Sequential data assimilation for streamflow forecasting using a distributed hydrologic model: particle filtering and ensemble Kalman filtering

SEONG JIN NOH[1, 2], YASUTO TACHIKAWA[3], MICHIHARU SHIIBA[3] & SUNMIN KIM[3]

1 Dept. of Urban and Env. Eng., Kyoto University, Kyoto 615-8540, Japan
seongjin.noh@gmail.com
2 Water Resources & Environment Research Department, Korea Institute of Construction Technology, 2311 Daewha-Dong, Ilsan-Gu, Goyang-Si, Gyeonggi-Do 411-712, Korea
3 Dept. of Civil and Earth Resources Engineering, Kyoto University, Kyoto 615-8540, Japan

Abstract Accurate streamflow predictions are crucial for mitigating flood damage and addressing operational flood scenarios. In recent years, sequential data assimilation methods have drawn attention due to their potential to handle explicitly the various sources of uncertainty in hydrologic models. In this study, we implement two ensemble-based sequential data assimilation methods for streamflow forecasting via the particle filters and the ensemble Kalman filter (EnKF). Among variations of filters, the ensemble square root filter (EnSRF) and the lagged regularized particle filter (LRPF) are implemented for a distributed hydrologic model. Two methods are applied for short-term flood forecasting in a small-sized catchment located in Japan (<1000 km^2). Soil moisture contents are perturbed by process noises and model ensembles are updated by streamflow observation at the outlet. In the case of the LRPF, state updating is performed through a lag-time window to take into account the different response times of hydrologic processes. For different flood events and various forecast lead times, LRPF forecasts outperform EnSRF forecasts and deterministic cases. The EnSRF shows limited performance in both forecasting accuracy and probabilistic intervals, which require introduction of a lag-time window in the filtering processes.

Key words sequential data assimilation; flood forecasting; particle filter; ensemble Kalman filter; distributed hydrologic model

INTRODUCTION

Data assimilation is a way to integrate information from a variety of sources to improve prediction accuracy while taking into consideration the uncertainty in both a measurement system and a prediction model. State-space filtering methods based on variations of the Kalman filter (KF) approach have been proposed and implemented because of their potential ability to explicitly handle uncertainties in hydrologic predictions (Vrugt *et al*., 2006). However, the KF approaches for a nonlinear system such as the extended Kalman filter (EKF) have limitations in practical application due to their instability with strong nonlinearity and the high computational cost of model derivative equations, especially for high-dimensional state-vector problems such as spatially distributed models. To cope with the drawbacks of EKF, Evensen (1994) introduced the ensemble Kalman filter (EnKF), which uses an ensemble of forecasts to estimate background-error covariances. Thus, no adjoint or linearized model is needed for error estimation, and the method provides great versatility, as any number of variables can be included in the update procedure. However, in the analysis step of the conventional EnKF, perturbation of measurements is used to update ensemble members and is an additional source of uncertainty. Thus, the ensemble square root filter (EnSRF) has been developed to avoid sampling issues associated with the use of "perturbed observations" in stochastic analysis ensemble update methods (Whitaker & Hamill, 2002; Clark *et al*., 2008).

Another alternative is particle filters, which are a Bayesian learning process in which the propagation of all uncertainties is carried out by a suitable selection of randomly-generated particles without any assumptions made about the nature of the distributions (Gordon *et al*., 1993; Arulampalam *et al*., 2002). Unlike the Kalman filter-based methods, which are basically limited to the linear correction step and the assumption of Gaussian distribution errors, the particle filters have the advantage of being applicable to non-Gaussian state-space models. In recent years, these methods have received considerable attention in hydrology and earth sciences (e.g. Moradkhani *et al*., 2005; Weerts & El Serafy, 2006; Noh *et al*., 2011a,b).

In this study, we implement and compare two sequential data assimilation methods for flood forecasting using a distributed hydrologic model. Recently-proposed elaborate schemes are selected: the ensemble square root filter modified from the original EnKF and the lagged regularized particle filter from the particle filters. A distributed hydrologic model, the water and energy transfer processes (WEP) model (Jia *et al*., 2009), is applied to the Katsura River catchment, Japan. The paper is organized as follows: The next section outlines the Bayesian filtering theory, ensemble Kalman filtering, and particle filtering followed by a section that presents the case study results, which demonstrate the applicability of the applied filtering methods. The EnSRF and the LRPF are evaluated for hindcasting of streamflow in the Katsura River catchment using the WEP model. Finally, the results section summarizes the results and conclusions.

