AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR019512

Bayesian calibration of groundwater models with input data uncertainty

Tianfang Xu[1,2], Albert J. Valocchi[1], Ming Ye[3], Feng Liang[4], and Yu-Feng Lin[5]

[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, [5]Illinois State Geological Survey, Prairie Research Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

Key Points:
- Neglecting uncertainty in inputs such as pumping and recharge rates may undermine the prediction power of calibrated groundwater models
- A marginalizing Bayesian method accounts for input uncertainty and yields more accurate prediction for a synthetic case study
- Based on variance decomposition analysis, input uncertainty can be the dominant source of prediction uncertainty

Supporting Information:
- Supporting Information S1

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

Citation:
Xu, T., A. J. Valocchi, M. Ye, F. Liang, and Y.-F. Lin (2017), Bayesian calibration of groundwater models with input data uncertainty, *Water Resour. Res.*, *53*, doi:10.1002/2016WR019512.

Received 14 JUL 2016
Accepted 26 MAR 2017
Accepted article online 31 MAR 2017

Abstract Effective water resources management typically relies on numerical models to analyze groundwater flow and solute transport processes. Groundwater models are often subject to input data uncertainty, as some inputs (such as recharge and well pumping rates) are estimated and subject to uncertainty. Current practices of groundwater model calibration often overlook uncertainties in input data; this can lead to biased parameter estimates and compromised predictions. Through a synthetic case study of surface-ground water interaction under changing pumping conditions and land use, we investigate the impacts of uncertain pumping and recharge rates on model calibration and uncertainty analysis. We then present a Bayesian framework of model calibration to handle uncertain input of groundwater models. The framework implements a marginalizing step to account for input data uncertainty when evaluating likelihood. It was found that not accounting for input uncertainty may lead to biased, overconfident parameter estimates because parameters could be over-adjusted to compensate for possible input data errors. Parameter compensation can have deleterious impacts when the calibrated model is used to make forecast under a scenario that is different from calibration conditions. By marginalizing input data uncertainty, the Bayesian calibration approach effectively alleviates parameter compensation and gives more accurate predictions in the synthetic case study. The marginalizing Bayesian method also decomposes prediction uncertainty into uncertainties contributed by parameters, input data, and measurements. The results underscore the need to account for input uncertainty to better inform postmodeling decision making.

1. Introduction

Over the last decade, awareness has grown that water resources systems are changing in response to changes in climate and human activities. For the foreseeable future, various studies suggest shifting patterns of precipitation and groundwater recharge, as well as increasing extremes of flooding and drought [*Held and Soden*, 2006; *Trenberth*, 2011; *Famiglietti*, 2014]. These changes threaten sustainable groundwater supply, which is intertwined with food security and energy production [*Famiglietti*, 2014]. As numerical groundwater models and integrated hydrologic models are increasingly used to inform water resources management decisions and policies under global change, there is increasing need to improve prediction accuracy and assess uncertainty of these models [*Gorelick and Zheng*, 2015].

Existing groundwater uncertainty quantification literature has mainly focused on parameter uncertainty. Various least squares regression-based and Bayesian methods have been developed to infer parameters that characterize the heterogeneous subsurface systems and to propagate parametric uncertainty to prediction uncertainty [e.g., *Doherty*, 2003; *Fienen et al.*, 2009; *Finsterle and Kowalsky*, 2011; *Hill and Tiedeman*, 2007; *Tonkin and Doherty*, 2009]. A common implicit assumption underlying these methods is that the input data are accurate, so that the residual, or the mismatch between model simulation results and corresponding observations, is dominated by measurement error. Such an assumption, however, is often violated in practice.

Input data uncertainty [*Liu and Gupta*, 2007] could arise in groundwater modeling when an indirect method or another model has been used to estimate forcings such as precipitation recharge, percolation from irrigation, evapotranspiration, and well pumping rates. As an example, in the Republican River Compact

Administration (RRCA) model [*McKusick*, 2003], the irrigation pumping rates were estimated based on irrigation acreage, farm efficiency, crop water requirement, and other relevant information. We compared the estimated total annual pumping rate at the county level with metered pumping rate for three counties (Perkins, Chase, and Dundy) in Nebraska from 1980 to 2006 (data courtesy of Dr. Nicholas Brozovic at the University of Nebraska, personal communication). It was found that the estimated annual pumping rate is 1–26% lower than the metered pumping rate; in 21 out of 27 years, the percent error of annual pumping rate is above 10%.

Besides pumping rate, groundwater recharge is a well-known input forcing that is difficult to estimate accurately [*Healy*, 2010; *Scanlon et al*., 2002]. The amount and timing of precipitation recharge can be estimated by various methods such as watershed models and water-budget methods [*Healy*, 2010; *Hsieh et al*., 2007; *Lee and Risley*, 2002]. The estimated input data can contain nonnegligible error, which in turn may lead to spatial and temporal correlation in model residual. While correlation in model residual could be theoretically allowed in least squares regression, groundwater calibration applications often assumes that the model residuals are independent [*Lu et al*., 2013].

Meanwhile in rainfall-runoff modeling, input data uncertainty has been recognized as a source of systematic bias. For example, *Kavetski et al*. [2006a] reasoned that the application of traditional least squares regression ignoring the high spatial and temporal uncertainty of precipitation can lead to biased parameter estimates. This is because parameters could be overly adjusted to compensate for rainfall error when uncertainty in observed rainfall is not taken into account during the calibration process [*Del Giudice et al*., 2013, 2016]. Parameter compensation can have a deleterious impact on model prediction under scenarios that are different from historical conditions reflected by calibration data [*White et al*., 2014], especially in the context of climate change.

One way to account for input uncertainty is to drive a hydrologic model with different calibration forcing data sets [*Mendoza et al*., 2015; *Sapriza-Azuri et al*., 2015] or an ensemble of stochastically generated rainfall time series [*Mockler et al*., 2016]. Alternatively, the uncertain input can be jointly inferred with model parameters in a Bayesian framework. *Kavetski et al*. [2006a] introduced storm event-based multipliers to characterize variability of rainfall, and inversely inferred the multipliers with rainfall-runoff model parameters. This method has been used in later studies [*Huard and Mailhot*, 2008; *Vrugt et al*., 2008; *McMillan et al*., 2011; *Renard et al*., 2010, 2011]. Recently, *Del Giudice et al*. [2016] proposed a stochastic input process (SIP) method and describes rainfall as a stochastic process, the prior of which is updated during calibration. The above studies have found that accounting for precipitation error significantly altered the shape of the posterior distributions of model parameters, highlighting the importance of explicit treatment of input uncertainty.

Various studies in surface hydrology have investigated the role of input data in prediction uncertainty. Within the Bayesian framework, *Renard et al*. [2011] used partial predictive distributions (PPDs) to assess the contributions to runoff prediction uncertainty from rainfall, model structure, remnant error, and runoff measurement error. The PPD of one source is derived as the posterior distribution of the prediction conditioned on the modal values of all other sources. In a case study based on the Yzeron catchment (France) and the conceptual rainfall-runoff model GR4J, it was found that while model structural error is the largest source of uncertainty, rainfall biases induced nonnegligible prediction uncertainty. This Bayesian paradigm gives graphical uncertainty decomposition results that are easy to understand and interpret. However, the decomposition may be affected by the choice of conditioning values. On the contrary, variance-based methods aim to provide quantitative measures of the relative importance of sources of uncertainty. Such measures are global in the sense that they are calculated averaging over the parameter space. Therefore, the variance decomposition results given by these methods do not rely on the choice of conditioning values. However, variance-based methods only provide summary statistics and do not portray the entire distribution of predictions. *Bosshard et al*. [2013] used an analysis of variance model to separate uncertainties due to meterological inputs calculated by climate models, the methods used to postprocess meterological inputs, and conceptual rainfall-runoff models. A case study on the Alpine Rhine (Eastern Switzerland) revealed that none of these uncertainty sources is negligible. *Mockler et al*. [2016] used a similar approach to assess uncertainties in streamflow predictions arising from rainfall and the identification of behavioral parameter sets. For all 32 catchments considered in their study, rainfall is the dominant source of streamflow prediction uncertainty.

In groundwater applications, recharge multipliers can be specified in zones or as pilot points in a way that is similar as rainfall multipliers for rainfall events; the multipliers are then jointly estimated along with other model parameters [*McKusick*, 2003; *White et al*., 2014] using least squares regression-based techniques. However, calibrating all uncertain inputs often results in a high-dimensional inverse problem. For groundwater models, it would take a great number of parameters to describe inputs that vary temporally and spatially. The resulting high-dimensional inverse problem is computationally challenging for both least squares regression and Bayesian calibration. In the context of least squares regression, techniques such as truncated singular value decomposition and Tikhonov regularization [*Tonkin and Doherty*, 2005; *White et al*., 2014] can be used to reduce the number of parameters or impose constraints to mitigate overfitting. Nevertheless, nonuniqueness or nonidentifiability issues may still arise from correlation among model inputs, parameters and output [*Demissie et al*., 2014; *Huard and Mailhot*, 2008; *Renard et al*., 2010, 2011]. In addition, when calibrating input data during the calibration process, the inputs (or parameterized as multipliers) may be overly adjusted to compensate for other sources of error, e.g., model structural error. Since true input is often unknown in modeling practice, validation of the inferred input-associated parameters, and therefore the model parameters and outputs, is difficult [*Ajami et al*., 2007].

Recently, *Demissie et al*. [2014] proposed an input uncertainty weighted least-squares (IUWLS) method to account for pumping rate uncertainty. The IUWLS method assumes that the pumping rates follow a Gaussian distribution with fixed mean and variance. It then propagates the pumping rate variance to calibration targets in order to determine their least-square weights in the objective function. In this way, the IUWLS method avoids the nonidentifiability issue due to high dimensionality and the correlation among pumping rates and model parameters (such as transmissivity). The method was demonstrated through a synthetic case study of cyclic pumping in a confined homogeneous aquifer. The IUWLS method requires that an unbiased estimate or observation of pumping rates along with quantitative assessment of associated uncertainty is available before carrying out calibration. In circumstances where the initially estimated or observed input data could be biased, and/or associated uncertainty cannot be quantified, it is more desirable to infer the level of bias and uncertainty through calibration. More studies are needed to better understand the impacts of input data uncertainty on calibration and prediction in more general settings.

With the motivation of reducing the dimensionality of the rainfall multiplier approach in *Kavetski et al*. [2006b], *Ajami et al*. [2007] treated multipliers as independent samples drawn from an identical Gaussian distribution. The mean and variance (hyperparameters) of the Gaussian distribution are estimated along with model parameters during Bayesian calibration. *Renard et al*. [2009] further theoretically discussed an "expected likelihood," which is defined as the likelihood integrated over rainfall multipliers as latent variables. This study is the first attempt to extend this formulation for uncertain inputs in groundwater models. Since the calibration process does not infer each individual uncertain input (or parameterized as multiplier), the nonidentifiability issue due to high dimensionality and the correlation among pumping rates and model parameters is mitigated. In addition to the different application fields, this study differs from preceding studies in rainfall-runoff modeling [*Ajami et al*., 2007, 2009; *Huard and Mailhot*, 2008; *Renard et al*., 2010, 2011] in that rather than to sample the multipliers, we implement a marginalizing step when evaluating the likelihood in order to account for input data uncertainty.

The marginalizing Bayesian method of this study allows for an assessment of prediction uncertainty from various sources including measurement errors, model parameters, and input data. As the second focus of this study, we perform variance decomposition analysis using the Bayesian inference results. By identifying the primary source(s) of prediction uncertainty, the variance decomposition results could inform future data collection efforts on how to best direct resources toward reducing prediction uncertainty. The variance decomposition scheme used in this study provides quantitative measures of the relative importance of sources of uncertainty. Similar approaches have been used in hydrological studies [*Bosshard et al*., 2013; *Dai and Ye*, 2015; *Mockler et al*., 2016] to separate uncertainties arising from different sources. However, these applications did not explicitly consider the impact of input uncertainty on calibrated parameters.

The main objectives of this paper are (1) investigate how uncertainties in groundwater pumping and recharge rates would affect parameter estimates and resulting predictions, (2) present a method to alleviate negative effects, and (3) perform variance decomposition analysis to quantify the contribution of input uncertainty to prediction uncertainty. In section 2, we present a marginalizing Bayesian calibration and uncertainty analysis method tailored for groundwater models that are subject to input data uncertainty. We

also present a variance decomposition method to analyze how much prediction uncertainty is contributed by measurement error, model parameters, and uncertain calibration inputs. In section 3, we evaluate the marginalizing method using a more realistic synthetic case study of surface-ground water interaction under changing pumping and land use conditions. The results reported and discussed in section 4 show that explicit treatment of uncertainties in input data (groundwater pumping rate and recharge rate) yields substantially different parameter estimates and more robust predictions when compared to classical Bayesian calibration that does not account for input data uncertainty. Finally, section 5 concludes and provides recommendations.

