1-1-2010

Evolutionary modeling for streamflow forecasting with minimal datasets: a case study in the West Malian River, China

Qingwei Ni
Dalian University of Technology Dalian China

Li Wang
Shenyang University of Chemical Technology, Shenyang, China.

Renzhen Ye
Agricultural University of Huazhong

Fenglin Yang
Dalian University of Technology Dalian China

Muttucumaru Sivakumar
University of Wollongong, siva@uow.edu.au

Recommended Citation

Ni, Qingwei; Wang, Li; Ye, Renzhen; Yang, Fenglin; and Sivakumar, Muttucumaru: Evolutionary modeling for streamflow forecasting with minimal datasets: a case study in the West Malian River, China 2010, 377-385.
https://ro.uow.edu.au/engpapers/2681

ENVIRONMENTAL ENGINEERING SCIENCE
Volume 27, Number 5, 2010
© Mary Ann Liebert, Inc.
DOI: 10.1089/ees.2009.0082

Evolutionary Modeling for Streamflow Forecasting with Minimal Datasets: A Case Study in the West Malian River, China

Qingwei Ni,[1] Li Wang,[2,*] Renzhen Ye,[3] Fenglin Yang,[1] and Muttucumaru Sivakumar[4]

[1]School of Environmental and Biological Science and Technology, Dalian University of Technology, Dalian, China.
[2]College of Environmental & Biological Engineering, Shenyang University of Chemical Technology, Shenyang, China.
[3]Department of Mathematics, Agricultural University of Huazhong, Wuhan, China.
[4]School of Civil, Mining, and Environmental Engineering, University of Wollongong, Wollongang, New South Wales, Australia.

Received: March 1, 2009 *Accepted in revised form: March 7, 2010*

Abstract

A large dataset is generally needed when modeling hydrological processes. However, for developing countries such as China, datasets are often unavailable in remote areas. An attempt to apply a novel genetic programming (GP) technique was made to model the relationship between streamflow of the West Malian River and the impact of climate change in the northeastern part of China. Available annual streamflow and climatic data were used for training and testing of the GP model. Data from the years between 1982 and 2002 were used for automatic selection of the model relationship. Prediction of the model was undertaken for the period 2003–2006 and the results were compared with measured data. Predicted annual streamflow of the West Malian River agreed with measured data to an acceptable degree of accuracy even with a small amount of dataset. For comparison, a multilayer perceptron method with back propagation algorithm, a gray theory model, and a multiple linear regression model were selected to conduct the prediction with the same dataset. Results showed that the performance of GP method was generally better than other statistical methods such as multilayer perceptron, gray theory model, and multiple linear regression model. Further, the results also showed that the GP method is a useful tool for water resource management, especially in developing countries, to evaluate the potential impacts of climate change on the streamflow when large datasets are unavailable.

Key words: annual streamflow; evaporation; GP algorithm; precipitation; statistical methods; West Malian River

Introduction

Global climate is changing because of anthropogenic activities (Loáiciga *et al.*, 2000; IPCC, 2001a, 2001b; Corfee-Morlot and Höhne, 2003). From the present time to 2050, the average global air temperature could rise by 1.2°C or more, and during this same period, in eastern and entire northern region in China, the increase in temperature could reach between 1.0°C and 2.0°C by 2050 (Yang, 1999). The direct effect of a global rise in average temperature will be the redistribution of surface water resources on the earth (Waggoner, 1990). Therefore, understanding and assessing the impact of climate change and its effect on hydrological systems is critical for sustainable water resource management. Streamflow is a fundamental component of the water cycle, and it is often related to fresh water availability for human and natural ecosystems (Makkeasorn *et al.*, 2008). Hence, measurement of annual streamflow, its variability due to climate change, and its prediction will provide crucial information for adaptive water resource management.

Hydrological models are useful tools that have been widely applied in investigating streamflow. Researchers have used different timescales for streamflow estimation, such as real-time prediction (Deo and Thirumalaiah, 2000), hourly prediction (García-Bartual, 2002), daily prediction (Coulibaly *et al.*, 2000; Legesse *et al.*, 2003; Collischonn *et al.*, 2005), weekly prediction (Zealand *et al.*, 1999), and monthly prediction (Castellano-Méndez *et al.*, 2004; Amisigo *et al.*, 2008). Generally, the mathematical models used for modeling hydrological processes require relatively long-time series of historical data (Trivedi and Singh, 2005). However, availability of large amount of data for model's training and validation is not a problem in the developed countries as their hydrological data banks are relatively massive. On the contrary, in the developing countries, such as China, large hydrological dataset is often unavailable, especially in remote areas. Therefore, choosing a suitable approach for streamflow prediction with relatively few available datasets will be of much practical importance.

*Corresponding author: College of Environmental & Biological Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China. *Phone:* 86-24-81839602; *Fax:* 86-24-89389166; *E-mail:* wanglijohn@hotmail.com

