Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter

Xianhong Xie a,⁎, Dongxiao Zhang a,b,c

a Department of Energy and Resources Engineering, College of Engineering, Peking University, Beijing 100871, China
b Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, California, USA
c Mark Family Department of Chemical Engineering and Materials Sciences, University of Southern California, Los Angeles, California, USA

1. Introduction

Understanding the response of a river catchment to atmospheric forcing is critically important to climate studies, agriculture (irrigation planning and vegetation and crop growth), natural hazards prevention and mitigation (e.g. floods and droughts), and other water resources management (e.g. water transfer and storages). These studies are highly dependent on advanced simulation models and large amounts of environmental data that are increasingly being made available. Hydrological models are built based on a set of principles coupled with a number of assumptions and imperfectly defined parameters, and measurements are usually scarce in space and discontinuous in time. These result in great uncertainties about the measurements, model structures and parameters. In applications, numerical hydrological modeling generally requires estimation of model parameters through calibration with observed data to mitigate the output uncertainties. Many calibration methods have been developed, including the automatic calibration with multiple objectives or criteria [20,24,32]. But most of the calibration methods are to find a set of reasonable parameter values and to attribute all errors to parameter uncertainties [23]. Moreover, general calibration methods are difficult to completely utilize multiple sources of data from remote sensing and automated ground-based sensors. These limitations can be remedied with the recently developed data assimilation techniques.

The hydrological data assimilation method shares the basic tenet of merging models and observations and accounts for uncertainties from different sources of information [21]. The idea has been inspired by and adapted from atmospheric and oceanic data assimilation systems that concern estimation of initial conditions. In contrast, land surface dynamics, including hydrological processes at the catchment scale, are fundamentally damped in nature, and the process assimilation is all about estimating errors in uncertain meteorological forcing conditions and model parameterizations [29]. So the sequential data assimilation methods, such as the Kalman filter (KF) and the...
ensemble Kalman filter (EnKF), are more suitable for hydrological applications and it can continually update the hydrological states and model parameters when new data become available. Especially, the EnKF is becoming popular in many areas of earth sciences because it is easy to use, is flexible, and makes relatively few restrictive assumptions [9,10]. It can also account for nonlinearities and partially non-Gaussianity.

The EnKF has been successfully applied in the land surface data assimilation [17,39,40]. The modern adaptive filtering techniques were investigated to address the error estimation problem that is a crucial component in the EnKF [6,29]. However, most of these studies focused on the regional or global scale of interest in hydrometeorology, reflecting on the land surface–atmosphere interactions, while the horizontal movements of water, including the overland flow, stream flow routing and water interactions between sub-catchments, are often omitted in the land surface models. On the other hand, the fundamental operative unit for water resources management is the catchment or river basin. There is therefore a need to extend the applicability of data assimilation in hydrology from the regional land surface processes to catchment scale hydrological issues [37]. At this more local scale, the main issues are to estimate hydrologic parameters (e.g. the permeability) and to characterize hydrologic responses such as runoff, soil water movement, evapotranspiration and groundwater movement.

There have been a few of studies with encouraging results concerning the local scale hydrological state estimation with data assimilation. Kitanidis and Bras [14,15] originally reformulated a nonlinear (jumped) conceptual catchment model into a form amenable to linear estimation for the KF application, and they found that the assimilating feedback significantly improved the real-time forecasting of river discharges even when the model and input error statistics are not perfectly known. Schuurmans et al. [33] applied the KF with a constant Kalman gain to assimilate the remotely sensed latent heat flux for improving the water balance computation, and demonstrated that data assimilation has much potential for analyzing and improving distributed hydrological model predictions. Aubert et al. [1] investigated the extended KF in an operational forecasting context by introducing soil moisture data into stream flow modeling. They also indicated remote sensing data coupled with sequential assimilation to be well adapted to streamflow forecasting. Komma et al. [16] examined the benefits of updating soil moisture with the EnKF in forecasting large floods. Das et al. [8] used an ensemble square root filter (EnSRF) scheme to assimilate the aircraft-based soil moisture observations in a distributed hydrological model.

These studies greatly increase the potential to successfully apply ensemble filtering methods in catchments for hydrological modeling that allows coupled physical processes including precipitation, overland flow, infiltration, evapotranspiration, groundwater and streamflow. The representation of these processes resorts to empirical and physical nonlinear equations instead of well-defined governing equations like the general groundwater models or the land surface models. In distributed hydrological models, unavoidably, there are a large number of states and parameters [28,35], which pose difficulties for assimilating different types of observations even with the EnKF. Hence data assimilation approaches are often used to estimate the dynamic states while the parameters are excluded from the assimilation update group [23]. For examples, Pauwels et al. [26] perturbed parameters associated with TOPMODEL but those parameters for every member were not updated when assimilating runoff to estimate soil moisture, and Clark et al. [5] used streamflow observations to update dynamic hydrological states based on prior calibrated parameters. This could obtain reasonable historical values, but may not be applicable to the long-term hydrological predictions or the environment with time-varying model parameters [23].

The filter techniques implicitly provide a combined estimation for the dynamic state and the parameters by augmenting a joint state vector [28]. The standard KF is limited to a linear dynamic system [23], while the EnKF bypasses this limitation and it has been successfully used in hydrogeology for the parameter estimation with synthetic experiments [4,36] and with field scale flow and transport experimental data sets [18]. In order to reduce the degree of freedom in the joint state vector, a dual state-parameter estimation strategy was developed based on the EnKF and has been demonstrated in conceptual rainfall-runoff models [23], while this dual estimation approach neglects the effect of cross-state and parameter dependencies [3], and it may be limited due to its overwhelming computational burden in the presence of large numbers of computational units.

Even though many studies focus on joint estimation using the EnKF with the state augmentation, little attention has been paid to the distributed hydrological modeling in which the state and parameters are rather high dimensional. The applicability of the state augmentation is also not clear when assimilating multi-site observations with biased error estimation. Therefore, the potential of the EnKF needs further exploration to address the issues about the combined estimation for distributed catchment models.

