Stochastic simulation of daily runoff in the middle reaches of the Yangtze river based on SVM-Copula model

Pingyi Wang, Jie Zhang, Meili Wang, Yuchun Liang and Jian Li

National Inland Waterway Regulation Center, Chongqing Jiaotong University, Chongqing, People’s Republic of China; School of Architecture and Urban Planning, Chongqing Jiaotong University, Chongqing, People’s Republic of China; China Metallurgical Construction Engineering Group Co., Ltd., Chongqing, People’s Republic of China

ABSTRACT

Hydrological stochastic simulation is an important method for hydrologic analysis which needs a large amount of measured data to determine the variable distribution function. That impedes the development of the random simulation in small sample cases. The characteristics of daily runoff series in the middle reaches of Yangtze River has changed a lot since the Three Gorges dam was built in 2003. The measured hydrologic data since 2003 is insufficient for stochastic simulation with conventional methods. In order to overcome the difficulty, a simulation model coupling the support vector machine with the Copula function was established in this paper. The measured data of Hankou Hydrology Station from 2003 to 2015 was used as an applied example to verify the applicability of the model. The research showed that the model we established can maintain the statistical characteristics of the original daily runoff series well, it could be a new way for the random simulation in the case of insufficient data.

1. Introduction

Hydrological processes are quite complicated due to the influence of meteorology, terrain and human activities. The stochastic simulation method has been an important mean of understanding and managing complex hydrology by studying the characteristics of random hydrological processes. A reasonable model is the key to effective stochastic simulation. The runoff is important hydrological element, the runoff simulation sequences generated by stochastic models are used to obtain the required hydrology data (such as the flood discharge, firm capacity, reservoir design flood level, etc.). Therefore, the simulation of runoff plays an important role in engineering design, hydrology and water resources management. With the continuous development of random hydrology, many different types of stochastic models have been formed, such as the autoregressive models, the dissolving models, the speckle noise models, wavelet models, stochastic models based on Copula function, etc. The stochastic model based on Copula functions can effectively describe the relationship between variables and construct multi-variable joint distribution.

It has been widely used in runoff stochastic simulation since Sklar first proposed the Copula theory in 1959 (Sklar, 1973). In recent years, a lot of researches study on the theory and its application of Copula simulation model in hydrology. Pereira, Álvaro, Erhardt, and Czado (2016) developed a statistical model which considers spatial and temporal dependencies of runoff for runoff scenario simulation, an applied study showed that the model could reproduce the temporal and spatial dependence structure. Chen and Guo (2015) proposed a Copula-based method for generating multi-site monthly and daily runoff data, the method could realize multisite stochastic simulation by establishing trivariate Copula. Liu, Zhou, Chen, and Guan (2015) proposed a multivariate conditional model for runoff prediction and the refinement of spatial prediction estimates. The model consisted of high-dimensional vine Copulas, conditional bivariate Copula simulations, and a quantile-Copula function. A maximum entropy-Gumbel-Hougaard Copula model for daily runoff simulation was established by Kong, Huang, Fan and Li, the model was proved to be effective by an example study. Hao and Singh (2012) proposed an entropy-Copula model for single-site monthly runoff simulation, the joint distribution of adjacent monthly runoff was constructed by using the Copula method. Lee and Salas (2014) explored the applicability of the Copula method for stochastic runoff simulation and investigate the pros and cons of different Copula time series models.
Jeong and Lee (2015) developed a stochastic simulation model based on Copula method for intermittent monthly flow arid regions, and applied the model to Colorado River intermittent runoff for water management. Mei, Yin, Hong, Li, and Li (2016) compared the synthesis effect on annual runoff of the double model simulation method and the typical method of solution set decomposition, the sample mean features keep well by two methods. Yang, Si, Fan, Kong, and Lin (2019) constructed the correlation analysis between rainfall and runoff in Xiangxi River by using Archimedean Copula method, and the joint distribution of monthly rainfall and runoff was obtained.

In these studies, the determination of the variable marginal distribution and the joint distribution between variables are the two keys to construct an appropriate stochastic model. The determination of the joint distribution function between variables can be achieved by selecting the appropriate Copula function, and the variable marginal distribution function usually determined by probabilistic methods which requires a sufficient sample size of the historical sample data. Therefore, the above models are usually stochastically simulated based on measured runoff data of rivers for at least several decades.

