Journal of Hydrology 554 (2017) 743–757

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Examining dynamic interactions among experimental factors influencing hydrologic data assimilation with the ensemble Kalman filter

S. Wang [a,*], G.H. Huang [b], B.W. Baetz [c], X.M. Cai [d], B.C. Ancell [e], Y.R. Fan [b]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China
[b] Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, Saskatchewan, Canada
[c] Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada
[d] Ven Te Chow Hydrosystems Laboratory, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[e] Department of Geosciences, Texas Tech University, Lubbock, TX, USA

ARTICLE INFO

Article history:
Received 9 March 2017
Received in revised form 24 September 2017
Accepted 27 September 2017
Available online 28 September 2017
This manuscript was handled by G. Syme, Editor-in-Chief.

Keywords:
Ensemble Kalman filter
Data assimilation
Hydrologic ensemble prediction
Interaction
Streamflow
Uncertainty

ABSTRACT

The ensemble Kalman filter (EnKF) is recognized as a powerful data assimilation technique that generates an ensemble of model variables through stochastic perturbations of forcing data and observations. However, relatively little guidance exists with regard to the proper specification of the magnitude of the perturbation and the ensemble size, posing a significant challenge in optimally implementing the EnKF. This paper presents a robust data assimilation system (RDAS), in which a multi-factorial design of the EnKF experiments is first proposed for hydrologic ensemble predictions. A multi-way analysis of variance is then used to examine potential interactions among factors affecting the EnKF experiments, achieving optimality of the RDAS with maximized performance of hydrologic predictions. The RDAS is applied to the Xiangxi River watershed which is the most representative watershed in China's Three Gorges Reservoir region to demonstrate its validity and applicability. Results reveal that the pairwise interaction between perturbed precipitation and streamflow observations has the most significant impact on the performance of the EnKF system, and their interactions vary dynamically across different settings of the ensemble size and the evapotranspiration perturbation. In addition, the interactions among experimental factors vary greatly in magnitude and direction depending on different statistical metrics for model evaluation including the Nash–Sutcliffe efficiency and the Box–Cox transformed root-mean-square error. It is thus necessary to test various evaluation metrics in order to enhance the robustness of hydrologic prediction systems.

1. Introduction

Since hydrologic models are mathematical representations of complex watershed processes, uncertainty is pervasive throughout hydrologic predictions (DeChant and Moradkhani, 2014). Uncertainty in hydrologic predictions originates from various sources, including the descriptions of boundary and initial conditions, the errors in model forcing data, difficulty in obtaining accurate parameter estimates, and model structural deficiencies (Ajami et al., 2007). Therefore, efficient quantification and reduction of uncertainty are necessary to provide reliable hydrologic predictions (Moradkhani et al., 2012; Wang et al., 2015c).

Over the past few decades, tremendous efforts have been made in the development and application of sequential data assimilation techniques for explicitly dealing with various sources of uncertainty in hydrologic modeling (Weerts and El Serafy, 2006; Liu and Gupta, 2007; Ryu et al., 2009; Gharamti et al., 2013; Panzeri et al., 2014; Randrianasolo et al., 2014; Khan and Valeo, 2016; Wang et al., 2017). Sequential data assimilation techniques continuously update model states when new observations become available to improve the forecast accuracy (Vrugt et al., 2005). The Kalman filter (KF) is the most commonly used sequential data assimilation technique, which was developed in the 1960s for optimal control of linear dynamic systems (Kalman, 1960). For nonlinear dynamics, the extended Kalman filter (EKF) can be used, which linearizes the error covariance equation by using a tangent linear operator. However, EKF produces unstable results when the nonlinearity in dynamic systems is strong and requires considerable computational effort due to the error covariance propagation (Evensen, 1992). As a result, the ensemble Kalman filter (EnKF) was introduced by Evensen (1994). The EnKF takes advantage of the Monte Carlo method to approximate the error covariance

* Corresponding author.
E-mail addresses: shuo.s.wang@polyu.edu.hk, sshuo.wwang@gmail.com (S. Wang).

https://doi.org/10.1016/j.jhydrol.2017.09.052
0022-1694/

evolution equation used in the EKF, which is capable of providing the forecast error estimate with a significantly lower computational cost and without any closure problem in the error covariance evolution equation.

Due to its attractive features of real-time adjustment and efficient implementation, the EnKF has been extensively used for recursive estimation of hydrologic model parameters and state variables (Xie and Zhang, 2010; Cammalleri and Ciraolo, 2012; DeChant and Moradkhani, 2012; Rafieeinasab et al., 2014; Gharamti et al., 2015; Liu et al. 2016a,b; Pathiraja et al., 2016b). For example, Moradkhani et al. (2005) proposed a dual state-parameter estimation approach based on the EnKF for sequential estimation of both parameters and state variables of a hydrologic model. Wang et al. (2009) proposed a constrained EnKF framework for simultaneous state estimation and sequential parameter learning in hydrologic modeling, in which the naive method, the projection and accept/reject methods were used to deal with inequality constraints. Samuel et al. (2014) evaluated the variations of streamflow and soil moisture by using the EnKF with dual state-parameter estimation for streamflow assimilation, soil moisture assimilation, and combined assimilation of streamflow and soil moisture. Pathiraja et al. (2016a) investigated the potential for data assimilation by using the EnKF to detect known temporal patterns in model parameters from streamflow observations. The EnKF is recognized as a powerful data assimilation technique that generates an ensemble of model variables through stochastic perturbations of forcing data and observations (inputs and outputs). Thus, identification of perturbation factors and selection of the ensemble size are key features of the EnKF (Moradkhani et al., 2005). However, relatively little guidance exists in literature with regard to the proper specification of the magnitude of the perturbation and the ensemble size, posing a significant challenge in optimally implementing the EnKF (Crow and Loon, 2006).

