Long-term runoff prediction for reservoir based on Mahalanobis distance discrimination

Zhongliang Cheng¹, Yong Liu²-³, Cheng Gao¹, Jian Hu¹, Tingting Cui²

¹ College of Hydrology and Water Resources, Hohai University, Nanjing city, 210098, China
² State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing city, 210029, 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. An accurate and timely forecast of medium and long-term runoff forecast is of great significance to reservoir safety and water resources scheduling. In order to improve the long-term runoff forecast accuracy of the reservoir, a long-term runoff forecasting model was constructed based on the principle of Mahalanobis distance discrimination analysis. The data sequence from 1952 to 2008 of Danjiangkou reservoir was selected, the correlation coefficient method and AIC criterion were used to sift out the highly correlated and independent factors, a long-term runoff forecasting model was constructed based on the principle of Mahalanobis distance discrimination analysis. The result showed that under the permutation error of 10%, the pass rate during the simulation period was 93.9%, and the pass rate during the inspection period was 87.5%. The research results serve as a reference for the operation of Danjiangkou reservoir.

1 Introduction

Runoff forecasting is one of the important application fields of hydrology, and it is also the premise of reservoir operation scheduling, flood control, drought resistance and water resources emergency dispatch [1]. Long-term runoff forecast is difficult to meet people's demand for social and economic production arrangements because of its long foresight period and low forecasting accuracy, and it has always been one of the difficulties in the field of hydrology research. There are many hydrological prediction methods, which can be roughly divided into two categories: process drive and data drive [2]. The process-driven method is a model based on the mechanism of production and flow, and is a development direction of runoff prediction. However, because the runoff is affected by many uncertain factors such as climatic meteorology, underlying surface and human activities[3], the formation mechanism and laws have not been fully grasped, and the prediction accuracy is not high, which makes the application of this method very difficult. With the development of data acquisition ability and computing power, the application of data-driven model in hydrological forecasting is more and more extensive. The prediction methods such as neural network[4] and support vector machine[5] have also achieved certain results in practical applications. In recent years, the establishment of runoff forecasting model based on physical genetic background is an important direction of current research by exploring the relationship between future runoff and large-scale climatic-hydrological variables such as pre-rainfall, sea surface temperature and atmospheric circulation index [6-9]. With the global climate change, underlying surface changes and high-intensity human activities, the applicability of traditional runoff forecasting methods has gradually deteriorated, which poses a challenge to the accurate forecasting of meteorological hydrology[10]. In order to further improve the accuracy and reliability of the forecast, this paper intends to use the combination of qualitative and quantitative forecasting methods to select forecasting factors from the physical causes of long-term runoff, and use AIC criteria to screen key factors. The long-term runoff prediction model was constructed by using Mahalanobis distance discrimination.

2 Research methods

2.1 Single correlation coefficient

Correlation coefficients are often used in hydrological mid and long-term predictions to investigate whether linear correlation between predictors and forecast objects is used as a basis for factor selection [3]. The formula for the single correlation coefficient is:

\[ r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(\bar{y} - \bar{y})^2}} \]  

(1)

Where \(X_i\) and \(Y_i\) are the series of factors and forecast objects respectively; \(\bar{X}\) and \(\bar{Y}\) are their mean values; \(n\) is the length of the sequence, \(r\) is a single correlation
coefficient, and the degree of significance is commonly used by t test to examine the reliability. The $t_\alpha$ can be found from the t-distribution table after $\alpha$ is determined. When $\alpha=t_\alpha$, it is considered that the two are linearly correlated under this reliability, otherwise it is considered to be linearly uncorrelated.

