Mid-long term runoff forecasting model based on RS-RVM

Wen Zhang[1,2], Jian Hu[3], Yintang Wang[1], Leizhi Wang[1], Lingjie Li[1] and Shiyi Cao[1]

[1] Nanjing Hydraulic Research Institute, State Key Lab of Hydrology-Water Resources and Hydraulic Engineering, 210029, Nanjing city, China
[2] Hohai University, College of Hydrology and Water Resources, 210029, Nanjing city, China
[3] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing city, 100101, China

Abstract. In view of the two key problems in hydrological mid-long term runoff forecasting- the selection of key forecasting factors and the construction of forecasting models, an analysis is made on, taking Danjiangkou Reservoir as an example, the basis of preliminarily identifying the sea-air physical factors such as atmospheric circulation, sea surface temperature and Southern Oscillation, et al. The rough set theory is used to establish the data decision table and reduce the factors, and the relevance vector machine method is adopted to establish the mid-long term runoff forecasting model based on reduced factor set. Meanwhile, this paper simulates and predicts the amount of runoff of the reservoir in September and October during the autumn floods from 1952 to 2008, and makes comparison with the model adopting support vector machine. The result shows that the relevance vector machine has better robustness and generalization performance. According to the standard of 20% annual variation, the simulation accuracy of September and October reaches 93.9% and 95.9%, respectively, and the accuracy of the trial forecasting is all up to standard. Moreover, this model better reflects the characteristics of ample flow period and low water period of the forecasting years.

1 Introduction

High-precision mid-long term hydrological forecasting plays an important role in the safety and flood control of reservoirs, effective drought relief, scientific arrangement of water resources scheduling, and improvement of hydropower generation stability and efficiency. It is also an indispensable component of efficient operation of reservoirs. Due to the complexity of the mid-long term runoff changes, there are many factors that cause it to change in the future, including a series of physical factors such as atmospheric circulation, ocean, underlying surface, astronomical earth and human activities, et al.[1-4] Due to the limitations of current science and technology, the specific physical mechanism of the above factors affecting the hydrological process is still unclear. It takes time to establish a complete physical process drive model. Therefore, at the beginning of the physical causes that affect the long-term runoff change process in the basin, finding the key predictors that effectively reflect mid-long term runoff changes, and then developing a prediction method that fully expresses the relationship between predictors and mid-long term runoff are the key to improving the accuracy of mid-long term hydrological forecast.

At present, the mid-long term hydrological forecasting is generally based on the linear correlation coefficient between the factors and the forecasting objects. There may be a nonlinear correlation between the forecasting factors and the forecasting objects, and there may be an approximate linear relationship between the forecasting factors, that is, there is complex collinearity. This causes imperfections in the extraction of forecast information, which in turn leads to instability of forecast results or deviations in forecast accuracy. Commonly used hydrological mid-long term forecasting techniques are involved in statistical methods such as multiple regression, autoregression, and time series analysis. [5-8] these techniques are based on linear changes in hydrological processes. In fact, the mid-long term hydrological process is a highly dynamic and highly nonlinear process. If only considering the linear angle or the approximate linear problem, the forecast bias will inevitably occur.

In order to improve the accuracy of mid-long term hydrological forecasting, this paper starts from the two aspects - the selection of key forecasting factors and the construction of forecasting models, and uses rough set(RS) theory and Relevance vector machine(RVM) method in statistical learning theory to explore mining key forecast information. The new approach establishes a hydrological forecasting model that can fully absorb the information of multiple forecasting factors and adaptively eliminate information redundancy. Finally, the case study of runoff forecast in the Danjiangkou Reservoir during autumn floods is carried out.

[a] Corresponding author: Jian Hu: 63621382@qq.com

2 Data and methods

2.1 Study area

This paper is illustrated with an example of the Danjiangkou Reservoir, and it forecasts the amount of runoff of the reservoir in September and October during the autumn floods.

As the source of the South-to-North Water Diversion Project, Danjiangkou Reservoir is located in the famous autumn rain area in western China. It has a total capacity of 30 billion cubic metres, a usable storage of 16.36 to 19 billion cubic metres and a control basin area of 95217 square kilometers, which is 60% of the catchment area of Hanjiang River Basin (Figure 1).

