Advances in Water Resources 60 (2013) 75–88

Contents lists available at ScienceDirect

Advances in Water Resources

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

Dual states estimation of a subsurface flow-transport coupled model using ensemble Kalman filtering

Mohamad El Gharamti [a], Ibrahim Hoteit [a,b,*], Johan Valstar [c]

[a] Earth Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia
[b] Applied Mathematics and Computational Sciences, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia
[c] Subsurface and Groundwater Systems, Deltares, Princetonlaan 6, 3584 CB, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 26 January 2013
Received in revised form 15 July 2013
Accepted 15 July 2013
Available online 27 July 2013

Keywords:
Flow and transport coupled models
Ensemble Kalman filter
Joint states estimation
Dual states estimation

ABSTRACT

Modeling the spread of subsurface contaminants requires coupling a groundwater flow model with a contaminant transport model. Such coupling may provide accurate estimates of future subsurface hydrologic states if essential flow and contaminant data are assimilated in the model. Assuming perfect flow, an ensemble Kalman filter (EnKF) can be used for direct data assimilation into the transport model. This is, however, a crude assumption as flow models can be subject to many sources of uncertainty. If the flow is not accurately simulated, contaminant predictions will likely be inaccurate even after successive Kalman updates of the contaminant model with the data. The problem is better handled when both flow and contaminant states are concurrently estimated using the traditional joint state augmentation approach. In this paper, we introduce a dual estimation strategy for data assimilation into a one-way coupled system by treating the flow and the contaminant models separately while intertwining a pair of distinct EnKFs, one for each model. The presented strategy only deals with the estimation of state variables but it can also be used for state and parameter estimation problems. This EnKF-based dual state-state estimation procedure presents a number of novel features: (i) it allows for simultaneous estimation of both flow and contaminant states in parallel; (ii) it provides a time consistent sequential updating scheme between the two models (first flow, then transport); (iii) it simplifies the implementation of the filtering system; and (iv) it yields more stable and accurate solutions than does the standard joint approach. We conducted synthetic numerical experiments based on various time stepping and observation strategies to evaluate the dual EnKF approach and compare its performance with the joint state augmentation approach. Experimental results show that on average, the dual strategy could reduce the estimation error of the coupled states by 15% compared with the joint approach. Furthermore, the dual estimation is proven to be very effective computationally, recovering accurate estimates at a reasonable cost.

1. Introduction and motivation

Geologists, subsurface engineers and hydrologists have been working for decades to understand the physical and dynamical processes that lead to understanding the delicate nature and configuration of the earth's subsurface. To this end, one method has been to build numerical hydrologic models that captures the complicated behaviors of fluids and solids in the subsurface [1]. The focus of these models is to simulate and predict dynamical fluxes and energies, defined as state variables, as accurately as possible based on selected time-invariant parameters that describe the subsurface geometry, fluid and rock properties, and surface–subsurface interactions [2]. Assessing the quality of the output of these models then became an issue.

Indeed, because of the large number of poorly understood parameters and of the presence of model uncertainties, the outputs of these numerical models are not always accurate. To improve these models, efforts have been directed toward calibrating their parameters using optimization-based techniques [3–8]. These calibration and inverse methodologies, which generally rely on adjoint-based or statistical Monte Carlo type approaches, aim at minimizing long-term prediction errors using historical batches of hydraulic head data, phase saturations, contaminant concentrations, etc. The limitations associated with these methods include their extensive computational burden and data storage, lack of an efficient approach to update the estimated parameters with

* Corresponding author at: Earth Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.
E-mail address: ibrahim.hoteit@kaust.edu.sa (I. Hoteit).
URL: http://assimilation.kaust.edu.sa/ (I. Hoteit).

0309-1708/$ - see front matter
http://dx.doi.org/10.1016/j.advwatres.2013.07.011

new observations, and the inability to take into account different sources of uncertainty in the system [2,3,9].

As an attempt to go beyond the calibration methods, sequential data assimilation techniques were successfully introduced as efficient solutions to improve hydrological models [10]. Data assimilation schemes have the advantage of accounting for model errors that are not only present in the uncertain parameters but also in the model structure and inputs, such as external forcings (recharge, river-aquifer interactions, etc.) [9]. Sequential data assimilation also does not require storage of all past information about the states and parameters by allowing online updating of the system variables using incoming observations.

There are different assimilation techniques that vary in structure depending on the nonlinearity, complexity and dimensions of the systems. From these, we mention the particle filter (PF) [11], which has been recently used in different hydrologic studies for tackling state-parameter estimation problems. PF can handle any type of statistical distribution (not necessarily Gaussian) and is well suited for strongly nonlinear dynamics [12–19]. It is based on forming a large set of particles that sample the state and parameter space according to a given prior distribution. The particles are propagated in time with the dynamical model and then weighted with available measurements according to the data-prediction misfit. Because of the high computational requirements of the PF, the more popular assimilation tool known as the Ensemble Kalman Filter (EnKF) [20–22], which is also based on Monte Carlo sampling, is much often used. The EnKF and its variants, which feature a Gaussian Kalman update [14,15], are widely used in hydrology, e.g., [2,6,8,9,23–33]. The EnKF has been shown to be efficient in handling nonlinear models with Gaussian uncertainties. It is further easy to implement and includes a general framework for state and parameter estimation.

