View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PolyU Institutional Repository

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Rainfall-Runoff Modelling Using Artificial Neural Network Coupled with Singular Spectrum Analysis

C. L. Wu[1, 2] and K. W. Chau[1,]*

[1] Dept. of Civil and Structural Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, People's Republic of China

[2] Changjiang Institute of Survey, Planning, Design and Research, Changjiang Water Resources Commission, 430010, Wuhan, HuBei, People's Republic of China

*Email: cekwchau@polyu.edu.hk

ABSTRACT

Accurately modelling rainfall-runoff (R-R) transform remains a challenging task despite that a wide range of modeling techniques, either knowledge-driven or data-driven, have been developed in the past several decades. Amongst data-driven models, artificial neural network (ANN)-based R-R models have received great attentions in hydrology community owing to their capability to reproduce the highly nonlinear nature of the relationship between hydrological variables. However, a lagged prediction effect often appears in the ANN modeling process. This paper attempts to eliminate the lag effect from two aspects: modular artificial neural network (MANN) and data preprocessing by singular spectrum analysis (SSA). Two watersheds from China are explored with daily collected data. Results show that MANN does not exhibit significant advantages over ANN. However, it is demonstrated that SSA can considerably improve the performance of prediction model and eliminate the lag effect. Moreover, ANN or MANN with antecedent runoff only as model input is also developed and compared with the ANN (or MANN) R-R model. At all three prediction horizons, the latter outperforms the former regardless of being coupled with/without SSA. It is recommended from the present study that the ANN R-R model coupled with SSA is more promising.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

KEYWORDS

Prediction; rainfall and runoff; artificial neural network; modular model; singular spectrum analysis

1. Introduction

The rainfall-runoff relationship is one of the most complex hydrological phenomena to comprehend, owing to the tremendous spatial and temporal variability of watershed characteristics and precipitation patterns, and to the number of variables involved in the modeling of the physical process (Kumar et al., 2005). Since the rational method for peak of discharge was developed by Mulvany (1850), numerous hydrologic models have been proposed. Based on the description of the governing processes, these models can be classified as either physically-based (knowledge-driven) or system theoretic (data-driven). Physically-based models involve a detailed interaction of various physical processes controlling the hydrologic behavior of a system. However, system theoretic models are instead based primarily on observations (measured data) and seek to characterize the system response from those data using transfer functions. As an example of system theoretic models, ANN-based R-R models have received great attentions in the last two decades due to their capability to reproduce the highly nonlinear nature of the relationship between hydrological variables.

The potential of ANN in hydrological modeling was reviewed, for example, by the ASCE Task Committee on Application of the ANNs in hydrology (ASCE, 2000), Maier and Dandy (2000), and Dawson and Wilby (2001). Most applications for river flow prediction consist in modeling the R-R transformation, providing input of past flows and precipitation observations. They have proved that ANNs are able to outperform traditional statistical R-R modeling (Hsu et al, 1995; Shamseldin, 1997; Sajikumar and

Thandaveswara, 1999; Tokar and Johnson, 1999; Coulibaly et al., 2000; Sudheer et al., 2002) and to offer promising alternatives for conceptual R-R models (Hsu et al, 1995; Tokar and Johnson, 1999; Coulibaly et al., 2000; Dibike and Solomatine, 2001; Birikundavyi et al., 2002; de Vos and Rientjes, 2005; Toth and Brath, 2007). Hsu *et al*. (1995) showed that the ANN model provided a better representation of the rainfall-runoff relationships than the ARMAX time series model or the conceptual SAC-SMA (Sacramento soil moisture accounting) models. Coulibaly et al. (2000) used the early stopping method, to train multi-layer perceptrons (MLP) for real-time reservoir inflow prediction. Results show that MLP can provide better model performance compared to benchmarks from the classic autoregressive model coupled with a Kalman filter (ARMAX-KF) and a conceptual model (PREVIS). Birikundavyi et al. (2002) investigated the ANN models for daily streamflow prediction and also showed that ANNs outperformed PREVIS and ARMAX-KF. Toth and Brath (2007) investigated the impact of the amount of the training data on model performance using ANN and a conceptual model (ADM). ANN was proved to be an excellent tool for the R-R simulation of continuous periods, provided that an extensive set of hydro-meteorological data was available for calibration purposes. However, compared with ANN, ADM may allow a significant prediction improvement when focusing on the prediction of flood events and especially in case of a limited availability of the training data.

Improvement of model performance is a long-term topic of interest by researchers when ANN is used to simulate the R-R relationship. It is recognized that the ANN model is data dependent and has a flexible structure, which leaves huge room for the improvement of ANN in the context of R-R prediction. The ANN model is highly sensitive to the studied data, which means that the structure of ANN is totally different with the change of the training data. Besides, the training algorithms, model configuration, and data preprocessing techniques also impose wide influences on the model performance. Hsu et al. (1995) found that the ANN models underestimated low flows and overestimated medium flows when they were used to simulate the R-R relationship. They further mentioned that this might have been due to the models not being able to capture the nonlinearity in the rainfall-runoff process and suggested that there is still room for improvement in applying different algorithms, such as stochastic global optimization and genetic algorithms, to reach near global solutions, and achieve better model performances. Hence, a more effective and efficient ANN R-R model was developed by Jain and Srinivasulu (2004) where ANN was trained by using real-coded GAs. Results showed that the proposed approach could significantly improve the estimation accuracy of the low-magnitude flows.

On the other hand, Zhang and Govindaraju (2000) recently pointed out that the rainfall-runoff mapping in a watershed can be fragmented or discontinuous with significant variations over the input space because of the functional relationships between rainfall and runoff being quite different for low, medium, and high magnitudes of streamflow. They found a single ANN to be rigid in nature and not suitable in capturing a fragmented input-output mapping. In order to overcome this problem they designed a modular neural network (MANN) consisting of three different ANN models for low-, medium-, and high-magnitude flows. Inspired by this study, many modular (or hybrid) models have been developed for R-R simulation. Solomatine and Xue (2004) applied an approach where separate ANN and M5 model-tree basin models were built for various hydrological regimes (identified on the basis of hydrological domain knowledge). Jain and Srinivasulu (2006) also applied decomposition of the flow hydrograph by a certain threshold value and then built separate ANNs for low and high flow regimes. Corzo and Solomatine (2007) investigated three modular ANNs for simulating two decomposed flow regimes, base flow and exceeding flow, depending on three different partitioning schemes: automatic classification based on clustering, temporal segmentation of the hydrograph based on an adapted baseflow separation technique, and an optimized baseflow separation filter. The modular models were shown to be more accurate than the global ANN model. The best modular model was the one using the optimized baseflow filtering equation. Evidently, all studies demonstrated that modular models generally made higher accuracy of prediction than global models built to represent all possible regimes of the modeled system. Hence, MANN continues to be examined in the present study.

When a rainfall or runoff (streamflow or discharge) time series is viewed as a combination of quasi-periodic signals contaminated by noises to some extent, a cleaner time series can be filtered by appropriate data preprocessing techniques such as singular spectrum analysis (SSA). Obviously, the predictability of a system can be improved by predicting the important oscillations (periodic components) taken from the system. For the purpose of cleaning rainfall or runoff series, many data preprocessing techniques, including Moving average (MA), Principal component analysis (PCA), wavelet analysis (WA), and singular spectrum analysis (SSA), have been employed in hydrology field by researchers (Sivapragasam et al., 2001; Marques et al., 2006; Hu et al., 2007; Partal and Kişi, 2007; Sivapragasam et al., 2007; Wu et al., 2010). Hu et al. (2007) employed PCA as an input data preprocessing tool to improve the prediction accuracy of the rainfall-runoff neural network models. The use of WA to improve rainfall forecasting was conducted by Partal and Kişi (2007). Their results indicated that WA was promising. Wu et al. (2010) compared MA, PCA and SSA as data preprocessing methods using ANN for rainfall predictions and found that SSA is preferred. SSA has also been recognized as an efficient preprocessing algorithm to avoid the effect of discontinuous or intermittent signals, coupled with neural networks (or similar approaches) for time series forecasting

(Lisi et al., 1995; Sivapragasam et al., 2001; Baratta et al., 2003). For example, Lisi et al. (1995) applied SSA to extract the significant components in their study on southern oscillation index time series and used ANN for prediction. They reconstructed the original series by summing up the first ‘‘p” significant components. Sivapragasam et al. (2001) proposed a hybrid model of support vector machine (SVM) and SSA for rainfall and runoff predictions. The hybrid model resulted in a considerable improvement in the model performance in comparison with the original SVM model. However, few studies employ SSA to filter rainfall and streamflow so as to generate cleaner inputs for an R-R model. Therefore, one of main purposes in this study is to develop an ANN (or MANN) R-R model coupled with SSA. To evaluate its performance, a linear regression (LR) R-R model and an ANN-based time series model (using antecedent runoff as only input variables) are developed as benchmarks. To ensure wider applications of conclusions, two river basins from China, Wuxi and Luishui, are explored.

This paper is structured in the following manner. Followed by Introduction, the study areas are described and modeling methods are presented. Section 3 presents their applications to two watersheds. The optimal model is identified and the implementation of SSA is described. In Section 4, main results are shown along with necessary discussions. Section 5 summarizes main conclusions in this study.

2. Methodology

2.1 Study Area and Data

Two river basins from China, Daning and Lushui, are considered as case studies.

The Daning River, a first-order tributary of the Yangtze River, is located in the northeast of Chongqing city. The collected daily data includes rainfall, runoff (or streamflow), and evaporation. The data period spans 20 years from January 1, 1988 to December 31, 2007. The daily rainfall data are measured at six rain gauges located at the upstream of the basin. The upstream part is controlled by "Wuxi" hydrology station, with a drainage area of around 2 000 km^2. The data of runoff and evaporation are gathered at "Wuxi" station (hereafter the studied area is denoted by "Wuxi"). The Lushui River, located in the southeast of Hubei province, is also a first-order tributary of the Yangtze River. The collected daily data includes runoff and rainfall. The data period covers a 4-year long duration (January 1, 2004 - December 31, 2007). The runoff data from Lushui River are collected at "Chongyang" hydrology station. The daily rainfall data are measured at eight rain gauges located at the drainage area controlled by Chongyang hydrology station. The drainage area controlled by the station is around 1 700 km2 (hereafter the studied area is referred to as "Chongyang"). Figure 1 demonstrates rainfall and runoff (or streamflow) time series in two basins. The data represents various types of hydrological conditions, and flow range from low to very high.

Each prediction model is a lumped type, namely, the watershed is considered as a whole, the input rainfall being the mean areal precipitation over the watershed by Thiessen polygon method and the output being the runoff measured at the control hydrology station. The entire input-output dataset in each watershed is partitioned into three data subsets as training set, cross-validation set and testing set: the first half of the entire data as training set and the first half of the remaining data as cross-validation set and the other half as testing set. The training set serves the model training and the testing set is used to evaluate the performances of models. The cross-validation set has dual functions: one is to implement an early stopping approach so as to avoid overfitting of the training data, and another is to select some best predictions from a large number of ANN's runs. Ten best predictions are selected from twenty ANN's runs in the present study. Moreover, ANN employs the hyperbolic tangent function as transfer functions in both hidden and output layers. Table 1 presents statistical information on rainfall and streamflow data, including mean (μ), standard deviation (S_x), coefficient of variation (C_{v}), skewness coefficient (C_{s}), minimum ($\mathrm{X}_{\min}$), and maximum ($\mathrm{X}_{\max}$). Obviously, the training data cannot fully include the cross-validation and testing data in terms of Wuxi. It's recommended that all data be scaled to the interval [-0.9, 0.9] instead of [-1, 1] which is the range of the hyperbolic tangent function. The advantage of using [-0.9, 0.9] is that some extreme data occurring outside the range of the training data may be accommodated in the mapping of ANN.

2.2 Singular spectrum analysis

According to Golyandina et al. (2001), the basic SSA consists of two stages: decomposition and reconstruction. The decomposition stage involves two steps: embedding and singular values decomposition (SVD); the reconstruction stage also comprises two steps: grouping and diagonal averaging. Consider a real-valued time series $F = \{x_1, x_2, \cdots, x_N\}$ of length $N(>2)$. Assume that the series is a nonzero series, viz. there exists at least one i such that $x_i \neq 0$. Four steps are briefly presented as follows.

1st step: embedding

The embedding procedure maps the original time series to a sequence of multi-dimensional lagged vectors. Let L be an integer (window length), $1 < L < N$, and τ be the delayed time as the multiple of the sampling period. The embedding procedure forms $n = N - (L-1)\tau$ lagged vectors $\mathbf{x}_i = \{x_i, x_{i+\tau}, x_{i+2\tau}, \cdots, x_{i+(L-1)\tau}\}^T$, where $\mathbf{x}_i \in \mathrm{R}^L$, and $i = 1, 2, \cdots, n$. The 'trajectory matrix' of the time series is denoted by $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_i \ \cdots \ \mathbf{x}_n]$ having lagged vectors as its columns. In other words, the trajectory matrix is

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

$$\mathbf{X}=\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ x_{1+\tau} & x_{2+\tau} & x_{3+\tau} & \cdots & x_{n+\tau} \\ x_{1+2\tau} & x_{2+2\tau} & x_{3+2\tau} & \cdots & x_{n+2\tau} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1+(L-1)\tau} & x_{2+(L-1)\tau} & x_{3+(L-1)\tau} & \cdots & x_N \end{pmatrix} \tag{1}$$

If $\tau=1$, the matrix $\mathbf{X}$ is called Hankel matrix since it has equal elements on the 'diagonals' where the sum of subscripts of row and column is equal to a constant. If $\tau>1$, the equal elements in $\mathbf{X}$ are not definitely in the 'diagonals'.

