Climate Variability and Change—Hydrological Impacts (Proceedings of the Fifth FRIEND World Conference held at Havana, Cuba, November 2006), IAHS Publ. 308, 2006.

Application of a statistical method for medium-term rainfall prediction

ZHONGMIN LIANG[1], BO LU[2] & XIAOFAN ZENG[3]

1 *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, P.R. China*
zmliang@hhu.edu.cn

2 *Department of Civil & Architectural Engineering, Drexel University, Pennsylvania 19104, USA*

3 *Nanjing Institute of Geography and Limnology, Nanjing 210098, P.R. China*

Abstract This paper presents a statistical approach for prediction of medium-term rainfall class based on the correlation between rainfall and meteorological indicators, including geopotential height, temperature, dew-point deficit, wind direction and wind velocity. Total rainfall of a 10-day period during the rain season (July and August) is classified into two types (dry or wet) using the *K*-mean method. Meteorological indicators that influence rainfall are selected by the *F*-test method for rainfall type and rainfall class, and then used with a bi- or multi-discriminant method to establish prediction models. This procedure is applied to predict rainfall class with a three-day lead time for the Yuecheng basin in China. Results show that it is effective in medium-term rainfall prediction with relatively little data requirement.

Key words medium-term rainfall prediction; antecedent influencing factor (AIF); *K*-mean method; *F*-test method; multi-discriminant method

INTRODUCTION

Medium-term rainfall prediction, normally with a lead time of 3–10 days, plays an important role in flood prevention and drought reduction, and thus has economic benefits for water resources management. Nowadays, medium-term rainfall prediction is still developing and has not yet reached the skills of short-term hydrological forecasting, due to the uncertainty in the driving factors with increased lead time. Even though methods based on numerical weather forecasting have improved, their strict dependence on meteorological and climatic data limits their practical applications in most cases (Zwiers & Van Storch, 2004), thus a method with little data requirement should be of more practical use.

In this region of China (the Yuecheng basin), rainfall types (i.e. wet and dry) are considered to be different representations of different rainfall mechanisms. Therefore, the method suggested here first classifies rainfall into two types, and then, for each of these types, defines distinct models to predict how heavy/light the rainfall is.

The approach used for medium-term rainfall forecasting with a lead time of three days is based on statistical relations between rainfall and meteorological indicators. The procedure includes the following steps: (a) Calculate mean areal rainfall from the historical rainfall data of stations located within the study basin using the Thiessen polygon method (Thiessen, 1911), and calculate rainfall totals for separate 10-day periods of July and August. (b) Classify each 10-day rainfall into dry or wet type using the *K*-mean method. (c) For each 10-day period, select the Antecedent Influencing Factors (AIF) amongst the meteorological indicators showing the highest correlation with rainfall totals by the *F*-test method. (d) Define rainfall type prediction equations using the bi-discriminant method (Hand, 1981) with the AIF. (e) Establish daily rainfall class prediction models for both dry and wet rainfall types for each 10-day period. Four rainfall classes (P) were defined: no rain (i.e. $P < 0.1$mm), little rain ($0.1 \le P < 10$ mm), medium rain ($10 \le P < 25$ mm) and heavy rain ($P > 25$mm). For a given 10-day period, for example the first ten days of July, two groups (wet or dry type) are defined based on measured rainfall totals, and for each dry and wet group the corresponding AIF are selected. Daily rainfall class prediction models are then derived from a multi-discriminant approach (Joachimsthaler & Stam, 1988). This method can be applied for any day in July and August using the rainfall type prediction equation of the 10-day period, including that day to predict rainfall type first, and then using the rainfall class prediction equation in accord with the predicted rainfall type to predict the rainfall class of that day. It was tested on the Yuecheng reservoir basin.

METHODOLOGY

Mean areal rainfall calculation

Daily mean areal rainfall is computed by the Thiessen method with historical rainfall data from stations within the basin. The whole basin is divided into as many Thiessen polygons as rainfall stations, and areal rainfall for each polygon is considered equal to the rainfall of the station located within this polygon. The mean areal rainfall of the whole basin is given by:

$$\overline{P} = \frac{1}{A}\sum_{i=1}^{n} P_i a_i \tag{1}$$

where P_i and a_i are the mean areal rainfall and area of the polygon i respectively, while A represents the total area of the basin.