Since the Three Gorges Reservoir was impounded in 2003, the runoff of the hydrological station in the middle and lower reaches of the Yangtze River has undergone significant changes which are affected by reservoir dispatching and rainfall. The daily runoff sequence during one year changed greatly especially. The changes are expressed in a decrease during flood season and an increase during dry season. The random simulation of the runoff sequence after the Three Gorges reservoir has become a new focus of the Three Gorges reservoir’s dispatching and flood controlling. Since the Three Gorges water storage, the accumulated hydrological data of the hydrological station in the middle and lower reaches of the Yangtze river are less than 20 years which are far from the sample size to determine the variable’s marginal distribution by the frequency statistical method. Finding a new effective marginal distribution identification method has become the key point for establishing a runoff simulation model. Support vector machine (SVM) is a machine learning method based on statistical learning theory which was developed in the mid-90s. It seeks to increase the generalization ability of learning machines and get the minimization of the risk of experience and the confidence range by seeking the minimization of the structural risk. Hence it has a good performance in solving the problems of small samples and nonlinearities. Zhou, Jun, Zhang, and Jun (2017) proposed a model combining wavelet transform (WT) with bacterial colony chemotaxis algorithm and support vector machine (BCC–SVM) algorithm (WT–BCC–SVM), they applied the method to diagnosis of pesticide residues to provided it could work. SVM classification method has been widely used in mechanical engineering, medicine, artificial intelligence and other fields. Diyana, Yee, Tadaaki, Hiroshi, and Fumiori (2017) presented a two-stage tongue’s multi-colour classification based on SVM method for precise tongue colour diagnosis, an application showed that the accuracy of diagnosis was satisfactory. Chen, Wang, Yang, and Zhou (2006) proposed a novel approach based on rough set theory and SVM method for facial expression recognition with considering the geometric feature only, and the approach could get high recognition ratio. Jangid, Dhir, Rani, and Singh (2011) also used SVM classifier with RBF kernel to classify handwritten numbers in their study. Jing, Cao, Hu, and Gao (2015) proposed a reliability life distribution pattern recognition method based on multi-classification fuzzy support vector machine (MFSVM), simulation experiment results showed that this model can overcome the shortcomings of traditional support vector machines and can recognize the commonly used life distribution patterns intelligently. A large number of practices have proved that SVM classification method can effectively identify different types of images and texts. Based on the SVM pattern recognition method and the Copula function of constructing joint distribution, the SVM-Copula coupled stochastic model is established in this paper. It provides a new way for the random simulation of the daily runoff sequence in the small sample cases.

2. Identification of the marginal distributions of daily runoff based on the SVM theory

The purpose of SVM classification is to develop a computationally efficient way to find the optimal classification hyperplane in high-dimensional feature space. At present, the SVM classification has been widely used in text classification, image recognition, handwritten recognition, biometric classification and recognition (Cristiani & Shawe, 2001).

2.1. The algorithm of SVM

The goal of classification is to find the optimal hyperplane, the target hyperplane is required not only to correctly separate the points of the sample but also to maximize the classification interval which means the minimum structural risk (Cherkassky, 1995). In the case of linear inseparability, the nonlinear mapping algorithm is used to transform the linear inseparable samples of the low-dimensional input space into high-dimensional feature spaces to make them linearly separable.
2.1.1. SVM bivariate classification algorithm

In the case of two separable sets, we can separate the two sets with an infinite hyperplane which is just at the centre between the two sets. The basic fundamental of the SVM algorithm is to find the support plane, the bivariate classification algorithm is as follow (Sakthivel, Sugumaran, & Nair, 2010).

Set the training sample set as

\[ D = \{ (x_i, y_i) \mid i = 1, 2, \ldots, n \}, x \in R^d, y \in \{-1, +1\}, \]

where \( n \) is the number of training samples, \( d \) is the dimension of each training sample vector, and \( y \) is the category label (if \( y = +1 \), the sample belongs to one class; if \( y = -1 \), the sample belongs to another class).

In linear indivisible cases, transform function \( \Phi : x \rightarrow \Omega \) maps training samples from input space to the high-dimensional feature space, the kernel function \( K(x_i, y_i) = \Phi(x_i)\Phi(y_i) \) is the inner product in the corresponding high dimensional feature space \( \Omega \).

Designing an SVM bivariate classifier is to find out the optimal hyperplane that maximizes the classification decision function of each training sample vector, and \( y \) is the category label (if \( y = +1 \), the sample belongs to one class; if \( y = -1 \), the sample belongs to another class).

