Improving forecasting accuracy of river flow using gene expression programming based on wavelet decomposition and de-noising

Xiaorong Lu, Xuelei Wang, Liang Zhang, Ting Zhang, Chao Yang, XinXin Song and Qing Yang

ABSTRACT

Due to the effects of anthropogenic activities and natural climate change, streamflows of rivers have gradually decreased. In order to maintain reliable water supplies, reservoir operation and water resource management, accurate streamflow forecasts are very important. Based on monthly flow data from five hydrological stations in the middle and lower parts of the Hanjiang River Basin, between 1989 and 2009, we consider an efficient approach of adopting the gene expression programming model based on wavelet decomposition and de-noising (WDDGEP) to forecast river flow. Original flow time series data are initially decomposed into one sub-signal approximation and seven sub-signal details using the dmey wavelet. A wavelet threshold de-noising method is also applied in this study. Data that have been de-noised after decomposition are then adopted as inputs for WDDGEP models. Finally, the forecasted sub-signal results are summed to formulate an ensemble forecast for the original monthly flow series. A comparison of the prediction accuracy between the two models is based on three performance evaluation measures. Results show that the new WDDGEP models can effectively enhance accuracy in forecasting streamflow, and the proposed wavelet-based de-noising of the observed non-stationary time series is an effective measure to improve simulation accuracy.

Key words | flow forecasting, gene expression programming, wavelet transformation, WDDGEP models

Xiaorong Lu
Xuelei Wang (corresponding author)
Liang Zhang
Chao Yang
XinXin Song
Institute of Geodesy and Geophysics, Chinese Academy of Sciences,
Wuhan 430077, Hubei, China
and
Key Laboratory for Environment and Disaster Monitoring and Evaluation,
Wuhan 430077, Hubei, China
E-mail: *xueleiwang@whigg.ac.cn*

Xiaorong Lu
Ting Zhang
Chao Yang
XinXin Song
University of Chinese Academy of Sciences,
Beijing 100049, Beijing, China

Qing Yang
Key Laboratory of Pollution Ecology and Environmental Engineering,
Shenyang 110016, Liaoning, China

Ting Zhang
School of Resource and Environmental Science and Engineering, Hubei University of Science and Technology,
Xianning 437100, Hubei, China

INTRODUCTION

Hydrological time series forecasting is an important predictive approach for watershed management, maintaining efficient water supplies and reservoir regulation (Liu *et al.* 2014). Hydrological time series commonly contain different, complicated components (Osorio *et al.* 2014; Rebora *et al.* 2016), which have stochastic fuzzy, nonlinear, non-stationary and multi-temporal scale characteristics (Rebora *et al.* 2016), features which significantly increase the difficulty of hydrological prediction. Mathematical modelling of hydrological time series to reproduce the underlying stochastic structure of this type of hydrological process has been previously extensively performed. Randomness and deterministic always coexist and they are not separable or additive components, thus making it difficult to separate noise from inherent random variability. To improve prediction accuracy, a wider understanding of variable factors,

doi: 10.2166/nh.2017.115

such as long-term trends, seasonal changes, cyclic variations and irregular changes, is needed through probability and statistical models to identify the change of the variable distribution (Koutsoyiannis 2009). Although there is no absolute deterministic model in the simulation of hydrological time series, different models have different requirements for data stability. Conventional time series models, such as auto-regressive (AR), auto-regressive and moving average models (ARMA), and autoregressive integrated moving average models (ARIMA), having exogenous input and multiple linear regressions, are linear models that have a high requirement for data stability. These models, however, are not suitable for non-stationary data and non-linearity involved in hydrological processes (Farajzadeh *et al.* 2014). The recent use of artificial intelligence techniques and data-driven approaches in water resource engineering, such as genetic programming (GP) and genetic algorithms (GAs) based on evolutionary computing, have emerged as powerful modelling tools suitable for solving hydrological and water resource problems (Kisi *et al.* 2012). The basic idea of GP is similar to traditional GA, i.e. to randomly generate the initial population which are suitable for a given environment. However, the adaptive evolution of the simulation program of GP enhances structural complexity. As the spatial traversal of GPs is better than the traditional heuristic search, and the method of custom and dynamic process multiplexing are leaded into GP, GP algorithms have therefore increased in use in dynamic model systems (Yaseen *et al.* 2015). Gene expression programming (GEP), similar to GAs and GPs, exhibits better performance than other data-driven approaches, such as the artificial neural network (ANN) and adaptive neuro-fuzzy inference systems (ANFIS) (Pereira & Saraiva 2011); GEP is a simple and efficient method of evolution modelling which is widely used in many fields. The three different data-driven approaches (ANN, ANFIS and GEP) were applied to model rainfall-runoff, results of which indicated that the GEP model is superior to the ANFIS or ANN models (Nayak *et al.* 2013). To forecast monthly discharge time series, ANFIS techniques, ARMA models, GP models, support vector machine (SVM) methods and ANN approaches were applied, results of which indicated that GP, ANFIS and SVM provided the best performance (Wang *et al.* 2009). To estimate daily evaporation, the coupled ANFIS-GEP model was used as an alternative approach, results showing the GEP model to be more superior to the ANFIS model (Terzi 2013). In addition, a conceptual rainfall-runoff model was developed for hydropower plant site assessment in the Hurman River watershed, results of which showed strong agreement between simulated and observed data (Al-Juboori & Guven 2016).

Although many studies have suggested that the GEP model is a more appropriate alternative to other time series forecasting models, the accuracy of GEP models can still be improved. The wavelet transform (WT) technique has been recently utilised as an effective method to obtain further information about data characteristics of water resources and environmental time series. This technique has been used to quantify stream flow variability (Smith *et al.* 1998; Coulibaly & Burn 2004), to predict regional runoff and daily reservoir inflow (Coulibaly *et al.* 2000), and to analyse 55 large river discharge fluctuations (Labat 2008). Zhou *et al.* (2008), by using WT techniques, decomposed monthly discharge time series into a particular number of detail signals and an approximated signal; an ARMA model was then applied to this decomposed data to predict each WT sub-series.

However, the majority of studies on WGEP simulations are limited as they use data after wavelet decomposition as input data in the models; they do not consider effective de-noising methods. All hydrological time series data are, to some extent, contaminated by noise deriving from the measuring device, the influence of random intrinsic system events, or by feedback processes wherein the system is disturbed by a small random amount at each time step. This noise acts to limit the performance of many modelling and prediction techniques which can significantly influence the outcome of these methods. Since hydrologic series usually show complex non-stationary and nonlinear characteristics, traditional wavelet decomposition methods have many disadvantages and cannot meet specific accuracy requirements of prediction models. Therefore, an effective wavelet aided de-noising approach of hydrologic series is needed to improve the prediction accuracy of wavelet-assisted GEP models. This study presents river flow forecasting utilising WDDGEP models based on wavelet decomposition and de-noising, and thereby not only using the decomposed signal component as the model input. In

order to remove trends and seasonal components to isolate the stationary component, to fit the model and add seasonality and trend components to predict the non-stationary time series, the wavelet-assisted method was used to separate stationary components from unstable components of the flow time series. Original monthly flow time series data are decomposed into several sub-signal time series before wavelet de-noising methods are used with the flow data, and a conjunction model (WDDGEP) was developed to predict stream flow for the five hydrological stations. The performance of the WDDGEP models is also compared with results from simple GEP models developed without WT and de-noising for the research stations.