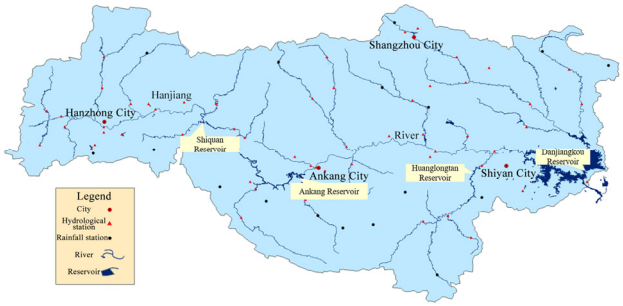

Figure 1. Schematic diagram of Danjiangkou Reservoir Control Basin

2.2 Data used

The data used in this paper mainly include inflow runoff data and climate hydrological data of Danjiangkou Reservoir, as shown below.

(1) Monthly runoff data of Danjiangkou Reservoir. Years: 1952~2008.

(2) Global 500hPa monthly average height reanalysis data provided by the US National Center for Environment Prediction (NCEP) and the US National Center for Atmospheric Research (NCAR). Years: 1948~2008, spatial resolution: 2.5°×2.5°.

(3) Global monthly sea surface temperature data provided by the National Oceanic and Atmospheric Administration (NOAA). Years: 1948~2008, spatial resolution: 2.0°×2.0°.

(4)Circulation feature data provided by the National Climate Center (NCC). Years: 1951~2008.

(5)Large-scale sea-air factor data such as ENSO, NAO, PDO, AMO, et al. Years: 1948~2008.

2.3 Methods

2.3.1 Rough set method (RS)

Rough set theory (RST) was a new data analysis tool proposed by Polish scientist Pawlak in 1982 to deal with fuzzy and uncertain information,[9,10] which has gradually attracted attention of scholars all over the world since 1990, and has become one of the most active research fields in information science. RST can not only effectively analyze and deal with inaccurate, inconsistent and other incomplete information, but also discover hidden knowledge and reveal potential laws.

This paper mainly uses the concept of approximate quality and reduction in rough set theory to identify massive hydrometeorological factors.[11] Let $X=\{X_1,X_2,\cdots,X_n\}$ be a division on the domain U, where $X_i(i=1,2,\cdots,n)$ is a category of X, $P\subseteq A$, then the approximate quality of X is defined as:

$$\gamma_p(X)=\frac{\sum_{i=1}^{n}\left|\underline{\mathrm{apr}_p}(X_i)\right|}{|U|} \tag{1}$$

Where $|\bullet|$ represents the basis of the set, and the approximate quality $\gamma_p(X)$ represents the ratio of the number of objects correctly classified by the attribute set P to the number of all objects in the system.

If the approximate quality $\gamma_p(X)=1$, then the knowledge X is completely dependent on P. If $0\le r_p(X)\le 1$, then the knowledge X relies partly on P, which reveals that only some attributes in P are available, or the attribute set is initially defective. Besides, the complementarity of $\gamma_p(X)$ gives a contradictory measure of the subset of selected data sets. If $r_p(X)=0$, then the knowledge X is completely independent of P.

When an attribute is removed from a specified set of conditional attributes, the importance of the attribute can be defined by calculating the change in the dependency. Given $Q\subseteq A$, $p\in P$, $P\subseteq A$, the attribute importance $\mathrm{sgf}(p,Q)$ is:

$$\mathrm{sgf}(p,Q)=\gamma_p(Q)-\gamma_{P-\{p\}}(Q) \tag{2}$$

The greater the dependency change, the more important p is. Therefore, attribute selection refers to excluding attributes that have no significant impact on the current pattern classification task.

Common used attribute reduction algorithm includes discernibility matrix algorithm, quick reduction algorithm, attribute reduction algorithm, and genetic algorithm, et al.

2.3.2 Relevance vector machine method (RVM)

A brief introduction of theoretical basis of RVM for Regression is provided in this section. A more detailed description on the subject is available in the paper by Tipping. [12,13] The idea of learning machines was firstly proposed by Turing (1950). Vapnik (1995) discussed the feature of learning machines and proposed Support Vector Machine (SVM) based on statistical learning. [14] Tipping (2000) put forward a Sparse Bayesian learning model like SVM.[15] However it can derive more accurate prediction and utilize dramatically fewer basis functions than SVM.[16] And when being applied in regression prediction, it can output the distribution function of predicting variable because its training is in the Bayesian probabilistic framework.

