Rainfall–runoff modeling using principal component analysis and neural network

Tiesong Hu, Fengyan Wu and Xiang Zhang

State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, Hubei Province, China. E-mail: *tshu@whu.edu.cn*

Received 4 September 2006; accepted in revised form 8 February 2007

Nordic Hydrology **Vol 38 No 3 pp 235–248** © IWA Publishing 2007

Abstract The predictive accuracy of a Rainfall–Runoff Neural Network (RRNN) model depends largely on the suitability of its structure. Unfortunately, the procedures for selecting an appropriate structure for the RRNN have not been thoroughly examined. Inclusion of too many input neurons in the RRNN may complicate its structure, and thereby decrease its generalization performance. The objective of this study is to evaluate the potential of a Principal Component Analysis (PCA) method, i.e. by extracting the principal components from lagged input hydrometeorological data, in improving the predictive accuracy of the RRNN. The Darong River watershed located in Guangxi Province of China, with a drainage area of 722 km^2, has been selected to demonstrate the PCA method for modeling the hourly Rainfall–Runoff (RR) relationship. Comparative tests on the forecasting accuracy were conducted among the RRNNs configured with both basin-averaged and spatially distributed rainfall information. Experimental results revealed that, when calibrating the RRNNs with spatially distributed rainfall, the RRNNs using the PCA as an input data-preprocessing tool were found to provide a generally better representation of the RR relationship for the Darong River watershed. However, variable results were observed if the neural networks had been calibrated with basin-averaged rainfall.

Keywords Neural network; principal component analysis; rainfall–runoff modeling

Introduction

Rainfall–runoff (RR) modeling has been receiving an immense attention by practicing hydrologists due to the demands for more timely and accurate forecasts. During the last decade, Artificial Neural Networks (ANNs) have emerged as a promising alternative for RR modeling (ASCE Task Committee on Artificial Neural Networks in Hydrology 2000*a,b*).

While a wide variety of issues in RR modeling have been addressed (Lorrai and Sechi 1995; Smith and Eli 1995; Mason *et al.* 1996; Minns and Hall 1996; Shamseldin 1997; Dawson and Wilby 1998; Sajikumar and Thandaveswara 1999; Tokar and Johnson 1999), and encouraging results have been reached (Hsu *et al.* 1995; Shameseldin *et al.* 1997; Campolo *et al.* 1999*a,b*; Jain *et al.* 1999; Coulibaly *et al.* 2000*a,b*; Imrie *et al.* 2000; Liong *et al.* 2000; Luk *et al.* 2000), the ANN applications are still restricted to the research environment and their potentials are yet to be fully exploited. A major concern is to improve the model's predictive performance. As in the identification of conceptual RR models, difficulties associated with model structure identification and parameter estimation, such as the existence of numerous multi-local minima on the response surface, the absence of effective and efficient structure identification and parameter estimation algorithms, and time-consuming training process, still exist in the ANN-based hydrologic modeling (Hu *et al.* 2005).

Structure identification plays an extremely important role in the generalization performance, and has therefore gained the maximum attention by hydrological researchers.

doi: 10.2166/nh.2007.010

If the network is too small, it may not have sufficient degrees of freedom to capture the highly nonlinear RR relationship. Conversely, if the network is too large, it may memorize the fluctuations in training data that are not representative of the watershed under investigation, and this could eventually lead to poor generalization performance.

In recent years, a number of techniques have been investigated to determine network structure with varying degrees of effectiveness including the A Information Criterion (AIC), B Information Criterion (BIC), Schwartz Information Criterion (SIC), principal component analysis (PCA), etc. (Karunanithi *et al.* 1994; Roadknight *et al.* 1997; Jain *et al.* 1999; Zealand *et al.* 1999; Coulibaly *et al.* 2000*b*; Zhang and Govindaraju 2000). The AIC and BIC were employed by Hsu *et al.* (1995) to assist in selecting an appropriate number of hidden nodes. Roadknight *et al.* (1997) applied the PCA to the inputs of the ANN model for determining the critical levels of ozone for visible injury to occur under various microclimatic conditions. Guhathakurta *et al.* (1999) investigated long-range forecasts of Indian summer monsoon rainfall using ANNs with the PCA as a preprocessing tool. In a study by Shamseldin (1997), in order to circumvent the problem of complex network architecture due to large memory length, Rainfall Index (RI), i.e. a weighted sum of the *m* most recent rainfall values, was the sole external input factor to the network in one of the four input scenarios investigated. Lachtermacher and Fuller (1995) determined the number of input units by the number of autoregressive terms and the number of differencing operations of the calibrated ARIMA model in their hydrological time series research. Minns and Hall (1996) used a Feed-Forward Neural Network (FFNN) to model the RR process. The number of nodes in the intervening hidden layer was chosen to be approximately half of the number of input nodes, which consisted of the concurrent and 14 antecedent rainfall depths and 3 antecedent flow ordinates.

Irrespective of the various approaches being reported, no well recognized approaches are available for identifying a proper network structure in the context of hydrologic modeling. In this regard, the need for timely and accurate hydrologic forecasting calls for the emergence of an efficient and effective structure identification paradigm, especially for situations where long records of rainfall and runoff data are lacking, as this may prohibit the use of a large network structure.

This paper is mainly concerned with the identification of a suitable number of input neurons, with an objective to evaluate the potential of the PCA method, i.e. by extracting the principal components from lagged input hydrometeorological data, in improving the predictive accuracy of RR modeling.

