Computers & Geosciences 108 (2017) 21–32

Contents lists available at ScienceDirect

Computers & Geosciences

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

Research paper

Implicit sampling combined with reduced order modeling for the inversion of vadose zone hydrological data

Yaning Liu*, George Shu Heng Pau, Stefan Finsterle

Earth & Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

ARTICLE INFO

Keywords:
Inverse modeling
Implicit sampling
Vadose zone
TOUGH2
Polynomial chaos expansion
Sparse Bayesian learning

ABSTRACT

Bayesian inverse modeling techniques are computationally expensive because many forward simulations are needed when sampling the posterior distribution of the parameters. In this paper, we combine the implicit sampling method and generalized polynomial chaos expansion (gPCE) to significantly reduce the computational cost of performing Bayesian inverse modeling. There are three steps in this approach: (1) find the maximizer of the likelihood function using deterministic approaches; (2) construct a gPCE-based surrogate model using the results from a limited number of forward simulations; and (3) efficiently sample the posterior distribution of the parameters using implicit sampling method. The cost of constructing the gPCE-based surrogate model is further decreased by using sparse Bayesian learning to reduce the number of gPCE coefficients that have to be determined. We demonstrate the approach for a synthetic ponded infiltration experiment simulated with TOUGH2. The surrogate model is highly accurate with mean relative error that is <0.035% in predicting saturation and <0.25% in predicting the likelihood function. The posterior distribution of the parameters obtained using our proposed technique is nearly indistinguishable from the results obtained from either an implicit sampling method or a Markov chain Monte Carlo method utilizing the full model.

1. Introduction

Hydrological models are crucial to the understanding and description of water cycles. Hydrological model parameters, such as site-specific material properties and process-related parameters, as well as boundary conditions and site geometry play a major role in the model's ability to predict the hydrological states. These parameters can be large scale, highly uncertain and difficult to measure (Abubakar et al., 2009; Liu and Gupta, 2007). Thus, inverse modeling is typically performed to infer the model parameter values based on the sparse observations of some observables, by matching the numerical model (the forward model) to measured data at discrete spatial and temporal points.

Inverse modeling techniques, in general, fall into two categories: deterministic and probabilistic inversions. The deterministic approach aims to find a single set of parameter values that represent the "best fit" given the observations and a criterion that measures the closeness between the model response and the observations. Therefore, the essence of a deterministic inversion is the minimization of the objective function, which measures the difference between the model and the observed data. Commonly used objective functions are derived based on the assumptions that the model is correct and the errors in the measured data are normally or exponentially distributed, resulting in the maximum likelihood and L_1 estimators, respectively (Tarantola, 2004). A broader selection of objective functions can be found in Finsterle and Najita (1998) and Schoups and Vrugt (2010). For detailed theory and computational methods for deterministic inversions, the readers are referred to Vogel (2002), Neto and da Silva Neto (2012), and Ramm (2005). Deterministic inversion usually requires some form of regularization since inverse problems are often ill-posed (Kabanikhin, 2008). The well-known Tikhonov regularization, for instance, is typically used in least-square problems. In addition, estimation of the inversion uncertainty within a deterministic framework requires strong assumptions about the error structure of the observation and parameters, and about the linearity of the forward problem (Carrera and Neuman, 1986).

Accurate characterization of the inversion uncertainty is desirable in many applications as it provides an informed representation of the distribution of parametric uncertainty that can be propagated through the forward model, as is often done in uncertainty quantification (UQ) (Mondal et al., 2010; Liu et al., 2015a). With probabilistic inversion methods, the inversion result is presented in the form of joint probability distributions instead of a single set of parameter values. A common way to achieve this is through the Bayesian probability theory, which relates the parameter posterior distribution conditioned on the observations to the product of the prior distribution

* Corresponding author.
E-mail addresses: yaningliu@lbl.gov (Y. Liu), gpau@lbl.gov (G.S.H. Pau), safinsterle@lbl.gov (S. Finsterle).

http://dx.doi.org/10.1016/j.cageo.2017.04.001
Received 28 May 2016; Received in revised form 18 February 2017; Accepted 3 April 2017
Available online 06 April 2017

and likelihood function. Since the resulting posterior densities can be complex, they are typically discretely represented by a set of samples obtained through sampling methods. With the Markov chain Monte Carlo (MCMC) method (e.g., Andrieu et al., 2003), these samples are generated by an acceptance-rejection approach. However, evaluating the acceptance-rejection criteria in MCMC, which involves a forward simulation for each proposed sample, can be extremely inefficient if the fraction of the rejected samples is large. The "burn-in" period for a MCMC sampling can also be long, resulting in a large number of samples being wasted. In addition, the significant (high-probability) region of the posterior distribution may be small, which further decreases the efficiency of MCMC sampling. Recent advances in MCMC methods address some of these difficulties. For example, Vrugt et al. (2009) proposed a differential evolution adaptive Metropolis scheme (DREAM) that explores the optimized proposal distribution in parallel and extends its applicability to estimating multi-modal posterior densities; Goodman and Weare (2010) suggest a family of multi-particle MCMC samplers with an affine invariance property that can offer significantly improved performance over standard single-particle methods; and Martin et al. (2012) present a stochastic Newton MCMC method by constructing a proposal density based on local Gaussian approximation that is especially efficient for large-scale inversions.

As a powerful alternative, particle filters (PF) (Liu and Chen, 1998; Arulampalam et al., 2002) are sequential Monte Carlo methods used in data assimilation to update the discrete representation of posteriors of the state variables in the form of particles (samples) with associated weights as new observable data become available. PF are "embarrassingly parallel" and the computational complexity is independent of the dimensionality of the system. However, PF can suffer from "particle collapse" (sample impoverishment), where only a small fraction of the particles have non-negligible weights. A large number of particles are thus needed for a meaningful approximation of the posterior.

