groundwater flow model. We constructed computationally frugal surrogate models to emulate the response of the groundwater model with respect to its parameters. With this strategy, a 150-fold speedup was obtained, and Bayesian calibration of the complicated groundwater model becomes feasible.

In the SVRP case study, the model outputs of concern, namely groundwater head and stream gain-and-loss, possess relatively low degree of nonlinearity with respect to model parameters. When strong nonlinearity or even discontinuity is present, it may be challenging to achieve high surrogate accuracy [99].

The results are consistent with observations in the second case study (Chapter 5). More specifically, it was demonstrated that the Bayesian with error model method yielded parameter posterior pdfs that are substantially different from posteriors obtained using the classical Bayesian that does not account for model structural error. As for prediction performance, not accounting for model structural error led to biased and overconfident head predictions. In contrast, integrating a GP error model effectively improves the prediction accuracy and yielded prediction intervals that are consistent with validation data.

This study considers the temporal prediction scenario; the groundwater model and error models are calibrated using head at all observation wells before October 2004 and make forecasts since October 2004 at the same well locations. In groundwater modeling practice, it is often desirable to predict head at an unsampled locations. When used under the spatial prediction scenario, the GP error model's bias correction capability may decrease, as the GP posterior reduces to essentially zero when the prediction location is outside of the correlation range from training wells. As discussed in our preceding work [98], in the SVRP case study the density of monitoring wells is not sufficient for spatial prediction in most parts of the basin. The spatial prediction capability of the Bayesian with error model method needs further investigation in more real-world case studies with denser observation network.

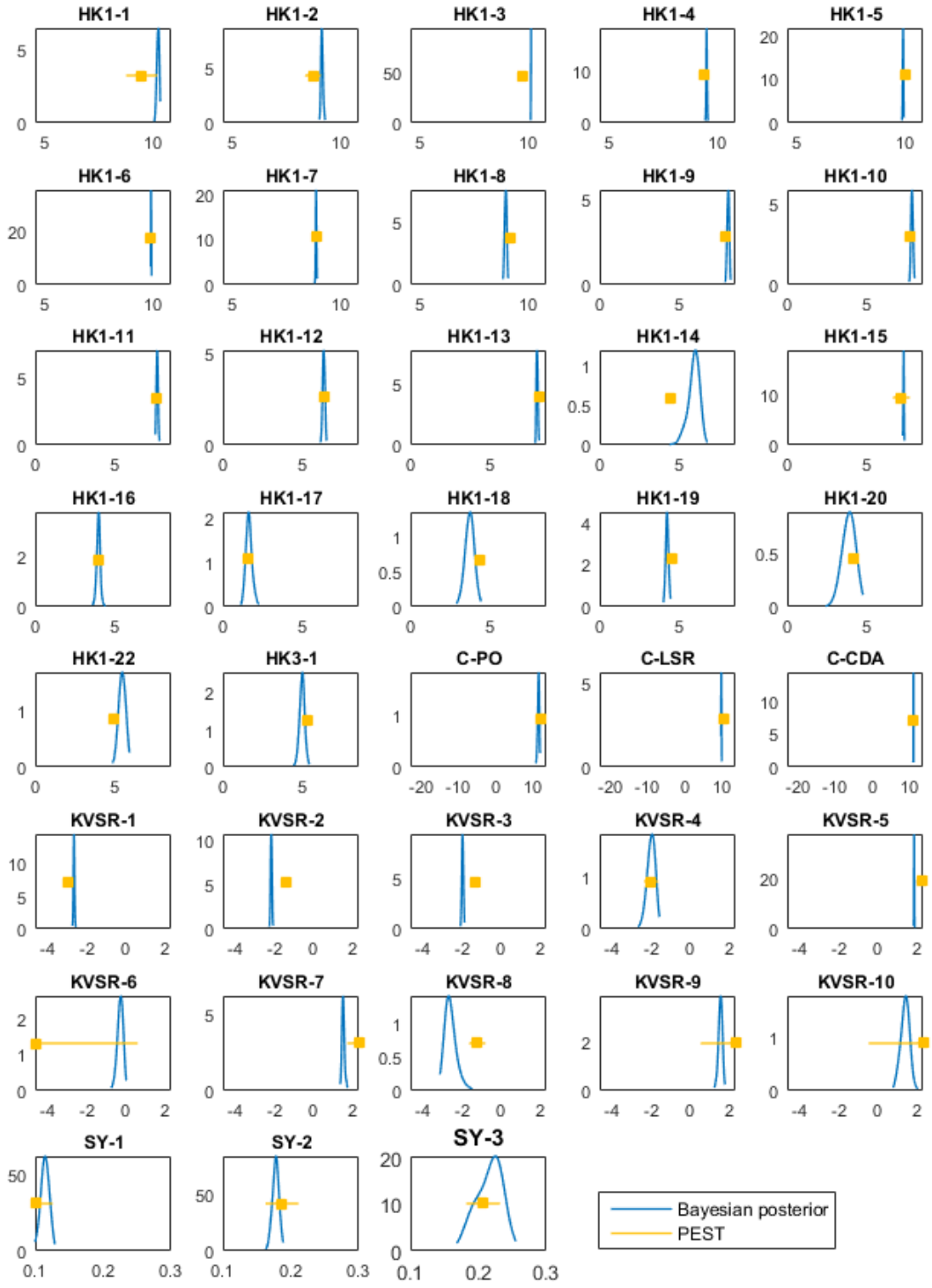

Figure 6.3: Comparison between PEST estimated parameters with 95% confidence interval (yellow bars) and Bayesian posterior marginal distributions (blue curves) for 38 parameters. The Bayesian results were obtained using the same observations as the PEST calibration. The hydraulic conductivity parameters are natural logarithm transformed. The ranges of x axes represent the lower and upper bounds enforced during calibration.

