Medium- and long-term runoff forecasting based on a random forest regression model

Chen Shijun, Wei Qin, Zhu Yanmei, Ma Guangwen, Han Xiaoyan and Wang Liang

ABSTRACT

Medium- and long-term runoff forecasting is closely related to the generation capacity forecasting of cascade hydropower stations, which is of great significance to power plants when arranging production plans and assisting market decisions. In order to improve the accuracy of runoff forecasting, an attempt was made to use random forest regression (RFR) to model the medium- and long-term runoff forecasting and to further make a verification based on the actual monthly runoff data of Mupo and Chuntangba stations. By comparison with the forecast results attained through a support vector machine (SVM) and an integrated autoregressive moving average model (IARMA), the results showed that the RFR model had the lowest mean square error (MSE) among the three methods. In addition, the coefficients of determination \( R^2 \) of the RFR for the two stations increased by 0.0261 and 0.0295 compared with the SVM model, and the \( R^2 \) rose by 0.1134 and 0.1332 compared with the IARMA model. The comparison of the three methods showed that the RFR had higher forecasting accuracy as well as stronger reliability and practicability than the IARMA model and the SVM model, so the RFR provided a new idea and method for the study of runoff forecasting.

Key words | Bagging algorithm, classification and regression tree, mean square error, random forest regression model, runoff forecasting

HIGHLIGHTS

- A random forest regression model for runoff forecasting is established.
- The actual runoff data of Mupo station and Chuntangba station are used for runoff forecasting.
- SVM model and IARMA model are used for comparison.

INTRODUCTION

Runoff forecasting is always a difficulty in the research field of the utilization of hydrology and water resources. Accurate and reliable runoff forecasting is considered a data foundation for major decisions in terms of various aspects, for example, water resource planning and distribution, flood control, disaster reduction, and drought relief (Xing et al. 2016). The formations and the changes of medium- and long-term runoff, as complex hydrological phenomena, are affected by many factors. Thus, the exploration of a model for runoff forecasting with high accuracy and strong applicability has theoretical and practical significance. At present, there are many methods for medium- and long-term runoff forecasting. Traditional statistical models include the time series analysis model (Yurekli et al. 2005; Wang et al. 2011), wavelet analysis model (Guo 2007), and multiple linear regression model (Liu et al. 2018). Due to the emergence of big data mining and artificial intelligence, some new machine learning methods have been extensively applied in runoff forecasting. These methods mainly involved the artificial neural network (ANN) model (Kísi
2008; Taormina et al. 2012; Wang 2017), support vector machine (SVM) model (Behzad et al. 2009; Zhang et al. 2010; Samsudin et al. 2011), and nearest neighbor bootstrap regressive (NNBR) model (Li et al. 2015; Ye et al. 2017). As a new learning method with strong generalization ability, the random forest (RF) model is widely used in the field of machine learning due to its favorable resistance to overfitting and its strong robustness (Zhen et al. 2015; Liu et al. 2019). At present, RF is extensively applied in feature recognition, classification, and optimization, while it is scarcely used in medium- and long-term runoff forecasting (Xu et al. 2016; Xie et al. 2019). Therefore, RF is expected to improve the accuracy of runoff forecasting based on the model, which will hopefully provide a new method and idea for exploring medium- and long-term runoff forecasting.

**BASIC PRINCIPLE OF RF**

RF is an ensemble learning method proposed by Breiman based on the proposed Bagging algorithm and a classification and regression tree (CART) (Breiman 2001).

**Bagging algorithm**

**Implementation steps**

The Bagging (Bootstrap aggregating) algorithm is a famous ensemble learning algorithm proposed by Leo Breiman in 1996 (Breiman 1996). The algorithm aims to integrate multiple weak models to average errors and further strengthen the generalization ability and accuracy of a model. The Bagging algorithm is performed according to the following steps:

1. For an initial sample set with \( n \) samples, a new sample set is attained by conducting \( n \) random samplings with replacement.
2. \( K \) new sample sets corresponding to \( K \) weak models are obtained by repeating Step 1 \( K \) times.
3. \( K \) weak models are trained to acquire \( K \) output values for the models. Then the final output result of the models is attained based on voting.

