# Data-driven runoff forecasting for Minjiang River: a case study

Yuqiang Wu, Qinhui Wang, Ge Li and Jidong Li

## ABSTRACT

Long-term runoff forecasting has the characteristics of a long forecast period, which can be widely applied in environmental protection, hydropower operation, flood prevention and waterlogging management, water transport management, and optimal allocation of water resources. Many models and methods are currently used for runoff prediction, and data-driven models for runoff prediction are now mainstream methods, but their prediction accuracy cannot meet the needs of production departments. To this end, the present research starts with this method and, based on a support vector machine (SVM), it introduces ant colony optimization (ACO) to optimize its penalty coefficient *C*, Kernel function parameter *g*, and insensitivity coefficient *p*, to construct a data-driven ACO-SVM model. The validity of the method is confirmed by taking the Minjiang River Basin as an example. The results show that the runoff predicted by use of ACO-SVM is more accurate than that of the default parameter SVM and the Bayesian method.

**Key words** | ACO, data-driven model, Mingjiang River, runoff forecasting, SVM

**Yuqiang Wu**
**Qinhui Wang**
**Jidong Li** (corresponding author)
College of Water Conservancy and Hydropower Engineering,
Sichuan Agricultural University,
Ya'an 625014, Sichuan,
China
E-mail: *survivers@126.com*

**Ge Li**
Consulting Hydrologist, Chongqing Tongwang Water Conservancy and Hydropower Engineering Design Ltd,
Chongqing 401120,
China

## HIGHLIGHTS

- Grid-point precipitation data are used for runoff prediction.
- Support vector machine parameters are optimized.
- A new ACO-SVM coupling model is established.

## INTRODUCTION

River runoff is an important component of the hydrological cycle of a riverine basin, and runoff prediction is an important part of the hydrological system prediction. The long-term runoff prediction results are affected by many factors such as climatic conditions, anthropogenic inputs, soil, loam texture, vegetation coverage, and topography. At present, many models are used in the study of runoff prediction, which are roughly classified as process-driven models and data-driven models (Zhang *et al.* 2018). The set of factors that affect runoff is difficult to determine and there is a lack of relevant data, so it remains difficult to predict runoff using a process-driven model, and data-driven models are widely used due to their operability. Commonly used data-driven methods include: support vector machines (SVMs), fuzzy analysis, grey system analysis, artificial neural networks, and wavelet analysis (Wu 2010).

The support vector machine (SVM) (Wu 1999; Zhang *et al.* 2009a) has been gradually applied in water resources research for runoff forecasting (Cheng *et al.* 2015), because of its reliable global optimum nature and good generalization ability; however, when the kernel function is selected, the runoff forecasting value of the default parameter SVM differs from the measured value, so it is necessary to optimize the parameters of SVM when used for runoff forecasting (Wang 2015).

Ant colony optimization (ACO) (Dirk & Christian 2012) is a heuristic biological evolutionary algorithm proposed by the Italian scholar Dorigo in the 1990s. Its inspiration comes

doi: 10.2166/ws.2020.134

from the study of the collective foraging behavior of ants in their natural habitat. The ACO algorithm adopts a random search strategy based on population evolution, which is characterized by parallelism, robustness, and global search; the solution time is short and it is easy to implement via computer (Zhu & Li 2016). In recent years, it has solved many classical optimization problems such as the travelling salesman problem (TSP) (Dorigo *et al.* 1996) and secondary allocation tasks (Maniezzo & Colorni 1999).

In the present work, ACO is introduced to optimize some parameters of an SVM, and the ACO-SVM coupling model is established. The ACO-SVM coupling model is applied to the monthly runoff forecasting of the upper reaches of Minjiang River. Compared with the monthly runoff forecasting value predicted by a default-parameter SVM and Bayesian Statistical Forecasting Theory (BSFT), the feasibility of ACO of SVM parameters for the monthly runoff forecasting of the Minjiang River is verified.

## METHODOLOGY

### SVM regression

SVM regression is similar to a BP neural network. The model is trained through samples in advance, and then, for the trained model prediction, given the input data, the corresponding prediction output can be obtained. Let there be a sample set $\{(x_i, y_i), i = 1, 2, \cdots, l\}$, in which $x_i \in \mathrm{R}^n$ is the input value of the $i^{\mathrm{th}}$ learning sample, which is also an $n$-dimensional column vector. The corresponding target value is $x_i = [x_i^1, x_i^2, \cdots, x_i^n]^{\mathrm{T}}$, $y_i \in \mathrm{R}$ is the corresponding output value, and $l$ is the number of samples. The essence of the problem is to use SVM to find a curve fitting for the sample point set, requiring the curve to be as flat as possible (Figure 1). In the case of linear data sets, the decision function $f$ is assumed to be in the following form:

$$\mathrm{f}(x) = w^T x + \mathrm{b} \tag{1}$$

where $w^{\mathrm{T}}x$ represents the inner product of vectors, $x \in \mathrm{R}^n$, $w \in \mathrm{R}^n$ is the weight vector; $b \in \mathrm{R}^n$ represents the offset phase.

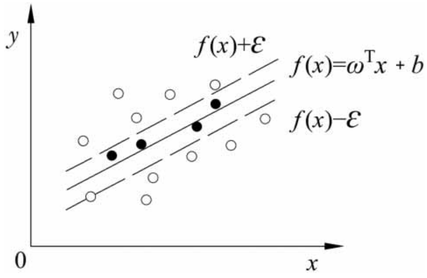

**Figure 1** | Support vector regression model.

With the introduction of $\varepsilon$-insensitive loss function by Vapink, the use of the SVM has been extended and applied to non-linear regression estimation and curve fitting; that is, support vector regression (SVR). Like the classification problem, the kernel function method is introduced to transform the non-linearity of the input sample space into a high-dimensional linear feature space, where the linear method is adopted to solve the non-linearity problem (Vafakhah & Khosrobeigi 2019). In non-linear SVMs, the Gaussian radial basis function is a commonly used kernel function:

$$K(\mathrm{x_i}, \mathrm{x_j}) = \exp(-\gamma \|\mathrm{x_i} - \mathrm{x_j}\|^2), \quad \gamma > 0 \tag{2}$$

By searching for the optimum $w$ and $b$, the optimization problem obtained by minimizing the confidence interval under the condition that formula (1) remains unchanged is as follows:

$$\begin{cases} \min\limits_{\mathrm{w}} \dfrac{1}{2} w^{\mathrm{T}} w \\ \text{s.t.} \quad y_i - f(x_i) \le \varepsilon \\ \qquad\; f(x_i) - y_i \le \varepsilon \end{cases}; \tag{3}$$

where $\varepsilon$ is the error.

