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.

© 2015 Elsevier Ltd. All rights reserved.

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/© 2015 Elsevier Ltd. All rights reserved.

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.

Download English Version:

https://daneshyari.com/en/article/506962

Download Persian Version:

https://daneshyari.com/article/506962

Daneshyari.com