*Water* **2015**, *7*, 4144-4160; doi:10.3390/w7084144

OPEN ACCESS

*water*

**ISSN 2073-4441**
www.mdpi.com/journal/water

*Article*

# Daily Runoff Forecasting Model Based on ANN and Data Preprocessing Techniques

**Yun Wang [1,2], Shenglian Guo [1,*], Lihua Xiong [1], Pan Liu [1] and Dedi Liu [1]**

[1] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China; E-Mails: wyun.1987@gmail.com (Y.W.); Xionglh@whu.edu.cn (L.X.); liupan@whu.edu.cn (P.L.); Dediliu@whu.edu.cn (D.L.)

[2] China Yangtze Power Co., Ltd., Yichang, Hubei 443002, China

* Author to whom correspondence should be addressed; E-Mail: slguo@whu.edu.cn; Tel./Fax: +86-27-68773568.

Academic Editor: Kwok-wing Chau

*Received: 10 June 2015 / Accepted: 20 July 2015 / Published: 28 July 2015*

---

**Abstract:** There are many models that have been used to simulate the rainfall-runoff relationship. The artificial neural network (ANN) model was selected to investigate an approach of improving daily runoff forecasting accuracy in terms of data preprocessing. Singular spectrum analysis (SSA) as one data preprocessing technique was adopted to deal with the model inputs and the SSA-ANN model was developed. The proposed model was compared with the original ANN model without data preprocessing and a nonlinear perturbation model (NLPM) based on ANN, *i.e.*, the NLPM-ANN model. Eight watersheds were selected for calibrating and testing these models. Comparative study shows that the learning and training ability of ANN models can be improved by SSA and NLPM techniques significantly, and the performance of the SSA-ANN model is much better than the NLPM-ANN model, with high foresting accuracy. The SSA-ANN1 model, which only considers rainfall as model input, was compared with the SSA-ANN2 model, which considers both rainfall and previous runoff as model inputs. It is shown that the Nash-Sutcliffe criterion of the SSA-ANN2 model is much higher than that of the SSA-ANN1 model, which means that the proper selection of previous runoff data as rainfall-runoff model inputs can significantly improve model performance since they usually are highly auto-correlated.

**Keywords:** daily runoff forecasting; data preprocessing; linear perturbation model; singular spectrum analysis; artificial neural network

## 1. Introduction

Real-time hydrological forecasting plays an important role in flood control and reservoir operation, and higher forecasting precision can increase the utilization efficiency of water resources. Traditionally, hydrological simulation modeling systems are classified into three main groups, namely empirical black box, lumped conceptual, and distributed physically-based models [1]. The last two groups focus on understanding hydrological processes and involve various physical phenomena. Owing to the complexity of the rainfall-runoff process, these physical process simulations and model calibrations require large amounts of hydrological data. On the contrary, black-box modeling does not require a deep knowledge of the underlying physics and also can solve the problem of the scarcity of the data. Several black-box models have been developed and used in hydrological forecasting, such as fuzzy theory [2,3], artificial neural network [4,5], chaos [6], genetic programming [7], support vector machine [8], and so on.

Artificial neural network, inspired by research into the biological neural networks, has a flexible structure, and self-learning and self-adaptive features. In 2000, the American Society of Civil Engineering (ASCE) Task Committee explicitly reviewed the application of artificial neural networks in hydrology [9,10]. Hsu *et al.* [5] mentioned that the artificial neural network (ANN) model can identify the complex nonlinear relationship between runoff and rainfall time series, even though the model structure and parameters cannot represent the physical process of the catchments. Maier and Dandy [11] reviewed using ANN models to deal with water resource variables prediction, outlined the steps that should be followed in the development of ANN models, and concluded that the ANN model has advantages in hydrological forecasting. Currently, ANN is still a research hot point and has been successfully applied in hydrological forecasting [12–22].

Due to the highly seasonal variation, and nonlinear and noisy characteristics of hydrological time series, preprocessing input data becomes an effective way to improve model precision [23–28]. Considering the highly seasonal variation of rainfall and runoff time series, Nash and Brasi [23] developed the linear perturbation model (LPM) based on the assumption that subtraction of the seasonal means from the original series would remove much of the non-linearity of the rainfall-runoff process. The relationship between the departures is simulated by the linear response function, but only part of the nonlinearity of the rainfall-runoff process can be removed by subtracting the seasonal means. Pang *et al.* [16] used the ANN model to replace the linear response function and proposed a nonlinear perturbation model (NLPM) based on ANN (NLPM-ANN). The advantage of the NLPM-ANN model is that it is capable of obtaining satisfactory results even if the explicit form of the relationship between the involved variables is unknown.

Considering that the hydrological time series can be viewed as a combination of quasi-periodic signals contaminated by noises to some extent [29], the singular spectrum analysis (SSA) proposed by Vautard *et al.* [30] can decompose the time series into a sum of a small number of interpretable components, such as a slowly varying trend, oscillatory components, and a "structureless" noise [31]. By performing a spectrum analysis on the input data, eliminating the noises, and inverting the remaining components to

yield a "filtered" time series, the model performance could be improved. Sivapragasam *et al.* [25] proposed a prediction technique based on SSA coupled with support vector machines to predict runoff and rainfall, and showed that the proposed technique yields a significantly higher prediction accuracy than that of the nonlinear prediction method. Wu and Chau [29] also found that SSA can considerably improve the performance of the rainfall-runoff model and it is promising in hydrological forecasting.

In this paper, an approach of improving daily runoff forecasting accuracy in terms of data preprocessing and the selection of predictive factors is discussed. The artificial neural network (ANN) is used for rainfall-runoff simulation. The SSA and LPM techniques are adopted to deal with data preprocessing. Then SSA-ANN models are developed and compared with the NLPM-ANN model based on the daily data from the eight watersheds used by Pang *et al.* [16]. A comparative study is also conducted involving two different types of model inputs, namely considering rainfall as an input and considering both rainfall and runoff as inputs.

## 2. Data-Driven Models

### *2.1. NLPM-ANN Model*

The structure of the NLPM-ANN model as shown in Figure 1 was proposed by Pang *et al.* [16] to consider the influence of seasonal changes and the nonlinearity of the rainfall-runoff process. The model input is divided into two parts. The first is the series of the seasonal expectations of the input ($p_d$) that is transformed to the series of the seasonal expectations of the output ($q_d$) through an undefined relation. The second part, which is the input perturbations ($P_i$-$p_d$), is transformed into the output perturbations ($Q_i$-$q_d$) through ANN. The total output is the sum of the seasonal expectations of the output and the output perturbations.

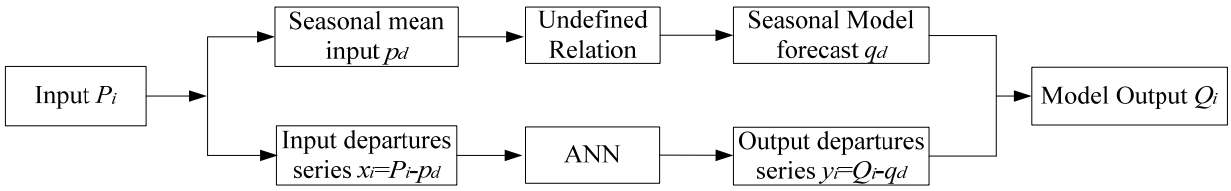


**Figure 1.** Schematic diagram of the NLPM-ANN model.

### *2.2. Singular Spectrum Analysis*

Singular spectrum analysis (SSA) is a suitable analysis method for researching the period oscillatory behavior. It is also a statistical technique starting from a dynamic reconstruction of the time series and is associated with empirical orthogonal function (EOF). Generally, SSA can be considered as a special application of EOF decomposition. The main purpose of SSA is converting a one-dimensional time series into a multi-dimensional matrix with a given window length, and then the orthogonal decomposition of this matrix is obtained. If the obvious pairs of eigenvalues are produced and the corresponding EOF is almost periodic or orthogonal, this corresponding EOF can be considered the oscillatory behavior of the time series.

