Article ID: 1006-8775(2019) 04-0519-09

A PREDICTION SCHEME FOR THE PRECIPITATION OF SPR BASED ON THE DATA MINING ALGORITHM AND CIRCULATION ANALYSIS

LI Chao (李　超) [1], SHI Da-wei (史达伟) [2], CHEN Yu-tian (陈誉天) [3], ZHANG Hong-hua (张红华) [2],
GENG Huan-tong (耿焕同) [4], WANG Peng (王　鹏) [2]

(1. Jiangsu Meteorological Observatory, Nanjing 210008 China; 2. Meteorological Bureau of Lianyungang City, Lianyungang 222000 China; 3. College of Geography, Nanjing University of Information Science and Technology, Nanjing 210044 China; 4. College of Continuing Education, Nanjing University of Information Science and Technology, Nanjing 210044 China)

Abstract: Based on the 74 circulation indexes provided by National Climate Center of China (NCC) and the 24 indexes compiled by National Oceanic and Atmospheric Administration (NOAA) of the US, the study used the C4.5 algorithm in data mining to establish a decision tree prediction model to predict whether the spring persistent rains (SPR) of 55 years (from 1961 to 2015) is more than the normal, and obtained 5 rules to determine whether the SPR is more than the normal. The accuracy rate of the test set, namely “whether the SPR is more than the normal”, is 98.18%. After the evaluation of the model by conducting ten 10-fold cross validations to take the average value, the test accuracy rate gained is 84%. There are differences between the three types of years with a SPR more than the normal when it comes to intensity and distribution. In spring, they have respective anomalous 850hPa monthly mean wind fields and water-vapor flux distribution, and 700hPa forms the zone where the vertical speed is anomalously negative. As indicated by the results, the SPR prediction model based on the C4.5 algorithm has a high prediction accuracy rate, the model is reasonably and effectively constructed, and the decision rules take comprehensive factors into consideration. The anomalous rainfall and circulation distribution characteristics obtained based on the decision classification results provide new ideas and methods for the prediction of SPR.

Key words: spring persistent rains; data mining; C4.5 algorithm; prediction model; model analysis

CLC number: P457.6 **Document code:** A

doi: 10.16555/j.1006-8775.2019.04.008

1 INTRODUCTION

As China is located in the East Asian monsoon area, the climate in East China is greatly impacted by the monsoon rainfall caused by the monsoon burst. Therefore, the East Asian monsoon rainfall has always been one of the important subjects greatly concerned by meteorologists in China and relevant countries. As early as the 1930s, ZHU Ke-zhen, a famous Chinese meteorologist, first proposed the impact of the East Asian summer monsoon on the rainfall in China (Zhu [1]). Gao et al. pointed out that East Asian summer monsoon is closely related to rainfall in various regions of China, and the rainy season in these regions usually begins when the summer monsoon arrives every year [2]. Chen et al. studied the division of rainfall in East China and characteristics of drought and flood variations in different regions in details. With 110° E as the boundary, to the east of 110° E, the main rain band in April and May generally hovers between areas to the south of the Yangtze River and Qinling Mountains, while southern China is in the annually first rainy season[3]. The main rain band moves to the middle and lower reaches of the Yangtze River in June to cause the plum rain season. As the main rain band moves north to the Yellow River Basin in July, and the Huang-Huai area and the northeast China Plain enter the rainy season, summer drought appear in the middle and lower reaches of the Yangtze River and areas to the south of the Yangtze River. In August, when the main rain band withdraws to the northernmost tip of south China, south China is in the annually second flood season. As the main rain band withdraws south quickly, except for the lower reaches of the Yangtze River and south China, the rainy season in east China ends.