When the constraint condition cannot be realized, the optimization problem is transformed into the following framework by introducing slack variables $\xi_i$ and $\xi_i^*$:

$$\begin{cases} \min\limits_{\mathrm{w},\xi,\xi^*} = \dfrac{1}{2} w^{\mathrm{T}} w + C \sum\limits_{i=1}^{n} (\xi_i + \xi_i^*) \\ y_i - f(x_i) \le \varepsilon + \xi_i \\ f(x_i) - y_i \le \varepsilon + \xi_i^* \end{cases} \tag{4}$$

where $\xi_i \ge 0$, $\xi_i^* \ge 0$, $i = 1, 2, \cdots, n$.

Using the Lagrange multiplier method to solve convex quadratic programming problems, the results are as follows:

$$\begin{cases} \min\limits_{a,a^*} \dfrac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}(a_i-a_i^*)(a_j-a_j^*)K(x_i,x_j)+\varepsilon\sum\limits_{i=1}^{n}y_i(a_i-a_i^*) \\ s.t. \quad \sum\limits_{i=1}^{n}(a_i-a_i^*)=0 \\ \quad 0\le a_i\le C \\ \quad 0\le a_i^*\le C \end{cases} \tag{5}$$

The regression estimation function obtained by learning is used for non-linear classification or regression prediction. The regression estimation function is as follows:

$$\begin{cases} f(x)=\sum(a_i-a_i^*)K(x_i,x_j)+b \\ b=\dfrac{1}{N_{SV}}\left\{\sum\limits_{0<a_i<C}\left[y_i-\sum\limits_{x_i\in SV}(a_j-a_j^*)K(x_i,x_j)-\varepsilon\right]+\sum\limits_{0<a_i^*}\left[y_i-\sum\limits_{x_i\in SV}(a_j-a_j^*)K(x_i,x_j)+\varepsilon\right]\right\} \end{cases} \tag{6}$$

where $N_{SV}$ represents the number of standard support vectors, $a_i^*$ and $a_j^*$ are the Lagrange multipliers, $C$ is the penalty coefficient, and $K(\mathrm{x_i},\mathrm{x_j})$ denotes the kernel function.

## The principle of ACO

Assume that $Z=\{c_1, c_2, \cdots, c_n\}$ is a set of $n$ cities. $L=\{l_{uv}|c_u, c_v\subset Z\}$ is the set of two-connected elements (cities) in $Z$. $d_{uv}$ $(u,v=1, 2, \cdots, n)$ is the Euclidean distance of $l_{uv}$. $G(Z, L)$ represents a directed graph. The goal of ACO is to find the shortest length of the tour path in $G$. The optimization in an ACO algorithm is implemented on the directed graph using three criteria (transition probability criterion, local adjustment criterion, and global pheromone adjustment criterion).

When ant $k$ comes to city $u$ at time $t$, the ant will calculate a transition probability according to the number of pheromones and the heuristic information of the path from city $u$ to all of the next cities. $P_{uv}(t)$ is used to describe the probability of ant $k$ transferring from element $u$ to element $v$ at time $t$, then:

$$P_{uv}^k(t)=\begin{cases} \dfrac{[\mathrm{\tau}_{uv}(\mathrm{t})]^\alpha[\mathrm{\eta}_{uk}(\mathrm{t})]^\beta}{\sum_{\mathrm{N}\subset allowed_\mathrm{k}}[\mathrm{\tau}_{ur}(\mathrm{t})]^\alpha[\tau_{ur}(\mathrm{t})]^\beta}, & v\in allowed_k, \quad \mathrm{r}\subset allowed_k \\ 0, & others \end{cases} \tag{7}$$

where $allowed_k$ indicates all the cities that ant $k$ will choose next; $\alpha$ is the information heuristic factor; and $\tau_{uv}$ is the path pheromone from city $u$ to city $v$. $\beta$ represents the relative importance of visibility. $\eta_{uv}$ is an elicitation function; that is, $\eta_{uv}(t)=1/d_{uv}$, and $d_{uv}$ represents the path length from city $u$ to city $v$.

When the ant chooses a path, it needs to leave pheromone information to guide subsequent ants to optimum effect. The global pheromone adjustment rules are expressed as follows:

$$\begin{aligned} \tau_{uv}(t+n)&=(1-\rho)\tau_{uv}(t)+\Delta\tau_{uv}(t) \\ \Delta\tau_{uv}(t)&=\sum_{k=1}^{n}\Delta\tau_{uv}^k(t) \end{aligned} \tag{8}$$

where $\rho$ is the volatilization coefficient; $\rho\subset[0, 1]$; $(1-\rho)$ represents the pheromone residual factor. $\Delta\tau_{\mathrm{uv}}(t)$ is the increment of the number of pheromones on the path $(u, v)$ in this cycle, and $\Delta\tau_{\mathrm{uv}}(t)$ represents the amount of information left on the path $(u,v)$ by the $k^{\mathrm{th}}$ ant in this cycle.

An ant cycle model is adopted for the calculation of $\Delta\tau_{\mathrm{uv}}^k(t)$, as defined by:

$$\Delta\tau_{uv}^k(t)=\begin{cases} \dfrac{Q}{\mathrm{L_k}}, \mathit{If\ ant\ k\ passes\ the\ path}(u, v)\ \mathit{in\ this\ cycle} \\ 0, \ other \end{cases} \tag{9}$$

where $Q$ denotes the pheromone intensity and $\mathrm{L_k}$ denotes the total length of the path travelled by ant $k$ in this cycle.

### Establishment of the model

By using pattern recognition and regression toolbox LIBSVM-3.23, ACO-SVM, a medium and long-term runoff forecasting model of the upper reaches of Minjiang River based on ACO and an SVM, is established in MATLAB™ 2018a. The modelling steps are as follows:

(1) Normalization of training samples and prediction samples is carried out over a normalized range of [0, 1] such that:

$$y = \frac{(y_{max} - y_{min})(x - x_{min})}{(x_{max} - x_{min})} + y_{min} \tag{10}$$

where $x$, $y$ are the sample data before and after processing; $x_{\max}$, $x_{\min}$, $y_{\max}$, and $y_{\min}$ represent the maximum and minimum values of the corresponding processed data.