Given a set of training data $\{x_n, t_n\}_{n=1}^{N}$, (x_n is the input vector, t_n is independence target value and N is total number of data patterns), the output for RVM is as follows:

$$y(x;w)=\sum_{i=1}^{N}\omega_i K(x,x_i)+\omega_0 \quad (3)$$

Where $w=(\omega_1,\omega_2,\cdots,\omega_N)^T$ are adjustable parameters (or 'weights'), $K(x,x_i)$ is a kernel function.

$$t_n=y(x_n;w)+\varepsilon_n \quad (4)$$

Where ε_n are independent samples from some noise process which is further assumed to be mean-zero Gaussian with variance σ^2 .The likelihood of the complete data set can be written as:

$$p(t|w,\sigma^2)=(2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\|t-\Phi w\|^2\right\} \quad (5)$$

Where $t=(t_1\cdots t_N)^T$, $w=(\omega_0\cdots\omega_N)^T$ and Φ is the $N\times(N+1)$ 'design' matrix with $\Phi=[\phi(x_1),\phi(x_2),\cdots,\phi(x_N)]^T$, wherein $\phi(x_N)=[1,K(x_n,x_1),K(x_n,x_2),\ \cdots,K(x_n,x_1)]^T$.

With many parameters in the model, we would expect maximum likelihood estimation of w and σ^2 from (5) to lead to severe over-fitting. To avoid this, we impose some additional constraint on the parameters. We encode a preference for smoother (less complex) functions by making the popular choice of a zero-mean Gaussian prior distribution over w:

$$p(t|\alpha)=\prod_{i=0}^{N}N(\omega_i|0,\alpha_i^{-1}) \quad (6)$$

Where α a vector of N+1 hyperparameters.

Having defined the prior, Bayesian inference proceeds by computing from Bayes'rule, the posterior over all unknowns given the data:

$$p(w,\alpha,\sigma^2|t)=\frac{p(t|w,\alpha,\sigma^2)p(w,\alpha,\sigma^2)}{p(t)} \quad (7)$$

Then given a new test point, x_*, predictions are made for the corresponding target t_* , in terms of the predictive distribution:

$$p(t_*|t)=\int p(t_*|w,\alpha,\sigma^2)p(w,\alpha,\sigma^2|t)dwd\alpha d\sigma^2 \quad (8)$$

After finding the most optimizing hyperparameters α_{MP} and σ_{MP}^2, we can compute the predictive distribution.

$$p(t_*|t,\alpha_{MP},\sigma_{MP}^2)=\int p(t_*|w,\alpha,\sigma_{MP}^2)p(w|t,\alpha_{MP},\sigma_{MP}^2)dw \quad (9)$$

Since both terms in the integrand are Gaussian, this is readily computed, giving:

$$p(t_*|t,\alpha_{MP},\sigma_{MP}^2)=N(t_*|y_*,\sigma_*^2) \quad (10)$$

With

$$y_*=\mu^T\phi(x_*) \quad (11)$$

$$\sigma_*^2=\sigma_{MP}^2+\phi(x_*)^T\Sigma\phi(x_*) \quad (12)$$

So the predictive mean is intuitively $y(x_*;\mu)$, or the basis functions weighted by the posterior mean weights, many of which will typically be zero. The predictive variance (or 'error-bars') comprises the sum of two variance components: the estimated noise on the data and that due to the uncertainty in the prediction of the weights.

2.3.3 Research on mid-long term hydrological forecasting model based on RS-RVM

This paper proposes a method of combining RS with RVM. The RS is used to pre-process the input data, which means the RS network is used as the pre-system in advance, and then the information is predicted based on the structure of the RS. The information prediction system based on RS-RVM is shown in Figure 2.

