Computers & Geosciences 85 (2015) 124–136

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Data-driven methods to improve baseflow prediction of a regional groundwater model

Tianfang Xu*, Albert J. Valocchi

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ARTICLE INFO

Article history:
Received 15 July 2014
Received in revised form
16 January 2015
Accepted 27 May 2015
Available online 4 June 2015

Keywords:
Statistical learning
Baseflow
Predictive error

ABSTRACT

Physically-based models of groundwater flow are powerful tools for water resources assessment under varying hydrologic, climate and human development conditions. One of the most important topics of investigation is how these conditions will affect the discharge of groundwater to rivers and streams (i.e. baseflow). Groundwater flow models are based upon discretized solution of mass balance equations, and contain important hydrogeological parameters that vary in space and cannot be measured. Common practice is to use least squares regression to estimate parameters and to infer prediction and associated uncertainty. Nevertheless, the unavoidable uncertainty associated with physically-based groundwater models often results in both aleatoric and epistemic model calibration errors, thus violating a key assumption for regression-based parameter estimation and uncertainty quantification. We present a complementary data-driven modeling and uncertainty quantification (DDM-UQ) framework to improve predictive accuracy of physically-based groundwater models and to provide more robust prediction intervals. First, we develop data-driven models (DDMs) based on statistical learning techniques to correct the bias of the calibrated groundwater model. Second, we characterize the aleatoric component of groundwater model residual using both parametric and non-parametric distribution estimation methods. We test the complementary data-driven framework on a real-world case study of the Republican River Basin, where a regional groundwater flow model was developed to assess the impact of groundwater pumping for irrigation. Compared to using only the flow model, DDM-UQ provides more accurate monthly baseflow predictions. In addition, DDM-UQ yields prediction intervals with coverage probability consistent with validation data. The DDM-UQ framework is computationally efficient and is expected to be applicable to many geoscience models for which model structural error is not negligible.

1. Introduction

Effective water resource management policies and practices require clear understanding of the interaction between ground water aquifers and surface-water bodies such as streams and rivers. Of particular interest is baseflow, which represents the net groundwater discharge to the stream. Accurate quantification of baseflow is critical when dealing with issues such as water supply reliability, low flow requirements for in-stream ecology, and water allocation and trading. Physically-based models of groundwater flow are powerful tools to simulate and predict baseflow under varying hydrologic, climate and human development conditions.

However, predictions made from groundwater models are subject to error and uncertainty. The inherent error and uncertainty in groundwater modeling has been widely recognized in the literature as arising from multiple sources, including structure, parameter, input data and measurements used to evaluate the model (Caers, 2011; Dou et al., 1997; Hunt and Welter, 2010; Liu and Gupta, 2007). As a result, the model simulation is subject to both aleatoric and epistemic errors that cannot be fully attributed to measurement error. The model residuals (i.e. the difference between model simulation and observations) may have complex statistical characteristics, such as temporal and spatial correlation and non-normality (Doherty and Welter, 2010; Honti et al., 2013). Common practice is to use least squares regression to estimate model parameters and associated uncertainty from historical observation data; the calibrated model is then used for subsequent prediction and uncertainty analysis (Doherty et al., 1994; Hill and Tiedeman, 2007). A fundamental assumption of least squares regression is that model residuals can be described by a noise term corresponding to measurement error and that the noise term is uncorrelated and Gaussian distributed. This assumption is often violated when the groundwater model has significant input and

* Corresponding author.
E-mail address: txu3@illinois.edu (T. Xu).

http://dx.doi.org/10.1016/j.cageo.2015.05.016
0098-3004/

structural errors. As a result, simulations made with the calibrated model could be biased and the resulting predictive uncertainty intervals may be unreliable (Honti et al., 2013).

The limitation of classic least squares calibration highlights the need for proper treatment of model residuals in order to reliably assess predictive uncertainty. Methods have been proposed to accommodate correlated and/or non-Gaussian residuals of surface and ground water models, typically relying on an error model. Correlation in model residuals can be inferred using the first-order-second-moment method (Tiedeman and Green, 2013) or simulated using autoregressive models (Bates and Campbell, 2001; Kuczera, 1983; Lu et al., 2013). Traditionally, the Gaussianity of residuals can be improved using power transformations (Bates and Campbell, 2001; Box and Cox, 1964; Kuczera, 1983). Schoups and Vrugt (2010) proposed a generalized likelihood function based on a universal statistical error model to explicitly handle residual errors that are correlated, heteroscedastic and non-Gaussian. This and similar approaches have been applied to modeling rainfall-runoff (Schoups and Vrugt, 2010), unsaturated flow (Erdal et al., 2012) and groundwater contaminant transport (Shi et al., 2014). Kennedy and O'Hagan (2001) proposed a generic Bayesian formulation that integrates a Gaussian process error model to characterize predictive uncertainty of numerical simulation models. An application of this approach in river water quality modeling can be found in Reichert and Schuwirth (2012).

The error model is sometimes inferred jointly with the parameters of one or more hydrologic models having different structures (Kennedy and O'Hagan, 2001; Reichert and Schuwirth, 2012; Schoups and Vrugt, 2010). In this way, the joint inference approach can assess the contribution to predictive uncertainty from parameter, model structural, input data and measurement uncertainty. However, the interactions among different uncertainty sources pose challenges to the identification of these contributions (Kennedy and O'Hagan, 2001). In addition, the computational cost associated with joint inference is often high and even infeasible for complex models having long evaluation time. On the contrary, postprocessor approaches (Evin et al., 2014) estimate the error model from the residuals of a single calibrated hydrologic model (Lu et al., 2013; Pianosi and Raso, 2012; Solomatine and Shrestha, 2009; Weerts et al., 2011). It is assumed that the uncertainties arising from structural, parametric and data errors are represented implicitly by the model residuals. As reported in Evin et al. (2014), a postprocessor method yielded predictive uncertainty estimates comparable to a joint inference approach in a synthetic case study, and performed more robustly in a real-world case study. These findings suggest that postprocessor approaches comprise a computationally efficient alternative for post-calibration predictive uncertainty analysis. Therefore this study adopts a postprocessor approach to estimate the error model.

Existing postprocessor methods focus on time series data, and most of them rely on relatively simple statistical description of the model residual distribution (Evin et al., 2014; López López et al., 2014; Pianosi and Raso, 2012; Weerts et al., 2011). The challenge lies in how to configure the form of the error model to be capable of characterizing the distribution of complicated spatiotemporal residual fields of groundwater models. Fortunately, the statistical characterization of model residuals can be approached from an inductive, data-driven modeling prospective. Statistical learning techniques such as artificial neural networks, model trees and locally weighted regression have been successfully applied to uncertainty analysis of rainfall-runoff models (Dogulu et al., 2014; Shrestha and Solomatine, 2006; Solomatine and Shrestha, 2009). These algorithms do not require explicit assumption about the residual distribution. Instead, given a set of historical data, they are able to learn complex relations between the response variable (i.e. model residual or its quantiles, in the context of error modeling) and selected input variables. Besides the above mentioned uncertainty analysis applications, data-driven error models based on statistical learning techniques have proven effective for bias correction (also commonly referred to as error correction) of rainfall-runoff (Abebe and Price, 2003; Goswami et al., 2005) and groundwater models (Demissie et al., 2009; Gusyev et al., 2013; Xu et al., 2014).

However, previous groundwater applications of data-driven error models (Demissie et al., 2009; Gusyev et al., 2013; Xu et al., 2014) focus on using deterministic statistical learning methods for bias correction and cannot provide information about prediction uncertainty. This study fills the gap of integrating advanced statistical learning techniques into the postprocessor approach to statistically characterize groundwater model residuals, which are usually spatiotemporal and substantially more complicated than time series data. We present a complementary data-driven modeling and uncertainty quantification (DDM-UQ) framework to reduce the predictive bias of physically-based groundwater models and to provide more robust prediction intervals. First, we develop data-driven models (DDMs) based on statistical learning techniques to account for the bias of the calibrated groundwater model. By learning from the historical error of the groundwater model, the DDMs are capable of correcting its bias when the model is used for forecasting or extrapolation purposes. Two statistical learning techniques, random forests and support vector machine, are used to build the DDMs. Second, we estimate prediction uncertainty due to the aleatoric component of groundwater model residuals using both parametric and non-parametric distribution estimation methods. We then calculate the prediction interval by imposing the aleatoric error distribution on the DDMs-corrected prediction of interest. The DDM-UQ framework is tested on baseflow prediction of a real-world case study of the Republican River Basin.

The remainder of this paper is organized as follows. Section 2 briefly reviews the statistical learning techniques used in the DDM-UQ framework. Section 3 introduces the proposed DDM-UQ framework as well as performance assessment metrics. Next the DDM-UQ framework is tested on a real-world case study; the data and application procedures are described in Section 4. The results are presented and discussed in Section 5. Finally, Section 6 provides conclusions and recommendations.

2. Overview of statistical learning techniques

This section briefly reviews three statistical learning techniques used in this study. In contrast to physically-based groundwater models, statistical learning techniques learn inductively from the data. Based on a set of *training* data, a statistical learning algorithm learns a mapping from the input variables to the output (or response) variable that can be generalized to predict on a separate set of *testing* data. Cross validation (CV) is the most widely used tool to assess generalization error for tuning hyperparameters of statistical learning algorithms (further details are in Sections 2.2 and 2.3). Ten-fold CV is carried out in this study. The training dataset is randomly partitioned into 10 subsets of approximately equal size. Every time, a DDM is trained using nine subsets and tested on the remaining one to assess the generalization error or testing error. This step is repeated 10 times until every subset has been used once as testing data. The CV process can be repeated using varying hyperparameter values; the hyperparameter set that yields lowest generalization error (averaged over 10 subsets) is selected. Finally the DDM is retrained using the whole training data with the selected hyperparameter set.

