Real-time flood forecast using the coupling support vector machine and data assimilation method

Xiao-Li Li, Haishen Lü, Robert Horton, Tianqing An and Zhongbo Yu

ABSTRACT

An accurate and real-time flood forecast is a crucial nonstructural step to flood mitigation. A support vector machine (SVM) is based on the principle of structural risk minimization and has a good generalization capability. The ensemble Kalman filter (EnKF) is a proven method with the capability of handling nonlinearity in a computationally efficient manner. In this paper, a type of SVM model is established to simulate the rainfall–runoff (RR) process. Then, a coupling model of SVM and EnKF (SVM + EnKF) is used for RR simulation. The impact of the assimilation time scale on the SVM + EnKF model is also studied. A total of four different combinations of the SVM and EnKF models are studied in the paper. The Xinanjiang RR model is employed to evaluate the SVM and the SVM + EnKF models. The study area is located in the Luo River Basin, Guangdong Province, China, during a nine-year period from 1994 to 2002. Compared to SVM, the SVM + EnKF model substantially improves the accuracy of flood prediction, and the Xinanjiang RR model also performs better than the SVM model. The simulated result for the assimilation time scale of 5 days is better than the results for the other cases.

Key words | ensemble Kalman filter, rainfall–runoff simulation, support vector machine, Xinanjiang rainfall–runoff model

Xiao-Li Li
Haishen Lü (corresponding author)
Tianqing An
Zhongbo Yu
College of Science,
State Key Lab of Hydrology-Water Resources & Hydraulic Engineering,
Hohai University,
Nanjing 210098,
China
E-mail: *haishenlu@gmail.com*

Xiao-Li Li
College of Electronics and Information,
Nanjing University of Technology,
Nanjing 210009,
China

Robert Horton
Department of Agronomy,
Iowa State University,
Ames, IA 50011,
USA

INTRODUCTION

Accurate real-time flood forecasts are crucial for water resources planning and management, and reservoir and river regulation (Chang & Chen 2001; Rajurkara *et al.* 2004). Many approaches in artificial intelligence have been exploited for hydrological forecasting, such as artificial neural networks (Coulibaly *et al.* 2000; Taormina *et al.* 2012), genetic algorithm (Cheng *et al.* 2002), fuzzy theory (Nayak *et al.* 2005) and support vector machine (SVM) (Yu *et al.* 2006). Hydrological applications of the SVM have been investigated. Sivapragasam *et al.* (2001) used the SVM model to perform one-lead-day rainfall–runoff forecasting. Choy & Chan (2003) determined objectively the framework of the radial basis function network using the SVM, and applied such a network to simulate the relationship between rainfall and runoff. Yu *et al.* (2004) combined chaos theory and the SVM to forecast daily runoff. Bray & Han (2004) developed a runoff prediction method based on the SVM by identifying an appropriate model structure and relevant parameters. In order to achieve an optimal training data set, Sivapragasam & Liong (2005) divided the flow range into three zones, including high, medium, and low zones, and used different SVM models to forecast daily flows in different zones. Yu *et al.* (2006) predicted hourly flood stages in real time using the SVM, and performed a sensitivity analysis on lagged input variables of the SVM.

Although the SVM model has been broadly used in flood forecasting, the SVM model predictions usually substantially deviate from observations (Sivapragasam *et al.* 2001; Sivapragasam & Liong 2005). Sivapragasam & Liong (2005) showed that the error is mainly due to a lack of training data in the model preparation. The data assimilation method can help to reduce prediction error and results in the best state estimates (Reichle *et al.* 2002). The ensemble Kalman filter

doi: 10.2166/hydro.2013.075

(EnKF), an important data assimilation method, has been widely applied in meteorological, oceanographic, and hydrological forecasting (Weerts & Serafy 2006; Lü *et al.* 2010a, 2010b). Reichle *et al.* (2002) used the EnKF to forecast soil moisture and obtained satisfactory results compared to the variational assimilation method. Crow & Wood (2003) applied the EnKF to assimilate soil brightness temperature into a land-surface model and found that this assimilation method was more competitive than other approaches. Moradkhani & Hsu (2005) used a conceptual hydrologic model coupling the EnKF data assimilation method to forecast the RR process and found that the updating procedure using online measured runoff improved runoff forecasts. Kashif Gill *et al.* (2007) simulated soil moisture changes by combining the SVM and the EnKF. The results showed that the simulated results were very consistent with the observed results. To our knowledge, SVM and EnKF have not been coupled to forecast the rainfall–runoff process.

In this paper, the goal is to simulate the RR process by coupling the SVM and EnKF methods. The three main research problems are: (1) to evaluate the SVM model in the study region; (2) to couple the SVM and the EnKF assimilation method; and (3) to compare the simulated results among the SVM method, the SVM + EnKF method, and the Xinanjiang RR model. The study region is located in the Luo River Basin, Guandong Province, China.

METHODOLOGY

Study area and data set description

The main study area (see Figure 1) is the Luo River Basin located in Guangdong, China. The area is about 150 km^2. Four rain gauges and evaporation stations were selected in the main streams of the basins: Nangao, Luoyan, Dingyang, and Nanwan. Precipitation and evaporation data from the four rain gauges, and four evaporation stations in the area were used for daily runoff simulations. Daily rainfall and evaporation data are available from these four stations, and daily streamflow data are available from the Nangao hydrologic control stations for the nine-year period of 1994–2002. Type E-601 evaporation pans were used to observe the amount of evaporation during this period. Annual

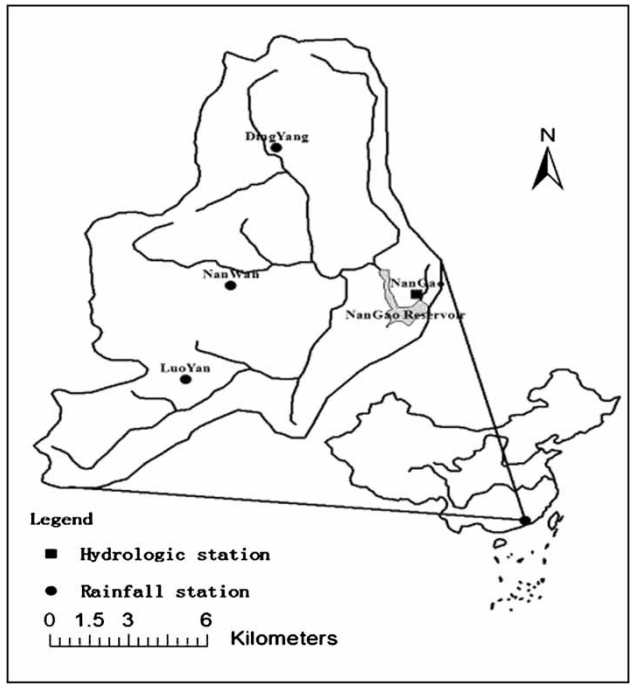

Figure 1 | Location map for the study area.

average rainfall is about 2,330 mm (during the flood season from April to September, rainfall is about 1,890 mm, 81% of the total annual precipitation). The variance of annual precipitation is about 1,090 mm^2. The average streamflow into the Nangao Reservoir is 8.76 m^3/s. Mean annual volume is 2.76×10^8 m^3. The variance of annual average streamflow is 3.40 (m^3/s)^2.

The main goal of this paper is to develop the SVM model and the SVM + EnKF model to forecast the streamflow at the Nangao Reservoir. It is well known that the appropriate input variables contain important features about the complex autocorrelation structure in the data sets. In order to simplify the calculations, we used the precipitation (P) as an input variable for the models and the discharge (Q) as the output variable. In the model, the value of P is substituted by $[P–EP]$, where P represents the mean precipitation and EP is the mean potential evapotranspiration.

To ensure that all variables receive equal weights during the training process, it is necessary to normalize the raw data (precipitation) to the interval from −1 to 1 or from 0 to 1. Therefore, the SVM and SVM + EnKF models process

the scaled data, and the output data are returned to their original scale. The data are normalized between 0.1 and 0.9. The scaling and reverse scaling formulas are as follows:

$$P_n(t) = 0.1 + 0.8 \times \left(\frac{P(t) - P_{\min}}{P_{\max} - P_{\min}}\right) \quad (1)$$

$$Q_{\mathrm{n,obs}}(t) = 0.1 + 0.8 \times \left(\frac{Q_{\mathrm{obs}}(t) - Q_{\min}}{Q_{\max} - Q_{\min}}\right) \quad (2)$$

$$Q_{\mathrm{sim}}(t) = Q_{\min} + \frac{1.0}{0.8} \times (Q_{\mathrm{n}}(t) - 0.1) \times (Q_{\max} - Q_{\min}) \quad (3)$$

where $P(t)$ is the observed precipitation data; $P_{\mathrm{n}}(t)$ is the scaled precipitation data at time t; $P_{\min}$ and $P_{\max}$ are the minimum and maximum of precipitation data series during the simulation period; $Q_{\mathrm{obs}}(t)$ is observed streamflow; $Q_{\mathrm{n,obs}}(t)$ is normalized observed streamflow; $Q_{\mathrm{n}}(t)$ is the simulated streamflow using the SVM model; $Q_{\mathrm{sim}}(t)$ is the reverse scaled streamflow; $Q_{\min}$ and $Q_{\max}$ are the observed minimum and maximum of the streamflow.

