A hybrid model to simulate the annual runoff of Kaidu River in northwest China

Jianhua Xu [1], Yaning Chen [2], Ling Bai [1], Yiwen Xu [1]

1 The Research Center for East-West Cooperation in China, School of Geographic Sciences, East China Normal University, Shanghai 200241, China.

2 State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China.

Correspondence to: J. H. Xu (jhxu@geo.ecnu.edu.cn)

Abstract

The fluctuant and complicated hydrological processes can result in the uncertainty of runoff forecasting. Thus, it is necessary to apply the multi-method integrated modelling approaches to simulate runoff. Integrating the ensemble empirical mode decomposition (EEMD), the back propagation artificial neural network (BPANN) and the nonlinear regression equation, we put forward a hybrid model to simulate the annual runoff (AR) of the Kaidu River in northwest China. We also validate the simulated effects by using the coefficient of determination (R^2) and the Akaike information criterion (AIC) based on the observed data from 1960 to 2012 in Dashankou hydrological station. The average absolute and relative errors show the high simulation accuracy of the hybrid model. R^2 and AIC both illustrate that the hybrid model has a much better performance than the single BPANN. The hybrid model and integrated approach elicited by this study can be applied to simulate the annual runoff of the similar rivers in northwest China.

1 Introduction

The description of hydrological processes is the basis of hydrological modelling and simulation. Many models have been developed for describing hydrological processes over the past decades. From different perspectives, these hydrologic models can be classified as stochastic and deterministic models according to their mathematical property, or classified as conceptual and physically based models according to the physical processes involved in modelling, or classified as lump and distributed models according to the spatial description of the watershed process (Refsgaard, 1996; Moglen and Beighley, 2002).

Among the hydrologic models, distributed hydrological models are widely used. The soil-water-atmosphere-plant (SWAP) model has been intensively validated during the past two decades (van Dam et al., 1997; Gusev and Nasonova, 2003; Kroes et al., 2003; Gusev et al., 2011; Ma et al., 2011). Different versions of SWAP are validated against various observed hydrothermal characteristics. The validations are performed both for "point" experimental sites and for catchments and river basins with different areas (from 10^{-1} to 10^{5} km^{2}) on a long-term basis and under different environmental conditions (Nasonova and Gusev, 2007). The soil and water assessment tool (SWAT) model is a continuation of almost 30 years of modelling efforts conducted by the USDA Agricultural Research Service, and widely used in the world. A number of scientists have used SWAT model for simulating streamflow and related hydrologic analyses (Gan and Luo, 2013; Levesque et al, 2008; Liu et al., 2008, 2014; Luo et al., 2012; Shope et al., 2014; Lin et al., 2015; Yang and Musiake, 2003). According to the investigation by Gassman et al. (2007), there have been hundreds of published articles including SWAT applications, reviews of SWAT components, or other researches of SWAT in the past decades.

However, the application prerequisite of the distributed hydrological model is to successfully obtain a large number of parameters (such as temperature, precipitation, evapotranspiration, topography, land use, soil moisture, and vegetation coverage) at each grid cell (Yang et al., 2015). But for a large river basin with sparse meteorological and hydrological sites as well as lacking of observed data, it is difficult to obtain the large number of parameters mentioned above at each grid cell. Therefore, more studies are required to explore the hydrological processes from different perspectives by means of different methods.

In fact, hydrologists have used many methods for understanding the variation pattern of streamflow in the recent two decades. Various methods such as grey model (Yu et al., 2001; Trivedi and Singh, 2005), functional-coefficient time series model (Shao et al. 2009), wavelet analysis (Labat et al., 2000a, 2000b; Lane, 2007; Sang, 2012), genetic algorithm (Seibert, 2000), and artificial neural network (Hsu et al. 1995; Hu et al. 2008; Tokar and Johnson, 1999; Modarres, 2009) have been widely used for hydrologic analysis and streamflow simulation. Specially, hybrid models have been paid more attentions (Nourani et al., 2009; Zhao et al., 2009; Sahay and Srivastava 2014; Xu et al., 2014; Yarar 2014).

The water resource in northwest China which can be utilized is mainly from the streamflow of inland rivers. Hence the runoff variation of inland rivers has aroused more and more attention (Chen et al. 2009; Li et al. 2008; Wang et al. 2010; Xu et al., 2011). However, the runoff variation pattern of inland rivers in northwest China has not been clearly comprehended because of the complexity of the hydrological process (Xu et al. 2009, 2010). To understand the runoff variation pattern of inland rivers in northwest China, this study select the Kaidu River as a typical case of an inland river in northwest China, and integrated the ensemble empirical mode decomposition (EEMD), the back propagation artificial neural network (BPANN) and nonlinear regression equation to conduct a hybrid model for simulating annual runoff (AR).

2 Study basins and data

2.1 Study Area

The Kaidu River is situated at the north fringe of Yanqi Basin on the south slope of the Tianshan Mountains in Xinjiang, and is enclosed between latitudes 42°14′–43°21′N and longitudes 82°58′–86°05′E (**Fig. 1**). The river starts from the Hargat Valley and the Jacsta Valley in Sarming Mountain with a maximum altitude of 5000 m (the middle part of the

Tianshan Mountain), and ends in Bosten Lake, which is located in the Bohu County of Xinjiang. This lake is the largest lake in Xinjiang (also once the largest interior fresh water lake in China) and immediately starts another river known as the Kongque River. The catchment area of the Kaidu River above Dashankou, is 18,827 km^2, with an average elevation of 3100m (Chen et al., 2013).

Bayanbuluke wetland, which is in Kaidu River Basin, is the largest wetland of the Tianshan Mountain area. The large areas of grassland and marshes in Bayanbuluke wetland have provided favorable conditions for swan survival and reproduction. For this reason, it becomes the China's sole state-level swan nature reserve. The annual average temperature is only -4.6 ^{o}C and the extreme minimum temperature is -48.1 ^{o}C. The snow cover days are as many as 139.3 d and the largest average snow depth is 12 cm. As a unique high alpine cold climate and topography, it cultivates various alpine grassland and meadow ecosystems, having abundant aquatic plants, animals and good grassland resources. It is the birthplace and water saving place of the Kaidu River and plays a crucial role in regulating, reserving water and maintaining water balance. It also plays an utmost important role in protecting the Bosten Lake, its surrounding wetlands, and the ecological environment and green corridor of the lower reaches of Tarim River.

2.2 Data

The purpose of this study is to well understand the internal variation pattern by simulation method, so we used the annual runoff (AR) time series data from 1960 to 2012, which are observed at the Dashankou hydrological station. To analyze the correlation between the AR and regional climate change, the data of precipitation and temperature in the same period at the Bayinbuluke meteorological station were used. The two stations locate in the mountainous area (the source area of the river) where human activities are relatively rare. Therefore, it was assumed that the observed data reflect the natural conditions (Chen et al., 2013). In order to compare the hydrological cycle of Kaidu River and the El Niño meteorological phenomena, we also used the NINO3.4 index from NOAA Earth System Research Laboratory (http://www.esrl.noaa.gov/psd/data/climateindices/list/#Nina34).

3 Methods

To simulate the AR, we made a hybrid model by integrating ensemble empirical mode decomposition (EEMD), back propagation artificial neural network (BPANN) and regression equation. We firstly used the EEMD method to decompose the AR into four intrinsic mode functions (i.e. IMF1, IMF2, IMF3 and IMF4) and a trend (RES). Then we simulated IMFs by the back propagation artificial neural network (BPANN), and simulated RES (trend) by a nonlinear regression equation. Finally, the simulated values for AR are obtained from the summation of the simulated results of the trend (RES) and IMFs. The framework of the hybrid model is showed in **Fig. 2**.

3.1 EEMD method

The ensemble empirical mode decomposition (EEMD) is a new noise-assisted data analysis method based on the empirical mode decomposition (EMD), which defines the true IMF components as the mean of an ensemble of trials, each consisting of a signal plus white noise of finite amplitude (Wu and Huang, 2009).

The EMD has been developed for non-linear and non-stationary signal analysis, though only empirically. The EMD decompose a signal into several intrinsic mode functions (IMFs), then the frequencies of the IMFs are arranged in decreasing order (high to low), where the lowest frequency of the IMF components represents the overall trend of the original signal or the average of the time series data (Huang et al., 1998). Most importantly, each of these IMFs must satisfy two conditions: (1) the number of extrema and the number of zero crossings must be equal or differ at most by one; (2) at any point, the mean value of the envelope defined by the local maxima and local minima must be zero.

The EMD processing is as follows.

For the original signal $x(t)$, first we find out all the local maxima and minima, and then use cubic spline interpolation method to form the upper envelope $u_1(t)$ and the lower envelope $u_2(t)$; the local mean envelope $m_1(t)$ can be expressed as:

$$m_1(t) = \frac{1}{2}\left(u_1(t) + u_2(t)\right) \quad (1)$$

The first component $h_1(t)$ can be obtained by subtracting the local mean envelope $m_1(t)$ from the original signal $x(t)$, with the mathematical expression as follows:

$$h_1(t) = x(t) - m_1(t) \quad (2)$$

If $h_1(t)$ does not satisfy the IMF conditions, regard it as the new $x(t)$, and repeat the steps in formula (1) and (2) k times until $h_{1k}(t)$ is obtained as an IMF.

$$h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t) \quad (3)$$

Designate $C_1 = h_{1k}$, and select a stoppage criterion defined as follows:

$$SD = \sum_{t=0}^{T}\left[\frac{h_{1(k-1)}(t) - h_{1k}(t))}{h_{1(k-1)}(t)}\right]^2 \quad (4)$$

Here, the standard deviation (SD) is smaller than a predetermined value. If the above process is repeated too many times, the IMF will become a pure frequency modulation signal with constant amplitude in the actual operation, possibly resulting in loss of its actual meaning.

