A multi-model integration method for monthly streamflow prediction: modified stacking ensemble strategy

Yujie Li, Zhongmin Liang, Yiming Hu, Binquan Li, Bin Xu and Dong Wang

ABSTRACT

In this study, we evaluate elastic net regression (ENR), support vector regression (SVR), random forest (RF) and eXtreme Gradient Boosting (XGB) models and propose a modified multi-model integration method named a modified stacking ensemble strategy (MSES) for monthly streamflow forecasting. We apply the above methods to the Three Gorges Reservoir in the Yangtze River Basin, and the results show the following: (1) RF and XGB present better and more stable forecast performance than ENR and SVR. It can be concluded that the machine learning-based models have the potential for monthly streamflow forecasting. (2) The MSES can effectively reconstruct the original training data in the first layer and optimize the XGB model in the second layer, improving the forecast performance. We believe that the MSES is a computing framework worthy of development, with simple mathematical structure and low computational cost. (3) The forecast performance mainly depends on the size and distribution characteristics of the monthly streamflow sequence, which is still difficult to predict using only climate indices.

Key words | elastic net regression, eXtreme Gradient Boosting, monthly streamflow forecasting, random forest, stacking ensemble strategy, support vector regression

Yujie Li
Zhongmin Liang
Yiming Hu (corresponding author)
Binquan Li
College of Hydrology and Water Resources,
Hohai University,
Nanjing 210098, China
E-mail: *yiming.hu@hhu.edu.cn*

Yujie Li
Department of Infrastructure Engineering,
University of Melbourne,
Melbourne, VIC 3010, Australia

Bin Xu
School of Earth Sciences and Engineering,
Hohai University,
Nanjing 210098, China

Dong Wang
Bureau of Hydrology,
Changjiang Water Resources Commission,
Wuhan 430010, China

INTRODUCTION

Monthly runoff prediction with high and stable performance is of great strategic significance and application value in formulating the rational allocation and optimal operation of water resources (Dai *et al.* 2011; Bennett *et al.* 2017; Liu *et al.* 2018) and improving the breadth and depth of hydrological forecasting integrated services (Bennett *et al.* 2016; Schepen *et al.* 2016). With the vigorous development of water conservancy informatization, hydrological data have gradually shown characteristics of being massive and having multiple sources, multiple structures, high value and sparse value density, as well as strong spatial and temporal attributes (Shortridge *et al.* 2016; Yaseen *et al.* 2016). In the era of big data, while strengthening the collection and collation of basic streamflow observation data, there is theoretical and practical significance in learning how to use new mathematical models and computer technology to explore the intrinsic value and relationship between meteorological and hydrological data (Zhou *et al.* 2011) and to learn how to establish accurate and reliable medium- and long-term streamflow forecasting methods. These two 'how to' topics are the pioneer research fields in developing and expanding hydrological forecasting (Singh *et al.* 2014; Ye & Wu 2018).

Meteorological forecasts coupled with hydrological models, and data-driven methods are two main approaches for monthly streamflow forecasting. The former approach uses monthly meteorological forecasts (e.g. precipitation and evaporation) to drive hydrological models to obtain the monthly streamflow forecast (Martinez & Gupta 2010; Wang *et al.* 2011). The uncertainty, originally from precipitation forecasts, may be further amplified during the hydrological model simulation, which may lead to obstacles

doi: 10.2166/hydro.2019.066

in the application (Humphrey *et al.* 2016; Xiong *et al.* 2018). Data-driven models based on various machine learning algorithms directly build the relationship between predictors (e.g. large-scale climate indices) and predictand (e.g. streamflow) (Vojinovic *et al.* 2003; Wang & Babovic 2016; Yang *et al.* 2019). Hydrologists have introduced a large number of data-driven models into streamflow forecasting (Babovic 2005). Artificial neural network (Babovic *et al.* 2001), support vector machine (SVM) (Liang *et al.* 2017), extreme learning machine (Yaseen *et al.* 2016), relevance vector machine (Liu *et al.* 2017), gradient boosting decision tree (GBDT) (Lu *et al.* 2018), and random forest (RF) (Liang *et al.* 2018) models have shown the potential to produce streamflow forecasts with good performance.

Meanwhile, integration methods have been developed rapidly to effectively use multiple model results. Logistic regression (Šípek & Daňhelka 2015), non-homogeneous regression (Suchetana *et al.* 2019), Bayesian model averaging (Liang *et al.* 2011), and quantile model averaging (Schepen & Wang 2015) are widely regarded as effective combination methods and have demonstrated excellent simulation performance. A stacking ensemble strategy (SES), as an efficient multi-model performance integration framework, has been widely applied in machine learning and effectively reduces bias and variance (Breiman 1996; Ting & Witten 1999; Wolpert & Macready 1999; Seewald 2002). Although several researchers have used the SES in, for example, PM2.5 forecasting (Zhai & Chen 2018), forecasting annual river ice breakup dates (Sun & Trevor 2018) and short-term electricity consumption forecasting (Divina *et al.* 2018), it has not been applied to streamflow forecasting at the monthly scale.

In this paper, elastic net regression (ENR), support vector regression (SVR), RF and eXtreme Gradient Boosting (XGB) models are employed as forecasting models to realize the monthly streamflow prediction. In addition, we propose an improved multi-model integration method named modified SES (MSES). Performance evaluation indices, including the relative error (RE), mean absolute relative error (MAPE), relative root mean square error (RRMSE), quantitative qualification rate (QR1) and qualitative qualification rate (QR2), are employed to compare and analyze the simulation results of the different models. We apply the above methods to the Three Gorges Reservoir in the Yangtze River Basin. The structure of this paper is as follows. The methodologies of ENR, SVR, RF, XGB and the MSES are described in the next section followed by a section which presents the case study. The results and conclusions are presented in the final two sections, respectively.

METHODOLOGY

Elastic net regression

ENR is an enhanced form of multiple line regression (Comber & Harris 2018), which combines two types of norms. It is common knowledge that a regression equation should contain two important parts: loss function and regularization, which is also named the penalty term. When the Euclidean norm (L2 norm) is employed as the penalty term, the equation becomes the Ridge regression and its loss function can be described as follows (Ma *et al.* 2017; Chu *et al.* 2018):

$$\hat{\beta}^{\text{Ridge}} = \arg\min_{\beta} \sum_{i=1}^{N} \left[\left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2 \right] \quad (1)$$

where β_0 is a constant and λ_1 is a penalty term. As the value of λ_1 increases, the shrinkage of the regression coefficient also increases accordingly. When $\lambda_1 = 0$, Ridge regression becomes least squares regression. The L2 norm assumes that the parameters follow a Gaussian distribution, which is beneficial to prevent over-fitting of the simulation.

When the Taxicab norm (L1 norm) is employed as the penalty term, the equation becomes least absolute shrinkage and selection operator regression (Lasso regression) and the loss function of Lasso regression can be described as follows:

$$\hat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \sum_{i=1}^{N} \left[\left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j| \right] \quad (2)$$

The L1 norm assumes that the parameters follow a double exponential distribution, which is conducive to ensuring the sparseness of the weight vector.

ENR combines the advantages of the above two approaches, and the loss function can be defined as follows:

$$\hat{\beta}^{\text{Elastic Net}} = \arg\min_{\beta} \sum_{i=1}^{N} \left[\left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j| \right] \tag{3}$$

Furthermore, another expression can be obtained by the following transformation:

$$\varepsilon = \lambda_1 + \lambda_2,\ \theta = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$
$$\hat{\beta}^{\text{Elastic Net}} = \arg\min_{\beta} \sum_{i=1}^{N} \left[\left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 + \varepsilon \left(\theta \sum_{j=1}^{P} |\beta_j| + (1-\theta) \sum_{j=1}^{P} \beta_j^2 \right) \right] \tag{4}$$

Therefore, the penalty term ε of ENR is just a convex linear combination of the abovementioned models. When $\varepsilon = 0$, it becomes a Ridge regression; when $\varepsilon = 1$, it becomes a Lasso regression.