The SVM model and model construction

The original SVM, introduced in 1992 (Boser *et al.* 1992; Cortes & Vapnik 1995; Vapnik 1998), can be characterized as a supervised learning algorithm capable of solving linear and nonlinear classification problems. The main building blocks of SVM are structural risk minimization, originating from statistical learning theory, which was mainly developed by Vapnik & Chervonenkis (1974), nonlinear optimization and duality and kernel induced features spaces, underlining the technique with an exact mathematical framework. Meanwhile, several extensions to the basic SVM have been introduced, e.g., for multiclass classification as well as regression and clustering problems, making the technique broadly applicable in the data mining area. Recently, some applications of the SVM model or SVM regression model have been used in the prediction of rainfall–runoff process, rainfall, and river flow (Sivapragasam *et al.* 2001; Bray & Han 2004; Sivapragasam & Liong 2004, 2005; Yu *et al.* 2004; Lin *et al.* 2006; Wu *et al.* 2008).

For the linear SVM regression model, suppose that the training sample set is

$$S = \{(x_1, y_1), \ldots, (x_l, y_l)\} \in (X, Y)^l, \quad X = R^n, Y = R \quad (4)$$

The SVM regression for the linear case is to find a linear regression function, $f(x) = \langle w, x \rangle + b$, which can best approximate the actual output vector Y, with an error tolerance ε, and is concurrently as flat as possible. Here, $w \in R^n$, $b \in R$ are the parameter vectors of the regression function. The regression function is such that

$$|y_i - f(x_i)| \le \varepsilon, \quad i = 1, \ldots, l \quad (5)$$

The linear regression problem can be expressed as the following convex optimization problem:

$$\min_{(w,b,\xi,\xi^*)} \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \quad (6)$$

subject to

$$\begin{aligned} &f(x_i) - y_i \le \xi_i^* + \varepsilon, \quad i = 1, \ldots, l \\ &y_i - f(x_i) \le x\xi_i + \varepsilon, \quad i = 1, \ldots, l \\ &\xi_i, \xi_i^* \ge 0, \quad i = 1, \ldots, l \end{aligned} \quad (7)$$

where C indicates the capacity parameter cost, ξ_i and ξ_i^* determine the degree to which sample points are penalized if the error is larger than ε.

Most real-world problems are nonlinear. The method of solving this limitation is to map the input data into a higher dimensional feature space, and then perform the linear regression in this feature space. The decision function can be expressed as Equation (8):

$$f(w, b) = w \cdot \phi(x) + b \quad (8)$$

where $\phi(x)$ is a nonlinear function that maps x into a feature space. Similarly, the nonlinear regression problem can be expressed as the following optimization problem:

$$\min_{(w,b,\xi,\xi^*)} \frac{1}{2}w^T w + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \quad (9)$$

subject to

$$\begin{cases} y_i - (w \cdot \phi(x_i) + b) \le \varepsilon + \xi_i \\ (w \cdot \phi(x_i) + b) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots, l \end{cases} \tag{10}$$

In Equation (9), minimizing the first term is equivalent to minimizing the confidence interval of the learning machine, and minimizing the second term corresponds to minimizing the empirical risk. The parameter C controls the flatness of the regression function. An increase penalizes large errors and, consequently, leads to a decrease in approximation error. Increasing the weight vector norm $\|w\|$ can also achieve this result. Figure 2 explains that the principle of nonlinear SVM, ε is the Vapnik's insensitive loss function, ξ_i and ξ_i^* are the slack variables. $\oplus$ denotes the support vector, while Δ depicts the data within the margin. Figure 2 presents the separation ability of SVM which depends on both training error and margin. A large margin results in a low training error, but overfitting might occur. A small margin brings out a large training error. Hence, the ideal model is to reach the tradeoff between margin determination and training error control.

A technique for solving the optimization problem of Equations (9) and (10) is to use the dual form by introducing a dual set of Lagrange multipliers, α^* and α. The dual form of the nonlinear SVM can be expressed as

$$\min_{(\alpha_i^*, \alpha_i)} \frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \langle \phi(x_i) \cdot \phi(x_j) \rangle + \varepsilon \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} y_i (\alpha_i^* - \alpha_i) \tag{11}$$

subject to

$$\sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0; 0 \le \alpha_i \le C, i = 1, \ldots, l; \quad 0 \le \alpha_i^* \le C, i = 1, \ldots, l. \tag{12}$$

where $i = 1, \ldots, l$ is the sample size.

In Equation (11), it is difficult to compute $\langle \phi(x_i) \cdot \phi(x_j) \rangle$ because little knowledge may be available to select an appropriate nonlinear function, ϕ. An advantage of SVM is that the nonlinear function $\phi(x)$ need not be used. The computation in input space can be performed using the kernel function: $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Any functions that satisfy Mercer's theorem can be used as a kernel. Some commonly used kernels in SVM are linear kernel, polynomial kernel, sigmoid kernel, and radial basis function kernel. Dibike *et al.* (2001) applied different kernels in the SVM model to simulate the rainfall–runoff process and demonstrated that the radial basis function outperformed other kernel functions. In this paper, the Gaussian radial basis function is used as the kernel function:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \text{for } \gamma > 0 \tag{13}$$

Because a detailed analysis for solving Equations (11) and (12) can be found in Cristianini & Shawe-Taylor (2000), it is omitted here. The kernel function allows the decision function of the nonlinear SVM to be

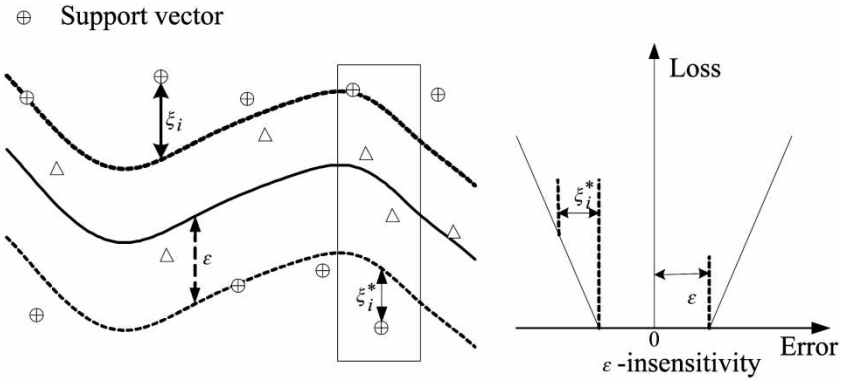

Figure 2 | Nonlinear SVM with Vapnik's ε-insensitive loss function (Yu *et al.* 2006).

expressed as follows:

$$f(x) = \sum_{i=1}^{l_k} (\alpha_i^* - \alpha_i) K(x_i, x) + b \tag{14}$$

where $l_k (l_k < l)$ is the number of selected sample points or support vectors, and K is the kernel function.

The key point for the nonlinear SVM is to find the following three parameters: the radius of the insensitive tube, ε, the cost constant, C, and the kernel parameter, γ. These parameters are mutually dependent. The detailed meanings of parameters C and ε can be found in Yu *et al.* (2006). Determining appropriate values of C and ε is often a heuristic trial-and-error process. The main methods for selecting the values of C, ε, and γ are the empirical method and the grid search method. The optimal values of SVM parameters may vary substantially among cases (Yu *et al.* 2006). In the hydrologic model, some experiences of parameter selection in the applications of SVM can be found (Mattera & Haykin 1999; Liong & Sivapragasam 2002; Choy & Chan 2003; Cherkassky & Ma 2004). The key points for parameters C and ε are: (1) C is sensitive to the results between 0 and 100, and C and ε converge quickly in the trial-and-error process (Liong & Sivapragasam 2002); (2) C should be equal to the range of output data and parameter ε should be selected such that the percentage of support vector (SV) in the SVM regression model is around half of the number of samples (Mattera & Haykin 1999); (3) C does not markedly affect the optimization results (Choy & Chan 2003). Detailed information about the parameters C and ε can be found in Cherkassky & Ma (2004).

