Flood simulation using parallel genetic algorithm integrated wavelet neural networks

Yuhui Wang [a,*], Hao Wang [b], Xiaohui Lei [b], Yunzhong Jiang [b], Xinshan Song [a]

[a] School of Environmental Science and Engineering, Donghua University, 2999 North Renmin Road, Shanghai 201620, PR China
[b] China Institute of Water Resources and Hydropower Research, 1 South Yuyuantan Road, Haidian, Beijing 100038, PR China

ARTICLE INFO

ABSTRACT

The conventional means of flood simulation and prediction using conceptual hydrological model or artificial neural network (ANN) has provided promising results in recent years. However, it is usually difficult to obtain ideal flood reproducing due to the structure of hydrological model. Back propagation (BP) algorithm of ANN may also reach local optimum when training nodal weights. To improve the mapping capability of neural networks, wavelet function was adopted (WANN) to strengthen the non-linear simulation accuracy and generality. In addition, genetic algorithm is integrated with WANN (GAWANN) to avoid reaching local optimum. Meanwhile, Message Passing Interface (MPI) subroutines are introduced for distributed implement considering the time consumption during nodal weights training. The GAWANN was applied in the flood simulation and prediction in arid area. The test results of 4 independent cases were compared to reveal the relations between historical rainfall and runoff under different time lags. The simulation was also carried out with Xinanjiang model to demonstrate the capability of GAWANN. The numerical experiments in this paper indicated that the parallel GAWANN has strong capability of rain-runoff mapping as well as computational efficiency and is suitable for applications of flood simulation in arid areas.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Flood simulation and prediction is one of the most active researching areas in surface water hydrology. Flood takes place whenever there is a heavy or a long period of precipitation. An accurate prediction of flood under changeable meteorological and layer conditions can not only help in the water resources management especially in hydropower, but also reduce the loss of lives and property to the minimum in floodplain areas. While with the fast increasing economic activities in floodplain and rivers especially in arid areas in northeast China, further requirements are raised on more accurate flood prediction precision.

The actual rain-runoff (RR) process is such a non-linear problem as until now no explicit formula can describe the process perfectly. During the last decades, great progress in flood prediction has been made by taking several techniques such as empirical model, statistical model, and physical based conceptual or distributed hydrological model into account. Empirical models can provide hydrograph in a special basin but long time series of observations are badly needed before carrying out simulation [1].

Besides, empirical methods are always not capable of generating a runoff hydrograph with complete information on timing of peak. Physical based hydrological model are more popular since the fast development in computer science and technology in the 1940s [2–5]. Newly developing spatial technology like remote sensing (RS) can also support physical based models by providing large sets of spatial data such as leaf area index (LAI) and land use category, even soil moisture [6–8]. On the other hand, hydrological models commonly used in flood prediction can be divided into two categories, i.e. conceptual and distributed model. Developed by Zhao [9–11], Xinanjiang model is a most widely accepted conceptual lumped hydrological model that performs in flood simulation especially suitable in wet areas. The model was conceptualized to divide the runoff into surface and ground flow using Horton's theory [12]. Lin et al. [13,14] coupled Xinanjiang model and meteorological system in flood prediction of Huaihe River Basin. Compared to distributed hydrological models, less data inputs are needed in conceptual model. Inputs of such conceptual model are mainly about precipitation, evapotranspiration, etc. Researches on distributed hydrological models that started in the 1960s have been applied in many major basins around the world due to its strong power in taking the underlayer viabilities into consideration [15]. Braud et al. analyzed the flash flood event using two distributed hydrological model: CVN

and MARINE. Different terms of the two models are discussed in the peak discharges [16]. Results of such models are generally ideal; however many of them involve ecology, atmosphere, human activities, and can also be affected by basic science including physics, bio-process, or chemistry. Others may differ or orient for various simulation purposes [17]. On the other hand, physics based models need a great number of spatial and observation data as inputs to implement simulation and calibration including temperature, wind speed, sunshine hour, etc.

Compared to the aforementioned methods, black-box models, artificial neutral network (ANN), have the advantage of fast data fitting and have became the preferred approach since its development in the 1980s. Ju et al. [18] employ division-based backpropagation neural networks in rainfall–runoff simulation and compared with the Xinanjiang conceptual RR hydrological model. Ahmad et al. [19] adopted ANN model for prediction peak flow in Red River timing and shape of runoff hydrograph together with the employment of meteorological parameters including antecedent precipitation index and melt index.

In addition to the application using ANN issues above, many improvements have also been made to strengthen the performance of ANN. This includes integrating data preprocessing techniques with ANN, such as moving average (MA) and singular spectrum analysis (SSA) [20]. The coupled ANN has the ability to get rid of the effects of white noise that may add errors to the weights training. For another instance, in order to improve the drawbacks of the conventional optimal process, Chen and Chang [21] proposed a novel evolutionary artificial neural network (EANN) for time series forecasting.

In flood prediction, however, a much higher runoff time series forecasting accuracy is often requested. Besides, the forecasting ability, especially the forecasting interval length, is also a significant criterion to be taken into consideration. In this paper, an ANN model hybrid wavelet function was proposed and applied in the simulation and prediction process of flood period in Nen River Basin of China. The nodal weight training of ANN model was improved by the hybrid of parallel genetic algorithm to reach global optimal before carrying out BP at the last iteration. Forecasting ability was then further discussed by taking three different prediction solutions. The objective of this study lies in (1) developing wavelet artificial neural networks and their applications to flood simulation in arid floodplain of Northern China and (2) enhancing the computational efficiency of GAWANN using MPI technique.

The remaining paper outlines a framework for developing a flood prediction system using genetic algorithm hybrid back propagation wavelet neutral networks and can be organized as follows. Section 2 gives a brief introduction to the WANN and the improvements done by integrating genetic algorithm with MPI technology. Section 3 follows the study case with data preprocessing. Results of different prediction solutions were drawn with further discussions and comparisons in Section 4. Meanwhile simulation results of the Xinanjiang model were also used as a comparison to GAWANN. In the last section, conclusion and remarks are summarized.

2. Methodology

2.1. Back propagation artificial neural networks

As RR is an incidental, non-linear process and always affected by the variation of under-layer conditions, it is difficult to derive a single accurate formula to describe all the physical processes. Artificial neural networks (ANN), belonging to the class of blackbox models, can be explored in RR simulation as alternative due
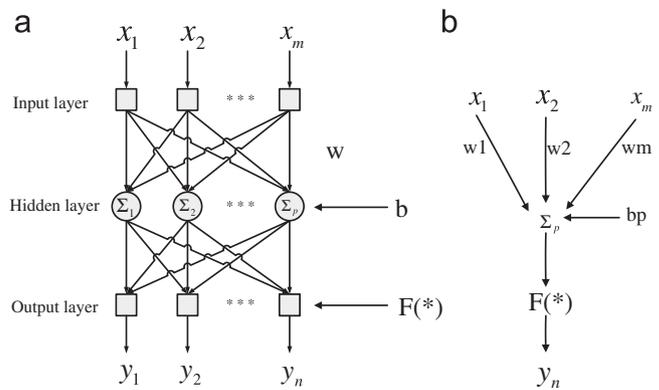
Fig. 1. (a) General architecture of a three-layer feed forward artificial neural network and (b) single neuron calculation of a three-layer feed forward artificial neural network.

to its strong non-linear mapping capability. ANN model is able to learn the underlying relationship between input and output signals of a sequential process with no need to take explicit physical rules into consideration. The training process of ANN is to reach optimal nodal weights. Back Propagation (BP) neural networks are multilayer feed forward networks, which can be trained with the error back propagation learning algorithm. Three-layer structural neural networks containing an input layer, hidden layer, and output layer are currently the most popular and widely adopted neural network pattern. The three-layer linear neural network has its connection neurons between the input and output layer units. Fig. 1(a) is a general architectural description of a three-layer feed forward ANN model with multi-inputs and multi-outputs, which contains only one hidden layer. As shown in Fig. 1(b), neuron p describes the function of biological neuron that is weighting, summation from mapping the inputs to the hidden layer, and the transition to give the output. $\sum_p$ is the summation of x_i multiplied with the weight w_i:

$$\sum_p = \sum w_i x_i + b_j \tag{1}$$

The output neuron y_i is the mapping of $\sum_p$ using $f(*)$ transformation, which is often a monotonic increasing and bounded function, that is

$$y_i = f\left(\sum_p\right) \tag{2}$$

Back propagation algorithm is a widely accepted algorithm in training ANN models. The forward propagation of weight training in ANN will produce simulated results through mapping inputs and outputs, and then the error signal between simulation and observation is back propagated to each neuron to execute gradient descent neuron weight modification in the weight vector space. The weight vectors that reach the minimum value of error function in the network will be adopted after a group of iteration.