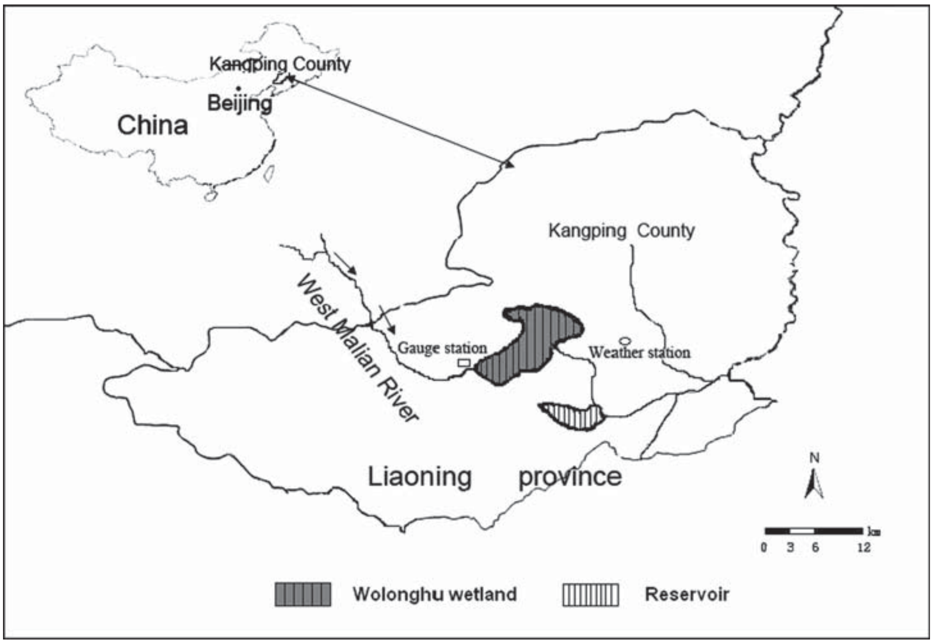

FIG. 1. Study area in the Liaoning province, China.

Genetic programming (GP) (Koza, 1992) is an extension of genetic algorithms, which has been considered as one of the best methods for searching, optimization, and modeling of complex natural systems (McKay, 2001; Whigham and Crapper, 2001; Kamal and Eassa, 2002; Duyvesteyn and Kaymak, 2005; Muttil and Lee, 2005; Lopes, 2007). Nath *et al.* (1997) and Muttil and Lee (2005) found that GP has good forecasting ability even when the amount of dataset is relatively small. Moreover, GP can capture the relationship between the inputs and the outputs automatically without the need of prior knowledge of the underlying physics.

Compared with three other popular forecasting methods, namely the multilayer perceptron (MLP), gray theory model (GM), and multiple linear regression model (MLR), a novel GP method was applied in this study to predict the annual streamflow response to the climate change in the West Malian River in the northeast China with relatively small datasets. With real coding mode, a specific and simple prediction model automatically evolved from the GP with good performance. This article is organized as follows: in the next section, a description of the study area and dataset is presented so as to allow readers to understand the background of the application. In the Genetic Programming section, the key principles of GP are outlined. In the Methods section, applications of GP, MLP, GM, and MLR for streamflow prediction are carried out. In the Results and Discussion section, the results from GP method and other methods are compared and discussed. And finally, the conclusions are drawn in the Conclusions section.

Study Area and the Datasets

The research was carried out in the West Malian River basin, which is located at the remote northern part of the Liaoning province in northeast China. The annual mean precipitation and evaporation of this region are 514 and 1933.2 mm, respectively, and the annual mean air temperature is 7.6°C. The West Malian River originates from the Keerqin desert region of Inner Mongolia and runs about 30 km southeast and finally flows east and joins the Wolonghu wetland, as shown in Fig. 1. The West Malian River is the primary source of recharge to the Wolonghu wetland. The Wolonghu wetland is the largest natural inland–wetland in the Liaoning province, and its area is covered with large biodiversity, which has to be preserved as a significant provincial natural reserve. Because of the semiarid climate and limited water availability for recharge, the wetland has developed serious loss of biodiversity in the recent past, which greatly impeded the sustainable development of the ecosystem. To restore this important ecosystem, certain amount of water, the so-called environmental flow, should be guaranteed in the wetland. Therefore, correct estimation of the fluctuations of the West Malian River streamflow will help the local managers in making the plan of adaptive water resource allocation for the wetland.

The West Malian River watershed is a rural area, where the streamflow gauging station is sparse and its records are typically short owing to the cost difficulties. Thus, 25 (1982–2006) available annual streamflow records of the West Malian River from the local hydrological gauging station and the corresponding climatic data from the only local weather station were collected and used for the GP model development.

Genetic Programming

GP (Koza, 1992), a member of evolutionary algorithms, is a relatively new approach to model the inputs and the outputs automatically. It conducts its research in the space of computer programs (solutions of a problem) whose structures are represented by varying size and shape binary trees. For example, an expression of $x_1 * \ln(x_2) + \cos(x_3)$ is represented by the program (Fig. 2). The computer programs consist of an internal node $F = \{$arithmetical functions, relation functions, etc.$\}$ and a terminal node $T = \{$numerical constants, variables, etc.$\}$. GP starts by randomly creating a population of com-