Methodology

Artificial neural network

An ANN is a distinctively parallel-distributed processor that has a natural propensity for storing experimental knowledge and making it available for use. It has been successfully applied in a number of diverse fields including pattern recognition, prediction and optimization. Owing to such characteristics as parallel processing and distributed representation, and also to the capability to be universal approximators for arbitrary finite-input environment measures provided that as many hidden units as required for internal representation are employed, ANNs are attractive to hydrologic research. Its applications include rainfall prediction (Luk *et al.* 2000), RR process modeling (Hsu *et al.* 1995; Dawson and Wilby 1998), river flow rate and stage prediction (Campolo *et al.* 1999*a,b*; Coulibaly *et al.* 2000*a*), confidence intervals approximation (Whitley 1999), hydrological time series forecasting (Hu *et al.* 2001), reservoir operating policies deriving (Raman and Chandramouli 1996), groundwater remediation (Rogers and Dowla 1994), synthetic inflow generation (Raman and Sunilkumar 1995), as well assessment of a stream's hydrological and ecological response to climate change (Poff *et al.* 1996). Three kinds of neural networks have been investigated for hydrologic applications, including (i) the FFNNs, such as the Back-Propagation (BP) network and radial basis function network, (ii) recurrent neural networks, and (iii) self-organizing ANNs, in which the FFNNs are the most popular in RR relationship modeling.

Studies on the application of the FFNN for modeling RR process usually assumed that the runoff at a certain time period is related to a combination of factors such as precipitation, temperature, evaporation and runoff at different time intervals, i.e.

$$R(t) = f_{\mathrm{FNN}}\{P(t), \ldots, P(t-n_p), E(t), \ldots, E(t-n_E), R(t-1), \ldots, R(t-n_R)\} \quad (1)$$

where $R(t)$, $P(t)$, $E(t)$ are runoff, precipitation and evaporation at time t, respectively. n_P, n_E and n_R are the number of antecedent precipitation, evaporations and runoffs contributing to the present runoff. $f_{\mathrm{FNN}}(\bullet)$ denotes the nonlinear relationship inherent in the hydrometeorological data of the watershed being investigated, which could be extracted by solving the following optimization problem (Hu 1997):

$$\begin{cases} \min F = \sum_{p=1}^{P} \sum_{k=1}^{N_O} (TO_{pk} - OO_{pk})^2 \\ \text{s.t. } OO_{pk} - f\left(\sum_{j=1}^{N_H} w_{kj}^{\mathrm{O}} OH_{pj} - \theta_k^{\mathrm{O}}\right) = 0 \quad (k = 1, 2, \ldots, N_O) \\ OH_{pj} - f\left(\sum_{i=1}^{N_I} w_{ji}^{\mathrm{H}} OI_{pi} - \theta_j^{\mathrm{H}}\right) = 0 \quad (j = 1, 2, \ldots, N_H) \end{cases} \quad (2)$$

where N_I, N_H, N_O are the numbers of neurons in the input layer, the hidden layer and the output layer, respectively, in which NI $= n_P + n_E + n_R$. N_O equals one if only one-time-step-ahead prediction is required. w_{kj}^{O}, w_{ji}^{H} are the connection strength between neuron k of the output layer and neuron j of the hidden layer, and between neuron j of the hidden layer and neuron i of the input layer, respectively; θ_k^{O}, θ_j^{H} are the bias of the neuron k and j; OO_{pk}, OH_{pj}, OI_{pi} are the output of neurons k, j, i for the pth training pattern, in which OI_{pi} is identical to the input to the ith input node, i.e. I_{pi}. TO_{pk} is the target output for the kth output neuron of the pth training pattern, which means observed runoffs in the context of RR modeling. $(TO_{pk}, I_{pi}, p = 1, 2, \ldots, P)$ constitutes the P training pairs used for parameter estimation of the FFNN model. $f(\ )$ denotes a nonlinear neuron transfer function such as the generally used logistic function, which is that also used in the present study. A number of optimization techniques are available to solve this problem with varying degrees of effectiveness, including gradient descent algorithm, conjugated gradient algorithm and genetic algorithm, among which application of the gradient descent algorithm gave birth to the popular BP network.

The steps involved in the application of the FFNN to RR modeling are (i) to identify a suitable structure of the neural network model, i.e. to identify the structural parameters $N_I(n_P, n_E, n_R)$, N_H and N_O; (ii) to select the training pairs $(TO_{pk},\ I_{pi},\ p = 1, 2, \ldots, P)$ that are well representative of the characteristics of a watershed and meteorological patterns for calibration and verification; (iii) to estimate the network parameters, i.e. $w_{kj}^{\mathrm{O}}, w_{ji}^{\mathrm{H}},\ \theta_k^{\mathrm{O}}, \theta_j^{\mathrm{H}}$ using the proper optimization algorithm; and (iv) to verify the identified neural network model with test data.

Principal component analysis

The transformation of rainfall into runoff involves many complex components, which could be influenced by many factors. These factors can be classified into three main groups: (i) geomorphological factors of the watershed, such as watershed topography, vegetation cover, soil types, etc.; (ii) climatic factors of the region, such as rainfall intensity and distribution, temperature, snow melting, evaporation and transpiration, etc.; and (iii) human activities, such as land use and even the amount of carbon dioxide and other greenhouse gases released, which are believed to induce global warming. The combined effect of physical and artificial factors further complicates the relationship between rainfall and runoff. The practical factors that have been used for modeling the RR process are rainfall, temperature, evaporation, snowmelt equivalent and runoff at previous times. Other factors being ignored may be largely a result of both the difficulties in the measurement of some factors such as land use, soil moisture, and groundwater characteristics, and of the inefficiency of conventional models in handling too many input factors.

