Water Supply

Water Supply Vol 24 No 3, 799 doi: 10.2166/ws.2024.033

Integrated forecasting method of medium-and long-term runoff by ridge regression based on optimal sub-model selection

Binbin Chen [a,*], Zhengdong Chen[a], Chuping Song[a] and Yanhong Song[b]

[a] College of Information Engineering, Nanjing Polytechnic Institute, Nanjing 210048, China
[b] College of Computer and Information, Hohai University, Nanjing 211100, China
*Corresponding author. E-mail: chenbinbin@njpi.edu.cn

BC, 0000-0002-8454-6697

ABSTRACT

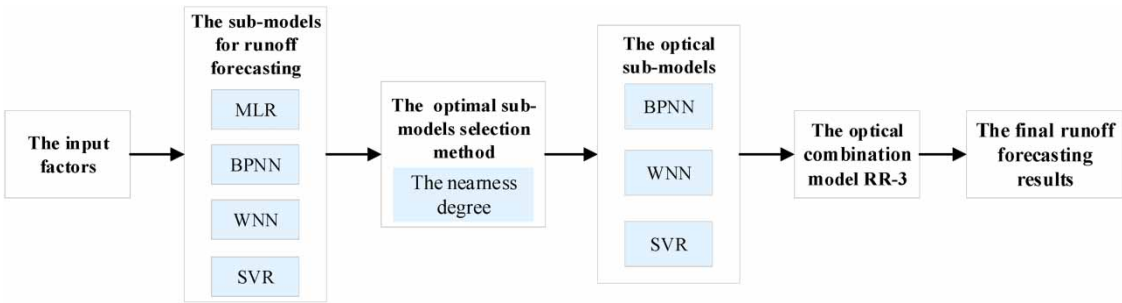
Numerous studies have demonstrated that the combination models can improve the runoff forecast performance compared to individual forecasts. However, some models do not take into account the effects of inappropriate sub-models on the combination models. Based on this, a medium-and long-term runoff integrated forecasting method based on optimal sub-models selection was proposed. First, the sub-models, including linear regression (MLR), BP neural network (BPNN), wavelet neural network (WNN), and support vector regression (SVR), are optimally selected based on the nearness degree. Second, ridge regression (RR) is used to combine the optimal sub-models to predict runoff. Finally, the Guandi hydropower station is taken as an example to verify the effect of the integrated forecasting model. The results show that SVR, BPNN, and WNN are the optimal sub-models, and RR-3 is the optimal integrated forecasting model composed of the optimal sub-models. In addition, compared with the other two combination models, the RR-3 performs better.

Key words: integrated forecast, medium- and long-term runoff, nearness degree, ridge regression, sub-model optimization

HIGHLIGHTS

- The nearness degree was proposed to select the optimal sub-models in the medium- and long-term runoff combination forecasting.
- A combination prediction method using RR to predict the medium- and long-term runoff is established.
- The sub-models can affect the accuracy of runoff combination forecasting.

GRAPHICAL ABSTRACT



1. INTRODUCTION

With the development of the national economy and the adjustment of national water control policy, the gap between the existing hydrologic medium- and long-term runoff forecasting methods and the demand for production and application have been further widened (Ai *et al.* 2022). Therefore, as the most important task in hydrology forecasting, accurate and reliable medium- and long-term runoff forecasting is essential to improve response efficiency of flood disaster preparedness and disaster resistance (Lv *et al.* 2020).

At present, many forecasting models have been proposed and widely applied in medium-and long-term runoff, which include multiple linear regression (MLR) (Maniquiz *et al.* 2010; Ahani *et al.* 2018), autoregressive moving average (ARMA) (Valipour *et al.* 2013), back propagation neural network (BPNN) (Chang & Li 2017), wavelet neural network (WNN) (Shoaib *et al.* 2018), support vector regression (SVR) (Kisi & Parmar 2016; Adnan *et al.* 2022), extreme learning machine (ELM) (Yaseen *et al.* 2019), deep belief networks (DBN) (Yue *et al.* 2023), long–short term memory network (LSTM) (Gao *et al.* 2020; Xiang *et al.* 2020), etc. Although these models have merits on their own, due to the intrinsic weaknesses of all models, and the uncertainty and complexity of runoff change, one forecasting model cannot improve runoff accuracy fundamentally.

However, to combine information and disperse errors from different models, combination forecasting model was proposed, and many studies have proposed that it can improve the forecast performance by a combination of multiple models. For instance, Xu *et al.* (2013) applied the adaptive federated filtering algorithm to combine the forecasting models, and demonstrated that multi-model information fusion can enhance the stability and accuracy of prediction. Chu *et al.* (2017) developed a Bayesian model averaging (BMA)-based multi-model, and performed better than those of the other models. Ai *et al.* (2022) proposed a combination prediction method using ELM to predict the medium- and long-term runoff for better prediction performance and stronger robustness. While they proved that forecast combination improves accuracy, there is limited research on the sub-models selection in the medium-and long-term runoff combination forecasting. Due to the introduction of inappropriate sub-models in the combined model, the prediction accuracy will be reduced. Thus, how to select the optimal sub-models from the available individual models became important in combination forecast.

Recently, the methods adopted for sub-models selection were mainly mutual information (MI) (Cang & Yu 2014), max-linear-relevance and min-linear-redundancy (Che 2015), the nearness degree (Su *et al.* 2019) and neighborhood MI with a maximum relevance and minimum redundancy algorithm (Xiao *et al.* 2019). Among them, the nearness degree has the advantages of simplicity and quickness, and has been widely used in the selection of individual and combination approaches. Therefore, in this study, we propose the nearness degree to select the optimal sub-models for the medium- and long-term runoff forecasting. In addition, the ridge regression (RR) is used as the combination method to combine optimal sub-models, because it has been verified to be feasible in terms of weather (Feng & Wu 1985), ozone concentration (Ji & Cheng 2018), infrared spectrum (Ding 2019), etc.

Aiming to solve the above problems, this paper proposes a integrated forecasting method of medium- and long-term runoff by RR based on optimal sub-models selection. For this reason, the nearness degree firstly is used to select the optimal sub-models. Second, based on the optimal sub-models, a combination prediction method using RR to predict the medium-and long-term runoff is established. Third, the method is applied to the Guandi hydropower station.

2. METHODS AND DATA

In this section, the methods and data are described in detail, including the data source, the data normalization, the nearness degree of optimal sub-models selection, the RR integrated forecasting, and the framework of the proposed model.

2.1 . Data source and data normalization

2.1.1. Data source

The Yalong River basin is located at 96°52′–102°480 E, 26°32′–33°58′ N, with an area of about 136,000 km^2 . The river is 1,571 km long with a drop of 3,830 m and is one of the rivers with the most abundant water energy resources in China (Yue *et al.* 2020). The Guandi hydropower station, located in the lower reaches of the Yalong River, is a large hydropower hub dominated by power generation, with a total storage capacity of 760 million m^3 and a normal water level of 1,330.00 m.