Since an inappropriate specification of factors affecting the EnKF experiment can degrade the performance of the data assimilation system, sensitivity experiments are often carried out for identifying error parameters and estimating the ensemble size (Clark et al., 2008; Sun et al., 2009; McMillan et al., 2013; Rasmussen et al., 2015). However, the sensitivity analysis experiments are limited in determining optimal settings of the EnKF applied to a particular problem. Thus, Yin et al. (2015) used a series of mathematical derivations to derive the optimal ensemble size of the EnKF used for a sequential soil moisture data assimilation system. In the EnKF, stochastic perturbations account for uncertainties in model parameters, inputs, and outputs. Specification of perturbation factors is a key feature of the EnKF, which plays a crucial role in the performance of sequential data assimilation experiments (Clark et al., 2008).

As a recursive scheme for estimating state variables and model parameters, the experimental factors involved in the EnKF are actually correlated with each other, and their interactions have a remarkable influence on the behavior of nonlinear dynamic systems. For example, many of the highly sensitive factors may provide redundant and misleading information regarding the variability of response variables since their sensitivities may be correlated with those of the other factors. As a result, failure to account for potential interactions among experimental factors can degrade the performance of the EnKF system (Crow and Loon, 2006; Thiboult and Anctil, 2015). It is thus necessary to examine the interactions among experimental factors and to quantify their contributions to the variation in model responses in order to maximize the predictive performance.

In this paper, we develop a robust data assimilation system (RDAS) with a factorial experimental design framework to enhance the effectiveness and robustness of the EnKF for hydrologic ensemble predictions. A multi-factorial EnKF method will be proposed by combining the strengths of multivariate hypothesis testing and sequential data assimilation techniques. In the RDAS, the EnKF will be carried out under various combinations of factors with different scenarios, leading to a diverse set of EnKF experiments. The multi-way analysis of variance (ANOVA) will then be used to uncover dynamic interactions among factors involved in the EnKF experiments, which provides meaningful insights for advancing the understanding of the sequential data assimilation process and maximizing the EnKF performance. The RDAS will be applied to predict daily streamflow in the Xiangxi River watershed in China since daily streamflow predictions play a key role in flood risk assessment and management.

This paper is organized as follows. Section 2 introduces the framework of the proposed RDAS for hydrologic ensemble predictions. Section 3 provides details on the study area and the experimental setup. Section 4 presents a systematic analysis of multi-factorial EnKF experiments along with a thorough discussion of interactions among experimental factors affecting the performance of the EnKF system. Finally, conclusions are drawn in Section 5.

2. Development of robust data assimilation system

The RDAS takes into account potential interactions among experimental factors influencing the EnKF data assimilation and quantifies their joint effects on the EnKF performance. An overview of the steps involved within the RDAS framework is provided as follows: 1) selection of the EnKF experimental factors, 2) factorial design of the EnKF experiments, 3) execution of ensemble data assimilation experiments, 4) multi-way ANOVA, 5) examination of dynamic interactions among experimental factors, 6) quantification of the joint effects of experimental factors on the EnKF performance, and 7) hydrologic ensemble predictions. The aforementioned steps can be categorized into four parts: EnKF, multi-factorial ANOVA, multi-factorial EnKF, and selection of statistical metrics for model evaluation.

2.1. Ensemble Kalman filter

The EnKF is a sequential data assimilation technique that makes use of Monte Carlo integration methods to approximate the error covariance matrix by a stochastic ensemble of model realizations (Evensen, 2003). In contrast to the extended Kalman filter (EKF), the EnKF represents the error covariance evolution through a set of model realizations rather than an explicit mathematical expression, which is particularly useful for nonlinear dynamic models. As a result, the ensemble of model states is integrated forward in time to predict error statistics (DeChant and Moradkhani, 2012). The model forecast can be made through the EnKF as follows:

$$x_{i,t+1}^{-} = f\left(x_{i,t}^{+}, u_{i,t+1}, \theta_{i,t+1}^{-}\right) + \omega_{i,t+1}, \omega_{i,t+1} \sim N\left(0, \sum_{t+1}^{m}\right), \tag{1}$$

where i and t denote the ensemble number and the time step, respectively; $x_{i,t}^{+}$ and $x_{i,t+1}^{-}$ represent the posterior model states at the previous time step and the predicted model states at the current time step, respectively; f represents the forward model that propagates the system states from time t to $t+1$ in response to model inputs $u_{i,t+1}$ and parameters $\theta_{i,t+1}$; and $\omega_{i,t+1}$ represent the model errors that follow a Gaussian distribution with zero mean and covariance $\sum_{t+1}^{m}$. As for the recursive parameter estimation through the EnKF, it is assumed that model parameters are perturbed by a small random noise in order to maintain diversity in posterior parameters:

$$\theta_{i,t+1}^{-} = \theta_{i,t}^{+} + \tau S(\theta_{i,t}^{-}), \tag{2}$$