Mid and long-term hydrological classification forecasting model based on KDE-BDA and its application research

LIU Yong¹, YIN Zhijie², ZHANG Ying³, WANG Qin¹

¹ Nanjing Hydraulic Research Institute, State Key Lab of Hydrology-Water Resources and Hydraulic Engineering, Nanjing, 210029, China;
² Information Center (Hydrology Monitor and Forecast Center), Ministry of Water Resources, Beijing, 100053, China;
³ Hydraulic Engineering Planning Bureau of Jiangsu Province, Nanjing, 210029, China

*Corresponding author: LIU Yong: yongliu@nhri.cn

Abstract. In view of the uncertainties in the current mid and long-term hydrological forecast results, a mid and long-term hydrological classification forecasting model based on KDE-BDA is built and its application research is performed. The kernel density estimation method (KDE) is used to improve the distribution density function estimation method in the conventional Bayesian discriminant method (BDA). Based on this, a mid and long-term hydrological classification prediction model based on KDE-BDA was built. An example of its application is demonstrated in the runoff forecast of the Danjiangkou Reservoir in the autumn of September. The results show that different forecasting factors have different forecasting ability for runoff classification (flood, normal, dry). Among them, the 500hPa height field factor has higher forecasting ability for the Danjiangkou runoff category (flood, normal, dry), and the 100hPa height field factor, SST and circulation characteristics have a good indication of the runoff’s flood state. The results of Half-Brier score show that the combined forecasting model combines the forecasting advantages of each factor, so the forecasting effect is the best. The pass rate in the simulation period was 89.8%, and the pass rate in the inspection period was 87.5%. The simulation and forecast results were relatively stable.

1. Introduction
Accurate and timely mid and long-term hydrological forecasting is of great significance for the initiative to fight against floods and droughts. It is of great significance to formulate scientific water resources dispatching schemes to ensure the safe operation and scheduling of water conservancy facilities and to exert their economic benefits. Scholars at home and abroad have been conducting extensive research on mid and long-term hydrological forecasting [1-10]. There are many methods such as regression analysis [1, 2], time series analysis [3, 4], neural network [5, 6], support vector machine [7, 8], fuzzy set theory [9, 10]. However, most of the method prediction results are deterministic quantitative prediction values. The simulation of mid and long-term hydrological processes is affected by a large number of deterministic and uncertain factors such as astronomy, meteorology, underlying surface and human activities. In addition, the hydrological model is only a simulation of the objective hydrological process, which is not completely accurate. Therefore, there is a great deal of uncertainty as a quantitative forecast value for the model output. According to a certain
principle, the forecast object is divided into multiple attribute levels, and different attribute levels are regarded as different modes, and the attribute level of the forecast object is determined by a pattern recognition or reasoning method such as neural network, fuzzy reasoning, etc. Discriminating the attribute level of the forecast object based on reasonable forecasting factors using pattern recognition or reasoning methods such as neural network and fuzzy inference. It plays a very important role in the efficient use of water resources. In view of this, this paper uses the kernel density estimation method to improve the distribution density function estimation in Bayesian discriminant analysis method. On this basis, a mid and long-term hydrological classification prediction model based on KDE-BDA was constructed and applied in Danjiangkou Reservoir to verify the effectiveness of the proposed method.

2. Bayesian discriminant method and its improvement

2.1. Bayesian discriminant method

Discriminant analysis [11], which was produced in the 1930s, is a statistical method for discriminating models from unknown categories using samples of known categories. The basic idea of the Bayesian discriminant method is: assuming that there is a certain understanding of the research object, the prior probability distribution is often used to describe this understanding; Then, a sample is extracted, and the existing prior probability distribution is corrected by the sample, and the posterior probability distribution is obtained. Combined with the misjudgment loss function, the expected discriminant loss is obtained, and the discriminant method that minimizes the estimated cost of misclassification (ECM) is called Bayes discriminant analysis (BDA).