2.1. Clustering

Cluster analysis partitions data into groups with the goal of maximizing the similarity of data within the same group and minimizing the similarity of data across groups. Similarity is often defined based on a distance measure, and smaller distance indicates higher similarity. This study employs the widely used k-means clustering algorithm. This algorithm starts from a random initialization of cluster assignment and iteratively minimizes the *within-cluster point scatter*. The within-cluster point scatter is defined as the sum of the distance between every pair of data points assigned to the same cluster (Hastie et al., 2001). In this study, *MATLAB*® subroutine `kmeans` is used to implement k-means clustering with squared Euclidean distance.

2.2. Random forests

Decision trees are a conceptually simple yet powerful non-parametric classification and regression tool (Hastie et al., 2001). A tree-based classification and regression algorithm, classification and regression trees (CART) is briefly described here (Breiman et al., 1984). CART recursively partitions the feature space into rectangular regions and fits a constant value in each region. The resulting tree is analogous to a piecewise constant function. Let $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, n$ denote a set of training data, where $\mathbf{x}_i = \{x_{i,1}, \ldots, x_{i,p}\}^T$ is an input data point, p is the dimension of the input feature space and y_i is the corresponding response variable. A binary split at $x_j = s$ partitions the space into two regions; the average values of observations in each region are fitted to the two regions, respectively. The splitting variable x_j and split point s at every non-terminal node are chosen to maximize the goodness-of-fit at this node. The splitting process is repeated on the resulting regions, and a maximal tree is grown until some minimum node size is reached at the leaves (terminal nodes). Subsequently, the maximal tree is usually pruned to a subtree to prevent overfitting.

One disadvantage of decision trees is statistical instability, i.e. small changes in the training data may induce large changes in the tree structure (Hastie et al., 2001). Random forests (RF) are an ensemble learning method proposed by Breiman (2001) to overcome the instability of decision trees. A RF consists of an ensemble of CARTs. Each CART is grown on a bootstrap sample of the training data $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, n$. A bootstrap sample is obtained by randomly drawing n observations from the training data with replacement. Each bootstrap sample leaves out about 1/3 of the data, which are called the out-of-bag observations. After training, the prediction for an unseen data point $\mathbf{x}^*$ is given by averaging the predictions from all individual CARTs, or more rigorously (Meinshausen, 2006)

$$\hat{y}(\mathbf{x}^*) = E[y(\mathbf{x}^*)|\mathbf{x}^*] = \sum_{i=1}^{n} w_i(\mathbf{x}^*) y_i, \tag{1a}$$

$$w_i(\mathbf{x}^*) = k^{-1} \sum_{t=1}^{k} w_{i,t}(\mathbf{x}^*), \tag{1b}$$

$$w_{i,t}(\mathbf{x}^*) = \frac{\mathbf{1}\{\mathbf{x}_i - R_t(\mathbf{x}^*)\}}{\#\{j|\mathbf{x}_j \in R_t(\mathbf{x}^*)\}}. \tag{1c}$$

In the above equations, k is the number of trees in RF, $R_t(\mathbf{x}^*)$ is the leaf of the t-th tree that contains the testing data point $\mathbf{x}^*$, and $\mathbf{1}\{\mathbf{x}_i \in R_t(\mathbf{x}^*)\}$ is an indicator function that equals 1 if $\mathbf{x}_i$ is in the leaf $R_t(\mathbf{x}^*)$ and 0 otherwise. The weights $w_{i,t}(\mathbf{x}^*)$, $i = 1, \ldots, n$, defined for the t-th tree in the ensemble, sum to one and are positive if $\mathbf{x}_i \in R_t(\mathbf{x}^*)$ and 0 otherwise. The weight $w_i(\mathbf{x}^*)$ is obtained by averaging $w_{i,t}(\mathbf{x}^*)$ over all trees. In this way, the conditional mean prediction $\hat{y}(\mathbf{x}^*)$ can be considered as the weighted average of observations y_i, $i = 1, \ldots, n$.

Because of bootstrap aggregating, RF is not prone to overfitting. Therefore pruning of individual CART is not necessary. Furthermore, at each split during the construction of a single CART, the splitting variable is selected among a randomly chosen subset of input variables. The size of the random set and the node size (i.e. minimum number of observations in the leaves) are two tuning hyperparameters. In this study, they are conventionally set to $p/3$ and 10, respectively. It is noteworthy that the performance of RF changes very little over a wide range of the two hyperparameters (Meinshausen, 2006; Svetnik et al., 2003). Random forests also provide the rank of importance of input variables. The importance of $\mathbf{x}_j$ is computed by averaging the increase of out-of-bag error after permuting $\mathbf{x}_j$ over all CARTs.

Random forests (RF) accurately estimate the mean of response variable y conditioned on input data $\mathbf{x}$. Since their introduction, they have gained popularity in various fields such as meteorology (Cloke and Pappenberger, 2008), soil science (Ließ et al., 2012) and natural hazard assessment (Catani et al., 2013). Later, a variation of RF, namely the quantile regression forests (QRF) (Meinshausen, 2006) was introduced to approximate the conditional distribution of a response variable. The family of quantile regression estimates the quantiles of the response variable conditioned on input variables, usually solved by minimization of a loss function (Koenker, 2005). The quantile regression forest algorithm takes a different approach; it is designed to exploit the information contained in random forests about the variability of the response variable. Given an unseen data point $\mathbf{x}^*$, the conditional distribution function of y is defined as

$$F(y|\mathbf{x}^*) = P(Y \le y|\mathbf{x}^*) = E[\mathbf{1}\{Y \le y\}|\mathbf{x}^*]. \tag{2}$$

Analogous to Eq. (1), the QRF estimates of $F(y|\mathbf{x}^*)$ is given by

$$\hat{F}(y|\mathbf{x}^*) = \sum_{i=1}^{n} w_i(\mathbf{x}^*) \mathbf{1}\{y_i \le y\}. \tag{3}$$

With the estimated conditional distribution of a prediction, it is then straightforward to construct prediction intervals with specified confidence levels.

The packages `randomForest` and `quantregForest` in R environment are used to construct random forests and quantile regression forests.

2.3. Support vector machine regression

Support vector regression (SVR) (Vapnik, 1998) is a relatively new class of learning algorithm that has been applied in many fields including rainfall-runoff modeling (Rasouli et al., 2012), radioactive soil contamination (Kanevski et al., 2004) and groundwater hydrology (Asefa et al., 2005; Xu et al., 2014). The SVR algorithm has good generalization performance, because it seeks to minimize an upper bound of the generalization error rather than minimize the training error. In addition, the solution of SVR is globally optimal under conditions that can be easily satisfied, while many other statistical learning tools (e.g. artificial neural network) may converge to local minima. This section briefly overviews ε-SVR (Vapnik, 1998).

Given a set of training data $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, n$, where $\mathbf{x}_i$ denotes input and y_i denotes output that has been observed, the idea of SVR is to first project input $\mathbf{x}$ to a higher dimensional feature space by the map $\Phi: \mathcal{X} \to \mathcal{F}$, and then carry out a linear regression of y in the feature space $\Phi(\mathbf{x})$:

$$f(\mathbf{x}) = w \cdot \Phi(\mathbf{x}) + b. \tag{4}$$

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

$$\text{minimize} \quad \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^* \right) \tag{5a}$$

$$\text{subject to} \quad (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) - y_i \leq \varepsilon + \xi_i, \tag{5b}$$

$$y_i - \left(\mathbf{w}^T \Phi(\mathbf{x}_i) + b \right) \leq \varepsilon + \xi_i^*, \tag{5c}$$

$$\xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, n. \tag{5d}$$

The first term in Eq. (5a) represents the complexity of the regression model and therefore acts as regularization. The second term represents goodness-of-fit to training data; the slack variables ξ_i, ξ_i^* are introduced to cope with otherwise infeasible constraints of the optimization problem. They are derived from the ε-insensitive loss function $|\xi|_\varepsilon = \max\{0, |y_i - f(\mathbf{x}_i)| - \varepsilon\}$. The constant C in Eq. (5a) determines the trade-off between the flatness of f and deviations exceeding ε.

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

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

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2). \tag{6}$$

The regularization hyperparameter C is chosen according to the training data following the recommendations of Cherkassky and Ma (2004). The loss function hyperparameter ε and kernel width hyperparameter γ are tuned by 10-fold cross validation. The LIBSVM toolbox (Chang and Lin, 2011) is used to implement ε-SVR.

3. Data-driven uncertainty quantification framework

3.1. Framework overview

The data-driven uncertainty quantification (DDM-UQ) framework (Fig. 1) presented in this study extends the complementary bias-correcting data-driven models developed in Demissie et al. (2009) and Xu et al. (2014). The DDM-UQ framework works with an existing calibrated model, which follows the common practice in groundwater modeling. The framework is motivated by the observation in groundwater modeling that a Gaussian distributed error term with zero mean and small variance is generally not achievable via regression-based calibration. Often, the simulation results of even well-calibrated models contain both aleatoric and epistemic errors (Doherty and Welter, 2010; Xu et al., 2014). The overall residual of the calibrated model can be viewed as a combination of the errors associated with model structure, parameters, input stress and measurements used for calibration. Proper description of the residual of physically-based models is vital for accurate prediction and associated uncertainty analysis.