(2) Setting ACO parameters and inputting training set samples, multiple groups of the main parameters $c$, $g$, and $P$ of the LIBSVM learning model are obtained.

The main ACO parameters are: the number of ants $k = 50$, time $t = 100$, information volatilization coefficient $r = 0.9$, transition probability constant $p_0 = 0.2$, upper bound on the penalty coefficient $C \in (0.1, 100)$, Kernel function parameter $g \in (0.01, 10)$, and insensitive loss coefficient $p \in (0, 0.05)$.

(3) Select the type of SVM and kernel functions in the training model of LIBSVM, train the data on the training samples, establish the learning model, use the established learning model to predict the training set, analyze its relative error and qualified rate, and determine the optimal parameters $C$, $g$, and $P$ (Figure 2).

The parameters of the training model in LIBSVM include: SVM type, kernel function type, set value of kernel function, and so on. To facilitate the calculation and improve the prediction accuracy of model, the SVM type in the ACO-SVM training model parameters chosen in this work is a multi-level classification and regression SVM, the kernel function type is the Gaussian radial basis function, and other parameters in the training model are all set to their default values (Wang *et al.* 2018).

(4) According to the learning model established by the optimal parameters $c$, $g$, and $P$, the prediction sample set is regressed and predicted.

## CASE STUDY

### Watershed overview

The Minjiang River is the largest tributary of the Yangtze River and the basis of the ecological environment of the Chengdu Plain. The average annual precipitation of the downstream reaches 1,100 to 1,600 mm, but the average annual precipitation in the upstream is only 400 to 700 mm. Meanwhile, the flood season of Minjiang River Basin is concentrated between June and September, and the rainfall in summer and autumn can account for more than 80% of the annual total rainfall (Du & He 2015). The uneven distribution of precipitation in the year brings significant challenges to the operation of reservoirs in the basin. Therefore, the monthly runoff forecast of the upper reaches of the Minjiang River has important practical value for analyzing the joint optimization of cascade reservoirs, improving the utilization efficiency of hydropower resources, and improving the local environment (Li *et al.* 2016).

### The selection of a predictor

Influenced by many factors, such as topography, geology, vegetation, meteorological conditions, and human activities, there are often complex non-linear relationships among 'rainfall-runoff', 'pre-runoff-future runoff', and 'upstream runoff-downstream runoff' (Zhao & Yang 2011). In this monthly runoff forecasting of Minjiang River Basin, precipitation data are difficult to obtain and are affected by the underlying surface due to a small amount of precipitation in the predicted month, and the abundant precipitation during the wet season, so the confluence time increases, which results in little effect on the current monthly runoff but a greater effect on the next monthly runoff, the effect of the precipitation in the previous month should be considered without considering that of the current month. In addition, due to the underlying surface, the runoff of last month will affect the predicted monthly runoff, so the effect of the runoff of previous month should be considered. In summary, the precipitation and runoff from the previous month in the forecasted month are selected as the predictors for this runoff forecasting model.

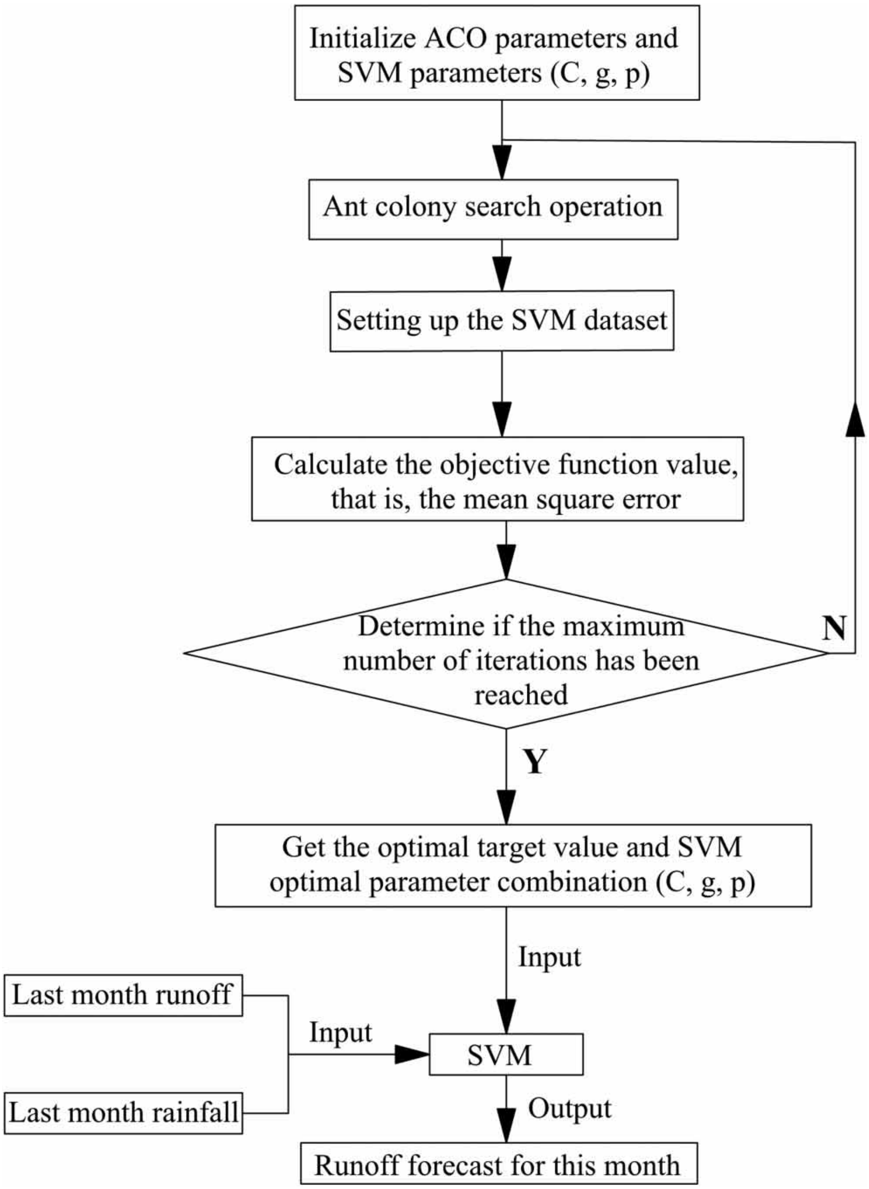
Initialize ACO parameters and SVM parameters (C, g, p)
Ant colony search operation
Setting up the SVM dataset
Calculate the objective function value, that is, the mean square error
Determine if the maximum number of iterations has been reached
N
Y
Get the optimal target value and SVM optimal parameter combination (C, g, p)
Input
Last month runoff
Input
SVM
Last month rainfall
Output
Runoff forecast for this month