**Estimation of generalization error**

It is supposed that each sample \((x, y)\) in the dataset \(\Gamma\) is attained through independent sampling with the general probability distribution \(P\); it is further supposed that a single forecasting model is expressed as \(\Phi(x, \Gamma)\) and the expected forecast value of a model is \(\Phi_A(x) = E_\Gamma \Phi(x, \Gamma)\), which represents the mean of the forecast values of different single models for a specific sample \(x\). Hence, the expected generalization error of a model is attained, shown as follows:

\[
E_\Gamma (y - \Phi(x, \Gamma))^2 = E_\Gamma ((y - \Phi_A(x) + \Phi_A(x) - \Phi(x, \Gamma))^2)
\]

\[
= E_\Gamma ((\Phi(x, \Gamma) - \Phi_A(x))^2) + E_\Gamma ((\Phi_A(x) - y)^2)
\]

\[
+ E_\Gamma (2(\Phi(x, \Gamma) - \Phi_A(x))(\Phi_A(x) - y))
\]

\[
= E_\Gamma ((\Phi(x, \Gamma) - \Phi_A(x))^2) + E_\Gamma ((\Phi_A(x) - y)^2)
\]

(1)

where \(E_\Gamma ((\Phi(x, \Gamma) - \Phi_A(x))^2)\) and \(E_\Gamma ((\Phi_A(x) - y)^2)\) refer to the variance and bias of a model, respectively.

The voting mode is simplified to calculate the mean of various weak models in the Bagging method; that is, various weak models have the same weight. Therefore, the expected forecast value of the entire model is expressed as follows:

\[
E_\Gamma \left(\frac{1}{K} \sum_{k=1}^{K} \Phi(x, \Gamma_k)\right) = \frac{1}{K} E_\Gamma \left(\sum_{k=1}^{K} \Phi(x, \Gamma_k)\right) = \mu,
\]

(2)

where \(\mu\) stands for the expected forecast value of a single weak model. The expected forecast value of the entire model approximates to that of a single model; that is, the bias of the former is approximated to that of the latter. Thus, it is common to select a learning method with a low bias as a weak model. Because various sub-sample sets in Bagging are attained through sampling with replacement and duplicate samples are likely to occur, a correlation coefficient \(\rho\) (0 < \(\rho\) < 1) is employed to characterize the correlation between weak models corresponding to various sub-sample sets. Hence, the variance of a model can be obtained:

\[
\text{Var}_\Gamma \left(\frac{1}{K} \sum_{k=1}^{K} \Phi(x, \Gamma_k)\right) = \frac{\sigma^2}{K} + \frac{K-1}{K} \rho \sigma^2,
\]

(3)
where $\sigma^2$ represents the variance of a single weak model. With increasing weak models, the first and second terms in Equation (3) converge to 0 and $\rho \sigma^2$, respectively. Thus, the Bagging method decreases the variance of a weak model, which avoids the overfitting phenomenon and theoretically improves the generalization ability of the model to some extent.

**CART**

As a basic unit for an RF model, a CART is essentially used as a binary decision tree. For an input feature space $T$, the optimal partitioning feature $t(t \in T)$ and partitioning point $j(j \in J_t)$ are found at each node of the CART based on the partition principle with minimum error, where $J_t$ stands for the data range of the feature $t$ and the objective of the regression and decision tree; that is, the minimum error of the node partition can be expressed as follows:

$$\min_{t_j} \sum_{x_t \in R_{m}} (y_t - f(x_t))^2$$

(4)

where $R_m$ represents the $m$th feature space after partition, which can be expressed as $R_l$ and $R_r$ in a binary tree-CART; $f(x_t) = \text{ave}(y_j | x_t \in R_m)$ stands for the forecast value of a model, equivalent to the mean in various spaces. Hence, the objective of a node partition based on a CART is shown as follows:

$$\min_{t_j} \left[ \sum_{x_t \in R_l(t,j)} (y_t - f(x_t))^2 + \sum_{x_t \in R_r(t,j)} (y_t - f(x_t))^2 \right]$$