2.2. Wavelet function

Although BP neural networks have the advantages of strong non-linear mapping, accuracy and better generality, there may still exist some disadvantages such as dropping into local minimum easily and slow convergence speed. However, it is a worthwhile attempt to transform the traditional BP neural

networks into wavelet neural networks. The major advantages of employing wavelet base function lies in its excellent performance in non-stationary signal analysis and non-linear mapping [22]. According to the wavelet transformation theory, assume that function $\Psi \in L^2(\mathbf{R})$ follows the formula below:

$$C_\Psi = \int \frac{|\Psi_F(\omega)|^2}{|\omega|} d\omega < \infty \tag{3}$$

In Eq. (3), C_ψ is the admissibility constant and Ψ_F is the Fourier transform of Ψ.

The wavelet function follows the following form:

$$\psi_i = |a_i|^{-(1/2)} \Psi\left(\frac{x-b_i}{a_i}\right), \quad a_i, b_i \in \mathbf{R}; \quad i \in Z \tag{4}$$

In Eq. (4), ψ_i is a sub-function from $\Psi(x)$, where a_i is the scalar parameter while b_i is the translation parameter. Then output of the wavelet neural network y_i can be described as the formula given below:

$$y_i = f\left[\sum_{j=1}^p \omega_{ij}\left[\sum_{k=1}^m x_k(t)\psi\left(\frac{t-b_j}{a_j}\right)\right]\right], \quad i = 1, 2, \ldots, n \tag{5}$$

where f is the sigmoid transition function, ψ is the wavelet function, x_k denotes the kth input, y_i denotes the ith output, p represents the number of neurons, and ω_{ij} is the weight between the hidden layer and the output layer. In this study, the Morlet wavelet function was used as the transformation function in wavelet neural networks.

2.3. Nodal weights training with parallelized genetic algorithm

Back propagation training algorithm has advantages of fast convergence and higher accuracy in reaching the optimal criterion. However, back propagation algorithm may have the problem of dropping in local optimum [21]. Such problem of BP algorithm can be solved by integrating genetic algorithm (GA). The genetic algorithm was first proposed by Holland based on Darwin's evolutionary concept [23]. GA has the abilities of global searching, evolutionary adaption properties and can be used as a supplement to help avoiding the insufficient WANN in global optimum searching. GA regards the nodal weights of artificial neural networks as natural evolutionary spices and chromosome in each individual transfers through crossover, mutation to simulate the evolutionary process. Through such a way the training process of the WANN nodal weights can be explained as the evolutionary process of each individual, achieving the better adaptation under certain environment.

GA used in the application of neural network weights training has been widely explored in the last decades and has been successfully carried out as applications referred in many papers [24–26]. But modeling a RR process with WANN model always demands a WANN structure as complicated as the numbers of factors influencing the RR process increase. As a result, a higher number of nodes in WANN model might be requested to avoid bad accuracy and the population size of GA should be correspondingly enlarged to get a uniformly distributed initialization. Moreover, short time step in flood simulation with WANN involves large metrological and runoff observation data leading to a long time in weight training process.

Fortunately, in recent years, the fast development of computing technologies offers the possibility of solving optimization problem that is high-dimensional [27], especially in Fortran MPI

distributed computer system using the power of multiple processors. Parallel computing, however, has not been widely spread in the field in training nodal weights using evolutionary algorithm. In this study, a parallel computing implementation of genetic algorithm for global optimum searching on WANN weights training was proposed to speed up the training process. A detailed description of the method is given below and can be illustrated in Fig. 2.

(1) MPI initialization:
Count the total number of processors (NW) in the MPI communication world and select $N+1$ processors in the communication world to form a processor group in which only one processor is selected as the master node and the rest N processors are labeled as the slave nodes. Then transfer all selected processor groups into a sub-communication world SW where data packages can be sent and received.

(2) Population initialization:
Master—Sample S combinations $\{\alpha_1, \alpha_2 \ldots \alpha_S\}$ using Latin Hypercube sampler [28] while α_S is an array vector of nodal weights that are sequentially arranged. Pack a subset of S and send the package to the slaves in the communication world. Then collect and unpack the evaluated value objective function sent back by the slaves as packages.
Slaves—Receive the package of nodal weights from the Master. Unpack and evaluate the WANN model before sending the packed results back to the Master.

(3) Individual ranking:
Master—Sort the S individuals in the order of increasing evaluation values and store the individuals as parent candidates (Ic).

(4) Parent selection:
Master—Select the parents for reproduction to generate offsprings. A binary tournament selection method is used in the selection. The number of selected individuals from parent candidates is defined as Ip.

(5) Crossover and mutation:
Master—The real-coded genetic algorithm adopts simulated binary crossover (SBX) method [29] using crossover probability pc. The mutation was operated using the polynomial mutation method [29,30] and the mutation probability in this paper is set to a constant pm. The new individuals (In) can be the total individuals generated through crossover and mutation operator.

(6) Evaluation on new individuals:
Master—Divide the new In individuals into subsets familiar to step (2), pack and send them to slaves. Wait and receive the result packages from slaves.
Slave—Receive the package of unevaluated individuals from the master, unpack and evaluate the WANN model before packing and return the results finally.

(7) Off-spring selection:
Master—The off-spring is a combination of new individuals chosen from the current generation. The size of the off-spring may exceed the size of the population S. Sort the off-spring with increasing objective values and the best S off-spring can be selected as the new generation.

(8) Check iteration:
Master—If the current iteration exceeds the maximum iteration number, stop, carry on the BP algorithm. Otherwise return to step (4).