In this paper, we assess the performance of the EnKF with the state augmentation for a physical-based distributed hydrological model, SWAT, and focus on the parameter estimation of the curve number (CN2) in the SWAT, as well as the estimation of prognostic variables such as the runoff, soil water content and evapotranspiration. The parameters are assembled into an augmented state vector because there are strong relations between parameters and dynamic states. We do so on the basis of synthetic examples. Multiple types of measurements are introduced and their effects are discussed. In order to better demonstrate its performance, sensitivities from the error prescription, the initial realization and the ensemble size are investigated.

The remainder of the paper is organized as follows. After a brief review of the SWAT model and the EnKF method in Section 2, we describe the design of the data assimilation system and illustrate the synthetic experiments in Section 3. Diagnostics of the capability and sensitivity of EnKF are discussed in Sections 4 and 5, respectively. Conclusions and a summary of future work are given in Section 6.

2. Sequential data assimilation

2.1. Dynamic hydrological process

With the time evolving, a hydrological system is integrated with dynamic states and static parameters under driving forces. A dynamic hydrological model can be expressed as a nonlinear stochastic process [2].

$$X_{t+1} = f(X_t, U_t + 1) + \omega_t + 1$$

where $t$ denotes the time step, $X$ is an augmented state vector consisting of dynamic variables (e.g., water content in the soil profile) and static parameters (e.g., hydraulic conductivity), $f$ is the nonlinear hydrological model forward operator, $U$ is a set of externally specified time-dependent forcing variables (e.g., precipitation), and the noise term $\omega$ accounts for model errors that represent all uncertainties related to model structure, forcing variables and parameters.

Observations using different instruments or techniques can be obtained from the hydrological system. It could be directly or indirectly related to the hydrological states and be written as,

$$Y_t = H(X_t) + \epsilon_t$$

where $Y$ denotes the observation vector from measurement instruments, $H$ is the linear or nonlinear observation operator specifying deterministic relationship between the observation data and the true state $X$, and $\epsilon$ is a noise term accounts for both measurement error
where the probability density of model states is represented by a large ensemble of model states, including parameters, and each member of the ensemble is integrated forward in time by the model independently. Similar to Eq. (1), the model forecast is executed in the EnKF for each ensemble member as follows:

\[ X_{t+1}^- = f(X_t, \mu_i, U_{t+1}) + \omega_{t+1}, \quad \omega_{t+1} \sim N(0, Q_{t+1}), \quad i = 1, \ldots, n \]  

(3)

\[ U_{t+1} = U_{t+1} + \xi_{t+1}, \quad \xi_{t+1} \sim N(0, R_{t+1}) \]  

(4)

where \( n \) is the ensemble size, namely the number of ensemble members, \( X_{t+1}^- \) is the component of the \( i \)th ensemble member forecast at time \( t + 1 \), \( X_t \) is the \( i \)th updated ensemble member at time \( t \), \( U_t \) is the \( i \)th perturbed forcing variables, \( \omega_{t+1} \) and \( \xi_{t+1} \) are independent white noises for the forecast model and forcing terms, drawn from multi-normal distributions with zero mean and specified covariance \( Q_{t+1} \) and \( R_{t+1} \). At time \( t + 1 \), the observation ensemble member can be written as,

\[ Y_{t+1} = HX_{t+1}^- + \mu_{t+1}, \quad \mu_{t+1} \sim N(0, S_t) \]  

(5)

where \( HX_{t+1}^- \) is a simplification of \( H(X_t, \mu_i) \), which is the original observation vector obtained from the true hydrological catchment state \( X_{t+1}^- \), and \( \mu_{t+1} \) is the noise term with zero mean and specified covariance \( S_t \).

With the model forecasts and observations being available, the assimilation or updating process can be expressed as,

\[ X_{t+1}^- = X_{t+1}^- + K_{t+1}(Y_{t+1} - HX_{t+1}^-) \]  

(6)

\[ K_{t+1} = P_{t+1}H^T(HP_{t+1}H^T + S_t)^{-1} \]  

(7)

\[ P_{t+1} \approx \frac{1}{N_t} \sum_{i=1}^{N} \left[ (X_{t+1}^- - X_{t+1}^-)(X_{t+1}^- - X_{t+1}^-)^T \right]^{-1} \]  

(8)

\[ X_{t+1} = \frac{1}{N_t} \sum_{i=1}^{N} X_{t+1}^- \]  

(9)

where \( X_{t+1}^+ \) is the new optimal estimate vector of hydrological states after assimilation, \( K_{t+1} \) is the Kalman gain, which determines the weight between the modeling and observation states in assimilation processes, \( P_{t+1} \), approximated with the ensemble, is the prior model error covariance. Similarly, the updated (posterior) error covariance could be estimated with the updated ensemble members. This covariance is not indispensable for the EnKF, whereas its estimation is necessary in the standard Kalman filter.

As shown in Eqs. (8) and (9), the EnKF algorithm uses the first and second order moments to represent the probability density, which means the Gaussian hypothesis is implicit for the model states, including parameters. Even though this hypothesis is often violated for the hydrological modeling, the EnKF appears to provide a good approximation for nonlinear and non-Gaussian land surface problems [40].

### 3. Data Assimilation Setup

#### 3.1. SWAT model and hydrological state

The data assimilation algorithm is implemented on a popular distributed hydrological model, Soil and Water Assessment Tool (SWAT). Since it is physically based and computationally efficient, uses readily available inputs, and enables users to study long-term impacts, this model has been widely used to predict impacts of land management practices on water, sediment, and agricultural chemical yields in large, complex watersheds with varying soils, land use, and management conditions over long periods of time [11,19,25]. For modeling purposes, a catchment is partitioned into a number of subcatchments or sub-basins according to the property of Digital Elevation Model (DEM) data, and then input information for each subbasin is grouped or organized into different hydrologic response units (HRUs) that are comprised of unique land cover, soil, and management combinations.