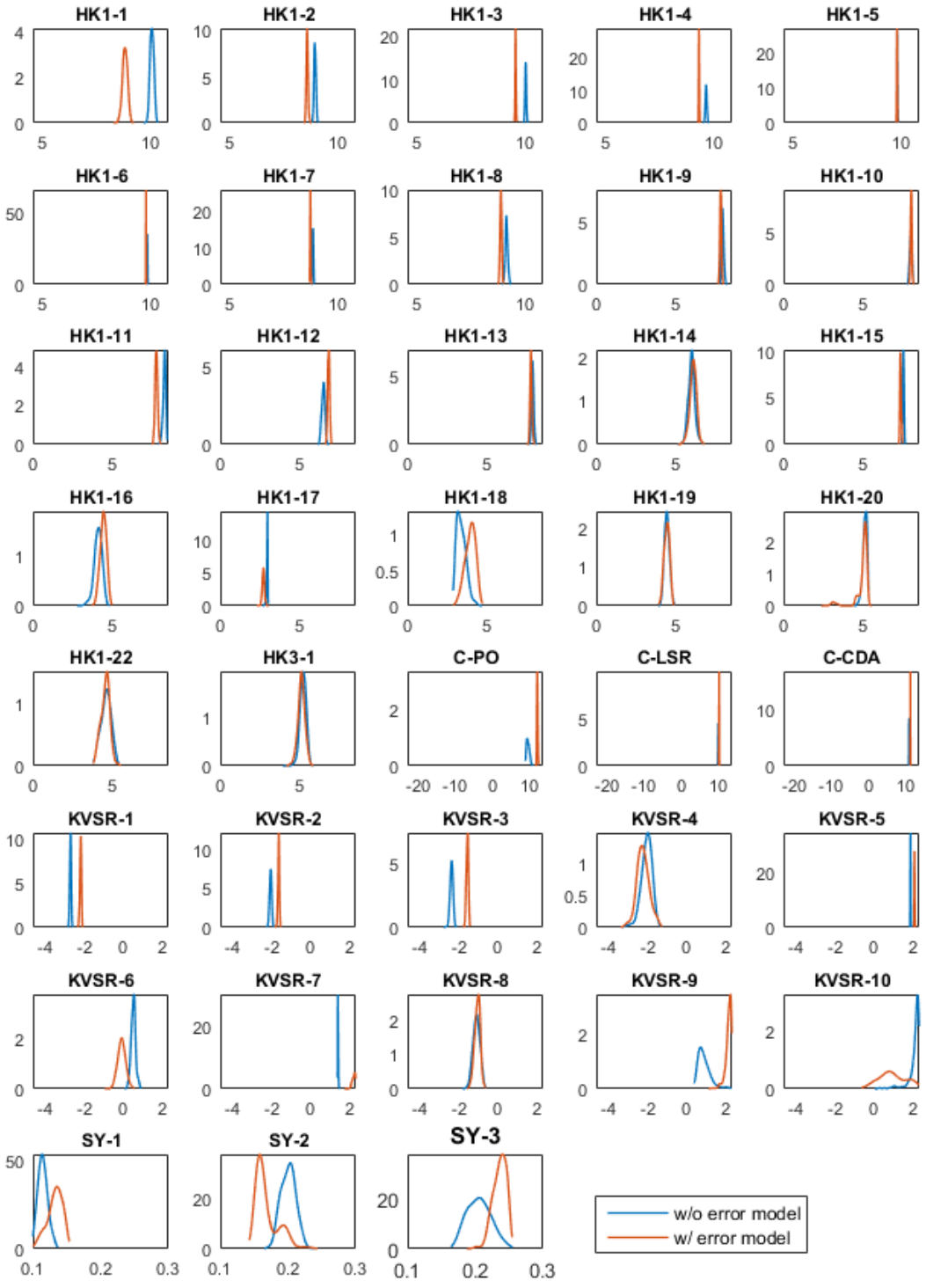

Figure 6.4: Marginal posterior distributions given by the classical Bayesian (blue) and the proposed Bayesian with error model approach (orange) of 38 SVRP model parameters. The hydraulic conductivity parameters are natural logarithm transformed. The ranges of x axes represent the lower and upper bounds enforced during calibration.

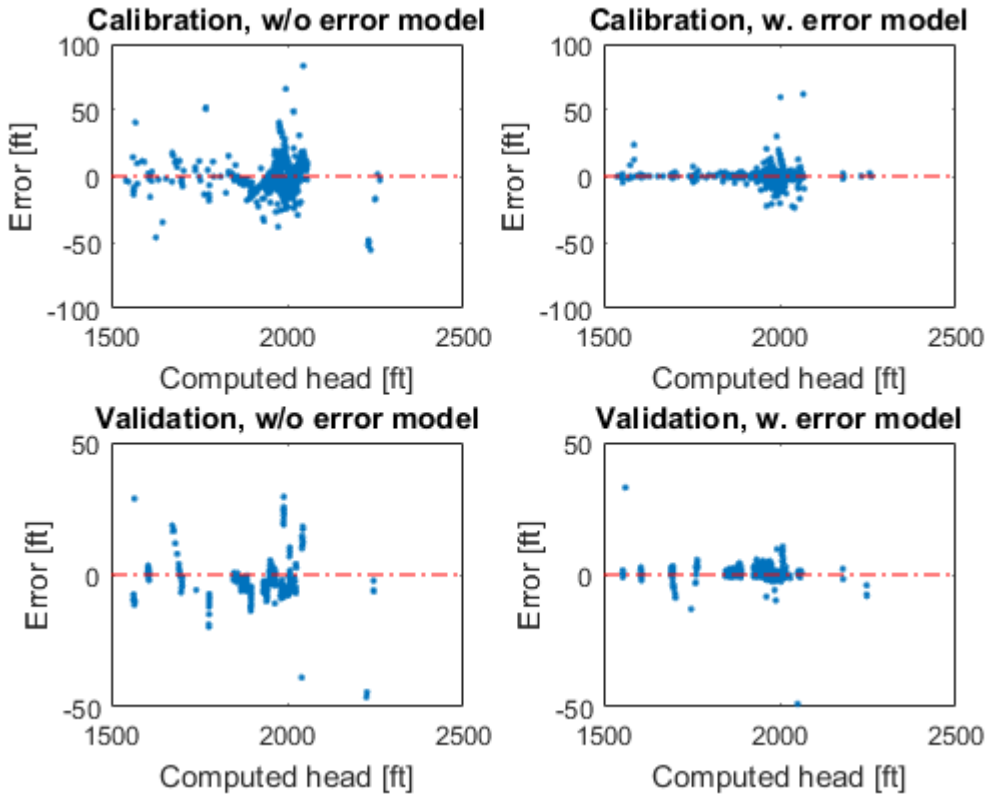

Figure 6.5: Head simulation error plotted versus posterior mean given by the classical Bayesian (left) and the proposed Bayesian with error model (right) methods. Calibration period is from October 1995 to September 2004, and the validation period spans October 2004 to September 2005.