METHODS OF SEQUENTIAL DATA ASSIMILATION

Bayesian filtering theory

To define the problem of Bayesian filtering, consider a general dynamic state-space model, which is described as:

$$x_k = f(x_{k-1}, u_k) + \omega_k \qquad \omega_k \sim N(0, W_k) \tag{1}$$

$$y_k = h(x_k) + v_k \qquad v_k \sim N(0, V_k) \tag{2}$$

where x_k is the n_x dimensional vector denoting the system state at time k. The operator $f : \Re^{n_x} \to \Re^{n_x}$ expresses the system transition in response to the forcing data u_k (e.g. rainfall, weather data). $h : \Re^{n_x} \to \Re^{n_y}$ expresses the measurement function. ω_k and v_k represent the model error and the measurement error, respectively, and W_k and V_k represent the covariance of the error. In the Bayesian recursive estimation, if the system and measurement models are nonlinear and non-Gaussian, it is not possible to analytically derive the posterior probability density function (pdf) of the current state x_k on the measurement.

Ensemble Kalman filtering

The ensemble Kalman filter (EnKF) is a suboptimal estimator, where the error statistics are predicted using a Monte Carlo method. The EnKF consists of update and prediction steps. The ensemble mean of the states is defined as:

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_k^i \tag{3}$$

where x_k^i denotes the state of ith ensemble and n is the number of ensemble members. If the measurements are a nonlinear combination of state variables, the Kalman gain is calculated as:

$$K = P_{xy_k}(P_{yy_k} + V_k)^{-1} \tag{4}$$

$$P_{xy_k} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_k^i - \bar{x}_k\right)\left(h(x_k^i) - \overline{h(x_k)}\right)^T \tag{5}$$

$$P_{yy_k} = \frac{1}{n-1}\sum_{i=1}^{n}\left(h(x_k^i) - \overline{h(x_k)}\right)\left(h(x_k^i) - \overline{h(x_k)}\right)^T \tag{6}$$

In the conventional Kalman filter, perturbed observations, which can have a detrimental effect in the analysis step, are used in the update equation (Clark *et al*., 2008). Whitaker & Hamill (2002) introduced the ensemble square root filter (EnSRF), which provides the correct analysis error covariance without perturbing the observations. In this method, the ensemble is broken into mean and anomaly portions, and updating is performed separately for the ensemble mean and anomalies:

$$\bar{x}_k^{up} = \bar{x}_k + K\left(\bar{y}_k - \overline{h(x_k)}\right) \tag{7}$$

$$x_k'^{up,i} = x_k'^{i} + K'\left(y_k'^{i} - h(x_k^i)'\right) \tag{8}$$

where the prime denotes the deviations of each ensemble from the ensemble mean. The ensemble mean is updated with the traditional gain equation given above, while anomalies are updated with a reduced gain given by:

$$K' = P_{xy_k}\left[\left(\sqrt{P_{yy_k} + V_k}\right)^{-1}\right]^T\left[\left(\sqrt{P_{yy_k} + V_k} + \sqrt{V_k}\right)\right]^{-1} \tag{9}$$

In equation (8), $y_k'^{i} = 0$, which means no perturbation of observation in anomalies. Whitaker & Hamill (2002) showed that the sampling error associated with perturbed observations makes the EnSRF more accurate than the conventional EnKF.

Particle filtering

The particle filters are simulation-based methods that provide a flexible approach to computing posterior distributions without any assumptions about the nature of the distributions. The key idea of the particle filters is based on point mass ("particle") representations of probability densities with associated weights:

$$p(x_k \mid y_{1:k}) \approx \sum_{i=1}^{n} w_k^i \delta(x_k - x_k^i) \tag{10}$$

where w_k^i denotes the ith weight, and $\delta(\cdot)$ denotes the Dirac delta function. After several computational steps, weight updating becomes:

$$w_k^i \propto w_{k-1}^i p(y_k \mid x_k^i) \tag{11}$$

where $p(y_k \mid x_k^i)$ is the so-called likelihood of each ensemble. Detailed descriptions and introductions of the particle filters can be found in Arulampalam *et al*. (2002) and Moradkhani *et al*. (2005). In conventional particle filters, the resampling step is performed to avoid degeneracy phenomenon, in which, after a few iterations, all but one particle will have negligible weight (van Leeuwen, 2009). However, the particles resampled from high weights are statistically selected many times. This leads to another problem, known as sample impoverishment, which means a loss of diversity among the particles because the resultant sample will contain many repeated points (Ristic *et al*., 2004). An alternative solution is to introduce the regularization step when sample impoverishment becomes severe (Musso *et al*., 2001). The main idea of the RPF consists of changing the discrete approximation of posterior distribution to a continuous approximation, so the resampling step is changed into simulating an absolutely continuous distribution, thus producing a new particle system with *n* different particle locations. The RPF can be used with the Markov chain Monte Carlo (MCMC) move step (Gilks & Berzuini, 2001) based on the Metropolis-Hastings algorithm to approximate the posterior distribution properly.

In a distributed hydrologic model, there are many types of state variables, each of which interacts with others based on different time scales, which need to be considered in the data assimilation. In this study, we implement the lagged regularized particle filter (LRPF) to deal with the delayed response, which originates from different time scales of hydrologic processes in a distributed model. A detailed description of the LRPF can be found in Noh *et al*. (2011a).

IMPLEMENTATION

Study area and model setup

Two sequential data assimilation methods are applied to the Katsura River catchment for short-term river flow forecasting. This catchment is located in Kyoto, Japan, and covers an area of

1100 km^2 (887 km^2 at Katsura station) (see Fig. 1). The elevation in the catchment ranges from 4 to 1158 m, with an average of about 325 m. The controlled outflow record from the dam reservoir is given as inflow to the hydrologic model, and the model simulates rainfall–runoff processes for the downstream of the dam.

The hydrologic model used is the water and energy transfer processes (WEP) model, which was developed for simulating spatially variable water and energy processes in catchments with complex land covers (Jia & Tamai, 1998; Jia *et al*., 2009). State variables of WEP include soil moisture content, surface runoff, groundwater tables, discharge and water stage in rivers, heat flux components, etc. (Fig. 2). Runoff routing on slopes and in rivers is carried out by applying a 1D kinematical wave approach from upstream to downstream. This model has been applied in several watersheds in Japan, Korea, and China with different climate and geographic conditions (Jia *et al*., 2001, 2009; Kim *et al*., 2005).

The model setup uses a 250 m grid resolution and an hourly time step. We use hourly observe d rainfall from 13 observation stations organized by the Ministry of Land, Infrastructure, Transport

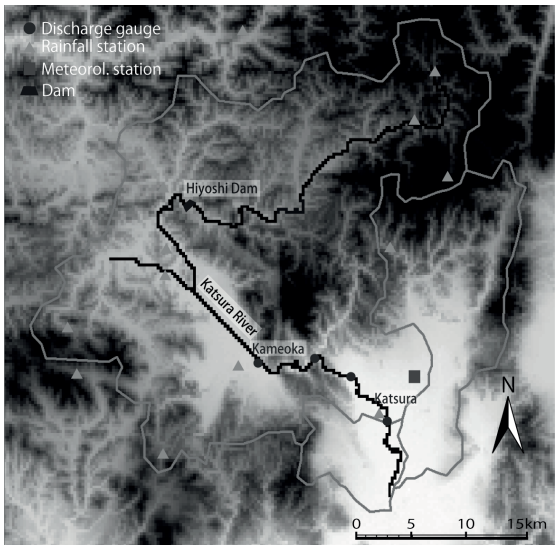

Fig. 1 The Katsura River catchment.

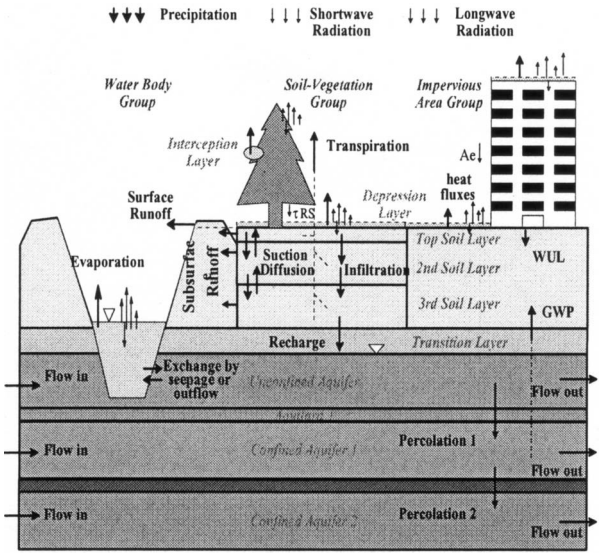

Fig. 2 A schematic view of WEP model (Jia *et al*., 2001).