Brief operating procedures of SSA are summarized as follows. Assume that the series is a nonzero series $F = \{f_0, f_1, \ldots, f_{N-1}\}$ ($f_i \neq 0$), the length of series is $N$ (>2). Given a window length $L$, the

one-dimensional time series can be transferred into a sequence of $L$-dimensional vectors $\mathbf{X}_i = \{f_{i-1}, \ldots, f_{i+L-2}\}^T$, ($I = 1, \ldots, K = N-L+1$). The $K$ vectors $\mathbf{X}_i$ will form the columns of the ($L \times K$) trajectory matrix:

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{bmatrix} \quad (1)$$

Then the singular value decomposition (SVD) of the trajectory matrix $\mathbf{X}$ is conducted. Let $\mathbf{S} = \mathbf{X}\mathbf{X}^T$. The eigenvalues and eigenvectors of $\mathbf{S}$ can be calculated, and these eigenvalues range in the decreasing order of magnitude. According to the conventional computation of EOF, an expansion of the matrix $\mathbf{X}$ is represented as:

$$x_{i+j} = \sum_{k=1}^{L} a_i^k \mathbf{E}_j^k \quad (2)$$

where $i = 1, 2, \ldots, N - L + 1, j = 1,2, \ldots, L, k = 1, 2, \ldots, L$, $a_i^k$ is the time principal components (T-PC), $\mathrm{E}_j^k$ is the corresponding eigenvector denoted by T-EOF. The key step of SSA is to reconstruct a new one-dimensional series of length $N$ using each component of the T-PC and T-EOF. The formula is expressed as follows:

$$x_i^k = \begin{cases} \dfrac{1}{i}\displaystyle\sum_{j=1}^{L} a_{ij}^k \mathrm{E}_j^k & 1 \le i \le L-1 \\ \dfrac{1}{L}\displaystyle\sum_{j=1}^{L} a_{ij}^k \mathrm{E}_j^k & L \le i \le N-L+1 \\ \dfrac{1}{N-i+1}\displaystyle\sum_{j=i-N+L}^{L} a_{ij}^k \mathrm{E}_{ij}^k & N-L+2 \le i \le N \end{cases} \quad (3)$$

Equation (3) produces an $N$-length time series $F_k$, thus the initial series $F$ is decomposed into the sum of $L$ series:

$$F = \sum_{k=1}^{L} F_k \quad (4)$$

If the number of contributing components is $p$, then the filtered series is the sum of $p$ series:

$$F = \sum_{k=1}^{p} F_k \quad (5)$$

The sum of the remaining series is noise. As mentioned above, these reconstructed components can be associated with the trend, oscillations, or noise of the original time series with proper choices of $L$ and $p$.

### *2.3. Artificial Neural Network*

ANN can be categorized as single-layer, bilayer, and multilayer according to the number of layers, and as feed-forward, recurrent, and self-organizing according to the direction of information flow and processing [9]. Among these different architectures, the multilayer feed-forward networks, which consist of an input layer, several hidden layers, and an output layer, have been widely used. Each layer has different nodes, and the number of hidden layers and the hidden nodes of each hidden layer are usually determined by trial-and-error method.

Assuming the three-layer ANN denoted by $m \times h \times 1$, where $m$ stands for the number of input nodes, namely the number of predictive factors, and $h$ is the number of nodes in the hidden layer, the ANN prediction model can be formulated as:

$$\widehat{Q_{t+T}} = f(\mathbf{X}_t, w, \theta, m, h) = \theta_0 + \sum_{j=1}^{h} w_j^{out} \varphi\left(\sum_{i=1}^{m} w_{ji} \mathbf{X}_t + \theta_j\right) \tag{6}$$

where $X_t$ is the input data; $T$ is the length of lead time; φ denotes transfer functions; $w_{ji}$ are the weights defining the link between the $i$th node of the input layer and the $j$th of the hidden layer; $\theta_j$ are biases associated with the $j$th node of the hidden layer; $w_j^{out}$ are the weights associated with the connection between the $j$th node of the hidden layer and the node of the output layer; and $\theta_0$ is the bias at the output node. The Levenberg–Marquardt algorithm is chosen to adjust the values of $w$ and θ in this study [32].

### *2.4. Proposed SSA-ANN Models*

The SSA-ANN models are proposed with the aim of analyzing the effect of data processing. The flowchart of SSA-ANN models is illustrated in Figure 2, where the original series is decomposed into oscillations and noise by SSA, firstly. Then the reconstructed series is selected as the ANN model input. If the input is the rainfall data series only, the SSA-ANN1 model is built to simulate the relationship between rainfall and runoff. If the input contains both the rainfall and runoff data series, the SSA-ANN2 model is built to simulate the relationship between rainfall and previous runoff with forecasting runoff.

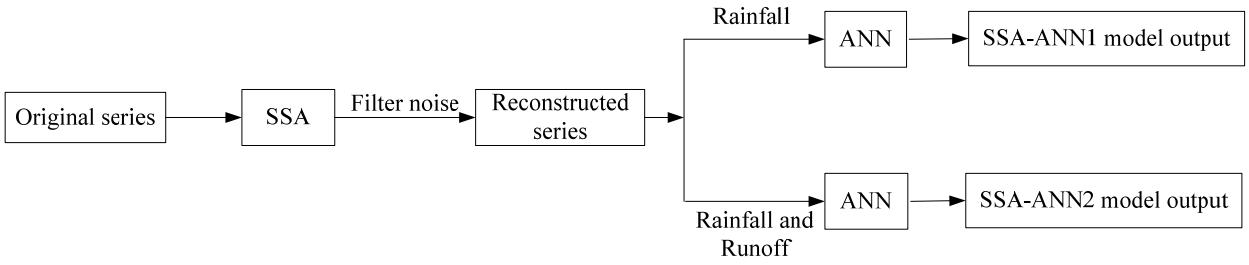


**Figure 2.** Schematic diagram of SSA-ANN models.

### *2.5. Evaluation of Model Performances*

Two criteria are selected to evaluate the prediction performance based on Chinese Hydrological Forecasting (or prediction) guidelines (2008), they are:

(1) Determination coefficient (or Nash-Sutcliffe criterion) ($R^2$)

$$R^2 = (1 - \frac{\sum_{t=1}^{n}(Q_t - Q_t^{'})^2}{\sum_{t=1}^{n}(Q_t - \overline{Q_t})^2}) \tag{7}$$

(2) Water balance coefficient (WB)

$$WB = \frac{\sum_{t=1}^{n} Q_t^{'}}{\sum_{t=1}^{n} Q_t} \tag{8}$$

where $n$ is the number of year, $Q_t$ and $Q_t^{'}$ are the observed and forecasted inflows, respectively, $\overline{Q_t}$ is the average value of observed flow; if the values of $R^2$ and $WB$ are closer to one, the better the prediction results that are obtained.

## 3. Comparative Study

### *3.1. Data*

To compare the proposed SSA-ANN models with the NLPM-ANN model, eight watersheds in China used by Pang *et al.* [16] were selected as case studies in this paper. The data include the daily rainfall and runoff data. Each of data series is divided into three parts, *i.e.*, training set, cross-validation set, and testing set. The training set is used to train the network and the cross-validation set is used to check the progress of the network and implement an early stopping approach in order to avoid the over-fitting of the training set. The testing set serves as model evaluation. Table 1 lists statistical information about all watersheds, including mean (μ), standard deviation ($S_x$), maximum ($X_{max}$), and minimum ($X_{min}$). As shown in Table 1, the training data does not cover the cross-validation or testing data totally. In order to ensure the extrapolation ability of ANN and avoid numerical difficulties during calculation, all data are scaled to the interval [−0.9, 0.9] by normalization.