Let z_i, $i = 1, 2, \ldots, n$ denote a quantify of interest; in the context of groundwater modeling, it can be the groundwater head, flux from aquifer to stream, solute concentration and other types of response of the groundwater system. Further letting M_i denote the simulation result of the calibrated physically-based groundwater model, the relation between z_i and M_i can be written as

$$z_i = M_i + y_i = M_i + \hat{y}(\mathbf{x}_i, \phi) + \epsilon_i. \tag{7}$$

In the above equation, y_i denotes model residual lumped from all sources of error. We assume that the residual can be decomposed into two components, namely the epistemic term (bias) and the aleatoric term. This decomposition is similar as in Pianosi and Raso (2012), where the bias term is handled using a first-order autoregressive error corrector and the aleatoric term is assumed Gaussian with time-varying standard deviation. Although Eq. (7) adopts the natural additive representation of residual components, the proposed framework can easily incorporate other forms, e.g. multiplicative. Unlike Pianosi and Raso (2012) which focuses on time series hydrologic forecast, in this study we characterize the epistemic error $\hat{y}(\mathbf{x}_i, \phi)$ using deterministic data-driven models (DDMs) based on statistical learning techniques with input $\mathbf{x}_i$ and hyperparameters ϕ. Once the DDMs are trained with a set of training data $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, n$, where $y_i = z_i - M_i$, they are capable of correcting the bias when the model is used for forecasting or extrapolation purposes (Xu et al., 2014). The two statistical learning techniques described in Section 2, the random forests (RF) and support vector machine regression (SVR), are used to build the bias-correcting DDMs. Because the DDMs compensate for epistemic error that has temporal and spatial correlation structures, the aleatoric errors ϵ_i's can be considered as independently distributed random noises. Next, the probability distribution of ϵ_i is estimated using methods described in the following section. Then prediction intervals can be derived by imposing the aleatoric error distribution on the DDMs-corrected prediction of the physically-based groundwater model. In this way, the DDM-UQ framework is expected to not only reduce the predictive bias of the groundwater model, but also provide more robust prediction intervals. An optional preprocessing step is to use clustering algorithms (Section 2.1) to divide data points into groups such that data belonging to the same cluster would have similar residual characteristics. Then local DDMs and prediction intervals can be constructed for individual groups, and they are expected to be more robust than a global DDM and prediction interval fitted on the whole dataset.

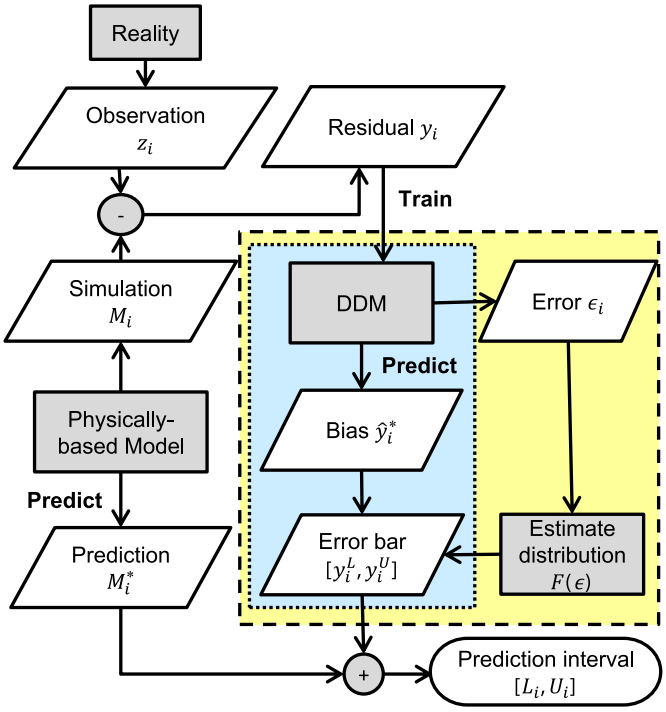

Fig. 1. The framework of the data-driven modeling uncertainty quantification (DDM-UQ) method. The blue dotted box corresponds to the DDM-UQ(QRF) implementation, and the yellow dashed box corresponds to the DDM-UQ(SVR) implementation, as further explained in Section 3.2.

3.2. Constructing prediction interval

In this study, both parametric and non-parametric distribution estimation methods are implemented in order to construct prediction intervals. The parametric approach is implemented with SVR, and this implementation will be referred to as DDM-UQ(SVR). The parametric approach is based on a direct method proposed in Lin et al. (2004). Although originally introduced to construct predictive probability intervals for SVR predictions, this generic method can be used with almost any statistical learning technique.

In order to estimate the distribution of ϵ_i's, 10-fold cross validation is carried out once more after tuning SVR hyperparameters. A distribution is then fitted to the CV generalization errors on all 10 subsets, i.e. $z_i - M_i - \hat{y}(\mathbf{x}_i, \phi)$, $i = 1, \ldots, n$. The distribution type needs to be chosen on a case-by-case basis. In this study, three types of distributions are considered as candidates, namely Gaussian, Laplace and Cauchy distributions. The Gaussian distribution is the most widely used noise model. Laplace and Cauchy distributions (Delleur et al., 1976; Kotz et al., 2001; Lin et al., 2004) are also considered in this study because preliminary analysis revealed that their probability density functions resemble the shape of the histogram of aleatoric errors. Each of Gaussian, Laplace and Cauchy distributions has two parameters in their CDFs. For Gaussian and Laplace distributions, the two parameters control the mean and variance; for the Cauchy distribution (the mean and variance are undefined) the two parameters specify the median and interquartile range. Because the Cauchy distribution has heavy tails, using it as the error model allows for outliers (Delleur et al., 1976). Given the CV generalization errors, the parameters can be inferred using maximum likelihood estimation (MLE). Once their parameters are estimated, the goodness-of-fit of various distribution families can be assessed by comparing the likelihood corresponding to the estimated parameters. The distribution possessing the highest likelihood is preferred. Let $F(\epsilon)$ denote the cumulative distribution function (CDF) of the fitted distribution, and F^{-1} denote the inverse of the CDF. The prediction interval $[L_i, U_i]$ for a quantity of interest z_i can then be constructed with a specified confidence level $1 - \alpha$:

$$L_i = M_i + y_i^L = M_i + \hat{y}_i + F^{-1}(\alpha/2), \tag{8a}$$

$$U_i = M_i + y_i^U = M_i + \hat{y}_i + F^{-1}(1 - \alpha/2). \tag{8b}$$

Besides DDM-UQ(SVR), this study also implemented non-parametric distribution estimation method to construct prediction intervals. The non-parametric method is based on the quantile regression forests (QRF) algorithm described in Section 2.2 that builds upon the ensemble nature of random forests. As mentioned therein, this non-parametric approach falls within the category of quantile regression. Examples of using quantile regression to assess predictive uncertainty of flood forecasting models can be found in López López et al. (2014) and Weerts et al. (2011), where the quantiles are estimated as a linear or piecewise linear function of a deterministic forecast value after transformation into the Gaussian domain. The present study adopts the QRF algorithm to better simulate the nonlinear relationship between spatio-temporal residuals of groundwater models and multiple input variables. The resulting version of the data-driven uncertainty quantification framework is hereby abbreviated as DDM-UQ(QRF). Following the notation of Eq. (7), DDMs based on random forests account for the epistemic term $\hat{y}(\mathbf{x}_i, \phi)$. The QRF algorithm then estimates the distribution of model residual y_i rather than ϵ_i. This implementation difference with DDM-UQ(SVR) is also reflected in Fig. 1. However, it is noteworthy that estimating the distribution of y_i is equivalent to estimating the distribution of ϵ_i. To see this equivalence, simply replace y in Eq. (3) with ϵ and rewrite the equation into the following format:

$$\hat{F}(\epsilon|\mathbf{x}^*) = \sum_{i=1}^{n} w_i(\mathbf{x}^*)\mathbf{1}\{\epsilon_i \le \epsilon\}, \tag{9}$$

where $\epsilon_i = z_i - M_i - \hat{y}(\mathbf{x}_i, \phi)$. Subsequently, prediction intervals $[y_i^L, y_i^U]$ and $[L_i, U_i]$ can be constructed similarly as in Eqs. (8a) and (8b).

3.3. Performance assessment

Given a set of training data $\{\mathbf{x}_i, z_i\}$, $i = 1, \ldots, n$, the statistical description of predictive error of the physically-based groundwater model is constructed following the procedures outlined in Section 3.2. Then the performance of the DDM-UQ framework is evaluated on a testing dataset $\{\mathbf{x}_i^*, z_i^*\}$, $i = 1, \ldots, m$ that is independent from the training data; the asterisk differentiates the testing scenario from training data. For any quantity of interest z_i^*, the DDM-UQ framework generates two outputs: the DDMs corrected prediction $\hat{z}_i^* = M_i^* + \hat{y}_i^*$ and the associated prediction interval $[L_i, U_i]$. Correspondingly, the performance of DDM-UQ is assessed in two aspects, namely the efficiency of bias-correcting DDMs in reducing predictive bias of the physically-based model and the quality of the prediction intervals. This section overviews the statistics used to assess these two aspects.

If a new measurement z_i^* of true system response is available, the error of DDMs corrected prediction of interest $\hat{z}_i^* = M_i^* + \hat{y}_i^*$ is written as $e_i^* = z_i^* - \hat{z}_i^*$. The predictive accuracy of DDM-UQ is evaluated using three statistics based on e_i^*, $i = 1, 2, \ldots, m$. The percent bias (PBIAS) measures the tendency of systematic overestimation or underestimation (Gupta et al., 1999). It is defined as

$$\text{PBIAS} = \frac{\sum_{i=1}^{m} e_i^*}{\sum_{i=1}^{m} z_i^*} \times 100\%.$$

The mean absolute error (MAE) is computed as

$$\text{MAE} = \frac{1}{m}\sum_{j=1}^{m} |e_i^*|.$$

The Nash–Sutcliffe efficiency (NSE) is computed using the following equation:

$$\text{NSE} = 1 - \frac{\sum_{j=1}^{m} (e_i^*)^2}{\sum_{j=1}^{m} (z_i^* - \bar{z}_i^*)^2}, \tag{10}$$

where $\bar{z}_i^* = \sum_{j=1}^{m} z_i^*$. The range of NSE varies between $-\infty$ and 1.0 (perfect fit). An NSE coefficient less than zero implies that mean of observations is a better predictor than the model. A disadvantage of NSE is that it is not very sensitive to systematic overestimation or underestimation of the quantity of interest (Krause et al., 2005).