The collected datasets include: (1) monthly inflow runoff data of the Guandi station from January 1960 to December 2011 were provided by the Hydrographic Bureau of the Yangtze River Water Conservancy Commission; (2) 130 atmospheric circulation index data from the national climate center (https://www.ncc-cma.net/Website/index.php), the length of the sequence for January 1951–May 2020. Considering the time consistency and missing items of the series data, this paper finally selects 96 meteorological factors as alternative factors, and the research period is January 1962–December 2011. Taking the monthly inflow runoff of the Guandi hydropower station as an example, January 1962–December 2001 is selected as the model training period and January 2002–December 2011 as the model inspection period. In relation to this, Figure 1 shows the dataset structure of the Guandi station. In addition, this paper makes a statistical analysis of the runoff dataset (see in Table 1), which summarizes the statistical characteristics of the division of runoff series in different time periods.

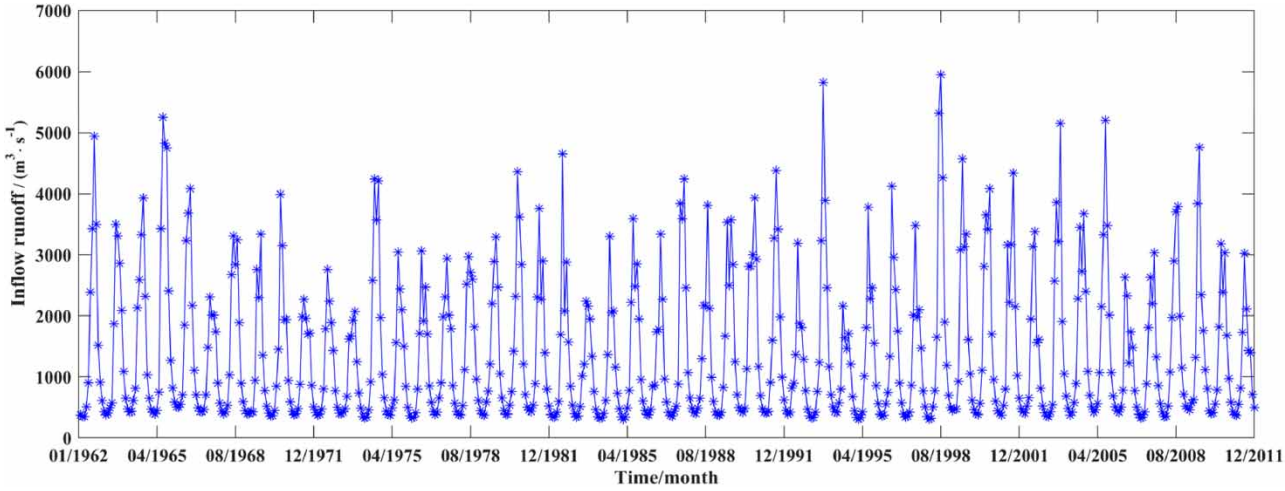


Figure 1 | Dataset structure of the Guandi station.

Table 1 | Statistical indicators of runoff data

			Statistical indicators($m^3 \cdot s^{-1}$)				
Period of records	**Samples**	**Numbers**	**Max.**	**Min.**	**Mean**	**Std.**	**Median**
1962.01–2011.12	All samples	600	5,950.00	303.00	1,417.02	1,191.10	864.50
1962.01–2001.12	Training	480	5,950.00	303.00	1,414.12	1,193.73	864.50
2002.01–2011.12	Testing	120	5,200.00	329.00	1,428.65	1,185.42	872.00

2.1.2. Data normalization

To eliminate the influence of data dimensionality and enhance the predictive performance of the proposed model, data were normalized into a standardized range by the following equation (Gao *et al.* 2020):

$$x_{\text{norm}} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where x_i is the observed values of input factors or runoff values; $x_{\min}$ represents the minimum values of input factors or runoff; $x_{\max}$ is the maximum values of input factors or runoff.

2.2. The nearness degree of optimal sub-models selection

Sub-models F mean that the number of single models participating in the forecast is at least 2, that is, $F = \{f_1, f_2, \cdots, f_p\}(p \geq 2)$. If an inappropriate single forecast model is introduced when combining them, the forecast accuracy may be reduced (Ling & Zhang 2019). Therefore, in order to improve the prediction accuracy, it is necessary to select the optimal sub-models before combining them. In this paper, the optimal prediction accuracy vector will be constructed by the relative error, and the sub-models will be selected by the close degree between the vectors. The best prediction accuracy vector is defined in the following (Su *et al.* 2019).

Suppose $e_i(t)$ is the relative error of model i at t, and its expression is

$$e_i(t) = \left| \frac{(f_i(t) - \hat{f}_i(t))}{f_i(t)} \right| (i = 1, 2, \cdots, p) \tag{2}$$

where $f_i(t)$ is the observed value of model i at t; $\hat{f}_i(t)$ is the simulated value of model i at t; $e_i(t) \in [0, 1]$ is the relative error vector of the forecast model i.

It is assumed that $e(t) = \min\{e_1(t), e_2(t), \mathrm{L}, e_p(t)\}$ is the minimum relative error of p prediction models at time t, and $e = \{e(1), e(2), \mathrm{L}, e(n)\}$ is called the best prediction accuracy vector. The smaller the value is, the higher the prediction accuracy of the model is. If $e_i(t)$ and e are very close, it indicates that the prediction accuracy of the forecast model is high. Therefore, the nearness between $e_i(t)$ and e can be used to determine the prediction accuracy of the forecast model. The concept of proximity is as follows:

Let the relative error vector of the ith prediction model be $e_i = \{e_i(1), e_i(2), \cdots, e_i(n)\}$, then the approximation degree between it and the best prediction accuracy vector $e = \{e(1), e(2), \cdots, e(n)\}$ is

$$\sigma(e_i, e) = 1 - \frac{\sum_{t=1}^{n} |e(t) - e_i(t)|}{\sum_{t=1}^{n} (e(t) + e_i(t))} \tag{3}$$

where $\sigma(e_i, e)(i = 1, 2, \cdots, p)$ is sorted from largest to smallest, and the higher the order, the higher the prediction accuracy of the forecast model (Ling & Zhang 2019). The sub-models optimization method is also suitable for RR integrated forecasting.

2.3. RR integrated forecasting

Based on the optimal sub-models selection, RR is adopted for integrated forecasting. The principle is as follows (Feng & Wu 1985):

RR integrated forecasting is proposed on the basis of general regression integrated(GRI) forecasting, and its expression is

$$Y = F\beta + \varepsilon \tag{4}$$

where Y is the observed value; F is the data matrix formed by the selected model $f_1, f_2, \ldots, f_p$, and p is the number of prediction models; β is the regression coefficient vector; ε is the codifference vector.

If there is a similarity in the $f_1, f_2, \ldots, f_p$ models, it will lead to the prediction results of each model with a great degree of correlation, which will easily make the $\mathrm{F^TF}$ close to singularity, resulting in large mean square error and instability. Therefore, RR is proposed as an integrated forecast. In this method, the F^TF in $\hat{\beta} = (F^TF)^{-1}F^TY$ is estimated by replacing $F^TF + \lambda I$ with the least square of β. The RR here is estimated to be

$$\hat{\beta}(\lambda) = (F^TF + \lambda I)^{-1}F^TY \tag{5}$$

where F^T is the transpose matrix of F; $(F^TF)^{-1}$ is the inverse matrix of F^TF; $\lambda \geq 0$ is the ridge parameter, when $\lambda = 0$, it's a least squares estimate; I. is the identity matrix.