2. Methods

In order to better establish the mathematical formulation of the problem that we aim to tackle in the context of groundwater modeling, we first introduce an illustrative simple example to demonstrate the impacts of ignoring input data uncertainty on calibration and prediction. We then present a marginalizing Bayesian method of calibration and uncertainty analysis with the presence of input data uncertainty. We then demonstrate the use of the marginalizing method through the illustrative example. Lastly, in this section, we present a variance decomposition method to analyze how much prediction uncertainty is contributed by measurement error, model parameters, and uncertain calibration inputs.

2.1. An Illustrative Simple Example

As an illustrative example, we consider steady state flow in an ideal confined aquifer to a group of 10 wells distributed on a circle. The drawdown at the center of the circle of wells can be calculated by superposition based on the Thiem equation:

$$s = \frac{\sum_i Q_i}{2\pi T} \ln\left(\frac{R}{r}\right), \tag{1}$$

where drawdown s is the quantity of interest in this example; R denotes the radius of influence, i.e., the distance beyond which the groundwater head is not affected by pumping; r is the radius of the well circle, i.e., the distance from the drawdown monitoring location to any well; Q_i is the pumping rate of the ith well; T denotes the transmissivity of the confined aquifer. Assume that an observation s_0 is taken from the "true" system:

$$s_0 = \frac{\sum_i Q_{i,0}}{2\pi T_0} \ln\left(\frac{R}{r}\right) + \epsilon. \tag{2}$$

where $Q_{i,0}, i=1,\ldots,10$, denotes the true yet unknown pumping rates, T_0 is the true yet unknown transmissivity, and ϵ denotes measurement error that follows $N(0,\sigma_\epsilon^2)$. Here σ_ϵ^2 is the measurement error variance, which is usually known for drawdown observations. In this illustrative example, we will generate a noise-free measurement s_0 to exclude the impact of random measurement error of the single drawdown measurement we will be using. The variance of measurement error, σ_ϵ^2, will still be used when evaluating the likelihood.

Next, suppose that estimated pumping rates, $\hat{Q}_i, i=1,\ldots,10$, are available. We further introduce some notations to simplify the form of equations (1) and (2). Let $\theta = 1/T$ and $k = \frac{1}{2\pi}\ln\left(\frac{R}{r}\right)$, then the true system response is $s_0 = k\sum_i Q_{i,0}\theta_0$. In order to estimate θ, we calibrate the following model using the observation:

$$s = k\sum_i \hat{Q}_i\,\theta. \tag{3}$$

Here it is assumed that k is exactly known. It can be seen that s is now linear with respect to both the inputs $\hat{Q}_i, i=1,\ldots,10$ and the new parameter θ.

Next, we assume that the estimated pumping rates are exact and use Bayesian calibration to infer θ. Using Bayes' theorem, the posterior distribution of the parameter θ is

$$p(\theta|s_0) \propto L(\theta|s_0)\cdot p(\theta), \tag{4}$$

and the likelihood is given by

$$s_0|\theta \sim N\left(k\sum_i \hat{Q}_i\theta, \sigma_\epsilon^2\right). \tag{5}$$

Using a uniform distribution over a wide range for θ as the prior, i.e., $p(\theta) \propto 1$, the posterior is

$$\theta|s_0 \sim N\left(\frac{\sum_i Q_{i,0}}{\sum_i \hat{Q}_i}\theta_0, \frac{\sigma_\epsilon^2}{k^2\left(\sum_i \hat{Q}_i\right)^2}\right). \tag{6}$$

The derivation of the above equation can be found in Appendix A. The *maximum a posterior* (MAP) estimate of θ will be biased if the estimated pumping rate is overall biased, i.e., $\sum_i \hat{Q}_i \neq \sum_i Q_{i,0}$. In other words, the parameter θ is overly adjusted to compensate for the biased estimate of pumping rate. Furthermore, as input uncertainty is neglected, the variance term is determined by head measurement error variance. With more data used for calibration, the posterior variance of θ will become smaller, i.e., we become more sure about the wrong estimation of θ. As a result, the credible interval may fail to cover the true value.

Next we use the inferred posterior distribution $p(\theta|s_0)$ to make prediction of the steady state drawdown at a different pumping rate Q^*. It follows that the true prediction $s_0^*=k\sum_i Q_i^*\theta_0$, and the model prediction is

$$s^*|s_0 \sim N\left(\frac{\sum_i Q_{i,0}}{\sum_i \hat{Q}_i}\cdot s_0^*, \frac{\left(\sum_i Q_i^*\right)^2}{\left(\sum_i \hat{Q}_i\right)^2}\cdot\sigma_\epsilon^2\right). \tag{7}$$

It can be seen that the mean of the prediction will be biased if $\sum_i \hat{Q}_i \neq \sum_i Q_{i,0}$. Similarly as in equation (6), the variance term is small when the pumping rates for prediction and calibration are of similar magnitudes and will become smaller with more calibration data. Therefore, the prediction conditioned on calibration data, $s^*|s_0$, may not encompass the true value s_0.

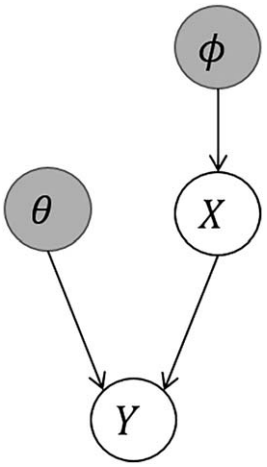

Figure 1. A Bayesian network showing the hierarchical conceptualization of the input data uncertainty. The nodes Y, $\boldsymbol{\theta}, X, \phi$ denote model output, parameters, input data, and input uncertainty hyperparameters as random variables. Arrows connecting nodes represent conditional dependencies among the random variables. In the calibration process, the posterior distributions of the two shaded random variables ($\boldsymbol{\theta}$ and ϕ) are to be inferred using observations of X and Y.

Finally, it is worth mentioning that for linear and moderately nonlinear systems, the Bayesian MAP and credible interval of θ can be numerically identical to least squares regression results when observation errors are multivariate Gaussian distributed and consistent priors are used [*Lu et al.*, 2012]. Therefore, the biasedness and overconfidence issues also occur in least squares-based calibration.

2.2. A Marginalizing Bayesian Method

In this section, we introduce the general formulation of the marginalizing Bayesian method of calibration and uncertainty analysis. Assume that a groundwater system can be represented as

$$y=f(\mathbf{x},\boldsymbol{\theta})+\epsilon, \tag{8}$$

where y is the quantity of interest that can be observed, f denotes a model with inputs $\mathbf{x}$ and parameter $\boldsymbol{\theta}$, and ϵ is the measurement error. Both y and f can be vectors that denote the system output at various times and locations. Here we omit input data that are known well, and hence $\mathbf{x}$ only refers to uncertain inputs. Parameters $\boldsymbol{\theta}$ typically include hydraulic conductivity, storativity, and dispersivity, among other hydrogeologic properties.

Based on the observation that input data can often be estimated from other sources of information with small to medium degree of bias and uncertainty, we formulate the input uncertainty hierarchically as shown in Figure 1. The Bayesian network assumes independency between model parameters $\boldsymbol{\theta}$ and the uncertain inputs $\mathbf{x}$. It is further assumed that given $\mathbf{x}$, the model output is independent of hyperparameters describing the input uncertainty, $\boldsymbol{\phi}$. A similar conceptualization, albeit with different motivation and implementation, can be found in *Ajami et al*. [2007] for rainfall-runoff modeling.

With $\mathbf{x}=\{x_i\}, i=1,\ldots,p$ denoting the true yet unknown inputs and $\hat{\mathbf{x}}=\{\hat{x}_i\}, i=1,\ldots,p$ denoting the corresponding observation or estimation, we introduce multipliers

$$x_i=\psi_i\hat{x}_i. \tag{9}$$

For example, $x_i, i=1,\ldots,p$ can denote the pumping rates at various wells. We then assume that the multipliers are independent and follow a Gaussian distribution:

$$\psi_1,\ldots,\psi_i,\ldots \overset{iid}{\sim} N(\mu_x,\sigma_x^2), \tag{10}$$

where hyperparameters $\boldsymbol{\phi}=\{\mu_x,\sigma_x\}$ are introduced to describe possible bias and uncertainty associated with one class of inputs. For example, $\mu_x<1$ suggests that this class of inputs is overall overestimated, while a larger σ_x allows the multipliers to deviate more from the mean μ_x. One set of $\boldsymbol{\phi}$ is specified for each class or type of inputs. In section 3, we will have one set of hyperparameters for pumping rates and another for recharge rates. In other cases, it may be desirable to divide, e.g., pumping rates into groups, and specify a set of hyperparameters for each group.

Similar to the multiplier approach in rainfall-runoff modeling [*Kavetski et al.*, 2006b], the multiplicative form in equation (9) allows for heteroscedastic errors. For inputs that are known to have low to medium level of bias and uncertainty, relatively informative prior can be specified to constrain the hyperparameters μ and σ.

Let $\mathbf{y}=\{y_i\}, i=1,\ldots,n$ denote a set of calibration data, and assume that the associated measurement errors $\epsilon_1,\ldots,\epsilon_n$ follow a multivariate Gaussian distribution with zero mean and a covariance matrix Σ_ϵ. It follows that the probability of observing $\mathbf{y}$ given the true input $\mathbf{x}$ and parameters $\boldsymbol{\theta}$ is given by

$$\mathbf{y}|\mathbf{x},\boldsymbol{\theta}\sim N(f(\mathbf{x},\boldsymbol{\theta}),\Sigma_\epsilon). \tag{11}$$

In practice, it is often assumed that measurement errors associated with the same type of observations are i.i.d. Gaussian with zero mean and a constant variance $\sigma_{\epsilon,k}^2$, where k denotes kth type of observations. This leads to a diagonal covariance matrix. In the calibration process, we infer the posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ simultaneously. In order to account for the uncertainty associated with $\mathbf{x}$, we derive the *marginal* likelihood:

$$L(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{y})=p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\phi})=\int p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\phi})d\mathbf{x}. \tag{12}$$

Because of the integration step to derive the marginal likelihood in the above equation, the presented method is referred to as the marginalizing method hereafter. The calculation of the marginal likelihood is described in Appendix B. Following equation (12), the posterior distribution of parameters $\boldsymbol{\theta}$ and input error model hyperparameters $\boldsymbol{\phi}$ can be written as

$$p(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{y})\propto\int p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\phi})d\mathbf{x}\cdot p(\boldsymbol{\theta})p(\boldsymbol{\phi}). \tag{13}$$

In practical applications, the posterior distribution in equation (13) 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 developed in *Vrugt et al.* [2009] and *Laloy and Vrugt* [2012] to sample from the posterior distribution $p(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{y})$. During the MCMC sampling, the marginal likelihood in equation (12) is used to calculate the acceptance ratio of a proposed sample. This is a key difference between this study and preceding studies with respect to numerical calculation of the posterior distribution [*Ajami et al.*, 2007; *Huard and Mailhot*, 2008; *Renard et al.*, 2010, 2011].

Next, the model is used to provide prediction at an unsampled location, in a future time, and/or under a new scenario that is different from the calibration data. Letting $\mathbf{x}^*$ denote uncertain inputs in a prediction scenario, a prediction can be calculated as $y^*=f(\mathbf{x}^*,\boldsymbol{\theta})+\epsilon$. Here it is assumed that given $\mathbf{x}^*$ and $\boldsymbol{\theta}$, y^* is not dependent on $\mathbf{y}$. For conciseness, $\mathbf{x}^*$ includes uncertain input data in the past, because they can have a persistent effect on model states in the prediction period. The prediction posterior distribution can be derived by marginalizing $\mathbf{x}^*$:

$$p(y^*|\boldsymbol{\theta},\boldsymbol{\phi})=\int p(y^*|\mathbf{x}^*,\boldsymbol{\theta})p(\mathbf{x}^*|\boldsymbol{\phi})d\mathbf{x}^*. \tag{14}$$

The distribution of a prediction y^* can then be inferred by integrating over the posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$:

$$p(y^*|\mathbf{y}) = \int p(y^*|\boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) d\boldsymbol{\theta} d\boldsymbol{\phi}. \tag{15}$$

Using posterior samples $\{\boldsymbol{\theta}_i, \boldsymbol{\phi}_i\}, i=1,\ldots,N$ generated by MCMC, Bayesian inference of prediction uncertainty can be performed following the procedures described in Appendix C. We first run the model to calculate $f(\mu_{x,i}\mathbf{x}^*, \boldsymbol{\theta}_i)$ and the derivatives of f with respect to $\mathbf{x}^*$. Next, we derive the marginal posterior $p(\mathbf{y}^*|\boldsymbol{\theta}_i, \boldsymbol{\phi}_i)$; $\mathbf{y}^*$ is a vector of predictions at various locations and time. We then draw one realization $\mathbf{y}_i^*$ from $p(\mathbf{y}^*|\boldsymbol{\theta}_i, \boldsymbol{\phi}_i)$. This process is repeated for every sample $\{\boldsymbol{\theta}_i, \boldsymbol{\phi}_i\}, i=1,\ldots,N$ and yields $\mathbf{y}_i^*, i=1,\ldots,N$. Finally, the posterior mean is given by $\overline{\mathbf{y}^*} = \sum_{i=1}^{N} \mathbf{y}_i^*$. Prediction credible intervals can be derived by sorting $\mathbf{y}_i^*, i=1,\ldots N$ to find quantiles.