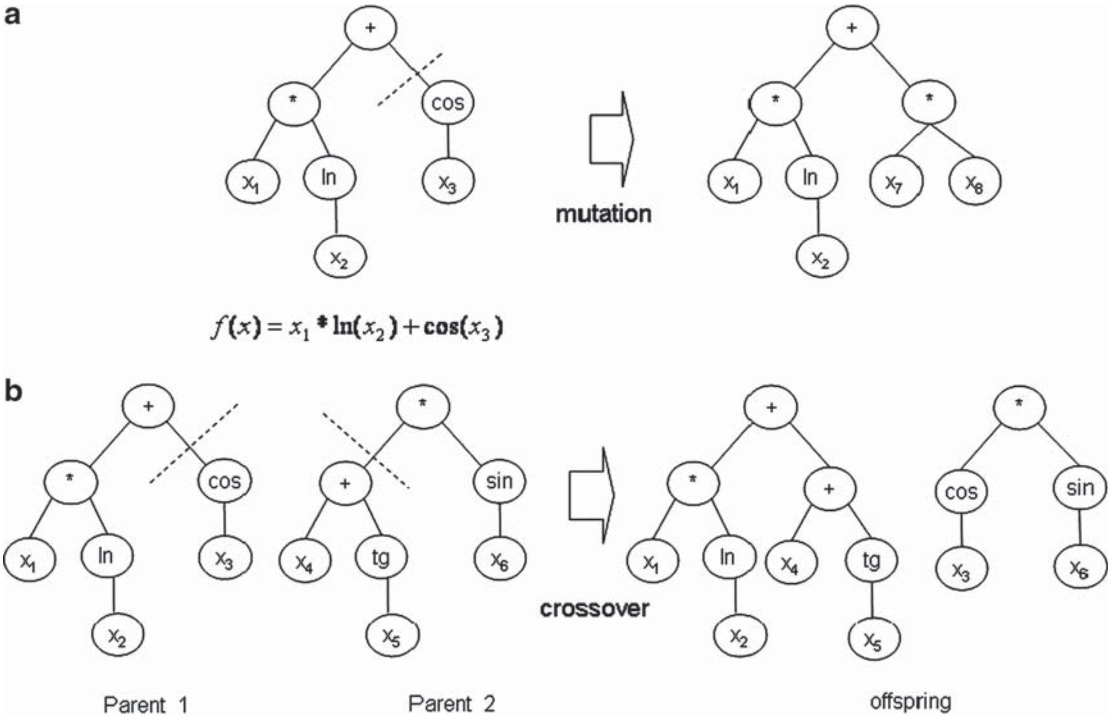

FIG. 2. (**a**) Genetic programming (GP) tree example and mutation and (**b**) crossover operation.

puter programs consisting of the available nodes from *F* and *T*, and each population represents a function *f* (*x*) of the given data. And then, GP genetically evolves the population by using the Darwinian principle of natural selection in which the selection, crossover, and mutation are the main operations. Thus, GP provides a way to find a suitable solution to a problem. The primary operators in GP are as follows:

1. Selection: Pairs of parent trees are selected based on their fitness values for reproduction.
2. Crossover: This process is performed by selecting a node randomly and then exchanging the associated subtrees to produce a pair of offspring trees (Fig. 2b).
3. Mutation: This process is performed by replacing a node selected at random with a newly created subtree (Fig. 2a). This process can prevent the GP model from falling into the local optima.

More information about the selection, crossover, and mutation can be found in Wang and Cao (2002).

Before running a GP algorithm, the following five steps should be adopted:

1. Determining the terminal node set, T;
2. Determining the internal node set, F;
3. Determining the fitness function;
4. Parameters and variables for controlling the run; and
5. Criterion for terminating the GP algorithm.

Methods

In this section, the GP algorithm for streamflow prediction is elaborated and this is followed by three other methods (including the MLP, GM, and MLR) that are used for comparison.

Application

GP algorithm for streamflow prediction. The main task here is to evolve a model automatically capable of simulating the annual streamflow response to the climate change with a small amount of datasets. The real values have been used here as the elements of the terminal node set T. Ten functions were used in the internal node set F for the purpose of the complex nonlinear relationship between streamflow and the climatic data. Of the F, four are arithmetic operators (+, −, *, \) and the rest are functional ones (exp, ln, cos, sin, tan, ctg).

In GP, the fitness function was used to estimate the solutions in terms of the problem. The final optimal model was obtained based on the fitness function. In this study, the function *f* was used as the measure of the fitness function:

$$f = \left\| y_t^{(0)}(x) - \hat{y}_t^{(0)}(x) \right\|^2 \qquad (1)$$

where the $y_t^{(0)}(x)$ represents the measured streamflow value, $\hat{y}_t^{(0)}(x)$ the simulated value, and *x* the input variable. The optimal model derived from the GP method is the one which has the relatively least *f* value.

As described by researchers and in some textbooks in the area of river hydrology, precipitation (Collischonn *et al.*, 2005; Chen, 2006) or both precipitation and evaporation (Li, 1992) are the main factors that affect the streamflow in a river, especially in the arid and semiarid regions. The fluctuations of the streamflow are also related to soil moisture, length of stream, elevation, geography of the catchment, and others. Owing to the limitation of the dataset, the available datasets of precipitation and evaporation were used to be the inputs for GP algorithm to capture the relationship with the output (streamflow). Figure 3 shows a plot of streamflow,

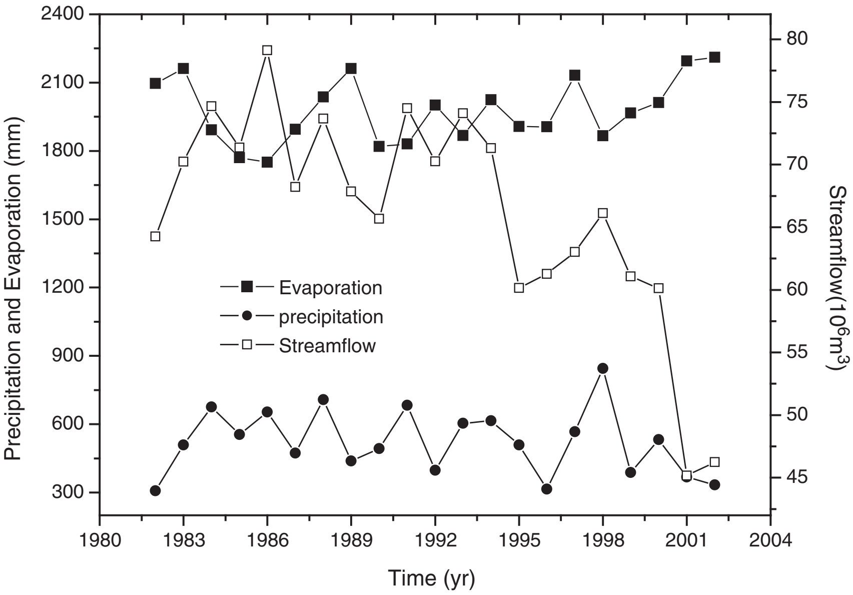

FIG. 3. Time series of the estimated streamflow and recorded precipitation and evaporation.