Support vector regression

SVR is the regression form of SVM which was proposed by Vapnik in 1995 (Cortes & Vapnik 1995) and has been widely used in hydrological forecasting. The basic idea of SVR is to use a small number of support vectors to represent the entire sample set and to convert the low-dimensional non-linear estimation into a high-dimensional linear estimation by using the non-linear mapping function $\varphi(x)$. The regression function can be defined as follows (Liang *et al.* 2017; Mosavi *et al.* 2018; Seo *et al.* 2018):

$$f(x) = \langle \omega_i, \varphi(x_i) \rangle + b \tag{5}$$

where ω_i and b are the weight vectors and bias, respectively, and $<\cdot,\cdot>$ is the vector product operator. In solving the optimal function, SVR introduces the relaxation factors ξ_i and ξ_i^* in the structural risk theory and the regression function is transformed as follows:

$$\begin{cases} \omega = \arg\min_{\beta} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ \text{s.t.} \quad y_i - \langle \omega_i, \varphi(x_i) \rangle - b \le \varepsilon + \xi_i \\ \qquad \langle \omega_i, \varphi(x_i) \rangle + b - y_i \le \varepsilon + \xi_i^* \\ \qquad \xi_i \ge 0,\ \xi_i^* \ge 0,\ i = 1, 2, \cdots, n \end{cases} \tag{6}$$

where C represents the risk experience and complexity factor. ε denotes the allowable error value. By introducing the Lagrange equation and the KKT condition, the above question is transformed into a quadratic programming problem:

$$\begin{cases} L = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} (\xi_i + \xi_i^*) - \sum_{i=1}^{n} \alpha_i (\xi_i + \varepsilon - y_i + \langle \omega_i, \varphi(x_i) \rangle + b) \\ - \sum_{i=1}^{n} \alpha_i^* (\xi_i^* + \varepsilon - y_i - \langle \omega_i, \varphi(x_i) \rangle - b) - \sum_{i=1}^{n} \eta_i (\xi_i + \xi_i^*) \\ \min: \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_i - a_i^*)(a_j - a_j^*) K\langle x_i, x_j \rangle - \sum_{i=1}^{n} y_i (a_i - a_i^*) \\ \qquad + \sum_{i=1}^{n} \varepsilon (a_i + a_i^*) \\ \text{s.t.} \quad \sum_{i=1}^{n} (a_i - a_i^*) = 0 \end{cases} \tag{7}$$

According to the above equation, the optimal result can be obtained as $a = (a_1, a_1^*, \cdots, a_n, a_n^*)$ and the final regression equation can be calculated as follows:

$$f(x) = \sum_{i,j=1}^{n} (a_i - a_i^*) K(x_i, x_j) + b \tag{8}$$

where $K\langle x_i, x_j \rangle$ is a kernel function that should match Mercer's condition (Li *et al.* 2018). In this study, the radial basis function is chosen as the kernel function and defined as follows, where σ represents the Gaussian noise level of the standard deviation:

$$K\langle x_i, x_j \rangle = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2} \right) \tag{9}$$

Random forest

RF was presented by Breiman (2001) as a typical Bagging method based on a decision tree algorithm, whose main

idea is to construct a strong learner by building and integrating a large quantity of weak learners. In this paper, the forecasting process contains three parts (Liang *et al.* 2018; Lai *et al.* 2018).

(1) The bootstrap sampling method is used to extract sub-training sets.

The bootstrap is a returnable sampling method. Assuming that the original dataset D1 contains m samples, random sampling is carried out for m times with replacement, so that the dataset D2 can be obtained. D2, which also contains m samples, exhibits a situation where some samples appear several times and some samples do not appear, and the probability that the samples are not collected in m times of sampling is estimated as follows:

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m=\frac{1}{e}\approx 36.8\% \tag{10}$$

Namely, approximately 36.8% of the samples in D1 did not appear in D2, so we can use D2 to train the model and use D1/D2 to test the model. In other words, both the actual evaluation model and the expected evaluation model use m samples, and at the same time, approximately 36.8% of the samples that do not appear in the training set are still tested. Such test results are called out-of-bag estimates. In general, the bootstrap sampling method is very useful for hydrological datasets that are small and difficult to effectively divide for training and testing according to a certain strategy.

(2) The classification and regression tree (CART) algorithm is employed as a weak learner to obtain sub-forecasting results.

For the generated CART, the category of each leaf node is the average of the labels falling on that leaf node. Assuming that the feature space is divided into M parts, i.e., there are now M leaf nodes $R_1, R_2, \cdots, R_M$, and the corresponding data quantity is $N_1, N_2, \cdots, N_M$, the predicted values of the leaf nodes are:

$$c_m=\frac{1}{N_m}\sum_{x_i\in R_m} y_i \tag{11}$$

The CART is also a binary tree, which is divided according to the value of the feature every time. If the value S of the feature J is segmented, the two regions after segmentation are:

$$R_1(j,s)=\{x_i|x_i^j\le s\},\ R_2(j,s)=\{x_i|x_i^j>s\} \tag{12}$$

Calculate the estimated values $C1$ and $C2$ of $R1$ and $R2$, respectively, and then calculate the loss after splitting according to (j, s):

$$\min_{j,s}\left[\sum_{x_i\in R_1}(y_i-c_1)^2+\sum_{x_i\in R_2}(y_i-c_2)^2\right] \tag{13}$$

(3) Simple average method (also known as voting) is adopted to obtain the final forecasting values:

$$H(x)=\frac{1}{T}\sum_{t=1}^{T}h_t(x) \tag{14}$$

where $h(x)$ and $H(x)$ are the values of the basic learners and ensemble results, respectively.

eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGB) is presented by Chen & Guestrin (2016) as a type of Boosting method based on the GBDT algorithm. It has gradually developed into a distributed gradient enhancement framework. The main idea of XGB is to step up a training process, using all samples for each round of training and changing the weights of the samples. The loss function is used to fit the residual error. The goal of each training round is to fit the residual error of the previous round, and the prediction result is the weighted average of the prediction results of each round when the residual error is small enough or reaches a certain number of iterations (Folberth *et al.* 2019).

The XGB is essentially an additional model, and its sub-decision tree model only uses the CART. Assuming the training sample is $N=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$, and x can be a multi-dimensional vector. $f^t(x)$ represents the predicted

value after the tth iteration, and $f_t(x)$represents the increment of the tth iteration:

$$f^t(x) = f^{t-1}(x) + f_t(x) = \sum_{k=1}^{t} f_k \tag{15}$$

Model learning is a non-deterministic polynomial process that finds the optimal solution and always uses a heuristic strategy. Assuming that y_i is the observed values of the training samples in the iterative process; $\hat{y}_i$ is the forecasting values; L is the loss function and Ω is the regularization function, then the objective function OBJ of XGB can be defined as follows:

$$\begin{aligned} Obj &= \sum L(y_i, \hat{y}_i) + \sum \Omega(g_k) \\ \Omega(f) &= \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \end{aligned} \tag{16}$$

In the regularization function, T is the number of leaf nodes; ω is the fraction of leaf nodes; λ is the regularization parameter and γ is the minimum loss required to further divide the leaf nodes. Penalizing the number of leaf nodes is equivalent to pruning in training, so that the model is not prone to over-fitting.

Modified stacking ensemble strategy

As shown in Figure 1, the data structure and calculation process of the MSES can be summarized in the following steps. The whole dataset has been divided into training and testing, 37 years (1965–2001) and 15 years (2002–2016), respectively. Besides, we call the model used in the second layer meta-model, as in the original SES (Zhai & Chen 2018).

Step 1: In the first training period, the four models mentioned above (ENR, SVR, RF and XGB) are calibrated independently by using the same training dataset and adopting the leave-one-out cross-validation (loocv) strategy to generate the validation values. Specifically, 36 years are used to calibrate the models, and the remaining 1 year is used for validation. Therefore, for a certain model, we can obtain 37 years long-validation results (orange, yellow, green and red parts).

Step 2: In the first testing period, we use the whole training dataset (37 years) to calibrate the models and then generate the predictions (15 years, light blue parts).

Step 3: All of the validations are composed sequentially into a new training set for the second layer. Although there is only one test prediction, the results of the 37 predictions are slightly different (RE is less than 1%, not shown). Therefore, all the predictions are built into a new testing set for the second layer by the simple average method.

Step 4: In the second training period, the dataset is composed of observations and four-model-validated values (validation in the first layer) that are 37 years long.

Step 5: Similarly, the dataset in the second testing period is composed of observations and four-model-predicted values (prediction in the first layer) that are 15 years long. After that, the meta-model is employed to recalibrate and repredict the final streamflow values.

Step 6: Then, the meta-model is employed to calibrate the multi-model aggregate simulation (using the second training period dataset of Step 4) and is applied to the second testing dataset (Step 5) to generate the final multi-model aggregate prediction.

Compared with the original SES (Divina *et al.* 2018; Sun & Trevor 2018; Zhai & Chen 2018), the MSES is improved in two parts:

(i) The reconstruction of the data structure.

In this paper, we propose to use the loocv strategy, taking the place of k-fold-cross-validation, to calibrate the models in the training period, which is very important for a small sample dataset. On the one hand, using the loocv make the best use of short sequence data; on the other hand, it can reduce the chance of causing different divisions.