2.3. Tested on the Illustrative Example

Based on the general formulation in section 2.2, now we consider the pumping rates $Q_i, i=1,\ldots,10$ as uncertain inputs:

$$Q_i|\hat{Q}_i, \mu, \sigma \sim N\left(\mu\hat{Q}_i, \sigma^2\hat{Q}_i^{\,2}\right). \tag{16}$$

For the linear model $s = k\sum \mu\hat{Q}_i\theta$, given θ, μ, σ^2, the variance of the output s is $k^2\sum_i \hat{Q}_i^2\sigma^2\theta^2$. Therefore, the likelihood is

$$L(\theta, \mu, \sigma^2|s_0) = p(s_0|\theta, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\left(k^2\sum_i \hat{Q}_i^2\sigma^2\theta^2 + \sigma_\epsilon^2\right)}} \exp\left[-\frac{(s_0 - k\hat{Q}\mu\theta)^2}{2\left(k^2\sum_i \hat{Q}_i^2\sigma^2\theta^2 + \sigma_\epsilon^2\right)}\right]. \tag{17}$$

Applying Bayes' theorem, the joint posterior distribution of θ, μ, σ^2 is

$$p(\theta, \mu, \sigma^2|s_0) \propto L(\theta, \mu, \sigma^2|s_0) p(\theta, \mu, \sigma^2). \tag{18}$$

As an example, we first generated a synthetic measurement s_0 using $Q_{1,0} = \ldots = Q_{10,0} = 10m^3/day$, $R = 10000m$, $r = 10m$, $\theta_0 = 1/T_0 = 0.1m^{-1}$. As explained in section 2.1, we did not add noise to the synthetic measurement as we would like to focus on input uncertainty. Next, we generated "estimated" pumping rates $\hat{Q}_1, \ldots, \hat{Q}_{10}$ independently from the Gaussian distribution with mean 8 and variance 1.6^2. Compared to the true pumping rates $Q_{1,0}, \ldots, Q_{10,0}$, it can be seen that $\hat{Q}_1, \ldots, \hat{Q}_{10}$ contain errors (i.e., the estimated rates are biased lower than the true pumping rate). We then estimated θ using the classical Bayesian calibration method. The measurement error variance σ_ϵ^2 is set to 0.05^2m^2, a common value for head observations. Applying equation (6), the posterior of θ peaks at 0.125 and does not encompass the true value $\theta_0 = 0.1$, as shown in Figure 2. This is consistent with the analysis in section 2.1 that when neglecting input uncertainty the parameter estimate can be biased and overconfident.

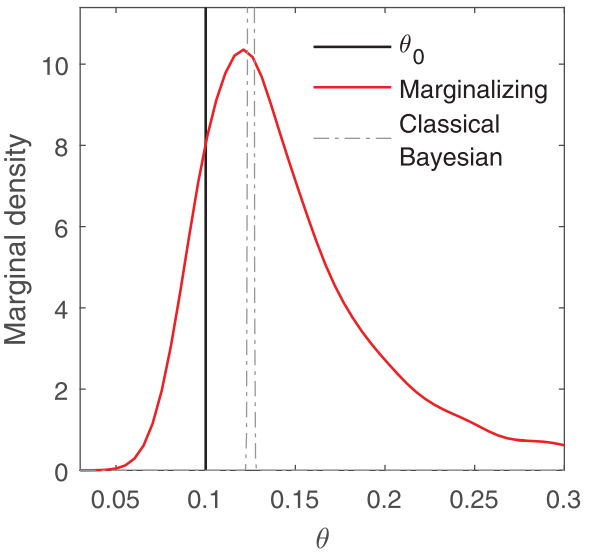

Figure 2. The posterior distribution of θ computed using methods described in sections 2.1 and 2.2, respectively. The peak of the posterior given by the classical Bayesian method (dashed grey) is out of scope of the plot.

Lastly we applied the marginalizing method. For illustration purpose, we assume *a priori*

that the parameters θ and hyperparameters μ, σ^2 are independent, i.e., $p(\theta, \mu, \sigma^2) = p(\theta)p(\mu)p(\sigma^2)$. When prior knowledge about their correlation is available, a joint prior distribution can be specified. We choose $p(\theta) \propto 1$ and for σ^2 the Gamma distribution with mean 0.04 and mode 0.02. For the linear model $s = k\sum \mu \hat{Q}_i \theta$, θ and μ are correlated: there are infinite number of combinations of θ and μ that can produce good fit to s_0. Therefore, we specify a Gaussian prior $\mu \sim N(1, \sigma_\mu^2)$ such that it is just sufficiently informative to reduce nonidentifiability of θ and μ. To obtain the marginal posterior of θ, we need to integrate out μ and σ from equation (18). Unfortunately, $p(\theta|s_0)$ does not have a closed form. Figure 2 shows the posterior obtained numerically using DREAM-ZS [*Laloy and Vrugt*, 2012; *Vrugt et al*., 2009]. Compared to the result given by classical Bayesian calibration, the posterior of the marginalizing method is flatter and encompasses the true value θ_0. For the linear system, similar findings are expected for prediction.

2.4. Variance Decomposition

Based on the inference results of the marginalizing Bayesian method, we investigate how much the total prediction variance can be explained by uncertainties in input data, model parameter, and measurement error. Applying the law of total variance and following the notations in section 2.2, the variance of a prediction y^* can be written as [*Dai and Ye*, 2015]

$$\mathbb{V}[y^*] = \mathbb{E}_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{x}^*|\boldsymbol{\theta}}\mathbb{V}[y^*|\boldsymbol{\theta}, \mathbf{x}^*] + \mathbb{E}_{\boldsymbol{\theta}}\mathbb{V}_{\mathbf{x}^*|\boldsymbol{\theta}}\mathbb{E}[y^*|\boldsymbol{\theta}, \mathbf{x}^*] + \mathbb{V}_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{x}^*|\boldsymbol{\theta}}\mathbb{E}[y^*|\boldsymbol{\theta}, \mathbf{x}^*] \tag{19}$$

In the above equation, we dropped the notations indicating conditioning on the calibration data for the convenience of mathematical expression. On the right-hand side, $\mathbb{V}[y^*|\boldsymbol{\theta}, \mathbf{x}^*]$ and $\mathbb{E}[y^*|\boldsymbol{\theta}, \mathbf{x}^*]$ are the mean and variance of y^* given by the model with associated parameters and input data. The three terms on the right-hand side represent uncertainties of measurement, input data, and model parameters, respectively. The first term $\mathbb{E}_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{x}^*|\boldsymbol{\theta}}\mathbb{V}[y^*|\boldsymbol{\theta}, \mathbf{x}^*]$ is the variance of measurement error. The second term first evaluates the variance of the model output conditioned on $\boldsymbol{\theta}$. The conditional variance itself is a random variable, whose value depends on $\boldsymbol{\theta}$. By calculating the expected value (over $\boldsymbol{\theta}$) of $\mathbb{V}_{\mathbf{x}^*|\boldsymbol{\theta}}\mathbb{E}[y^*|\boldsymbol{\theta}, \mathbf{x}^*]$, this term quantifies the variance of y^* attributed to uncertain input data. Meanwhile, $\mathbb{E}_{\mathbf{x}^*|\boldsymbol{\theta}}\mathbb{E}[y^*|\boldsymbol{\theta}, \mathbf{x}^*]$ is the expected value of y^* conditioned on $\boldsymbol{\theta}$, and the third term represents prediction variance as explained by model parameters.

The second and third terms are approximated using the following procedures. As explained in Appendix C, noise-free predictions $y^*_{0,i}, i = 1, 2, \ldots$ are generated based on posterior samples $\{\boldsymbol{\theta}_i, \boldsymbol{\phi}_i\}$; $y^*_{0,i}$ is the same as y^*_i except that the latter contains measurement error. Then the posterior samples of model parameters $\boldsymbol{\theta}_i, i = 1, \ldots, N$, are divided into groups. Accordingly, the corresponding predictions $y^*_{0,i}, i = 1, \ldots, N$ are divided into bins. We then calculate the mean and variance of $y^*_{0,i}$ within each group. Lastly, the input data uncertainty term $\mathbb{E}_{\boldsymbol{\theta}}\mathbb{V}_{\mathbf{x}^*|\boldsymbol{\theta}}\mathbb{E}[y^*|\boldsymbol{\theta}, \mathbf{x}^*]$ is approximated by the mean of bin-wise variance, and the parameter uncertainty term $\mathbb{V}_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{x}^*|\boldsymbol{\theta}}\mathbb{E}[y^*|\boldsymbol{\theta}, \mathbf{x}^*]$ is approximated by the variance of bin-wise mean.

Meanwhile, the left-hand side can be calculated as the variance of the posterior samples $y^*_i, i = 1, \ldots, N$. We then calculate the fraction of the total prediction variance that is explained by uncertainties in measurement, input data, and parameters, respectively, by dividing the three terms on the right-hand side of equation (19) with $\mathbb{V}[y^*]$.

3. Synthetic Case Study

In this section, we describe a synthetic case study used to investigate the impact of input data uncertainty on calibration and prediction and test the performance of the proposed Bayesian approach in a more realistic setting. The case study is based on a synthetic groundwater model. We first use the model, with a set of "true" parameters, to generate synthetic observations. Hereafter, we will refer to the model with true parameters as virtual reality. We then perform calibration experiments, in which the true parameter values are unknown and are to be estimated using synthetic observations. Next, we use the calibrated model to make forecasts under changing scenarios. The synthetic case study simulates the effect of pumping on two-dimensional groundwater flow in an unconfined aquifer that is hydraulically connected to a stream. The case study described in this section assumes no model structure error, i.e., the model is known perfectly except for the value of model parameters and input data. In supporting information Text S1, we discuss another synthetic case study based on *Xu and Valocchi* [2015], in which the model is a corrupted version of the virtual reality to reflect imperfect knowledge that is common in groundwater modeling applications.

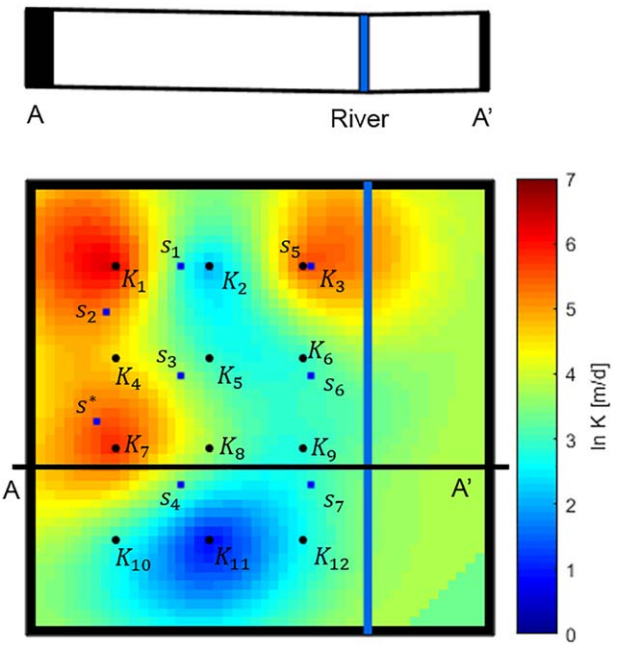

Figure 3. Modeling domain and cross section showing the unconfined unit with a stream running from north to south. Blue squares indicate locations of drawdown calibration targets $s_1, \ldots, s_7$ and evaluation data s^*. Color encodes the K field of the virtual reality. In the model, the K field is interpolated from the K values at 12 pilot points ($K_1, \ldots, K_{12}$).

There, we examine the applicability of the marginalizing method with the presence of model structural error.

3.1. General Setup

The synthetic model is a transient MODFLOW2000 [*Harbaugh et al*., 2000] model of an unconfined aquifer with impermeable bottom and surrounding boundaries. The model has a single layer, 50 $\times$ 50 grid cells and a uniform grid size of $200\times200m^2$ (Figure 3), running for 6 years with monthly stress period and time step. The model has straight surrounding boundaries and linearly inclined bottom elevation. The slope in the north to south direction is 0.0005, and the slope from sides to stream is 0.001.

The virtual reality has a homogeneous specific yield of 0.2. The natural logarithm of conductivity field *lnK* was interpolated from 12 pilot points using Ordinary Kriging and a spherical variogram with a sill of 2 and a range of 4 km. The conductivity field in the virtual reality is shown in Figure 3. In the model, the log conductivity values at pilot points will be calibrated.

The stream is simulated using the MODFLOW SFR1 package [*Prudic et al*., 2004]. The stream stage is routed using Manning's formula at each time step with Manning's *n* of 0.03 and a streambed slope of 0.0005. A rectangular streambed cross section is used, and the channel width is 14 m. The streambed hydraulic conductivity is uniform and equals 1 m/d in the virtual reality. Seasonally, varying inflow is specified at the inlet at the north boundary. For the model, the inflow is generated by randomly perturbing the inflow in the virtual reality according to a coefficient of variation (*CV*) of 0.01 for streamflow measurements.