Common practice in RR modeling with ANNs is that all practical factors are equally used as inputs to the ANN models. However, our experimental analysis of runoff autocorrelation and cross-correlation indicated that strong inter-correlations exists among runoff and its affecting factors, and multi-collinearity arising from these inter-correlations could lead to poor generalization performance as a result of the increased complexity of model structure (Hu 1997). In addition, too many runoff-affecting factors relative to a limited number of hydrometeorological data that is available for model calibration may render the RR model training computationally impossible. In this perspective, there is a need for an effective and efficient method which is capable of converting the correlated runoff-affecting factors into uncorrelated ones and reducing the dimensionality of the input data, while preserving as much information in watershed characteristics and rainfall patterns as possible.

PCA is a technique attempting to look for linear combinations of original runoff-affecting factors that can be used to summarize the data, with an objective of losing as little information as possible throughout the process. These new linear combinations are known as Principal Components (PC) (Rencher 1998). In cases where input space is not high dimensional, PCA can still be a very useful preprocessing technique for selecting optimal features in input data (Hotelling 1933). PC transformation is an orthogonal transformation that transforms any set of variables into a set of variables that are uncorrelated with each other. In particular, if the original variables are normally distributed, the new variables are not only uncorrelated but also independent.

Supposing that we have observations of P variables X on each of n individuals, i.e. $X = (x_1, x_2, \ldots, x_P)$ such as runoff-affecting factors and have to find a set of new variables $\xi = (\xi_1, \xi_2, \ldots, \xi_P)$, which are linear functions of the X's but are themselves uncorrelated with a decreasing variance from first to last:

$$\xi_i = \alpha_{i1}x_1 + \alpha_{i2}x_2 + \cdots + \alpha_{ij}x_j + \cdots + \alpha_{iP}x_P \tag{3}$$

In order to impose the condition that the transformation is self-orthogonal, the following constraints are required:

$$\sum_{i=1}^{P} \alpha_{ij}\alpha_{ik} = 0 \quad j \neq k \tag{4}$$

$$\sum_{i=1}^{P} \alpha_{ij}\alpha_{ik} = 1 \quad j = k \tag{5}$$

The orders of PCs depend on the degree of importance, i.e. the first PC should give the most significant relation in the original variables directed according to the largest variance, the second PC should give the second most significant relation and is orthogonal to the first PC, etc. The variances of the later principals would be relatively small if high correlation in the original variables occurs. In this connection, an investigation with the PCA can give an indication of how to reduce the number of modeled input variables and describe the most significant explanation in a fewer number of transformed PCs.

Applications

Description of study area

The Darong River watershed (Figure 1) is located in Guangxi Province, China. It has a drainage area of 722 km^2 and a continental climate as characterized by a wide range of flood magnitude, lighter winter precipitation, and strong and long duration floods. The Darong River is a rainfall-dominated stream exhibiting the behavior of sudden and high fluctuation in runoff and a highly nonlinear relationship between rainfall and runoff. The annual average rainfall of the watershed is approximately 2 000 mm. Within the watershed, the land use is substantially agricultural without significant change throughout our investigating period. There are five automatic rainfall-gauging stations, which were installed at Darong on the Darong River and a number of its tributaries, namely the Hua River, Chuan River, and Qingshui River. Hourly measurements of rainfall at the five gauge stations along with those of runoff at Darong were available for 46 isolated storms from 1968 to 1994.

The Darong Watershed was selected to demonstrate the proposed methodology for modeling hourly RR relationship using the PCA to preprocess the input data for a neural network.

Description of experiments

To assess the effects and accuracy of the RRNN models after performing the PCA, comparisons were made between the neural networks with the PCA as input preprocessing tool and those without employing the PCA. Moreover, comparisons were further made between the neural networks configured with both basin-averaged and spatially distributed

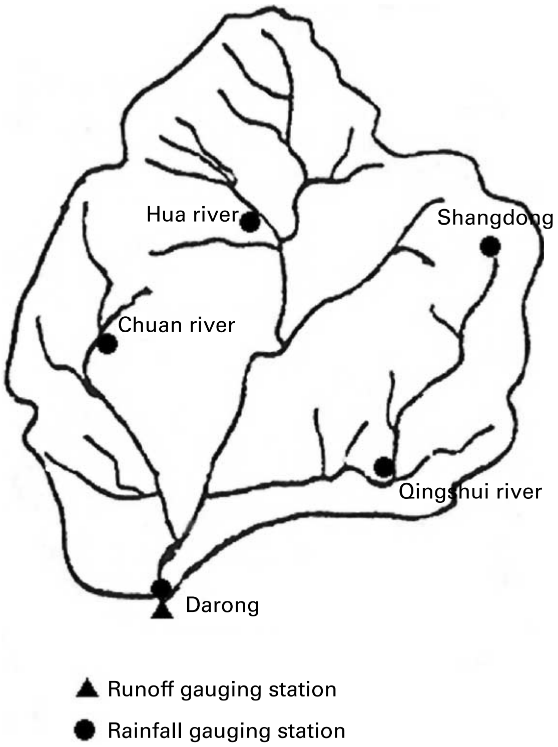

Figure 1 The Darong watershed in Guangxi Province, China

rainfall information. In order to get better understandings of the effect using PCA on the model complexity and performance, additional experiments were also carried out to check if similar results could be obtained using a number of actual inputs equivalent to the number of PCs. In this regard, another two comparative investigations were made between the networks using PCA as input data-preprocessing tools and networks using correlation analysis to choose the actual inputs that are strongly correlated to the output, in both of which the number of input nodes is equal to the number of PCs. Different combinations of rainfall information and input preprocessing and selection tools led to a total of six experiments designed in this study, which are shown in Table 1.