(ii) The selection of the meta-model.

In the previous works (Sikora 2015; Divina *et al.* 2018; Sun & Trevor 2018; Zhai & Chen 2018), authors analyzed the weights of different models to calculate the final predictions. In other words, the weighting method often means a certain linear relationship between the subs and final predictions, which can be combined with multiple linear regression (MLR). However, we believe that there is not

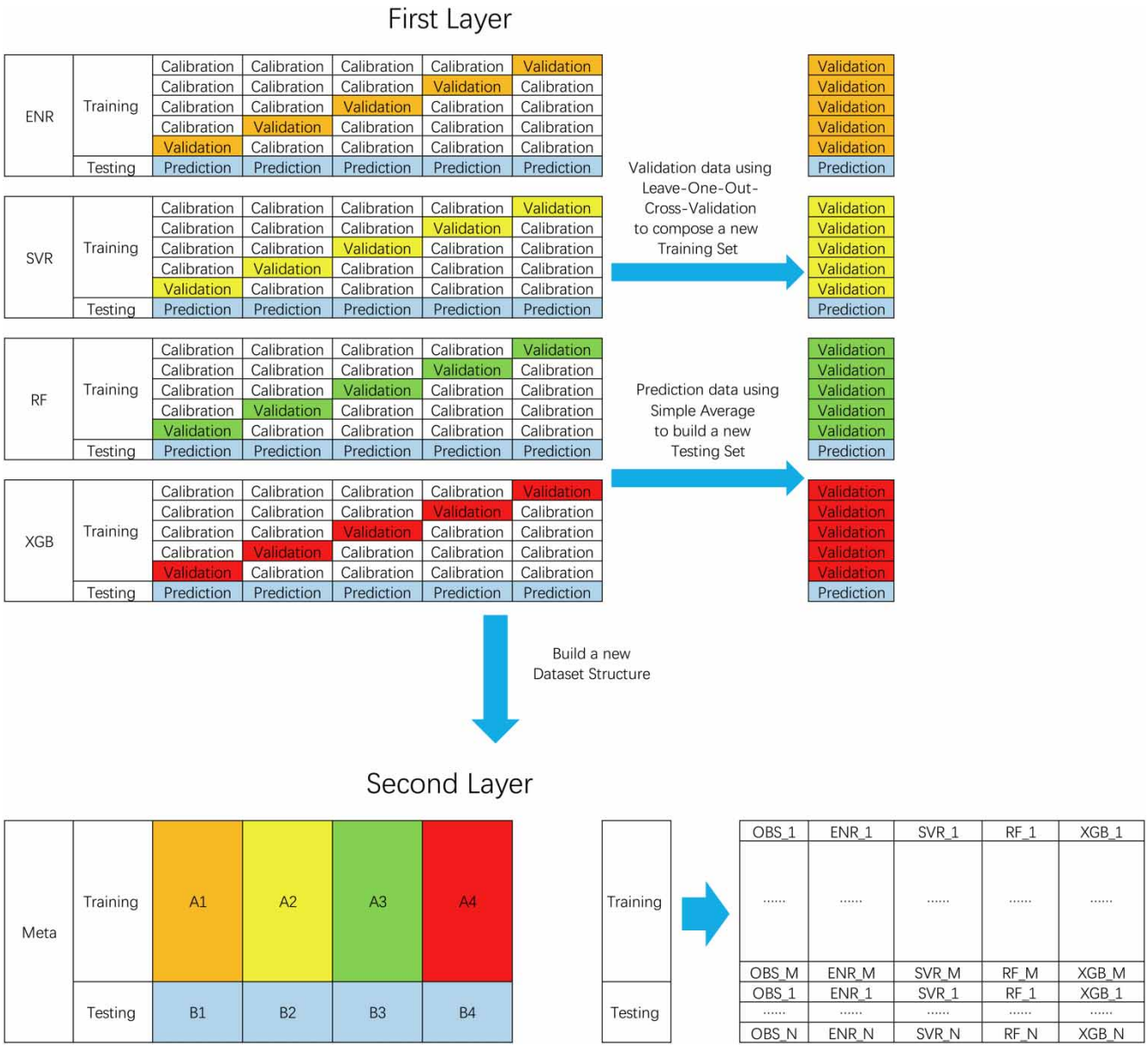

Figure 1 | Structure of the MSES.

necessarily a simple linear or non-linear relationship between the sub-prediction models and the predictand, i.e., the relationship may not have an explicit mathematical expression. Therefore, we abandon the concept of weight and propose to employ the machine learning model with the best performance in the first layer to take the place of the MLR in the second layer as the meta-model. To make the results more convincing, we compare a total of nine kinds of prediction results.

CASE STUDY

Study area

In this paper, we apply the proposed methodology to the Three Gorges Reservoir located in the Yangtze River Basin. The Yangtze River is the largest and longest river in China, with a basin area of 1.8×10^6 km^2, an annual average rainfall of 1,100 mm and a multi-year average total water

resources of 9.96×10^{11} m^3, accounting for approximately 35% of the total water resources in the country. The theoretical reserve of hydropower resources is 3.05×10^8 kW, and the average annual power generation is 2.67×10^{12} kWh, accounting for approximately 40% in the country. There are more than 3,600 navigable rivers in the Yangtze River system, with a total navigable length of approximately 7.1×10^4 km, accounting for 56% of the national inland navigable mileage. Influenced by a monsoon climate, the Yangtze River Basin has large spatial and temporal variabilities in rainfall and uneven distribution throughout the year, mainly concentrated from May to October, which is also the corresponding flood season period (Bai *et al.* 2016; Xu *et al.* 2018).

The Three Gorges Reservoir, located in Yichang City, is the largest water conservancy project in China. It has a total storage capacity of 4.51×10^{10} m^3 and a flood control capacity of 2.22×10^{10} m^3, with a design flood level of 175 m and a check flood river of 180 m (Huang *et al.* 2017). The completion of the Three Gorges Project, it has provided strong protection to flood control, water supply, shipping and power generation. At the same time, it also played a role in providing clean energy, reducing environmental pollution, improving river water quality, and protecting the ecological environment.

Dataset

The monthly mean streamflow data of the Three Gorges Reservoir are provided by the Changjiang Water Resources Commission (China) for the period 1965–2016. Streamflow here means the reconstructed natural inflow data to the reservoir, which is estimated by applying the regulation rules of the upstream cascade reservoirs and the principle of basin water balance. As shown in Figure 2, the box chart sequentially includes abnormal values (black circles), maximum non-abnormal values, upper quartile values, mean values (purple triangles), median values (blue lines), lower quartile values, minimum non-abnormal values and abnormal values (black circles) from top to bottom. This clearly illustrates two features. First, the variation in streamflow in flood months (May to October) is extremely large, and the values between the same months often show multiple differences. Second, during the other months, although the absolute value of the amplitude variation is not large, the increase in abnormal points shows an obvious interannual difference. The above two points have increased the difficulty of monthly streamflow forecasting.

Predictors

The predictor dataset is based on 130 climate indices provided by the National Climate Center (China). It includes three categories: 88 atmospheric circulation indices, 26 sea surface temperature indices and 16 other indices. The specific predictors can be found at http://cmdp.ncc-cma.net/Monitoring/cn_index_130.php. Since the release time of the 130 climate indices is the first of each month, we also give the monthly streamflow forecast for the next month later on the same day. In this paper, we assume that the effect of the climate indices lasts up to 1 year. The prediction structure can be defined as follows:

$$Q(t) = f(Q(t-1), \overrightarrow{\mathrm{CI}}_i\,(t-1), \mathrm{CI}_i\,(t-2), \cdots, \mathrm{CI}_i\,(t-12)) \tag{17}$$

where $Q(t)$ is the forecast streamflow of the month t. The predictors consist of the previous monthly observations $Q(t-1)$ and the climate indices of the preceding 12 months $\overrightarrow{\mathrm{CI}}_i\,(t-1), \mathrm{CI}_i\,(t-2), \cdots, \mathrm{CI}_i\,(t-12)$. CI_i is a vector that consists of 130 climate indices as $\mathrm{CI}_i = [\mathrm{CI}_1, \mathrm{CI}_2, \cdots, \mathrm{CI}_{130}]$. Therefore, there are 1,561 primary predictors. From these 1,561 predictors, a sub-set of the most important ones for a good modeling result has to be selected.

Predictor selection is an equally important step as a model construction. Since there is no unified standard for the method of selection and choosing the number of predictors, in this study, we use the regression mechanism inherent in the four models for predictor selection. We have calculated all the schemes with the number of selected predictors ranging from 10 to 50, and the results indicate (not shown, the selected predictors are analyzed in a separate paper which has been submitted to *Theoretical and Applied Climatology, 2019*) that when the number of predictors is between 15 and 30, the models have the best simulation in the loocv period. To improve the calculation efficiency, here we decide to set the number of predictors to 15.