After the optimal solutions of \( \alpha_i^* \) and \( b^* \) are obtained by solving the above constrained optimization problem, the classification function is as follows:

\[ f(x) = \text{sign} \left[ \sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^* \right] \]

where \( \alpha_i^* \) and \( b^* \) are the optimal solutions of \( \alpha_i \) and \( b \) separately.

The SVM classification algorithms are different with the different selected kernel functions \( K(\cdot, \cdot) \). The commonly used kernel functions are: linear kernel function; polynomial kernel function; radial basis kernel function and sigmoid kernel function.

2.1.2. SVM multivariate classification algorithm

It is necessary to construct a suitable multi-class classifier when deal with the classification with multivariate types. At present, multi-classifiers are mainly constructed by a combination of multiple two-classifiers which can be established by the one-to-one method and the one-to-all method. In practical applications, the one-to-one method is more commonly used. It needs to construct a classifier between any two types, and complete classification by voting.

2.2. Identification of the daily runoff marginal distribution based on SVM-classifier

Practice shows that the distribution patterns commonly used in hydrological analysis are Normal distribution, Lognormal distribution, Weibull distribution, Exponential distribution, Pearson distribution and Gamma distribution.

The characteristics of the distribution can be measured and described in three aspects. The first one is the concentration trend which reflects the degree of data closing to or aggregating from the central value such as median, quantile, average, harmonic mean, geometric mean, etc. The second one is the degree of dispersion which reflects the trend of each data away from the central value such as heterogeneous ratio, interquartile range, standard deviation and variance, standard deviation coefficient, discrete coefficient, etc. The third one is the skewness and the kurtosis reflecting the shape. Combined with the main features of hydrological data, seven characteristic parameters including quartile (upper quartile and lower quartile), interquartile range, standard deviation, standard deviation coefficient, skewness coefficient and kurtosis coefficient were selected.

The above distribution patterns were subdivided into 573 distribution patterns as different distribution parameters which are shown as Table 1.

A set of distribution feature parameter values corresponding to each distribution pattern constituted a training sample. The training sample value were calculated by...
Table 1. The distribution patterns.

| Distribution type       | Distribution function parameter | Number of segments |
|-------------------------|----------------------------------|--------------------|
| Normal Distribution     | $m=0$, std $= 1 \sim 100$, interval $1$ | $100$              |
| Lognormal Distribution  | $a=0$, $b=0.1 \sim 10$, interval $1$ | $100$              |
| Exponential Distribution| $a=1$                           | $100$              |
| Weibull Distribution    | $a=1$, $b=0.5 \sim 50$, interval $0.5$ (leave out $b=1$) | $99$              |
| Pearson Distribution    | $a=1 \sim 5$, $b=3 \sim 27$     | $75$              |
| Gamma Distribution      | $a=1$, $b=0.1 \sim 10$ (leave out $b=1$) | $99$              |

$m$ is the average, $s$ is the standard deviation, $a$ is the scale parameter, $b$ is the shape parameter.

using 100,000 computer simulation random sequences for each distribution mode. Part of training sample values are shown in Table 2.

The SVM two-classifiers were constructed with two of the six distribution patterns. A total of 15 classifiers are used to identify the distribution patterns of the test samples by using the voting method. The distribution type with the largest number votes is the goal marginal distribution that the test samples satisfied. And then the marginal distribution function of the test sample can be determined.

3. SVM-Copula stochastic model of daily runoff

3.1. Copula functions

The concept of Copula was first introduced by Sklar in 1959 to answer M. Fréchet’s question about the relationship between multi-dimensional distribution functions and low-dimensional edges. The Copula function is a uniformly distributed multi-dimensional joint distribution function defined on $[0, 1]$. It can join the marginal distributions of multiple random variables to construct a joint distribution (Nelsen, 1999).

$$H(X_1, X_2, \ldots, X_d) = C(F_1(X_1), F_2(X_2), \ldots, F_d(X_d)), \quad (6)$$

where $X_i$ is random variable; $F_i(X_i)$ is the marginal distribution function of random variables $X_i$.

For the two-dimensional case:

$$H(x, y) = C_0(F_X(x), F_Y(y)) \overset{\Delta}{=} C_0(u, v), \quad (7)$$

where $\theta$ is a parameter of the Copula function; $u = F_X(x), v = F_Y(y)$, which are the marginal distributions of the random variables $(X, Y)$ respectively.

