Available online at www.sciencedirect.com

SCIENCE DIRECT®

Advances in Water Resources 28 (2005) 135–147

Advances in Water Resources

www.elsevier.com/locate/advwatres

# Dual state–parameter estimation of hydrological models using ensemble Kalman filter

Hamid Moradkhani [a,*], Soroosh Sorooshian [a], Hoshin V. Gupta [b], Paul R. Houser [c]

[a] *Department of Civil and Environmental Engineering, University of California, Irvine, CA 92697-2175, USA*
[b] *Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, USA*
[c] *Hydrological Sciences Branch, NASA-GSFC Code 974, Greenbelt, MD 20771, USA*

Received 25 March 2004; received in revised form 20 September 2004; accepted 21 September 2004

**Abstract**

Hydrologic models are twofold: models for understanding physical processes and models for prediction. This study addresses the latter, which modelers use to predict, for example, streamflow at some future time given knowledge of the current state of the system and model parameters. In this respect, good estimates of the parameters and state variables are needed to enable the model to generate accurate forecasts. In this paper, a dual state–parameter estimation approach is presented based on the Ensemble Kalman Filter (EnKF) for sequential estimation of both parameters and state variables of a hydrologic model. A systematic approach for identification of the perturbation factors used for ensemble generation and for selection of ensemble size is discussed. The dual EnKF methodology introduces a number of novel features: (1) both model states and parameters can be estimated simultaneously; (2) the algorithm is recursive and therefore does not require storage of all past information, as is the case in the batch calibration procedures; and (3) the various sources of uncertainties can be properly addressed, including input, output, and parameter uncertainties. The applicability and usefulness of the dual EnKF approach for ensemble streamflow forecasting is demonstrated using a conceptual rainfall-runoff model.


*Keywords:* Streamflow forecasting; Stochastic processes; Data assimilation; Ensemble Kalman filter; Dual estimation; Kernel smoothing

## 1. Introduction and scope

Hydrologic models are defined largely by parameters and states, parameters being physical and generally time-invariant descriptions of surface and subsurface characteristics, and states being fluxes and storages of water and energy that are propagated in time by the model physics. In practice, in addition to model simulation, reliable operation of a watershed system requires a continuous correction of the forecast as observational data become available. This entails the critical need to extend the applicability of data assimilation in hydrology as emphasized by Troch et al. [39]. However, the successful use of data assimilation relies on unbiased model state prediction, which is largely dependent on accurate parameter estimation. During the past two decades, much effort has been directed toward the estimation of hydrologic model parameters (calibration) to improve the forecast accuracy [7,8,11,32]. Conceptual hydrologic models are usually deterministic representations, which typically do not contain descriptions of the various sources of uncertainties. Although it has been common to translate the inability of a model to generate accurate streamflow forecasts into parameter uncertainty, other sources of uncertainties, such as model structural error, input, and output measurement errors, also need to be accounted for [16,17]. Several

* Corresponding author. Tel.: +1 949 824 8821; fax: +1 949 884 8831.
*E-mail address:* moradkha@uci.edu (H. Moradkhani).

0309-1708/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2004.09.002

authors have studied the uncertainties associated with parameter estimation, and procedures have been developed for the statistical analyses of parameter uncertainties [18–20,33,34,42,43].

The aforementioned calibration procedures generally minimize long-term prediction error using a historical batch of data assuming time-invariant parameters, and thus make no attempt to include information from new observations. Batch calibration requires a set of historical data to be kept in storage and processed enmasse while lacking the flexibility to investigate possible temporal evolution of the model parameters. Thiemann et al. [35] emphasized another limitation of batch calibration in hydrological prediction of an ungauged watershed where the lack of sufficient historical data makes the batch method infeasible. These limitations, as well as an interest in inferring the uncertainty in the estimated parameters, motivated Thiemann et al. [35] and Misirli et al. [25] to develop a recursive scheme for model prediction and parameter estimation in the online mode. From another perspective, Kitanidis and Bras [15] stated that adaptive estimation might be suitable when the forecast lead-time is short in comparison to the response time of the watershed. They explained that it would be the case when the error in input is large while the error in output measurement is small.

Much of the efforts in simulation-based methods of hydrologic system analyses have been focused on (1) improved methods for parameter estimation wherein state variable uncertainties were not explicitly taken into account or (2) improved procedures for estimating time-varying state variables wherein the parameters were assumed to be known in advance. The commonly used batch calibration techniques only address parameter uncertainty while uncertainties in input, output and model structure are ignored. The main weakness of such approaches is that they attribute all errors from input, output and model structure to model parameter uncertainty. Sequential data assimilation procedures have the potential to overcome this drawback in simulation-based methods by explicitly taking into account all the sources of uncertainty. The Kalman filter [14], a recursive data-processing algorithm, is the most commonly used sequential data assimilation technique, which results in optimal estimation for linear dynamic models with Gaussian uncertainties.

Although filtering techniques can address the various sources of uncertainties in modeling, the typical presumption of these procedures is that the parameters are to be specified in advance and sequential estimation is applied only to the state variables. Because there is no guarantee that model behavior does not change over time, the model adjustment through the time variation of parameters together with state variables is incisive. Therefore a procedure that can provide the simultaneous estimate of states and parameters is required. The development of interactive (dual) state–parameter estimation using standard Kalman filter, in the context of hydrology, is traced back to [36,37] and later to the joint state–parameter by state augmentation technique [3,4] (see Section 3 for detail). Those techniques, however, were limited to linear dynamic systems. For nonlinear dynamics, the extended Kalman filter (EKF), which relies on linearization of model using first order approximation of Taylor series, can be used. As reported by Refs. [9,29,30] the EKF can lead to unstable results when the nonlinearity in the system is strong. To cope with the drawbacks of the EKF, a Monte Carlo-based Kalman filter called ensemble Kalman filter (EnKF) was introduced by Evensen [9]. One of the advantages of the EnKF comparing to the standard KF is that the estimation of priori model covariance (see Section 2.1) is not needed for the updating (analysis) step although its calculation using the model ensemble is straightforward.

The EnKF was originally developed for dynamic state estimation while in this paper its applicability to static state (parameter) estimation by dual state–parameter estimation strategy is extended and its usefulness on streamflow forecasting is examined.

The organization of the paper is as follow. In Section 2, the general framework for sequential data assimilation is explained, where the mathematical formulation of the EnKF as a special type of Monte Carlo procedure for state estimation is elaborated. A systematic approach for identifying the perturbation factor, as a key feature in the EnKF, and for tackling the uncertainties in forcing data (input) and observation (output) is suggested. In Section 3, the dual EnKF algorithm that deals simultaneously with both model parameters and state variables is explained and kernel smoothing of parameters is employed for parameter sampling to avoid the over-dispersion of parameters through random walk. In Section 4, the applicability of dual EnKF on a conceptual rainfall-runoff model and the power of this algorithm in streamflow forecasting is demonstrated.

## 2. General framework for sequential data assimilation

Over the past decade, a rapid increase in earth system science data assimilation activities has been witnessed. Similarly, hydrologic data assimilation techniques have garnered a great deal of attention of hydrologists in the sense that by taking advantage of real time observation, more accurate forecast can be made [22,23,29,30,39,40]. The mathematical framework of estimation theory provides the tools required to approach a variety of data assimilation problems. The basic objective of data assimilation is to characterize the state of a system at some future time from the knowledge of the initial state. The state of a hydrological system, $x_t$, at time $t$

could be conditioned on the observation, $y_{1:t}$, through the probability density function:

$$P(x_t \mid y_{1:t}) \tag{1}$$

Following [13], the generic discrete-time nonlinear stochastic-dynamic system can be expressed in the form of:

$$x_{t+1} = f(x_t, u_t, \theta) + \omega_t, \quad \omega_t \sim N(0, \Sigma_t^m) \tag{2}$$

where $x_t$ is an $n$-dimensional vector representing the system state variables (for example catchment soil moisture content) at time $t$. The nonlinear propagator $f(.)$ contains the model input vector (deterministic forcing data, $u_t$, e.g., mean areal precipitation), and the (possibly) time-invariant model parameter $\theta$. The model error is displayed by $\omega_t$ with covariance $\Sigma_t^m$ and represents all the uncertainties related to model structure and the forcing data. Cohn [6] explained that the model error is generally state-dependent even if the operator $f(.)$ is linear. The state-dependence and even dependence upon the parameters as part of the uncertain propagator cause the model error to be unknown. For simplicity, it is therefore appropriate to represent the model error as a stochastic perturbation in Eq. (2).