3.2. Input Data

The synthetic case study considers two common types of uncertain input data: groundwater pumping and precipitation recharge. Recharge rates are specified with four zones (Figure 4). Zones 1, 2, and 3 receive recharge from precipitation only. Zone 4 represents farm land; during the growing season it receives precipitation recharge and groundwater irrigation return flow, which is assumed to equal 20% of total pumping rates in that zone divided by the area. The return flow rate 20% is chosen according to commonly reported irrigation efficiency [*McKusick*, 2003]. Figure 4c shows the monthly varying recharge rates for four zones in the virtual reality, denoted as $R_{1,0}, R_{2,0}, R_{3,0}, R_{4,0}$; the subscripts 0 denote true but unknown values, which are different from "estimated" values described in section 4. The values of $R_{1,0}, R_{2,0}, R_{3,0}, R_{4,0}$ are specified based on typical recharge condition in the Nebraska portion of the Republican River basin [*McKusick*, 2003].

The case study simulates four pumping wells; their locations are marked in Figure 4b. Among the four wells, A and C are municipal supply wells and are pumped at a constant rate, while B and D are irrigation wells and are turned on during the growing season. Wells A and B start pumping from the first transient stress period. The other two wells, C and D start pumping from the evaluation period. The pumping rates at four wells $Q_{A,0}, Q_{B,0}, Q_{C,0}, Q_{D,0}$ are shown in Figure 4d. Likewise for recharge rates, the subscripts 0 denote true but unknown pumping rates used in the virtual reality.

3.3. Calibration and Evaluation Data

The simulation starts from a steady state stress period with no groundwater pumping, which mimics the natural equilibrium state before development. The virtual reality then runs for 6 years and generates quarterly synthetic drawdown (*s*) at locations shown in Figure 3 and stream gain-and-loss (ΔQ) observations, the

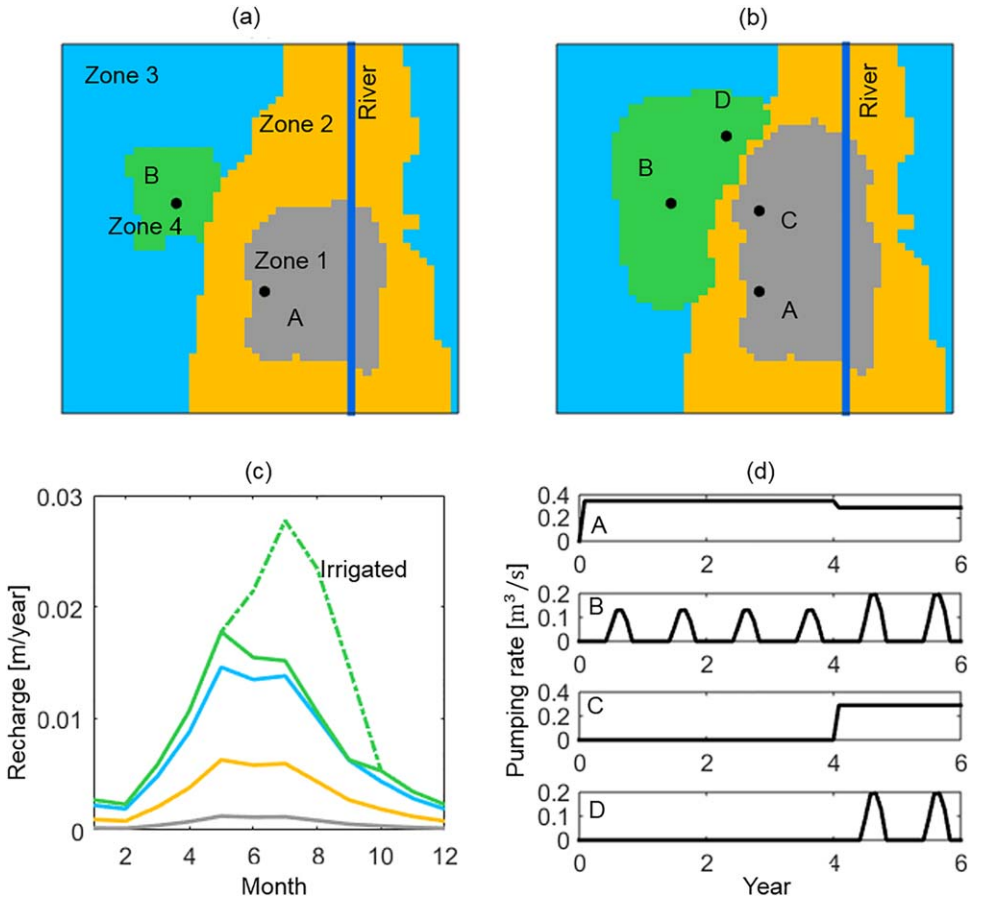

Figure 4. Recharge zones and location of pumping wells during the (a) calibration and (b) evaluation periods. The recharge rates for each zone are shown in Figure 4c, and pumping rates at four wells are shown in Figure 4d. In Figure 4c, colors identify the corresponding zones as shown in Figure 4a; the solid line plots the precipitation recharge, and the dashed line shows the total recharge including irrigation return flow during crop growth season.

most commonly used types of observation when calibrating a groundwater flow model. Drawdown targets are computed by subtracting the groundwater head at a time step from the head at steady state. The stream gain-and-loss (ΔQ) is calculated as the total flow rate from the stream to the aquifer, summed across the whole reach.

The synthetic observations in the first 4 years are contaminated with measurement error and used to calibrate the model. The drawdown measurement error is assumed to be independent and Gaussian distributed with zero mean and a standard deviation of 0.02 m. The streamflow measurement is also independent and Gaussian distributed with zero mean and a coefficient of variation ($CV_{\Delta Q}$) of 0.01. The streamflow measurement error variance is computed by summing up the variance of upstream inflow and downstream outflow [*Hill and Tiedeman*, 2007]. A relatively low streamflow measurement $CV_{\Delta Q}$ is assumed because the case study is intended to focus on uncertainties other than measurement error. We denote the calibration targets as $D=\{s_{i,t}, \Delta Q_t\}$, where $i=1,\ldots,7$ corresponds to the seven drawdown observation locations and $t=1,\ldots,16$ is the number of drawdown (per location) or ΔQ observations.

Synthetic data in the remaining 2 years are reserved for evaluation. As indicated in section 3.2, the total pumping rate during the evaluation period is higher than during the calibration period. The evaluation period represents an increased groundwater demand scenario that is substantially different from the calibration period.

3.4. Three Calibration Experiments

As an illustration of input data error, we consider the case that during the calibration period the pumping rates used in the model are overestimated, while the recharge rates are underestimated. More specifically,

$\hat{Q}_A = 1.2Q_{A,0}, \hat{Q}_B = 1.4Q_{B,0}, \hat{R}_1 = R_{1,0}, \hat{R}_2 = 0.6R_{2,0}, \hat{R}_3 = 0.8R_{3,0}, \hat{R}_4 = 0.75R_{4,0}$. In this situation, the bias from overestimation of pumping rate and the bias from underestimation of recharge rate cannot cancel out and are expected to induce parameter compensation.

We carry out three sets of experiments, each using a different calibration strategy. Experiment A serves as the benchmark; we perform classical Bayesian calibration using inaccurate pumping and recharge rates, $\hat{Q}_A, \hat{Q}_B, \hat{R}_1, \ldots, \hat{R}_4$. The synthetic data during the first 4 years as described in section 3.3 were used to calibrate 16 parameters, namely the specific yield (S_y), natural logarithm of the hydraulic conductivity of the streambed ($\ln K_{rb}$) and at locations given by the pilot points ($\ln K_1, \ldots, \ln K_{12}$), the drawdown measurement error standard deviation (σ_s), and the stream gain-and-loss measurement coefficient of variation ($CV_{\Delta Q}$). Here σ_s and $CV_{\Delta Q}$ are likelihood parameters that will be jointly inferred with $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The likelihood function is given by

$$p(D|\mathbf{x}, \boldsymbol{\theta}, \sigma_s, CV_{\Delta Q}) = \prod_{i=1}^{7}\prod_{t=1}^{16} \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left\{-(s_{i,t} - \hat{s}_{i,t})^2 / 2\sigma_s^2\right\} \cdot \prod_{t=1}^{16} \frac{1}{\sqrt{2\pi}\sigma_{\Delta Q,t}} \exp\left\{-(\Delta Q_t - \Delta\hat{Q}_t)^2 / 2\sigma_{\Delta Q,t}^2\right\}. \quad (20)$$

The above equation is a special case of equation (11) for the two kinds of data. Here $D = \{s_{i,t}, \Delta Q_t\}, i = 1, \ldots, 7, t = 1, \ldots, 16$ denotes the calibration data as defined in section 3.3; $\hat{s}_{i,t}$ and $\Delta\hat{Q}_t$ denote the model simulation results given parameters $\boldsymbol{\theta}$ and input data $\mathbf{x}$; $\sigma_{\Delta Q,t}^2$ is the measurement error variance of the tth stream gain-and-loss calibration targets, and is given by $CV_{\Delta Q}^2$ multiplied by the sum of the square of upstream inflow and the square of downstream outflow.

Table 1 summarizes the priors of the 16 parameters. The priors of σ_s and $CV_{\Delta Q}$ follow an exponential distribution with a mean of 0.05. The prior distribution of specific yield S_y is specified as a Gaussian distribution with a mean of 0.18 and a standard deviation of 0.036 as S_y is usually well constrained *a priori*. Relatively vague prior distributions are specified for hydraulic conductivity. The joint prior distribution of $\ln K_1, \ldots, \ln K_{12}$ is specified as a multivariate Gaussian distribution with a mean of 4.1 and a covariance matrix Σ_K. The covariance matrix is computed using the variogram used to interpolate the log hydraulic conductivity from pilot points (a spherical variogram with a range of 4 km and a sill of 2). In Table 1, the mean of prior distributions of *lnK* at pilot points are different from the true value in the virtual reality to reflect imperfect prior knowledge that is common in groundwater modeling applications.

In experiment B, we follow a common practice in groundwater modeling that calibrates recharge while pumping rates are fixed at assumed known values. More specifically, the inaccurate pumping rates $\hat{Q}_A$ and $\hat{Q}_B$ are used as exact input data, while four recharge multipliers $\psi_i, i = 1, \ldots, 4$ are introduced to be calibrated along with the specific yield, hydraulic conductivities and measurement error parameters. The recharge multiplier of one zone is defined as the ratio of the true recharge to the estimated recharge rate in that zone, i.e., $R_i = \psi_i \hat{R}_i, i = 1, \ldots, 4$. The prior marginal distributions of $\psi_i, i = 1, \ldots, 4$ are listed in Table 1. In total 20 parameters are calibrated. The calibrated model is then run for the whole simulation span of 6 years.

As the third calibration strategy, experiment C uses the marginalizing method and assumes that $Q_A \sim N(\mu_Q \hat{Q}_A, (\sigma_Q \hat{Q}_A)^2), Q_B \sim N(\mu_Q \hat{Q}_B, (\sigma_Q \hat{Q}_B)^2)$. For recharge rates, it is assumed that $R_i \sim N(\mu_R \hat{R}_i, (\sigma_R \hat{R}_i)^2), i = 1, \ldots, 4$. Here we introduce a set of hyperparameters $\{\mu_Q, \sigma_Q, \mu_R, \sigma_R\}$ to describe uncertainties in pumping and recharge rates. The marginalizing method then infers the joint posterior distribution of 20 parameters, the priors of which are listed in Table 1. In this study, we use one set of hyperparameters $\{\mu_Q, \sigma_Q\}$ to describe pumping rate uncertainty for both municipal supply and agricultural irrigation wells. Depending upon specific application, sets of hyperparameters can be assigned to different groups of wells so that different levels of bias and uncertainty can be handled.

Table 1. Prior Distributions of Calibrated Parameters and Assumed Distribution of Inputs

Notation	Unit	Distribution
S_y	m	$N(0.18, 0.036^2)$
$\ln K_{rb}$	m/d	$N(0.69, 0.69^2)$
$[\ln K_1, \ldots, \ln K_{12}]^T$	m/d	$N\left([4.1, \ldots, 4.1]^T, \Sigma_K\right)$
$CV_{\Delta Q}$		$Exp(0.05)$
σ_s	m	$Exp(0.05)$
μ_Q		$N(1, 0.2^2)$
σ_Q		$Exp(0.25)$
μ_R		$N(1, 0.25^2)$
σ_R		$Exp(0.25)$
$\psi_i, i = 1, \ldots, 4$		$N(1, 0.25^2)$

In this study, we used DREAM-ZS [*Laloy and Vrugt*, 2012; *Vrugt et al.*, 2009] to sample from the

posterior distributions of parameters. For experiment C, In each iteration of MCMC sampling, the marginal likelihood is calculated using the method described in Appendix B. Sensitivity analysis showed that model simulated drawdown and stream depletion are almost linear with respect to pumping and recharge rates (supporting information Text S3 and Figures S7 and S8). Therefore, the first-order Taylor approximation in equation (B3) is expected to have adequate accuracy.

Based on the $\hat{R}$ statistic [*Gelman and Rubin*, 1992], visual inspection of trace plots and other diagnostics [*Cowles and Carlin*, 1996], the chains converge after ~50,000, ~90,000, and ~300,000 model evaluations in experiment A, B, and C, respectively. Experiment C requires more model evaluations because one calculation of the marginal likelihood requires running the model $6+1=7$ times, where 6 is the total number of pumping and recharge rates. After convergence, 15,000 samples were generated from the joint posterior distribution of parameters after convergence.