STUDY AREA AND DATA

The Hanjiang River, located in Hubei province, is the longest tributary of the Yangtze River in China. The Middle Route of the South-to-North Water Diversion Project is a significant strategic project that implements the optimal allocation of water resources in this area, aiming to alleviate water shortage problems in the north. This project also improves flood-control standards of the middle and lower sections of the Hanjiang River, and ensures regional water security. However, the implementation of the diversion project can result in severe water shortages in the middle and lower sections of the Hanjiang River: due to climate change and over-utilisation of water resources in the basin, flow in the middle and lower parts of the river have reduced.

Monthly flow data from five hydrological stations (Huangjiagang, Huangzhuang, Shayang, Xiantao and Xiangyang), located in different sections of the middle and lower parts of the Hanjiang River (Figure 1), were selected for the application of our model. The middle and lower sections of the Hanjiang River are located in a typical subtropical monsoon climate area, providing strong seasonal flow variation with flooding often occurring during the wet season (June–September). The flow rate in the river gradually recedes to extremely low levels during the dry

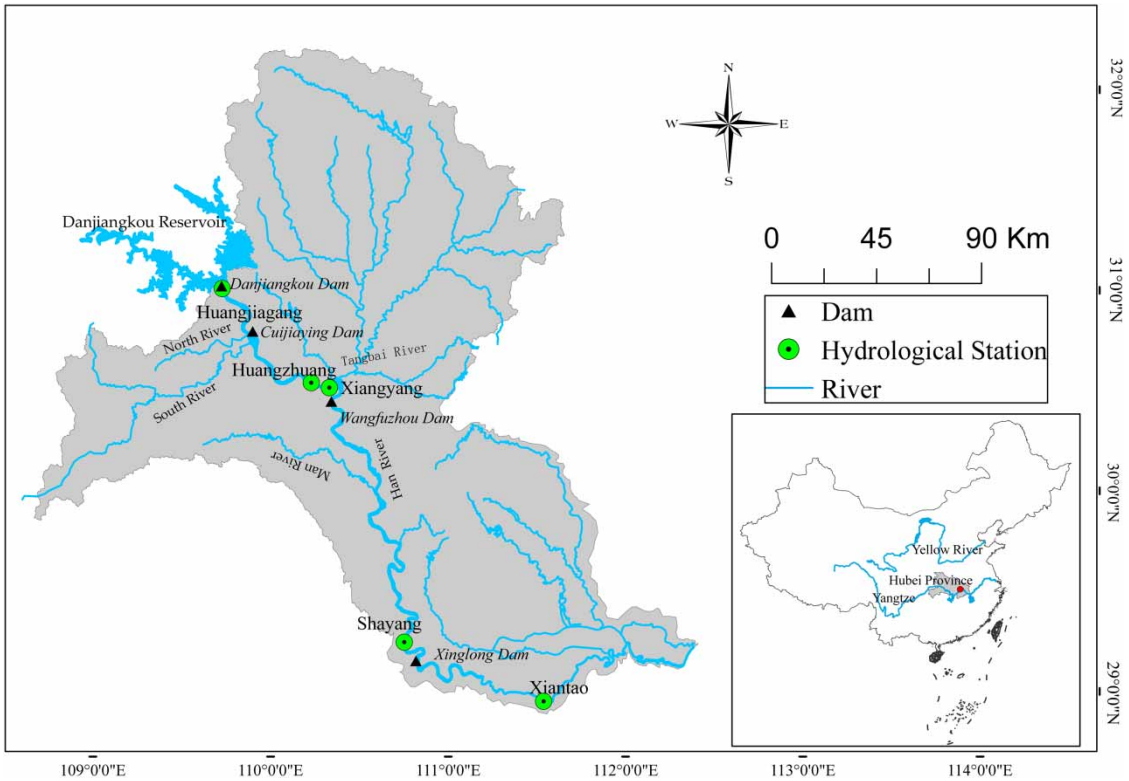

Figure 1 | Map of the middle and lower parts of the Hanjiang River catchment showing the location of the five hydrological stations.

season (November–May of the following year) until the rain starts again in June. The maximum average monthly flow in the wet period is 3365.84 m^3/s in Shayang, and the minimum average monthly flow in the dry period is 369.07 m^3/s in Huangjiagang. Monthly flow data from each hydrological station (1989–2009) were applied in two models (Table 1). Data from the first 172 months were used to train the models and data from the final 80 months were used to validate model performance. Stream flow data from each station show that the maximum monthly flow was 7558.67 m^3/s (Huangzhuang station) and the minimum monthly flow was 237.71 m^3/s (Huangjiagang station).

METHODOLOGY

Wavelet transform

Hydrologic time series forecasting, the central issue in hydrology and ecology, encounters many challenges due to influences from different uncertainties (Willems 2009), such as complicated factors, multi-timescale problems and noise in time series data. It is therefore very difficult to obtain accurate results from hydrologic time series predictions. Under these circumstances, WT and decomposition are widely used for hydrologic time series to improve forecasting accuracy (Guyennon *et al.* 2014; Shamshirband *et al.* 2016).

WT, a data mining tool applied in signal analysis, decomposes data flow into a number of sub-series of approximation and details. WT not only attempts to obtain insight into the characteristics of the raw data, it can also help to de-noise a particular data set (Belayneh *et al.* 2016). Fourier transform (FT) and short time FT are the first known popular transforms. However, FT has noticeable drawbacks in terms of time information and high resolution (Prahlada & Deka 2015). FT is a type of global transformation that is either completely in the time domain or completely in the frequency domain, therefore making it unsuitable to describe time-frequency local properties of a signal. In contrast, WT has the ability to characterise the local signal in the time and frequency domains; the time and frequency windows can be adjusted according to the concrete signal forms. WT can be applied in the continuous WT (CWT) or in the discrete WT (DWT). CWT is defined as (Maity *et al.* 2016):

$$W(a, b) = |a|^{-1/2} \int_{-\infty}^{\infty} x(t)\Psi^*\left(\frac{t-b}{a}\right)dt \; a, b \in R, a \neq 0 \quad (1)$$

where a is the temporal scale; b is the translation of the wavelet function along the time axis; $\psi^*(t)$ is the complex conjugate; and $W(a, b)$ is the CWT. The scaling and translation factor of the CWT in the continuous variation are real numbers which cannot deal with a digital signal. CWT is therefore primarily applied in theoretical analysis and demonstration whilst DWT is frequently adopted in forecasting applications (Belayneh *et al.* 2016). DWT can be obtained by the discrete scale factor a and shift factor b in Equation (1), and the parameters a and b can be determined based on Equation (2):

$$a = a_0^m, b = na_0^m b_0, a_0 > 1, b_0 \in R \quad (2)$$

where n and m are integer numbers that control the wavelet dilation and translation, respectively. Thus, DWT is often expressed as:

$$Wx(m, n, \Psi) = a_0^{-m/2} \int_{-\infty}^{\infty} f(t)\Psi^*(a_0^{-m}t - nb_0)dt \quad (3)$$