Yeh et al. argue that there are only two natural seasons throughout the year, that is, winter and summer; winter is relatively long and the transition season is so short that it is almost negligible [4]. Therefore, among

Received 2018-08-05; **Revised** 2019-08-06; **Accepted** 2019-11-15

Foundation item: National Key Research and Development Plan (2017YFC1502104); Nature Science Foundation of Jiangsu Province (BK20161603, BK20171095); National Natural Science Foundation of China(41575081); Forecaster Special Project of China Meteorological Administration (CMAYBY2017-026); Huaihe Basin Meteorological Open Research Foundation (HRM201602); Social Development Project of Jiangsu Provincial Science and Technology Department (BE2011720); Lianyungang City Science and Technology Plan (SH1634)

Biography: LI Chao, Master, primary undertaking research on weather forecast technology and air sea interaction.

Corresponding author: ZHANG Hong-hua, e-mail: zhhziteng@163.com

climate researches, there are few analyses on the characteristics of spring rainfall and circulation in areas to the south of middle and lower reaches of the Yangtze River. Only the continuous rainy weather in areas to the south of middle and lower reaches of the Yangtze River will be paid attention to as a research object for synoptic meteorology or short-term weather forecast (Wu et al. [5]; Wang et al. [6]). Different from the main rainy season in most parts of China, the proportions of spring rainfall and summer rainfall in the total annual rainfall in areas to the south of middle and lower reaches of the Yangtze River are basically the same, thus it is necessary to study the spring rainfall in areas to the south of middle and lower reaches of the Yangtze from the perspective of average climate(Yang et al. [7]). However, it was not until the end of the 1990s that Tian proposed the concept of SPR for the first time and studied it as a climate event[8]. Wan et al. systematically analyzed the spatiotemporal distribution of SPR, pointing out that it was appropriate to take the 13th and 27th pentad as the establishment and termination time of SPR respectively, meanwhile southeast China, including areas to the south of the middle and lower reaches of the Yangtze River (30N°) and areas to the east of 110E°, was defined as the spatial range of SPR [9]. Wan et al. pointed out that the dynamic and thermodynamic effects of the Tibetan Plateau play an important role in the establishment and maintenance of SPR [10]. According to Zhang et al. [11]. and Shang et al. [12], abnormal ocean temperature in the western Pacific will influence the amount of SPR. To sum up, most of the previous studies have studied SPR from the perspectives of external forcing factors such as dynamic and thermodynamic effects of the Tibetan Plateau, ocean temperature, etc.

With the continuous improvement of computer performance, it has become increasingly common to apply the data mining technology in researches on meteorological problems. Zhang et al. used the C4.5 algorithm to build a model to accurately predict whether the typhoon in the northwest Pacific Ocean was landing and whether it was turned [13–14]. Geng et al. applied the finite mixing model (FMM) algorithm and the classification and regression tree (CART) algorithm to predict the path and frequency of the tropical cyclone landed in China, achieving a good predictive effect [15]. Based on radar data, satellite data and model output parameter, David et al. used the random forest (RF) algorithm to establish a mesoscale convective system (MCS) prediction model [16]. In this study, the C4.5 decision tree algorithm was used to establish the SPR rainfall prediction model based on a total of 100 circulation indexes compiled by NCC and NOAA. Moreover, the rainfall and circulation characteristics of each type were analyzed based on the decision classification results, providing new ideas and methods for the climatic prediction of SPR.

2 DATA AND METHODS

2.1 *Data*

(1) The study used the reanalysis data set of the NCEP-NCAR, with the basic elements including zonal wind *u*, meridional wind *v*, vertical speed *ω* and specific humidity shum, the spatial resolution being 2.5°×2.5°, and the download link is http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis.html.

(2) The circulation indexes used in this study were extracted from the following two sources (Geng et al. [15]): (1).26 climatic signals such as the ENSO index and the PDO index compiled by NOAA; (2).74 circulation indexes organized and compiled by NCC. Both are monthly average data.

Table 1. 100 climate indexes used in this study (partial indexes Indexes).