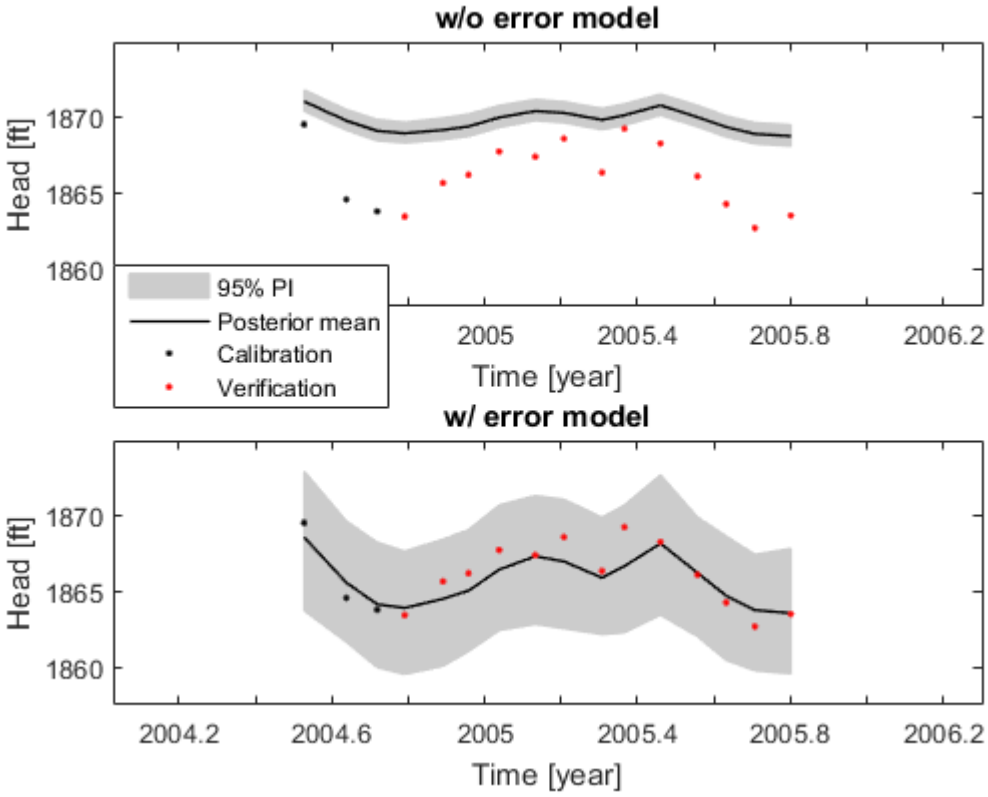

Figure 6.6: Head prediction at well W1 given by the classical Bayesian (up) and the proposed Bayesian with error model approach (bottom). Grey shades show 95% prediction interval, black dots are calibration data, and red dots are verification data. Well location is shown in Figure 6.1.

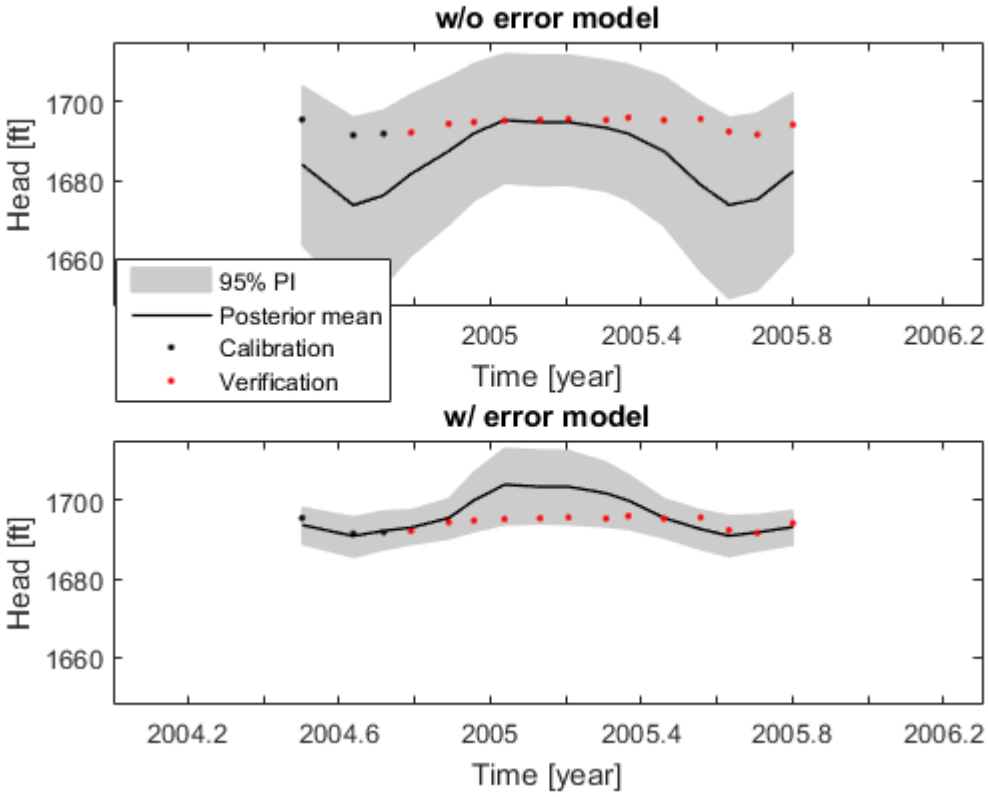

Figure 6.7: Head prediction at well W2 given by the classical Bayesian (up) and the proposed Bayesian with error model approach (bottom). Grey shades show 95% prediction interval, black dots are calibration data, and red dots are verification data. Well location is shown in Figure 6.1.

