Research Article

Jie Wang, Guoqing Wang*, Amgad Elmahdi, Zhenxin Bao, Qinli Yang, Zhangkang Shu, and Mingming Song

Comparison of hydrological model ensemble forecasting based on multiple members and ensemble methods

<https://doi.org/10.1515/geo-2020-0239>

received July 21, 2020; accepted March 02, 2021

Abstract: Ensemble hydrologic forecasting which takes advantages of multiple hydrologic models has made much contribution to water resource management. In this study, four hydrological models (the Xin'anjiang model (XAJ), Simhyd, GR4J, and artificial neural network (ANN) models) and three ensemble methods (the simple average, black box-based, and binomial-based methods) were applied and compared to simulate the hydrological process during 1979–1983 in three representative catchments (Daixi, Hengtangcun, and Qiaodongcun). The results indicate that for a single model, the XAJ model and the GR4J model performed relatively well with averaged Nash and Sutcliffe efficiency coefficient (NSE) values of 0.78 and 0.83, respectively. For the ensemble models, the results show that the binomial-based ensemble method (dynamic

weight) outperformed with water volume error reduced by 0.8% and NSE value increased by 0.218. The best performance on runoff forecasting occurs in the Hengtang catchment by integrating four hydrologic models based on binomial ensemble method, achieving the water volume error of 2.73% and NSE value of 0.923. Finding would provide scientific support to water engineering design and water resources management in the study areas.

Keywords: hydrological model, ensemble forecasting, method comparison, flood forecasting

1 Introduction

Hydrological models have played essential roles in many applications such as flood-control and disaster reduction [1], water resources utilization [2], hydraulic engineering construction [3], and pollution evaluation [4]. Hydrologic models have developed from the experience-based stage to the model study stage, including the lumped, semi-distributed, and distributed models. Although loads of efforts have been made and diverse hydrological models [5,6] have been proposed, hydrological models are still remains challenges on describing the hydro-physical process more accurately [7,8].

Different hydrological models consider different emphases in water cycle simulation, which results in different simulation results. Furthermore, different models have different applicabilities in different climatic zones and underlying surface conditions (such as terrain, vegetation, urbanization) due to the calculation methods in hydrological processes [9–11]. Liu et al. compared the applicability of the Xinanjiang model [12], the unsaturated runoff model, and time-variant gain model (TVGM) in arid and semiarid region of China. The results indicated that the TVGM model considering rainfall intensity was the most applicable model in the area. Wang et al. compared the applications of the Xinanjiang, Topmodel,

* **Corresponding author: Guoqing Wang**, State Key Laboratory of Hydrology – Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China; Research Center for Climate Change, Ministry of Water Resources, Nanjing 210029, China; Yangtze Institute for Conservation and Development, Nanjing 210029, China, e-mail: gqwang@nhri.cn

Jie Wang: State Key Laboratory of Hydrology – Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China; Research Center for Climate Change, Ministry of Water Resources, Nanjing 210029, China; Yangtze Institute for Conservation and Development, Nanjing 210029, China

Amgad Elmahdi: MENA Region-International Water Management Institute-IWMI, 3000 Cairo, Egypt

Zhenxin Bao, Zhangkang Shu, Mingming Song: State Key Laboratory of Hydrology – Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China; Research Center for Climate Change, Ministry of Water Resources, Nanjing 210029, China

Qinli Yang: State Key Laboratory of Hydrology – Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China; School of Resources and Environment, University of Electronic Science and Technology of China, No. 2006, Xiyuan Avenue, Chengdu 611731, China

and artificial neural network (ANN) models in the small and medium-sized river basins in China [13]. After comparing the ANN, Hydrologiska Byråns Vattenbalansavdelning-D (HBV-D), and Soil and Water Integrated Model (SWIM) models, Gao *et al.* concluded that different models were suitable for different scales of time and space [14,15]. The results showed that the ANN performed better on the monthly scale and the other two models are more applicable on the daily scale. Seann *et al.* found that hydrological models always had higher uncertainty for small catchments than large or medium catchments [16]. Hassan and Lahcen compared the simulation results of the lumped and distributed models, namely, the Hydrologic Engineering Center's-Hydrologic Modeling System (HEC-HMS) and the L'Atelier Hydrologique Spatialisé model (ATHYS), in the Issen basin [17].

In order to reduce the uncertainty in the application of hydrological models and improve hydrologic forecasting accuracy, the ensemble hydrological prediction was developed by taking the advantages of different models into account [18,19]. Ensemble simulation approach could be classified as input ensemble, process ensemble, and result ensemble. For input ensemble approach, the ensembled input data which are based on multiple inputs are used to drive model and produce output. This approach has been widely applied in hydrological forecasting. For example, numerous studies on the ensemble flood forecasting are based on ensemble precipitation forecasts driving hydrological model [20,21]. In the process ensemble, ensemble methods consist of different approaches to each hydrologic cycle subprocess. For example, Bao *et al.* developed a distributed model by coupling a mixed runoff generation model and 2D overland flow routing model (GMKHM-2D) [22]. In the result ensemble, the simulation results of multiple models are ensemble [23].

In this paper, the authors mainly focus on the result ensemble. For example, Jasper *et al.* forecasted the inflow of Lake Maggiore using the five-model ensemble [24]. Davolio forecasted floods by using six different rainfall-runoff models in northern Italy [25]. Cloke and Pappenberger developed the Ensemble Forecasting System (EPS) in flood forecasting, which is based on the Monte Carlo structure [26]. A hydrologic model output statistical method (HMOS) is used for integrated forecasting of short-term runoff [27]. Arsenault *et al.* compared nine ensemble methods to integrate the results of four lumped hydrological models and it was found that the Granger-Ramanathan average variant C (GRC) performed the best [28]. Huo *et al.* used seven models to simulate runoff in the semi-humid area of northern China and then calculated the ensemble runoff using the Bayesian model, which

indicated that XAJ model performed better and simulation accuracy was improved after ensemble [29,30]. This paper will also discuss the influence of model numbers. Arsenault *et al.* improved the traditional ensemble method and proposed a new idea combining multi-inputs and a multi-model ensemble [31]. Globally, most countries and regions have adopted the ensemble forecasting method. For example, Lorenzo *et al.* evaluated the ensemble streamflow prediction in Europe using the European Flood Awareness System (EFAS), which is a multi-model ensemble approach [32]. The result was that the simulation precision of small catchments was obviously lower than that of large basins.