precipitation, and evaporation data from 1982 to 2002. This figure shows that the streamflow had a declining trend since 1992 along with the declining precipitation and rising evaporation. This trend is particularly striking between 1998 and 2002. Simple statistical analysis by using SPSS tool showed that annual streamflow was correlated to precipitation with $r = 0.537$, and to precipitation and evaporation ($|P - \mathrm{ET}|$) with a higher correlation ($r = 0.624$). The results indicated that the streamflow changes had close relationship with the joint effect of precipitation and evapotranspiration. Therefore, the values of $|P - \mathrm{ET}|$ was chosen as the input and the annual streamflow as the output. The relationship can be expressed mathematically in Eq. (2):

$$\hat{y}_t^{(0)}(x) = U(|P - \mathrm{ET}|) \quad (2)$$

where t refers to time of year, P and ET represent the annual precipitation and annual evapotranspiration of the tth year, respectively. The value of $|P - \mathrm{ET}|$ is equal to x in Eq. (1).

The steps involved in setting up the GP algorithm are as follows:

1. Generate an initial population of models randomly for the streamflow;
2. Run a tournament, which picks two models randomly out of the population; then apply the search operators (selection, crossover or mutation) to produce an offspring (new model) in the following way:
   (a) with crossover frequency, apply crossover operation to produce offspring,
   (b) with mutation frequency, mutate the models;
3. Compare all of the models and remove the loser based on the fitness measure function f;
4. Repeat the above procedures 2 and 3 until the termination criterion has been satisfied;
5. Display the model with the best fitness evolved from the GP.

The terminal criterion in this study is that the least fitness function value is unchanged after 50 repeats of procedure 4. Otherwise, the program will perform the procedure continuously. The parameters used in this GP algorithm are given in Table 1.

Streamflow and climatic data covering the years 1982–2006 were divided into two parts, of which 21-year data (1982–2002) was used during the training phase and the rest 4-year data for the testing phase.

The program of the GP algorithm was developed in language C and was run on a PC with a 1.74 GHz and 512 MB RAM memory, and the running time was within 10 min.

MLP for streamflow prediction.

Artificial neural network (ANN) techniques are popular methods capable of identifying correlated patterns between the input data and the corresponding target values. Therefore, they have been widely and successfully applied in hydrological modeling systems (Abrahart and Kneale, 1997; Dawson and Wilby, 1998; Imrie *et al.*, 2000; Baratti *et al.*, 2003; Castellano-Méndez *et al.*, 2004; Riad *et al.*, 2004; Valença *et al.*, 2005; Aqil *et al.*, 2007). Here, the MLP method, as one of the most popular ANN architectures for hydrological simulations (Castellano-Méndez *et al.*, 2004), was chosen for streamflow prediction and for comparing its prediction ability with the GP method.

In MLP, the first layer i is called the input layer and the last one k is called output layer, whereas the layer j between them is named the hidden layer, where the x_i ($i = 1, \ldots, n$), y_j ($j = 1, \ldots, m$), and o represent the input variables, the hidden variables, and the output, respectively (Fig. 4). In this

Table 1. Parameters Used in the Genetic Programming Algorithm

Parameter	*Value*
Initial population creation method	Grow method
Maximum tree depth size	6.0
Crossover rate	0.6
Mutation rate	0.1
Population size	100

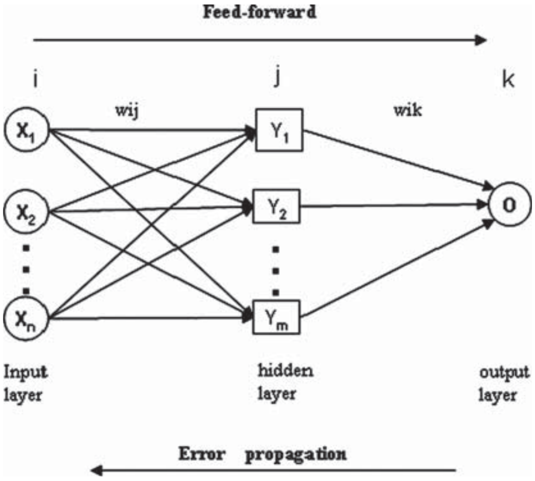

FIG. 4. Three-layer feed-forward backpropagation multilayer perceptron (MLP) network.

study, two types of architecture of MLP have been used. One has one node in the input layer, the other has two nodes. The MLP is trained using the error back propagation learning algorithm. The appropriate number of nodes in the hidden layer was determined by the rule of from $2n^{1/2}+m$ to $2n+1$ (Fletcher and Goss, 1993), which is helpful to prevent the MLP algorithm from overfitting the data (Huang and Foo, 2002), where n is the number of input nodes and $m$is the number of the output nodes.

To ensure that the MLP can generalize and will perform well when confronted with fresh data, the available data were also divided into two groups (data 1 and data 2) for model cross-validation procedure. The data 1 with 21-year data (1982–2002) was used to train the parameters in the MLP network. Data 2 with 4-year data (2003–2006) was used to verify the MLP performance.

GM for streamflow prediction. Gray system theory (Deng, 1989) provides a way to investigate the relationship of input–output process with relatively little data (as low as four), and it has been successfully applied in the fields of finance, engineering, economics, and hydrology (Chen and Lin 1996; Chiao *et al.*, 1997; Hao *et al.*, 2006). In this article, the GM (1, 2) model was used as another comparative tool to forecast the annual streamflow of the West Malian River. [Detailed construction process of the gray model is given by Hao *et al.* (2006).] For the GM (1, 2) model, 1 represents the data that need to be predicted (streamflow), and 2 represents the data series that was mainly impacted by 1. Two groups of dataset have been used to set up the GM (1, 2) model. One group is from 1982 to 2002 and the other group of data is from 1993 to 2002.