Classify rainfall type of historical data

Historical mean areal rainfall for each 10-day period of July and August are calculated, defining six sample series. The *K*-mean method is used to classify the rainfall type (dry or wet type) for each 10-day period. For example, consider the first ten days of July. Suppose there are m historical years, the series of mean areal rainfall samples being denoted as $x_1,x_2,\ldots,x_m$. Define the two samples x_1,x_2, as the two groups (wet and dry) centres; the larger value is attributed to the wet type. Calculate the absolute distance between the group centre and all the samples values, i.e. $|x_i - x_1|$ and $|x_i - x_2|$ ($i = 3,4,...,m$). The smaller distance to x_1 or x_2 determines which type the sample belongs to. If a sample is at the same distance from x_1 and x_2, then it can belong to either group type. Subsequently, the samples can be divided into two groups. Set the mean sample value of each group as a new group centre and repeat the previous steps. This process is iterated until the difference between the last two group centres of the same type can be neglected, e.g. smaller than 0.001. This procedure is repeated for all six 10-day periods.

Antecedent influencing factors

The AIF are selected from meteorological indicators of ten upper-air stations, including geopotential height, temperature, dew-point deficit, wind direction and wind velocity at the 850 hPa, 750 hPa and 500 hPa pressure fields. For a three-day lead forecast, observations three days ahead of each day of the 10-day period of July or August are considered.

For example, to select the AIF for the first 10-day period of July, first, compute the *F*-statistic value for each meteorological factor on each pressure field from each upper-air station, e.g. temperature on 700 hPa pressure field of Nanning station as:

$$F = \frac{\sum_{g=1}^{G} N_g (\bar{x}_g - \bar{x})^2 / (N - G)}{\sum_{g=1}^{G}\sum_{k=1}^{N_g} (x_{kg} - \bar{x}_g)^2 / (N - 2)} \tag{2}$$

where, g represents two classified types and G equals to two; N is the total number of years, while N_1 is dry-type years and N_2 is wet-type years ($N_1 + N_2 = N$); $\bar{x}$ is the mean indicator value of all years, while $\bar{x}_g$ is the mean indicator value of dry or wet type years; x_{kg} is the sample indicator value, i.e. temperature on the 700 hPa pressure field of Nanning upper-air station on 28 June in the kth year of g type group. The larger the *F*-value, the larger the difference between $\bar{x}_g$ and $\bar{x}$, and the better the classification, thus the more effective the factor to determine the rainfall type. A factor is selected as AIF if its *F*-statistic is larger than the criterion value F_α, taken from the *F*-distribution.

Those selected factors could be different meteorological indicators of the same upper-air station, or the same meteorological indicators at a different upper-air station, and they were not classified as dry-type or wet-type factors.

Rainfall type prediction equation

With the selected AIF, the bi-discriminant method (Kim & Moy, 2002) is used to define rainfall type prediction equation for each 10-day period. The basic discriminant function is a linear composition of m factors, $x_1,\ldots,x_m$ with weight coefficient $c_k(k = 1,\ldots m)$ respectively, expressed as:

$$y = \sum_{k=1}^{m} c_k x_k \tag{3}$$

where c_k is determined based on Fisher rule, and y is the forecasted rainfall type.

All the historical years being sorted into two categories dry type (A) and wet type (B) for each 10-day period, the discriminant criterion y_c is computed as:

$$y_c = \frac{N_A \bar{y}^A + N_B \bar{y}^B}{N_A + N_B} \tag{4}$$

where $\bar{y}^A, \bar{y}^B$ are mean discriminant function values y of years of dry and wet type N_A, N_B respectively.

The rainfall type of a given 10-day period can be forecast by comparing the values of y and y_c. Two cases are included:

(a) when $\bar{y}^A > \bar{y}^B$, if $y > y_c$, then y belongs to category A, i.e. dry type; if $y < y_c$, then y belongs to category B, i.e. wet type.

b) when $\bar{y}^A < \bar{y}^B$, if $y < y_c$, then y belongs to category A, i.e. dry type; if $y > y_c$, then y belongs to category B, i.e. wet type.

