Journal of Hydrology 535 (2016) 120–134

Contents lists available at ScienceDirect

Journal of Hydrology

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

Evaluating two sparse grid surrogates and two adaptation criteria for groundwater Bayesian uncertainty quantification

Xiankui Zeng [a,b], Ming Ye [b], John Burkardt [b], Jichun Wu [a,*], Dong Wang [a], Xiaobin Zhu [a]

[a] Key Laboratory of Surficial Geochemistry, Ministry of Education, School of Earth Sciences and Engineering, Nanjing University, Nanjing, China
[b] Department of Scientific Computing, Florida State University, Tallahassee, FL, USA

ARTICLE INFO

Article history:
Received 28 October 2015
Received in revised form 31 December 2015
Accepted 22 January 2016
Available online 1 February 2016
This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Felipe de Barros, Associate Editor

Keywords:
Log-likelihood surrogate
State-variable surrogate
Adaptive sparse grid
Markov chain Monte Carlo
Bayesian inference
Groundwater modeling

SUMMARY

Sparse grid (SG) stochastic collocation methods have been recently used to build accurate but cheap-to-run surrogates for groundwater models to reduce the computational burden of Bayesian uncertainty analysis. The surrogates can be built for either a log-likelihood function or state variables such as hydraulic head and solute concentration. Using a synthetic groundwater flow model, this study evaluates the log-likelihood and head surrogates in terms of the computational cost of building them, the accuracy of the surrogates, and the accuracy of the distributions of model parameters and predictions obtained using the surrogates. The head surrogates outperform the log-likelihood surrogates for the following four reasons: (1) the shape of the head response surface is smoother than that of the log-likelihood response surface in parameter space, (2) the head variation is smaller than the log-likelihood variation in parameter space, (3) the interpolation error of the head surrogates does not accumulate to be larger than the interpolation error of the log-likelihood surrogates, and (4) the model simulations needed for building one head surrogate can be recycled for building others. For both log-likelihood and head surrogates, adaptive sparse grids are built using two indicators: absolute error and relative error. The adaptive head surrogates are insensitive to the error indicators, because the ratio between the two indicators is hydraulic head, which has small variation in the parameter space. The adaptive log-likelihood surrogates based on the relative error indicators outperform those based on the absolute error indicators, because adaptation based on the relative error indicator puts more sparse-grid nodes in the areas in the parameter space where the log-likelihood is high. While our numerical study suggests building state-variable surrogates and using the relative error indicator for building log-likelihood surrogates, selecting appropriate type of surrogates and error indicators depends on the shapes of response surfaces. The shapes should be approximated and examined before building sparse grid surrogates.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Uncertainty analysis has become a common practice in groundwater modeling in the last several decades for evaluating model predictive performance, improving model structures, and supporting science-informed decision-making (Gupta et al., 2012; Matott et al., 2009; Tartakovsky, 2013). Among various methods developed for uncertainty analysis, Bayesian approaches are one of the most popular methods. However, in comparison with other methods of uncertainty analysis that are computationally frugal (Hill et al., 2015), Bayesian approaches are computationally expensive, because they always involve Markov chain Monte Carlo (MCMC) simulations, in which tens to hundreds of thousands of model executions are necessary for estimating the probability distributions of model parameters and predictions. To alleviate the computing burden, one solution is to replace a model by its surrogate that is sufficiently accurate but computationally cheap, and a review article of surrogate modeling is given by Razavi et al. (2012). Among various methods of building surrogates, the sparse grid (SG) stochastic collocation methods are used in this study. Although the SG methods have become popular, using them for Bayesian uncertainty quantification has been reported only in a limited number of groundwater studies (Zeng et al., 2012; Zhang et al., 2013, 2015). In other uses of SG methods (e.g., Lin and Tartakovsky, 2009, 2010; Lin et al., 2010; Shi and Yang, 2009; Zhang et al., 2010; Dai and Ye, 2015), SG methods are used to estimate the distributions or moments (e.g., mean and covariance) of groundwater state variables (e.g., hydraulic head and solute concentration). These studies assumed known parameter

* Corresponding author.
E-mail address: jcwu@nju.edu.cn (J. Wu).

http://dx.doi.org/10.1016/j.jhydrol.2016.01.058
0022-1694/© 2016 Elsevier B.V. All rights reserved.

distributions, and did not estimate the distributions using Bayesian approaches.