(9) Back propagation
After the maximum iteration has been achieved retrain the WANN again with back propagation algorithm using initial weights of the selected best individual from the population of the last GA iteration.
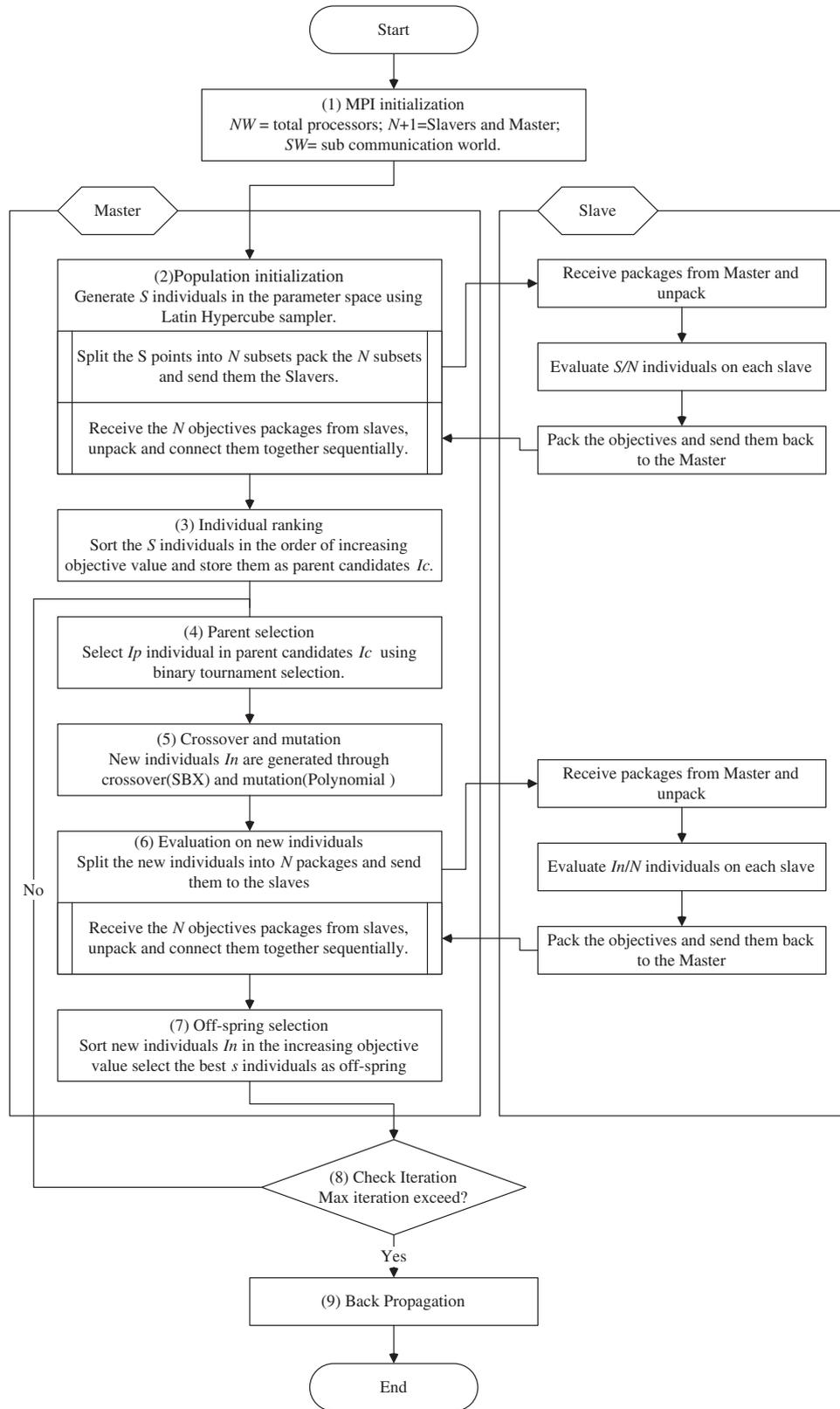
Fig. 2. Flowchart of a parallelized implement of GA algorithm. The master node implement most algorithmic steps, the slave nodes evaluate the WANN model with weight vectors sent from the master.

The GA algorithm is implemented using the MPI parallel technology. The platform of MPICH2 was developed by national laboratory, which can be downloaded from the website freely (http://www.mcs.anl.gov/research/projects/mpich2/). The detailed explanations on how to use parallel functions of MPI, if interested, should be further explored in the manual by users. In this study,

only a few most important MPI subroutines involved are used to realize the communication between nodes of the Master and the Slaves:

(1) MPI_Send ( ): send packaged individuals containing nodal weights to slaves or packaged results of objectives evaluated by the slaves are sent back;
(2) MPI_Get_Elements ( ): to get the size of packages; and
(3) MPI_Recv ( ): to receive the packages from the senders.

3. Application

3.1. Study area

Nen River basin is one of the largest basins located in the northeast of China. The study area, Taonan sub-basin, is the upper area of the Taonan hydrological station on the left part of Nen River basin, as shown in Fig. 3. It is relatively flat on the right part of Taonan sub-basin topologically while quite undulating on the opposite side. The total area of Taonan sub-basin is approximately counted up to 28,452 square kilometers. Taonan sub-basin can be divided into 3 catchments (catchment boundary with violet dash shown in Fig. 3) according to the 3 hydrological station considered that are Taonan, Zhenxi, and Chaersen station. The purpose of the spatial division of the Taonan sub-basin into catchments is to take full advantage of runoff data from each hydrological station and verify the applicability of the GAWANN model. While there scattered totally 36 rainfall stations within the boundary of study area shown in Fig. 3.

3.2. Data preprocessing

The Taonan sub-basin is characterized with arid in winter and rainy in summer. Thus in flood prediction, we carry out simulation only in the flood season from month 6 to 9 yearly. The year 1998 has the heaviest accumulative rainfall amount during flood season in history. It is recorded (the rainfall records from the rainfall stations scattered on the Taonan sub-basin shown in Fig. 3) that the heaviest precipitation in flood season in 1998 reached at 51.45 mm and the maximum discharge at Taonan hydrological station is 1870 m³/s. Thus the simulation input data of year 1998 should be included in GAWANN simulation as the more the extreme historical conditions included in the simulation the more the accuracy it can reproduce in prediction. Hence, the rainfall, stream flow data of 6-h time step for a 14-year period is used in this study. The data from year 1994 to 1998 is then selected for neutral network training while the rest are used for validation.

The observed rainfall data from rainfall station is point representative and when used on catchment scale, the data should be properly interpolated onto the catchments. In terms of the minimum distance interpolation, the Thiessen polygon [31] method is adopted in the interpolation, which is simple, accurate, and computationally time saving. The areal rainfall amount of the catchment can then be obtained by calculating the areal weighted rainfall in terms of the area of the Thiessen polygon within the catchment boundary (shown in Fig. 4).

Another important factor of flood prediction is the evapotranspiration in the catchment. In many research reports, the Penman equation is widely used [32] to calculate areal evapotranspiration (ET). The ET can be described as a function of solar radiation, leaf area index, wind speed, etc. But in this study, the Julian date is adopted to represent the complicated factors of ET because ET in a year changes regularly. As the GAWANN is a black-box model, it does not matter that the Julian date can be used as a replacement of the actual ET.

The data in simulation should be normalized to a specific range as white noise often exits in the time series. The range we chose here is between (0, 1), and the normalization formula can be expressed as

$$P(i) = (p(i) - P_{min})/(P_{max} - P_{min}) \tag{6}$$
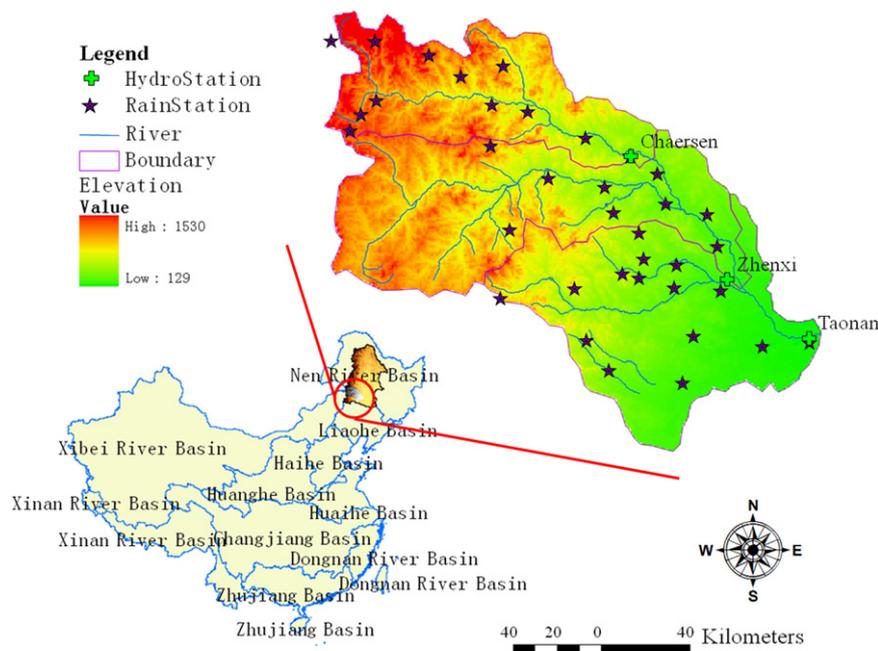
Fig. 3. Study area: Taonan sub-basin located in the upstream of Nen River Basin.

where $P(i)$ is the normalized ith element in time series vector p, P_{min}, and P_{max} are the minimum and the maximum value of vector p, respectively.