Suppose that a set of scalar observations is taken at time $t+1$ and that we intend to assimilate the vector of observations into the model. The output variables of the model are functions of both the model state variables and the parameters characterizing the model. The observation process in general form can be written as:

$$y_{t+1} = h(x_{t+1}, \theta) + v_{t+1}, \quad v_{t+1} \sim N(0, \Sigma_{t+1}^y) \tag{3}$$

where propagator $h(.)$ relates the state variables to the measured variables (in our case streamflow) and yields the expected value of the prediction given the model states and parameters. All sources of errors in the observation are reflected by $v_{t+1}$, which will be assumed here to be Gaussian and independent of model error $\omega_t$.

### 2.1. Ensemble Kalman filter (EnKF)-state estimation

Sequential data assimilation, also known as filtering, consists of model state estimation at each observation time based only on the observations up to present. In the linear case, this problem is solved by the well-known Kalman filter [14] as an optimal recursive data-processing algorithm. In the case of nonlinear dynamics, one can linearize the current state vector to use the so-called extended Kalman filter (EKF) [13]. The EKF has many well-known drawbacks such as computational demand owing to the error covariance propagation and closure approximation by neglecting the higher order derivatives of the model, which correspondingly may produce instabilities or even divergence [10,13,24]. The ensemble Kalman filter (EnKF) as an alternative to the traditional EKF was first introduced by Evensen [9] and later clarified by Burgers et al. [5] and Van Leewen [41]. The EnKF is based upon Monte Carlo or ensemble generations where the approximation of forecast (a priori) state error covariance matrix is made by propagating an ensemble of model states using the updated states (ensemble members) from the previous time step. The key point in the performance of the EnKF according to [5,29,30] is to generate the ensemble of observations at each update time by introducing noise drawn from a distribution with zero mean and covariance equal to the observational error covariance matrix; otherwise the updated ensemble will possess a very low covariance. A schematic representation of the EnKF is demonstrated in Fig. 1. As seen in Fig. 1, the EnKF propagates an ensemble of state vectors in parallel such that each state vector represents one realization of generated model replicates. Similar to Eq. (2), the model forecast is made in the EnKF for each ensemble member as follows:

$$x_{t+1}^{i-} = f(x_t^{i+}, u_t^i, \theta, t) + \omega_t^i, \quad i = 1, \ldots, n \tag{4}$$

where $x_{t+1}^{i-}$ is the $i$th ensemble member forecast at time $t+1$ and $x_t^{i+}$ is the $i$th updated ensemble member at time $t$. In addition to representing the additive process noise, which is common in standard Kalman filtering, the EnKF represents the multiplicative model errors through forcing data perturbations. The forcing data perturbations are made by adding the $\zeta_t^i$ noise with covariance $\Sigma_t^u$ to the forcing data at each time step:

$$u_t^i = u_t + \zeta_t^i, \quad \zeta_t^i \sim N(0, \Sigma_t^u) \tag{5}$$

Now, we form the expression for the error covariance matrix associated with the forecasted (a priori) estimate. If the true state variables are known, we can use the following expectation to estimate the a priori model error covariance:

$$P_{t+1}^- = E[(x_{t+1}^- - x_{t+1}^{\text{true}})(x_{t+1}^- - x_{t+1}^{\text{true}})^{\mathrm{T}}] \tag{6}$$

However, because the true state is generally unknown, it is convenient to calculate the ensemble covariance matrix:

$$P_{t+1}^- = E[X_{t+1} X_{t+1}^{\mathrm{T}}] = \frac{1}{n-1} X_{t+1} X_{t+1}^{\mathrm{T}} \tag{7}$$

where, $X_{t+1} = [x_{t+1}^{i-} - \bar{x}_{t+1}^-, \ldots, x_{t+1}^{n-} - \bar{x}_{t+1}^-]$ and $\bar{x}_{t+1}^- = E[x_{t+1}^{i-}] = \frac{1}{n}\sum_{i=1}^n x_{t+1}^{i-}$.

The updated (a posteriori) error covariance could be estimated similarly after updating all of the ensemble members.

With the assumption of a priori estimate (forecasted states $x_{t+1}^{i-}$), we now seek to use the observation $y_{t+1}$ to obtain the posterior estimate (updated states $x_{t+1}^{i+}$). A linear correction equation is used according to standard Kalman filter to update forecasted state ensemble members:

$$x_{t+1}^{i+} = x_{t+1}^{i-} + K_{t+1}(y_{t+1}^i - \hat{y}_{t+1}^i) \tag{8}$$

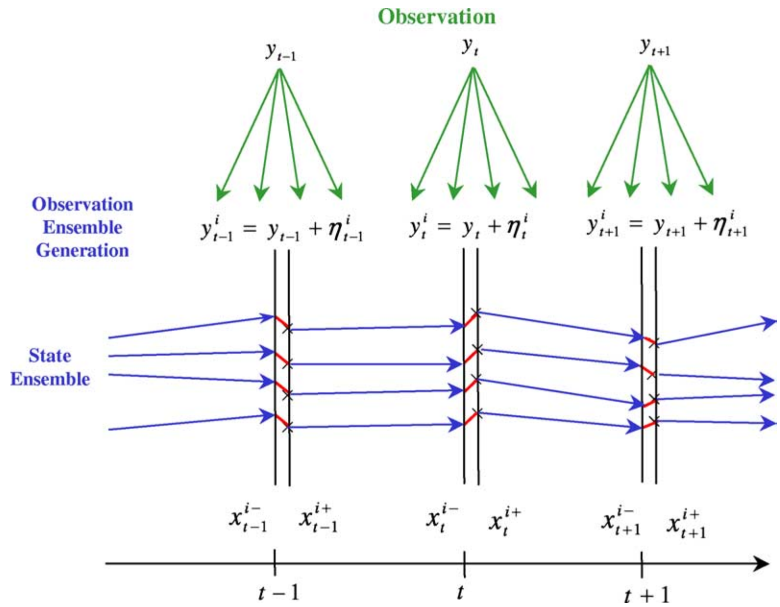


Fig. 1. Ensemble Kalman filter (EnKF) schematic. $x^{i-}_t$: Forecasted state ensemble member and $x^{i+}_t$: updated state ensemble member.

where $y^i_{t+1}$ is the $i$th trajectory of the observation replicates generated by adding the noise of $\eta^i_{t+1}$, with covariance $\Sigma^y_{t+1}$, to the actual observation:

$$y^i_{t+1} = y_{t+1} + \eta^i_{t+1}, \quad \eta^i_{t+1} \sim N(0, \Sigma^y_{t+1}) \tag{9}$$

This is one of the features of the EnKF in which observations (Eq. (9)) are treated as random variables by generating an observation ensemble with mean equal to the actual observation at each time and a predefined covariance. One may want to consider an alternate strategy in updating step (Eq. (8)) by using the ensemble square root filter (EnSRF) [47] such that the perturbation of observation is not needed. Whitaker and Hamill [47] justified the applicability of EnSRF to the linear observation models while, in this study we are interested in filtering the nonlinear model dynamics; hence we develop our strategy according to the version of the EnKF that treats the observation as a random variable and perturbation of observation is required [5,29,30]. Therefore, in the next section, we will elaborate on a systematic procedure to tune the magnitude of the forcing data and observation covariances in order to generate a reliable ensemble while ensemble size can be determined correspondingly.