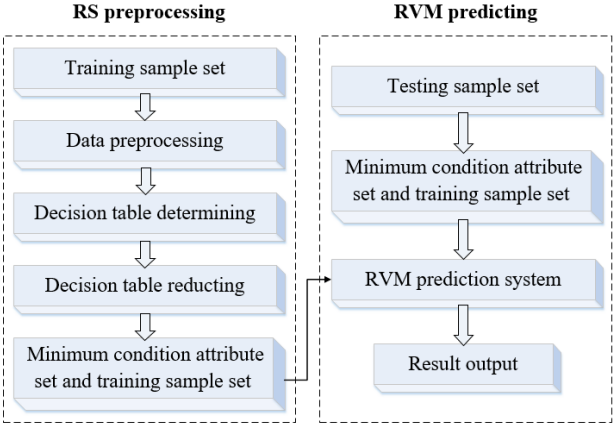

Figure 2. Information prediction system based on RS-RVM

3 Results and Discussion

3.1 Determination of preliminarily forecasting factor set

Based on the analysis of the physical background of the correlation between the early physical background and the historical runoff process of Danjiangkou Reservoir in September and October during autumn floods. Meanwhile, it preliminarily identifies the sea-air physical factors such as atmospheric circulation, sea surface temperature and Southern Oscillation(ENSO), et al. Finally, the number of primary selection factors for the runoff forecast of the Danjiangkou Reservoir in September and October was 39 and 44, respectively.

3.2 Reduction and determination of predictors

On the basis of preliminarily identifying predictors, this paper used rough set theory to establish the data decision table and reduce the factors.

The values in the decision table are generally required to be represented by symbolic data when rough set theory is used to deal with decision tables. However, the forecasting factors and runoff data used in this paper are all numerical, so the related data needs to be preprocessed first. When a conventional discretization algorithm such as equal width and equal frequency is used to convert a numerical attribute into a symbolic attribute, information loss is inevitably brought about. The result of the computational processing is highly dependent on the effect of the discretization. In order to

solve this problem, this paper uses the neighbourhood relationship model to granulate the data of each factor. When a non-empty finite set on a given real space $U=\{x_1,x_2,\cdots x_n\}$ is given, for any object x_i on U, its neighborhood of δ is: $\delta(x_i)=\{x|x\in U,\Delta(x,x_i)\le\delta\}$. Where $\delta\ge 0$, $\delta(x_i)$ is the neighborhood particle of object x_i The value range of δ is between 0.05 and 0.5, and δ=0.125 is used in this paper. The 39 factors in September and the 44 factors in October are granulated, according to the granulation data, a decision relationship table corresponding to the monthly forecast is formed.

If the number of attributes in a decision table is N, it is found that all reductions of the decision system need to test 2^N-1 subsets of attributes. When the number of attributes is too large, the amount of calculation is not tolerable. For this reason, the forwarded greedy algorithm based on attribute importance is used to reduce the granulated decision table. The steps of the numerical attribute reduction algorithm based on the neighborhood rough set model are as follows:

Input: decision table matrix and neighborhood radius; output: attribute reduction table and importance;

Step 1: Granulate the decision table;

Step 2: Initialize the reduction matrix;

Step 3: Calculate the importance of all remaining attributes;

Step 4: Select the attribute with the highest attribute importance value to add to the reduction matrix;

Step 5: If the dependency value of the reduction matrix does not change after the new attribute is added, proceed to step 6; otherwise, go to step 3;

Step 6: The program ends.

According to the calculation, the number of factors after the reduction of runoff forecast in September and October is seven and nine, respectively. The basic forecasting factors in September and October are shown in Table 1 and 2.

It can be seen from Table 1 and 2 that the correlation of each factor after reduction had passed the bilateral test with 0.05 confidence (the critical value of the correlation coefficient is 0.273 at this time), indicating that there was a significant correlation between each factor and the forecast object. The multiple correlation coefficient of the factor in September was 0.908, and in October it was 0.929, indicating that the selected factors had higher predictability. Meanwhile, according to the attribute importance of each factor, the sum of the factor attribute importance of runoff in September and October reached 0.995 and 0.992, respectively, indicating that the basic factor set after reduction basically contained all the forecasting information, and used the least predictors.

