Advances in Water Resources 66 (2014) 8–18

Contents lists available at ScienceDirect

Advances in Water Resources

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

Comparison of Ensemble Kalman Filter groundwater-data assimilation methods based on stochastic moment equations and Monte Carlo simulation

M. Panzeri [a,*], M. Riva [a,b], A. Guadagnini [a,b], S.P. Neuman [b]

[a] Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milano, Italy
[b] Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, USA

A R T I C L E I N F O

Article history:
Received 30 October 2013
Received in revised form 30 January 2014
Accepted 31 January 2014
Available online 10 February 2014

Keywords:
Ensemble Kalman Filter
Random hydraulic conductivity field
Moment equations
Data assimilation
Transient groundwater flow
Filter inbreeding

A B S T R A C T

Traditional Ensemble Kalman Filter (EnKF) data assimilation requires computationally intensive Monte Carlo (MC) sampling, which suffers from filter inbreeding unless the number of simulations is large. Recently we proposed an alternative EnKF groundwater-data assimilation method that obviates the need for sampling and is free of inbreeding issues. In our new approach, theoretical ensemble moments are approximated directly by solving a system of corresponding stochastic groundwater flow equations. Like MC-based EnKF, our moment equations (ME) approach allows Bayesian updating of system states and parameters in real-time as new data become available. Here we compare the performances and accuracies of the two approaches on two-dimensional transient groundwater flow toward a well pumping water in a synthetic, randomly heterogeneous confined aquifer subject to prescribed head and flux boundary conditions.

1. Introduction

Kalman Filter (KF) is a well-known inverse technique used to assimilate incoming data into physical system models sequentially in real time. It was originally introduced by Kalman [1] to integrate data corrupted by white Gaussian noise in linear dynamic models the outputs of which include additive noise of a similar type. KF entails two steps: a forward modeling (or forecasting) step that propagates system states in time until new measurements become available, and an updating step that modifies system states optimally in real time on the basis of such measurements. Some modern versions of KF update system states (e.g., hydraulic heads) and parameters (e.g., hydraulic conductivities) jointly based on measurements of one or both variables (e.g., [2]).

Gelb [3] proposed an Extended Kalman Filter (EKF) to deal with nonlinear system models. EKF linearizes the model and propagates the first two statistical moments of target model variables in time. As such it is not suitable for strongly non-linear systems of the kind encountered in the context of groundwater flow or transport in any but mildly heterogeneous media. EKF further requires large amounts of computer storage which limits its use to relatively small-size problems. Evensen [4] and Burgers et al. [5] proposed to overcome these limitations through the use of Monte Carlo (MC) simulation. Their so-called Ensemble Kalman Filter (EnKF) approach utilizes sample mean values and covariances to perform the updating. The development of sensors and measuring devices capable of recording massive amounts of data in real time has made EnKF popular among hydrologists, climate modelers and petroleum reservoir engineers [6,7]; assimilating such massive data sets in batch rather than sequential mode, as is common with classical inverse frameworks such as Maximum Likelihood, would not be feasible. Applications of EnKF to groundwater and multiphase flow problems include the pioneering works of [8,9]; for more recent reviews see [6,7,10].

A crucial factor affecting EnKF is the size of the ensemble, i.e., the number (*NMC*) of MC simulations (sample size) employed for moment evaluation. Whereas to estimate mean and covariance accurately requires many simulations, working with large *NMC* tends to be computationally demanding. Chen and Zhang [11] showed that a few hundred *NMC* appear to provide accurate estimates of mean log-conductivity fields. They pointed out, however, that obtaining covariance estimates of comparable accuracy would require many more simulations, a task they had not carried through. Efforts to reduce the dimensionality of the problem through orthogonal decomposition of state variables have been reported by Zhang et al. [12] and Zeng et al. [13,14].

* Corresponding author. Tel.: +39 0223996256.
E-mail address: marco.panzeri@polimi.it (M. Panzeri).

http://dx.doi.org/10.1016/j.advwatres.2014.01.007
0309-1708/

Small sample sizes give rise to filter inbreeding [6] whereby EnKF systematically understates parameter and system state estimation errors; rather than stabilizing as they should, these errors appear to continue decreasing indefinitely with time, giving a false impression that the quality of the parameter and state estimates likewise keeps improving. There is no general theory to assess, *a priori*, the impact that the number *NMC* of MC simulations would have on the accuracy of moment estimates. The rate at which the sample mean, variance and associated confidence intervals of a random variable converge with the number of Monte Carlo runs is found, for example, in [15] and references therein. It suggests that increasing *NMC* by a factor of a few hundred, as is often done, would likely not lead to marked improvements in accuracy. A practical solution is to continue running MC simulations till the sample mean and variance stabilize or, if computer time is at a premium, till their rates of change slow down markedly.