As for the choice of λ value, $\lambda = \hat{\sigma}^2/\max v_i^2$ is used this paper. Among

$$v = D^T\beta = (v_1, v_2, \cdots, v_i)^T \tag{6}$$

where D is an orthogonal matrix, obtained from $F^TF = D^T\Lambda D$, Λ is a diagonal matrix, whose principal diagonal element is the eigenvalue of F^TF. ; σ is the variance.

Since σ^2 and β are not reassurable, in order to obtain numerical results, this paper utilizes subsample estimation to represent them. That is, $\hat{\beta}$ represents β, and σ^2 is estimated by the following formula

$$\hat{\sigma}^2 = (Y - F\hat{\beta})^T(Y - F\hat{\beta})/(N - P) \tag{7}$$

where N is the number of samples; P is the number of independent variables.

Based on the principles of RR, an explanatory figure that schematizes the prototype of medium- and long-term runoff is provided, as depicted in Figure 2.

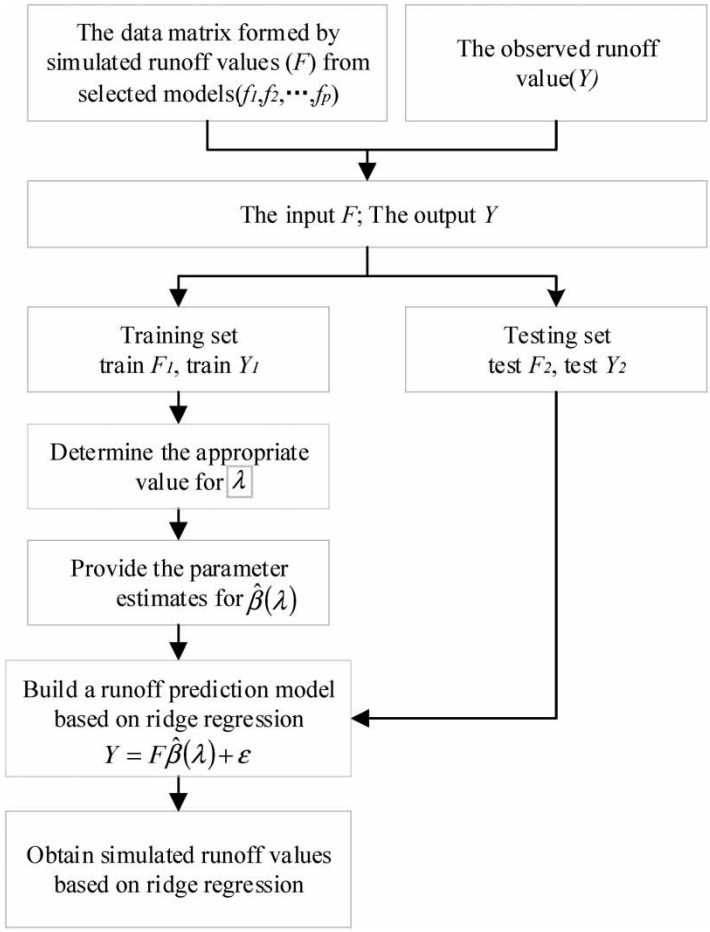


Figure 2 | An explanatory figure that schematizes the prototype of medium- and long-term runoff.

2.4. Flowchart of the proposed model

In this paper, before using RR to establish integrated forecasting, sub-models are sorted by using proximity degree. Then, according to the priority ranking, a single model participates in the RR integrated forecast, and the constructed RR integrated forecast model is optimized, in order to determine the optimal RR integrated forecast model and sub-models. The detailed steps are as follows:

S1. The relative error vector $e_i = \{e_i(1), e_i(2), \cdots, e_i(n)\}(i = 1, 2, \cdots, p)$ is obtained by calculating the $e_i(t)$ of sub-models F;
S2. Determine the best prediction accuracy vector $e = \{e(1), e(2), \cdots, e(n)\}$;
S3. Calculate the approximation degree $\sigma(e_i, e)$;
S4. The $\sigma(e_i, e)$ is sorted from large to small, and if $\sigma(e_1, e) > \sigma(e_2, e) > \cdots > \sigma(e_p, e)$, the corresponding model is sorted as $f_1 > f_2 > L > f_p$. According to the principle that the forecasting effect of the forecasting model ahead of the priority is better than that of randomly selecting the same number of forecasting models (Ling & Zhang 2019), RR is used to do the integrated forecast according to the $f_1 - f_2, f_1 - f_2 - f_3, \cdots, f_1 - f_2 - f_3 - f_4 - \cdots - f_p.$, which is recorded as $RR - 2, RR - 3,$ L, $RR - p$;
S5. By using the method of subsample estimation, the estimated value $\hat{\beta}_j(\lambda)$ of RR is obtained according to the Formulas (4)–(6), and the RR integrated forecast equation is obtained by bringing it in (3);
S6. On the basis of obtaining the $RR - j(j = 2, 3, \cdots, p)$ integrated prediction results, the nearness degree $\sigma(e_{RR-j}, e_{RR})$ is calculated according to step S1–S3, and the model is optimized according to its value. Where e_{RR} is the best prediction accuracy vector of RR model, and e_{RR-j} is the relative error vector of the j th RR model;

S7. If $\sigma(e_{RR-j}, e_{RR}) = \max\{\sigma(e_{RR-2}, e_{RR}), \text{L}, \sigma(e_{RR-p}, e_{RR})\}$, *RR-j* is the optimal integrated forecast model, and $f_1, f_2, \cdots, f_j$ is the optimal single forecast model participating in the integration.

Based on the above knowledge, the flow chart of the method is drawn, as shown in Figure 3.

2.5. Evaluation indicators

Currently, various performance metrics are used in evaluation the forecasting performance of the model. However, there is no single standard to determine which evaluation metrics is the most accurate assessment method. Thus, to evaluate the performance of the proposed model, five common evaluation indexes in medium- and long-term runoff prediction are adopted in this paper, including root mean square error (RMSE), certainty coefficient (DC), qualification rate (QR), mean absolute error (MAE) (Chu *et al.* 2017; Gao *et al.* 2020), and mean absolute percentage error (MAPE) (Yue *et al.* 2020, 2023). The specific calculation formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(f(t) - \hat{f}(t))^2} \tag{8}$$

$$\text{DC} = 1 - \frac{\sum_{t=1}^{n}(f(t) - \hat{f}(t))^2}{\sum_{t=1}^{n}(f(t) - \bar{f}(t))^2} \tag{9}$$