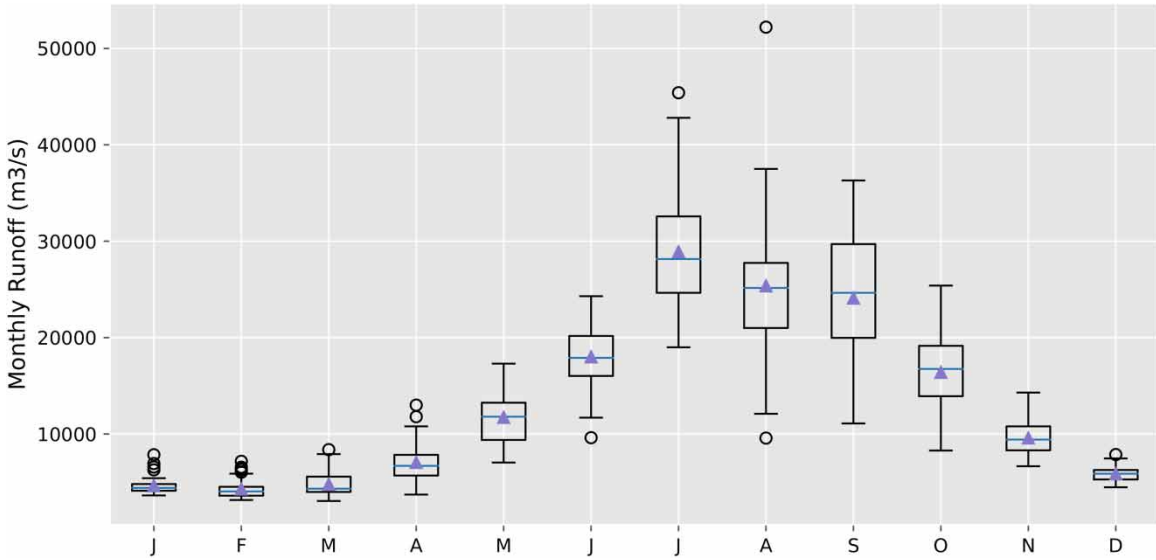

Figure 2 | Monthly mean inflow sequence of the Three Gorges Reservoir over the 52-year period 1965–2016.

Performance evaluation indices

The model accuracy analysis is based on the following performance indices, which have been widely used to evaluate the goodness-of-fit of the hydrologic models. We use the RRMSE, RE, mean absolute percentage error (MAPE) (Chadalawada & Babovic 2019) and qualification rate (QR). In the following descriptions, $Q_{i,o}$, $Q_{i,s}$, $\overline{Q_o}$ and $\overline{Q_s}$ are observed values, forecasting values, the mean of observed sequences and the mean of forecasting sequences, respectively. The values of n and N are the qualified length and total length of the dataset, respectively.

(1) Relative root mean square error

The RRMSE is based on the RMSE. Since the RMSE is related to the magnitude of the streamflow values, it cannot be directly used to compare simulation errors between different months. However, the value range of the RRMSE is 0–1, so it is adopted to solve the issue and evaluate the performance of the models in the flood season and non-flood season. The RRMSE is calculated as follows (Lin & Chen 2004):

$$\mathrm{RRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Q_{i,s}-Q_{i,o}}{Q_{i,o}}\right)^2} \tag{18}$$

(2) Relative error and mean absolute percentage error

The RE is a conventional metric used to show the results at each data point. The MAPE represents the average level of the RE. They are given by the following:

$$\mathrm{RE} = \frac{Q_{i,s}-Q_{i,o}}{Q_{i,o}} \tag{19}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Q_{i,s}-Q_{i,o}}{Q_{i,o}}\right| \tag{20}$$

(3) Qualification rate

The QR is used to evaluate the eligibility of the streamflow sequence in both calibration and verification. It is issued by the Ministry of Water Resources of China and defined as follows:

$$\mathrm{QR} = \frac{n}{N} \tag{21}$$

$$\mathrm{Anomaly} = \frac{Q_{i,o}-\overline{Q_o}}{\overline{Q_o}} \tag{22}$$

The standard for hydrological prediction in China includes qualitative and quantitative prediction. In the

qualitative forecast, the results can be divided into five levels: dry (Anomaly $< -20\%$), partially dry ($-20\% \leq$ Anomaly $< -10\%$), normal ($-10\% \leq$ Anomaly $\leq 10\%$), partially wet ($10\% <$ Anomaly $\leq 20\%$) and wet (Anomaly $> 20\%$). If the forecasting level is the same as the observed level, it is counted as a qualified point; otherwise, it is unqualified. For the quantitative forecast, a permissible error is first defined as 20% of the multi-year amplitude in the same period for many years. In this paper, the multi-year amplitude is the result of the maximum streamflow (from 1965 to 2016) minus the minimum streamflow (from 1965 to 2016), and the permissible error is the multi-year amplitude times 20%. Second, if the forecasting error is lower than the permissible error, it is counted as a qualified point; otherwise, it is unqualified. In this paper, the above quantitative QR (QR1) and qualitative QR (QR2) methods are both used. QR1 considers the changes in streamflow over the years and evaluates different months with different standards. QR2 illustrates the estimation ability for the different distributions of streamflow magnitude. These two eligibility criteria are widely used in China, because they effectively combine the error characteristics based on the different streamflow sequence lengths, multi-year variations and extreme value distributions.

RESULTS AND DISCUSSION

Training period

We evaluate the above four models in the loocv training period (1965–2001) using a quantile–quantile (Q–Q) plot (Figure 3). Compared with all four models, XGB is the best model for the simulation results, which are closest to the 1:1 line and relatively evenly distributed on both sides. The quantiles of SVR are farther from the 1:1 line, followed by RF and ENR. When the streamflow values exceed 30,000 m^3/s, almost all of the points in Figure 3 are above the red line, which shows that the four models are underpredicting, and when the streamflow values reach 55,000 m^3/s, the maximum forecasting value is only 35,000 m^3/s.

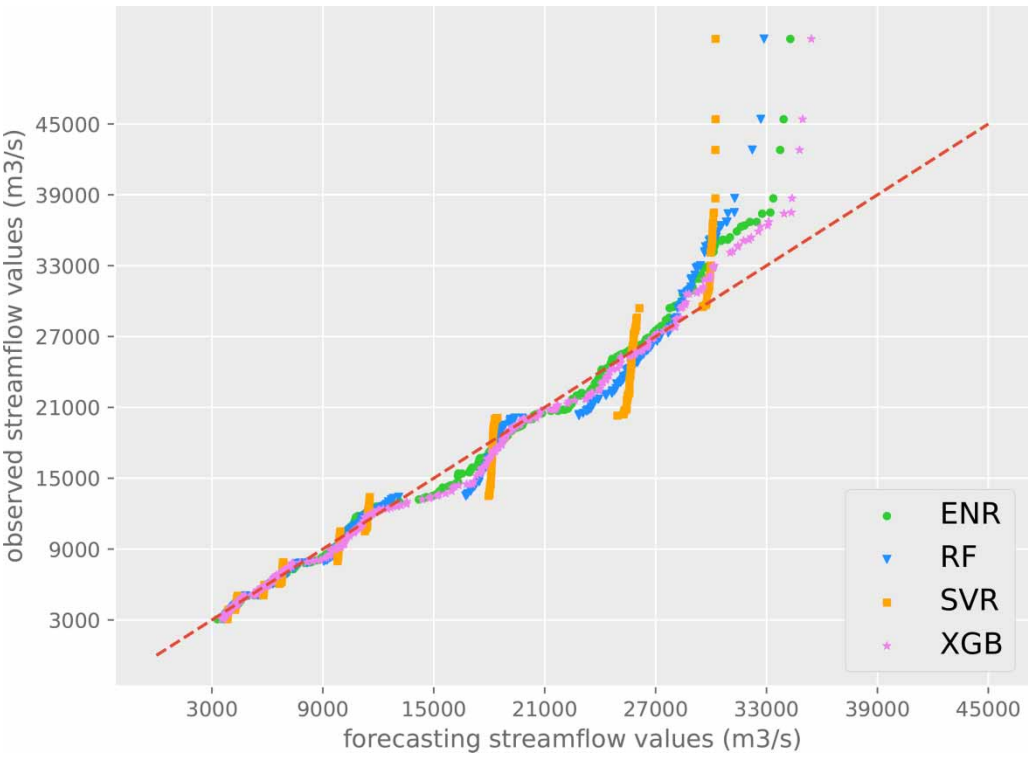

Figure 3 | Q–Q plot shows the relationship between the observation and forecasting streamflow data in the training period. The red line represents the 1:1 line which means a perfect fit. Please refer to the online version of this paper to see this figure in color: http://dx.doi.org/10.2166/hydro.2019.066.