In this study, a SVM model is developed to simulate the streamflow at the Nangao Reservoir. The streamflow responds to the precipitation and runoff from the rainfall–runoff process. The spatial distribution of precipitation is not considered. The average rainfall data were calculated using the Thiessen polygon method. The average rainfall data were used as the input data of the SVM model. In the SVM model, we introduce a new concept, N: the time of precipitation impact (TPI), which is a parameter in the model. The TPI (N) is often larger than the lag time and the concentration time. The TPI (N) is related to the lag time, concentration time, soil type, and vegetation. Therefore, N is determined as a parameter, which can be selected in the following model structure. The streamflow at time t correlates with the past streamflow at times $t-1$, $t-2$, …. So, the current and the observed times are $t-1$, $t-2$, etc., and the future (forecasted) time is t, and the future precipitation (i.e., $P(t)$) is assumed to be known in this model. Precipitation data that are assumed to be known at times $t, t-1, \ldots, t-N+1$ (N day: TPI) and streamflow data (simulated) at times $t-1, t-2, \ldots, t-N+1$ ($N-1$ day) are used to predict the streamflow at t (where t denotes the day). The forecasting model in a real time fashion can be expressed as the following equation:

$$\begin{aligned} Q_n(t) = f_{svr}(&P_n(t), P_n(t-1), \ldots, P_n(t-N+1) \\ &Q_n(t-1), Q_n(t-2), \ldots, Q_n(t-N+1)) \end{aligned} \tag{15}$$

where the function f_{svr} indicates the SVM model; t is time (day); $P_n(t)$, $P_n(t-1)$, and $P_n(t-N+1)$ are the normalized precipitation data at times t, t–1, …, t–N+1 (N day); $Q_n(t-1), Q_n(t-2), \ldots, Q_n(t-N+1)$ are the simulated normalized streamflow at time t − 1, …, $t-N+1$, (N–1 day); $Q_n(t)$ is the simulated normalized streamflow at the future time t. N is TPI of the runoff process.

Ensemble Kalman filter

The well-known extended Kalman filter (EKF) can be used for nonlinear procedures, but it is notoriously unstable if the nonlinearities are strong. Error covariance integration for large-scale environmental systems results in relatively large computational demand (Gelb 1974). To overcome these limitations, the EnKF was introduced by Evensen (1994). The EnKF has an advantage over the other filtering techniques as it uses a limited number of model states and results in faster convergence. The EnKF has been widely used in data assimilation of hydrological fields for several reasons. The sequential structure is convenient for processing remotely sensed measurements in real time, it provides information on the accuracy of its estimates, and it is relatively easy to implement even if the land surface model and measurement equations include thresholds and other nonlinearities. Moreover, the EnKF is able to account for a wide range of possible model error, but it relies on a number of assumptions and approximations that may compromise its performance in certain situations.

The framework of the EnKF mainly includes two processes, forecasting and updating. In the EnKF, the forecasting model is executed for each ensemble member i.

$$Q_{\mathrm{n}}^{i-}(t) = f_{\mathrm{svr}}(P_{\mathrm{n}}^{i}(t),\ \ldots,\ P_{\mathrm{n}}^{i}(t-N+1),\ Q_{\mathrm{n}}^{i+}(t-1),\ \ldots,\ Q_{\mathrm{n}}^{i+}(t-N+1)) + W^{i}(t), \quad i = 1,\ \ldots,\ M \tag{16}$$

$$P_{\mathrm{n}}^{i}(\tau) = P_{\mathrm{n}}(\tau) + \zeta^{i}(\tau),\ \zeta^{i}(\tau) - N(0, R(\tau)) \tag{17}$$

where, the nonlinear function f_{svr} is a regression function (see Equation (15)); $Q_{\mathrm{n}}^{i-}(t)$ is the ith ensemble member prediction at time t. Q_{n}^{i+} stands for the ith updated ensemble member; $P_{\mathrm{n}}^{i}(\tau)$ is the ith perturbed forcing variables at time τ, $\tau = t, t-1, \ldots, t-N+1$; $W^{i}(t)$ $(-N(0,\Phi(t)))$ and $\zeta^{i}(\tau)(-N(0,R(\tau)))$ represents the forecast model noise and forcing noise which are assumed to be independent of white noises and normal distributions; $\Phi(t)$ and $R(\tau)$ show forecast noise covariance and observation noise covariance, respectively; M is the ensemble member. The observation map can be expressed using the following equation; the '+' superscript denotes that the estimate is a posteriori, the '$-$' superscript denotes that the estimate is *a priori*.

$$Q_{\mathrm{n,obs}}^{i}(t) = HQ_{\mathrm{n}}(t) + \varepsilon^{i}(t), \varepsilon^{i}(t) - N(0, S(t)) \tag{18}$$

where H is an observation map that converts the simulated hydrological catchment state $Q_{\mathrm{n}}(t)$ to the observation $Q_{\mathrm{n,obs}}^{i}(t)$ and $\varepsilon^{i}(t)$ represents the noise term with zero mean and specified covariance $S(t)$. The other process of the EnKF is updating, which is also defined as a corrector. When observational data are available, the corrector incorporates a new measurement into the priori estimate to obtain an improved posteriori estimate at the same time. The related calculations are as follows:

$$Q_{\mathrm{n}}^{i+}(t) = Q_{\mathrm{n}}^{i-}(t) + K_t(Q_{\mathrm{n,obs}}^{i}(t) - Q_{\mathrm{n}}^{i-}(t)) \tag{19}$$

$$K_t = F_t^{-}H^{T}(HF_t^{-}H^{T} + S_{t+1})^{-1} \tag{20}$$

$$F_t^{-} = \frac{1}{M-1}\sum_{i=1}^{M}[(Q_{\mathrm{n}}^{i-}(t) + <Q_{\mathrm{n}}^{-}(t)>)(Q_{\mathrm{n}}^{i-}(t) - <Q_{\mathrm{n}}^{-}(t)>)]^{-1} \tag{21}$$

$$<Q_{\mathrm{n}}^{-}(t)> = \frac{1}{M}\sum_{i=1}^{M}<Q_{\mathrm{n}}^{i-}(t)> \tag{22}$$

where Equation (19) calculates the new optimal estimate vector of hydrological states after assimilation; K_t is the gain-term (also known as Kalman gain) which calculates the weight between the forecast model and observation states; F_t^{-} is the prior estimate of error covariance and approaches zero.

Here, we will couple the SVM model and the EnKF method to estimate the state variables and determine the impact of the assimilation time scale (ATS) on the simulation results. When the precipitation and evapotranspiration are available daily, but the streamflow data are not available at all times, the impact of the assimilation time scale on the simulation results is an important problem. In this paper, we suppose that the streamflow data become available in ATS-day sets (it is not available for the next ATS days and so on). The ATS $= j$ day means that there is not assimilation for j continuous days, then there is assimilation for continuous j days (assimilation at every time step), respectively. The longer the continuous assimilation time, the more improved the model structure. The algorithm of the SVM +EnKF model is described in the followings steps (Figure 3).

Step 1: Initializing state estimate sample (M), TPI (N) and the length of the simulation time (T). It is necessary to choose an ensemble of initial state estimates that captures the initial probability distribution. The precipitation and the discharge are initialized through perturbations via adding noise to original values (Georgakakos 1986; Weerts & Serafy 2006).

$$P(t) = P_{\mathrm{obs}}(t) + (0.15 \times P_{\mathrm{obs}}(t) + 0.2)^{2} \tag{23}$$

$$Q(t) = Q_{\mathrm{obs}}(t) + (0.1 \times Q_{\mathrm{obs}}(t))^{2} \tag{24}$$

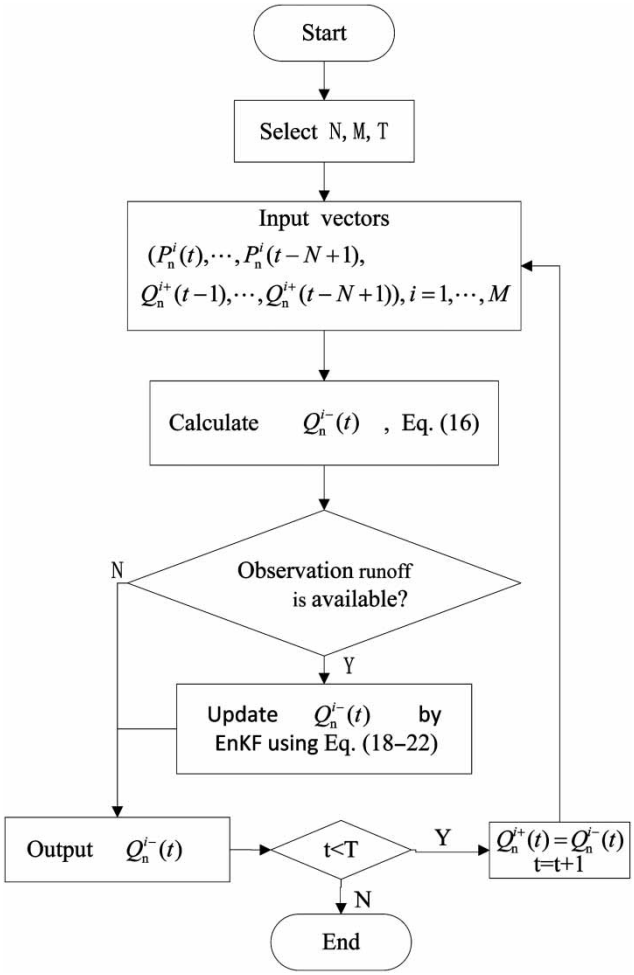

Figure 3 | Flowchart of the SVM + EnKF model.