**Table 1.** List of the watershed statistical information.

| Watershed and Datasets | | | Statistical Parameters | | | | Data Period |
|---|---|---|---|---|---|---|---|
| | | | μ | $S_x$ | $X_{max}$ | $X_{min}$ | |
| Jiahe area: 5578 km$^2$ | rainfall (mm) | whole data | 2.3 | 5.9 | 71.4 | 0 | January 1980–December 1990 |
| | | training data | 2.3 | 6.0 | 68.9 | 0 | |
| | | cross-validation data | 2.3 | 6.2 | 71.4 | 0 | |
| | | testing data | 2.1 | 5.3 | 44.2 | 0 | |
| | runoff (m$^3$) | whole data | 58.7 | 125.1 | 2620 | 6.5 | |
| | | training data | 61.9 | 141.6 | 2620 | 6.5 | |
| | | cross-validation data | 55.3 | 99.6 | 1220 | 7.9 | |
| | | testing data | 50.7 | 76.4 | 1080 | 10.1 | |
| Laoguanhe area: 4217 km$^2$ | rainfall (mm) | whole data | 2.2 | 6.4 | 69.4 | 0 | January 1980–December 1990 |
| | | training data | 2.3 | 6.8 | 69.2 | 0 | |
| | | cross-validation data | 2.0 | 5.8 | 56.0 | 0 | |
| | | testing data | 2.0 | 5.7 | 69.4 | 0 | |
| | runoff (m$^3$) | whole data | 27.1 | 73.6 | 1460 | 0.1 | |
| | | training data | 33.5 | 84.1 | 1460 | 0.4 | |
| | | cross-validation data | 16.8 | 50.6 | 586 | 0.1 | |
| | | testing data | 14.8 | 46.1 | 793 | 0.2 | |

**Table 1.** *Cont.*

| Watershed and Datasets | | | Statistical Parameters | | | | Data Period |
|---|---|---|---|---|---|---|---|
| | | | μ | $S_x$ | $X_{max}$ | $X_{min}$ | |
| Baohe area: 3415 km$^2$ | rainfall (mm) | whole data | 2.5 | 6.9 | 80.6 | 0 | January 1980–December 1990 |
| | | training data | 2.5 | 7.1 | 80.6 | 0 | |
| | | cross-validation data | 2.2 | 6.0 | 51.3 | 0 | |
| | | testing data | 2.6 | 6.8 | 80.5 | 0 | |
| | runoff (m$^3$) | whole data | 46.5 | 129.4 | 4020 | 0 | |
| | | training data | 49.7 | 150.7 | 4020 | 1.2 | |
| | | cross-validation data | 31.4 | 54.8 | 523 | 3.8 | |
| | | testing data | 50.3 | 96.8 | 2010 | 0.0 | |
| Mumahe area: 1224 km$^2$ | rainfall (mm) | whole data | 3.2 | 8.8 | 132.8 | 0 | January 1980–December 1990 |
| | | training data | 3.2 | 8.6 | 132.8 | 0 | |
| | | cross-validation data | 3.3 | 9.3 | 98.6 | 0 | |
| | | testing data | 2.9 | 9.1 | 94.4 | 0 | |
| | runoff (m$^3$) | whole data | 39.3 | 80.3 | 1270 | 1.2 | |
| | | training data | 41.0 | 80.8 | 1270 | 1.2 | |
| | | cross-validation data | 40.6 | 82.1 | 796 | 4.6 | |
| | | testing data | 32.1 | 76.4 | 990 | 2 | |
| Nianyushan area: 924 km$^2$ | rainfall (mm) | whole data | 3.8 | 11.6 | 269.5 | 0 | January 1975–December 1999 |
| | | training data | 3.9 | 12.2 | 269.5 | 0 | |
| | | cross-validation data | 3.3 | 9.3 | 102.5 | 0 | |
| | | testing data | 3.7 | 10.8 | 144.7 | 0 | |
| | runoff (m$^3$) | whole data | 18.5 | 62.1 | 2095 | 0 | |
| | | training data | 19.8 | 68.3 | 2095 | 0 | |
| | | cross-validation data | 13.5 | 33.2 | 508 | 0 | |
| | | testing data | 17.6 | 55.9 | 822 | 0 | |
| Gaoguan area: 303 km$^2$ | rainfall (mm) | whole data | 4.2 | 12.5 | 179.1 | 0 | January 1984–December 1999 |
| | | training data | 4.4 | 12.8 | 179.1 | 0 | |
| | | cross-validation data | 3.5 | 11.3 | 143.8 | 0 | |
| | | testing data | 4.2 | 12.7 | 116.0 | 0 | |
| | runoff (m$^3$) | whole data | 5.8 | 15.1 | 246 | 0 | |
| | | training data | 5.7 | 14.2 | 237 | 0 | |
| | | cross-validation data | 5.1 | 13.5 | 246 | 0 | |
| | | testing data | 7.7 | 20.5 | 214 | 0 | |
| Shimen area: 271.25 km$^2$ | rainfall (mm) | whole data | 3.8 | 11.4 | 141.3 | 0 | January 1989–December 1999 |
| | | training data | 3.5 | 10.1 | 114.9 | 0 | |
| | | cross-validation data | 5.1 | 15.1 | 141.3 | 0 | |
| | | testing data | 3.8 | 11.8 | 116.8 | 0 | |
| | runoff (m$^3$) | whole data | 4.9 | 15.2 | 296 | 0 | |
| | | training data | 3.7 | 9.9 | 150 | 0 | |
| | | cross-validation data | 8.7 | 25.1 | 296 | 0 | |
| | | testing data | 5.5 | 17.9 | 172 | 0 | |

**Table 1.** *Cont.*

| Watershed and Datasets | | | Statistical Parameters | | | | Data Period |
|---|---|---|---|---|---|---|---|
| | | | μ | $S_x$ | $X_{max}$ | $X_{min}$ | |
| Tiantang area: 220 km$^2$ | rainfall (mm) | whole data | 3.7 | 12.1 | 193.4 | 0 | January 1973–December 1984 |
| | | training data | 3.6 | 11.6 | 175.0 | 0 | |
| | | cross-validation data | 3.7 | 11.4 | 151.7 | 0 | |
| | | testing data | 4.2 | 14.7 | 193.4 | 0 | |
| | runoff ($m^3$) | whole data | 6.1 | 18.4 | 535 | 0 | |
| | | training data | 5.6 | 16.5 | 400 | 0 | |
| | | cross-validation data | 5.6 | 16.5 | 378 | 0.3 | |
| | | testing data | 8.2 | 25.6 | 535 | 0.3 | |

### *3.2. Determination of Model Inputs*

The suitable predictive factors have an important impact on model performance. If the model input is only rainfall, it can be expressed as:

$$y_i = f(x_i, x_{i-1}, \cdots, x_{i-n+1}) \tag{9}$$

where $x$ is the rainfall series, $y$ is the runoff series, and $n$ is the number of antecedent rainfall components. In Pang *et al.*'s paper [16], only rainfall was selected as model input, so the SSA-ANN1 model, which only uses rainfall as model input, was developed. In order to ensure the comparability of model performance, the same $n$ values for the SSA-ANN1 model and the NLPM-ANN model were selected. From Pang *et al.*'s results of the NLPM-ANN model [16], the values of $n$ are 8, 6, 6, 8, 10, 8, 6, and 10 for Jiahe, Laoguanhe, Baohe, Mumahe, Nianyushan, Gaoguan, Shimen, and Tiantang, respectively.

As we know, the autocorrelation of the runoff series is strong and the impact of previous runoff on current runoff cannot be ignored, so the SSA-ANN2 model which uses rainfall and runoff as model inputs was developed in this paper. It can be expressed as:

$$y_i = f(y_{i-1}, \cdots, y_{i-m+1}, x_i, x_{i-1}, \cdots, x_{i-n+1}) \tag{10}$$

where $m$ is the number of previous runoff data. The values of $n$ for the SSA-ANN2 model are the same as the SSA-ANN1 model. In view of the convenience of operation and simplicity of computation, the autocorrelation function (ACF) is used to determine $m$. The smaller the values of correlation, the poorer the relationship is. Figure 3 plots the ACF values of the runoff series at the one-step prediction horizon. Then the number of model inputs can be taken with the values of 5, 5, 5, 3, 2, 3, 2, and 1 for Jiahe, Laoguanhe, Baohe, Mumahe, Nianyushan, Gaoguan, Shimen, and Tiantang, respectively. It can be seen that the number of previous daily runoff is obviously related with the watershed area.