Table 1 | The performance evaluation indices (RRMSE, MAPE, QR1 and QR2) of ENR, SVR, RF and XGB models for 12 months in the training period

Month	Jan				Feb				Mar				Apr			
Model	**ENR**	**SVR**	**RF**	**XGB**	**ENR**	**SVR**	**RF**	**XGB**	**ENR**	**SVR**	**RF**	**XGB**	**ENR**	**SVR**	**RF**	**XGB**
RRMSE	0.094	0.101	0.096	0.076	0.108	0.108	0.103	0.086	0.205	0.235	0.207	0.197	0.271	0.271	0.245	0.173
MAPE (%)	7.4	8.3	7.7	6.0	8.9	8.9	8.5	6.8	15.5	17.9	15.8	15.4	21.6	20.9	20.3	13.0
QR1 (%)	56.8	43.2	54.1	69.2	56.8	54.1	59.5	66.7	59.0	51.4	56.8	54.1	73.0	70.3	67.6	82.1
QR2 (%)	54.1	62.2	62.2	69.2	62.2	62.2	62.2	61.5	38.5	43.2	48.6	48.6	35.1	37.8	24.3	35.9
Month	**May**				**Jun**				**Jul**				**Aug**			
RRMSE	0.327	0.247	0.244	0.278	0.261	0.277	0.243	0.233	0.218	0.220	0.211	0.190	0.369	0.304	0.334	0.220
MAPE (%)	25.7	19.1	20.2	22.6	19.5	19.9	16.8	16.8	16.7	17.8	16.7	13.9	23.0	20.9	21.6	14.7
QR1 (%)	43.2	54.1	40.5	46.2	43.6	51.4	62.2	62.2	59.0	59.5	62.2	73.0	83.8	75.7	81.1	87.2
QR2 (%)	29.7	32.4	29.7	12.8	30.8	40.5	45.9	45.9	23.1	32.4	35.1	35.1	35.1	40.5	37.8	46.2
Month	**Sep**				**Oct**				**Nov**				**Dec**			
RRMSE	0.260	0.303	0.275	0.230	0.196	0.180	0.180	0.195	0.193	0.163	0.171	0.173	0.140	0.127	0.125	0.115
MAPE (%)	19.2	22.4	20.5	17.0	15.0	14.2	13.9	15.5	14.3	13.1	13.3	13.2	11.0	10.4	10.2	8.4
QR1 (%)	56.4	40.5	51.4	67.6	56.4	59.5	59.5	54.1	54.1	59.5	54.1	48.7	51.3	59.5	54.1	64.9
QR2 (%)	20.5	32.4	32.4	35.1	28.2	48.6	48.6	37.8	43.2	54.1	54.1	30.8	38.5	56.8	56.8	56.8

As we know, constructing a dedicated forecast model for each calendar month is an effective method to reduce simulation errors and improve prediction performance, i.e., in this paper, we have built 12 forecasting models for 12 months. We compare the forecast performances of the above evaluation indices for the four models in 12 months (Table 1).

In general, XGB shows the best forecasting performance, accounting for the 9 smallest RRMSEs, 9 smallest MAPEs, 8 highest QR1 values and 7 highest QR2 values in 12 months. RF is the second-best model, which accounts for the 2 smallest and 6 second-smallest RRMSEs, 2 smallest and 5 second-smallest MAPEs, 2 highest and 4 second-highest QR1 values, and 7 highest and 3 second-highest QR2 values. The accuracies of ENR and SVR are similar, but considerably less than the accuracy of the other two models. Taking the RE as the key evaluation index, the heatmap in Figure 4 demonstrates the detailed values of each model in each simulation period and reflects two key points. First, for the different months of the year, the accuracy in the non-flood season is far better than that in the flood season, which is understandable, because the streamflow in the non-flood season has a smaller variation and a more stable trend. Second, for the same month, different forecast models show similar forecast results. In other words, when using meteorological predictors to forecast the monthly streamflow, different models forecast similar change trends in a certain month, such as abundant water or dryness, but the specific numbers are different.

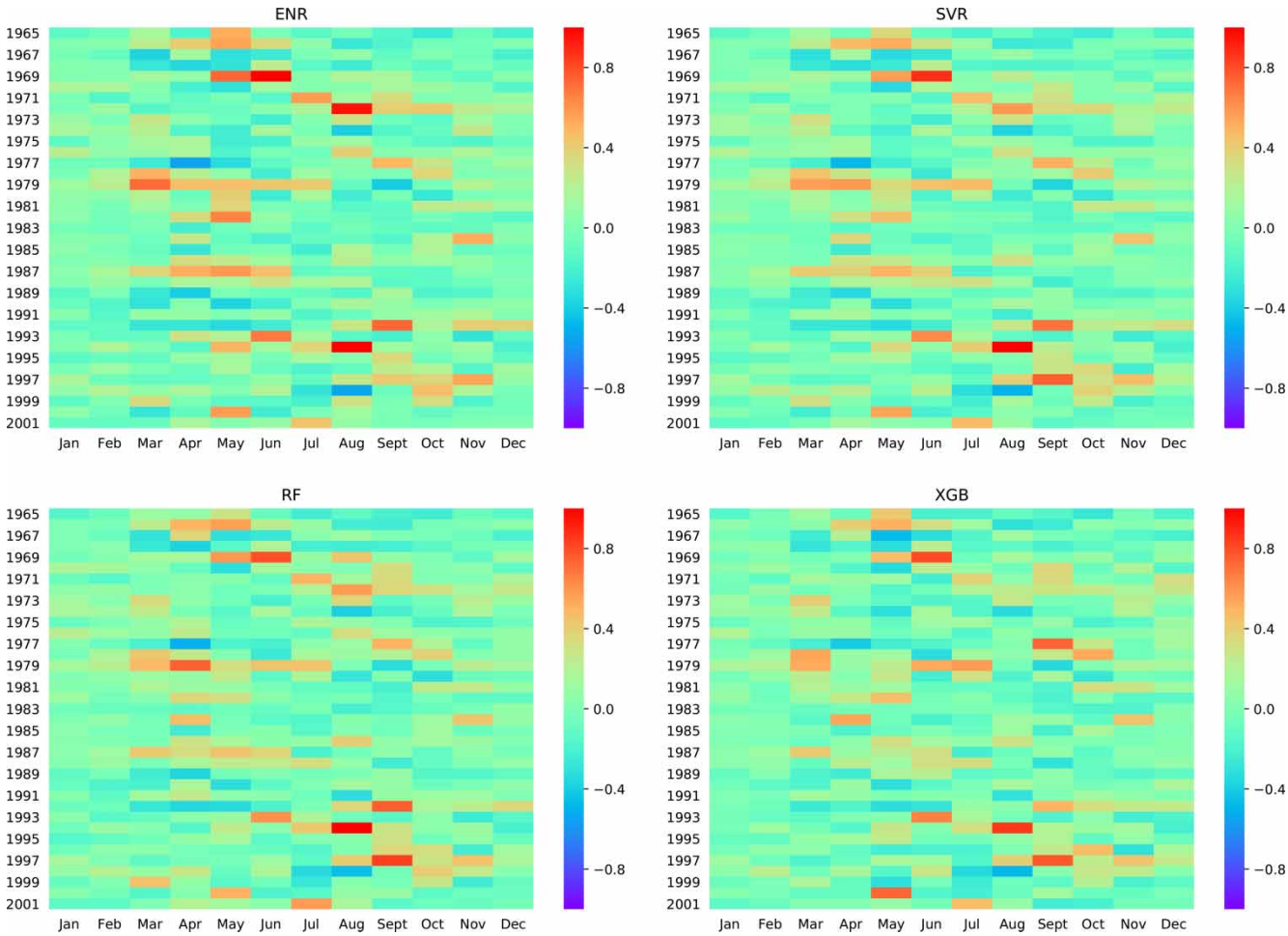

Figure 4 | Heatmap plot presenting the RE of each model in each month in the training period. The legend is set to +100% at the maximum and −100% at the minimum. As the color in the grid gets darker, the error increases accordingly.

It is worth noting that although the simulation accuracy of XGB has reached an acceptable level according to the MAPE, if we focus on QR1 and QR2, monthly streamflow for this case study is still difficult to predict.