Step 2: Running the SVM model. The SVM model is used to predict the future model state as the prediction propagates the state forward in time. In this study, N precipitation values and $N-1$ runoff values are used as input data in the SVM model to predict runoff. Only the first $N-1$ runoff observation datum is applied to predict the Nth runoff value, then the predicted value is added to the input sequence to predict the next step.

Step 3: Updating by the EnKF. When the observed runoff is available, the EnKF is applied to update the predictive runoff as Equations (18)–(22). There is a difference between the EnKF and the classical filter. In EnKF, the measurements are treated as random variables obtained from an ensemble. The ensemble is generated by the Monte Carlo method from a distribution with mean equal to the first guess observation and the covariance equal to W.

Step 4: Coupling the SVM and the EnKF. The result of previous calibration by the EnKF is fed back into the SVM model as an input to perform the next time step in the forecast.

The Xinanjiang RR model and parameter calibration method

The Xinanjiang RR model as the traditional conceptual RR model is employed as a comparison with the SVM model and the SVM+EnKF model. Thus the streamflows are also simulated by the Xinanjiang RR model. The Xinanjiang RR model was first published in 1980 (Zhao 1992). It is a conceptual watershed model. The main feature of the Xinanjiang RR model is the concept of runoff formation on repletion of storage, which means the runoff production occurs only when the soil moisture content of the aeration zone reaches field capacity. The Xinanjiang RR model consists of four major parts: evapotranspiration, runoff production, runoff separation, and flow routing. Figure 4 shows the structure of the Xinanjiang RR model.

All symbols inside the blocks are variables while those outside the blocks are parameters. The variables include inputs, outputs, state variables, and internal variables. The inputs are areal mean rainfall, P, and measured pan evaporation, EP. The discharge, TR, from the whole basin and the actual evapotranspiration, EA, which includes three components EU, EL, and ED are outputs. The state variables include the areal mean free water storage, S, and the areal mean tension water storage, W which has three components WU, WL, and WD in the upper, lower, and deeper layers, respectively. In addition, the FR is a runoff contributing area factor relating to W. The remaining symbols are internal variables, which include R, RB, and $QSIM$. R indicates the runoff produced from the previous area, which is divided into three components RS, RI, and RG regarded as surface runoff, interflow, and groundwater runoff, respectively. The three components are further separated into TRS, TRI, and TRG and together form the total inflow to the channel network of the sub-basin. The $QSIM$ is the outflow of the sub-basin. K is the ratio of potential evapotranspiration to pan evaporation. WM shows the areal mean tension water

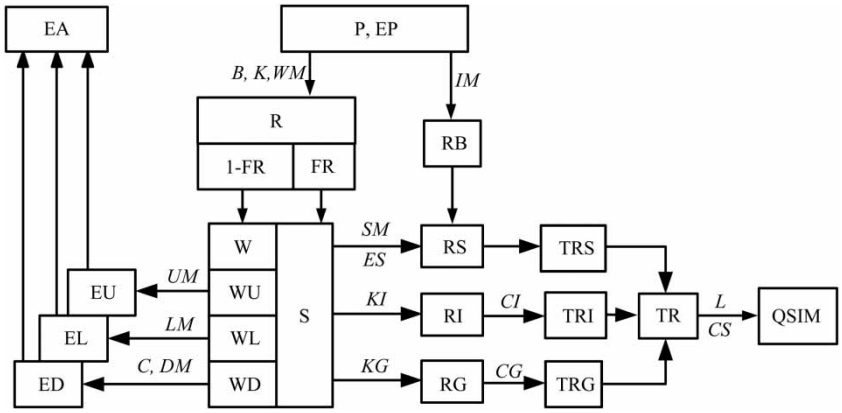

Figure 4 | Structure of the Xinanjiang RR model (Zhao 1992).

capacity including three components which are *UM*, *LM*, and *DM*. The tension water capacity distribution curve is indicated by *B*. *IM* shows the factor of an impervious area. *SM* is the areal mean free water capacity, and *ES* is the free water capacity distribution curve. *KI* and *KG* show coefficients relating to *RI* and *RG*. The other parameters, *CI*, *CG*, *L*, and *CS* are for flow routing. In this paper, a basic genetic algorithm (GA) was employed to calibrate the Xinanjiang RR model parameters. The flowchart of GA can be found in Gorges-Schleuter (1989).

Some parameters of GA need to be chosen in order to obtain good performance, such as the moderate population size (P_{size}), a high crossover probability (P_c), and a low mutation probability (P_m). P_{size} critically affects the efficiency and solution quality of the GA. Generally, P_{size} is set to be a value between 150 and 300. P_c controls the frequency of crossover operation. In general, P_c is chosen between 0.5 and 0.8. P_m is a critical factor in extending the diversity of the population. P_m is often chosen between 0.001 and 0.1.

Statistical analysis

The root mean square error (RMSE), the Pearson's correlation (R), and the mean bias error (MBE) of the observed and simulated streamflow are used to assess the performance of the data assimilation scheme. The Nash–Sutcliffe coefficient (Nash & Sutcliffe 1970) of efficiency (CE) is used to analyze the simulated results. The CE has been widely used to evaluate the performance of hydrologic models. It is the ratio of the mean square error to the variance in the measured data subtracted from unity. It can range from $-\infty$ to 1. An efficiency of 1 (CE=1) corresponds to a perfect match of modeled discharge to the observed data. An efficiency of 0 (CE=0) indicates that the model predictions are as accurate as the mean of the observed data, whereas an efficiency less than zero ($CE < 0$) occurs when the observed mean is a better predictor than the model or, in other words, when the model residual (described by the numerator), is larger than the data variance (described by the denominator). The CE is an improvement over RMSE for model evaluation purposes because it is sensitive to differences in the measured and model estimated means and variances.

$$\text{RMSE} = \left(\frac{1}{LL} \sum_{t=1}^{LL} (Q_{\text{sim}}(t) - Q_{\text{obs}}(t))^2 \right)^{1/2} \tag{25}$$

$$R = \frac{\sum_{t=1}^{LL} (Q_{\text{sim}}(t) - \bar{Q}_{\text{sim}})(Q_{\text{obs}}(t) - \bar{Q}_{\text{obs}})}{\sqrt{\sum_{t=1}^{LL} (Q_{\text{sim}}(t) - \bar{Q}_{\text{sim}})^2 \sum_{t=1}^{n} (Q_{\text{obs}}(t) - \bar{Q}_{\text{obs}})^2}} \tag{26}$$

$$\text{MBE} = \frac{1}{LL} \sum_{t=1}^{LL} (Q_{\text{sim}}(t) - Q_{\text{obs}}(t)) \tag{27}$$

$$CE = 1 - \frac{\sum_{t=1}^{LL} (Q_{\text{obs}}(t) - Q_{\text{sim}}(t))^2}{\sum_{t=1}^{LL} (Q_{\text{obs}}(t) - \bar{Q}_{\text{obs}})^2} \tag{28}$$

where $Q_{\text{sim}}(t)$ is the reverse scaled streamflow of $<Q_n^{i+}(t)>$ using Equation (3) and is the simulated streamflow at time

t, $\bar{Q}_{sim}$ is the average value of $Q_{sim}(t)$, $Q_{obs}(t)$ is the measured streamflow at time t, and $\bar{Q}_{obs}$ is the average of $Q_{obs}(t)$. LL is the total number of observations.

RESULTS AND DISCUSSION

Calibration of the SVM model

The streamflow is directly affected by the soil moisture, rainfall, runoff, and evapotranspiration. It is difficult for the linear SVM to handle land surface data. Hence, in this paper, a nonlinear SVM is used for flood forecast and the Gaussian radial basis function is employed as a kernel function. Calibration is a process of standardizing predicted values, using deviations from observed values for a particular area to derive correction factors that can be applied to generate predicted values that are consistent with the observed values. The calibration process can provide important insight into both local conditions and model performance. If correction factors are large or inconsistent across several study areas, it suggests that some significant component of the hydrologic system or its controls is being neglected. Several methods of calibration are available based on methods such as artificial neural networks, multiple objective optimal methods, and nonlinear regression models. Choosing an approach depends on the purpose of the model, the model parameters or variables involved. The calibration of long-term rainfall–runoff models is based on long-term trends rather than on individual events. In this paper, the trial-and-error technique (the grid search method) is employed for the SVM model calibration.