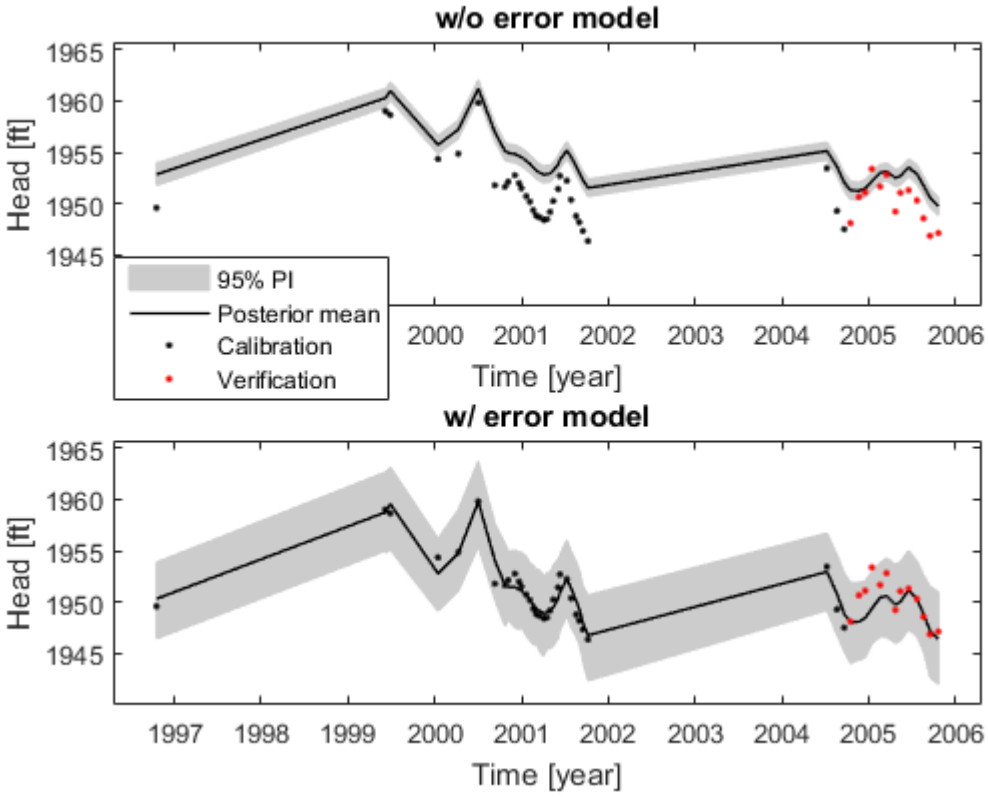

Figure 6.8: Head prediction at well W3 given by the classical Bayesian (up) and the proposed Bayesian with error model approach (bottom). Grey shades show 95% prediction interval, black dots are calibration data, and red dots are verification data. Well location is shown in Figure 6.1.

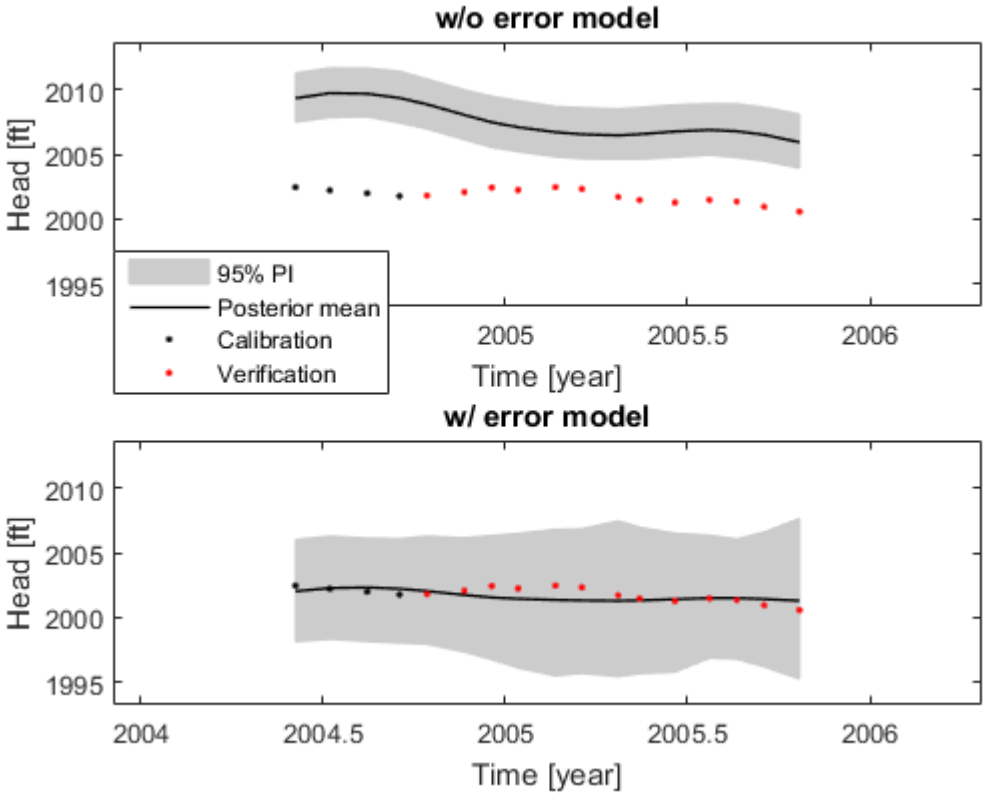

Figure 6.9: Head prediction at well W4 in southern Rathdrum Prairie, given by the classical Bayesian (up) and the proposed Bayesian with error model approach (bottom). Grey shades show 95% prediction interval, black dots are calibration data, and red dots are verification data. Well location is shown in Figure 6.1.

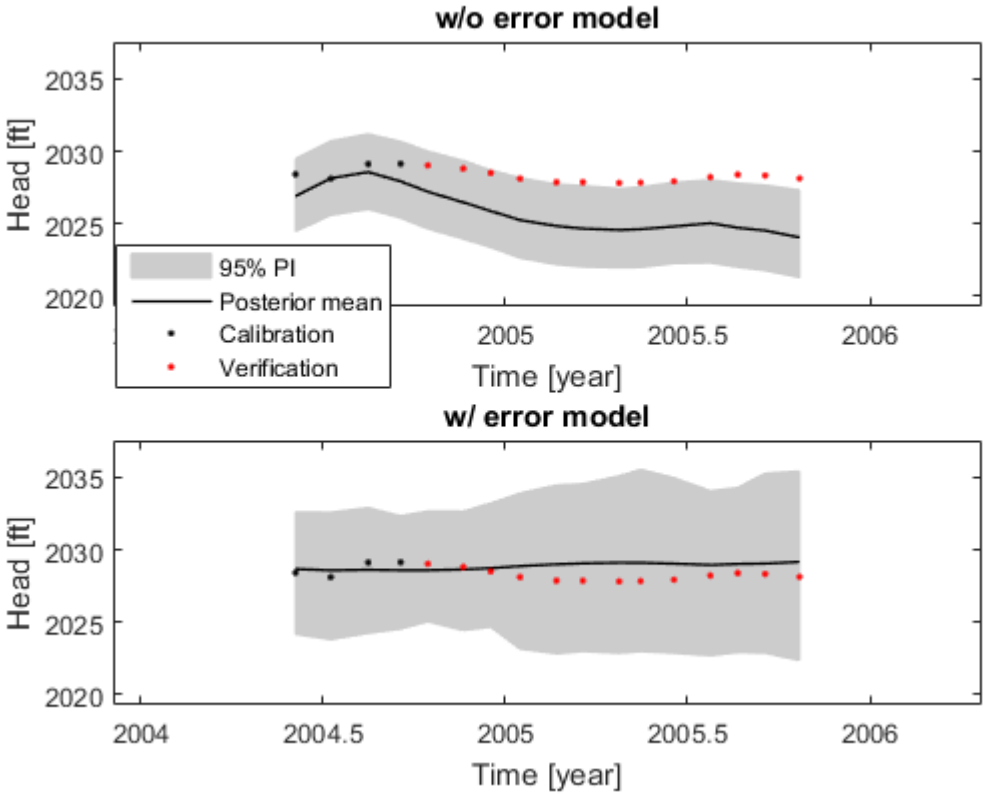

Figure 6.10: Head prediction at well W5 in northern Rathdrum Prairie, given by the classical Bayesian (up) and the proposed Bayesian with error model approach (bottom). Grey shades show 90% prediction interval, black dots are calibration data, and red dots are verification data. Well location is shown in Figure 6.1.