The quality of prediction intervals (PIs) yielded by DDM-UQ is evaluated using the prediction interval coverage probability (PICP), which is defined as the percentage of total observations that fall into the estimated prediction interval (Lin et al., 2004; Solomatine and Shrestha, 2009). Let $[L_i, U_i]$ denote the prediction interval with desired confidence level, and z_i^* denote the observed value. The PICP can be computed using

$$\text{PICP} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left\{L_i \le z_i^* \le U_i\right\},$$

where $\mathbf{1}\left\{L_i \le z_i^* \le U_i\right\}$ is an indicator function that equals 1 if the observation falls between the interval and 0 otherwise. As an example, a prediction interval with 90% confidence level should theoretically cover 90% of observation data. A PICP smaller than

90% suggests that the estimated PIs are biased and/or too narrow. In contrast, a PICP greater than 90% indicates that the estimated PIs are too wide and predictive uncertainty is overestimated.

4. RRCA model case study

4.1. Study area and data

The bias-correcting DDMs were applied to a regional groundwater flow model of the Republic River Basin spanning across Colorado, Kansas and Nebraska (Fig. 2). The Republican River Compact Administration (RRCA) model was developed in 2002 and updated annually to resolve the water conflicts and litigation among the states as growing water demand for irrigation and other uses led to dramatically increased groundwater pumping and decline in streamflow (Szilagyi, 2001). The input and output data of the model from 1918 to 2007 are available via the RRCA website (http://www.republicanrivercompact.org/). The model, constructed using a modified version of MODFLOW2000 (Harbaugh et al., 2000), has a single confined layer, a uniform grid size of 1 square mile (1.61×1.61 km^2) and monthly stress period. Hydraulic conductivity, saturated thickness, recharge and evapotranspiration rates were calibrated based on head measurements at over 10,000 wells and baseflow estimated from streamflow data at 65 gages from 1918 to 2000 using "trial and error" and automated calibration techniques (McKusick, 2003); however, more detailed information concerning the calibration process is not available.

The calibrated RRCA model is used to simulate baseflow under pumping and no pumping conditions. The simulation results are processed through a series of accounting procedures to evaluate the effect of groundwater irrigation pumping on streamflow. Therefore, accurate prediction of baseflow and associated uncertainty is of vital importance in this case study. The RRCA model was investigated in our previous study where statistical learning techniques were applied to reduce the model's head prediction bias (Xu et al., 2014). The focus of the present study is to explore the potential of using data-driven methods to improve baseflow prediction of the RRCA model. Baseflow prediction is generally considered more challenging than head prediction because of the complexity of river and aquifer interaction as well as scarcity of baseflow observations.

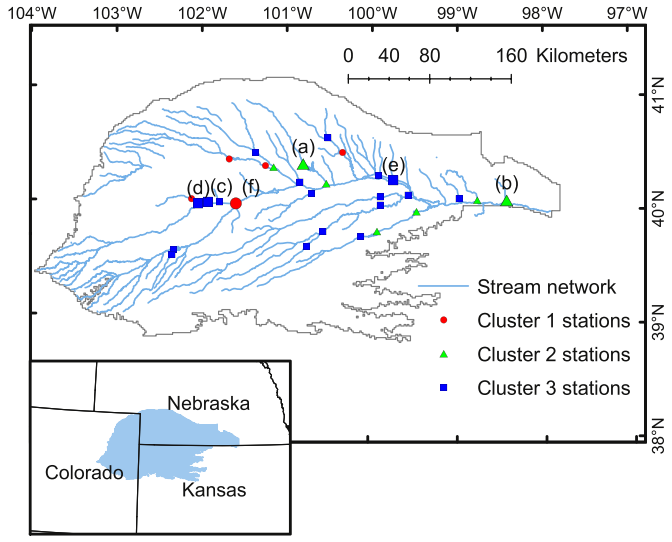

Fig. 2. Republican River Basin covering portions of eastern Colorado, northwest Kansas and southwest Nebraska. The locations of streamflow gaging stations are shown; color encodes which cluster a station belongs to (see Section 4.3 and Fig. 4). Enlarged symbols with labels indicate representative gaging stations for further investigation in Section 5. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Monthly baseflow calibration targets at 65 gaging stations are available via the RRCA website. The baseflow targets had been estimated by partitioning total streamflow into surface water and groundwater components (McKusick, 2003). In this study, the baseflow estimates are considered "observations" because there were no direct measurements of baseflow. The dataset in this study includes only gaging stations that have at least 10 years of monthly baseflow observations. In addition, this study focuses on headwater catchments under positive baseflow conditions, in which the aquifer discharges to the streams; streamflow losses are not included in the dataset. In addition, time periods corresponding to zero baseflow, either observed or simulated by the RRCA model, are removed. In total, the dataset is comprised of monthly baseflow at 30 gaging stations in the Republican River Basin from 1941 to 2000. Data during 1941–1989 are used to train the DDMs and fit the error distribution, and the remaining data during 1990–2000 are reserved for validation. Because baseflow separation results after the year 2000 are not available, it is not possible to further test the extrapolation capability of the proposed DDM-UQ framework for longer forecast horizon.

4.2. Residual analysis

Viewed globally, the model was calibrated to satisfactory accuracy given the complexity of hydrogeologic conditions it aimed to represent. In order to further assess the quality of baseflow computed by the RRCA model, baseflow residuals were calculated by subtracting computed baseflow from "observed" baseflow (i.e. estimated from measured streamflow via the hydrograph separation procedure). During the training period (1941–1989), the percent bias (PBIAS) is 8.94%, and the mean absolute residual (MAE) is 0.15 cubic meter per second (m^3/s). Subsequently, residual analysis was carried out to examine baseflow residuals, and the results are summarized in Fig. 3. Fig. 3(a) and (b) shows bias and heteroscedasticity in baseflow residuals. Fig. 3(c) is a quantile–quantile plot of the baseflow residuals versus a Gaussian distribution. The plot deviates from linear, indicating that the baseflow residuals are not Gaussian distributed. Fig. 3(d) shows Durbin–Watson (DW) statistics (Durbin and Watson, 1971) of one-month lagged baseflow residuals at each gaging station. The DW statistics of all stations are substantially smaller than two, suggesting autocorrelation. These systematic patterns in baseflow residuals point to the need for reducing baseflow prediction error and better characterization of associated uncertainty.

4.3. Framework implementation

The DDM-UQ framework described in Section 3 was used to improve predictive accuracy and derive prediction intervals of the regional RRCA groundwater model. Residual analysis revealed varying patterns in the baseflow residuals of the RRCA model among gaging stations. Cluster analysis can be employed to divide data into groups in order to construct a local error model for each group separately. For rainfall-runoff modeling, Shrestha and Solomatine (2006) and Solomatine and Shrestha (2009) employed fuzzy clustering in identified input space to divide the hydrograph into groups that correspond to the various mechanisms of the runoff generation process. Quantiles of the model error were then estimated for each cluster. In this study, the 30 gaging stations were grouped into three clusters using the standard *k*-means clustering algorithm as shown in Fig. 4; bias correction and uncertainty analysis were then performed for gaging stations within the same cluster. Clustering was based on the mean and standard

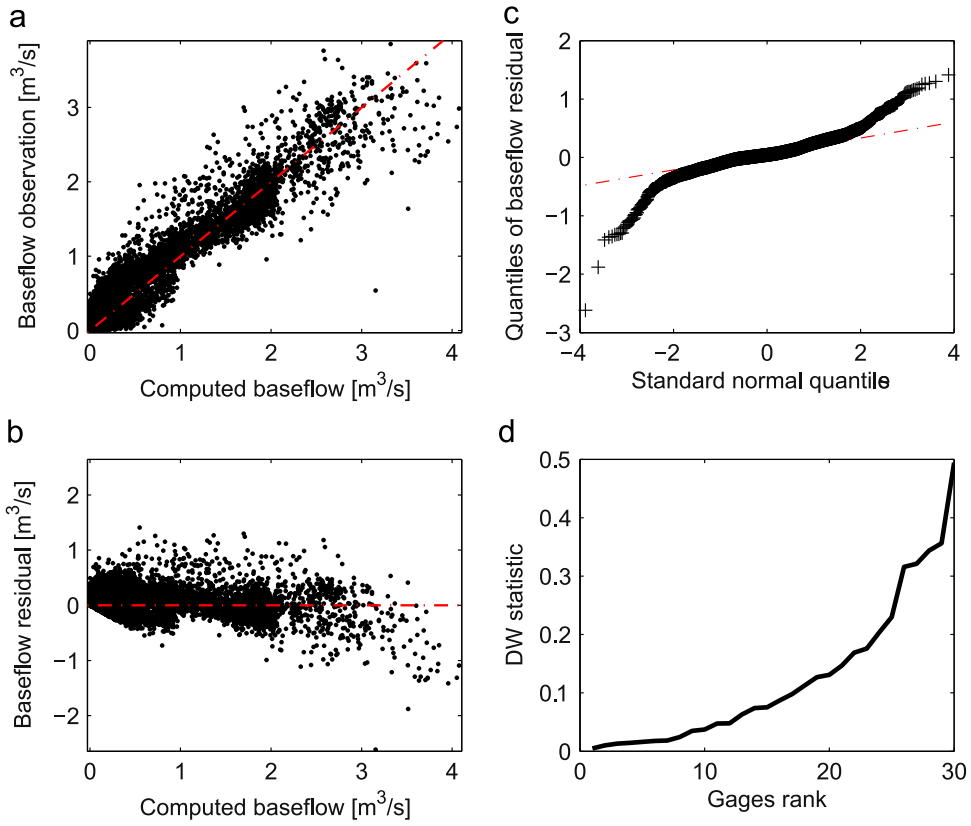

Fig. 3. (a) Plot of baseflow observations versus RRCA model computed counterparts during the training period (1941–1989). (b) Plot of residuals versus RRCA model computed baseflow. (c) Normal *Q*–*Q* plot of baseflow residuals. (d) Durbin–Watson statistics of one-month lagged residual at each gaging stations, sorted from lowest to highest.