The number N reflects that the forecasted runoff is related to the precipitation. When the watershed area, watershed topography, watershed underlying surface, and the concentration time are different, N is different too. For Luo River Basin, the concentration time is estimated using the Kerby equation (Chow *et al.* 1988):

$$t = G(Lr/S^{0.5})^{0.467} \tag{29}$$

where t is time of concentration; G is regional constant; L is longest watercourse length in the watershed (ft); r is Kerby retardance roughness coefficient; S is average slope of the watercourse, (m/m). In the Luo River basin: $G = 0.83$; $L = 52{,}000$ m; $S = 0.03$ m/m; and $r = 0.6$. The calculated concentration time is about 6.9 hours. Thus, in this paper, because the watershed area is not large and the concentration time is less than 1 day, we only study the case of $N = 2$ days (see Equation (15)).

Obtaining the optimal prediction depends on having suitable model parameters. The nonlinear SVM model parameters include the capacity parameters cost C, the radius of the insensitive tube ε, and kernel parameter γ. The coefficient C which affects the trade-off between complexity and the proportion of non-separable samples, must be selected by the user. If it is too large, there is a penalty for non-separable points, and there may be overfitting with the support vectors. If it is too small, there may be underfitting. The value of ε determines the level of accuracy of the approximated function. It relies entirely on the target values in the training set. If ε is larger than the range of the target values, the simulated results may be poor. If ε is zero, the results may be overfitted. Therefore, ε is selected to reflect the data in some way. The best combination of (C, γ, ε) is often determined by a grid search with exponentially growing sequences (Hsu *et al.* 2003). The final model, which is used for testing new data, is then trained on the whole training set using the selected parameters (Luchetta & Manetti 2003; Goswami *et al.* 2005).

The code of the SVM model is written in Matlab language, and the SVM model is trained on eight years of data (year, 1994–2001) for every case. Here, the trial-and-error technique and grid search method are employed to select parameters C, ε, and γ for each case. The grid search method is a straightforward and exhaustive method. This method may be time-consuming, so Hsu *et al.* (2003) suggested the application of a two-step grid search method, applied on exponentially growing grids of parameters. First, a coarse grid search (for example, $C = 2^{-5}, 2^{-3}, \ldots, 2^{9}$; $\varepsilon = 2^{-11}, 2^{-9}, \ldots, 2^{-1}$; $\gamma = 2^{-7}, 2^{-5}, \ldots, 2^{3}$) was used to determine the best region of these three-dimensional grids. Then, a finer grid search (for example, $C = 2^{0.5}$, $2^{0.75}, \ldots, 2^{1.25}$; $\varepsilon = 2^{-3}, 2^{-2.75}, \ldots, 2^{-1}$; $\gamma = 2^{-1.5}, 2^{-1.25}, \ldots, 2^{1.5}$) was conducted to find the optimal parameters. The RMSE was used to optimize the parameters.

The optimal results including length of training sample, percentage of SV, parameters (C, ε, γ), RMSE, and R for SVM ($N = 2$ days) are as follows: length of training sample is

2,910, percentage of SV is 53.5%, $C = 3.414$, $\varepsilon = 0.361$, $\gamma = 1.125$, RMSE = 5.05 m^3/s, and $R = 0.996$. Figure 5 gives the simulated results during the calibration period and the correlation coefficient using eight years of data. It is obvious that the SVM ($N = 2$) can simulate accurately the runoff process during the calibration period, but simulations of flood peaks are overestimated. The Pearson's correlation coefficient is 0.996. During the calibration, the parameter C is selected based on RMSE. When C is too small, the SVM is underfitting and RMSE is large. The error gradually reduces with increasing C. When C becomes too large, the SVM is overfitting and the error gradually increases. There is an interval for suitable C. In this interval, the value of C has little impact on the simulation error. So, for fixed ε and γ, C belongs to a fixed interval that provides good simulation results. For selecting of ε, a suitable interval exists. In this interval, the change of ε has little impact on RMSE. If ε is large, the length of SV is short and the regression function is simple, the computing speed is fast, and the RMSE is large. This is an underfitting condition.

Calibration of the Xinanjiang RR model

There are 14 parameters in the Xinanjiang RR model. Some have a range of values, and some have a fixed value (Zhao 1992). Output results are quite sensitive to some of the parameters, and a little change may bring about a large change in the simulated results. The other parameters have little impact on the model results. Many methods can be used to calibrate hydrologic model parameters. Xu *et al.* (2013) used three different optimization methods to calibrate the Xinanjiang streamflow model, including genetic algorithm (GA), shuffled complex evolution of the University of Arizona (SCE-UA), and the recently developed shuffled complex evolution Metropolis algorithm of the University of Arizona (SCEM-UA). In the Xinanjiang RR model's parameters, L (Figure 4) is an integer, equal to 0, 1, 2, or 3. According to the characteristics of the study area, we took $L = 0$ (constant) in this study. Lü *et al.* (2013) studied the sensitivity of the Xinanjiang RR model parameters in this watershed. They found that the parameters B, S_m, K_g, C_g, and C_i were sensitive, and the other parameters were not sensitive (Lü *et al.* 2013). So, the parameters B, S_m, K_g, C_g, and C_i were adjusted using the basic genetic algorithm. The other parameters were selected from the literature (Ju *et al.* 2009; Lü *et al.* 2013). Nangao Reservoir rainfall–runoff data from 1994 to 2001, a total of eight continuous years, are used for calibration. In this paper, the parameters in GA are selected as follows: $P_{size} = 100$, $P_c = 0.65$, and $P_m = 0.05$ (Gorges-Schleuter 1989). The evolution number of generation Gmax = 2,000. The GA provides a good objective value when the evolution number of generation is near 1,000. Table 1 shows the calibrated Xinanjiang RR model parameter values. Figure 6 shows the simulated discharge vs observed discharge during the calibration for the Xinanjiang RR model. It is very clear that the Xinanjiang RR model as a traditional hydrological model still exploits its advantages, and its R value is 0.999.

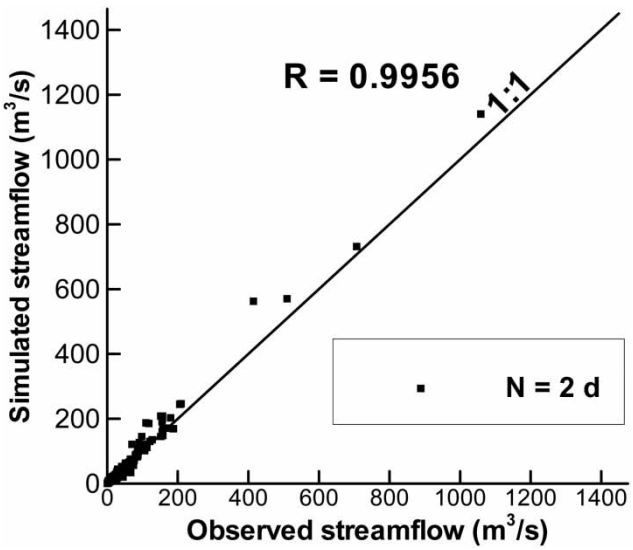

Figure 5 | Calibration results of the SVM model for $N = 2$ days: comparison between observed and simulated daily runoff (1994–2001).

Simulation of the SVM model for $N = 2$

In this study, the rainfall–runoff process from DOY (day of year) 206 (July 25, 2002) to DOY 288 (Oct. 15, 2002) was used to test the SVM model, the EnKF + SVM model, and the Xinanjiang RR model. The simulated results of the rainfall–runoff process for the three models are compared, and the advantages and disadvantages for the three models are analyzed. For the sake of convenience, Figure 7 shows the time series of the mean precipitation data between DOY 206 and DOY 288, 2002 at the Nangao experimental site.

Table 1 | Parameter values of the Xiananjian RR model

Parameter	Physical meaning	Value
U_M (mm)	Averaged soil moisture storage capacity of the upper layer	20
L_M (mm)	Averaged soil moisture storage capacity of the lower layer	60
D_M (mm)	Averaged soil moisture storage capacity of the deep layer	40
B	Exponential of the distribution to tension water capacity	0.3
I_m (%)	Percentage of impervious and saturated areas in the catchment	0.02
C	Coefficient of the deep layer, that depends on the proportion of the basin area covered by vegetation with deep roots	0.15
S_m (mm)	Areal mean free water capacity of the surface soil layer, which represents the maximum possible deficit of free water storage	40
E_s	Exponent of the free water capacity curve influencing the development of the saturated area	1.2
K_g	Outflow coefficients of the free water storage to groundwater relationships	0.39
K_i	Outflow coefficients of the free water storage to interflow relationships	0.31
C_g	Recession constant of the groundwater storage	0.992
C_i	Recession constant of the lower interflow storage	0.7
C_s	Recession constant in the lag and route method for routing through the channel system within each sub-basin	0.3
L (d)	Lag in time empirical value	0

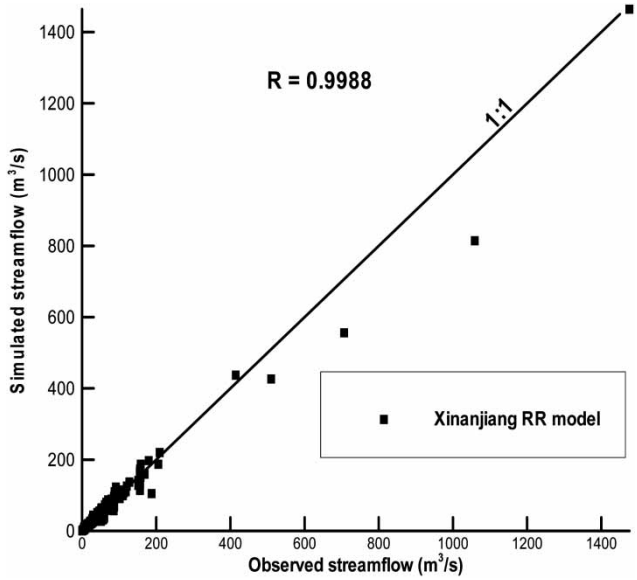

Figure 6 | Calibration results for the Xinanjiang RR model: comparison between observed and simulated daily runoff (1994–2001).