Table 1. Set of basic forecasting factors for September runoff

Serial number	Forecast factors	Correlation coefficient	Attribute importance	Factor description
1	74_11_39	0.552	0.367	November of the previous year, Eastern Pacific Subtropical High Northern Boundary Position Index
2	74_11_47	0.484	0.261	November of the previous year, Pacific Polar Vortex Area Index
3	74_5_29	0.353	0.120	May of that year, North American Subtropical High Ridge Position Index
4	500hPa_4_105	-0.352	0.102	April of that year, 500hPa potential height at point 105
5	74_10_24	-0.339	0.061	October of the previous year, North African Subtropical High Ridge Position Index
6	74_2_57	-0.335	0.042	February of that year, Northern Hemisphere Polar Vortex Central Intensity Index
7	100hPa_4_182	0.33	0.041	April of that year, 100hPa potential height at point 182

Table 2. Set of basic forecasting factors for October runoff

Serial number	Forecast factors	Correlation coefficient	Attribute importance	Factor description
1	Q_9	0.597	0.246	September of that year, inflow of reservoir
2	100hPa_3_149	-0.57	0.185	March of that year, 100hPa potential height at point 149
3	SST_9_187	-0.522	0.127	September of the previous year, North Pacific SST Index at point 187
4	100hPa_9_276	-0.406	0.098	September of the previous year, 100hPa potential height at point 276
5	SST_7_189	-0.395	0.089	July of the previous year, North Pacific SST Index at point 189
6	74_7_52	0.386	0.071	July of that year, Pacific Polar Vortex Intensity Index
7	SST_8_357	-0.366	0.066	October of the previous year, North Pacific SST Index at point 357
8	74_2_46	0.352	0.056	February of that year, Asia Polar Vortex Area Index
9	74_3_17	0.337	0.054	March of that year, Eastern Pacific Subtropical High Intensity Index

3.3 Relevance vector machine modeling

The relevance vector machine method was adopted to establish the mid-long term runoff forecastingmodel based on reduced factor set. The structure of the model is shown in Figure 3:

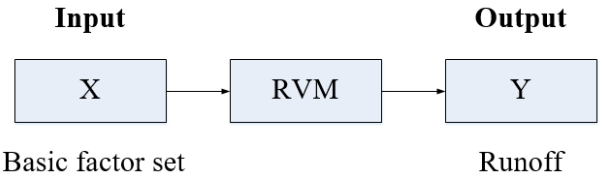

Figure 3. Mid-long term runoff forecasting model structure based on correlation vector machine

The input of the model was the reduced factor set, and the output was the runoff of the autumn floods(September and October). In order to avoid the magnitude difference between the various factors, the input data needed to be normalized first to eliminate the influence of each factor due to different dimensions and units:

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{13}$$

Where: x_i and y_i are the variables before and after normalization; $x_{\max}$ and $x_{\min}$ are the maximum and minimum values, respectively.

In the relevance vector machine, selecting different kernel functions will form different algorithms. Commonly used kernel functions include polynomial, Gaussian radial basis kernel function, B-spline kernel function and so on. Experience has shown that the Gaussian radial basis kernel function has good nonlinear processing ability. Therefore, the Gaussian radial basis kernel function is selected in this study. The specific formula is shown above. Based on the Cross-validation method, the width of the Gaussian kernel function of the runoff forecast in September and October is $\sigma = 3.5$ and $\sigma = 1.8$. This paper simulated the amount of runoff of the reservoir in September and October during the autumn floods from 1952 to 2000, and predicted the runoff from 2001 to 2008. Meanwhile, it made comparison with the model adopting support vector machine.

3.4 Forecast result analysis

According to the established model, the runoff in September and October from 1952 to 2000 was simulated, and the runoff from 2001 to 2008 was predicted, and made comparison with the model adopting support vector machine. Figures 4 and 5 show the comparison of measured and simulated values of runoff in September and October from 1952 to 2000, respectively. Figures 6 and 7 show comparison of reported and actual values of the runoff in September and October from 2001 to 2008, respectively.