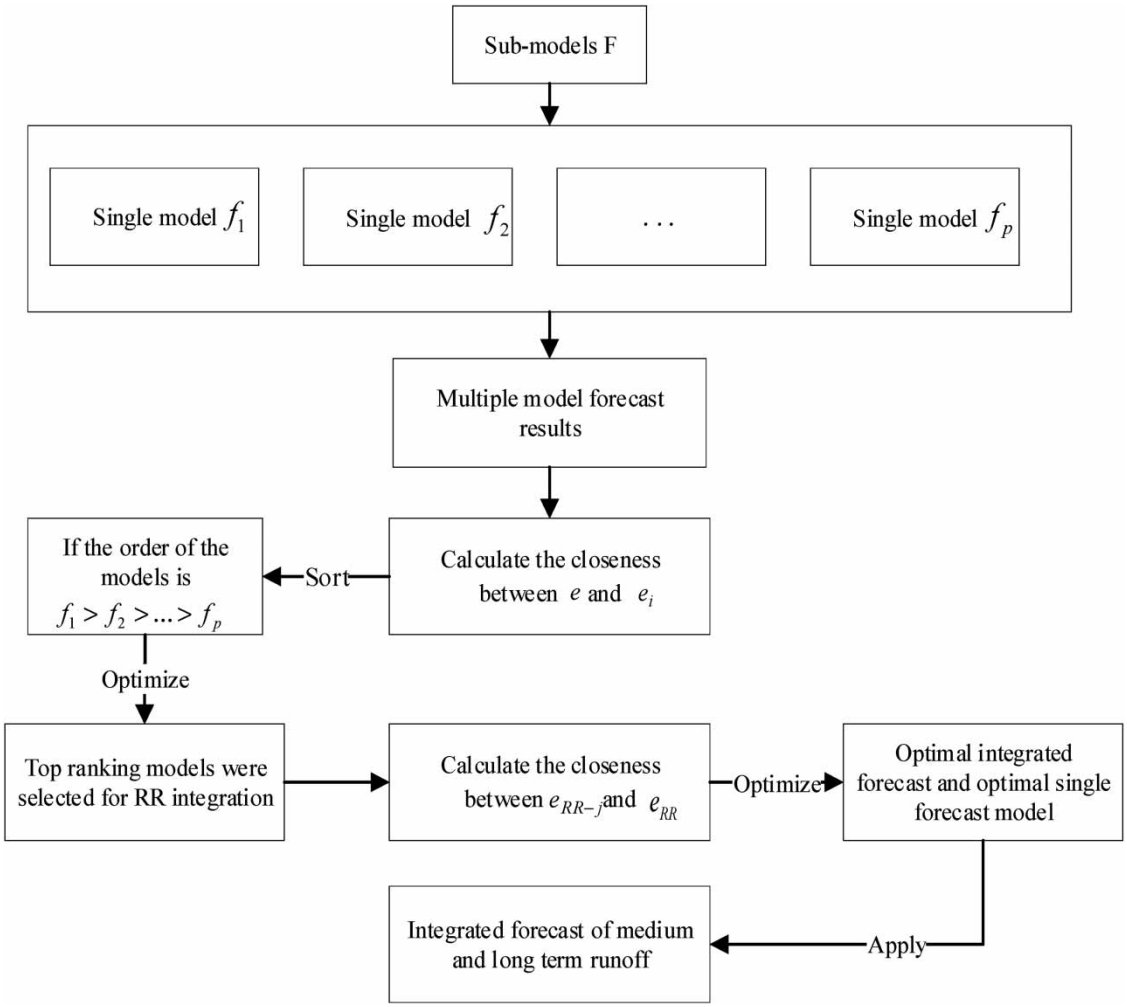


Figure 3 | The flow chart of the method proposed in this paper.

$$\mathrm{QR} = \frac{m}{N} \times 100\% \tag{10}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |f(t) - \hat{f}(t)| \tag{11}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{f(t) - \hat{f}(t)}{f(t)} \right| \tag{12}$$

where $f(t)$ is the observed value at time t, $m^3 \cdot s^{-1}$; $\hat{f}(t)$ is the simulated value at time t, $m^3 \cdot s^{-1}$; $\bar{f}(t)$ is the mean value at time t, $m^3 \cdot s^{-1}$; n the total number of samples; m is the number of qualified simulations (the relative error of simulation $<20\%$ is qualified); N is the total number of simulations.

The RMSE is utilized to assess the accuracy of the simulated values, with a closer value to zero indicating a better match between the simulated and observed values. The closer the DC value is to 1, the more consistent the proposed forecast model is with the actual situation. The higher the QR value, the higher the accuracy of the proposed forecast model. The MAE represents the average absolute deviation between individual observations and the arithmetic mean, providing an accurate reflection of the actual prediction error. The MAPE, being the most frequently used statistical index, is employed to examine the error between the predicted and observed values (Ai *et al.* 2022). Among them, smaller values of RMSE, MAE, and MAPE indicate better forecasting performance, whereas higher values of DC and QR reflect improved overall model performance.

3. RESULTS AND DISCUSSION

3.1. The input factors

Due to the lag effect of climate-related factors on runoff (Cheng *et al.* 2019), this paper takes 2a lag as the term to apply 2,304 (96*24) alternative input variables to 96 climatic factors of the Guandi hydropower station and normalizes the data according to Equation (1). Based on the method of combining correlation coefficient and stepwise regression (Ai *et al.* 2022), reasonable input factors are selected, and the results are shown in Table 2.

3.2. Models structure and parameter selection

Based on the literature review, an individual prediction model cannot perform optimally in any environment for precise and stable mid- to long-term runoff prediction. Therefore, we consider as many sub-models as possible for the adaptive sub-model selection to ensure the prediction accuracy of the integrated forecasting method. In this study, four individual forecast models are used, including MLR, BP, WNN, and SVR; in addition, there are three combination methods, simple average (SA), GRI and RR, are selected for the comparison. To determine the most acceptable model, it is necessary to find the suitable parameters for each of the above algorithms, as follows.

The MLR model adopted in this paper is a linear model, and the regression parameter is (−4,491.3200, −41.0268, 0.9124, 0.0649, 15.9310, −0.8822, 57.7334, −17.7694, 2.3502, 64.2055) by model training. Substitute into Equation (1), and the

Table 2 | The input results of different models

Variable code	The input factor	Lag time/month
x_1	North Africa Subtropical High Area Index (20W-60E)	t-1
x_2	Area Index of North American Atlantic Subtropical High (110W-20W)	t-1
x_3	Polar Vortex Strength Index in the Northern Hemisphere (zone 5,0–360)	t-1
x_4	Precipitation	t-1
x_5	Tibet Plateau (30N–40N,75E–105E)	t-7
x_6	Cold air frequency	t-11
x_7	IOWPA Warm Pool Area Index for the Indian Ocean	t-16
x_8	Tibet Plateau (25N–35N,80E–100E)	t-18
x_9	WPWPA Warm Pool Area Index (WPWPA)	t-23

equation is obtained as follows:

$$f_1 = -4491.3200 - 41.0268x_1 + 0.9124x_2 + 0.0649x_3 + 15.9310x_4 - 0.8822x_5 + 57.7334x_6 - 17.7694x_7 + 2.3502x_8 + 64.2055x_9$$

The BP and WNN models are nonlinear models, and the number of neurons in the input layer, hidden layer and output layer is 9, l and 1, respectively. The number of hidden layer nodes is $l = \sqrt{m+n} + \alpha$, ($\alpha \in [0, 10]$), where n is the number of input layer nodes and m is the number of input layer nodes. It is known from the formula l that the value interval of l is [2,14]. One unit per interval is taken as the number of nodes in the hidden layer, and the RMSE of the model is compared to select the number of nodes in the hidden layer. Figure 4 shows the relationship between the number of hidden layer nodes and RMSE in BP and WNN models, respectively. It is known from Figure 4(a) that when the number of nodes in the hidden layer is 7, the RMSE of the BP neural network model is the smallest, so the structure of the model is 9-7-1. As shown in Figure 4(b), when the number of nodes in the hidden layer of the WNN model is 12, the RMSE is the smallest, so this study selected 9-12-1 as the structure of the model.

