Improving ANN model performance in runoff forecasting by adding soil moisture input and using data preprocessing techniques

Huanhuan Ba, Shenglian Guo, Yun Wang, Xingjun Hong, Yixuan Zhong and Zhangjun Liu

ABSTRACT

This study attempts to improve the accuracy of runoff forecasting from two aspects: one is the inclusion of soil moisture time series simulated from the GR4J conceptual rainfall–runoff model as (ANN) input; the other is preprocessing original data series by singular spectrum analysis (SSA). Three watersheds in China were selected as case studies and the ANN1 model only with runoff and rainfall as inputs without data preprocessing was used to be the benchmark. The ANN2 model with soil moisture as an additional input, the SSA-ANN1 and SSA-ANN2 models with the same inputs as ANN1 and ANN2 using data preprocessing were studied. It is revealed that the degree of improvement by SSA is more significant than by the inclusion of soil moisture. Among the four studied models, the SSA-ANN2 model performs the best.

Key words | artificial neural network, data preprocessing, runoff forecasting, singular spectrum analysis, soil moisture

Huanhuan Ba
Shenglian Guo (corresponding author)
Xingjun Hong
Yixuan Zhong
Zhangjun Liu
State Key Laboratory of Water Resources and Hydropower Engineering Science, Hubei Provincial Collaborative Innovative Center for Water Resources Security,
Wuhan University,
Wuhan 430072,
China
E-mail: *slguo@whu.edu.cn*

Yun Wang
China Yangtze Power Co., Ltd,
Yichang 443000,
China

INTRODUCTION

Runoff forecasting plays an important role in flood control and reservoir water resources management. In general, runoff forecast models can be grouped into two categories: knowledge-driven models and data-driven models. Knowledge-driven models focus on modeling the rainfall–runoff process with many mathematical formulations and parameters which require detailed understanding of the underlying physical process of the watershed system. Since the rainfall–runoff process in a watershed is complex and is not easy to describe, knowledge-driven models suffer from drawbacks of parameter estimation difficulty and large calibration data requirements (Hsu *et al.* 1995; Tokar & Johnson 1999; De Vos & Rientjes 2005; Noori & Kalin 2016). On the contrary, in data-driven models, the complex and nonlinear relationship between inputs and outputs can be identified based on the statistical analysis of historical hydrological data, requiring no detailed understanding of the rainfall–runoff process. The regression analysis models and artificial intelligence models including multiple linear regression, artificial neural network (ANN), support vector machine (SVM), adaptive neuro-fuzzy inference system belong to this category (Dawson & Wilby 2001; Chang & Chang 2006; Jain & Kumar 2007; Budu 2013; Makwana & Tiwari 2014; Rezaeianzadeh *et al.* 2014; Li *et al.* 2015).

The ANN model, regarded as a black-box model, has been proven to be an effective and powerful approach for hydrological simulation and forecasting (Hsu *et al.* 1995; Tokar & Jhonson 1999; Pang *et al.* 2007; Aksoy & Dahamsheh 2009; Chen *et al.* 2013; Tsai *et al.* 2014; Cheng *et al.* 2015; Chang & Tsai 2016; Li *et al.* 2017; Noori & Kalin 2016; Gorgij *et al.* 2017). The potential of ANN models for modeling the rainfall–runoff process was presented by Hsu *et al.* (1995), who showed that the performance of the ANN approach is superior to that of the ARMAX time

doi: 10.2166/nh.2017.048

series approach or the conceptual SAC-SMA model. Govindaraju (2000b) examined the application of ANN in various branches of hydrology including rainfall–runoff modeling and streamflow prediction, and found that ANN is a robust tool for modeling the complex nonlinear hydrology processes. Dawson & Wilby (2001) reviewed the application of ANN in rainfall–runoff modeling and flood forecasting, and mentioned that the ANN model performance possibly can be improved further by determining appropriate inputs and data cleansing techniques.

In order to improve the accuracy of runoff forecasts through ANN, finding out appropriate input variables is a crucial step. In general, antecedent runoff and rainfall data are usually used as potential predictors for the ANN model by many researchers (Minns & Hall 1996; Wu & Chau 2011; Cheng *et al.* 2015; Wang *et al.* 2015). Robertson *et al.* (2013) stated that using observations of antecedent streamflow and rainfall to represent the initial catchment conditions can potentially have limitations under some circumstances. Therefore, other input variables including soil moisture, evaporation, ground water, and snow accumulation are also incorporated into ANN for streamflow forecasting in many studies (Zealand *et al.* 1999; Nilsson *et al.* 2006; Rosenberg *et al.* 2011; Robertson & Wang 2012; Robertson *et al.* 2013; Noori & Kalin 2016). Since the soil moisture can be used to represent initial catchment conditions, soil moisture also has been incorporated into the ANN model input. Gautam *et al.* (2000) found that the soil moisture data can be useful predictors for ANN models. Due to the lack of soil moisture observations, many studies also have proved that time series of soil moisture estimations can be successfully used as an ANN model input (Anctil *et al.* 2004; De Vos & Rientjes 2005; Nilsson *et al.* 2006; Humphrey *et al.* 2016). Anctil *et al.* (2004) combined the conceptual model with ANN to enhance the runoff forecasting performance. The ANN model was first optimized by using streamflow and rainfall as inputs, and the soil moisture calculated through the lumped conceptual rainfall–runoff model GR4J was used as an auxiliary input. Results revealed that the soil moisture estimation is useful to improve 1-day-ahead stream flow forecasting. Nilsson *et al.* (2006) also came to the conclusion that the soil moisture and snow accumulation simulated from the conceptual model can provide useful information for the ANN model to improve simulation skills. Recently, the possibility of considering soil moisture simulated from the lumped GR4J rainfall–runoff model as an ANN additional input to enhance streamflow forecast performance was also investigated by Humphrey *et al.* (2016). The results showed that adding the simulated soil moisture as ANN input can provide additional independent information to represent initial catchment conditions. Therefore, simulated soil moisture data from the GR4J conceptual model together with rainfall and runoff are considered as ANN model inputs in this study.

The quality of the input data has a great influence on the ANN model performance (Humphrey *et al.* 2016). However, the hydrological time series obtained from available observations or simulated from hydrological models are usually polluted by various noises which may result in poor model performance. Reducing noise of the input series can be effective for improving the model predictability. Therefore, a signal filter technique for the purpose of filtering out noise in the original hydrological series is required. Singular spectrum analysis (SSA), as an alternative efficient signal filter technique, has been successfully applied in the hydrology field (Sivapragasam *et al.* 2001; Zhang *et al.* 2011; Wang *et al.* 2014, 2015). The SSA is a non-parametric time series analysis technique (Golyandina *et al.* 2001). By applying the SSA algorithm to the input data, the noise may be removed and a filtered series can be generated for model input which can result in the improvement of the model performance. Sivapragasam *et al.* (2001) employed SSA to filter the input series first and then used a SVM model to learn the filtered series for rainfall and runoff predictions. The results demonstrated that the SVM model performance can be significantly improved by SSA data preprocessing techniques. Wu *et al.* (2009) investigated the effects of SSA on five models' performances, and found that SSA can greatly improve each model performance for 1-month-ahead forecasting. Zhang *et al.* (2011) proposed a hybrid model of autoregressive integrated moving average (ARIMA) coupled with SSA for annual runoff forecasting, which resulted in a great model performance.

Based on the above, the main objectives of this study are: (1) to establish an ANN model with soil moisture data simulated from GR4J model as an additional input, and test the influences of considering estimated soil moisture on ANN model performance; (2) to employ SSA to filter the noise of all the input series and generate cleaner inputs for ANN to enhance the ANN model performance.

Three watersheds in China were selected as case studies and the performances of the different ANN models are compared and discussed.

This paper is organized as follows. Following the Introduction, the SSA method, ANN model structure, and the GR4J conceptual model are introduced in detail. The next section describes the study areas and data, the model inputs selection, and the implementation of SSA. Then, the main results are given along with the necessary discussion and, finally, the main conclusions are summarized.

METHODS

Singular spectrum analysis

Single spectrum analysis is well known as a nonparametric time series spectrum analysis technique, which is able to extract the trend components, harmonic components, and noise components from a time series. In this study, SSA, as a data preprocessing method, is applied to filter out the noise components or the high frequency components to obtain the filtered series. For the implementation of SSA, refer to Vautard *et al.* (1992) and Golyandina *et al.* (2001). The basic SSA consists of four steps, namely, embedding, singular value decomposition (SVD), grouping, and reconstruction. The concrete steps of SSA are illustrated as follows.