The land phase of the hydrologic cycle is simulated at the HRU scale based on a water balance equation with a daily step:

\[ SW_t = SW_0 + \sum_{i=1}^{n} \left( R_t - Q_{surf,i} - ET_t - W_{s,0,i} - Q_{gw,i} \right) \]  

(10)

where \( SW \) is the final soil water content (mm H2O), \( SW_0 \) is the initial soil water content (mm H2O), \( t \) is the time (days), \( R_t \) is the amount of precipitation on day \( t \) (mm H2O), \( Q_{surf,i} \) is the amount of surface runoff on day \( t \) (mm H2O), \( ET_t \) is the amount of evapotranspiration on day \( t \) (mm H2O), \( W_{s,0,i} \) is the amount of percolation and bypass flow exiting the soil profile bottom on day \( t \) (mm H2O), and \( Q_{gw,i} \) is the amount of return flow on day \( t \) (mm H2O). In this equation, the term of \( Q_{surf,i} \) is the main component that determines streamflow in reaches and the soil moisture in soil profiles. It can be expressed with an empirical model, SCS runoff equation:

\[ Q_{surf,i} = \left( \frac{(R_t - 0.2S_t)^2}{R_t + 0.8S_t} \right) \]  

(11)

where \( S_t \) is the retention storage of soil profiles (mm H2O). It is empirically defined as:

\[ S_t = \frac{25400}{CN_i} - 254 \]  

(12)

where \( CN_t \) is the temporal curve number for the day. In order to capture the temporal–spatial variations of the retention storage (\( S_t \)) due to changes in soil water content, soil properties, land use, management and slope, the temporal curve number is implicitly expressed as:

\[ CN_t = f(CN_2, SW_t) \]  

(13)

where \( CN_2 \) is the static curve number corresponding to the average moisture condition, and it depends on the soil's permeability, land use and antecedent soil water conditions [25]. Its value is often between 30 and 100. It is a dominant parameter for the surface runoff generation in an HRU, and consequently influences the streamflow process and the soil moisture states.
After the surface runoff generation, other processes including soil percolation, evapotranspiration, groundwater and seepage are carried out at the HRU level and subsequently water from these processes is aggregated to the subbasin level. The total runoff generated in a subbasin is routed through the channel network using the variable storage routing method or the Muskingum River routing method [19,25].

Although the SWAT model is capable of simulating other processes, such as the plant growth and the sediment movement, here we exclusively focus on the water movement in catchments. Moreover, we select 11 primary state variables to diagnose the performance of data assimilation on this model, even though there are hundreds of state variables to support the model run.

As shown in the Table 1, the first seven general states, including water stored or lagged in the soil profile and the drainage reaches, are model-dependent variables to characterize water storage conditions in HRUs or in subbasins. The subsequent three dynamic variables, Runoff (R), Soil water content (SW) and Evapotranspiration (ET) are viewed as observable or prognostic states since they could be measured with specific instruments or techniques. In addition, there are a large number of static variables (parameters) needed to be identified for a simulation application, and each of them may make a different contribution. However, the SCS runoff curve number, CN2, is the most sensitive parameter and its uncertainty could induce great effects on runoff modeling and other hydrological processes [13,24,32,34]. It is often needed to be estimated by calibration methods or data assimilation methods. In the SWAT model, it should be mentioned that the curve number is used to compute the maximum retention value and other two shape coefficients [25]. This computational process is implicitly expressed with Eq. (13). The three parameters will be continuously corrected with the CN2 updated in the data assimilation process.

### 3.2. Experimental area description

In this study, an experimental catchment is adopted in the Lake Fork Watershed in Northeast Texas, USA. This catchment area covers 489.85 km² and the altitudes of the basin vary between 106 m and 195 m above sea level. The pasture (51.3%), range-grasses (28.4%) and deciduous forest (16.0%) are the three main land use types.

The data set including topographic information, land use information, soil data, daily precipitation, and climate records is taken from the SWAT visual software modular [19]. As a synthetic data assimilation experiment, nevertheless, the data about stream runoff (streamflow), soil water content and actual evapotranspiration in subbasins are drawn from a reference simulation that will be described in the next subsection.

Based on a digital elevation model, the catchment is partitioned into 20 subbasins and subsequently 49 HRUs according to the land use and soil type information. The soil profile is divided into seven layers with different soil properties. As shown in Fig. 1, the serial number of subbasins is not in a natural sequence because some outlets are added to get approximately uniform subbasin areas. This is beneficial to the hydrological simulation and the data assimilation. Consequently, the joint state vector consists of 717 such variables as stated in Table 1.

### 3.3. Data assimilation procedure

Synthetic experiments are designed to assess the capability and sensitivity of the assimilation process to model parameterizations and physical representations, and this is also a general and effectual way to investigate data assimilation performance on predictive systems as shown by Chen and Zhang [4] and Kumar et al. [17]. Since the true parameter values are known in the experiments, we can easily assess the performance of the assimilation approach. In these experiments, one reference field is randomly picked up from a predefined Gaussian distribution with given statistics. And then the hydrological model is integrated to obtain the reference state (or “true” state), referred to as the “control run” or reference modeling. At the measurement locations, the observations (e.g. the runoff) to be used in the assimilation system are drawn from the reference state. Finally, the assimilation integrations are conducted by introducing the synthetic observations on the basis of a stochastic modeling platform.