3.3. Model structure

Runoff time series has strong autocorrelation as runoff is a kind of wave process and the lag factor has strong effects on the

Fig. 4. Thiessen polygon on each catchment division of Taonan sub-basin.

runoff process. Therefore, it is necessary to select a proper sequential number of time series to describe the time lag in catchment routing. Also the rainfall is critical in runoff generation and may affect the hydrograph at the hydrological station near the catchment outlet. As a result, in terms of the times series theory, the runoff (Q_i, Q_{i+1}, …,Q_{i+N}) at a hydrological station can be written as below:

$$(Q_i, Q_{i+1}, \ldots, Q_{i+N}) = f(Q_{i-1}, Q_{i-2}, \ldots, Q_{i-n}, P_{i-1}, P_{i-2}, \ldots, P_{i-n}, J_i)$$
(7)

In formula (7), ($Q_i,Q_{i+1},…,Q_{i+N}$) represents the output runoff vector containing N days after Q_i, Q_{i-n} is the runoff at interval $i-n$, P_{i-n} is the precipitation amount at interval $i-n$, and J_i is the Julian date at interval i.

The optimal network structure was determined by taking trials, and the result with higher accuracy of training and validation results were preferred. In this essay, only one hidden layer in GAWANN was considered and the number of nodes in hidden layer was set to $Nn+2$, where Nn is the number of inputs. The structure of GAWANN can be three-layer forward neural networks with Nn inputs, $Nn+2$ nodes, and $N+1$ outputs.

3.4. Termination criterion

Error criterion in nodal training is the summation of the root mean square errors (RMSE), which are the square errors

Table 1
Simulation and validation results of GAWNN, where $E_{ss,i}$ and $CE_{ss,i}$ represent the summed RMSE and the mean CE values of all the cases of time lags during simulation, while $E_{sp,i}$ and $CE_{sp,i}$ represent the summed RMSE and mean CE value of all cases of the time lags during validation (i is the catchment index: 1=Chaersen, 2=Zhenxi. and 3=Taonan.), respectively.

Case	Input	Output	Simulation						Validation					
			$E_{ss,1}$ (%)	$E_{ss,2}$ (%)	$E_{ss,3}$ (%)	$CE_{ss,1}$	$CE_{ss,2}$	$CE_{ss,3}$	$E_{sp,1}$ (%)	$E_{sp,2}$ (%)	$E_{sp,3}$ (%)	$CE_{sp,1}$	$CE_{sp,1}$	$CE_{sp,1}$
1	$Q_{t-1},\ldots,Q_{t-2},P_{t-1},\ldots,P_{t-3}$	Qt	0.38	1.05	2.69	0.99	0.99	0.98	0.88	12.51	24.40	0.95	0.77	0.88
		$Qt,Qt+1$	0.54	4.51	1.59	0.98	0.96	0.99	0.62	28.22	22.39	0.91	0.69	0.82
		$Q_t,\ldots,Q_{t+2}$	0.86	7.96	2.13	0.96	0.86	0.97	0.76	39.86	26.33	0.87	0.55	0.78
		$Q_t,\ldots,Q_{t+3}$	0.48	4.00	2.68	0.98	0.93	0.97	2.87	29.62	31.40	0.88	0.61	0.73
		$Q_t,\ldots,Q_{t+4}$	0.51	3.67	3.80	0.98	0.94	0.97	5.22	27.99	40.13	0.85	0.57	0.71
		$Q_t,\ldots,Q_{t+5}$	1.45	4.28	7.57	0.91	0.94	0.96	2.83	34.94	10.60	0.79	0.52	0.89
		$Q_t,\ldots,Q_{t+6}$	3.58	5.95	6.84	0.84	0.93	0.97	6.54	41.93	64.49	0.74	0.42	0.47
		$Q_t,\ldots,Q_{t+7}$	5.93	5.79	10.84	0.81	0.94	0.92	8.68	39.34	74.77	0.76	0.47	0.32
2	$Q_{t-1},\ldots,Q_{t-3},P_{t-1},\ldots,P_{t-4}$	Qt	0.34	0.84	1.57	0.99	0.99	0.98	0.36	11.19	8.77	0.93	0.79	0.82
		$Qt,Qt+1$	0.94	1.17	1.79	0.97	0.98	0.93	0.97	14.42	9.12	0.92	0.81	0.80
		$Q_t,\ldots,Q_{t+2}$	0.92	2.36	1.87	0.99	0.94	0.98	0.76	19.64	10.50	0.85	0.72	0.82
		$Q_t,\ldots,Q_{t+3}$	0.89	3.18	2.14	0.98	0.96	0.91	2.98	24.83	18.77	0.85	0.61	0.78
		$Q_t,\ldots,Q_{t+4}$	1.25	5.58	2.37	0.95	0.95	0.98	2.27	20.67	30.76	0.83	0.83	0.69
		$Q_t,\ldots,Q_{t+5}$	1.15	6.22	3.03	0.97	0.93	0.94	3.40	13.47	28.55	0.86	0.88	0.71
		$Q_t,\ldots,Q_{t+6}$	2.55	7.38	3.24	0.87	0.87	0.92	5.34	38.79	36.78	0.82	0.68	0.63
		$Q_t,\ldots,Q_{t+7}$	1.96	8.11	5.56	0.88	0.83	0.84	6.81	44.43	49.06	0.80	0.51	0.47
3	$Q_{t-1},\ldots,Q_{t-4},P_{t-1},\ldots,P_{t-5}$	Qt	0.31	11.71	6.89	0.98	0.93	0.97	0.81	14.36	7.33	0.93	0.89	0.97
		$Qt,Qt+1$	0.87	7.55	10.38	0.96	0.92	0.96	1.62	18.92	15.65	0.91	0.85	0.91
		$Q_t,\ldots,Q_{t+2}$	1.92	8.47	9.42	0.97	0.96	0.96	2.23	18.55	22.31	0.91	0.85	0.86
		$Q_t,\ldots,Q_{t+3}$	3.95	9.28	17.62	0.86	0.95	0.85	1.92	25.54	27.30	0.92	0.76	0.85
		$Q_t,\ldots,Q_{t+4}$	3.12	11.72	19.62	0.87	0.96	0.82	3.53	61.42	43.32	0.73	0.33	0.58
		$Q_t,\ldots,Q_{t+5}$	2.19	16.56	25.54	0.91	0.91	0.78	4.52	30.45	64.43	0.83	0.73	0.44
		$Q_t,\ldots,Q_{t+6}$	4.23	22.76	24.11	0.85	0.88	0.78	6.13	35.43	45.34	0.82	0.71	0.51
		$Q_t,\ldots,Q_{t+7}$	3.34	32.45	29.36	0.86	0.82	0.72	5.31	33.46	47.65	0.80	0.68	0.53
4	$Q_{t-1},\ldots,Q_{t-5},P_{t-1},\ldots,P_{t-6}$	Qt	0.92	14.97	6.26	0.98	0.86	0.96	3.91	20.53	8.88	0.86	0.79	0.94
		$Qt,Qt+1$	0.95	18.88	17.76	0.99	0.78	0.91	3.94	18.57	20.01	0.87	0.75	0.92
		$Q_t,\ldots,Q_{t+2}$	0.96	22.79	40.43	0.98	0.73	0.79	4.15	23.78	41.77	0.89	0.71	0.79
		$Q_t,\ldots,Q_{t+3}$	1.87	19.35	29.80	0.99	0.77	0.86	3.67	17.51	68.71	0.83	0.78	0.60
		$Q_t,\ldots,Q_{t+4}$	3.57	25.45	21.96	0.99	0.72	0.82	4.07	29.62	81.43	0.82	0.75	0.38
		$Q_t,\ldots,Q_{t+5}$	2.54	28.99	35.71	0.99	0.64	0.80	3.89	34.54	57.34	0.81	0.66	0.64
		$Q_t,\ldots,Q_{t+6}$	4.42	28.54	44.51	0.99	0.64	0.81	3.75	36.08	68.22	0.79	0.65	0.61
		$Q_t,\ldots,Q_{t+7}$	5.59	34.87	50.31	0.93	0.52	0.71	4.18	48.06	65.13	0.77	0.37	0.59