Various hydrogeological studies have been conducted to tackle the state-parameter estimation problem with the EnKF. Two approaches were mainly considered: joint and the dual state-parameter approaches. The standard joint approach concurrently estimates the state and the parameters in a single augmented vector and has been extensively applied to hydrological problems, e.g., [9,25,27,34,35]. Advocates of the dual approach argue that the joint approach introduces important inconsistencies that could lead to unstable estimation, particularly with large dimensions and in highly nonlinear systems [2,36–39]. In the dual formulation, two filters are run in parallel; one for the parameters and the other for the states [40,41]. The parameters could be set to take some random walk while waiting to be updated indirectly by the state variables data [2,37]. It is worth mentioning that the dual estimation has been also applied in the variational assimilation framework, e.g., [42–44].

In this work, we investigate and tackle an essential estimation problem that is similar in concept to the state-parameter one. This problem arises in coupled models where at least two models, such as subsurface flow and contaminant transport, are dynamically coupled. In this one-way coupled system, reliable simulations require accurate flow and transport simulators. However, these models, especially the one for flow, are often far from being perfect because of many sources of uncertainties. If the flow model is uncertain, this would affect the solution of the transport model because uncertain flow outputs, such as hydraulic heads and Darcy velocities, are used as inputs to the transport model, which simply sees these outputs as "time-varying" parameters. If the EnKF is to be employed in such a framework at every assimilation cycle, estimates of the flow and transport states could be handled using either a joint or, as we will introduce here, a dual approach. The key point is that observations from both models are available, unlike in the state-parameter case in which only data from state variables are used to update the static parameters. Previous estimation studies with similar coupled hydrological systems have been conducted following the joint EnKF approach only, e.g., [26,27]. In this work, we introduce a dual-EnKF framework for efficient sequential data assimilation into one-way coupled models. This method can be seen as an extension of the dual state-parameter estimation method to a dual state-state estimation method for one-way coupled models. We formulate the dual states estimation procedure from a Bayesian perspective and present its EnKF-based solution. We then implement and evaluate the joint and dual estimation strategies with a coupled flow and transport system in a confined aquifer using the standard (stochastic) EnKF [20,21].

The remainder of this paper is organized as follows: Section 2 presents the general ensemble data assimilation framework. Section 3 discusses the joint states estimation method before introducing the dual estimation strategy for one-way coupled models. In Section 4, a synthetic example of a coupled subsurface transport model with results from assimilation experiments is presented. Final conclusions are given in Section 5.

2. Ensemble data assimilation

Sequential data assimilation aims at optimally estimating the state of a dynamical system using all available observations up to the estimation time. Modeling the system's uncertainties as random variables, this procedure is often implemented following the standard recursive Bayesian estimation framework [22]:

- Start from the probability distribution function (pdf) of the state x_{k-1} at time t_{k-1} given all available observations up to t_{k-1} denoted as $y^o_{1:k-1}$. This is referred to as the analysis pdf at t_{k-1} and denoted as $p(x_{k-1}|y^o_{1:k-1})$.
- For the forecast, propagate the analysis pdf to the time of the next available observation t_k. Assuming Markovian processes, this is done using the Chapman–Kolmogorov equation [46]. The resulting forecast pdf at time t_k is $p(x_k|y^o_{1:k-1})$.
- Update the forecast distribution with incoming observations, y^o_k, to obtain the new analysis distribution, $p(x_k|y^o_{1:k})$, using the Bayes rule.
- Once the update step is done, start a new forecast step.

The optimal estimate of the state x given observation $y, p(x|y)$ can be understood as the value that best describes a realization of x according to the given observation, y. From a statistical point of view, different optimal estimators can be formulated [45,46] and the most common ones are the minimum variance estimator (MV), the maximum a posteriori estimator (MAP) and the maximum likelihood estimator (ML). These estimators differ in terms of their objective functions. For instance, the MV estimator minimizes the spread around the state, x, whereas the MAP estimator looks for x that maximizes the conditional pdf, $p(x|y)$.

Consider now the generic discrete space–time dynamical system

$$x_k = \mathscr{M}_k(x_{k-1}, u_k) + \eta_k, \tag{1}$$

where $\mathscr{M}_k$ is a nonlinear dynamical operator integrating the state forward from time t_{k-1} to t_k, u_k is an input forcing to the system, and η_k is a random variable assumed to be uncorrelated in time describing the system noise (or model error) with a normal distribution, $\mathcal{N}(0, Q_k)$. The observation equation is written as

$$y_k = H_k(x_k) + \varepsilon_k, \tag{2}$$

where H_k is the observational operator mapping the state variables to the measured ones and ε_k is a random observational error that is assumed to be uncorrelated in time and also uncorrelated with the model noise, with a Gaussian distribution, $\mathcal{N}(0, R_k)$. Denote by N_x