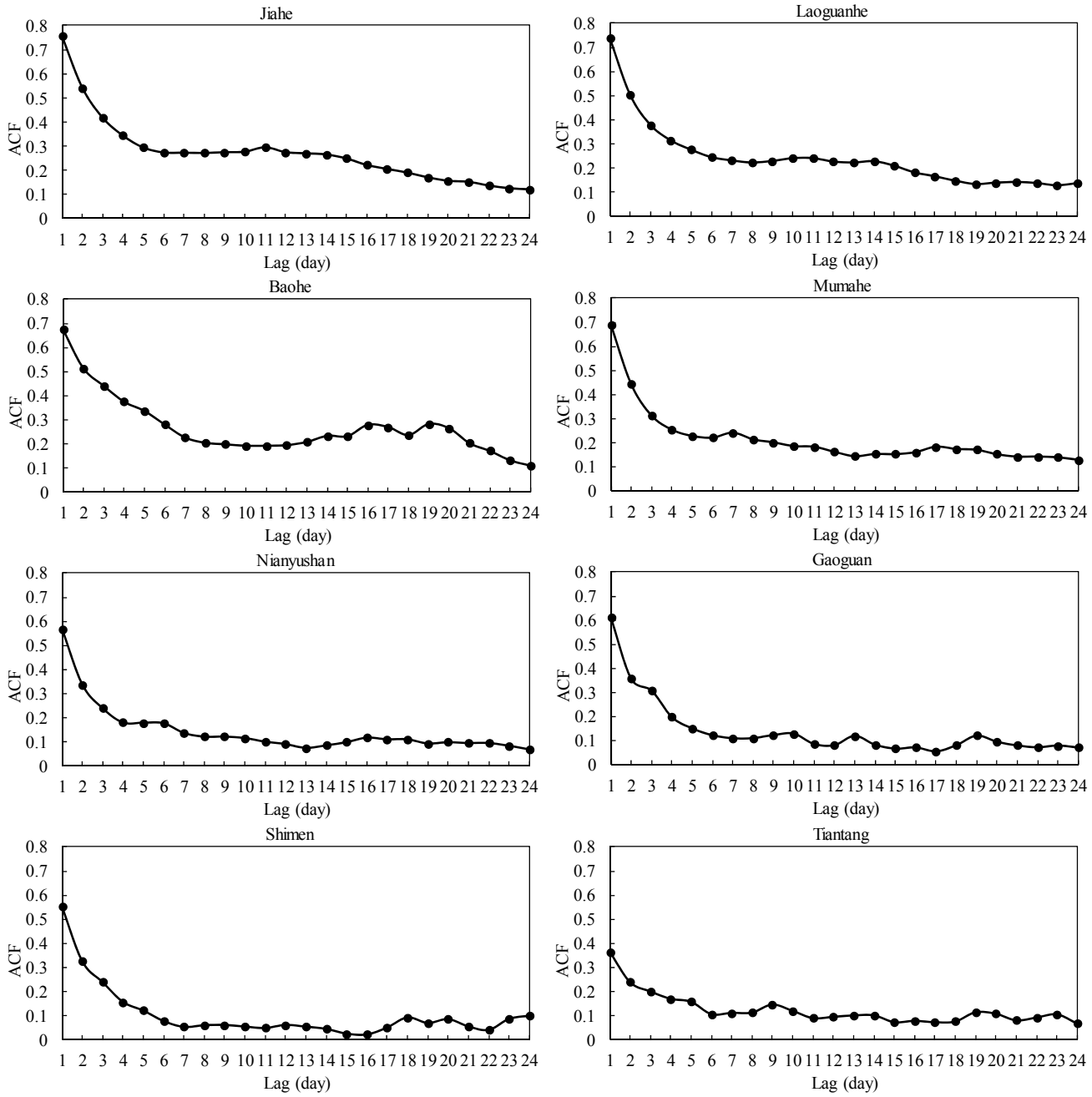


**Figure 3.** Autocorrelation function (ACF) values of runoff series for all watersheds.

### *3.3. Data Preprocessing*

According to the theory of the SSA, the decomposition procedure requires identifying the parameter $L$. The value of an appropriate $L$ should be able to clearly resolve different oscillations hidden in the original signal. In the current study, a small interval of [2,12] is examined to choose $L$ [28]. $L$ is considered as the target only if the singular spectrum can be markedly distinguished [33]. Figures 4 and 5 present the relation between singular values and singular numbers for the rainfall and runoff series, respectively, where the singular values associated with the appropriate $L$ are highlighted by the dotted solid line. It can be seen that $L$ is selected as 8, 8, 8, 8, 9, 10, 9, and 7 for the rainfall series, and $L$ is set as 9, 8, 9, 10, 9, 10, 9, and 7 for the runoff series in the Jiahe, Laoguanhe, Baohe, Mumahe, Nianyushan, Gaoguan, Shimen, and Tiantang watersheds, respectively.

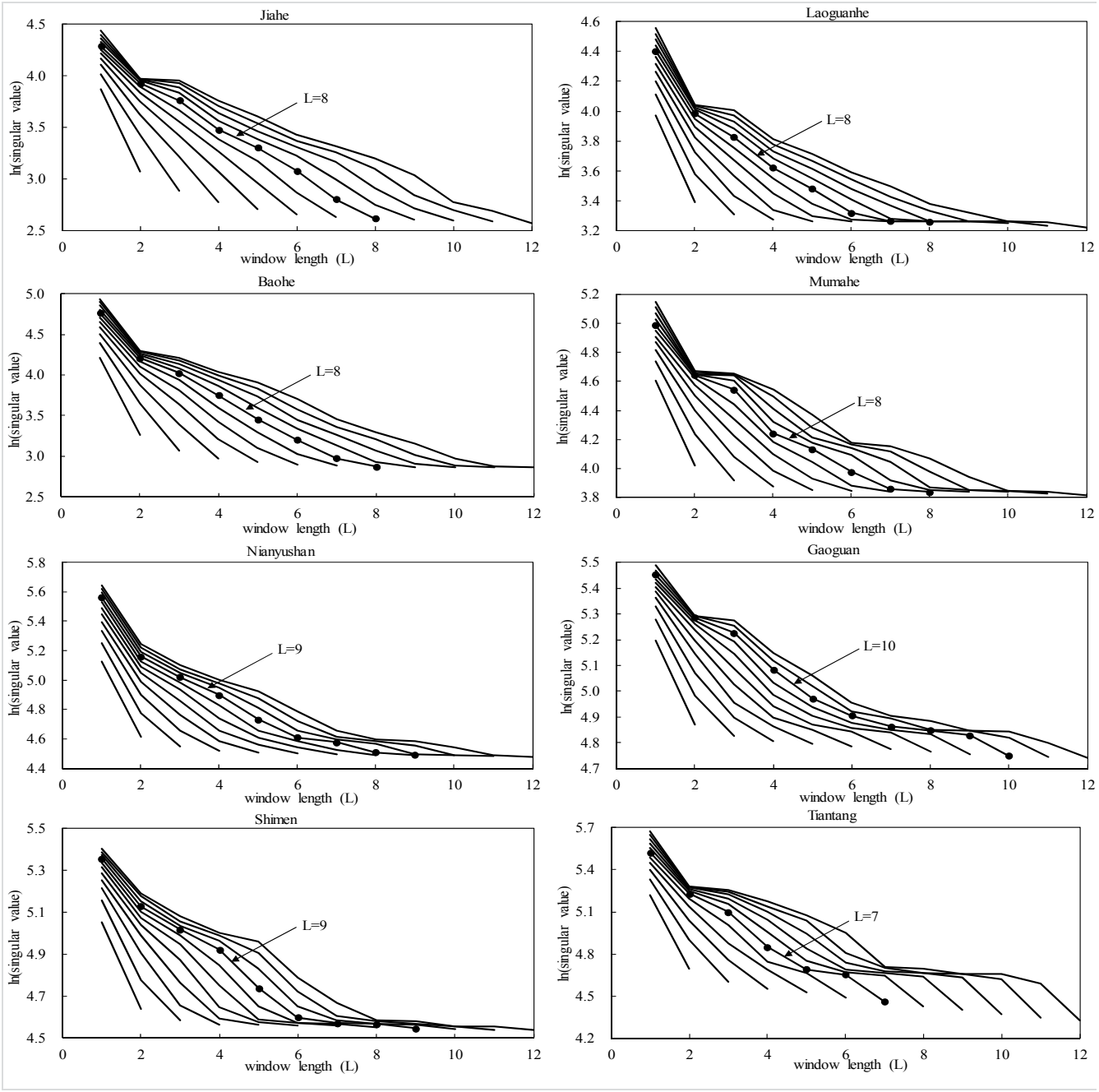


**Figure 4.** Singular values as a function of different window length *L* for rainfall series.

Once the original series is decomposed into L components, the subsequent task is to identify noise, choose the contributing components and reconstruct a new series as model inputs. This paper applied the cross-correlation function (CCF) to find the number of contributing components p ($\leqslant$L). From the perspective of linear correlation, the positive or negative CCF value indicates that the component makes a positive or negative contribution to the output of model. Table 2 listed all CCF values between each decomposed component and original series for all watersheds. Take Jiahe rainfall series as an example; the last four components have positive CCF values, which mean that they have positive correlation with the original series. So the number of contributing components p is equal to 4 and the sum of the last four components is reconstructed series. Meanwhile, the reconstructed series of other time series can be obtained by the same way.