Chapter 7

SUMMARY AND CONCLUSIONS

We present a fully Bayesian calibration and uncertainty quantification framework tailored for groundwater models. The framework implements a marginalizing step to account for input data error when evaluating the likelihood. A data-driven error model is integrated into the Bayesian framework to correct for spatiotemporal groundwater model structural error.

We demonstrated the Bayesian approach using synthetic case studies of surface-ground water interaction under changing stress conditions as well as a real-world case study. In the first synthetic case study, we investigated the impact of errors in input data on calibration and prediction. The case study uses a virtual reality to generate synthetic observations of pumping-induced drawdown and stream depletion under "true" groundwater pumping and precipitation recharge rates. The synthetic observations are used to calibrate a model driven by biased pumping and recharge rates. The performance of the proposed marginalizing method is compared with classical Bayesian method that does not account for input error and the augmentation method, which estimates inputs together with model parameters during calibration. It is found that explicit treatment of errors in input data has substantial impact on the posterior distribution of groundwater model parameters. The classical Bayesian method yielded biased and overconfident prediction due to parameter compensation, while, by accounting for input data error, the proposed marginalizing method gave more accurate predictions.

One limitation of the marginalizing approach is that it only recognizes the uncertainty associated with input data, but does not correct for potential input bias. As a result, for quantities that are closely related to biased input data, such as streamflow gain-and-loss in this case study, the prediction made by the marginalizing approach may still be biased, although the prediction interval encompasses validation data. For these quantities, the augmentation method may yield more accurate prediction. However, as discussed in Section 4.4.3, the augmentation method may overly adjust input data to compensate for model structural error (when present). We recommend the marginalizing approach for modeling problems in which (1) substantial knowledge is available to specify a reasonably tight input distribution, and (2) augmentation method may not work due to identifiability issues. In the future, we will investigate a variant of the marginalizing method that jointly infers the hyperparameters of the input data error model with other parameters during the calibration process.

The second synthetic case study investigated the role of model structural error in calibration and prediction. The case study uses a virtual reality to generate synthetic observations; the observations are used to calibrate a simplified model which differs from the virtual reality. The differences reflect common types of model inadequacy in groundwater modeling practice, including simplified geometry, underrepresented heterogeneity of hydraulic conductivity, idealized stream cross section, inaccurate stream inflow and other aspects. While necessarily restricted by use of a specific complex numerical model to represent reality, the case study can nevertheless provide insights into the potential of presented approach to handle various types of commonly encountered model structural error.

In the second case study, we first demonstrated that both the conventional least squares regression (LSR) and classical Bayesian method yielded biased (and often overconfident) predictions under a scenario differing from the calibration period. It was then shown that integrating error models into Bayesian calibration reduces the degree of parameter compensation, leading to parameter posteriors that differ substantially from results not considering

model structural error. In terms of prediction accuracy, the Bayesian framework with error model delivered overall better performance than LSR and classical Bayesian methods.

We also presented a new recalibration strategy that aims to circumvent a well-known drawback of using an error model. Using the example of groundwater flow models, typical quantities of interest include groundwater head and flow interaction with surface water bodies. An error model adjusting the physically-based model simulated groundwater head may violate mass balance, because such physical constraints are not enforced on the data-driven error model. As a remedy, we "recalibrate" the groundwater model against prediction made by the Bayesian approach (which has been corrected by the error model). In this way, the recalibration strategy utilizes the Bayesian prediction while preserving mass conservation.

As the third case study, the Bayesian with error model approach was further tested on a real-world groundwater flow model. By using computationally frugal surrogate models as substitute and with the help of high performance computing (HPC) resources, Bayesian calibration becomes feasible even for a complicated regional-scale MODFLOW model with 38 parameters to be estimated. The surrogate models were constructed using support vector regression, a powerful machine learning algorithm.

Similarly as in the second case study, it was shown in the third case study that the joint inference of groundwater model parameters and model structural error led to parameter posterior pdfs that are substantially different from posteriors obtained using classical Bayesian that does not account for model structural error. When using the posterior parameter samples to make forecast beyond the calibration horizon, the classical Bayesian approach yielded biased and overconfident predictions. In contrast, the Bayesian with error model method delivered significantly more accurate prediction along with prediction intervals that are consistent with validation data. The results suggest that the proposed approach could be a robust method in real-world modeling problems. As a followup study, we will explore the poten-

tial of the proposed method in other modeling studies and under various prediction scenarios.

The presented framework constructs error models in an inductive, data-driven way, which sets this study apart from other work in the hydrology literature. In the second and third case studies, the error model inputs include simulation results of the physically-based groundwater model. This allows using the error model to extrapolate under conditions different from the calibration period. In addition, other relevant information not directly used in the development of the physically-based model can also be incorporated into the inputs to further improve the robustness of the error model. Selecting input for the error model is problem specific and should be guided by residual analysis. For example, feature selection techniques borrowed from information theory and statistical learning can detect dependency between model residual and other possibly relevant data [19, 97]. However, it should be noted that the predictability challenge in forecasting dynamic changes [48] still remains for the GP error model. This is because all machine learning methods including Gaussian process regression are essentially empirical. These inductive methods can be powerful tools in learning complex functional relationships, however they cannot predict dynamics that are not reflected in the training dataset.

A premise of our Bayesian method with error model is that model structural error is visible through the calibration process and cannot be fully compensated by adjusting parameters. The presented framework works the best with a parsimoniously parameterized model. For a highly parameterized model, model structural error is sometimes not discernible. It is noteworthy that using an error model for an unbiased model does not necessarily impair calibration results [70]. Furthermore, the overfitting issue can be alleviated by specifying priors that incorporate soft expert knowledge.