Similarly, $\hat{y}^i_{t+1}$ is the $i$th predictive variable at time $t+1$:

$$\hat{y}^i_{t+1} = h(x^{i-}_{t+1}, \theta) \tag{10}$$

In Eq. (8), $K_{t+1}$ is the Kalman gain matrix which, in adaptation to the ensemble based approach can easily be proven to be as:

$$K_{t+1} = \Sigma^{xy-}_{t+1}[\Sigma^{yy}_{t+1} + \Sigma^y_{t+1}]^{-1} \tag{11}$$

where $\Sigma^{yy}_{t+1}$ is the forecast error covariance matrix of the prediction $\hat{y}^i_{t+1}$, and $\Sigma^{xy-}_{t+1}$ is the forecast cross covariance of the state variables $x^{i-}_{t+1}$ and prediction $\hat{y}^i_{t+1}$. The above form of Kalman gain is its modified version of the standard Kalman gain represented as ($K_{t+1} = P^-_{t+1}H^{\mathrm{T}}[HP^-_{t+1}H^{\mathrm{T}} + R_{t+1}]^{-1}$) where $P^-_{t+1}$ is defined in (6), $H$ is the observation transition operator after linearization of observation model (2) and $R_{t+1}$ is the same as $\Sigma^y_{t+1}$ defined by (9). One of the advantages of EnKF is that the estimation of $P^-_{t+1}$ is not needed (although possible as explained in (7)), whereas its estimation in the standard KF is necessary.

### 2.2. Identification of hyper-parameters and estimation of ensemble size

In general, the performance of most ensemble forecasts (EF) is influenced by the quality of the ensemble generation method, the forecast model and also the analysis scheme. A large number of procedures exist to evaluate the ensemble forecasts [12,27,38,48]. The key feature of the EnKF, however, is the perturbation of forcing data to generate replicates of the model state variables, and then the correction of the forecasted ensemble members through the analysis (update) step (Eq. (8)). A question may arise on how to perturb the system to construct a reliable ensemble where the spread of the ensemble is within a meaningful range. Another issue in EF is the efficiency of the procedure, which is highly related to the ensemble size. As shown by Eqs. (5) and (9), the perturbation of forcing data and observation (input and output) are made by adding noise to the variable of interest. A fundamental limitation here is connected with the identifiability of the noise variance such that it can tackle the uncertainty in input and output. Stochastic noises are assumed to be Gaussian with predetermined variances, which are assumed to be heteroscedastic (variance changing) [31]. Therefore the variance of the noises introduced to the input and output variables (Eqs. (5) and (9)) are proportional to the magnitude of the variables as follows:

$$\Sigma_t^u = \gamma \cdot u_t \tag{12}$$

$$\Sigma_{t+1}^y = \rho \cdot y_{t+1} \tag{13}$$

Here, we call the proportionality factors of $\gamma$ and $\rho$ as hyper-parameters.

The magnitudes of these unknown hyper-parameters determine the ensemble spread. A model's failure to properly fit the observations is a measure of model error, for which an obvious approximation is a comparison between the spread of an ensemble and the ensemble mean forecast error. The deficiency in spread might be a measure of the uncertainty associated with the ensemble mean. Anderson [1] discussed a simple procedure based on [27] for evaluating the similarity of truth versus randomly selected members of the ensemble. According to this method, the ratio of the time-averaged RMSE of the ensemble mean, $R_1$, to the mean RMSE of the ensemble members $R_2$, is calculated:

$$Ra = \frac{R_1}{R_2} \tag{14}$$

$$R_1 = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\left[\left(\frac{1}{n}\sum_{i=1}^{n}\hat{y}_t^i\right) - y_t^i\right]^2} \tag{15}$$

$$R_2 = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t^i - y_t^i)^2} \tag{16}$$

where $n$ and $T$ are the ensemble size and period of analysis respectively.

If the actual observation is statistically indistinguishable from $n$ ensemble members the expected value of the RMSE ratio $Ra$, as explained in [1,27] should be

$$E[Ra] = \sqrt{\frac{(n+1)}{2n}} \tag{17}$$

The ratio of $Ra$ to $E[Ra]$ is referred to as the Normalized RMSE Ratio (NRR),

$$\mathrm{NRR} = \frac{Ra}{E[Ra]} \tag{18}$$

In order to evaluate the ensemble performance, the normalized RMSE ratio is used, while NRR > 1 indicates that the ensemble has too little spread, and NRR < 1 is an indication of an ensemble with too much spread. Ideal ensemble generation should produce a NRR value close to unity. As seen in Eqs. (13) and (14), hyper-parameters control the ensemble spread through the perturbation variance. Tuning of these new parameters results to the meaningful ensemble generation while input and output errors are taken into account.

### 2.3. Tuning of hyper-parameters

To demonstrate the EnKF hyper-parameter tuning procedure presented above, the conceptual Hydrologic MODel (HyMOD) described by Refs. [2,44] (see Fig. 2) was used. HyMOD originates in the probability distributed moisture model (PDM) [26], an extension of some of the lumped storage models developed in 1960s, and later to the case of multiple storages representing a spatial distribution of different storage

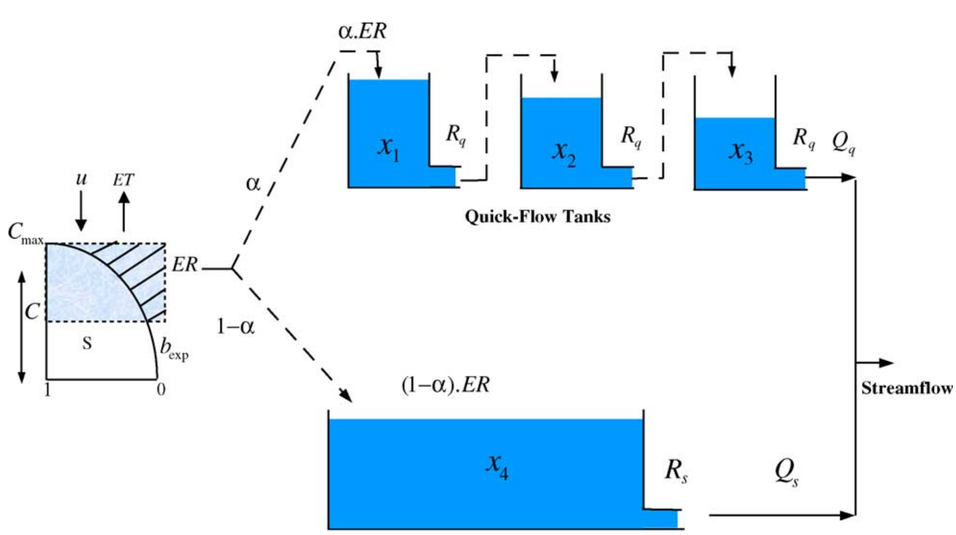


Fig. 2. Hydrologic MODel (HyMOD) conceptualization.

capacities in a watershed. Boyle [2] described HyMOD as a rainfall excess model through a nonlinear tank connected with two series of linear tanks (three identical quick-flow tanks) in parallel to a slow-flow tank representing the groundwater flow. From the definitions of state variables given in Section 2, state variables in this system are $S$: storage in the nonlinear tank representing the watershed soil moisture content, $x_1$, $x_2$ and $x_3$: the quick-flow tank storages representing the temporary (short-time) detentions, e.g., depression storages, and $x_4$: the slow-flow tank storage (subsurface storage). Correspondingly parameters of this model are $C_{\max}$, as the maximum storage capacity within the watershed, $b_{\exp}$, the degree of spatial variability of the soil moisture capacity within the watershed, $\alpha$, a factor for partitioning the flow between two series of tanks, $R_q$ and $R_s$ as the residence time parameters of quick-flow and slow-flow tanks, respectively.

The above procedure was applied to tune the hyper-parameters and ensemble size for streamflow forecasting of the Leaf River watershed, a humid watershed with an area 1944 km$^2$ area located north of Colins, Mississippi which has been a test basin in numerous studies [2,11,32,35,42]. The data consist of potential evapotranspiration, ET (mm/d), mean areal precipitation, $P$ (mm/d) as forcing data, and streamflow (m$^3$/s) as observation.

According to Eq. (19), NRR was calculated for a range of observation perturbation factor (observation hyper-parameter) and forcing data perturbation (input hyper-parameter) with 50 ensemble members, where four combinations of input and observation noise magnitudes were investigated (Fig. 3). The light shaded area in each subplot shows the acceptable bound for NRR and a range of acceptable hyper-parameters. As seen, NRR is more sensitive to the observation hyper-parameter than the input hyper-parameter, implying that accurate estimation using HyMOD model is highly dependent on the observation replicate generation.