**Figure 2** | Research process diagram.

## Data analysis and selection

To eliminate the effect of the upstream reservoir and power station operation, the measured runoff data are restored and we finally obtain the natural runoff data used in this research. The annual distribution of the runoff series and precipitation series showed consistent seasonal characteristics (Figure 3), combined with the comprehensive analysis of local climate characteristics, it can be considered that the seasonality of the runoff series formed naturally. Taking the inflow of the dam site of Zipingpu Reservoir as the target value, the monthly precipitation (hereinafter referred to as 'station precipitation') data of five meteorological stations (Hongyuan, Songpan, Dujiangyan, Xiaojin, and Barkam) in the upper reaches of Minjiang River in each month from 1967 to 2011 and natural runoff are taken as the predictors (Figure 4). The SVM-3 default parameters are adopted to predict monthly runoff. We average the precipitation recorded at five stations during each month over the years to obtain the mean surface

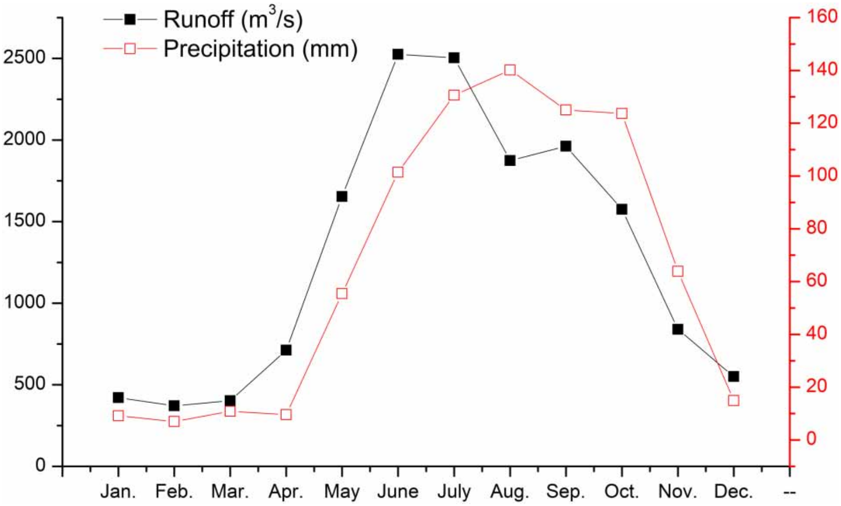

**Figure 3** | Annual distribution of precipitation and runoff.

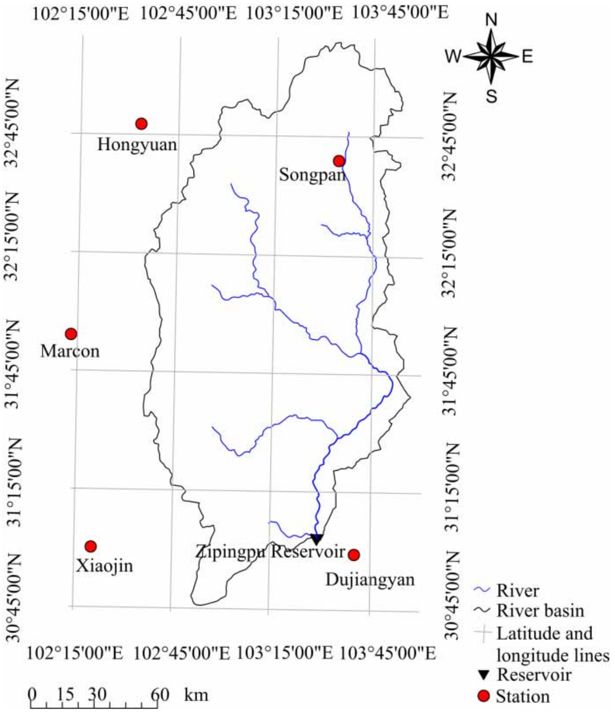

**Figure 4** | Map showing the five meteorological stations.

precipitation. The Pearson's correlation coefficient between monthly inflow of this month and monthly mean surface precipitation of last month is calculated, and the correlation coefficient between monthly inflow of this month and monthly precipitation of last month of each station over the years is derived. The average correlation coefficient of precipitation at each of the five stations is taken as the average correlation coefficient (Table 1).

Similarly, the monthly precipitation (hereinafter referred to as 'grid precipitation') data and natural runoff of the 0.5° × 0.5° grid of the National Meteorological Information Center in the upper reaches of Minjiang River at 102°45′00″E–103°45′00″E and 31°15′00″N–32°45′00″N from each month of next year over the years are taken as predictors (Figure 5). The SVM-3 default parameters are also employed to predict monthly runoff. We average the precipitation across 12 grids in each month over the years and take this as the mean surface precipitation. The Pearson's correlation coefficient between monthly inflow of this month and monthly mean surface precipitation of the previous month is calculated, and the Pearson's correlation coefficient between monthly inflow of this month and monthly precipitation of the previous month of each grid over the years is deduced. The average correlation coefficient over all precipitation data from all 12 grids is taken as the average correlation coefficient (Table 2).