van Leeuwen [16] showed theoretically that filter inbreeding is caused by (a) updating a given set (ensemble) of model output realizations with a gain computed on the basis of this same set and (b) spurious covariances associated with gains based on finite numbers *NMC* of realizations. Remedies suggested in the literature are generally *ad hoc*. Houtkamer and Mitchell [17] proposed splitting the set of MC runs into two groups and updating each subset with a Kalman gain obtained from the other subset. Hendricks Franssen and Kinzelbach [18] proposed alleviating the adverse effects of filter inbreeding by (a) dampening the amplitude of log-conductivity fluctuations, (b) correcting the predicted covariance matrix on the basis of a comparison between the predicted ensemble variance and the average absolute error at measurement locations, and (c) running a large number of realizations (in their case *NMC* = 1000) during the first simulation step and a subset of realizations (*NMC* = 100) thereafter; a procedure similar to the latter was also suggested in [19]. To select an optimal subset one would minimize some measure of differences between cumulative sample distributions of hydraulic heads obtained in the first step with (say) *NMC* = 1000 and *NMC* = 100. This, however, brings about an artificial reduction in variance [18]. Hendricks Franssen and Kinzelbach [18] obtained best results with a combination of all three techniques. Hendricks Franssen et al. [20] observed filter inbreeding when analyzing variably saturated flow through a randomly heterogeneous porous medium with *NMC* = 100 even after dampening log-conductivity fluctuations by a factor of 10. Several authors (e.g., [21–24]) have seen a reduction in filter inbreeding effects through covariance localization and covariance inflation. Covariance localization is achieved upon multiplying each element of the updated state covariance matrix by an appropriate localization function to reduce the effect of spurious correlations [17,25]. In the covariance inflation methods, the forecast ensemble is inflated through multiplication of each state by a constant or variable factor (e.g., [23,24,26]).

To eliminate the need for repeated MC simulations and associated filter inbreeding effects, we [27] proposed a new EnKF approach based on stochastic moment equations (MEs) of transient groundwater flow [28,29]. Solving these deterministic equations yields direct estimates of theoretical ensemble moments required for EnKF. We tested our new approach on a synthetic two-dimensional flow problem, showing it to yield accurate estimates of log-conductivity and their variance across the flow domain. MEs have been used successfully to analyze steady state [30] and transient flow [29] as well as particle travel times and trajectories [31,32] in randomly heterogeneous media. Second-order approximations of these equations have yielded accurate predictions of flows in heterogeneous media with unconditional variances of (natural) log hydraulic conductivity as high as 4.0 [30]. A transient algorithm based on the Laplace transform due to [29] was shown to be more efficient when computing transient hydraulic head variance than the traditional Monte Carlo method. A detailed comparison between ensemble- and simulation-based inversion methods in the case of steady-state groundwater flow was presented by Hendricks Franssen et al. [33].

While the theoretical elements and the numerical algorithms associated with our new ME-based EnKF framework have been presented in [27], a detailed comparison between MC- and ME-based EnKF variants in domains having various degrees of heterogeneity is still lacking. In this paper we compare the performances and accuracies of these two approaches on synthetic problems of two-dimensional transient groundwater flow toward a well pumping water from a randomly heterogeneous confined aquifer subject to prescribed head and flux boundary conditions. Problems differ from each other in the variance and (integral) autocorrelation scale of the log hydraulic conductivity field. The paper is organized as follows. Section 2 casts the Kalman Filter updating algorithm for groundwater data assimilation within a Bayesian framework (e.g., [34–36]). Section 3 presents the flow problem and describes the two EnKF procedures based on ME and MC. Section 4 illustrates and discusses some of our key results and Section 5 presents our conclusions.

2. Bayesian representations of ME- and MC-based EnKF

We consider transient groundwater flow in a saturated domain Ω governed by stochastic partial differential equations of mass balance and Darcy's law

$$S_S \frac{\partial h(\mathbf{x},t)}{\partial t} + \nabla \cdot \mathbf{q}(\mathbf{x},t) = f(\mathbf{x},t) \quad (1)$$

$$\mathbf{q}(\mathbf{x},t) = -K(\mathbf{x})\nabla h(\mathbf{x},t) \quad (2)$$

subject to initial and boundary conditions

$$h(\mathbf{x}, t=0) = H_0(\mathbf{x}) \quad \mathbf{x} \in \Omega \quad (3)$$

$$h(\mathbf{x},t) = H(\mathbf{x},t) \quad \mathbf{x} \in \Gamma_D \quad (4)$$

$$-\mathbf{q}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = Q(\mathbf{x},t) \quad \mathbf{x} \in \Gamma_N \quad (5)$$

where $h(\mathbf{x}, t)$ is hydraulic head and $\mathbf{q}(\mathbf{x}, t)$ the Darcy flux vector at point $(\mathbf{x}, t)$ in space–time, $K(\mathbf{x})$ is an autocorrelated random field of scalar hydraulic conductivities, S_S is specific storage treated here as a deterministic constant, $H_0(\mathbf{x})$ is (generally) a random initial head field, $f(\mathbf{x}, t)$ is (generally) a random source function of space and time, $H(\mathbf{x}, t)$ and $Q(\mathbf{x}, t)$ are (generally) random head and normal flux conditions on Dirichlet boundaries Γ_D and Neumann boundaries Γ_N, respectively, and $\mathbf{n}$ is a unit outward normal to Γ_N.

Our goal is to determine the posterior probability distribution of the random augmented (i.e., containing both model parameters and state variables) state vector

$$\mathbf{s} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{h} \end{bmatrix} \quad (6)$$

conditioned on measurements of the random vectors $\mathbf{Y}$ and $\mathbf{h}$. The parameter vector $\mathbf{Y}$ contains N_Y log-conductivities and the state vector $\mathbf{h}$ includes N_h hydraulic head values satisfying (1)–(5), so that $\mathbf{s}$ has dimension $N_s = N_Y + N_h$. In our finite element solver of (1)–(5), described below, N_Y is the number of elements (or collections of elements) in which hydraulic conductivity is taken to be uniform and N_h is the number of nodes at which heads are computed.

We denote the state vector $\mathbf{s}$ at the end of time interval $(T_{k-1}, T_k]$, before new measurements become available at time $t = T_k$, by $\mathbf{s}^{f,T_k}$. In line with [34–36] we consider $\mathbf{s}^{f,T_k}$ to be multivariate Gaussian with *prior* probability density (pdf)