Next, the calibrated models are run repeatedly using the posterior parameter samples for the whole simulation period of 6 years to provide forecast in the evaluation period (years 5 and 6). For experiments A and B, biased pumping rates are used during the calibration period. In experiment C, on the other hand, during years 1–4 the marginalizing method accounts for pumping rate uncertainty by marginalizing over the inferred hyperparameters (equation (14)). For all the experiments, true pumping rates are used in the evaluation period (years 5 and 6). This is because in real-world applications groundwater models are often used to provide forecast under one or more presumed water demand scenarios.

The specification of recharge rate during the evaluation period differs in the three experiments. In experiment A, the same biased recharge as in the calibration period is used in the evaluation period. In experiment B, recharge is specified according to the posterior distributions of recharge multipliers $\psi_i, i=1,\ldots,4$. In experiment C, marginalizing is implemented throughout years 1–6 using posterior distribution of hyperparameters (equation (14)). The reason for such an implementation is that in groundwater modeling practice the recharge rate is often estimated based on precipitation and soil characteristics corresponding to different land uses or soil types; these characteristics do not change in time [*Lin et al*., 2009]. Note, however, that the marginalizing method is generic and not limited to the specific implementation of pumping and recharge rates adopted here.

4. Results and Discussion

This section compares the parameter estimates and prediction results from experiments A, B, and C on the synthetic case study described in section 3. Using the posterior samples of prediction obtained in experiment C, we perform variance decomposition analysis (section 2.4) to disaggregate prediction uncertainty into different sources. Results will be discussed in section 4.3.

4.1. Parameter Estimates

The marginal posterior distributions of specific yield, natural logarithm hydraulic conductivity of streambed and pilot points, and measurement error parameters are shown in Figure 5. As described in section 3.4, the three calibration strategies use identical prior distributions for these parameters. It can be seen from Figure 5 that the marginalizing method (experiment C) gives overall the best parameter estimates. In experiment A, the classical Bayesian method gives biased marginal posterior distributions for most of the parameters; the modes of the posterior samples deviate from the true values (marked by the horizontal line in Figure 5) in the virtual reality. This is not surprising: with the overestimation of pumping rate and underestimation of recharge rate, the calibration process tries to "fill in" the missing water, while matching the observed drawdown generated by the virtual reality under true pumping and recharge rates. Therefore, the specific yield S_y is overestimated. Similarly, classical Bayesian calibration (experiment A) gives streambed conductivity ln K_{rb} that is higher than the true value; higher k_{rb} produces more inflow from the stream to the aquifer to make up for the missing water.

In experiment A, the posterior samples of streamflow measurement coefficient of variation $CV_{\Delta Q}$ and drawdown measurement error standard deviation σ_s are mostly higher than the true values used to generate synthetic calibration data (section 3.3). The root-mean-square-error (RMSE) averaged over drawdown used for calibration is 0.0326 m; the RMSE is similar to the maximum a posterior (MAP) of σ_s and is higher than 0.02 m, the value used to generate synthetic calibration data. The RMSE of streamflow gain-and-loss is

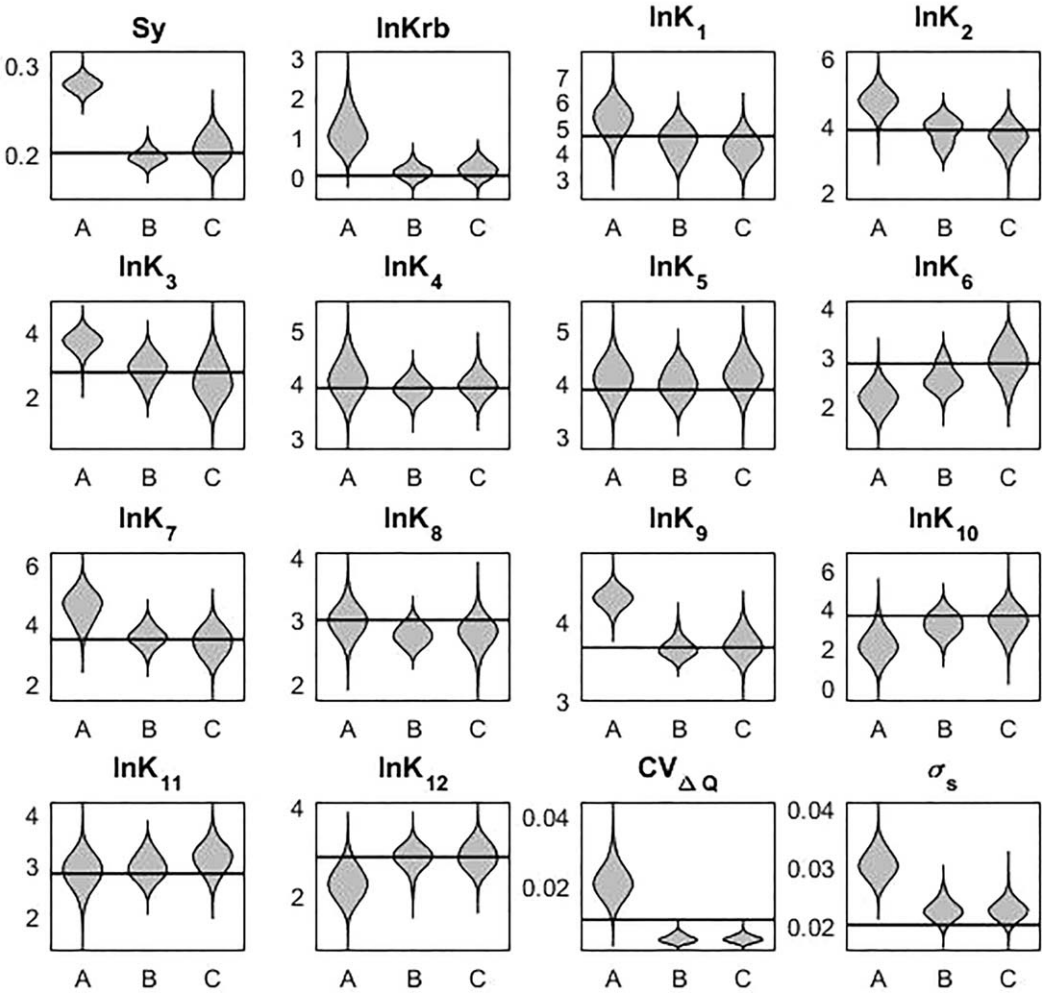

Figure 5. Violin plot of marginal posterior distributions of S_y, $\ln K_{rb}$, $\ln K_1, \ldots, \ln K_{12}$, σ_s, $CV_{\Delta Q}$ of the calibrated models in experiments A, B, and C. The width of shaded violins represents the posterior probability density at different values of parameters as indicated by the vertical axis. The horizontal lines mark the true value of the parameters in the virtual reality.

0.0607 cfs, or 5.65% of temporally averaged streamflow. It can be seen that calibration error cannot be fully attributed to measurement error. In this synthetic case study, model parameters are overly adjusted to compensate for input data errors. However, input data errors cannot be fully absorbed by parameter compensation.

In experiment B, we adopt the conventional groundwater model calibration strategy and estimate recharge multipliers, leading to 20 parameters to be inferred. Similarly as in experiment A, this leads to biased posterior of most of the model parameters. For S_y and $\ln K_{rb}$, the posteriors in experiment B are closer to the true values than those in experiment A, suggesting that estimating recharge multipliers can partially reduce parameter compensation due to input data errors. The posteriors of measurement error parameters $CV_{\Delta Q}$ and σ_s are closer to the true values, indicating improved goodness-of-fit of model simulation to calibration data. The calibration error RMSE of drawdown and stream gain-and-loss are 0.025 m and 0.0191 cfs (1.8% of mean streamflow), respectively. Both are smaller than in experiment A and closer to the true values.

To further investigate the performance of the calibration strategy in experiment B, Figure 6 shows the posterior distributions of recharge multipliers. Zones 2 and 3 have the largest area and highest total amount of recharge (Figure 4), and thus have more impacts on model parameters and prediction. Posterior modes of ψ_2 and ψ_4 are smaller than the true values, while the posterior mode of ψ_3 is greater than the true value $R_{3,0}/\hat{R}_3$. The calibration partially corrects the underestimation for zones 2 and 4, while overestimating the recharge multiplier for zone 3.

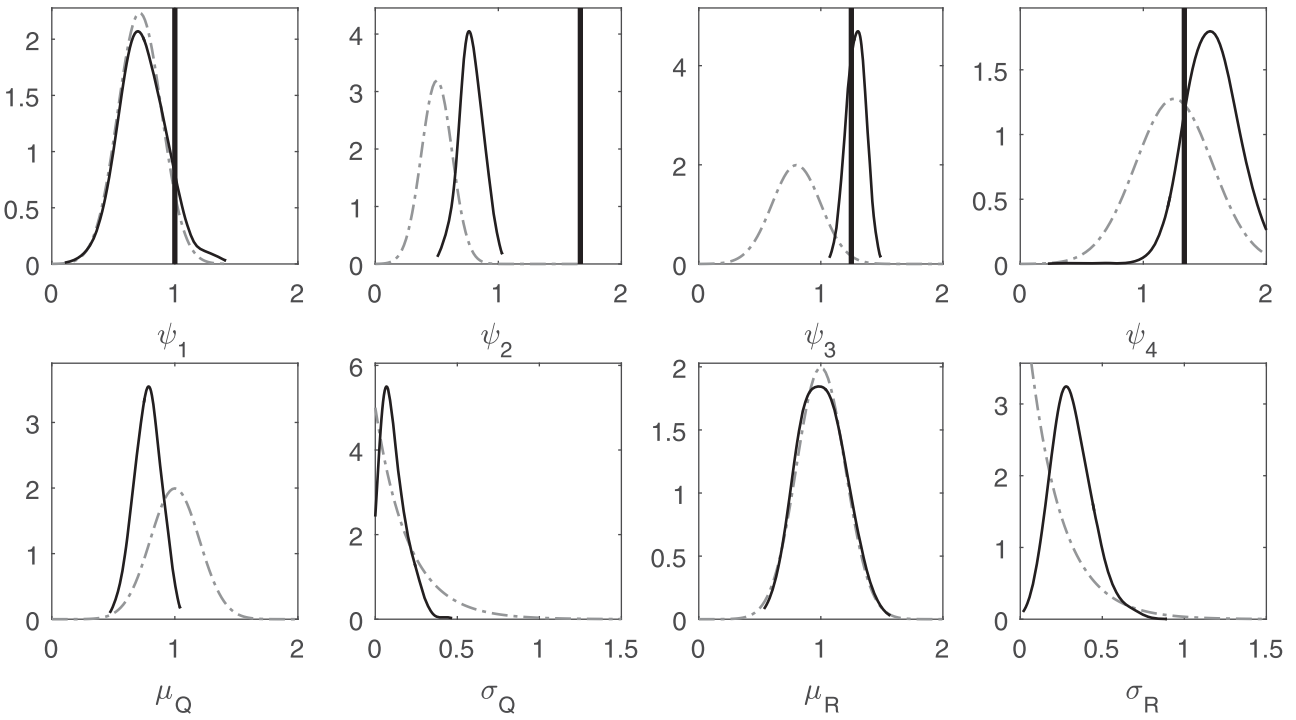

Figure 6. (top) Prior (grey, dashed) and marginal posterior (black, solid) distributions of recharge multipliers $\psi_1, \ldots, \psi_4$ as inferred in experiments B. The vertical lines mark the true value. (bottom) Prior and marginal posterior distributions of hyperparameters $\mu_Q, \sigma_Q, \mu_R, \sigma_R$ as inferred in experiment C.

Compared to experiments A and B, the marginalizing method in experiment C yields parameter estimates that are the most consistent with true values. By adjusting the hyperparameters μ_Q and μ_R which control the overall level of bias in pumping and recharge rates, the marginalizing method alleviates the problem of the classical Bayesian method overly adjusting other parameters (particularly for the specific yield S_y) to compensate for input data uncertainty. Figure 5 also shows that the posterior samples of the measurement error parameters yielded by the marginalizing method are smaller than those in experiments A and are closer to the true values. The calibration error RMSE of drawdown and stream gain-and-loss ΔQ are 0.0922 m and 0.0143 cfs (1.33% of mean streamflow), respectively. The RMSE of ΔQ is higher than that in experiment B, and the RMSE of drawdown is higher than in both experiments A and B. This is not surprising, because the marginalizing Bayesian method uses the marginal likelihood function (equation (12)). The marginal likelihood accounts for variance propagated from uncertain input data, and hence allows larger calibration error. In experiments A and B, on the other hand, the likelihood function only contains measurement error, and the calibration process may overly adjust model parameters to force smaller calibration error up to the level of measurement error.