Testing period

According to the above analysis, XGB is the model with the best simulation performance and the most stable prediction performance in the first layer, so it is employed as the meta-learner in the second layer. The modified stacked ensemble strategy (MSES) simulation result is evaluated for the testing period (2002–2016). We also evaluate the four individual model results for the testing period. Figure 5 reveals a situation similar to that in Figure 3 and can be summarized as follows. First, from the initial four models, RF and XGB display better forecast performances than SVR and ENR. The two machine learning models are closer to the 1:1 line and more evenly distributed, which is reflected in the simulation of the extreme streamflow values. Second, SVR shows an unsuitable curve both in training and testing. During the loocv period, we have further changed the two core parameters of SVR, i.e., *C* and gamma, but this hardly improved the simulation results. Therefore, we believe that SVR may not be suitable for the case of a large number of predictors. Third, the simulation curve of the MSES is between the curve of the RF and XGB, from which we cannot directly judge the performance. Consequently, in Table 2, we again compare the accuracy and stability of the models through the above four performance evaluation indices. In addition, we add the results of the original SES (OSES), whose meta-model is MLR, to compare the performances for the two kinds of stacking. As the scatter plots of the OSES and MSES are very close, we only provide the values in Table 2 rather than drawing the OSES results in Figure 5. Moreover, it is notable that the results may overpredict in testing and underpredict in training. Follow-up research will be done to find out if this is a structural phenomenon, and if so, what can be the reasons.

It is found that the MSES is overall the best-performing method, as shown in Table 2, accounting for the 10 smallest

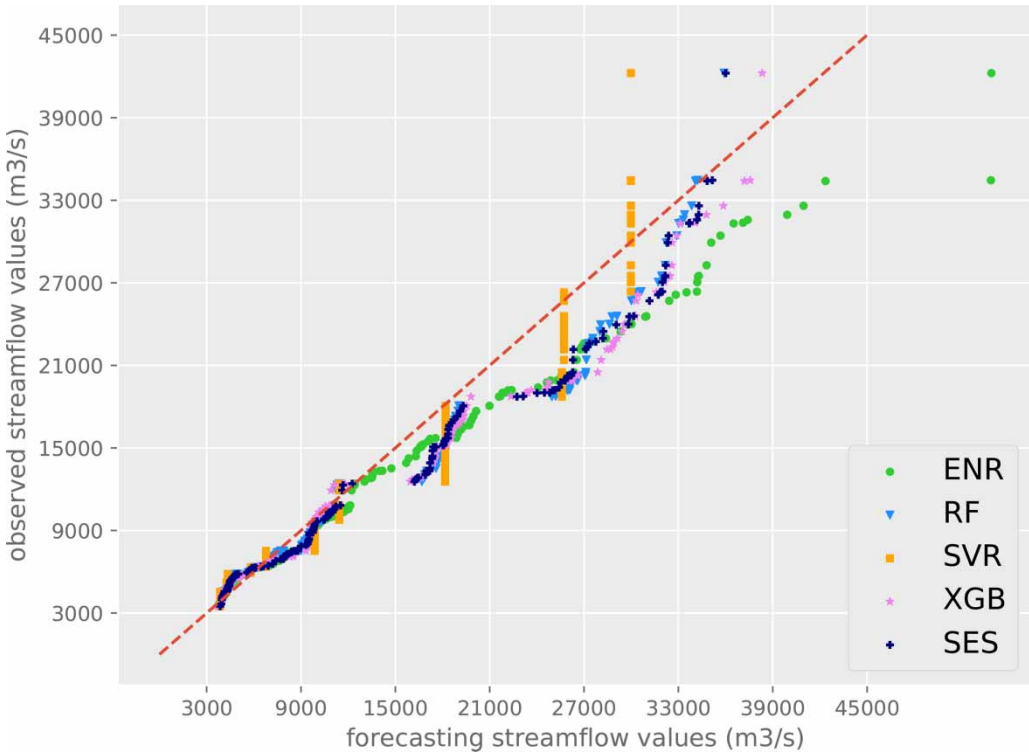

Figure 5 | Q–Q plot showing the relationship between the observed and forecasting streamflow data in the testing period. The red line represents the 1:1 line which means a perfect fit. Please refer to the online version of this paper to see this figure in color: http://dx.doi.org/10.2166/hydro.2019.066.

Table 2 | The performance evaluation indices (RRMSE, MAPE, QR1 and QR2) of ENR, SVR, RF, XGB, OSES and MSES models in 12 months in the testing period (2002–2016)

Month	Jan						Feb					
Model	ENR	SVR	RF	XGB	MSES	OSES	ENR	SVR	RF	XGB	MSES	OSES
RRMSE	0.236	0.247	0.229	0.232	0.227	0.236	0.156	0.170	0.151	0.148	0.135	0.156
MAPE (%)	19.0	20.4	18.5	18.8	18.4	19.0	13.4	14.8	12.7	12.7	12.0	13.4
QR1 (%)	46.7	46.7	46.7	46.7	46.7	46.7	53.3	46.7	66.7	60.0	53.3	53.3
QR2 (%)	26.7	13.3	26.7	26.7	33.3	26.7	20.0	13.3	26.7	26.7	40.0	20.0
Month	**Mar**						**Apr**					
RRMSE	0.169	0.210	0.179	0.164	0.159	0.159	0.239	0.268	0.175	0.182	0.168	0.266
MAPE (%)	14.4	18.0	14.7	13.9	13.4	13.5	18.7	18.0	15.1	14.3	14.6	17.8
QR1 (%)	40.0	40.0	46.7	46.7	53.3	53.3	40.0	73.3	40.0	60.0	60.0	73.3
QR2 (%)	26.7	20.0	33.3	33.3	26.7	26.7	46.7	53.3	26.7	33.3	26.7	53.3
Month	**May**						**Jul**					
RRMSE	0.273	0.230	0.208	0.206	0.180	0.180	0.537	0.425	0.438	0.410	0.325	0.409
MAPE (%)	21.4	19.2	17.9	18.0	15.3	15.3	42.5	33.6	34.2	34.0	26.0	34.0
QR1 (%)	33.3	40.0	40.0	46.7	53.3	53.3	40.0	33.3	40.0	26.7	46.7	26.7
QR2 (%)	33.3	33.3	33.3	33.3	33.3	33.3	13.3	26.7	20.0	6.7	20.0	6.7%
Month	**Jun**						**Aug**					
RRMSE	0.334	0.188	0.196	0.207	0.202	0.207	0.854	0.634	0.612	0.646	0.492	0.645
MAPE (%)	26.0	15.5	16.1	17.1	17.0	17.1	63.1	48.2	42.0	50.5	31.0	50.5
QR1 (%)	33.3	33.3	33.3	46.7	46.7	46.7	26.7	20.0	26.7	13.3	40.0	13.3
QR2 (%)	26.7	46.7	40.0	26.7	33.3	26.7	6.7	6.7	20.0	6.7	13.3	6.7
Month	**Sep**						**Oct**					
RRMSE	0.489	0.568	0.512	0.476	0.434	0.567	0.397	0.321	0.342	0.318	0.208	0.319
MAPE (%)	34.8	40.8	36.8	32.9	31.3	40.7	32.5	26.6	28.3	26.9	15.9	27.0
QR1 (%)	33.3	40.0	46.7	46.7	53.3	40.0	13.3	13.3	13.3	6.7	46.7	6.7
QR2 (%)	26.7	20.0	20.0	13.3	13.3	20.0	6.7	6.7	13.3	6.7	26.7	6.7
Month	**Nov**						**Dec**					
RRMSE	0.293	0.289	0.259	0.225	0.256	0.225	0.108	0.094	0.096	0.099	0.087	0.099
MAPE (%)	23.7	24.1	21.1	18.0	21.0	18.0	9.2	7.7	7.6	7.5	7.0	7.5
QR1 (%)	73.3	73.3	86.7	93.3	93.3	93.3	20.0	40.0	40.0	53.3	40.0	53.3
QR2 (%)	20.0	33.3	40.0	20.0	33.3	20.0	53.3	73.3	73.3	73.3	73.3	73.3

and 2 second-smallest RRMSEs, 9 smallest and 2 second-smallest MAPEs, 9 highest and 2 second-highest QR1 values and 5 highest and 4 second-highest QR2 values. The OSES accounts for the 3 smallest and 1 second-smallest RRMSEs, 2 smallest and 2 second-smallest MAPEs, 7 highest and 0 second-highest QR1 values and 3 highest and 3 second-highest QR2 values, making it the second-best method. The performances of the remaining four models are similar to those in the loocv period. Therefore, we can summarize the following conclusions. First, the

machine learning models represented by Bagging method (RF) and Boosting method (XGB) can be used to forecast the monthly streamflow with higher accuracy and better stability than the traditional SVR and ENR models. Second, as an ensemble strategy, i.e., the multi-model integration method, the MSES has the advantages of a clear calculation structure and low calculation cost. Compared with the OSES, the MSES can effectively reconstruct the original training data in the first layer to optimize the nonlinear machine learning model in the second layer and improve the prediction performance.