and Tourism in Japan and hourly observed meteorological data, including air temperature, relative humidity, wind speed, and duration of sunlight, from the Kyoto station, which is organized by Japan Meteorological Agency. The nearest-neighbour interpolation method is used for representation of spatial distribution of rainfall. Ensemble simulation is conducted on a multiprocessing computer (96 cores in the supercomputing system of Kyoto University) via parallel-computing techniques of an open message passing interface (open MPI) (http://www.open-mpi.org/).

Processes and measurement error models

Because there are numerous state variables in a distributed hydrologic model, it is not practical to treat the uncertainty of all state variables with a limited number of ensembles. In this study, we select soil moisture content in each grid as hidden state variables and streamflows at the Katsura station as an observable variable for data assimilation. Global multipliers are introduced to perturb state variables for the EnKF and the LRPF. The total soil moisture depth at the previous time step S_{k-1} is aggregated for three soil layers within the catchment as:

$$S_{k-1} = \sum_{l=1}^{3} \sum_{j=1}^{m} \theta_j^l d_j^l \tag{12}$$

where θ_j^l and d_j^l are the volumetric soil moisture content (m^3/m^3) and the soil depth (m) in each layer, and l and m represent the number of soil layers and the total number of grids within the catchment, respectively. Process noise of the soil moisture content w_{soil_k} is then added to the aggregated state variable S_{k-1} as:

$$\hat{S}_k = S_{k-1} + w_{soil_k} \tag{13}$$

w_{soil_k} is assumed as Gaussian distribution $N(0, \sigma_{soil_k}^2)$ having a standard deviation of:

$$\sigma_{soil_k} = \alpha_{soil} S_{k-1} + \beta_{soil} \tag{14}$$

In the above equation, α_{soil} and β_{soil} are adaptable parameters. The multiplicative factor γ_s and the perturbed states of soil moisture $\hat{\theta}_j^l$ are calculated as follows:

$$\gamma_s = \frac{\hat{S}_k}{S_{k-1}} \tag{15}$$

$$\hat{\theta}_j^l = \gamma_s \theta_j^l \tag{16}$$

The measurement error of the discharge is assumed to be a Gaussian distribution $N(0, \sigma_{obs_k}^2)$ as in previous studies (Georgakakos, 1986; Weerts & El Serafy, 2006; Salamon & Feyen, 2010). The standard deviation of the measurement error is chosen as:

$$\sigma_{obs_k} = \alpha_{obs} y_k + \beta_{obs} \tag{17}$$

where α_{obs} is set to 0.1 and the constant coefficient β_{obs} is applied as 5 (m^3/s) to consider uncertainty in periods of low flow, such as artificial water use and dam reservoir control.

RESULTS

We implement two sequential data assimilation methods, the EnSRF and the LRPF, for the hindcasting of streamflow using the WEP model. Simulation periods and observation are shown in Table 1. Hourly observed discharges at the Katsura station are used for data assimilation, and forecasted discharges are predicted up to a 24-h lead time. The lag time of 8 h is applied in the LRPF. A5-day warm-up period is added before the data assimilation starts.

Table 1 Simulation periods and observations.

Simulation period	Max. observed flow at Katsura (m^3s^{-1})	Total areal rainfall (mm)
1 Jun.–31 Jul. 2007	336.9	491
1 Jun.–31 Aug. 2003	361.6	729