Assuming that there are m units: \(G_1, G_2, \ldots, G_m\), the prior probabilities are \(q_1, q_2, \ldots, q_m\), the probability density function is \(f_1(x), f_2(x), \ldots, f_m(x)\), according to the Bayesian formula, for any sample \(x\) the posterior probability from the \(g\)th unit is:

\[
P(g/x) = \frac{q_g f_g(x)}{\sum_{g=1}^{m} q_g f_g(x)}, \quad g = 1, 2, \ldots, m
\]

When the sample \(x\) is judged to come from a certain population, it is likely to come from any of the \(m\) populations. The average loss from the \(h\)-th overall of any sample, ie the expected loss is defined as:

\[
E(h/x) = \sum_{g=1}^{m} P(g/x) L(h/g)
\]

Among them, \(L(h/g)\) as the loss function, the sample which is originally the \(g\)-th overall is misjudged as the loss of the \(h\)-th overall. When analyzing actual problems, the losses are often considered to be exactly equal, such as:

\[
L(h/g) = 1, \quad h \neq g
\]

Therefore, the expected loss of \(x\) to the \(h\)-th overall is:

\[
E(h/x) = \sum_{g=1}^{m} \frac{q_g f_g(x)}{\sum_{g=h}^{m} q_g f_g(x)}
\]

To minimize the average loss:

\[
\frac{q_h f_h(x)}{\sum_{g=1}^{m} q_g f_g(x)} \rightarrow \text{maximum}
\]

Because the different types of denominators are consistent, the above formula is equivalent to find out the maximum of (6)

\[
q_h f_h(x) \rightarrow \text{maximum}
\]
2.2. Improvement of Bayesian Discriminant Method Based on Kernel Density Estimation

In the above discriminant criteria, the commonly used solution method assumes that the general types obey a certain distribution and generally take a normal distribution. In fact, the statistical analysis assumes that the premise of normal distribution is based on large number of samples. However, large samples are not only related to research problems, but it is also difficult to obtain large enough samples due to conditional constraints. Therefore, if the probability distribution hypothesis is inconsistent with the actual overall situation, a false positive will occur. There is a positive relationship between the false positive rate and the distribution probability deviation.

Therefore, based on the kernel density estimation theory, the distribution density function estimation in Bayesian discriminant method is improved for the defects existing in Bayesian discriminant method. The Kernel Density Estimation (KDE) algorithm is a simple and effective non-parametric density estimation method. KDE does not introduce a prior assumption about the data distribution. It only obtains the distribution characteristics of the data from the training samples themselves. It can be used to estimate the density function of any shape, and has obvious advantages compared with the parametric method and the semiparametric method. The following is a brief introduction to the nuclear density estimation method.

Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed random variable of value and $R$, and the distribution density function it obeys is $f(x), x \in R$, define the function:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - X}{h}\right), x \in R$$

(7)

The above equation is called the kernel density estimate of the density function $f(x)$, where $K(\cdot)$ is the kernel function; $h$ is a predetermined positive number, commonly referred to as window width or smoothness parameters.

For convenience, let $K_h(u) = K(u/h)h$, the above formula can be expressed as:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x_i - x), x \in R$$

(8)

It can be seen from the formula that the kernel density estimation of the distribution density function $f(x)$ is not only related to a given set of sample points, but also related to the selection of the kernel function and the selection of the bandwidth parameters. Among them, the bandwidth parameter $h$ controls the degree of influence of the sample points on the point density at different distances when the approximate density at the point $h$ is obtained.

In theory, any function can be used as a kernel function, but in order to estimate the convenience and rationality of the density function, the kernel function is usually required to satisfy the following conditions:

$$K(-u) = K(u)$$

(9)

$$\sup_u K(u) = K(u) < \infty, \quad \int_{-\infty}^{\infty} K(u)du = 1$$

(10)

Commonly used kernel density functions include Gaussian kernel function, Epanechnikov kernel function, Biweight kernel function and Triweight kernel function.