Embedding

Let the original time series be a nonzero series $x = \{x_1, x_2, \cdots, x_N\}$, the length of x is N. The first step is to transfer the time series x to a sequence of multidimensional lagged vectors. Given L $(1 < L < N)$, called window length, let $K = N-L+1$, L-lagged vector $X_i = (x_j, x_{j+1}, \cdots, x_{j+L-1})^T$, $j = 1, 2, \ldots, K$ is defined. Then, the columns of $(L \times K)$ trajectory matrix $\boldsymbol{X}$ are generated through the K vectors X_i, which is denoted by:

$$\boldsymbol{X} = [X_1, X_2, \cdots, X_K] = \begin{bmatrix} x_1 & x_2 & \cdots & x_{i+1} & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{i+2} & \cdots & x_{K+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_L & x_{L+1} & \cdots & x_{i+L} & \cdots & x_N \end{bmatrix} \quad (1)$$

Considering Equation (1), the trajectory matrix $\boldsymbol{X}$ has an important property that the elements on the anti-diagonals of $\boldsymbol{X}(i+j = \text{constant})$ are equal. It means that the trajectory matrix is a Hankel matrix. In the embedding procedure, the window length L is an important parameter.

Singular value decomposition

In this stage, the SVD of the trajectory matrix $\boldsymbol{X}$ is performed. Let $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^T$, which is covariance matrix, the eigenvalues, $\lambda_1, \lambda_2, \cdots, \lambda_L$ and eigenvectors, $\boldsymbol{U}_1, \boldsymbol{U}_2, \cdots, \boldsymbol{U}_L$ of S then are computed. The eigenvalues are sorted in a decreasing order $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L \geq 0)$. Then, the SVD of the trajectory matrix $\boldsymbol{X}$ can be expressed as follows:

$$\boldsymbol{X} = \boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_L \quad (2)$$

where $\boldsymbol{X}_i = \boldsymbol{U}_i\sqrt{\lambda_i}\boldsymbol{V}_i^T$ are elementary matrices and have rank 1, $\boldsymbol{V}_i = \boldsymbol{X}_i^T\boldsymbol{U}_i/\sqrt{\lambda_i}$ (equal to the *ith* eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$) are the *ith* left singular vectors of $\boldsymbol{X}$, whereas, $\boldsymbol{U}_i$ attained by calculating the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$ are the *ith* right singular vectors of the $\boldsymbol{X}$, $\sqrt{\lambda_i}$ are the *ith* singular value of $\boldsymbol{X}$ (equal to the square root of the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^T$). The collection $\{\boldsymbol{U}_i, \lambda_i, \boldsymbol{V}_i\}$ is referred to as *ith* eigentriple of the trajectory matrix $\boldsymbol{X}$.

Eigentriple grouping

One can identify the trend component, harmonic component, and noise component by grouping component in this step. If one does not want to accurately extract the hidden information, this step can also be skipped.

In the grouping procedure, one divides the submatrices into m disjoint subsets of $I_1, I_2, \ldots, I_m$. Let $I_k = \{i_{k,1}, \ldots, i_{k,p}\}$, the resultant matrix $\boldsymbol{X}_{I_k}$ related to the group I_k can be defined as $\boldsymbol{X}_{I_k} = \boldsymbol{X}_{i_{k,1}} + \cdots + \boldsymbol{X}_{i_{k,p}}$. Then, the trajectory matrix can be regrouped by the sum of m resultant matrices.

$$\boldsymbol{X} = \boldsymbol{X}_{I_1} + \cdots + \boldsymbol{X}_{I_m} \quad (3)$$

The selection of the sets $I_1, I_2, \ldots, I_m$ is referred to as eigentriple grouping.

Diagonal averaging

The last step is to transform each resultant matrix of the grouped decomposition into a new one-dimensional series with the same length of the original time series. This procedure is called diagonal averaging. It is assumed that the $(L \times K)$ resultant matrix in Formula (3) with elements y_{ij}, $1 \le i \le L$, $1 \le j \le K$, the k elements of the new series is equal to the average of the i row j column elements in the resultant matrix where $i + j = k + 1$. The equation of diagonal average can be expressed as follows:

$$y_k = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} y_{m,k-m+1} & 1 \le k < L^* \\ \dfrac{1}{L^*}\displaystyle\sum_{m=1}^{L^*} y_{m,k-m+1} & L^* \le k \le K^* \\ \dfrac{1}{N-k+1}\displaystyle\sum_{m=k-K^*+1}^{N-K^*+1} y_{m,k-m+1} & K^* < k \le N \end{cases} \quad (4)$$

where $L^* = \min(L, K)$, $K^* = \max(L, K)$. Then, diagonal averaging of each resultant matrix will form m new series $R_1, R_2, \ldots, R_m$ with the length of N, thus the original time series x is decomposed into the sum of the m reconstruction components:

$$x = R_1 + R_2 + \cdots + R_m \quad (5)$$

Through the data preprocessing technique of SSA, these construction components can be associated with trend component, harmonic component, and noise component with the proper window length L and the sets of $I_1, I_2, \ldots, I_m$. Certainly, without the third step, the original time series will be decomposed to L reconstructed components (RCs).

Artificial neural network

The ANN model, as a 'black-box' model, has a flexible mathematic structure and is a robust approach for modeling the complex nonlinear rainfall–runoff process. According to the direction of information flow, the ANNs can be grouped into two types, namely, feed-forward networks and recurrent networks (Govindaraju 2000a). The multilayer perceptron and the radial basis function belong to feed-forward network types. In the feed-forward networks, the information transmits in one direction through the network from input layer, hidden layer, and finally to the output layer (Dawson & Wilby 2001). The feed-forward multilayer perception network type is one of the most suitable types of ANN in the hydrology field, which usually uses the error backpropagation algorithm for training the network. A three-layer feed-forward ANN structure that was used in this study to model the rainfall–runoff process is shown in Figure 1.

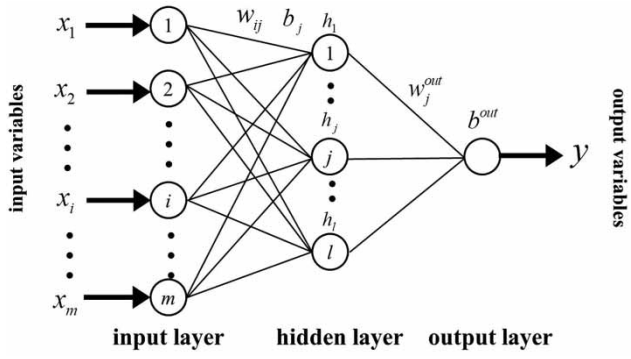

Figure 1 | Schematic diagram of three-layer feed-forward ANN.

As shown in Figure 1, the ANN consists of an input layer with m input nodes, a hidden layer with l hidden nodes, and an output layer with one output node. m is equal to the number of the model inputs, l is determined by a trial-and-error method, the ANN model only has one output, i.e., daily runoff at outlet of the watershed. For the node h_j in the hidden layer, it receives the incoming signals from each node x_i in the input layer. The effective incoming signal, defined as h'_j, can be calculated as:

$$h'_j = b_j + \sum_{i=1}^{m} w_{ij} \cdot x_i, \quad 1 \le j \le l \quad (6)$$

where b_j is the bias associated with the node h_j in the hidden layer, w_{ij} is the weight associated with the connection between the input node x_i and the hidden node h_j.

The outgoing signal of the node h_j can be produced through a nonlinear activation function. The tangent sigmoid function is applied in the hidden layer in this study, which is defined as:

$$h_j = f(h'_j) = \frac{2}{1 + \exp(-2 \cdot h'_j)} - 1 \quad (7)$$

Then, the output of the ANN model can be calculated as:

$$y = b^{out} + \sum_{j=1}^{l} w_j^{out} \cdot h_j \tag{8}$$

where b^{out} is the bias at the output node, w_j^{out} is the weight associated with the connection between the hidden node h_j and the output node.

The purpose of the network training is to adjust the interconnection weights and bias to minimize the error between the ANN output and the desired output. The Levenberg–Marquardt training algorithm is chosen here for the network training to obtain the optimal values of the interconnection weights and bias. For the ANN model simulations, in order to avoid the effect of different scales of input variables and ensure that all the variables can receive equal attention during the training process, all input variables should be normalized (Maier & Dandy 2000). Besides, to ensure extrapolation ability of the ANN model, all data are normalized in the range of [−0.9, 0.9] instead of [−1, 1] (Dawson & Wilby 2001). A linear transformation formula is defined as:

$$x_i' = -0.9 + \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} * 1.8 \tag{9}$$

where x_i' are the normalized input, x_i are the original input, $x_{\max}$ are the maximum of the original input data, $x_{\min}$ are the minimum of the original input data.