Figure 6 shows the marginal distributions of hyperparameters $\mu_Q, \sigma_Q, \mu_R, \sigma_R$. For μ_Q, the MAP is smaller than 1, which is reasonable because the pumping rates are overestimated. When the degree of bias of all pumping rates are the same, i.e., $Q_A/\hat{Q}_A = Q_B/\hat{Q}_B = \lambda$, the posterior of μ_Q should ideally be λ. Since the pumping rates at two wells are biased differently ($\hat{Q}_A = 1.2Q_{A,0}, \hat{Q}_B = 1.4Q_{B,0}$), there is no true value for μ_Q. Meanwhile, σ_Q controls the variability of the latent variables ψ_A, ψ_B around μ_Q as defined in equations (9) and (10). The posterior pdf of σ_Q shrinks compared to the prior pdf, suggesting that the calibration provided some information to reduce input uncertainty. On the other hand, the marginal posterior distribution of μ_R are not deviating significantly from the prior. Notably from Figure 5, the marginalizing method posterior of S_y is more spread out as a result of interaction among S_y, μ_Q, and μ_R. The lack of identifiability could be intrinsic to this problem and may not necessarily be removed by increasing the number of observations. Under this and similar conditions, it is critical to specify a prior that is as informative as possible [*Huard and Mailhot*, 2008; *Renard et al.*, 2010, 2011]. An alternative approach is to *a priori* fix $\mu_Q = \mu_R = 1$ and specify a reasonable value for σ_Q and σ_R based on prior knowledge about the level of input data uncertainty. In the calibration

process, one still implements the marginalizing step to infer model parameters while fixing the hyperparameters. This approach is suitable when there is no evidence for systematic bias in input data, and that the inputs are estimated with low to medium levels of uncertainty.

4.2. Prediction

In the prediction (i.e., evaluation) phase, the model is run repeatedly using the posterior samples for the whole simulation period of 6 years as described in section 3.4. The posterior mean and 95% prediction intervals of drawdown at three locations (Figure 3) and stream gain-and-loss are shown in Figure 7. In experiment A, the darker shades indicate prediction uncertainty contributed by postcalibration uncertainty of parameters $S_y, K_{rb}, K_1, \ldots, K_{12}$ and the intervals are obtained by running the model repeatedly using the posterior samples of parameters. In experiment B, the darker shades correspond to prediction uncertainty from the posterior uncertainty of $S_y, K_{rb}, K_1, \ldots, K_{12}$ and recharge multipliers $\psi_1, \ldots, \psi_4$. In both

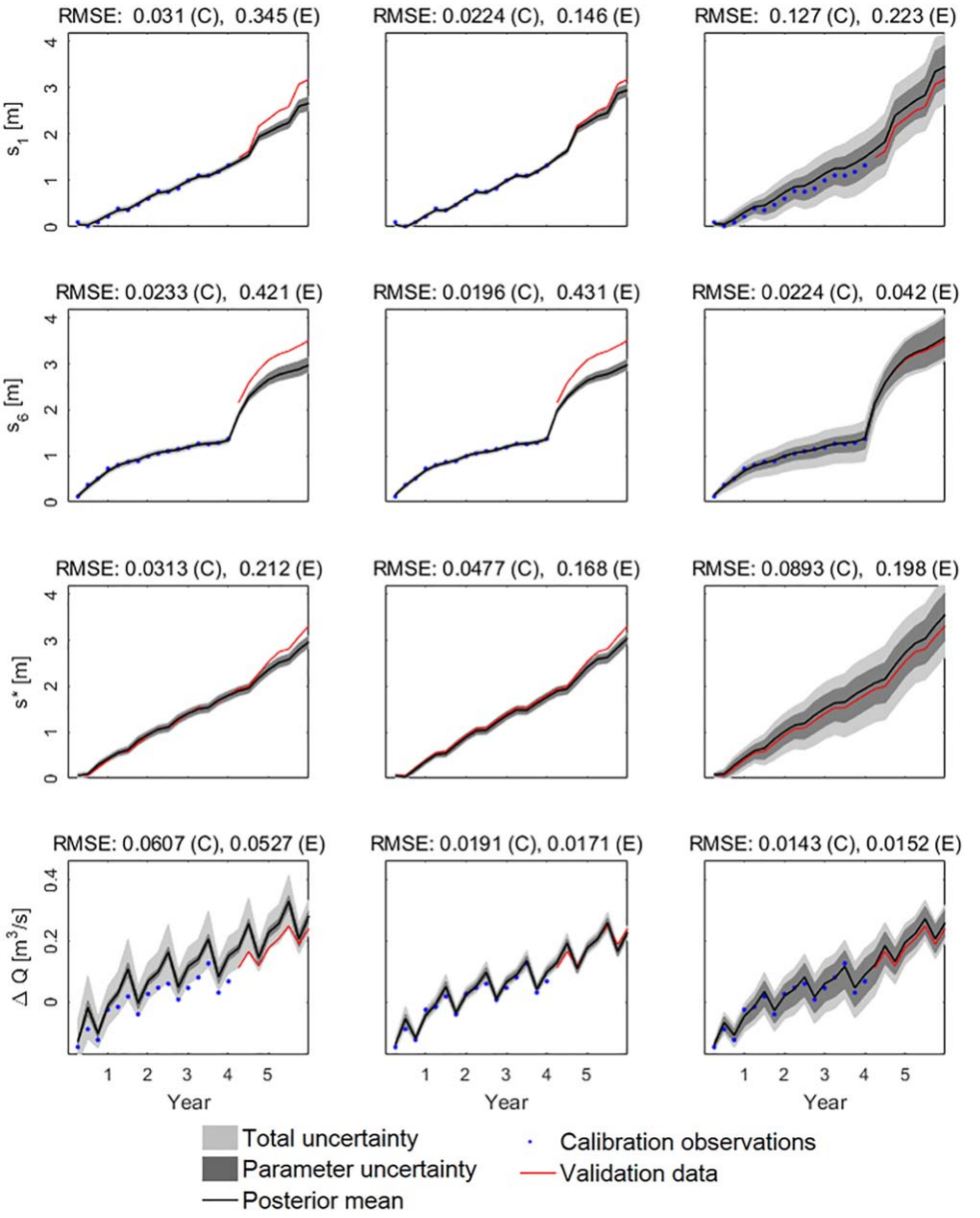

Figure 7. Simulation results of drawdown s_1, s_6, s^* and stream gain-and-loss ΔQ using the calibrated models in experiments A (left), B (middle), and C (right). Locations of drawdown observations are shown in Figure 3. Shades show 95% prediction intervals. Also shown are the root-mean-square-error (RMSE) statistics calculated for the calibration (C) period and the evaluation (E) period, respectively.

experiments A and B, the lighter shades indicate total prediction uncertainty including drawdown and streamflow measurement uncertainty.

In experiment C, the darker shades correspond to prediction uncertainty contributed by postcalibration uncertainty of model parameters and input uncertainty hyperparameters $\mu_Q, \sigma_Q, \mu_R, \sigma_R$. The darker shades are obtained by sorting and finding the quantiles of $f(\mu_{Q,i}\hat{Q}_A, \mu_{Q,i}\hat{Q}_B, \mu_{R,i}\hat{R}_1, \mu_{R,i}\hat{R}_2, \mu_{R,i}\hat{R}_3, \mu_{R,i}\hat{R}_4, \boldsymbol{\theta}_i)$, where f denotes the model. The darker shades are the combined effects of posterior parameter uncertainty and systematic input biases. The lighter shades show the total uncertainty contributed from parameter, input data, and measurement uncertainties. They are calculated based on samples $y_i^*, i=1,\ldots N$, where y_i^* is drawn from $y^*|\boldsymbol{\theta}_i, \mu_{Q,i}, \sigma_{Q,i}, \mu_{R,i}, \sigma_{R,i}$. The lighter shades include the effects of random input errors, or deviation of input data from their mean value. The procedures of deriving the prediction intervals are described in detail in Appendix C.

For the model calibrated in experiment A, despite relatively small calibration error, the model predictions have significant bias and that the 95% prediction intervals do not encompass the evaluation data. This is consistent with the observation in section 4.1 that parameters are overly adjusted to compensate for input data errors. The RMSE is higher than the magnitude of measurement error. Similarly, the model calibrated in experiment B yields biased and overconfident prediction at s_6. In contrast, the marginalizing method (experiment C) reduces the RMSE and gives overall less biased posterior mean during the evaluation period. Figure 8 further compares the prediction error of spatially varying drawdown among the three models calibrated in experiments A, B, and C. The models calibrated in experiments A and B underestimate the extent of the depression cone around pumping wells A and C. In contrast, the model calibrated by the marginalizing method yields drawdown prediction that is more consistent with the virtual reality. For the drawdown prediction in June, 6th year, the mean error of the marginalizing method is 81% smaller than the RMSE in experiment A, and 78% smaller than the RMSE in experiment B.

As for stream gain-and-loss ΔQ, Figure 7 shows that the classical Bayesian calibrated model A yields biased prediction. This is expected because experiment A does not correct bias in input data and resulted in biased

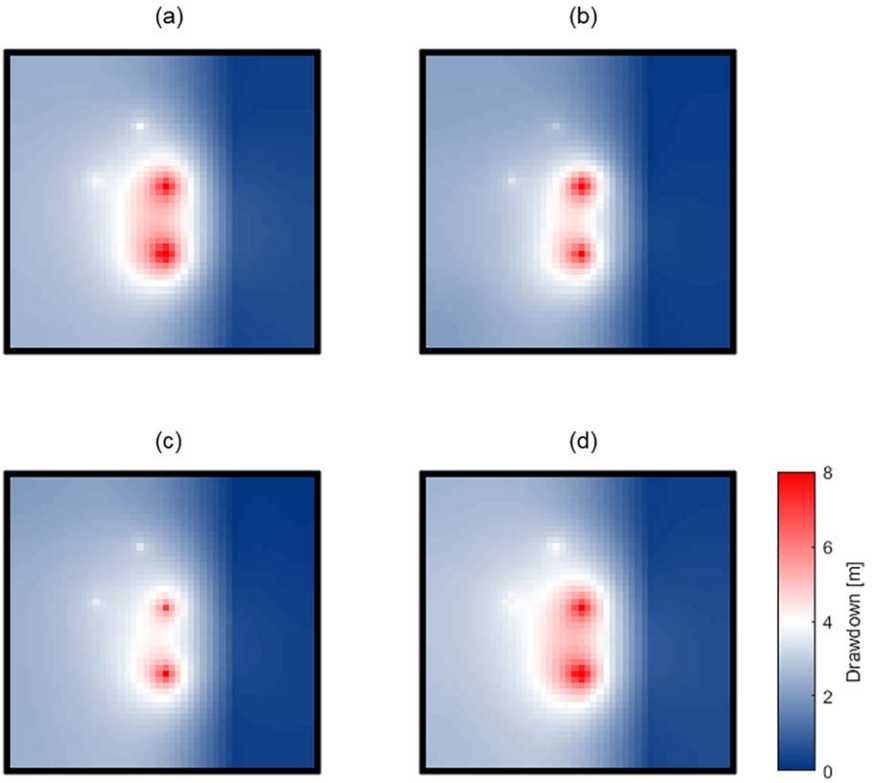

Figure 8. Drawdown in June, (a) 6th year as simulated by the virtual reality and the models calibrated in experiments (b) A, (c) B, and (d) C.

estimates of model parameters to which the stream gain-and-loss is sensitive. The model calibrated in experiment B gives smaller bias with narrow uncertainty bounds. The main reason is that experiment B obtains a reasonable estimate of the recharge multiplier of zone 2 (Figure 6). The stream runs within zone 2 (Figure 4), and therefore is sensitive to recharge in that zone. However, the streambed hydraulic conductivity parameter $\ln K_{rb}$ estimated in experiment B is biased (Figure 5), and the recharge multiplier ψ_3 is overly adjusted (Figure 6). Such parameter compensation may have deleterious effects on prediction accuracy under a different future scenario. Finally, the marginalizing method (C) yields the smallest RMSE, mainly because the overall reasonable estimates of model parameters and input data hyperparameters.

Figure 7 also shows that the 95% prediction intervals given by calibrated model B are narrower than those derived from the classical Bayesian method. The reason is that the augmentation method achieves better fit to calibration data with the increased freedom from four additional parameters (recharge multipliers). The prediction intervals appear overconfident during the evaluation period for drawdowns. It is noteworthy, however, that in experiment B recharge multipliers are adjusted while pumping rates are fixed at estimated values, following the common practice of groundwater modeling.

On the other hand, the marginalizing method gives wide prediction intervals mainly due to relatively large posterior uncertainty of S_y, μ_Q, μ_R. Another reason is that the inferred hyperparameter σ_R, i.e., the standard deviation of recharge multipliers, is high. As mentioned earlier in this section and in Appendix C, the lighter shades in Figure 7 are derived by marginalizing using posterior samples of σ_R and σ_Q and hence include the effects of random input errors. Higher values of σ_R mean higher recharge rate uncertainty, which propagates to prediction uncertainty.

4.3. Variance Decomposition

Lastly, variance decomposition was carried out to further analyze the contribution to prediction uncertainty from input data, parameter, and measurement uncertainty. The fractions are calculated for drawdown $s_1, \ldots, s_7, s^*$ and stream gain-and-loss ΔQ at the end of the evaluation period and are plotted in Figure 9. Theoretically, the three fractions should sum to one for each prediction; as can be seen the sum is close to one, suggesting that the error resulting from the binning approximation is acceptable. More sophisticated methods [e.g., *Saltelli et al.*, 2010; *Dai and Ye*, 2015] can also be applied to estimate the variance components. Further investigation on the estimation error of the variance components is beyond the scope of this study.