From the literature, it can be seen that the hydrological ensemble uncertainty of small and medium-sized scale catchments is higher than large scale catchment [33]. Therefore, it is necessary to conduct in-depth studies on small and medium scales catchments. In this paper, the three small and medium-sized catchments located in the mountainous area of western Zhejiang Province are selected as study cases. By studying and analyzing the applicability of the four models, the simulation accuracy under different ensemble methods and the number of ensemble members are compared. This paper is aiming to not only improve the accuracy of the hydrological forecasting, but also contribute to revealing the hydrological laws characteristics of the small and medium-sized catchments. The objectives of this study are to: (1) evaluate the performance of four individual hydrologic models on runoff forecasting in three selected areas, and (2) evaluate and compare the performances of ensemble models on runoff forecasting by taking different ensemble methods.

2 Data and methods

2.1 Study area and data

In this study, three catchments were selected as the study areas, namely, the Hengtangcun, Qiaodongcun, and Daixi catchments. They are all located in the mountainous areas of the western Tai Lake basin and in the western part of Zhejiang Province, China, with catchment areas of 1,373, 162, and 233 km², respectively. They belong to the typical subtropical monsoon climate zone where precipitation is unevenly distributed within the year. The southeast monsoon prevails in summer, with abundant water vapor and precipitation. The flood season is

Figure 1: Location of hydro-meteorological stations in the study areas.

generally from May to September. There are 25 rainfall stations and three hydrologic stations located within the study areas (Figure 1).

In this study, we use STRM (Shuttle Radar Topography Mission) Digital Elevation Model (DEM) with a resolution of 90 m to produce river system and catchment boundary. The STRM DEM was derived from Geospatial Data Cloud (<http://www.gscloud.cn/>). Daily precipitation data from the 25 rainfall stations from 1979 to 1983 were obtained through the water yearbook of the Ministry of Water Resources of the People’s Republic of China. The observed streamflow data of the four hydrologic stations and observed evaporation data were also obtained from the water yearbook (Table 1).

Previous studies indicate that the Taihu River basin is located in the eastern Asian Monsoon climate zone with a high variability of annual runoff (180–1,450 mm) [34]. Statistical results for data series from 1979 to 1983 in this study show that the annual runoff ranges from 200

to 1,300 mm which approximately covers most of the range of annual variability in previous studies. We therefore believe that the 5-year data series (1979–1983) of runoff has a certain representativeness.

2.2 Hydrological models

Four hydrologic models were used to simulate runoff in this study: the Xinanjiang (XAJ), Simhyd, GR4J, and Back Propagation (BP) neural network models.

The XAJ model is a conceptual rainfall-runoff model developed by HoHai University, China [35–37]. The XAJ model is based on the mechanism of saturation excess. The key of the model is the storage capacity curve. This model consists of four parts, namely, runoff generation, evapotranspiration, water source partition, and flow concentration. In this paper, a classical XAJ model (three

Table 1: Data information

Data type	Period	Temporal resolution	Spatial resolution	Data source
DEM data	—	—	90 m	Geospatial Data Cloud (http://www.gscloud.cn/)
Precipitation data	1979–1983	Daily	—	Water yearbook of the Ministry of Water Resources of the People’s Republic of China
Streamflow data	1979–1983	Daily	—	
Evaporation data	1979–1983	Daily	—	

components: surface runoff, underground runoff, and interflow) with sixteen parameters is applied.

The Simhyd model is a conceptual daily rainfall and runoff model [38,39], which takes both infiltration excess runoff and saturation excess runoff into consideration. The model is mainly composed of six parts: interception, evaporation, infiltration, soil water storage, basic flow, and flow concentration. Total runoff is the sum of surface runoff, interflow, and basic flow, with a total of seven parameters.

The GR4J model is a conceptual hydrological model [40,41]. The calculation process of this model can be generalized to two nonlinear reservoirs, called runoff generation reservoir and confluent reservoir, with only four parameters.

The BP neural network model is a black box model with a strong deep-learning ability [42]. The model consists of an input layer, a hidden layer (calculation layer), and an output layer, as shown in Figure 2. The number of neurones depends on the number of input data (i.e., the number of rainfall station for this study) and objective output (i.e., runoff of watershed outlet). The number of neurones in hidden layer will depend on empirical formula based on previous study experience. In this paper, a three-layer BP model is applied with one hidden layer. And the numbers of neurones in input, hidden, and output layers are 13, 10, and 1, respectively (taking Hengtang River Basin as an example). So the structure of BP model is 13–10–1. There are four matrices as parameters to connect the calculation between layers, including the two threshold matrices of hidden layer (β_j , β_m) and output layer and two connection weights between input and hidden and hidden and output layer (W_{ij} , W_{jm}).

It is well-known that the over-fitting and under-fitting are normal problems in many machine learning algorithms. In this paper, we divide the input data into three parts by random sampling, including training data,

validation data, and test data. Then we can judge the generalization ability model according to the different simulation performances in different data periods. After model built and calculation, we noticed that the training error is close to the validation set error, and both are relatively small. It can be regarded that there is almost no over-fitting and under-fitting problem in this paper. So, the BP model can be well applied to hydrological simulation.

The parameters of the four models are shown in Table 2. Experience shows that the XAJ model is well developed and more suitable for humid areas, but has many parameters [35–37]. The structure of the GR4J model is simple with few parameters, but is not widely used. The BP model has only four parameters, but it does not have a specific physical mechanism. The Simhyd model has seven parameters, making it more suitable for semi-humid and semiarid areas.