Unsurprisingly, the RE in the testing period shown in Figure 6 is generally much larger than that in the loocv period. By comparing the distribution of errors, it can be seen that the simulation accuracy depends on the size and distribution characteristics of the streamflow. In other words, as the distribution range becomes wider, the prediction performance gradually becomes unstable. For example, the accuracy of the flood season (May to October) forecast is much lower than that of the non-flood season. In addition, some studies (Wang *et al.* 2013; Liang *et al.* 2017; Liang *et al.* 2018) show that the uncertainty of autumn rain (a special weather phenomenon) in West China is one of the reasons for the difficulties in monthly streamflow forecasting in the Yangtze River Basin from September to November. It is suggested that meteorological forecasts need to be introduced to improve the accuracy of monthly streamflow forecasts.

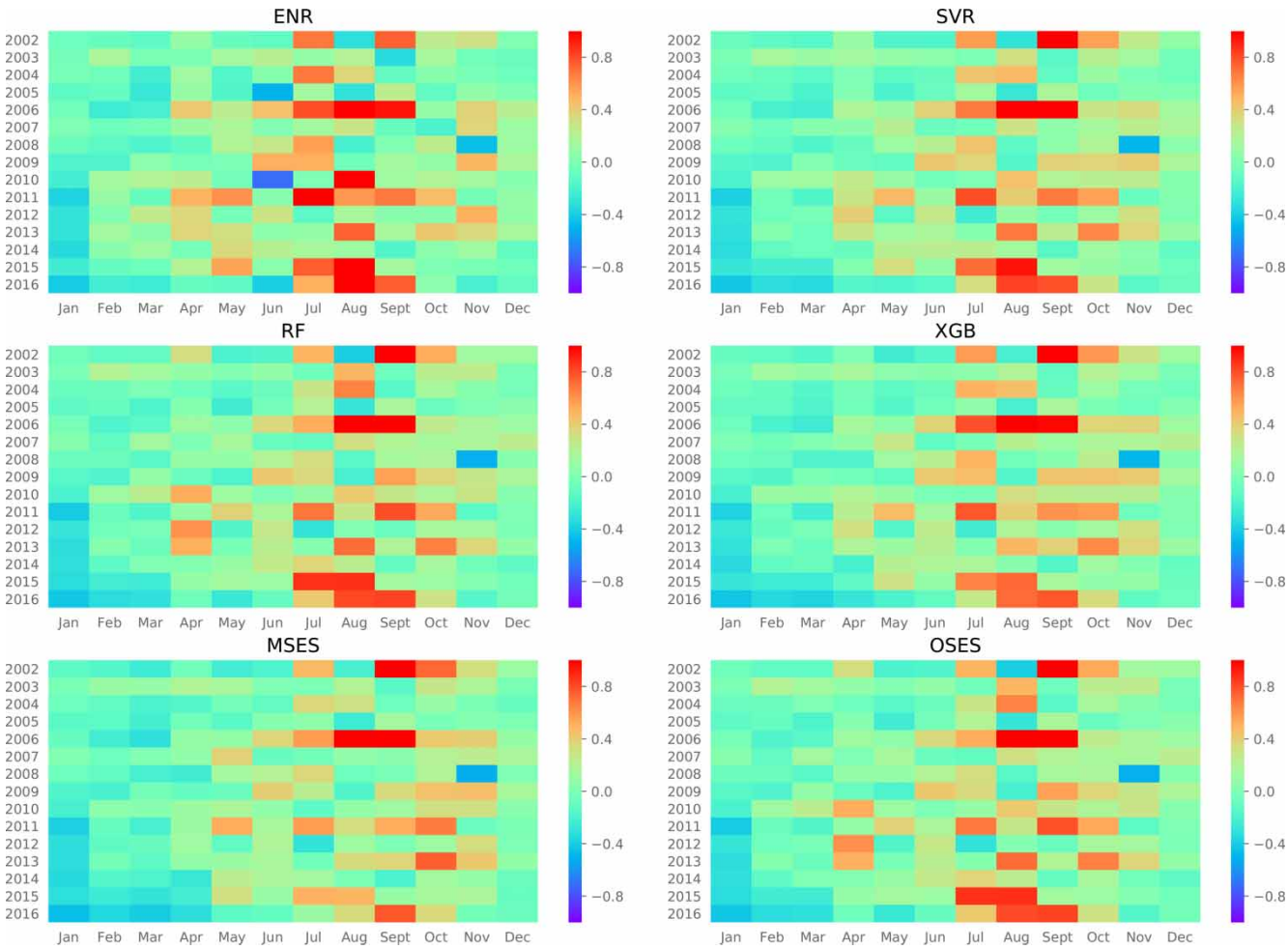

Figure 6 | Heatmap plot presenting the RE of each model and the MSES result in each month in the testing period. The legend is set to +100% at the maximum and −100% at the minimum. As the color in the grid gets darker, the error increases accordingly.

CONCLUSIONS

New data-driven modeling methods are being developed continuously and widely applied to monthly runoff forecasting. Developing consistently accurate multi-model integration methods for these data-driven models is becoming increasingly important. In this paper, we evaluate for different types of data-driven models. Specifically, the ENR model based on the MLR, the SVR based on the statistical learning theory, the RF model based on the Bagging algorithm and the XGB model based on the Boosting algorithm (both based on machine learning) are employed as monthly streamflow forecasting models. We then propose an ensemble integration method based on the SES named the MSES. We apply the above forecasting models and the ensemble integration method to realize monthly runoff prediction to the Three Gorges Reservoir in the Yangtze River Basin to realize monthly streamflow prediction. Five evaluation metrics (RRMSE, RE, MAPE, QR1 and QR2) are employed to measure the forecast performance. Through the simulation results in the periods of training and testing, the following conclusions can be obtained.

(1) The different models often predict similar tendencies, such as wet or normal or dry, but the specific values differ greatly. The RF and XGB present better forecasting performances and higher and more stable accuracies than ENR and SVR. It can be said that the regression models based on machine learning have the potential for application in monthly streamflow forecasting.
(2) The MSES, as a modified stacking ensemble integration method, has the advantages of a clear calculation structure and low calculation cost. Compared with the OSES, the MSES reconstructs more effectively the original training data in the first layer and optimize the nonlinear machine learning model in the second layer to reduce prediction error and improve prediction performance. We believe that the MSES is a multi-model computing framework worth testing on other catchments and hydrological forecasting problems.
(3) However, by comparing the distribution of errors, it can be inferred that the simulation performance mainly depends on the size and distribution characteristics of the streamflow. In other words, as the distribution range becomes wider, the prediction performance gradually becomes unstable. We believe that if only large-scale climate indices and the previous monthly streamflow are used, it will still be difficult to make accurate monthly forecasts.

ACKNOWLEDGEMENTS

The study is financially supported by the National Key Research and Development Program of China (2016YFC0402706, 2016YFC0402707), Fundamental Research Funds for the Central Universities (2018B611X14), Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX18_0584) and Chinese Government Scholarship. We gratefully acknowledge the anonymous editors and reviewers for their insightful and professional comments, which greatly improved this manuscript.

REFERENCES

Babovic, V. 2005 Data mining in hydrology. *Hydrological Processes: An International Journal* **19** (7), 1511–1515.
Babovic, V., Cañizares, R., Jensen, H. R. & Klinting, A. 2001 Neural networks as routine for error updating of numerical models. *Journal of Hydraulic Engineering* **127** (3), 181–193.
Bai, Y., Xie, J., Wang, X. & Li, C. 2016 Model fusion approach for monthly reservoir inflow forecasting. *Journal of Hydroinformatics* **18** (4), 634–650.
Bennett, J. C., Wang, Q., Li, M., Robertson, D. E. & Schepen, A. 2016 Reliable long-range ensemble streamflow forecasts: combining calibrated climate forecasts with a conceptual runoff model and a staged error model. *Water Resources Research* **52** (10), 8238–8259.
Bennett, J. C., Wang, Q. J., Robertson, D. E., Schepen, A., Li, M. & Michael, K. 2017 Assessment of an ensemble seasonal streamflow forecasting system for Australia. *Hydrology and Earth System Sciences* **21** (12), 6007–6030.
Breiman, L. 1996 Stacked regressions. *Machine Learning* **24** (1), 49–64.
Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.
Chadalawada, J. & Babovic, V. 2019 Review and comparison of performance indices for automatic model induction. *Journal of Hydroinformatics* **21** (1), 13–31.
Chen, T. & Guestrin, C. 2016 Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

Chu, H., Wei, J. & Qiu, J. 2018 Monthly streamflow forecasting using EEMD-Lasso-DBN method based on multi-scale predictors selection. *Water* **10** (10), 1486.