MLR for streamflow prediction. MLR model was the third comparative method used for streamflow prediction. The significant advantage of the MLR is the fact that it is easy to calculate and it is also very robust (Geladi *et al.*, 1999). The common pattern of the model is as follows:

$$Y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{3}$$

where Y_t represents the streamflow to be modeled; x_1 the evapotranspiration; x_2 the precipitation; β_0 a constant; β_1, β_2 the linear regression coefficient; and ε a residual, often assumed to contain the noise. A simple MLR for the prediction of streamflow is given by Eq. (4).

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \tag{4}$$

where $\hat{Y}_t$ is the predicted value of Y_t; and $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the estimated values of β_0, β_1, and β_2, respectively, described above.

Performance evaluation

In this study, two criteria were used to evaluate the performances of the forecasting models.

The first measurement is the mean absolute percent error (MAPE):

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{P_t - A_t}{A_t}\right| \times 100 \tag{5}$$

where P_t is the predicted value at time t, A_t is the actual value at time t, and n is the number of predictions.

The second criterion is the correlation coefficient r.

Results and Discussion

The model derived automatically from the GP algorithm is given in Eq. (6).

$$\hat{y}_t^{(0)}(x) = \text{tg}((\text{tg}(x) + 441.071x) + \text{tg}(257.391x)) \\ - x/(\sin(\ln(|x|))/(((-102.483)/x) + 4.994)) \tag{6}$$

where x represents $|P - \text{ET}|$ and $\hat{y}_t^{(0)}(x)$ represents the predicted values of the streamflow in the West Malian River. Through Eq. (6), once the precipitation and evaporation values of a given year is known, we can use $|P - \text{ET}|$ as the x to calculate the corresponding value of $\hat{y}_t^{(0)}(x)$, which is the total amount of annual streamflow of the West Malian River. The results of streamflow calculated from Eq. (6) are presented in Table 2 and Fig. 5.

In Table 2, the minimum average relative error is 0.25% in 1994 and the maximum average relative error is 9.65% in 1992 during the training phase. Table 3 shows that GP had a good performance with relatively small MAPE (5.48) and high correlation ($r = 0.88$) during the training phase. Also, the simulation curve is in substantial agreement with the measured data as shown in Fig. 5. Using the data of $|P - \text{ET}|$ from 2003 to 2006 (the black bold numbers in Table 2) as input to Eq. (6), the predicted annual streamflows of the West Malian River are calculated to be 48.57, 54.63×10^6, 67.27×10^6, and 69.36×10^6 m^3, respectively. The relative errors are 7.01%, 8.92%, 9.89%, 9.71%, for 2003, 2004, 2005, and 2006, respectively, which showed the best performance of GP over other methods with the least MAPE (8.89) and the highest correlation ($r = 0.99$).

The performances of other methods including the MLP, GM, and MLR are presented in Table 3 and Fig. 6.

In Table 3, the MLP method has the lowest MAPE (0.67, 0.85) and the highest correlation coefficient r (0.96, 0.93) in the training phase. The MAPE and correlation for the GP,

Table 2. Testing of the Genetic Programming Model of the Streamflow in the West Malian River

Year	$y_t^{(0)}(x)$ *Raw streamflow* *($\times 10^6$ m^3)*	$\hat{y}_t^{(0)}(x)$ *Simulated streamflow* *($\times 10^6$ m^3)*	$\varepsilon(x)=y_t^{(0)}(x)-\hat{y}_t^{(0)}(x)$ *Residual error*	$\Delta(x)=\frac{\lvert\varepsilon(x)\rvert}{y_t^{(0)}(x)}\times 100$ *Relative error (%)*
1982	64.25	61.70	2.55	3.97
1983	70.23	68.07	2.16	3.08
1984	74.65	76.43	−1.78	2.38
1985	71.35	68.07	3.28	4.60
1986	79.13	71.90	7.23	9.14
1987	68.22	67.13	1.09	1.60
1988	73.66	68.58	5.08	6.90
1989	67.84	61.43	6.41	9.45
1990	65.67	65.25	0.42	0.64
1991	74.49	72.88	1.61	2.16
1992	70.26	63.48	6.78	9.65
1993	74.09	69.14	4.95	6.68
1994	71.32	71.14	0.18	0.25
1995	60.15	63.74	−3.59	5.97
1996	61.26	65.84	−4.58	7.48
1997	63.03	68.94	−5.91	9.38
1998	66.13	69.23	−3.10	4.69
1999	61.07	65.18	−8.11	6.73
2000	60.12	65.04	−4.92	8.18
2001	45.16	46.68	−1.52	3.37
2002	46.25	42.20	4.05	8.76
2003	52.23	48.57	3.66	7.01
2004	59.98	54.63	5.35	8.92
2005	74.66	67.27	7.39	9.89
2006	76.82	69.36	7.46	9.71

GM (1, 2) with data 1, GM (1, 2) with data 2, and MLR methods are (5.48, 0.88), (34.16, −0.45), (26.79, −0.3), and (9.90, 0.66), respectively. The results of the training phase showed that the GP method outperformed other methods except for the MLP. In the testing phase, however, GP method showed the best performance among all the methods compared. The MAPE and *r* of GP method are 8.89 and 0.99, respectively. However, the MAPE and *r* of the MLP with one input node and two input nodes are 12.18, 0.33 and 14.31, 0.16, respectively. The MAPE and *r* for the GM (1, 2) with data 1, GM (1, 2) with data 2, and MLR are (38.97, −0.98), (20.55, −0.98), and (13.00, 0.70), respectively.

