Journal of Hydrology 557 (2018) 826–837

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

A surrogate-based sensitivity quantification and Bayesian inversion of a regional groundwater flow model

Mingjie Chen *, Azizallah Izady, Osman A. Abdalla, Mansoor Amerjeed

Water Research Center, Sultan Qaboos University, Muscat, Oman

ARTICLE INFO

Article history:
Received 16 November 2017
Accepted 30 December 2017
Available online 2 January 2018

Keywords:
MODFLOW
Bayesian
Inverse modeling
Surrogate
Bagging MARS

ABSTRACT

Bayesian inference using Markov Chain Monte Carlo (MCMC) provides an explicit framework for stochastic calibration of hydrogeologic models accounting for uncertainties; however, the MCMC sampling entails a large number of model calls, and could easily become computationally unwieldy if the high-fidelity hydrogeologic model simulation is time consuming. This study proposes a surrogate-based Bayesian framework to address this notorious issue, and illustrates the methodology by inverse modeling a regional MODFLOW model. The high-fidelity groundwater model is approximated by a fast statistical model using Bagging Multivariate Adaptive Regression Spline (BMARS) algorithm, and hence the MCMC sampling can be efficiently performed. In this study, the MODFLOW model is developed to simulate the groundwater flow in an arid region of Oman consisting of mountain-coast aquifers, and used to run representative simulations to generate training dataset for BMARS model construction. A BMARS-based Sobol' method is also employed to efficiently calculate input parameter sensitivities, which are used to evaluate and rank their importance for the groundwater flow model system. According to sensitivity analysis, insensitive parameters are screened out of Bayesian inversion of the MODFLOW model, further saving computing efforts. The posterior probability distribution of input parameters is efficiently inferred from the prescribed prior distribution using observed head data, demonstrating that the presented BMARS-based Bayesian framework is an efficient tool to reduce parameter uncertainties of a groundwater system.

1. Introduction

Numerical simulation of a groundwater flow system requires knowledge of hydrologic characteristics of the aquifers, such as hydraulic conductivity and specific storage of hydrogeologic units. Complete knowledge of these properties are usually very hard to obtain in the subsurface, and the incurred uncertainty will reduce the predictive capability of the groundwater model and impair the reliability of groundwater resources management based on model predictions. To improve the model accuracy and reduce uncertainties, many inverse modeling methods have been developed to estimate hydrologic parameters by matching modeled results to corresponding field measurements (e.g., Carrera, 1988; Dai and Samper, 2004; Carrera et al., 2005; Hill and Tiedeman, 2007; Tonkin and Doherty, 2009; Sadeghi-Tabas et al., 2016; Shang et al., 2016). Whereas, applications of these methods are usually limited by considerable number of forward model simulations required during inverse modeling (Wagner, 1995; Jordan et al., 2015).

Owing to the insufficient knowledge of the aquifers, considerable uncertainties might be introduced into groundwater modeling systems yielding ill-posed inverse problems with non-unique solutions (de Marsily et al., 1999). To cope with this issue, Bayesian inference has been studied and applied extensively as an effective inverse framework yielding a parameter posterior probability distribution instead of a deterministic solution from its prior distribution (Oliver et al., 1997; Kavetski et al., 2006; Fu and Gomez-Hernandez, 2009; Tarantola, 2004). The analytical solutions of posterior distribution are usually very difficult to obtain in Bayesian inversion (Khaleghi et al., 2013). Instead, Markov Chain Monte Carlo (MCMC) methods provide a practical tool to draw samples from the specified target distribution to numerically approximate posterior distributions for groundwater model parameters (Hassan et al., 2009; Efendiev et al., 2011; Laloy et al., 2013). However, when dealing with large-scale high-fidelity groundwater models, MCMC methods may become computationally burdensome because of the considerable number of uncertain parameters to be estimated and many times of repeated forward model calls

* Corresponding author at: Sultan Qaboos University, P.O. Box 17, Postal Code 123, Al-Khoud, Oman.

E-mail addresses: cmj1014@gmail.com, mingjie@squ.edu.om (M. Chen).

https://doi.org/10.1016/j.jhydrol.2017.12.071
0022-1694/

for posterior distribution to reach convergence (Smith and Marshall, 2008). There are primarily four promising ways to address this issue include (1) parallel calculation and grid computing (e.g., Kerrou et al., 2010), (2) improvement of MCMC sampling algorithms (e.g., Vrugt et al., 2009; Rajabi and Ataie-Ashtiani, 2016), (3) reduction of inverse model dimension, that is, excluding insensitive physical properties and only keep key parameters in Bayesian inversion (e.g., Kerrou and Renard, 2010; Chen et al., 2014), and (4) development of surrogate statistic model to substitute high-fidelity physical model in MCMC sampling (e.g., Zeng et al., 2012). In this study, we will combine the latter two approaches to develop a surrogate-based Bayesian framework, in which the primary behavior of a regional groundwater flow model is replicated with a cheaper surrogate model using fewer but sensitive parameters for application in MCMC sampling.