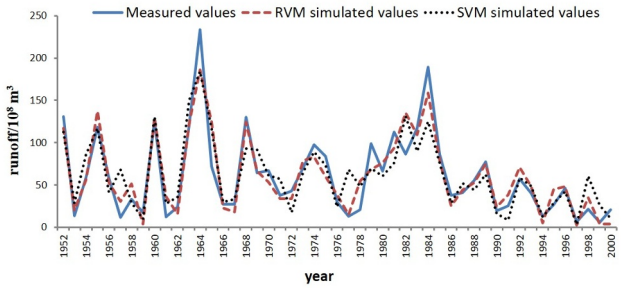

Figure 4 Comparison of simulated and measured values of runoff in September from 1952 to 2000

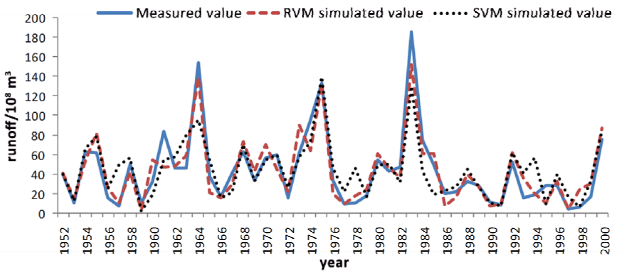

Figure 5. Comparison of simulated and measured values of runoff in October from 1952 to 2000

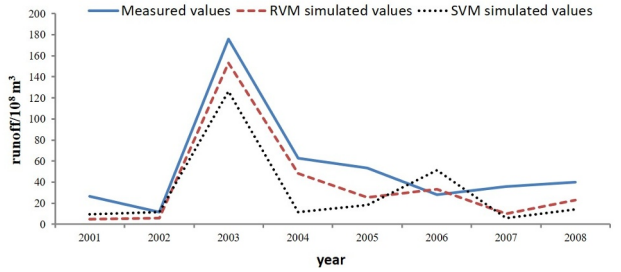

Figure 6. Comparison of predicted and measured values of runoff in September from 2001 to 2008

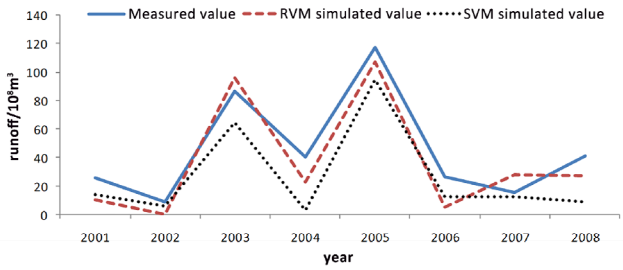

Figure 7. Comparison of predicted and measured values of runoff in October from 2001 to 2008

It can be seen from the figures that for the mid-long term forecasting of runoff of the Danjiangkou Reservoir in September and October during the autumn floods, the relevance vector machine model and the support vector machine model can both achieve good simulation, and the fitted curves of the simulated and measured values are both consistent.

In order to evaluate the prediction accuracy and performance of the prediction model, the following error criteria are used to analyze the prediction accuracy, including correlation coefficient (R), root mean square error (RMSE), and Nash efficiency coefficient (E). Meanwhile, combined with the scheme for assessing the accuracy of mid-long term forecast in *Forecasting norm*

for hydrology intelligence (SL250-2000)[17] as the evaluation criteria for forecasting model: For quantitative forecasting, the water level (flow) is 10% of annual variation, the other elements are 20% of annual variation, and the occurrence time of the elemental extremum is 30% of annual variation as the permissible error. It is expected to fully reflect the performance of the model in terms of accuracy, efficiency and error response. The correlation coefficient (R), root mean square error (RMSE), and Nash efficiency coefficient (E) are shown below.

correlation coefficient (R):

$$R=\frac{\frac{1}{n}\sum_{i=1}^{n}(Q_0(i)-\overline{Q_o})(Q_f(i)-\overline{Q_f})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_0(i)-\overline{Q_o})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_f(i)-\overline{Q_f})^2}} \tag{14}$$

root mean square error (RMSE)

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_f(i)-Q_0(i))^2} \tag{15}$$

Nash efficiency coefficient (E)

$$E=1-\frac{\sum_{i=1}^{n}(Q_0(i)-Q_f(i))^2}{\sum_{i=1}^{n}(Q_0(i)-Q_0)^2} \tag{16}$$

In the above three formulas, n is the length of the time series, $Q_0(i)$ is the measured runoff, $Q_f(i)$ is the predicted runoff, $\overline{Q_o}$ and $\overline{Q_f}$ are the mean of the measured runoff and the predicted runoff, respectively. The larger the correlation coefficient (R) and the Nash efficiency coefficient (E), the smaller the root mean square error (RMSE), indicating that the better the prediction effect is.