2nd step: SVD

Let $\mathbf{S}=\mathbf{X}\mathbf{X}^T$. Denoted by $\lambda_1, \lambda_2, \cdots, \lambda_L$ the eigenvalues of $\mathbf{S}$ taken in the decreasing order of magnitude ($\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_L \geq 0$) and by $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_L$ the orthonormal system of the eigenvectors of the matrix $\mathbf{S}$ corresponding to these eigenvalues. If we denote $\mathbf{v}_i=\mathbf{X}_i^T\mathbf{u}_i \big/ \sqrt{\lambda_i}$ ($i=1,\cdots,L$) (equivalent to the ith eigenvector of $\mathbf{X}^T\mathbf{X}$), then the SVD of the trajectory matrix $\mathbf{X}$ can be written as

$$\mathbf{X}=\mathbf{X}_1+\cdots+\mathbf{X}_L \tag{2}$$

where $\mathbf{X}_i=\sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^T$. The matrices $\mathbf{X}_i$ have rank 1; therefore they are elementary matrices. The collection (λ_i, $\mathbf{u}_i$, $\mathbf{v}_i$) will be called ith eigentriple of the SVD. Note that $\mathbf{u}_i$ and $\mathbf{v}_i$ are also ith left and right singular vectors of $\mathbf{X}$, respectively.

3rd step: grouping

The purpose of this step is to appropriately identify the trend component, oscillatory components with different periods, and structureless noises by grouping components. This step can be also skipped if one does not want to precisely extract hidden information by regrouping and filter of components.

The grouping procedure partitions the set of indices $\{1,\cdots,L\}$ into m disjoint subsets $I_1,\cdots,I_m$, so the elementary matrix in Eq. (2) is regrouped into m groups. Let

$I = \{i_1, \cdots, i_p\}$. Then the resultant matrix $\mathbf{X}_I$ corresponding to the group I is defined as $\mathbf{X}_I = \mathbf{X}_{i_1} + \cdots + \mathbf{X}_{i_p}$. These matrices are computed for $I_1, \cdots, I_m$ and substituting into Eq. (2) one obtains the new expansion

$$\mathbf{X} = \mathbf{X}_{I_1} + \cdots + \mathbf{X}_{\mathrm{I_m}} \tag{3}$$

The procedure of choosing the sets $I_1, \cdots, I_m$ is called the eigentriple grouping.

4th step: Diagonal averaging

The last step in the Basis SSA transforms each resultant matrix of the grouped decomposition (3) into a new series of length N. The diagonal averaging is to find equal elements in the resultant matrix and then to generate a new element by averaging over them. The new element has the same position (or index) as that of these equal elements in the original series. As mentioned in the step 1, the concept of 'diagonal' is not true for $\tau > 1$. Regardless of the value of τ larger than or equal 1, the principle of reconstruction is the same. For $\tau = 1$, the diagonal averaging can be carried out by formula recommended by Golyandina et al. (2001). Let $\mathbf{Y}$ be a ($L \times n$) matrix with elements y_{ij}, $1 \le i \le L$, $1 \le j \le n$. Make $L^* = \min(L, n)$, $n^* = \max(L, n)$ and $N = n + (L-1)\tau$. Let $y^*_{ij} = y_{ij}$ if $L < n$ and $y^*_{ij} = y_{ji}$ otherwise. Diagonal averaging transfers matrix $\mathbf{Y}$ to a series $\{y_1, y_2, \cdots, y_N\}$ by the following equation:

$$y_k = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} y^*_{m,k-m+1} & for\, 1 \le k < L^* \\ \dfrac{1}{L^*}\displaystyle\sum_{m=1}^{L^*} y^*_{m,k-m+1} & for\, L^* \le k \le K^* \\ \dfrac{1}{N-k+1}\displaystyle\sum_{m=k-K^*+1}^{N-K^*+1} y^*_{m,k-m+1} & for\, L^* < k \le N \end{cases} \tag{4}$$

Eq. (4) corresponds to averaging of the matrix elements over the 'diagonals' $i+j=k+1$. The diagonal averaging, applied to a resultant matrix $\mathbf{X}_{I_k}$, produces a $N-$length series F_k, and thus the original series F is decomposed into the sum of m series:

$$F = F_1 + \cdots + F_m \tag{5}$$

As mentioned above, these reconstructed components (RCs) can be associated with the trend, oscillations or noise of the original time series with proper choices of L and the sets of $I_1, \cdots, I_m$. Certainly, if the third step (namely, grouping) is skipped, F can be decomposed into L RCs.

2.3 Model development

A representative data-driven R-R model can be defined as

$$\hat{Q}_{t+T} = f(\mathbf{X}_t) = f(Q_{t+1-l_1}, R_{t+1-l_2}, S_{t+1-l_3}) \tag{6}$$

where $\hat{Q}_{t+T}$ stands for the predicted flow at time instance $t+T$; T (with $T=1,2,3$ for the present study) refers to how far into the future the runoff prediction is desired; Q_{t+1-l_1} is the antecedent flow (up to $t+1-l_1$ time steps), R_{t+1-l_2} is the antecedent rainfall (up to $t+1-l_2$ time steps) and S_{t+1-l_3} (up to $t+1-l_3$ time steps) represents any other factors contributing to the true flow Q_{t+T}, such as evaporation or temperature; l_1 , l_2 , and l_3 respectively stand for the number of previous flow, rainfall and other factors. The predictability of future behavior is a consequence of the correct identification of the system transfer function of $f(\bullet)$. Herein, linear regression and nonlinear regression (e.g. ANN) techniques are respectively used to approximate the $f(\bullet)$.

(1) LR

The LR model herein is actually called stepwise linear regression (SLR) model because the forward stepwise regression is used to determine optimal input variables. The basic idea of SLR is to start with a function that contains the single best input variable and to subsequently add potential input variables to the function one at a time in an attempt to improve model performance. The order of addition is determined by using the partial F-test values to select which variable should enter next. The high partial F-value is compared to a (select or default) F-to-enter value. After a variable has been added, the function is examined to see if any variable should be deleted. More details can be found in Draper and Smith (1998) and McCuen (2005).

(2) ANN

The multilayer perceptron network is by far, among ANN paradigms, the most popular, which usually uses the technique of error back propagation to train the network configuration. The architecture of the ANN consists of a number of hidden layers and a number of neurons in the input layer, hidden layers and output layer. ANNs with one hidden layer are commonly used in hydrologic modeling (Dawson and Wilby, 2001; de Vos and Rientjes, 2005) since these networks are considered to provide enough complexity to accurately simulate the nonlinear-properties of the hydrologic process. The three-layer ANN can be denoted by $m \times h \times 1$ where m stands for number of neuron in the input layer and h is the number of neuron in the hidden layer. According to Eq. (6), $m = l_1 + l_2 + l_3$. The ANN prediction model is formulated as

$$\hat{Q}_{t+T} = f(\mathbf{X}_t, w, \theta, m, h) = \theta_0 + \sum_{j=1}^{h} w_j^{out} \varphi(\sum_{i=1}^{m} w_{ji} \mathbf{X}_t + \theta_j) \tag{7}$$

where φ denotes transfer functions; w_{ji} are the weights defining the link between the ith node of the input layer and the jth of the hidden layer; θ_j are biases associated to the jth node of the hidden layer; w_j^{out} are the weights associated to the connection between the jth node of the hidden layer and the node of the output layer; and θ_0 is the bias at the output node. To apply Eq. (7) to runoff predictions, appropriate training algorithm is required to optimize w and θ.

(3) MANN

To construct MANN, the training data have to be divided into several clusters according to cluster analysis techniques, and then each single model is applied to each cluster. The fuzzy c-means (FCM) clustering technique is adopted in the present study (e.g., Bezdek, 1981, Wang et al., 2006). It is able to generate either soft or crisp clusters. Predictions from a modular model can be conducted in two ways: soft and hard. Soft prediction means that the testing data can belong to each cluster with different weights. As a consequence, the modular model output would be a weighted average of the outputs of several single models fitted for each cluster of training data. Hard prediction is that the modular model output is directly from the output of only triggered local model. ANN (or similar techniques) is unable to extrapolate beyond the range of the data used for training. Otherwise, poor predictions or predictions can be expected when a new input data is outside the range of those used for training. Hard prediction method is, therefore, adopted in this study.

Figure 2 displays the schematic diagram of MANN where the training data is partitioned into three clusters. Once input-output pairs are obtained, they are first split into three subsets by the FCM technique, and then each subset is approximated by a single ANN. The final output of the modular model results directly from the output of one of three local models.

2.4 Implementation framework of R-R prediction

Figure 3 illustrates the implementation framework of rainfall-runoff prediction where four prediction models can be conducted in two modes: without/with three data preprocessing methods (dashed box). These acronyms in the column of “methods for model inputs" represent five methods to determine model inputs: LCA (linear correlation analysis, Sudheer et al., 2002), AMI (average mutual information, Fraser and Swinney, 1986), PMI (partial mutual information, May et al., 2008), SLR (stepwise linear regression), and MOGA (ANN based on multi-objective genetic algorithm, Giustolisi and Simeone, 2006).

2.5 Evaluation of model performances

The Pearson's correlation coefficient (r) or the coefficient of determination ($R^2 = r^2$), have been identified as inappropriate measures in hydrologic model evaluation by Legates and McCabe (1999). The coefficient of efficiency (CE) (Nash and Sutcliffe, 1970) is a good alternative to r or R^2 as a "goodness-of-fit" or relative error measure in that it is sensitive to differences in the observed and predicted means and variances. Legates and McCabe (1999) also suggested that a complete assessment of model performance should include at least one absolute error measure (e.g., RMSE) as necessary supplement to a relative error measure. Besides, the Persistence Index (PI) (Kitanidis And Bras, 1980) was adopted here for the purpose of checking the prediction lag effect. Three measures are therefore used in this study. They are listed below.

$$\mathrm{CE} = 1 - \sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2 / \sum_{i=1}^{n}(Q_i - \bar{Q})^2 \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(Q_i - \hat{Q}_i)^2} \tag{9}$$

$$\mathrm{PI} = 1 - \sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2 / \sum_{i=1}^{n}(Q_i - Q_{i-l})^2 \tag{10}$$

In these equations, n is the number of observations, $\hat{Q}_i$ stands for predicted flow, Q_i represents observed flow, $\bar{Q}$ denotes average observed flow, and Q_{i-l} is the flow estimate from a persistence model (or termed naïve model) that basically takes the last flow observation (at time i minus the lead time l) as the prediction. CE and PI values of 1 stands for perfect fits. A small value of PI may imply the occurrence of the lag prediction.

3. Applications of Models

3.1 Potential input variables

In the process of determining model inputs, the first step is to find out appropriate input variables (causal variables) for Eq. (6). In general, causal variables in the R-R relationship can be rainfall (precipitation), previous flows, evaporation, temperature etc. Depending on the availability of data, the input variables tend to be varied in previous studies. Most studies employed rainfall and previous flow (or water level) as inputs (Campolo et al., 1999; Liong et al., 2002; Xu and Li, 2002; Sivapragasam et al., 2007) whereas input variables in some studies also included additional factors such as temperature or evaporation (Abrahart et al., 1999; Tokar and Johnson, 1999; Zealand et al, 1999; Zhang and Govindaraju, 2000; Coulibaly et al., 2001; Abebe and Price, 2003; Solomatine and Dulal, 2003;Wilby et al., 2003; Hu et al., 2007; Toth and Brath, 2007; Solomatine and Shrestha, 2009).

The necessity of previous flows in model inputs was widely recognized by researchers (Campolo et al., 1999; de Vos and Rientjes, 2005). Campolo et al. (1999) made use of distributed rainfall data observed at different raingauge stations for the prediction of water levels at the catchment outlet. Poor predicted results were achieved when only water levels were used as input. However, the accuracies of predictions were improved when rainfall and previous water levels were included in inputs. de Vos and Rientjes (2005) employed different model inputs as hydrological state representation of ANN. Results also showed that the ANN model with rainfall input variable only had the worst performance compared to those whose input variables consisting of rainfall, flow and/or other states.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

However, some studies pointed out that evaporation (or temperature) as input variable seemed to be unnecessary (Abrahart et al., 2001; Anctil et al., 2004; Toth and Brath, 2007). Anctil et al. (2004) found that potential evapotranspiration failed to improve the MLP performance when it was introduced into the initial model inputs consisting of rainfall and streamflow for R-R modeling. Results from Toth and Brath (2007) also indicated that the inclusion of potential evapotranspiration values in inputs did not improve the prediction results, but gave rise to a slight deterioration in comparison with the use of precipitation data alone. That result may be explained by the fact that the addition of evapotranspiration (or temperature measures) input nodes increases the network complexity, and therefore the risk of overfitting. In the present experiments, analyses of LCA, AMI, and PMI between evaporation and streamflow indicate that evaporation can be excluded since the dependence relation is not significant. Therefore, rainfall and streamflow are identified as final input variables.

3.2 Selection of model inputs

Having chosen appropriate input variables, the next step is the determination of appropriate lags for each variable to form model inputs. ANN, equipped with Levernberg-Marquardt training algorithm and hyperbolic tangent transfer functions, is used as the benchmark model to examine five input methods.

Figure 4 demonstrates the results of LCA of the runoff series for Wuxi and Chongyang. The partial auto-correlation function (PACF) value decayed within the confidence band around at lag 5 for Wuxi and lag 4 for Chongyang. Therefore, the number l_1 of lags of flow was initially set at the value of 5 for Wuxin and 4 for Chongyang. The number l_2 of lags of rainfall is generally determined according to time of concentration of the watershed. The time of concentration used herein is estimated between the center of hyetograph and the peak flow. The average time of concentration is approximately 1 day for Wuxi and Chongyang. To take account of delay between rainfall and runoff, the value of l_2 is originally set to 5 for both Wuxi and Chongyang. Table 2 presents the results of ANN with different model inputs determined by LCA, AMI, PMI, SLR and MOGA. These results are based on one-step-ahead flow prediction (i.e. $\hat{Q}_{t+1}$ where t represents the present time instance). In terms of RMSE, there is no salient difference among all five methods. However, our experiments reveal that the ANN with inputs from LCA outperforms the others in the SSA scenario. Moreover, LCA can significantly reduce the effort and computational time requirement in developing an ANN model. The LCA method is therefore adopted for the later analysis. Figure 5 illustrates cross correlation functions (CCFs) between rainfall and streamflow for Wuxi and Chongyang. The past five rainfall observations have significant relations (correlation coefficient > 0.2) with the present streamflow. The most significant correlation occurs at the first lag which indicates the time of response of watershed being about 1 day.