Soil moisture estimation

Anctil *et al.* (2004) and Humphrey *et al.* (2016) have shown that the time series of soil moisture simulated from the lumped GR4J rainfall–runoff model to represent initial catchment condition can be used as ANN input to improve the model performance. Therefore, a simple soil moisture reservoir, taken from the GR4J model suggested by Perrin *et al.* (2003), was adopted here. The daily rainfall and potential evaporation observations are used to calculate the soil moisture. The schematic diagram of the conceptual soil reservoir derived from the GR4J model is shown in Figure 2. The reservoir maximum capacity is A, which is the only parameter required to be estimated. Based on the precipitation intensities P_t are weaker or stronger than the potential

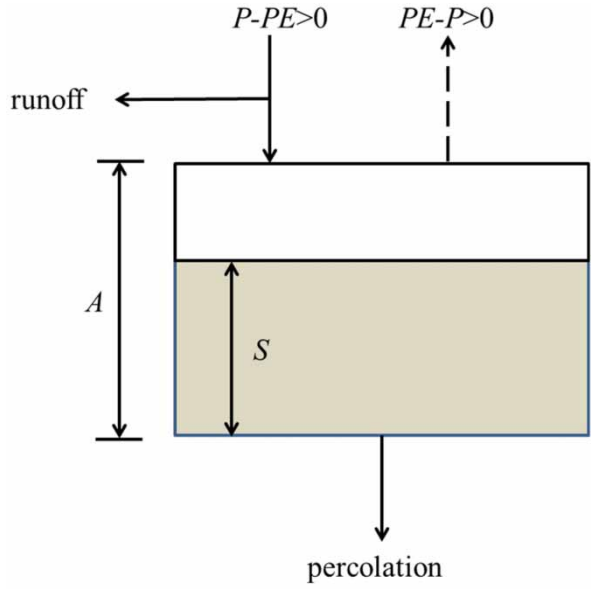

Figure 2 | Schematic diagram of the conceptual soil reservoir derived from the GR4J lumped rainfall–runoff model.

evapotranspiration PE_t, the soil water content S inside the reservoir is calculated by different formulas.

If $P_t \le PE_t$, the water in the soil moisture reservoir will be taken away by actual evapotranspiration (AE), then the soil water content S^* can be calculated as follows:

$$\begin{aligned} S* &= S_{t-1} - AE_t \\ &= S_{t-1} - \frac{S_{t-1}(2A - S_{t-1}) \tanh\left((PE_t - P_t)/A\right)}{A + (A - S_{t-1}) \tanh\left((PE_t - P_t)/A\right)} \end{aligned} \tag{10}$$

If $P_t > PE_t$, a portion of the effective rainfall (I) will supplement the soil moisture reservoir, then S^* can be obtained according to the following formula:

$$S* = S_{t-1} + I_t = S_{t-1} + \frac{(A^2 - S_{t-1}^2) \tanh\left((P_t - PE_t)/A\right)}{A + S_{t-1} \tanh\left((P_t - PE_t)/A\right)} \tag{11}$$

where S^* can never exceed the reservoir maximum storage capacity A. Considering the outflow from the reservoir due to percolation, the soil moisture S^* is thus adjusted using the following formula to get S_t:

$$S_t = S^* \left[1 + \left(\frac{4S^*}{9A}\right)^4\right]^{-1/4} \tag{12}$$

The conceptual soil moisture reservoir used here to obtain simulated soil moisture data requires only one parameter, i.e., the reservoir maximum capacity A, which can be optimized by trial and error method.

Proposed models

In general, rainfall and antecedent runoff are commonly used as potential predictors to the ANN model for runoff forecasting. In this study, simulated soil moisture data from the GR4J conceptual model as an important predictor to represent initial catchment conditions are also considered as the ANN model inputs. The purpose of this study is to test the runoff forecasting improvement resulting from the inclusion of simulated soil moisture input and data preprocessing techniques according to four model performance indices. Four models, namely ANN1, ANN2, SSA-ANN1, and SSA-ANN2 are proposed in this study. A flowchart of these four models is shown in Figure 3. The ANN1 model using rainfall and runoff data only is selected as the benchmark, while the ANN2 model is fed by an additional soil moisture input. The SSA-ANN1 and SSA-ANN2 are generated by using the same input data series preprocessed by SSA techniques.

Model performance evaluation criteria

To provide comprehensive assessments on model performance, Legates & McCabe (1999) recommended that apart from widely used goodness-of-fit measures, at least one absolute error measure (e.g., *RMSE* or *MAE*) should be included to supplement model evaluation criteria. Therefore, in this study, four performance evaluation criteria, Nash–Sutcliffe efficiency (*NSE*), root mean square error (*RMSE*), persistence index (*PI*), and mean absolute error (*MAE*) are selected. These formulas are given as follows:

1. Nash–Sutcliffe efficiency (*NSE*) defined by Nash & Sutcliffe (1970) has been widely used to evaluate the goodness-of-fit of hydrologic models, which is defined as:

$$NSE = \left(1 - \frac{\sum_{t=1}^{n} (Q_{ot} - Q_{ft})^2}{\sum_{t=1}^{n} (Q_{ot} - \overline{Q_{ot}})^2}\right) \times 100\% \tag{13}$$

   where n is the length of observed flow series, Q_{ft} and Q_{ot} represent the forecasted and observed flows at time t, respectively, $\overline{Q_{ot}}$ is the average value of observed flows. The *NSE* value ranges from $-\infty$ to 1. The closer the value of *NSE* to 1, the better the model performs.
2. Root mean square error (*RMSE*) is used to measure the differences between the forecasted values and the observed values. The *RMSE* can provide an evaluation of the error in the units of the flows which often can provide more information about the model performance. It is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (Q_{ot} - Q_{ft})^2} \tag{14}$$

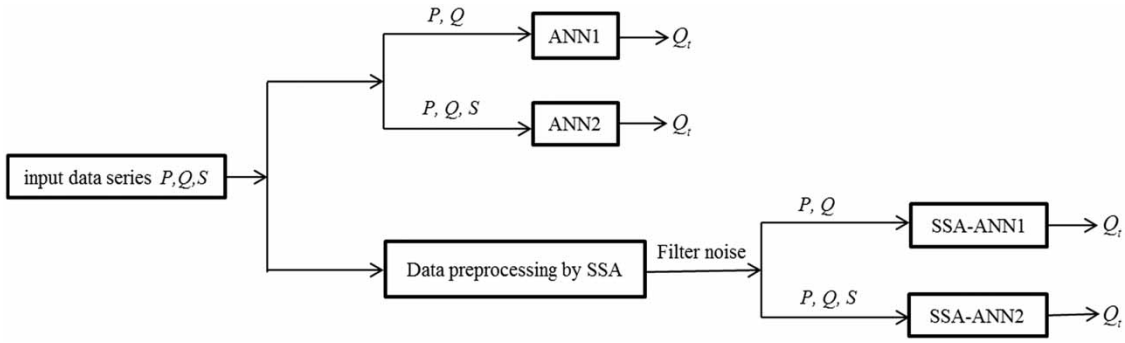

Figure 3 | Flowchart of the four proposed models.

The *RMSE* value closer to 0 indicates a better prediction.

3. Persistence index (*PI*) defined by Kitanidis & Bras (1980) is also adopted here for assessing the prediction lag effect. It is defined as:

$$PI = 1 - \frac{\sum_{t=1}^{n} (Q_{ot} - Q_{ft})^2}{\sum_{t=1}^{n} (Q_{ot} - Q_{ot-1})^2} \quad (15)$$

where Q_{ot-1} is the runoff estimation from a so-called persistence model that basically takes the last runoff observation (at time t minus the lead time T) as a prediction. If the *PI* equals to 1 it reflects a perfect fit between forecasted values and observed values.

4. Mean absolute error (*MAE*). As an absolute error measure, it is used here to describe the difference between the forecasted values and observed values. The average absolute error is given by:

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |Q_{ot} - Q_{ft}| \quad (16)$$

A smaller *MAE* value means a better prediction.

CASE STUDY

Study area and data

Two watersheds, as shown in Figure 4, were selected for case studies. The Baohe and Mumahe watersheds are both located in the upper Han River basin, which is the source