All the test values in Table 1 are selected as the best result of 10 independent trials.

obtained between the observed and the simulated values after the iteration. The total summed RMSE of all the time lags E_s is defined as

$$E_s = \sum_{j=i}^{i+N} e_j \qquad (8)$$

$$e_j = \sqrt{\frac{1}{N_{day,j}} \sum_{t=1}^{N_{day,j}} (Q_{sim(j,t)} - Q_{obs(j,t)})^2} \qquad (9)$$

where $Q_{sim(j,t)}$ and $Q_{obs(j,t)}$ are the simulated and the observed runoff at time t, respectively, e_j is the RMSE value of $Q_{sim(j)}$ and $Q_{obs(j)}$ at lag j, and $N_{day,j}$ is the total number of observations at day lag j.
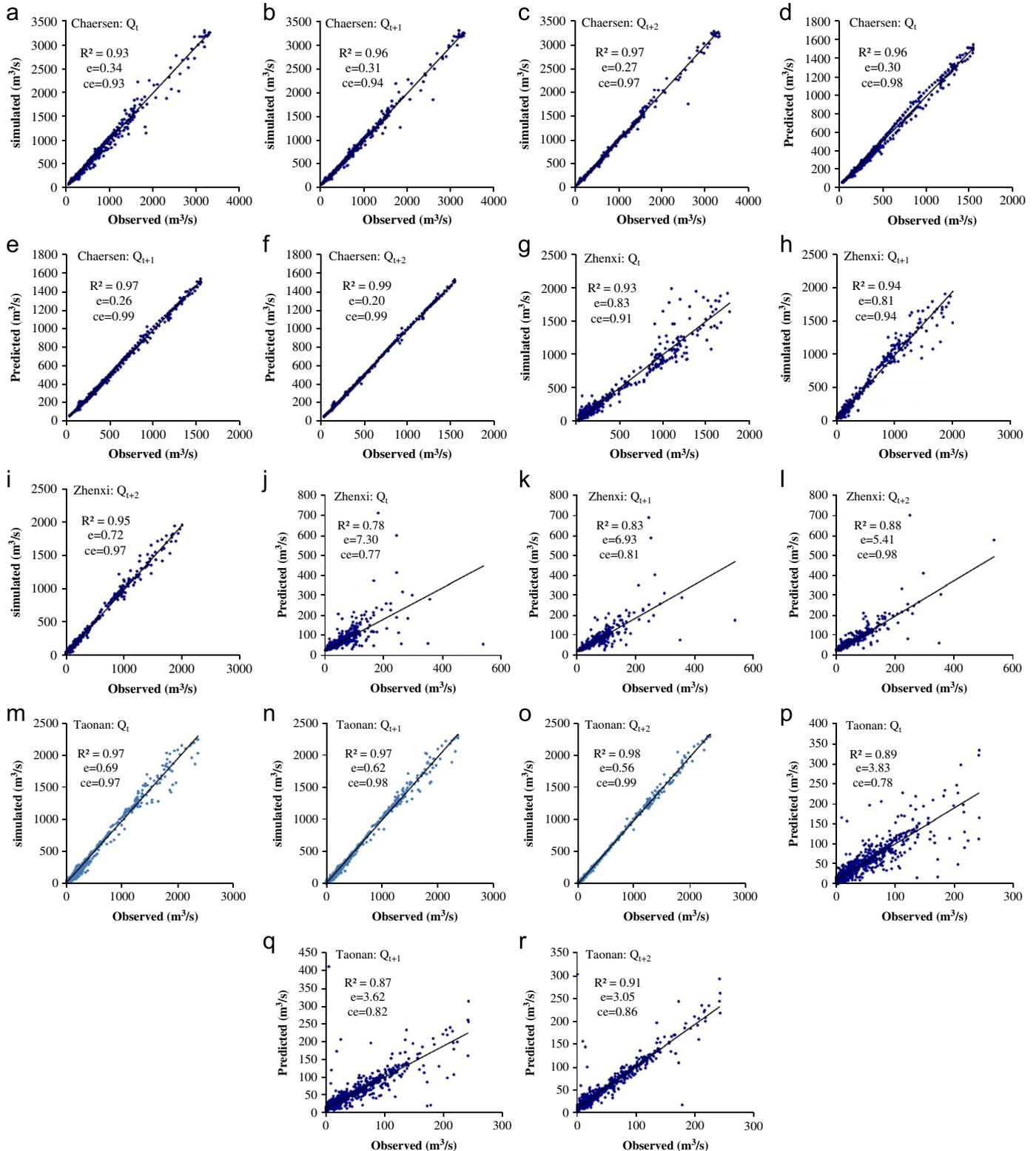
Fig. 5. (a–r) Simulation and prediction results of advisable GAWANN input–output structure: $(Q_t, Q_{t+1}, Q_{t+2}) = f(Q_{t-1}, \ldots, Q_{t-3}, P_{t-1}, \ldots, P_{t-4}, J_i)$, where R^2 is the auto correlation coefficient.

However, another important criterion in flood simulation and prediction is regularly adopted to evaluate the calibration performance, which is called the Nash & Sutcliffe [33] coefficient of efficiency (CE). The CE is widely used in runoff simulation as it is capable of evaluating the higher discharge rate in runoff process when a flood occurs. Considering the different time lags used, the mean CE value of all time lags CE_s can be written as the mean value of ce_j on every output runoff at lag j, defined as

$$CE_s = \sum_{j=i}^{i+N} ce_j / N \qquad (10)$$

$$ce_j = 1 - \sum_{t=1}^{N_{day,j}} (Q_{obs(j,t)} - Q_{sim(j,t)})^2 \Big/ \sum_{t=1}^{N_{day,j}} (Q_{obs(j,t)} - Q_{sim(j,t)})^2 \qquad (11)$$

where $Q_{m,j}$ is the mean value of the observations at lag j. A higher CE value represents a more accurate simulation in runoff value in a flood process as more attention is often paid to the accuracy of higher discharge process of a flood. A CE value of 0.9 and above indicates the best rank in flood simulation while CE below 0.6 is often considered to be poor.