In the first experiment, the PCA was not employed, and the RR modeling was made with real hourly-averaged values of rainfall over the entire watershed, which have been summed and averaged from the five rainfall gauge stations. In the second experiment, the PCA was applied to the input data before proceeding to the training process. All other configurations for the two experiments were kept the same as much as possible. As for the third and fourth experiments, distributed rainfall information at the five rainfall gauge stations was used as the inputs. The complexity of the last two neural network models was therefore significantly increased when compared to the first two experiments. The difference between the third and fourth experiments is that the PCA was applied in the fourth experiment but not in the third experiment. The number of input nodes of experiments 5 and 6 are equal to those of experiments 2 and 4, respectively, but actual inputs that were highly correlated with output (runoff) were directly chosen as inputs to the RRNN models instead of using PCA to produce inputs in experiments 2 and 4.

In order to avoid the selection of imbalanced training set, which could otherwise deteriorate the generalization performance, considerable attention had been paid when selecting the flood records. The selection process aimed to ensure a proper portion of different storm events with varying duration, total depth and profile and occurring at irregular intervals, etc., exist in the sample. Furthermore, different shapes of storm profiles, such as early-peaked and late-peaked, as well as symmetrical events of both single-storm and multi-storm types were also included.

As a result, a total of 32 storm events were selected from the period of 1968 to 1994 with a mean of 26 h and a deviation of 8 h. These storm events were used for neural network parameter estimation and model performance assessment. A total of 1 035 training pairs were generated from the 32 storm events. In order to examine the effects of insufficient training data on the performance of the RRNN model, especially in the situations where long historical data records are lacking (which is typically in developing countries), another 6 flooding events with similar statistical properties to those of the 32 events were selected from 1968 to 1986. Another training set comprising 312 pairs of input–output patterns were generated from the six flooding events. Obviously, there were much lesser pairs in the latter

Table 1 Description of experiments

Experiment number	Performing PCA or not	Calibrated with basin averaged rainfall	Does the number of input nodes equal its corresponding experiment employing PCA?
1	No	Yes	No
2	Yes	Yes	–
3	No	No	No
4	Yes	No	–
5	No	Yes	Yes
6	No	No	Yes

case. The RRNN model calibrations were made with both the large data set of 1 035 input–output training patterns and the small data set of 312 pairs.

The six experiments and the two input data scenarios (large and small training datasets) result in a total of 12 different RRNN configurations.

Determination of the network structure

Inputs to the 12 different RRNN configurations were selected as antecedent discharges at Darong along with the precipitation and temperature of the examined hour, precipitation of preceding hours at the five gauging stations, and antecedent hourly basin-averaged air temperature. Based on the investigation done by Darong Water Resources Planning Office, the rain gauge input of the Hua River station was lagged by 4 h. The time lags for other stations, except for the Darong itself, are listed in Table 2.

In modeling the RR relationship, a frequently encountered problem is to determine what is the suitable length of time series from each gauging station with which to form the RRNNs input patterns. Physically, the length of logged rainfall or runoff is affected by a number of factors including the degree of watershed wetness, size of watershed, and intensity and distribution of rainfall.

In this investigation, cross-correlation analyses were performed to gain intuitive understanding on the degree of relationship between runoff at Darong and rainfall of the five gauging stations, and to identify the appropriate number of antecedent rainfall values as inputs. For example, the result of cross-correlation analysis suggested that the discharge of Darong at a certain time could be affected by antecedent rainfall at Hua River from up to 6 h previous excluding the 4-h time lag. Results of cross-correlation analyses for other tributaries are shown in Table 2. In addition, autocorrelation analysis of runoff was carried out to determine the appropriate number of preceding runoff ordinates (Table 2). Similar correlation analyses were also been conducted between runoff at Darong and hourly basin-averaged rainfall for preparing the approximate inputs for experiments 1 and 2 (see Table 3).

A trial-and-error procedure was used in this study for determining the optimal number of hidden nodes in all the experiments on the ground as no well-defined algorithm can determine the optimal number of hidden nodes in all circumstances. In this regard, the number of hidden nodes of the networks, which have been finally selected, differs from experiment to experiment, which are shown in Tables 4 and 5.

Table 2 Input variables to the RRNNs calibrated with distributed rainfall information

		Input variables		
Gauging stations	**Time lags**	**Precipitation**	**Temperature**	**Discharge**
Hua River	4	$R_{t-5}, R_{t-6}, \ldots, R_{t-9}$	$T_t, T_{t-1}, \ldots, T_{t-3}$	–
Chuan River	2	$R_{t-3}, R_{t-4}, \ldots, R_{t-7}$	$T_t, T_{t-1}, \ldots, T_{t-3}$	–
Shangdong	4	$R_{t-5}, R_{t-6}, \ldots, R_{t-8}$	$T_t, T_{t-1}, \ldots, T_{t-3}$	–
Qingshui	2	$R_{t-3}, R_{t-6}, \ldots, R_{t-8}$	$T_t, T_{t-1}, \ldots, T_{t-3}$	–
Darong	0	$R_t, R_{t-1}, \ldots, R_{t-7}$	$T_t, T_{t-1}, \ldots, T_{t-3}$	$D_{t-1}, D_{t-2}, \ldots, D_{t-4}$

Table 3 Input variables to the RRNNs calibrated with basin-averaged rainfall information

Precipitation	**Temperature**	**Discharge**
$R_t, R_{t-1}, \ldots, R_{t-9}$	$T_t, T_{t-1}, \ldots, T_{t-3}$	$D_{t-1}, D_{t-2}, \ldots, D_{t-4}$

Table 4 Calibration statistics

	Large dataset (32 storm events)			Small dataset (6 storm events)		
Experiment number	Network structure	RMSE (m^3/s)	R^2	Network structure	RMSE (m^3/s)	R^2
1	18-9-1	62.5	0.8686	18-7-1	37.8	0.8737
2	8-5-1	113.3	0.7868	8-4-1	85.7	0.8137
3	32-15-1	12.9	0.9943	32-12-1	11.3	0.9984
4	14-5-1	22.3	0.9386	14-5-1	31.4	0.9475
5	8-6-1	57.8	0.8467	8-5-1	30.5	0.8854
6	14-8-1	26.4	0.9272	14-7-1	22.7	0.9126