For the SVR model, the model is trained with three kernel default parameters, and its DC value is shown in Figure 5(a)–5(c). From Figure 5(a)–5(c), it is known that the best kernel function is the radial basis, whose DC value is the largest, 0.8872. Therefore, the radial basis is chosen as the kernel function of the SVR model. For the parameters gamma and C of SVR model, GridSearch is used (Ai *et al.* 2022). The method is an exhaustive search, in all the candidate parameters, through loop traversal, trying each possibility. In this paper, we traverse gamma in the interval [0.1, 2], with 0.1 as the interval, and C in the interval [0.5, 5]. Through the search, the optimal parameter is gamma = 0.1, C = 4.5, and the DC value of the corresponding radial basis kernel function is 0.9174, as shown in Figure 5(d). It can be seen from Figure 5(a)–5(d) that the prediction effect of SVR model is significantly improved after adjusting parameters.

For the combination forecasting models, the SA combination method can be expressed as $Y_t = \sum_{j=1}^{m} W_j \hat{y}_t^{(j)}$ where $w_j = 1/M$, $\hat{y}_t^{(j)}$ is the forecast value (output) from the jth single forecasting model and Y_t is the combined forecast model at time t, M is the total number of individual forecasting models. Thus, the weight W of SA in this paper is 1/3, 1/3, and 1/3, respectively. According to the Equation (3), the parameters of GRI is $\beta = (0.9040,\ 0.0190,\ 0.0944)$. And, the ridge estimate $\hat{\beta}_3$ of RR is 0.8790, 0.0384, and 0.1007, respectively.

3.3. Sub-models optimization selection

Based on the steps S1–S3 and formula (3), the nearness of sub-models e_i (i = 1,2,3,4) and e in the verification period are 0.5463, 0.6121, 0.5761, and 0.7002, respectively. In terms of the nearness value, the SVR exhibits the highest prediction accuracy, followed by BPNN and WNN, while MLR shows the worst performance. This outcome can be attributed to the numerous factors influencing medium-and long-term hydrological processes, most of which are expressed through complex nonlinear relationships. The SVR model is well-suited for addressing nonlinear problems by providing an effective mapping between input and output data in a higher-dimensional feature space, thereby enhancing forecasting accuracy. Moreover, the SVR algorithm operates based on the structural risk minimization criterion. To minimize the expected risk, it is crucial to simultaneously minimize the empirical risk and the confidence range. This approach involves maintaining a fixed training

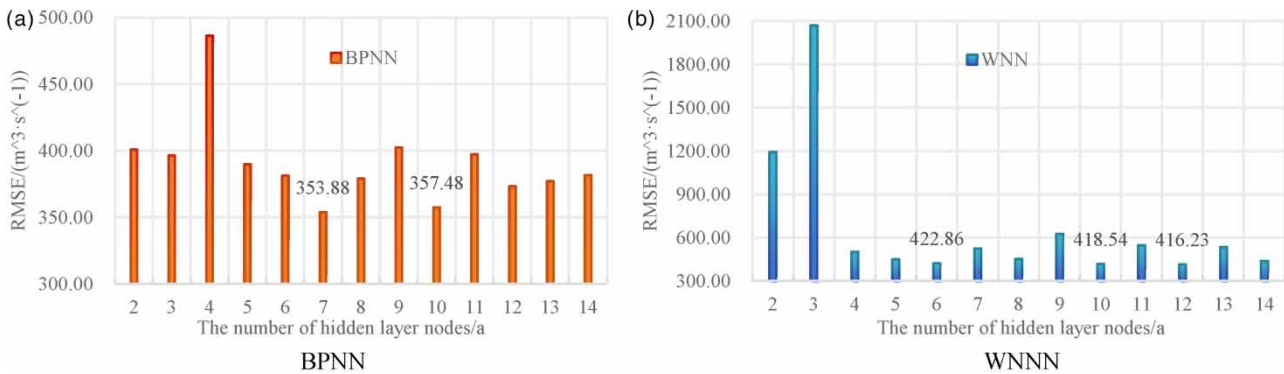


Figure 4 | The relationship between the number of hidden layer nodes and the performance: (a) BP and (b) WNN.

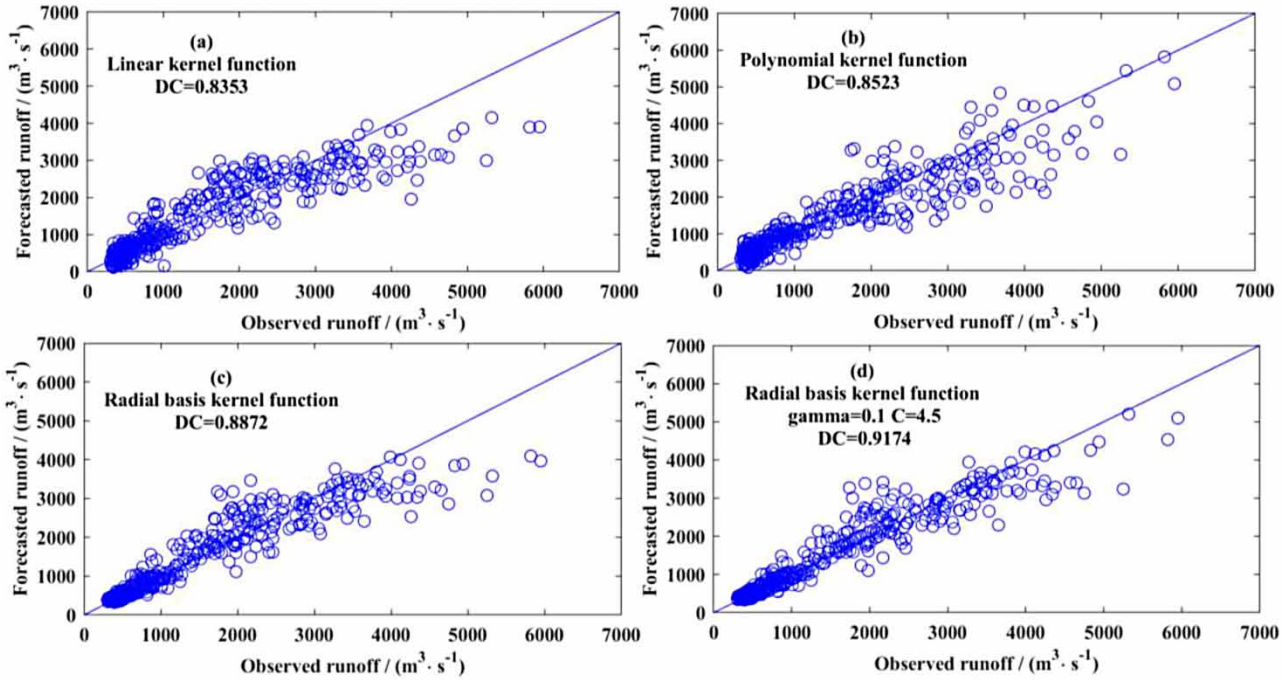


Figure 5 | The performance of SVR kernel function and parameter selection.

error while minimizing the confidence range, effectively addressing over-learning issues and enhancing the model's ability to generalize across samples. Given the linear nature of MLR, it is considered less suitable for capturing these complexities. Therefore, BPNN, WNN, and SVR are considered as the individually optimal nonlinear models.