It is difficult to model hydrological and hydraulic processes in river catchment with limited data scenarios. Using the GP algorithm, it is shown here that it is possible to simulate the changes of streamflow at the West Malian River with readily available but limited dataset. From Figs. 5 and 6, it can be seen that the simulation and prediction curves matched well with the measured ones, particularly during the periods from 2001 to 2003 and from 2004 to 2006 during which the studied area

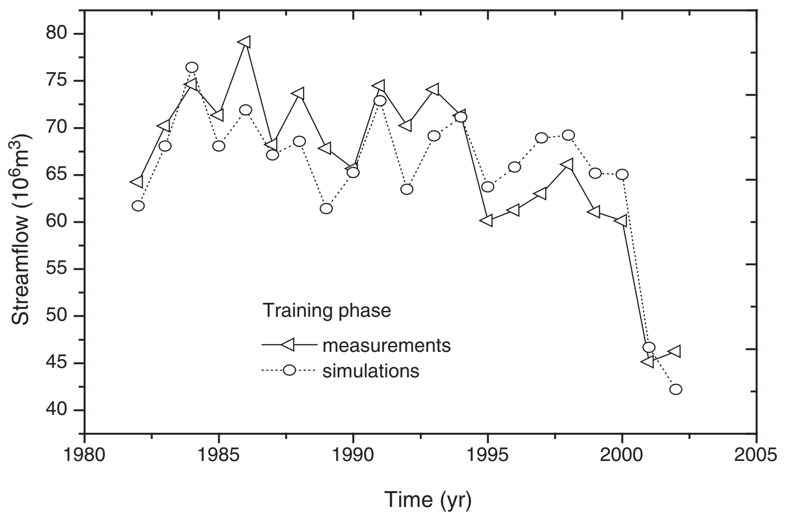

FIG. 5. Results of the simulated streamflow by GP.

Table 3. Relative Performance of the Streamflow Forecast of the West Malian River

	Genetic programming	*Multilayer perceptron with one input node*	*Multilayer perceptron with two input nodes*	*Gray theory model (1, 2) with data 1*	*Gray theory model (1, 2) with data 2*	*Multiple linear regression model*
Training						
Mean absolute percent error	5.48	0.67	0.85	34.16	26.79	9.90
r	0.88	0.96	0.93	−0.45	−0.3	0.66
Testing						
Mean absolute percent error	8.89	12.18	14.31	38.97	20.55	13.00
r	0.99	0.33	0.16	−0.98	−0.98	0.70

experienced a 3-year period of drought and a 3-year period of relatively heavy rainfall, respectively. The combined effects of annual precipitation and evaporation obviously affected the fluctuation of annual streamflow. Based on the statistic analysis and the results of the estimation from the GP model, it can be concluded that as expected the streamflow is positively correlated with precipitation and negatively correlated with evapotranspiration. The rise in annual average temperature in the northern China due to climate change will affect both precipitation and evaporation, and hence, the West Malian River annual flow will likely to be significantly affected by climate variability.

The poor performance of the MLP model most probably is due to overfitting. This usually occurs when the network is not fully trained and large amounts of datasets are unavailable. For a common MLP model, more training always makes the network memorize detail features of the dataset. Moreover, one of ANN's defects is that it is easy to run around the local optimal. In this study, Fig. 6 showed that the MLP had good simulation performance during the training phase with high correlation and the least error. However, in the testing phase, the performances were poor with relatively large errors and low correlations because of overfitting. On the contrary, the GP method, taking the errors as its inner motivation, is able to obtain as close to the global optimal solution as possible. Compared with the MLP, the GP method emphasizes the integrated characteristics of the dataset, which will reduce overfitting significantly. Table 3 shows that the GP method has good simulation ability during training phase (MAPE = 5.48, $r = 0.88$) and good prediction ability during testing phase (MAPE = 8.89, $r = 0.99$). Hence, based on the results of the GP method during the training phase and testing phase the overfitting problem was able to be determined.

Figure 6 indicates that the GM (1, 2) models did not perform as desirable as other case studies have reported (Hao *et al.*, 2006). Despite the fact that they showed high correlation ($r = -0.98$) during the testing phase, the MAPE values of the GM (1, 2) models were the largest compared with other methods (Table 3). As indicated earlier, the GM model is a potential tool for modeling with less dataset (as few as four). However, the data quality often affected the outcomes of the method. The GM often performed well when the data series satisfied the gray exponential law (Fu, 1992). It also meant that peaks in the dataset often led to poor performance. In this study, we had peaks in the data of streamflow even though we have divided the dataset into two parts (data 1 and data 2). Therefore, high accuracy of prediction could not be anticipated by this model (Wu and Chen, 2005). In the case of Hao *et al.* (2006), their data used for prediction were much smooth and this could have accounted for their good forecasting. For the MLR model, the results were also not as good as that of the GP model. The MAPE and *r* of the MLR model are 9.90 and