Table 5 Verification statistics

	Large dataset (32 storm events)			Small dataset (6 storm events)		
Experiment number	Network structure	RMSE (m^3/s)	R^2	Network structure	RMSE (m^3/s)	R^2
1	18-9-1	76.3	0.8271	18-7-1	89.7	0.8123
2	8-5-1	67.5	0.8242	8-4-1	34.7	0.8669
3	32-15-1	127.7	0.8054	32-12-1	178.6	0.7573
4	14-5-1	37.6	0.8976	14-5-1	54.4	0.8544
5	8-6-1	65.4	0.8249	8-5-1	42.1	0.8647
6	14-8-1	69.1	0.8156	14-7-1	117.8	0.7675

Extraction of principal components

In practice, PC extraction generally involves the computation of input data covariance matrix and then the application of a diagonalization procedure to extract the eigenvalues and the corresponding eigenvectors. For a detailed extraction procedure the reader is referred to Rencher (1998).

As indicated in Tables 2 and 3, the total numbers of input variables in experiments 1 and 3 comprising the precipitation, temperature and discharge, are 18 and 32, respectively. In order to extract PCs and reduce the dimension of the input space, the PCA of both the 18 and 32 input variables with the data from 1968 to 1994 was performed. Figures 2 and 3 show the results of the PCA with basin-averaged rainfall and distributed rainfall information, respectively. In experiment 2, i.e. with basin-averaged rainfall, it was observed that the first 8 PCs explained 85.7% of the total variance, and as a result, the input neurons was set as 8. Similarly, the number of input nodes in experiment 4 was set as 14 since the first 14 PC accounted for 84.8% of the total variance.

Input and output data normalization

Input and output data normalization is an important factor affecting the model performance, and they have to be carried out even though the PCA had been performed with original input data, and that was the case of experiments 2 and 4. According to Azoff (1994), the along-channel normalization, across-channel normalization, mixed-channel normalization and external normalization are the four methods for normalization. In order to avoid computational problems and to facilitate network learning (Hu 1997), the following normalization formula was used for all three kinds of input, i.e. precipitation, temperature and discharge:

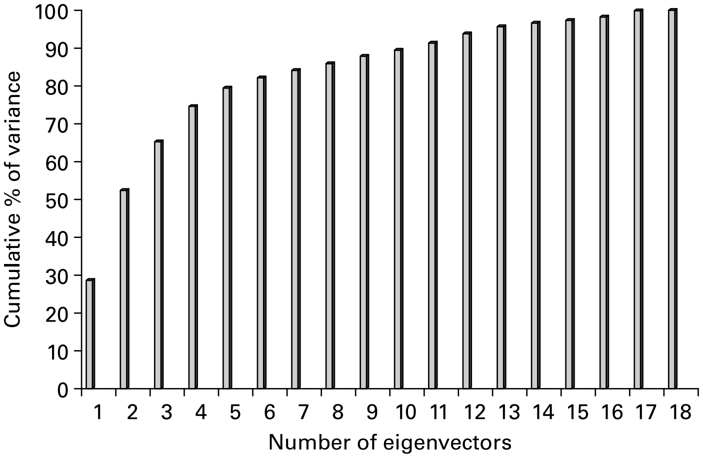

Figure 2 Result of principal component analysis with basin-averaged rainfall information

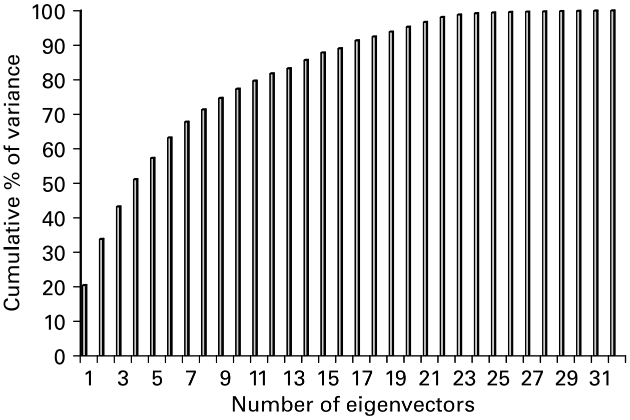

Figure 3 Result of principal component analysis with distributed rainfall information

$$x' = \frac{x - \mu}{\gamma\sigma} \tag{6}$$

where x' is the normalized value of input data, x is the original input value of these hydrometeorological data, μ,σ are the mean and standard deviation of the input data. γ is the parameter controlling the mapping range. About 95% of the input variable data maps in $[-1,1]$ range when the input variable follows a normal distribution and γ is 1.96. In this paper, a value of $\gamma = 1.96$ is used. As far as the normalization of output data is concerned, the output data is scaled down linearly in the range of [0.1, 0.9] in this paper, which makes the extrapolation of discharge possible and eases the tension of slow convergence to some extent.

Performance criteria

A variety of verification criteria could be used for evaluating the model performance. With an intention to gain a quantitative indication of the model error in terms of dimensioned quantity, the Root Mean Square Error (RMSE)) was selected, which is given by the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{p=1}^{\mathrm{NOV}} (TO_{pk} - OO_{pk})^2}{\mathrm{NOV}}} \quad (k = 1) \tag{7}$$

where NOV is the number of data elements in the verification period. In addition to the RMSE, the model efficiency criterion R^2 was also selected to assess the model performance on the ground that the lengths of calibration periods and verification periods in a large and small dataset could be quite different, and that the RMSE cannot really indicate the performance of the models for different calibration record lengths:

$$R^2 = 1 - \frac{\sum_{p=1}^{P} (TO_{pk} - OO_{pk})^2}{\sum_{p=1}^{P} (TO_{pk} - \overline{TO})^2} \quad (k = 1) \tag{8}$$

where $\overline{TO}$ is the average value of the observed runoff for the calibration period. The higher the value of R^2, the better is the forecasting performance. If R^2 is equal to 1, it implies that the forecast replicates observation 100% of the time. If all forecast values equal to the long-term observed mean, R^2 would assume the value of 0. If R^2 is less than 0, however, it implies that the forecast is worse than the long-term observed mean.