3. Construction of mid and long-term hydrological classification forecasting model based on KDE-BDA

The KDE-BDA-based mid and long-term hydrological classification prediction model can be used to classify and forecast runoff. The main steps are as follows:

1. Selecting forecasting factor: according to the forecast object, using the method of combining statistical and physical genesis in the literature to screen out the forecasting factor;
② Divide the runoff category: sort the forecast objects from large to small, calculate the corresponding experience frequency, and classify the runoff into flood, normal and dry according to Table 1;

| Grading   | Dry       | Normal    | Flood     |
|-----------|-----------|-----------|-----------|
| Frequency (P) | P ≥ 75% | 25% ≤ P < 75% | P < 25% |

③ Prior probability calculation: According to the runoff classification, calculate the prior probability $p_i (i = 1, 2, 3)$ of each category: $p_1 = P(Q \geq Q_1)$ , $p_2 = P(Q_0 \leq Q < Q_1)$ , and $p_3 = P(Q < Q_0)$.

④ The posterior probability calculation: the corresponding forecasting factors are divided into three categories. For each factor and forecast object, the Bayesian discriminant method (BDA) is used to establish the forecast model, and the probability of occurrence of different types of runoff in the case of the occurrence of the factor is calculated. The formula is as follows:

$$P(Q_i | X) = \frac{p_i f_i(x)}{\sum_{i=1}^{r} p_i f_i(x)},$$

(11)

Where: $Q_i$ is the runoff of various categories, $r$ is the number of runoff categories, $X$ is the forecast factor, $f_i(x)$ is the probability density of the corresponding category factor. Using the kernel density method of the previous section to estimate, the Gaussian kernel function is used here.

⑤ Integration of forecast results: Calculate the optimal combined forecast results $P_{OCF}$ based on the forecast results calculated for each factor.

$$P_{OCF} = \sum_{i=1}^{m} a_i \times P_{X_i},$$

(12)

Where: $m$ is the number of factors, $a_i$ is the parameters optimized for each factor, $P_{X_i}$ is the posterior probability of each factor forecast. The parameters of each predictor were optimized using accelerated genetic algorithm (RAGA).

⑥ Forecast result score: The Half-Brier scoring method proposed by Brier and Allen is used to evaluate the forecast results and combined forecast results of each factor. The formula is as follows:

$$B = \frac{1}{N} \sum_{j=1}^{r} \sum_{i=1}^{N} (\delta_{ij} - \phi_j)^2,$$

(13)

Where: $N$ is the year of the data for runoff, $r$ is the number of runoff classifications, $\phi_j$ is the probability that the event occurred in the $j$ category, $\delta_{ij}$ is 1 or 0, depending on whether the event occurred in the $j$ category.

The Half-Brier score is in fact a root means square error for a classification forecast. The value is 0 when the forecast is completely correct and the value is 2 when the forecast is completely wrong.

The calculation process of the mid and long-term hydrological classification forecasting model based on KDE-BDA can be summarized as the following figure 1.
4. Case study

4.1. Forecast object and factor selection

Taking the Danjiangkou Reservoir inflow runoff forecast as an example, a case study was carried out. The Danjiangkou Reservoir (Fig. 2) is located in the autumn rain area of Huaxi, China, and is the water source for the Middle Route of the South-to-North Water Transfer Project. Each year, September and October are the transition seasons from midsummer to early winter. The Danjiangkou Reservoir Basin often has continuous rainy weather and autumn floods. Rainfall has the characteristics of high intensity and concentrated rain area. If the water source is improperly dispatched, it will pose a threat to the reservoir, reservoir area safety and the normal operation of the entire South-to-North Water Transfer Central Line project. Therefore, the practical significance of carrying out the mid and long-term forecast of runoff in the Danjiangkou Reservoir during the autumn flood season is significant.
The used forecasting factors include three element fields: the North Pacific Sea Surface Temperature Field (SST), the 100hpa monthly mean height field (H100hPa) in the Northern Hemisphere, the 500hpa monthly average height field (H500hPa) in the Northern Hemisphere. 74 meteorological factors of circulation characteristics were used as screening influence factors.