Once the first IMF component is determined, the residue $r_1(t)$ can also be obtained by separating C_1 from the rest of the data, i.e.

$$r_1(t) = x(t) - C_1 \quad (5)$$

By taking the residue $r_1(t)$ as new data and repeating steps (1)-(5), a series of IMFs, namely, $C_2, C_3, \cdots, C_n$ can be obtained.

The sifting process finally stops when the residue, $r_n(t)$, becomes a monotonic function or a function with only one extremum from which no more IMF can be extracted. Finally, the original signal $x(t)$ can be reconstructed by n IMFs (i.e. $C_i(t)$) and a residue $r_n(t)$ as follows:

$$x(t) = \sum_{i=1}^{n} C_i(t) + r_n(t) \qquad (6)$$

Although EMD has many merits, there is a shortcoming of mode mixing in EMD. To overcome the mode mixing problem, the EEMD has been developed for non-linear and non-stationary signal analysis (Wu and Huang 2009).

The principle of EEMD is that: adding white noise to the data, which distributes uniformly in the whole time–frequency space, the bits of signals of different scales can be automatically designed onto proper scales of reference established by the white noise.

The EEMD algorithm is straightforward and can be described as follows: first, add a white noise series to the original signal

$$x_i(t) = x(t) + n_i(t) \qquad (7)$$

Where $x_i(t)$ is the new signal after adding *i*th white noise to the original signal data $x(t)$, $n_i(t)$ is the white noise. Then, decompose the signal with added white noise into IMFs using EMD according to the steps of (1)-(5) equation, the corresponding IMF components $C_{ij}(t)$ and residue component $r_i(t)$ of the decompositions were obtained. Finally, adopt the means of the ensemble corresponding to the IMFs of the decompositions as the final result, namely

$$C_j(t) = \frac{1}{N}\sum_{i=1}^{N} C_{ij}(t) \qquad (8)$$

where $C_j(t)$ is the final *j*th IMF component, N is the number of white noise series, $C_{ij}(t)$ denotes the *j*th IMF from the added white noise trial.

Wu and Huang (2009) noted that the amplitude size of the added noise exerts little influence on the decomposition results on the condition that it is limited, is not vanishingly small or very large and can include all possibilities. Therefore, the application of the EEMD method does not rely on subjective involvement; it is an adaptive data analysis method.

The significance test in EEMD can be carried out by means of white noise ensemble disturbance, to get each IMF credibility (Wu and Huang 2009; Huang and Shen 2005).

In addition, to solve the overshooting and undershooting phenomenon of the impact of the boundary on the decomposition process, mirror-symmetric extension (Huang and Shen 2005; Xue et al. 2013) was used to address the EEMD decomposition boundary problem.

The residue of EEMD is a monotonic function that intrinsically presents the overall trend of a time series (Wu et al., 2007, 2009). Thus, the reconstruction of signal $x(t)$ based on EEMD can be obtained as following:

$$x(t) = IMF1 + IMF2 + \cdots + IMF_n + RES(trend) \qquad (9)$$

Where RES is the residue of EEMD, i.e. the trend of signal $x(t)$.

In this study, we decomposed the AR time series to a trend (RES) and four IMFs.

The MATLAB programs for EEMD are provided by RCADA, National Central University, which can be downloaded at the website (http://rcada.ncu.edu.tw/research1_clip_ex.htm).

3.2 BPANN

In the back propagation artificial neural network (BPANN), a number of smaller processing elements (PEs) are arranged in layers: an input layer, one or more hidden layers, and an output layer (Hsu et al., 1995). The input from each PE in the previous layer (x_i) is multiplied by a connection weight (w_{ji}). These connection weights are adjustable and may be likened to the coefficients in statistical models. At each PE, the weighted input signals are summed and a threshold value (θ_j) is added. This combined input (I_j) is then passed through a transfer function ($f(\cdot)$) to produce the output of the PE (y_j). The output of one PE provides the input to the PEs in the next layer. This process can be summarized (Maier and Dandy, 1998) in equation (13) and (14) and illustrated in **Fig. 3**.

$$I_j = \sum w_{ji} x_i + \theta_j \qquad (10)$$

$$y_i = f(I_j) \tag{11}$$

The error function of network at *t*th moment is defined as follows:

$$E(t) = \frac{1}{2}\sum_{j=1}^{q}[y_j(t) - d_j(t)]^2 \tag{12}$$

where $y_i(t)$ is the actual output and $d_i(t)$ is the desired output respectively corresponding to *i*th neuron at *t*th moment. When $E(t) \le \varepsilon$ (ε is a given error in advance), the network will stop training and the network model at this time is just what we need.

We used the BPANN with a four-tier structure to simulate IMF1, IMF2, IMF3 and IMF4 of the AR based on the results from the EEMD. The four-tier structure of the BPANN for each IMF is as follows (**Fig. 4**): an input layer with three variables, i.e. *(t-1)*-th, *(t-2)*-th and *(t-3)*-th value of the IMF; two hidden layers, in which the first layer contains three neurons and the second layer contains four neurons; an output layer with a variable, i.e. *t*th value of the IMF.

The transfer function from the input layer to two hidden layers is tansig, i.e. the hyperbolic tangent sigmoid transfer function (http://www.mathworks.com/help/nnet/ref/tansig.html). The transfer function from the hidden layers to the output layer is purelin, i.e. the linear function (http://www.mathworks.com/help/nnet/ref/purelin.html).

The purpose of our BPANN is to capture the relationship between a historical set of inputs and corresponding outputs. As mentioned above, this is achieved by repeatedly presenting examples of the input/output relationship to the model and adjusting the model coefficients (i.e. the connection weights) in an attempt to minimize an error function between the historical outputs and the outputs predicted by the model. This calibration process is generally referred to as ‘training’. The aim of the training procedure is to adjust the connection weights until the global minimum in the error surface has been reached. The network training process (Moghadassi et al., 2009) is summarized in **Fig. 5**.

The back-propagation process is commenced by presenting the first example of the desired relationship to the network. The input signal flows through the network, producing an output signal, which is a function of the values of the connection weights, the transfer function and the network geometry. The output signal produced is then compared with the desired (historical) output signal with the aid of an error (cost) function.

Because it can train any network as long as its weight, net input, and transfer functions have derivative functions (Kermani et al., 2005), we selected the Levenberg-Marquardt optimization, i.e. trainlm (http://www.mathworks.com/help/nnet/ref/trainlm.html) as a network training function in the computing environment of MATLAB.

3.3 Nonlinear regression

In order to simulate the trend of AR, we fitted a quadratic polynomial by using the nonlinear regression based on the results from the EEMD. We conducted the quadratic polynomial regression equation as follows:

$$y = at^2 + bt + c \qquad (13)$$

where the independent variable (t) is the time variable, and the dependent variable (y) represent the trend of AR, which is the RES obtained from the EEMD. The coefficients (a, b and c) are obtained by method of least squares (Lancaster and Šalkauskas, 1986).

3.4 Simulated effect test

In order to identify the uncertainty of the simulated results, the coefficient of determination was calculated as follows:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (14)$$

where R^2 is the coefficient of determination; $\hat{y}_i$ and y_i are the simulate value and actual data of AR respectively; $\bar{y}$ is the mean of $y_i (i = 1,2,...,n)$; $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares; $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares. The coefficient of determination is a measure of how well the simulate results represent the actual data. A bigger coefficient of determination indicates a higher certainty and lower uncertainty of the estimates (Xu, 2002).

To compare the goodness between our hybrid model and single BPANN, we also used the measure of Akaike information criterion (AIC) (Anderson et al., 2000). The formula of AIC is as follows:

$$AIC = 2k + n\ln(RSS / n) \qquad (15)$$

where k is the number of parameters estimated in the model; n is the number of samples; RSS is the same as in formula (14). A smaller AIC indicates a better model (Burnham and Anderson, 2002).

4 Results and discussion

4.1 Decomposition for AR

Fig. 6 reveals anomaly fluctuations of the AR time series in the Kaidu River during 1960 -2012. It is clear that the AR shows a strong nonlinear and non-stationary variation. Because of the nonlinear and non-stationary characteristics, it is difficult to show the change law of the AR time series.

To discover intrinsic modes in the signal of AR, we decomposed the AR time series by the EEMD method. For decomposing the AR time series, the ensemble number is 100, and the added noise has amplitude that is 0.2 times the standard deviation of the corresponding data, and four IMF components (IMF1-4) and a trend component (RES) were obtained. The decomposed results are showed in **Fig. 7**.

The significance test showed that IMF2, IMF3 and IMF4 reach above the 95% confidence level, while IMF1 reach above 90% confidence level. The variance contribution rate of IMF1, IMF2, IMF3, IMF4 and RES (trend) is 28.29%, 19.61%, 10.11%, 8.58% and 33.41% respectively. The summation of IMF1, IMF2, IMF3, IMF4 and RES represent of the reconstruction for AR time series, which is very highly correlative with its original data series. It can be seen that the reconstruction for AR series with the original data series is almost exactly the same (**Fig. 8**). This result illustrates that the decomposition of the AR time series by EEMD got a good prospective effect.

Each IMF component in **Fig. 7** has its own physical meaning, which reflects the inherent oscillation at a characteristic scale. The four IMF components (IMF1-4) reflect the fluctuation characteristics from high frequency to low frequency. IMF1 presents the highest frequency fluctuation and IMF4 shows the lowest frequency fluctuation. Whereas the fluctuation frequency of IMF2 is higher than that of IMF3 but lower than that of IMF1, and the fluctuation frequency of IMF3 is higher than that of IMF4 but lower than that of IMF2. The residual (RES) of EEMD is a monotonic function that presents the overall trend of the AR time series.