Comparing Tables 1 and 2, the average relative error between runoff forecast values and measured values and the mean variance of grid precipitation data are shown to be smaller than that in the corresponding forecast and measured values of data from all five stations: the relative error decreases by 8.2% to 26.5%, and the mean square error (MSE) decreases by 6.8% to 29.2%. The qualified

**Table 1** | Partial forecast results using grid precipitation data

| | Calibration period | July | Aug. | Sep. | Dec. | Jan. | Feb. |
|---|---|---|---|---|---|---|---|
| Five-station Ave. | Min. relative value | 0.0121 | 0.0040 | 0.0112 | 0.0014 | 0.0123 | 0.0031 |
| | Max. relative value | 0.4224 | 0.4764 | 0.4119 | 0.2407 | 0.2048 | 0.1400 |
| | Ave. relative value | 0.1730 | 0.1382 | 0.1671 | 0.0660 | 0.0684 | 0.0572 |
| | Correlation coefficient | 0.4513 | 0.5437 | 0.5484 | 0.9356 | 0.9274 | 0.8971 |
| | MSE | 0.0454 | 0.0331 | 0.0415 | 0.0073 | 0.0062 | 0.0048 |
| | NSE | 0.1924 | 0.2930 | 0.2543 | 0.8452 | 0.8330 | 0.7488 |
| | Qualified rate | 0.6389 | 0.7500 | 0.6667 | 0.9722 | 0.9722 | 1.0000 |
| | **Inspection period** | **July** | **Aug.** | **Sep.** | **Dec.** | **Jan.** | **Feb.** |
| | Min. relative value | 0.0080 | 0.0705 | 0.0221 | 0.0122 | 0.0295 | 0.0597 |
| | Max. relative value | 0.4713 | 0.5013 | 0.2983 | 0.2708 | 0.3449 | 0.4134 |
| | Ave. relative value | 0.1679 | 0.2488 | 0.1624 | 0.1280 | 0.1279 | 0.1412 |
| | Correlation coefficient | 0.4673 | 0.2206 | 0.2633 | 0.9260 | 0.8578 | 0.9645 |
| | MSE | 0.0578 | 0.0826 | 0.0335 | 0.0241 | 0.0251 | 0.0297 |
| | NSE | 0.1073 | 0.0284 | −0.5913 | 0.6321 | 0.6945 | 0.6036 |
| | Qualified rate | 0.7000 | 0.5000 | 0.7000 | 0.8000 | 0.8000 | 0.9000 |

*NSE* is a certainty coefficient, *MSE* denotes the mean square error.

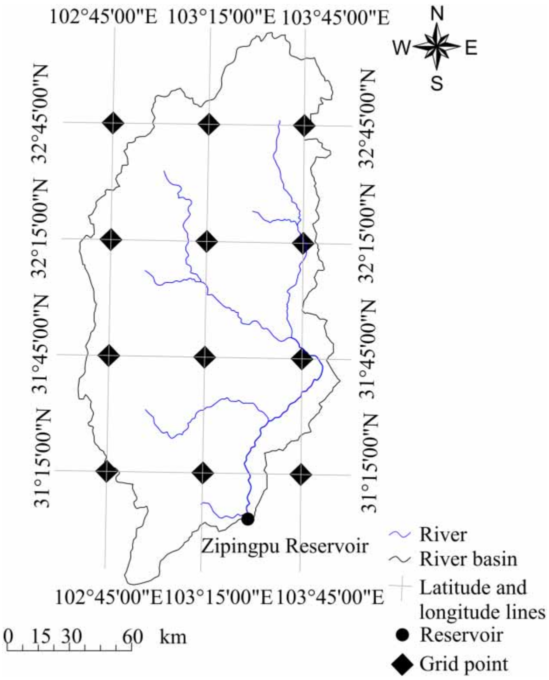

**Figure 5** | Map showing the 12 latitude and longitude grid squares.

rate of the grid data is greater than that of the tabulated station data; therefore, in this runoff forecast, the grid precipitation data for the previous month corresponding to the predicted month are selected as the precipitation in the previous month to forecast the runoff.

## Model parameters

According to the steps outlined in the section Establishment of the model, a moderate function is established with the goal of minimizing the periodic mean square error of the calibration period, and using the ACO algorithm to optimize the three SVM parameters, and the main parameters of ACO-SVM forecasting function model from November to April of next year are obtained (Table 3).

## Selection of training and prediction samples

The measured monthly runoff data from Dujiangyan hydrological station for each November-April of the following year from 1966 to 2011 for 45 years are selected, and the natural runoff data obtained from the original measured data after excluding the influence of upstream reservoir operation by the restoration calculation are taken as the sample set (Chen *et al.* 2009): this treats the first 35 years as the calibration period and the last 10 years as the inspection period.

## Training of prediction models

The parameters of the ACO-SVM model from November to April of the following year are obtained by inputting and fitting the runoff and precipitation data in the calibration period (Table 4).

**Table 2** | Partial forecast result

| | **Calibration period** | **July** | **Aug.** | **Sep.** | **Dec.** | **Jan.** | **Feb.** |
|---|---|---|---|---|---|---|---|
| 12-grid Ave. | Min. relative value | 0.0076 | 0.0184 | 0.0062 | 0.0019 | 0.0027 | 0.0006 |
| | Max. relative value | 0.4190 | 0.5698 | 0.3756 | 0.2441 | 0.1649 | 0.1634 |
| | Ave. relative value | 0.1711 | 0.1422 | 0.1571 | 0.0615 | 0.0627 | 0.0526 |
| | Correlation coefficient | 0.4885 | 0.4991 | 0.6420 | 0.9395 | 0.9335 | 0.8873 |
| | MSE | 0.0449 | 0.0354 | 0.0344 | 0.0069 | 0.0051 | 0.0047 |
| | NSE | 0.2011 | 0.2444 | 0.3813 | 0.8531 | 0.8621 | 0.7507 |
| | Qualified rate | 0.6667 | 0.7222 | 0.6389 | 0.9722 | 1.0000 | 1.0000 |
| | **Inspection period** | **July** | **Aug.** | **Sep.** | **Dec.** | **Jan.** | **Feb.** |
| | Min. relative value | 0.0131 | 0.0438 | 0.0263 | 0.0067 | 0.0150 | 0.0356 |
| | Max. relative value | 0.4905 | 0.5155 | 0.3091 | 0.2540 | 0.3446 | 0.4315 |
| | Ave. relative value | 0.1823 | 0.2194 | 0.1193 | 0.1175 | 0.1315 | 0.1274 |
| | Correlation coefficient | 0.4205 | 0.4264 | 0.4362 | 0.9320 | 0.8356 | 0.9580 |
| | MSE | 0.0628 | 0.0712 | 0.0237 | 0.0208 | 0.0270 | 0.0277 |
| | NSE | 0.0304 | 0.1634 | −0.1259 | 0.6817 | 0.6710 | 0.6304 |
| | Qualified rate | 0.7000 | 0.6000 | 0.8000 | 0.8000 | 0.8000 | 0.9000 |