3.4. Optimization of integrated forecasting model

According to the step S4 and sub-models sequencing, in this paper, RR is used for integrated forecasting according to the methods of SVR-BPNN, SVR-BPNN-WNN, and SVR-BPNN-WNN-MLR, denoted as *RR-j* ($j = 1, 2, 3, 4$). On the basis of the integrated prediction results, according to steps S1–S3 and formula (3), the nearness degrees of e_{RR-j} and e_{RR} of RR-2, RR-3 and RR-4 integrated prediction models are 0.9223, 0.9299 and 0.9077, respectively. It can be seen that RR-3 model has the highest prediction accuracy, followed by RR-2 and RR-4 models. The main reason should be that the prediction accuracy of MLR in the single model involved in integration is the worst, which leads to the unsatisfactory prediction results of the integrated forecast, indicating that it is necessary to select the single model involved in integration when constructing the integrated forecast model (Ling & Zhang 2019).

It is known that SVR, BPNN and WNN are the optimal sub-models, and RR-3 is the optimal integrated forecast model. The specific integration process of the RR-3 model is as follows. Based on obtaining SVR(f_1), BPNN(f_2), WNN(f_3) prediction results, the ridge estimate of the RR-3 integrated prediction model is obtained according to step S5. It can be concluded that SVR model has the best prediction effect, followed by WNN and BPNN. Then, by substituting $\hat{\beta}_3$ into Equation (4), the equation of integrated prediction RR-3 is obtained as $Y = 0.8790 \times f_1 + 0.0384 \times f_2 + 0.1007 \times f_3$.

3.5. Integrated forecast results

The selected optimal sub-models and RR-3 are applied to the integrated forecast of medium- and long-term runoff of the Guandi hydropower station. Table 3 shows the performance comparison between sub-models and the combination models for the training and testing period. From Table 3, the following results can be obtained:

(a) During the training period, compared with the other three single models and two combination forecasting methods, the RR-3 model has the maximum DC and QR, with the values of 0.9184 and 75.54%; the minimum MAE, MAPE, and RMSE values of 201.06, 0.1510, and 340.67 $m^3 \cdot s^{-1}$ respectively. Next, for the other three single models, the SVR model with MAE, MAPE, RMSE, DC and QR values of 209.84 $m^3 \cdot s^{-1}$, 0.1569, 342.77 $m^3 \cdot s^{-1}$, 0.9174, and 74.58%, respectively,

Table 3 | Performance comparison between each model and RR integrated model

Model		Training					Testing				
		MAE/ ($m^3 \cdot s^{-1}$)	MAPE	RMSE/ ($m^3 \cdot s^{-1}$)	DC	QR/%	MAE/ ($m^3 \cdot s^{-1}$)	MAPE	RMSE/ ($m^3 \cdot s^{-1}$)	DC	QR/%
Optimal sub-models	WNN	274.69	0.2138	373.10	0.9021	61.67	333.70	0.2413	531.28	0.7974	49.17
	BPNN	253.60	0.1873	462.47	0.8496	48.54	298.61	0.2214	444.15	0.8584	53.33
	SVR	209.84	0.1569	342.77	0.9174	74.58	285.60	0.1912	495.81	0.8236	65.00
The combination methods	SA	221.62	0.1615	357.81	0.9100	67.27	271.56	0.1824	447.22	0.8565	64.17
	GRI	201.54	0.1512	340.68	0.9184	73.33	277.11	0.1821	480.51	0.8343	65.83
	RR-3	201.06	0.1510	340.67	0.9184	75.54	270.24	0.1820	437.62	0.8661	66.67

performs better than the other two models. In the other two combination forecasting methods, the GRI model has a better fitting effect than the SA model. Based on the five evaluation metrics, the order of all the combination models from good to bad is RR-3, GRI and SA. Figure 6 shows the result of the scatter figures of the RR-3 model and other models.

(b) During the testing period, the RR-3 model, that contains the optimal MAE, MAPE, RMSE, DC, and QR values of 270.24 $m^3 \cdot s^{-1}$, 0.1820, 437.62 $m^3 \cdot s^{-1}$, 0.8661, and 66.67% respectively, can get the best forecasting result. Next, for the other

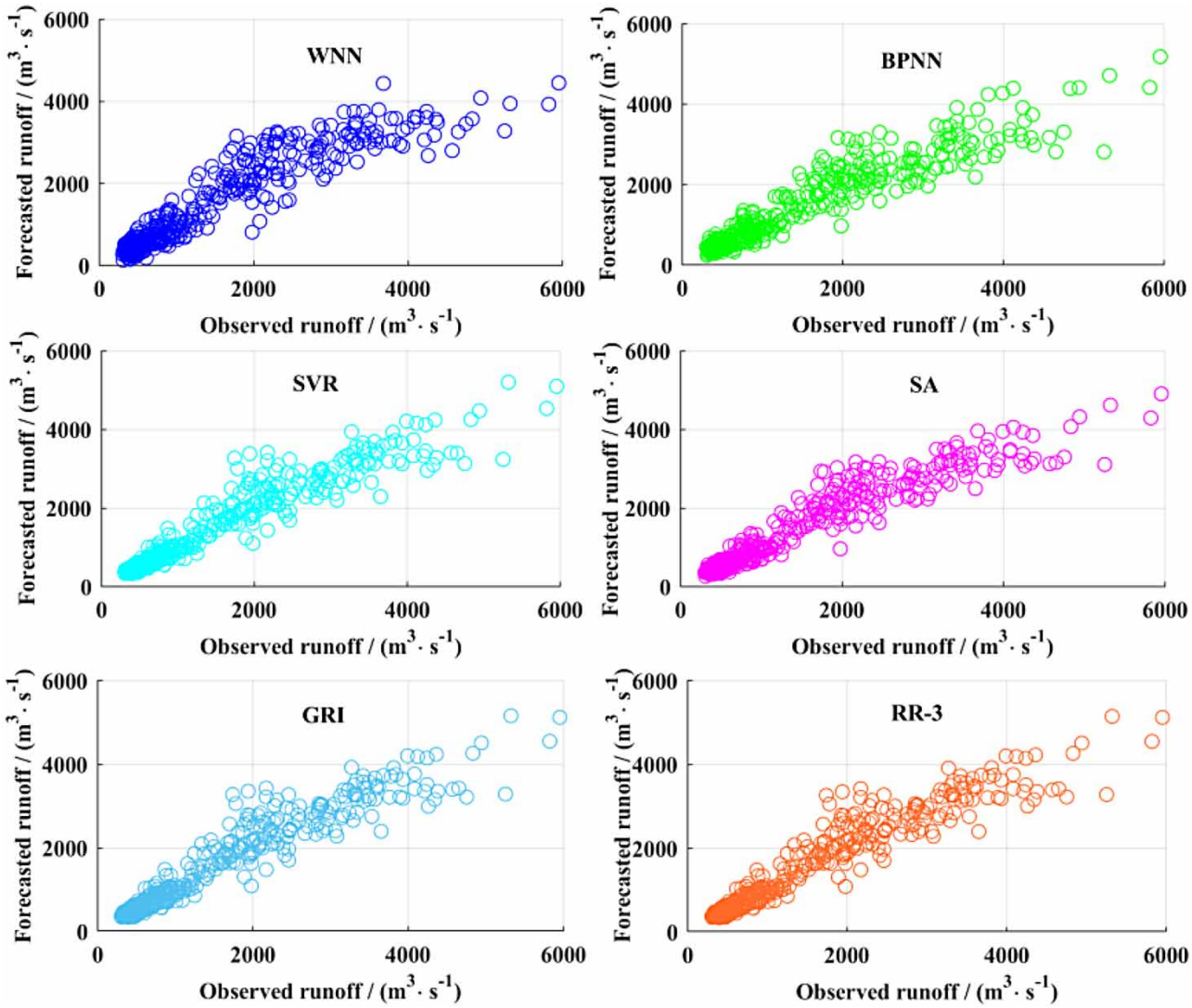


Figure 6 | The scatter plot between observed and forecasted runoff of different models in training.