3.3 Identification of models

The model identification of a prediction model is to determine the structure by using training data to optimize relevant model parameters once model inputs are already obtained.

(1) LR

LR can be viewed as a model-driven model which has known model structure. Model identification only consists in optimizing the coefficient of each input. The stepwise linear regression (SLR) technique was used to concurrently determine the model inputs and the corresponding coefficients. With model inputs already obtained by SLR in Table 2, the LR model at one-step lead for Wuxi and Chongyang can expressed respectively as Eq. (11),

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

$$\hat{Q}_{t+1} = -0.019Q_{t-4} + 0.025Q_{t-2} + 0.016Q_{t-1} + 0.469Q_t + 0.046R_{t-4} + 0.07R_{t-3} + 0.027R_{t-2} + 0.121R_{t-1} + 0.272R_t \tag{11}$$

and Eq. (12),

$$\hat{Q}_{t+1} = 0.032Q_{t-3} + 0.526Q_t + 0.099R_{t-3} + 0.053R_{t-2} + 0.037R_{t-1} + 0.454R_t \tag{12}$$

(2) ANN and MANN

As a three-layer MLP was adopted, the identification of ANN's structure is to optimize the number of hidden nodes h in the hidden layer when the model inputs have been determined by LCA and there is a unique model output. The optimal size h of the hidden layer is found by systematically increasing the number of hidden neurons from 1 to 10 until the network performance on the cross-validation set no longer improves significantly. The identified configurations of ANN were 10-8-1 for Wuxi and 9-9-1 for Chongyang, respectively (presented in Table 2). The same method is used to identify three local ANNs in MANN. As a consequence, the structures of MANN are 10-4/4/2-1 for Wuxi and 9-3/3/1-1 for Chongyang, respectively.

In order to perform multi-step-ahead predictions, two methods are available: (1) re-using a one-step-ahead prediction as input into the network, after which it predicts the two-step-ahead prediction, and so forth, and (2) by directly having the multi-step-ahead prediction as output. The former and the latter are respectively termed the dynamic model and static model. For simplification, the static model is adopted herein.

3.4 Decomposition of rainfall and runoff series by SSA

To filter raw rainfall and runoff series, each series needs to be decomposed into components with the aid of SSA. The decomposition by SSA requires identifying the parameter pair (τ, L). The choice of L represents a compromise between information content and statistical confidence (Elsner and Tsonis, 1996). The value of an appropriate L should be able to clearly resolve different oscillations hidden in the original signal.

However, the present study does not require accurately resolving the raw rainfall signal into trends, oscillations, and noises. A rough resolution can be adequate for the separation of signals and noises where some leading eigenvalues should be identified. To select L, a small interval of [3, 10] was examined in the present study.

A target L can be empirically determined in accordance with a specified criterion: the singular spectrum under the target L can be distinguished markedly, i.e. singular values forming the singular spectrum are quite different from each other. Figure 6 illustrates the sensitivity analysis of the singular spectrum on L for rainfall and streamflow series from two basins of Wuxi and Chongyang. Singular values of both rainfall and flow series in the Wuxi watershed are clearly separated. Clearly, in terms of the criterion, L can be arbitrarily chosen from 3 to 10. To obtain a more robust ANN model, it is recommended that a larger L be taken which results in more combinations of RCs in the process of seeking the optimal model inputs. Thus, the final L is set at the value of 9 for the Wuxi rainfall, 7 for the Wuxi flow, 7 for both Chongyang rainfall and flow. Figure 6 highlights the singular spectrum curve associated with the selected L in the dotted line.

Figure 7 shows the results of sensitivity analysis of the singular spectrum on the lag time τ using SSA with the chosen L. The singular spectrum can be clearly distinguished at $\tau = 1$. Therefore, the final parameter pair (τ, L) in SSA was set as (1, 9) for the Wuxi rainfall, and (1, 7) for the other three series. Thus, each rainfall or flow series can be decomposed into RCs with these identified parameter pair.

3.5 Combination of models with SSA

Once an input (rainfall or runoff) time series is decomposed into RCs, the subsequent task is to filter RCs by finding contributing RCs from all existing RCs to model output, and then reconstruct a new input series by summing these contributing RCs.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

There is no practical guide on how to identify a contributing or noncontributing component to the improvement of accuracy of prediction. Apparently, a single higher-frequency component may be noncontributing. However, the situation may become complicated with the combination of components and change of the prediction horizon. For example, one component viewed as contribution to one-step-ahead prediction may have a negative impact on two-step-lead prediction. Nevertheless, the combined signal of several high-frequency RCs may yield a better input/output mapping than a low-frequency RC. Therefore, an enumeration method is recommended where all input combinations from rainfall (or runoff) are examined. If the number of RCs is L, there are 2^L combinations. For instance, there are 2^9 combinations for the Wuxi rainfall series in view of $L = 9$. It should also be noticed that the enumeration method may be computationally intensive if L is a large number, say 20 or 30.

Since input variables consist of rainfall and flow, the filtering procedure has to be conducted separately for each variable. Taking ANN with SSA (hereafter referred to as ANN-SSA) as an example, two new ANN models need to be established for the purpose of RCs' filtering, one for rainfall input and the other for runoff input. For the convenience of identification, the ANN model for rainfall input filtering is denoted by ANN-RF, and the ANN model for runoff input filtering is referred to as ANN-QF. ANN-RF has the same model output as that of the original ANN model and its model input is the same as the rainfall part of the original ANN model inputs. Likewise, the ANN-QF model input is from the runoff part of the original ANN model inputs, and both of them have the same model output variable. Depending on trial and error, ANN-RF and ANN-QF can be identified. For example, ANN-RF was 5-3-1 for Wuxi and 5-4-1 for Chongyang, respectively, and ANN-QF was 5-4-1 for Wuxi and 4-1-1 for Chongyang, respectively. Similarly, LR-RF and LR-QF were also developed for the RCs filtering of both rainfall and runoff series in the context of LR. Table 3 presents the RCs filtering results of input variables of rainfall and runoff for both LR-SSA and ANN-SSA (or MANN-SSA). Two basic conclusions can be drawn from Table 3 in the context of SSA: one is that ANN-SSA outperforms LR-SSA with the same model inputs; the model with only runoff input, either LR-SSA or ANN-SSA, performs better than that with only rainfall input. Therefore, inclusion of flow in model inputs proves to be imperative in R-R prediction.

4. Results and Discussions

Results of R-R prediction are respectively presented according to the normal mode and SSA mode. In each mode, three models of LR, ANN, and MANN are compared by three model performance indices. In the SSA mode, three models are referred to as LR-SSA, ANN-SSA, and MANN-SSA.

4.1 Predictions in normal mode

As observed from Table 4, all models except for LR for Chongyang have made one-step-ahead predictions with a high CE over 0.7. This indicates that causal variables of model output have been accurately selected for this prediction horizon. The performance of each model deteriorates abruptly with the increase of prediction horizons, which may indicate the adoption of inappropriate model inputs. Basically, it is intuitive that a poor prediction on the testing set may result from the lack of similar patterns between the training set and testing set. Conversely, an excellent prediction probably means that there are a large number of similar patterns between them. For example, all models perform better using the Wuxi data than using the Chongyang data since the former has a large size training data (ten years) which allows models to be appropriately trained. A conclusion can also be drawn that ANN (or MANN) tends to be superior to LR if the mapping relation is identified appropriately. The superiority of MANN over ANN seems to be dependent on the studied data.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Figure 8 illustrate representative details of hydrographs and whole scatter plots of one-step-ahead prediction using three prediction models for Wuxi and Chongyang, respectively. The scatter plot from the LR model with high spread at low magnitude flows indicates poor predictions of low flows compared with scatter plots from both ANN and MANN. ANN and MANN fairly underestimate or overestimate peak flows, but reproduce low flows appropriately because low flows are more frequent in the data set than large flows.

In order to set up a relative optimal model for runoff prediction, some researchers carried out runoff predictions depending on ANN (or similar techniques) with two different inputs: inputs with antecedent runoffs only; and inputs with both antecedent rainfalls and runoffs. For example, Minns (1998) observed a phase shift error in prediction outputs when antecedent discharge values were the only inputs used to predict present discharge. However, models developed using discharge and rainfall inputs were not observed to exhibit phase shift errors. Sivapragasam et al. (2007) respectively used GP (genetic programming) and ANN to predict river flows from one- up to four-step leads with the two types of inputs. Results indicated that the model with rainfall and flow as inputs, regardless of GP or ANN, made more accurate prediction than that with only flow input. In this study, we will extend this comparison from the normal mode to the SSA mode.

According to the same method to construct ANN or MANN in the context of rainfall-runoff transformation as mentioned procedures in Section 3, identified ANNs with only runoff inputs are 5-3-1 for Wuxi, and 4-8-1 for Chongyang, and identified MANNs with only runoff inputs are 5-10/10/4-1 for Wuxi, and 4-8/8/5-1 for Chongyang. In the SSA mode, parameter pair (τ, L) is also (1, 7) for each of them.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Table 5 presents comparison of runoff predictions using ANN and MANN with two types of inputs: past flow as the only input variable, and previous rainfall and flow as input variables. It can be observed that, for the study case of Wuxi, the inclusion of rainfall in input results in the improvement of model performance irrespective of ANN and MANN. However, the degree of the improvement mitigates with the increase of prediction leads. This may indicate that the influence of rainfall on runoff gradually weakens with the increase of prediction horizons. An opposite result was found by Sivapragasan et al. (2007) in which the influence of rainfall on runoff (the time resolution of the data is fortnightly) gradually increased with increasing prediction lead. Employing the data with an hourly time resolution, Toth and Brath (2007) investigated the performance of ANN in two types of inputs. They found that ANN with the inclusion of rainfall in input outperformed ANN with only flow as input at all prediction leads from 1 hour up to 12 hours. Actually, whether or not rainfall is introduced to input heavily relies on the characteristic of the studied watershed. In general, inclusion of rainfall in input could be helpful in improving accuracy of predictions if the prediction lead is less than the average time of concentration. The time of concentration can be roughly identified by the AMI (or CCF) analysis between available rainfall and flow data, and it approximately equals the maximum AMI (or CCF). As shown in Figure 5, the time of concentration in each basin is around one day. If the time resolution of data is hourly-based, the time of concentration can be approximated to hours but days. Therefore, the inclusion of rainfall in input has led to a noticeable improvement of accuracy of one-day-ahead prediction. In this regard, a more detailed analysis will be addressed in the section of discussions.

The hydrograph of one-step-ahead prediction is presented in Figure 9. The ANN model with only flow input makes the lagged predictions whereas the ANN model with rainfall and flow as inputs eliminates the lag effect. However, with the increase of prediction leads, each of two types of ANN yields a prediction lag effect as shown in Figure 10, which indicates the effect of rainfall on model output being markedly mitigated.

4.2 Predictions in SSA mode

Table 6 presents the results of R-R predictions for Wuxi and Chongyang using three prediction models coupled with SSA. Compared with the results of Table 4, the SSA technique brings about a significant improvement of model performance at all three prediction horizons. Models of ANN and MANN outperform the LR model, but the MANN model exhibits no superiority over the ANN model.

The representative details of hydrograph and whole scatter plots of one-step-ahead prediction for Wuxi and Chongyang are shown in Figure 11. These results show that three models with SSA are able to make good predictions because the predicted hydrograph perfectly reproduces the actual hydrograph and the scatter plots are close to the exact line with rather a low spread. It can be observed from the hydrograph that the LR-SSA model produces some negative predictions for the low flows and ANN-SSA and MANN-SSA occasionally make negative predictions at the low-flow points. The peak values are still overestimated or underestimated although each model with SSA exhibits excellent overall performances.

Table 7 presents comparison of two types of model inputs feeding ANN-SSA and MANN-SSA. ANN-SSA (or MANN-SSA) fed by rainfall and flow performs better than the corresponding model fed by only flow at all prediction leads. It is observed that the advantage of models with rainfall and flow inputs over those with flow input only becomes more obvious with increasing prediction leads, which indicates that SSA improves the dependence relation more significantly between rainfall and flow than that between flows itself. The model output may therefore depend more on rainfall inputs instead of flow itself when the prediction lead is larger than one day.

Figure 12 illustrates one-step-ahead prediction hydrographs for Wuxi and Chongyang using ANN-SSA in two types of inputs. ANN-SSA with rainfall and flow inputs better captures the peak flows, and reproduces the actual hydrograph more smoothly whereas the hydrograph from ANN-SSA with flow input only is serrated at some locations. It is found that there is no time shift between the predicted hydrograph and the actual one. Figure 13 demonstrates the results of lag effect analysis at all three prediction horizons by depicting CCF between observation and prediction. SSA eradicates the prediction lag effect in the ANN model regardless of model input types. Moreover, it can be observed that the CCF curve in ANN-SSA with rainfall and flow inputs is more symmetrical than that in ANN-SSA with only flow input, which reveals that predictions in the former is in better agreement with the observations in time.

4.3 Discussions

The following discussions focus on two aspects: investigating the difference between two types of model inputs for runoff prediction, and investigating the effect of SSA on the R-R ANN model inputs.

a) Analysis of model inputs

As shown in Table 5, ANN with rainfall and flow inputs performs better than that with flow input only at all prediction leads, but the improvement of model performance decreases abruptly at a two-step lead. A direct explanation for that phenomenon is that the impact of rainfall on runoff weakens suddenly at two-step-ahead prediction, which can be examined by AMI and CCF between model inputs and output.