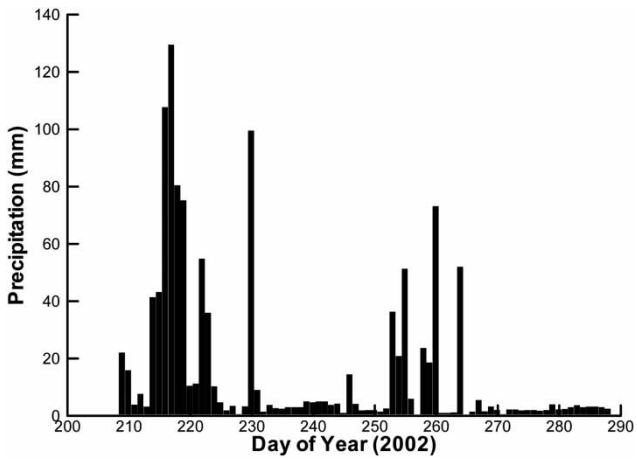

Figure 7 | Recorded precipitation at the Nangao experimental site (DOY 206–288, 2002).

The average daily precipitation in this period was 12.1 mm, and three major rainfalls occurred, 107 mm on DOY 216, 130 mm on DOY 217, and 100 mm on DOY 230. During DOY 214 to DOY 224, the number of rainfall days in this period was 11 days. It was 13% of the total simulated period. The average daily rainfall was 54.6 mm and the average observed runoff was 78.9 m^3/s during this period.

Figure 8 shows the simulated RR results between DOY 206 and DOY 288, 2002 using the SVM model without EnKF data assimilation. According to the calibration results, the SVM model for $N = 2$ days is considered. The parameters and the length of the SV are presented in the calibration section. The trend of the simulated streamflow is consistent with the observed results. The time of the flood peak is accurately captured. The maximum precipitation which occurred on DOY 217 is 130 mm. The time of maximum simulated discharge also occurred on DOY 217. There are small differences between the observed discharge and the simulated discharge after the flood peak. At DOY 217, the observed discharge is 263 m^3/s, and the simulated discharge is 336 m^3/s. On the other hand, the flood peak discharge is overestimated at DOY 230. For the remaining precipitations during the simulation period, the peak discharges are also

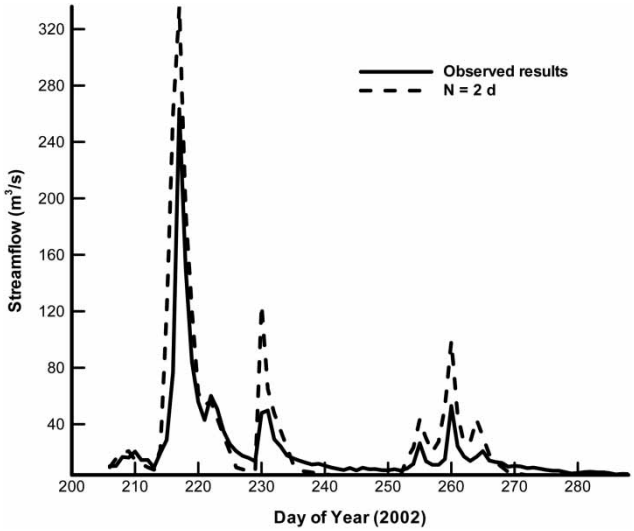

Figure 8 | Comparison between observed and simulated daily runoff (DOY 206–288, 2002) using the SVM model for $N = 2$ days.

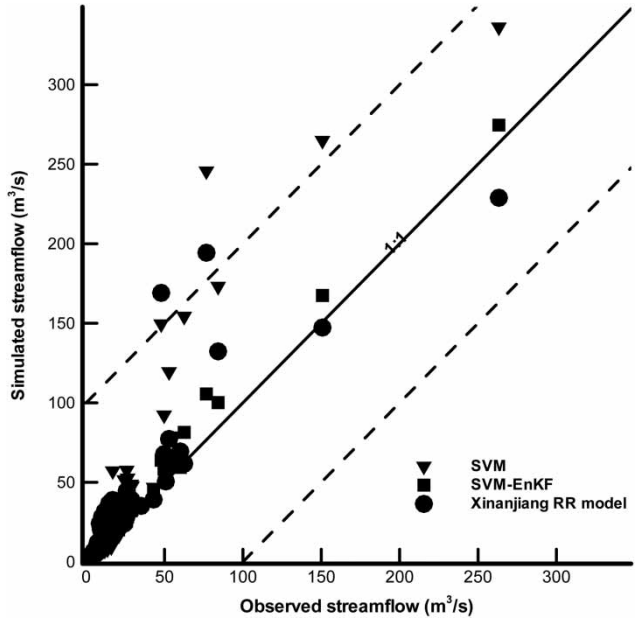

Figure 9 | Comparison between observed and simulated daily runoff (DOY 206–288, 2002) using the SVM ($N = 2$ days), the SVM-EnKF ($N = 2$ days, ATS = 1 day), and the Xinanjiang RR models. The solid line is $y = x$ and the dashed lines are $y = x - 100$ and $y = x + 100$.

overestimated. So the SVM model ($N = 2$) often overestimates the flood peak. The reasons for overestimating are: (1) spatial distribution of precipitation is not considered; (2) TPI ($N = 2$) may be small. When the TPI (N) is small, the SVM model does not consider the impact of the baseflow on the current discharge. When a storm occurs, the discharge increases rapidly and when the storm finishes, the discharge rapidly regresses; and (3) during the SVM calibration period, the simulations overestimate the streamflow for flood peak. Thus, the model parameters for SVM reflect an overestimation for flood peak (Figure 5). Hence, during the simulated period, the simulation often overestimated the flood peak.

After the flood peak, the simulation value is similar to the observation value. So, the TPI (N) has an important impact on the recession limb in the Luo River Basin. The statistical characteristics of the simulated daily runoff during the whole simulated period are shown in the following: Min: 5.17 m^3/s, Max: 336 m^3/s, RMSE: 30.6 m^3/s, R: 0.928, MBE: 10.9 m^3/s, *CE*: 0.202. It is obvious from MBE that the discharges are overestimated during the simulation period.

The differences between the SVM model, the SVM + EnKF model, and the Xinanjiang RR model

Figure 9 shows a scatter plot of the simulated streamflow vs the observed streamflow for the SVM, the SVM + ENKF, and the Xinanjiang RR model. The simulated period is also DOY 206 to DOY 288, 2002. To compare the differences between the SVM and the SVM + EnKF models, we still select $N = 2$ days for the SVM model. The simulated results are calibrated using the EnKF assimilation method. The assimilation time scale is 1 day (or ATS = 1 day). ATS = 1 day means the SVM simulation data are assimilated in the first day. The data are not assimilated in the second day. The Xinanjiang RR model is used as a comparison with the SVM ($N = 2$ day) model and the SVM + EnKF ($N = 2$ day, ATS = 1 day). The parameters of the Xinanjiang RR model have been calibrated as described in the above section. Figure 9 indicates that the simulation overestimated the streamflow for all methods. The SVM + EnKF model performed better than the SVM model. For the simulated discharge of 83 days (the whole test period), all of the scatter plot point for the SVM + EnKF method fell in the strip region (Figure 9). Two points fell outside the strip region for the Xinanjiang RR model, and five points fell outside the strip region or on the boundary of the strip region for the SVM model. The statistical characteristics (Table 2) show that the ratios of the RMSE are 30.6 (SVM) vs 13.5 (SVM + EnKF) vs 20.8 (Xinanjiang RR model). The RMSE

Table 2 | Statistical characteristics of simulated daily runoff (DOY 206–288, 2002) using three different methods

Method	Min (m^3/s)	Max (m^3/s)	RMSE (m^3/s)	R (–)	MBE (m^3/s)	CE (–)
SVM ($N = 2$)	5.17	336	30.6	0.928	10.9	0.202
SVM-EnKF ($N = 2$, ATS = 1)	4.66	275	13.5	0.947	4.25	0.839
Xinanjiang	5.50	229	20.8	0.876	7.40	0.631

of the SVM + EnKF model is the smallest. The R of the three methods is relatively large with a minimum value of 0.876 for the Xinanjiang RR model. From the MBE analysis, the three methods overestimated the streamflow. The MBE of the SVM + EnKF is the smallest. The CE of the SVM + EnKF is the largest, the CE of the Xinanjiang RR model is the second largest, and the CE of the SVM methods is the smallest. Overall, the SVM + EnKF model provides the best simulated results. These results indicate that data assimilation can improve model structure and enhance predicting precision.