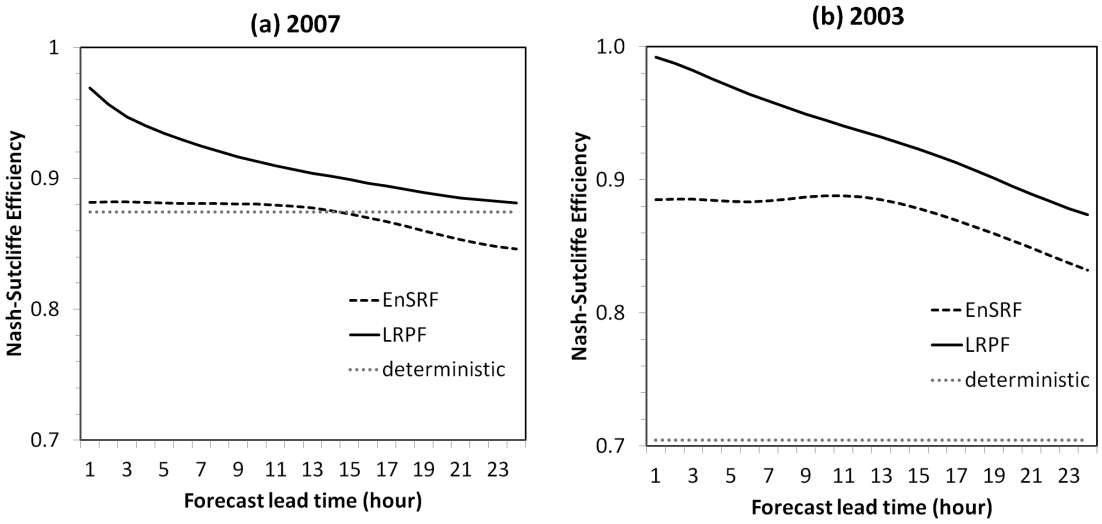

Fig. 3 Nash-Sutcliffe model efficiency for varying forecast lead time. The black lines represent the LRPF. The dashed lines represent the EnSRF. The dotted lines represent a deterministic modelling case.

Figure 3 shows Nash-Sutcliffe efficiency of each particle filter for varying lead times in the years 2004 and 2003 calculated as:

$$\mathrm{NSE} = 1 - \frac{\sum_{k=1}^{T} \left(y_k - y_{sim_k}\right)^2}{\sum_{k=1}^{T} \left(y_k - \bar{y}\right)^2} \tag{18}$$

where y_k is observation, $\bar{y}$ is the mean of observation, y_{sim_k} is the forecasted streamflow at the measurement site, and T is the total number of time steps.

As shown in Fig. 3, the LRPF forecasts are superior to the EnSRF forecasts and the deterministic modelling cases in both simulation periods. NSE scores of the LRPF are higher than 0.87 even if the forecast lead time reaches 24 h, which shows that state updating via LRPF reproduces the measurement state variable properly and the effects of updating are still valid even 24 h later. While the LRPF shows the highest scores in short lead times, the accuracy of the EnSRF is not improved in short lead times and shows a nearly flat shape within a 12-h lead time. The limited performance of the EnSRF seems to be related to the short interval of updating. The ensemble Kalman filter updates state variables directly using linear updating rules based on covariance information of model states and observations. One-hour frequency of updating may be too frequent to estimate correct soil moisture states using streamflow observations because the response time is usually larger than that.

Figures 4 and 5 illustrate 6-h lead forecasts via the EnSRF and the LRPF compared to deterministic cases for selected events among simulation periods. The mean of forecasted streamflow via the LRPF shows good conformity between observation and simulation, while deterministic modelling shows significant underestimation, especially in high flood. The mean of the EnSRF shows unstable fluctuations and large variations compared to streamflow observations. Confidence intervals also show different characteristics in both filters. The LRPF has stable and narrow confidence interval, while those of the EnSRF increase very sharply during flood events. As shown in Fig. 5(a), confidence intervals of the EnSRF increase very sharply between 1070 and

