AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR019831

Key Points:
- We present an error-explicit Bayesian uncertainty quantification method that accounts for spatiotemporal model structural error
- Tested on a real-world groundwater flow model, the error-explicit method gives more accurate prediction than neglecting structural error
- The method provides an assessment of prediction uncertainty contributed from various sources

Correspondence to:
A. J. Valocchi,
valocchi@illinois.edu

Citation:
Xu, T., A. J. Valocchi, M. Ye, and F. Liang (2017), Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model, *Water Resour. Res.*, *53*, doi:10.1002/2016WR019831.

Received 22 SEP 2016
Accepted 28 APR 2017
Accepted article online 8 MAY 2017

Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model

Tianfang Xu[1,2], Albert J. Valocchi[1], Ming Ye[3], and Feng Liang[4]

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, [2]Now at Department of Earth and Environmental Sciences, Michigan State University, East Lansing, Michigan, USA, [3]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA, [4]Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

Abstract Groundwater model structural error is ubiquitous, due to simplification and/or misrepresentation of real aquifer systems. During model calibration, the basic hydrogeological parameters may be adjusted to compensate for structural error. This may result in biased predictions when such calibrated models are used to forecast aquifer responses to new forcing. We investigate the impact of model structural error on calibration and prediction of a real-world groundwater flow model, using a Bayesian method with a data-driven error model to explicitly account for model structural error. The error-explicit Bayesian method jointly infers model parameters and structural error and thereby reduces parameter compensation. In this study, Bayesian inference is facilitated using high performance computing and fast surrogate models (based on machine learning techniques) as a substitute for the computationally expensive groundwater model. We demonstrate that with explicit treatment of model structural error, the Bayesian method yields parameter posterior distributions that are substantially different from those derived using classical Bayesian calibration that does not account for model structural error. We also found that the error-explicit Bayesian method gives significantly more accurate prediction along with reasonable credible intervals. Finally, through variance decomposition, we provide a comprehensive assessment of prediction uncertainty contributed from parameter, model structure, and measurement uncertainty. The results suggest that the error-explicit Bayesian approach provides a solution to real-world modeling applications for which data support the presence of model structural error, yet model deficiency cannot be specifically identified or corrected.

1. Introduction

As numerical groundwater models are increasingly used to inform water resources management decisions and policies under future scenarios of climate and land use change, there is a need to ensure accuracy and quantify the intrinsic uncertainties of these models. In the last two decades, there have been significant advances in least square regression-based calibration and uncertainty quantification techniques [e.g., *Doherty et al*., 2010; *Hill and Tiedeman*, 2007; *Lin et al*., 2016; *Tonkin et al*., 2007]. However, common practices often focus on parameter uncertainty while neglecting model structural error, which is ubiquitous in groundwater models due to simplified or even incorrect conceptualiztion of the real hydrogeologic system [*Cooley*, 2004; *Gupta et al*., 2012; *Refsgaard et al*., 2006].

Model structural error may lead to systematic bias and underestimated uncertainty in the output for a calibrated model [*Ye et al*., 2004]. In addition, various studies have highlighted the importance of apportioning overall prediction uncertainty among different sources thereby identifying the main causes of uncertainty, which leads to improving the modeling process, and eventually reducing prediction uncertainty [*Baroni and Tarantola*, 2014; *Freni and Mannina*, 2010; *Liu and Gupta*, 2007; *Renard et al*., 2010]. Meaningful uncertainty disaggregation requires adequate characterization of error sources [*Renard et al*., 2010; *Salamon and Feyen*, 2010], which is particularly challenging for unknown model structural error.

Multimodel methods provide a way to account for conceptualization uncertainty by recognizing and combining multiple competing model representations (hypotheses) [*Clark et al*., 2015; *Neuman*, 2003; *Schöniger*

et al., 2015; *Ye et al*., 2004]. Meanwhile, a parallel line of research has been focusing on statistically characterizing the systematic error in model residuals, i.e., the difference between observations and corresponding model simulations. Because of model structural error, the model residual may exhibit temporal and spatial correlation structures [*Doherty and Welter*, 2010; *Xu et al*., 2014]. For least squares regression-based calibration, correlated residuals can be handled by using a full error covariance matrix (i.e., with nonzero off-diagonal entries) in the objective function [*Lu et al*., 2013]. In Bayesian calibration, customized likelihood functions are used to characterize model residuals that are non-Gaussian, biased, skewed, heteroscedastic, and correlated [*Beven and Freer*, 2001; *Erdal et al*., 2012; *Nearing et al*., 2016; *Schoups and Vrugt*, 2010; *Shi et al*., 2014]. The above methods have been applied to various fields, including rainfall-runoff, unsaturated flow and groundwater reactive transport modeling.

However, existing applications deal with time series data and rely on relatively simple autoregressive models to describe the temporal correlation of model residuals. For groundwater models that simulate spatio-temporally varying quantities such as groundwater piezometric head and contaminant concentration, it is more challenging to configure the form of the full error covariance matrix or the likelihood function, as the correlation structure due to model structural error is typically unknown.

Xu and Valocchi [2015a] presented a Bayesian approach that incorporates a data-driven error model to account for model structural error. The approach is adapted from the general framework proposed by *Kennedy and O'Hagan* [2000] and is tailored for groundwater applications. The error model is based on Gaussian process regression, a machine learning algorithm that has become popular in the last decade [*Bishop and Nasrabadi*, 2006; *Liang et al*., 2007; *Rasmussen and Williams*, 2006]. The Bayesian framework jointly infers model parameters and the error model. In this way, it provides a complete assessment of uncertainties contributed from parameters, model structure, and measurements. The Gaussian process error model corrects for model structural error revealed by the model residual [*Kennedy and O'Hagan*, 2000], therefore reducing the risk of parameters being overly adjusted to compensate for model structural error [*Doherty and Welter*, 2010]. In addition, the error models are constructed in an inductive, data-driven way such that they can learn functional relationships between output (i.e., model structural error) and a selected set of inputs, or explanatory variables. The 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. By learning from the historical error of the groundwater model, the Gaussian process error model can be used for forecasting. *Xu and Valocchi* [2015a] demonstrated the effectiveness of Gaussian process error model for a hypothetical test problem. More studies are needed to further examine the robustness of this method for real-world field problems, which possess both theoretical and computational challenges beyond the scope of a hypothetical test case.

In Bayesian inference, sampling from the posterior distribution often requires hundreds of thousands of model evaluation. For complex models having long evaluation time, the computational cost of MCMC sampling may be prohibitively high. In addition, the joint inference of model parameters and structural error may decrease the convergence rate of sampling and further increase the computational expense, due to the increased dimension of sampling space and the interaction among different uncertainty sources [*Xu and Valocchi*, 2015a]. For Bayesian calibration of complex models, computationally frugal surrogate models can be used as a substitute for the original model [*Asher et al*., 2015; *Razavi et al*., 2012].

Surrogate models can be constructed from the original model by reducing numerical resolution, relaxing convergence tolerance, and/or omitting processes [*Asher et al*., 2015]. However, the parameters of reduced-order models may not be defined exactly the same as in the original model, making the inference of the 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, radial basis functions were used to approximate the Nash-Sutcliffe index [*Mugunthan and Shoemaker*, 2006] and Bayesian posterior [*Bliznyuk et al*., 2012]. Similarly, Gaussian process regression was used to emulate the logarithm of likelihood for the calibration of a rainfall-runoff model [*Wang et al*., 2014].

Alternatively, surrogate models can be built to emulate the system state variables or outputs such as groundwater head. State-variable surrogates may outperform log likelihood surrogates when the model state response surface is smoother than that of the log likelihood in parameter space [*Zeng et al*., 2016]. Generalized polynomial chaos expansion (gPC) [*Marzouk and Xiu*, 2009], sparse grid, and Gaussian process

regression methods have been used to construct state-variable surrogate models in various hydrology applications including groundwater modeling [*Deman et al.*, 2016; *Laloy et al.*, 2013; *Zeng et al.*, 2016; *Zhang et al.*, 2016].

In this study, we use machine learning algorithms, namely random forest [*Breiman*, 2001] and support vector regression (SVR) [*Vapnik*, 1995], to construct surrogate models. The surrogate models use as input the parameters that need to be calibrated and emulate the outputs of the groundwater model, such as head and stream gain-and-loss at various times and locations. In the calibration process, the surrogates are used jointly with Gaussian process error model to evaluate likelihood.

The primary goal of this paper is to investigate the impact of model structural error on model calibration and prediction in the context of real-world groundwater modeling. Using a fast surrogates-facilitated Bayesian method, we construct a data-driven error model to explicitly account for model structural error. While it has been noted in the literature that the interactions among different uncertainty sources could render joint inference methods less robust than postprocessor approaches [*Evin et al.*, 2014], we find in the case study that cautious specification of error model priors helps alleviate the identifiability issue due to interaction, delivering reasonable uncertainty analysis performance even for a complicated regional groundwater model with 38 parameters to be estimated. The error-explicit Bayesian method allows for a complete assessment of prediction uncertainty contributed by uncertainties in measurement, model structure and parameters. The second focus of this study is, therefore, to perform variance decomposition analysis using the Bayesian inference results. In this way, the error-explicit Bayesian method provides meaningful insights towards diagnosing the primary sources of prediction uncertainty and suggestions for model improvement.

The remainder of this paper is organized as follows. In section 2, we outline the error-explicit Bayesian method, the variance decomposition technique and the machine learning algorithms used to construct the surrogate models. The error-explicit method is tested on a real-world case study in section 3. For comparison, we also perform a classical Bayesian calibration that does not account for model structural error. Section 4 compares and discusses the parameter estimation and prediction results obtained using the two methods. Variance decomposition results are also analyzed in this section. Finally, we draw conclusions and provide recommendations in section 5.

2. Methods

2.1. The Error-Explicit Bayesian Method: Calibration

In this study, we extend the error-explicit Bayesian method presented in *Xu and Valocchi* [2015a] to calibrate a real-world, complicated groundwater model. The Bayesian method integrates the data-driven error modeling technique [*Xu et al.*, 2014] into the Bayesian calibration framework [*Kennedy and O'Hagan*, 2001] and introduces an additive error model:

$$y = M(\boldsymbol{\theta}) + b(\mathbf{x}, \boldsymbol{\phi}) + \epsilon, \tag{1}$$

where y is the system output at a given time and location that can be measured, M denotes the groundwater model with parameters $\boldsymbol{\theta}$, and ϵ is random measurement error. When denoting multiple outputs at varying time and locations, the vector form $\mathbf{y}, \mathbf{M}, \mathbf{b}$ can be used. The model structural error term $b(\mathbf{x}, \boldsymbol{\phi})$ is represented as a function of its own inputs $\mathbf{x}$ and hyperparameters $\boldsymbol{\phi}$. Specifically, the error model input $\mathbf{x}$ may consist of the physically based model's output $M(\boldsymbol{\theta})$ 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 forecast under conditions different from the calibration period [*Xu et al.*, 2014; *Xu and Valocchi*, 2015a].

Next, we place a Gaussian process (GP) prior on the error model, that is, the prior of $\mathbf{b}$ is a multivariate Gaussian distribution $N(\boldsymbol{\mu}, \Sigma)$. The prior mean $\boldsymbol{\mu}$ is specified by a mean function, i.e., $\boldsymbol{\mu} = \mu(\mathbf{x}, \boldsymbol{\phi}) = \mathbb{E}[b(\mathbf{x})]$. The covariance matrix Σ is determined by a covariance function $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[b(\mathbf{x}) - \mu(\mathbf{x}))(b(\mathbf{x}') - \mu(\mathbf{x}')]$; the ijth entry is $\Sigma_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. Other nonparametric Bayesian kernel regression methods [*Liang et al.*, 2007; *Pillai et al.*, 2007; *Smola and Schölkopf*, 2003] can also be used depending on specific applications.