A Part of Climate Index from NCC	A Part of Climate Index from NOAA
The subtropical high ridge in the Northern Hemisphere of the previous winter	Nino3
Subtropical high northern boundary in the South China Sea of the previous winter	TSA (Tropical southern Atlantic index)
Position of the center of the polar vortex in the Northern Hemisphere of the previous winter	AMO (Atlantic multidecadal Oscillation)
The subtropical high ridge in North Africa of the previous winter (20W-60E)	SOI (Southern Oscillation Index)
…	…

(3) The daily rainfall data of 2200 stations compiled by NCC.

(4) Data about the global monthly average rainfall (Zi et al. [18]) provided by the Global Rainfall Climatology Centre (hereinafter referred to as GPCC), with the spatial resolution being 1.0°×1.0°.

2.2 *Definition*

(1) The average climatic data in this study refer to

the average data from 1961 to 2015. Since the circulation factors of the previous winter are involved, all the data are in the period from 1960 to 2015.

(2) According to Wan et al.[9], the study defined the SPR period to be from the 13th to the 27th pentad (March 1-May 15), with the SPR zone being 23°–30°N, 110°–120°E. 316 stations are chosen in this study, and the average daily rainfall of each station from March 1 to May 15 from 1961 to 2015 is taken as the SPR sample.

(3) The previous winter was defined as December of the previous year, and January and February of the same year.March and April of the same year were defined as spring.

(4) As spring takes up 80% of the SPR period, for statistical convenience, based on the method applied by Wan et al., the study used the monthly average rainfall data in spring issued by GPCC and the monthly average reanalysis data issued by NCEP/NCAR to analyze the rainfall and circulation features of the types with SPR more than the normal in the same period, so that the analysis results obtained by the study will not be greatly affected[9].

2.3 *The C4.5 decision tree algorithm*

As a classification and prediction algorithm, the C4.5 decision tree algorithm was invented in 1979 by JR Quinlan who also proposed the ID3 algorithm which mainly aimed at discrete attribute data, and then after constant improvement, the C4.5 algorithm was developed. It discretized the continuous attributes on the basis of ID3 (Quinlan [19]).

This algorithm selected the classification attribute on each node according to the size of the attribute information gain and split the sample per the attribute that can bring the maximum information gain, and recursively split until the stop conditions were met (Han and Kamber [20]). Finally, the results were tested to trim and remove the sample sets without significant contributions to the model.

If S is the training set containing s data samples, $S(C_i)$ is the number of samples that belong to Type C_i in S (i=1, 2, L, m), then the probability that the samples in the training set belong to the ith type is

$$P(C_i)=\frac{S(C_i)}{s} \tag{1}$$

At this time, the information (entropy) of the training set S is defined as:

$$info(S)=-\sum_{i=1}^{m}P(C_i)\log_2 P(C_i)=-\sum_{i=1}^{m}\frac{S(C_i)}{s}\log_2[\frac{S(C_i)}{s}] \tag{2}$$

Next, S should be divided into $\{S_1, S_2, L\ S_v\}$ by attribute A, and the information entropy of the leaf node on classified information is:

$$info(A|S)=-\sum_{j=1}^{v}\frac{S_j}{S}info(S) \tag{3}$$

Then, information gain can be calculated as follows:

$$gain(A|S)=info(S)-info(A|S) \tag{4}$$

The information gain is:

$$gain\ ratio(A|S)=\frac{gain(A|S)}{info(A|S)} \tag{5}$$

Developed based on the ID3 algorithm, one of the major improvements of the C4.5 algorithm is that it can deal with continuous data, the general process for the C4.5 algorithm to process continuous attribute data is as follows: (1) sort the attribute data in the nodes; (2) dynamically divide the training set data with different threshold values; (3) take the midpoint of the two values among the input data as a threshold value and determine a new threshold value when the input changes; (4) determine two categories according to the threshold and divide all the data samples into these two categories; (5) obtain all possible thresholds and calculate the information gains and gain rates for all categories; (6) Finally, each successive attribute is divided into two categories by the threshold (greater than or equal to the threshold and less than the threshold). For further details about this clustering method, see the study by Zhang et al.[13].