Bayesian inference often requires tens to hundreds of thousands of evaluations of the groundwater model. The interaction between groundwater model parameters and the error model

may deteriorate the convergence rate of sampling and further increase the computational expense. Beside the surrogate modeling strategy used in the third case study, the computational cost associated with Bayesian inference can be reduced in various ways, such as reducing the dimension of the parameter space and implementing multiple-chain, multiple-try and multiple-stage sampling algorithms [49, 57, 94]. Whether to implement a fully Bayesian calibration with its computational burden is a problem specific decision and can be informed by various diagnostics [36]. In general, we recommend the presented Bayesian approach when evidence supports the presence of input and model structural errors, yet model deficiency cannot be identified, and robust prediction uncertainty assessment is critical for post-modeling decision making.

Admittedly, like other empirical methods, the GP error models lack physical basis. Therefore, the presented framework cannot replace thoughtful modeling analysis and additional field observations toward improved understanding of specific groundwater systems. For example, posterior check of the error model posterior would identify regions (spatially or in the input space of the error model) that have significant predictive bias and high uncertainty. The gained information will help to locate spatial and temporal domains where the groundwater model does not perform satisfactorily. In addition, the data series associated with biased model simulation indicates that the related process may not be adequately represented in the groundwater model. These potential extensions will be further discussed in future work.

In the case studies, we used one GP error model with a simple covariance function (for each type of output) to emulate the model structural error lumped from various model deficiencies. Depending on specific applications, a modeler can use a more complex covariance function, which is a combination of several different kinds of simple covariance functions [67]. Each simple covariance function handles an individual property of the model structural error. Similarly, a mixture of GP models can be used to allow for more flexibility. Analysis

of posterior of hyperparameters of these covariance functions may shed light on the decomposition of model structural error contributed by various underlying processes at different time and spatial scales.

In summary, the results in the case studies highlight the importance of proper treatment of input data and model structural errors in circumstances where subsequent decision making and risk analysis require accurate prediction and uncertainty quantification. The Bayesian approach allows for disaggregation of uncertainty among various error sources. This could inform future model improvement and data collection efforts on how to best direct resources towards reducing predictive uncertainty. The presented Bayesian framework brings together the strength of physically-based groundwater models and inductive data-driven statistical learning techniques, and is in harmony with new trends towards increased data availability and promotion of environmental observatories. The presented framework can be used for subsurface solute transport models and other environmental modeling applications. Follow-up studies will further investigate the feasibility of joint inference of input and model structural errors, particularly for real-world modeling practice.

REFERENCES

[1] AJ Abebe and RK Price. Managing uncertainty in hydrological models using complementary models. *Hydrological sciences journal*, 48(5):679–692, 2003.

[2] Newsha K Ajami, Qingyun Duan, and Soroosh Sorooshian. An integrated hydrologic bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, 43(1), 2007.

[3] JG Arnold, DN Moriasi, PW Gassman, KC Abbaspour, MJ White, Raghavan Srinivasan, Chinnasamy Santhi, RD Harmel, Ann Van Griensven, MW Van Liew, et al. Swat: Model use, calibration, and validation. *Transactions of the ASABE*, 55(4):1491–1508, 2012.

[4] T. Asefa, M. Kemblowski, G. Urroz, and M. McKee. Support vector machines (SVMs) for monitoring network design. *Ground water*, 43(3):413–422, 2005.

[5] MJ Asher, BFW Croke, AJ Jakeman, and LJM Peeters. A review of surrogate models and their application to groundwater modeling. *Water Resources Research*, 51(8), 2015.

[6] Paul M Barlow and Stanley A Leake. Streamflow depletion by wells: Understanding and managing the effects of groundwater pumping on streamflow. Technical report, U.S. Geological Survey Circular 1376, 2012.

[7] Maria J Bayarri, James O Berger, Rui Paulo, Jerry Sacks, John A Cafeo, James Cavendish, Chin-Hsu Lin, and Jian Tu. A framework for validation of computer models. *Technometrics*, 49(2), 2007.

[8] K. Beven and A. Binley. The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, 6(3):279–298, 1992.

[9] K. Beven and J. Freer. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, 249(1):11–29, 2001.

[10] Keith J Beven, Paul J Smith, and Jim E Freer. So just why would a modeller choose to be incoherent? *Journal of hydrology*, 354(1):15–32, 2008.

[11] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[12] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[13] Jennỳ Brynjarsdóttir and Anthony O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30(11):114007, 2014.

[14] C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[15] V. Cherkassky and Y. Ma. Practical selection of *svm* parameters and noise estimation for *svm* regression. *Neural Networks*, 17(1):113–126, 2004.

[16] Richard L Cooley and Steen Christensen. Bias and uncertainty in regression-calibrated models of groundwater flow in heterogeneous media. *Advances in water resources*, 29(5):639–656, 2006.

[17] R.L. Cooley. *A theory for modeling ground-water flow in heterogeneous media*. US Dept. of the Interior, US Geological Survey, 2004.

[18] Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[19] Yonas K Demissie, Albert J Valocchi, Barbara S Minsker, and Barbara A Bailey. Integrating a calibrated groundwater flow model with error-correcting data-driven models to improve predictions. *Journal of Hydrology*, 364(3):257–271, 2009.

[20] A Dietzel and P Reichert. Bayesian inference of a lake water quality model by emulating its posterior density. *Water Resources Research*, 50(10):7626–7647, 2014.

[21] J. Doherty, L. Brebber, and P. Whyte. PEST: Model-independent parameter estimation user manual. Technical report, Watermark Computing, Corinda, Australia, 2010.

[22] J. Doherty and S. Christensen. Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resources Research*, 47(12), 2011.

[23] J. Doherty and D. Welter. A short exploration of structural noise. *Water Resources Research*, 46(5), 2010.

[24] John Doherty. Ground water model calibration using pilot points and regularization. *Groundwater*, 41(2):170–177, 2003.

[25] D Erdal, I Neuweiler, and JA Huisman. Estimating effective model parameters for heterogeneous unsaturated flow using error models for bias correction. *Water Resources Research*, 48(6), 2012.

[26] Guillaume Evin, Mark Thyer, Dmitri Kavetski, David McInerney, and George Kuczera. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research*, 50(3):2350–2375, 2014.

[27] DR Feldman, KN Liou, RL Shia, and YL Yung. On the information content of the thermal infrared cooling rate profile from satellite instrument measurements. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 113(D11), 2008.

[28] Michael Fienen, R Hunt, D Krabbenhoft, and Tom Clemo. Obtaining parsimonious hydraulic conductivity fields using head and transport observations: A Bayesian geostatistical parameter estimation approach. *Water resources research*, 45(8), 2009.