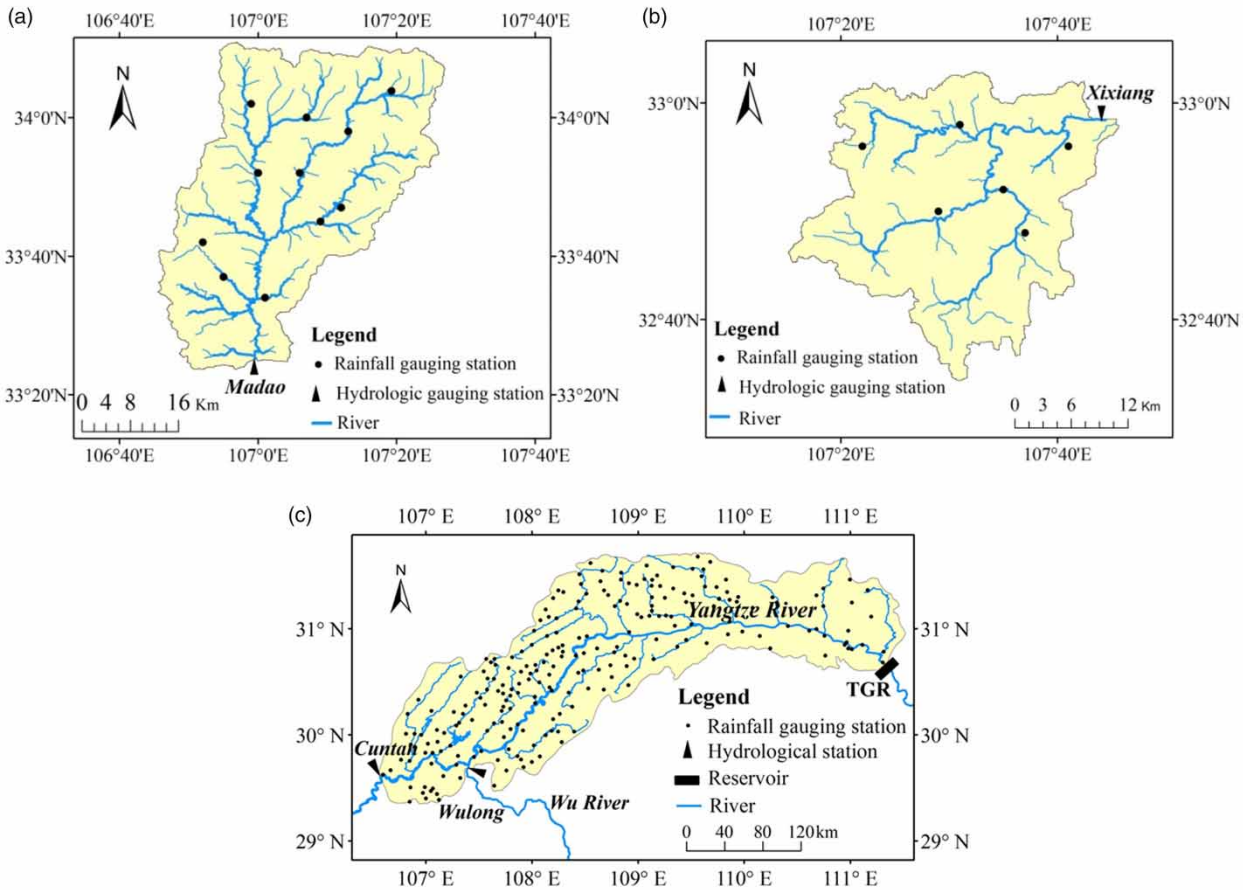

Figure 4 | Sketch maps of (a) Baohe and (b) Mumahe watersheds as well as (c) the TGR intervening basin.

of water of the South-to-North Water Diversion Project in China.

Baohe watershed, a tributary of the left bank of the upper Han River, is located in the southwest of Shanxi province. The watershed with a drainage area of 3,415 km^2 belongs to the humid region. The mean annual rainfall and runoff are 910 mm and 429 mm, respectively. The Mumahe watershed is a tributary of the right bank of the Han River. The mean annual rainfall and runoff are 1,070 mm and 687 mm, respectively. The watershed area controlled by the Xixiang hydrological station is 1,224 km^2. Daily rainfall, runoff, and potential evaporation data from 1981 to 1990 in these two watersheds were collected. The observed rainfall data from 11 and six rainfall stations in the Baohe and Mumahe watersheds were used to calculate the areal average rainfall, respectively. The used soil moisture data were calculated through the GR4J lumped conceptual rainfall–runoff model with the daily rainfall data and potential evaporation data.

The Three Gorges Reservoir (TGR) intervening basin located at the upper stream of the Yangtze River was also selected in this study. The upper Yangtze River is intercepted by the TGR with a drainage area of 10^6 km^2. The TGR is the largest and most important water conservancy project in China with comprehensive benefits of flood control, power generation, and navigation improvement. The TGR intervening basin which is 658 km long has a drainage area of 55,907 km^2. The inflow of TGR consists of three components, the main upstreamflow inflow controlled by Cuntan gage station, the tributary inflow from the Wu River controlled by Wulong gage station, and the rainfall runoff from the TGR intervening basin as shown in Figure 4. The observed streamflows of Cuntan and Wulong hydrological stations from 2004 to 2015 (12 years) are available for the TGR inflow forecasting. The observed rainfall data from all the rainfall gauging stations in the TGR intervening basin are used to calculate the areal average rainfall. The collected daily rainfall and potential evaporation observation data series are used to estimate the soil moisture data series through the GR4J lumped conceptual rainfall–runoff model.

In the process of the establishment of the model, the whole data set was divided into three parts, i.e., training set, validation set, and testing set. First, the training set is used for training the network to obtain a number of different networks. Then, the validation set is used to simulate the performance of models built in the training stage and select the best model. The testing set is then used to evaluate the selected model performance. For the Baohe and Mumahe watersheds, the data from 1981 to 1990 (10 years) are available. The first 6 years' data are used for training, the next 2 years for validation, and the remaining 2 years for testing. For the TGR intervening basin, the data from 2004 to 2015 (12 years) are used in this study. The training set includes the first 6 years' data, the validation set the next 3 years' data, and the testing set the remaining 3 years' data.

ANN model formulation and inputs

Before establishing the ANN model structure, one should determine appropriate lags for each input variable. The statistical analysis techniques based on the linear correlation method including cross-, auto-, and partial-auto-correlation of the data series were suggested by Sudheer *et al.* (2002) for identifying the appropriate input vector. Wu & Chau (2011) compared five different input selection methods including linear correlation analysis (LCA), average mutual information (AMI), partial mutual information (PMI), stepwise linear regression (SLR), multi-objective genetic algorithm (MOGA) and found out there is no significant difference among these methods. Therefore, considering the simplicity and easy calculation of the linear correlation analysis, the LCA method was adopted here to determine the suitable lags for each input variable.

For the Baohe and Mumahe watersheds, the auto-correlation function (ACF) of the runoff series was computed and the results are shown in Figure 5. It can be seen that the number of lags are selected as six and three for runoff input in the Baohe and Mumahe watersheds, respectively. The cross-correlation function (CCF) between the rainfall series, simulated soil moisture series and runoff series for various time lags were also calculated. Figure 5(c) and 5(d) illustrate the previous five and four rainfall observations have large CCF values (correlation coefficient >0.3) for the Baohe and Mumahe watersheds, respectively. As well, from Figure 5(c), it can be observed that the rainfall at lag 1 has the largest correlation coefficients with

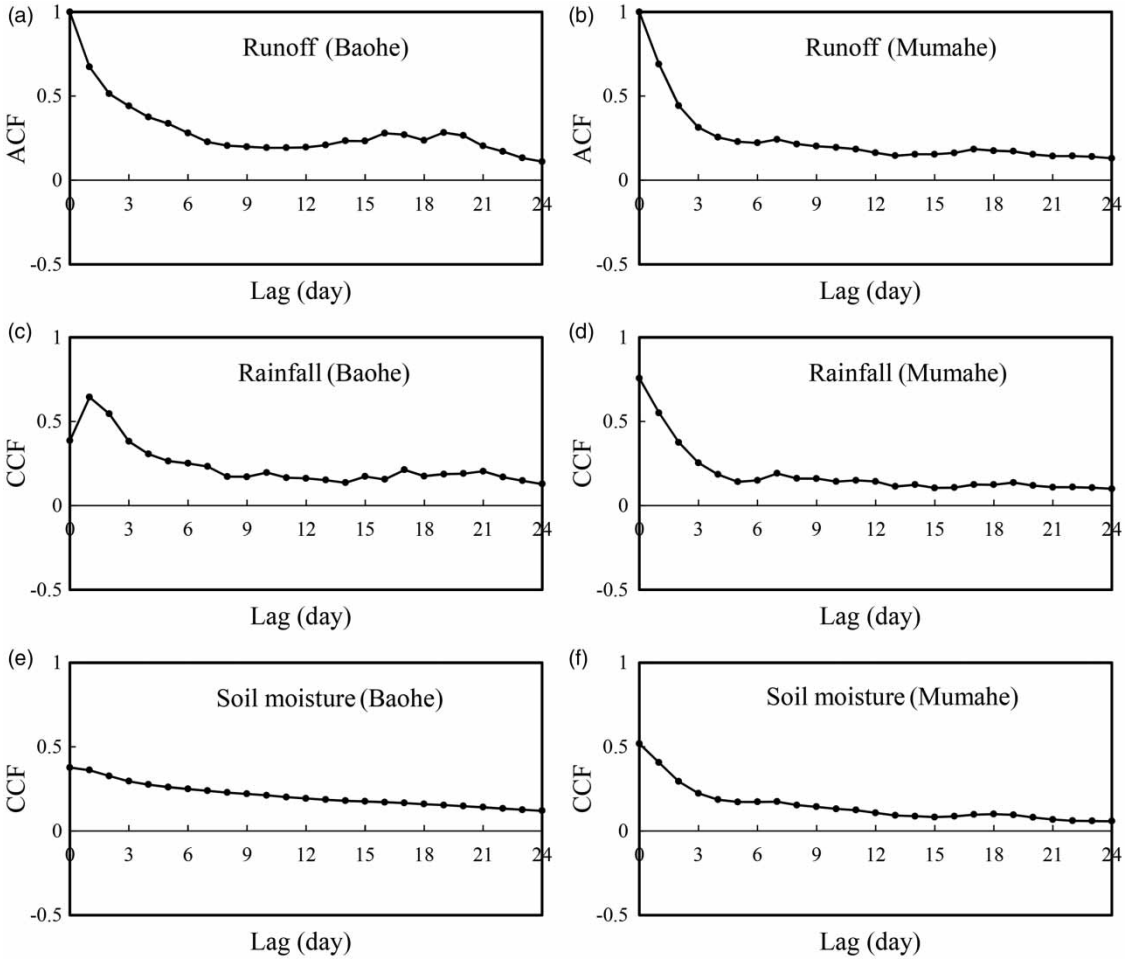

Figure 5 | Cross-correlation function (CCF) values and auto-correlation function (ACF) values for each input variable for the (a), (c) and (e) Baohe and (b), (d) and (f) Mumahe watersheds.

present runoff, which indicates that runoff concentration time of Baohe watershed is 1 day. Figure 5(e) and 5(f) show that the numbers of time lag are both 4 for the soil moisture variables.