Figure 14(a) presents AMI between each input and output of ANN in two model input scenarios for the Wuxi study case. The number of model inputs in the abscissa axis consists of 5 previous flow data and 4 previous rainfall data. The former 5 inputs stand for 5 past flows and the latter 5 inputs denote 5 past rainfall observations. In contrast, all 10 model inputs (actually 5) in the flow input scenario are the past 10 flow observations. First of all, it is clearly shown from all three sub-plots that AMI associated with each model input decreases significantly with an increase in the prediction lead, which may indicate decrease of the overall dependence relation between model inputs and output. Therefore, it provides a potential explanation for the trend in Table 6 that the accuracy of the prediction decreases with the increase of prediction horizons. Secondly, the nearest rainfall observation (the sixth model input in each plot) to the prediction horizon has the maximum AMI, so inclusion of such input improves the prediction. Some of the other rainfall inputs also have reasonably larger AMI compared to that of flow inputs, and they also contribute to the improvement of model performance.

Figure 14(b) shows AMI of each input and output of ANN with two types of inputs for the Chongyang study case. Regarding ANN in rainfall and flow inputs, the first 4 model inputs in the abscissa axis are from the past flows and the latter 5 inputs represent the 5 last rainfall observations. As far as ANN with flow input only is concerned, the first 4 model inputs in the abscissa axis are the actual inputs. It can be observed that, AMI of each model input and output between two-step-ahead and three-step-ahead predictions is similar and very small regardless of the input scenario. Moreover, the holistic AMI from rainfall inputs does not dominate over the overall AMI from flow inputs. Therefore, inclusion of such rainfall inputs may only make the training process computation intensive without any tangible improvement in prediction accuracy. As a consequence, the model performance of ANN with two types of inputs is similarly poor for both two- and three-step-ahead predictions (depicted as Table 5). On the contrary, for one-step-ahead prediction, the nearest two rainfall inputs have large AMIs which are only smaller than the AMI of the immediate past flow input. As expected, their inclusion in model inputs improves the overall mapping between inputs and output of ANN, making one-step-ahead prediction with good accuracy.

The static multi-step prediction method is adopted in this study. The poor prediction at two- or three-step-ahead horizon using ANN with rainfall and flow as inputs may be improved by adopting a dynamic ANN model instead of the current static ANN model. In the dynamic ANN model, the predicted flow and rainfall in the last step are used as the nearest flow and rainfall inputs in the present prediction step, and then a multi-step prediction becomes a repeated one-step prediction. However, de Vos and Rientjes (2005) mentioned that for both the daily and hourly data the two multi-step prediction methods performed nearly similar up to a lead time of respectively 4 days and 12 hours. Similarly, the results from Yu et al. (2006) for hourly data also showed that two methods could yield similar predictions.

b) Investigation of the SSA effect on model inputs

Herein, the effect of SSA on inputs of an ANN R-R model is investigated by AMI between each input and output of model. Results of prediction from the ANN R-R model with the normal mode (shown in Tables 4 or 5) show that the flows at one-step lead are predicted appropriately whereas poor predictions are obtained at two- or three-step lead. Correspondingly, it can be observed from Figure 15(a) that AMI associated with each model input for one-step prediction is far larger than the counterparts for two- or three-step predictions. Figure 15(b) shows that SSA improved AMI of each input at all three prediction horizons. The AMI curve of filtered inputs between one- and two-step predictions is very similar, which may indicate similar model performance (shown in Tables 6 or 7 where the model performance at the two prediction leads is also quite similar). Therefore, the AMI analysis proves to be able to reveal the suitability of a prediction model to some extent. Figure 15(b) also reveals that AMI at one-step prediction is far larger than that at two- and three-step leads. So the prediction accuracy at the former is markedly superior to that in the latter (shown in Tables 4 or 5). In the SSA mode, AMI of each input is considerably improved at all prediction horizons, which renders the ANN-SSA R-R model good predictions (shown in Tables 6 or 7) in comparison to that in the normal mode.

5. Conclusions

This study has predicted daily rainfall-runoff transformation from two different watersheds, namely Wuxi and Chongyang, through three models (viz. LR, ANN and MANN) in conjunction with SSA. Rainfall and runoff are firstly identified as appropriate input variables, and then model inputs are selected by LCA after comparison with the other four methods of determining model inputs. The model performance seems to be sensitive to the studied case in the normal mode. For Wuxi, the MANN R-R model (namely, rainfall and runoff as inputs) outperforms the ANN R-R model and the ANN R-R model performs better than the LR R-R model at all three prediction horizons. For Chongyang, the ANN R-R model performs the best among three models at one-step lead. However, they are similar at the other two prediction horizons. In the SSA mode, the performance of each model is significantly improved. Both ANN-SSA and MANN-SSA have similar performance and achieve better results than LR-SSA.

The ANN R-R model is also compared with the ANN model with only runoff input. The ANN R-R model outperforms the ANN model with only flow input in both the normal mode and SSA mode. The degree of superiority tends to mitigate with the increase of prediction leads in the normal mode. However, situation becomes reverse in the SSA mode where the advantage of the ANN R-R model seems to be more remarkable with the increase of prediction leads. It is recommended from the present study that the ANN R-R model coupled with SSA is more promising.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

References:

Abebe, A.J., and Price, R.K. (2003), Managing uncertainty in hydrological models using complementary models. Hydrological Sciences Journal-Journal des Sciences Hydrologiques, 48 (5), 679-692.

Abrahart, R.J., See, L.M., and Kneale, P.E. (1999), Using pruning algorithms and genetic algorithms to optimise network architectures and forecasting inputs in a neural network rainfall-runoff model. Journal of Hydroinformatics, 1, 103-114.

Abrahart, R.J., See, L.M., and Kneale, P.E. (2001), Applying saliency analysis to neural network rainfall-runoff modelling. Computers and Geosciences, 27, 921-928.

Anctil, F., Perrin, C., and Andréassian, V. (2004), Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models. Environmental Modeling and Software, 19, 357-368.

ASCE. (2000), Artificial neural networks in hydrology 2: Hydrology applications. Journal of Hydrologic Engineering, 5(2), 124-137.

Baratta et al., Baratta, D., Cicioni, G., Masulli, F. and Studer, L. (2003), Application of an ensemble technique based on singular spectrum analysis to daily rainfall forecasting. Neural Networks, 16, 375-387.

Bezdek, J.C. (1981), Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Birikundavyi, S., Labib, R., Trung, H.T., and Rousselle, J. (2002), Performance of neural networks in daily streamflow forecasting. Journal of Hydrologic Engineering, 7(5), 392-398.

Campolo, M., Andreussi, P., and Soldati, A. (1999), River flood forecasting with a neural network model, Water Resources Research, 35 (4), 1191-1197.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Corzo, G., and Solomatine, D. (2007), Baseflow separation techniques for modular artificial neural network modelling in flow forecasting. Hydrological Sciences–Journal–des Sciences Hydrologiques, 52(3), 491-507.

Coulibaly, P., Anctil, F., and Bobée, B. (2000), Daily reservoir inflow forecasting using artificial neural networks with stopped training approach Journal of Hydrology, 230, 244-257.

Coulibaly, P., Anctil, F., and Bobée, B. (2001), Multivariate reservoir inflow forecasting using temporal neural networks. Journal of Hydrologic Engineering, 6 (5), 367-376.

Dawson, C.W., and Wilby, R.L. (2001), Hydrological Modeling Using Artificial Neural Networks. Progress in Physical Geography, 25(1), 80-108.

de Vos, N.J. and Rientjes, T.H.M. (2005), Constraints of artificial neural networks for rainfall -runoff modeling: trade-offs in hydrological state representation and model evaluation. Hydrology and Earth System Sciences, 9, 111-126.

Dibike, Y. B. and Solomatine, D. P. (2001), River flow forecasting using artificial neural networks, Physics and Chemistry of the Earth (B), 26 (1), 1-7, 2001.

Draper, N. R. and Smith, H. (1998), Applied regression analysis, 3rd ed. New York: Wiley.

Elsner, J., and Tsonis, A. (1996), Singular Spectrum Analysis. A New Tool in Time Series Analysis. New York: Plenum Press.

Fraser, A.M. and Swinney, H.L. (1986), Independent coordinates for strange attractors from mutual information, Physical Review A, 33(2), 1134-1140.

Giustolisi, O., and Savic, D. A. (2006), A symbolic data-driven technique based on evolutionary polynomial regression, Journal of Hydroinformatics, 8(3), 207-222.

Golyandina, N. Nekrutkin,V., and Zhigljavsky, A. (2001), Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Hsu, K.L., Gupta, H.V., and Sorooshian, S. (1995), Artificial neural network modeling of the rainfall–runoff process. Water Resources Research, 31(10), 2517-2530.

Hu, T.S.,Wu, F.Y., and Zhang, X. (2007), Rainfall-runoff modeling using principal component analysis and neural network. Nordic Hydrology, 38(3), 235-248.

Jain, A., and Srinivasulu, S. (2004), Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques, Water Resource Research, 40, W04302.

Jain, A., and Srinivasulu, S. (2006), Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. Journal of Hydrology, 317, 291-306.

Kitanidis, P. K. and Bras, R. L. (1980), Real-time forecasting with a conceptual hydrologic model, 2, applications and results, Water Resources Research, 16 (6), 1034–1044.

Kumar, A.R.S., Sudheer, K.P., Jain, S.K., and Agarwal, P.K. (2005), Rainfall-runoff modelling using artificial neural networks: comparison of network types. Hydrological Processes, 19 (6), 1277-1291.

Legates, D. R., and McCabe, Jr, G. J. (1999), Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation, Water Resources. Research, 35(1), 233- 241.

Liong, S. Y., Gautam, T. R., Khu, S. T., Babovic, V., and Muttil, N. (2002), Genetic programming: A new paradigm in rainfall-runoff modeling. Journal of American Water Resources Association, 38(3), 705-718.

Lisi, F., Nicolis, and Sandri, M. (1995), Combining singular-spectrum analysis and neural networks for time series forecasting. Neural Processing Letters, 2(4), 6-10.

Maier H.R., and Dandy G.C. (2000), Neural networks for the prediction and forecasting

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

of water resources variables: a review of modeling issues and applications. Environmental Modeling and Software, 15, 101-23.

Marques, C.A.F., Ferreira, J., Rocha, A., Castanheira, J., Gonçalves, P., Vaz. N., and Dias, J.M. (2006), Singular spectral analysis and forecasting of hydrological time series. Physics and Chemistry of the Earth, 31, 1172-1179.

May, R.J., Maier, H.R., Dandy, G.C., and Fernando, T.M.K. (2008), Non-linear variable selection for artificial neural networks using partial mutual information. Environmental Modeling & Software, 23, 1312-1328.

McCuen, R. H. (2005), Hydrologic analysis and design (3rd ed.), Upper Saddle River, NJ: Pearson/Prentice Hall.

Minns, A. W. (1998), Artificial Neural Networks as Subsymbolic Process Descriptors. Balkema, Rotterdam, The Netherlands.

Mulvany, T. J. (1850), On the use of self-registering rain and flood gauges. Proc. Inst. Civ. Eng., 4(2), 1–8.

Nash, J. E. and Sutcliffe, J. V. (1970), River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology, 10 (3), 282-290.

Sajikumar, N., and Thandaveswara, B.S. (1999), A non-linear rainfall-runoff model using artificial neural networks. Journal of Hydrology, 216, 32-55.

Shamseldin, A.Y. (1997), Application of a neural network technique to rainfall–runoff modelling. Journal of Hydrology 199, 272-294.

Sivapragasam, C., Liong, S.Y. and Pasha, M.F.K. (2001), Rainfall and runoff forecasting with SSA-SVM approach. Journal of Hydroinformatics, 3(7), 141-152.

Sivapragasam, C., Vincent and, P., and Vasudevan, G. (2007), Genetic programming model for forecast of short and noisy Data. Hydrological Processes, 21, 266-272.

Solomatine, D., and Dulal, K. (2003), Model trees as an alternative to neural networks in

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

rainfall–runoff modelling. Hydrological Sciences Journal, 48(3), 399-411.

Solomatine, D.P., and Shrestha, D.L. (2009), A novel method to estimate model uncertainty using machine learning techniques. Water Resources Research, 45, W00B11, doi:10.1029/2008WR006839.

Solomatine, D. P. and Xue, Y. I. (2004), M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. Journal of Hydrological Engineering, 9(6), 491-501.

Sudheer, K. P., Gosain, A. K., and Ramasastri, K. S. (2002), A data-driven algorithm for constructing artificial neural network rainfall-runoff models. Hydrological Processes, 16, 1325-1330.

Tokar, A.S., and Johnson, P.A. (1999), Rainfall–runoff modeling using artificial neural networks. Journal of Hydrologic Engineering, 4(3), 232-239.

Toth, E., and Brath, A. (2007), Multistep ahead streamflow forecasting: Role of calibration data in conceptual and neural network modeling. Water Resources Research, 43(11), art. no. W11405.

Wang,W., Van Gelder, P.H.A.J.M., Vrijling, J.K. and Ma, J. (2006), Forecasting Daily Streamflow Using Hybrid ANN Models. Journal of Hydrology, 324, 383-399.

Wilby, R.L., Abrahart, R.J., and Dawson, C. W. (2003), Detection of conceptual model rainfall-runoff processes inside an artificial neural network. Hydrological Sciences Journal, 48(2), 163-181.

Wu, C.L., Chau, K.W., Fan, C. (2010), Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. Journal of Hydrology, 389 (1-2), 146-167.

Xu, Z.X., and Li, J.Y. (2002), Short-term inflow forecasting using an artificial neural network model. Hydrological Processes, 16(12), 2423-2439.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Yu, P.S., Chen, S.T., and Chang I.F. (2006), Support vector regression for real-time flood stage forecasting. Journal of hydrology, 328,704-716.