Results and discussions

The proposed neural networks were applied to model the RR relationship of Darong River watershed under the six-experiment scheme with both the large data set (32 storm events) and small dataset (6 storm events). A total of 12 neural networks were finally calibrated and tested for the effects of performing the PCA on model performance. Verification of these neural networks was done with 284 testing patterns of the last four storms that were not involved in the training sets. The main results are shown in Tables 4 and 5 and Figures 4–7.

Tables 4 and 5 present a summary of statistical performance under the six-experiment scheme with the neural network models calibrated on the large and small dataset for one hour-ahead prediction. An examination of Tables 4 and 5 shows that all 12 best-fit neural

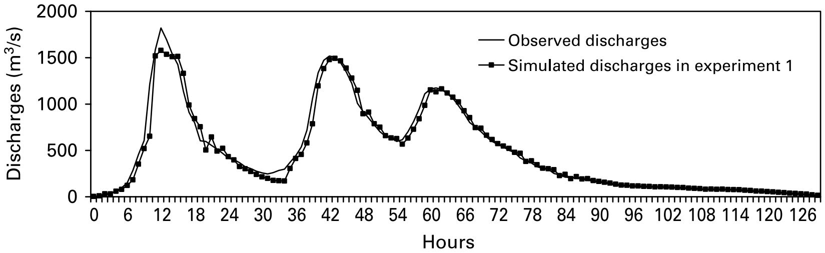

Figure 4 Scatter plot of simulated discharges of experiment 1 against observed discharges (14–21 July 1974)

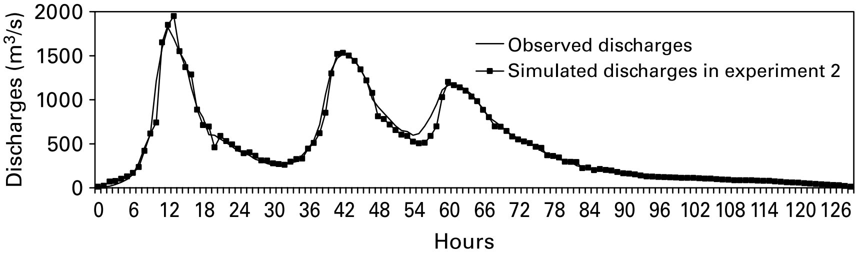

Figure 5 Scatter plot of simulated discharges of experiment 2 against observed discharges (14–21 July 1974)

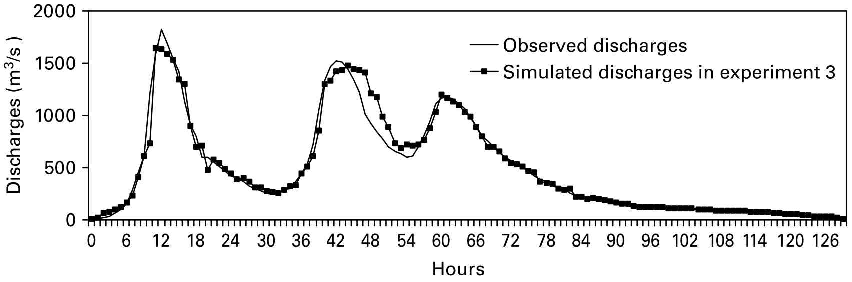

Figure 6 Scatter plot of simulated discharges of experiment 3 against observed discharges (14–21 July 1974)

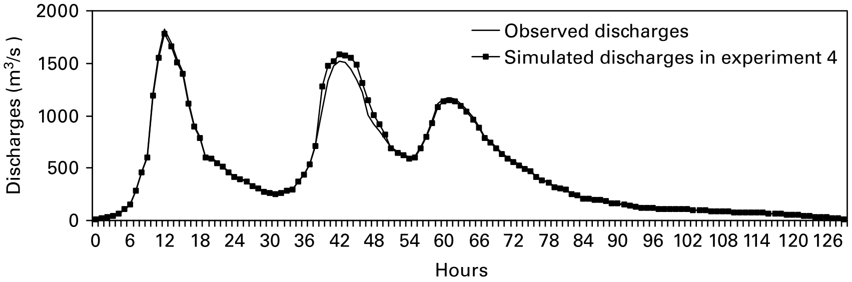

Figure 7 Scatter plot of simulated discharges of experiment 4 against observed discharges (14–21 July 1974)

networks were generally effective for simulating the RR relationship of the Darong River watershed, and most of the configurations were able to provide a fairly good model performance with the R^2 efficiency values being greater than 0.80 in both the calibration and verification periods.