Tables 3 and 4 show the accuracy evaluation result of the correlation coefficient (R), root mean square error (RMSE), and the Nash efficiency coefficient (E). Table 5 shows the accuracy evaluation result of the *Forecasting norm for hydrology intelligence* (SL250-2000).[17]

Table 3. Results of the trial forecasting and accuracy analysis of runoff in September during the autumn floods

Model	Simulation period			Forecast period		
	R	RMSE	E	R	RMSE	E
RVM	0.928	18.072	0.860	0.977	18.009	0.839
SVM	0.859	24.749	0.738	0.936	19.386	0.780

Table 4. Results of the trial forecasting and accuracy analysis of runoff in October during the autumn floods

Model	Simulation period			Forecast period		
	R	RMSE	E	R	RMSE	E
RVM	0.927	13.930	0.859	0.954	14.206	0.839
SVM	0.841	20.192	0.703	0.923	21.723	0.623

Table 5. Runoff simulation and forecasting accuracy during the autumn floods from 1952 to 2000

Pass rate	September		October	
	Simulation period	Forecast period	Simulation period	Forecast period
	RVM	SVM	RVM	SVM	RVM	SVM	RVM	SVM
10% variation	83.7%	77.3%	62.5%	50%	81.6%	73.4%	75%	50%
20% variation	93.9%	87.8%	100%	75%	95.9%	91.8%	100%	87.5%

As can be seen from the evaluation results of three kinds of forecasting accuracy indicators, the RVM model is superior to the SVM model for the mid-long term forecast of runoff of the Danjiangkou Reservoir in September and October during the autumn floods. Moreover, for the runoff in September, the RVM model used only 5 correlation vectors, while the SVM model used 16 support vectors; for the runoff in October, the RVM model used only 6 correlation vectors, while the SVM model used 19 support vectors, indicating that the RVM has better prediction performance in the same case.

From the accuracy evaluation results of the *Norm*, the advantages of the RVM model are also reflected. According to the standard of 10% annual variation, the fitting qualification rate of the RVM and SVM model of historical runoff in September reached 83% and 77%, respectively. In the 8 years of the forecast period, the RVM is qualified for 5 years and the SVM is qualified for 4 years. The fitting qualification rate of the RVM and SVM model of historical runoff in October reached 81% and 73%, respectively. In the 8 years of the forecast period, the RVM is qualified for 6 years and the SVM is qualified for 4 years. According to the standard of 20% annual variation, the fitting qualification rate of the RVM and SVM model of historical runoff in September reached 93,9% and 87.8%, respectively. In the 8 years of the forecast period, the RVM is qualified for all years and the SVM is qualified for 6 years. The fitting qualification rate of the RVM and SVM model of historical runoff in October reached 95.9% and 91.8%, respectively. In the 8 years of the forecast period, the RVM is qualified for all years and the SVM is qualified for 7 years.

Overall, the RVM has better robustness and generalization performance, both the simulation and the trial report accuracy are satisfactory. Moreover, this model better reflects the characteristics of ample flow period and low water period of the forecasting years.

4 Conclusions

(1) The selection of forecasting factors is to extract effective information from a large amount of forecasting information as an input to the model. The traditional statistical methods have great defects in dealing with such problems. The rough set theory has the advantages of dealing with incomplete information, reducing massive data information and obtaining key knowledge expression. The attribute reduction algorithm using rough set theory can directly obtain the basic forecasting factor set with the highest forecasting effect, which can provide an effective forecasting information source for model establishment later. The results of the Danjiangkou Reservoir show that the basic factor set after reduction basically contains all the forecasting information and uses the least forecast factor.

(2)As a new machine learning method of statistical learning theory, RVM can effectively deal with the nonlinear relationship between forecasting factors and forecasting objects , and the nonlinear relationship among forecasting factors. The simulation results show that the RVM is, under the same noise condition, similar to the SVM for the sinc(x) function. However, the number of relevance vectors in the RVM is significantly smaller than the number in the SVM, which indicates that the RVM has better robustness and generalization performance.