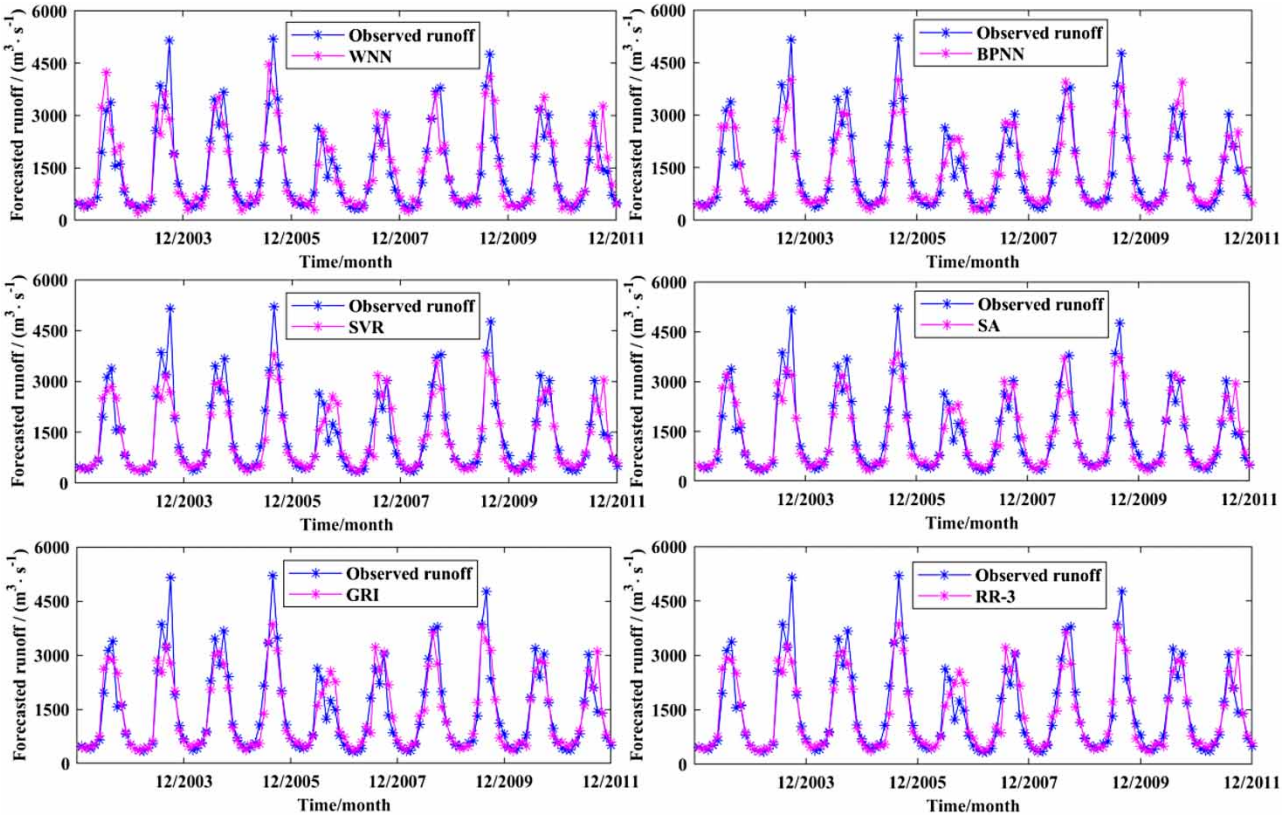


Figure 7 | The prediction results of different models in testing.

three single models, there isn't a unified law in terms of the performance metrics, based on RMSE and DC, BPNN performs better than the other two models; based on MAE, MAPE and QR, SVR can get better forecasting results. However, what is certain is that the worst-performing model among them is WNN. Then, for the other two combination models, there is no standardized legislation regarding performance metrics, the SA model has the minimum MAE and RMSE value of 271.56 and 447.22 $m^3 \cdot s^{-1}$, and the maximum DC and QR, with the values of 0.8565 and 64.17%, indicating that the SA model performs better than the GRI model. Figure 7 shows the forecasting results of the RR-3 model and other models.

Based on the comprehensive comparison and analysis of the forecasting effects of BPNN, WNN, SVR, SA, GRI, and RR-3 models during the training period and verification period, it is found that the RR-3 can get best forecasting performance. This is mainly because the BPNN model has a strong nonlinear mapping ability and can reflect the nonlinear characteristics of runoff. WNN model combines the advantages of wavelet analysis and BPNN model, and has high prediction accuracy, but its qualified rate decreases obviously in the verification period, indicating that WNN model has the phenomenon of over-fitting and the reliability of the model is reduced. After training, BPNN and WNN models establish a network model based on empirical risk minimization, which has some shortcomings such as local minimum and instability. For example, during the model verification period, the evaluation indexes RMSE and DC of BPNN model perform relatively well, but its qualified rate QR is relatively poor. The SVR model utilizes kernel functions and employs structural risk minimization as the guiding principle, ultimately yielding a unique solution that can address some of the shortcomings of the aforementioned models (Liang *et al.* 2020), resulting in a better simulation prediction effect compared to BPNN and WNN.

On the other hand, in the SA combination model, each individual predictive model contributes equally (with the same weight) to the combined value, but is less reliable. The GRI combination model is a kind of unequal weight method which finds the weight by linear fitting. However, if the prediction results of single models are highly correlated, the mean

square error of weights will be very large and unstable. For example, during the testing period, the prediction effect of GRI combination model is worse than that of SA. Conversely, the RR-3 combination model avoids the above problems by adding ridge parameters, obtaining appropriate weights, and demonstrating better prediction performance. To sum up, the RR-3 model integrates the advantages of each single model and improves the accuracy of medium-and long-term runoff forecasting.

4. CONCLUSIONS

To consider the influence of sub-models on combination models, a medium- and long-term runoff integrated forecasting method based on optimal sub-models selection was proposed in this paper. And, it is applied to the medium- and long-term runoff forecast of the Guandi hydropower station in Yalong River Basin. The main findings can be briefly concluded as follows. First, SVR, BPNN and WNN are identified as the optimal sub-models, with RR-3 being the optimal integrated forecasting model composed of the aforementioned three single forecasting models. Second, through the comparative analysis of the runoff forecasting results, it is evident that the RR-3 integrated forecasting model demonstrates a good forecasting effect, which proves that the sub-models can affect the accuracy of runoff combination forecasting.