For the TGR intervening basin, the input variables consist of Cuntan, Wulong inflows, areal average rainfall, and simulated soil moisture. Due to the high auto-correlation relation of TGR reservoir inflow series, the antecedent TGR reservoir inflow has a significant influence on present inflow forecasting. Therefore, the TGR reservoir inflow is also selected as potential predictor. The ACF values of the reservoir inflow series and the CCF values between the first four input variables and the present reservoir inflow were computed. Similarly to the above analysis, the numbers of time lag are finally taken as the values of 2, 2, 2, 4, and 2 for reservoir inflow, Cuntan inflow, Wulong inflow, rainfall, and soil moisture variables, respectively.

SSA decomposition of raw data

Before applying the ANN model, the SSA data preprocessing techniques are used to remove noise from the raw input time series and obtain the filtered series for model input. The procedure of SSA data preprocessing consists of SSA decomposition, selecting the contributed components, and RCs. In the process of the implementation of SSA decomposition, the only one parameter required to be identified is the window length L. Chau & Wu (2010) pointed out that when SSA, used as the signal filter technique, is just performed to remove noise components from the original

time series without the need to precisely extract potential trend and oscillations hidden in the original signal, a rough resolution can be sufficient for the extracting of effective information and noise. Therefore, a small interval [2,12] is examined to select L in this study (Wang *et al.* 2014). A target L can be identified only if the singular spectrum formed by the singular values can be distinguished markedly (Wu & Chau 2011). The singular spectrums at different window lengths L for all input time series in the Baohe and Mumahe watersheds are plotted in Figure 6. Taking rainfall time series in the Baohe watershed as an example, as shown in Figure 6(a), when L is larger than 8, the small singular values cannot be distinguished clearly, which means the noise components can be identified by SSA when L is set a value of 8 for rainfall series in the Baohe watershed. According to the above empirical determination, the final L are set at the values of 8, 9, and 10 for rainfall, runoff, and soil moisture series in the Baohe and Mumahe watersheds, respectively. Similarly, for all input series of the TGR intervening basin, the values of L are all set at the value of 10.

Once the appropriate window length L is determined, original time series can be separated into L RCs by SSA decomposition. In order to filter out noise and obtain the filtered series as model input, one should be able to search the effective RCs. The CCF filtering method, adopted by Wang *et al.* (2014), is also applied in this study to find the number of effective RCs because of its simplicity and convenience. The CCF values between each RC and original series under the determined window length L are computed and the results are listed in Table 1, where P is the number of effective RCs. Taking rainfall series in the Baohe

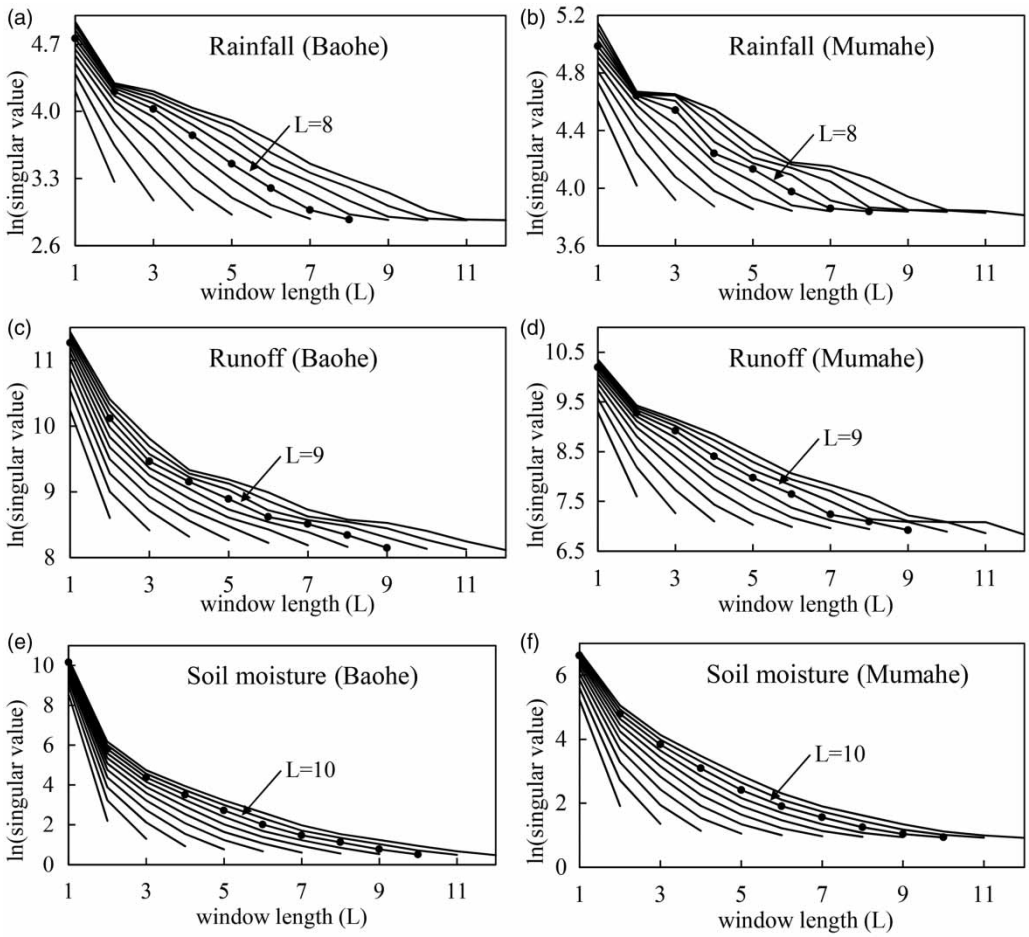

Figure 6 | Singular spectrum as a function of different window lengths L for the (a), (c) and (e) Baohe and (b), (d) and (f) Mumahe watersheds.

Table 1 | Cross-correlation function (CCF) values between each reconstructed component and original series

Watersheds		Reconstructed components										*L*	*P*
		1	2	3	4	5	6	7	8	9	10		
Baohe	P	−0.26	−0.26	−0.18	−0.04	**0.14**	**0.35**	**0.50**	**0.60**	–	–	8	4
	Q	−0.18	−0.20	−0.16	−0.08	**0.04**	**0.16**	**0.33**	**0.54**	**0.76**	–	9	5
	S	−0.02	−0.01	−0.01	−0.01	**0.01**	**0.02**	**0.04**	**0.08**	**0.21**	**0.99**	10	6
Mumahe	P	−0.34	−0.32	−0.22	−0.06	**0.13**	**0.34**	**0.47**	**0.52**	–	–	8	4
	Q	−0.16	−0.18	−0.13	−0.06	**0.05**	**0.21**	**0.39**	**0.56**	**0.72**	–	9	5
	S	−0.06	−0.07	−0.06	−0.03	**0.00**	**0.06**	**0.16**	**0.29**	**0.53**	**0.91**	10	6
TGR	P	−0.25	−0.27	−0.23	−0.14	−0.02	**0.11**	**0.23**	**0.40**	**0.48**	**0.50**	10	5
	Q_{CT}	−0.02	−0.02	−0.02	−0.01	**0.01**	**0.05**	**0.12**	**0.21**	**0.34**	**0.95**	10	6
	Q_{WL}	−0.02	−0.03	−0.03	−0.02	**0.02**	**0.07**	**0.15**	**0.30**	**0.47**	**0.91**	10	6
	Q_{IN}	−0.02	−0.03	−0.02	−0.01	**0.01**	**0.05**	**0.12**	**0.22**	**0.35**	**0.95**	10	6
	S	−0.01	−0.01	−0.01	−0.01	**0.00**	**0.02**	**0.04**	**0.09**	**0.21**	**0.99**	10	6

Note: Q_{CT}, Q_{WL}, and Q_{IN} represent the runoff at Cuntan and Wulong hydrological stations as well as TGR inflow.
Bold values denote the reconstructed components which have positive CCF values.

watershed as an example, with the determined window length of eight, original rainfall series is decomposed into eight RCs. As shown in Table 1, the last four RCs with positive correlation coefficients are considered as effective RCs, which mean these RCs can make positive contributions to the output of the model. Then the remaining RCs with negative CCF values are recognized as noise. Finally, the filtered series used as model inputs can be obtained by summing up the effective RCs. Similarly, the number of effective RCs for other input variables can be obtained, which are also listed in Table 1.