In the hydrological modeling, a warming up process is often needed to initialize the model state and harmonize hydrologic responses with the meteorological forcing, and thus the warming up period should be set before the assimilation performs. However, the computation may be expensive to run an ensemble with a large number of modeling members through a long warming up period, which is often more than one year. Alternatively, we could run only one member of the ensemble to warm up (spin up) the model and then perturb the meteorological forcing terms and generate stochastic realizations of the parameters. In this study, three running periods are

<table>
<thead>
<tr>
<th>Order</th>
<th>State variable</th>
<th>Description</th>
<th>Scale</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Qsurfine</td>
<td>Amount of surface runoff stored or lagged</td>
<td>HRU</td>
<td>V</td>
</tr>
<tr>
<td>2</td>
<td>Qlateral</td>
<td>Amount of lateral flow stored or lagged</td>
<td>HRU</td>
<td>V</td>
</tr>
<tr>
<td>3</td>
<td>Qshl</td>
<td>Amount of shallow water stored or lagged</td>
<td>HRU</td>
<td>V</td>
</tr>
<tr>
<td>4</td>
<td>Qchrg</td>
<td>Amount of recharge entering the aquifer</td>
<td>HRU</td>
<td>V</td>
</tr>
<tr>
<td>5</td>
<td>Qgswinh</td>
<td>Amount of groundwater flow into the main channel</td>
<td>HRU</td>
<td>V</td>
</tr>
<tr>
<td>6</td>
<td>W_soil</td>
<td>Amount of water stored in the soil layer for each HRU</td>
<td>HRU × Nlay</td>
<td>V</td>
</tr>
<tr>
<td>7</td>
<td>W_r</td>
<td>Amount of water stored in the reach</td>
<td>subbasin</td>
<td>V</td>
</tr>
<tr>
<td>8</td>
<td>R</td>
<td>Amount of water flow (Runoff) out of a reach</td>
<td>subbasin</td>
<td>V</td>
</tr>
<tr>
<td>9</td>
<td>SW</td>
<td>Amount of water in soil in a subbasin</td>
<td>subbasin</td>
<td>V</td>
</tr>
<tr>
<td>10</td>
<td>ET</td>
<td>Amount of actual evapotranspiration in subbasin</td>
<td>subbasin</td>
<td>V</td>
</tr>
<tr>
<td>11</td>
<td>CN2</td>
<td>SCS runoff curve number for moisture condition II</td>
<td>HRU</td>
<td>P</td>
</tr>
</tbody>
</table>

Note: The class V denotes general state variables, V_o denotes the observable or prognostic variables, and P denotes principle parameters; Nlay is the number of layers of soil profiles in a hydrologic response unit (HRU).

Fig. 1. Subbasin distribution of the experimental catchment from the Lake Fork Watershed in Northeast Texas, USA.
specified for the data assimilation procedures: (1) Warming up period: the model propagates forward with the same boundary conditions (meteorological driving force) as the "control run", and the parameters (here is the curve number, \( C_{N2} \)) could be specified randomly; (2) Perturbation period: At the beginning of this period an ensemble of stochastic parameter fields are generated using a predefined Gaussian distribution, and the meteorological condition is perturbed using additive white noise throughout this period. Each member of the ensemble propagates forward independently and the hydrological states are simulated given the continuously perturbed the driving force; (3) Data assimilation period: After the ensemble has been generated and initialized, the hydrological states are updated by assimilating observations drawn from the "control run". The modeling states, as well as the driving force, are perturbed with white noises throughout this period. Clearly, these procedures are not only applicable to the synthetic experiments in this study but also could be carried out in a real-world application. Furthermore, when there is a poor knowledge about the initial hydrological conditions of a catchment for short-period modeling, the first two procedures are able to create reasonable distributions of hydrological states to approximate the initial conditions and be beneficial to the subsequent data assimilation operations. For long period modeling, however, there is no need to perturb meteorological conditions during the perturbation period, since the impact from poor initial conditions on assimilation performances may not be significant owing to the damping nature of hydrological systems [29].

In this study, the length of the control run is three years, 1095 days in total. As to the data assimilation modeling, correspondingly, the first 445 days is set as warming up period, and then it is the perturbation period (from 446th to 455th day). Without exceptional specification, subsequently, the assimilation is carried out from the 456th to the 1095th time step, 640 days in total. It should be reminded that the data assimilation method used here focuses on the estimations of \( C_{N2} \) and prognostic variables exclusively, and other parameters in the control run and assimilation run hold identical values.

In terms of the landuse and soil types in our study area, the values of the curve numbers (\( C_{N2} \)) for each HRU are around 75.0 [25], and uncorrelated stochastic realizations of \( C_{N2} \) are generated for the 49 HRUs by specifying Gaussian distributions, \( N(75.0, 5.0^2) \). The assumption of the Gaussianity is made for the convenience and should be subject to further investigation. Moreover, the standard deviation is empirically set as 5.0 that is large enough to account for their uncertainties, because the range of \( C_{N2} \) are not beyond (60, 90) in this area according to the reference values [25]. Based on this statistics, the Latin hypercube method [12,27] is used to generate a set of realizations. Fig. 2 shows 200 realizations and the black squares are randomly selected as a reference parameter field to simulate the "true" states under the control run.

Since the first seven general variables in Table 1 are temporary and their model errors are difficult to be identified, we assume they are free of model errors in our experiments. In contrast, the prognostic variables, namely the runoff (\( R \)), evapotranspiration (\( ET \)) and soil water content (\( SW \)), are perturbed by additive white noises to represent the possible measurement and model errors. These noises are mutually uncorrelated in space and time with means equaling to zero (indicating unbiased estimations) and standard deviations scaling to the current values of variables. Likewise, the precipitations are perturbed, and the standard deviations also equal to the products of a scaling factor and the current precipitation values. Furthermore, two different scaling factors for precipitation perturbations could be used in the perturbation period (period 2) and the assimilation period (period 3). Specifically, in the period 2, giving a larger scaling factor (here it is 0.30) can generate broad spreads of precipitation and hydrological response to account for uncertainty in the initial conditions when the prior knowledge is poor. Without exceptional specification, the ensemble size is 200 for all cases, and detailed specifications are shown in Table 2.