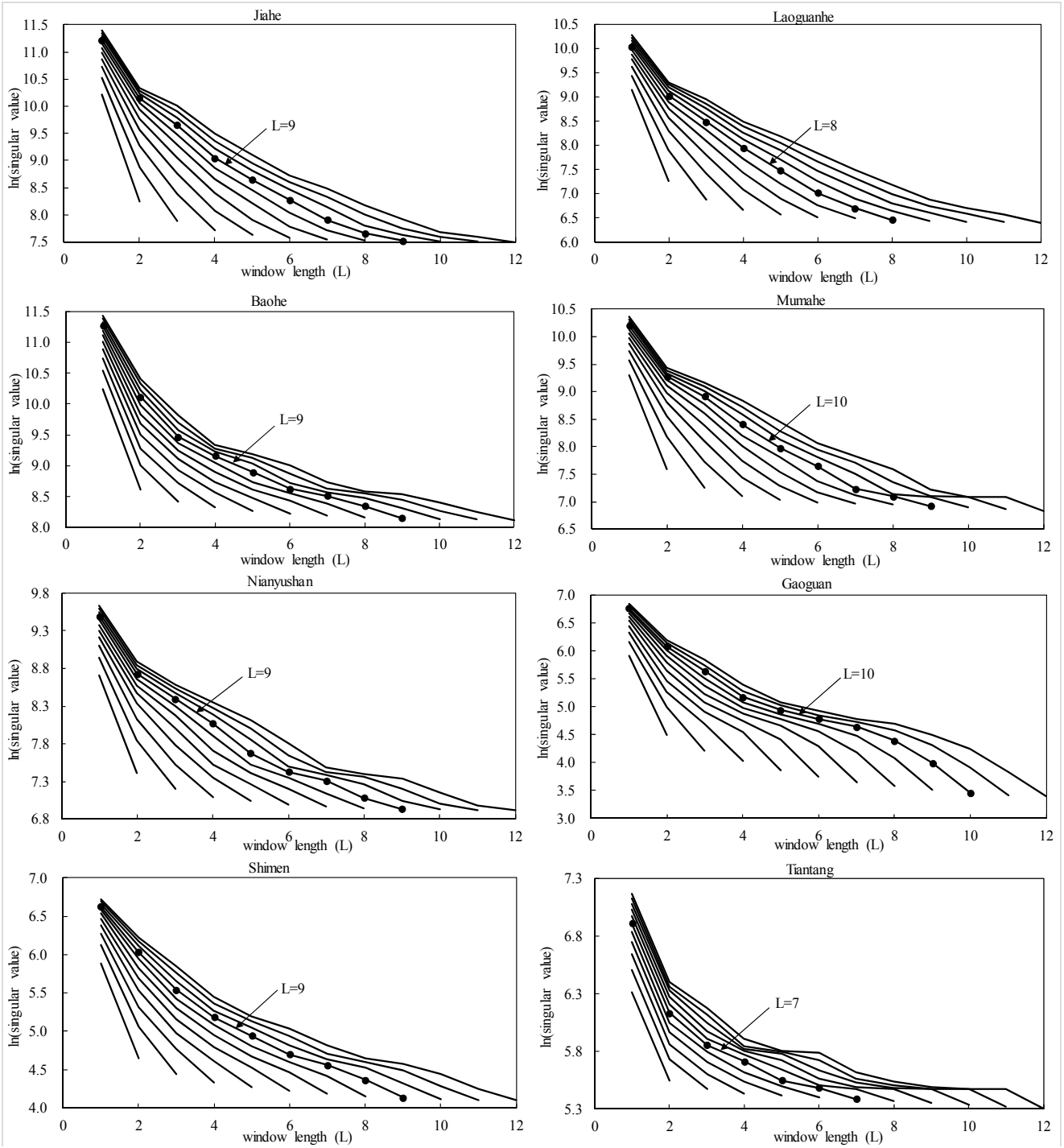


**Figure 5.** Singular values as a function of different window length *L* for runoff series.

**Table 2.** Cross-correlation function (CCF) values between each decomposed component and original series.

| Watershed | | Decomposed Components | | | | | | | | | | *L* | *p* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | | |
| Jiahe | rainfall | −0.26 | −0.27 | −0.19 | −0.05 | 0.13 | 0.36 | 0.50 | 0.55 | – | – | 8 | 4 |
| | runoff | −0.14 | −0.15 | −0.11 | −0.05 | 0.05 | 0.18 | 0.39 | 0.55 | 0.77 | – | 9 | 5 |
| Laoguanhe | rainfall | −0.36 | −0.33 | −0.24 | −0.06 | 0.12 | 0.33 | 0.47 | 0.53 | – | – | 8 | 4 |
| | runoff | −0.15 | −0.15 | −0.10 | 0.00 | 0.14 | 0.35 | 0.55 | 0.77 | – | – | 8 | 4 |

**Table 2.** *Cont.*

| Watershed | | Decomposed Components | | | | | | | | | | *L* | *p* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | | |
| Baohe | rainfall | −0.26 | −0.26 | −0.18 | −0.04 | 0.14 | 0.35 | 0.50 | 0.60 | – | – | 8 | 4 |
| | runoff | −0.18 | −0.20 | −0.16 | −0.08 | 0.04 | 0.16 | 0.33 | 0.54 | 0.76 | – | 9 | 5 |
| Mumahe | rainfall | −0.34 | −0.32 | −0.22 | −0.06 | 0.13 | 0.34 | 0.47 | 0.52 | – | – | 8 | 4 |
| | runoff | −0.15 | −0.18 | −0.14 | −0.09 | −0.01 | 0.11 | 0.25 | 0.41 | 0.56 | 0.71 | 10 | 5 |
| Nianyushan | rainfall | −0.33 | −0.33 | −0.26 | −0.13 | 0.02 | 0.19 | 0.35 | 0.47 | 0.51 | – | 9 | 5 |
| | runoff | −0.22 | −0.22 | −0.16 | −0.03 | 0.15 | 0.34 | 0.54 | 0.68 | – | – | 8 | 4 |
| Gaoguan | rainfall | −0.32 | −0.37 | −0.30 | −0.18 | −0.07 | 0.09 | 0.23 | 0.37 | 0.46 | 0.43 | 10 | 5 |
| | runoff | −0.14 | −0.19 | −0.17 | −0.12 | −0.03 | 0.09 | 0.23 | 0.42 | 0.58 | 0.67 | 10 | 5 |
| Shimen | rainfall | −0.34 | −0.34 | −0.32 | −0.28 | 0.01 | 0.19 | 0.35 | 0.47 | 0.48 | – | 9 | 5 |
| | runoff | −0.21 | −0.23 | −0.18 | −0.09 | 0.04 | 0.19 | 0.39 | 0.58 | 0.66 | – | 9 | 5 |
| Tiantang | rainfall | −0.32 | −0.34 | −0.19 | 0.03 | 0.28 | 0.46 | 0.53 | – | – | – | 7 | 4 |
| | runoff | −0.31 | −0.31 | −0.16 | 0.03 | 0.25 | 0.46 | 0.62 | – | – | – | 7 | 4 |

## 4. Results Analysis

Table 3 summarized the model performances for each watershed during calibration and testing periods. The ANN model is the benchmark in which the input is the original rainfall series without data preprocessing. It is shown that the model performance is improved significantly by data preprocessing techniques. During the testing period, the mean values of $R^2$ and WB of eight watersheds are 70.16% and 0.879 by ANN, and are increased to 75.86% and 1.155 by NLPM-ANN, and 80.62% and 1.04 by SSA-ANN1, respectively. In the Tiantang watershed, the performance of the NLPM-ANN and SSA-ANN1 models is improved significantly, so the $R^2$ value increased from 59.79% to 81.96% and 79.54%, respectively, during the testing period.