In this study, the prior mean is set to constant zero based on the prior belief that the model does not have bias. The covariance matrix Σ is calculated as the summation of a nugget term σ_ϵ^2 and an isotropic squared exponential covariance function [*Rasmussen and Williams*, 2006]:

$$k(\mathbf{x}_i,\mathbf{x}_j)=\sigma^2\ \exp\left[-\frac{1}{\lambda^2}(\mathbf{x}_i-\mathbf{x}_j)^T(\mathbf{x}_i-\mathbf{x}_j)\right]+\sigma_\epsilon^2\mathbb{1}(i=j), \tag{2}$$

where $\mathbb{1}(i=j)$ is an indicator function that equals one if $i=j$ and zero otherwise. The squared exponential covariance function reflects the prior belief that the model structural error is smooth [*Rasmussen and Williams*, 2006]. The first term on the right-hand side decreases slowly as $|\mathbf{x}_i-\mathbf{x}_j|$ increases when $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to one another. In equation (2), λ, σ^2, and σ_ϵ^2 are three hyperparameters: σ^2 controls the marginal variance of $b(\mathbf{x})$, λ is the characteristic length scale hyperparameter, and the nugget term σ_ϵ^2 describes uncorrelated errors, such as measurement error. In the geostatistics literature, σ^2 and λ are often called *sill* and *range*, respectively.

Next, let $\mathbf{y}=\{y_1,\ldots,y_n\}$ denote a set of n observation data and $\mathbf{M}$ the vector of predicted model outputs of interest with parameters $\boldsymbol{\theta}$. Assume that the measurement errors are i.i.d. normal, i.e., $\epsilon\sim N(0,\sigma_\epsilon^2)$. Depending on specific applications, σ_ϵ can either be fixed, or inferred in the calibration process. Here without the loss of generality, the model parameters $\boldsymbol{\theta}$ and Gaussian process error model hyperparameters $\boldsymbol{\phi}=\{\lambda,\sigma^2,\sigma_\epsilon^2\}$ will be inferred jointly to allow for a complete assessment of uncertainty from parameter, model structure, and measurement. We first specify the prior distribution $p(\boldsymbol{\theta},\boldsymbol{\phi})=p(\boldsymbol{\theta})p(\boldsymbol{\phi})$. Here we assume that the prior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are independent [*Brynjarsdóttir and O'Hagan*, 2014]. The error-explicit Bayesian method can be easily adapted for correlated cases. According to Bayes' theorem,

$$p(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{y})\propto p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\phi})p(\boldsymbol{\theta})p(\boldsymbol{\phi}). \tag{3}$$

The likelihood $p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\phi})$ is given by [*Rasmussen and Williams*, 2006; *Xu and Valocchi*, 2015a]

$$\log p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\phi})=-\frac{1}{2}(\mathbf{y}-\mathbf{M}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y}-\mathbf{M}-\boldsymbol{\mu})-\frac{1}{2}\log|\Sigma|-\frac{n}{2}\log 2\pi, \tag{4}$$

where $\boldsymbol{\mu}=\mu(\mathbf{x},\boldsymbol{\phi})$ is the prior mean of the error model. The posterior $p(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{y})$ usually does not have a closed form. In this study, we use DREAM-ZS (DiffeRential Evolution Adaptive Metropolis algorithm), a Markov chain Monte Carlo (MCMC) sampler [*Vrugt et al*., 2009; *Laloy and Vrugt*, 2012; *Vrugt*, 2016] to estimate the distribution.

2.2. The Error-Explicit Bayesian Method: Predictive Uncertainty

Once sufficient samples $\{\boldsymbol{\theta}_i,\boldsymbol{\phi}_i\},i=1,\ldots,N$ are generated using MCMC, Bayesian inference of prediction uncertainty can be carried out. We first run the model in prediction mode with $\boldsymbol{\theta}_i$ to obtain $\mathbf{M}_i^*$; the asterisk denotes prediction. The groundwater model output will be used as one of the inputs of the GP error model. Next we evaluate model structural error using the following equations [*Rasmussen and Williams*, 2006]:

$$\mathbf{b}^*|\mathbf{y},\boldsymbol{\phi}\sim N\left(\overline{\mathbf{b}^*},C_{bb}(\mathbf{b}^*)\right), \tag{5}$$

where

$$\overline{\mathbf{b}^*}=E[\mathbf{b}^*|\mathbf{y}-\mathbf{M},\boldsymbol{\phi}]=\boldsymbol{\mu}^*+\Sigma^{*T}\Sigma^{-1}(\mathbf{y}-\mathbf{M}-\boldsymbol{\mu}), \tag{6}$$

$$C_{bb}(\mathbf{b}^*)=\Sigma^{**}-\Sigma^{*T}\Sigma^{-1}\Sigma^*. \tag{7}$$

In the above equations, $\boldsymbol{\mu}^*$ denotes the prior mean $\mu(\mathbf{x}^*,\boldsymbol{\phi})$, and Σ^{**} denotes the prior covariance matrix. Σ^* and Σ^{**} are calculated as $\Sigma^*_{i,j}=k(\mathbf{x}_i,\mathbf{x}_j^*)$ and $\Sigma^{**}_{i,j}=k(\mathbf{x}_i^*,\mathbf{x}_j^*)$, where $k(\cdot,\cdot)$ is the covariance function (equation (2)), and $\mathbf{x}$ and $\mathbf{x}^*$ denote the GP error model input during the calibration period and in the prediction period, respectively. The model output $\mathbf{M}_i$ and $\mathbf{M}_i^*$ will be used in this step if they are included in the inputs $\mathbf{x}$ and $\mathbf{x}^*$. More details on the derivation of equations (5)–(7) can be found in *Rasmussen and Williams* [2006] and *Xu and Valocchi* [2015a]. According to equation (7), the posterior uncertainty of predictions reflected by $C_{bb}(\mathbf{b}^*)$ will generally be smaller than the prior uncertainty reflected by Σ^{**}. It can be seen that we are essentially allowing observation data to "sculpt" the relatively vague prior we placed over a family of possible functions into a posterior. Because of this data-driven feature,

GP is more flexible 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 *Rasmussen and Williams* [2006].

With $\{\boldsymbol{\theta}_i, \boldsymbol{\phi}_i\}$ and $\mathbf{M}_i^*$, we draw one realization $\mathbf{b}_i^*$ from the posterior of the error model according to equation (5). Note that $\mathbf{b}_i^*$ is conditioned on calibration residuals, which have been calculated during the calibration phase. For measurement error ϵ_i, we draw from $N(0, \sigma_{\epsilon,i}^2)$. Finally, the posterior mean of predictions is given by $\overline{\mathbf{y}^*} = \sum_{i=1}^{N} \mathbf{y}_i^*/N$, where $\mathbf{y}_i^* = \mathbf{M}_i^* + \mathbf{b}_i^* + \epsilon_i$ and N is the number of posterior samples. Here $\mathbf{y}_i^*$ is a vector of predictions at various locations and time. Predictive quantiles $\mathbf{y}_{\alpha/2}^*, \mathbf{y}_{1-\alpha/2}^*$ corresponding to a specified confidence level α can be derived by sorting $\mathbf{y}_i^*, i = 1, \ldots, N$.

2.3. Variance Decomposition

The Bayesian method described in sections 2.1 and 2.2 is capable of providing an assessment of prediction uncertainty contributed by various sources. We perform variance decomposition to investigate how much of the total prediction variance can be explained by model structural error in addition to parametric uncertainty and measurement error.

Consider a prediction $y^* = M^* + b^* + \epsilon$, e.g., groundwater head at a certain location and time. The variance of y^* can be written as

$$\mathbb{V}[y^*] = \mathbb{E}_{\boldsymbol{\theta},b^*,\sigma_\epsilon}\left[\mathbb{V}_{\epsilon|\boldsymbol{\theta},b^*,\sigma_\epsilon}[y^*|\boldsymbol{\theta}, b^*, \sigma_\epsilon]\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{b^*|\boldsymbol{\theta}}\left[\mathbb{E}_{\sigma_\epsilon,\epsilon|\boldsymbol{\theta},b^*}[y^*|\boldsymbol{\theta}, b^*]\right]\right] + \mathbb{V}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*,\sigma_\epsilon,\epsilon|\boldsymbol{\theta}}[y^*|\boldsymbol{\theta}]\right]. \quad (8)$$

The derivation can be found in Appendix A. In the above equation, the operator $\mathbb{E}$ denotes expectation, the operator $\mathbb{V}$ denotes variance, the subscripts denote the random variables over which the expectations and variance are evaluated.

The prediction procedures described in section 2.2 give posterior samples $\boldsymbol{\theta}_i, b_i^*, \sigma_{\epsilon,i}^2, i = 1, \ldots, N$, and corresponding predictions y_i^*. The left-hand side of equation (8) represents the total prediction uncertainty and can be calculated as the variance of the posterior samples $y_i^*, i = 1, \ldots N$. On the right-hand side of equation (8), the first term represents prediction variance explained by measurement error. As explained in Appendix A, this term is calculated as $\frac{1}{N}\sum_{i=1}^{N} \sigma_{\epsilon,i}^2$. Meanwhile, we have $\mathbb{E}_{\sigma_\epsilon,\epsilon|\boldsymbol{\theta},b^*}[y^*|\boldsymbol{\theta}, b^*] = M^* + b^*$ (Appendix A). Therefore, the second term on the right-hand side first calculates the conditional variance $\mathbb{V}_{b^*|\boldsymbol{\theta}}[M^* + b^*]$ due to the variability of b^* but given the value of $\boldsymbol{\theta}$, and then averages the conditional variance over possible values of $\boldsymbol{\theta}$. This term can be interpreted as the prediction variance explained by model structural error. Lastly, the third term first calculates the conditional expectation of y^* given $\boldsymbol{\theta}$, and then calculates the variance of the conditional expectation due to the variability of $\boldsymbol{\theta}$. Hence, this term is interpreted as the prediction variance contributed by groundwater model parameters $\boldsymbol{\theta}$.

In order to reduce computational cost, we use a binning method to approximate the second and third terms in equation (8). The binning method and pseudocode to implement it are given in Appendix A. Lastly, we calculate the fraction of the total prediction variance explained by uncertainties in various sources, by dividing the three terms on the right-hand side of equation (8) with $\mathbb{V}[y^*]$.

2.4. Surrogate Modeling

Markov chain Monte Carlo often requires hundreds of thousand model evaluations to generate sufficient samples from the posterior. Therefore, Bayesian calibration can be infeasible for a complicated groundwater flow model such as the one used in the case study (section 3). In order to reduce the computational cost, we construct computationally frugal surrogate models to mimic the model outputs that vary with parameter values. The surrogate models take as inputs the parameters to be inferred, and as output the original model's simulation results corresponding to calibration data, such as head and stream gain-and-loss at various times and locations. In this study, we construct surrogate models using two machine learning algorithms: random forest and support vector regression (SVR). For each output, we first use random forest to select a subset of calibration parameters to which the model outputs are the most sensitive. We then train the SVR algorithm using the selected parameters as inputs. The input selection step is important because complex real-world groundwater models often have a large number of parameters; one output can be sensitive to some parameters while being insensitive to others. The accuracy of a surrogate model for a particular output could be undermined if insensitive or redundant parameters are included in the inputs.

Random forest [*Breiman*, 2001] is an ensemble learning method that constructs multiple decision trees and outputs the mean prediction of individual trees. Each tree is fitted to a bootstrap sample of the training data set, and the out-of-bag error (i.e., prediction error at data points not selected by bootstrapping) provides an estimate of generalization error. Since its introduction, random forest has gained popularity in various fields such as meteorology [*Cloke and Pappenberger*, 2008] and soil science [*Ließ et al*., 2012]. In the area of hydrology, random forest regression was employed in a complementary data-driven modeling and uncertainty quantification (DDM-UQ) framework to improve predictive accuracy of a regional groundwater flow model and to provide robust prediction intervals [*Xu and Valocchi*, 2015b]. In this study, we use the MATLAB TreeBagger library to perform random forest regression analysis and calculate the input variable importance measure. The importance measure of an input variable is computed by averaging the increase of the out-of-bag error after permuting this variable over all trees (see Appendix B for more details). Based on this measure, those important variables (i.e., groundwater model parameters) are then selected and used as inputs for the surrogate models (SVR, in this application).