Copula functions can be roughly divided into elliptical Copula, Archimedean Copula, quadratic Copula. Thereinto, Archimedean Copula functions are widely used in the field of hydrology. The expressions and applicable conditions of the common two-dimensional Archimedean Copula are as follows. Gumbel-Hougaard (GH) Copula function:

$$C(u, v) = e^{-[(\ln u)^\theta + (\ln v)^\theta]^\frac{1}{\theta}}, \quad \theta \in [-1, \infty). \quad (8)$$

Clayton Copula function:

$$C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}, \quad \theta \in (0, \infty). \quad (9)$$

Ali-Mikhail-Haq (AMH) Copula function:

$$C(u, v) = \frac{uv}{1 - \theta(1 - u)(1 - v)}, \quad \theta \in [-1, 1). \quad (10)$$

Frank Copula function:

$$C(u, v) = -\frac{1}{\theta} \ln \left[1 + \frac{(e^{-\theta u - 1}) (e^{-\theta v} - 1)}{e^{-\theta} - 1}\right], \theta \neq 0. \quad (11)$$

3.2. Construction of SVM-Copula stochastic model

The daily runoff sequence is a non-stationary hydrological time series. $x_t$ and $x_{t-1}$ have different marginal distributions with $F_t(x)$ and $F_{t-1}(x)$, the joint distribution function

Table 2. Partial training sample values.

| Distribution type | Q0.25 | Q0.75 | Q  | $\sigma$ | $\nu$ | Skew | Kur |
|-------------------|-------|-------|----|----------|-------|------|-----|
| N(0, 1)           | 0.391 | 0.546 | 0.155 | 0.115 | 0.244 | 0.001 | 2.980 |
| N(0, 100)         | 0.417 | 0.576 | 0.160 | 0.118 | 0.238 | 0.001 | 2.979 |
| L(0, 0.1)         | 0.319 | 0.472 | 0.153 | 0.114 | 0.286 | 0.305 | 3.171 |
| L(0, 10)          | 0.000 | 0.000 | 0.000 | 0.003 | 298.437 | 315.806 | 99819.507 |
| E(1)              | 0.027 | 0.131 | 0.104 | 0.095 | 1.003 | 1.972 | 8.581 |
| E(100)            | 0.024 | 0.115 | 0.091 | 0.083 | 1.003 | 2.023 | 9.177 |
| W(1, 0.5)         | 0.001 | 0.015 | 0.014 | 0.034 | 2.236 | 6.470 | 76.309 |
| W(1, 50)          | 0.708 | 0.831 | 0.123 | 0.098 | 0.129 | 1.002 | 4.762 |
| P(1, 3)           | 0.039 | 0.402 | 0.363 | 0.250 | 1.002 | 1.005 | 3.009 |
| P(5, 27)          | 0.000 | 0.000 | 0.000 | 0.167 | 4.722 | 5.013 | 27.136 |
| G(1, 0.1)         | 0.023 | 0.113 | 0.090 | 0.082 | 1.003 | 1.993 | 8.842 |
| G(1, 10)          | 0.025 | 0.121 | 0.096 | 0.087 | 1.001 | 1.989 | 8.818 |

$N(-,-)$ represents the normal distribution, $L(-,-)$ represents the lognormal distribution, $E(-,-)$ represents the exponential distribution, $W(-,-)$ represents the weibull distribution, $P(-,-)$ represents the pearson distribution, $G(-,-)$ represents the gamma distribution, $Q0.25$ represents the lower quantile, $Q$ represents the interquartile range, $Q = Q0.25 - Q0.75$, $\sigma$ represents the standard deviation, $\nu = \frac{Q}{\sigma}$ represents the standard deviation coefficient which is also called the discrete coefficient, Skew represents the skewness, Kur represents the kurtosis.
can be express as:

\[ G(x_t, x_{t-1}) = C_0(F(x_t), F(x_{t-1})) = C_0(z_t, z_{t-1}), \quad (12) \]

where \( z_t, z_{t-1} \) are the marginal distributions of \( x_t \) and \( x_{t-1} \). The conditional distribution function of \( x_t \) with respect to \( x_{t-1} \) is:

\[
G_\theta(x_t|x_{t-1}) = \frac{\partial C_\theta(z_t, z_{t-1})}{\partial z_{t-1}}, \quad (13)
\]

After the conditional distribution function of \( x_t \) with respect to \( x_{t-1} \) has been gotten, a random simulation of the daily runoff time series can be performed according to the inverse function of the generated sample.

\[
z_{t,i} = G^{-1}_\theta(e_{t,i}|z_{t-1,i}),
\]

\[
x_{t,i} = F^{-1}(z_{t,i}), \quad (14)
\]

where \( t = 1, 2, \ldots, m \), \( m \) is the number of sections; \( i = 1, 2, \ldots, n \), \( n \) is the sequence length; \( \varepsilon_{t,i} \) is random number uniformly distributed in accordance with \((0, 1)\). The previous section value of the first section value of the \( i \)-th year is the \( m \)-th section of the \((i-1)\)-th year which is \( Z_{m,1-1} \). \( \theta_t \) represents the relationship between the first section and the \( m \)th section values of the previous year. The marginal distribution function \( F(x) \) of random variable can be determined by SVM classifier.