Equation (3) shows that a_0 is a specified fixed dilation step larger than 1, and that b_0 is the location parameter that must be larger than zero. This decomposition process can then be iterated with successive approximations being decomposed in turn, and the signal can be broken down

Table 1 | Hydrological station information and the length of the hydrological series

Station	Longitude	Latitude	Hydrological series
Huangjiagang	111°29'	32°32'	1989.1–2009.12
Huangzhuang	112°02'	32°03'	1989.1–2009.12
Shayang	112°35'	30°41'	1989.1–2009.12
Xiangyang	112°08'	32°01'	1989.1–2009.12
Xiantao	113°26'	30°23'	1989.1–2009.12

into a number of lower resolution components. The approximations are the high-scale, low-frequency components of the signal, while detail represents the low-scale, high-frequency components. In Equation (3), the appropriate choices for a_0 and b_0 depend on the wavelet function; different wavelet functions are characterised by their distinctive features, including the region of support and the number of vanishing moments. The wavelet used for DWT must therefore meet the condition in Equation (4) (Galiana-Merino *et al.* 2014):

$$\int_{\infty}^{+\infty} t^k \Psi(t) d_t = 0, \ \ldots, \mathrm{N}-1 \tag{4}$$

The condition in Equation (4) can be met by five wavelet functions (by calculation and verification): Symlets (symN), Daubechies (dbN), DMeyer (dmey), Coiflets (coifN) and ReverseBior (rbioM.N) (Roy *et al.* 2015; Lu *et al.* 2016). The selected wavelet function must be localised in the appropriate frequency to guarantee the removal of noise from the signal without causing damage to the usefulness of the signal, as well as being able to detect spectral and temporal information contained in the raw data. Analysis by Shoaib *et al.* (2015) on the most appropriate wavelet function for the development of wavelet coupled GEP models applied ten different mother wavelet functions from different wavelet families. These wavelet functions were: Haar, Daubechies (db2), Daubechies (db4), Daubechies (db8), Symlets (sym2), Symlets (sym4), Symlets (sym8), Coiflets (coif2), Coiflets (coif4) and the discrete approximation of the meyer (dmey) wavelet. Results recommended that the dmey wavelet function was the most appropriate due to its suitable time–frequency localisation and wider supporting length properties. The dmey wavelet function was therefore applied in our investigation to decompose the raw monthly flow series into sub-signals.

Gene expression programming

GEP, a population-based evolutionary algorithm evolved from GA and GP, was initially proposed by Ferreira (2001). The GEP algorithm is a GA that uses linear chromosomes with a fixed length and non-linear parse trees with different sizes and shapes obtained from the GA and GP. The chromosomes in the GEP comprise of multiple genes, and each gene codes a subprogram. The expression trees (ETs) hold the genetic information programmed in the chromosomes with a set of rules in the information decoding process. For example, the ET of an algebraic formula in Equation (5) is shown in Figure 2:

$$\sqrt{(a+b)/(c-d)} \tag{5}$$

This ET is considered as a phenotype in GEP, whereas the genotype can be easily inferred from the phenotype given by Equation (6):

$$Sqrt/+-abcd \tag{6}$$

Equation (6) is an open reading frame that starts at '*Sqrt*' and terminates at '*d*'. The GEP chromosomes are generally composed of more than one gene of equal length. Time series analysis typically has six steps: select and set the function, set the terminal, fitness function, control the GEP parameters, specify the termination condition and create the chromosomes (Kisi & Shiri 2011).

There are a number of advantages associated with using GEP. First, chromosomes used in the model are simple entities: linear, compact, relatively small and easily manipulated genetically (i.e. replicate, mutate, recombine and transpose). Secondly, ETs are exclusively the expression of their respective chromosomes; they are the entities upon which selection

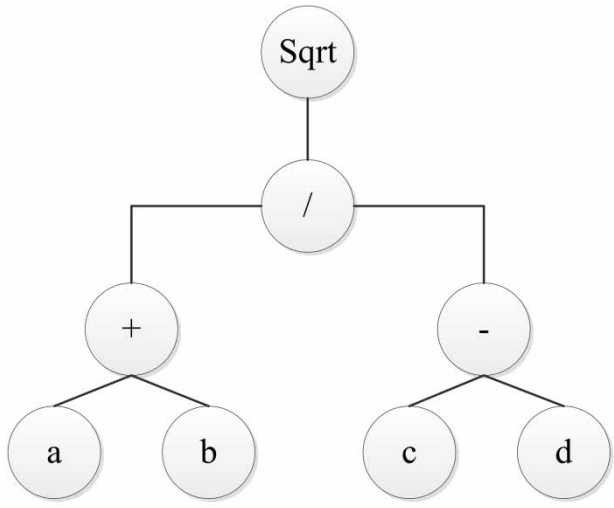

Figure 2 | Expression trees of Equation (5).

acts and they are selected to reproduce with modification according to fitness. In the GEP, it is chromosomes of individuals which are reproduced with modification and transmitted to the next generation during reproduction, not the chromosomes of the ETs.

GeneXpro Tools 5.0 was used in our investigation to model monthly flow in the middle and lower parts of Hanjiang River. Different mathematical functions (i.e. Sqrt, Exp, Ln, Log, 1/x, x2, x3, Cube root, Sin, Cos and Arctgx) and basic arithmetic operators (+, −, * and /) were used to evolve the desired GEP model. The set of terminals *T* was composed of time-lagged flow data and the genetic operators were subsequently selected: chromosome number, head size, gene number and linking function. The GEP parameters applied in this study are summarised in Table 2; all of the parameters were the default values used in the model.