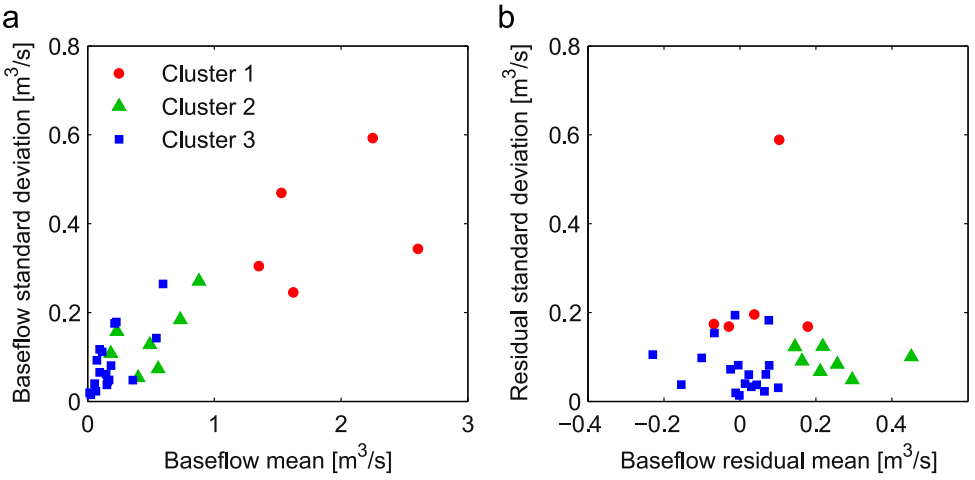

Fig. 4. Gaging stations were grouped into three clusters and plotted in the space of mean and standard deviation of (a) baseflow observations and (b) baseflow residuals during the period 1941–1989.

Table 1
The percent bias (PBIAS), mean absolute error (MAE) and Nash–Sutcliffe efficiency (NSE) statistics of baseflow prediction before and after corrected by DDMs and PICPs of estimated 90% prediction intervals.

	Training (1941–1989)				Testing (1990–2000)			
Method	PBIAS	MAE (m^3/s)	NSE	PICP	PBIAS	MAE (m^3/s)	NSE	PICP
MODFLOW	8.94%	0.15	0.90	–	18.13%	0.14	0.83	–
DDM-UQ(QRF)	−0.02%	0.04	0.99	97.50%	0.08%	0.07	0.93	82.45%
DDM-UQ(SVR)	0.32%	0.06	0.98	94.84%	2.30%	0.07	0.93	92.89%

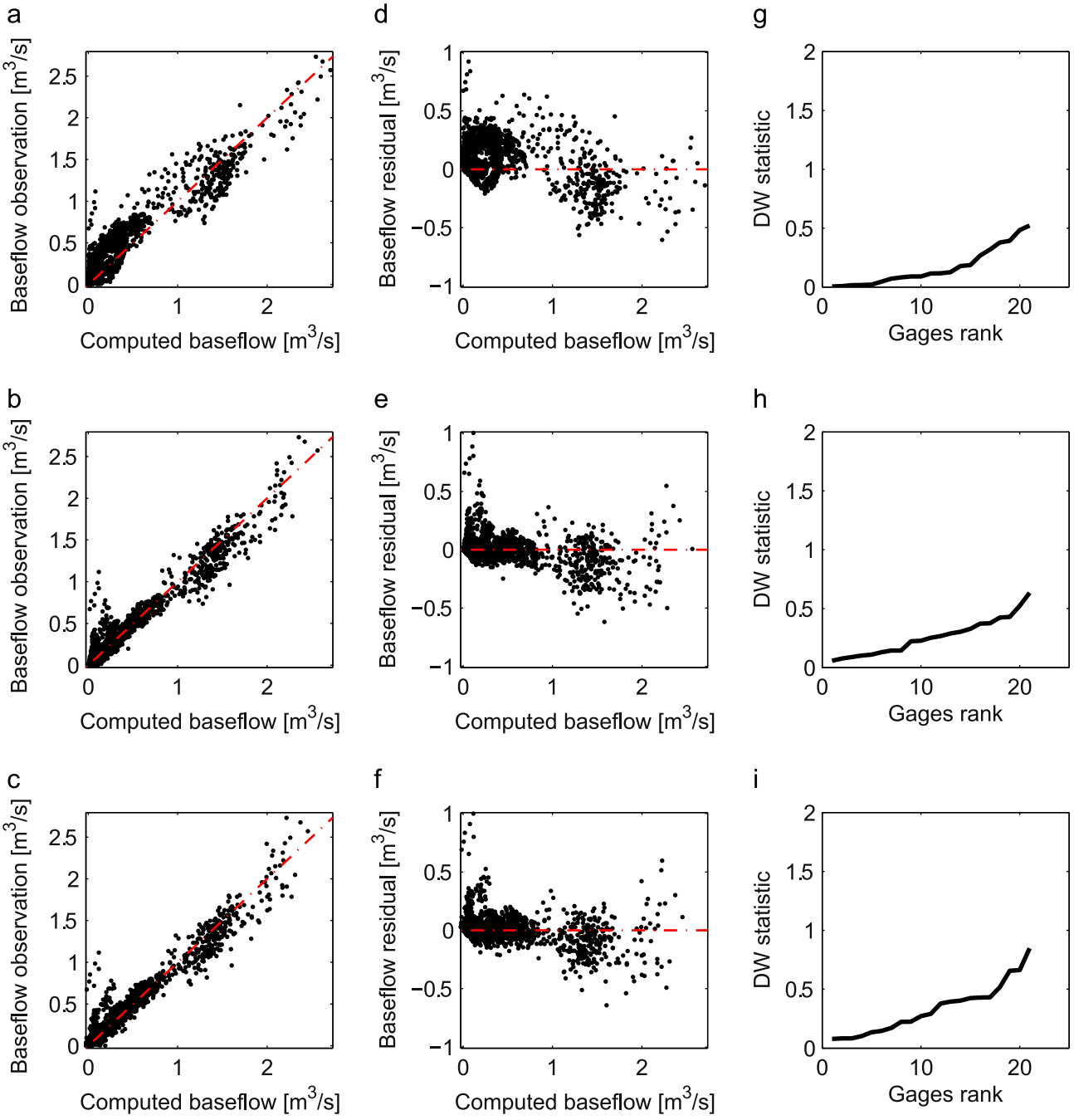

Fig. 5. Left column: plot of baseflow observations versus baseflow computed by the RRCA model (a), updated by RF (b) and SVR (c). Center column: plot of baseflow residuals versus baseflow computed by the RRCA model (d), updated by RF (e) and SVR (f). Right column: Durbin–Watson statistics of one-month lagged residual at each gaging stations, sorted from lowest to highest. Top row is based on the RRCA model simulation results, middle and bottom rows are, respectively, based on RF and SVR updated results. All for the testing period 1990–2000.

deviation of observed baseflow and residuals during the training period (1941–1989). The number of clusters was selected heuristically based on preliminary residual analysis results; optimization of the cluster number is beyond the scope of this study. As can be seen from Fig. 4, cluster 1 corresponds to big rivers with relatively high baseflow rate, while the other two clusters correspond to smaller streams. For gaging stations in cluster 2, the RRCA model significantly underestimates the baseflow, as indicated by the positive residuals. Arguably, the baseflow residuals at gaging stations that belong to the same cluster have similar patterns, and this is beneficial for applying DDMs for individual clusters.

Next DDMs were constructed to correct the baseflow prediction bias of the RRCA model. One random forest (RF) and one support vector regression (SVR) model were built for each cluster using data during the training period (1941–1989). Potential input features of bias-correcting DDMs consisted of month, location of gaging stations, precipitation rate, evapotranspiration (ET) rate, groundwater pumping rate, the difference between groundwater head and streambed elevation, and baseflow computed by the RRCA model. A set of neighboring grids were identified such that they are located within a distance of three miles (4.83 km) from the stream segment corresponding to a particular gaging station. Preliminary results showed that using different distances did not significantly change the results of DDM-UQ, mainly because the small gradient of precipitation, ET and groundwater head. The local precipitation rate, ET rate, pumping rate and groundwater

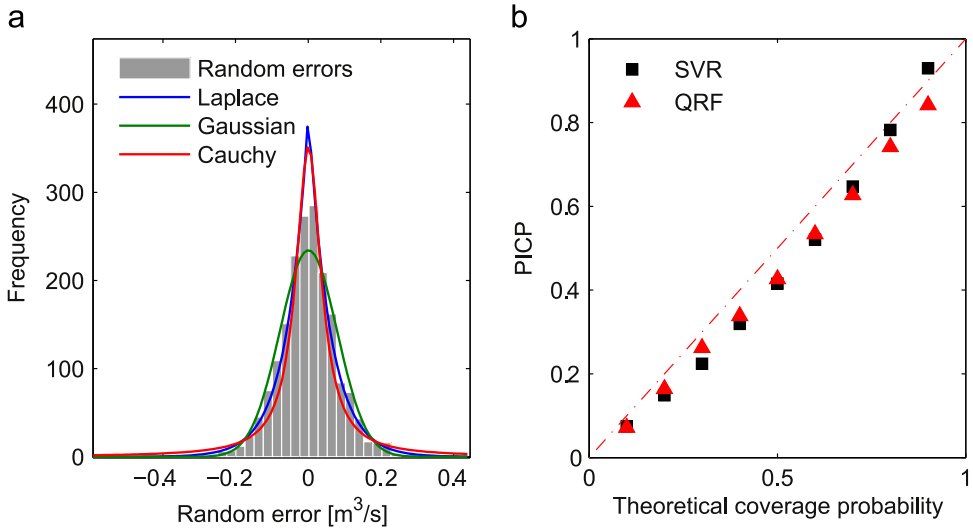

Fig. 6. (a) Histogram of ϵ_i's of cluster 2 during testing period 1941–1989. The histogram is fitted using Laplace, Gaussian and Cauchy distributions with MLE parameters. (b) The actual coverage probability of prediction intervals with varying confidence level. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

head for every gaging stations were then averaged over the neighboring grids.