2.3 Multi-model ensemble methods

Three ensemble methods were used to integrate the results of each model: the simple average method, the black box ensemble method (BP neural network), and the induction of binomial coefficient ensemble method.

- The simple average method is not complex. This method only needs to average the simulation results of the four models.
- The black box ensemble method used in this paper is the BP neural network, which was mentioned in section 2.2. The theory of this method is the same in the application of hydrological models and ensemble methods, so it is not repeated here. The difference is that the input and output of the former are the precipitation data of the rainfall stations and simulated runoff, respectively, while the simulation results of different models and ensemble streamflow are applied to the latter. When BP model was used as ensemble method, the number of neurones depends on the number of input data (the number of individual model) and objective output (ensemble runoff of watershed outlet), and then the number of neurones is calculated as mentioned empirical methods above. In this paper, the structure of BP ensemble model is 4–9–1, 3–7–1 for three-model ensemble and four-model ensemble, respectively.
- The induction of binomial coefficient ensemble method uses the pre-precision to rank and weigh the simulation results at the next time to get the ensemble forecast results at the next time. For the integration method of

Figure 2: Framework of BP model.

Table 2: The parameters of the four models

Model	Meaning (Unit)	Symbol	Meaning (Unit)	Symbol
XAJ model	Reduction coefficient of evaporation	Kc	Storage capacity of the upper layer (mm)	WLM
	Storage capacity of free water (mm)	SM	Storage capacity of the deeper layer (mm)	WDM
	Outflow coefficient of free water storage to groundwater	Kg	Impervious area proportion	Imp
	Outflow coefficient of free water storage to interflow	Ki	Exponent of storage capacity curve	B
	Regression coefficient of groundwater storage	Cg	Evapotranspiration coefficient of deeper layer	C
	Regression coefficient of interflow storage	Ci	Exponent of free water storage capacity curve	EX
	Water storage recession coefficient of river network	CR	Coefficient of hysteresis calculation (h)	L
GR4J model	Storage capacity of the upper layer (mm)	WUM	Regression coefficient of surface runoff	CS
	Capacity of runoff generation reservoir (mm)	X1	Capacity of runoff flow concentration reservoir (mm)	X3
BP model	Exchange coefficient of groundwater (mm)	X2	Flow concentration time of unit hydrograph	X4
	Threshold matrix of hidden layer	β_i	Connection weight between input and hidden layer	W_{ij}
	Threshold matrix of output layer	β_m	Connection weight between hidden and output layer	W_{jm}
Simhyd model	Interception ability (mm)	INSC	Loss coefficient of infiltration	SQ
	Maximum loss of infiltration (mm)	COEFF	Proportional coefficient of groundwater equation	CRAK
	Soil water-holding capacity (mm)	SMSC	Recession coefficient of basic flow	Kg
	Proportional coefficient of interflow equation	SUB		

N set members, and assuming an ensemble method with N members, the basic idea is as follows.

- (1) The weights of N models are calculated according to the following formula and arranged from the smallest to the largest:

$$W_n = C_{2N-1}^{n-1} / 2^{2(N-1)}, \quad n = 1, 2, \dots, N \quad (1)$$

where W_n is the weight of every model, C is binomial coefficient.

- (2) Calculate the forecasting accuracy of the n_{th} model at time t , and arrange them in order from smallest to largest:

$$A_{n,t} = \begin{cases} 1 - \frac{|Q_{RE,t} - Q_{n,t}|}{Q_{RE,t}}, & \frac{|Q_{RE,t} - Q_{n,t}|}{Q_{RE,t}} < 1 \\ 0, & \frac{|Q_{RE,t} - Q_{n,t}|}{Q_{RE,t}} \geq 1 \end{cases} \quad (2)$$

where $A_{n,t}$ is the accuracy of the n_{th} model at time t ; $Q_{RE,t}$ is the observed runoff at time t , m^3/s ; $Q_{n,t}$ is the simulated runoff at time t of n_{th} model, m^3/s .

- (3) According to the forecasting accuracy at time t , the modified simulation value at time $t + 1$ was adjusted to select and sort:

$$Q_{n,t+1} = Q_{m,t+1}, \quad \frac{A_{m,t} - A_{n,t}}{A_{m,t}} > 0.5 \quad (3)$$

where $A_{m,t}$ is the maximum value of $A_{n,t}$.

- (4) Calculate the ensemble forecasting runoff at time $t + 1$:

$$Q_{t+1} = \sum_{n=1}^N Q_{n,t+1} \times W_n \quad (4)$$

Comparing these ensemble methods, the simple average method is much easier and the weight of each member is identical and unchanged. The weights of the BP ensemble method do not change after they are determined. With the binomial ensemble method based on induced ordering, the ranking is continuously adjusted according to the prediction accuracy at each moment.

2.4 Parameter calibration method and precision estimation method

The SCE-UA intelligent optimization algorithm was used to optimize the parameters of each model in this paper [43]. And the relative water volume error (ER) and the NSE were taken as objective functions as the basis for the end of parameter optimization.

$$ER = \frac{\sum_{t=1}^T Q_{RE}(t) - \sum_{t=1}^T Q_{SI}(t)}{\sum_{t=1}^T Q_{RE}(t)} \times 100\% \quad (5)$$

$$NSE = 1 - \frac{\sum_{t=1}^T (Q_{RE}(t) - Q_{SI}(t))^2}{\sum_{t=1}^T (Q_{RE}(t) - \overline{Q_{RE}})^2} \quad (6)$$

where $Q_{RE}(t)$, $Q_{SI}(t)$ represent the observed and simulated runoff on day t , respectively, m^3/s ; $\overline{Q_{RE}}$ is the average of observed runoff, m^3/s ; T is the total simulation time, day.