### 2.2 Akaike information criterion

The selection of the number of factors $p$ has a great influence on the prediction accuracy and stability of the model. Take a smaller $p$, the degree of fitting is poor; taking a larger $p$, it is too much affected by accidental changes, and even over-fitting phenomenon, the forecast effect will be affected. AIC is a standard for measuring the goodness of statistical model fitting. It was proposed by Japanese statistician in 1974. It is based on the concept of entropy and provides a standard for weighing the complexity of the model and the goodness of the fitted data. Now define a criterion function as:

$$AIC(p) = ln\sigma^2(p) + \frac{2p}{n}$$  \hspace{1cm} (2)

Where $\sigma^2(p)$ is the variance of the error; $p$ is the number of factors; $n$ is the number of samples. The first term represents the goodness of fit, the second term represents the penalty after the factor is added, the both are weighed, and the smallest AIC($p$) is chosen as the reasonable number of factors $p$.

### 2.3 Mahalanobis distance discrimination

The Mahalanobis distance was proposed by the Indian statistician Mahalanobis. It represents the covariance distance of the data. The Mahalanobis distance discriminant analysis method is a statistical analysis method for classifying and recognizing newly acquired samples based on the observed number of samples, and discriminating the type of the samples to effectively calculate the similarity of the sample sets.

There are $k$ m-dimensional populations $G_1, G_2, ..., G_k$, the mean vectors are $\mu_1, \mu_2, ..., \mu_k$, respectively, the covariance matrices are $\Sigma_1, \Sigma_2, ..., \Sigma_k$, $\alpha=1, 2, ..., k$, then the sample $X$ is to each group.

$$D^2(X, G_\alpha) = (X - \mu_\alpha)^T \Sigma^{-1}(X - \mu_\alpha)$$  \hspace{1cm} (3)

Discriminant rule: If $D^2(X, G_\beta) = \min_{1 \leq \alpha \leq k} D^2(X, G_\alpha)$ then $X \in G_\alpha$. For practical problems, when $\mu_1, \mu_2, ..., \mu_k$ and $\Sigma_1, \Sigma_2, ..., \Sigma_k$ are unknown, they can be replaced by the corresponding sample estimates. The specific calculation steps of the Mahalanobis distance are as follows.

Let $X^\alpha_1, X^\alpha_2, ..., X^\alpha_n$ be samples from the total $G_\alpha (\alpha=1, 2, ..., k)$, then $\mu_\alpha$ and $\Sigma(\alpha=1, 2, ..., k)$ can be estimated as:

$$\bar{X}_\alpha = \frac{1}{n_\alpha} \sum_{i=1}^{n_\alpha} X^\alpha_i \hspace{1cm} \alpha = 1, 2, ..., k$$  \hspace{1cm} (4)

$$\Sigma_\alpha = \frac{1}{n_{\alpha-1}} S_\alpha \hspace{1cm} \alpha = 1, 2, ..., k$$  \hspace{1cm} (5)

$$S_\alpha = \sum_{i=1}^{n_\alpha} (X^\alpha_i - \bar{X}_\alpha)(X^\alpha_i - \bar{X}_\alpha)^T$$  \hspace{1cm} (6)

### 2.4 Multiple linear regression

Multiple linear regression is a method for studying the correlation between a random variable and multiple variables. Multiple regression equations are established using multiple factors $X_1, X_2, ..., X_m$ and object $Y$:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m$$  \hspace{1cm} (7)

Where $b_0, b_1, ..., b_m$ are regression coefficients, and the regression coefficients of the multiple linear regression equation are determined by least squares.

### 2.5 Model accuracy assessment method

The accuracy and stability of the model are evaluated by the deterministic coefficient and the average absolute error in the Hydrological Information Forecasting Specification (GB/T22482-2008)[11]. The formula for determining the deterministic coefficient $D_c$ and the mean absolute error $e$ is as follows (8) and (9).

$$D_c = 1 - \frac{\sum_{i=1}^{n} |Y_i - \bar{Y}_0|}{\sum_{i=1}^{n} |Y_i - Y_0|}$$  \hspace{1cm} (8)

$$e = \frac{\sum_{i=1}^{n} |Y_i - \bar{Y}_0|}{\sum_{i=1}^{n} |Y_i - Y_0|}$$  \hspace{1cm} (9)

Where $D_c$ is the deterministic coefficient; $e$ is the average absolute error; $n$ is the sequence length; $Y_0$ is the predicted value; $\bar{Y}_0$ is the measured value.