**Table 3** | SVM parameters optimized by ACO

| **Month/ Parameter** | **Penalty coefficient *C*** | **Kernel function parameter *g*** | **Insensitive loss coefficient *p*** | **MSE** |
|---|---|---|---|---|
| May | 81.0591 | 0.2761 | 0.0330 | 0.0547 |
| June | 26.6547 | 1.4281 | 0.0438 | 0.0629 |
| July | 11.7687 | 0.2197 | 0.0177 | 0.0521 |
| Aug. | 99.3124 | 8.7540 | 0.0469 | 0.0418 |
| Sep. | 3.0301 | 0.1913 | 0.0355 | 0.0447 |
| Oct. | 1.4452 | 6.2699 | 0.0357 | 0.0322 |
| Nov. | 36.4620 | 0.4337 | 0.0185 | 0.0102 |
| Dec. | 11.1414 | 2.7837 | 0.0388 | 0.0068 |
| Jan. | 11.2128 | 0.1777 | 0.0342 | 0.0055 |
| Feb. | 42.0280 | 1.9030 | 0.0406 | 0.0043 |
| Mar. | 38.2115 | 1.0832 | 0.0201 | 0.0120 |
| Apr. | 4.5256 | 8.2652 | 0.0152 | 0.0437 |

## RESULTS AND DISCUSSION

### Forecasting results

The precipitation and runoff data in the inspection period are input into the model to acquire the forecasting results of the runoff in the inspection period. The errors and related statistical parameters are listed in Table 5. To test the optimization effect of ACO on the SVM parameters, and the monthly runoff forecasting effect of the ACO-SVM model, the error and related statistical parameters of the monthly runoff forecast result, which use the default parameter SVM, are compared with the ACO-SVM runoff forecast result, and the forecasting results of SVM default parameters are listed in Table 2.

### Analysis of results

#### Optimization effect

By comparing the data in Tables 2 and 5, it can be seen that the average relative error and mean square error of each month's calibration and inspection period in the dry season have been decreased after ACO is applied to the SVM parameters, and the certainty coefficient and average qualified rate have been improved significantly compared with the predicted values of SVM default parameters.

In the calibration period, the average relative error between the predicted and measured values of the runoff of the ACO-SVM model is reduced by 1.2% to 17.9% compared with the predicted result of the default SVM, and mean square error is reduced by 1.7% to 66.7%, the certainty coefficient NSE is increased by 4.5% to 205.7%, and the qualified rate generally increases, showing that the goodness-of-fit of ACO-SVM to the sample data set is higher than that of the default parameter SVM. During the inspection period, the average relative error decreases by 1.7% to 34.0%.

**Table 4** | ACO-SVM prediction model

| **Parameter/Month** | **May** | **June** | **July** | **Aug.** | **Sep.** | **Oct.** |
|---|---|---|---|---|---|---|
| *nu* | 0.567545 | 0.800994 | 0.508677 | 0.902342 | 0.937500 | 0.466456 |
| *obj* | −100.559618 | −121.793804 | −85.452432 | −87.357344 | −12.493618 | −60.566177 |
| *rho* | −0.002701 | −0.622608 | −0.599988 | 0.101549 | −0.685184 | −0.524438 |
| $N_{SV}$ | 31 | 32 | 29 | 31 | 31 | 27 |
| $N_{bsv}$ | 12 | 20 | 11 | 28 | 29 | 8 |
| MSE | 0.126849 | 0.175542 | 0.126959 | 0.0428112 | 0.0447425 | 0.0648428 |
| Squared correlation coefficient | 0.0137583 | 0.0122813 | 0.00559318 | 0.108904 | 0.207437 | 0.0401919 |
| **Parameter/Month** | **Nov.** | **Dec.** | **Jan.** | **Feb.** | **Mar.** | **Apr.** |
| *nu* | 0.412548 | 0.703099 | 0.681795 | 0.373345 | 0.798729 | 0.806393 |
| *obj* | −51.784314 | −27.923460 | −47.504455 | −40.900303 | −158.778634 | −250.631323 |
| *rho* | 0.904255 | −0.278622 | −0.407973 | −0.036655 | −0.575151 | 0.240353 |
| *nSV* | 15 | 31 | 32 | 23 | 31 | 28 |
| $N_{bsv}$ | 11 | 18 | 16 | 6 | 22 | 20 |
| MSE | 0.00665455 | 0.0224863 | 0.0150721 | 0.0160867 | 0.0227255 | 0.0638941 |
| Squared correlation coefficient | 0.861286 | 0.544593 | 0.761683 | 0.523841 | 0.556827 | 0.0997706 |

*nu* is the parameter of the kernel function type, *obj* represents the minimum value obtained by the quadratic programming solution converted from the SVM file, *rho* is the bias term *b* of the decision function, $N_{SV}$ is the number of standard support vectors, and $N_{bSV}$ denotes the number of support vectors on the boundary.