**Table 3.** Summary of model performances during calibration and testing periods.

| Watershed | | ANN | | NLPM-ANN | | SSA-ANN1 | | SSA-ANN2 | |
|---|---|---|---|---|---|---|---|---|---|
| | | $R^2$ (%) | WB | $R^2$ (%) | WB | $R^2$ (%) | WB | $R^2$ (%) | WB |
| Jiahe | calibration | 68.19 | 1.023 | 85.46 | 1.015` | 80.97 | 0.982 | 96.09 | 1.013 |
| | testing | 61.48 | 0.866 | 61.31 | 1.119 | 74.91 | 0.975 | 92.40 | 1.013 |
| Laoguanhe | calibration | 69.72 | 1.048 | 85.66 | 1.042 | 82.29 | 0.972 | 96.31 | 1.186 |
| | testing | 60.42 | 1.058 | 68.25 | 1.412 | 78.44 | 1.464 | 93.20 | 1.407 |
| Baohe | calibration | 64.75 | 0.975 | 70.93 | 1.039 | 88.50 | 1.029 | 94.01 | 1.006 |
| | testing | 68.62 | 0.667 | 69.38 | 0.893 | 74.03 | 0.927 | 94.31 | 0.956 |
| Mumahe | calibration | 80.64 | 0.950 | 90.18 | 1.050 | 87.86 | 0.976 | 95.08 | 1.019 |
| | testing | 80.17 | 0.913 | 85.6 | 1.410 | 92.41 | 1.108 | 94.71 | 1.053 |
| Nianyushan | calibration | 75.8 | 0.941 | 83.44 | 1.084 | 84.89 | 0.910 | 85.86 | 1.020 |
| | testing | 82.38 | 0.803 | 85.39 | 1.329 | 88.30 | 0.939 | 88.39 | 1.077 |
| Gaoguan | calibration | 66.16 | 1.035 | 77.6 | 1.045 | 80.17 | 1.002 | 93.24 | 1.005 |
| | testing | 76.38 | 0.957 | 77.97 | 0.894 | 80.43 | 0.840 | 89.85 | 0.962 |
| Shimen | calibration | 65.03 | 0.848 | 64.85 | 1.068 | 73.85 | 1.141 | 94.53 | 1.084 |
| | testing | 72 | 0.772 | 75.72 | 1.281 | 76.90 | 1.089 | 87.99 | 1.055 |
| Tiantang | calibration | 65.47 | 0.985 | 73.06 | 1.049 | 78.08 | 0.960 | 88.66 | 1.131 |
| | testing | 59.79 | 0.895 | 81.96 | 0.956 | 79.54 | 1.015 | 91.32 | 1.043 |
| Mean | calibration | 69.47 | 0.976 | 78.41 | 1.046 | 82.08 | 1.00 | 92.97 | 1.06 |
| | testing | 70.16 | 0.879 | 75.86 | 1.155 | 80.62 | 1.04 | 91.52 | 1.07 |

The mean values of $R^2$ and WB for the SSA-ANN1 model are 82.08% and 80.62%, and 1.0 and 1.04, during calibration and testing periods, respectively, which are much better than that of the NLPM-ANN model. It means that the reconstructed series obtained by SSA has a strong regularity and is easy to simulate. It also demonstrated that the impact of noise in hydrological time series on model performance is bigger than the seasonal hydrological behavior. Therefore, SSA is an effective way to improve runoff forecasting accuracy. The mean values of $R^2$ for the SSA-ANN2 model are 92.97% and 91.52%, which are much better than those of the SSA-ANN1 model. It is concluded that considering previous runoff as a model input can improve model efficiency greatly.