The SVM model and the Xinanjiang model only consider the calibrated parameters using past data to simulate the future rainfall–runoff process. The SVM + EnKF model is adjusted by using current observed data. Thus, the simulated results of the SVM + EnKF model during the test period are better than those from the other methods.

The impact of the ATS on the SVM + EnKF

In the actual study of rainfall–runoff simulation using the SVM + EnKF model, data assimilation is only performed on the occasions when the observations are available. Due to various conditions, the observations are not always available. Therefore, in the following experiments, we assume the observations on every ATS day (ATS = 2, 3, 5 days) are nonexistent. Thus, ATS days with available observations and ATS days void of observations alternate regularly. Figure 10 shows simulated results of three cases for the SVM + EnKF model at $N = 2$ days and ATS = 2, 3, or 5 days. Figures 10(a) and 10(b) describe the simulated results around two larger peaks. Figure 10(c) shows the hydrograph in the entire simulation period (DOY 206–288, 2002). Figure 10(d) also expresses the simulated results of three cases by a scatter

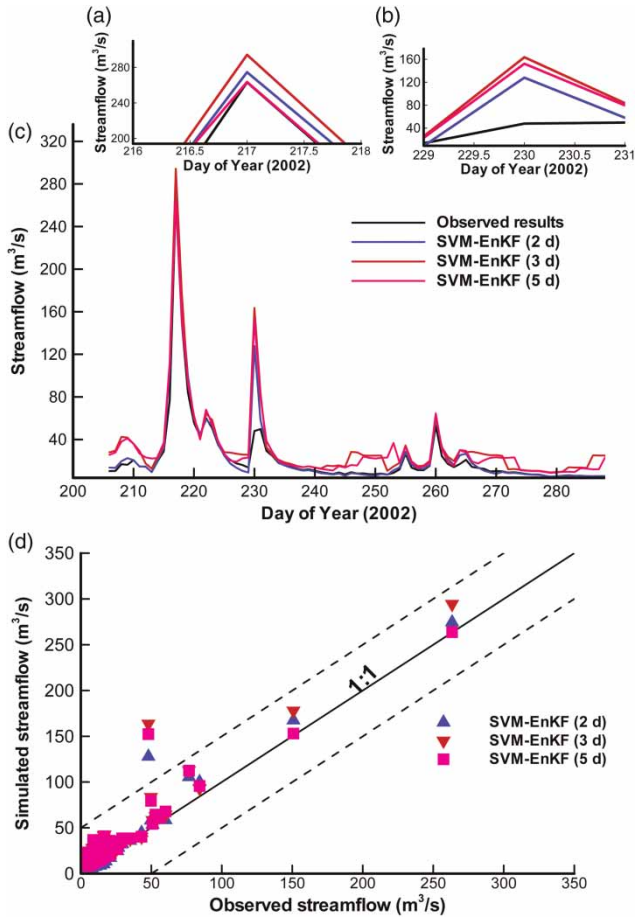

Figure 10 | Comparison of the simulated results (DOY 206–288, 2002) among the SVM-EnKF model for $N = 2$ days, ATS = 2, 3, and 5 days: (a) simulated streamflow during the period from DOY 216 to 218; (b) simulated streamflow during the period from DOY 229 to DOY 231; (c) simulated streamflow during the period from DOY 206 to day 290; (d) scatter plot in the simulated period. The solid line is $y = x$ and the dashed lines are $y = x - 50$ and $y = x + 50$.

plot. The statistical characteristics for the simulated results are shown in Table 3. From Figures 10(a) and 10(c), we find that the simulations of the first flood peak at DOY 217, 2002 are very close to the observations for all of the cases, but slight overestimates still exist. For example, the actual discharge is 263 m^3/s, the simulated values are 274 m^3/s for ATS = 2 days, 291 m^3/s for ATS = 3 days and 264 m^3/s for ATS = 5 days. For the other flood peak (Figure 10(b)) occurring at DOY 230, 2002, the runoff is also overestimated. The observation is 48 m^3/s, but the simulation is 128 m^3/s for ATS = 2 days, 161 m^3/s for ATS = 3 days, and 152 m^3/s for ATS = 5 days. The SVM model is known to often overestimate discharge. If the flood peak

Table 3 | Statistical characteristics of simulated daily runoff (DOY 206–288, 2002) using the SVM + EnKF method for $N = 2$ and different ATS

ATS	RMSE (m^3/s)	R (–)	MBE (m^3/s)	CE (–)
2	13.9	0.959	4.35	0.834
3	12.7	0.961	3.95	0.863
5	11.7	0.948	2.69	0.884

occurs at the period of assimilation, the flood peak is slightly overestimated. If the flood peak occurs at the non-assimilation period, the flood peak is often overestimated. When ATS = 2, 3, DOY 217 belongs to the period of assimilation, the discharge is slightly overestimated. When ATS = 5, DOY 217 belongs to the period of non-assimilation, but, DOY 211–215 belongs to the period of assimilation, the model can be improved after a continuous 5 days' assimilation. So, the simulated flood peak is very close to the observed flood peak at DOY 217. For the second peak, DOY 230 belongs to the period of non-assimilation, and the simulated discharge overestimated the observed discharge.

The ATS has an obvious impact on the simulation results. The statistical analysis (Table 3) shows that the errors of simulation using the SVM + EnKF model with different ATS (ATS = 2, 3, 5 days) vary. The R values range from 0.948 to 0.961. The ratio of the RMSE is 13.9 m^3/s vs 12.7 m^3/s vs 11.7 m^3/s, and the ratio of the CE is 0.835 vs 0.863 vs 0.884 for ATS = 2, 3, 5 days. These results indicate that all cases of AST = 2, 3, and 5 days can effectively simulate the streamflow in the study area. However, the case of ATS = 5 days is the best. The reason is that the continuous calibrated time (ATS = 5 days) is longer than that for the other cases, so the model structure is improved. This improves significantly the simulation results. The SVM + EnKF model for any assimilation time scale behaves better than the SVM model.

CONCLUSIONS

Within the watershed scale, the rainfall–runoff process is very complex. Although there are many RR models which are widely applied to predict stream flow, due to nonlinearities and frequent human activities, uncertainty in the model parameters forces data uncertainty, and the traditional method to determine model parameters using historical data is constrained. The SVM has a strict theoretical foundation, good generalization capability, and no local minima. The SVM can solve multi-dimensional problems. Thus the SVM is attractive, and becomes a focus of machine learning after neural network, and is used to solve the system identification problem after its successful application in pattern recognition, regression, and control theory. Recent developments of ensemble-based techniques and filtering strategies for sequential data assimilation make it possible to implement techniques that properly account for data (input and output), state variables, parameters, and model structural uncertainties, although the details of how the implementations can be achieved are still an active area of research.

In this paper, the SVM RR models ($N = 2$ days) are established to simulate the streamflow at the Nangao Reservoir. Eight years of RR data sets are used to train the SVM model. An 83 day RR data set (DOY 206, 2002 to DOY 288, 2002) is used to test the SVM model. The grid search method is employed to obtain the SVM parameters. Then, the coupled SVM and the EnKF (SVM + EnKF) models are used to simulate the RR process at the same catchment and the same time period (DOY 206, 2002 to DOY 288, 2002). For the SVM + EnKF, different assimilation time scales (ATS = 1, 2, 3, 5 days) are considered. The simulation results are compared with the SVM model. A traditional RR model (Xinanjiang RR model) is used as a reference of the SVM model and the SVM + EnKF model. The parameters of the Xinanjiang RR model are optimized by the GA method. The calibration period and the test period are the same as for the SVM model. RMSE, MBE, R, and CE are applied to evaluate the models.

During the calibration period, the grid search method is used to obtain a group of suitable SVM model parameters. The number of SV in the SVM model is maintained at about 50%. The GA method also is applied to obtain a group of suitable parameters for the Xinanjiang RR model. During the training period, the R is very high.