**Table 5** | Analysis of partial ACO-SVM prediction results

| **ACO_3 (calibration period)** | **July** | **Aug.** | **Sep.** | **Dec.** | **Jan.** | **Feb.** | **Mar.** |
|---|---|---|---|---|---|---|---|
| Min. relative value | 0.0065 | 0.0011 | 0.0235 | 0.0005 | 0.0042 | 0.0020 | 0.0075 |
| Max. relative value | 0.4011 | 0.3664 | 0.3822 | 0.2252 | 0.1771 | 0.1982 | 0.2142 |
| Ave. relative value | 0.1690 | 0.0749 | 0.1592 | 0.0505 | 0.0542 | 0.0459 | 0.0710 |
| Correlation coefficient | 0.4878 | 0.8645 | 0.6434 | 0.9488 | 0.9327 | 0.8923 | 0.8878 |
| MSE | 0.0441 | 0.0118 | 0.0360 | 0.0048 | 0.0048 | 0.0041 | 0.0081 |
| NSE | 0.2161 | 0.7472 | 0.3527 | 0.8989 | 0.8689 | 0.7847 | 0.7823 |
| Qualified rate | 0.6667 | 0.9167 | 0.6667 | 0.9722 | 1.0000 | 1.0000 | 0.9722 |
| **ACO_3′ (inspection period)** | **July** | **Aug.** | **Sep.** | **Dec.** | **Jan.** | **Feb.** | **Mar.** |
| Min relative value | 0.0137 | 0.0172 | 0.0381 | 0.0234 | 0.0172 | 0.0300 | 0.0457 |
| Max relative value | 0.4532 | 0.7980 | 0.2990 | 0.3000 | 0.3479 | 0.1376 | 0.3333 |
| Ave. relative value | 0.1635 | 0.1934 | 0.1280 | 0.1154 | 0.1225 | 0.0841 | 0.1585 |
| Correlation coefficient | 0.5188 | 0.4055 | 0.4034 | 0.9227 | 0.8371 | 0.9777 | 0.7638 |
| MSE | 0.0546 | 0.0845 | 0.0239 | 0.0209 | 0.0258 | 0.0080 | 0.0334 |
| NSE | 0.1579 | 0.0059 | −0.1332 | 0.6804 | 0.6854 | 0.8934 | 0.4438 |
| Qualified rate | 0.7000 | 0.7000 | 0.8000 | 0.8000 | 0.8000 | 1.0000 | 0.7000 |

Except for a small (less than 1.0%) increase in individual months, the mean square error decreases by between 4.4% and 71.1% in other months. The certainty coefficient is increased by 2.1% to 419.4%, and the qualified rate is generally improved. This analysis shows that the runoff prediction effect of ACO-SVM is better than that of the default parameter SVM.

It is worth noting that the certainty coefficient of NSE and the average qualified rate in March are relatively small, which may be due to the fact that this is in the transitional stage from the dry season to the wet season, when groundwater recession is relatively significant, and the ground runoff is supplemented by groundwater and snow melting (Li & Xue 2017), resulting in inaccurate forecasting results.

### Comparison of model actual application effect

To examine the actual application effect of ACO-SVM, the runoff forecast model (Zhang *et al.* 2009b) established by Bayesian statistical forecasting theory (BSFT) is selected as the reference model. The model first uses Bayesian method to determine the precipitation uncertainty based on establishing a correlation model between precipitation and runoff. Based on this, the probability density distribution function of the uncertain input data is solved, then the first-order autoregressive model of runoff between adjacent two months is established to determine the prior probability distribution of monthly runoff, and then the data pertaining to $h_i$ and $p_i$ predicted by the precipitation runoff model and first-order runoff autoregressive model are used to build a linear model. Thereafter, regression analysis is used to determine the parameters of the linear model between $h_i$ and $p_i$, and obtain the sample likelihood function and the posterior probability distribution of runoff with the predicted value $p_i$ as the condition and determine the Bayesian model parameters. Finally, by combining the input uncertainty function, the explicit solution of runoff distribution density predicted by the Bayesian model is derived, and the expected value is taken as the final prediction result.

According to the prediction results of both ACO-SVM and BSFT, the average relative error and mean square error between predicted and measured value and qualified rate are calculated respectively, and the comparison results are plotted (Figures 6–8). Compared with BSFT, the average relative error and mean square error of ACO-SVM are generally reduced (Figures 6 and 7), and the qualified rate is significantly improved (Figure 8), indicating that the prediction accuracy