The precipitation rate for every 1 mile by 1 mile (1.61 km) model grid was interpolated from monthly PRISM (Daly et al., 2008) precipitation data with a resolution of 30 arc-seconds (approximately 800 m). The evapotranspiration (ET) rate was extracted from the input file of RRCA model, which was calculated using the ET rate at three U.S. National Weather Service climate stations at Akron, McCook and Red Willow. Groundwater heads in neighboring cells were extracted from the simulation results of the RRCA model. The extracted head was then updated by bias-correcting DDMs developed in Xu et al. (2014). Subsequently, the head was averaged over neighboring cells and subtracted by the mean streambed elevation of the stream segment corresponding to the gaging station. The subtraction result is used as an indicator of the overall tendency of local groundwater discharging to the stream. Cell-by-cell pumping rate was extracted from the input of the RRCA model. In a forecast scenario, projected precipitation, ET and pumping rates can be used. In this study the "real" historical values were used as input data in the testing dataset in order to assess the DDM-UQ performance against historical baseflow observations.

As input variable selection is embedded in the tree growing process, the performance of a random forest is considered not sensitive to the presence of irrelevant input variables (Svetnik et al., 2003). Therefore all potential input variables were fed into the RF learning algorithm. The predictor importance estimated by RF was then used to select input data for SVR. Potential input variables were sorted from highest importance to lowest importance and added to the SVR one by one. A variable is accepted as an input variable if CV error decreases after adding it, and dropped off if CV error increases. Among the potential input variables, month, gaging station location, baseflow computed by the RRCA model and the difference between groundwater head and streambed elevation are selected for all clusters. Precipitation, ET and groundwater pumping rates are found to be important predictors in some clusters. Input variable selection depends on specific problem of concern; introduction to advanced input selection techniques can be found in Abebe and Price (2003), Catani et al. (2013), Galelli and Castelletti (2013), Solomatine and Shrestha (2009) and references therein.

5. Results and discussion

The results of DDMs updating are summarized in Table 1 and Fig. 5. Table 1 shows the percent bias (PBIAS), mean absolute error (MAE) and Nash–Sutcliffe efficiency (NSE) statistics of the baseflow simulated by the RRCA MODFLOW model and after DDMs bias correction during training and testing periods. For both training and testing periods, the DDMs effectively improved the baseflow prediction accuracy. For the testing period, RRCA model simulated baseflow is systematically smaller than observations (PBIAS = 18.13%); the bias is mostly removed by RF and SVR. The MAE is reduced by 50%, and the Nash–Sutcliffe efficiency (NSE) increases. The performance metrics in Table 1 are further illustrated in Fig. 5 during the testing period 1990–2000. As shown in Fig. 5(a)–(f), bias-correcting DDMs reduced the bias and magnitude of error. Comparing Fig. 5(g)–(i), it can be seen that DW statistics of baseflow residuals are increased for most gaging stations after application of DDMs. This suggests that the baseflow residuals after corrected by DDMs are less temporally correlated.

We then compute prediction intervals $[L_i, U_i]$ for each cluster using both parametric and non-parametric distribution estimation methods outlined in Section 3.2. For the DDM-UQ(SVR) implementation, Fig. 6(a) shows the histogram of the CV errors (ϵ_i's) of cluster 2 along with probability density functions (pdfs) of Laplace, Gaussian and Cauchy distributions with MLE parameters. For this cluster, Gaussian distribution has the smoothest peak, but does not fit the histogram well in the shoulder (area between peak and tails). In contrast, Cauchy distribution exhibits the heaviest tails. Overall, Laplace distribution with MLE parameters yields the highest likelihood value and is therefore selected to compute the prediction interval for this cluster. Laplace and Cauchy distributions are used to model the random error of clusters 1 and 3, respectively.

As reported in Table 1, the coverage probability of estimated prediction intervals (PICP) for the testing period given by DDM-UQ (QRF) is slightly lower than the theoretical value 90%; the PICP given by DDM-UQ(SVR) is close to 90%. In order to further scrutinize the goodness-of-fit of the estimated distribution to random errors (ϵ's), PIs of varying confidence levels were generated and associated PICPs computed. Fig. 6(b) plots the actual PICPs versus desired confidence levels. The red dashed line represents the ideal scenario in which the distribution model resembles exactly the real distribution of random error. For example, the 90% prediction

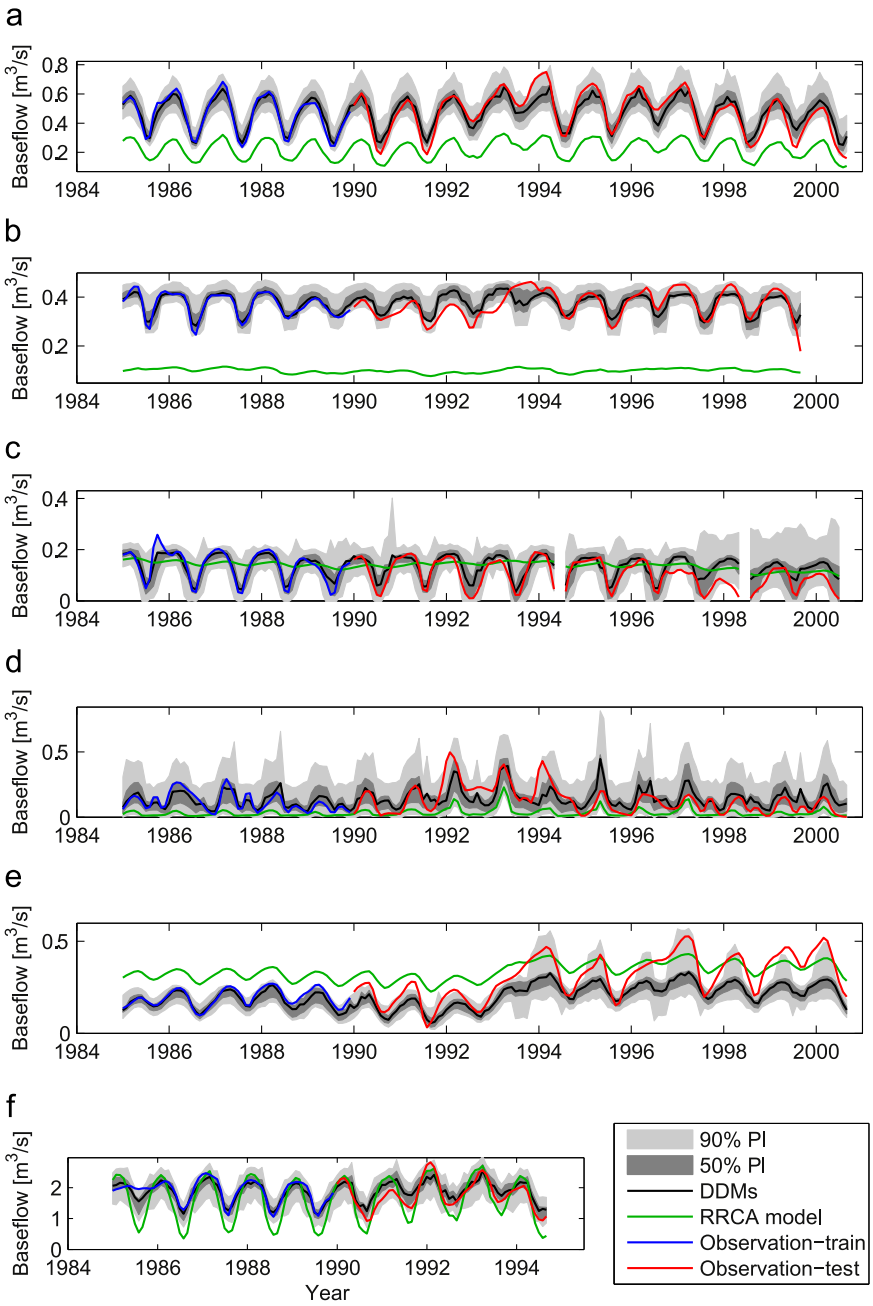

Fig. 7. Observed and RRCA model simulated baseflow, along with DDM-UQ(QRF) simulated mean and 50% and 90% prediction intervals of baseflow at six gaging stations. The station locations are plotted in Fig. 2 and their ID numbers and names are listed in Table 2.

interval should theoretically have a coverage probability of 90%. It can be seen from Fig. 6(b) that the PICPs are in general consistent with theoretical values. Notably, the discrepancy between PICPs of DDM-UQ(QRF) and theoretical values is larger for PIs with higher confidence level. This is possibly due to baseflow variability that is not explained by the random forests. To the contrary, PICPs of DDM-UQ(SVR) are lower than theoretical value in the middle region while above the theoretical value for 90% PI. The performance of DDM-UQ(SVR) could potentially be improved by adopting non-parametric methods to estimate the distribution of ϵ's, however the gain may not be sufficient to offset the increased complexity.