In addition, considering the correlation of runoff in the early stage, the previous runoff is also used as an alternative factor to establish a set of factors for runoff forecasting. The forecasting factor set basically reflects the general trend of the formation and development of previous weather processes and has a good physical background.

### 4.2. Selection of key predictors

The selection predictor is the process of selecting effective predictors from a large number of physical factors. By analyzing the correlation coefficient between the Danjiangkou Reservoir inflow runoff and the previous influencing factors, the factors with large correlation coefficient (beyond certain confidence) and having the genetic factors are selected as the candidate factors. In this study, the grid data of SST, 500hpa and 100hpa, the data of 74 circulation characteristics and the previous runoff data were used to conduct a general survey with the Danjiangkou Reservoir in September. At the time of the census, all the forecasting factors from January of the previous year to the previous month of the forecast month were correlated with the runoff of the forecast month. At the same time, combined with the method in the literature [21], the key predictors were selected. The basic factor set of the September runoff forecast for the Danjiangkou Reservoir is shown in Table 2.

**Table 2. Set of basic forecasting factors for runoff in Danjiangkou Reservoir in September**

| Number | Forecast factor | Correlation coefficient | Factor description |
|--------|-----------------|-------------------------|--------------------|
| 1      | 100hpa_4_252    | 0.458                   | 100hPa geopotential height at 252th point in April of that year |
| 2      | 100hpa_4_182    | 0.471                   | 100hPa geopotential height at 182th point in April of that year |
| 3      | 500hpa_11_126   | -0.468                  | 500hPa geopotential height at 126th point in November of last year |
| 4      | 500hpa_4_105    | -0.451                  | 500hPa geopotential height at the 105th point in April of that year |
| 5      | 500hpa_2_266    | -0.402                  | 500hPa geopotential height at the 266th point in February of that year |
The sea temperature at 412 points in June of that year

The sea temperature at 418 points in March of that year

Pacific vortex area index in November of last year

North Africa's subtropical high ridge line in October of last year

Inbound logistics in May of that year

4.3. Performance analysis of forecasting factors

Four different predictive factor fields were used, and the effective information provided by them differed. In order to make full use of the effective forecast information of each factor, a single factor is selected for classification and forecasting, and then the linear combination is used to optimally combine the classification and forecasting of each factor. In order to show the ability of different forecasting factors to predict and forecast the forecasting objects, the probability density function of different factors of the Danjiangkou Reservoir classification forecast is plotted below, as shown in Figure 3.

Fig 3. Probability density function diagram of different factors for classification and forecasting of Danjiangkou Reservoir
It can be seen from Fig. 3 that for the Danjiangkou Reservoir, the 500hPa height field factor has a higher forecasting ability for the three classifications in the four factor fields. When the density function is left-biased, that is, the height value is too small, the runoff will be flood; When the density function is right-biased, that is, when the height value is too large, the runoff is excessive. When the density function is centered, the later runoff tends to be normal. The 100hPa height field factor is also more capable of forecasting runoff and aperiodic years, but not good for forecasting the years of dry. The forecasting ability of the North Pacific SST and the circulation characteristics is similar, and both can effectively predict the flood of runoff, but it cannot be effectively distinguished between dry and normal conditions.

4.4. Analysis of forecast results
It can be seen from the above analysis that the classification and prediction capabilities of different factors are different. Therefore, the classification combined forecasting model is used to predict the optimal combination of the forecast results of each factor. Firstly, the Half-Brier scoring method is used to evaluate the classification forecast results of each factor, and the higher forecasting factor is selected as the optimal combined forecasting factor. Then, the accelerated genetic algorithm (RAGA) is used to optimize the parameters of each forecasting factor in the optimal combined model. Finally, five optimal forecasting factors were selected for Danjiangkou Reservoir, and Table 3 is the improved parameters for optimal forecasting factors.