Zealand, C. M., Burn, D. H., and Simonovic, S. P. (1999), Short term stream flow forecasting using artificial neural networks. Journal of Hydrology, 214, 32-48.

Zhang, B., and Govindaraju, R. S. (2000), Prediction of watershed runoff using Bayesian concepts and modular neural networks. Water Resources Research, 36(3), 753-762.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Figure Captions

Figure 1 Daily rainfall-runoff time series: (a) Wuxi and (b) Chongyang

Figure 2. Flow chart of MANN

Figure 3. Implementation framework of forecasting models

Figure 4. Plots of ACF and PACF of the runoff series with the 95% confidence bounds (the dashed lines), (a) and (c) for Wuxi, and (b) and (d) for Chongyang

Figure 5. CCFs between rainfall and flow series with the 95% confidence bounds (the dashed lines): (a) for Wuxi, and (b) for Chongyang.

Figure 6. Singular Spectrum as a function of lag using varied window lengths L from 3 to 10: (a) and (c) for Wuxi, and (b) and (d) for Chongyang.

Figure 7. Sensitivity analysis of singular Spectrum on varied τ : (a) and (c) for Wuxi and (b) and (d) for Chongyang.

Figure 8. Hydrographs (representative details) and scatter plots of one-step-ahead prediction for (a) Wuxi and (b) Chongyang.

Figure 9. Hydrographs for one-step-ahead prediction using ANN with two types of inputs: (a) Wuxi, and (b) Chongyang.

Figure 10. Lag analysis of observation and forecasts of ANN with two types of inputs: (a) and (c) for Wuxi, and (b) and (d) for Chongyang.

Figure 11. Hydrographs (representative details) and scatter plots of one-step-ahead prediction in SSA mode for (a) Wuxi and (b) Chongyang.

Figure 12. Hydrographs for one-step-ahead prediction using ANN-SSA with two types of inputs: (a) Wuxi, and (b) Chongyang.

Figure 13. Lag analysis of observation and forecasts of ANN-SSA with two types of inputs: (a) and (c) for Wuxi, and (b) and (d) for Chongyang.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Figure 14. AMIs between model inputs and output for ANN with two types of inputs using the data of (a) Wuxi and (b) Chongyang.

Figure 15. AMIs between model inputs and output for ANN and ANN-SSA in the context of R-R forecasting using the data of (a) Wuxi and (b) Chongyang.

Journal of Hydrology, Vol. 399, No. 3-4, 2011, pp. 394–409.

Table Captions

Table 1. Statistical information on rainfall and streamflow data

Table 2. Comparison of methods to determine mode inputs using ANN

Table 3. Optimal p RCs of rainfall and runoff input variables at various forecast horizons

Table 4. R-R Model performances at three forecasting horizons in the normal mode

Table 5. Performances of ANN and MANN in two types of input variables

Table 6. Performances of R-R models in the SSA mode

Table 7. Performances of ANN-SSA and MANN-SSA using two types of input variables

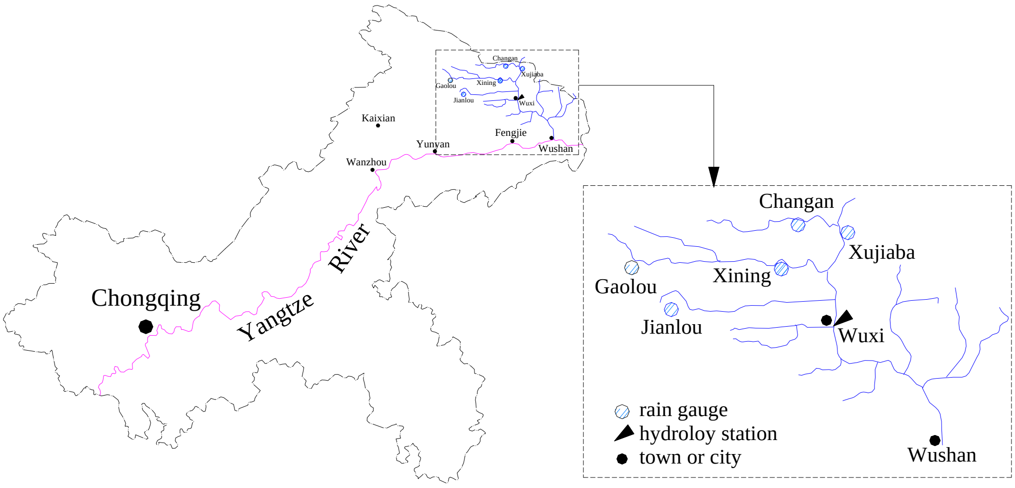

Changan
Xujiaba
Gaolou
Xining
Jianlou
Wuxi
Kaixian
Fengjie
Yunyan
Wanzhou
Wushan
Chongqing
Yangtze
River
Changan
Xujiaba
Gaolou
Xining
Jianlou
Wuxi
rain gauge
hydroloy station
town or city
Wushan

Figure1.


Puqi
Xinzhuang
Shadun
Huangtubang
Bainiqiao
Maojiaqiao
Chongyang
Tangkou
Qingshan
Beigang
Dashaping
Gaojianpu
Tongcheng
Maishi
Shijiadun
hydrology station
rain gauge
city
reservior
river or creek

Figure 2

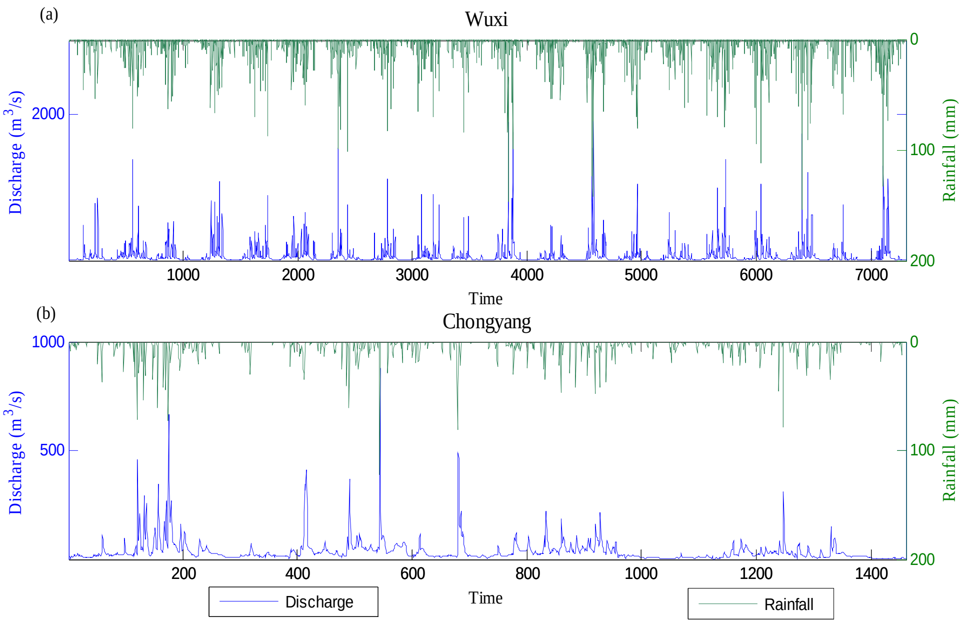

(a)
Wuxi
Discharge (m3/s)
Rainfall (mm)
2000
0
100
200
1000
2000
3000
4000
5000
6000
7000
Time
(b)
Chongyang
1000
500
0
100
200
200
400
600
800
1000
1200
1400
Time
Discharge
Rainfall

Figure 3

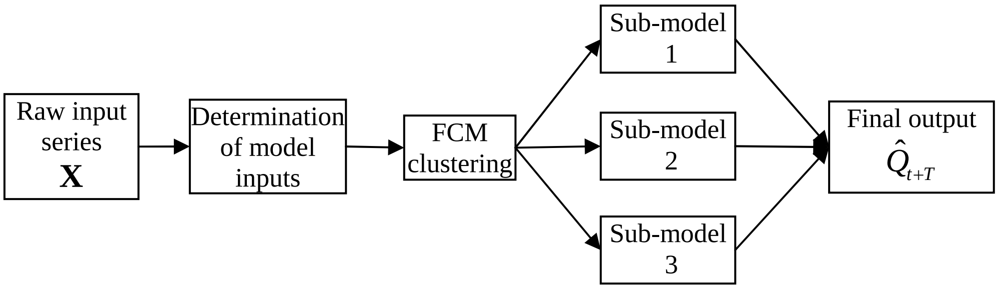

Raw input series
X
Determination of model inputs
FCM clustering
Sub-model 1
Sub-model 2
Sub-model 3
Final output
$\hat{Q}_{t+T}$

Figure 4.

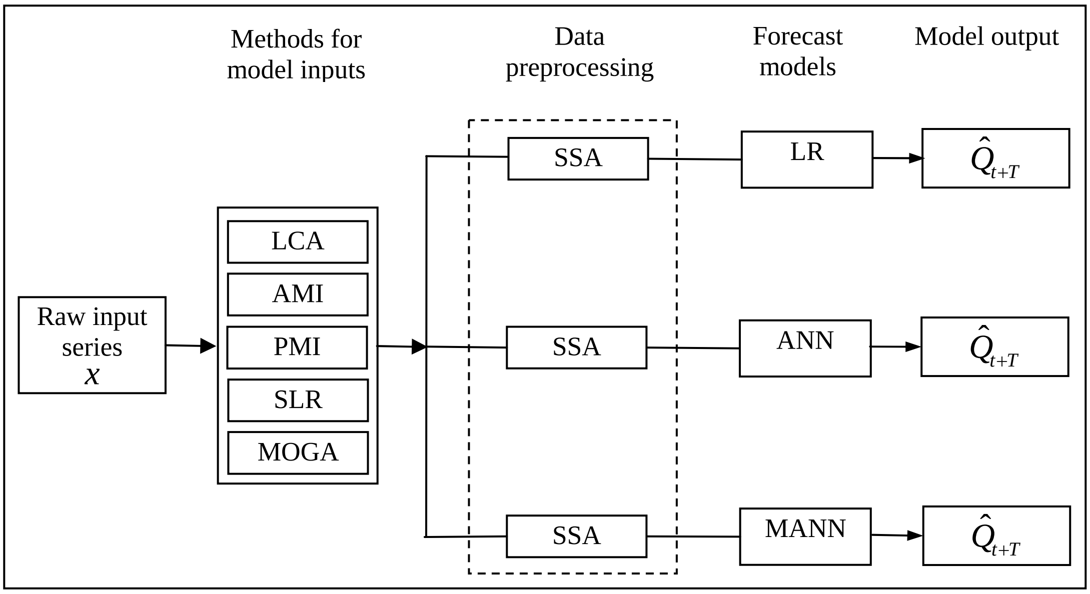

Methods for model inputs
Data preprocessing
Forecast models
Model output
Raw input series x
LCA
AMI
PMI
SLR
MOGA
SSA
SSA
SSA
LR
ANN
MANN
$\hat{Q}_{t+T}$
$\hat{Q}_{t+T}$
$\hat{Q}_{t+T}$

Figure 5.

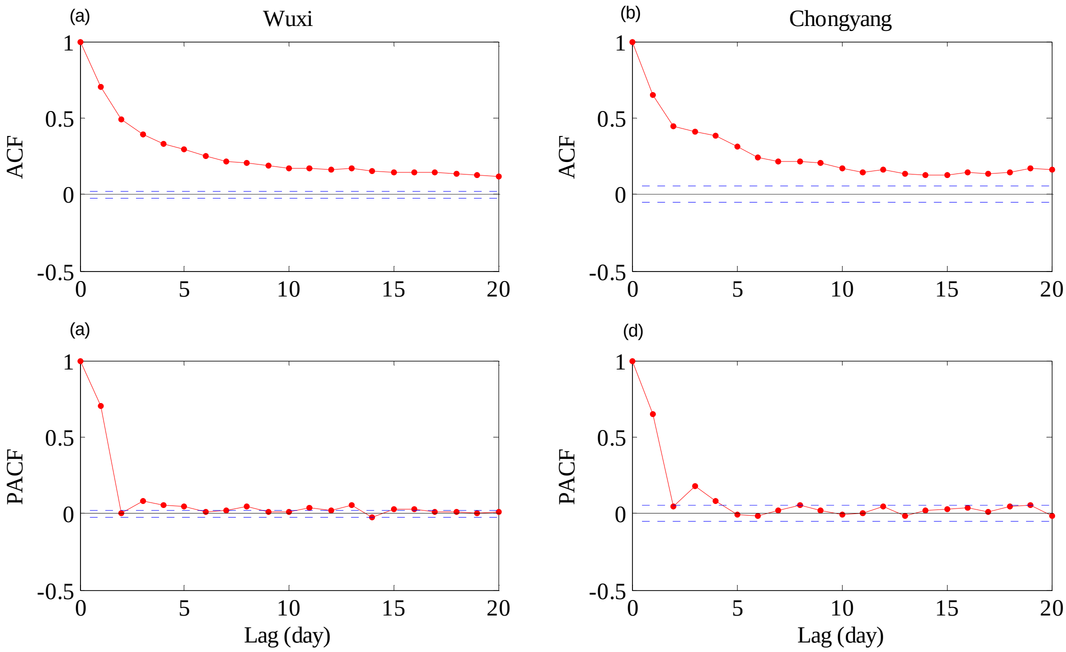


Figure 6.