2.4 *K-fold cross validation*

K-Fold Cross Validation is a commonly used model assessment method in the field of data mining and machine learning. The data sets are randomly divided into *K* subsets that are disjointed and of the same size, the *K* combinations are repeated one by one using the *K*-1 subset training models and the remaining one subset test model, the mean of *K* test accuracy rates is obtained and it is used as an estimate of the generalization test accuracy rate. In most cases, scientists use 10-fold cross validation to evaluate models.

3 CONSTRUCTION OF THE MODEL TO PREDICT THE MORE THAN THE NORMAL SPR BASED ON THE C4.5 ALGORITHM

The C4.5 algorithm was used to establish the decision tree model with the SPR samples from 1961 to 2015 as the training sets. A total of 26 circulation indexes of each spring and the previous winter issued by NOAA and the 74 circulation indexes issued by NCC were selected as the learning attributes of the model. “Whether the SPR is more than the normal” was set as the target variable. When the normalized anomaly was >0.5, the SPR in that year would be more than the normal. The decision tree model was obtained after the selection and calculation of the C4.5 algorithm, as shown in Fig. 1. The root node is Pre-winter Nino3, which is the most important attribute for SPR classification. In the leaf nodes, T (F) denotes the SPR that are more than the normal (not more than the normal SPR). A, B and C in the bracket represent the three types of SPRs that are more than the normal and D and E stand for the two cases of normal SPRs. As indicated by the observation, the climatic signals involved in the modeling include Pre-winter Nino3, the subtropical high ridge in the Northern Hemisphere (SHRNH) of the same period, the tropical southern Atlantic index (TSA) of the same period, the subtropical high northern boundary in the South China Sea of the previous winter (Pre-winter SHNB). The 3 categories of SPRs more than the normal were named Type A, B and C, and the two types of SPRs that were not more than normal were named Type D and E.

The model was used to classify and predict the data samples collected during the 55 years. The learning accuracy rate of the model is 98.18% and one sample was not correctly classified (1991). In order to generalize the test accuracy of the estimation model, the model was evaluated by taking the mean of the 10-fold cross validation carried out for 10 times, with the accuracy rate being 84%.

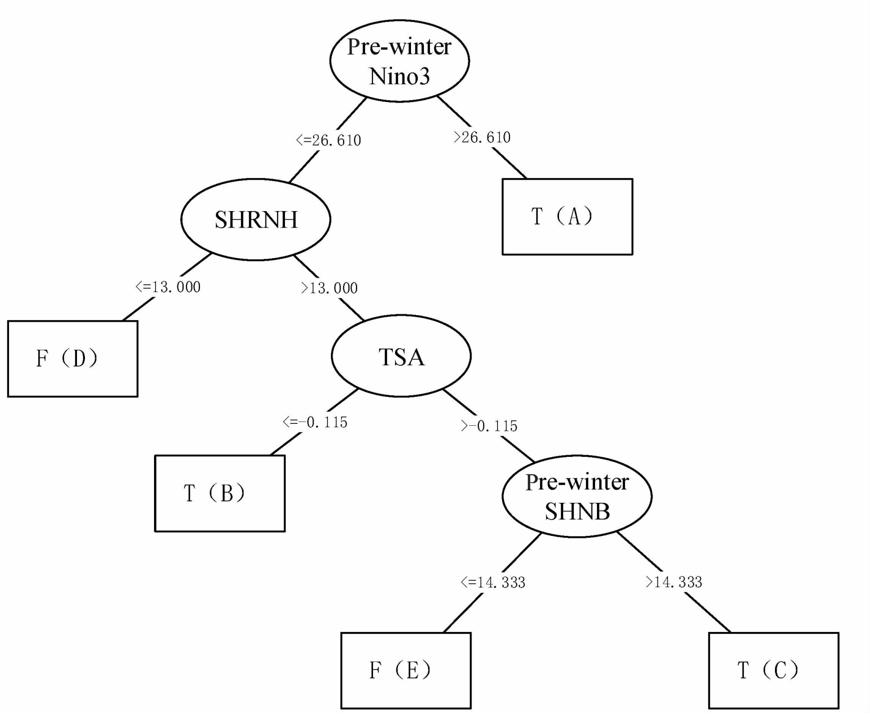

Figure 1. The decision tree model to predict whether SPR is more than the normal based on the C4.5 algorithm.