To further investigate the variation of NRR with respect to ensemble size, NRR was examined for a range of observation hyper-parameter ($\rho$ = 5–25%) and input hyper-parameter ($\gamma$ = 10%) (Fig. 4). As seen in case A of Fig. 3, the minimum noise in input and observation

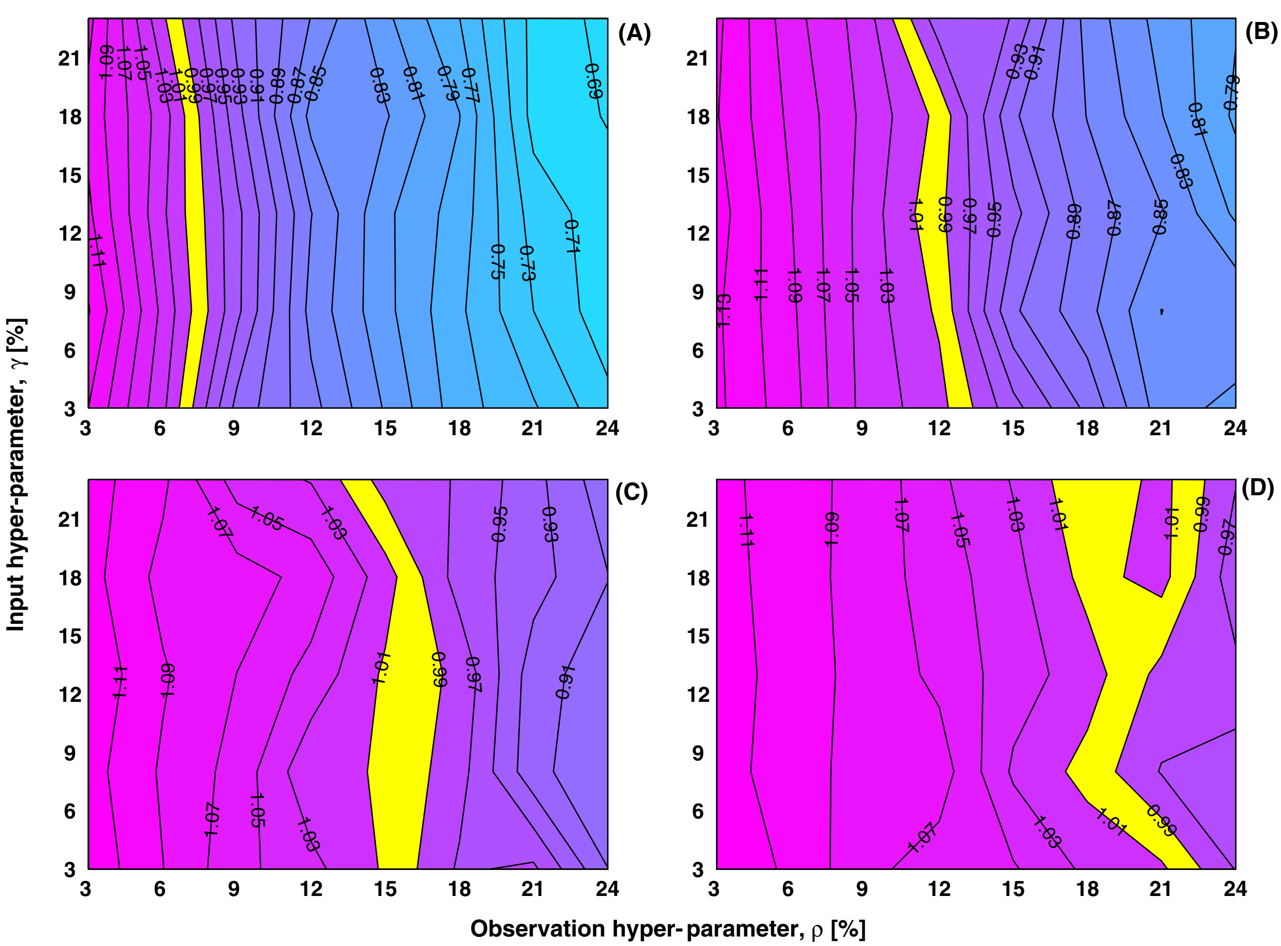


Fig. 3. NRR space with respect to hyper-parameters with 50 ensemble members for different noise scenarios, (A) input noise = 5%, observation noise = 5%, (B) input noise = 20%, observation noise = 5%, (C) input noise = 5%, observation noise = 20%, and (D) input noise = 20%, observation noise = 20%.

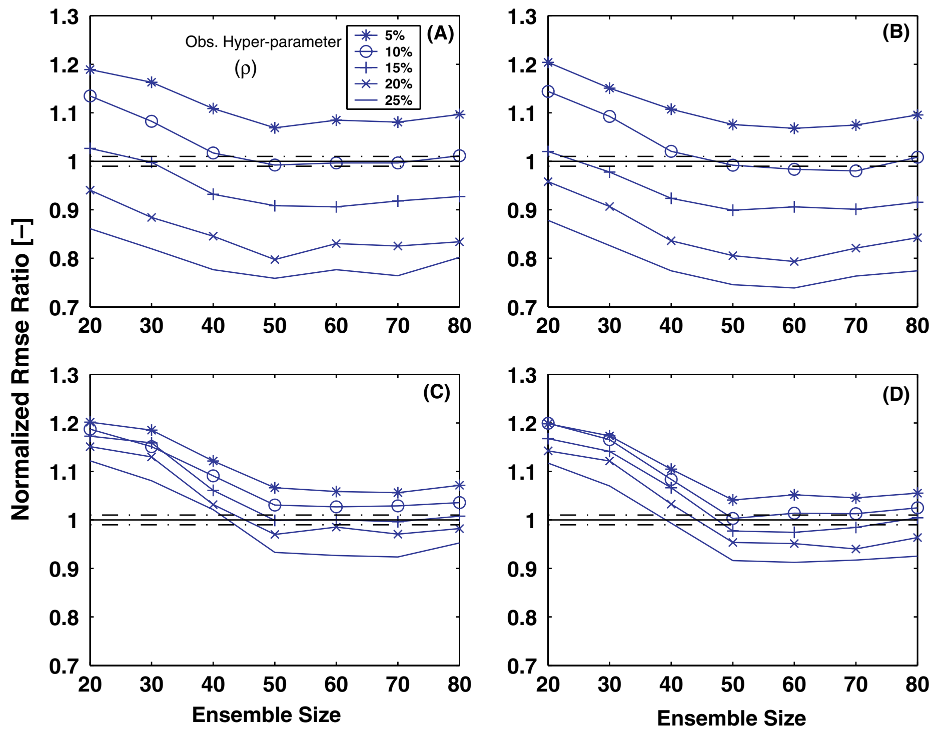


Fig. 4. Normalized RMSE ratio (NRR) against ensemble size for input hyper-parameter (input perturbation factor, $\gamma = 10\%$) and different observation hyper-parameter (observation perturbation factors, $\rho = 5$–$25\%$). Four input and output error combinations: (A) Input noise = 5%, observation noise = 5%, (B) input noise = 20%, observation noise = 5%, (C) input noise = 5%, observation noise = 20%, and (D) input noise = 20%, observation noise = 20%. Dash lines demonstrate the upper and lower bound of acceptable NRR (0.99–1.01) for this study.

(each 5%), the minimum ensemble size of 40 and observation hyper-parameter, $\rho = 10\%$ will keep the NRR within the acceptable range (0.99–1.01). Although for $\rho = 15\%$ with the ensemble size of 30, NRR still lies within the acceptable range, the reliability on the ensemble size for this case is low because by increasing the ensemble size the NRR drops suddenly, implying that the ensemble spread becomes too large. By increasing the input noise in case B, and keeping the observation noise same as A, minor changes in NRR with respect to observation hyper-parameter $\rho$ and ensemble size are seen, whereas in case C, by just increasing the observation noise and having the input noise the same as case A, more significant changes in the output hyper-parameter and required ensemble size is seen.