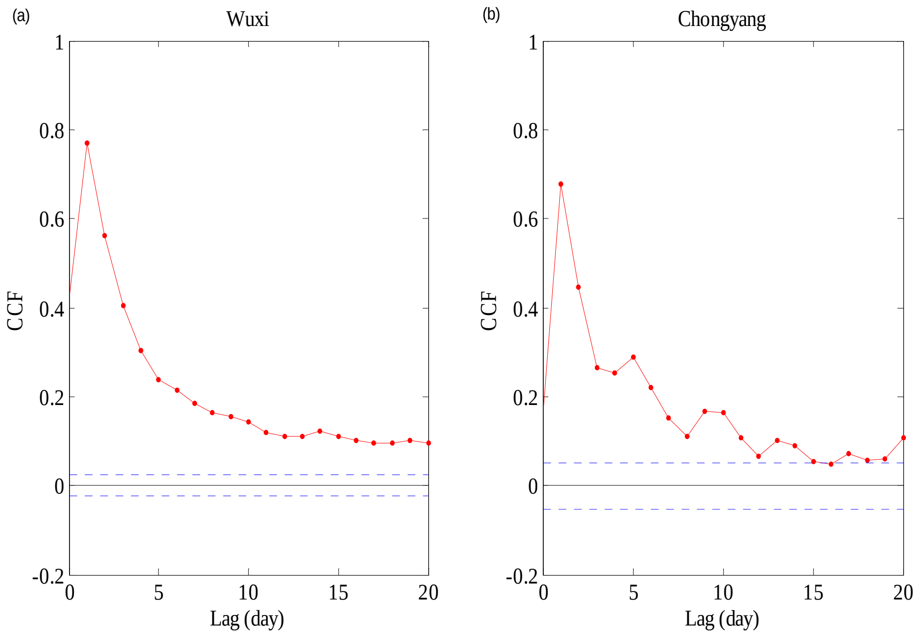

(a)
Wuxi
CCF
1
0.8
0.6
0.4
0.2
0
-0.2
0
5
10
15
20
Lag (day)
(b)
Chongyang
CCF
1
0.8
0.6
0.4
0.2
0
-0.2
0
5
10
15
20
Lag (day)

Figure 7.

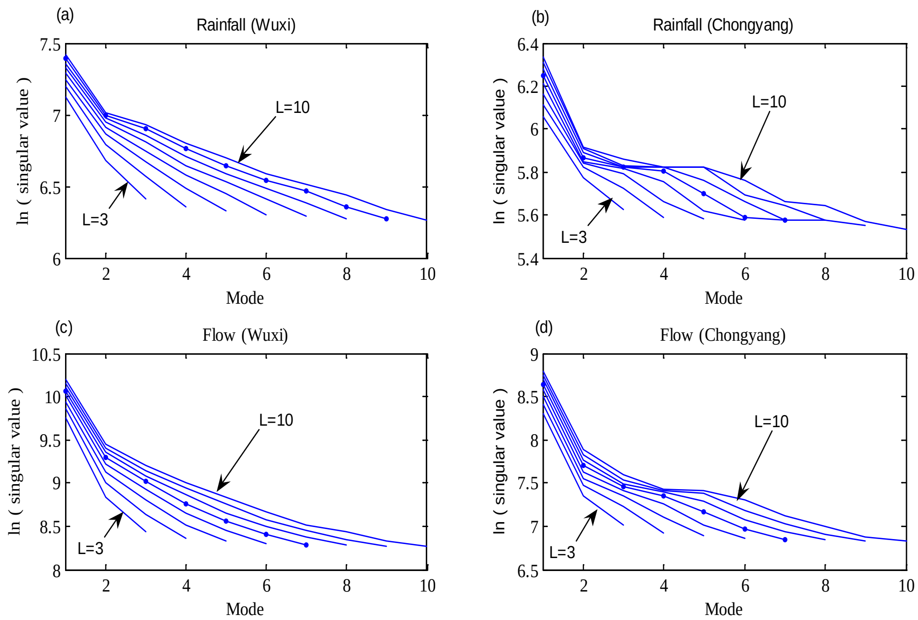
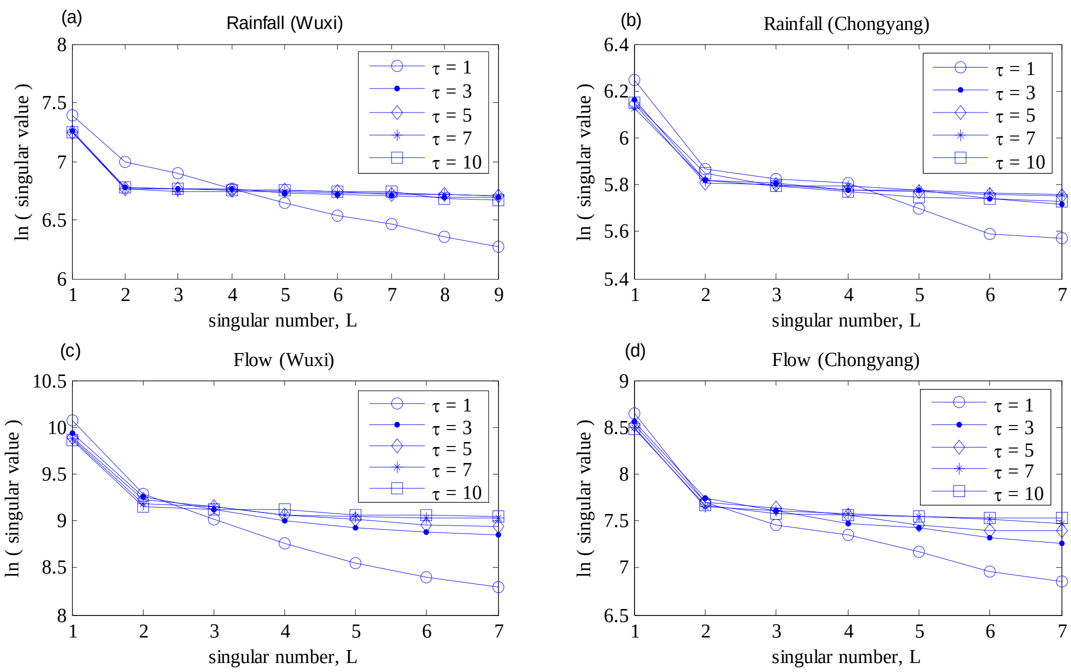

(a)
Rainfall (Wuxi)
L=10
L=3
ln (singular value)
Mode
(b)
Rainfall (Chongyang)
L=10
L=3
ln (singular value)
Mode
(c)
Flow (Wuxi)
L=10
L=3
ln (singular value)
Mode
(d)
Flow (Chongyang)
L=10
L=3
ln (singular value)
Mode

Figure 8.



(a)
Rainfall (Wuxi)
(b)
Rainfall (Chongyang)
(c)
Flow (Wuxi)
(d)
Flow (Chongyang)
ln (singular value)
singular number, L
τ = 1
τ = 3
τ = 5
τ = 7
τ = 10

Figure 9

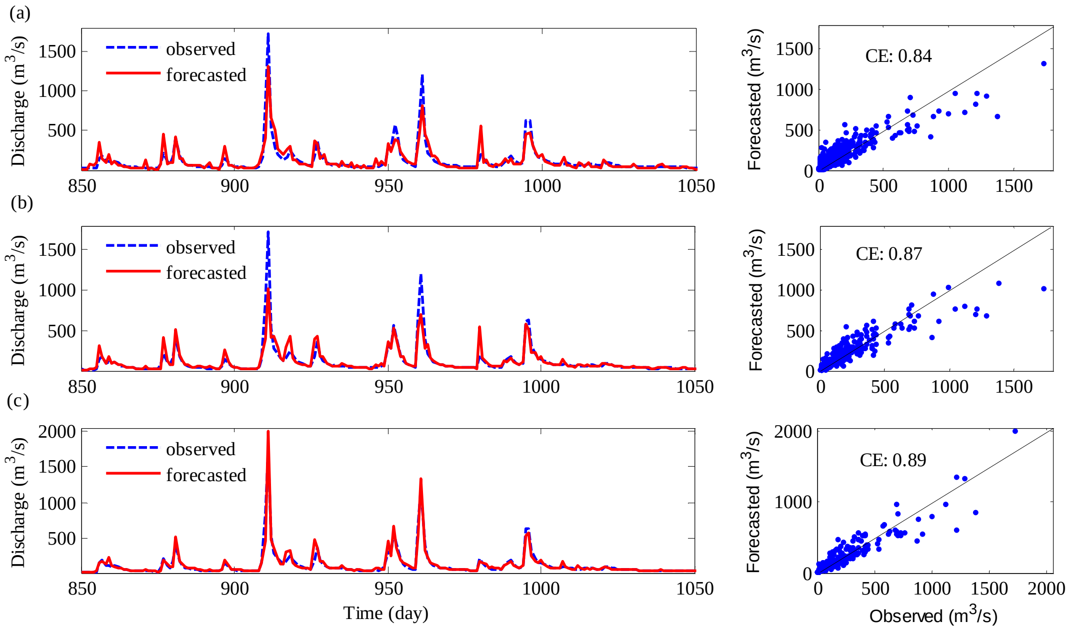

(a)
(b)
(c)
observed
forecasted
Discharge (m3/s)
Forecasted (m3/s)
Observed (m3/s)
Time (day)
CE: 0.84
CE: 0.87
CE: 0.89

Figure 10.

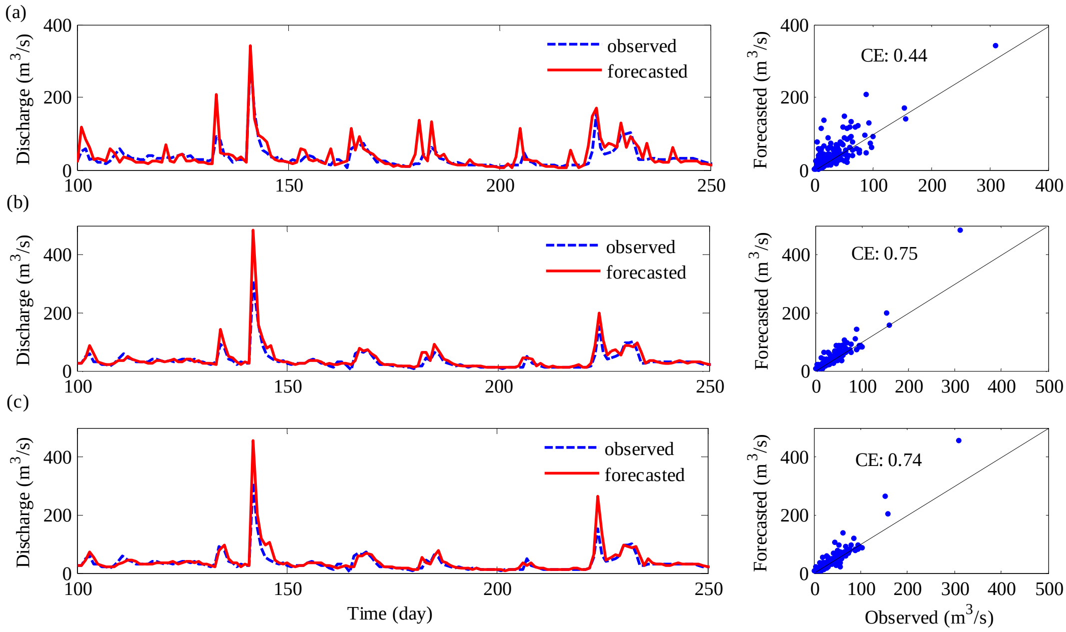

(a)
(b)
(c)
Discharge (m3/s)
observed
forecasted
Time (day)
Forecasted (m3/s)
Observed (m3/s)
CE: 0.44
CE: 0.75
CE: 0.74

Figure 11.

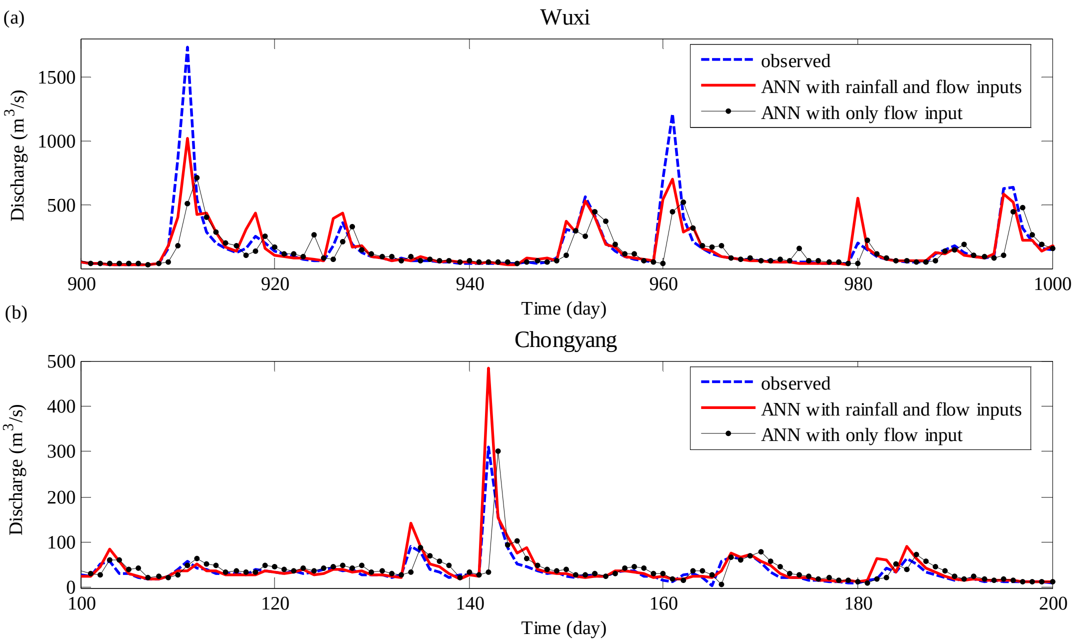
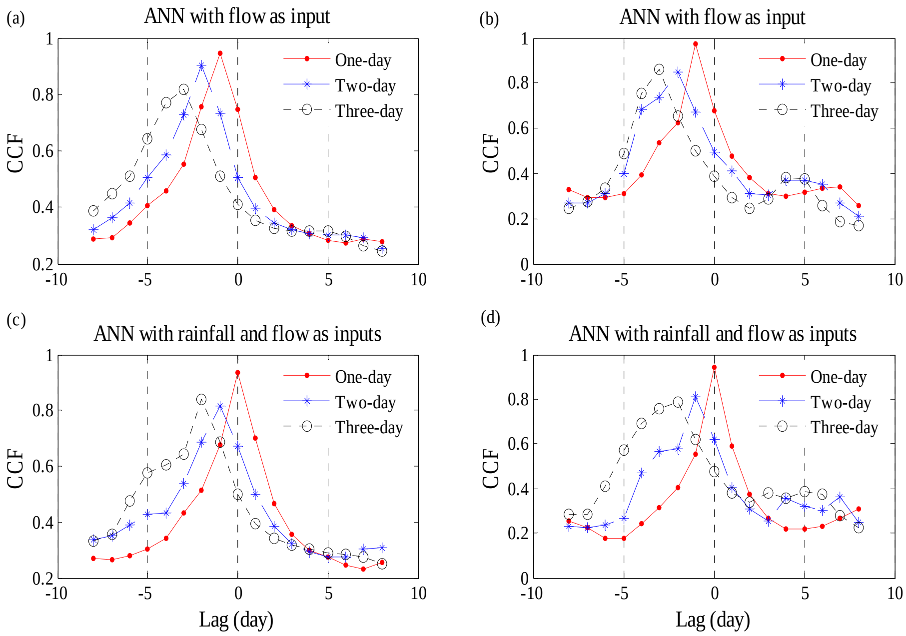

(a)
Wuxi
Discharge (m^3/s)
1500
1000
500
900
920
940
960
980
1000
Time (day)
observed
ANN with rainfall and flow inputs
ANN with only flow input
(b)
Chongyang
Discharge (m^3/s)
500
400
300
200
100
0
100
120
140
160
180
200
Time (day)
observed
ANN with rainfall and flow inputs
ANN with only flow input

Figure 12.