The multi-scale oscillations of runoff in the Kaidu River reflect not only the periodic changes of the climatic system under external forcing but also the non-linear feedback of the climatic system. To compare the hydrological cycle of Kaidu River and the El Niño meteorological phenomena, we also decomposed the NINO3.4 index data series in the same period by using the EEMD method. The results show that the four IMF components (IMF1-4) of the NINO3.4 index data series respectively display quasi-3-year, quasi-6-year, quasi-11-year and quasi-28-year periodic fluctuation (**Fig. 9**), whereas the four IMF components (IMF1-4) of the AR series in the Kaidu River respectively show quasi-3-year, quasi-6-year, quasi-11-year and quasi-27-year cyclic variation (**Fig. 7**). Although the two cycles are not complete same, they show some comparability. A study showed that there was a possible variability in droughts and wet spells over China on the multi-year or decadal scale when one strong El Niño event happened, but it does not mean that each El Niño event must cause a wet-dry change (Su and Wang, 2006). Similarly, the larger fluctuations of runoff in the Kaidu River on the multi-year or decadal scale possibly relate to strong El Niño events, but it does not mean that a big change of runoff certainly corresponds to a strong El Niño event. The possible reason is that the influencing factors include not only El Niño event but also other factors.

In fact, there are many other factors affecting the runoff, such as the varied topography, vegetation cover and construction of water conservancy project (Chen et al., 2013). Our previous study showed that the runoff process of the Kaidu River is closely related to the regional climate change (Xu et al., 2014; Bai et al., 2015). To compare the cycles between the runoff in Kaidu River and the regional climatic factors in the study period, we used the EEMD method to decompose the data series of annual precipitation (AP) and annual average temperature (AAT) into four IMF components (IMF1-4) and a trend. The results are similar to that of the AR: the AP and AAT on the whole show an upward trend, meanwhile, a) the AP presents quasi-3-year, quasi-6-year, quasi-11-year and quasi-27-year cycles, and b) the AAT displays quasi-3-year, quasi-6-year, quasi-13-year and quasi-27-year cycles. To further analyze the correlation between runoff and precipitation and temperature, we reconstructed inter-annual and inter-decadal precipitation and temperature variations, in which the inter-annual precipitation/temperature was obtained by IMF1 and IMF2, while the inter-decadal precipitation/temperature was obtained by IMF3 and IMF4. The results of multi-scale correlation analysis among annual runoff, annual precipitation and annual average temperature are shown in **Table 1**. Evidently, although there are differences in the length and strength of the periods among the precipitation, temperature and runoff changes, the positive correlation between runoff and precipitation, temperature are still significant except for inter-annual precipitation v.s. inter-decadal runoff, suggesting that the precipitation and temperature are two main causes of runoff variation. Furthermore, the higher correlation between runoff and climate factors is precipitation, followed by temperature at both the inter-annual and inter-decadal scales.

4.2 Simulation for IMFs

In order to capture the relationship between the historical data and real time output, we constructed the BPANN with a four-tier structure to simulate IMF1, IMF2, IMF3 and IMF4 of the AR based on the results from the EEMD. Using the MATLAB software (http://www.mathworks.com/products/matlab/), we selected the transfer function for input layer to the hidden layer and the hidden layer to the output layer as the tangent sigmoid function (tansig) and the linear function (purelin) respectively, and chose 'trainlm' as a training function to train the network. We set the learning rate as 0.01 and the training error accuracy as 0.01, and randomly extracted 70%, 15% and 15% of the data in the time series of each IMF as the training, testing, and validation samples. We finally obtained the optimized network for each IMF after thousands of training. Using the optimized networks, we obtained the simulated results for IMF1, IMF2, IMF3 and IMF4 respectively (**Fig. 10**).

Table 2 presents the R^2 and AIC value of the simulation model (the optimized networks) for each IMF. The big coefficient of determination (R^2) indicates that the simulated accuracy for each IMF is very high. The smaller AIC value means the better simulation effect, which indicates that the simulated effect of IMF4 is the best, and then follows are IMF3, IMF2, and IMF1 respectively.

4.3 Simulation for the trend

As above mentioned, the residue (RES) of EEMD presents the overall trend of the AR time series. Because it is a monotonic function, we can simulate the trend by a regression equation. Based on the data of RES from EEMD, we obtained the regression equation by using the method of least squares as the following quadratic polynomial:

$$y = 0.002t^2 - 7.7975t + 7632.6 \qquad (16)$$

where *t* is the time, which is measured by year; and *y* is the simulated value for the trend of the AR time series.

The coefficient of determination of formula (16) is as high as 0.9999. It is evident that the simulated effect of the RES (trend) is even better than that of IMF1, IMF2, IMF3 and IMF4 (also see **Table 2**). The simulated results for the trend of AR time series calculated by formula (16) are shown as **Fig. 11**.

4.4 Simulation for AR

Based on the idea and framework of the hybrid model mentioned previously in the methodology of this study, we can calculate the simulated value of AR at each year by summing the simulated value of IMF1, IMF2, IMF3, IMF4 and RES. By summing the simulated value of IMF1, IMF2, IMF3, IMF4 and RES at each year, we calculated the simulated value of AR for each year.

For calibration and validation purposes, we divided the whole data series into two periods, the calibration period, i.e. 1960-1989, and the validation period, i.e. 1990-2012. The calibration period is used for parameter estimation for the EEMD, BPANN and nonlinear regression equation. The validation period is used for validating the effectiveness of the hybrid model. The simulation results show the excellent performances of the model for both the calibration (1960-1989) and validation (1990-2012) periods with R^2 and AIC value (**Table 3**), which is highly acceptable. **Fig. 12** shows the observed data of AR and its simulated values by the hybrid model.

In order to compare and validate the simulated results from the hybrid model, we also simulated the AR series by using a single BPANN. **Table 3** shows the simulated effect comparisons between the hybrid model and the single BPANN. It can be seen that the coefficient of determination (R^2) of the hybrid model is as high as 0.9747, whereas that of the single BPANN is only 0.4037. Moreover, the AIC value of the hybrid model (6.2550) is far smaller than that of the single BPANN (171.7801). It is clear that both R^2 and AIC value indicate that the simulated effect of the hybrid model is much better than that of the single BPANN. Furthermore, the average absolute and relative error show the high simulation accuracy of the the hybrid model.

All the indices illustrate that the hybrid model is much better than the single BPANN. The reason is that the hybrid model concentrated on the advantages of both EEMD and BPANN. Where the EEMD precisely decompose the non-linear and non-stationary signal of AR into intrinsic mode functions (IMFs), and the BPANN well recognize and accurately simulate the IMFs. Because the non-linear and non-stationary AR signal contains many components and each component has its own intrinsic mode, a single BPANN can not accurately recognized and simulated the all change patterns in AR series. For this reason, this study used an integrated approach to conduct the hybrid model. In order to identify the pattern of each component in the non-linear and non-stationary AR signal, we firstly used the EEMD to decompose the AR series into four intrinsic mode functions (i.e. IMF1, IMF2, IMF3 and IMF4) and a trend (RES). Then we used the BPANN to accurately recongnize the pattern of each IMF by net learning and training, while using the nonliner regression to exactly simulate the pattern of the trend (RES). The above simulated results have already proved that our hybrid model is effective.

5 Conclusions

Integrating the ensemble empirical mode decomposition, the back propagation artificial neural network and the nonlinear regression equation, we conducted a hybrid model to simulate the annual runoff of the Kaidu River in northwest China. The main conclusions of this study are as follows:

(1) The comparison between simulated values of annual runoff and its original data shows the high simulation accuracy of the hybrid model. Both of the small average absolute and relative errors illustrate the high simulation accuracy of the hybrid model. The big R^2 and small AIC both indicate that the simulated effect of the hybrid model is much better than that of the single back propagation artificial neural network.

(2) This study elicited an integrated approach to simulate annul runoff of inland rivers, and the framework of the hybrid model conducted by this study can be applied to other inland rivers in northwest China.

Acknowledgements

This work is supported by the Open Foundation (No. G2014-02-07) of State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences.

References

Anderson, D. R., Burnham, K. P., and Thompson, W. L.: Null hypothesis testing: problems, prevalence, and an alternative, J. Wildlife Manage., 64, 912-923, 2000.

Bai, L., Chen, Z. S, Xu, J. H., and Li, W. H.: Multi-scale response of runoff to climate fluctuation in the headwater region of Kaidu River in Xinjiang of China, Theor. Appl. Climatol., DOI: 10.1007/s00704-015-1539-2, 2015.

Burnham, K. P. and Anderson, D. R.: Model Selection and Multimodel Inference: a practical information-theoretic approach, 2nd edn. Springer-Verlag, New York, 49–97, 2002.

Chen, Y. N., Xu, C. C., Hao, X. M., Li, W. H., Chen, Y. P., Zhu, C. G., and Ye, Z. X.: Fifty-year climate change and its effect on annual runoff in the Tarim River Basin, China, Quatern. Int., 208, 53-61, 2009.

Chen, Z. S., Chen, Y. N., and Li, B. F.: Quantifying the effects of climate variability and human activities on runoff for Kaidu River Basin in arid region of northwest China. Theor. Appl. Climatol., 111, 537-545, 2013.

Gan, R. and Luo, Y.: Using the nonlinear aquifer storage–discharge relationship to simulate the base flow of glacier-and snowmelt-dominated basins in northwest China, Hydrol. Earth Syst. Sci., 17, 3577-3586, doi:10.5194/hess-17-3577-2013, 2013.

Gassman, P. W., Reyes, M. R., Green, C. H., and Arnold, J. G.: The soil and water assessment tool: historical development, applications, and future research directions, T. ASABE, 50, 1211-1250, 2007.

Gusev, E. M., Nasonova, O. N., Dzhogan, L. Ya., and Kovalev, E. E.: Northern Dvina runoff simulation using land-surface model SWAP and global databases. Water Res., 38, 470–483, 2011.