Table 2 | GEP model parameters

Chromosome number	30
Head size	10
Gene number	9
Fitness function error type	Root mean square error (RMSE)
Inversion	0.1
Mutation	0.044
Insertion sequence transposition rate	0.1
Root insertion sequence transposition	0.1
Linking function	Addition
IS transposition	0.1
RIS transposition	0.1
One-point recombination	0.3
Two-point recombination	0.3
Gene transposition	0.1
Gene recombination	0.1
Gene transposition	0.1
Constants per gene	2
RNC mutation	0.01
DC mutation	0.044
DC inversion	0.1
DC IS transposition	0.1
Iteration number	10,000

WDDGEP models development

The majority of hydrologic data used for forecasting models are non-linearity and non-stationary which results in undesirable forecasting performance. Therefore, to improve the forecasting accuracy of monthly flow, previous studies have examined the efficiency and accuracy of stream-flow models utilising a wavelet based model (Kisi *et al.* 2012; Liu *et al.* 2014). These models have a commonality, in that the models initially decompose a time series into multiple levels of element, and then the signal's main frequency components are identified by implementing a multi-resolution analysis, as well as abstracting local information from the time series. The appropriate subseries were then used in the wavelet based model.

In this paper, the WDDGEP models can be described as follows: First, detailed information about the flow data was obtained by decomposing original flow time series data into a sub-series of approximation and details; the DWT of the input flow data was obtained using the dmey wavelet function. The WDDGEP model was then constructed using a GEP model that utilised the decomposed sub-time series extracted with DWT on the flow data. The time series of the stream flow in this study was decomposed into several multi-frequency time series (i.e. *d1*, *d2*… *di* and *ai*) by DWT, where *d1*, *d2* … and *di* are the detail series, and *ai* is the approximation of the stream flow time series. As shown in Figure 3, this decomposition process was iterated with successive approximations being decomposed in turn, and the signal was broken down into many lower resolution components. Thus, the WDDGEP models were developed utilising the transformed and de-noising flow data as input.

To identify the mapping function, a fitness function was initially selected for the GEP process. The fitness function can be a measure of the error indicating the difference between the outcome and the actual expected value. For our WDDGEP models, the mean squared error was selected as the fitness function. The GEP gene was then created by selecting a set of terminals *T* and functions *F*. In this study, the set of terminals *T* composed of time lagged flow data. The set of functions *F* included arithmetic operators, testing and Boolean functions. The chromosomal architecture was selected and it comprised the head size, the number of genes and the linking function. Subsequently, the last step

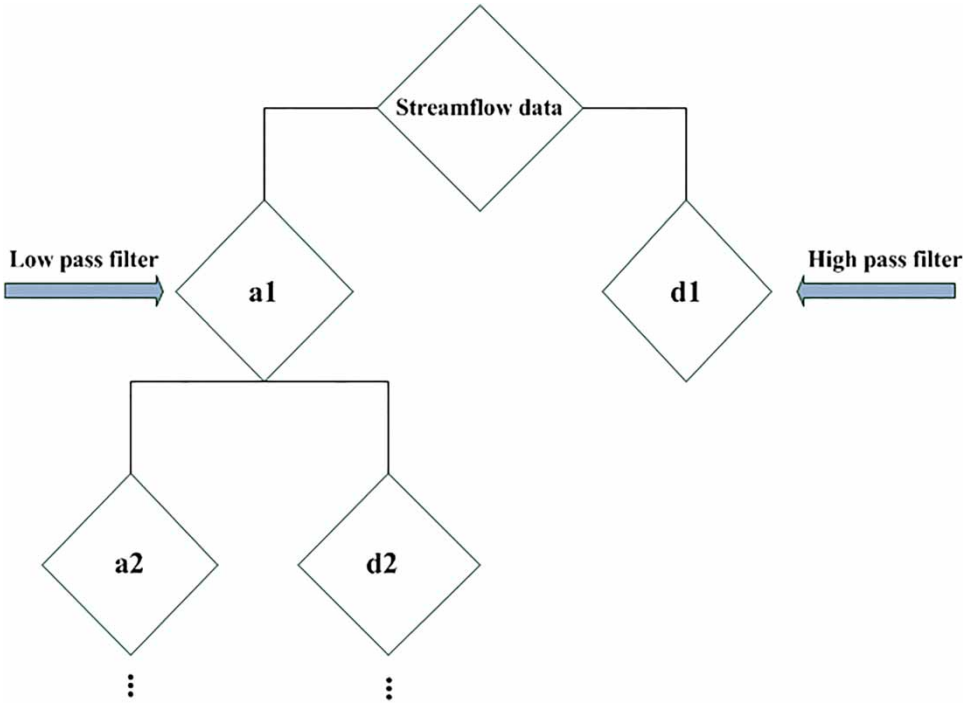

Figure 3 | n-Level decomposition of the flow data.

was the selection of the genetic operators. The detailed diagram of the WDDGEP models is shown in Figure 4.

Performance evaluation of the proposed model

The forecasting ability of the developed models can be evaluated in terms of statistical measures of goodness of fit through a number of evaluation indicators. Three statistical assessment criteria, adapted from similar investigations (Legates & McCabe 1999), were used in this study to evaluate model performance: coefficient of correlation (R), RMSE and Nash-Sutcliffe efficiency coefficient (NSEC). The formulae are:

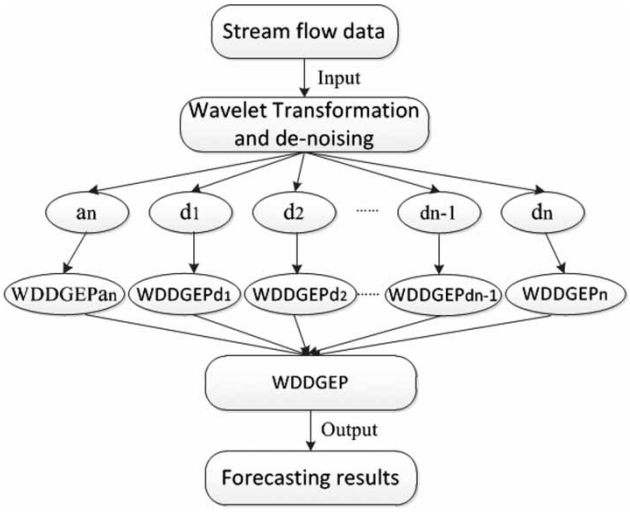

Figure 4 | Flow chart of the WDDGEP models.

$$R = \frac{\sum_{i=1}^{n}(m_i - \bar{m})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(m_i - \bar{m})^2}\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2}} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - o_i)^2} \quad (8)$$

$$NSEC = 1 - \frac{\sum_{i=1}^{n}(m_i - o_i)^2}{\sum_{i=1}^{n}(m_i - \bar{o}_i)^2} \quad (9)$$

where m_i and o_i are the modelled and observed data, respectively; and n is the pattern number in the data set. R,

commonly used to model evaluation as a measure of the strength of linear dependence between two variables, has been used here to measure the correlation between the two variables m_i and o_i. RMSE indicates discrepancy between modelled and observed values; the lower the RMSE value, the more accurate the prediction becomes. NSEC represents the relative magnitude of the residual compared to the observed data variance (Schultz *et al.* 2016). R, RMSE and NSEC were used to evaluate model performance in this study.