3 Results

3.1 The simulation results of individual models

During the study period, the observed runoff ranges from about 200 to 1,300 mm per year, which indicates that the data of these three catchments have certain representativeness. Four hydrological models were used to simulate the daily runoff process in the three catchments. The warm-up period, calibration period, and validation period of the model were 1978 (using 1979 data), 1979–1981, and 1982–1983, respectively. The SCE-UA algorithm was used for parameter optimization. After calculating and estimating, the simulation results of each ensemble member were obtained as shown in Table 3. And the comparison of the simulated and recorded hydrographs is shown in Figure 3.

Overall, the XAJ and GR4J models performed much better than the other models. The NSEs of the simulation results of the three catchments were above 0.6 and the average value was 0.78 and 0.83. The average relative error of water volume was relatively low, at 9.4 and 10.6% respectively, which shows that the two models have good adaptability in the three catchments. Compared with the XAJ and GR4J models, the simulation accuracy of the Simhyd and BP models was relatively low. The relative error of water volume in the Qiaodongcun catchment was within the permissible error (20%) in the period of calibration and validation, and the NSE was greater than 0.5 [44]. But the water volume error exceeded the permissible range and the NSE values were low in other two catchments, indicating that the two models had relatively poor adaptability in the study area.

The study areas in this paper are humid areas, and the results are consistent with previous studies that indicate that the Simhyd model is more suitable for semi-humid and semiarid areas. At the same time, the BP model is an ANN, which does not belong to the category of hydrological model in terms of the model structure. The theory of this model was just used to calculate in this paper without a clear physical mechanism, so the uncertainty and error range may be relatively large.

In addition, the simulation results in the first year, 1979, are generally worse than those in other years, and the error is relatively large. It is considered that the simulation sequence length is a little short and the initial state of the model is unstable, while the influence of the initial

Table 3: Simulation results of each model

Catchment	Period	Year	R_ob/mm	ER/%				NSE			
				XAJ	Simhyd	BP	GR4J	XAJ	Simhyd	BP	GR4J
Daixi	C	1979	274.95	−44.6	−53.8	−94.3	−62.8	0.76	0.25	0.33	0.80
		1980	670.78	−7.8	20.8	−34.2	−17.0	0.85	0.48	0.54	0.87
		1981	955.66	−0.9	11.0	22.0	−7.5	0.78	0.59	0.61	0.85
	V	1982	512.23	−15.5	12.9	5.9	−24.4	0.79	0.59	0.64	0.85
		1983	1155.78	4.1	−1.7	−9.8	−2.8	0.75	0.65	0.52	0.88
Hengtang	C	1979	432.11	24.2	−4.7	−22.2	6.8	0.67	0.32	0.53	0.68
		1980	879.83	−7.2	14.6	−2.3	−9.5	0.81	0.63	0.69	0.81
		1981	914.68	−5.0	2.6	4.3	−4.1	0.88	0.47	0.43	0.88
	V	1982	697.06	1.9	19.2	8.0	1.4	0.87	0.52	0.53	0.82
		1983	1215.12	10.4	7.5	−5.2	6.1	0.90	0.70	0.76	0.89
Qiaodong	C	1979	517.69	15.2	−9.0	−8.0	5.4	0.72	0.59	0.68	0.81
		1980	1010.36	1.7	1.6	−3.8	−2.2	0.72	0.62	0.56	0.82
		1981	1107.4	1.0	5.2	−1.6	−0.4	0.67	0.59	0.38	0.80
	V	1982	677.72	0.2	17.1	−17.9	4.1	0.70	0.52	0.37	0.83
		1983	1138.46	1.1	−4.8	−3.3	−5.1	0.78	0.78	0.79	0.86

Note: C means calibration period; V means validation period in the table.

Figure 3: Comparison between simulated and observed hydrograph (taking Hengtang catchment as an example). (a) XAJ, (b) Simhyd, (c) BP, and (d) GR4J.

state of the model on the prediction could not be eliminated in the warm-up period setting.

3.2 Simulation results of the multi-model ensemble

According to the simulation results of the four ensemble members, three ensemble methods were selected for calculation. In this part, all model simulation results were used as ensemble members, which means a four-model ensemble here. The results of each ensemble method are shown as follows.

3.2.1 The simple average ensemble method

After calculating by the simple average ensemble method, the water volume errors of the Daixi, Hengtang, and Qiaodong catchments were 15.0, 2.3, and 0.2%, respectively, and the NSE values are 0.802, 0.814, and 0.749, respectively. Figure 4a shows that in terms of the relative error of water volume, the simple average ensemble method had the greatest improvement on the BP model, and the water volume errors in 1979 and 1980 in Daixi catchment decreased by 34.5 and 24.6%, respectively. The other two catchments also made improvement, but the improvement was relatively small.

For the GR4J model, there was little change, which is not obvious. Apart from the error of the Daixi catchment

decreasing 19.1% in 1982, the improvement basically revolves around the circle represented by zero value. For the XAJ model and Simhyd model, water error increased significantly in a few years (in 1979), but slightly in other years. The effect of the ensemble on water error in 1979 was greater than that in other years, and in the three catchments, the ensemble improved the results of the BP model in 1979, while there was just a little improvement for the other three models in 1979. In addition, overall, the water volume errors in the Daixi and Hengtang catchments were reduced more, in comparison to Qiaodong catchment.

On the other hand, the NSE (Figure 4b) and comparing the simulation results of the Simhyd and BP models both indicated that simple average ensemble method increased accuracy of simulations for all three catchments, with an average NSE increases of 0.238 and 0.237, respectively. The increases in the Daixi and Hengtang catchments were more obvious. The NSE of the XAJ model has shown little improvement, falling within the range of ± 0.08 . While, for GR4J, most NSEs decreased, with an average decrease of 0.04.

3.2.2 Ensemble method based on the black box

In this case, the black box ensemble method is the BP neural network. The data from 1979 to 1981 were used for the training network and the calibrating parameter matrix, while the data from 1982 to 1983 were used for validation. After the black box ensemble, the average

Figure 4: Improvement of the simple average ensemble method compared with ensemble members. Note: (a) represents ER Improvement; (b) represents NSE improvement. The radius of the circle in the figure represents the improvement of the ensemble method. The positive value represents an increase in accuracy compared with ensemble members, while the negative value represents a decrease in accuracy. In other words, the circle periphery with a value of 0 represents an increase in accuracy after ensemble, while the circle interior represents a decrease. There are 15 black radii here. The green, pink, and blue outlines represent the simulation results of the Daixi, Hengtang, and Qiaodong catchments, respectively. The results from 1979 to 1983 are in anticlockwise order, the same as the Figures 4 and 5.