Recently, a variant of PF, named implicit particle filter (IPF) was developed by Chorin and Tu (2009) and Chorin et al. (2010, 2013) as a remedy for sample impoverishment. IPF searches samples in the high probability region of the posterior by connecting the target particles with a reference distribution through a mapping of one's choice. The quality of the particles can therefore be significantly enhanced, as the proportion of the particles with non-negligible weights increases, and the overall number of particles needed is reduced. A number of improvements and applications of IPF have been made since its inception. Morzfeld et al. (2012) proposed a random map procedure and applied it to assimilating data for a stochastic Lorenz attractor; Morzfeld and Chorin (2012) applied IPF to geomagnetic data assimilation with partial noise; Atkins et al. (2013) established the connection of IPF with variational data assimilation. More recently, Morzfeld et al. (2015) implemented IPF for estimating parameters for subsurface flow modeled by Darcy's law; they observed a faster convergence compared to Metropolis MCMC. We henceforth use the term "implicit sampling" adopted therein in the case of inverse modeling independent of time, while "implicit particle filter" is used for the data assimilation processes where the systems are dynamic and the parameters are time-dependent.

Implicit sampling (IS) involves the mapping from samples drawn from a different distribution, called the reference distribution, to the target particles, which may require solving a nonlinear equation for each given sample, depending on the type of the mapping chosen, and thus requires a large number of forward simulations. In view of this, we propose using a reduced order model (ROM) that serves as a surrogate for the forward model (see Razavi et al. (2012), Pau et al. (2014), Liu et al. (2016a, 2016b)), among others). The ROM is constructed with an initial set of forward simulations (training set), and subsequently substitutes the forward model. We develop our ROM based on the widely used generalized polynomial chaos expansion (gPCE) (Xiu and Karniadakis, 2002). The construction of gPCE requires determining the coefficients of the expansion terms once the type of the polynomial basis and expansion order are selected. However, the optimal polynomial expansion order is not a priori known. A low-order expansion may not accurately represent the response surface. On the other hand, a high-order expansion leads to exponentially large number of expansion terms. As a consequence, the number of forward simulations needed to estimate the gPCE coefficients also increases exponentially. The error associated with the overall gPCE can also increase since the estimation errors associated with the estimated gPCE coefficients may increase substantially as the expansion order increases.

However, the gPCE coefficients can be sparse, i.e., only a small number of the coefficients are non-zero. The sparsity is due to the following reasons: higher-order parameter interactions may not exist (Rabitz et al., 1999); the model response is smooth and so are the higher-order derivatives, leading to a fast decrease in the magnitudes of the coefficients as the polynomial orders increase; and the model response is, by nature, the superposition of only a sparse subset of all the polynomial bases up to a given order. Since the set of non-negligible gPCE coefficients is not a priori known, we can pose the problem as a sparse Bayesian learning (SBL) problem (Sargsyan et al., 2014) (also known as relevance vector machine and Bayesian compressive sensing (Tipping, 2001; Tipping and Faul, 2003)), where the model outputs are characterized by a hierarchical form of Gaussian likelihood and prior. Babacan et al. (2010) further demonstrated that using Laplace prior to model the sparsity improved the performance. With SBL, the sparsity is obtained by updating one gPCE coefficient at a time using a greedy algorithm that iteratively selects the most contributing coefficients until a prescribed stopping criterion is reached. The initial set of coefficients is set to be empty, and an efficient algorithm to update the coefficient set, either by including a new one, revising the value of an existing coefficient, or deleting an existing one, is described in Tipping and Faul (2003). Sargsyan et al. (2014) empirically demonstrated that good estimates of the gPCE coefficients can be obtained if the number of training samples is about five times that of non-zero coefficients. Depending on the ratio of the number of non-zero coefficients to the number of coefficients required by a particular expansion order, the number of forward model simulations to construct the ROM can be greatly reduced.

In the present work, we implement implicit sampling in inverse modeling for a synthetic ponded infiltration experiment. The goal of the inversion is to determine the permeability distribution of the vadose zone based on saturation measurements. We also demonstrate that the performance of the inversion can be further enhanced by constructing and utilizing a reduced order model based on generalized polynomial chaos expansion and sparse Bayesian learning. The proposed inversion method is compared to a state-of-the-art Markov chain Monte Carlo simulator described in Goodman and Weare (2010).

The rest of this paper is structured as follows. In the next section, we introduce the vadose zone hydrological forward model, followed by a mathematical description of the implicit sampling method and details about the construction of the reduced order model using generalized polynomial chaos expansion and sparse Bayesian learning. In Section 4, the main results and discussion of the inversion of the hydrological model are presented. In the end, we conclude the paper with possible improvements for future work.

2. A hydrological inverse modeling problem

The hydrological problem used to demonstrate our approach is a synthetic field experiment shown in Fig. 1. An infiltration pond releases water into a heterogeneous but structured vadose zone whose water table is at the depth of 3 m. The saturation distribution is initially in gravity-capillary equilibrium, and the infiltration rate is controlled in order for the water level in the pond to stay at 2 cm for 1 day. After that, the experiment proceeds without infiltration for another day. Water saturation is measured at 36 monitoring points (circles in Fig. 1) initially and every 2 h from 34th hour till the end of day 2; the amount of water flowing out of the pond is also measured at the same times. The subsurface flow is modeled by Richards' equation (Richards, 1931) as implemented in the integral finite difference simulator TOUGH2