(5)

The objective is generally solved by using the traversing method. The range of the feature selection at each node of a CART in RF also learns from the Bagging method. By setting the maximum number $m$ of features, it is feasible to stochastically sample $m$ features from the general features to perform partitioning during each selection. The CART can be constantly split based on the method. To get rid of the overfitting of the CART, the stopping conditions are set, which include the maximum number of leaf nodes, the maximum depth of the tree, and the minimum sample size of the leaf nodes. Due to the application of the Bagging method in RF, the CART can constantly grow and thereby reduce the bias of the models while Bagging decreases the variance of the models.

**ESTABLISHMENT OF THE MODEL**

**Modeling for medium-term and long-term runoff forecasting**

Medium- and long-term runoff is a continuous process. The runoffs of the first three months and the runoffs in the same month for the first three years (a total of six-dimensional input features) could be set beforehand as the forecasting inputs of the model, while the output of the model was the predicted monthly runoff. Hence, the inputs and the output of the model were constructed as shown in Figure 1.

**Calibration of model parameters**

A new sample set was generated through sampling with replacement based on the Bagging method in RF. Hence, in terms of the probability, the probability that each sample was not drawn was expressed as follows:

$$\lim_{m \to \infty} \left( 1 - \frac{1}{m} \right)^m = \frac{1}{e} \approx 36.79\%$$

(6)

where $m$ refers to the number of samples; that is, in terms of probability, the samples took up 36.79% of all samples that could not be drawn when forming a new sample set. This
part of the samples referred to out-of-bag (OOB) samples. By calculating the mean of the results of all of the OOB samples in the corresponding CART as the model output of these OOB samples, the model error of the OOB samples (i.e., OOB error) was attained. It was verified that the OOB error corresponded to the unbiased estimation of the generalization error of the model (Lv & Feng 2019). Therefore, the model parameters were calibrated by utilizing the OOB error. The specific steps are described as follows. The range of the model parameters was set to traverse all combinations. Then the optimal parameter combination was selected by taking the OOB error as the evaluation index.

### Evaluation method for the results of the model

The evaluation of the results of the model was intended to assess the accuracy of the model for the test sets. The evaluation indices for the accuracy of regression prediction generally involve the mean square error (MSE), mean absolute error (MAE), and coefficient of determination \( R^2 \). The MSE and \( R^2 \) were chosen as the evaluation indices, which were separately calculated as follows:

\[
\text{MSE} = \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2, \tag{7}
\]

\[
R^2 = 1 - \frac{\sum_{n=1}^{N} (\hat{y}_n - y_n)^2}{\sum_{n=1}^{N} (y_n - \bar{y}_n)^2}. \tag{8}
\]

### CASE VERIFICATION AND ANALYSIS

To verify the feasibility and practicability of the model, the data for the monthly runoffs at Mupo and Chuntangba stations (Xiaojin County, Sichuan, China; shown in Figure 2) over a period of 60 years from 1959 to 2018 were applied. At first, the original samples were processed by taking the runoff data in the same month in the first three years and the data for the first three months selected beforehand as the input values. Afterwards, by taking January 2015 as a boundary, the processed samples were divided into a training set and a test set. That is, monthly runoffs from January 1959 to December 2014 were used as the training samples while those for the period of January 2015–December 2018 were taken as the test samples. The test was carried out by utilizing the random forest regression (RFR) model. Because the SVM model was a mature tool for predicting factors including the power, load, electricity price, and runoff, an SVM was also selected to perform tests for comparison with the RFR model (Peng et al. 2013; Wang et al. 2014, 2015). Additionally, a traditional integrated autoregressive moving average (IARMA) model was also selected for comparison. The training and test samples of the two models were the same as those of the RFR model. The calculation of the three methods was achieved with Python programming.

By taking the OOB error as the evaluation index, the RFR model traversed the parameter combinations to attain the optimal combination and its corresponding OOB error, as shown in Table 1.