Gusev, Ye. M. and Nasonova, O. N.: Modelling heat and water exchange in the boreal spruce forest by the land-surface model SWAP, J. Hydrol., 280, 162–191, 2003.

Hu, C., Hao, Y., Yeh, T. C. J., Pang, B., and Wu, Z.: Simulation of spring flows from a karst aquifer with an artificial neural network, Hydrol. Process., 22, 596-604, 2008.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., N.-C., Yen, Tung, C. C., and Liu, H. H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, P. Roy. Soc. Lond. A Mat., 454, 903-995, 1998.

Hsu, K., Gupta, H. V., and Sorooshian, S.: Artificial neural network modeling of the rainfall-runoff process, Water Resour. Res., 31, 2517-2530, 1995.

Huang, N. E., Shen, Z., and Long, S. R.: A new view of nonlinear water waves: the Hilbert Spectrum, Annu. Rev. Fluid Mech., 31, 417-457, 1999.

Kroes, J. G., Wesseling, J. G., and Van Dam, J. C.: Integrated modelling of the soil ‐ water ‐ atmosphere ‐ plant system using the model SWAP 2.0 an overview of theory and an application, Hydrol. Process., 14, 1993-2002, 2000.

Kermani, B. G., Schiffman, S. S., and Nagle, H. G.: Performance of the Levenberg–Marquardt neural network training method in electronic nose applications, Sensor. Actuat. B-Chem., 110, 13–22, 2005.

Labat, D., Ababou, R., and Mangin, A.: Rainfall-runoff relations for karstic springs. Part I: convolution and spectral analyses, J. Hydrol., 238, 123-148, 2000a.

Labat, D., Ababou, R., and Mangin, A.: Rainfall–runoff relations for karstic springs. Part II: continuous wavelet and discrete orthogonal multiresolution analyses, J. Hydrol., 238, 149-178, 2000b.

Lancaster, P. and Šalkauskas, K.: Curve and Surface Fitting: An Introduction, Academic

Press, London, 1986.

Lane, S. N.: Assessment of rainfall-runoff models based upon wavelet analysis, Hydrol. Process., 21, 586-607, 2007.

Levesque, E., Anctil, F., Van Griensven, A. N. N., and Beauchamp, N.: Evaluation of streamflow simulation by SWAT model for two small watersheds under snowmelt and rainfall, Hydrolog. Sci. J., 53, 961-976, 2008.

Li, Z. L., Xu, Z. X., Li, J. Y., and Li, Z. J.: Shift trend and step changes for runoff time series in the Shiyang River basin, northwest China, Hydrol. Process., 22, 4639-4646, 2008.

Lin, B., Chen, X., Yao, H., Chen, Y., Liu, M., Gao, L., and James, A.: Analyses of landuse change impacts on catchment runoff using different time indicators based on SWAT model, Ecol. Indic., 58, 55-63, 2015.

Liu, Y. B., Yang, W., and Wang, X.: Development of a SWAT extension module to simulate riparian wetland hydrologic processes at a watershed scale, Hydrol. Process., 22, 2901-2915, 2008.

Liu, Y. B., Yang, W., Yu, Z., Lung, I., Yarotski, J., Elliott, J., and Tiessen, K.: Assessing Effects of Small Dams on Stream Flow and Water Quality in an Agricultural Watershed, J. Hydrol. Eng., 19, 05014015, doi:10.1061/(ASCE)HE.1943-5584.0001005, 2014.

Luo,Y., Arnold, J., Allen, P., and Chen X.: Baseflow simulation using SWAT model in an inland river basin in Tianshan Mountains, Northwest China, Hydrol. Earth Syst. Sci., 16, 1259-1267, doi:10.5194/hess-16-1259-2012, 2012.

Ma, Y., Feng, S., Huo, Z., and Song, X.: Application of the SWAP model to simulate the field water cycle under deficit irrigation in Beijing, China, Math. Comput. Model., 54, 1044-1052, 2011.

Maier, H. R. and Dandy, G. C.: The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study, Environ. Modell. Softw., 13, 193–209, 1998.

Modarres, R.: Multi-criteria validation of artificial neural network rainfall-runoff modeling, Hydrol. Earth Syst. Sci., 13, 411-421, doi:10.5194/hess-13-411-2009, 2009.

Moglen, G. E. and Beighley, R. E.: Spatially explicit hydrologic modeling of land use change, Journal of the American Water Resources Association, 38, 241-253, 2002.

Moghadassi, A. R., Parvizian, F., Hosseini, S. M., and Fazlali, A. R.: A new approach for estimation of PVT properties of pure gases based on artificial neural network model. Braz. J. Chem. Eng., 26, 199–206, 2009.

Nasonova, O. N. and Gusev Y. M.: Can a land surface model simulate runoff with the same accuracy as a hydrological model?. Quantification and reduction of predictive uncertainty for sustainable water resources management, in: Proceedings of Symposium HS2004 at IUGG2007, Perugia, July 2007, IAHS Press, 258–265, 2007.

Nourani, V., Komasi, M., and Mano, A.: A multivariate ANN-wavelet approach for rainfall–runoff modeling, Water Resour. Manag., 23, 2877-2894, 2009.

Refsgaard, J.C.: Terminology, modelling protocol and classification of hydrologic model codes, in: Abbott, M. B. and Refsgaard, J. C. (Eds.), Distributed Hydrologic Modelling, Kluwer Academic Publishers, dordrecht, 41-54, 1996.

Sahay, R. R. and Srivastava, A.: Predicting monsoon floods in rivers embedding wavelet transform, genetic algorithm and neural network, Water Resour. Manag., 28, 301-317, 2014.

Sang, Y. F.: A practical guide to discrete wavelet decomposition of hydrologic time series, Water Resour. Manag., 26, 3345-3365, 2012.

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrol. Earth Syst. Sci., 4, 215-224, doi:10.5194/hess-4-215-2000, 2000.

Shope, C. L., Maharjan, G. R., Tenhunen, J., Seo, B., Kim, K., Riley, J., Arnhold, S., Koellner, T., Ok, YS., Peiffer, S., Kim, B., Park, J. -H., and Huwe, B.: Using the SWAT model to improve process descriptions and define hydrologic partitioning in South Korea, Hydrol. Earth Syst. Sci., 18, 539-557, doi:10.5194/hess-18-539-2014, 2014.

Su, M. F. and Wang, H. J.: Relationship and its instability of ENSO Chinese variations in droughts and wet spells, Sci China Ser D-Earth Sci., 50, 145-152, 2007.

Tokar, A. S. and Johnson, P. A.: Rainfall-runoff modeling using artificial neural networks, J. Hydrol. Eng., 4, 232-239, 1999.

Trivedi, H. V. and Singh, J. K.: Application of grey system theory in the development of a runoff prediction model, Biosyst. Eng., 92(4), 521-526, 2005.

van Dam, J. C., Huygen, J., Wesseling, J. G., Feddes, R. A., Kabat, P., van Walsum, P. E. V., Groenendijk, P., and van Diepen, C. A.: Theory of SWAP version 2.0: simulation of water flow, solute transport and plant growth in the Soil-Water-Atmosphere-Plant environment. Technical Document 45, DLO Winand Staring Centre, Report 71, Department Water Resources, Wageningen Agriculture University, Wageningen, 1997.

Wang, J., Li, H., and Hao, X.: Responses of snowmelt runoff to climatic change in an inland river basin, Northwestern China, over the past 50 years, Hydrol. Earth Syst. Sci., 14(10): 1979–1987, doi:10.5194/hess-14-1979-2010, 2010.

Wu, Z. H. and Huang, N. E.: A study of the characteristics of white noise using the empirical mode decomposition method, P. Roy. Soc. Lond. A Mat., 460, 1597-1611, 2004.

Wu, Z. H., Huang, N. E., Long, S. R., and Peng, C. K.: On the trend, detrending, and variability of nonlinear and nonstationary time series, P. Natl. Acad. Sci. USA, 104, 14889-14894, 2007.

Wu, Z. H. and Huang, N. E.: Ensemble empirical mode decomposition: A noise-assisted data analysis method, Advances in Adaptive Data Analysis, 01, 1-41, 2009.

Wu, Z. H., Huang, N. E., Wallace, J. M., Smoliak, B. V., and Chen, X. Y.: On the time-varying trend in global-mean surface temperature, Clim. Dynam., 37, 759-773, 2011.

Xu, J. H.: Mathematical methods in contemporary geography, Higher Education Press, Beijing, 37–105, 2002.

Xu, J. H., Chen, Y. N., Li, W. H., Ji, M. H., and Dong, S.: The complex nonlinear systems with fractal as well as chaotic dynamics of annual runoff processes in the three headwaters of the Tarim Rive, J. Geogr. Sci., 19, 25-35, 2009.

Xu, J. H., Li, W. H., Ji, M. H., Lu, F., and Dong, S.: A comprehensive approach to characterization of the nonlinearity of runoff in the headwaters of the Tarim River, western China, Hydrol. Process., 24, 136-146, 2010.

Xu, J. H., Chen, Y. N., Li W. H., Yang, Y., and Hong, Y. L.: An integrated statistical approach to identify the nonlinear trend of runoff in the Hotan River and its relation with climatic factors, Stoch. Env. Res. Risk A., 25, 223–233, 2011

Xu, J. H., Chen, Y. N., Li, W. H., Nie, Q., Song, C. N., and Wei, C. M.: Integrating wavelet analysis and BPANN to simulate the annual runoff with regional climate change: a case study of Yarkand River, Northwest China, Water Resour. Manag., 28, 2523-2537, 2014.

Yang, D. W. and Musiake, K.: A continental scale hydrological model using the distributed approach and its application to Asia, Hydrol. Process., 17, 2855-2869, 2003.

Yang, D. W., Gao, B., Jiao, Y., Lei, H. M., Zhang, Y. L., Yang, H. B., and Cong, Z. T.: A distributed scheme developed for eco-hydrological modeling in the upper Heihe River, Sci China-Earth Sci., 58, 36-45, 2015.