This study investigates an important problem for SG-based Bayesian uncertainty quantification, i.e., how to evaluate the likelihood function used in Bayesian inference. Consider a Bayesian inference problem for a nonlinear model, f, used to simulate state variables (e.g., hydraulic head and solute concentration),

$$\boldsymbol{d} = f(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{d}$ is a vector dataset of state variable, $\boldsymbol{\theta}$ is a vector of model parameters, and $\boldsymbol{\varepsilon}$ is a vector of residuals that may include errors in data, model parameters, and model structures. The goal of Bayesian inference is to estimate the posterior distributions, $p(\boldsymbol{\theta}|\boldsymbol{d})$, of model parameters, $\boldsymbol{\theta}$, given data, $\boldsymbol{d}$, using Bayes' theorem (Box and Tiao, 1992)

$$p(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{L(\boldsymbol{\theta}|\boldsymbol{d})p(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}|\boldsymbol{d})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{2}$$

where $p(\boldsymbol{\theta})$ is the prior distribution and $L(\boldsymbol{\theta}|\boldsymbol{d})$ is the likelihood function to measure goodness-of-fit between model simulations, $f(\boldsymbol{\theta})$, and data, $\boldsymbol{d}$. The prior distribution can be specified using data from previous studies or expert judgment. When prior information is lacking, a common practice is to assume uniform distributions with relatively large parameter ranges so that the prior distributions do not affect the estimation of posterior distributions. Defining a likelihood function appropriate to a specific problem is still an open question, and it has been shown that the likelihood function has substantial impacts on the results of Bayesian inference (Evin et al., 2014; Lu et al., 2013; Schoups and Vrugt, 2010; Shi et al., 2014; Smith et al., 2010). While SG methods can work with various likelihood functions (Zhang et al., 2013), this study uses the commonly used Gaussian likelihood function,

$$L(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{d} - f(\boldsymbol{\theta}))^T \Sigma^{-1}(\boldsymbol{d} - f(\boldsymbol{\theta}))\right), \tag{3}$$

where N is the number of data (i.e., the dimension of $\boldsymbol{d}$), and $\sum$ is the covariance matrix of the residuals, $\boldsymbol{\varepsilon}$. Because analytical expressions for $p(\boldsymbol{\theta}|\boldsymbol{d})$ are unavailable for nonlinear models, Markov chain Monte Carlo (MCMC) methods are often used for estimating $p(\boldsymbol{\theta}|\boldsymbol{d})$. In MCMC, a large number (tens to hundreds of thousands) of parameter samples are drawn; for each sample, the nonlinear function, $f(\boldsymbol{\theta})$, and the likelihood function, $L(\boldsymbol{\theta}|\boldsymbol{d})$, are evaluated. If the nonlinear function is computationally expensive, the computational cost for the Bayesian inference may be unaffordable. This necessitates the use of SG surrogates.