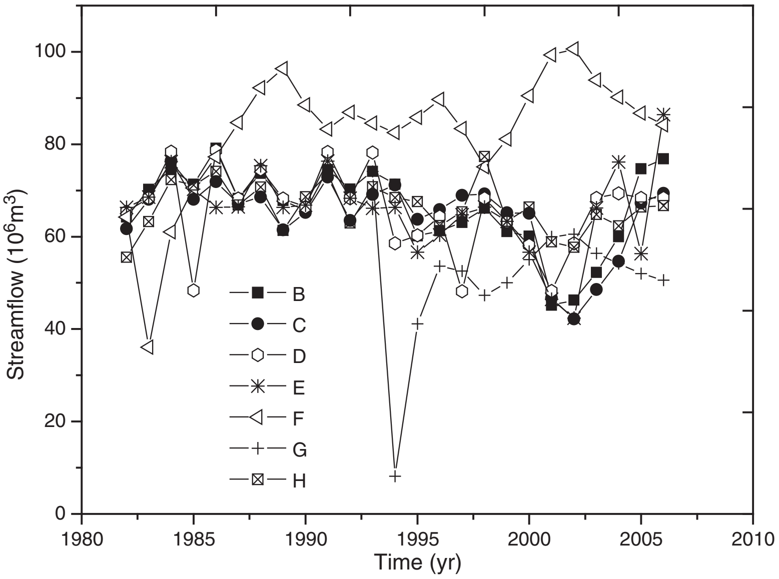

FIG. 6. Performance estimation of the methods. B, raw streamflow data; C, GP; D, MLP with one input node; E, MLP with two input nodes; F, gray theory model (1, 2) with data 1; G, gray theory model (1, 2) with data 2; H, multiple linear regression model.

0.66 during the training phase and 13.00 and 0.70 during the testing phase, respectively. The reason for the poor performance is probably due to the complex nonlinear relationship that exists between the streamflow and climatic variables and the MLR method may not pick up the correct relationship.

In this study, three possible reasons may cause the discrepancy between the observed data and the predicted ones for the GP model. First, some calculation errors were from the data standardization process and their amplificatory response of noise. And second, inaccuracy may occur because of the improper choice of the input variables. Here, the precipitation and evaporation were chosen as the inputs for the GP model. As mentioned earlier, these are the main factors that affect the streamflow of the West Malian River, and more importantly, these data were readily available. In fact, other factors, such as the soil moisture, length of stream, elevation, and geography of the catchment, also impacted the streamflow. A lack of a comprehensive dataset surely would have affected the accuracy of prediction by the GP method. However, the GP model was able to provide good results because GP can capture the relationship between the inputs and the outputs automatically without the need of prior knowledge of the underlying physics. The final reason is probably due to the use of a small amount of data. Although GP can perform well in capturing the underlying relationship between the inputs and the outputs even with less datasets, use of a large amount of data is expected to increase its accuracy of prediction.

Conclusions

Streamflow is the one of the most critical variables that influence the dynamics of any river ecosystem. Accurate estimation of annual streamflow fluctuation in response to climate change will be helpful for adaptive regional water resource management. It is shown here that the GP algorithm provided a quick and flexible means of creating a model automatically between the inputs and the outputs and was successfully applied to predict the annual streamflow fluctuations of the West Malian River in China with a small amount of dataset. Using climatic data of $|P-\mathrm{ET}|$ as the input, and annual streamflow as the output, a simple but specific model was set up automatically by the GP algorithm. Using this model, the predicted annual streamflows of the West Malian River were 48.57×10^6, 54.63×10^6, 67.27×10^6, and 69.36×10^6 m^3 in the years 2003–2006, respectively. The predicted results from the GP model were in substantial agreement with the measured data (MAPE = 8.89, $r = 0.99$) (Table 3, Fig. 6). Comparison of the results from various models such as the MLP, GM, and MLR indicated that GP's prediction is superior to other methods despite the fact that the dataset used was small. Hence, the GP algorithm can be a cost-effective and easy-to-use alternative tool for the estimation of annual streamflow fluctuation. In particular, this model can be a useful tool for managers in the developing countries to evaluate the potential impacts of climate changes on the streamflow where long-term dataset is unavailable.

In this case, the traditional agriculture with small industries is the main economic activity in the research area. On the basis of the analysis of the local water resource protection plan from the local environmental protection agency and the field investigation, the water intake of agriculture is the main part but the amount is relatively a constant. Therefore, streamflow prediction is able to be made in an acceptable accuracy through the model without the data on water intake from the river because the data have partly included the water intake for agriculture. However, with the local economic development, the amount of water intake from the river will surely increase. To achieve the accuracy of streamflow prediction, data on the amount of water intake from the river for agriculture and industry should be collected accordingly to improve the prediction by auto water resource monitoring systems in the future.

Acknowledgments

This work was supported by the National Water Special Project of China (no. 2009ZX07528-006-02) and the National Education Ministry Foundation of China (no. 705011).

Author Disclosure Statement

No competing financial interests exist.

References

Abrahart, R.J., and Kneale, P.E. (1997). Exploring neural network rainfall-runoff modelling. BHS 6th National Symposium, Salford, United Kingdom, pp. 9.35–9.44.

Amisigo, B.A., Van de Giesen, N., Rogers, C., Andah, W.E.I., and Friesen, J. (2008). Monthly stream flow prediction in the Volta Basin of West Africa: a SISO NARMAX polynomial modeling. *Phys. Chem. Earth* 33, 141.

Aqil, M., Kita, I., Yano, A., and Nishiyama, S. (2007). Analysis and prediction of flow from local source in a river basin using a Neuro-fuzzy modeling tool. *J. Environ. Manag.* 85, 215.

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

Castellano-Méndez, M., González-Manteiga, W., Febrero-Bande, M., Prada-Sánchez, J.M.,and Lozano-Calderón, R. (2004). Modelling of the monthly and daily behaviour of the runoff of the Xallas River using Box-Jenkins and neural networks methods. *J. Hydrol.* 296, 38.

Chen, J.S. (2006). River water quality theory and river water quality in China. Beijing: Science Press, pp. 57–62 (in Chinese).

Chen, J.Y., and Lin, Y.H. (1996). Design of fuzzy sliding mode controller with grey predictor. *J. Grey Syst.* 8, 147.