[29] Michael N Fienen, Marco D'Oria, John E Doherty, and Randall J Hunt. Approaches in highly parameterized inversion: bgaPEST, a Bayesian geostatistical approach implementation with PEST: documentation and instructions. Technical report, US Geological Survey, 2013.

[30] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[31] J Jaime Gómez-Hernánez, Andrés Sahuquillo, and José E Capilla. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data–1. theory. *Journal of Hydrology(Amsterdam)*, 203(1):167–174, 1997.

[32] M. Goswami, KM O'Connor, KP Bhattarai, AY Shamseldin, et al. Assessing the performance of eight real-time updating models and procedures for the brosna river. *Hydrology and Earth System Sciences*, 9(4):394–411, 2005.

[33] MA Gusyev, HM Haitjema, CP Carlson, and MA Gonzalez. Use of nested flow models and interpolation techniques for science-based management of the Sheyenne National Grassland, North Dakota, USA. *Groundwater*, 51(3):414–420, 2013.

[34] Richard W Healy. *Estimating groundwater recharge*. Cambridge University Press, 2010.

[35] Dave Higdon, Marc Kennedy, James C Cavendish, John A Cafeo, and Robert D Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

[36] Mary C Hill, Dmitri Kavetski, Martyn Clark, Ming Ye, Mazdak Arabi, Dan Lu, Laura Foglia, and Steffen Mehl. Practical use of computationally frugal model analysis methods. *Groundwater*, 2015.

[37] MC Hill and CR Tiedeman. Effective calibration of groundwater models, with analysis of data, sensitivities, predictions, and uncertainty. *John Wiely, New York*, 2007.

[38] M Honti, C Stamm, and P Reichert. Integrated uncertainty assessment of discharge predictions with a statistical error model. *Water Resources Research*, 49(8):4866–4884, 2013.

[39] P.A. Hsieh, M.E.Barber, B.A.Contor, A.Hossain, G.S.Johnson, J.L.Jones, and A.H.Wylie. Ground-water flow model for the Spokane Valley-Rathdrum Prairie Aquifer, Spokane County, Washington, and Bonner and Kootenai Counties, Idaho. Technical report, USGS Scientific Investigations Report 2007-5044, 2007.

[40] David Huard and Alain Mailhot. Calibration of hydrological model gr2m using bayesian uncertainty analysis. *Water Resources Research*, 44(2), 2008.

[41] David Insua, Fabrizio Ruggeri, and Mike Wiper. *Bayesian analysis of stochastic process models*, volume 978. Wiley. com, 2012.

[42] M Kanevski, Roman Parkin, Aleksey Pozdnukhov, Vadim Timonin, Michel Maignan, V Demyanov, and Stéphane Canu. Environmental data mining and modeling based on machine learning algorithms and geostatistics. *Environmental Modelling & Software*, 19(9):845–855, 2004.

[43] Dmitri Kavetski, George Kuczera, and Stewart W Franks. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, 42(3), 2006.

[44] Dmitri Kavetski, George Kuczera, and Stewart W Franks. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, 42(3), 2006.

[45] Elizabeth H Keating, John Doherty, Jasper A Vrugt, and Qinjun Kang. Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. *Water Resources Research*, 46(10), 2010.

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

[47] George Kuczera, Dmitri Kavetski, Stewart Franks, and Mark Thyer. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, 331(1):161–177, 2006.

[48] Praveen Kumar. Typology of hydrologic predictability. *Water Resources Research*, 47(3), 2011.

[49] Eric Laloy, Bart Rogiers, Jasper A Vrugt, Dirk Mallants, and Diederik Jacques. Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. *Water Resources Research*, 2013.

[50] SA Leake, JP Hoffmann, and Jesse E Dickinson. Numerical ground-water change model of the c aquifer and effects of ground-water withdrawals on stream depletion in selected reaches of Clear Creek, Chevelon Creek, and the Little Colorado River, northeastern Arizona. Technical Report 2005-5277, U.S. Geological Survey, 2005.

[51] Karl K Lee and John C Risley. Estimates of ground-water recharge, base flow, and stream reach gains and losses in the willamette river basin, oregon. Technical report, US Department of the Interior, US Geological Survey, 2002.

[52] Feng Liang, Kai Mao, Ming Liao, Sayan Mukherjee, and Mike West. Nonparametric Bayesian kernel models. *Department of Statistical Science, Duke University, Discussion Paper*, pages 07–10, 2007.

[53] Dan Lu, Ming Ye, and Mary C Hill. Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification. *Water Resources Research*, 48(9), 2012.

[54] Dan Lu, Ming Ye, Philip D Meyer, Gary P Curtis, Xiaoqing Shi, Xu-Feng Niu, and Steve B Yabusaki. Effects of error covariance structure on estimation of model averaging weights and predictive performance. *Water Resources Research*, 49(9):6029–6047, 2013.

[55] Regan R.S. Hay L.E. Viger R.J. Webb R.M.T. Payn R.A. Markstrom, S.L. and J.H. LaFontaine. PRMS-IV, the precipitation-runoff modeling system, version 4: U.s. geological survey techniques and methods, book 6. Technical Report 2328-7055, U.S. Geological Survey, 2015.

[56] Youssef Marzouk and Dongbin Xiu. A stochastic collocation approach to Bayesian inference in inverse problems. *Communications in Computational Physics*, 6:826–847, 2009.

[57] Youssef M Marzouk and Habib N Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.

[58] V. McKusick. Final report for the special master with certificate of adoption of RRCA groundwater model. Technical report, State of Kansas v. State of Nebraska and State of Colorado, in the Supreme Court of the United States, 2003.

[59] Catherine Moore and John Doherty. Role of the calibration process in reducing model predictive error. *Water Resources Research*, 41(5), 2005.

[60] Pradeep Mugunthan and Christine A Shoemaker. Assessing the impacts of parameter uncertainty for computationally expensive groundwater models. *Water Resources Research*, 42(10), 2006.

[61] SP Neuman. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment*, 17(5):291–305, 2003.

[62] SP Neuman and PJ Wierenga. A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites. Technical report, University of Arizona Report NUREG/CR-6805, 2003.

[63] F. Pianosi, L. Raso, and L. Raso. Dynamic modelling of predictive uncertainty by regression on absolute errors. *Water Resources Research*, 48(3), 2012.

[64] Natesh S Pillai, Qiang Wu, Feng Liang, Sayan Mukherjee, and Robert L Wolpert. Characterizing the function space for Bayesian kernel models. *Journal of Machine Learning Research*, 8(8), 2007.