In SG applications for Bayesian inference, two kinds of SG surrogates have been used. One is for the logarithm of the likelihood function used to directly replace $L(\boldsymbol{\theta}|\boldsymbol{d})$ during Bayesian inference; the other is for the state variables used to replace $f(\boldsymbol{\theta})$ for evaluating the likelihood. Building the state variable surrogates is common in the literature of not only SG collocation (Ma and Zabaras, 2009b; Zeng et al., 2012; Zhang et al., 2015) but also other stochastic collocation methods of Bayesian inference (Marzouk et al., 2007; Marzouk and Xiu, 2009; Liao and Zhang, 2013; Laloy et al., 2013). While building log-likelihood surrogates is less common (Zhang et al., 2013), it is theoretically superior to building state-variable surrogates for two reasons. First, only one log-likelihood surrogate is needed regardless of the number of observations, whereas one state-variable surrogate is needed for each observation. When the number of observations is large, the computational cost of building multiple state-variable surrogates can be significantly higher than that of building a single log-likelihood surrogate. In addition, each state-variable surrogate has its SG interpolation error, and the error may accumulate and become large when the surrogates are used for evaluating the likelihood function. However, building the log-likelihood surrogates has its own disadvantages as discussed in the numerical example below. It is therefore necessary to evaluate the two kinds of surrogates to determine which kind of surrogate is more appropriate for Bayesian inference.

To the best of our knowledge, there has been no reported reference on comparing the state-variable surrogates and the log-likelihood surrogates. The study of Petvipusit et al. (2014) is the only reference related to the comparison that we are aware of. The study compared two surrogates used for optimization of CO_2 sequestration. One surrogate was built for a break-even tax credit function, and the other for the moments (i.e., mean and variance) of the function. The comparative study of Petvipusit et al. (2014) showed that building the moment surrogate is computationally more efficient than building the function surrogate. However, their study is irrelevant to comparison between the log-likelihood and state-variable surrogates. The two kinds of surrogates are compared in this study in terms of accuracy and efficiency. The accuracy is evaluated by comparing the posterior distributions obtained using the two kinds of surrogates with the reference distributions obtained using the original model without any surrogates. The computational efficiency is evaluated by directly comparing the number of model executions needed for building the log-likelihood and state-variable surrogates. The comparative evaluation is done by conducting a numerical study for a synthetic groundwater flow model. The conclusions drawn from the synthetic study through the quantitative and comprehensive evaluations are expected to be applicable to other groundwater studies, given that the complexity of the synthetic model is representative for groundwater modeling.

This study also addresses another important issue for building adaptive SG, i.e., whether absolute or relative error should be used as the indicator for adaptation. Building adaptive SG is common for saving computational cost by adding SG nodes only in the areas where SG interpolation error is larger than a user-specified tolerance value (Barthelmann et al., 2000; Klimke, 2006; Ma and Zabaras, 2009a; Pfluger, 2010; Zhang et al., 2013). The absolute error (difference between a model simulation and its surrogate) is the interpolation error itself, and has been used widely (Ma and Zabaras, 2009a, 2009b; Stoyanov, 2013a, 2013b; Webster et al., 2014; Zeng et al., 2012; Zhang et al., 2013), because it directly controls SG accuracy. However, it should be noted that having an accurate SG surrogate is insufficient to having an accurate Bayesian inference, i.e., obtaining accurate posterior parameter distributions. For example, adding adaptive SG points in low likelihood regions to reduce SG error is useless to Bayesian inference, because only parameter samples generated from high likelihood regions are accepted during MCMC simulation; this is demonstrated below using the numerical examples based on the synthetic groundwater model. The key question is where to add adaptive SG nodes in Bayesian inference, and this problem is resolved empirically in this study by using relative error, i.e., absolute error divided by the model simulation. We explore whether the relative error outperforms the absolute error by using both absolute and relative error indicators to build adaptive log-likelihood and state-variable surrogates. As discussed below in Section 4, the two error indicators lead to significantly different SG node locations when building the log-likelihood surrogates, but not the case when building the state-variable surrogates. As a result, the two error indicators have substantial impacts on the accuracy of estimating the posterior distributions of model parameters and predictions.

It should be noted that this study is focused on using SG for Bayesian uncertainty quantification; other uses of SG are beyond

Download English Version:

https://daneshyari.com/en/article/6410087

Download Persian Version:

https://daneshyari.com/article/6410087

Daneshyari.com