Support vector regression (SVR) [*Vapnik*, 1995] is a powerful machine learning algorithm which can be used to construct the surrogate models. Because of its good generalization performance, SVR has been applied to many fields including rainfall-runoff modeling, soil contamination and groundwater hydrology [*Asefa et al*., 2005; *Kanevski et al*., 2004; *Rasouli et al*., 2011; *Xu et al*., 2014]. Appendix C provides an overview of SVR; more details can be found in *Vapnik* [1995]. In this study, the LIBSVM toolbox [*Chang and Lin*, 2011] is used to implement ε-SVR.

When evaluating the likelihood during MCMC sampling, we will use SVR as surrogates of the original groundwater model. This introduces additional error to the inference process. Letting f_i denote the SVR prediction for the ith output M_i, we have $M_i = f_i + e_i$, where e_i is the surrogate error. We assume that $e_i, i = 1, 2, \ldots n$ are independent and Gaussian with mean zero and a constant variance $\sigma^2_{SVR,i}$ (similar to measurement error). This assumption will be further evaluated and discussed in the case study (section 3.3). In order to incorporate the surrogate error, we add Σ_{SVR}, a diagonal matrix with elements $\sigma^2_{SVR,i}, i = 1, 2, \ldots, n$ to the error covariance matrix Σ in equation (4).

3. A Real-World Case Study

In this section, we describe a real-world regional-scale groundwater flow model used as a case study to test the capability of the error-explicit Bayesian calibration and uncertainty quantification method. The regional model was developed by a multiinstitution team and parameters were calibrated by conventional least squares regression [*Hsieh et al*., 2007]. This case study is motivated by the findings in our previous work [*Xu et al*., 2014] that systematic bias exists in groundwater head simulated by the least squares calibrated model, indicating presence of model structural error.

3.1. The Spokane Valley-Rathdrum Prairie Model

This case study is based on a regional-scale groundwater flow model, namely the Spokane Valley-Rathdrum Prairie (SVRP) model. The SVRP aquifer covers approximately 326 mi^2 (844.3 km^2)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 [*Hsieh et al*., 2007]. We have used the SVRP model as a case study in our previous paper [*Xu et al*., 2014] in which data-driven error models based on machine learning techniques were used as postprocessors to improve the model's head prediction accuracy.

Figure 1 shows the SVRP model domain. The model has a uniform cell size of 1320 by 1320 ft^2 (402.3 by 402.3 m^2), and stress period of 1 month from September 1990 through September 2005. The SVRP aquifer is conceptualized as an unconfined layer (layer 1) except in Hillyard Trough 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 1. Inflow to the aquifer from surrounding tributary basins and lakes are simulated as a prescribed flow boundary. Groundwater lateral outflow at the west edge of layer 3 is simulated as a head dependent boundary. The model bottom and boundary grids that do not contain the inflow and outflow points are specified as no-flow boundaries.

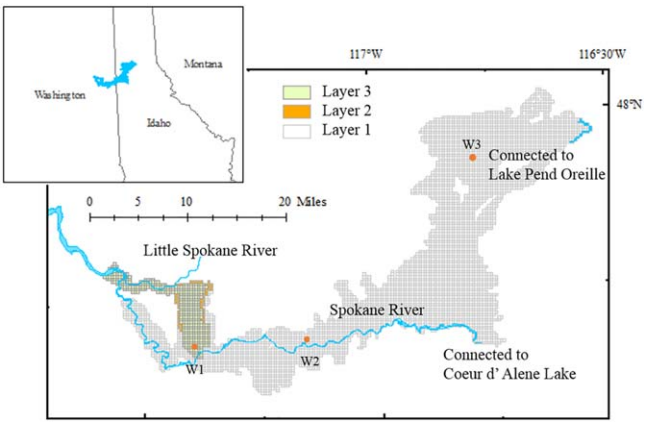

Figure 1. The Spokane Valley-Rathdrum Prairie aquifer on the border of Washington and Idaho. The three layers are shown in different colors. The grids represent the spatial discretization of the MODFLOW model. The Spokane River and the Little Spokane River are shown in blue. Also shown are the model grids simulated using RIV package to represent the hydraulic connection with Lake Pend Oreille and Coeur d'Alene Lake. The locations of representative monitoring wells as discussed in section 4.4 are marked. Adapted from *Xu et al*. [2014].

The parameterization of the model is summarized in Table 1 and explained in the following paragraphs. Additional details are given in *Hsieh et al*. [2007]. 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 is negligible and set to zero.

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 (Figure 1) are simulated using the 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 the streamflow-routing package (SFR) [*Prudic et al*., 2004]. The Spokane River within the model domain is divided into 11 sections (the stream sections are shown in *Hsieh et al*. [2007, Figure 35]), and one streambed conductance is assigned to each section (denoted by KVSR-1 through KVSR-11).

The model was calibrated using PEST [*Doherty et al*., 2010] by the model developers [*Hsieh et al*., 2007]. 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 not physically reasonable. Therefore, these three parameters were not adjusted in the calibration process but fixed. In total, there are 38 calibrated parameters [*Hsieh et al*., 2007] as listed in Table 1.

The PEST calibration data were composed of more than 1500 head measurements and 313 measurements of streamflow gain-and-loss along segments of the Spokane River and Little Spokane River from October 1995 to September 2005. The 5 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 [*Hsieh et al*., 2007] 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 is mostly within the error bounds of the measured quantities, mainly because of the relatively large measurement error of streamflow. However, residual analysis revealed that there is some bias 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 previous work [*Xu et al*., 2014] found that the calibration error is correlated temporally and spatially, indicating presence of model structural error.

3.2. Calibration and Evaluation Data

Postaudit of the SVRP model is difficult if not impossible 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

Table 1. Model Calibration Parameters and Lower and Upper Bounds[a]

		Narrow LHS Bounds		Calibration Bounds Wide LHS Bounds	
Parameter	Units	Lower	Upper	Lower	Upper
HK1–1	ft/d	3,000	25,000	100	50,000
HK1–2	ft/d	3,000	10,000	100	50,000
HK1–3	ft/d	3,000	25,000	100	50,000
HK1–4	ft/d	3,000	25,000	100	50,000
HK1–5	ft/d	3,000	25,000	100	50,000
HK1–6	ft/d	3,000	25,000	100	50,000
HK1–7	ft/d	3,000	10,000	100	50,000
HK1–8	ft/d	3,000	10,000	100	50,000
HK1–9	ft/d	1,000	4,000	1	5,000
HK1–10	ft/d	1,000	4,000	1	5,000
HK1–11	ft/d	1,000	4,000	1	5,000
HK1–12	ft/d	1	1,000	1	5,000
HK1–13	ft/d	1,000	4,000	1	5,000
HK1–14	ft/d	1	1,000	1	5,000
HK1–15	ft/d	1,000	4,000	1	5,000
HK1–16	ft/d	1	1,000	1	5,000
HK1–17	ft/d	1	1,000	1	5,000
HK1–18	ft/d	1	1,000	1	5,000
HK1–19	ft/d	1	1,000	1	5,000
HK1–20	ft/d	1	1,000	1	5,000
HK1–22	ft/d	1	1,000	1	5,000
HK3–1	ft/d	1	1,000	1	5,000
C-PO	ft^2/d	10,000	1,000,000	10^{-10}	10^{10}
C-LSR	ft^2/d	10,000	100,000	10^{-10}	10^{10}
C-CDA	ft^2/d	10,000	200,000	10^{-10}	10^{10}
KVSR-1	ft/d	0.01	0.5	0.01	10
KVSR-2	ft/d	0.01	0.5	0.01	10
KVSR-3	ft/d	0.01	0.5	0.01	10
KVSR-4	ft/d	0.01	0.5	0.01	10
KVSR-5	ft/d	5	10	0.01	10
KVSR-6	ft/d	0.01	5	0.01	10
KVSR-7	ft/d	5	10	0.01	10
KVSR-8	ft/d	0.01	1	0.01	10
KVSR-9	ft/d	1	10	0.01	10
KVSR-10	ft/d	1	10	0.01	10
SY-1		0.1	0.2	0.1	0.3
SY-2		0.15	0.3	0.1	0.3
SY-3		0.15	0.3	0.1	0.3

[a]Adapted from *Hsieh et al.* [2007, Table 8].

rates) requires a variety of information, such as land use maps, irrigation amount for both agricultural and recreational lands, and domestic and public supply pumping records. Not all of the required information is readily available for this study. Therefore, it is not feasible 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 5 years as warm-up period. Groundwater piezometric head and stream gain-and-loss measurements from October 1995 to September 2004 are then used as calibration data, while measurements from October 2004 to September 2005 are reserved for evaluating (testing) the calibrated model. The period from October 2004 to September 2005 corresponds to a dry period since the precipitation recharge is lower than in preceding years [*Hsieh et al.*, 2007, Figure 9]. In other words, the evaluation period exhibits somewhat different hydrogeologic conditions from the conditions reflected by the calibration data set.

The whole data set of model calibration and evaluation is composed 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 3.1), as well as additional head observations that became available via the USGS Water Data for the Nation (NWIS) online database (http://waterdata.usgs.gov/nwis/gw) after model construction and PEST calibration in 2006. In total, calibration data set includes 1552 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 evaluation data set is composed 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. For head measurements, the standard deviation of measurement error is calculated using the information provided in NWIS including the accuracy of site land surface altitude, accuracy of depth to groundwater measurement and site status (e.g., if pumped recently) [*Hsieh et al*., 2007]. The resulting head measurement standard deviation can vary in space and time and ranges from 0.01 to 25.5 ft (0.003 to 7.77 m). A high measurement error standard deviation can be the result of low accuracy of land surface elevation and/or if the well has not returned to equilibrium after being recently pumped. Meanwhile, 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 [*Hsieh et al*., 2007]. The stream gain-and-loss measurement error standard deviation ranges from 9 to 1.2×10^3 cfs.

3.3. Building Surrogate Models

One forward run for the simulation period (from 1990 to 2005) of the SVRP MODFLOW model takes approximately 2 min and 20 s (depending on parameter values) on a single core of a multicore 2.00GHz processor. In order to reduce the computational cost of Bayesian calibration, we construct computationally frugal surrogate models to mimic the model outputs that vary with parameter values (section 2.4).

In order to generate the training data of the surrogate models, the SVRP model was run repeatedly using 6400 sets of the 38 calibration parameters drawn using Latin Hypercube sampling. Among the 6400 parameter sets, 3200 sets are drawn uniformly according to the lower and upper bounds enforced during calibration in Table 1. The remaining 3200 sets are drawn uniformly from a narrower interval (Table 1) that are believed a priori to be more reasonable based on the hydrogeologic conditions of the region [*Xu*, 2016]. In this way, the LHS samples span across a wide range of parameter space, while being denser in a more probable region. As the 6400 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. For some parameter sets, the SVRP model fails to converge. The remaining LHS samples, about 3550 parameter sets (subject to the randomness of LHS sampling) are retained to train the surrogate models.

Sensitivity analysis reveals that for many observations, the change in the values of some parameters does not significantly alter the model outputs. The most influential parameters for one output may have little impact on another output. Therefore, we performed feature selection for every model output based on the variable importance measure provided by the random forest algorithm as described in section 2.4. We divide the importance measures for 38 parameters by the sum of measures. Parameters that correspond to a normalized measure greater than 0.01 are selected as inputs of the surrogate model. For different observed outputs, the number of selected inputs varies from 3 to 38 with an average of 17.6.