The simultaneous equations of Equation (13) and Equation (14) are the general forms of the SVM-Copula-based stochastic model of the daily runoff sequence:

\[
G_\theta(x_t|x_{t-1}) = \frac{\partial C_\theta(z_t, z_{t-1})}{\partial z_{t-1}}
\]

\[
z_{t,i} = G^{-1}_\theta(e_{t,i}|z_{t-1,i}),
\]

\[
x_{t,i} = F^{-1}(z_{t,i}), \quad (15)
\]

where \( t = 1, 2, \ldots, m \), \( i = 1, 2, \ldots, n \).

### 3.3. Parameter estimation

The parameters of SVM-Copula daily runoff series stochastic model can be divided into two categories: marginal distribution function parameters and the Copula function parameters. The parameters of the marginal distribution function can generally be obtained by the moment method, the linear moment method or the maximum likelihood method. The correlation index method is usually used to determine the Copula function parameters. The parameter \( \theta \) can be calculated by the relationship with the Kendall rank correlation coefficient. The relationships between \( \theta \) and \( \tau \) are as follows: GH Copula function:

\[ \tau = 1 - \frac{1}{\theta}, \theta \in [1, \infty). \quad (16) \]

Clayton Copula function:

\[ \tau = \frac{\theta}{\theta + 1}, \theta \in (0, \infty). \quad (17) \]

AMH Copula function:

\[ \tau = \left( 1 - \frac{2}{3\theta} \right) - \frac{2}{3} \left( 1 - \frac{1}{\theta} \right)^{\frac{2}{3}}, \theta \in [-1, 1). \quad (18) \]

Frank Copula function:

\[ \tau = 1 + \frac{4}{\theta} \left[ \int_0^\theta \frac{t}{e^t - 1} dt - 1 \right], \theta \neq 0. \quad (19) \]

### 3.4. Model algorithm

The algorithm of the SVM-Copula model consists of two parts: the marginal distributions which are determined by SVM classifier and the conditional distribution functions which are calculated by Copula functions. On the basis of these, the daily runoff simulation sequence can be obtained by the inverse function method. The simulation steps of the SVM-Copula daily runoff stochastic model are:

Step 1. Determine the marginal distribution. Identify the marginal distribution of each day’s runoff by SVM classifier, determine the marginal distribution function \( F_t(x) \) and its parameters \( \mu_t, \sigma_t \).

Step 2. Compute the joint distribution. Choose an appropriate Copula function by analysing the measured daily runoff sequence, determine the parameter \( \theta_t \), and solve the Equation (13) to get the conditional distribution function of \( x_t \) with respect to \( x_{t-1} \). put the result into Equation (14) to determine \( z_{t,i} \) and \( x_{t,i} \).

Step 3. Give the original number. Generate a random number \( \varepsilon_{1,1} \) uniformly distributed in accordance with \((0, 1)\), put them into the first formula of Equation (12) to get \( z_{1,1} \), and then solve the second formula of Equation (12) to get the simulation value of the first section of first year \( x_{1,1} \).

Step 4. Simulate the random runoff sequence of the first year. Put \( z_{1,1} \) and a random number \( \varepsilon_{2,1} \) to Equation (15) to get \( z_{1,1} \) and the simulation value \( x_{2,1} \) of the second section of the first year, and so on. Get the \( m \) section values \( z_{3,1}, z_{4,1}, \ldots, z_{m,1} \) and the simulation values \( x_{3,1}, x_{4,1}, \ldots, x_{m,1} \).

Step 5. Simulate the random runoff sequence of \( n \) years. Put \( z_{m,1} \) and a random value \( \varepsilon_{1,2} \) into Equation (12) to get \( z_{1,2} \) and \( x_{1,2} \), repeat step 4 to get the simulation values of \( m \) sections of the second year. And so on to obtain the simulation values of daily runoff sequence of \( n \) years.
4. Application study

The current runoff data of the Hankou Hydrological Station from 2003 to 2015 was used for the purpose of testing the feasibility and effectiveness of the SVM-Copula stochastic model.