Daily rainfall class prediction equation

Prediction equations for daily rainfall class are constructed for both dry and wet types. Similarly to the rainfall type prediction equation, select the AIF for predicting the rainfall class of each rainfall type for each 10-day period. Take the dry type of the first ten days in July as an example: from the rainfall totals, determine the daily rainfall class of 1 July to 10 July in dry-type years, select the AIF amongst the rainfall type 3-day lead meteorological indicators (28 June to 7 July) using the F-test method. In equation (2), G represents the number of rainfall classes, equal to four; N is the dry-type years, while N_g is the sample number of g (g = 1,…4) rainfall classes. The final AIF are selected from the F-statistic values. The multi-discriminant method is used to define the daily rainfall class prediction equation from the selected AIF. Similarly to the bi-discriminant method, a linear function is constructed, i.e. a prediction function for rainfall class, as in equation (3), in which the discriminant coefficient c_k is determined by the Fisher rule. For a given day, y is used to determine the rainfall class category as follows:

$$M^{(g)} = \frac{1}{N_g} \frac{\left| y - \bar{y}^{(g)} \right|}{\sigma^{(g)}} \qquad \bar{y}^{(g)} = \sum_{i=1}^{m} c_i \bar{x}_i^{(g)} \qquad \sigma^{(g)} = \sqrt{\frac{1}{N_g - 1} c s^{(g)} c'} \tag{5}$$

where c' is the transposed of matrix (c_1, c_2,…c_m) and s^g is the matrix of correlation coefficient computed in determining c_k. The category of the predicted value is that corresponding to the smallest value of M.

APPLICATION

Basin description

The Yuecheng reservoir is situated on the mainstream of the Zhanghe River, which crosses the boundary between Hebei and He'nan provinces in China (Figs 1 and 2). The controlling area is about 18 100 km^2, which takes up nearly 99.4% of the Zhanghe basin. It is located in the eastern Asian monsoon region of the temperate zone where the climate is influenced by the topography, thus hot and rainy in summer, with storms concentrating in July and August, and dry and cold in winter.

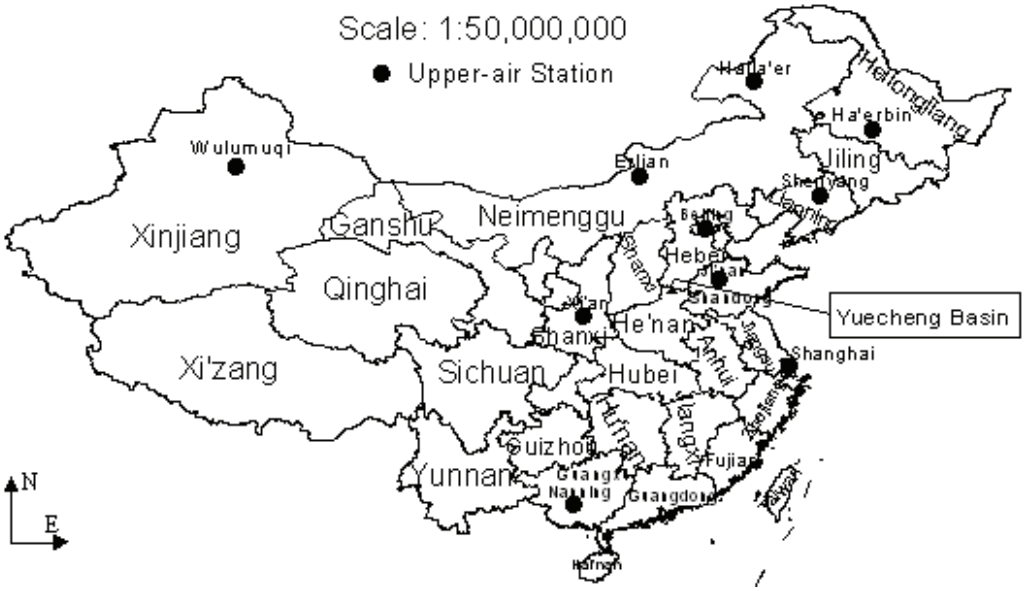

Fig. 1 Distribution of upper-air stations selected for obtaining meteorological data.