The above procedure to quantify the input and output perturbation factors and their impact on ensemble generation has been carried out separately from the uncertainty associated with state and parameter estimates. In the following section, an interactive procedure in the context of EnKF to provide a probabilistic estimate of states and parameters is developed.

## 3. Dual state–parameter estimation with EnKF

Although the parameters of a hydrologic model can be estimated in a batch-processing scheme, there is no guarantee that model behavior does not change over time; therefore model adjustment over time may be required. Additionally, due to the multiplicative nature of errors in forcing data and observation, it is prudent to assemble the parameter adaption in the state evolution and forecasting system [40]. The need for real time state–parameter estimation of hydrological models is not free from empiricism and has been reported in several studies [3,4,36,37,40].

Section 2.1 illustrated that recursive state estimation in a stochastic-dynamic system is carried out such that the parameters are assumed to be a time-invariant system description. In this section, we consider the combined estimation problem, in which both model state variables and parameters are estimated simultaneously given erroneous forcing data and observations. One approach for combined estimation is provided by joint estimation where state and parameter vectors are concatenated into a single joint state vector (state augmentation) [3,4,29,36,37,40]. The drawback of such a strategy is that, by increasing the number of unknown model states and parameters, the degree of freedom in the system increases and makes the estimation unstable and intractable especially in the nonlinear dynamic models. An alternative approach to joint estimation is dual estimation, designed as two interactive filters motivated either by the need to estimate state from the model (parameters) or by the need to estimate the model from state. Examples of dual estimation are the dual extended Kalman filter (EKF) developed by Refs. [28,45] for estimating neural networks model signal (state) and weights (parameter). The dual EKF requires separate state-space representation for the state and parameters through two parallel filters. To extend the applicability of the EnKF to simultaneous state–parameter estimation, we need to treat the parameters similar to state variables with a difference that parameter evolution is set up artificially, i.e., it is assumed that the parameters follow a random walk; therefore, in EnKF, parameter samples can be made as follows:

$$\theta_{t+1}^{i-} = \theta_t^{i+} + \tau_t^i, \quad \tau_t^i \sim N(0, \Sigma_t^\theta) \tag{19}$$

Using the artificially forecasted parameters and forcing data replicates, a model state ensemble and predictions are made, respectively:

$$x_{t+1}^{i-} = f(x_t^{i+}, u_t^i, \theta_{t+1}^{i-}) \tag{20}$$

$$\hat{y}_{t+1}^{i} = h(x_{t+1}^{i-}, \theta_{t+1}^{i-}) \tag{21}$$

Updating the parameter ensemble members is made according to the standard Kalman filter equation:

$$\theta_{t+1}^{i+} = \theta_{t+1}^{i-} + K_{t+1}^{\theta}(y_{t+1}^i - \hat{y}_{t+1}^i) \tag{22}$$

here, $K_{t+1}^{\theta}$ is the Kalman gain for correcting the parameter trajectories and is obtained by:

$$K_{t+1}^{\theta} = \Sigma_{t+1}^{\theta y}[\Sigma_{t+1}^{yy} + \Sigma_{t+1}^{y}]^{-1} \tag{23}$$

here $\Sigma_{t+1}^{\theta y}$ is the cross covariance of parameter ensemble and prediction ensemble.

Now using the updated parameter, the new model state trajectories (state forecasts) and prediction trajectories are generated:

$$x_{t+1}^{i-} = f(x_t^{i+}, u_t^i, \theta_{t+1}^{i+}) \tag{24}$$

$$\hat{y}_{t+1}^{i} = h(x_{t+1}^{i-}, \theta_{t+1}^{i+}) \tag{25}$$

Model states ensemble is similarly updated as follows:

$$x_{t+1}^{i+} = x_{t+1}^{i-} + K_{t+1}^{x}(y_{t+1}^i - \hat{y}_{t+1}^i) \tag{26}$$

where $K_{t+1}^{x}$ is the Kalman gain for correcting the state trajectories and is obtained by:

$$K_{t+1}^{x} = \Sigma_{t+1}^{xy}[\Sigma_{t+1}^{yy} + \Sigma_{t+1}^{y}]^{-1} \tag{27}$$

where, $K_{t+1}^{x}$ is the cross covariance of states ensemble and prediction ensemble.

### 3.1. Kernel smoothing of parameter samples

The artificial parameter evolution at each time step by adding small random perturbation provides a new parameter set in simulation and has been performed by many authors, from which [36] is one of the earliest in hydrologic application. The drawback of such parameter sampling is the over-dispersion of parameter samples and loss of information between time points when the parameters are considered to be fixed. In other words, loss of information results in posterior distribution of parameters that are too diffuse when compared to the posteriors of fixed parameters [21]. One remedy to this problem is the Kernel smoothing of parameter samples introduced by West [46]. Suppose that the $\theta_t^i$, and their weights $w_t^i$, $i = 1, \ldots, n$ denote a random measure $\{\theta_t^i, w_t^i\}$ to characterize the discrete Monte Carlo approximation to posterior density of parameters as $P(\theta_{t+1}|y_1, \ldots, y_t)$ with mean $\bar{\theta}_t$ and the variance matrix $V_t$. West [46] explained that the smooth kernel density form of the Monte Carlo approximation to $P(\theta_{t+1}|y_1, \ldots, y_t)$ could be a mixture of Gausian densities with mean $m_t^i$ and variance $h^2 V_t$, weighted by sample weights $w_t^i$:

$$P(\theta_{t+1} \mid y_1, \ldots, y_t) \sim \sum_{i=1}^{n} w_t^i N(\theta_{t+1} \mid m_t^i, h^2 V_t) \tag{28}$$

where, $h$ is the smoothing or variance reduction parameter. The standard kernel method considers that $m_t^i = \theta_t^i$, however, this results to an over-dispersed kernel density relative to posterior samples. West [46] and later Liu [21] suggested that this flaw can be corrected by shrinkage of kernel locations:

$$m_t^i = a\theta_t^i + (1-a)\bar{\theta}_t \quad \text{with } a = \sqrt{1-h^2} \tag{29}$$

If the Monte Carlo approximation to posterior density $P(\theta_{t+1}|y_1, \ldots, y_t)$ has mean $\bar{\theta}_t$ and variance matrix $V_t$, the parameter evolution in Eq. (19) with independent perturbation $\tau_t^i$, which was assumed to be independent of $\theta_t$, has the correct mean $\bar{\theta}_t$ but variance matrix $V_t + \Sigma_t^\theta$. This problem was reported as loss of information in [21,46]. Therefore Liu [21] showed that the artificial evolution needs to be modified by considering the correlations between $\theta_t$ and the perturbation $\tau_t^i$. By doing so the conditional evolution density of parameters is written as follows:

$$P(\theta_{t+1} \mid \theta_t) \sim N(\theta_{t+1}^i \mid a\theta_t^i + (1-a)\bar{\theta}, h^2 V_t) \tag{30}$$

where, $a = \frac{3\delta - 1}{2\delta}$ and $\delta$ is a factor in (0 1], which is typically around 0.95–0.99. For more information about the derivation of the conditional density and the parameters associated with it, please refer to Liu [21].

A dual state–parameter estimation flowchart using EnKF with kernel smoothing of parameters is shown in Fig. 5.