4.1. Model establishment

According to the measured runoff data of Hankou Station from 2003 to 2015, the characteristic parameters are extracted from the daily runoff time series, the daily runoff characteristic parameters of each day constitute the test sample set. Partial test samples are shown in Table 3.

The distributed pattern recognition is performed by the one-to-one SVM multi-classifier which was constructed in the foregoing, and the voting result is shown in Table 4.

From Table 4, the logarithmic normal distribution has the highest vote, and the daily runoff is considered to be in the Lognormal distribution \( L(\mu, \sigma) \) whose probability density function (PDF) is

\[
f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right).
\]

Its cumulative distribution function (CDF) is

\[
F(x) = \int_0^x \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right) \, dx,
\]

where parameters \( \mu \) and \( \sigma \) are the mean and standard deviation of \( \ln(x) \) respectively.

Therefore, the parameters of the marginal distributions of different days are different. Partial Lognormal distribution parameters are shown in Table 5.

It has been proved by practice that the Frank Copula function simulates the runoff sequence better than other

### Table 3. Partial test sample values.

| Date | Q0.25 | Q0.75 | Q | σ | Skew | Kur |
|------|-------|-------|---|---|------|-----|
| D1   | 0.156 | 0.530 | 0.374 | 0.308 | 0.785 | 0.933 | 2.715 |
| D50  | 0.138 | 0.287 | 0.149 | 0.246 | 1.022 | 2.197 | 7.667 |
| D100 | 0.159 | 0.402 | 0.243 | 0.271 | 0.788 | 1.324 | 4.018 |
| D150 | 0.360 | 0.892 | 0.532 | 0.344 | 0.586 | -0.516 | 1.926 |
| D200 | 0.281 | 0.657 | 0.376 | 0.292 | 0.681 | 0.457 | 2.335 |
| D250 | 0.291 | 0.715 | 0.424 | 0.319 | 0.632 | -0.012 | 0.048 |
| D300 | 0.314 | 0.814 | 0.500 | 0.312 | 0.788 | 1.022 | 2.309 |
| D360 | 0.176 | 0.663 | 0.487 | 0.294 | 0.716 | 0.657 | 2.309 |
| D365 | 0.176 | 0.663 | 0.487 | 0.294 | 0.716 | 0.657 | 2.309 |

Q0.75 represents the upper quartile; Q0.25 represents the lower quartile; \( q \) represents the interquartile range; \( \sigma \) represents the standard deviation; \( \nu \) represents the standard deviation coefficient which is also called the discrete coefficient; Skew represents the skewness; Kur represents the kurtosis.

### Table 5. Partial lognormal distribution parameters.

| Date | \( \mu \) | \( \sigma \) |
|------|----------|----------|
| D1   | 9.232    | 0.178    |
| D50  | 9.257    | 0.217    |
| D100 | 9.592    | 0.172    |
| D150 | 10.204   | 0.270    |
| D200 | 10.538   | 0.198    |
| D250 | 10.335   | 0.342    |
| D300 | 9.673    | 0.190    |
| D365 | 9.251    | 0.176    |

### Table 4. Voting result.

| Distribution type | Normal | Lognormal | Exponential | Weibull | Pearson | Gamma |
|-------------------|--------|-----------|-------------|---------|---------|-------|
| votes             | 3      | 5         | 0           | 4       | 2       | 1     |

### Table 6. Partial measured daily runoff data rank correlation coefficients \( \tau \).

| Date | \( \tau \) |
|------|----------|
| D1   | 0.9780   |
| D2   | 0.9724   |
| D3   | 0.9945   |
| D4   | 0.9503   |
| D360 | 0.9990   |
| D361 | 0.9780   |
| D362 | 0.9341   |

### Table 7. Partial Copula function parameters \( \theta \).

| Date | \( \theta \) |
|------|-------------|
| D1   | 177         |
| D2   | 140         |
| D3   | 715         |
| D4   | 77          |
| D360 | 400         |
| D361 | 177         |
| D362 | 57          |

Copula functions (Jeong & Lee, 2015). Beyond that, parts of the Kendall rank correlation coefficients \( \tau \) of the daily runoff data are shown in Table 6. As we can see, the rank correlation coefficients between adjacent two days runoff are large. Therefore, the Frank Copula function whose function form has been shown in Equation (14) is selected as the coupling function of the model, and the relationship between \( \theta \) and \( \tau \) has been shown in Equation (19).