Surrogate models have been developed to approximate physical models using a variety of statistical algorithms, such as kriging, radial basis functions (RBFs), polynomials, support vector machines (SVMs), artificial neural networks (ANNs), sparse grid interpolation (e.g., Simpson and Mistree, 2001; Regis and Shoemaker, 2007; Fen et al., 2009; Zhang et al., 2009; Behzadian et al., 2009; Zeng et al., 2012). Razavi et al. (2012) reviewed 48 applications of surrogate models in hydrology covering most of these approximation techniques. Chen et al. (2013, 2015) introduced multivariate adaptive regression spline (MARS) (Friedman, 1991) into a simulation-optimization methodology for underground gas and geothermal reservoir modeling. Dai et al. (2014) and Keating et al. (2016) further applied MARS surrogate model to a risk assessment of carbon geological storage. MARS algorithm, which adaptively develops local models in local regions for flexible regression modeling of high dimensional data, shows super performance in handling high-dimensional data, and its stability can be enhanced by bootstrap aggregating, namely Bagging (Bühlmann and Yu, 2002). Bagging is a kind of smoothing operation which could be used to improve MARS model prediction (Bühlmann, 2003). In our recent study, Bagging MARS (BMARS) surrogate model was coupled with an optimizer to numerically calibrate a complicated MODFLOW for simulation of groundwater flow in mountain hardrock and unconfined alluvial aquifers in northwest Oman (Chen et al., 2017). In this study, we extend from deterministic optimization to integrate BMARS modeling approach to MCMC sampling process to achieve an efficient Bayesian inversion of a regional-scale groundwater flow model.

Sensitivity measures input parameters' relative contributions to the uncertainty of an output. Sensitivity Analysis (SA) is very useful in model parameter importance ranking, uncertainty quantification and management (Wei et al., 2013). Local SA methods have been widely used in groundwater modeling studies (e.g., Sanz and Voss, 2006). They had been implemented in standard calibration softwares UCODE (Poeter and Hill, 1999) and PEST (Doherty, 2005) and become readily accessible for users. Compared to traditional local SA, the emerging global SA methods take into account the entire uncertain range of input parameters, as well as their interactions between them (Oladyshkin et al., 2012). Since input parameters (e.g., hydraulic conductivity) of groundwater models are usually varying across several orders of magnitude, the global SA is best suitable and essential for comprehensive sensitivity evaluations. However, demanding computational cost has prevented the global SA from widespread use as local SA methods. To tackle this challenge, Polynomial Chaos Expansions (PCEs) have been extensively studied to efficiently calculate the global sensitivity indices (e.g., Sudret, 2008; Sochala and Le Maître, 2013; Rajabi et al., 2015). In this study, the developed BMARS model is introduced into Sobol' method (Sobol', 1993) for highly efficient global SA of MODFLOW model parameters. According to importance ranking, the insensitive parameters are thus identified and excluded from Bayesian inversion. Without losing key physical characteristics of the original model, the dimension of inverse modeling is reduced and computing efforts are focused on the most important parameters.

The presentation of the study is organized as follows. In Section 2, BMARS based Bayesian inverse modeling framework methodology is illustrated and its major components including physical MODFLOW model, BMARS model, Bayesian inference, Sobol' sensitivity are described. Section 3 gives details on the MODFLOW model for the Al-Fara area in Oman. Results and subsequent analysis are presented in Section 4, and concluding remarks are drawn in Section 5.

2. Methodology

The purpose of this study is to develop a surrogate-based Bayesian inversion of a MODFLOW model system using BMARS algorithm, entailing high-dimensional parameter space sampling, groundwater flow model development, BMARS model construction, Sobol' sensitivity analysis, and Bayesian inference using BMARS model. In this section, the methodology framework is illustrated firstly, followed by detailed description of core components.

2.1. BMARS-based Bayesian inversion framework

The proposed inversion framework proceeds as follows (Fig. 1):

1. M input parameters are chosen with uncertainties defined by their distributions and ranges, which are representative of prior knowledge of the groundwater flow system. In this study, the parameters are hydrogeologic characteristics of the aquifers, including hydraulic conductivity, specific storage and specific yield.
2. N input M-vectors are sampled from their prior probability density functions (PDF) using Latin-Hypercube (L-H) method (McKay et al., 1979). These N realizations are used to generate input files to run N MODFLOW model simulations and their corresponding responses (in our case below, heads at observation wells) are obtained.
3. The N input vector-response pairs (shaded in Fig. 1) are used to train and cross-validate BMARS surrogate model. Leave-one-out cross validation (LOOCV) approach is used for surrogate model validation (Picard and Cook, 1984).
4. The importance of the input parameters of the MODFLOW model is ranked according to Sobol' total order sensitivity indices (Sobol' 1993), which is efficiently computed using cheap BMARS model. Only top sensitive parameters are considered influential to the hydrogeologic system and included in Bayesian inversion.
5. The BMARS model is utilized in place of the MODFLOW model within a Bayesian inversion scheme, from which posterior distributions of the sensitive parameters are inferred by comparisons with observed head data. The posterior PDFs represent a subset of the prior data that yield indeterminate solutions with highest probabilities.

The BMARS-driven Bayesian inference framework is programmed in Python script by coupling MODFLOW model with various algorithms mentioned above, i.e., L-H sampling, BMARS approximation, LOOCV, Sobol' method, and Bayesian inference. The suite of MODFLOW simulations are distributed to and executed on high performance computing facility at Sultan Qaboos University in Oman.