[65] EP Poeter, MC Hill, D Lu, and SW Mehl. Ucode_2014, with new capabilities to define parameters unique to predictions, calculate weights using simulated values, estimate parameters with svd, and evaluate uncertainty with mcmc. *International Ground Water Modeling Center Report, to appear*, 2014.

[66] David E Prudic, Leonard F Konikow, and Edward R Banta. A new streamflow-routing (SFR1) package to simulate stream-aquifer interaction with modflow-2000. Technical report, US Department of the Interior, US Geological Survey, 2004.

[67] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. The MIT Press, Cambridge, MA, USA, 2006.

[68] K. Rasouli, W.W. Hsieh, and A.J. Cannon. Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, 414-415:284–293, 2011.

[69] J.C. Refsgaard, J.P. Van der Sluijs, J. Brown, and P. Van der Keur. A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources*, 29(11), 2006.

[70] P Reichert and N Schuwirth. Linking statistical bias description to multiobjective model calibration. *Water Resources Research*, 48(9):W09543, 2012.

[71] Nicolas Remy, Alexandre Boucher, and Jianbing Wu. *Applied geostatistics with SGeMS: a user's guide*. Cambridge University Press, 2009.

[72] Benjamin Renard, Dmitri Kavetski, George Kuczera, Mark Thyer, and Stewart W Franks. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46(5), 2010.

[73] Mojtaba Sadegh and Jasper A Vrugt. Approximate Bayesian computation using markov chain monte carlo simulation: DREAM (ABC). *Water Resources Research*, 50(8):6767–6787, 2014.

[74] Bridget R Scanlon, Claudia C Faunt, Laurent Longuevergne, Robert C Reedy, William M Alley, Virginia L McGuire, and Peter B McMahon. Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley. *Proceedings of the national academy of sciences*, 109(24):9320–9325, 2012.

[75] Gerrit Schoups and Jasper A Vrugt. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, 46(10), 2010.

[76] Xiaoqing Shi, Ming Ye, Gary P Curtis, Geoffery L Miller, Philip D Meyer, Matthias Kohler, Steve Yabusaki, and Jichun Wu. Assessment of parametric uncertainty for groundwater reactive transport modeling. *Water Resources Research*, 50(5):4416–4439, 2014.

[77] Alexander J Smola and Bernhard Schölkopf. Bayesian kernel methods. In S. Mendelson and A.J. Smola, editors, *Advanced lectures on machine learning*, pages 65–117. Springer, 2003.

[78] D.P. Solomatine and D.L. Shrestha. A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Research*, 45(1), 2009.

[79] Jery R Stedinger, Richard M Vogel, Seung Uk Lee, and Rebecca Batchelder. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, 44(12), 2008.

[80] Claire R Tiedeman and Christopher T Green. Effect of correlated observation error on parameters, predictions, and uncertainty. *Water Resources Research*, 49(10):6339–6355, 2013.

[81] Matthew James Tonkin and John Doherty. A hybrid regularized inversion methodology for highly parameterized environmental models. *Water Resources Research*, 41(10), 2005.

[82] Matthew James Tonkin and John Doherty. Calibration-constrained Monte Carlo analysis of highly parameterized models using subspace techniques. *Water Resources Research*, 45(12), 2009.

[83] Tadeusz J Ulrych, Mauricio D Sacchi, and Alan Woodbury. A Bayes tour of inversion: A tutorial. *Geophysics*, 66(1):55–69, 2001.

[84] V.N. Vapnik. *The nature of statistical learning theory*. Springer, New York, 1995.

[85] Jasper A Vrugt, Cajo JF Ter Braak, Martyn P Clark, James M Hyman, and Bruce A Robinson. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, 44(12), 2008.

[86] Jasper A Vrugt, Cajo JF Ter Braak, Hoshin V Gupta, and Bruce A Robinson. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026, 2009.

[87] Jasper A Vrugt, CJF Ter Braak, CGH Diks, Bruce A Robinson, James M Hyman, and Dave Higdon. Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3):273–290, 2009.

[88] Brian J Wagner and Marshall W Gannett. Evaluation of alternative groundwater-management strategies for the bureau of reclamation klamath project, oregon and california. Technical Report 2014-5054, U.S. Geological Survey, 2014.

[89] Chen Wang, Qingyun Duan, Wei Gong, Aizhong Ye, Zhenhua Di, and Chiyuan Miao. An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environmental Modelling & Software*, 60:167–179, 2014.

[90] AH Weerts, HC Winsemius, and JS Verkade. Estimation of predictive hydrological uncertainty using quantile regression: examples from the national flood forecasting system (England and Wales). *Hydrology and Earth System Sciences*, 15(1):255–265, 2011.

[91] Jeremy T White, John E Doherty, and Joseph D Hughes. Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resources Research*, 50(2):1152–1173, 2014.

[92] C. K. I. Williams. Guassian processes. In Michael A Arbib, editor, *The handbook of brain theory and neural networks*. The MIT press, Cambridge,MA, second edition, 2003.

[93] Carl Wunsch and Patrick Heimbach. Practical global oceanic state estimation. *Physica D: Nonlinear Phenomena*, 230(1):197–208, 2007.

[94] Hua Xie, J Wayland Eheart, Yuguo Chen, and Barbara A Bailey. An approach for improving the sampling efficiency in the bayesian calibration of computationally expensive simulation models. *Water Resources Research*, 45(6), 2009.

[95] T. Xu. Use of data-driven models to improve prediction of physically based groundwater models. Master's thesis, University of Illinois at Urbana-Champaign, 2012.

[96] Tianfang Xu and Albert J Valocchi. A Bayesian approach to improved calibration and prediction of groundwater models with structural error. *Water Resources Research*, 51(11):9290–9311, 2015.

[97] Tianfang Xu and Albert J Valocchi. Data-driven methods to improve baseflow prediction of a regional groundwater model. *Computers & Geosciences*, 2015.

[98] Tianfang Xu, Albert J Valocchi, Jaesik Choi, and Eyal Amir. Use of machine learning methods to reduce predictive error of groundwater models. *Groundwater*, 52(3):448–460, 2014.

[99] Xiankui Zeng, Ming Ye, John Burkardt, Jichun Wu, Dong Wang, and Xiaobin Zhu. Evaluating two sparse grid surrogates and two adaptation criteria for groundwater bayesian uncertainty quantification. *Journal of Hydrology*, 535:120–134, 2016.

[100] Guannan Zhang, Dan Lu, Ming Ye, Max Gunzburger, and Clayton Webster. An adaptive sparse-grid high-order stochastic collocation method for bayesian inference in groundwater reactive transport modeling. *Water Resources Research*, 49(10):6871–6892, 2013.