According to the relationship between Frank Copula function parameter \( \theta \) and the Kendall rank correlation coefficient \( \tau \), the parameters of Copula function \( \theta_t \) of the model can be obtained. Parts of the parameters are as shown in Table 7.

According to Equation (16), the conditional distribution function of \( x_t \) is

\[
G_{\theta_t}(x_t|x_{t-1}) = \frac{\partial c_{\theta_t}(z_t,z_{t-1})}{\partial z_{t-1}} = \frac{1 - e^{-\theta_t}}{1 - e^{-\theta_t}z_{t-1} - e^{-\theta_t}} - 1^{-1}.
\]

(22)
### Table 8. Partial mean and mean variance of measured values and simulated values.

| Date | Mean | Relative Error (%) | Mean Squared Error | Relative Error (%) |
|------|------|--------------------|--------------------|--------------------|
| 1    | 10392| 0.42               | 2024               | 0.6                |
| 2    | 10346| 0.18               | 2036               | 0.76               |
| 3    | 10291| 0.01               | 1923               | 0.83               |
| 4    | 10144| 0.06               | 1831               | 0.96               |
| 5    | 10137| 0.02               | 1780               | 0.17               |
| 360  | 10972| 0.05               | 2385               | 4.64               |
| 361  | 10885| 0.36               | 2334               | 2.28               |
| 362  | 10786| 0.11               | 2240               | 2.82               |
| 363  | 10726| 0.03               | 2145               | 0.78               |
| 364  | 10657| 0.17               | 2065               | 3.47               |
| 365  | 10576| 0.19               | 2003               | 0.41               |

The units of the mean values of measured data and simulated data are $m^3/s$.

### Table 9. Partial coefficients of variation and Kendall rank coefficients.

| Date | Coefficients of Variation | Kendall Rank Coefficient |
|------|---------------------------|--------------------------|
| 1    | 0.19                      | 0.98                     |
| 2    | 0.2                       | 0.97                     |
| 3    | 0.19                      | 0.99                     |
| 4    | 0.18                      | 0.95                     |
| 5    | 0.18                      | 0.97                     |
| 300  | 0.19                      | 0.95                     |
| 301  | 0.19                      | 0.97                     |
| 302  | 0.19                      | 0.94                     |
| 303  | 0.19                      | 0.85                     |
| 304  | 0.19                      | 0.95                     |
| 305  | 0.21                      | 0.93                     |
| 360  | 0.22                      | 0.93                     |
| 361  | 0.21                      | 0.95                     |
| 362  | 0.21                      | 0.96                     |
| 363  | 0.21                      | 0.98                     |
| 364  | 0.19                      | 0.96                     |
| 365  | 0.19                      | 0.87                     |

With the simultaneous equations of Equation (22), Equation (11) and Equation (21), we can establish a stochastic model of the daily runoff based on the SVM-Copula function:

\[
F(x) = \int_0^x \frac{1}{x\sqrt{2\pi}} \exp \left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) dx
\]

\[
x_{t+1} = F_t^{-1} \left( -\frac{1}{\theta} \ln \left[ 1 - \frac{1-e^{-\theta t}}{1 + e^{-\theta t} F_t(x_{t-1})} \right] \right),
\]

where \( t = 1, 2, \ldots, 365 \).

In the above formula, \( \epsilon_{t+1} \) is a random number uniformly distributed in accordance with \((0, 1)\). It should be noted that \( x_{t-1} \) and \( F_{t-1}(x_{t-1}) \) indicate the daily runoff corresponding to the 365th day and its cumulative distribution function respectively when \( t = 1 \).

### 4.2. Model applicability test

In order to test the applicability of the model, the long-sequence method was used. We run random simulations of daily runoff based on the established model and ended up with 10000 simulation sequences. The mean values, the mean square errors, the coefficients of variations and the Kendall rank correlation coefficients of the simulated value and the measured value are calculated respectively to verify whether the simulation data of the model could maintain the characteristics of the measured data. The partial results are shown in Table 8 and Table 9.

As we can see, the errors between measured value and simulated value is relatively small for the mean values, it means that the simulated data can maintain the characteristics of the measured data well in terms of the numerical value. That the mean variances and the relative errors of variation coefficient of simulated data and the measured data are small indicates that the model accurately identifies the marginal distribution pattern of daily runoff and maintains the original characteristics of measured data in terms of the degree of dispersion. The rank correlation coefficient error of the adjacent two days between the simulated runoff value and the measured value is quite small, it indicates that the simulated daily runoff sequence generated by the model
can maintain the statistical characteristics of the measured runoff sequence in terms of autocorrelation.