It should be mentioned that these errors may be prescribed in a number of ways. One way is to specify time-dependent errors [9,30] or even to use a modern adaptive filter algorithm [6,29]. But the other way described above is much more straightforward and operational and has been widely used in data assimilation works [5,17]. In addition, the hydrological system in our study is free of errors, since both of the reference modeling and the data assimilation modeling are implemented based on the same model (SWAT), and the observations are directly drawn from the reference simulation. In fact, the success of EnKF is quite dependent on the way of error prescription that should be consistent with the real uncertainties of the model and the measurement [9,10,26]. Therefore, the errors prescribed with scaling factors (Table 2) have overestimated the real uncertainties of the interested hydrological system, and this inconsistency is favorable to examine the robustness of EnKF.

4. Capability of EnKF

In practice, runoff is a comprehensive response to the hydrological cycle and it could be easily obtained from hydrological stations. We take it as a preferred observation data in the assimilation work as it is the most general type of measurement in catchments. Moreover, the remote sensing products, for example the soil moisture retrieval data, play more and more important roles in hydrological modeling with the recent technology development and they could be considered as complementary data sets in data assimilations.

4.1. Runoff measurement

We first take five runoff observations in the lower catchments as shown in Fig. 1, from the subbasins 7, 8, 9, 10 and 14. The runoff observations are drawn from reference modeling and assimilated at every time step. The initial realizations of curve number (\( C_{N2} \)) are generated with Gaussian distribution \( N(70, 5.5^2) \), which is biased with respect to the true solution, representing an initial guess to (prior knowledge of) the unknown true properties. It is also feasible to specify relatively accurate estimation of \( C_{N2} \) for each HRU according to prior knowledge (soil and land cover properties), due to the assumption of independence for all HRUs. But here we simply set identical distributions for all HRUs. In order to assess the capability of EnKF, we also prescribe relatively small scaling factors for the standard deviations of the errors: 0.001 for the observed runoff and 0.01 for the precipitation and the three model-based prognostic variables. These detailed specifications are exhibited as Case 1 in
4.1.1. Parameter estimation

After assimilation under the conditions described above, scatter-plots for the ensemble means of the \( CN_2s \) vs. the reference field are depicted in Fig. 3 for the initial, the 60th, the 180th and the 640th assimilation steps. It is apparent that the initial estimates of \( CN_2s \) are about 70 and do not have any resemblance of the reference field. Based on this set of initial realizations, the model replicates run to the next time step and are assimilated with the five runoff observations. The scatter points of the final result are around the 45-degree line, which indicates a good agreement between the ensemble means of \( CN_2s \) and the true values. Even just at the 180th assimilation step, the

<table>
<thead>
<tr>
<th>Case</th>
<th>Measurement type</th>
<th>( CN_2 s ) distribution</th>
<th>( S_o ) for observation</th>
<th>( S_m ) for model</th>
<th>( S_p ) for precipitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>RL</td>
<td>( N(75, 5.02) )</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>RL</td>
<td>( N(70, 5.52) )</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>RL</td>
<td>( N(70, 5.52) )</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>RL + SW</td>
<td>( N(70, 5.52) )</td>
<td>0.001 0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>4</td>
<td>RL + ET</td>
<td>( N(70, 5.52) )</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>RL</td>
<td>( N(70, 5.52) )</td>
<td>0.05</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>6</td>
<td>RL</td>
<td>( N(70, 5.52) )</td>
<td>0.1</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>RL</td>
<td>( N(70, 5.52) )</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>8</td>
<td>RL</td>
<td>( N(70, 5.52) )</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>9</td>
<td>RL</td>
<td>( N(70, 5.52) )</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>10</td>
<td>RL</td>
<td>( N(70, 5.52) )</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>11</td>
<td>RL</td>
<td>( N(70, 7.02) )</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>12</td>
<td>RL</td>
<td>( N(65, 7.02) )</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>13</td>
<td>RL</td>
<td>( N(65, 5.52) )</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Note: \( S_o \), \( S_m \) and \( S_p \) are the scaling factors for error standard deviations corresponding to variables; RL, R U and RO denote the runoff observations from five lower, five upper streams and the catchment outlet, respectively; SW and ET denote observations of the soil water content and the evapotranspiration; “+” denotes a combination of two kinds of measurements. The SW and the ET observations are only assimilated at the 1st, 31st, 61st, 91st and 121st assimilation steps.

Fig. 3. Assimilation estimates of the \( CN_2s \) vs true (reference) values for Case 1: (a) ensemble means of initial realizations; (b) ensemble means at the 60th assimilation step; (c) ensemble means at the 180th assimilation step, and (d) ensemble means at the 640th assimilation step (final results).
ensemble means capture the major features of the reference field, and there is a slight improvement for the CN₂ estimation in the rest of assimilation steps.

It is of interest to note that some estimates are still discrepant to the true values although the prognostic variables are matched very well. This implies the well-known nonuniqueness in the inverse problems. Specifically, the biased estimation values of CN₂s could provide acceptable hydrological responses, such as the runoff estimation. As typical examples taken from the 7th HRU in subbasin 5 and the 28th HRU in subbasin 10 (Fig. 4), the ensemble means of the CN₂s match the true values very well after 120 assimilation steps. In contrast, for the 31th HRU in the subbasin 11, the value of CN₂ is underestimated, and the ensemble spreads to a broad range.

4.1.2. Prognostic variable estimation

Getting accurate estimation for the prognostic variables is an objective of the hydrological simulation and the data assimilation. To compare the performance of the assimilation integrations, open loop simulations are conducted using the same initial realizations of the CN₂s as well as the uncertain precipitation inputs but without assimilating the model with observations. Thus each member of the ensemble integrates forward independently without updating parameters and variables. Root mean square errors (RMSEs) with respect to the true simulation are computed based on the ensemble means of the three prognostic variables to assess the estimation accuracy.