3.5. Training settings

In order to draw ideal training results of GAWANN, the training epoch is set to be 500 generations in the evolution process of genetic algorithm. The population size S was set to 1000 and the size of new individual selection pool is set to 500.
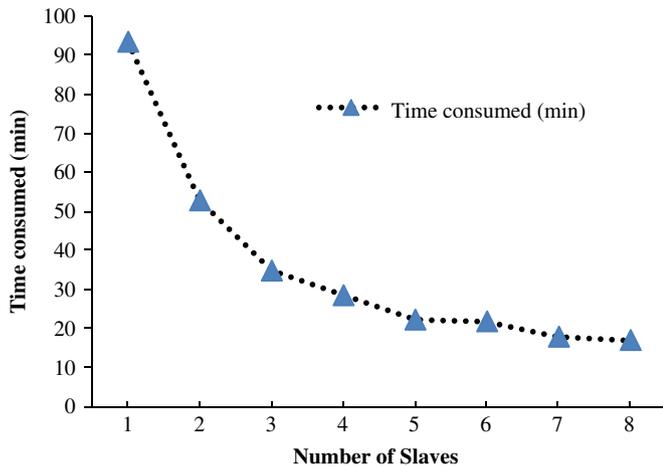
Fig. 6. Computational time needed against the number of slave processors.

The probability of crossover pc and the mutation pm are set to 0.5 and 0.2, respectively. After the last generation, the BP algorithm is implemented using initial nodal weights optimized by GA algorithm at last iteration. The number of iteration in BP algorithm should be set large enough unless the minimum criterion changing rate is reached. In this test, the minimum criterion changing rate is set to 0.1% and the maximum iteration of BP algorithm is set to 4000 avoiding over-fitting.

4. Results and discussion

4.1. Simulation results

Test cases are carried out among 32 input–output conditions. The data from year 1994 to 1998 (5 years) are used for nodal weight training while data from 1999 to 2007 are used in validation. As the 6-h rainfall–runoff data were recorded during flood period mainly from June to September, the total number of data records engaged in the nodal weighs training counts up to 2433. Evaluation results and the performance of the GAWANN model during the flood period is presented in Table 1. Apparently, it infers that GAWANN model has strong ability in flood process simulation and validation. There also exits well non-linear relationship between the summations of RMSE at different day lags and the specified number of inputs meaning that the increasing number of outputs results in the rising trend of RMSE while by contrast causing decreases in the CE value. In application, it is known that a larger number of forecasting time steps mean stronger predictability of GAWANN. However, in general 3 forecasting time steps corresponding to the output $Q_t,\ldots,Q_{t+2}$ (meaning 18 h) are advisable in this study according to the test results as the output series $Q_t,\ldots,Q_{t+2}$ in different cases has comparatively lower RMSE values and higher CE values.

The number of inputs among different cases also has effects on the simulation and prediction results of GAWANN because too many or too few inputs cannot describe the flood process in a right way. The purpose of the test considering different inputs is to determine the number inputs that can represent the flood process well. It can also be seen from Table 1 that among the 4 cases of different cases, the input structure $Q_{t-1},\ldots,Q_{t-3}$, $P_{t-1},\ldots,P_{t-4}$ has better simulation results than the rest cases. The final advisable test result of input–output structure is decided as $(Q_t, Q_{t+1}, Q_{t+2}) = f(Q_{t-1},\ldots,Q_{t-3}, P_{t-1},\ldots,P_{t-4}, J_i)$.

The simulation and prediction results of the flood process are shown in Fig. 5(a–r) from which Fig. 5(a–f) shows the simulated and predicted runoff of Chaersen catchment located on the most upstream of the areal division that have the best results
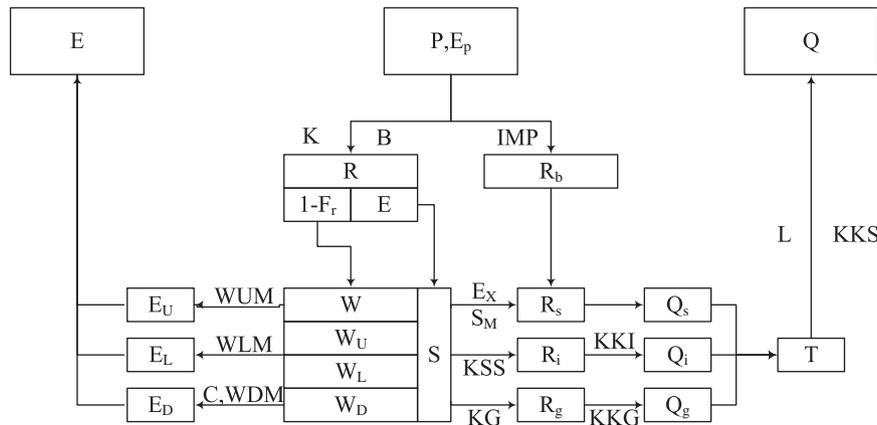
Fig. 7. Structure of Xinanjiang Hydrological Model.

Table 2
Calibration results of Xinanjiang model in Chaersen, Zhenxi and Taonan catchment.

Parameter	Range	Catchment			Explanation
		Chaersen	Zhenxi	Taonan	
K	[0.01, 1.0]	0.42	0.46	0.80	Ratio of the evapotranspiration to the pan evaporation
WUM (mm)	[5, 100]	20.0	30.0	28	Average soil moisture storage capacity of the upper layer
WLM (mm)	[50, 300]	100.0	100.0	100.0	Average soil moisture storage capacity of the lower layer
WDM (mm)	[5, 100]	50.0	60.0	60.0	Average soil moisture storage capacity of the deep layer
SM (mm)	[5, 100]	98.0	96.0	92.0	Areal mean free water capacity
KG	[0.05, 0.7]	0.42	0.34	0.50	Free water division coefficient to groundwater
KSS	[0.2.0.5]	0.12	0.50	0.48	Out flow ratio to the interflow
KKG	[0.9, 0.999]	0.96	0.98	0.98	Recession coefficient of groundwater storage
KKI	[0.7, 0.9]	0.86	0.86	0.79	Recession coefficient of interflow storage
KKS	[0.2, 0.6]	0.35	0.35	0.32	Recession coefficient of surface channel flow
B	Usually follows: $B = \log_{10}(\text{area})/10$	0.33	0.35	0.35	Non-uniform spatial distribution of soil moisture storage capacity over the catchment
C	[0.01, 0.15]	0.08	0.08	0.08	Evapotranspiration coefficient of deeper layer
IMP	[0.01%, 2%]	0.02	0.14	0.04	Percentage of imperious areas in the catchment
E_X	1.20	1.20	1.20	1.20	Exponent of the free water capacity curve
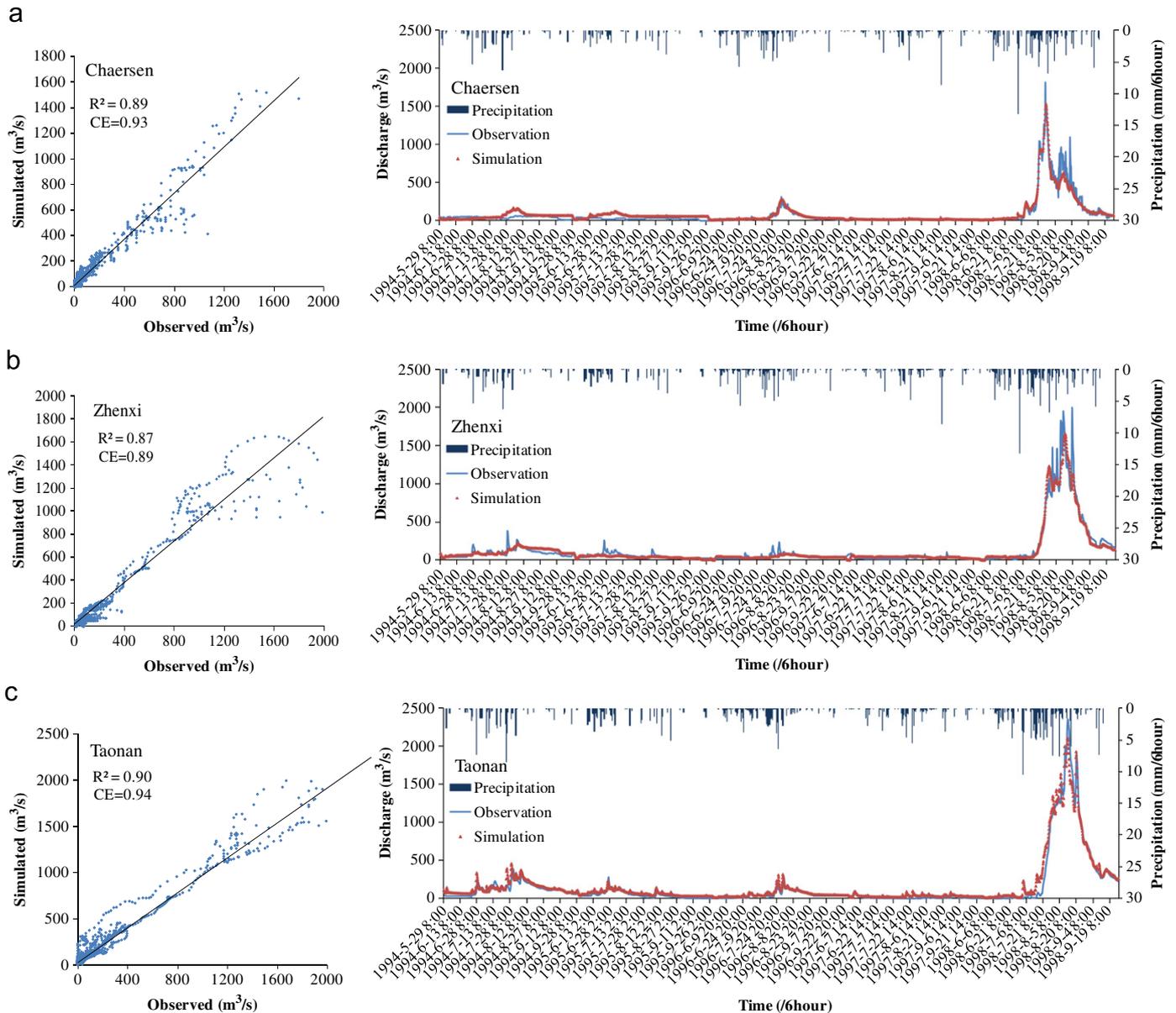
Fig. 8. (a)–(c) Simulated hydrograph using Xinanjiang model with calibrated model parameters.