Based on the accomplished results, there are still some areas that need to be improved. In the future, a nonlinear combination method will be considered to improve the accuracy of medium- and long-term runoff forecasting. Besides, for the state-of-the-art single model in the combination model, we will adopt deep learning methods, such as LSTM and Convolutional Neural Network (CNN). Moreover, the latest data would be updated and supplemented, including human activity, evaporation, temperature, and so on.

ACKNOWLEDGEMENTS

This work was supported by 'Accurate Extraction Research and Product Realization of Water Information About Impervious Surface from High-resolution Images'(Grant No. NJPI-RC-2023-06) and 'Jiangsu Province Vocational Education Teaching Reform Key Project Funding' (Grant No. ZZZ18).

AUTHOR CONTRIBUTIONS

C.B. wrote the original draft, methodology, formal analysis. C.Z. was involved in conceptualization, writing-reviewing, and funding acquisition. S.C. performed methodology and data curation. S.Y. collected resources, and was involved in data curation, writing – review and editing.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Adnan, R. M., Kisi, O., Mostafa, R. R., Ahmed, A. N. & El-Shafie, A. 2022 The potential of a novel support vector machine trained with modified mayfly optimization algorithm for streamflow prediction. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques* **67** (2), 161–174. doi:10.1080/02626667.2021.2012182.

Ahani, A., Shourian, M. & Rad, P. R. 2018 Performance assessment of the linear, nonlinear and nonparametric data driven models in river flow forecasting. *Water Resources Management* **32** (2), 383–399. doi:10.1007/s11269-017-1792-5.

Ai, P., Song, Y., Xiong, C., Chen, B. & Yue, Z. 2022 A novel medium- and long-term runoff combined forecasting model based on different lag periods. *Journal of Hydroinformatics* **24** (2), 367–387. doi:10.2166/hydro.2022.116.

Cang, S. & Yu, H. 2014 A combination selection algorithm on forecasting. *European Journal of Operational Research* **234** (1), 127–139. doi:10.1016/j.ejor.2013.08.045.

Chang, C.-L. & Li, M.-Y. 2017 Predictions of diffuse pollution by the HSPF model and the back-propagation neural network model. *Water Environment Research* **89** (8), 732–738. doi:10.2175/106143017x14902968254665.

Che, J. 2015 Optimal sub-models selection algorithm for combination forecasting model. *Neurocomputing* **151**, 364–375. doi:10.1016/j.neucom.2014.09.028.

Cheng, Q., Zuo, X., Zhong, F., Gao, L. & Xiao, S. 2019 Runoff variation characteristics, association with large-scale circulation and dominant causes in the Heihe River Basin, Northwest China. *Science of The Total Environment* **688**, 361–379. doi:10.1016/j.scitotenv.2019.05.397.

Chu, H., Wei, J., Li, J., Qiao, Z. & Cao, J. 2017 Improved medium- and long-term runoff forecasting using a multimodel approach in the Yellow River headwaters region based on large-scale and local-scale climate information. *Water* **9** (8), 1–16. doi:10.3390/w9080608.

Ding, H. 2019 *Application of Ridge Regression and its Improved Algorithm in Infrared Spectral Data*. Wenzhou University. doi:CNKI:CDMD:2.1018.285309

Feng, Y. & Wu, D. 1985 Application of ridge regression in forecasting integration. *Meteorology* **11**, 2–4. doi:10.7519/j.issn.1000-0526.1985.11.001.

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M. & Lin, Q. 2020 Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. *Journal of Hydrology* **589**, 1–11. doi:10.1016/j.jhydrol.2020.125188.

Ji, M. & Cheng, L. 2018 Application of ridge regression algorithm in ensemble prediction of ozone concentration. *Journal of Anhui University (Natural Science Edition)* **42** (4), 93–102. doi:10.3969/j.issn.1000-2162.2018.04.016.

Kisi, O. & Parmar, K. S. 2016 Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *Journal of Hydrology* **534**, 104–112. doi:10.1016/j.jhydrol.2015.12.014.

Liang, H., Huang, S., Meng, E. & Huang, Q. 2020 Runoff prediction based on multiple hybrid models. *Journal of Hydraulic Engineering* **51** (1), 112–125. doi:10.13243/j.cnki.slxb.20190434.

Ling, L. & Zhang, D. 2019 A review of construction and application of combination forecast model. *Statistics & Decision* **35** (1), 18–23. doi:10.13546/j.cnki.tjyjc.2019.01.004.

Lv, N., Liang, X., Chen, C., Zhou, Y., Li, J., Wei, H. & Wang, H. 2020 A long short-Term memory cyclic model with mutual information for hydrology forecasting: A Case study in the xixian basin. *Advances in Water Resources* **141**, 1–10. doi:10.1016/j.advwatres.2020.103622.

Maniquiz, M. C., Lee, S. & Kim, L.-H. 2010 Multiple linear regression models of urban runoff pollutant load and event mean concentration considering rainfall variables. *Journal of Environmental Sciences* **22** (6), 946–952. doi:10.1016/s1001-0742(09)60203-5.

Shoaib, M., Shamseldin, A. Y., Khan, S., Khan, M. M., Khan, Z. M., Sultan, T. & Melville, B. W. 2018 A comparative study of various hybrid wavelet feedforward neural network models for runoff forecasting. *Water Resources Management* **32** (1), 83–103. doi:10.1007/s11269-017-1796-1.

Su, L., Song, Y. & He, H. 2019 Variable weight combination forecasting method considering weight uncertainty. *Statistics & Decision* **35** (11), 60–63. doi:10.13546/j.cnki.tjyjc.2019.11.014.

Valipour, M., Banihabib, M. E. & Behbahani, S. M. R. 2013 Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology* **476**, 433–441. doi:10.1016/j.jhydrol.2012.11.017.

Xiang, Z., Yan, J. & Demir, I. 2020 A rainfall-runoff model with LSTM-Based sequence-to-sequence learning. *Water Resources Research* **56** (1), 1–17. doi:10.1029/2019wr025326.

Xiao, L., Wang, C., Dong, Y. & Wang, J. 2019 A novel sub-models selection algorithm based on max-relevance and min-redundancy neighborhood mutual information. *Information Sciences* **486**, 310–339. doi:10.1016/j.ins.2019.01.075.

Xu, W., Zhang, C., Peng, Y., Wang, B. & Lu, D. 2013 Study on medium and long-term hydrological forecasting based on data fusion. *Journal of Hydroelectric Engineering* **32** (6), 11–18.

Yaseen, Z. M., Sulaiman, S. O., Deo, R. C. & Chau, K.-W. 2019 An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology* **569**, 387–408. doi:10.1016/j.jhydrol.2018.11.069.

Yue, Z., Ai, P., Yuan, D. & Xiong, C. 2020 Ensemble approach for mid-long term runoff forecasting using hybrid algorithms. *Journal of Ambient Intelligence and Humanized Computing* **13**, 5103–5122. doi:10.1007/s12652-020-02345-9.

Yue, Z., Liu, H. & Zhou, H. 2023 Monthly runoff forecasting using particle swarm optimization coupled with flower pollination algorithm-based deep belief networks: A case study in the Yalong River Basin. *Water* **15**, 1–25. doi:10.3390/w15152704.

First received 11 October 2023; accepted in revised form 9 February 2024. Available online 24 February 2024