Comparing the calibration results of experiments 3 with those of experiment 1 (Table 4) indicates that a calibration of the RRNNs with distributed rainfall information was generally found to be rewarding in that it significantly improved the calibration accuracy when compared with those which had been calibrated with basin-averaged rainfall information. This is reflected in the higher values of the two parameters in experiment 3–namely the RMSE and the model efficiency R^2. Similar improvements in calibration accuracy were observed when comparing the results in experiment 4 with those in experiment 2 (Table 4). Among the six experiments, the neural networks used in experiment 3 performed the best in terms of both the RMSE and R^2 during the calibration period. Unfortunately, it was noticed that the high calibration accuracy in these experiments did not lead to high verification accuracy as the authors originally expected, and in contrast the testing accuracy presented in experiment 3 was the lowest among the six experiments (Table 5). This is likely to be the result of having too many input neurons and relatively inadequate training data in this experiment, thereby deteriorating the generalization performance. The dramatic deterioration of validation statistics for experiment 3 supports the conclusion that there is an optimal limit of temporal and spatial information that may be considered for accurate ANN-based forecasting (Luk *et al.* 2000).

In the verification period, when the results of experiment 4 are compared with those of experiment 3, it is conspicuous that the results of experiment 4, in terms of both RMSE and R^2, are favorably comparable with those of their counterparts. In the case of the large dataset, for example, R^2 of 0.8976 of the neural network, configured with a large dataset, spatially distributed rainfall as well as the PCA as an input data preprocessing tool, is significantly higher than that of 0.8054, which is the result of the neural network without performing the PCA. This observation suggested that performing the PCA could significantly improve the generalization performance of the RRNN model under situation wherein spatially distributed rainfall information is used for the description of rainfall input. However, improvement in the predictive accuracy of the neural networks calibrated with basin-averaged rainfall was not as high as those calibrated with spatially distributed rainfall. Moreover, the R^2 of 0.8242 of the RRNN model calibrated with the large dataset in experiment 2 was marginally lower than that of of its counterpart (0.8271), indicating that the additional computational efforts required in performing the PCA does not appear to be justified for circumstances where the RRNN models are calibrated with a large dataset and basin-averaged rainfall.

Comparison of the results of experiments 2 and 5, both of which utilize 8 input nodes, shows that, in the case of basin-averaged rainfall, there is no clear indication that the network using PCA as an input preprocessing tool is better than those using correlation analysis as an input selection tool, where the R^2 value of 0.8242 of experiment 2 is just marginally lower than that of 0.8269 from experiment 5. Conversely, examination of the results of experiments 4 and 6 suggests that, in the case of distributed rainfall, the network using PCA as an input preprocessing tool performs better than the network using correlation analysis technique as an input selection tool.

In the case of the small dataset, performing the PCA before calibrating the RRNNs was generally found to be rewarding with both basin-averaged rainfall and spatially distributed rainfall in that results of neural networks performing the PCA compared favorably with that of neural networks not performing the PCA.

Figures 4–7 present the comparisons of observed (one sequence of three storm events from 14 July to 21 July 1974) and simulated hourly discharges obtained with the RRNN models calibrated with the large dataset. An examination of Figures 4–7 shows that the RRNN models tended to forecast the magnitude of base flow quite well while experiencing relatively greater difficulty in predicting the magnitude and timing of the three peak flows. Most likely, it may be attributable to both the fact that most of the data used for training is on the long recessions and the lower nonlinearities inherent in the long recessions. Comparison of the peak flow predictions between Figures 4 and 5 indicated that performing the PCA allowed the possibility of overestimation of the peak flows. A similar result can be revealed by comparing the results presented in Figures 6 and 7.

Referring to the small dataset, the results of experiments 1 and 2 compared favorably with those of experiments 3 and 4, respectively, even though the R^2 value of 0.8669 of experiment 2 was not notably different from that of 0.8544 of experiment 4. These results indicate that, whether the PCA were used or not, in the case of the small dataset, the RRNN models calibrated with basin-averaged rainfall were shown to provide a better RR relationship of the Darong River watershed than the models calibrated spatially distributed rainfall. It may be attributable to the limited number of data used for model identification, which could not accommendate the requirement for calibrating the distributed RRNN model. Conversely, in the case of the large dataset, comparison of the results of experiment 2 with those of experiment 4 reveals that spatially distributed rainfall information was found to be helpful in improving the model generalization performance.

Conclusions

The potential of PC analysis as an input data preprocessing tool for neural networks in the context of RR modeling has been investigated in this study. Comparison tests on the calibration and verification accuracy were made between the RRNNs performing the PCA and those not performing the PCA, with both spatially distributed rainfall and basin-averaged rainfall information. Furthermore, these RRNNs were calibrated with both the large and small datasets to give some indications of the effects of performing the PCA on model performance under different scenarios of training data lengths.

The results of comparison tests indicated that the ANNs performing the PCA as an input data preprocessing tool were generally found to provide a better representation of RR relationship of the Darong River watershed than those without performing the PCA. Significant improvements in the model performance were observed under situations whereby spatially distributed rainfall information was used for the description of the rainfall input, thereby substantially increasing the complexity of the model structure. However, some variable results were also observed, and additional computational efforts required in performing the PCA did not appear to be justified for circumstances where the RRNN models are applied with a large dataset and basin-averaged rainfall information in the case of the Darong River.

The implication of the above results is that a balance between the complexity of neural network structure and the length of data available for calibrating a neural network model should be maintained. It is shown that spatially distributed rainfall information and the large dataset do not always lead to higher model performance. In this regard, performing the PCA may be important, especially for circumstances where reliable long recorded hydrological data is not available for modeling.

Acknowledgements

The authors gratefully acknowledge the helpful comments and suggestions from two anonymous reviewers, which are reflected in the present paper. This research was funded by both the Natural Science Foundation of China (project No. 50479039) and the He Yindong Foundation (project No. 71073).

References

ASCE Task Committee on Artificial Neural Networks in Hydrology (2000a). Artificial neural networks in hydrology. I. Preliminary concepts. *J. Hydrol. Engng.*, **5**(2), 115–123.

ASCE Task Committee on Artificial Neural Networks in Hydrology (2000b). Artificial neural networks in hydrology. II. Hydrologic applications. *J. Hydrol. Engng.*, **5**(2), 124–137.