Table 3
R^2 and CE values of GAWANN and Xinanjiang model outputs considering normal flow year and extreme flow year.

Period	Chaersen				Zhenxi				Taonan			
	GAWANN		Xinanjiang		GAWANN		Xinanjiang		GAWANN		Xinanjiang	
	R^2	CE	R^2	CE	R^2	CE	R^2	CE	R^2	CE	R^2	CE
Normal flow year (1994–1997)	0.93	0.92	0.59	0.62	0.91	0.91	0.55	0.46	0.97	0.96	0.85	0.77
Extreme flow year (1998)	0.98	0.97	0.94	0.93	0.98	0.97	0.93	0.93	0.98	0.98	0.95	0.96

($R^2 > 0.95$) among the three test cases. This is mainly because that in Chaersen catchment, there scattered 13 rainfall gages, the areal gage containing rate of which is far more than those in Zhenxi (Fig. 5(g–n)) and Tanoan (Fig. 5 (o–r)) catchment. Thus the areal precipitation of Chaersen catchment is uniformly averaged among rainfall gages while the areal rainfall information is not enough and heterogeneous in Zhenxi and Taonan catchments. From Fig. 5, it is also indicated that low flow discharges (base flow) are generally slightly under-estimated while some of the peak flow discharges are over-estimated. On the whole, the results reveal that the admirable effectiveness of the GAWANN is beneficial to model for flood simulation and prediction.

4.2. Parallel implement efficiency

To examine the computational efficiencies of the parallel GAWANN algorithm in neural network nodal weight training, 8 CPUs (2.0 GHz) are used for this test case. The structure of this parallel testing is derived as the aforementioned: $(Q_t, Q_{t+1}, Q_{t+2}) = f(Q_{t-1}, \dots, Q_{t-3}, P_{t-1}, \dots, P_{t-4}, J_i)$. Fig. 6 presents a plot of the computational time consumed to evaluate GAWANN versus the number of slave processors. The curve represents the average value of 10 independent trials. It shows that GAWANN in parallel implement contributes to a considerable time saving. As represented, up to 82.7% computational time can be reduced. But the test shows that the time consumption cannot be further reduced to a large extent when the number of processors exceed 5.

4.3. Simulation with Xinanjiang model

To demonstrate the capability of GAWANN model, the conventional Xinganjiang RR model is used in flood simulation. The structure of Xinanjiang model can be illustrated in Fig. 7. Main components of Xinanjiang model include evapotranspiration, infiltration, surface flow, interflow, and ground flow routing. There exit mainly 14 parameters in Xinanjiang model, the calibration of which is always conducted manually. Calibrated optimum and the explanation of Xinanjiang model parameters on the three catchments are listed in Table 2, showing that there are only slight differences of parameters among the catchments. The simulation results of the three catchments are illustrated in Fig. 8(a–c). The CE values of catchments Chaersen, Zhenxi, and Taonan, on the whole, are quite acceptable, reaching 0.93, 0.89 and 0.94, respectively. Compared with GAWANN (using output time series of $Q_t, \dots, Q_{t+2}$), results of Chaersen, Zhenxi, and Taonan catchments show that GAWANN has yielded even better CE values than Xinanjiang model. The comparison of R^2 is quite the same.

However, differences can be further distinguished by separating the whole time series into normal flood period (year 1994–1997) and extreme flood period (year 1998). CE and R^2 values of GAWNN and Xinanjiang model are listed in Table 3. As indicated, Xinanjiang model has not resulted in ideal CE values ($CE=0.62$ in Chaersen, $CE=0.46$ in Zhenxi, and $CE=0.77$ in Taonan) during normal flow year. But with GAWANN, the capability of hygrograph reproduction remains strong reaching up to 0.92 in

Chaersen, 0.91 in Zhenxi, and 0.96 in Taonan. This difference may be caused because Xinanjiang model is originally designed suitable for simulation in humid areas. It indicates that GAWANN is also suitable for simulation where soil moisture is low. In extreme flood year, both GAWANN and Xinanjiang model act well in mapping high flood flow process. The performance of mapping hydrograph in the three catchments during extreme flood year through GAWANN is still slightly higher than Xinanjiang model (0.97 in Chaersen, 0.97 in Zhenxi, and 0.98 in Taonan), showing some advantages of GAWANN in extreme flood simulation. The familiar comparison of R^2 can also be found in Table 3.

5. Summary and conclusions

In this paper, parallelized genetic algorithm integrated with wavelet neural networks is applied to flood simulation and prediction during flood period in arid area of China. The GA method is integrated with BP wavelet neural algorithm to avoid reaching local optimum; However, GA has increased the computational complexity and time consumption of the implement. Thus, the GAWANN was enhanced to be implemented using parallel computation (MPI). The comparison of the test cases indicates that the simulated runoff is strongly related to the rainfall with 4 time steps lag and observed runoff with 3 time steps lag. Meanwhile the prediction capability up to 3 time steps can also be concluded. Furthermore, the parallel implementation of the best test case shows that the proposed parallel GAWANN has good performance in increasing computational efficiency.