Fig. 5 shows average time series of RMSEs of prognostic variables for the entire modeling domain. It is clear that the data assimilation in Case 1 reduces the RMSEs of the runoff and soil water content estimations. Especially in peak runoff occurring periods, the RMSEs are lower than that in open loop simulations. However, the assimilation does not give any significant improvement for the evapotranspiration estimates as the two RMSE series keep at the same levels approximately. Evapotranspiration is a complex process containing the soil water evaporation, plant transpiration and canopy free water evaporation, and these processes are dominated by many parameters and variables. In this study, however, we only focus on the curve number (CN₂) estimation that dominates the surface runoff generation and soil water balance (Eqs. (10)–(13)) and indirectly influences the evapotranspiration processes. In order to diagnose the performance of evapotranspiration estimation, other parameters related to these processes may be needed under this assimilation condition.

Furthermore, taking subbasin 2 as a typical example which is far away from the observation locations, we exhibit the results monthly in order to distinguish the discrepancies of the EnKF and the open loop. It shows that the data assimilation estimates provide better matches with true processes than do the open loop simulations in which great discrepancies are present at the peak value points (Fig. 6). In fact, these results are general exhibitions that are shared by the other subbasins. Consequently, the data assimilation is capable of reducing the simulation uncertainties of these prognostic variables and improving their estimation accuracy.

4.2. Impact of runoff measurement sites

Runoff measurement stations may be assembled at the upper streams or the lower streams in a catchment. Runoff processes at different measurement sites are impacted by hydrological responses from their upper regions. Thus, it is necessary to examine the impact of runoff measurement sites on data assimilation. Two sets of data assimilation runs are conducted: (1) The runoff observations from the five upper streams, in subbasins 1, 3, 4, 15 and 19, are assimilated into the SWAT model at every simulation time step, and (2) the runoff observation only from the outlet at subbasin 14 is assimilated. As described in Case 2 and Case 3 in Table 2, other assimilation conditions...
are not changed. The initial ensemble means for the two cases are still about 70 as shown in Fig. 3a.

Fig. 7 exhibits the final results of Case 2 and Case 3. In Case 2, most of the points are below the 45-degree line, which indicates that the values of \( CN_2 \) are underestimated and that reasonable estimates for \( CN_2 \) are not obtained by assimilating the five observations from the upper streams (Fig. 7a). However, the Case 3 shows much better results since the assimilated-true points are around the 45-degree line (Fig. 7b).

When comparing the RMSE series of Case 1, Case 2 and Case 3 as shown in Fig. 8 (a), we can see that the RMSE level obtained with the five lower observations is lowest in the assimilation period, while that obtained with the upper observations is at the highest level. Furthermore, another favorable error measure, mean absolute error (MAE), is employed as it describes a natural error magnitude and avoids the effect of long-tail probability distributions [38]. However, the MAEs show identical patterns to the RMSEs (Fig. 8), which means long-tail probability distribution does not exist evidently and the parameter estimates can be improved by assimilation observations. Particularly, more reasonable results could be obtained by assimilating the runoff observations from the lower stream stations. This trait attributes to the fact that the runoff from lower streams, especially the catchment outlet, contains the full information about the hydrological states of the upper regions.

4.3. Multiple measurements

There are many types of measurements that can be combined to estimate the hydrological states. In the measurement family, hydrologic remote sensing plays an important role since it can provide much more information about the land surface conditions at large scales, for example the radiance, surface soil moisture and evapotranspiration. Even though the EnKF method is capable of assimilating the satellite radiances directly with nonlinear observation operators (as Eq. (2)), this treatment is too complex to get successful application for poor knowledge of radiative transfer processes [31]. Alternatively, it is possible to assimilate the remote sensing retrieval data that is considered as complementary information when the runoff observations are inadequate.

To test this idea, Case 4 and Case 5 are designed based on Case 3 as shown in the Table 2. In addition to the runoff observation from the outlet (subbasin 14), the soil water content and evapotranspiration of each subbasin, drawn from the reference modeling, are assimilated at the 1st, 31st, 61st, 91st and 121st time steps. Hence, the measurement groups are multiple: five-time data of soil water content are added to the single runoff observation for Case 4, and evapotranspiration for Case 5 likewise. As the treatment on the runoff observations, the scaling factors of standard deviations of these new observations are prescribed with 0.001 to account for their uncertainties in our synthetic experiments, even though this value may be too optimistic for practical measurements.

It is clear from the Fig. 8 that assimilating multiple types of measurements reduces the RMSEs and the MAEs of \( CN_2 \) estimates and leads to systematic improvements over the case of only assimilating the outlet runoff (Case 3). The improvement of Case 5 (adding evapotranspiration observations) is not as significant as that of Case 4 since the relationship between the \( CN_2 \) and the evapotranspiration is weak as explained before. Fig. 9 gives final results of Case 4 and Case 5 for the \( CN_2 \) estimates vs. the true values, indicating that their matches are acceptable despite the nonuniqueness and nonlinearity between parameters and hydrological responses. Moreover, Case 4 provides the best estimates about the three prognostic variables among the first five
cases after 100 assimilation steps even though its estimation accuracy for CN$_2$ is not comparable to Case 1 (Figs. 10 and 5). In fact, the soil water moisture condition plays an important role in controlling the surface runoff generation and the evapotranspiration processes, as expressed in Eq. (10) through Eq. (13). When its observations are assimilated, for example at the first assimilation step, the water storages in all subbasins approach the reference (true) values and these updated states result in acceptable estimates on the runoff and evapotranspiration at the second and subsequent steps. Therefore, the soil water content observation provides complementary data to the assimilation systems, even though it may not be available or utilized at every time step.

5. Sensitivity of EnKF

The EnKF is a Monte Carlo approximation based on a sequential Bayesian filtering process. When constructing a data assimilation framework, we should first consider various aspects, mainly including the model and observation errors, the prior knowledge about the parameter distribution as well as the ensemble size, because these factors have critical impacts on the accuracy and performance [4,17]. In this section, we diagnose sensitivities of the EnKF and focus on the performance with respect to the three factors. The detailed design for the cases is given in Table 2.