Azoff, E.M. (1994). *Neural Network Time Series Forecasting of Financial Markets*, John Wiley and Sons, Chichester.

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

Campolo, M., Soldati, A. and Andreussi, P. (1999b). Forecasting river flow rate during low-flow periods using neural networks. *Water Res. Res.*, **35**(11), 3547–3552.

Coulibaly, P., Anctil, A. and Bobee, B. (2000a). Daily reservoir inflow forecasting using ANNs with stopped training approach. *J. Hydrol.*, **230**, 245–257.

Coulibaly, P., Anctil, A. and Bobee, B. (2000b). Neural network-based long-term hydropower forecasting system. *Computer-Aided Civil Infrastruc. Engng.*, **15**, 355–364.

Dawson, C.W. and Wilby, R. (1998). An artificial neural network approach to rainfall-runoff modeling. *Hydrol. Sci. J.*, **43**(1), 47–66.

Guhathakurta, P., Rajeevan, M. and Thapliyal, V. (1999). Long range forecasting Indian summer monsoon rainfall by a hybrid principal component neural network model. *Meteorol. Atmos. Phys.*, **71**, 255–266.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Philos.*, **24**, 417–441.

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

Hu, T.S. (1997). *Neural Optimization and Prediction*, Dalian Maritime University Press, Dalian, China (in Chinese).

Hu, T.S., Lam, K.C. and Ng, S.T. (2001). River flow time series prediction with range-dependent neural network. *Hydrol. Sci. J.*, **46**(5), 729–745.

Hu, T.S., Lam, K.C. and Ng, S.T. (2005). A modified neural network for improving river flow prediction. *Hydrol. Sci. J.*, **50**(2), 299–318.

Imrie, C.E., Durucan, S. and Korre, A. (2000). River flow prediction using artificial neural networks: generalization beyond the calibration range. *J. Hydrol.*, **233**, 138–153.

Jain, S.K., Das, A. and Srivastava, D.K. (1999). Application of ANN for reservoir inflow prediction and operation. *J. Water Res. Plann. Mngmnt., ASCE*, **125**(5), 263–271.

Karunanithi, N., Grennery, W.J, Whitley, D. and Bovee, K. (1994). Neural networks for river flow prediction. *J. Comput. Civil Engng.*, **8**(2), 201–219.

Lachtermacher, G. and Fuller, J.D. (1995). Backpropagation in time-series forecasting. *J. Forecast.*, **14**, 381–393.

Liong, S.-Y., Lim, W.H. and Paudyal, G.N. (2000). River stage forecasting in Bangladesh: neural network approach. *J. Comput. Civil Engng.*, **14**(1), 1–8.

Lorrai, M. and Sechi, G.M. (1995). Neural nets for modeling rainfall-runoff transformation. *Water Res. Mngmnt.*, **9**(4), 299–313.

Luk, K.C., Ball, J.E. and Sharma, A. (2000). A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. *J. Hydrol.*, **227**, 56–65.

Mason, J.C., Tem'me, A. and Price, R.K. (1996). A neural network model of rainfall-runoff using radial basis function. *J. Hydrol. Res.*, **34**, 537–548.

Minns, A.W. and Hall, M.J. (1996). Artificial neural networks as rainfall-runoff models. *Hydrol. Sci. J.*, **4**(3), 399–417.

Poff, N.L., Tokar, S. and Johnson, P. (1996). Stream hydrological and ecological response to climate change assessed with an artificial neural network. *Limnol. Oceanogr.*, **41**(5), 857–863.

Raman, H. and Chandramouli, V. (1996). Deriving a general operating policy for reservoir using neural networks. *J. Water Res. Plann. Mngmnt., ASCE*, **122**(5), 342–347.

Raman, H. and Sunilkumar, N. (1995). Multivarirate modeling of water resources time series using artificial neural networks. *Hydrol. Sci. J.*, **40**(2), 145–163.

Rencher, A.C. (1998). *Multivariate Statistical Inference and Applications*. Wiley-Interscience, New York.

Roadknight, C.M., Balls, G.R., Mills, G.E. and Palmer-Brown, D. (1997). Modeling complex environmental data. *IEEE Trans. Neural Networks*, **8**(4), 852–862.

Rogers, L.L. and Dowla, F.U. (1994). Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling. *Water Res. Res.*, **30**(2), 457–481.

Sajikumar, N. and Thandaveswara, B.S. (1999). A non-linear rainfall-runoff model using an artificial neural network. *J. Hydrol.*, **216**, 32–55.

Shamseldin, A.Y. (1997). Application of a neural network technique to rainfall-runoff modeling. *J. Hydrol.*, **199**, 272–294.

Shameseldin, A.Y., O'Connor, K.M. and Liang, G.C. (1997). Methods for combining the outputs of different rainfall-runoff models. *J. Hydrol.*, **197**, 203–229.

Smith, J. and Eli, R.N. (1995). Neural network models of rainfall-runoff process. *J. Water Res. Plann. Mngmnt., ASCE*, **121**(6), 499–508.

Tokar, A.S. and Johnson, P.A. (1999). Rainfall-runoff modeling using artificial neural networks. *J. Hydrol. Engng.*, **4**(3), 232–239.

Whitley, R. (1999). Approximate confidence intervals for design floods for a single site using a neural network. *Water Res. Res.*, **35**(1), 203–209.

Zealand, C.M., Burn, D.H. and Simonovic, S.P. (1999). Short term streamflow forecasting using artificial neural network. *J. Hydrol.*, **214**, 32–48.

Zhang, B. and Govindaraju, R.S. (2000). Prediction of watershed runoff using Bayesian concepts and modular neural networks. *Water Res. Res.*, **36**(3), 753–763.