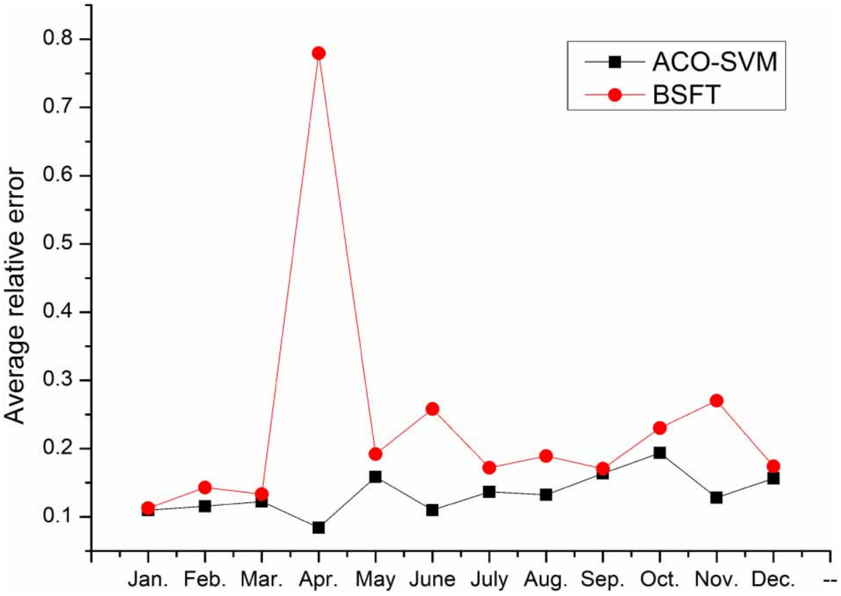

**Figure 6** | Average relative error between predicted and measured values of ACO-SVM and BSFT.

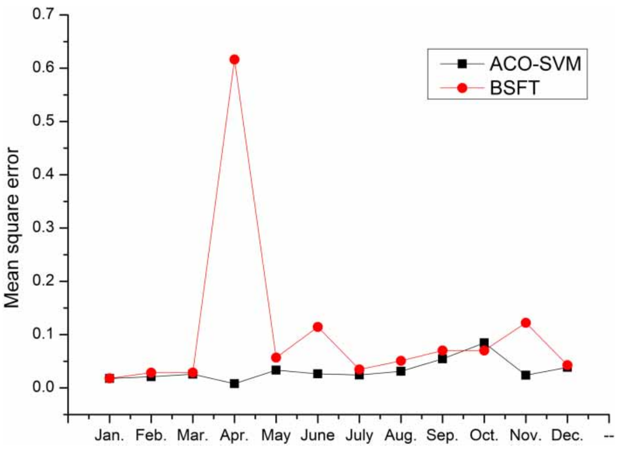

**Figure 7** | Mean square error between predicted and measured values of ACO-SVM and BSFT.

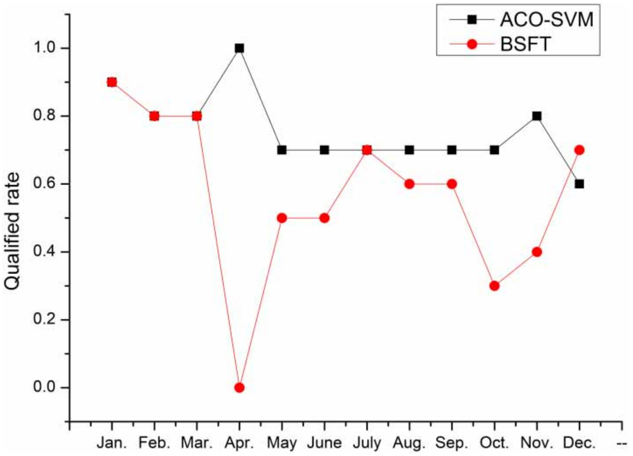

**Figure 8** | Qualified rate of ACO-SVM and BSFT.

of ACO-SVM is higher than that of BSFT. It can be considered that ACO-SVM has good practical application effect.

## CONCLUSION

Through the aforementioned research, the following conclusions can be drawn:

(1) It is feasible to use grid precipitation as a predictor of an SVM model to predict monthly runoff, and the effect is better than that of station precipitation.
(2) The SVM monthly runoff forecasting model optimized by ACO algorithm can achieve better results.
(3) When forecasting runoff, the influence of the monthly change trend of the water recession behavior on the confluence and the effect of snow melting supplement on the runoff cannot be ignored.

The complexity of topography, geology, hydrology, and meteorology in the upper reaches of Minjiang River gives rise to many influencing factors, which poses difficulties to medium and long-term runoff forecasting. A regression support vector machine (RSVM) is developed on the basis of a classified support vector machine (CSVM), which inherits the advantages of the good generalization ability of CSVM and is often used to deal with the problems of regression prediction. Aimed at the problem of low runoff forecasting accuracy of traditional SVM, the use of ACO is proposed to make the parameters of SVM approximate to a reasonable value, and an appropriate data-driven model is established to improve the accuracy of runoff forecasting, which provides a reference for future runoff forecasting research.

## ACKNOWLEDGEMENT

We thank the staff in Sichuan Province responsible for the award of a major science and technology project grant: 'Power Network Intelligent Key Technology Research and Application Demonstration' (Project Code: 2018GZDX0043).

## DATA AVAILABILITY STATEMENT

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

## REFERENCES

Chen, S. C., Yan, Y. Q., Li, Z. H. & Li, J. 2009 Study on design flood dispatching of Changtan Reservoir considering the influence of upstream small and medium reservoirs. In: *Shanghai Maritime Exchange Association, China. Proceedings of the 14th National Maritime Technical Symposium*, Shanghai, pp. 90–97 (in Chinese).

Cheng, C. T., Feng, Z. K., Niu, W. J. & Liao, S. 2015 Heuristic methods for reservoir monthly inflow forecasting: a case study of Xinfengjiang Reservoir in Pearl River, China. *Water* **7** (8), 4477–4495.

Dirk, S. & Christian, T. 2012 Running time analysis of Ant Colony Optimization for shortest path problems. *Journal of Discrete Algorithms* **10**, 165–180.

Dorigo, M., Maniezzo, V. & Colorni, A. 1996 Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics* **26** (1), 29–41.

Du, H. M. & He, S. Y. 2015 The analysis on characteristics of precipitation and trends in drought and flood disasters in Minjiang River Basin. *Research of Soil and Water Conservation* **22** (1), 153–157 (in Chinese).

Li, Y. H. & Xue, C. 2017 Inter-annual variation characteristics of reservoir runoff in Zipingpu Reservoir. *Sichuan Water Power* **36** (S2), 113–115. +132 (in Chinese).

Li, J. D., Huang, W. B., Zhao, Q. X. & Ma, G. 2016 Water level control plan under joint operation of cascade reservoirs. *Systems Engineering – Theory & Practice* **36** (06), 1625–1632.

Maniezzo, V. & Colorni, A. 1999 The ant system applied to the quadratic assignment problem. *IEEE Transactions on Knowledge & Data Engineering* **11** (5), 769–778.

Vafakhah, M. & Khosrobeigi, B. S. 2019 Regional analysis of flow duration curves through support vector regression. *Water Resources Management* **34** (14), 1–12.

Wang, J. J. 2015 Predication of annual runoff in Kaidu river based on modified support vector machine. *Northwest Hydropower* **04**, 1–5 (in Chinese).

Wang, Q. H., Li, J. D., Chen, S. J. & Wang, X. 2018 SVM-based Implicit stochastic scheduling mode for cascade hydropower stations. Beijing. *MATEC Web of Conferences.* **246**, 2046–2052.

Wu, Y. H. 1999 Statistical learning theory. *Technometrics* **41** (4), 377–378.

Wu, C. L. 2010 Hydrological predictions using data-driven models coupled with data preprocessing techniques. PhD thesis, *Hong Kong Polytechnic University (Hong Kong)* 16–79.

Zhang, B. L., Qian, L. F., Cao, J. J. & Ren, G. 2009a Parameter optimization of support vector machine based on ant colony optimization algorithm. *Journal of Nanjing University of Science and Technology (Natural Science Edition)* **04**, 464–468 (in Chinese).

Zhang, M., Li, C. J. & Zhang, Y. C. 2009b Application of the Bayesian statistic hydrological forecast system to middle-and long-term runoff forecast. *Advances in Water Science.* **020** (001), 40–44.

Zhang, Z., Zhang, Q., Singh, V. P. & Shi, P. 2018 River flow modeling: comparison of performance and evaluation of uncertainty using data-driven models and conceptual hydrological model. *Stochastic Environmental Research and Risk Assessment* **32** (9), 2667–2682.

Zhao, S. & Yang, D. W. 2011 Mutual information-based input variable selection method for runoff-forecasting neural network model. *Journal of Hydroelectric Engineering* **30** (01), 24–30 (in Chinese).

Zhu, H. P. & Li, X. H. 2016 Research on a new method based on improved ACO algorithm and SVM model for data classification. *International Journal of Database Theory and Application.* **9** (1), 217–226.

First received 5 March 2020; accepted in revised form 13 June 2020. Available online 26 June 2020