Yarar, A.: A hybrid wavelet and neuro-fuzzy model for forecasting the monthly streamflow data, Water Resour. Manag., 28, 553-565, 2014.

Yu, P. S., Chen, C. J., Chen, S. J., and Lin, S. C.: Application of grey model toward runoff forecasting, J. Am. Water Resour. As., 37, 151–166, 2001.

Zhao, Q., Liu, Z., Ye, B., Qin, Y., Wei, Z., and Fang, S: A snowmelt runoff forecasting model coupling WRF and DHSVM, Hydrol. Earth Syst. Sci., 13, 1897-1906, doi:10.5194/hess-13-1897-2009, 2009.

Table 1 Correlations between runoff and climate factors

Time scale	Precipitation vs. runoff	Temperature vs. runoff
Inter-annual scale	0.666**	0.416**
Inter-annual v.s. inter-decadal scale	0.205	0.441**
Inter-decadal v.s. inter-annual scale	0.279*	0.438**
Inter-decadal scale	0.822**	0.617**

Note: **correlation is significant at the 0.01 level (2-tailed); *correlation is significant at the 0.05 level (2-tailed).

Table 2 R^2 and AIC value of simulation models for the IMFs and trend of AR

IMFS	R^2	AIC
IMF1	0.9107	0.5789
IMF2	0.9619	-54.9342
IMF3	0.9859	-105.9041
IMF4	0.9980	-204.2977
Trend	0.9999	-405.1425

Table 3 Comparison of simulated effect between the hybrid model and the single BPANN

	Hybrid model	Single BPANN
R^2	0.9747	0.4037
AIC	6.2550	171.7801
Average absolute error ($10^8 m^3$)	0.9970	3.5477
Average relative error (%)	2.9107	10.1079

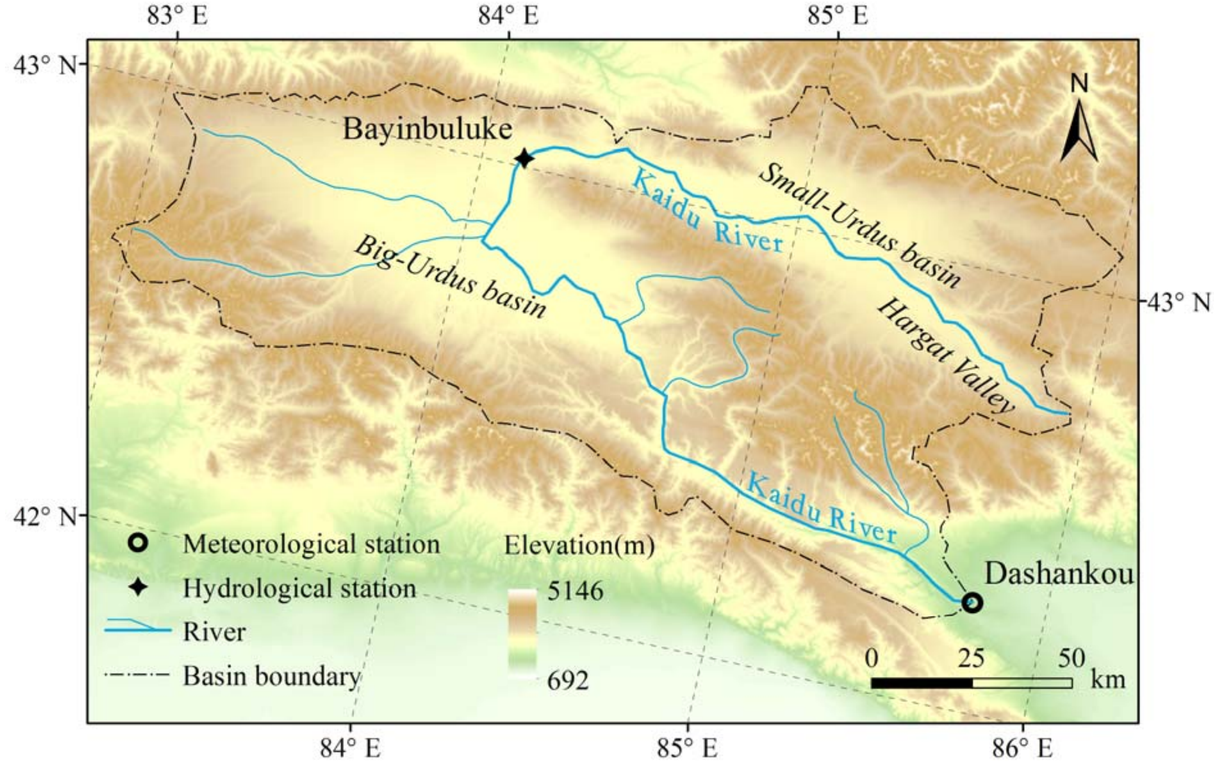

Figure 1 Location of the Kaidu River, northwest China

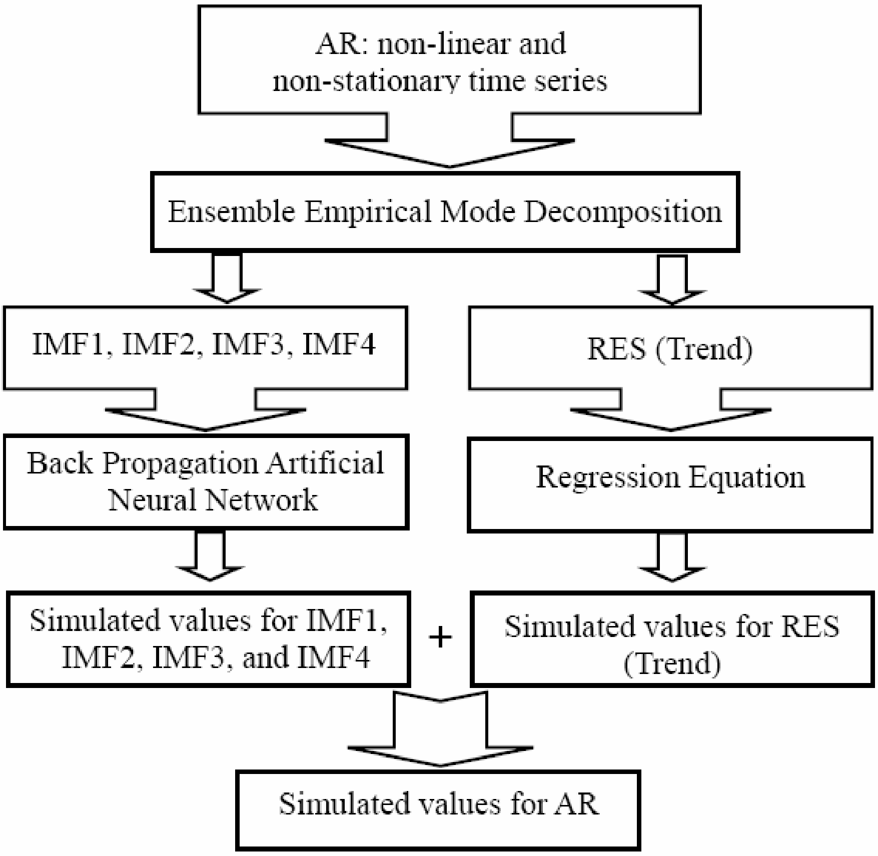
AR: non-linear and non-stationary time series
Ensemble Empirical Mode Decomposition
IMF1, IMF2, IMF3, IMF4
RES (Trend)
Back Propagation Artificial Neural Network
Regression Equation
Simulated values for IMF1, IMF2, IMF3, and IMF4
+
Simulated values for RES (Trend)
Simulated values for AR