### 3 Forecast application

#### 3.1 Research area

The Danjiangkou reservoir is located in the middle and upper reaches of the Han River. It is located between $106^\circ 12'-111^\circ 26'$ east longitude and $31^\circ 24'-34^\circ 11'$ north latitude. It is the water source for the Middle Route of the South-to-North Water Transfer Project. Danjiangkou reservoir has five functions: flood control, power generation, irrigation, shipping and aquaculture. It is one of the large-scale comprehensive utilization reservoirs in China, which the average annual reservoir water volume of the reservoir is 39.48 billion m$^3$ and the controlled watershed area is 95217 km$^2$.

#### 3.2 Primary selection of forecast factors

Using the collected reservoir runoff data of September from 1952 to 2008 in Danjiangkou reservoir, the monthly average sea surface temperature in the North Pacific, the 100 hPa and 500 hPa monthly mean height field in the Northern Hemisphere, and 74 monthly circulation characteristics data. The factors of the factor field were correlated and passed a significance test of...
0.05. The set of basic factors for runoff forecasting is screened out.

3.3 Model forecast

3.3.1 Mahalanobis distance discriminant analysis

The Danjiangkou reservoir's storage runoff data from September 1952 to September 2008 was divided into simulation period (1952-2000) and inspection period (2001-2008), and according to the Hydrological Information Forecasting Specification (GB/T22482-2008) [11], the inflow runoff of the Danjiangkou reservoir during the simulation period is divided into three sections according to the anomaly value, and the detailed division criteria is shown in Table 1.

| Classification | Element anomaly /% | wet | normal | dry |
|----------------|--------------------|-----|--------|-----|
|                | < -20              |     |        |     |
|                | [-20, 20]          |     |        |     |
|                | > 20               |     |        |     |

The selection of key predictors was carried out by using the AIC. The key factors selected are shown in Table 2.

| Runoff category | Serial number | factors         |
|-----------------|---------------|-----------------|
| Wet(1)          | X₁            | SST_3_460       |
|                 | X₂            | SST_4_541       |
|                 | X₃            | 74_2_57         |
| Normal(2)       | X₄            | 74_10_24        |
|                 | X₅            | 74_2_57         |
| Dry(3)          | X₆            | 100hPa_4_182    |
|                 | X₇            | 500hPa_2_226    |
|                 |               | SST_6_222       |

The piecewise linear regression equations established by the multiple linear regression method according to the AIC criteria are as follows:

\[
Y = 433.38 - 33.233X₁ + 41.611X₂ - 3.232X₃ + 80.42 + 3.618X₄ - 2.694X₅ - 0.512X₆ - 1643.37 + 0.053X₇ + 0.065X₈ + 17.855X₉ (10)
\]

According to the established regression equation, the simulation of the inflow runoff in September of the simulation period from 1952 to 2000 can be carried out. According to key factors, the discriminant analysis results are shown in Table 3.

| year | D²( X, G₁) | D²( X, G₂) | D²( X, G₃) |
|------|------------|------------|------------|
| 1952 | 5.96       | 91.84      | 6.67       |
| 1953 | 8.17       | 40.24      | 4.52       |
| 1954 | 16.49      | 7.10       | 5.71       |
| 1955 | 3.43       | 3.80       | 5.41       |
|      |            |            |            |
| 2005 | 35.82      | 13.26      | 8.97       |
| 2006 | 16.37      | 19.29      | 5.80       |
| 2007 | 18.04      | 7.13       | 5.03       |
| 2008 | 6.34       | 5.52       | 1.44       |

In the 49-year period of the training period from 1952 to 2000, 43 years pass the test, the pass rate is 87.8%. In the test period from 2001 to 2008, 7 years pass the test, and the pass rate is 87.5%.

3.3.2 Forecast result analysis

The simulation period forecast result is shown in Figure 1.