During the simulation period, the results of the SVM models show that the discharges of the flood peak are overestimated. The simulation is affected by the training data and the length of the SV. A comparison of the SVM model ($N = 2$ days), the SVM + EnKF model ($N = 2$ days, ATS = 1 day), and the Xinanjiang RR model shows that

the streamflows are overestimated by all of the models. The SVM + EnKF model provides the best simulated results. These results indicate that data assimilation can improve model structure and enhance predicting precision. For the SVM + EnKF model ($N = 2$ days and ATS = 1 day), the simulation results show the SVM + EnKF model is better than the SVM model. That is to say, the assimilation for the SVM model can improve the simulation results. The SVM + EnKF model ($N = 2$ days and ATS = 2, 3, 5 days) shows that: (1) the SVM + EnKF model for any assimilation time scale is better than the SVM model; and (2) the assimilation time scale has an important impact on the simulated results. The case for $N = 2$ days and ATS = 5 days is better than the other cases. If the continuous calibrated time is long, the model structure is improved and the simulation results are significantly improved.

The TPI (N) also affects the simulation results. When the TPI (N) is small, the impact of the past precipitation on the current discharge is not considered completely in the SVM model. When a storm occurs, the simulation discharge increases rapidly. After the storm finishes, the discharge rapidly regresses. Thus, the TPI (N) has an important impact on the recession limb in the watershed hydrology model.

ACKNOWLEDGEMENTS

This research is supported by the National Basic Research Program of China (2013CBA01806, 2010CB951101) and the NNSF of China (51190090, 41371049, 50939006, IWHR-SKL-201213).

REFERENCES

Boser, B. E., Guyon, I. & Vapnik, V. N. 1992 A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, 5, 144–152, July 27–29, ACM, Pittsburgh, PA.

Bray, M. & Han, D. 2004 Identification of support vector machines for runoff modeling. *J. Hydroinform.* **6**, 265–280.

Chang, F. J. & Chen, Y. C. 2001 A counter propagation fuzzy-neural network modeling approach to real time streamflow prediction. *J. Hydrol.* **245**, 153–164.

Cheng, C. T., Ou, C. P. & Chau, K. W. 2002 Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall-runoff model calibration. *J. Hydrol.* **268**, 72–86.

Cherkassky, V. & Ma, Y. 2004 Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Netw.* **17**, 113–126.

Chow, V. T., Maidment, D. R. & Mays, L. W. 1988 *Applied Hydrology*. McGraw-Hill Book Company, New York.

Choy, K. & Chan, C. 2003 Modeling of river discharges and rainfall using radial basis networks based on support vector regression. *Int. J. Syst. Sci.* **34**, 763–773.

Cortes, C. & Vapnik, V. 1995 Support vector networks. *Mach. Learn.* **20**, 1–25.

Coulibaly, P., Anctil, F. & Bobée, B. 2000 Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *J. Hydrol.* **230**, 244–257.

Cristianini, N. & Shawe-Taylor, J. 2000 *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK.

Crow, W. & Wood, E. 2003 The assimilation of remotely sensed soil brightness temperature imagery into a land-surface model using ensemble Kalman filtering: a case study based on ESTAR measurements during SGP97. *Adv. Water Resour.* **26**, 137–149.

Dibike, Y. B., Velickov, S., Solomatine, D. & Abbott, M. 2001 Model induction with support vector machines: introduction and applications. *J. Comput. Civil Eng.* **15**, 208–216.

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.

Gelb, A. 1974 *Applied Optimal Estimation*. The MIT Press, Cambridge, MA.

Georgakakos, K. 1986 A generalized stochastic hydrometeorological model for flood and flash-flood forecasting Part II. Case studies. *Water Resour. Res.* **22**, 2096–2106.

Gorges-Schleuter, M. 1989 Asparagos: a population genetics approach to genetic algorithms. M Venue: In: *Evolution and Optimization '89* (H. M. Voigt, H. Mhlenbein & H. P. Schwefel, eds). Akademie-Verlag, Berlin, pp. 86–94.

Goswami, M., O'Connor, K. M. & Shamseldin, A. 2005 Assessing the performance of eight real-time updating models and procedures for the Brosna River. *Hydrol. Earth Syst. Sci.* **9**, 394–411.

Hsu, C. W., Chang, C. C. & Lin, C. J. 2003 A Practical Guide to Support Vector Classification. Available at: www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Ju, Q., Yu, Z. B., Hao, Z. C., Ou, G. X., Zhao, J. & Liu, D. D. 2009 Division-based rainfall-runoff simulations with BP neural networks and Xinanjiang model. *Neurocomputing* **72**, 2873–2883.

Kashif Gill, M., Kemblowski, M. W. & McKee, M. 2007 Soil moisture data assimilation using support vector machines and ensemble Kalman filter. *Am. Water Resour. Ass.* **43**, 1004–1015.

Lin, J. Y., Cheng, C. T. & Chau, K. W. 2006 Using support vector machines for long term discharge prediction. *Hydrolog. Sci. J.* **51**, 599–612.

Liong, S. & Sivapragasam, C. 2002 Flood stage forecasting with support vector machines. *J. Am. Water Resour. Res.* **38**, 173–186.

Lü, H., Li, X., Yu, Z., Horton, R., Zhu, Y., Hao, Z. & Xiang, L. 2010a Using a H-∞ filter assimilation procedure to estimate root zone soil water content. *Hydrol. Process.* **24**, 3648–3660.

Lü, H., Yu, Z., Zhu, Y., Drake, S., Hao, Z. & Sudicky, E. A. 2010b Dual state-parameter estimation of root zone soil moisture by optimal parameter estimation and extended Kalman filter data assimilation. *Adv. Water Resour.* **34**, 395–406.

Lü, H., Hou, T., Horton, R., Zhu, Y., Chen, X., Jia, Y., Wang, W. & Fu, X. 2013 The streamflow estimation using the Xinanjiang rainfall runoff model and dual 3 state-parameter estimation method. *J. Hydrol.* **480**, 102–114.

Luchetta, A. & Manetti, S. 2003 A real time hydrological forecasting system using a fuzzy clustering approach. *Comput. Geosci.* **29**, 1111–1117.

Mattera, D. & Haykin, S. 1999 Support vector machines for dynamic reconstruction of a chaotic system. In: *Advances in Kernel Methods – Support Vector Learning* (B. Scholkopf, C. J. B. Burges & A. J. Smola, eds). MIT Press, Cambridge, MA, pp. 211–242.

Moradkhani, H. & Hsu, K.-L. 2005 Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water Resour. Res.* **41**, W05012.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models Part I – A discussion of principle. *J. Hydrol.* **10**, 282–290.

Nayak, P. C., Sudheer, K. P. & Ramasastri, K. S. 2005 Fuzzy computing based rainfall-runoff model for real time flood forecasting. *Hydrol. Process.* **19**, 955–968.

Rajurkara, M., Kothyarib, U. & Chaubec, U. 2004 Modeling of the daily rainfall-runoff relationship with artificial neural network. *J. Hydrol.* **285**, 96–113.

Reichle, R., McLaughlin, D. & Entekhabi, D. 2002 Hydrologic data assimilation with the ensemble Kalman filter. *Month. Weather Rev.* **130**, 103–114.

Sivapragasam, C. & Liong, S.-Y. 2004 Identifying optimal training data set – a new approach. In: *Proceedings of the Sixth International Conference on Hydroinformatics, Singapore, 21–24 June 2004* (S. Y. Liong, K. K. Phoon & V. Babovic, eds). World Scientific Publishing Co., Singapore.

Sivapragasam, C. & Liong, S. 2005 Flow categorization model for improving forecasting. *Nordic Hydrol.* **36**, 37–48.

Sivapragasam, C., Liong, S. Y. & Pasha, M. 2001 Rainfall and runoff forecasting with SSA-SVM approach. *J. Hydroinform.* **3**, 141–152.

Taormina, R., Chau, K. & Sethi, R. 2012 Artificial Neural Network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. *Engrg. Applic. Artif. Intell.* **25**, 1670–1676.

Vapnik, V. N. 1998 *Statistical Learning Theory*. Wiley, New York.

Vapnik, V. & Chervonenkis, A. 1974 *Theory of Pattern Recognition*. Nauka, Moscow (in Russian).

Weerts, A. H. & Serafy, Y. H. 2006 Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resour. Res.* **42**, W09403.

Wu, C. L., Chau, K. W. & Li, Y. S. 2008 River stage prediction based on a distributed support vector regression. *J. Hydrol.* **358**, 96–111.

Xu, D., Wang, W., Chau, K., Cheng, C. & Chen, S. 2013 Comparison of three global optimization algorithms for calibration of the Xinanjiang model parameters. *J. Hydroinform.* **15**, 174–193.

Yu, X., Liong, S. Y. & Babovic, V. 2004 EC-SVM approach for real time hydrologic forecasting. *J. Hydroinform.* **6**, 209–223.

Yu, P. S., Chen, S. T. & Chang, I. F. 2006 Support vector regression for real-time flood stage forecasting. *J. Hydrol.* **328**, 704–716.

Zhao, R. 1992 Xinanjiang model applied in China. *J. Hydrol.* **135**, 371–381.

First received 12 June 2013; accepted in revised form 18 September 2013. Available online 26 November 2013