RESULTS' ANALYSIS AND DISCUSSION

In order to investigate the improvement of model performance by the inclusion of soil moisture estimation and data preprocessing techniques, four indices were selected to evaluate the proposed models. Table 2 summarizes the four model performances for three watersheds in different periods. The ANN1 model only with runoff and rainfall as inputs without data preprocessing was used as the benchmark. The effects of the inclusion of the simulated soil moisture and SSA data

Table 2 | Summary of model performance at different stages

Watersheds	Model	Input	NSE (%)			RMSE			MAE			PI		
			Train.	Valid.	Test.	Train.	Valid.	Test.	Train.	Valid.	Test.	Train.	Valid.	Test.
Baohe	ANN1	P + Q	75.49	82.25	76.23	77.7	23.1	47.1	14.3	11.4	13.3	0.608	0.671	0.725
	ANN2	P + Q + S	78.30	86.36	83.53	73.1	20.2	39.2	11.6	9.4	12.7	0.653	0.747	0.810
	SSA-ANN1	P + Q	91.92	92.31	90.20	44.6	15.2	30.2	9.2	6.6	9.3	0.871	0.857	0.887
	SSA-ANN2	P + Q + S	94.40	95.32	93.37	37.2	11.9	24.9	7.8	6.1	8.8	0.910	0.913	0.923
Mumahe	ANN1	P + Q	90.55	88.04	83.01	26.1	28.4	31.4	9.6	10.1	9.3	0.835	0.805	0.797
	ANN2	P + Q + S	92.51	92.38	88.73	23.2	22.7	25.6	9.3	9.8	8.4	0.869	0.876	0.866
	SSA-ANN1	P + Q	94.89	94.41	93.71	19.2	19.4	19.1	8.2	8.2	7.7	0.911	0.909	0.925
	SSA-ANN2	P + Q + S	96.07	96.44	95.10	16.8	15.5	16.9	6.9	6.2	6.8	0.931	0.942	0.942
TGR	ANN1	P + Q	98.68	98.44	98.03	1,035.3	1,269.6	1,202.3	537.6	642.9	646.2	0.730	0.783	0.649
	ANN2	P + Q + S	98.78	98.31	98.06	996.9	1,319.7	1,192.6	509.4	646.5	674.7	0.750	0.766	0.655
	SSA-ANN1	P + Q	99.50	99.14	98.85	636.1	940.6	915.9	372.8	497.1	509.6	0.898	0.881	0.796
	SSA-ANN2	P + Q + S	99.57	99.18	98.89	593.8	918.8	901.2	348.9	493.8	503.6	0.911	0.887	0.803

preprocessing on the ANN model performance are discussed as follows.

Analysis of the model inputs

To analyze the effect of inclusion of soil moisture, the model performance of the ANN2 was compared with that of ANN1. As shown in Table 2, in the Baohe watershed, the values of *NSE*, *PI* are 76.23%, 0.725 by ANN1 and increase to 83.53%, 0.810 by the ANN2 model, and the values of *RMSE*, *MAE* are 47.1, 13.3 by ANN1 and decrease to 39.2, 12.7 by ANN2 during testing periods; while in the Mumahe watershed, the values of *NSE, PI* are 83.01%, 0.797 by ANN1 and increase to 88.73%, 0.866 by the ANN2 model, and the values of *RMSE, MAE* are 31.4, 9.3 by ANN1 and decrease to 25.6, 8.4 by the ANN2 model during testing periods. It is shown that the ANN2 model produces more accurate forecasting results with much higher values of *NSE*, *PI* and lower values of *RMSE*, *MAE* than those of the ANN1 model. The ANN2 model, which includes the soil moisture simulated from the GR4J model as an additional input variable, performs better than the ANN1 model. Results indicate that the soil moisture time series can provide additional independent information for runoff forecasting. The scatter plots of daily runoff forecasting by ANN1 and ANN2 models are plotted in Figure 7. As shown in Figure 7, the ANN1 model has larger deviations from the 1:1 line than the ANN2 model in the Baohe and Mumahe watersheds. The observed and forecasted flow hydrographs in the Baohe and Mumahe watersheds are shown in Figure 8(a) and 8(b). It can be seen that the prediction of the ANN1 model in the Baohe watershed is slightly lagged in comparison with the observed runoff series. Moreover, it can be observed that the peak flows forecasted by the ANN2 model are much closer to the observed values than those by the ANN1 model. From all the above, conclusions can be drawn that the ANN2 model with rainfall, runoff, and soil moisture as inputs produces better forecasting results than the ANN1 model with only rainfall and runoff as inputs. For the TGR intervening basin, only a slight improvement can be found by the inclusion of soil moisture in terms of performance indices, in Table 2. A direct explanation for this phenomenon is that 90% of the TGR reservoir inflow is controlled by Cuntan and Wulong stations, and the influence of intervening basin soil moisture on reservoir inflow is weak.

Analysis of the data preprocessing by SSA

For the analysis of the effect of SSA data preprocessing, a comparison was made between the SSA-ANN2 and the ANN2 model. The results of SSA-ANN2 and ANN2 models are also listed in Table 2. A significant improvement can be seen by SSA data preprocessing techniques in terms of *NSE, RMSE, PI*, and *MAE* no matter what types of input. It is shown that the SSA-ANN2 model has much lower

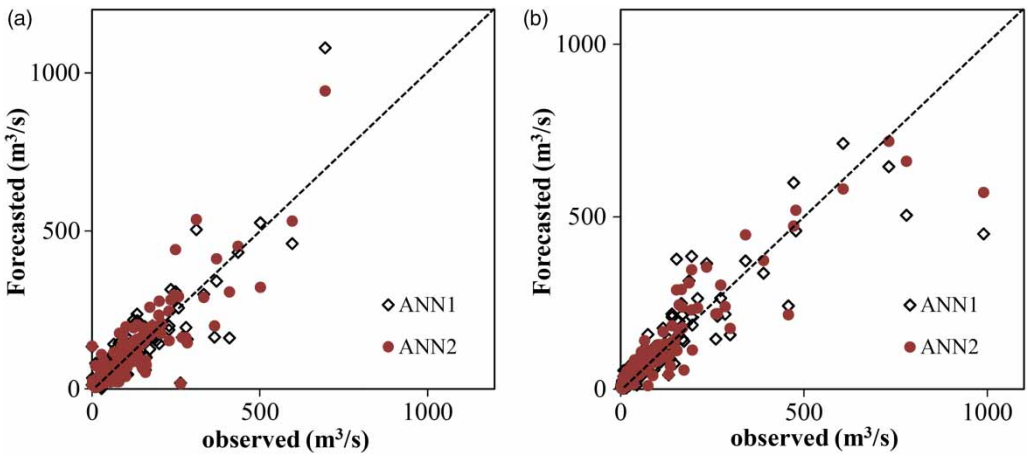

Figure 7 | Scatter plots of daily flow forecasting by the ANN1 and ANN2 models for the (a) Baohe and (b) Mumahe watersheds.

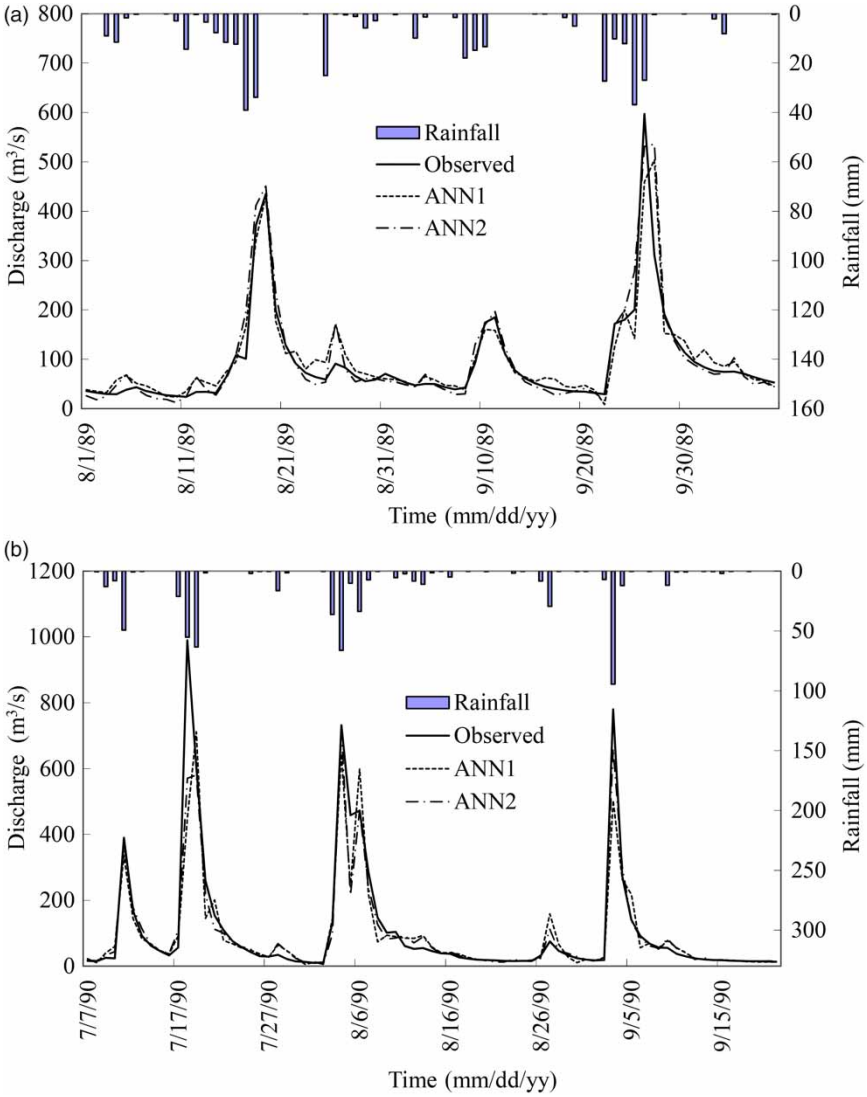

Figure 8 | Observed and forecasted flow hydrographs by the ANN1 and ANN2 models for the (a) Baohe and (b) Mumahe watersheds.