Comber, A. & Harris, P. 2018 Geographically weighted elastic net logistic regression. *Journal of Geographical Systems* **20** (4), 317–341.

Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Machine Learning* **20** (3), 273–297.

Dai, Z., Amatya, D. M., Sun, G., Trettin, C. C., Li, C. & Li, H. 2011 Climate variability and its impact on forest hydrology on South Carolina coastal plain, USA. *Atmosphere* **2** (3), 330–357.

Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M. & Torres, J. 2018 Stacking ensemble learning for short-term electricity consumption forecasting. *Energies* **11** (4), 949.

Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N, & Obersteiner, M. 2019 Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agricultural and Forest Meteorology* **264**, 1–15.

Huang, F., Huang, J., Jiang, S.-H. & Zhou, C. 2017 Prediction of groundwater levels using evidence of chaos and support vector machine. *Journal of Hydroinformatics* **19** (4), 586–606.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C. & Maier, H. R. 2016 A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology* **540**, 623–640.

Lai, C., Chen, X., Wang, Z., Xu, C.-Y. & Yang, B. 2018 Rainfall-induced landslide susceptibility assessment using random forest weight at basin scale. *Hydrology Research* **49** (5), 1363–1378.

Li, X., Sha, J., Li, Y.-m. & Wang, Z.-L. 2018 Comparison of hybrid models for daily streamflow prediction in a forested basin. *Journal of Hydroinformatics* **20** (1), 191–205.

Liang, Z., Wang, D., Guo, Y., Zhang, Y. & Dai, R. 2011 Application of Bayesian model averaging approach to multimodel ensemble hydrologic forecasting. *Journal of Hydrologic Engineering* **18** (11), 1426–1436.

Liang, Z., Li, Y., Hu, Y., Li, B. & Wang, J. 2017 A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework. *Theoretical and Applied Climatology* **1–2**, 1–13.

Liang, Z., Tang, T., Li, B., Liu, T., Wang, J. & Hu, Y. 2018 Long-term streamflow forecasting using SWAT through the integration of the random forests precipitation generator: case study of Danjiangkou Reservoir. *Hydrology Research* **49** (5), 1513–1527.

Lin, G.-F. & Chen, L.-H. 2004 A non-linear rainfall-runoff model using radial basis function network. *Journal of Hydrology* **289** (1–4), 1–8.

Liu, Y., Sang, Y.-F., Li, X., Hu, J. & Liang, K. 2017 Long-term streamflow forecasting based on relevance vector machine model. *Water* **9** (1), 9.

Liu, Y., Ye, L., Qin, H., Hong, X., Ye, J. & Yin, X. G. 2018 Monthly streamflow forecasting based on hidden Markov model and Gaussian mixture regression. *Journal of Hydrology* **561**, 146–159.

Lu, X., Sang, Y.-F., Li, X., Hu, J. & Liang, K. 2018 Daily pan evaporation modeling from local and cross-station data using three tree-based machine learning models. *Journal of Hydrology* **566**, 668–684.

Ma, M., Yan, R. & Cai, W. 2017 An extended STIRPAT model-based methodology for evaluating the driving forces affecting carbon emissions in existing public building sector: evidence from China in 2000–2015. *Natural Hazards* **89** (2), 741–756.

Martinez, G. F. & Gupta, H. V. 2010 Toward improved identification of hydrological models: a diagnostic evaluation of the 'abcd' monthly water balance model for the conterminous United States. *Water Resources Research* **46**, 8.

Mosavi, A., Ozturk, P. & Chau, K.-w. 2018 Flood prediction using machine learning models: literature review. *Water* **10** (11), 1536.

Schepen, A. & Wang, Q. 2015 Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia. *Water Resources Research* **51** (3), 1797–1812.

Schepen, A., Zhao, T., Wang, Q., Zhou, S. & Feikema, P. 2016 Optimising seasonal streamflow forecast lead time for operational decision making in Australia. *Hydrology and Earth System Sciences* **20** (10), 4117–4128.

Seewald, A. K. 2002 How to make stacking better and faster while also taking care of an unknown weakness. In: *Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pp. 554–561.

Seo, Y., Kim, S. & Singh, V. 2018 Machine learning models coupled with variational mode decomposition: a new approach for modeling daily rainfall-runoff. *Atmosphere* **9** (7), 251.

Shortridge, J. E., Guikema, S. D. & Zaitchik, B. F. 2016 Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences* **20** (7), 2611–2628.

Sikora, R. 2015 A modified stacking ensemble machine learning algorithm using genetic algorithms. In: *Handbook of Research on Organizational Transformations through Big Data Analytics*, IGi Global, Hershey, PA, pp. 43–53.

Singh, K. P., Gupta, S. & Mohan, D. 2014 Evaluating influences of seasonal variations and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning approaches. *Journal of Hydrology* **511** (4), 254–266.

Šípek, V. & Daňhelka, J. 2015 Modification of input datasets for the Ensemble Streamflow Prediction based on large-scale climatic indices and weather generator. *Journal of Hydrology* **528**, 720–733.

Suchetana, B., Rajagopalan, B. & Silverstein, J. 2019 Investigating regime shifts and the factors controlling total inorganic nitrogen concentrations in treated wastewater using non-homogeneous Hidden Markov and multinomial logistic

regression models. *Science of The Total Environment* **646**, 625–633.

Sun, W. & Trevor, B. 2018 A stacking ensemble learning framework for annual river ice breakup dates. *Journal of Hydrology* **561**, 636–650.

Ting, K. M. & Witten, I. H. 1999 Issues in stacked generalization. *Journal of Artificial Intelligence Research* **10**, 271–289.

Vojinovic, Z., Kecman, V. & Babovic, V. 2003 Hybrid approach for modeling wet weather response in wastewater systems. *Journal of Water Resources Planning and Management* **129** (6), 511–521.

Wang, X. & Babovic, V. 2016 Application of hybrid Kalman filter for improving water level forecast. *Journal of Hydroinformatics* **18** (5), 773–790.

Wang, Q., Pagano, T., Zhou, S., Hapuarachchi, H., Zhang, L. & Robertson, D. 2011 Monthly versus daily water balance models in simulating monthly runoff. *Journal of Hydrology* **404** (3–4), 166–175.

Wang, H., Wei, M., Li, G., Zhou, S. & Zeng, Q. 2013 Analysis of precipitable water vapor from GPS measurements in Chengdu region: distribution and evolution characteristics in autumn. *Advances in Space Research* **52** (4), 656–667.

Wolpert, D. H. & Macready, W. G. 1999 An efficient method to estimate bagging's generalization error. *Machine Learning* **35** (1), 41–55.

Xiong, L., Yang, H., Zeng, L. & Xu, C.-Y. 2018 Evaluating consistency between the remotely sensed soil moisture and the hydrological model-simulated soil moisture in the Qujiang catchment of China. *Water* **10** (3), 291.

Xu, B., Yao, H., Zhong, P.-A., Chen, J., Fu, J., Guo, L. & Deng, X. 2018 Exploration and attribution of synergistic gains from joint optimal operation of downstream Jinsha River cascade and Three Gorges cascade reservoirs for hydropower generation. *Journal of Hydroinformatics* **20** (5), 1042–1057.

Yang, Z., Yang, K., Su, L. & Hu, H. 2019 The multi-objective operation for cascade reservoirs using MMOSFLA with emphasis on power generation and ecological benefit. *Journal of Hydroinformatics* **21** (2), 257–278.

Yaseen, Z. M., Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J. & El-Shafie, A. 2016 Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. *Journal of Hydrology* **542**, 603–614.

Ye, X. & Wu, Z. 2018 Contrasting impacts of ENSO on the interannual variations of summer runoff between the upper and mid-lower reaches of the Yangtze River. *Atmosphere* **9** (12), 478.

Zhai, B. & Chen, J. 2018 Development of a stacked ensemble model for forecasting and analyzing daily average PM 2.5 concentrations in Beijing, China. *Science of The Total Environment* **635**, 644–658.

Zhou, H., Tang, G., Li, N., Wang, F., Wang, Y. & Jian, D. 2011 Evaluation of precipitation forecasts from NOAA global forecast system in hydropower operation. *Journal of Hydroinformatics* **13** (1), 81–95.

First received 23 March 2019; accepted in revised form 22 October 2019. Available online 7 November 2019