Next, support vector regression (SVR, section 2.4) is used to construct surrogate models for each head measurement and each stream gain-and-loss observation, resulting in a total of 1552 surrogate models for head and 264 for stream gain-and-loss. Compared to the SVRP model, the SVR surrogates are much faster to run and take approximately 4.6 s on a single core of a multicore 2.00 GHz processor. In addition, the surrogates for different output can be run in parallel. With eight parallel threads, using surrogate model achieves an 187-fold speedup.

Split-sample validation is carried out to examine the emulation accuracy of SVR surrogates. Model simulation results with LHS samples are randomly divided into a training data set (80% of the samples) and a testing data set (20% of the samples). Using the training data set, we tune the SVR hyperparameters via fivefold cross validation (section 2.4 and Appendix C) to optimize SVR performance. We then retrain the SVR surrogates using the whole training data set with the selected hyperparameters; the trained SVR surrogates are then tested on the testing data set which consists of about 710 data points for every drawdown and stream gain-and-loss output.

Figure 2 compares the surrogate model prediction results with the MODFLOW model outputs for the testing data set. Overall, the head emulation coefficient of determination R^2 is 0.999, and the RMSE is 1.3 ft. The LHS parameter sets cover a wide range, and the corresponding SVRP model simulated head can vary significantly (by more than 55 ft for half of calibration targets). For streamflow gain-and-loss on the Spokane River

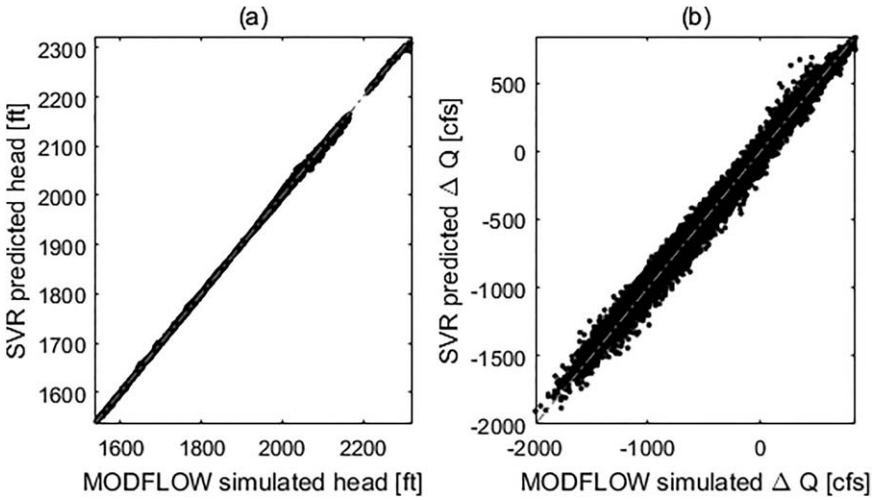

Figure 2. (a) Groundwater head simulated by the SVR surrogate models plotted versus the head simulated by the MODFLOW model. (b) Streamflow gain-and-loss simulated by the SVR surrogate models plotted versus the gain-and-loss simulated by the MODFLOW model.

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, we conclude that the SVR emulation accuracy can be considered acceptable.

As discussed in section 2.4, using SVR as surrogates of the MODFLOW model may introduce additional error to the inference process. As described earlier, we test the trained SVR models on a testing data set (20% of LHS samples) that is separate from the training data set. Let $f_{i,j}$ denote the SVR prediction with the jth draw for the ith MODFLOW output M_i; therefore we write $M_i=f_{i,j}+e_{i,j}$, where $e_{i,j}$ is the surrogate error. Comparing the SVR predicted values with the MODFLOW model 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 size of the testing data set. The MSE s_i^2 is a good estimate of the surrogate error variance $\sigma_{SVR,i}^2$. When evaluating the likelihood $p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\phi},\sigma_\epsilon)$ we include Σ_{SVR}, a diagonal matrix with elements $\sigma_{SVR,i}^2=s_i^2, i=1,2,\ldots,n$. Here we assume that the surrogate errors $e_{i,j}, j=1,\ldots,n$ are independent because residual analysis did not show any significant correlation. In addition, the surrogate errors approximately follow a Gaussian distribution with zero mean (not shown), suggesting that the assumption described in section 2.4 holds for this case study.

4. Results and Discussions

In this section, we perform calibration and prediction using both classical Bayesian which assumes no model structural error and our fully Bayesian approach with an error model. The parameter estimation and prediction results obtained using the two methods are assessed and compared. For the error-explicit Bayesian approach, we also discuss the variance decomposition analysis results.

4.1. Implementation of Classical Bayesian Calibration

As a 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 wide range that is considered as physically reasonable given the hydrogeologic conditions, as listed in Table 1.

As described in section 3.3, we add a surrogate error covariance matrix Σ_{SVR} to account for the error resulting from using the SVR surrogates. Meanwhile, variogram analysis on head residuals of the SVRP model calibrated using least squares regression reveals a nugget effect that is larger than the magnitude of head measurement error [*Xu et al*., 2014]. Therefore, we add a nugget term σ_0 to the standard deviation of head measurement errors when evaluating the likelihood during MCMC sampling. More specifically, σ_ϵ^2 in equation (2) is the sum of σ_0^2 and measurement error variance calculated using the method described in section 3.2. There is no nugget effect found for stream gain-and-loss residuals, so σ_0 is set to zero for ΔQ

observations. The nugget term is interpreted as random error that cannot be attributed to measurement error. Hereafter we will refer to this component of error as unresolvable. The head unresolvable error standard deviation σ_0 will be inferred along with the groundwater model parameters during calibration. We specify a vague prior as an exponential distribution with a mean of 10 ft, which slightly favors smaller σ_0. However, if the model output cannot match well with observations regardless of what values the parameters take, the posterior values of σ_0 will be high. When deriving the prediction intervals, we will add noise generated from the measurement and unresolvable errors. We do not add surrogate error because we will run the full SVRP model during the prediction phase.

The DREAM-ZS runtime setting is configured following the recommendations in *Vrugt et al*. [2008]. Ten Markov chains are used to generate 1600 samples from the joint posterior distribution of model parameters after convergence is determined based on the $\hat{R}$ statistic, visual inspection of trace plots, and other diagnostics [*Cowles and Carlin*, 1996]; about 300,000 model evaluations are required to converge (burn-in). The marginal posterior distributions of the 38 parameters are shown in Figure 3. The parameter estimation and prediction results will be discussed in sections 4.3 and 4.4.

4.2. Implementation of Error-Explicit Bayesian Calibration

Based on the residual analysis results reported in section 3.1, we construct an error model to describe model structural error in head. We do not construct error models for stream gain-and-loss because residual analysis did not reveal significant bias and correlation structure in streamflow residuals partially due to large streamflow measurement error. As variogram analysis reveals spatial correlation in head residual [*Xu et al*., 2014], the input of the error model should include spatial locations of head measurements, $\mathbf{u}=(u_x, u_y)$. Since in this 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 *Xu et al*. [2014] found that it was better to use the (surrogate) model simulated head (M_h) instead of time for temporal prediction. In summary, the input of the error model is $\mathbf{x}=\{\mathbf{u}, \mathbf{M}_h\}$. $\mathbf{u}$ is scaled to a range of $[0, 1]$; $\mathbf{M}_h$ changes with varying $\boldsymbol{\theta}$, and is scaled by a factor determined from the range of the outputs of LHS model runs. 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 could be explored in future studies.

The interaction between the physical model parameters and error model could potentially lead to identifiability issues [*Brynjarsdóttir and O'Hagan*, 2014; *Reichert and Schuwirth*, 2012]. 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. An isotropic squared exponential covariance function (equation (2)) is used to enforce smoothness and regularize the degree of complexity of inferred model structural error. The GP error model has three hyperparameters: characteristic length scale λ, standard deviation σ, and a nugget term σ_0. The length scale hyperparameter represents the degree of correlation in the space of GP input. We specify a uniform distribution on $(0, 1]$ for λ, which is a loose bound for the case study. The standard deviation represents the amount of model structural error that is acceptable. A larger σ allows the error model to take on a larger compensation role of the mismatch between observations and model simulation results. We specify an exponential distribution with a mean of 10 ft based on residual analysis results in *Xu et al*. [2014] for σ. In this 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. The nugget term σ_0 characterizes the unresolvable error in head measurements as explained in section 4.1. Different from the classical Bayesian calibration, here the nugget term is interpreted as the aleatoric error that the error model does not capture, including additional factors such as the interpolation error which occurs when we interpolate (both in space and time) grid-based model output to compare with observations. Such interpolation error is not included in the measurement error standard estimation (section 3.2). For the 38 MODFLOW parameters and σ_0, the same prior distributions as in section 4.1 are used.

In total, Bayesian calibration infers the joint posterior distribution of 41 parameters, using head and streamflow gain-and-loss during the calibration period (October 1995 to September 2004). Using DREAM-ZS, 10

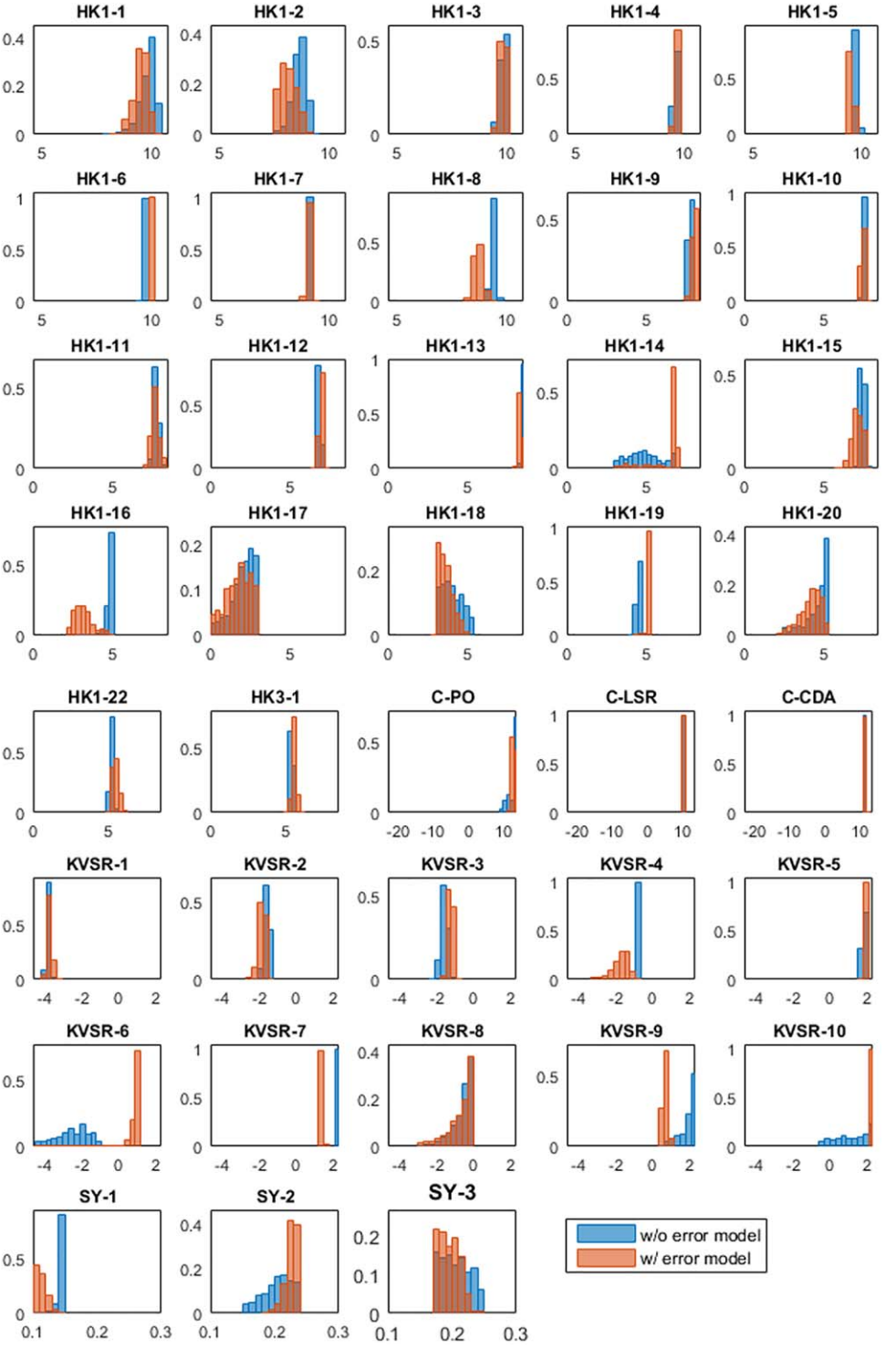

Figure 3. Normalized histogram of marginal posterior distributions given by the classical Bayesian (blue) and the error-explicit Bayesian approach (orange) of 38 SVRP model parameters. The posteriors given by the two methods have overlap, which is indicated by brown color. The ranges of *x* axes represent the lower and upper bounds enforced during calibration. The hydraulic conductivity parameters are natural logarithm transformed when generating the histograms.