RMSE and *MAE* values and higher *NSE* and *PI* values than those of the ANN2 model. Moreover, the performance obtained from the SSA-ANN2 model is consistently superior to that of the ANN2 model according to all indices in different periods. In the Baohe and Mumahe watersheds, the values of *NSE* by the SSA-ANN2 model increased 8.96% and 9.84%, 4.06% and 6.37%, in comparison with the ANN2 model during the validation and testing periods, respectively. For the TGR intervening basin, the values of *NSE* increased from 98.31% and 98.06% by ANN2 model to 99.18% and 98.89% by the SSA-ANN2 model during the validation and testing periods, respectively. It is easy to find that the original ANN2 model performance has been improved successfully by SSA data preprocessing techniques with high model efficiency and small relative errors. The reason for this significant improvement in model performance is that the cross-correlation between the input series and output series is improved by SSA, which means that the filtered series obtained by SSA decomposition can provide more useful information for the output series than the raw series. Figure 9 illustrates the scatter plots of daily runoff forecasting by the ANN2

and SSA-ANN2 models. It can be observed that SSA-ANN2 performs better than ANN2 without data preprocessing due to the scatter plots of SSA-ANN2 models with a low spread are much closer to the 1:1 line than those of the ANN2 models. The observed and forecasted flow hydrographs by ANN2 and SSA-ANN2 models are shown in Figure 10. It is observed that the ANN2 model underestimates a few peak flows whereas the SSA-ANN2 model captures the time and values of peak flow well. The flood peaks and hydrographs simulated by the SSA-ANN2 model are much closer to the observations than those by the ANN2 model. Therefore, conclusions can be drawn that the ANN model performance can be improved significantly by SSA data preprocessing technique.

Discussions

Through the above results' analysis, conclusions can be drawn that the soil moisture time series are useful for flow forecasting in the Baohe and Mumahe watersheds, whereas in the TGR intervening basin, only a slight improvement can be found. A catchment's response to a rainfall event is determined by not only rainfall intensity but also the hydrological state of a catchment. Therefore, the soil moisture as an important variable to represent initial catchment conditions can be used in ANN input successfully, which is also confirmed by Anctil *et al.* (2004). In addition, it demonstrates that SSA is an effective and powerful way to improve the model performance by filtering out the noise from the original data series. It is observed in Table 2 that the SSA-ANN2 model performs the best among these models in terms of four performance indices, and it is concluded that the ANN model performance can be improved by the inclusion of soil moisture and SSA. As well, it can be found that SSA-ANN1 performs much better than ANN2 in terms of four indices, which indicates that the degree of the improvement by SSA is more significant than by the inclusion of soil moisture as an additional model input.

CONCLUSIONS

The objective of this study was to investigate the effect of the inclusion of simulated soil moisture series and SSA on improving ANN model performance. An ANN model with soil moisture data simulated from a GR4J model as an auxiliary input variable was established. The SSA data preprocessing technique was adopted to remove the noise of the data series and generate cleaner inputs to enhance performance of the ANN model. Three watersheds from China were selected to test the proposed model performances. ANN model with only rainfall and runoff as inputs was used as the benchmark. The major findings of this study are summarized as follows:

1. The ANN2 model with soil moisture as an additional input is superior to the ANN1 model with only runoff

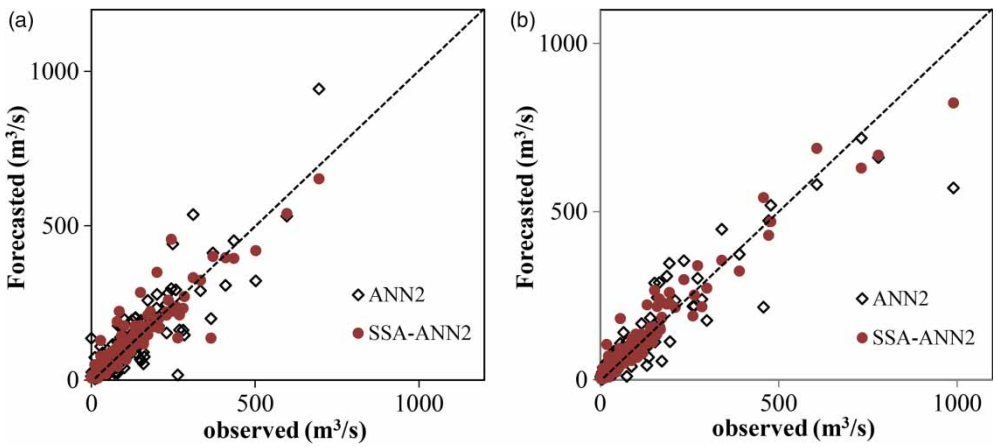

Figure 9 | Scatter plots of daily flow forecasting by the ANN2 and SSA-ANN2 models for the (a) Baohe and (b) Mumahe watersheds.

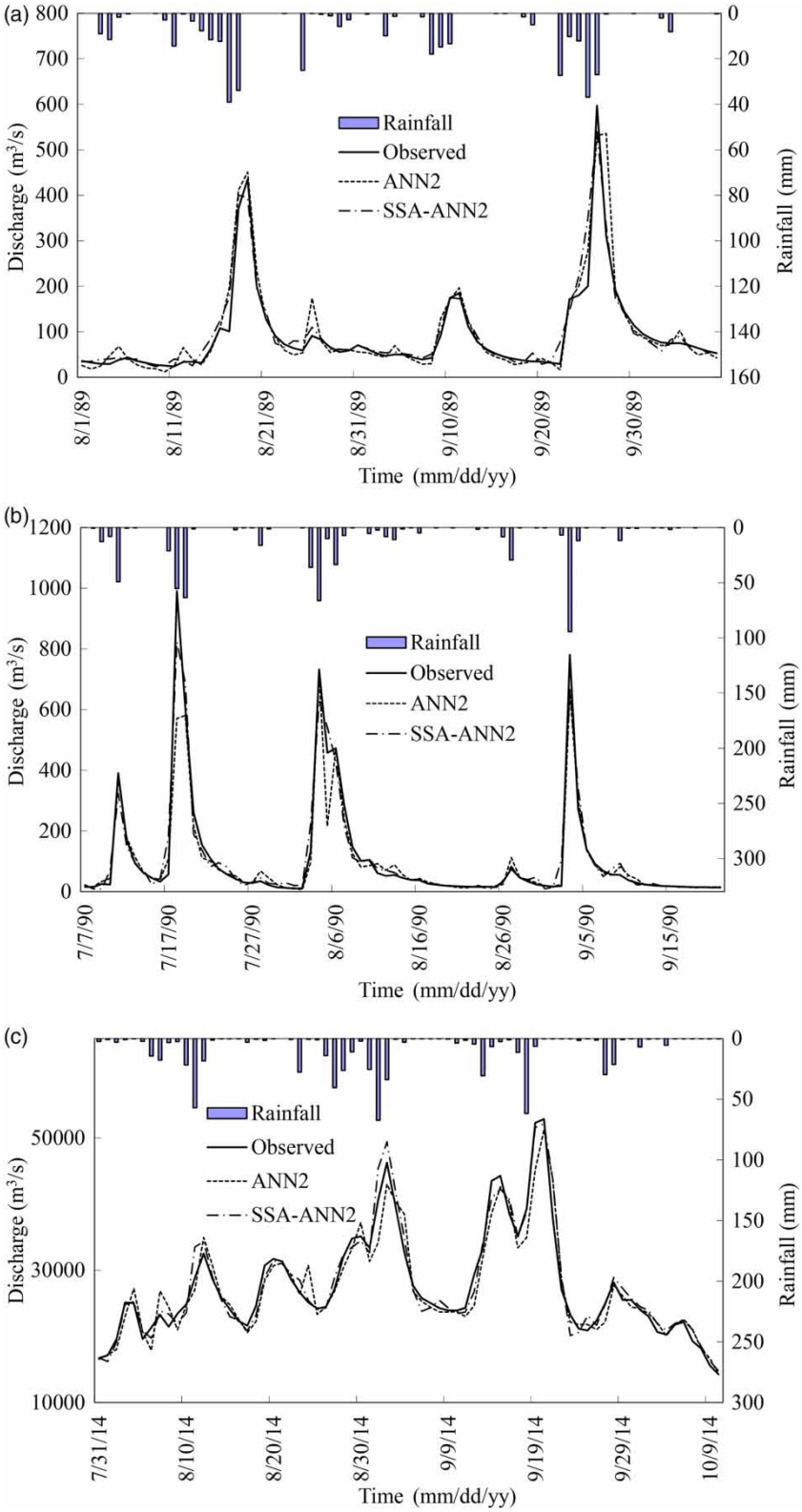

Figure 10 | Observed and forecasted flow hydrographs by the ANN2 and SSA-ANN2 models for the (a) Baohe and (b) Mumahe watersheds as well as (c) the TGR intervening basin.

and rainfall as inputs, which demonstrates that the inclusion of soil moisture estimation as an additional input can improve the ANN model accuracy.