RESULTS AND DISCUSSION

Decomposing and de-noising monthly flow time series

The first stage of the discrete process started from the original series, and the result includes detailed coefficients and approximation series under each level. Detailed coefficients represent the low-scale, high-frequency components of the signal, and approximations represent the high-scale, low-frequency components. The high-frequency components often contain the noise whilst the low-frequency components provide the signal with its identity. The detail signals can obtain trivial attributes of interpretational values in the data, whereas the approximation shows the background information of the data. As the prediction accuracy of the WDDGEP models is directly affected by the decomposition level, it is vital that suitable decomposition levels are selected. The formula shown in Equation (10) (a formula used in numerous previous studies), was used to select a suitable decomposition level:

$$L = \text{int}(\log(N)) \tag{10}$$

where N is the total number of data points; and L is the decomposition level in the data.

As per previous investigations, we used the dmey wavelet function to decompose the time series of the flow data into individual components. Seven detail series ($d1$, $d2$, $d3$, $d4$, $d5$, $d6$, $d7$) and one approximation series ($a7$) were obtained by this decomposition. However, due to the influence of random factors, such as the inevitability of measurement errors and dynamism of the natural environment, these sub-signals could be composed of noise components. To improve the simulation accuracy of this model, wavelet-based de-noising of observed non-stationary hydrological time series can be an effective measure. The signal energy of the flow time series is concentrated on the low-frequency domain, whereas noise energy exists in the high-frequency domain. Considering the sampling interval and practical significance of the data, a convenient wavelet de-noising method was applied. The wavelet threshold de-noising method, which aims to adjust the detail wavelet coefficients, used Equation (11):

$$W_f'(a, b) = \sigma(W_f(a, b), Ta) \tag{11}$$

where $W_f'(a,b)$ is the adjusted $W_f(a,b)$ value; Ta is the threshold under level b; and $\sigma()$ is the thresholding rule. After adjustment, this sub-signal noise can be removed before further analysis and simulation, thus the decomposed components can be utilised as inputs for the WDDGEP models. The specific process of decomposing monthly flow time series at the Huangzhuang hydrological station is shown in Figure 5.

Examination of the spectral and temporal information

Due to the noise composition of the hydrological data being very difficult to distinguish, examination of the spectral and temporal information is therefore very important in wavelet de-noising. The accuracy of noise reduction is mainly verified using prediction accuracy as the main diagnostic tool; the examination of spectral and temporal information after decomposition and noise reduction is often overlooked in previous similar studies. In order to detect the spectral and temporal information in the raw data after noise reduction, FT was used to evaluate WDDGEP of the monthly flow time series. The FT, referred to as the frequency domain representation of the original signal, decomposes a function of a signal into its frequencies. The term FT refers to both the frequency domain representation and the mathematical operation that associates the frequency domain representation to a function of time. By using this transform we can identify if the de-noising data have damaged any valuable information and detect spectral and temporal information contained in the raw data. The spectrum comparison

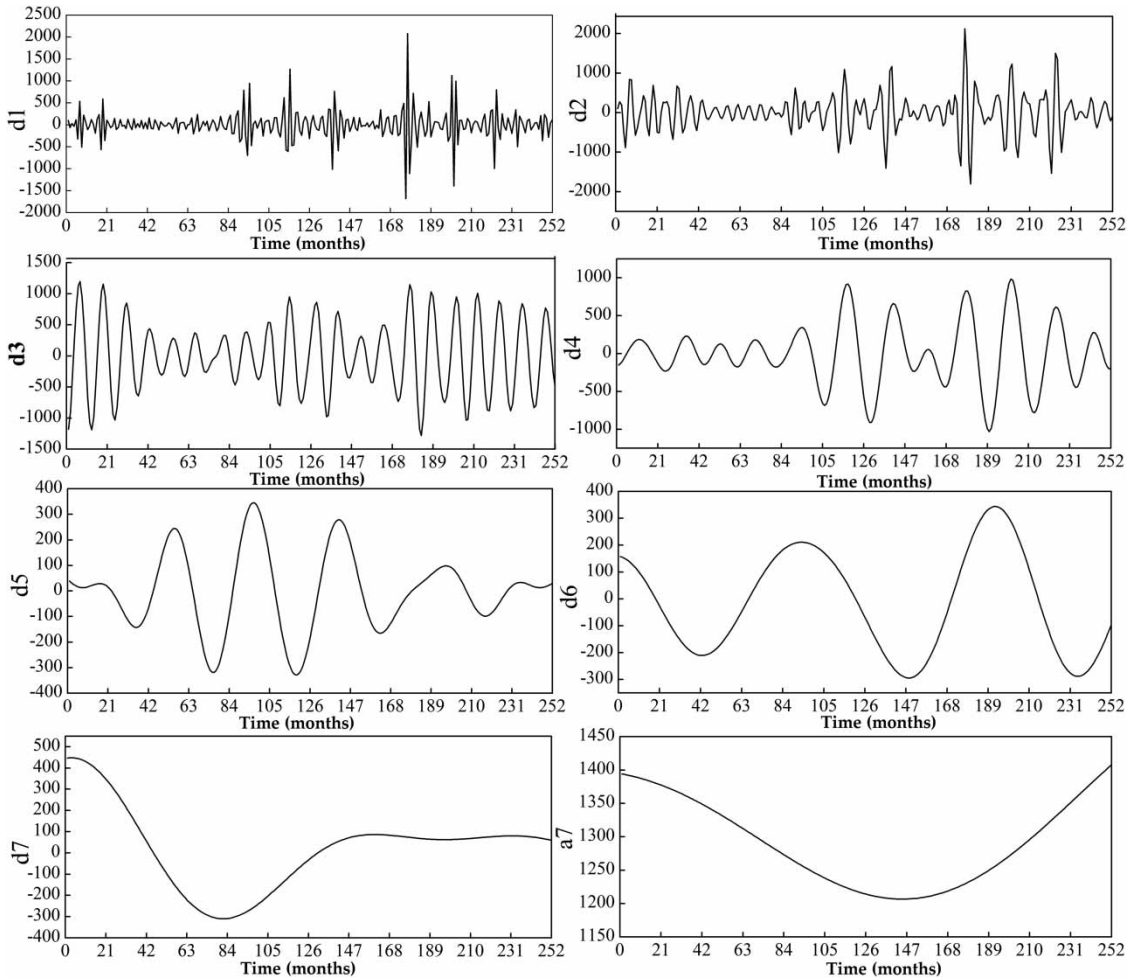

Figure 5 | Decomposition of the monthly flow time series at the Huangzhuang hydrological station.

before and after de-noising at the Huangzhuang station is shown in Figure 6; it is reasonable to decompose and de-noise data without reduction in the signal energy.