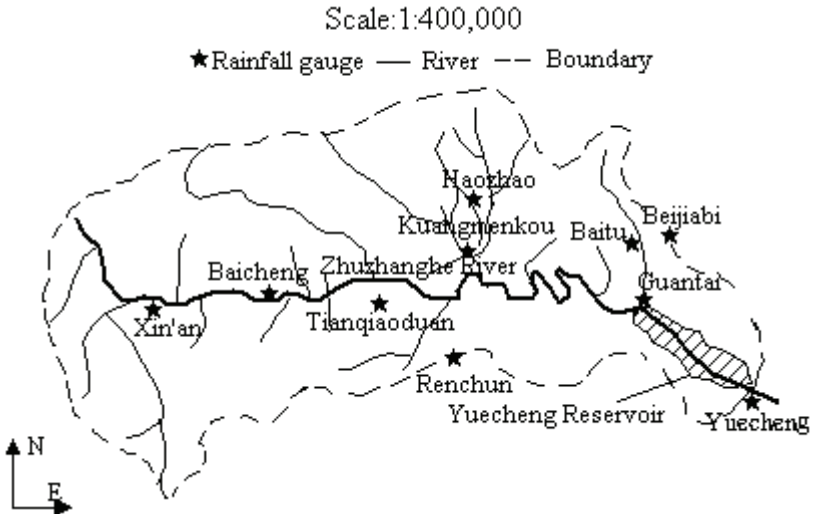

Fig. 2 Distribution of raingauges in the Yuecheng reservoir basin

Meteorological indicators

Meteorological indicators were obtained from ten upper-air stations around China, including Haila'er, Ha'erbin, Wulumuqi, Erlian, Shenyang, Beijing, Ji'nan, Xi'an, Shanghai and Nanning (Fig. 1). Daily upper-air meteorological records of geopotential height (HHHH), temperature (TTTT), dew-point deficit (UUU), wind direction (DDD) and wind velocity (FFF) on 850 hPa, 700 hPa and 500 hPa pressure fields were selected as potential antecedent factors for the prediction models. Meteorological data were collected for the rainfall record period, from 28 June to 29 August in 1980 to 1993. Records of the first twelve years were used for model calibration and that of the last two years (1992–1993) were used for model verification.

RESULTS ANALYSIS

Daily mean areal rainfall series are computed by averaging rainfall from all rainfall gauges using the Thiessen polygon method. Total rainfall of the six 10- (or 11-) day periods of July and August were calculated from 1980 to 1991 and daily rainfall class was identified into four classes. Rainfall data of each 10-day period were partitioned into dry or wet type by the *K*-mean method (Table 1). Meteorological factors discriminating rainfall type were selected by the *F*-test method. *F*-statistic

Table 1 Classification of rainfall type in July and August from 1980 to 1991 in Yuecheng.

Time	Rainfall type	In July	In August
First ten days	Dry type	1980–1982,1984–1986,1988–1991	1980,1982,1984–1986,1988–1991
	Wet type	1983,1987	1981,1983,1987
Second ten days	Dry type	1981–1986,1988–1989	1982–1986,1988–1991
	Wet type	1980,1987,1990–1991	1980–1981,1987
Last eleven days	Dry type	1980,1982–1986,1989–1991	1980–1983,1985–1990
	Wet type	1981,1987–1988	1984,1991

values were computed for all meteorological indicators and compared with the *F*-distribution table. Using the criteria α as 0.05 and 0.1 showed little difference in the number of selected AIF, and 12 indicators were finally selected for each 10-day period. For example, in the first 10-day period of July, the selected AIF are Xi'an_700_FFF (wind velocity on 700 hPa field of Xi'an upper-air station), Shenyang_850_FFF, Haila'er_500_HHHH, Ji'nan_500_FFF, Ji'nan_700_FFF, Beijing_500_UUU, Wulumuqi_850_HHHH, Ha'erbin_850_DDD, Haila'er_850_TTTT, Shanghai_850_TTTT, and Shanghai_500_HHHH. For each rainfall type in each period, the daily rainfall class prediction model is built similarly. Considering the practical application and accuracy requirement, four influencing factors for predicting rainfall classes were selected (Table 2) and the prediction functions were defined (Table 3). Equations for predicting rainfall types are not listed due to the space limitation. The accuracy of the prediction of daily rainfall classes for July and August 1992–1993 is about 86%, which satisfies the requirement for operational rainfall prediction.

Table 2 Factors selected to construct discriminant functions for rainfall class prediction.