Table 2 shows the 5 rule sets generated according to Fig. 1 to predict whether the SPR was more than the normal. The first column describes each rule in the form of "If-then", the second column describes the attribute variables involved to establish those rules, and the third column describes the learning accuracy of each classification rule. This accuracy rate is the ratio of the number of correctly classified samples in the corresponding leaf node to the total number of samples in this leaf node.

Table 2. The rule sets discovered by C4.5 to determine whether the SPR is more than the normal.

Rule	Decision Attribute	Learning Accuracy
Type A: If (Pre-winter Nino3 > 26.610), then the SPR is more than the normal	Pre-winter Nino3	6/6=100%
Type B: If (Pre-winter Nino3 <= 26.610 and SHRNH > 13 and TSA > –0.115 and Pre-winter SHNB > 14.333), then the SPR is more than the normal	Pre-winter Nino3, SHRNH, TSA, Pre-winter SHNB	3/4=75%
Type C: If (Pre-winter Nino3 <= 26.610 and SHRNH > 13 and TSA <= –0.115), then the SPR is more than the normal	Pre-winter Nino3, SHRNH, TSA	6/6=100%
Type D: If (Pre-winter Nino3 <= 26.610 and SHRNH <= 13), then the SPR is not more than normal	Pre-winter Nino3, SHRNH	28/28=100%
Type E: If (Pre-winter Nino3 <= 26.610 and SHRNH > 13 and TSA > –0.115 and Pre-winter SHNB <= 14.333), then the SPR is not more than normal	Pre-winter Nino3, SHRNH, TSA, Pre-winter SHNB	7/7=100%

Table 3 lists the years that Type A, B, C, D and E with SPR more than and not more than the normal belong to, and it can be seen that there are 16 samples with SPR more than the normal in the SPR samples collected during the 55 years, and there is only 1 misclassified sample among all the samples. It is thus clear that the C4.5 algorithm has a good effect on predicting whether the SPR is more than the normal, and it is quite scientific and easy to use because it can work out the concise rule sets for prediction. The C4.5 algorithm is a classic method of data mining. Therefore, using the C4.5 algorithm to study the SPR also provides a new idea for the analysis of nonlinear SPR climatic characteristics and climatic prediction.

Through statistics, we get the proportion of all types of precipitation in Table 3. It can be found that the proportion of type A and type C is 13.91% and 13.88% respectively in the case of more than normal SPR, and the proportion of precipitation is relatively high. In the case of not more than normal SPR, the precipitation of type D accounted for 46.31% of the total precipitation, accounting for the highest proportion.

Table 3. Distribution of years (A, B, C, D and E) with different types of SPR and percentage of precipitation related to the Type A, B, C, D, and E.

Types of SPR more than the normal	Year	The percentage of precipitation per type
Type A (with more than normal SPR)	1973, 1983, 1987, 1992, 1998, and 2010	13.91%
Type B (with more than normal SPR)	1961, 1984, **1991**, and 1999	8.31%
Type C (with more than normal SPR)	1970, 1975, 1979, 1980, 1981, and 2012	13.88%
Type D (with no more than normal SPR)	1965, 1966, 1967, 1968, 1969, 1971, 1972, 1974, 1976, 1978, 1982, 1988, 1989, 1990, 1993, 1994, 1995, 1997, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2009, 2011, and 2013	46.31%
Type E (with no more than normal SPR)	1962, 1963, 1964, 1977, 1985, 1986, 1996, 2000, 2008, 2014, and 2015	17.59%