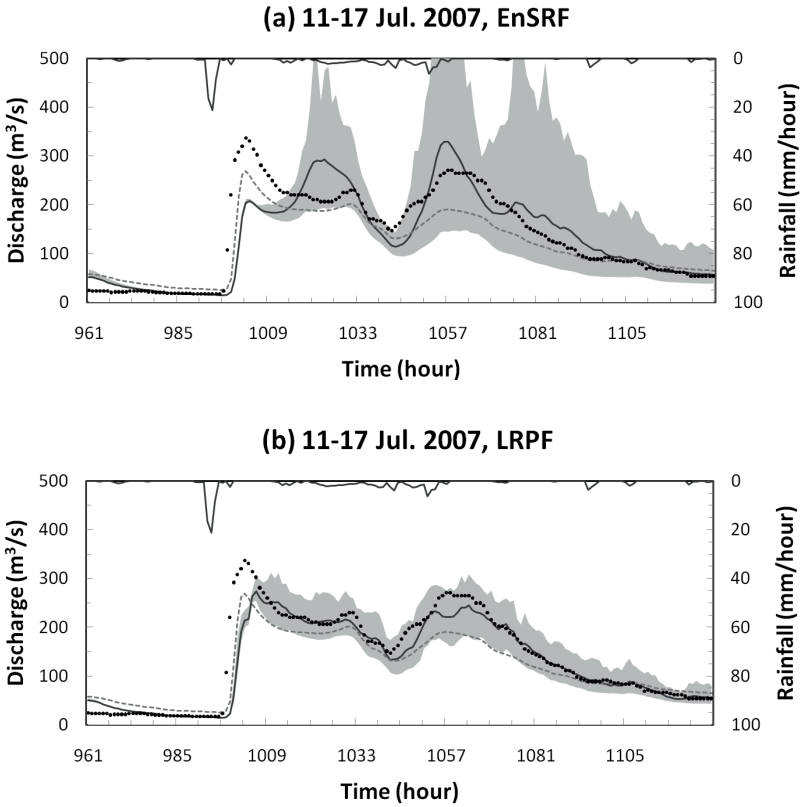

Fig. 4 Observed *versus* 6-h lead forecasts at the Katsura station via the EnSRF and the LRPF (11 to 17 July 2007). The solid lines and area represent the mean value and 90% confidence intervals, respectively. The dashed lines represent deterministic modelling cases. The dots represent observed discharge.

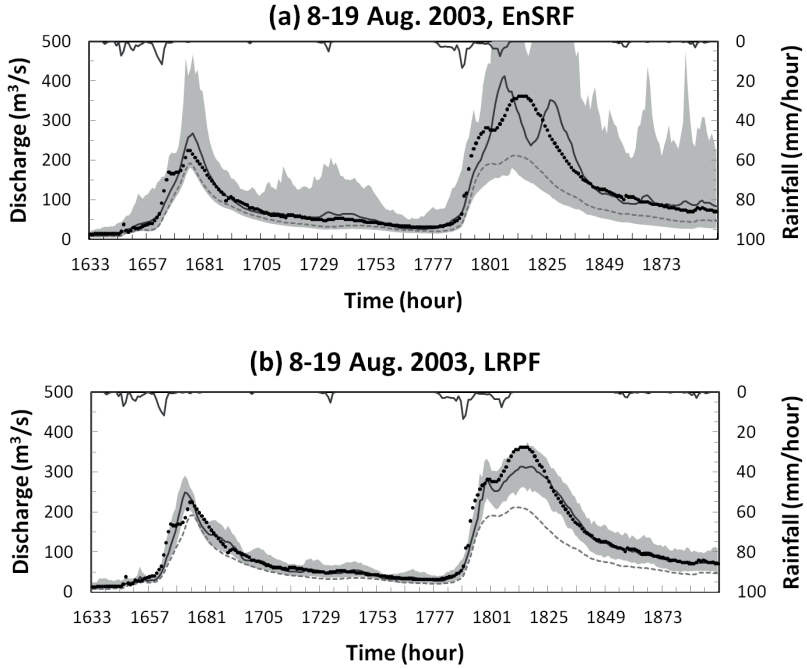

Fig. 5 Observed *versus* 6-h lead forecasts at the Katsura station via the EnSRF and the LRPF (8–19 August 2003). The solid lines and area represent the mean value and 90% confidence intervals, respectively. The dashed lines represent deterministic modelling cases. The dots represent observed discharge.

1090 hours, even when rainfall events stop. In the case of the LRPF, a lag-time window, which calculates an analysis from several previous time steps, is adopted. This lagged analysis step of the LRPF seems to contribute to enhancement of forecasting accuracy. Another advantage of the particle filters is that all information in an ensemble is duplicated in the resampling step or the regularization step, which reduces numerical instability and increases forecasting accuracy.