Table 3. Optimization index of optimal forecasting factors for runoff forecast of Danjiangkou Reservoir.

| Forecast factor | 100hpa_4_252 | 500hpa_4_105 | SST_6_412 | SST_3_418 | 74_11_47 |
|----------------|--------------|--------------|-----------|-----------|---------|
| coefficient    | 0.020        | 0.335        | 0.144     | 0.318     | 0.182   |

According to the established model, the simulated runoff of the Danjiangkou Reservoir in September from 1952 to 2000 was simulated and classified. The output is shown in Table 4. At the same time, the model was tested using the data from 2001 to 2008, and the results are shown in Table 5.

Table 4. Simulation results of runoff simulation of Danjiangkou Reservoir from 1952 to 2000

| Year | Measured value (10^8 m^3) | Actual category | Forecast probability | Accuracy assessment |
|------|---------------------------|-----------------|----------------------|---------------------|
|      |                           |                 | flood | normal | dry   |          |
| 1952 | 130.92                    | flood           | 0.5625 | 0.3117 | 0.1258 | √        |
| 1953 | 13.22                     | dry             | 0.0897 | 0.4085 | 0.5019 | √        |
| 1999 | 6.27                      | dry             | 0.1326 | 0.3964 | 0.4710 | √        |
| 2000 | 20.38                     | dry             | 0.1115 | 0.5676 | 0.3209 | ×        |

Table 5. Forecast results of runoff classification of Danjiangkou Reservoir from 2001 to 2008

| Year | Measured value (10^8 m^3) | Actual category | Forecast probability | Accuracy assessment |
|------|---------------------------|-----------------|----------------------|---------------------|
|      |                           |                 | flood | normal | dry   |          |
| 2001 | 26.625                    | normal          | 0.0694 | 0.6503 | 0.2803 | √        |
| 2002 | 11.664                    | dry             | 0.1879 | 0.3437 | 0.4683 | √        |
It can be seen from the results of Table 4 and Table 5: For the classification forecast of Danjiangkou Reservoir, there are 5 years of error in the simulation period of 49 years, and the qualified rate of fitting is 89.8%. In the 8 years of the inspection period, the only error was reported in 2008, and the pass rate was 87.5%. From the classification of the error report, 4 years in the 5 years of the error period is the state of runoff, and the runoff in 2008 is also dry, indicating that the forecast model has a higher forecast for the runoff of Danjiangkou Reservoir. From the classification of the error report, 4 years in the 5 years of the error period is in the state of runoff, and the runoff in 2008 is also dry, indicating that the forecasting model has a higher forecasting effect on the flood state of runoff in Danjiangkou Reservoir, while the forecasting performance is slightly worse for runoff. The reason may be that the factors affecting the early years of Danjiangkou dry are more complicated and lack of indicative predictors.

In addition, in order to compare the forecast performance differences between the single factor forecast model and the combined forecast model, the Half-Brier score table for the runoff forecast of Danjiangkou Reservoir is listed below. It can be seen that in the runoff forecast of each factor of Danjiangkou Reservoir, the 500hPa height field factor and sea temperature factor have better prediction effects. The forecast performance of the 100hPa geopotential and circulation feature quantities is the second, and the combined forecast combines the forecasting advantages of each factor, and its effect is better than the single factor.