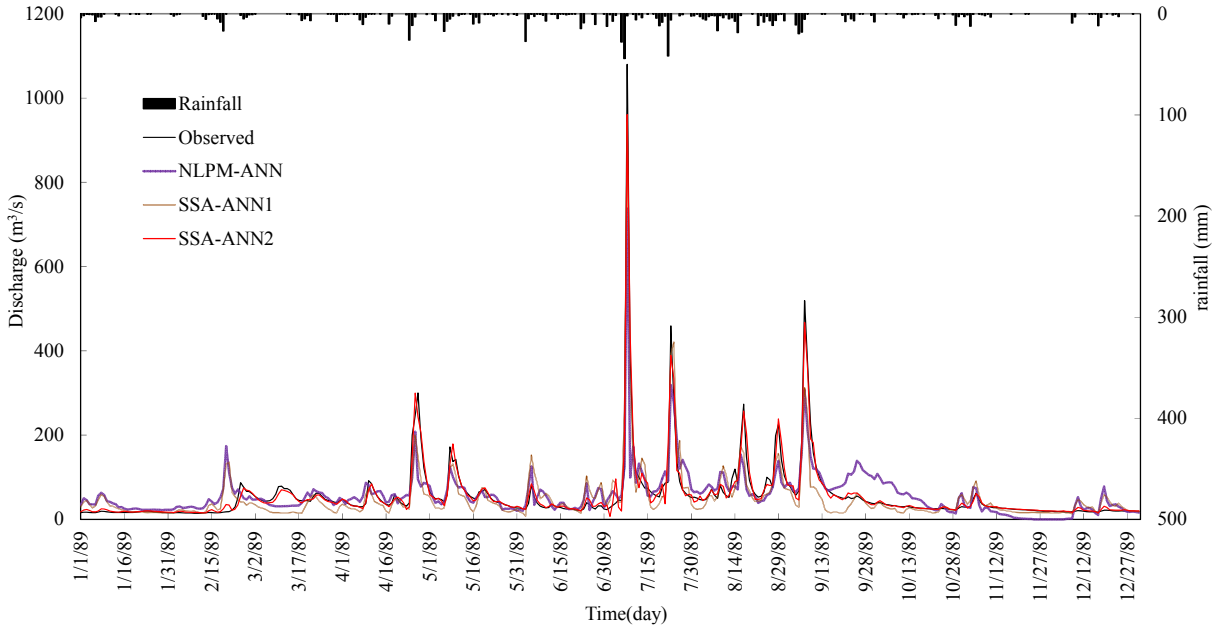


**Figure 6.** Observed and simulated runoff hydrographs by three models for Jiahe.

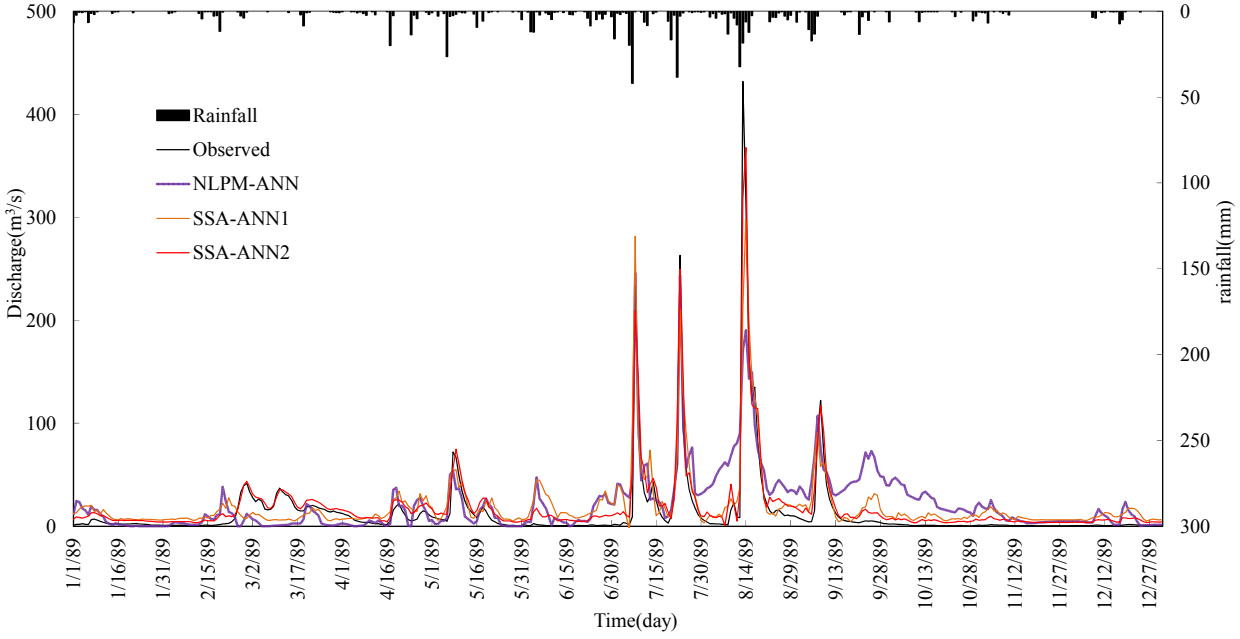


**Figure 7.** Observed and simulated runoff hydrographs by three models for Laoguanhe.

In order to compare the NLPM-ANN model, SSA-ANN1 model, and SSA-ANN2 model clearly and deeply, we selected one year during the testing period of four watersheds as an example, and the observed and simulated runoff hydrographs created by these three models for the Jiahe, Laoguanhe, Baohe, and Shimen watersheds are plotted in Figures 6–9, respectively. These figures show that the runoff hydrograph simulated by the SSA-ANN2 model is much closer to the observational one. The peak and minimum flows simulated by the SSA-ANN2 model are the best among these models. Therefore, the SSA-ANN2 model can predict daily runoff very well in practice.

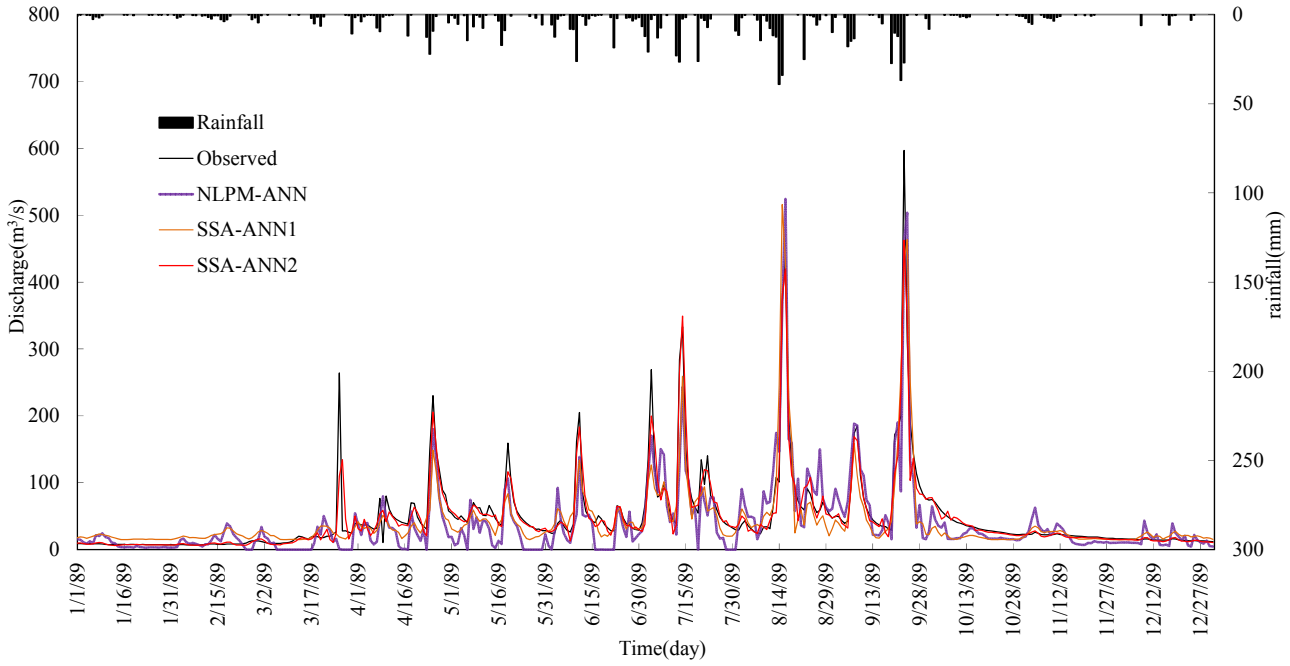


**Figure 8.** Observed and simulated runoff hydrographs by three models for Baohe.

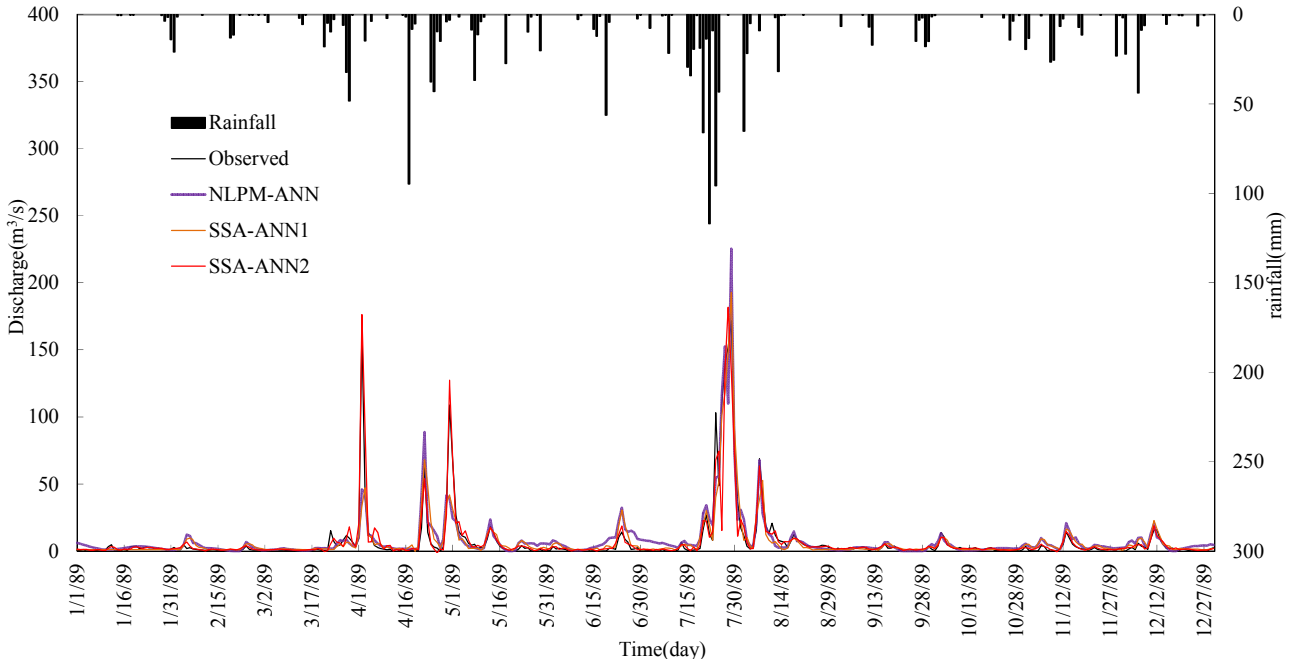


**Figure 9.** Observed and simulated runoff hydrographs by three models for Shimen.

## 5. Summary and Conclusions

The objective of this study is to investigate the approach of improving daily runoff forecasting in terms of data preprocessing and model input selection. The black-box model ANN is selected as the benchmark. Considering the subtraction of the seasonal means from the original series can remove the nonlinearity of the rainfall-runoff process, the NLPM method was used to preprocess model inputs. Considering the hydrological time series can be viewed as a combination of quasi-periodic signals contaminated by noises, the SSA method was used to filter the noise and choose reconstructed series as model inputs. These two data preprocessing techniques were compared and analyzed. Main findings and discussions were summarized as follows:

(1) The performance of the ANN model can be improved by data preprocessing techniques. SSA is more effective and it can improve the learning and training ability of the ANN type model significantly. Results also show that the impact of noise in hydrological time series on model performance is bigger than the seasonal hydrological behavior.
(2) Comparing the SSA-ANN1 model with the NLPM-ANN model, the mean values of $R^2$ and WB for the SSA-ANN1 model are 82.08% and 80.62%, and 1.0 and 1.04, during calibration and testing periods, respectively, which are much better than that of the NLPM-ANN model.
(3) The SSA-ANN2 model performs best for daily runoff forecasting for all selected watersheds. The effective way for increasing daily runoff forecasting accuracy is to preprocess data series by SSA and select both previous related rainfall and runoff as predictive factors.
(4) There are some limitations in this study. The method to select the contributing components relies on liner correlation analysis, which disregards the existence of nonlinearity in the hydrologic process. The sensitivities and uncertainties of model parameters are not analyzed. All of these will be the focus in our future research.