![Figure 1. measured and simulated runoff value in September.](https://doi.org/10.1051/matecconf/201824602028)
coefficient DC is 0.93 and average absolute error e is 0.13.

Table 4. Statistical table of measured and predicted values in the test period.

| Forecast year | Measured ($\times 10^8$ m$^3$) | Forecast ($\times 10^8$ m$^3$) | error | Accuracy assessment |
|---------------|-------------------------------|-------------------------------|-------|---------------------|
| 2001          | 26.00                         | 22.65                         | -3.35 | qualified           |
| 2002          | 11.48                         | 30.29                         | 18.81 | qualified           |
| 2003          | 175.93                        | 137.86                        | -38.07| Unqualified         |
| 2004          | 62.56                         | 64.32                         | 1.76  | qualified           |
| 2005          | 53.36                         | 43.15                         | -10.21| qualified           |
| 2006          | 28.80                         | 32.83                         | 4.03  | qualified           |
| 2007          | 35.48                         | 22.14                         | -13.34| qualified           |
| 2008          | 39.71                         | 28.69                         | -11.02| qualified           |

With 10% of the multi-year variable as the license error, in the 8 years of the inspection period, only the error in 2003 exceed the allowable value, the qualified rate of the forecast is 87.5%; when the permissible change is 20% of the multi-year change, all the license errors are met the inspection period in 8 years, the pass rate is 100%.

4 Conclusion
a) The forecast application result shows that the prediction model based on Mahalanobis distance has good prediction accuracy and stability.
b) The key to improving the accuracy of forecasting is to improve the representativeness of factors rather than just increasing the number of factors. Using AIC criteria to determine the number of factors can help improve the accuracy and stability of the forecast.
c) The correlation analysis between forecasting factors and the analysis of the causal mechanism between forecasting factors and forecasting objects should be strengthened to further improve the representativeness of forecasting factors and reduce the error of forecasting in the future.

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

References
1. Z. Tongtiegang, Y. Dawen, L. Mingliang, Exceedance probability method for mid-term and long-term streamflow prediction[J], Journal of Water Resources, 42(6), 692-699(2011).
2. W. Wen, M. Jun, Summary of several hydrological forecasting methods[J], Advances in Science and Technology of Water Resources, 25(1), 56-60(2005).
3. T. Chengyou, G. Xuewen, Z. Shiming, Modern medium and long term hydrological forecasting method and its application[M], China Water & Power Press, (2008).
4. Z. Wenxiu, Z. Xiaoli, L. Guohui, Research on the Long-Term Runoff Forecast Based on Random Forest Model and RBF Network[J], Yellow River, 37(2), 10-12(2015).
5. L. Jianyi, C. Chuntian, Application of support vector machine method to long-term runoff forecast[J], Journal of Water Resources, 37(6), 681-686(2006).
6. F. Xiaochong, W. Yintang, L. Yong, Monthly runoff forecast for Danjiangkou reservoir based on physical statistical methods[J], Journal of Hohai University(Natural Sciences), 39(3), 242-247(2011).
7. L. Yong, W. Yintang, C. Yuanfang, Long-term runoff forecasting for autumn flooding seasons in Danjiangkou reservoir based on analyzing the physical causes[J], Advances in Water Science, 21(6): 771-778(2010).
8. X. Xiaoyan, T. Haihua, W. Jianping, Runoff Forecasting Model in Flood Season Based on Weather Factors, Water Resources and Power, 33(10), 10-12(2015).
9. G. Zhaoxia, X. Mei, S. Yingling, Application of multi-factor stepwise regression cycle analysis in medium and long-term hydrological forecast[J], Journal of Hohai University(Natural Sciences), 37(3), 255-257(2009).
10. Z. Jianyun, Review and reflection on China's hydrological forecasting techniques[J], Advances in Water Science, 21(4), 435-443(2010).
11. GB/T22482-2008. Standard for hydrological information and hydrological forecasting[S], China Standards Press, (2009).