Time	Rainfall	X_1	X_2	X_3	X_4
July First ten-day	Dry	Xi'an_700_FFF	Shenyang_850_FFF	Haila'er_500_HHHH	Er'lian_850_HHHH
	Wet	Beijing_500_UUU	Wulumuqi_850_HHHH	Ha'erbin_850_DDD	Haila'er_850_TTTT
July Second ten-day	Dry	Shenyang_500_DDD	Haila'er_500_UUU	Er'lian_850_HHHH	Wulumuqi_850_HHHH
	Wet	Nanning_850_FFF	Nanning_700_FFF	Nanning_500_FFF	Shenyang_850_UUU
July Last eleven-day	Dry	Erlian_700_DDD	Nanning_700_HHHH	Haila'er_700_HHHH	Haila'er_500_HHHH
	Wet	Shenyang_850_DDD	Shanghai_500_HHHH	Shanghai_700_TTTT	Shanghai_700_HHHH
August First ten-day	Dry	Ha'erbin_700_UUU	Haila'er_700_FFF	Shanghai_850_UUU	Nanning_850_HHHH
	Wet	Shanghai_850_FFF	Shenyang_500_DDD	Shanghai_500_FFF	Nanning_850_DDD
August Second ten-day	Dry	Erlian_500_FFF	Haila'er_500_TTTT	Wulumuqi _500_FFF	Wulumuqi_700_HHHH
	Wet	Wulumuqi_700_DDD	Wulumuqi_500_FFF	Wulumuqi_700_UUU	Wulumuqi_700_TTTT
August Last eleven-day	Dry	Xi'an_500_DDD	Nanning_850_UUU	Nanning_500_FFF	Haila'er_700_FFF
	Wet	Beijing_850_UUU	Jinan_700_FFF	Shanghai_850_UUU	Haila'er_500_UUU

Table 3 Discriminant functions for rainfall class prediction of Yuecheng basin.

Time	Rainfall type	Discriminant function (in July)	Discriminant function (in August)
First ten–day	Dry type	$Y = -0.65X_1 + X_2 - 0.12X_3 - 0.08X_4$	$Y = -0.06X_1 + X_2 - 0.21X_3 + 0.28X_4$
	Wet type	$Y = -0.60X_1. + X_2 - 0.31X_3 - 0.54X_4$	$Y = -X_1 + 0.04X_2 - 0.15X_3 + 0.02X_4$
Second ten–day	Dry type	$Y = 0.69X_1 - 0.54X_2 + 0.98X_3 + X_4$	$Y = X_1 + 0.04X_2 - 0.02X_3 + 0.06X_4$
	Wet type	$Y = -X_1 + 0.23X_2 - 0.04X_3 + 0.10X_4$	$Y = -0.20X_1 + X_2 - 0.12X_3 - 0.19X_4$
Last eleven-day	Dry type	$Y = 0.16X_1$-$0.07X_2 - 0.49X_3 + X_4$	$Y = 0.11X_1 - 0.03X_2 + X_3 - 0.76X_4$
	Wet type	$Y = 0.03X_1 + X_2 - 0.08X_3 - 0.05X_4$	$Y = 0.41X_1 + X_2$-$0.27X_3 + 0.11X_4$

Table 4 Number of days that forecasted rainfall class matches with observation

Year	Month	Rainfall type of ten-day period: First	Second	Third	Number of days that forecast matches observation	Relative error (%)
1992	July	Dry	Wet	Wet	27	87.1
	August	Wet	Dry	Dry	28	90.3
1993	July	Wet	Wet	Wet	21	67.7
	August	Wet	Dry	Dry	27	87.1

CONCLUSIONS

In this study, a statistical model was designed for predicting medium-term rainfall classes. Rainfall types for 10-day periods in July and August from 1983 to 1991 were classified by the *K*-mean method, and four rainfall classes defined. The *F*-test method was adopted to select the meteorological indicators significant for predicting rainfall type and rainfall classes independently. With these selected factors, a rainfall-type forecasting model was built using the bi-discriminant method, while the model for rainfall class prediction was derived from a multi-discriminant analysis, thus predictions of both the rainfall type for each 10-day period and rainfall class for each day of the period were obtained.

The procedure was applied to rainfall prediction with a 3-day lead time for Yuecheng reservoir basin. Results show comparatively high forecasting skills with good forecast for about 86% of days. This approach does not require much data, and the required meteorological data could be easily obtained in operational forecast. It is thus a valuable alternative for medium-term rainfall forecast. With better selection of key meteorological factors, the forecasting accuracy of this approach could be improved.

REFERENCES

Hand, D. J. (1981) *Discrimination and Classification*. Wiley, New York, USA.

Joachimsthaler, E. A. & Stam, A. (1988) Four approaches to the classification problem in discriminant analysis: an experimental study. *Decision Sciences* **19**, 322–333.

Kim, F. L. & Moy, J. W. (2002) Combining discriminant methods in solving classification problems in two-group discriminant analysis. *European J. Operational Research* **138**, 294–301.

Thiessen, A. H. (1911) Precipitation for large areas. *Monthly Weather Review* **39**, 1082–1084.

Zwiers F. W. & Von Storch, J. (2004) On the role of statistics in climate research. *Int. J. Climatology* **24**, 665–680