The above results have shown the overall performance of the DDM-UQ framework on improving baseflow prediction accuracy and estimating prediction intervals. Figs. 7 and 8 further illustrate the mean and prediction interval of baseflow provided by DDM-UQ(QRF) and DDM-UQ(SVR) at individual gaging stations. Information about these representative stations is listed in Table 2. The whole testing period (1990–2000) and a short segment of the training period (1985–1989) are shown. In terms of the accuracy of DDMs updated baseflow, overall RF and SVR performed similarly although SVR slightly outperformed RF at gaging station (e). For gaging stations (a) and (b), the mean prediction (black lines) given by DDM-UQ agrees satisfactorily with validation data. The bias-correcting DDMs not only fixed the systematic underestimation of the RRCA model simulation results, but also altered the intra-annual baseflow fluctuation. Likewise for gaging stations (c), (d) and (f), the DDMs moderately to significantly improved the prediction accuracy of the RRCA model. In particular, as shown in (d), in many years there occurred a smaller baseflow peak in autumn following the major peak in spring. The second peak is probably a combined effect of summer months precipitation, groundwater pumping for irrigation and irrigation return flow infiltration. The DDMs-

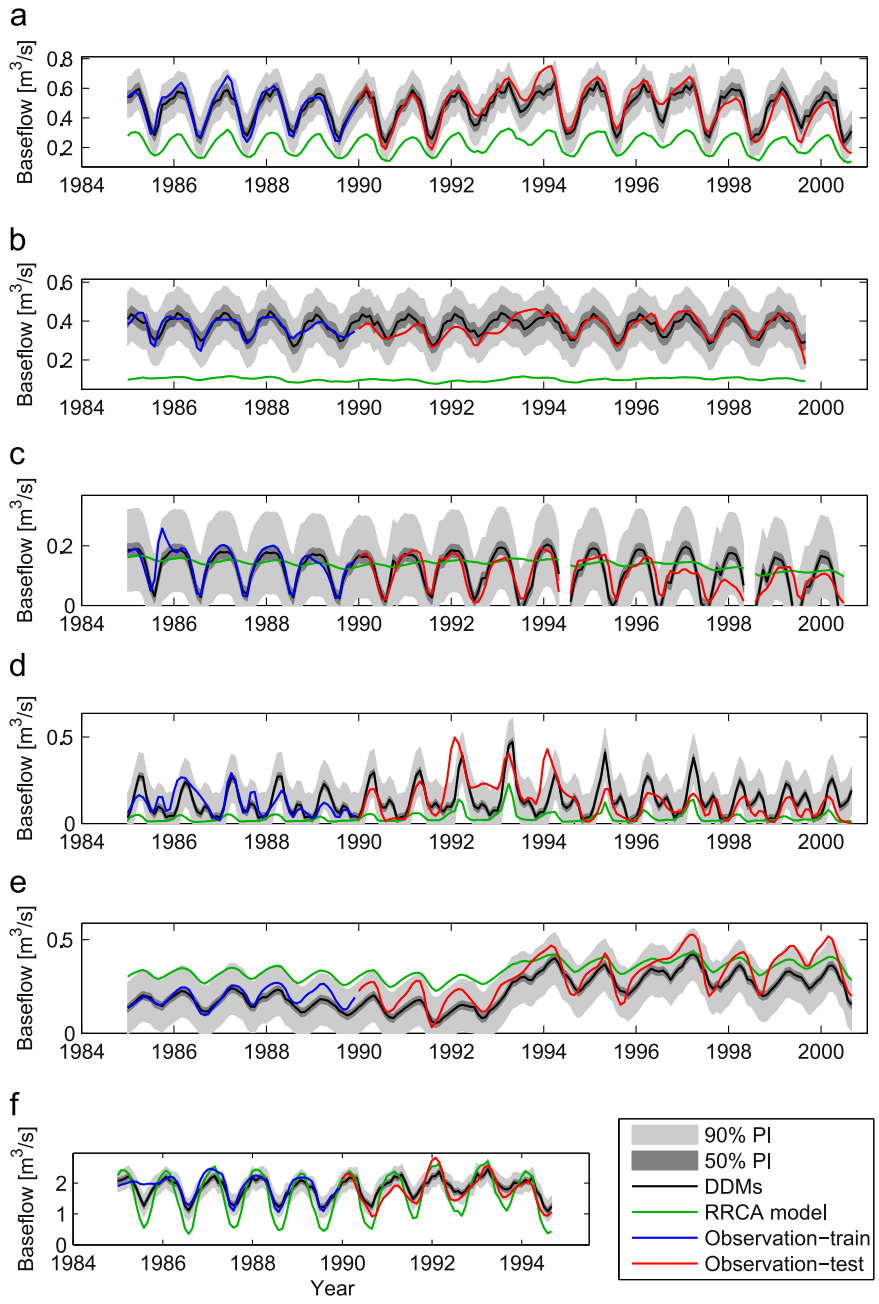

Fig. 8. Observed and RRCA model simulated baseflow, along with DDM-UQ(SVR) simulated mean and 50% and 90% prediction intervals of baseflow at six gaging stations. The station locations are plotted in Fig. 2 and their ID numbers and names are listed in Table 2.

Table 2
Gaging stations referred to in Figs. 2, 7 and 8.

	USGS ID	Gaging station name
(a)	06837300	Red willow Creek above Hugh Butler Lake
(b)	06852000	Elm Creek at Amboy
(c)	06823500	Buffalo Creek near Haigler
(d)	06821500	Arikaree River at haigler
(e)	06844210	Turkey Creek at Edison
(f)	06824500	Republican River at Benkelman

updated results reproduced these peaks missed by the RRCA model. However, the accuracy of the second peak correction is mixed in terms of timing and magnitude. Utilization of more information such as catchment scale and infiltration rate could further improve the performance. Overall the mean prediction given by both DDM-UQ based on RF and SVR reproduced intra-annual variability of baseflow to a satisfactory degree. Nevertheless, forecast of long-term inter-annual variability is challenging, as illustrated in Figs. 7(e) and 8(e). The RRCA model simulation systematically over-predicted baseflow, yet the bias decreases as there was an increase in baseflow in 1992–1994 due to unknown reasons. For this gaging station, it seems that the input data of DDMs did not capture the reasons leading to the change in baseflow pattern. Admittedly, this highlights one drawback of DDMs, namely that their effectiveness depends on the scope of data.

Besides the mean baseflow prediction, Figs. 7 and 8 also display the 90% and 50% prediction intervals generated by DDM-UQ(QRF) and DDM-UQ(SVR). Notably for DDM-UQ(SVR), the 90% PIs are much wider than the 50% PIs, especially in (c–e). The main reason is that the three gaging stations belong to cluster 3, for which the Cauchy distribution was used to fit the random error. The Cauchy

distribution has heavy tails, and thus leads to wide 90% PIs; see Fig. 6(a) for a plot of the Cauchy distribution. This also explains why DDM-UQ(SVR) gives wider 90% PIs than DDM-UQ(QRF) for gaging stations (c) and (e). Adopting a non-parametric distribution estimation method to work with SVR and/or using more clusters may alleviate the overestimation of uncertainty by DDM-UQ(SVR), however at the price of losing implementation simplicity. Furthermore, the width of PIs given by DDM-UQ(SVR) also depends on the goodness-of-fit of SVR model to historical residuals. Incorporating more input variables that help to explain baseflow variability might yield narrower PIs and thus more precise baseflow prediction.

While the DDM-UQ(SVR) approach provides PIs of the same width for data in the same cluster, DDM-UQ(QRF) gives PIs that vary pointwise. In general, those testing data points that are similar to historical training data would have narrower PIs thus smaller predictive uncertainty, while those significantly deviating from historical data would have higher predictive uncertainty. For gaging stations (b) and (c), DDM-UQ(QRF) provides more precise results in that the 90% PIs are narrower than 90% PIs given by DDM-UQ(SVR); however DDM-UQ(QRF) fell short at gaging station (e) for which the random forest did not explain well the baseflow variability. Finally, it is worth mentioning that the parametric distribution estimation method is generic and can be coupled with other statistical learning techniques beside SVR. In addition, this method is straightforward to apply at negligible computational cost. On the other hand, the non-parametric uncertainty quantification approach of QRF builds upon the ensemble structure of random forests. This approach comes at higher computational cost, however it is still much faster compared to other UQ approaches that require multiple evaluations of a time-consuming physically-based model.

6. Conclusions

The data-driven uncertainty quantification (DDM-UQ) framework has been applied to a real-world case study based on an existing precalibrated and well-documented regional groundwater flow model. The results show that data-driven models based on random forests and support vector machine regression algorithms effectively improved the baseflow prediction accuracy of the groundwater model. In addition, both the non-parametric distribution estimation method implemented in DDM-UQ(QRF) and the parametric method in DDM-UQ(SVR) provided robust baseflow prediction intervals that are consistent with validation data for a 10-year prediction horizon.

The DDM-UQ framework brings together the strength of physically-based groundwater models and inductive data-driven techniques, and it is in harmony with new trends in the field of hydrology towards increased data availability and promotion of environmental observatories. During the construction of bias-correcting DDMs, it was found that incorporating information that was not directly used by the groundwater model, such as precipitation rate, improves the DDMs performance at some gaging stations. In addition, by using information such as pumping, precipitation, evapotranspiration rates and groundwater heads in the inputs, the DDMs have the extrapolation power to make predictions under conditions different from the calibration period, which has not been addressed in the hydrological modeling literature (Reichert and Schuwirth, 2012). It is also worth noting that newly available data can be easily incorporated into the training dataset to retrain the DDMs and re-estimate the model residual distribution.

As a postprocessor approach, the presented DDM-UQ framework produces estimates of predictive uncertainty conditioned on the calibrated parameter set of the physically-based model. In addition, it focuses on the total model residual and does not distinguish different sources of uncertainty. Nevertheless, the satisfactory performance of this framework in the real-world case study suggests that DDM-UQ is a computationally efficient solution to assess predictive uncertainty when Monte Carlo based methods (Doherty and Christensen, 2011; Schoups and Vrugt, 2010; Tonkin and Doherty, 2009) become infeasible for computationally demanding models. The DDM-UQ framework is especially suitable when the simulation results of the precalibrated groundwater model systematically differ from (possibly new) observations, yet it is impossible or unaffordable to recalibrate or modify the model. Although it is illustrated using baseflow prediction in the application area of groundwater modeling, this generic framework is expected to be applicable to many geoscience models for which model structural error is not negligible. We have implemented the DDM-UQ framework in a MATLAB and R based toolbox. The toolbox, running instructions and sample datasets are available at http://wiki.cites.illinois.edu/wiki/display/mlgwm. Because of its model-independent and non-intrusive features, DDM-UQ can be conveniently coupled with existing codes for geoscience applications.