Discussion on the performance of the best models

The original monthly flow time series and the decomposed component were modelled using the WDDGEP and GEP models, respectively. Figure 7 shows that GEP based on the wavelet decomposed monthly flow time series can produce a suitable and close forecast; in comparison with the original monthly flow time series forecasting at the five stations, GEP did not have a close fit. Results in Figure 7 also show that WDDGEP and GEP recorded the best RMSE performance for the Huangjiagang station; WDDGEP had a RMSE value of 268.26 in the training phase and 751.49 in the validation phase, and an NSEC value of 0.58 in the validation phase. Compared with the Huangjiagang station, the flow prediction of the GEP model recorded a RMSE value of 824.77 and R values of 0.08 and 0.19 for the training and validation phases, respectively, at the Huangzhuang station. The predicted data for WDDGEP had a RMSE of 302.52 in the training phase and 888.94 in the validation phase, and R values of 0.92 and 0.69, respectively. A NSEC value of 0.45 was observed for WDDGEP with an R value of 0.69 in the validation phase, which performed better than the GEP at the Huangzhuang station. For the Xiantao station, the WDDGEP models had an NSEC value of 0.71 in the training phase and −0.32 in the validation phase, this not being significant when compared to simple GEP models, with the

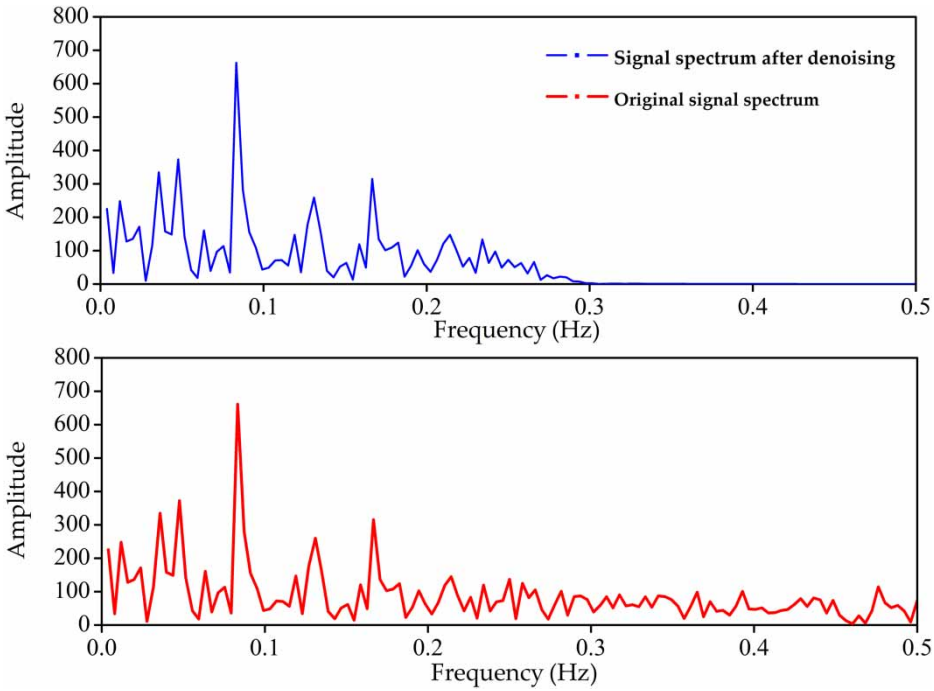

Figure 6 | Spectrum comparison before and after de-noising at the Huangzhuang station.

WDDGEP models obtaining the lowest R value (0.53) at this station. The Xiantao station is situated furthest downstream in the research area, located at the lower reaches of Xinglong Water Conservancy Project. The main function of this project is to divert water from the Yangtze River to the lower reaches of the Hanjiang River, a process which alters total runoff quantity and extreme runoff duration in the downstream station, thereby one possible reason for the failure of the simulation results. For the Xiangyang station, NSEC values of 0.36 and −0.55 were recorded in the training and validation phases using the WDDGEP models, respectively; RMSE values of 318.24 and 209.40 were recorded in the training and in the validation phase with R values of 0.04 and 0.194, respectively. The flow prediction of the WDDGEP models at the Shayang station recorded a NSEC value of 0.82 in the training phase and 0.38 in the validation phase; RMSE and R values of 317.29 and 890.09, and 0.92 and 0.62, were recorded in the training and validation phases, respectively.

Flow prediction results for the five stations utilising the two models were plotted against observed data using the GEP and WDDGEP models (Figure 8) to enable evaluation of the results. The testing period included the last 80 months of the observed data and time series of the predicted flow for the two models were plotted with the observed data. Results in Figure 8 show that the GEP based on the wavelet model can obtain better results than the simple GEP model, especially with large fluctuations of flow data. The flow forecasting results of the five stations indicate that the simple GEP model cannot trail the features of the observed flow in the middle and lower parts of the Hanjiang River. The proposed WDDGEP models were therefore suitable for decomposing the monthly flow time series and it can overcome the drawbacks of individual models. The results also show that the WDDGEP models tends to underestimate maxima and overestimate minima values, a performance similar to that of the GEP model. Underestimation of maxima and overestimation of minima values is a common limitation of statistical prediction models that requires further research.

CONCLUSIONS

In our investigation, we used GEP to predict the flow time series for five stations located in the middle and lower parts of the Hanjiang River. The WDDGEP models