5.1. Error specification

The EnKF requires estimates of the model and observation errors to properly merge model predictions with observations [17], while the accurate specification of the errors is generally difficult because the source and statistical structure of these errors are often unknown [7]. Adaptive filtering approaches addressing this problem are capable of estimating model and observation error covariance information during the online cycling of a data assimilation system [6]. To date, however, these approaches have received little attention and been rarely applied to hydrological modeling [29]. Here we hence prescribe the error noises explicitly by representing the standard deviations scaling to the current values of the variables, as described in Section 3.3. This treatment holds an implicit assumption that larger values of model and observation variables could introduce greater errors. In actual applications, it is preferable to overestimate rather than underestimate errors as the underestimation may result in filter divergences [5]. So here we examine the impacts of large errors by varying the scaling factors and focus on the parameter (CN$_2$) estimation.

The scaling factors ($S_o$ and $S_m$) for the observation and the model as well as the precipitation ($S_p$) are changed intentionally to represent different error magnitudes. Five cases (from Case 6 to Case 10) are designed based on Case 1 in order to identify their contributions to the assimilation system. Note that all errors in these cases are overestimated compared to the true ones in the error-free synthetic hydrological system.

Fig. 11a illustrates that the assimilation estimates of the CN$_2$s for Case 6 ($S_o = 0.05$) agree with their true values very well, and the result is comparable to that of Case 1 with small errors. When the three scaling factors increase to 0.1 (Case 8), the result is suboptimal as the matching points are symmetrically distributed around the 1:1 line (Fig. 11b); and further the assimilation estimates will only capture synoptic properties of the distribution of true values as the
scaling factors reach 0.4 (Case 10, Fig. 11c). Therefore, it is preferable that the overestimated level of the errors should be kept below 10% (corresponding to the 0.1 for the scaling factors), and larger levels may degrade the performance of EnKF for a catchment hydrological system [7]. These characteristics are further interpreted by Fig. 12 that describes the estimated errors of the \( CN_2 \) for the five cases. As to Case 6, for example, the RMSE and the MAE series drop dramatically within the first 100 assimilation steps, and the final errors approximate to the results of Case 1. Moreover, with the increase of the scaling factor, the series of the RMSE and the MAE would go up gradually.

In addition to the error magnitude, the data assimilation performance also depends on a consistent combination of errors for the forcing boundary (i.e. precipitation in this study), the observation and the model. For example, Case 7, whose scaling factor for runoff observation \( S_o = 0.1 \) is ten times the size of the model and the precipitation \( S_m = S_p = 0.01 \), exhibits quite discrepant estimates. Even though its RMSE series and MAE series follow the levels of Case 6 within the first 100 assimilation steps, they jump up subsequently and keep at high levels (Fig. 12). This property is not unique for Case 7, and
Case 1 also presents oscillatory behaviors as shown in Fig. 8. In fact, more unacceptable results can appear in the cases with inconsistent combination of errors, especially the case with large error magnitudes (e.g., for $S_C = 0.05$, $S_M = 0.01$ and $S_P = 0.05$, not shown). This kind of combination may not render negative impact at the beginning steps, but at the late assimilation steps it spoils the data assimilation performance or even induces filter divergences. Remedy to this issue may be made by prescribing time-correlated errors or using adaptive filter techniques [6,29].

5.2. Initial realization of parameters

Usually according to the prior knowledge of the landuse and soil type in a catchment, the initial realizations can be generated with raw statistical properties including means and variances. If the land cover type is pasture and the soil type is in “B group” with fair hydrological condition in an HRU, for example, the mean of $CN_2$ could be set as 69 and the variance would be small [25]. In practice, this kind of rough prescription may be far away from the true statistics of parameters, particularly when the HRU contains an erratic type of landuse and complex soil formations. So the subjectivity and randomness are often unavoidable. The goal of this subsection is to examine the impact of the biased prior statistics on the estimation results. The mean and variance are taken into account in Case 11, Case 12 and Case 13 whose means and variances are far away from the true values (Table 2).

Fig. 13 illustrates final results of the three cases. It is clear that these assimilation estimates have captured principal properties of the $CN_2$ distribution for the HRUs. Case 11 with relative small initial biases provides the best result while there are no significant differences among the three. This has also been demonstrated by the evolutions of errors. As shown in Fig. 14, the RMSEs and MAEs drop dramatically at the first five assimilation steps even for Case 12 and Case 13 with large initial biases. After 250 assimilation steps, each of them keeps constant approximately and there are small differences mutually. The overestimated errors of the observation, the model and the forcing boundary (Table 2) should be partially responsible for the estimated results, which are not as good as that of Case 1. Therefore, the data assimilation performance is not greatly dependent on the initial guess of the parameters. In addition, compared to Case 13, Case 12 with larger initial biases exhibits less RMSE and MAE within the beginning 90 assimilation steps and even its final errors are not comparable to that of Case 13. So it is preferable to enlarge the magnitudes of the variances if the prior means are far away from the truths (i.e., with poor prior knowledge), especially for a short-period assimilation. Wide range distributions of the initial realizations are beneficial for broadly covering true values and making each realization approach truths by assimilating observations.

5.3. Ensemble size

In the EnKF, a cluster of realizations (samples) are used to approximate the probability distribution of the states. Generally, increasing the ensemble size enables the algorithm to propagate the error information more accurately but at the same time it increases
the computational burden. Moreover, the assimilation estimation results also depend on specified random realizations even with the same ensemble size. Therefore, it is of great importance to explore the appropriate ensemble size to balance the estimation accuracy and the computational efficiency.