Acknowledgments

This work is supported by the National Science Foundation Hydrologic Science Program under Grant no. 0943627. The authors thank Dr. Yonas K. Demissie of Washington State University for sharing a part of the baseflow data used in the case study. The authors are grateful for the thoughtful review and suggestions by the two anonymous reviewers.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ecoenv.2015.06.031

References

Abebe, A.J., Price, R.K., 2003. Managing uncertainty in hydrological models using complementary models. Hydrol. Sci. J. 48 (5), 679–692.
Asefa, T., Kemblowski, M., Urroz, G., McKee, M., 2005. Support vector machines (SVMs) for monitoring network design. Ground Water 43 (3), 413–422.
Bates, Bryson C., Campbell, Edward P., 2001. A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resour. Res. 37 (4), 937–947.
Box, George E.P., Cox, David R., 1964. An analysis of transformations. J. R. Stat. Soc., Ser. B 26 (2), 211–252.
Breiman, Leo, 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Breiman, Leo, Friedman, Jerome, Stone, Charles J., Olshen, Richard A., 1984. Classification and Regression Trees. CRC Press, Boca Raton, Florida.
Caers, Jef, 2011. Modeling Uncertainty in the Earth Sciences. John Wiley & Sons, Ltd., Chichester, UK.
Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat. Hazards Earth Syst. Sci. 13 (11), 2815–2831.
Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2 (3), 27.
Cherkassky, V., Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw. 17 (1), 113–126.
Cloke, Hannah L., Pappenberger, Florian, 2008. Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. Meteorol. Appl. 15 (1), 181–197.
Daly, Christopher, Halbleib, Michael, Smith, Joseph I., Gibson, Wayne P., Doggett, Matthew K., Taylor, George H., Curtis, Jan, Pasteris, Phillip P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol. 28 (15), 2031–2064.
Delleur, Jacques W., Tao, P.C., Kavvas, M.L., 1976. An evaluation of the practicality

and complexity of some rainfall and runoff time series models. Water Resour. Res. 12 (5), 953–970.
Demissie, Yonas K., Valocchi, Albert J., Minsker, Barbara S., Bailey, Barbara A., 2009. Integrating a calibrated groundwater flow model with error-correcting data-driven models to improve predictions. J. Hydrol. 364 (3–4), 257–271.
Dogulu, N., López López, P., Solomatine, D.P., Weerts, A.H., Shrestha, D.L., 2014. Estimation of predictive hydrologic uncertainty using quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrol. Earth Syst. Sci. Discuss. 11 (9), 10179–10233.
Doherty, J., Brebber, L., Whyte, P., 1994. PEST: Model-Independent Parameter Estimation. Watermark Computing, vol. 122. Corinda, Australia.
Doherty, J., Christensen, S., 2011. Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. Water Resour. Res. 47 (12).
Doherty, J., Welter, D., 2010. A short exploration of structural noise. Water Resour. Res. 46 (5).
Dou, C., Woldt, W., Dahab, M., Bogardi, I., 1997. Transient ground-water flow simulation using a fuzzy set approach. Groundwater 35 (2), 205–215.
Durbin, J., Watson, G.S., 1971. Testing for serial correlation in least squares regression. III. Biometrika 58 (1), 1–19.
Erdal, D., Neuweiler, I., Huisman, J.A., 2012. Estimating effective model parameters for heterogeneous unsaturated flow using error models for bias correction. Water Resour. Res. 48 (6).
Evin, Guillaume, Thyer, Mark, Kavetski, Dmitri, McInerney, David, Kuczera, George, 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resour. Res. 50 (3), 2350–2375.
Galelli, S., Castelletti, A., 2013. Tree-based iterative input variable selection for hydrological modeling. Water Resour. Res. 49 (7), 4295–4310.
Goswami, M., O'Connor, K.M., Bhattarai, K.P., Shamseldin, A.Y., et al., 2005. Assessing the performance of eight real-time updating models and procedures for the Brosna River. Hydrol. Earth Syst. Sci. 9 (4), 394–411.
Gupta, Hoshin Vijai, Sorooshian, Soroosh, Yapo, Patrice Ogou, 1999. Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. J. Hydrol. Eng. 4 (2), 135–143.
Gusyev, M.A., Haitjema, H.M., Carlson, C.P., Gonzalez, M.A., 2013. Use of nested flow models and interpolation techniques for science-based management of the Sheyenne National Grassland, North Dakota, USA. Groundwater 51 (3), 414–420.
Harbaugh, Arlen W., Banta, Edward R., Hill, Mary C., McDonald, Michael G., 2000. MODFLOW-2000, the US Geological Survey Modular Ground-Water Model: User Guide to Modularization Concepts and the Ground-Water Flow Process. US Geological Survey Reston, VA, USA.
Hastie, Trevor, Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning. Springer, New York, NY.
Hill, M.C., Tiedeman, C.R., 2007. Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty. Wiley-Interscience, Hoboken, NJ.
Honti, M., Stamm, C., Reichert, P., 2013. Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resour. Res. 49 (8), 4866–4884.
Hunt, Randall J., Welter, David E., 2010. Taking account of “unknown unknowns”. Groundwater 48 (4), 477.
Kanevski, M., Parkin, Roman, Pozdnukhov, Aleksey, Timonin, Vadim, Maignan, Michel, Demyanov, V., Canu, Stéphane, 2004. Environmental data mining and modeling based on machine learning algorithms and geostatistics. Environ. Model. Softw. 19 (9), 845–855.
Kennedy, Marc C., O'Hagan, Anthony, 2001. Bayesian calibration of computer models. J. R. Stat. Soc., Ser. B (Stat. Methodol.) 63 (3), 425–464.
Koenker, Roger, 2005. Quantile Regression. Cambridge University Press, New York, NY.
Kotz, Samuel, Kozubowski, Tomasz, Podgorski, Krzysztof, 2001. The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance. Springer, New York, NY.
Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5 (5), 89–97.
Kuczera, George, 1983. Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. Water Resour. Res. 19 (5), 1151–1162.
Ließ, Mareike, Glaser, Bruno, Huwe, Bernd, 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and random forest models. Geoderma 170, 70–79.
Lin, Chih-Jen, Weng, Ruby C., et al., 2004. Simple Probabilistic Predictions for Support Vector Regression. Technical Report, Department of Computer Science, National Taiwan University, Taipei.
Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: toward an integrated data assimilation framework. Water Resour. Res. 43 (7), 1–18.
López López, P., Verkade, J.S., Weerts, A.H., Solomatine, D.P., 2014. Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison. Hydrol. Earth Syst. Sci. Discuss. 11 (4), 3811–3855.
Lu, Dan, Ye, Ming, Meyer, Philip D., Curtis, Gary P., Shi, Xiaoqing, Niu, Xu-Feng, Yabusaki, Steve B., 2013. Effects of error covariance structure on estimation of model averaging weights and predictive performance. Water Resour. Res. 49 (9), 6029–6047.
McKusick, V., 2003. Final Report for the Special Master with Certificate of Adoption of RRCA Groundwater Model. Technical Report, State of Kansas vs. State of Nebraska and State of Colorado, in the Supreme Court of the United States.
Meinshausen, Nicolai, 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.
Pianosi, F., Raso, L., 2012. Dynamic modelling of predictive uncertainty by regression on absolute errors. Water Resour. Res. 48 (3).
Rasouli, Kabir, Hsieh, William W., Cannon, Alex J., 2012. Daily streamflow forecasting by machine learning methods with weather and climate inputs. J. Hydrol. 414, 284–293.
Reichert, P., Schuwirth, N., 2012. Linking statistical bias description to multiobjective model calibration. Water Resour. Res. 48 (9).
Schoups, Gerrit, Vrugt, Jasper A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res. 46 (10).
Shi, Xiaoqing, Ye, Ming, Curtis, Gary P., Geoffery, Miller L., Philip, Meyer D., Kohler, Matthias, Yabusaki, Steve, Wu, Jichun, 2014. Assessment of parametric uncertainty for surface complexation modeling of groundwater uranium reactive transport. Water Resour. Res.
Shrestha, Durga L., Solomatine, Dimitri P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Netw. 19 (2), 225–235.
Solomatine, D.P., Shrestha, D.L., 2009. A novel method to estimate model uncertainty using machine learning techniques. Water Resour. Res. 45 (1).
Vladimir, Svetnik, Liaw, Andy, Tong, Christopher, Christopher Culberson, J., Sheridan, Robert P., Feuston, Bradley P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43 (6), 1947–1958.
Szilagyi, Jozsef, 2001. Identifying cause of declining flows in the Republican River. J. Water Resour. Plan. Manag. 127 (4), 244–253.
Tiedeman, Claire R., Green, Christopher T., 2013. Effect of correlated observation error on parameters, predictions, and uncertainty. Water Resour. Res. 49 (10), 6339–6355.
Tonkin, M., Doherty, J., 2009. Calibration-constrained Monte Carlo analysis of highly parameterized models using subspace techniques. Water Resour. Res. 45 (12), W00B10.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.
Weerts, A.H., Winsemius, H.C., Verkade, J.S., 2011. Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). Hydrol. Earth Syst. Sci. 15 (1), 255–265.
Xu, Tianfang, Valocchi, Albert J., Choi, Jaesik, Amir, Eyal, 2014. Use of machine learning methods to reduce predictive error of groundwater models. Groundwater 52 (3), 448–460.