CONCLUSIONS

The ensemble square root filter and the lagged regularized particle filter were implemented for flood forecasting using a distributed hydrologic model, WEP. The distributed soil moisture state was perturbed and assimilated with observed streamflow discharges every hour in both filters. Two data assimilation methods were also effectively parallelized and implemented in the multicore computing environment via the MPI library. Forecasts based on updated results were assessed for various lead times up to 24 h using Nash-Sutcliffe efficiency. The LRPF showed improved forecasts compared to the EnSRF and deterministic modelling in most data periods and forecast lead times. Updating effects via the LRPF lasted more than 24 h, while the EnSRF showed limited improvements, especially in short forecast lead times. In terms of probabilistic adequacy, confidence intervals of the LRPF showed stable bands, while those of the EnSRF increased during flood events and showed diffuse patterns. Alternatively, introduction of a lag-time window for the EnKF may improve performance. Sequential data assimilation has significant potential for high nonlinearity problems, especially for process-based distributed models, and the LRPF is expected to be used as one of the frameworks for sequential data assimilation of process-based distributed hydrological models.

REFERENCES

Arulampalam, M. S., Maskell, S., Gordon, N. & Clapp, T. (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Processes* 50, 174–188.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J. & Uddstrom, M. J. (2008) Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Adv. Water Resour.* 31, 1309–1324.

Evensen, G. (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* 99, 10143–10162.

Georgakakos, K. P. (1986) A generalized stochastic hydrometeorological model for flood and flash-flood forecasting. *Water Resour. Res.* 22, 2096–2106.

Gilks, W. R. & Berzuini, C. (2001) Following a moving target-Monte Carlo inference for dynamic Baysian models. *J. Roy. Statist. Soc.* B 63, 127–146.

Gordon, N. J., Salmond, D. J. & Smith, A. F. M. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Proc. Inst. Electr. Engng* 140, 107–113.

Jia, Y., Ding, X., Qin, C. & Wang, H. (2009) Distributed modeling of landsurface water and energy budgets in the inland Heihe river basin of China. *Hyrol. Earth Syst. Sci.* 13, 1849–1866.

Jia, Y., Ni, G., Kawahara, Y. & Suetsugi, T. (2001) Development of WEP model and its application to an urban watershed. *Hydrol. Processes* 15, 2175–2194.

Jia, Y. & Tamai, N. (1998) Integrated analysis of water and heat balance in Tokyo metropolis with a distributed model. *J. Japan Soc. Hydrol. Water Resour.* 11, 150–163.

Kim, H. J., Yoon, S. K., Noh, S. J. & Jang, C. H. (2005) The Cheonggye-cheon restoration project and hydrological cycle analysis. *Water Engng Res.* 6, 179–187.

Moradkhani, H., Hsu, K.-L., Gupta, H. & Sorooshian, S. (2005) Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water Resour. Res.* 41, W05012, doi: 10.1029/2004WR003604.

Musso, C., Oudjane, N. & LeGland, F. (2001) Improving regularized particle filters. In: *Sequential Monte Carlo in Practice* (ed. by A. Doucet, N. de Freitas & N. Gordon), 247–271, Springer-Verlag.

Noh, S. J., Tachikawa, Y., Shiiba, M. & Kim, S. (2011a) Applying sequential Monte Carlo methods into a distributed hydrologic model: lagged particle filtering approach with regularization. *Hydrol. Earth Syst. Sci.* 15, 3237–3251.

Noh, S. J., Tachikawa, Y., Shiiba, M. & Kim, S. (2011b) Dual state-parameter updating scheme on a conceptual hydrologic model using sequential Monte Carlo filters. *Ann. J. Hydraul. Engng JSCE* 55, 1–6.

Ristic, B., Arulampalam, S. & Gordon, N. (2004) *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House.

Salamon, P. & Feyen, L. (2010) Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation. *Water Resour. Res.* 46, W12501, doi: 10.1029/2009WR009022.

van Leeuwen, P. J. (2009) Particle filtering in geophysical systems. *Mon. Weather Rev.* 137, 4089–4114.

Vrugt, J. A., Gupta, H. V., Nuallain, B. O. & Bouten, W. (2006) Real-time data assimilation for operational ensemble streamflow forecasting. *J. Hydrolmeteorol.* 7, 548–565.

Weerts, A. H. & G. Y. H. El Serafy (2006) Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resour. Res.* 42, W09403, doi: 10.1029/2005WR004093.

Whitaker, J. S. & Hamill, T. M. (2002) Ensemble data assimilation without perturbed observation. *Mon. Weather Rev.* 130, 1913–1924.