## 4. Streamflow forecasting by applying dual EnKF on HyMOD model

The applicability and usefulness of the dual EnKF on state–parameter estimation of HyMOD for one-day

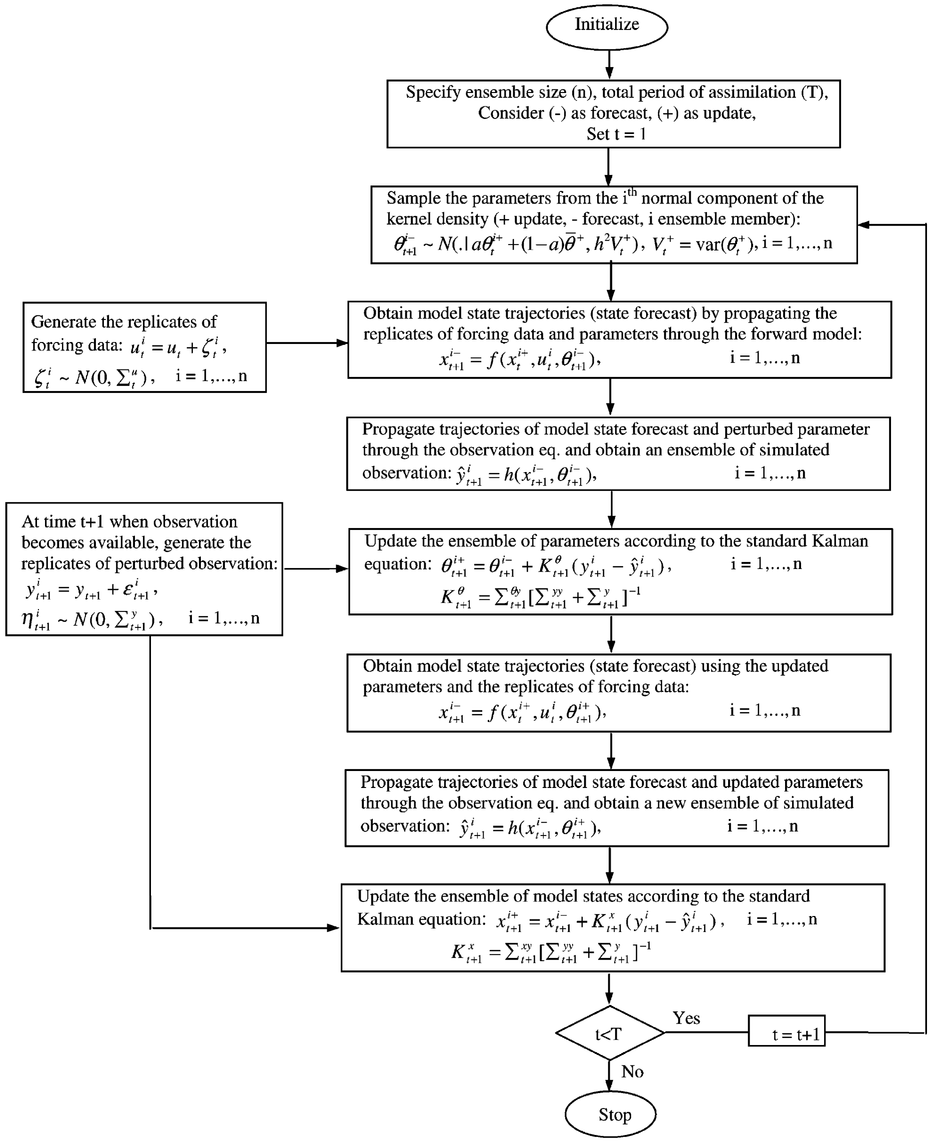


Fig. 5. Dual state–parameter estimation flowchart using the ensemble Kalman filter by kernel smoothing of parameters.

ahead streamflow forecasting in the Leaf River basin were investigated. The system was initialized by defining the prior uncertainty range associated with the parameters in Table 1. With regard to state variables, storages in the linear tanks have no threshold, and storage in nonlinear tank limited to the minimum and maximum bound defined for the nonlinear tank parameters which is found from model formulation [2] to be between 60 and 320. The starting point in the parameter space is sampled from the uniform distribution, then forecast-updates of all state variables and parameters are made simultaneously using the dual estimation. Owing to the stochastic-dynamic nature of the problem, it is required to run the model for the sufficient number of parameter samples to examine the time evolution of predictive uncertainties.

Table 1
Prior uncertainty associated with parameters in HyMOD model

| Parameter | Description | Minimum | Maximum |
|---|---|---|---|
| $R_q$ | Residence time for quick-flow tanks | 0.20 | 0.70 |
| $R_s$ | Residence time for slow-flow tank | 0.01 | 0.10 |
| $\alpha$ | Partitioning factor between tanks | 0.60 | 0.99 |
| $b_{exp}$ | Spatial variability of soil moisture capacity | 0.10 | 1.50 |
| $C_{max}$ | Maximum storage capacity of watershed | 150.00 | 350.00 |

For this experiment, 500 starting points were sampled in the parameter space, and dual ensemble filtering with ensemble size of 50 (resulted from the tuning of hyperparameters) was performed from each starting point. Fig. 6 displays the time evolution of HyMOD parameters for the water years of 1950–1953. Shaded areas in this figure represent the evolution of confidence intervals obtained from 500 trajectories, while each trajectory is the mean of 50 ensemble members. As seen all parameters converge smoothly to the certain region in parameter space where the uncertainty bounds stabilize.

It also appears from Fig. 6 that quick-flow tank parameter $R_q$ is the most identifiable parameter by showing the fastest convergence with a minimum degree of uncertainty comparing to the others. In contrast, the maximum storage capacity of the watershed displayed by $C_{max}$ is less identifiable than the others and shows the slowest convergence. It is apparent from the model configuration (Fig. 2) that $C_{max}$ and $b_{exp}$ are in high interaction such that one compensates for another, that is, the uncertainty in $C_{max}$ can be compensated with suitable degree of convexity or concavity of the nonlinear reservoir represented by $b_{exp}$ to provide the most accurate excess rainfall possible for parallel tanks. As a comparison, the expected values of parameters obtained using different algorithms are shown in Table 2. The last three columns in Table 2 give the algorithms developed at University of Arizona from which the SCE-UA [7,8] and SCEM-UA [42] are the global optimization algorithms suitable for the batch calibration of hydrologic model parameters. BaRE algorithm [25,35] in the last column is a Bayesian recursive estimation technique, which also investigates the time evolution of parameter probabilities.

Table 2
Expected value of the parameter sets in HyMOD estimated by different algorithms

| Parameter | Estimation algorithm | | | |
|---|---|---|---|---|
| | SCE-UA | SCEM-UA | BaRE | Dual EnKF |
| $R_q$ | 0.465 | 0.46 | 0.478 | 0.463 |
| $R_s$ | 0.01 | 0.0027 | 0.0295 | 0.0127 |
| $\alpha$ | 0.861 | 0.84 | 0.667 | 0.82 |
| $b_{exp}$ | 0.251 | 0.38 | 0.15 | 0.406 |
| $C_{max}$ | 282.51 | 257.33 | 181.91 | 258.34 |

Although the dual EnkF result is comparable with other algorithms and, to a higher extent, with the batch calibration schemes, it has some advantages over the above models such as:

(1) The capability of dual EnKF in interactive parameter and state estimation in which the updated parameters at each time step are used to update the model state. As an example, the estimation of one of the state variables as the storage of nonlinear tank, conceptually representing the watershed soil moisture storage, is demonstrated in Fig. 7. This is an unobservable quantity and the accuracy of its estimation is translated through the accuracy of streamflow forecasting as the observable and predictable variable.

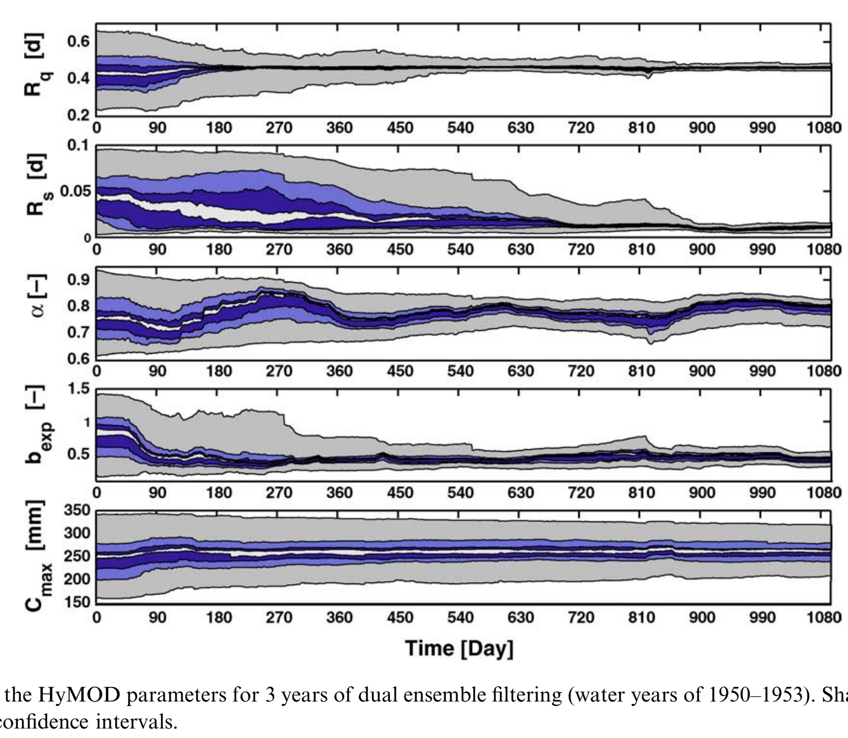


Fig. 6. Time evolution of the HyMOD parameters for 3 years of dual ensemble filtering (water years of 1950–1953). Shaded areas correspond to 95, 75, 68 and 10 percentile confidence intervals.