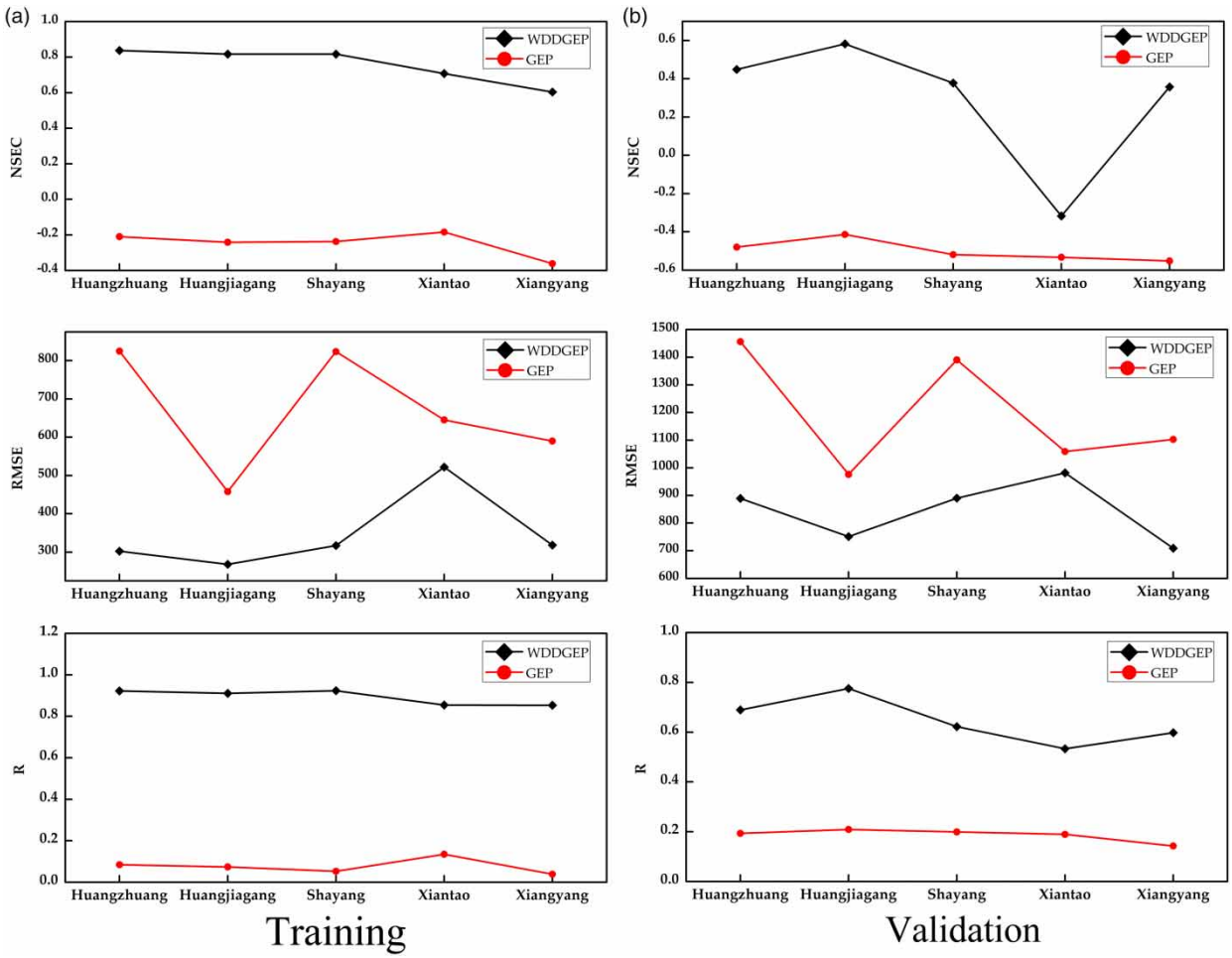

Figure 7 | Comparison of RMSE, NSEC and R values for the five stations.

were developed utilising the flow sub-time series; decomposed details and approximation data components were adopted as inputs to the WDDGEP models after signal de-noising. The capability of the WDDGEP models was compared with those of traditional GEP models and the results of the two developed models were evaluated on a statistical platform of error and performance metrics that were compared with actual experimental results. Analysis from the performance and error evaluation shows that, for the five hydrological stations, WDDGEP can more accurately predict monthly flow time series than simple GEP models developed utilising raw flow data. It can be further concluded that the simple GEP model is unable to trail the features of the observed flow time series. Furthermore, DWT can not only extract multi-scale characteristics of signals, it can also be applied to decompose monthly flow time series data. Wavelet de-noising methods can potentially be applied to help define and extract useful information from time series data, thereby improving the future performance of surface hydrological modeling.

The proposed methodologies can be adopted in future studies to construct and predict other hydrological applications in south-eastern China. Future studies can also investigate model performance using different input series constructed from effective or all wavelet components.

AUTHOR CONTRIBUTIONS

Xiaorong Lu conceived and designed this study. Xuelei Wang and Liang Zhang made substantial contributions to

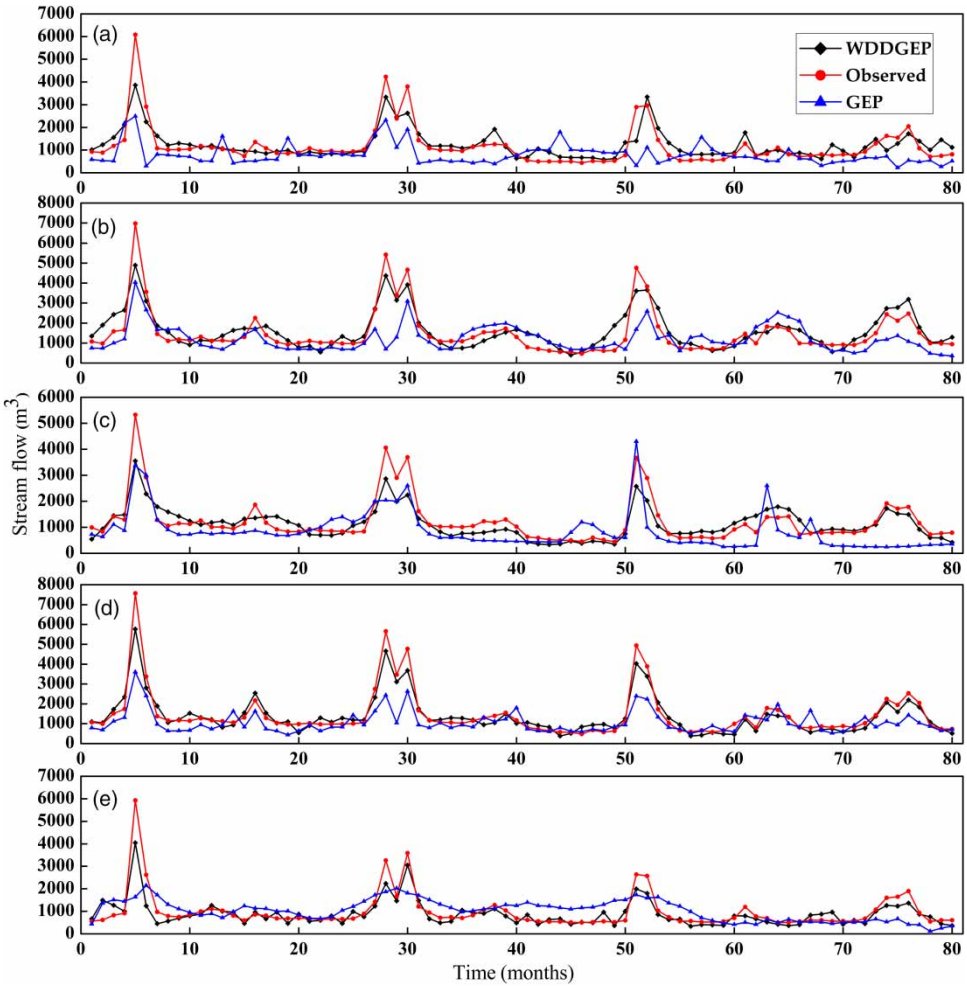

Figure 8 | GEP and WDDGEP forecasted and observed flows during the validation period at the five stations: (a) Xiangyang; (b) Shayang; (c) Xiantao; (d) Huangzhuang; (e) Huangjiagang.

acquisition, analysis and interpretation of the data. Xiaorong Lu contributed to the WDDGEP calculations. This manuscript was prepared by Ting Zhang and revised by Chao Yang. XinXin Song and Liang Zhang provided valuable suggestions for the revision and were involved in polishing the language. All authors read and approved the submitted manuscript, agreed to be listed and accepted the final version for publication.