## Acknowledgment

This study is financially supported by the National Natural Science Foundation of China (NSFC 51190094 and 51379148). The authors would like to thank the editor and anonymous reviewers whose comments and suggestions help to improve the manuscript.

## Author Contributions

The SSA-ANN model was proposed and used to simulate the rainfall-runoff relationship. Results show that this model can improve daily runoff forecasting accuracy and it is worth of applying in practice.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

1. Baratti, R.; Cannas, B.; Fanni, A.; Pintus, M.; Sechi, G.M.; Toreno, N. River flow forecast for reservoir management through neural networks. *Neurocomputing* **2003**, *55*, 421–437.

2. Chang, F.J.; Chen, Y.C. A counterpropagation fuzzy-neural network modeling approach to real time streamflow prediction. *J. Hydrol*. **2001**, *245*, 153–164.
3. Nayak, P.C.; Sudheer, K.P.; Ramasastri, K.S. Fuzzy computing based rainfall–runoff model for real time flood forecasting. *Hydrol. Process.* **2005**, *19*, 955–968.
4. French, M.N.; Krajewski, W.F.; Cuykendall, R.R. Rainfall forecasting in space and time using a neural network. *J. Hydrol*. **1992**, *137*, 1–31.
5. Hsu, K.L.; Gupta, H.V.; Sorooshian, S. Artificial neural network modeling of the rainfall–runoff process. *Water Resour. Res*. **1995**, *31*, 2517–2530.
6. Sivakumar, B.; Liong, S.Y.; Liaw, C.Y. Evidence of chaotic behavior in Singapore rainfall. *J. Am. Water Resour. Assoc*. **1998**, *34*, 301–310.
7. Whigam, P.A.; Crapper, P.F. Modelling rainfall–runoff relationships using genetic programming. *Math. Comput. Model*. **2001**, *33*, 707–721.
8. Liong, S.Y.; Sivapragasm, C. Hood stage forecasting with SVM. *J. Am. Water Resour. Assoc*. **2002**, *38*, 173–186.
9. Govindaraju, R.S. Artificial neural networks in hydrology. I: Preliminary concepts. *J. Hydrol. Eng.* **2000**, *5*, 115–123.
10. Govindaraju, R.S. Artificial neural networks in hydrology. II: Hydrological applications. *J. Hydrol. Eng*. **2000**, *5*, 124–137.
11. Maier, H.R.; Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environ. Modell. Softw*. **2000**, *15*, 101–124.
12. Dawson, C.W.; Wilby, R.L. Hydrological modeling using artificial neural networks. *Progr. Phys. Geogr.* **2001**, *25*, 80–108.
13. Sudheer, K.P.; Gosain, A.K.; Ramasastri, K.S. A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrol. Process*. **2002**, *16*, 1325–1330.
14. Xiong, L.H.; O'Connor, K.M.; Guo, S.L. Comparison of three updating schemes using Artificial Neural Network in flow forecasting. *Hydrol. Earth Syst. Sci*. **2004**, *8*, 247–255.
15. Kumar, A.R.S.; Sudheer, K.P.; Jain, S.K.; Agarwal, P.K. Rainfall-runoff modelling using artificial neural networks: Comparison of network types. *Hydrol. Process*. **2005**, *19*, 1277–1291.
16. Pang, B.; Guo, S.L.; Xiong, L.H.; Li, C.Q. A nonlinear perturbation model based on artificial neural network. *J. Hydrol.* **2007**, *333*, 504–516.
17. Rezaeian, Z.M.; Amin, S.; Khalili, D.; Singh, V.P. Daily outflow prediction by multilayer perceptron with logistic sigmoid and tangent sigmoid activation functions. *Water Resour. Manag*. **2010**, *24*, 2673–2688.
18. Rezaeian, Z.M.; Stein, A.; Tabari, H.; Abghari, H.; Jalalkamali, N.; Hosseinipour, E.Z.; Singh, V.P. Assessment of a conceptual hydrological model and artificial neural networks for daily out-flows forecasting. *Int. J. Environ. Sci. Technol.* **2013**, *10*, 1181–1192.
19. Shamseldin, A.Y. Artificial neural network model for river flow forecasting in a developing country. *J. Hydroinform*. **2010**, *12*, 22–35.
20. Wu, J.S.; Han, J.; Annambhotla, S.; Bryant, S. Artificial neural networks for forecasting watershed runoff and stream flows. *J. Hydrol. Eng.* **2005**, *10*, 216–222.
21. Taormina, R.; Chau, K. Neural network river forecasting with multi-objective fully informed particle swarm optimization. *J. Hydroinform.* **2015**, *17*, 99–113.

22. Wu, C.L.; Chau, K.W.; Li, Y.S. Methods to improve neural network performance in daily flows prediction. *J. Hydrol*. **2009**, *372*, 80–93.
23. Nash, J.E.; Brasi, B.I. A hybrid model for flow forecasting on large catchments. *J. Hydrol*. **1983**, *65*, 125–137.
24. Liang, G.C.; Nash, J.E. Linear models for river flow routing on large catchments. *J. Hydrol*. **1988**, *103*, 157–188.
25. Sivapragasam, C.; Liong, S.Y.; Pasha, M.F.K. Rainfall and runoff forecasting with SSA-SVM approach. *J. Hydroinform*. **2001**, *3*, 141–152.
26. Marques, C.A.F.; Ferreira, J.; Rocha, A.; Castanheira, J.; Goncalves, P.; Vaz, N.; Dias, J.M. Singular spectral analysis and forecasting of hydrological time series. *Phys. Chem. Earth*. **2006**, *31*, 1172–1179.
27. Wang, W.S.; Jin, J.L; Li, Y.Q. Prediction of inflow at Three Gorges Dam in Yangtze River with wavelet network model. *Water Resour. Manage*. **2009**, *23*, 2791–2803.
28. Wang, Y.; Guo, S.L.; Chen, H.; Zhou, Y.L. Comparative study of monthly inflow prediction methods for the Three Gorges Reservoir. *Stoch. Environ. Res. Risk Assess*. **2014**, *28*, 555–570.
29. Wu, C.L.; Chau, K.W. Rainfall-runoff modeling using artificial neural network coupled with singular spectrum analysis. *J. Hydrol*. **2011**, *399*, 394–409.
30. Vautard, R.; Yiou, P.; Ghil, M. Singular-spectrum analysis: A toolkit for short, noisy and chaotic signals. *Physica. D*. **1992**, *58*, 95–126.
31. Golyandina, N.; Nekrutkin, V.; Zhigljavsky, A. *Analysis of time Series Structure: SSA and the Related Techniques*; CRC Press: Boca Raton, FL, USA, 2001.
32. Toth, E.; Brath, A.; Montanari, A. Comparison of short-term rainfall prediction models for real-time flood forecasting. *J. Hydrol*. **2000**, *239*, 132–147.
33. Wu, C.L.; Chau, K.W.; Li, Y.S. Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resour. Res*. **2009**, *45*, 2263–2289.