Markov chains are used to generate 1600 samples from the joint posterior distribution of 41 parameters after convergence is determined based on the $\hat{R}$ statistic [*Gelman and Rubin*, 1992], visual inspection of trace plots and other diagnostics [*Cowles and Carlin*, 1996]. As burn-in, 400,000 samples are discarded.

In the prediction phase, the groundwater head forecasts $\mathbf{h}_i^*$ could be evaluated using the surrogate models if surrogate models for prediction quantities have been trained following the procedures described in section 3.3. 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 allows for verifying that the parameter values obtained from the calibration of the surrogate model are also valid for the MODFLOW SVRP model. It was found that all the model runs converge. Repeated evaluation of the MODFLOW model is feasible because the model runs using different sets of parameters can be executed in parallel. The MODFLOW model is evaluated using 1600 samples drawn with DREAM-ZS. Depending on the accuracy requirement and computational expense of a specific application, more or fewer posterior samples may be needed.

The Gaussian process error model uses as input $\mathbf{x}^* = [\mathbf{u}, \mathbf{M}_h^*]$. The Bayesian framework yields an ensemble of head predictions $\mathbf{h}_i^* = \mathbf{M}_h^*(\boldsymbol{\theta}_i) + \mathbf{b}_i^* + \epsilon_{h,i}, i = 1, \ldots, N, N = 1,600$; $\mathbf{h}_i^*$ denotes groundwater head varying in both space and time; $\mathbf{M}_h^*(\boldsymbol{\theta}_i)$ denotes head simulated by the MODFLOW model using parameters $\boldsymbol{\theta}_i$; $\mathbf{b}_i^*$ is a noise-free vector drawn from the GP error model posterior; $\epsilon_{h,i}$ is randomly drawn from a normal distribution $N(0, \sigma_{0,i}^2 I + \Sigma_\epsilon)$, where $\sigma_{0,i}$ is the ith posterior sample of the unresolvable error standard deviation (i.e., nugget), I is an identity matrix, and Σ_ϵ is a diagonal matrix with head measurement error variance as diagonal entries. As mentioned in section 3.2, the head measurement error variance could vary in space and time. The Bayesian posterior of prediction can then be estimated by collecting the realizations in the ensemble, and the posterior mean is given by $\overline{\mathbf{h}}^*$.

Lastly, we perform variance decomposition to examine the proportion of prediction uncertainty contributed from different variance sources. Following section 2.3, for a prediction h^* the variance decomposition is written as

$$\mathbb{V}[h^*] = \sigma_\epsilon^2 + \mathbb{E}[\sigma_0^2] + \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{b^*|\boldsymbol{\theta}}\left[\mathbb{E}_{\sigma_\epsilon,\epsilon}[h^*|\boldsymbol{\theta}, b^*]\right]\right] + \mathbb{V}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*,\sigma_\epsilon,\epsilon|\boldsymbol{\theta}}[h^*|\boldsymbol{\theta}]\right], \tag{9}$$

Note that the first term on the right-hand side of equation (8) is expanded to include the head measurement error variance, which is fixed, as well as the unresolvable error variance given by the nugget term averaged over all posterior samples.

4.3. Results: Parameter Estimates

Figure 3 shows the marginal posterior distributions estimated by the classical Bayesian method without error model and the error-explicit Bayesian method. The names of the 38 SVR parameters were defined in section 3.1. All parameters except SY-1, SY-2, and SY-3 are natural logarithm transformed when generating the histogram plots. The priors of the parameters were specified as uniform distributions over a wide range (sections 4.1 and 4.2). The priors are not displayed in Figure 3 because they would be a horizontal line close to the horizontal axes.

In this real-world study, the "true" values of model parameters are unknown. Therefore, it is not possible to validate the correctness of parameter posteriors. For some parameters (e.g., KVSR-7, KVSR-9, and KVSR-10), the posteriors are on the upper bound. The same phenomena occurred in PEST calibration, and the resulting parameter estimates are considered reasonable [*Hsieh et al*., 2007]. Given the wide priors, the parameter posteriors determined by both the classical Bayesian and the error-explicit method can be considered as fairly constrained. While the posteriors given by the two methods have some overlap, it can be seen that the Bayesian with GP error model approach yields posteriors that are substantially different from the classical Bayesian method for many parameters, such as HK1–16, KVSR-6, SY-1, and SY-2. This suggests that parameter compensation is likely to occur when model structural error is present but not handled explicitly [*Doherty and Christensen*, 2011; *Xu and Valocchi*, 2015a].

4.4. 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 4 and Table 2 assess how the simulated head compares with observation data. Figure 4 plots the difference between observations and posterior mean given by the classical and the error-explicit Bayesian methods. For both the calibration and the evaluation periods, the proposed Bayesian method simulation error is smaller compared to classical Bayesian results. Table 2 summarizes the mean error, mean absolute percentage error, and root-mean-square error (RMSE) statistics. The

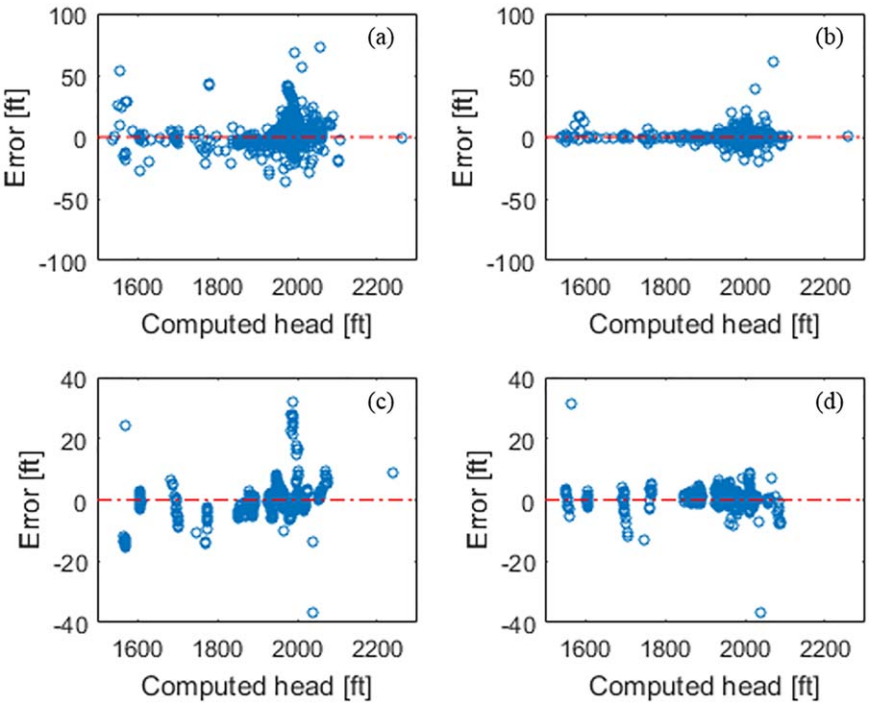

Figure 4. Head simulation error plotted versus posterior mean given by (a, c) the classical Bayesian and (b, d) the error-explicit Bayesian methods, calculated during the calibration (Figures 4a and 4b) and evaluation (Figures 4c and 4d) periods, respectively. The calibration period is from October 1995 to September 2004, and the evaluation period spans October 2004 to September 2005.

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 SVRP model, reducing the RMSE by more than 50% for the evaluation period. The error model also removed most of the global bias, reducing the mean error from −2.08 to 0.483 ft, and the mean absolute percentage error from −0.11% to 0.026%.

Figure 5 shows head simulation results at three representative wells; the locations of wells are plotted in Figure 1. Figures 5a–5d plot the head prediction at two wells located in the Spokane Valley (W1 and W2 in Figure 1). For both wells, the classical Bayesian approach simulation results are biased in that the posterior mean deviates from observation data. The error-explicit Bayesian approach magnified the seasonal fluctuation of head at well W1, yielding head prediction that better matches the evaluation data. Well W2 is close to the Spokane River. The GP error model is not able to fully recover the observed head rise starting from September, which is caused by the rise in Spokane River stage as the Post Falls Dam opens its gates [*Hsieh et al.*, 2007]. The performance could potentially be improved by incorporating relevant information, such as river stage, into the GP error model inputs.

Figure 5e shows that for well W3 in the northern Rathdrum Prairie, the classical Bayesian method results in good fit to calibration data. However, in the prediction period the simulated groundwater head is systematically lower than observations; the fluctuation character does not match measurements. A possible reason is that the temporal distribution of recharge for this region used in the model during 2004–2005 is not

Table 2. Head Simulation Error of the Classical Bayesian and Error-Explicit Bayesian Methods[a]

	Calibration		Evaluation	
	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
MAPE (%)	0.0595	−0.0018	−0.11	0.0258
RMSE (ft)	11.4	4.48	7.84	3.55

[a]Performance measures are calculated for the calibration period (October 1995 to September 2004) and the evaluation period (October 2004 to September 2005), respectively. MAPE is the mean absolute percentage error, and RMSE is the root-mean-square error.