It can be seen from Figure 9 that input data uncertainty is the major source of uncertainty for stream gain-and-loss and most drawdown predictions. In addition, the fraction varies for drawdown at different locations. For drawdown s_6 and s_7, input data uncertainty fraction is still the highest, yet parameter uncertainty explains about a third of the total prediction variance. A possible reason is that these two drawdown predictions are very sensitive to the parameter K_{rb}, and the posterior uncertainty of K_{rb} propagates to ΔQ prediction.

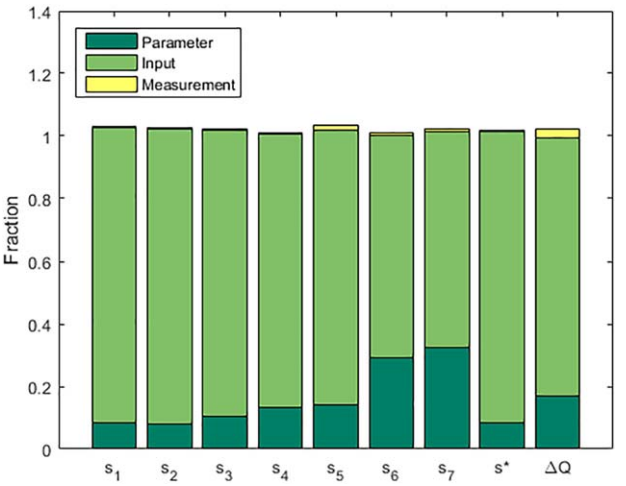

Figure 9. Fraction of prediction uncertainty contributed by parameter, input, and measurement uncertainty for drawdown $s_1, \ldots, s_7, s^*$ (locations shown in Figure 3) and stream depletion ΔQ at the end of the evaluation period.

The above variance decomposition analysis is conditioned on the assumptions made in the specification of input error model. The choice of the error model and the prior distribution of hyperparameters are inevitably prone to subjectivity. It should be emphasized that every effort should be made to specify a meaningful error model, preferably supported by data from applications that bare similarity with the specific problem of concern.

5. Conclusions

We demonstrated the Bayesian marginalizing method to account for input uncertainty through an example of Thiem equation and a synthetic case study of

surface-ground water interaction under changing pumping and land use conditions. We found in the numerical experiments that explicit treatment of uncertainty in input data (groundwater pumping and recharge rates) has substantial impact on the posterior distribution of groundwater model parameters. Using statistical models to explicitly account for input data uncertainty reduces predictive bias caused by parameter compensation. Compared to conventional calibration methods in experiments A and B, the marginalizing approach yields overall more accurate predictions of drawdown and stream gain-and-loss. The marginalizing method results indicate that input uncertainty increases parametric and prediction uncertainty. Based on the inference results of the marginalizing Bayesian approach, we performed variance decomposition to investigate prediction uncertainty from model parameters, input data and measurement error. We found in the synthetic case study that the input uncertainty is dominant among various sources of uncertainty. Such insights could inform future model improvement and data collection efforts on how to best direct resources toward reducing prediction uncertainty. In general, we recommend the marginalizing approach to be used for situations in which (1) substantial knowledge is available to specify reasonably informative priors for input uncertainty hyperparameters, and (2) joint inference of inputs and model parameter do not work due to identifiability issues.

The synthetic case study described in section 3 assumes that the model structure is known perfectly, which is often not tenable in real-world applications. Insupporting information Text S1, we tested the marginalizing approach in another synthetic case study, in which the model is subject to model structural error arising from an oversimplified hydraulic conductivity field, aquifer bottom elevation, and boundary geometry. We found that when model structural error is present but not accounted for, the Bayesian method in experiment B (i.e., calibrating recharge rates in each zone) tends to overly adjust the recharge rates in order to compensate for model structural error. This led to predictive bias. On the contrary, the marginalizing method gives overall more reasonable estimates of parameters and more accurate predictions than the other two Bayesian methods, which is consistent with the observations in the case study in section 3 without model structural error. Compared to the no structural error scenario, the variance decomposition showed a slightly higher predictive variance contributed by model parameters. A possible reason is that structural error resulted in larger calibration error, leading to higher posterior parametric uncertainty. Nevertheless, variance decomposition still identified input data error as the primary source of prediction variance. Overall, the results in supporting information suggest that the marginalizing method is applicable in the presence of model structural error considered in our case study.

While an in-depth investigation of handling both input data and model structural errors is beyond the scope of this study, the marginalizing Bayesian method does not preclude explicit treatment of model structural error. For example, one can specify an error model, and the hyperparameters of the error model can be estimated together with model parameters and input data error hyperparameters during Bayesian calibration. The Bayesian inference results can then be used to perform variance decomposition to separate all uncertainty sources. However, such joint inference often induces identifiability issues due to potential interactions between input data and model structural errors [*Renard et al*., 2010]. Supplying informative priors could help to identify the input data and model structural errors, however, at the risk of introducing subjective bias if the priors are not accurate. Further investigation is needed to test the applicability of the marginalizing method in real-world calibration problems with unknown model structural error, as well as for cases requiring the simultaneous identifiability of input data and model structural errors.

As described in Appendix B, because of the marginalizing step, the number of model evaluations per MCMC iteration is linearly dependent on the number of uncertain inputs. For modeling applications with thousands of pumping wells and many recharge zones, the high-dimensional integration involved in the evaluation of the marginal likelihood can be computationally intractable. One way to reduce the computational expense is to group wells (or recharge zones) into clusters. When calculating sensitivity, all pumping (or recharge) rates belonging to the same cluster are adjusted simultaneously. In this way, the integration dimension is reduced to the number of clusters. Another solution is to reuse the sensitivity matrix for multiple MCMC proposals and update only when the proposal moves to another region. The underlying hypothesis is that, the sensitivity with respect to inputs does not change significantly in a local neighborhood of the parameter space.

Bayesian inference often requires tens to hundreds of thousands of model evaluations. The computational cost can be reduced in various ways, e.g., using a fast surrogate of the time consuming model [*Asher et al*.,

2015] and implementing multiple-chain, multiple-try, and multiple-stage sampling algorithms [*Laloy et al*., 2013; *Marzouk and Najm*, 2009; *Xie et al*., 2009]. Depending on a specific problem, various diagnostics can be used to help decide whether to implement a fully Bayesian calibration at the expense of computational burden [*Hill et al*., 2015]. The results in the synthetic case study suggest that the marginalizing approach could be used when evidence supports the presence of input uncertainty, yet it is not feasible to collect input data of higher precision and accuracy. The presented framework can be used for other environmental modeling applications such as integrated hydrologic modeling. Follow-up studies will further investigate the potential of the proposed method particularly for real-world modeling practice.

Appendix A: Derivation of the Posterior in the Thiem Equation Example

Here we derive the parameter posterior distribution (equation (6)) in the Thiem equation example given by the classical Bayesian method. For illustration purpose, we specify a uniform prior on the interval $[a, b]$ for the parameter θ, i.e., $p(\theta)=1/(b-a)$. The interval $[a, b]$ is wide to reflect vague prior information. Therefore, the prior can be approximately written as $p(\theta) \propto 1$; we dropped scaling constant $b - a$ as it does not affect the shape of the prior and posterior distributions. Applying the Bayes rule with the likelihood $L(\theta|s_0)$ given by equation (5), we have

$$p(\theta|s_0) \propto \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[-\frac{\left(s_0 - k\sum_i \hat{Q}_i\theta\right)^2}{2\sigma_\epsilon^2}\right]$$
$$= \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[\frac{\left(\theta - \frac{s_0}{k\sum_i \hat{Q}_i}\right)^2}{\frac{2\sigma_\epsilon^2}{k^2\left(\sum_i \hat{Q}_i\right)^2}}\right]. \tag{A1}$$

Since $s_0 = k\sum_i Q_i\theta_0$,

$$p(\theta|s_0) \propto \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[\frac{\left(\theta - \frac{k\sum_i Q_i\theta_0}{k\sum_i \hat{Q}_i}\right)^2}{\frac{2\sigma_\epsilon^2}{k^2\left(\sum_i \hat{Q}_i\right)^2}}\right]$$
$$= \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[\frac{\left(\theta - \frac{\sum_i Q_i}{\sum_i \hat{Q}_i}\theta_0\right)^2}{\frac{2\sigma_\epsilon^2}{k^2\left(\sum_i \hat{Q}_i\right)^2}}\right] \tag{A2}$$

which is Gaussian distribution with mean $\frac{\sum_i Q_i}{\sum_i \hat{Q}_i}\theta_0$ and variance $\frac{\sigma_\epsilon^2}{k^2\left(\sum_i \hat{Q}_i\right)^2}$.

Appendix B: Calculation of Marginal Likelihood

In this appendix, we describe the implementation of the marginalizing step. Let $\mathbf{x}=\{x_i\}, i=1, 2, \ldots p$ denote the true yet unknown inputs and $\hat{\mathbf{x}}=\{\hat{x}_i\}, i=1, 2, \ldots$ denote the corresponding estimation. Based on equations (9) and (10), we have

$$x_i \overset{iid}{\sim} N(\mu_x\hat{x}_i, \sigma_x^2\hat{x}_i^2), i=1, 2, \ldots, p. \tag{B1}$$

Next, we approximate the model simulation outputs $f(\mathbf{x}, \boldsymbol{\theta})=\{f_i(\mathbf{x}, \boldsymbol{\theta})\}, i=1, \ldots, n$ with first-order Taylor series with respect to $\mathbf{x}$:

$$f_i(\mathbf{x},\boldsymbol{\theta}) \approx f_i(\mu_x\hat{\mathbf{x}},\boldsymbol{\theta}) + \sum_j \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{x}=\mu_x\hat{\mathbf{x}}} (x_j - \mu_x\hat{x}_j).$$
$$= \sum_j \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{x}=\mu_x\hat{\mathbf{x}}} x_j + f(\mu_x\hat{\mathbf{x}},\boldsymbol{\theta}) - \sum_i \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{x}=\mu_x\hat{\mathbf{x}}} \mu_x\hat{x}_j. \tag{B2}$$

This approximation yields satisfactory accuracy when the degree of nonlinearity with respect to input $\mathbf{x}$ is low. Next, let K denote the sensitivity matrix, i.e., $K_{i,j} = \frac{\partial f_i}{\partial x_j}|_{\mathbf{x}=\mu_x\hat{\mathbf{x}}}$, and write the above equation compactly as

$$f(\mathbf{x},\boldsymbol{\theta}) \approx K\mathbf{x} + f_0 - \mu_x K\hat{\mathbf{x}}. \tag{B3}$$

Applying the above equations to equation (12), we have

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{y}) &= \int\cdots\int \frac{1}{\sqrt{(2\pi)^n|\sigma_\epsilon^2 I_n|}} \exp\{-[\mathbf{y}-f(\mathbf{x},\boldsymbol{\theta})]^T[\mathbf{y}-f(\mathbf{x},\boldsymbol{\theta})]/2\sigma_\epsilon^2\} \\ &\cdot\prod_i \frac{1}{\sqrt{2\pi\sigma_x^2\hat{x}_i^2}} \exp\left[-\frac{(x_i-\hat{x}_i)^2}{2\sigma_x^2\hat{x}_i^2}\right] dx_1\cdots dx_p \\ &\approx \int\cdots\int \frac{1}{\sqrt{(2\pi)^n|\sigma_\epsilon^2 I_n|}} \exp\{-\frac{[(\mathbf{y}-f_0+\mu_x K\hat{\mathbf{x}})-K\mathbf{x}]^T[(\mathbf{y}-f_0+\mu_x K\hat{\mathbf{x}})-K\mathbf{x}]}{2\sigma_\epsilon^2}\} \\ &\cdot\prod_i \frac{1}{\sqrt{2\pi\sigma_x^2\hat{x}_i^2}} \exp\left[-\frac{(x_i-\hat{x}_i)^2}{2\sigma_x^2\hat{x}_i^2}\right] dx_1\cdots dx_p. \end{aligned} \tag{B4}$$

Lastly, applying the convolution theorem recursively and inverse Fourier transform [*Bromiley*, 2014],

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{y}) \approx\ & \frac{1}{\sqrt{(2\pi)^n \parallel \sigma_x^2\sum_{i=1}^p \hat{x}_i^2\mathbf{k}_i\mathbf{k}_i^T + \sigma_\epsilon^2 I_n \parallel}} \\ & \cdot\exp\{-[\mathbf{y}-f(\mu_x\hat{\mathbf{x}},\boldsymbol{\theta})]^T\left(\sigma_x^2\sum_{i=1}^p \hat{x}_i^2\mathbf{k}_i\mathbf{k}_i^T + \sigma_\epsilon^2 I_n\right)^{-1}[\mathbf{y}-f(\mu_x\hat{\mathbf{x}},\boldsymbol{\theta})]\}, \end{aligned} \tag{B5}$$

which is the probability density function of a Gaussian distribution with mean $f(\mu_x\hat{\mathbf{x}},\boldsymbol{\theta})$ and variance $\sigma_x^2\sum_{i=1}^p \hat{x}_i^2\mathbf{k}_i\mathbf{k}_i^T + \sigma_\epsilon^2 I_n$. Here $\mathbf{k}_i$ denotes the ith column of the sensitivity matrix K.