Figure 2 The framework of the hybrid model to simulate the annual runoff

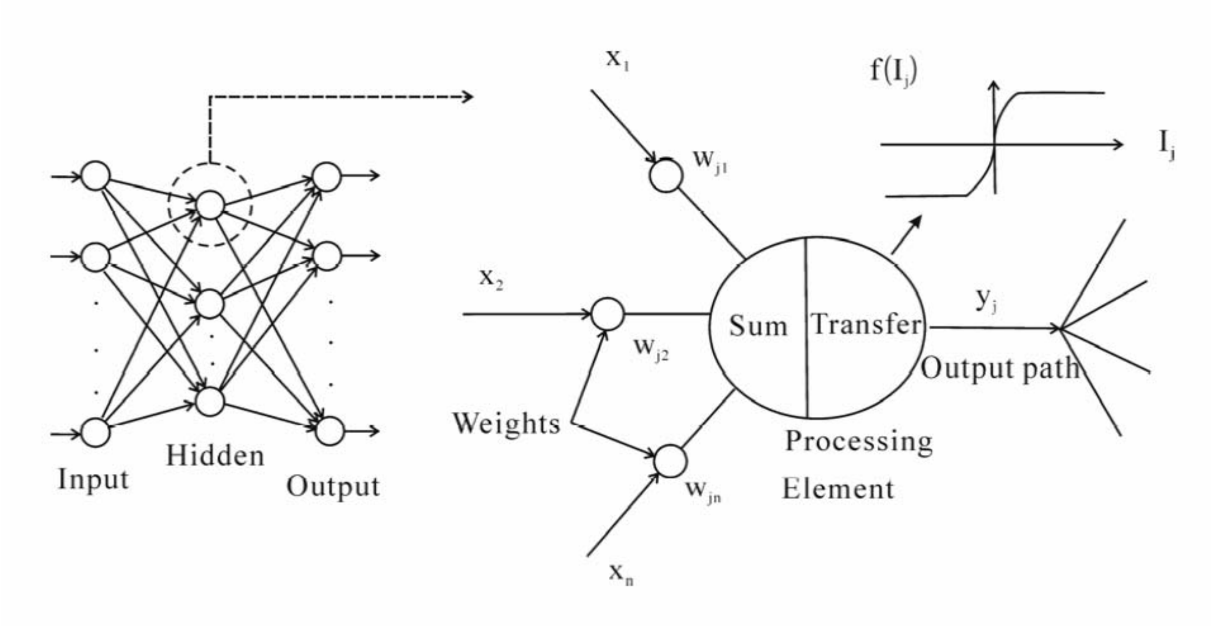

Figure 3 The back-propagation artificial neural network

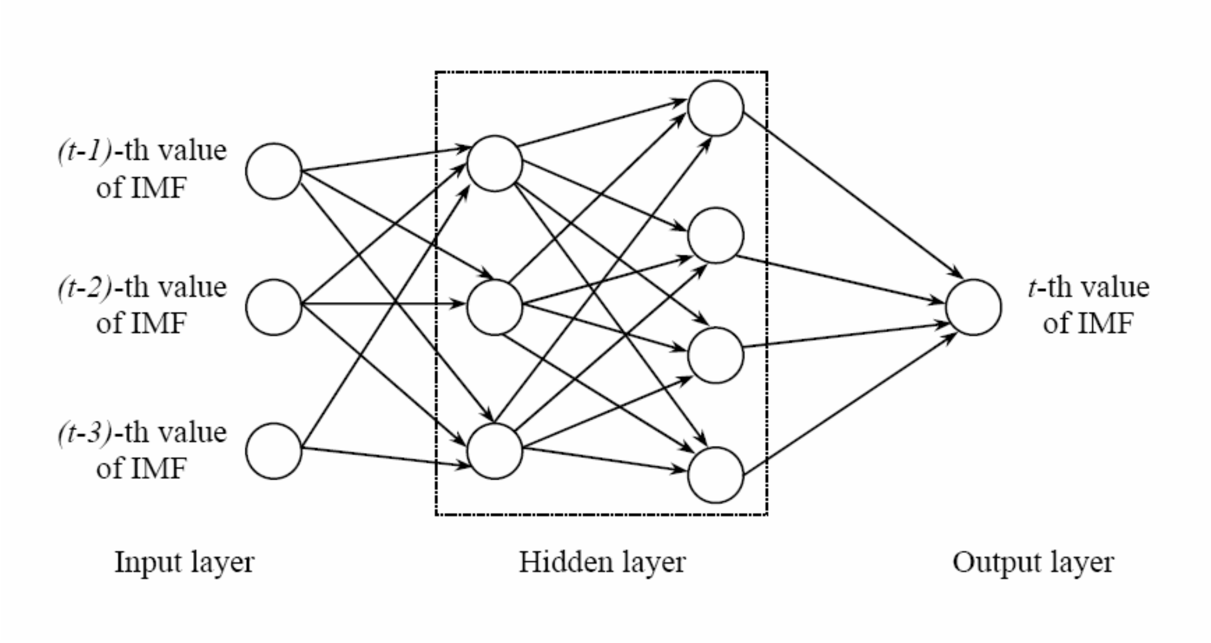

Figure 4 Four-tier structure BPANN to simulate the IMFs of AR

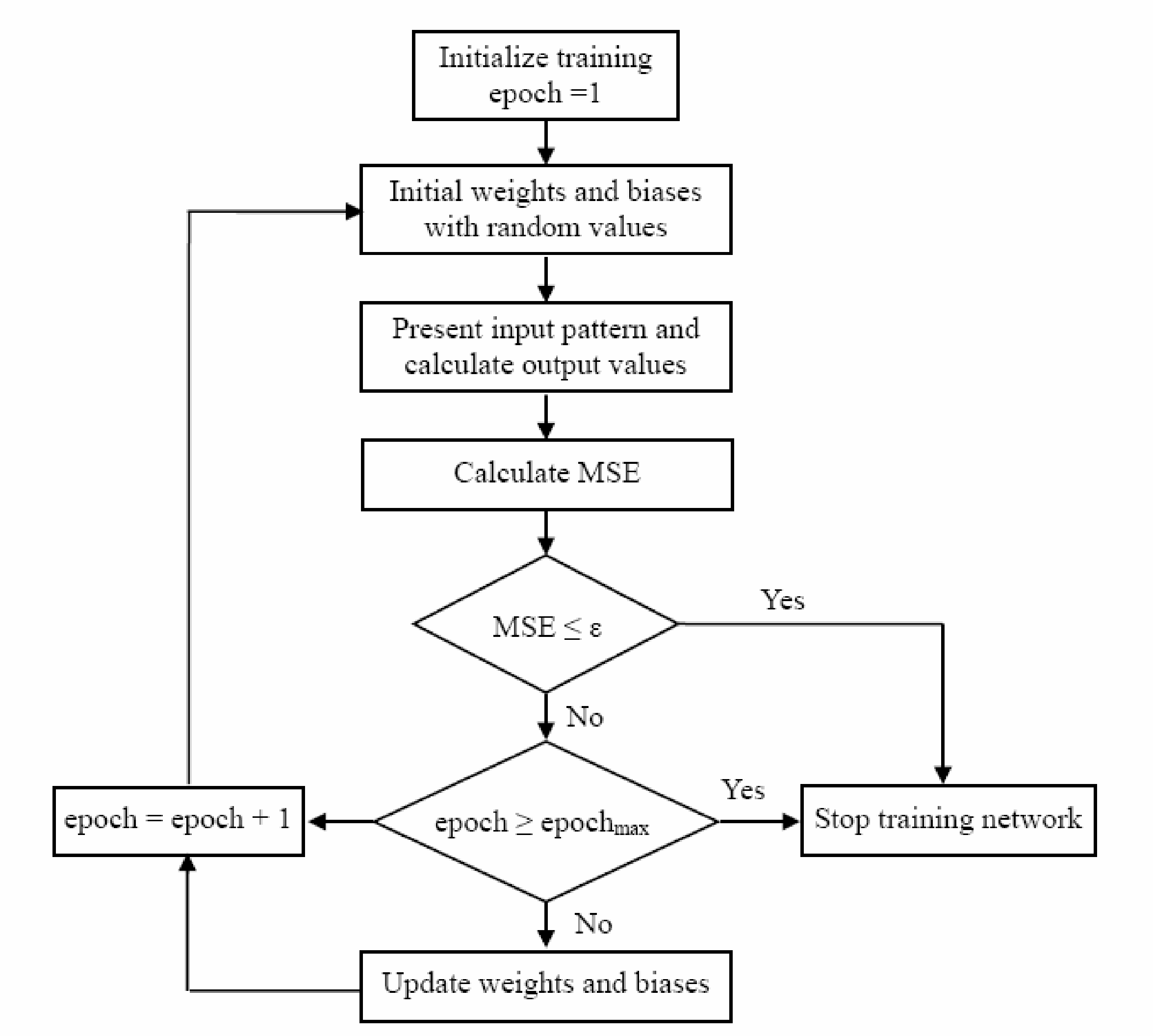
Initialize training epoch =1
Initial weights and biases with random values
Present input pattern and calculate output values
Calculate MSE
MSE ≤ ε
Yes
No
epoch ≥ epoch_max
Yes
Stop training network
epoch = epoch + 1
No
Update weights and biases