| Factor    | Score | 100hpa_4_252 | 500hpa_4_105 | SST_6_412 | SST_3_418 | 74_11_47 | Combined forecast |
|-----------|-------|--------------|--------------|-----------|-----------|----------|------------------|
| 2003      | 0.4972| 0.3504       | 0.1524       |           |           |          |                  |
| 2004      | 0.0649| 0.6239       | 0.3112       |           |           |          |                  |
| 2005      | 0.1205| 0.5174       | 0.3621       |           |           |          |                  |
| 2006      | 0.0686| 0.6236       | 0.3079       |           |           |          |                  |
| 2007      | 0.0740| 0.6205       | 0.3055       |           |           |          |                  |
| 2008      | 0.2733| 0.4563       | 0.2704       |           |           |          |                  |

5. Conclusion
In this paper, a mid and long-term hydrological classification forecasting model based on KDE-BDA is established, and an example of its application is demonstrated with regard to the runoff forecast of the Danjiangkou Reservoir in autumn. The main conclusions are as follows:

(1) Among the forecasting factors, the 500hPa height field factor has higher forecasting ability for the three classifications of Danjiangkou runoff (flood, normal, dry); The 100hPa geopotential factor has a strong ability to predict the flood of runoff and normal years, and the forecasting ability of the North Pacific SST and circulation characteristics is similar, which can effectively predict the flood of runoff.

(2) The forecasting model has achieved good results in the runoff forecast of the autumn flood season in Danjiangkou. Among them, the qualified rate of the classification forecasting period is 89.8%, and the passing rate of the inspection period is 87.5%. The forecast result is relatively stable.

(3) The forecasting model has a higher forecasting effect on the runoff state of the Danjiangkou Reservoir, while the forecasting performance is slightly worse for the dry of runoff. The main reason is that the influencing factors of Danjiangkou dry year are more complicated and lack of indicative predictors.

(4) The results of Half-Brier score show that the combined forecasting model combines the forecasting advantages of each factor, and its effect is better than that of single factor.
Acknowledgments
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).

References
[1] 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, 2007, 12(5):442-451.
[2] GE Zhaoxia, XUE Mei, SONG Yingling. Application of multi-factor stepwise regression cycle analysis in medium and long-term hydrological forecast [J]. Journal of Hohai University (Natural Sciences), 2009, 37(3): 255-257.
[3] TANG Chengyou, GUO Lijuan, WANG Rui. Application of prediction model for stochastic combination of stepwise regression of hydrologic time series [J]. Water Resources and Hydropower Engineering, 2007, 38(6): 1-4.
[4] CHENG C T, XIE J X, CHAU KW, et, al. A new indirect multi-step-ahead prediction model for a long-term hydrologic prediction [J]. Journal of Hydrology, 2008, 361: 118-130.
[5] JAIN A, KUMAR A M, Hybrid neural network models for hydrologic time series forecasting [J]. Applied Soft Computing, 2007, 7(2): 585-592.
[6] Hsu K, Gupta H Vand Sorroshians. Artificial neural network modeling of the rainfall runoff. Water Resources Research, 1995, 31(10): 2517-2530.
[7] LI Yanbin, HUANG Qiang, XU Jianxin, ZHANG Zezhong. Support vector machine model of mid-long term prediction of runoff in stream [J] Journal of hydroelectric engineering, 2008, 27(5): 28-32.
[8] LIN Jianyi, CHENG Chuntian. Application of support vector machine method to long-term runoff forecast [J]. Journal of Hydraulic Engineering, 2006, 37(6): 681-686.
[9] FIRAT M, GUNGOR M. Hydrological time-series modeling using an adaptive neuro-fuzzy inference system [J] Hydrological Processes. 2008, 22(13): 2122-2132.
[10] ZHU Yongying, ZHOU Huicheng, PENG Hui. Rough fuzzy inference model and its application for mid-long term hydrological forecast [J]. Journal of hydroelectric engineering, 2009, 28(1): 45-50.
[11] Piechota, T.C., Chiew, F.H.S., Dracup, J.A., McMahon, T.A., Development of exceedance probability streamflow forecast[J]. Journal of Hydrologic Engineering, 2001, 6 (1): 20-28.