56



57

58 Figure 13.

59

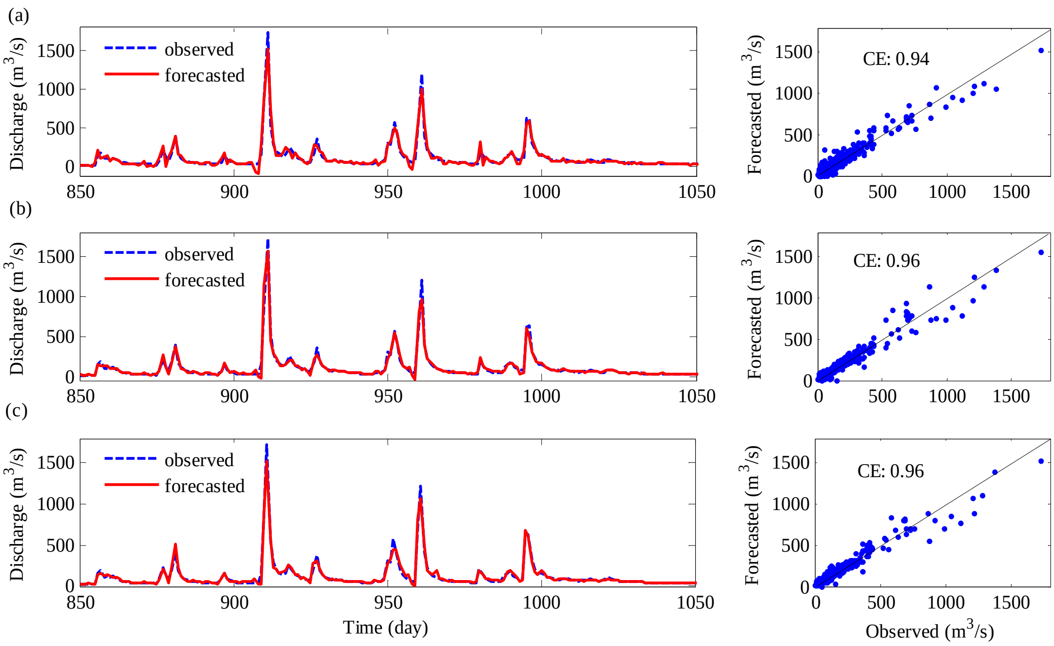

(a)
(b)
(c)
observed
forecasted
Discharge (m3/s)
Forecasted (m3/s)
Observed (m3/s)
Time (day)
CE: 0.94
CE: 0.96
CE: 0.96
0
500
1000
1500
850
900
950
1000
1050

Figure 14.

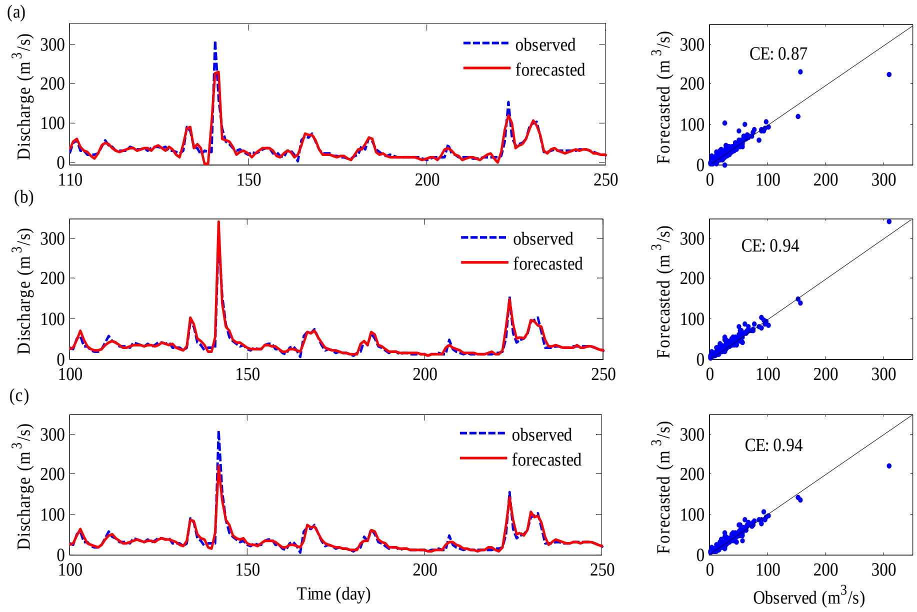
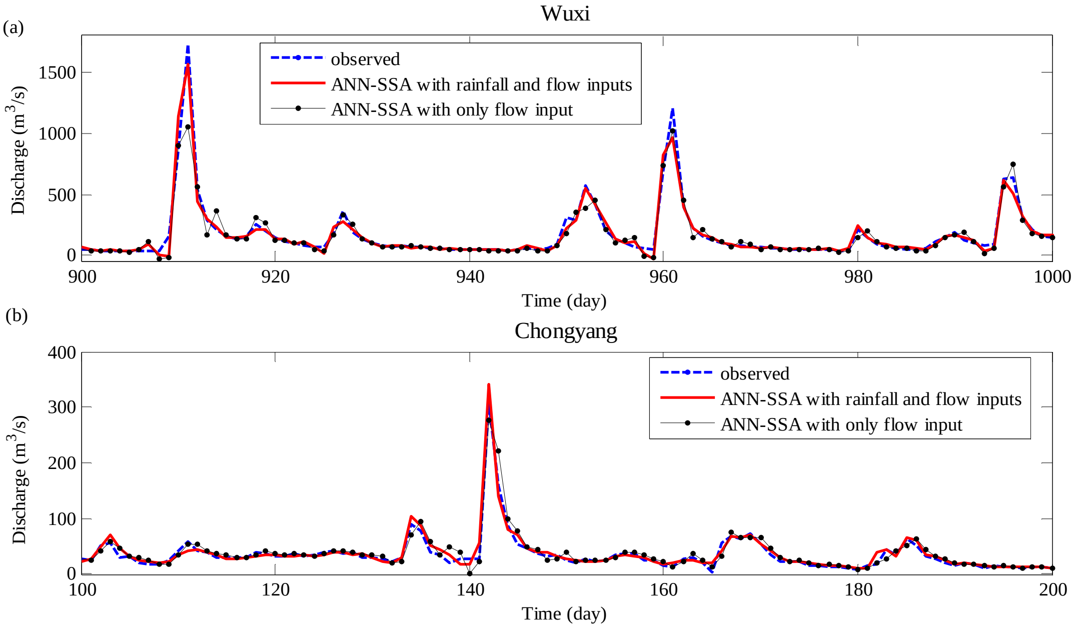

(a)
(b)
(c)
observed
forecasted
Discharge (m^3/s)
Forecasted (m^3/s)
CE: 0.87
CE: 0.94
CE: 0.94
Time (day)
Observed (m^3/s)

Figure 15.



(a)
Wuxi
observed
ANN-SSA with rainfall and flow inputs
ANN-SSA with only flow input
Discharge (m^3/s)
Time (day)
(b)
Chongyang
observed
ANN-SSA with rainfall and flow inputs
ANN-SSA with only flow input
Discharge (m^3/s)
Time (day)

Figure 16.

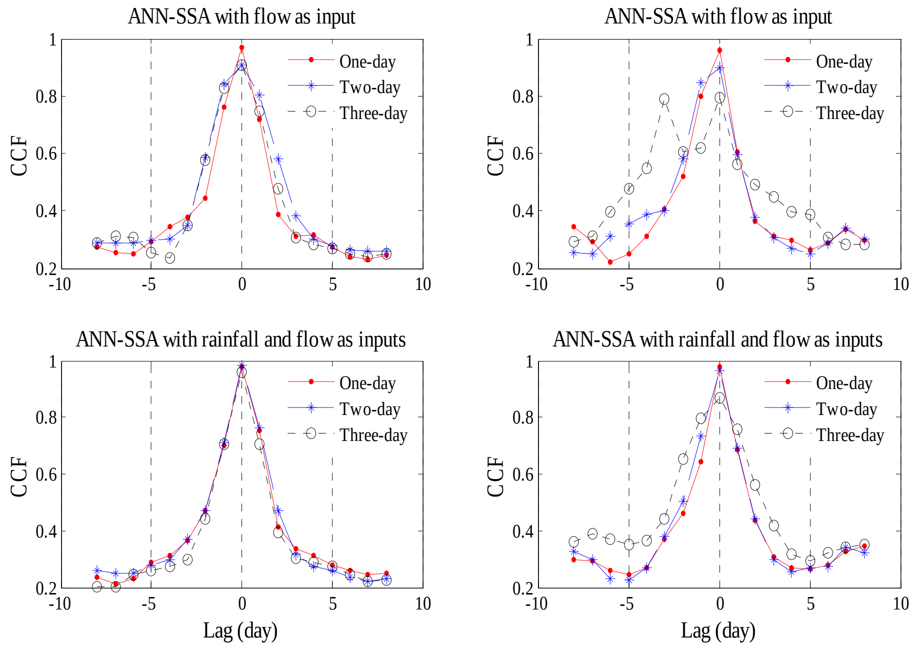

ANN-SSA with flow as input
ANN-SSA with flow as input
ANN-SSA with rainfall and flow as inputs
ANN-SSA with rainfall and flow as inputs
One-day
Two-day
Three-day
CCF
Lag (day)

Figure 17.

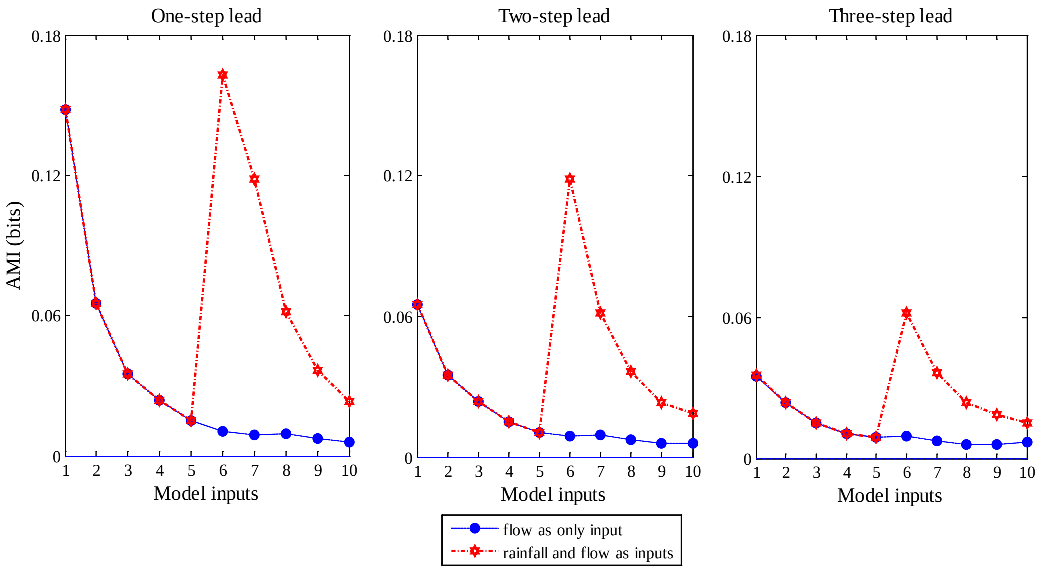

One-step lead
Two-step lead
Three-step lead
AMI (bits)
0.18
0.12
0.06
0
1 2 3 4 5 6 7 8 9 10
Model inputs
flow as only input
rainfall and flow as inputs

Figure 18.