Figure 5 Back-propagation training process

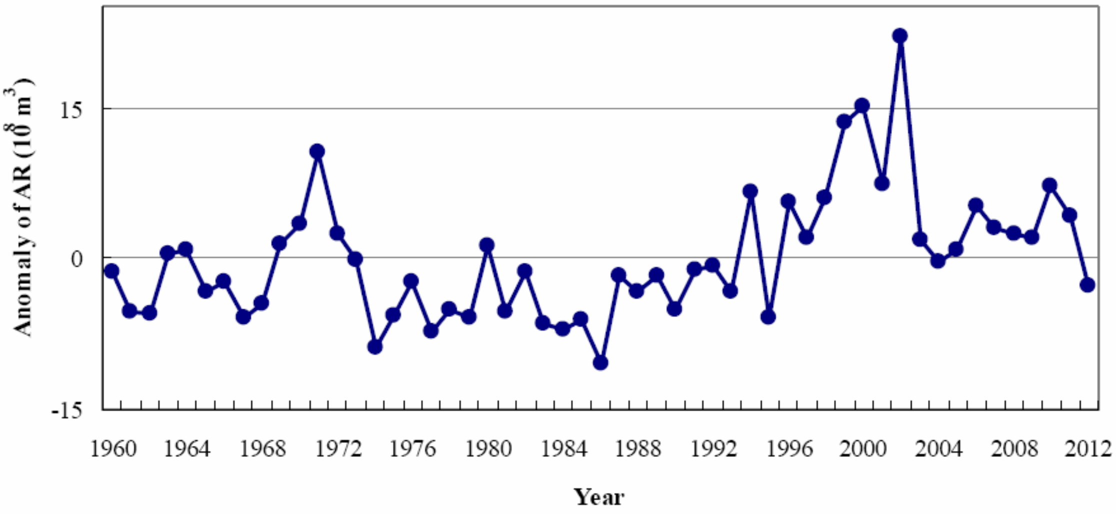

Figure 6 Annual runoff anomalies in the Kaidu River during 1960 - 2012

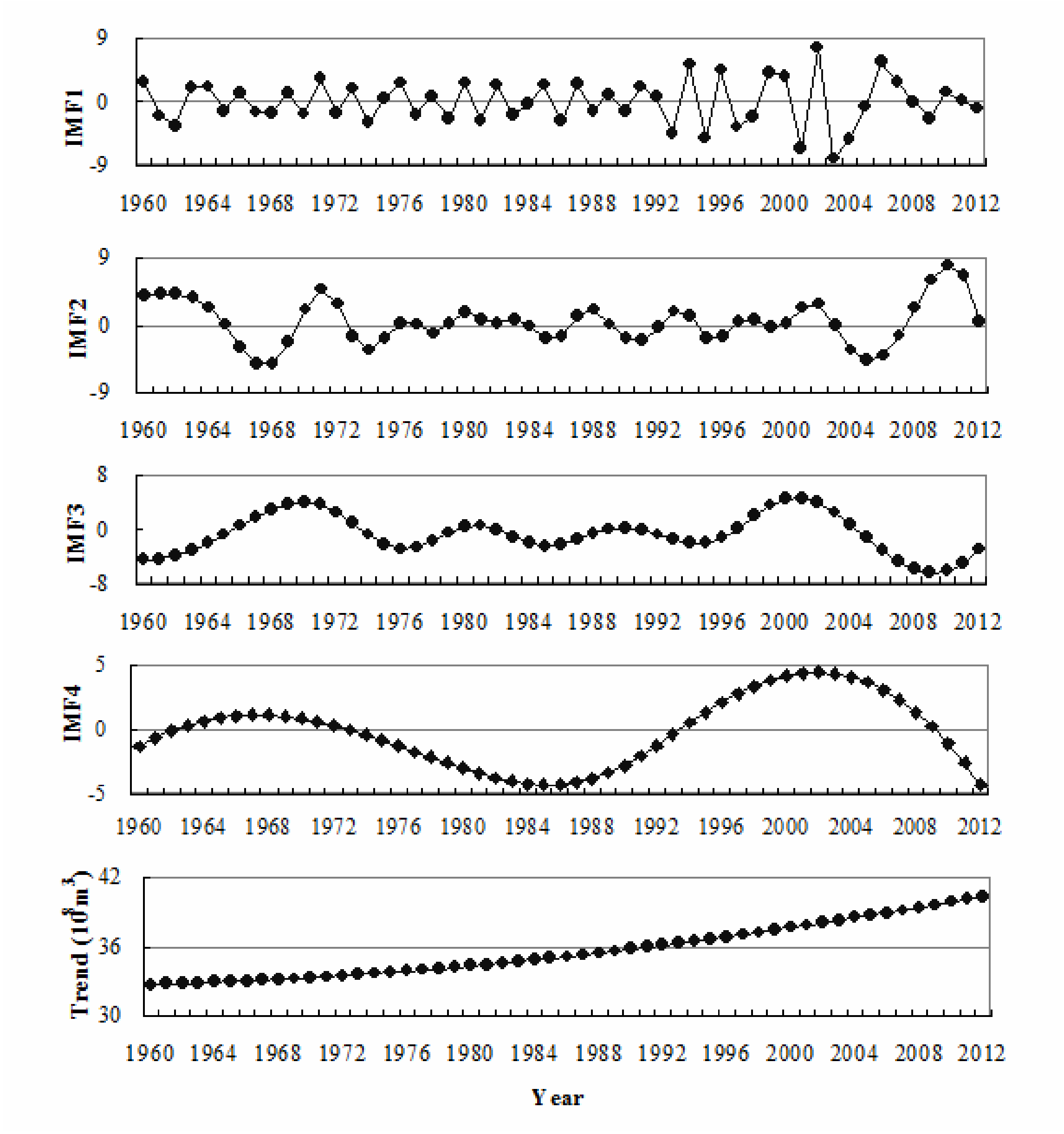

Figure 7 The EEMD results for the time series of AR in the Kaidu River

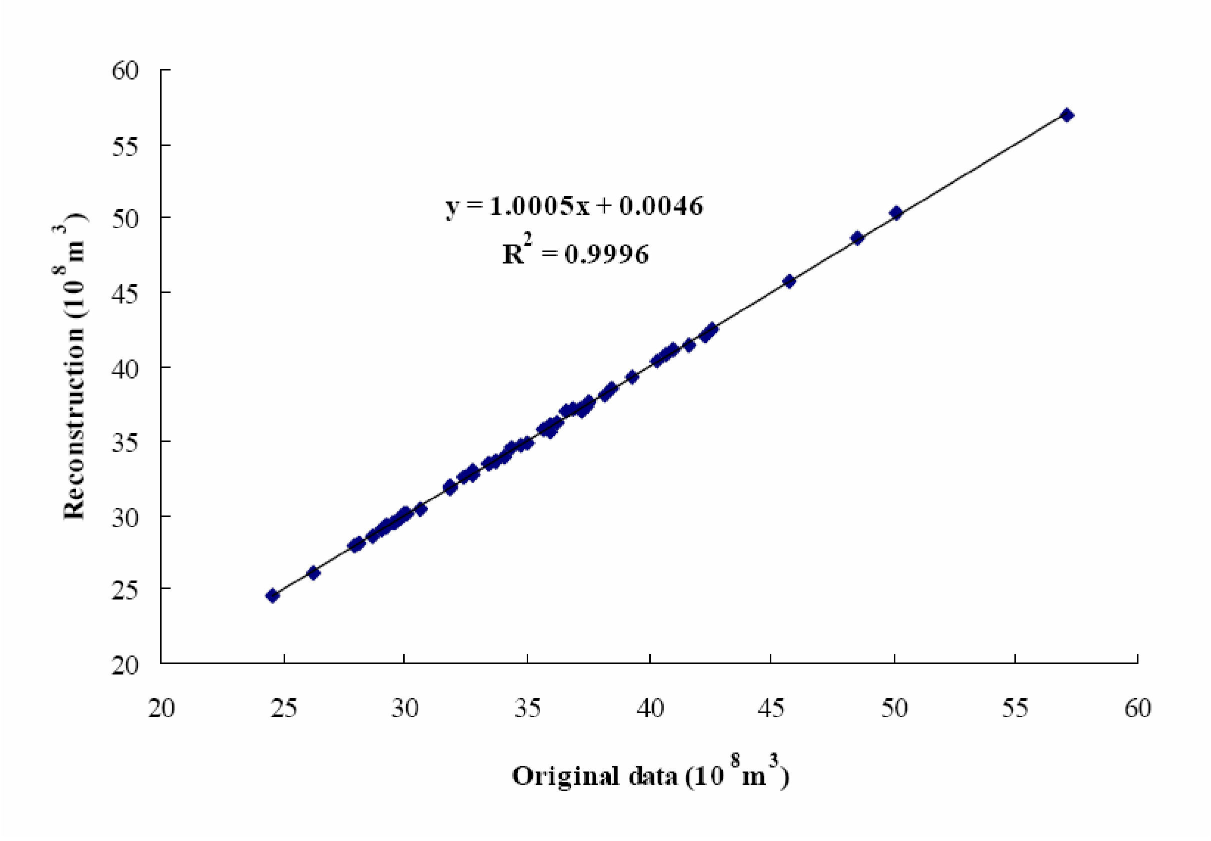

Figure 8 The correlation between the reconstruction of AR time series based on EEMD and its original data