Figure 5: Improvement of the black box ensemble method compared with ensemble members. (a) Represents ER Improvement; (b) represents NSE improvement.

water volume errors in the Daixi, Hengtong, and Qiaodong catchments were respectively -11.9 , -9.0 , and -8.2% and the NSEs were 0.852 , 0.831 , and 0.825 . The improvement of the black box ensemble forecasting method relative to the individual ensemble members is calculated, as shown in Figure 5.

From the aspect of water volume relative error, the black box ensemble method had a higher degree of improvement on the BP and GR4J models, and the improvement was mainly reflected in the decrease of error on the Daixi catchment (19.1 and 4.8%). The improvement of the XAJ model and Simhyd model was not observed, and even the error of a few years increases slightly after ensemble. In particular, the black box ensemble method significantly decreased the simulation error in 1979 (except the XAJ simulation results of the Hengtang catchment), which indicates that the ensemble method can reduce the error caused by the uncertainty of the initial state of a single model to a certain extent and plays a certain role in correction.

For the NSE values, the simulation results of three ensemble members in all study areas improved greatly. And there was a slight decrease in the NSE in some individual catchments, but the decrease was not more than 0.01 . In particular, the improvement was greatest for the Simhyd and BP models, and the NSE increased by about 0.29 and 0.27 on average. Because of the problem of model applicability, the NSE values are relatively low in the simulation results of these two models, which effectively improved after applying the ensemble method.

3.2.3 Ensemble method based on the induction of binomial coefficient

The binomial coefficient ensemble method based on induced ordering involves calculating the ensemble forecasting value of the next moment by using the prediction precision of the current moment to induce the weight of the

Figure 6: Improvement of the induction of binomial coefficient ensemble method compared with ensemble members. (a) Represents ER Improvement; (b) represents NSE improvement.

Figure 7: Comparison of improvement ((a) ER improvement; (b) NSE improvement) among different ensemble methods. (The simple average method, black box ensemble method, and induction of binomial coefficient ensemble method are referred to as average ensemble, BP ensemble, and binomial ensemble, the same below.)

next moment. After the ensemble, the relative errors of the average water volume in the Daixi, Hengtang, and Qiaodong catchments were -15.4 , 2.73 , and 1.46% , while the NSE values were 0.903 , 0.923 , and 0.865 , respectively. The improvement of this ensemble method is calculated and analyzed, as shown in Figure 6.

For the relative error of water volume, the simulation result of the XAJ model by the method of induced binomial coefficient has not shown much improvement, and the error was increased in a few years. This may be attributed to the relatively good simulation accuracy of the XAJ model; there is little room for improvement. The errors of the Simhyd model, BP model, and GR4J model significantly decreased, and the errors of the four models decreased by 3.7 , 5.3 , 9.4 , and 5.3% , respectively.

The ensemble method improved the NSE values of the four ensemble members in all study areas to different degrees, and the four models (XAJ, Simhyd, BP, GR4J) were respectively improved by 0.116 , 0.358 , 0.353 , and 0.061 . It can be seen that this ensemble method has

improved the Simhyd and BP models the most, while XAJ model and GR4J model were only slightly improved, similar to the case of the black box ensemble.

3.3 Comparison of different ensemble methods

3.3.1 Comparison among three ensemble methods

According to the ensemble results of each catchment, the binomial ensemble method showed the best performance among these three methods overall, as shown in Figure 7. The three ensemble methods showed similar performance in decreasing water volume error. The improvement of water volume error by average ensemble, BP ensemble, and binomial ensemble was 1.52 , -0.69 , and 0.77% , respectively. The greatest improvement was in the BP model, and the improvement of the other three models

Table 4: Comparison ER and NSE between three-model ensemble and four-model ensemble

	Average ensemble		BP-ensemble		Binomial ensemble	
	FOUR	THREE	FOUR	THREE	FOUR	THREE
ER						
Daixi	-15.02%	-12.66%	-11.94%	-1.65%	-15.38%	-11.95%
Hengtang	2.34%	4.27%	-8.97%	-2.75%	2.73%	4.97%
Qiaodong	-0.17%	2.07%	-8.23%	-6.50%	1.46%	2.16%
NSE						
Daixi	0.802	0.813	0.852	0.872	0.903	0.891
Hengtang	0.814	0.804	0.831	0.822	0.923	0.884
Qiaodong	0.759	0.759	0.825	0.811	0.865	0.835

Note: FOUR means four-model ensemble; THREE means three-model ensemble in the table.

was similar. In addition, the three methods improved the NSE values significantly, by 0.113, 0.157, and 0.218, respectively. But while it improved the performance of the Simhyd and BP models, it did not make a positive contribution to the GR4J model performance.

When applying the simple average ensemble method to forecasting, the first step is runoff prediction. This involves driving hydrological model using a series of numerical weather forecasting products and then calculating the forecasting runoff to obtain the ensemble results. Another approach is to embody prediction in the early simulation of ensemble members during the calibration and validation period. This paper applied the latter. The forecasting of the black box ensemble method is shown by using historical data to calibrate parameters or weights and predicting in the validation period. As for the binomial ensemble, the prediction of the method is shown in the unit period. The accuracy of the earlier period is used to modify the results of the next moment (time step), where all models are using daily time step.

Comparisons of these three ensemble methods are discussed; the simple average method takes the averages of the simulation results in each model directly and assumes that the performance of each model is the same, thus giving the same weight. This method is obviously not sufficiently considered and worthy of further improvement. The black box ensemble forecasting method (BP ensemble), which selects a set of parameters as the weight matrix to simulate the flow in the validation period, does not keep changing after the weight is determined, and it is called the static-weight or fixed-weight method. The static method cannot make full use of the advantages of the ensemble members and mainly relies on the data in the training stage.