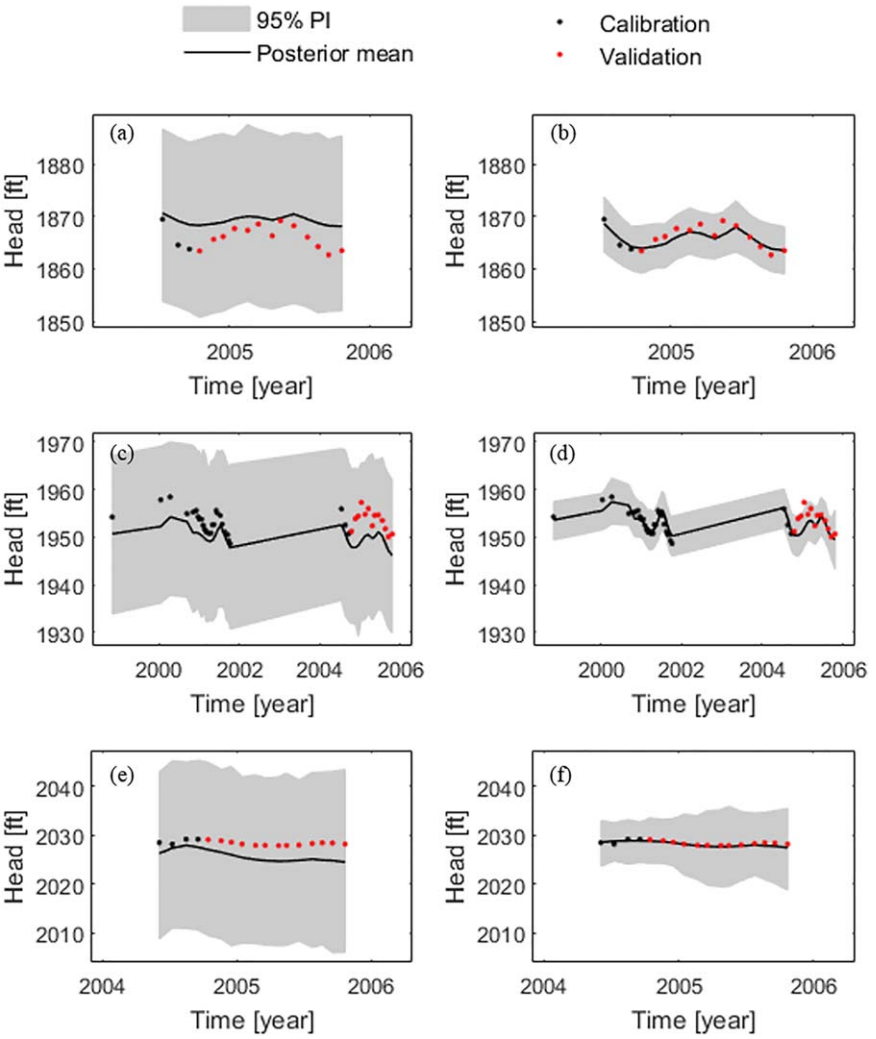

Figure 5. Head prediction at well (a, b) W1, (c, d) W2, and (e, f) W3 given by (left) the classical Bayesian and (right) the proposed error-explicit Bayesian approach. Well locations are shown in Figure 1. Grey shades show 95% prediction interval, black dots are calibration data, and red dots are verification data. All available observation data at the three wells from 1995 to 2005 are shown.

accurate [*Hsieh et al.*, 2007]. Using the GP error model (Figure 5f), we were able to correct this issue and improve the prediction accuracy.

It can be observed from Figure 5 that the classical Bayesian method yields very wide 95% prediction intervals. The main reason is that the classical Bayesian method gives large unresolvable error estimates, even though the prior favors smaller σ_0, which occurs because the model output cannot fit calibration data well due to model structural error. On the other hand, the error-explicit Bayesian method provides more reasonable prediction intervals. The GP error model resolves a major part of model structural error, resulting in lower posterior of the unresolvable error. The comparison observed in this study poses an interesting scenario different from our preceding work [*Xu and Valocchi*, 2015a] in which small calibration error was achieved due to overfitting despite model structural error. In that case, the classical Bayesian method tends to underestimate prediction uncertainty and provides prediction intervals that are too narrow to encompass evaluation data. In real-world applications, either scenario is possible, and parameter compensation occurs in both cases.

Lastly, the error-explicit Bayesian method is capable of providing a complete assessment of prediction uncertainty contributed from various sources. Through variance decomposition (section 2.3 and equation (9)), we calculate the fractions of the total head prediction variance explained by parametric uncertainty, model structural, unresolvable and measurement errors. Theoretically, the fractions should sum to one for

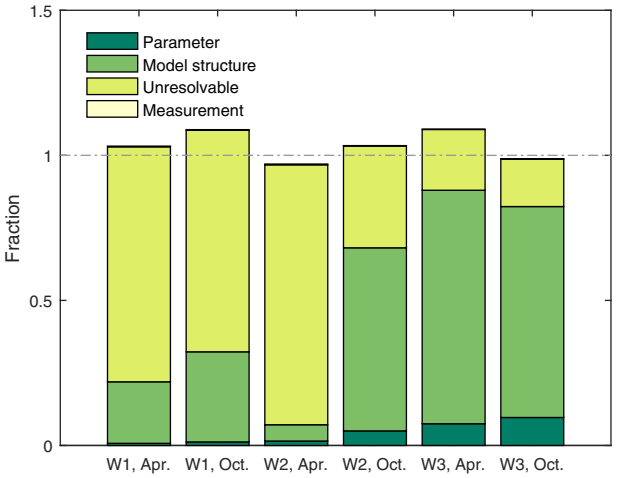

Figure 6. Bar plot showing the decomposition of prediction variance into variance contributed by parameter, model structure, unresolvable, and measurement uncertainty for groundwater head predictions at wells W1, W2, and W3 in April and October 2005.

each prediction; as can be seen from Figure 6, the sum is close to one, suggesting that error resulting from the binning approximation is acceptable.

It is found that the prediction variance decomposition among the four sources is different among wells and observation times. Figure 6 shows the variance decomposition calculated for head predictions at three wells W1, W2, and W3 in April and October 2005. For the three wells, measurement error is negligible, and parametric uncertainty contribution to total prediction variance is low. For the head predictions at W1 and at W2 in April 2005, the unresolvable error is the primary source of prediction variance, while prediction variance due to model structural error is relatively small. Meanwhile, Figure 5 shows that the classical Bayesian calibrated model has nonnegligible calibration error at W1 and W2, suggesting relatively large model structural error. The two observations are not inconsistent. Rather, the variance decomposition result indicates that for these predictions the model structural error is well constrained by the calibration data. In other words, we are relatively certain about the amount of model structural error associated with these predictions. However, uncertainty from model structural error is significantly higher for head prediction at W2 in October than in April. At this well, the posterior uncertainty of model structural error increases as the error model input data during the prediction period move away from input data during the calibration period. In addition, as mentioned earlier in this section, groundwater head rise at well W2 in autumn is not well reproduced by the model, thus leading to higher model structural uncertainty in October.

For W3 model, structural error is the largest source of prediction uncertainty. Recall that the classical Bayesian calibration fits calibration data better at W3 than at W1 and W2 (Figure 5). Head prediction at W3 is likely subject to the interaction between model parameters and structural error. Due to the interaction, the W3 prediction variance from model structural error and parameter uncertainty is higher than the equivalents for W1 and W2. Model structural error being the primary source of prediction uncertainty for W3 suggests that the region in which W3 is located might not be well represented in the SVRP model and therefore needs further investigation.

4.5. Discussion

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. For problems with strong nonlinearity or even discontinuity, it may be challenging to achieve high surrogate accuracy [*Zeng et al.*, 2016], especially when only a relatively small amount of training data is available due to high computational cost of the model. Under these circumstances, other strategies can be employed to reduce the computational cost associated with Bayesian inference, such as by reducing the dimension of the parameter space using Karhunen-Love truncation [*Laloy et al.*, 2013] and by multiple-try sampling [*Xie et al.*, 2009].

When the surrogate accuracy is not satisfactory for Bayesian inference, a two-stage MCMC simulation scheme [*Cui et al.*, 2011; *Elsheikh et al.*, 2014; *Laloy et al.*, 2013] can be used to prevent the sampling bias arising from using surrogate models. In the two-stage scheme, a surrogate model is first evaluated for every proposal generated by MCMC. Then in the second stage, only the proposals that are initially accepted are evaluated using the original model. However, the two-stage approach often requires tens of thousands of sequential forward run of the original model, and therefore could be computationally unaffordable for time-consuming models. The choice of surrogate modeling strategy should be made based on the tradeoff between computational cost and accuracy required for a specific application.

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 are used to make forecasts beyond October 2004 at the same well locations. In groundwater modeling practice, it is often necessary to predict head at an unsampled location. 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 inferred from training wells. As discussed in our preceding work [*Xu et al*., 2014], 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 error-explicit Bayesian method requires further investigation for additional real-world case studies with denser observation networks.

The simulation results of the groundwater model are included in the error model inputs. This makes it possible to use the Gaussian process error model to account for future scenario changes and predict during the evaluation period. During the evaluation period in the case study, the aquifer receives less precipitation recharge than in the preceding years. As discussed in section 4.4, the robustness of the error model can potentially be further improved by incorporating other relevant information not directly used in the development of the physically based model.

In this study, we used one GP error model with a simple isotropic covariance function to emulate the model structural error lumped from various model deficiencies. Depending on specific applications, a modeler can use an anisotropic covariance function or a combination of several different kinds of simple covariance functions [*Rasmussen and Williams*, 2006]. Each simple covariance function handles an individual property of the model structural error. 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. Similarly, a mixture of GP models and other types of Bayesian kernel methods [*Liang et al*., 2007; *Pillai et al*., 2007; *Smola and Schölkopf*, 2003] can be used to allow for more flexibility. When this is performed, cautious specification of error model priors are needed to alleviate the identifiability issues [*Brynjarsdóttir and O'Hagan*, 2014; *Reichert and Schuwirth*, 2012; *Renard et al*., 2010], which tend to be more prominent with increasing flexibility of error model. Nevertheless, the implication of the user-specified prior information (including those described in section 4.2) for the inference process remains an open question and requires further study.

5. Conclusions

Through a real-world groundwater flow modeling study, we investigated the impacts of model structural error on calibration and prediction. We tested the error-explicit Bayesian calibration and uncertainty quantification method [*Xu and Valocchi*, 2015a] in a real-world setting. A data-driven error model is integrated into the Bayesian framework to correct for spatiotemporal model structural error of groundwater head. We constructed computationally frugal surrogate models to emulate the response of the groundwater model with respect to its parameters. With this strategy, a 187-fold speedup was obtained, and Bayesian calibration becomes feasible for the complicated SVRP model with 38 parameters to be estimated.

It was demonstrated that the error-explicit Bayesian method yielded parameter posterior pdfs that are substantially different from posteriors obtained using classical Bayesian that does not account for model structural error. As for prediction performance, not accounting for model structural error led to biased head predictions. In contrast, integrating a GP error model effectively improved the prediction accuracy and yielded prediction intervals that are consistent with evaluation data. The results suggest that the proposed approach is a robust method in real-world modeling problems.

The error-explicit Bayesian method constructs error models in an inductive, data-driven way. This gives the error model predictive capability under conditions different from the hydrogeologic and development conditions reflected by calibration data. However, it should be noted that the predictability challenge in forecasting dynamic changes [*Kumar*, 2011] 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 or calibration data set. More specifically, in this study the data-driven error model would not be capable of predicting model structural error that occurs only in the prediction period and is not manifested by the calibration data. As such, data-driven methods cannot replace

thoughtful modeling analysis and additional field observations toward improved understanding of the physical processes.

In summary, the results of the case study suggest that the error-explicit Bayesian approach could be a robust practical method in real-world modeling problems. The results also underscore the importance of proper treatment of model structural errors to ensure prediction accuracy and robust uncertainty quantification. The Bayesian method provides a complete assessment of prediction uncertainty contributed by parameter, structural error, and measurement. Insights gained from the variance decomposition analysis could inform future model refinement and data collection efforts on how to best direct resources towards reducing predictive uncertainty. We recommend the error-explicit Bayesian method for applications in which data support the presence of model structural error, yet model deficiency cannot be diagnosed. Follow-up studies will further investigate the feasibility of joint inference of input and model structural errors, particularly for real-world modeling practice.