Appendix C: Inference of Predictions

This appendix describes in detail the procedures of generating posterior samples of prediction based on equation (14). For illustrative purposes, we introduce the procedures using the case study described in section 3.

In this case study, we consider six uncertain inputs, denoted as $\mathbf{x} = [Q_A, Q_B, R_1, \ldots, R_4]^T$. As described in section 3.4, we use hyperparameters $\boldsymbol{\phi} = \{\mu_Q, \sigma_Q, \mu_R, \sigma_R\}$ to describe uncertainties in these input data. Calibration gives N posterior samples of model parameters $\boldsymbol{\theta}$, hyperparameters $\boldsymbol{\phi}$, and likelihood parameters $\sigma_s, CV_{\Delta Q}$. Next, using $\{\boldsymbol{\theta}_i, \boldsymbol{\phi}_i, \sigma_{s,i}, CV_{\Delta Q,i}\}, i = 1,\ldots,N$, we derive the posterior samples of predictions $\mathbf{y}^*$. Similarly to equation (B5), we have

$$\mathbf{y}^*|\boldsymbol{\theta}_i, \boldsymbol{\phi}_i, \sigma_{s,i}, \sigma_{\Delta Q,i} \sim N\left(f(\boldsymbol{\mu}_x^T\hat{\mathbf{x}}, \boldsymbol{\theta}_i), \Sigma + \Sigma_{\epsilon,i}\right), \tag{C1}$$

where

$$\begin{aligned} \Sigma &= \sum_{i=1}^p \sigma_{x,i}^2\hat{x}_i^2\mathbf{k}_i^*\mathbf{k}_i^{*T} \\ &= \sigma_Q^2(\hat{Q}_A^2\mathbf{k}_A^*\mathbf{k}_{Q,A}^{*T} + \hat{Q}_B^2\mathbf{k}_{Q,B}^*\mathbf{k}_{Q,B}^{*T}) + \sigma_R^2\sum_{i=1}^4 \hat{R}_i^2\mathbf{k}_{R,i}^*\mathbf{k}_{R,i}^{*T}. \end{aligned} \tag{C2}$$

Here f denotes the model, $\boldsymbol{\mu}_{x,i} = [\mu_{Q,i}, \mu_{Q,i}, \mu_{R,i}, \mu_{R,i}, \mu_{R,i}, \mu_{R,i}]^T$, $\hat{\mathbf{x}} = [\hat{Q}_A, \hat{Q}_B, \hat{R}_1, \ldots, \hat{R}_4]^T$. The measurement error covariance matrix $\Sigma_{\epsilon,i}$ is usually diagonal. In the synthetic case study, the diagonal entries are $\sigma_{s,i}^2$ and

Algorithm C1: Draw posterior realizations of predictions

1: $\Sigma \leftarrow 0$
2: **for** $i = 1$ **to** N **do**
3: $\quad \hat{\mathbf{y}}_{0,i} \leftarrow f(\boldsymbol{\mu}_{x,i}^T \hat{\mathbf{x}}, \boldsymbol{\theta}_i)$ ▷ Run the model using a posterior sample of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$
4: $\quad \hat{\mathbf{y}}_1 \leftarrow f(\boldsymbol{\mu}_x^T \hat{\mathbf{x}} + \Delta\mathbf{x}, \boldsymbol{\theta}_i)$ ▷ Run the model with different input data
5: $\quad$ **for** $j = 1$ **to** p **do**
6: $\quad\quad \mathbf{k}_j \leftarrow (\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_{0,i})/\Delta\mathbf{x}$ ▷ Approximately calculate the derivatives
7: $\quad\quad \Sigma \leftarrow \Sigma + \sigma_{x,j}^2 \hat{x}_j^2 \mathbf{k}_j^* \mathbf{k}_j^{*T}$ ▷ Calculate the prediction covariance matrix
8: $\quad$ **end for**
9: $\quad \mathbf{y}_{0,i} \leftarrow$ sample from $N(\hat{\mathbf{y}}_0, \Sigma)$ ▷ Draw a noise-free sample
10: $\quad \epsilon_i \leftarrow$ sample from $N(0, \Sigma_{\epsilon,i})$ ▷ Measurement error
11: $\quad \mathbf{y}_i \leftarrow \mathbf{y}_{0,i} + \epsilon_i$ ▷ Generate measurement error-contaminated prediction
12: **end for**
13: **return** $\{\hat{\mathbf{y}}_{0,i}, \mathbf{y}_{0,i}, \mathbf{y}_i\}, i = 1, \ldots, N$

$\sigma_{\Delta Q}^2$ corresponding to drawdown and stream gain-and-loss observations, respectively; $\sigma_{\Delta Q,i}^2$ is calculated using $CV_{\Delta Q,i}$ as explained in section 3.4. The vector $\mathbf{k}_i^*$ indicates the sensitivity of predictions $\mathbf{y}^*$ with respect to input x_i; accordingly, $\mathbf{k}_{Q,A}^*$ means the derivative of $\mathbf{y}^*$ to pumping rate at well A, and $\mathbf{k}_{R,i}^*$ denotes the derivative of $\mathbf{y}^*$ to the recharge rate in zone i. The derivatives can be calculated numerically using methods such as forward finite difference. In general, for a given set of $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, evaluating equations (C1) and (C2) requires running the model $p + 1$ times, where p is the number of uncertain inputs. The pseudocode to implement the procedures is given in Algorithm C1. The notation $\leftarrow$ means assigning the value of right-hand side to the left-hand side.

Acknowledgments
This work was 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-SC0002687. Supporting data are available from the authors upon request.

References

Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, *43*, W01403, doi:10.1029/2005WR004745.
Ajami, N. K., Q. Duan, and S. Sorooshian (2009), Reply to comment by B. Renard et al. on "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction", *Water Resour. Res.*, *45*, W03604, doi:10.1029/2008WR007215.
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.
Bosshard, T., M. Carambia, K. Goergen, S. Kotlarski, P. Krahe, M. Zappa, and C. Schär (2013), Quantifying uncertainty sources in an ensemble of hydrological climate-impact projections, *Water Resour. Res.*, *49*, 1523–1536, doi:10.1029/2011WR011533.
Bromiley, P. A. (2014), Products and convolutions of Gaussian probability density functions, Technical report, Tina-Vision, Manchester, U. K.
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.
Dai, H., and M. Ye (2015), Variance-based global sensitivity analysis for multiple scenarios and models with implementation using sparse grid collocation, *J. Hydrol.*, *528*, 286–300.
Del Giudice, D., M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann (2013), Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, *Hydrol. Earth Syst. Sci.*, *17*(10), 4209–4225.
Del Giudice, D., C. Albert, J. Rieckermann, and P. Reichert (2016), Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation, *Water Resour. Res.*, *52*, 3162–3186, doi:10.1002/2015WR017871.
Demissie, Y., A. Valocchi, X. Cai, N. Brozovic, G. Senay, and M. Gebremichael (2014), Parameter estimation for groundwater models under uncertain irrigation data, *Ground Water*, *53*, 614–625.
Doherty, J. (2003), Ground water model calibration using pilot points and regularization, *Ground Water*, *41*(2), 170–177.
Famiglietti, J. (2014), The global groundwater crisis, *Nat. Clim. Change*, *4*(11), 945–948.
Fienen, M., R. Hunt, D. Krabbenhoft, and T. Clemo (2009), Obtaining parsimonious hydraulic conductivity fields using head and transport observations: A Bayesian geostatistical parameter estimation approach, *Water Resour. Res.*, *45*, W08405, doi:10.1029/2008WR007431.
Finsterle, S., and M. B. Kowalsky (2011), A truncated Levenberg-Marquardt algorithm for the calibration of highly parameterized nonlinear models, *Comput. Geosci.*, *37*(6), 731–738.
Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*(4), 457–472.
Gorelick, S. M., and C. Zheng (2015), Global change and the groundwater management challenge, *Water Resour. Res.*, *51*, 3031–3051, doi:10.1002/2014WR016825.

Harbaugh, A. W., E. R. Banta, M. C. Hill, and M. G. McDonald (2000), MODFLOW-2000, the US geological survey modular ground-water model: User guide to modularization concepts and the ground-water flow process, *Tech. rep., U.S. Geol. Surv. Open-File Rep., 00-92*, U.S. Geological Survey Reston, Va.

Healy, R. W. (2010), *Estimating Groundwater Recharge*, Cambridge Univ. Press, Cambridge, U. K.

Held, I. M., and B. J. Soden (2006), Robust responses of the hydrological cycle to global warming, *J. Clim.*, *19*(21), 5686–5699.

Hill, M., and C. Tiedeman (2007), *Effective Calibration of Groundwater Models, with Analysis of Data, Sensitivities, Predictions, and Uncertainty*, John Wiley, New York.

Hill, M. C., D. Kavetski, M. Clark, M. Ye, M. Arabi, D. Lu, L. Foglia, and S. Mehl (2015), Practical use of computationally frugal model analysis methods, *Ground Water*, *54*, 159–170, doi:10.1111/gwat.12330.

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, *Tech. rep., U.S. Geol. Sci. Invest. Rep., 2007-5044*, U.S. Geological Survey Reston, Va.

Huard, D., and A. Mailhot (2008), Calibration of hydrological model GR2M using Bayesian uncertainty analysis, *Water Resour. Res.*, *44*, W02424, doi:10.1029/2007WR005949.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resour. Res.*, *42*, W03408, doi:10.1029/2005WR004376.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

Laloy, E., and J. A. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try 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.

Lee, K. K., and J. C. Risley (2002), Estimates of ground-water recharge, base flow, and stream reach gains and losses in the Willamette river basin, Oregon, Technical report, US Dep. of the Inter., U.S. Geol. Surv., 01-4215, Portland, Oreg.

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.

Lin, Y.-F., J. Wang, and A. J. Valocchi (2009), PRO-GRADE: GIS toolkits for ground water recharge and discharge estimation, *Ground Water*, *47*(1), 122–128.

Lu, D., M. Ye, and M. C. Hill (2012), Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification, *Water Resour. Res.*, *48*, W09521, doi:10.1029/2011WR011289.

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. M., and H. N. Najm (2009), Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems, *J. Comput. Phys.*, *228*(6), 1862–1902.

McKusick, V. (2003), Final report for the special master with certificate of adoption of RRCA groundwater model, *Tech. Rep.* 126, State of Kansas v. State of Nebraska and State of Colorado, in the Supreme Court of the United States.

McMillan, H., B. Jackson, M. Clark, D. Kavetski, and R. Woods (2011), Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, *J. Hydrol.*, *400*(1), 83–94.

Mendoza, P. A., M. P. Clark, N. Mizukami, E. D. Gutmann, J. R. Arnold, L. D. Brekke, and B. Rajagopalan (2015), How do hydrologic modeling decisions affect the portrayal of climate change impacts?, *Hydrol. Processes*, *30*, 1071–1095, doi:10.1002/hyp.10684.

Mockler, E., K. Chun, G. Sapriza-Azuri, M. Bruen, and H. Wheater (2016), Assessing the relative importance of parameter and forcing uncertainty and their interactions in conceptual hydrological model simulations, *Adv. Water Resour.*, *97*, 299–313.

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, U.S. Dep. of the Inter., U.S. Geol. Surv.

Renard, B., D. Kavetski, and G. Kuczera (2009), Comment on "an integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction" by Newsha K. Ajami et al., *Water Resour. Res.*, *45*, W03603, doi:10.1029/2007WR006538.

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.

Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resour. Res.*, *47*, W11516, doi:10.1029/2011WR010643.

Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola (2010), Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, *Comput. Phys. Commun.*, *181*(2), 259–270.

Sapriza-Azuri, G., J. Jódar, V. Navarro, L. J. Slooten, J. Carrera, and H. V. Gupta (2015), Impacts of rainfall spatial variability on hydrogeological response, *Water Resour. Res.*, *51*, 1300–1314, doi:10.1002/2014WR016168.

Scanlon, B. R., R. W. Healy, and P. G. Cook (2002), Choosing appropriate techniques for quantifying groundwater recharge, *Hydrogeol. J.*, *10*(1), 18–39.

Tonkin, M. J., and J. Doherty (2005), A hybrid regularized inversion methodology for highly parameterized environmental models, *Water Resour. Res.*, *41*, W10412, doi:10.1029/2005WR003995.

Tonkin, M. J., and J. Doherty (2009), Calibration-constrained Monte Carlo analysis of highly parameterized models using subspace techniques, *Water Resour. Res.*, *45*, W00B10, doi:10.1029/2007WR006678.

Trenberth, K. E. (2011), Changes in precipitation with climate change, *Clim. Res.*, *47*(1–2), 123–138.

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.

White, J. T., J. E. Doherty, and J. D. Hughes (2014), Quantifying the predictive consequences of model error with linear subspace analysis, *Water Resour. Res.*, *50*, 1152–1173.

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