Chiao, J.H., Wang, W.Y., and Lu, M.J. (1997). A study for applying grey forecasting to improve the reliability of product. *Second National Conference on Grey Theory and Applications*, Zhanghua, Taiwan: National Yunlin University of Science and Technology Press, pp. 202–206.

Collischonn, W., Haas, R., Andreolli, I., and Tucci, C.E.M. (2005). Forecasting River Uruguay flow using rainfall forecasts from a regional weather-prediction model. *J. Hydrol.* 305, 87.

Corfee-Morlot, J., and Höhne, N. (2003). Climate change: long-term targets and short-term commitments. *Global Environ. Change* 13, 277.

Coulibaly, P., Anctil, F., and Bobée, B. (2000). Daily reservoir inflow forecasting using artificial neural network for stopping training approach. *J. Hydrol.* 230, 244.

Dawson, C.W., and Wilby, R. (1998). An artificial neural network approach to rainfall-runoff modelling. *Hydrol. Sci. J.* 43, 14.

Deng, J.L. (1989). Introduction to grey system theory. *J. Grey Syst.* 1, 1.

Deo, M.C., and Thirumalaiah, K. (2000). Real time forecasting using neural networks. In: R.S. Govindaraju, A. Ramachandra Rao, Eds., *Artificial Neural Networks in Hydrology*. London: Kluwer Academic, pp. 53–71.

Duyvesteyn, K., and Kaymak, U. (2005). Genetic programming in economic modelling. *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, Edinburgh, United Kingdom, 1025–1031.

Fletcher, D., and Goss, E. (1993). Forecasting with neural networks: an application using bankruptcy data. *Inform. Manage.* 24, 159.

Fu, L. (1992). *Grey System Theory and Application*. Beijing: Scientific and Technological Document Publishing House (in Chinese).

García-Bartual, R. (2002). Short term river flood forecasting with neural networks. Available at: www.iemss.org/iemss2002/proceedings/pdf/volume%20due/266_bartual.pdf.

Geladi, P., Hadjiiski, L., and Hopke, P. (1999). Multiple regression for environmental data: nonlinearities and prediction bias. *Chemometr. Intell. Lab.* 47, 165.

Hao, Y.H., Yeh, T.C.J., Gao, Z.Q., Wang, Y.R., and Zhao, Y. (2006). A gray system model for studying the response to climate change: the Liulin Karst springs, China. *J. Hydrol.* 328, 668.

Huang, W., and Foo, S. (2002). Neural Network modeling of salinity variation in Apalachicola River. *Water Res.* 36, 356.

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

IPCC. (2001a). *Climate Change 2001: The Scientific Basis. Summary for Policymakers. Intergovernmental Panel on Climate Change.* Cambridge, United Kingdom: Cambridge University Press.

IPCC. (2001b). *Climate Change 2001: Impacts Adaptation and Vulnerability. Summary Policymakers. Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom: Cambridge University Press.

Kamal, H.A., and Eassa, M.H. (2002). Solving curve fitting problems using genetic programming. *IEEE MELECON* 7, 316.

Koza, J.R. (1992). *Genetic Programming on the Programming of Computers by Means of Natural Selection, Seventh Printing*. Cambridge, MA: MIT Press, pp. 73–164.

Legesse, D., Vallet-Coulomb, C., and Gasse, F. (2003). Hydrological response of a catchment to climate and land use changes in Tropical Africa: case study South Central Ethiopia. *J. Hydrol.* 275, 67.

Li, W.S. (1992). *River Hydrology*. Beijing: China Water Power Press (in Chinese).

Loáiciga, H.A., Maidment, D.R., and Valdes, J.B. (2000). Climate-change impacts in a regional karst aquifer, Texas, USA. *J. Hydrol.* 227, 173.

Lopes, H.S. (2007). Genetic programming for epileptic pattern recognition in electroencephalographic signals. *Appl. Soft. Comput.* 7, 343.

Makkeasorn, A., Chang, N.B., and Zhou, X. (2008). Short-term stream flow forecasting with global climate change implications—a comparative study between genetic programming and neural network models. *J. Hydrol.* 352, 336.

McKay, R.I. (2001). Variants of genetic programming for species distribution modelling—fitness sharing, partial functions, population evaluation. *Ecol. Model*. 146, 231.

Muttil, N., and Lee, J.H.W. (2005). Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecol. Model.* 189, 363.

Nath, R., Rajagopalan, B., and Ryker, R. (1997). Determining the saliency of input variables in neural network classifiers. *Comput. Oper. Res.* 24, 767.

Riad, S., Mania, J., Bouchaou, L., and Najjar, Y. (2004). Rainfall-runoff model using an artificial neural network approach. *Math. Comput. Model.* 40, 839.

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

Valença, M., Ludermir, T., and Valença, A. (2005). River flow forecasting for reservoir management through neural networks. *International Conference on Hybrid Intelligent Systems (Proceedings of the Fifth International Conference on Hybrid Intelligent Systems).* Washington, DC: IEEE Computer Society.

Waggoner, P.E. (1990). *Climate Change and U.S. Water Resource*. New York: Wiley.

Wang, X.P., and Cao, L.M. (2002). *Genetic Algorithm: Theory Application and Software Implement [M].* Xi'an, China: Xian Jiaotong University Press, pp. 96–98 (in Chinese).

Whigham, P.A., and Crapper, P.F. (2001). Modelling rainfall-runoff using genetic programming. *Math. Comput. Model.* 33, 707.

Wu, W.Y., and Chen, S.P. (2005). A prediction method using the grey model GMC (1, n) combined with the grey relational analysis: a case study onInternet access population forecast. *Appl. Math. Comput.* 169, 198.

Yang, G.S. (1999). Global changes and the study of a natural disaster trend in China. *Adv. Earth Sci.* 14, 83 (in Chinese).

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