Appendix A: Derivation and Evaluation of the Variance Decomposition in Equation (8)

This appendix derives the variance decomposition in equation (8) and gives the pseudocode of the binning method we used to approximately calculate the variance decomposition. Consider a prediction $y^*=M^*+b^*+\epsilon$, e.g., groundwater head at a certain location and time. Applying the law of total variance three times, the variance of y^* can be written as

$$\begin{aligned}\mathbb{V}[y^*] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{b^*,\sigma_\epsilon,\epsilon|\boldsymbol{\theta}}[y^*|\boldsymbol{\theta}]\right]+\mathbb{V}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*,\sigma_\epsilon,\epsilon|\boldsymbol{\theta}}[y^*|\boldsymbol{\theta}]\right]\\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[\mathbb{V}_{\sigma_\epsilon,\epsilon|\boldsymbol{\theta},b^*}[y^*|\boldsymbol{\theta},b^*]\right]\right]+\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{b^*|\boldsymbol{\theta}}\left[\mathbb{E}_{\sigma_\epsilon,\epsilon|\boldsymbol{\theta},b^*}[y^*|\boldsymbol{\theta},b^*]\right]\right]+\mathbb{V}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*,\sigma_\epsilon,\epsilon|\boldsymbol{\theta}}[y^*|\boldsymbol{\theta}]\right]\\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[\mathbb{E}_{\sigma_\epsilon|\boldsymbol{\theta},b^*}\left[\mathbb{V}_{\epsilon|\boldsymbol{\theta},b^*,\sigma_\epsilon}[y^*|\boldsymbol{\theta},b^*,\sigma_\epsilon]\right]\right]\right]+\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[\mathbb{V}_{\sigma_\epsilon|\boldsymbol{\theta},b^*}\left[\mathbb{E}_{\epsilon|\boldsymbol{\theta},b^*,\sigma_\epsilon}[y^*|\boldsymbol{\theta},b^*,\sigma_\epsilon]\right]\right]\right]\\ &\quad+\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{b^*|\boldsymbol{\theta}}\left[\mathbb{E}_{\sigma_\epsilon,\epsilon|\boldsymbol{\theta},b^*}[y^*|\boldsymbol{\theta},b^*]\right]\right]+\mathbb{V}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*,\sigma_\epsilon,\epsilon|\boldsymbol{\theta}}[y^*|\boldsymbol{\theta}]\right].\end{aligned} \tag{A1}$$

In the above equation, the expectation operator $\mathbb{E}$ and the variance operator $\mathbb{V}$ are evaluated over the random variables denoted by the subscripts (section 2.3). The second term on the right-hand side equals zero because $\mathbb{E}_{\epsilon|\boldsymbol{\theta},b^*,\sigma_\epsilon}[y^*|\boldsymbol{\theta},b^*,\sigma_\epsilon]=M^*+b^*$, and $\mathbb{V}_{\sigma_\epsilon^2|\boldsymbol{\theta},b^*}[M^*+b^*]=0$. Next, using the law of total expectation for the first term, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[\mathbb{E}_{\sigma_\epsilon|\boldsymbol{\theta},b^*}\left[\mathbb{V}_{\epsilon|\boldsymbol{\theta},b^*,\sigma_\epsilon}[y^*|\boldsymbol{\theta},b^*,\sigma_\epsilon]\right]\right]\right] &= \mathbb{E}_{\boldsymbol{\theta},b^*,\sigma_\epsilon}\left[\mathbb{V}_{\epsilon|\boldsymbol{\theta},b^*,\sigma_\epsilon}[y^*|\boldsymbol{\theta},b^*,\sigma_\epsilon]\right]\\ &= \mathbb{E}_{\boldsymbol{\theta},b^*,\sigma_\epsilon}\left[\sigma_\epsilon^2\right]=\frac{1}{N}\sum_{i=1}^{N}\sigma_\epsilon^2,\end{aligned} \tag{A2}$$

where N is the number of posterior samples. It follows that the variance decomposition is reduced to equation (8). Also note that for the second term on the right-hand side of equation (8), using the law of total expectation, $\mathbb{E}_{\sigma_\epsilon,\epsilon|\boldsymbol{\theta},b^*}[y^*|\boldsymbol{\theta},b^*]=\mathbb{E}_{\sigma_\epsilon|\boldsymbol{\theta},b^*}\left[\mathbb{E}_{\epsilon|\sigma_\epsilon,\boldsymbol{\theta},b^*}[y^*|\boldsymbol{\theta},b^*,\sigma_\epsilon]\right]=M^*(\boldsymbol{\theta})+b^*$. Similarly, for the inner expectation of the last term we have $\mathbb{E}_{b^*,\sigma_\epsilon,\epsilon|\boldsymbol{\theta}}[y^*|\boldsymbol{\theta}]=\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[\mathbb{E}_{\sigma_\epsilon,\epsilon|\boldsymbol{\theta},b^*}[y^*|\boldsymbol{\theta},b^*]\right]=\mathbb{E}_{b^*|\boldsymbol{\theta}}[M^*(\boldsymbol{\theta})+b^*]$.

Finally, let $y_0^*=M^*(\boldsymbol{\theta})+b^*$ denote the noise-free prediction. In addition to y_i^*, the prediction procedures described in section 2.1 also give noise-free predictions $y_{0,i}^*=M^*(\boldsymbol{\theta}_i)+b_i, i=1,\ldots,N$. Here we add the subscript 0 in $y_{0,i}^*$ to differentiate it from y_i^*, which contains measurement error. Equation (8) can be rewritten as

$$\mathbb{V}[y^*]=\frac{1}{N}\sum_{i=1}^{N}\sigma_\epsilon^2+\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{b^*|\boldsymbol{\theta}}\left[y_0^*|\boldsymbol{\theta},b^*\right]\right]+\mathbb{V}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[y_0^*|\boldsymbol{\theta},b^*\right]\right]. \tag{A3}$$

Direct evaluation of the variance decomposition terms can be very computationally expensive, due to the nested structure of the various uncertainty sources. We use a binning method to approximate the second and third terms on the right-hand side of equations (8) and (A3). The binning method divides (e.g., using k-means clustering) the posterior samples $\boldsymbol{\theta}_i, i=1,\ldots,N$ into K bins ($\Omega_k, k=1,\ldots,K$) such that the variability of $\boldsymbol{\theta}$ within each bin is small. Accordingly, the corresponding $\{b_i^*, y_{0,i}^*\}, i=1,\ldots N$ are also divided into bins. Next we calculate the mean and variance of $y_{0,i}^*$ within each bin, denoted by E_k and V_k, where $k=1,\ldots,K$. The model structure variance term $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{b^*|\boldsymbol{\theta}}\left[y_0^*|\boldsymbol{\theta},b^*\right]\right]$ is approximated by the mean of the bin-wise variance V_k, and the parameter uncertainty term $\mathbb{V}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[y_0^*|\boldsymbol{\theta},b^*\right]\right]$ is approximated by the variance of bin-wise

mean E_k. The pseudocode to implement this procedure is given in Algorithm A1; the notation $\leftarrow$ means assigning the value of right-hand side to the left-hand side.

Algorithm A1 Binning method to approximately calculate variance decomposition

Divide $\{\boldsymbol{\theta}_i\}, i=1,\ldots,N$ into bins $\Omega_1,\ldots,\Omega_K$

for $k=1$ **to** K **do**

$N_k \leftarrow ||\Omega_k||$ ▷ Number of samples in the kth bin

$E_k \leftarrow \frac{1}{N_k}\Sigma_{i:\boldsymbol{\theta}_i\in\Omega_k} y^*_{0,i}$ ▷ Approximates $\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[y^*_0|\boldsymbol{\theta},b^*\right]$

$V_k \leftarrow \frac{1}{N_k}\Sigma_{i:\boldsymbol{\theta}_i\in\Omega_k}\left(y^*_{0,i}-E_k\right)^2$ ▷ Approximates $\mathbb{V}_{b^*|\boldsymbol{\theta}}\left[y^*_0|\boldsymbol{\theta},b^*\right]$

end for

$V_{b^*} \leftarrow \frac{1}{K}\Sigma_{k=1}^K V_k$ ▷ Approximates $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{b^*|\boldsymbol{\theta}}\left[y^*_0|\boldsymbol{\theta},b^*\right]\right]$

$V_{\boldsymbol{\theta}} \leftarrow \frac{1}{K}\Sigma_{k=1}^K\left(E_k-\frac{1}{K}\Sigma_{k=1}^K E_k\right)^2$ ▷ Approximates $\mathbb{V}_{\boldsymbol{\theta}}\left[\mathbb{E}_{b^*|\boldsymbol{\theta}}\left[y^*_0|\boldsymbol{\theta},b^*\right]\right]$

return $V_{b^*}, V_{\boldsymbol{\theta}}$

Appendix B: Random Forest

This appendix briefly reviews random forest, the machine learning technique used for variable selection of surrogate models in the case study and other applications as needed. Here we follow conventional mathematical notations in random forest literature; some notations in this section are defined differently from in section 2.1.

A random forest is composed of an ensemble of Classical and Regression Trees, or CARTs [*Breiman et al.*, 1984]. Let $\{\mathbf{x}_i, y_i\}, i=1,\ldots,n$ denote a set of training data, where $\mathbf{x}_i=\{x_{i,1},\ldots x_{i,p}\}^T$ is an input data point and p is the dimension of the input feature space. A CART recursively partitions the input space into rectangular regions through sequentially binary splitting; a constant value is fit to each region. At every nonterminal node, the splitting variable and split point are chosen to maximize the goodness-of-fit at this node. A maximal tree is grown until the node size, or the minimum number of data points in the leaves (terminal nodes) is below a threshold. The maximal tree is then pruned to prevent overfitting.

One disadvantage of CARTs is that small changes in the training data may lead to large changes in the trained tree structure. Random forest is proposed in *Breiman* [2001] to overcome this statistical instability of CARTs. The random forest algorithm grows multiple CARTs, each trained using a bootstrap sample of the training data. A bootstrap sample is generated by randomly drawing n training data points with replacement. At each split during the construction of a single CART the splitting variable is selected among a random subset of input variables. The node size and size of the candidate subset are two hyperparameters. They are conventionally recommended to be set at 10 and one third of the total number of input variables, respectively. The performance of random forest changes very little over a wide range of the two hyperparameters [*Meinshausen*, 2006; *Svetnik et al.*, 2003]. Pruning of individual CART is not necessary because random forest is not prone to overfitting due to bootstrap aggregation. Once all CARTs are trained, the prediction for an unseen data point $\mathbf{x}^*$ is calculated by averaging the predictions from all individual CARTs.

Bootstrapping leaves out about one third of the data. Comparing these so-called out-of-bag observations with corresponding predictions made by a trained CART, out-of-bag error can be calculated as a measure of the generalization error of the tree. The random forest algorithm then calculates an importance measure of an input variable $\mathbf{x}_j$ by averaging the increase of out-of-bag error after permuting $\mathbf{x}_j$ over all CARTs.

Appendix C: Support Vector Regression

In this appendix, we briefly overview support vector regression (SVR) following conventional mathematical notations in SVR literature; some notations in this section are defined differently from section 2.1. Given a set of training data $\{\mathbf{x}_i, y_i\}, i=1,\ldots,n$, 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 $\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{F}$, and then carry out a linear regression of y in the feature space $\boldsymbol{\phi}(\mathbf{x})$:

$$f(\mathbf{x}) = w \cdot \boldsymbol{\phi}(\mathbf{x}) + b. \tag{C1}$$

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

$$\text{min}imize \ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{C2}$$

$$subject\ to \ (\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + b) - y_i \leq \varepsilon + \xi_i, \tag{C3}$$

$$y_i - (\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + b) \leq \varepsilon + \xi_i^*, \tag{C4}$$

$$\xi_i, \xi_i^* \geq 0, \ i = 1, \ldots, n. \tag{C5}$$

The first term in equation (C2) 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 equation (C2) determines the trade-off between the flatness of f and deviations exceeding ε.

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

$$\langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j), \tag{C6}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2\right). \tag{C7}$$

The regularization hyperparameter C is chosen according to the training data following the recommendations in *Cherkassky and Ma* [2004]. The loss function hyperparameter ε and kernel width hyperparameter γ are tuned by fivefold cross validation.

Acknowledgments

This work is supported by the National Science Foundation Hydrologic Science program under grant 0943627. The first author was also supported by the Computational Science and Engineering Fellowship, College of Engineering, University of Illinois. The third author was supported in part by the NSF-EAR grant 1552329 and DOE Early Career Award DE-SC0008272. The authors thank Yu-Feng Forrest Lin of Illinois State Geological Survey for help with MODFLOW-NWT and Paul A. Hsieh of U.S. Geological Survey for sharing the input, output, and calibration data set of the SVRP model. The authors are grateful for the thoughtful review and suggestions by Hoshin Gupta, two anonymous reviewers, and the Associate Editor. Supporting data are available from the authors upon request.