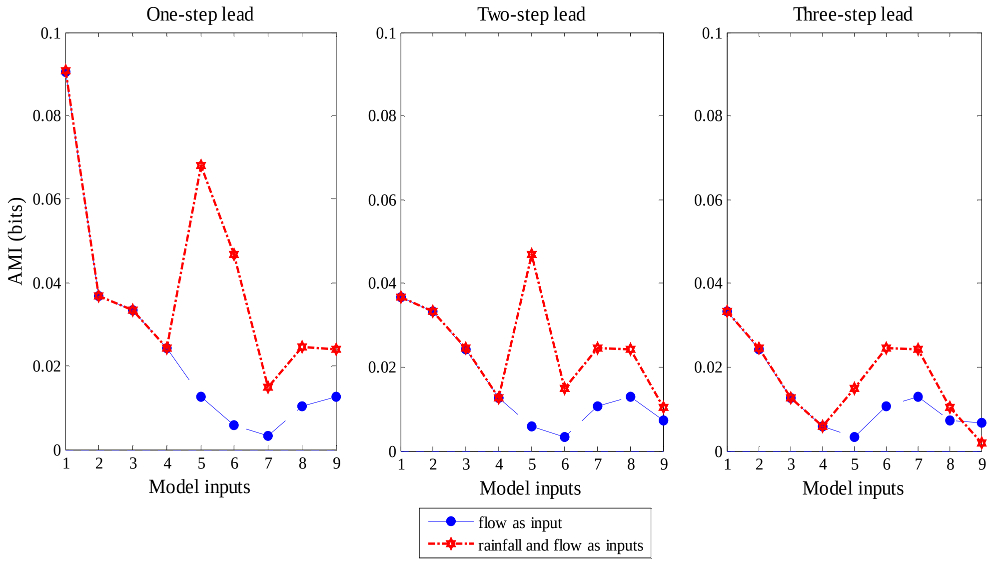

One-step lead
Two-step lead
Three-step lead
AMI (bits)
0.1
0.08
0.06
0.04
0.02
0
1 2 3 4 5 6 7 8 9
Model inputs
flow as input
rainfall and flow as inputs

Figure 19.

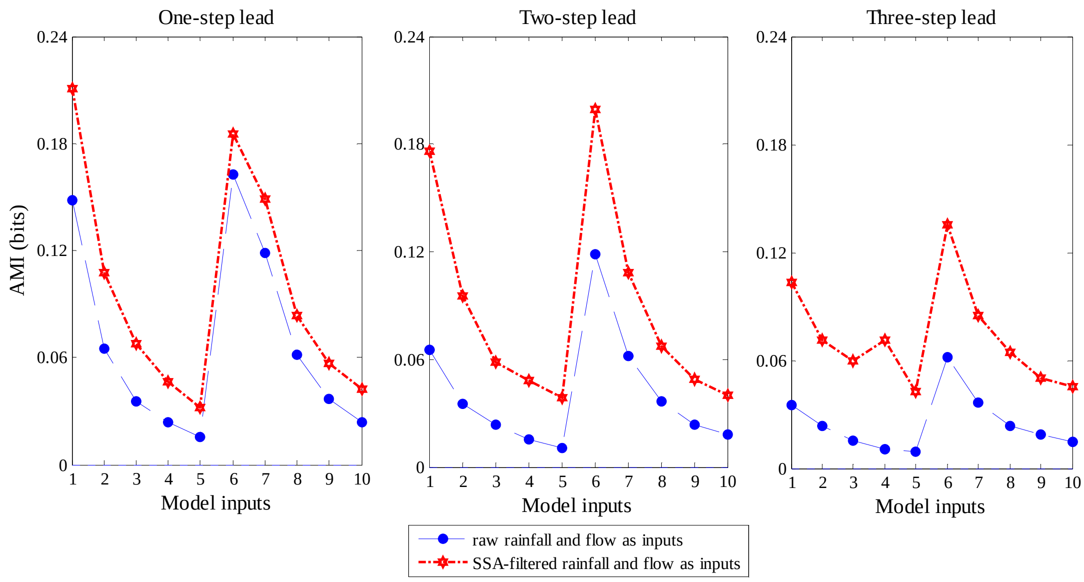

One-step lead
Two-step lead
Three-step lead
AMI (bits)
0.24
0.18
0.12
0.06
0
1 2 3 4 5 6 7 8 9 10
Model inputs
raw rainfall and flow as inputs
SSA-filtered rainfall and flow as inputs

Figure 20.

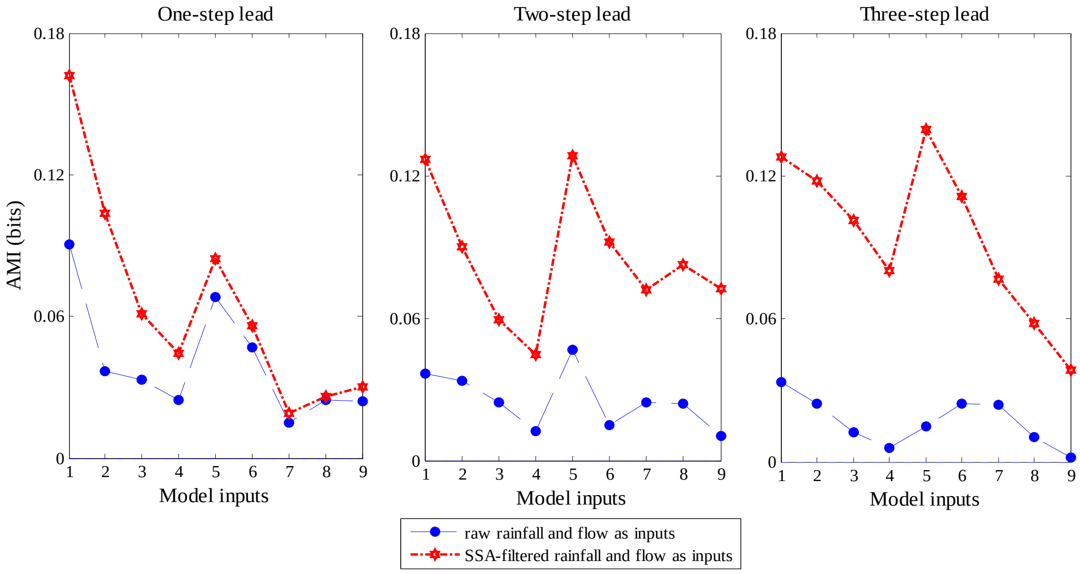

One-step lead
Two-step lead
Three-step lead
AMI (bits)
0.18
0.12
0.06
0
1 2 3 4 5 6 7 8 9
Model inputs
raw rainfall and flow as inputs
SSA-filtered rainfall and flow as inputs

Figure 21.

Table 1. Statistical information on rainfall and streamflow data

Watershed and datasets	**Statistical parameters**						**Watershed area and data period**
	μ	**S_x**	**C_v**	**C_s**	**X_{min}**	**X_{max}**	
Wuxi							
Rainfall(mm)							
Original data	3.7	10.1	0.36	5.68	0	154	Area:
Training	3.4	8.9	0.39	4.96	0	102	2 000 km^2
Cross-validation	3.8	10.9	0.35	6.27	0	147	Data period:
Testing	4.0	11.6	0.35	5.46	0	154	Jan., 1988- Dec., 2007
runoff(m^3/s)							
Original data	61.9	112.6	0.55	7.20	6	2230	
Training	60.6	95.6	0.63	5.90	8	1530	
Cross-validation	60.7	132.2	0.46	8.35	6	2230	
Testing	66.0	122.1	0.54	6.30	10	1730	
Chongyang							
Rainfall(mm)							
Original data	3.1	8.5	0.4	5.7	0.0	122	Area:
Training	3.5	9.8	0.4	5.7	0.0	122	1 700 km^2
Cross-validation	2.9	7.0	0.4	3.9	0.0	48	Data period:
Testing	2.6	7.0	0.4	5.6	0.0	78	Jan., 2004- Dec., 2007
runoff(m^3/s)							
Original data	39.1	54.8	0.7	6.4	2.1	881	
Training	48.1	70.1	0.7	5.5	6.9	881	
Cross-validation	35.6	33.7	1.1	2.3	4.4	226	
Testing	24.5	25.7	1.0	5.1	2.1	310	

Table 2. Comparison of methods to determine mode inputs using ANN

Watershed	Methods	τ	l_1	l_2	m	Effective inputs	Identified ANN	RMSE
Wuxi								
	LCA	1	5	5	10	all	(10-8-1)	41.98
	AMI	1	5	5	10	all	(10-8-1)	41.98
	PMI	1	5	5	10	all	(10-8-1)	41.98
	SLR	1	5	5	10	except for Rt-3	(9-5-1)	40.54
	MOGA	1	5	5	10	Rt, Rt-1, Rt-2, Rt-3, Rt-4, Qt, Qt-1, Qt-4	(8-6-1)	43.23
Chongyang								
	LCA	1	5	4	9	all	(9-9-1)	14.43
	AMI	1	5	4	9	except for Rt	(8-7-1)	14.18
	PMI	1	5	4	9	except for Rt	(8-7-1)	14.18
	SLR	1	5	4	9	except for Rt-1,t-2,t-4	(6-9-1)	13.54
	MOGA	1	5	4	9	Rt, Rt-1, Rt-2, Rt-4, Qt, Qt-2, Qt-3	(7-5-1)	13.57

Table 3. Optimal p RCs of rainfall and runoff input variables at various forecast horizons

Filter model	Prediction horizons	Wuxi		Chongyang	
		Optimal p RCs	RMSE	Optimal p RCs	RMSE
LR-RF					
	1	all RCs	57.13	1 3	25.88
	2	1 2 3 5[a]	58.37	1 2 6	25.81
	3	1 2 3	74.24	1 2 7	25.49
LR-QF					
	1	1 2 3	35.83	1 2 3	8.92
	2	1 2	55.94	1 2	13.41
	3	1	67.60	1	16.60
ANN-RF					
	1	1 3 4 6 7	49.72	1 3 5 7	18.45
	2	1 2 3 4 5	52.38	1 3	19.11
	3	1 2 3 4	60.01	1 2	21.72
ANN-QF					
	1	1 2 3 4	31.49	1 2 3	11.67
	2	1 2 7	45.39	1 2	14.97
	3	3 7	53.55	1	17.26

Note: [a] the numbers of "1, 2, 3, 5" stand for RC, RC2, RC4, and RC5, and RC1 is associated with the maximum eigenvalue, RC2 corresponds to the second largest eigenvalue, etc.

Table 4. R-R Model performances at three prediction horizons in the normal mode

Watershed	**Model**	**RMSE**			**CE**			**PI**		
		1*	**2***	**3***	**1**	**2**	**3**	**1**	**2**	**3**
Wuxi										
	LR	49.40	89.40	108.90	0.84	0.46	0.21	0.70	0.51	0.39
	ANN	43.97	87.32	104.94	0.87	0.49	0.26	0.76	0.54	0.43
	MANN	40.44	71.87	86.54	0.89	0.66	0.50	0.80	0.69	0.61
Chongyang										
	LR	19.18	22.74	25.53	0.44	0.22	0.01	0.17	0.29	0.24
	ANN	12.90	25.80	27.81	0.75	0.10	-0.15	0.63	0.10	0.13
	MANN	13.27	26.86	23.96	0.74	-0.07	0.14	0.61	0.03	0.35

* The number of “1, 2, and 3” denote one-, two-, and three-step-ahead forecasts

Table 5. Performances of ANN and MANN in two types of input variables

Watershed	Input variables	Model	RMSE			CE			PI		
			1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Wuxi											
	Rainfall+Flow										
		ANN	43.97	87.32	104.94	0.87	0.49	0.26	0.76	0.54	0.43
		MANN	40.44	71.87	86.54	0.89	0.66	0.50	0.80	0.69	0.61
	Flow										
		ANN	81.3	104.6	111.5	0.56	0.27	0.17	0.19	0.33	0.36
		MANN	75.7	93.7	97.1	0.62	0.41	0.37	0.30	0.46	0.51
Chongyang											
	Rainfall+Flow										
		ANN	12.90	25.80	27.81	0.75	0.10	-0.15	0.63	0.10	0.13
		MANN	13.27	26.86	23.96	0.74	-0.07	0.14	0.61	0.03	0.35
	Flow										
		ANN	20.3	26.1	27.8	0.38	-0.04	-0.18	0.08	0.06	0.10
		MANN	17.8	22.3	23.4	0.52	0.24	0.17	0.29	0.31	0.36

Table 6. Performances of R-R models in the SSA mode

Watershed	Model	RMSE			CE			PI		
		1	2	3	1	2	3	1	2	3
Wuxi										
	LR-SSA	29.02	44.42	58.34	0.94	0.87	0.77	0.90	0.88	0.82
	ANN-SSA	25.40	27.10	33.96	0.96	0.95	0.92	0.92	0.96	0.94
	MANN-SSA	25.08	26.87	34.05	0.96	0.95	0.92	0.92	0.96	0.94
Chongyang										
	LR-SSA	9.19	13.53	14.61	0.87	0.72	0.68	0.81	0.75	0.75
	ANN-SSA	6.22	7.08	11.12	0.94	0.93	0.82	0.91	0.93	0.86
	MANN-SSA	6.42	8.13	13.14	0.94	0.90	0.74	0.91	0.91	0.80

Table 7. Performances of ANN-SSA and MANN-SSA using two types of input variables

Watershed	Input variables	Model	RMSE			CE			PI		
			1	2	3	1	2	3	1	2	3
Wuxi											
	Rainfall+runoff										
		ANN-SSA	25.40	27.10	33.96	0.96	0.95	0.92	0.92	0.96	0.94
		MANN-SSA	25.08	26.87	34.05	0.96	0.95	0.92	0.92	0.96	0.94
	runoff										
		ANN-SSA	31.02	50.64	61.80	0.94	0.83	0.74	0.88	0.84	0.80
		MANN-SSA	26.20	41.02	48.69	0.95	0.89	0.84	0.92	0.90	0.88
Chongyang											
	Rainfall+runoff										
		ANN-SSA	6.22	7.08	11.12	0.94	0.93	0.82	0.91	0.93	0.86
		MANN-SSA	6.42	8.13	13.14	0.94	0.90	0.74	0.91	0.91	0.80
	runoff										
		ANN-SSA	7.93	11.15	15.72	0.91	0.81	0.63	0.86	0.83	0.72
		MANN-SSA	7.32	10.19	15.71	0.92	0.84	0.63	0.88	0.86	0.72