ACKNOWLEDGMENTS

This work was supported by: (1) the National Natural Science Foundation of China (Grant No. 41571202, 41171426); (2) the Key Program of Institute of Geodesy and Geophysics, Chinese Academy of Sciences; and (3) the Natural Science Foundation of Hubei Province, China (Project No. 2014CFB330).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Al-Juboori, A. M. & Guven, A. 2016 Hydropower plant site assessment by integrated hydrological modeling, gene expression programming and visual basic programming. *Water Resour. Manage.* **30** (7), 2517–2530.

Belayneh, A., Adamowski, J., Khalil, B. & Quilty, J. 2016 Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmos. Res.* **172–173**, 37–47.

Coulibaly, P. & Burn, D. H. 2004 Wavelet analysis of variability in annual Canadian streamflows. *Water Resour. Res.* **40** (3), W0315.

Coulibaly, P., Anctil, F. & Bobee, B. 2000 Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *J. Hydrol.* **230** (3), 244–257.

Farajzadeh, J., Fakheri Fard, A. & Lotfi, S. 2014 Modeling of monthly rainfall and runoff of Urmia lake basin using 'feed-forward neural network' and 'time series analysis' model. *Water Resour. Indus.* **7–8**, 38–48.

Ferreira, C. 2001 Gene expression programming: a new adaptive algorithm for solving problems. *Complex Systems* **13**, 87–129

Galiana-Merino, J. J., Pla, C., Fernandez-Cortes, A., Cuezva, S., Ortiz, J. & Benavente, D. 2014 Environmental wavelet tool: Continuous and discrete wavelet analysis and filtering for environmental time series. *Comput. Phys. Commun.* **185** (10), 2758–2770.

Guyennon, N., Valerio, G., Salerno, F., Pilotti, M., Tartari, G. & Copetti, D. 2014 Internal wave weather heterogeneity in a deep multi-basin subalpine lake resulting from wavelet transform and numerical analysis. *Adv. Water Resour.* **71**, 149–161.

Kisi, O. & Shiri, J. 2011 Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. *Water Resour. Manage.* **25** (13), 3135–3152.

Kisi, O., Dailr, A. H., Cimen, M. & Shiri, J. 2012 Suspended sediment modeling using genetic programming and soft computing techniques. *J. Hydrol.* **450**, 48–58.

Koutsoyiannis, D. 2009 HESS opinions 'A random walk on water'. *Hydrol. Earth Syst. Sci. Discuss.* **6** (5), 585–601.

Labat, D. 2008 Wavelet analysis of the annual discharge records of the world's largest rivers. *Adv. Water Resour.* **31** (1), 109–117.

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.

Liu, F., Yuan, L., Yang, Q., Ou, S., Xie, L. & Cui, X. 2014 Hydrological responses to the combined influence of diverse human activities in the Pearl River delta, China. *Catena* **113**, 41–55.

Lu, C., Troutman, J. R., Schmitz, T. L., Ellis, J. D. & Tarbutton, J. A. 2016 Application of the continuous wavelet transform in periodic error compensation. *Precis. Eng.* **44**, 245–251.

Maity, R., Suman, M. & Verma, N. K. 2016 Drought prediction using a wavelet based approach to model the temporal consequences of different types of droughts. *J. Hydrol.* **539**, 417–428.

Nayak, P., Venkatesh, B., Krishna, B. & Jain, S. K. 2013 Rainfall-runoff modeling using conceptual, data driven, and wavelet based computing approach. *J. Hydrol.* **493**, 57–67.

Osorio, V., Proia, L., Ricart, M., Pérez, S., Ginebreda, A., Cortina, J. L. & Barceló, D. 2014 Hydrological variation modulates pharmaceutical levels and biofilm responses in a Mediterranean river. *Sci. Total Environ.* **472**, 1052–1061.

Pereira, A. J. C. & Saraiva, J. T. 2011 Generation expansion planning (GEP) – A long-term approach using system dynamics and genetic algorithms (GAs). *Energy* **36** (8), 5180–5199.

Prahlada, R. & Deka, P. C. 2015 Forecasting of time series significant wave height using wavelet decomposed neural network. *Aqua. Procedia* **4**, 540–547.

Rebora, N., Silvestro, F., Rudari, R., Herold, C. & Ferraris, L. 2016 Downscaling stream flow time series from monthly to daily scales using an auto-regressive stochastic algorithm: streamFARM. *J. Hydrol.* **537**, 297–310.

Roy, S., Ghosh, A., Das, A. K. & Banerjee, R. 2015 Development and validation of a GEP model to predict the performance and exhaust emission parameters of a CRDI assisted single cylinder diesel engine coupled with EGR. *Appl. Energy* **140**, 52–64.

Schultz, M., Clevers, J. G., Carter, S., Verbesselt, J., Avitabile, V., Quang, H. V. & Herold, M. 2016 Performance of vegetation indices from landsat time series in deforestation monitoring. *Int. J. Appl. Earth Observ. Geoinform.* **52**, 318–327.

Shamshirband, S., Mohammadi, K., Khorasanizadeh, H., Yee, L., Lee, M., Petković, D. & Zalnezhad, E. 2016 Estimating the diffuse solar radiation using a coupled support vector machine–wavelet transform model. *Renew. Sustain. Energy Rev.* **56**, 428–435.

Shoaib, M., Shamseldin, A. Y., Melville, B. W. & Khan, M. M. 2015 Runoff forecasting using hybrid wavelet gene expression programming (WGEP) approach. *J. Hydrol.* **527**, 326–344.

Smith, L. C., Turcotte, D. L. & Isacks, B. L. 1998 Stream flow characterization and feature detection using a discrete wavelet transform. *Hydrol. Process.* **12** (2), 233–249.

Terzi, Ö. 2013 Daily pan evaporation estimation using gene expression programming and adaptive neural-based fuzzy inference system. *Neural Comput. Appl.* **23** (3–4), 1035–1044.

Wang, W.-C., Chau, K.-W., Cheng, C.-T. & Qiu, L. 2009 A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *J. Hydrol.* **374** (3), 294–306.

Willems, P. 2009 A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. *Environ. Model. Softw.* **24** (3), 311–321.

Yaseen, Z. M., El-Shafie, A., Jaafar, O., Afan, H. A. & Sayl, K. N. 2015 Artificial intelligence based models for stream-flow forecasting: 2000–2015. *J. Hydrol.* **530**, 829–844.

Zhou, H.-C., Peng, Y. & Liang, G.-H. 2008 The research of monthly discharge predictor-corrector model based on wavelet decomposition. *Water Resour. Manage.* **2** (2) 217–227

First received 22 June 2017; accepted in revised form 31 October 2017. Available online 28 November 2017