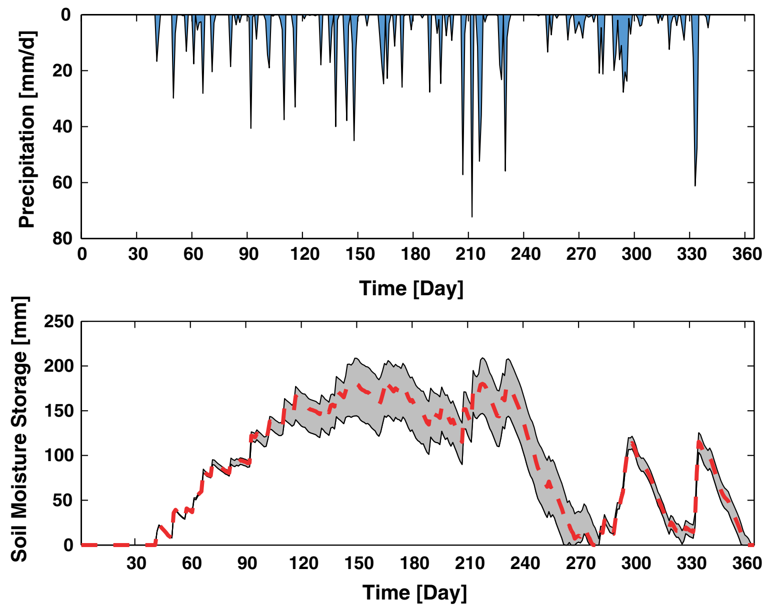


Fig. 7. Soil moisture storage variation (storage in the nonlinear tank of the HyMOD) in the Leaf River watershed (water year 1952–1953) with 95 percentile confidence interval. Dash line is the mean value of the ensemble.

(2) In addition to parameter uncertainty as the only source of uncertainty affecting the performance of the estimation considered in the above-mentioned procedures, the dual EnKF undertakes other sources of errors.
(3) Dual EnKF as a recursive procedure does not require keeping all of the data in storage; thus, by availability of observation at any time, the variables in the system can be adjusted for better conformity with the observation.
(4) Examining the stable uncertainty bounds in Fig. 6 determined by dual EnKF reveals that the parameters do not converge to single points and, therefore, degeneracy of parameter samples does not happen. This is the drawback that BaRE algorithm [25,35] suffers from, that is the parameter uncertainty bounds vanish in a short amount of time after the recursive estimation starts.

In keeping up with previous studies [25,42], the performance of the dual EnKF in streamflow forecasting is demonstrated in Fig. 8. The hydrograph simulation is the result of sequential dual estimation by assimilating streamflow everyday for the water year of 1952–1953 in the Leaf River basin using the HyMOD. The forecasting results with 95% confidence intervals are derived from model output ensemble at each time step. As seen the ensemble mean of daily streamflow forecasting is in very good agreement with the observations, implying that the dual EnKF is a reliable and effective approach for streamflow forecasting. The uncertainty bound also covers the observation in a consistent manner, despite the small negative bias in the rising limb and positive bias in the recession limb of hydrograph which can be seen in lower subplot in Fig. 8. This persistent bias can be explained as the role of model structural error that has not been considered in this study which could be included in future studies.

## 5. Summary and conclusion

Hydrologic models are still far from perfect, and hydrologists need to put the models in better compliance with observations prior to use in forecasting. Batch calibration procedures as the most commonly used techniques in hydrology, and even the recursive calibration schemes concern primarily the estimation of parameters and the identification of uncertainties associated with them. However, more general algorithms that account for the simultaneous interactions of model states and parameters are encouraging while different sources of errors are considered. In this study, an integrated and algorithmic framework for dual state–parameter estimation using EnKF was presented, which leads to the ensemble streamflow forecasting. Perturbation of input and output to generate and modify the ensemble of model variables and to determine the ensemble size are key features of the EnKF, and identification of the magnitude of perturbation in a systematic framework is desired and elaborated in this study.

In the hydrologic model (HyMOD) used for this study, the analysis certainly indicated the feasibility of sequential ensemble filtering that incorporates parameters of the model in addition to state variables. In essence, the dual EnKF use the ensemble of model

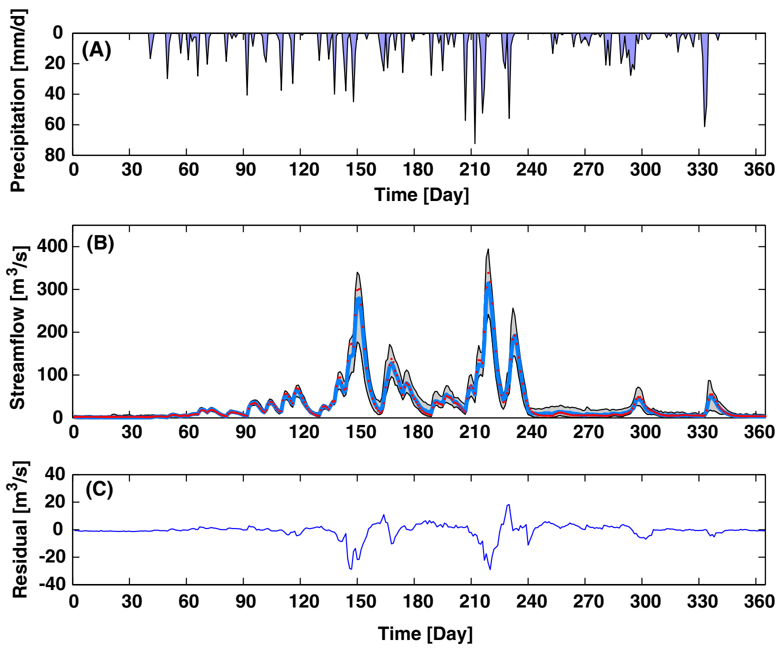


Fig. 8. Ensemble streamflow forecasting by dual state–parameter estimation of HyMOD for the Leaf River watershed (water year 1952–1953), (A) forcing data (precipitation), (B) daily assimilation of observed streamflow (solid dots), streamflow forecast (shaded area with 95% uncertainty bound), and mean value of the ensemble streamflow forecast (solid line), and (C) model residuals (observed—ensemble mean forecast).

trajectories in an interactive parameter–state space and provides the confidence interval of parameter–state estimation. Because the traditional random walk of parameters may result in over-dispersion/information loss and consequently the collapsing the parameter variance, the kernel smoothing of parameters can be employed.

Using the dual technique, the time evolution of parameter uncertainties and one of the state variables were demonstrated. The one-day ahead streamflow forecasting in the Leaf River watershed using the estimated states and parameters was performed, and result seemed to be very consistent with observation. From a filtering point of view, this study offers the following features which do not exist in nonensemble methods:

(1) It allows incorporating a wide range of uncertainties to the model.
(2) It provides a quantitative basis for probabilistic representation of estimates.
(3) The measure of confidence interval becomes possible using such ensemble technique.
(4) It provides a flexible and reliable strategy to deal with nonlinear dynamic models; the problem that could not be overcome entirely by even extended Kalman filter.
(5) It employs the kernel smoothing procedure that protects the parameter sampling from over-dispersion or loss of information while doing the forecasting. The concept of random walk of parameters was always suffering from this drawback.