2. The ANN model coupled with SSA data preprocessing techniques performs better than original ANN models according to all indices in different periods. It indicates that SSA is an effective and powerful way to improve the flow forecasting accuracy.
3. The SSA-ANN2 model with runoff, rainfall, and soil moisture as inputs and using data preprocessing technique performed the best. The degree of improvement by SSA was more significant than by the inclusion of soil moisture. Thus, the ANN model coupled with SSA is more promising for runoff forecast.

ACKNOWLEDGEMENTS

This study was financially supported by the National Natural Science Foundation of China (51539009) and the National Key Research and Development Plan of China (2016YFC0402206). We would like to thank the two anonymous referees for their instructive comments and Prof. Chang Fi-John for his technical English editing and proofreading, which have led to significant improvement in the presentation and quality of the paper.

REFERENCES

Aksoy, H. & Dahamsheh, A. 2009 Artificial neural network models for forecasting monthly precipitation in Jordan. *Stoch. Environ. Res. Risk A.* **23** (7), 917–931.

Anctil, F., Michel, C., Perrin, C. & Andréassian, V. 2004 A soil moisture index as an auxiliary ANN input for stream flow forecasting. *J. Hydrol.* **286** (1–4), 155–167.

Budu, K. 2013 Comparison of wavelet-based ANN and regression models for reservoir inflow forecasting. *J. Hydrol. Eng.* **19** (7), 1385–1400.

Chang, F. J. & Chang, Y. T. 2006 Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. *Adv. Water Resour.* **29** (1), 1–10.

Chang, F. J. & Tsai, M. J. 2016 A nonlinear spatio-temporal lumping of radar rainfall for modelling multi-step-ahead inflow forecasts by data-driven techniques. *J. Hydrol.* **535**, 256–269.

Chau, K. W. & Wu, C. L. 2010 A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *J. Hydroinform.* **12** (4), 458–473.

Chen, P. A., Chang, L. C. & Chang, F. J. 2013 Reinforced recurrent neural networks for multi-step-ahead flood forecasts. *J. Hydrol.* **497**, 71–79.

Cheng, C. T., Niu, W. J., Feng, Z. K., Shen, J. J. & Chau, K. W. 2015 Daily reservoir runoff forecasting method using artificial neural network based on quantum-behaved particle swarm optimization. *Water* **7** (8), 4232–4246.

Dawson, C. W. & Wilby, R. L. 2001 Hydrological modelling using artificial neural networks. *Prog. Phys. Geog.* **25** (1), 80–108.

De Vos, N. J. & Rientjes, T. H. M. 2005 Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation. *Hydrol. Earth Syst. Sci.* **2** (1), 365–415.

Gautam, M. R., Watanabe, K. & Saegusa, H. 2000 Runoff analysis in humid forest catchment with artificial neural network. *J. Hydrol.* **235** (1), 117–136.

Golyandina, N., Nekrutkin, V. & Zhigljavsky, A. 2001 *Analysis of Time Series Structure: SSA and the Related Techniques.* Chapman & Hall/CRC, Boca Raton, FL.

Gorgij, A. D., Kisi, O. & Moghaddam, A. A. 2017 Groundwater budget forecasting, using hybrid wavelet-ANN-GP modelling: a case study of Azarshahr Plain, East Azerbaijan, Iran. *Hydrol Res.* **48** (2), 455–467.

Govindaraju, R. S. 2000a Artificial neural networks in hydrology. I: Preliminary concepts. *J. Hydrol. Eng.* **5** (2), 115–123.

Govindaraju, R. S. 2000b Artificial neural networks in hydrology. II: Hydrological applications. *J. Hydrol. Eng.* **5** (2), 124–137.

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

Humphrey, G. B., Gibbs, M. S., Dandy, G. C. & Maier, H. R. 2016 A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network. *J. Hydrol.* **540**, 623–640.

Jain, A. & Kumar, A. M. 2007 Hybrid neural network models for hydrologic time series forecasting. *Appl. Soft Comput.* **7** (2), 585–592.

Kitanidis, P. K. & Bras, R. L. 1980 Real-time forecasting with a conceptual hydrologic model: 2. Applications and results. *Water Resour. Res.* **16** (6), 1034–1044.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **35** (1), 233–241.

Li, Y. L., Zhang, Q., Werner, A. D. & Yao, J. 2015 Investigating a complex lake-catchment-river system using artificial neural networks: Poyang Lake (China). *Hydrol. Res.* **46** (6), 912–928.

Li, X., Sha, J. & Wang, Z. L. 2017 A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. *Hydrol. Res.* **48** (5), 1214–1225. doi: 10.2166/nh.2016.149.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a

review of modelling issues and applications. *Environ. Modell. Softw.* **15** (1), 101–124.

Makwana, J. J. & Tiwari, M. K. 2014 Intermittent streamflow forecasting and extreme event modelling using wavelet based artificial neural networks. *Water Resour. Manage.* **28** (13), 4857–4873.

Minns, A. W. & Hall, M. J. 1996 Artificial neural networks as rainfall-runoff models. *Hydrolog. Sci. J.* **41** (3), 399–417.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *J. Hydrol.* **10** (3), 282–290.

Nilsson, P., Uvo, C. B. & Berndtsson, R. 2006 Monthly runoff simulation: comparing and combining conceptual and neural network models. *J. Hydrol.* **321** (1–4), 344–363.

Noori, N. & Kalin, L. 2016 Coupling SWAT and ANN models for enhanced daily streamflow prediction. *J. Hydrol.* **533**, 141–151.

Pang, B., Guo, S., Xiong, L. & Li, C. 2007 A nonlinear perturbation model based on artificial neural network. *J. Hydrol.* **333** (2), 504–516.

Perrin, C., Michel, C. & Andréassian, V. 2003 Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **279** (1), 275–289.

Rezaeianzadeh, M., Tabari, H., Yazdi, A. A., Isik, S. & Kalin, L. 2014 Flood flow forecasting using ANN, ANFIS and regression models. *Neural Comput. Appl.* **25** (1), 25–37.

Robertson, D. E. & Wang, Q. J. 2012 A Bayesian approach to predictor selection for seasonal streamflow forecasting. *J. Hydrometeorol.* **13** (1), 155–171.

Robertson, D. E., Pokhrel, P. & Wang, Q. J. 2013 Improving statistical forecasts of seasonal streamflows using hydrological model output. *Hydrol. Earth Syst. Sci.* **17** (2), 579–593.

Rosenberg, E. A., Wood, A. W. & Steinemann, A. C. 2011 Statistical applications of physically based hydrologic models to seasonal streamflow forecasts. *Water Resour. Res.* **47** (3), 1995–2021.

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

Sudheer, K. P., Gosain, A. K. & Ramasastri, K. S. 2002 A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrol. Process.* **16** (6), 1325–1330.

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

Tsai, M. J., Abrahart, R. J., Mount, N. J. & Chang, F. J. 2014 Including spatial distribution in a data-driven rainfall-runoff model to improve reservoir inflow forecasting in Taiwan. *Hydrol. Process.* **28**, 1055–1070.

Vautard, R., Yiou, P. & Ghil, M. 1992 Singular spectrum analysis: a toolkit for short, noisy and chaotic signals. *Physica D* **58**, 95–126.

Wang, Y., Guo, S., Chen, H. & Zhou, Y. 2014 Comparative study of monthly inflow prediction methods for the Three Gorges Reservoir. *Stoch. Environ. Res. Risk A.* **28** (3), 555–570.

Wang, Y., Guo, S., Xiong, L., Liu, P. & Liu, D. 2015 Daily runoff forecasting model based on ANN and data preprocessing techniques. *Water* **7** (8), 4144–4160.

Wu, C. L. & Chau, K. W. 2011 Rainfall–runoff modeling using artificial neural network coupled with singular spectrum analysis. *J. Hydrol.* **399** (3), 394–409.

Wu, C. L., Chau, K. W. & Li, Y. S. 2009 Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resour. Res.* **45** (8), 2263–2289.

Zealand, C. M., Burn, D. H. & Simonovic, S. P. 1999 Short term streamflow forecasting using artificial neural networks. *J. Hydrol.* **214** (1–4), 32–48.

Zhang, Q., Wang, B. D., He, B., Peng, Y. & Ren, M. L. 2011 Singular spectrum analysis and ARIMA hybrid model for annual runoff forecasting. *Water Resour. Manage.* **25** (11), 2683–2703.

First received 24 March 2017; accepted in revised form 16 June 2017. Available online 22 August 2017