By comparing the simulated data with the measured data, the simulated value of the SVM-Copula daily runoff stochastic model maintains the statistical characteristics of the measured daily runoff data in general. The model can be used for the random simulation of daily runoff.

5. Conclusion
The correct marginal distribution function of daily runoff and the appropriate correlation structure between adjacent two days runoff are the keys to realizing the random simulation of daily runoff series. In this paper, the SVM multi-classifier was constructed to automatically identify the marginal distribution pattern of each day’s runoff, and its distribution function was determined. The correlation structure between adjacent two days runoff was established based on the Frank Copula function, moreover, the SVM-Copula stochastic model of daily runoff was constructed. A case study showed that the simulated values generated by the SVM-Copula stochastic model maintain the statistical characteristic of the measured daily runoff series and the model is suitable for the stochastic simulation of daily runoff.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
The work described in this paper was supported by the Project of National Key R&D Program of China [grant number 2016YFC0402106].

References
Chen, L., & Guo, S. (2015). Copula-Based method for multisite monthly and daily streamflow simulation. Journal of Hydrology, 528, 369–384.
Chen, P., Wang, G., Yang, Y., & Zhou, J. (2006). Facial expression recognition based on rough set theory and SVM. International Conference on Rough Sets and Knowledge Technology, 4062, 772–777.
Cherkassky, V. (1995). The nature of statistical learning theory. New York, NY: Springer.
Cristianini, N., & Shawe, T. J. (2001). An introduction to support vector machines and other kernel-based learning methods. Kybernetes, 30(1), 103–115.
Diyana, K. N., Yee, O. C., Tadaaki, K., Hiroshi, O., & Fuminori, K. (2017). A fast SVM-based tongue’s colour classification aided by k-means clustering identifiers and colour attributes as computer-assisted tool for tongue diagnosis. Journal of Hydrology, 2017, 1–13.
Hao, Z., & Singh, V. P. (2012). An introduction to support vector machines and other kernel-based learning methods. Kybernetes, 30(1), 103–115.
Jangid, M., Dhir, R., Rani, R., & Singh, K. (2011). SVM classifier for recognition of handwritten Devanagari numeral. Proceedings: 2011 International Conference on Image Information Processing, 2011, 1–5.
Jeong, C., & Lee, T. (2015). Copula-based modeling and stochastic simulation of seasonal intermittent streamflow for arid regions. Journal of Hydro-Environment Research, 9(4), 604–613.
Jing, T., Cao, K. Q., Hu, L. M., & Gao, B. (2015). Application of M-FSVM in reliability life distribution pattern recognition. Journal of Astronautic Metrology and Measurement, 2015(4), 58–63.
Lee, T., & Salas, J. D. (2014). Using Copulas for stochastic streamflow generation. American Society of Civil Engineers, 316, 1–10.
Liu, Z., Zhou, P., Chen, X., & Guan, Y. (2015). A multivariate conditional model for streamflow prediction and spatial precipitation refinement. Journal of Geophysical Research: Atmospheres, 210(19), 10116–10129.
Mei, C., Yin, M. W., Hong, L., Li, M., & Li, D. Q. (2016). A comparison of two stochastic simulation methods for monthly runoff. China Rural Water and Hydropower, 12, 93–96.
Nelsen, R. B. (1999). An introduction to Copulas. New York, NY: Springer.
Pereira, G., Álvaro, V., Erhardt, T., & Czado, C. (2016). Spatial R-vine Copula for streamflow scenario simulation. Power Systems Computation Conference, 6, 20–24.
Sakthivel, N. R., Sugumaran, V., & Nair, B. B. (2010). Application of support vector machine (SVM) and proximal support vector machine (PSVM) for fault classification of monoblock centrifugal pump. International Journal of Data Analysis Techniques and Strategies, 2(1), 38–61.
Sklar, A. (1973). Random variables, joint distribution functions, and Copulas. Kybernetika (Praha), 9(6), 449–460.
Yang, A., Si, Y., Fan, Y., Kong, X., & Lin, X (2019). Correlation study of rainfall and runoff in Xiangxi river based on Archimedean Copula function. IOP Conference Series: Earth and Environmental Science, 223, 012055.
Zhou, X., Jun, S., Zhang, B., & Jun, W. (2017). Classification of different kinds of pesticide residues on lettuce based on fluorescence spectra and WT–BCC–SVM algorithm. Modern Physics Letters B, 31, 19–21.