Finally, it is safe to say that dual EnKF provides a more flexible approach compared to other estimation procedures explained in Section 4. It is a suitable technique for nonlinear models and together with kernel smoothing of parameter samples is a robust and effective algorithm that can tackle input, output and parameter uncertainties properly.

## Acknowledgments

Partial financial support for this research was provided by SAHRA (center for Sustainability of semi-Arid Hydrology and Riparian Areas) under the NSF-STC, Agreement EAR-9876800, and by the Hydrologic Laboratory of the National Weather Service under Agreement NA87WHO582.

## References

[1] Anderson JL. An ensemble adjustment filter for data assimilation. Monthly Weather Rev 2001;129:2884–903.
[2] Boyle DP. Multicriteria calibration of hydrological models. PhD Dissertation, Dep of Hydrol and Water Resour, Univ of Arizona, Tucson; 2000.
[3] Bras RL, Restrepo-Posada P. Real time automatic parameter calibration in conceptual runoff forecasting models. In: Proceedings of the Third International Symposium on Stochastic Hydraulics, 1980. p. 61–70.
[4] Bras RL, Rodriguez-Iturbe I. Random functions and hydrology. Reading, MA, USA: Addison Wesley; 1985. 559 pp.

[5] Burgers GP, Van Leeuwen J, Evensen G. Analysis scheme in the ensemble Kalman filter. Monthly Weather Rev 1998;126:1719–24.

[6] Cohn SE. An introduction to estimation theory. J Meteo Soc Jpn 1997;75(1B):147–1787.

[7] Duan Q, Sorooshian S, Gupta VK. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour Res 1992;28(4):1015–31.

[8] Duan Q, Gupta VK, Sorooshian S. A shuffled complex evolution approach for effective and efficient global minimization. J Optim Theory Appl 1993;76(3):501–21.

[9] Evensen G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res 1994;99:10143–62.

[10] Gelb A, Kasper JF, Nash RA, Price CF, Sutherland AA. Applied optimal estimation. The MIT Press; 1974. 374 pp.

[11] Gupta VK, Sorooshian S, Yapo PO. Towards improved calibration of hydrological models: multiple and noncomensurable measures of information. Water Resour Res 1998;34:751–63.

[12] Hamill TM. Notes and correspondence on "Interpretation of rank histograms for verifying ensemble forecasts". Monthly Weather Rev 2001;129:550–60.

[13] Jazwinski AH. Stochastic processes and filtering theory. San Diego: Academic; 1970.

[14] Kalman R. New approach to linear filtering and prediction problems. Trans AMSE J Basin Eng 1960;82D:35–45.

[15] Kitanidis PK, Bras RL. Adaptive filtering through detection of isolated transient errors in rainfall-runoff models. Water Resour Res 1980;16(4):740–8.

[16] Kitanidis PK, Bras RL. Real-time forecasting with conceptual hydrologic models, 1. Analysis of uncertainty. Water Resour Res 1980;16(6):1025–33.

[17] Kitanidis PK, Bras RL. Real-time forecasting with conceptual hydrologic models, 2. Applications and results. Water Resour Res 1980;16(6):1034–44.

[18] Kuczera G. On validity of first-order prediction limits for conceptual hydrological models. J Hydrol 1998;103:229–47.

[19] Kuczera G, Mroczkowski M. Assessment of hydrological parameter uncertainty and the worth of multiresponse data. Water Resour Res 1998;34(6):1481–9.

[20] Kuczera G, Parent E. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm. J Hydrol 1998;211:69–85.

[21] Liu F. Bayesian time series: analysis methods using simulation based computations. PhD Dissertation, Institutes of Statistics and Decision Sciences, Duke University, 2000.

[22] McLaughlin D. Recent developments in hydrologic data assimilation. Rev Geophys 1995(Suppl):977–84.

[23] McLaughlin D. An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering. Adv Water Res 2002;25:1275–86.

[24] Miller RN, Ghil M, Guathiez F. Advanced data assimilation in strongly nonlinear dynamical systems. J Atmos Sci 1994;51(8):1037–56.

[25] Misirli F, Gupta HV, Sorooshian S, Thiemann M. Bayesian recursive estimation of parameter and output uncertainty for watershed models: Calibration of Watershed Models. In: Duan Q, et al., editor. Water Sci Appl Ser, vol 6, 2003. p. 113–24.

[26] Moore RJ. The probability-distributed principle and runoff production at point and basin scales. Hydrol Sci J 1985;30(2):273–97.

[27] Murphy JM. The impact of ensemble forecast on predictability. Quart J Roy Meteor Soc 1988;114:89–125.

[28] Nelson AT. Nonlinear estimation and modeling of noisy time series by dual Kalman filtering methods. PhD Dissertation, Oregon Graduate Institute, 1999.

[29] Reichle RH, McLaughlin D, Entekhabi D. Hydrologic data assimilation with the ensemble Kalman filter. Monthly Weather Rev 2002;130:103–14.

[30] Reichle RH, Walker JP, Koster RD, Houser PR. Extended versus ensemble Kalman filtering for land data assimilation. J Hydromet 2002;3:728–40.

[31] Sorooshian S, Dracup JA. Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: correlated and heteroscedastic error cases. Water Resour Res 1980;16(2):430–42.

[32] Sorooshian S, Duan Q, Gupta VK. Calibration of rainfall-runoff models: application of global optimization to the soil moisture accounting model. Water Resour Res 1993;29(4):1185–94.

[33] Spear RC, Hornberger GM. Eutrophication in Peel Inlet, 2, Identification of critical uncertainties via generalized sensitivity analysis. Water Resour Res 1980;14:43–9.

[34] Spear RC, Grieb TM, Shang N. Parameter uncertainty and interaction in complex environmental models. Water Resour Res 1994;30(11):3159–69.

[35] Thiemann M, Trosset M, Gupta H, Sorooshian S. Bayesian recursive parameter estimation for hydrologic models. Water Resour Res 2001;37(10):2521–35.

[36] Todini E, Szollosi-Nagy A, Wood EF. Adaptive state–parameter estimation algorithm for real time hydrologic forecasting; a case study. In: IISA/WMO Workshop on the Recent Developments in Real Time Forecasting/Control of Water Resources Systems, Laxemburg (Austria), 1976.

[37] Todini, E. Mutually interactive state–parameter (MISP) estimation, Applications of Kalman filters to hydrology, hydraulics and water resources. In: Chiu C-L, editor. Proceedings of AGU Chapman Conference, 1978.

[38] Toth Z, Talagrand O, Candille G, Zhu Y. Probability and ensemble forecasts. In: Jolliffe IT, Stephenson DB, editors. Environmental forecast verification: a practitioner's guide in atmospheric science, John Wiley, in press.

[39] Troch PA, Paniconi C, McLaughlin D. Catchment-scale hydrological modeling and data assimilation. Adv Water Res 2003;26:131–5.

[40] Young PC. Advances in real-time flood forecasting. Philos Trans Roy Soc Lond 2002;360:1433–50.

[41] Van Leeuwen PJ. Comment on data assimilation using an ensemble Kalman filter technique. Monthly Weather Rev 1999;127:1374–7.

[42] Vrugt JA, Gupta HV, Bouten W, Sorooshian S. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour Res 2003;39(8):1–14.

[43] Vrugt JA, Gupta HV, Sorooshian S, Wagener T, Bouten W. Confronting parameter uncertainty in hydrologic modeling: application of the SCEM-UA algorithm to the Sacramento Soil Moisture Accounting Model. J Hydrol, in review.

[44] Wagener T, Boyle DP, Lees MJ, Wheater HS, Gupta HV, Sorooshian S. A framework for development and application of hydrological models. Hydrol Earth Syst Sci 2001;5(1):13–26.

[45] Wan EA, Nelson TA. Dual Kalman filtering methods for nonlinear prediction, estimation and smoothing. Adv Neural Informat Process Syst 1997;9.

[46] West M. Mixture models, Monte Carlo, Bayesian updating and dynamic models. Comput Sci Statist 1993;24:325–33.

[47] Whitaker JS, Hamill TM. Ensemble data assimilation without perturbed observations. Monthly Weather Rev 2002;130:1913–24.

[48] Wilks DS. Forecast verification, Statistical methods in the atmospheric sciences. Academic; 1995. p. 233–83.