The data used in this paper are only for 5 years and may not fully represent the full hydrological characteristics of the catchment. Therefore, the improvement of this method may be small and have certain uncertainty. The binomial ensemble method based on induced ordering involves continuously adjusting the ranking according to the prediction accuracy at each moment. Therefore, for each ensemble member, its weight is constantly changing, which is called dynamic-weight method. This dynamic-weight method can reflect the advantages and disadvantages of each model in different periods and can combine the advantages of each model, so as to achieve a real sense of ensemble.

3.3.2 Comparison of different ensemble numbers

In order to understand the influence of the number of ensemble members on the ensemble results, this paper separately calculates and discusses three-model ensemble and four-model ensembles when different methods were used. The BP model is not strictly a hydrological model without specific physical mechanism, so it is discussed separately. In other words, the three-model ensemble refers to the XAJ, Simhyd, and GR4J models, while four-model ensemble refers to the XAJ, Simhyd, GR4J, and BP models. Then the water volume relative error (ER) and NSE values are calculated (see Table 4). Observed and simulated ensemble hydrographs during the validation period are also shown in Figure 8.

For the simple average ensemble method, the three-model ensemble decreased the relative error of water volume in the Daixi catchment by 2.36% and increased the NSE by 0.011, compared with the four-model ensemble. At the same time, the error of Hengtang and Qiaodong catchments increased slightly (1.93 and 1.9%), while the NSE decreased slightly or remained unchanged (0.1, 0). For the BP ensemble method, the accuracy of the three-model ensemble improved significantly compared with that of the four-model ensemble, especially in terms of water volume errors, which are respectively decreased by 10.29, 6.22, and 1.73% in the three catchments. For the binomial ensemble, the result was similar to the simple average ensemble, improving the accuracy of the Daixi catchment, while increasing slightly the error of the Hengtang and Qiaodong catchments.

The results of the three different catchments were then compared. The three-model ensemble significantly decreased the relative water errors (3.43, 10.29, and 2.36%) in the Daixi catchment, compared with the four-model ensemble. Because the water volume error of the Daixi catchment is relatively large in the BP model simulation, removing this ensemble member effectively improves the accuracy. The errors of the Hengtang and Qiaodong catchments increased slightly.

Comparing the three-model ensemble and four-model ensemble under different ensemble methods, the three-model ensemble based on the BP ensemble method is significantly better than the four-model ensemble. At the same time, the three-model ensemble based on the simple average method and the binomial ensemble method is basically similar to the four-model ensemble results, with no obvious increase or decrease.

Figure 8: Observed (Q_{ob}) and simulated ensemble hydrographs in validation period. (a) The observed and ensemble runoff process of Qiaodong catchment using the average method; (b) the observed and ensemble runoff process of the Hengtang catchment using the BP ensemble method; (c) the observed and ensemble runoff process of the Daixi catchment using the binomial ensemble method.

4 Discussion

The induction of binomial coefficient ensemble method showed the best performance among these three ensemble methods. In fact, the simple average method ensembles runoff with fixed weights, which has the same weights of each model. Black box method is a static-weight method with different weights. Once the weights are determined, they don't change any more. Binomial coefficient ensemble method based on the induced ordering is a dynamic-weight method. The weights will change according to the change of accuracy. Of these three methods, the first one is the

simplest but least considered. The result of the second one mainly depends on the representativeness of the previous data and is restricted by the data. The third method can give full play to the advantages of individual members at different moments and combine them to improve the accuracy of forecasting. So, that is the reason why the third method performs best.

The difference between three-model and four-model ensemble is the reduction of BP model. In reality, the BP neural network is a black box model and each parameter matrix has no clear physical meaning. It is only a method for numerical simulation. The model structure

and parameters are uncertain. If the BP model is used to calculate runoff and then ensemble with the same method, a large error may be caused and the accuracy may be improved after removing this ensemble member. In conclusion, after removing the BP model from the ensemble members, the BP three-model ensemble method has the lowest error.

In addition, three basins simulated, including Hengtangcun (1,373 km²), Daixi (233 km²), and Qiaodongcun (162 km²) catchments. The size of Daixi catchment is close to similar to Qiaodongcun catchment, while the size of Hengtangcun catchment is larger (almost 6 times larger). The results of different models presented in Table 2 and the simulation results of different models are relatively more different in Daixi and Qiaodongcun catchment than those recorded in Hengtangcun catchment. The water volume error (ER), the difference between the maximum and the minimum of ER, is 20.74, 11.32, and 8.94% on average in Daixi, Hengtangcun, and Qiaodongcun catchment, respectively. While the difference between the maximum and the minimum of NSE is 0.34, 0.29, and 0.27, respectively. Although the catchment area is small and close to similar in Daixi and Qiaodongcun catchment, the differences in results are relatively large. This difference can be attributed to the level of uncertainty—the smaller the catchment scale, the greater the uncertainty. This study presented these preliminary results in relation to catchment scale and its uncertainty; without doubt, this needs further research in the future in relation to evaluation algorithm or methods to analyze uncertainty quantitatively which is out of the scope of this study and recommended for the following study. Uncertainty can be classified in terms of source, types, and nature; however, the main conclusion in this study is that uncertainty should be considered from the beginning of the study and not at the end of calibration and validation, particularly for integrated water resources arrangement purpose [45] which could result from external or internal data driving the model, the model structure, the technicalities of software or bugs, and model parameter values including statistical uncertainty.

5 Conclusions

Four hydrological models (XAJ, Simhyd, BP neural network, and GR4J models) were applied and tested to simulate the daily runoff in three catchments located in the western part of Zhejiang Province, China. The results of each model were obtained and used in ensemble using three methods (the simple average ensemble method, the

black box ensemble method, and the induction of binomial coefficient ensemble method). The accuracy of the different ensemble methods and different numbers of ensemble members were calculated and compared.