As a classical regression prediction model, the SVM model has many methods for parameter selection. The parameter selection results of the model for the Mupo and Chuntangba stations were separately attained by applying commonly used cross-validation, as shown in Table 2 (Ji et al. 2014).

The parameters of the IARMA model were mainly obtained by analyzing the stationarity, autocorrelation coefficient, and partial correlation coefficient of the sample series (Yang 2004). As a traditional and classical statistical analysis model, the model is extensively used and the calibration process of its parameters is complex. Thus, the parameters of the model were offered directly, as shown in Table 3.

The forecast results of the above models based on the test set are shown in Figures 3 and 4. The data statistics for the forecast results of various models are displayed in Table 4.

According to the above results, it could be seen that the runoff forecast results of Mupo station based on the three models for the test set were all superior to those of Chuntangba station. When predicting the runoff of Mupo station, it could be clearly found that RFR and SVM could more favorably predict the maximum runoff. However, the three models performed poorly when they were applied to predict the maximum runoff of Chuntangba station. As a whole, the RFR model presented a favorable forecasting effect for the test set. The \( R^2 \) of the runoff forecast results...
Figure 2 | The location diagram of Mupo and Chuntangba stations.

### Table 1 | Parameter selection and OOB error of the RFR model

<table>
<thead>
<tr>
<th>Station</th>
<th>Number of CART</th>
<th>Maximum feature number through node partition</th>
<th>OOB error $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mupo station</td>
<td>35</td>
<td>4</td>
<td>0.8454</td>
</tr>
<tr>
<td>Chuntangba station</td>
<td>85</td>
<td>5</td>
<td>0.8574</td>
</tr>
</tbody>
</table>

### Table 2 | Parameter selection based on the SVM model

<table>
<thead>
<tr>
<th>Station</th>
<th>Kernel function</th>
<th>C</th>
<th>gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mupo station</td>
<td>Gaussian kernel</td>
<td>400</td>
<td>1E-5</td>
</tr>
<tr>
<td>Chuntangba station</td>
<td>Gaussian kernel</td>
<td>1,200</td>
<td>1E-5</td>
</tr>
</tbody>
</table>
of the RFR model for Mupo and Chuntangba stations respectively increased by 0.0261 and 0.0295 relative to that by employing the SVM model, while the $R^2$ rose by 0.1134 and 0.1332, respectively, compared with the result obtained based on the IARMA model. However, the MSEs of the runoff forecast results of the RFR model for Mupo and Chuntangba stations were respectively reduced by 24.8 and 32.1 relative to the MSE found by employing the SVM model, while the MSEs were respectively reduced by 107.9 and 145.1 compared with the result obtained based on the IARMA model.

CONCLUSIONS AND PROSPECTS

The RFR model for predicting medium- and long-term runoffs was established and validated based on the measured data of the runoffs of two stations. The results showed that the RFR model exhibited higher forecasting accuracy compared with the SVM and IARMA models, which verified the feasibility and reliability of the model in the application of medium- and long-term runoff forecasting. The RFR model provided a new method for investigating medium- and long-term runoff forecasting.

Because the formation of runoffs is affected by multiple factors, it is difficult to accurately predict runoffs in the future by depending only on historical data for runoffs. Therefore, the accuracy of the runoff forecast results at Mupo and Chuntangba stations remains to be improved. Multiple types of variables can be input in the RF model, which is a tree model, and the runoff forecasting effect of the model on more complex input conditions remains to be further discussed.

ACKNOWLEDGEMENT

The work was supported by National Key Research and Development Project (2018YFB0905204).

DATA AVAILABILITY STATEMENT

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


Guo, H. 2007 Model for Medium and Long Term Runoff Forecasting Based on Wavelet Analysis. Zhengzhou University, Zhengzhou, China.


Xie, W., Sun, Y. & Huang, Y. 2019 Research on power line communication optimization algorithm based on improved random forest. Power System Protection and Control 47 (11), 22–29.


First received 6 July 2020; accepted in revised form 24 August 2020. Available online 21 September 2020.