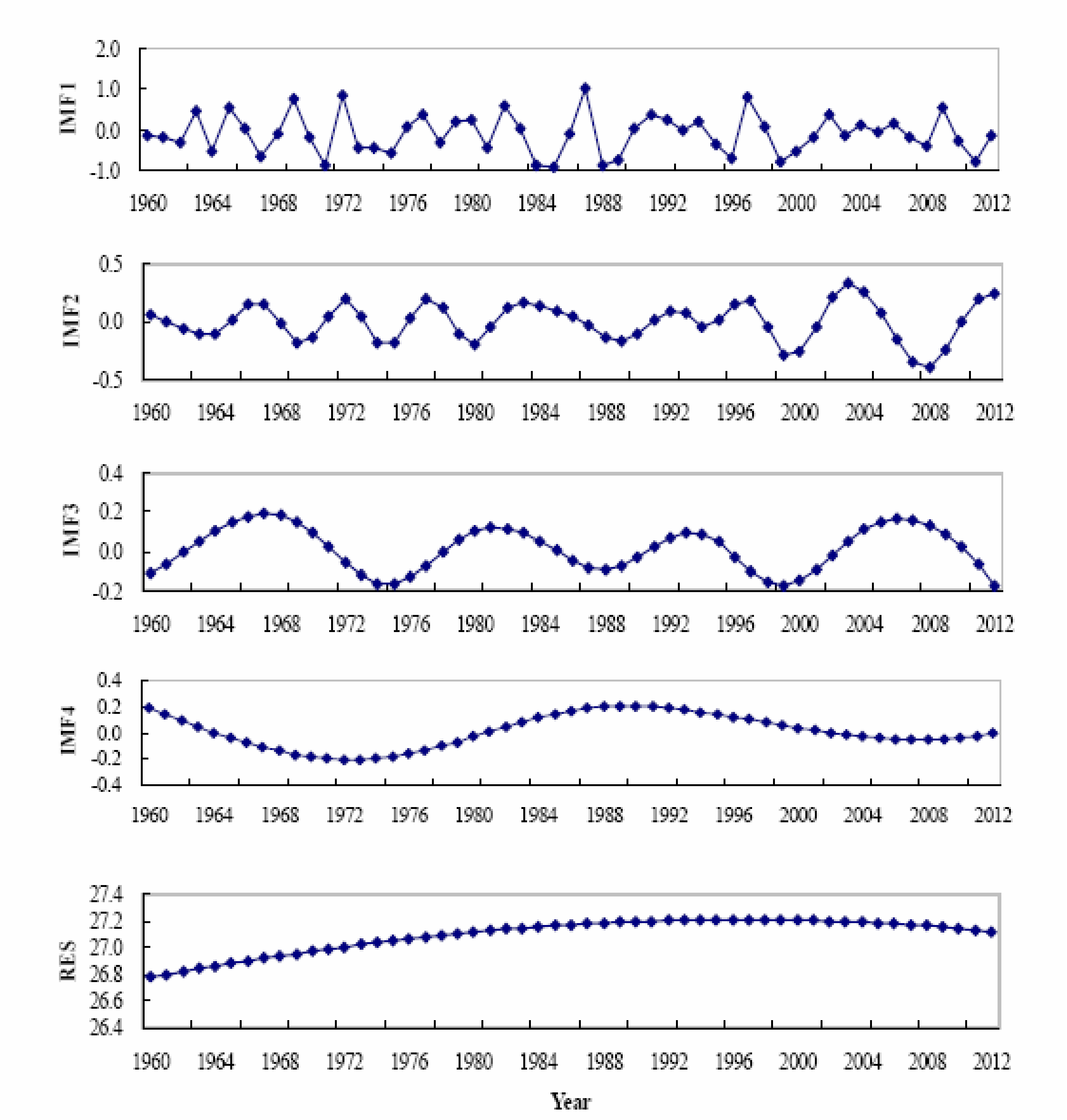

Figure 9 The EEMD results for the NINO3.4 index data series during 1960 - 2012

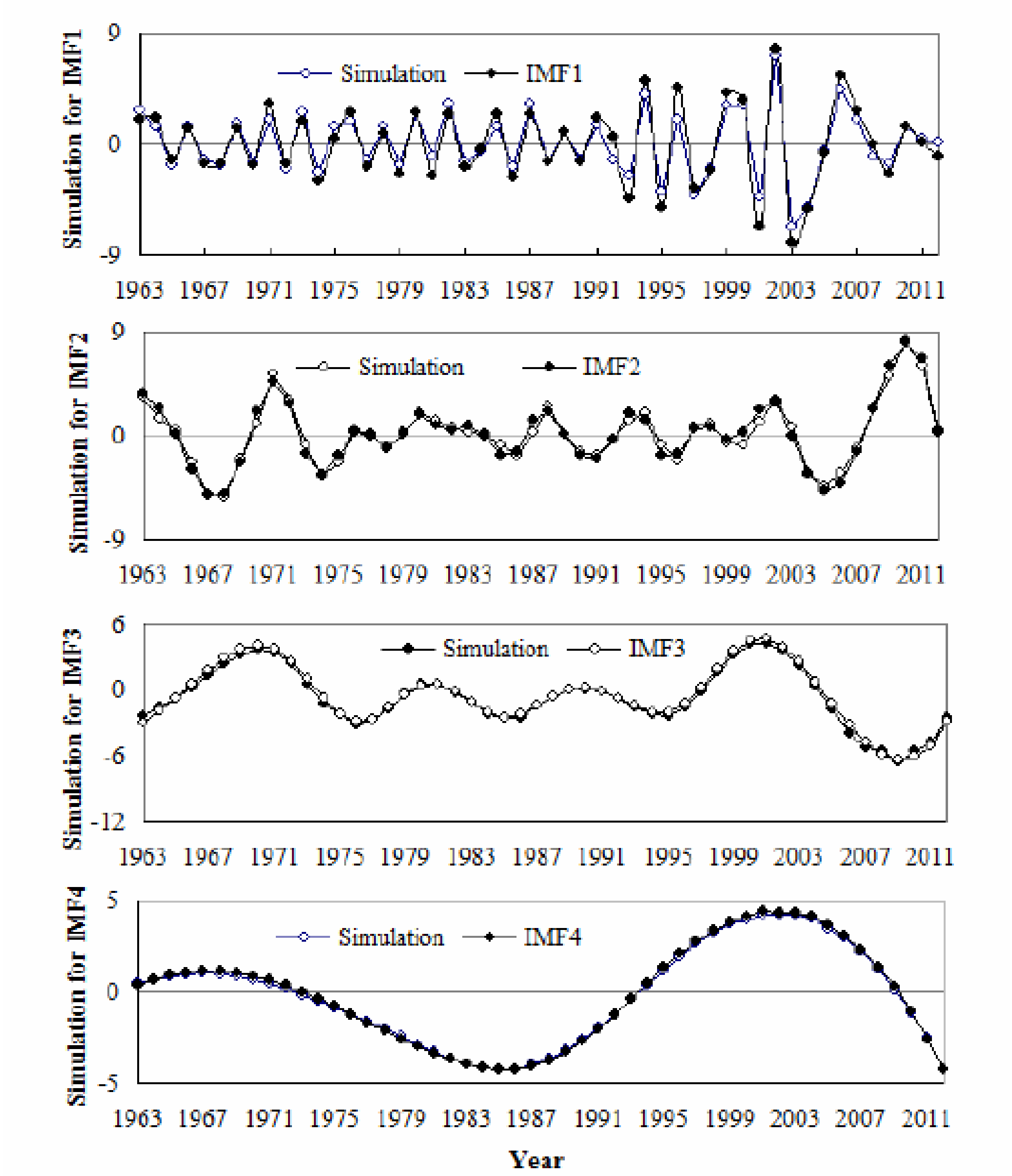

Figure 10 Simulation for the IMFS of AR by BPANN

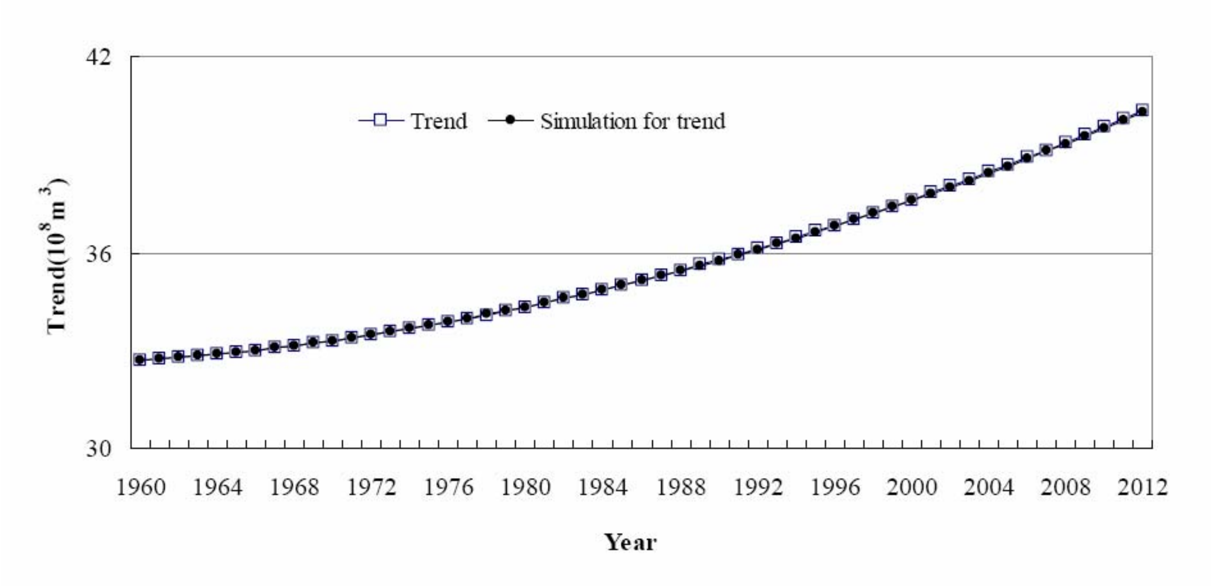

Figure 11 Simulation for the trend of AR by nonlinear regreesion equation

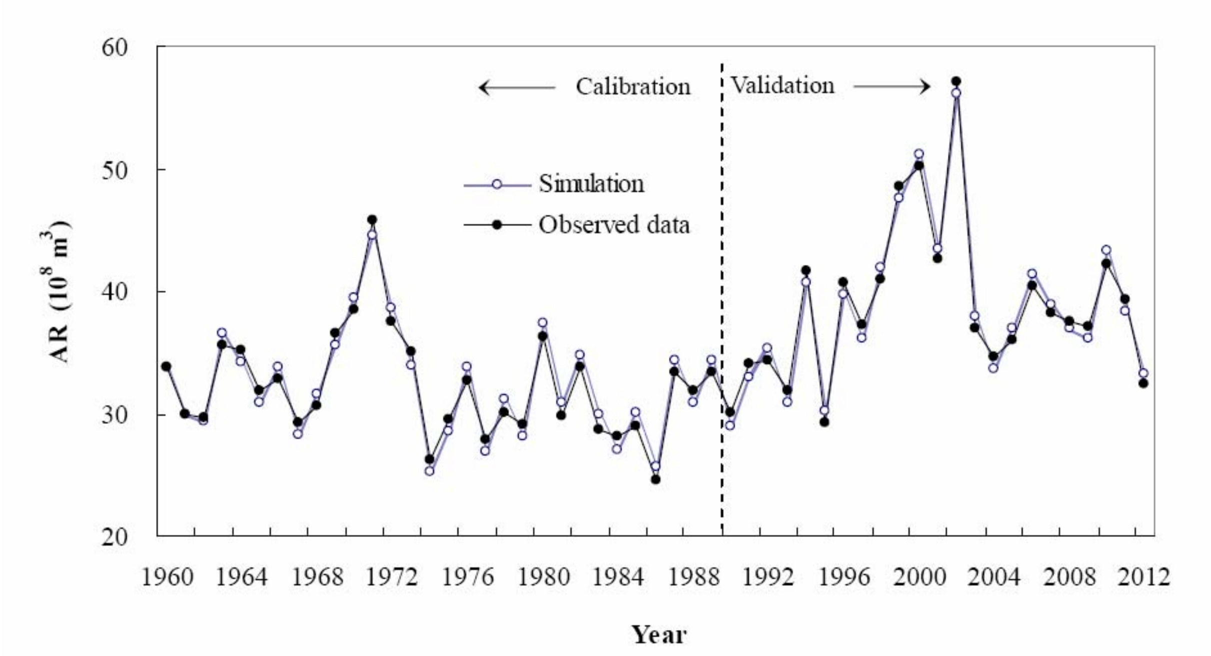

Figure 12 Comparisons between the observed data of AR and its simulated values for calibration period (1960–1989) and validation period (1990–2012)