The XAJ and GR4J models proved to be better applicable in the study catchments, while the Simhyd and BP models showed lower applicability. This can be attributed to the fact that the Simhyd model is more suitable for semiarid and semi-humid areas and the BP model does not have specific physical mechanism [46]; therefore, the water volume error of simulation results is relatively large.

The induction of binomial coefficient ensemble method showed the best performance among the three ensemble methods. The improvements of the three kinds of ensemble methods were quite similar for water volume error, but the binomial ensemble method has a great improvement in NSE value. In contrast, both of the first two methods have shortcomings in consideration or application. The third method (binomial coefficient ensemble) can give full play to the advantages of individual members at different moments and combine them to improve the accuracy of forecasting.

Through the comparison of the three-model and four-model ensemble, it can be concluded that the three-model ensemble, which was based on the BP method, had the lowest water volume error, while the four-model ensemble, which was based on binomial method, had the highest NSE value. In comparison with the four-model ensemble, the three-model ensemble clearly decreased the water volume error in the Daixi catchment. For the other two catchments, the accuracy of the three-model ensemble was slightly lower than that of the four-model ensemble or similar. If the BP model is used to calculate runoff and then ensemble with the same method, a large error may occur and cause declining of the accuracy. In conclusion, the results of ensemble mainly depend on the selection of ensemble members, number, and method, which still needs more thorough and detailed research to explore.

Acknowledgments: This study has been financially supported by the National Natural Science Foundation of China (Grants: 41830863, 51879162, 52079026, 41601025), the National Key Research and Development Programs of China (Grants: 2017YFA0605002, 2016YFA0601501, 2017YFC0404401, 2017YFC0404602), and The Belt and Road Fund on Water and Sustainability of the State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (Grant No. 2019nkzd02). Thanks also to the anonymous reviewers and editors.

Author contributions: Ms Jie Wang contributed results analysis and drafting the manuscript, Prof. Guoqing Wang and Prof. Zhenxin Bao structured the manuscript and contributed results discussion, Dr Amgad Elmahdi contributed methodology of the work, Prof. Qinli Yang contributed analysis on results reasonability, Mr Zhangkang Shu and Ms Mingming Song contributed data collection.

References

- [1] Chen SK, Chen RS, Yang TY. Application of a tank model to assess the flood-control function of a terraced paddy field. *Hydrol Sci J*. 2014;59(5):1020–31. doi: 10.1080/02626667.2013.822642.
- [2] Liu D, Guo S, Shao Q, Liu P, Xiong L, Wang L, et al. Assessing the effects of adaptation measures on optimal water resources allocation under varied water availability conditions. *J Hydrol*. 2018;556:759–74. doi: 10.1016/j.jhydrol.2017.12.002.
- [3] Daniela A, Federico G, Andrea C, Paolo B. Advancing reservoir operation description in physically based hydrological models. *Geophys Res Abstr*. 2016;18:10097.
- [4] Amin MGM, Veith TL, Collick AS, Karsten HD, Buda AR. Simulating hydrological and nonpoint source pollution processes in a karst watershed: A variable source area hydrology model evaluation. *Agric Water Manag*. 2017;180:212–23. doi: 10.1016/j.agwat.2016.07.011.
- [5] Wu X, Liu C. Some progresses of hydrological model. *Prog Geogr*. 2002;21(4):341–8.
- [6] Bao W. Hydrological forecasting. Beijing: China Water Power Press; 2006.
- [7] Yan X, Bao Z, Zhang J, Wang G, He R, Liu C. Quantifying contributions of climate change and local human activities to runoff decline in the upper reaches of the Luanhe River basin. *J Hydro-Environ Res*. 2020;28:67–74.
- [8] Bai P, Liu X, Xie J. Simulating runoff under changing climatic conditions: a comparison of the long short-term memory network with two conceptual hydrologic models. *J Hydrol*. 2020;592. doi: 10.1016/j.jhydrol.2020.125779.
- [9] Vrugt JA, Gupta HV, Nuallain B, Bouten W. Real-time data assimilation for operational ensemble streamflow forecasting. *J Hydrometeorol*. 2006;7(3):548–65.
- [10] Duan Q, Ajami NK, Gao X, Soroosh S. Multi model ensemble hydrologic prediction using Bayesian model averaging. *Adv Water Resour*. 2007;30(5):1371–86. doi: 10.1016/j.advwatres.2006.11.014.
- [11] Chen X, Yu JS. Multi-model ensemble hydrological simulation of Yalong River Basin based on artificial neural network. *South–North Water Transf Water Sci Technol*. 2018;16(2):74–80.
- [12] Liu S, Zhang L, She D, Wang Q, Hu C, Xia J. Applicability of catchment hydrologic models in arid and semi-arid regions. *Eng J Wuhan Univ*. 2019;52(5):384–90.
- [13] Wang J, Song X, Zhang J, Wang G, Liu J. Flood simulation of the small and medium-sized river catchment by using multiple hydrological models. *China Rural Water Hydropower*. 2019;7:72–6.
- [14] Gao C, Liu Q, Su B, Zhai J, Hu C. The applicability assessment of hydrological models with different resolution and database in the Huaihe River Basin. *China J Nat Resour*. 2013;28(10):1765–77.
- [15] Gao C, Yao MT, Wang YJ, Zhai S, Buda TF, Zeng XF, et al. Hydrological model comparison and assessment: criteria from catchment scales and temporal resolution. *Hydrol Sci J*. 2016;61(9–12):1941–51. doi: 10.1080/02626667.2015.1057141.
- [16] Seann R, Victor K, Michael S, Zhang Z, Fekadu M, Seo D, et al. Overall distributed model intercomparison project results. *J Hydrol*. 2004;298(1–4):27–60. doi: 10.1016/j.jhydrol.2004.03.031.
- [17] Hassan B, Lahcen B. Comparison of two hydrological models (lumped and distributed) over a pilot area of the Issen watershed in the Souss Basin, Morocco. *Eur Sci J*. 2016;12(18):1857–7881. doi: 10.19044/esj.2016.v12n18p347.
- [18] Epstein ES. Stochastic dynamic prediction. *Tellus*. 1969;21(6):739–59.
- [19] Leith CE. Theoretical skill of Monte Carlo forecast. *Mon Weather Rev*. 1974;102:409–18.
- [20] Hongjun B. Ensemble flood forecasting based on ensemble precipitation forecasts and distributed hydrological model. The 32nd conference on hydrology. The 98th AMS annual meeting. 2018.
- [21] Bao HJ, Zhao LN, He Y, Wetterhall F, Cloke HL, Pappenberger F, et al. Coupling ensemble weather predictions based on TIGGE database with Grid-Xinjiang model for flood forecast. *Adv Geosci*. 2011;29:61–7.
- [22] Bao HJ, Wang LL, Zhang K, Li Z. Application of a developed distributed hydrological model based on the mixed runoff generation model and 2D kinematic wave flow routing model for better flood forecasting. *Atmos Sci Lett*. 2017;18(7):284–93.
- [23] Wang H, Li Y, Ren L, Wang J, Yan D, Lu F. Uncertainty of hydrologic model and general framework of ensemble simulation. *Water Resour Hydropower Eng*. 2015;6:21–6.
- [24] Jasper K, Gurtz J, Lang H. Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and forecasts with a distributed hydrological model. *J Hydrol*. 2002;267(1–2):40–52. doi: 10.1016/S0022-1694(02)00138-5.
- [25] Davolio S. A meteo-hydrological prediction system based on a multimodel approach for precipitation forecasting. *Nat Hazards Earth Syst Sci*. 2008;8(1):143–59.
- [26] Cloke HL, Pappenberger F. Ensemble flood forecasting: a review. *J Hydrol*. 2009;375:613–26. doi: 10.1016/j.jhydrol.2009.06.005.
- [27] Satish KR, Dong JS, Bill L, James DB, Julie D. Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts – a hydrologic model output statistics (HMOS) approach. *J Hydrol*. 2013;497:80–96. doi: 10.1016/j.jhydrol.2013.05.028.
- [28] Arsenault R, Gatien P, Renaud B, Brissette F, Martel J. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *J Hydrol*. 2015;529:754–67. doi: 10.1016/j.jhydrol.2015.09.001.
- [29] Huo W, Li Z, Li Q. Hydrological models comparison and ensemble forecasting in semi-humid watersheds. *J Lake Sci*. 2017;29(6):1491–501.