(3)The RVM was adopted to simulate the runoff of the Danjiangkou Reservoir in September and October from 1952 to 2000, and conduct a trial report on runoff from 2001 to 2008. Meanwhile, this paper made comparison with the model adopting SVM. The results of four evaluation criteria show that, for the mid-long term runoff forecastingof the Danjiangkou Reservoir in September and October during autumn floods, the RVM model has better forecast performance in the same situation than the SVM model.

Acknowledgments:

This research was financially supported by the National Natural Science Foundation of China (No. 51609140, 51809252), The National Key Research and Development Program of China (No. 2016YFC0400910, 2016YFC0401502), Jiangsu Provincial Water Resources Technology Project (2017037).

References

1. Lima, Carlos HR, and Upmanu Lall. Climate informed long term seasonal forecasts of hydroenergy inflow for the Brazilian hydropower system.*Journal of hydrology*,**381**(1-2): 65-75. (2010)
2. Soukup T L., Aziz O A., Tootle G A., Piechota T C., & Wulff S S. Long lead-time streamflow forecasting of the North Platte River incorporating oceanic–atmospheric climate variability.*Journal of Hydrology*,**368**(1-4), 131-142. (2009)
3. F Gutierrez, J A Dracup. An analysis of the feasibility of long-range streamflow forecasting for Colombia using El Nino-Southern Oscillation indicators[J]. *Journal of Hydrology*, **246**: 181-196. (2001)
4. H S Yan, X D Yan, Impact of the Preceding Northern Hemisphere 500hPa Geopotential Height and Pacific SST Variation on the Flood Season Precipitation over China[J], *Chinese Journal of Atmospheric Sciences*, **28**(3): 405-413 (2004)
5. Araghinejad, Shahab, Donald H. Burn, and Mohammad Karamouz. Long-lead probabilistic forecasting of streamflow using ocean-atmospheric and hydrological predictors.*Water Resources Research*.**42**(3) (2006).
6. W C Wang, K W Chau, C T Cheng, et al. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series[J]. *Journal of hydrology*, **374**(3-4): 294-306. (2009)
7. GLENN A T, ASHOK K S, THOMAS C P, et, al. Long Lead-Time Forecasting of U.S. Streamflow Using Partial Least Squares Regression[J]. *Journal of Hydrologic Engineering*, **12**(5):442-451.(2007)
8. M Zhang, C J Li, Y C Zhang, Application of the Bayesian statistic hydrological forecast system to middle-and long-term runoff forecast[J], *Advances in Water Science*, **20**(1): 40-44, (2009)
9. Pawlak Z, Polkowski L, Skowron A. Rough Set Theory[J]. (2008)
10. Pawlak Z. Rough set theory and its applications to data analysis[J]. *Cybernetics & Systems*, **29**(7): 661-688. (1998)
11. T Y Lin, Q Liu. Rough approximate operators: axiomatic rough set theory[M]//*Rough Sets, Fuzzy Sets and Knowledge Discovery*. Springer, London, 256-260 (1994)
12. Tipping M E. The relevance vector machine[C]//*Advances in neural information processing systems*. 652-658. (2000)
13. Tipping M E. Sparse Bayesian learning and the relevance vector machine[J]. *Journal of machine learning research*, **1**(Jun): 211-244. (2001)
14. Cortes C, Vapnik V. Support-vector networks[J]. *Machine learning*, **20**(3): 273-297. (1995)
15. J Y Lin, C T Cheng. Application of support vector machine method to long-term runoff forecast[J]. *Journal of Hydraulic Engineering*, **37**(6): 681-686. (2006)
16. G R Yu, Z Q Xia. Prediction model of chaotic time series based on support vector machine and its application to runoff[J]. *Advances in Water Science*, **19**(1): 116-122. (2008)
17. W T Deng, Z B Sun, G Zeng, et al. Interdecadal variation of summer precipitation pattern over eastern China and its relationship with the North Pacific SST [J]. *Chinese Journal of Atmospheric Sciences*, **33**(4): 835-846. (2009)