In conclusion, the proposed parallel GAWANN in this study consists of the following features:

1. Wavelet function, which has the capability of strong non-linear mapping, was introduced in neural networks. Therefore, it can take the advantage of being strongly non-linear, accurate, and possessing better generality.
2. The genetic algorithm that has the ability of reaching global optimum has been integrated with wavelet neural networks to avoid local optimum that often occurs when using BP algorithm during the nodal weight training.
3. The distributed computation using MPI was further used to strengthen the implement efficiency GAWANN during weight training. At least 82.7% of time consumption can be reduced.
4. The parallelized GAWANN has successfully been applied to the flood simulation and predication in arid area. Meanwhile Xinanjiang model is also compared to demonstrate the capability of simulation precision. On the whole, the simulation and predication capability and computational efficiency of GAWANN are remarkable.

References

[1] R.K. Linsley Jr., et al., Hydrology for Engineers, 2nd ed, McGraw-Hill, New York, 1975.

[2] Kuniyoshi, et al., A BTOP model to extend TOPMODEL for distributed hydrological simulation of large basins, Hydrol. Process 22 (2008) 3236–3251.

[3] K.J. Beven, M.J. Kirkby, A physically based, variable contributing area model of hydrology, Hydrol. Sci. Bull. 24 (1979) 43–69.

[4] K. Beven, TOPMODEL: a critique, Hydrol. Process 11 (1997) 1069–1085.

[5] J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models, part 1. A discussion of principals, J. Hydrol. 10 (1970) 282–290.

[6] M. Lu, T. Koike, N. Hayakawa, A rainfall-runoff model using distributed data of radar and altitude, in: Proceedings of the JSCE No. 411/II-12, 1989, pp. 135–142.

[7] Yangwen Jia, et al., Development of WEP model and its application to an urban watershed, Hydrol. Process 15 (2001) 2175–2194.

[8] Oliver Coroza, et al., Enhancing runoff modeling with GIS, Landsc. Urban Plann. 38 (1997) 13–23.

[9] R.J. Zhao, The Xinanjiang model applied in China, J. Hydrol. 135 (1992) 371–381.

[10] Hongxia Li, et al., Predicting runoff in ungauged catchments by using Xinanjiang model with MODIS leaf area index, J. Hydrol. 370 (2009) 155–162.

[11] A.W. Jayawardena, A modified spatial soil moisture storage capacity distribution curve for the Xinanjiang model, J. Hydrol. 227 (2000) 93–113.

[12] R.J. Zhao, The Xinanjiang model applied in China, J. Hydrol. 135 (1992) 371–382.

[13] C.A. Lin, L. Wen, et al., Real-time forecast of the 2005 and 2007 summer severe floods in the Huaihe River Basin of China, J. Hydrol. 381 (2010) 33–41.

[14] C.A. Lin, L. Wen, et al., Atmospheric–hydrological modeling of severe precipitation and floods in the Huaihe River Basin, China, J. Hydrol. 330 (2006) 249–259.

[15] V.P. Singh, Hydrologic System: Watershed Modeling, vol. 2, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1989.

[16] I. Braud, H. Roux, et al., The use of distributed hydrological models for the Gard 2002 flash flood event: analysis of associated hydrological processes, J. Hydrol. 394 (1–2) (2010) 162–181.

[17] P. Nilsson, C.B. Uvo, et al., Monthly runoff simulation: comparing and combining conceptual and neural network models, J. Hydrol. 321 (2006) 344–363.

[18] Q. Ju, Z. Yu, et al., Division-based rainfall-runoff simulations with BP neural networks and Xinanjiang model, Neurocomputing 72 (2009) 2873–2883.

[19] S. Ahmad, S.P. Simonovic, An artificial neural network model for generating hydrograph from hydro-meteorological parameter, J. Hydrol. 315 (2005) 236–251.

[20] C.L. Wu, K.W. Chau, et al., Methods to improve neural network performance in daily flows prediction, J. Hydrol. 372 (2009) 80–93.

[21] Y. h. Chen, F.J. Chang, Evolutionary artificial neural networks for hydrological systems forecasting, J. Hydrol. 367 (2009) 125–137.

[22] Z. Jun, G.G. Walter, et al., Wavelet neural networks for function learning, Signal Process. 43 (1995) 1485–1497.

[23] J.H. Holland, Adaptation in Natural and Artificial Systems, 2nd ed., Massachusetts Institute of Technology, Cambridge, 1975.

[24] E.G. Bekele, J.W. Nicklow, Multi-objective automatic calibration of SWAT using NSGA-II, J. Hydrol. 341 (2007) 165–176.

[25] A. Abraham, Meta learning evolutionary artificial neural networks, Neurocomputing 56 (2004) 1–38.

[26] P. Chaves, F.J. Chang, Intelligent reservoir operation system based on evolving artificial neural networks, Adv. Water Resour. 31 (2008) 926–936.

[27] Jasper A. Vrugt, et al., Application of parallel computing to stochastic parameter estimation in environmental model, Comput. Geosci. 32 (2006) 1139–1155.

[28] A. Van Griensven, T. Meixner, et al., A global sensitivity analysis tool for the parameters of multi-variable catchment models, J. Hydrol. 324 (2006) 10–23.

[29] Kalyanmoy Deb, R.B. Agarwal, Simulated binary crossover for continuous search space, Complex Syst. 9 (1995) 115–148.

[30] M.M. Raghuwanshi, O.G. Kakde. Survey on multiobjective evolutionary and real coded genetic algorithms, in: Proceedings of the 8th Asia Pacific Symposium on Intelligent and Evolutionary Systems, 2004, pp. 150–161.

[31] T.E. Croley, H.C. Hartmann, Resolving Thiessen polygons, J. Hydrol. 76 (1985) 363–379.

[32] J.D. Valiantzas, Simplified versions for the Penman evaporation equation using routine weather data, J. Hydrol. 331 (2006) 690–702.

[33] J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models, I. A discussion of principles, J. Hydrol. 10 (1970) 282–290.

Yuhui Wang receives his B.S. and M.S. degrees from the School of Environmental Science and Engineering, Donghua University, Shanghai, China, in year 2003 and 2005, respectively. Since then, he carried out his Ph.D. in the same university till now. His main research interests include neural networks, distributed hydrological modeling, water quality simulation, water resources assessment.

Hao Wang is the academician of Chinese Academy of Engineering from China Institute of Water Resources and Hydropower Research (IWHR). He is a famous expert in hydrology and water resources planning and management.

Xiaohui Lei is an associate professor in China Institute of Water Resources and Hydropower Research, focusing on hydrological modeling, flood forecasting and water resources operation. He has developed a distributed hydrological model, EasyDHM, which is water resources and flood management oriented. The EasyDHM model system makes model building, calibration and analysis easy, which makes it available to practical water resources and flood management.

List of professional qualifications: 1998, B.E. from the Hydraulic Engineering, Tsinghua University, Beijing, China; 2001, M.E. from the Department of Hydrology and Water Resources, IWHR (China Institute of Water Resources and Hydropower Research), Beijing, China; 2007, PhD in Life and Environmental Science, University of Tsukuba, Ibaraki, China.

Yunzhong Jiang is a professor in China Institute of Water Resources and Hydropower Research, focusing on hydrological modeling, flood forecasting and water resources operation.

As one of the project leaders, Jiang has accomplished several China National Key-technologies R&D Programs. In addition, he has published several hydrological database standards and above 30 papers on water resources operation and hydrological modeling.

List of professional qualifications: 1991, B.S. from HoHai University, Nanjing, China; 1996, Ph.D. from the Dalian University of Technology, Dalian, China

Xinshan Song is the associate professor in the School of Environmental Science and Engineering, Donghua University, Shanghai, China. In recent years, he has published more than 20 papers and several books especially in environmental mathematical simulation, environmental assessment and planning.