- [30] Huo W, Li Z, Wang J, Yao C, Zhang K, Huang Y. Multiple hydrological models comparison and an improved Bayesian model averaging approach for ensemble prediction over semi-humid regions. *Stoch Environ Res risk Assess.* 2019;33(1):217–38.
- [31] Arsenault R, Gilles RC, Francois PB. Improving hydrological model simulations with combined multi-input and multimodel averaging frameworks. *J Hydrol Eng.* 2017;22(4):66–76.
- [32] Lorenzo A, Florian P, Fredrik W, Thomas H, David R, Peter S. Evaluation of ensemble streamflow predictions in Europe. *J Hydrol.* 2014;517(2):913–22. doi: 10.1016/j.jhydrol.2014.06.035.
- [33] He S, Guo S, Liu Z, Yin J, Chen K, Wu X. Uncertainty analysis of hydrological multi-model ensembles based on cbp-bma method. *Nordic Hydrol.* 2018;49(5–6):1636–51.
- [34] Lin H, Liu M. Revision of hydrological design for the Tai Lake basin. *J China Hydrol.* 2019;39(4):84–9 (in Chinese with an English abstract).
- [35] Zhao R. Hydrological simulation – Xin’anjiang model and Shanbei model. Beijing: Water Power Press; 1984.
- [36] Liu P, Hao Z, Wang G, Zhao S, Wang Y. Application of Xin’anjiang model and the improved BP neural network model in hydrological forecasting of Min river. *J Water Resour Water Eng.* 2017;28(1):40–4 (in Chinese with an English abstract).
- [37] Gu Y, Wang G, Hao Z, Bao Z, Chen Y. Hydrological modeling of Qujiang River Catchment using Xin’anjiang model. *J Water Resour Water Eng.* 2018;29(2):50–5 (in Chinese with an English abstract).
- [38] Chiew F, Peel MC, Western AW. Application and testing of the simple rainfall-runoff model SIMHYD. In: Singh VP and Frevert DK, editors. *Mathematical models of small watershed hydrology and applications.* Littleton, Colorado: Water Resources Publications; 2002. p. 335–67.
- [39] Wang G, Zhang J, He R. Impacts of environmental change on runoff in Fenhe river basin of the middle Yellow River. *Adv Water Sci.* 2006;6:853–8 (in Chinese with an English abstract).
- [40] Michel C. Un modèle pluie-débit journalier à trois paramètres. *La Houille Blanche.* 1989;2:113–22.
- [41] Deng P, Wang Y, Hu Q, Liu Y. Application of GR4J in daily runoff simulation for Ganjiang River Basin. *Hydrology.* 2014;34(2):60–5 (in Chinese with an English abstract).
- [42] Zhao Z. The foundation and application of fuzzy theory and neural network. Beijing: Tsinghua University Press; 1996.
- [43] Duan QY, Gupta VK, Sorooshian S. Shuffled complex evolution approach for effective and efficient global minimization. *J Optim Theoty Application.* 1993;76(3):501–21.
- [44] Ministry of water resources of the People’s Republic of China specification for geodesic survey in hydrology (SL58-2014); 2014.
- [45] Amgad E. A system approach to improve water productivity and environmental performance at the catchment level. PhD Thesis. Melbourne University, Australia; 2006.
- [46] Chen X, Zhang X, Liu C. Application of SIMHYD model in south and north catchments of China. China society of natural resources. Chinese hydraulic engineering society and the geography society of China theory and practice of human-water harmony; 2006. p. 414–7.