On the basis of Case 13, other three cases with ensemble size of 100, 400, 1000 are assembled with 300 assimilation time steps. For each of the four (Case 13 also included), we perform 30 sets of assimilation runs by varying the random seeds to generate different realizations. In addition to the RMSE, another measure of the goodness is the Ensemble Spread, which represents the estimated uncertainty. It should be close to the RMSE if the EnKF estimates the uncertainty of the state properly [4]. The confidence intervals for RMSE and Ensemble Spread at each assimilation step, with the 0.95 confidence intervals, are computed based on the 30 sets of assimilated results.

Table 3 displays the final results of the four cases. As expected, the spatial average RMSE decreases with the increase of ensemble size, while it exhibits slight improvement when the size is in excess of 400. For the case with a small ensemble size (e.g., 100, 200) the Ensemble Spread is smaller than the RMSE, which indicates that the realizations systematically underestimate the uncertainty of the state. This kind of underestimation can be eliminated by increasing the ensemble size. For example, in the case with size = 1000, the Ensemble Spread is close to the RMSE after 50 assimilation steps (Fig. 15). Nevertheless, this improvement comes with the price of a large computational time. With a computer of dual 3.50G HZ processor, the execution time increases from 78 s to 3977 s as the size goes from 100 to 1000, given the 300 time steps. The cost with size = 1000 is not quite expensive for our problems, but the computational burden will increase and become overwhelming if there are more than a hundred of computational units (sub-basins) in a different catchment. Certainly, the reasonable size depends on the nature of the problem. For example, the atmospheric estimation could be performed well with 64 ensemble members [22] and the land surface assimilation may require the size only about 10 [17], while the hydraulic conductivity estimation in geologic formations requires about 200 members [4].

With equal ensemble sizes, furthermore, different sets of parameter realizations will provide different estimation accuracy, especially for the small-sized cases. While the ensemble-spread confidence interval is negligible, there is a certain RMSE confidence interval for every case as shown in Table 3 and Fig. 15, and the case with size = 100 exhibits the broadest interval. This characteristic results from the basic principle of the EnKF which uses finite random samples to represent a notional probability distribution. So it is preferable to try different random realizations and compare their results when we conduct data assimilation with the EnKF.

6. Summary and conclusions

This paper has investigated the performance of the ensemble Kalman filter (EnKF) for the catchment scale hydrological data assimilation. Distributed hydrological models outperform lumped hydrological models in capturing the spatial patterns of the coupled hydrological processes in catchments [28,35], but their high dimensions of states and parameters will pose difficulties for obtaining optimal estimations. Our objective here has been to better understand the capability of combined state-parameter estimation with EnKF and to obtain a more complete understanding of its sensitivities. Accordingly, a data assimilation system has been constructed based on a physically based hydrological model (SWAT), and synthetic assimilation experiments have been performed to estimate a dominant parameter, CN2, as well as the three prognostic states, namely the runoff, soil water content, and evapotranspiration.

It has been found that by assimilating observations, such as the runoff, the EnKF can effectively update the hydrological states and progressively improve the parameter estimation. The hydrologic states (e.g., the soil water content) can be well reproduced even with biased prior knowledge of the parameter. The hydrologic parameters are more difficult to get perfect estimates owing to the nonuniqueness problem in distributed models.

Moreover, the runoff observations from different sites play different roles in the assimilation. The observations from lower streams, particularly from the outlet station, make the greatest contribution to the estimations. So the outlet runoff is a preferred observation for data assimilation if only a limited number of stations are available. In addition, taking the soil water content as complementary observations can obviously improve the inverse estimate of parameters, while the evapotranspiration observations are less effective owing to its weak correlation to the interested parameter (CN2).

Besides the measurement sites and types, the EnKF is also sensitive to other factors about the assimilation system setup. The success of the data assimilation is quite dependent on the error specification, which should be appropriate to the model capability and the measurement flexibility. Not only is the level of error overestimation
limited (less than 10% being recommended), but also the combination of errors (about the model, the observation and the forcing boundary) should be kept consistent to avoid spoiling the assimilation performance. Furthermore, the biased initial estimates of parameters do not pose significant impacts because they can be improved sequentially by assimilating observations, but relatively large variances are preferable when short-term data assimilation is conducted. In addition, before achieving statistical convergence the assimilation results also depend on the particular set of initial realizations because the EnKF is basically a Monte Carlo method. It may be a good idea to compare results from different sets of realizations for the case with a small ensemble size. A reasonable ensemble size should be taken into account to balance estimation accuracy and computational feasibility.

Our analysis is limited by several factors that could be addressed in the future work. In this study, first, the state-parameter estimation was based on the EnKF by concatenating the uncertain parameters into a single joint state vector (state augmentation). This approach highly relies on the state-parameter dependencies. It is not able to provide good estimates if the dependencies are weak (e.g., Case 5 in this study, see Fig. 8). To address this challenge, the novel variants of the EnKF, including the smoothed EnKF [3] and other dual estimate approaches [23], may be useful remedies. Furthermore, we focused only on one dominant parameter estimate in the synthetic data assimilation system by drawing observations from reference simulation and assumed the parameter to obey a Gaussian distribution. In reality, a number of additional parameters and states may control the hydrological processes, and the Gaussian assumption may be violated. These problems will need to be examined in real-world hydrological environment instead of synthetic experiments. In addition, the model and observation errors are assumed to vary with the magnitude of the variables according to the prescribed scaling factors. This simple treatment may be augmented with the adaptive filtering approach that estimates the error information during the online cycling of a data assimilation system [6,29]. Finally, a more general analysis would examine a broad range of observations including the in-situ measurements, multi-scale remote sensing soil moisture, evapotranspiration and land surface temperature. Our ongoing work is addressing some of these issues.

Acknowledgements

We are grateful to the supports by Natural Science Foundation of China (No. 50688901), the Chinese National Basic Research Program (No. 2006CB705800), and the China Postdoctoral Science Foundation (No. 20080440271). We would like to thank Dr. Witold F. Krajewski and the anonymous reviewers for their constructive comments.

References