References

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

Asher, M., B. Croke, A. Jakeman, and L. Peeters (2015), A review of surrogate models and their application to groundwater modeling, *Water Resour. Res.*, *51*, 5957–5973, doi:10.1002/2015WR016967.

Baroni, G., and S. Tarantola (2014), A general probabilistic framework for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study, *Environ. Modell. Software*, *51*, 26–34.

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

Bishop, C. M., and N. M. Nasrabadi (2006), *Pattern Recognition and Machine Learning*, vol. 1, Springer, New York.

Bliznyuk, N., D. Ruppert, C. Shoemaker, R. Regis, S. Wild, and P. Mugunthan (2012), Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation, *J. Comput. Graphical Stat.*, *17*, 270–294.

Breiman, L. (2001), Random forests, *Mach. Learn.*, *45*(1), 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984), *Classification and Regression Trees*, CRC Press, Boca Raton, Fla.

Brynjarsdóttir, J., and A. O'Hagan (2014), Learning about physical parameters: The importance of model discrepancy, *Inverse Probl.*, *30*(11), 114,007.

Chang, C., and C. Lin (2011), LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, *2*(3), 27.

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

Clark, M. P., et al. (2015), A unified approach for process-based hydrologic modeling: 1. Modeling concept, *Water Resour. Res.*, *51*, 2498–2514, doi:10.1002/2015WR017198.

Cloke, H. L., and F. Pappenberger (2008), Evaluating forecasts of extreme events for hydrological applications: An approach for screening unfamiliar performance measures, *Meteorol. Appl.*, *15*(1), 181–197.

Cooley, R. (2004), *A Theory for Modeling Ground-Water Flow in Heterogeneous Media*, U.S. Dep. of the Inter., U.S. Geol. Surv., Reston, Va.

Cowles, M. K., and B. P. Carlin (1996), Markov chain Monte Carlo convergence diagnostics: A comparative review, *J. Am. Stat. Assoc.*, *91*(434), 883–904.

Cui, T., C. Fox, and M. O'Sullivan (2011), Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance metropolis hastings algorithm, *Water Resour. Res.*, *47*, W10521, doi:10.1029/2010WR010352.

Deman, G., K. Konakli, B. Sudret, J. Kerrou, P. Perrochet, and H. Benabderrahmane (2016), Using sparse polynomial chaos expansions for the global sensitivity analysis of groundwater lifetime expectancy in a multi-layered hydrogeological model, *Reliab. Eng. Syst. Safety*, *147*, 156–169.

Doherty, J., and S. Christensen (2011), Use of paired simple and complex models to reduce predictive bias and quantify uncertainty, *Water Resour. Res.*, *47*, W12534, doi:10.1029/2011WR010763.

Doherty, J., and D. Welter (2010), A short exploration of structural noise, *Water Resour. Res.*, *46*, W05525, doi:10.1029/2009WR008377.
Doherty, J., L. Brebber, and P. Whyte (2010), PEST: Model-independent parameter estimation user manual, technical report, Watermark Comput., Corinda, Australia.
Elsheikh, A. H., I. Hoteit, and M. F. Wheeler (2014), Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates, *Comput. Methods Appl. Mech. Eng.*, *269*, 515–537.
Erdal, D., I. Neuweiler, and J. Huisman (2012), Estimating effective model parameters for heterogeneous unsaturated flow using error models for bias correction, *Water Resour. Res.*, *48*, W06530, doi:10.1029/2011WR011062.
Evin, G., M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera (2014), Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resour. Res.*, *50*, 2350–2375, doi:10.1002/2013WR014185.
Freni, G., and G. Mannina (2010), Uncertainty in water quality modelling: The applicability of variance decomposition approach, *J. Hydrol.*, *394*(3), 324–333.
Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*(4), 457–472.
Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, *48*, W08301, doi:10.1029/2011WR011044.
Hill, M., and C. Tiedeman (2007), *Effective Calibration of Groundwater Models, With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, John Wiley, New York.
Hsieh, P., M. E. Barber, B. A. Contor, A. Hossain, G. S. Johnson, J. L. Jones, and A. H. Wylie (2007), Ground-water flow model for the Spokane Valley-Rathdrum Prairie Aquifer, Spokane County, Washington, and Bonner and Kootenai Counties, Idaho, *U.S. Geol. Surv. Sci. Invest. Rep., 2007–5044*, 78 pp.
Kanevski, M., R. Parkin, A. Pozdnukhov, V. Timonin, M. Maignan, V. Demyanov, and S. Canu (2004), Environmental data mining and modeling based on machine learning algorithms and geostatistics, *Environ. Modell. Software*, *19*(9), 845–855.
Kennedy, M. C., and A. O'Hagan (2000), Predicting the output from a complex computer code when fast approximations are available, *Biometrika*, *87*(1), 1–13.
Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, *J. R. Stat. Soc. Ser. B*, *63*(3), 425–464.
Kumar, P. (2011), Typology of hydrologic predictability, *Water Resour. Res.*, *47*, W00H05, doi:10.1029/2010WR009769.
Laloy, E., and J. A. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try $\text{DREAM}_{(ZS)}$ and high-performance computing, *Water Resour. Res.*, *48*, W01526, doi:10.1029/2011WR010608.
Laloy, E., B. Rogiers, J. A. Vrugt, D. Mallants, and D. Jacques (2013), Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion, *Water Resour. Res.*, *49*, 2664–2682, doi:10.1002/wrcr.20226.
Liang, F., K. Mao, M. Liao, S. Mukherjee, and M. West (2007), Nonparametric Bayesian kernel models, discussion paper, pp. 07–10, Dep. of Stat. Sci., Duke Univ., Durham, N. C.
Ließ, M., B. Glaser, and B. Huwe (2012), Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and random forest models, *Geoderma*, *170*, 70–79.
Lin, Y., D. O'Malley, and V. V. Vesselinov (2016), A computationally efficient parallel Levenberg-Marquardt algorithm for highly parameterized inverse model analyses, *Water Resour. Res.*, *52*, 6948–6977, doi:10.1002/2016WR019028.
Liu, Y., and H. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*, W07401, doi:10.1029/2006WR005756.
Lu, D., M. Ye, P. D. Meyer, G. P. Curtis, X. Shi, X.-F. Niu, and S. B. Yabusaki (2013), Effects of error covariance structure on estimation of model averaging weights and predictive performance, *Water Resour. Res.*, *49*, 6029–6047, doi:10.1002/wrcr.20441.
Marzouk, Y., and D. Xiu (2009), A stochastic collocation approach to Bayesian inference in inverse problems, *Commun. Comput. Phys.*, *6*, 826–847.
Meinshausen, N. (2006), Quantile regression forests, *J. Mach. Learn. Res.*, *7*, 983–999.
Mugunthan, P., and C. A. Shoemaker (2006), Assessing the impacts of parameter uncertainty for computationally expensive groundwater models, *Water Resour. Res.*, *42*, W10428, doi:10.1029/2005WR004640.
Nearing, G. S., Y. Tian, H. V. Gupta, M. P. Clark, K. W. Harrison, and S. V. Weijs (2016), A philosophical basis for hydrological uncertainty, *Hydrol. Sci. J.*, *61*, 1666–1678.
Neuman, S. (2003), Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochastic Environ. Res. Risk Assess.*, *17*(5), 291–305.
Pillai, N. S., Q. Wu, F. Liang, S. Mukherjee, and R. L. Wolpert (2007), Characterizing the function space for Bayesian kernel models, *J. Mach. Learn. Res.*, *8*(8), 1769–1797.
Prudic, D. E., L. F. Konikow, and E. R. Banta (2004), A new streamflow-routing (SFR1) package to simulate stream-aquifer interaction with MODFLOW-2000, technical report 2004-1042, U.S. Dep. of the Inter., U.S. Geol. Surv., Carson City, Nev.
Rasmussen, C. E., and C. K. I. Williams (2006), *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Mass.
Rasouli, K., W. Hsieh, and A. Cannon (2011), Daily streamflow forecasting by machine learning methods with weather and climate inputs, *J. Hydrol.*, *414–415*, 284–293.
Razavi, S., B. A. Tolson, and D. H. Burn (2012), Review of surrogate modeling in water resources, *Water Resour. Res.*, *48*, W07401, doi:10.1029/2011WR011527.
Refsgaard, J., J. Van der Sluijs, J. Brown, and P. Van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Adv. Water Resour.*, *29*(11), 1586–1597.
Reichert, P., and N. Schuwirth (2012), Linking statistical bias description to multiobjective model calibration, *Water Resour. Res.*, *48*, W09543, doi:10.1029/2011WR011391.
Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.
Salamon, P., and L. Feyen (2010), Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation, *Water Resour. Res.*, *46*, W12501, doi:10.1029/2009WR009022.
Schöniger, A., T. Wöhling, and W. Nowak (2015), A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking, *Water Resour. Res.*, *51*, 7524–7546, doi:10.1002/2015WR016918.
Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.
Shi, X., M. Ye, G. P. Curtis, G. L. Miller, P. D. Meyer, M. Kohler, S. Yabusaki, and J. Wu (2014), Assessment of parametric uncertainty for groundwater reactive transport modeling, *Water Resour. Res.*, *50*, 4416–4439, doi:10.1002/2013WR013755.

Smola, A. J., and B. Schölkopf (2003), Bayesian kernel methods, in *Advanced Lectures on Machine Learning*, edited by S. Mendelson and A. Smola, pp. 65–117, Springer-Verlag, Berlin.

Svetnik, V., A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston (2003), Random forest: A classification and regression tool for compound classification and QSAR modeling, *J. Chem. Inf. Comput. Sci.*, *43*(6), 1947–1958.

Tonkin, M., J. Doherty, and C. Moore (2007), Efficient nonlinear predictive error variance for highly parameterized models, *Water Resour. Res.*, *43*, W07429, doi:10.1029/2006WR005348.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer, New York.

Vrugt, J. A. (2016), Markov chain Monte Carlo simulation using the dream software package: Theory, concepts, and Matlab implementation, *Environ. Modell. Software*, *75*, 273–316.

Vrugt, J. A., C. J. Ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Vrugt, J. A., C. Ter Braak, C. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 273–290.

Wang, C., Q. Duan, W. Gong, A. Ye, Z. Di, and C. Miao (2014), An evaluation of adaptive surrogate modeling based optimization with two benchmark problems, *Environ. Modell. Software*, *60*, 167–179.

Xie, H., J. W. Eheart, Y. Chen, and B. A. Bailey (2009), An approach for improving the sampling efficiency in the Bayesian calibration of computationally expensive simulation models, *Water Resour. Res.*, *45*, W06419, doi:10.1029/2007WR006773.

Xu, T. (2016), A fully Bayesian approach to uncertainty quantification of groundwater models, PhD thesis, Univ. of Ill. at Urbana-Champaign, Champaign.

Xu, T., and A. J. Valocchi (2015a), A Bayesian approach to improved calibration and prediction of groundwater models with structural error, *Water Resour. Res.*, *51*, 9290–9311, doi:10.1002/2015WR017912.

Xu, T., and A. J. Valocchi (2015b), Data-driven methods to improve baseflow prediction of a regional groundwater model, *Comput. Geosci.*, *85*, 124–136, doi:10.1016/j.cageo.2015.05.016.

Xu, T., A. J. Valocchi, J. Choi, and E. Amir (2014), Use of machine learning methods to reduce predictive error of groundwater models, *Ground Water*, *52*(3), 448–460.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.

Zeng, X., M. Ye, J. Burkardt, J. Wu, D. Wang, and X. Zhu (2016), Evaluating two sparse grid surrogates and two adaptation criteria for groundwater Bayesian uncertainty quantification, *J. Hydrol.*, *535*, 120–134.

Zhang, J., W. Li, L. Zeng, and L. Wu (2016), An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems, *Water Resour. Res.*, *52*, 5971–5984, doi:10.1002/2016WR018598.