4 ANALYSIS OF THE SPR MODEL

Based on the corresponding years of the three types of SPRs which are more than the normal (Type A, B and C) and the two types of normal SPRs (Type D and E), the anomalies of their spring rainfall and circulation (i.e. the anomaly obtained by this type minus the average climate) were analyzed respectively, providing a scientific basis for climatic prediction of SPR and circulation. Fig. 2 shows the GPCC daily spring rainfall of average climate and the three anomalies, namely A, B and C. As indicated in Fig. 2, in spring with average climate, there is a NE-SW rain band with a daily rainfall of 4mm/d in the southeastern part of China to the east of 110° E and the south of the middle and lower reaches of the Yangtze River (30° N), with the rainfall intensity in the center of the rain band reaching 6–7mm/d, which is basically consistent with the research results about the spatial range and intensity of SPR obtained by Wan et al. [10]. The anomalous rainfall intensity of Type A is 0.5–3mm/d, and its center is located in the southern region of the SPR area. The anomalous rainfall band of Type B with an intensity of 0.5–1.5mm/d covers the SPR area relatively uniformly. The anomalous rainfall intensity of Type C is 0.5–3mm/d, and its distribution pattern is relatively consistent with the average climate. To sum up, in spring, Type A, B and C are all featured by an SPR more than the normal, but abnormal rainfall intensity and distribution of the three types are different, which is worthy of attention in the climatic prediction of SPR.

According to Wan et al. and Li et al., the southwesterly airflow and water vapor transport in the lower reaches of the Yangtze River in spring are critical to the formation and maintenance of spring rain in the south of the Yangtze River [10, 21]. Fig. 3 shows the 850hPa average monthly wind field and water-vapor flux distribution of the average climate and Type A, B, C, D and E in spring. It can be seen from Fig. 3a, because of the plateau terrain, the westerly airflow in East Asia during spring is apparently divided into the north and south parts which join each other at (30°N, 120°E). In the southeastern part of the plateau, the southwest airflow produced by the "rounded flowing" joins the southwest airflow in the northwest Pacific anti-cyclone to form a southwest jet in the southeastern part of the plateau. The water vapor in the SPR area in spring comes from western Pacific and the South China Sea. The southwest water vapor transmitted in the SPR area is relatively strong, with the central intensity of the water-vapor flux reaching 700 $kg \cdot m^{-1} \cdot s^{-1} \cdot hPa^{-1}$, and its position is consistent with the area with a large value of the southwest jet in the southeastern part of the plateau. As shown by the 850hPa anomalous average monthly wind field and water-vapor flux distribution in spring (Fig. 3b), the SPR area is affected by the consistent southwest anomaly airflow. The main reason is that there is El Niño effect in the pre-winter of Category A, and there is abnormal anticyclone circulation development in the Northwest Pacific corresponding to 850 hPa in spring (Huang et al. [21]; Zhao [22]). Influenced by the anomalous southwest airflow, there is an SW-NE water vapor transfer band from the Indo-China Peninsula to the southern part of Japan, transmitting water vapor to the SPR area. As for Type B (Fig. 3c), with the circulation anomaly in spring, the western part of the SPR area is affected by the anomalous south airflow, while the eastern part is affected by the anomalous southeast airflow. Affected by such anomalous wind field, the anomalous water vapor in the SPR area comes from the eastern sea area of the Indo-China Peninsula and the middle latitude of the western Pacific. As for Type C (Fig. 3d), the rounded flowing in the southwestern part of the Qinghai-Tibet Plateau is extremely strong, and there is an anomalous southward airflow leading the anomalous water vapor to be transmitted to the SPR area in the South China Sea. In general, Type A, B, and C have their own abnormal 850hPa average monthly wind field and water-vapor flux distributions in spring, which provides reference for climatic prediction of SPR in terms of wind field and water vapor. To sum up, Type A, B, and C have their own abnormal 850hPa strongly southerly wind fields and water-vapor flux distributions in spring, which is consistent with the research findings obtained by Wan et al. [10] and Li et al. [23], whereas Type D and E lack the southerly wind field and water-vapor flux distribution in spring in the south of the Yangtze River, which provides reference for climatic prediction of SPR in terms of wind field and water vapor.

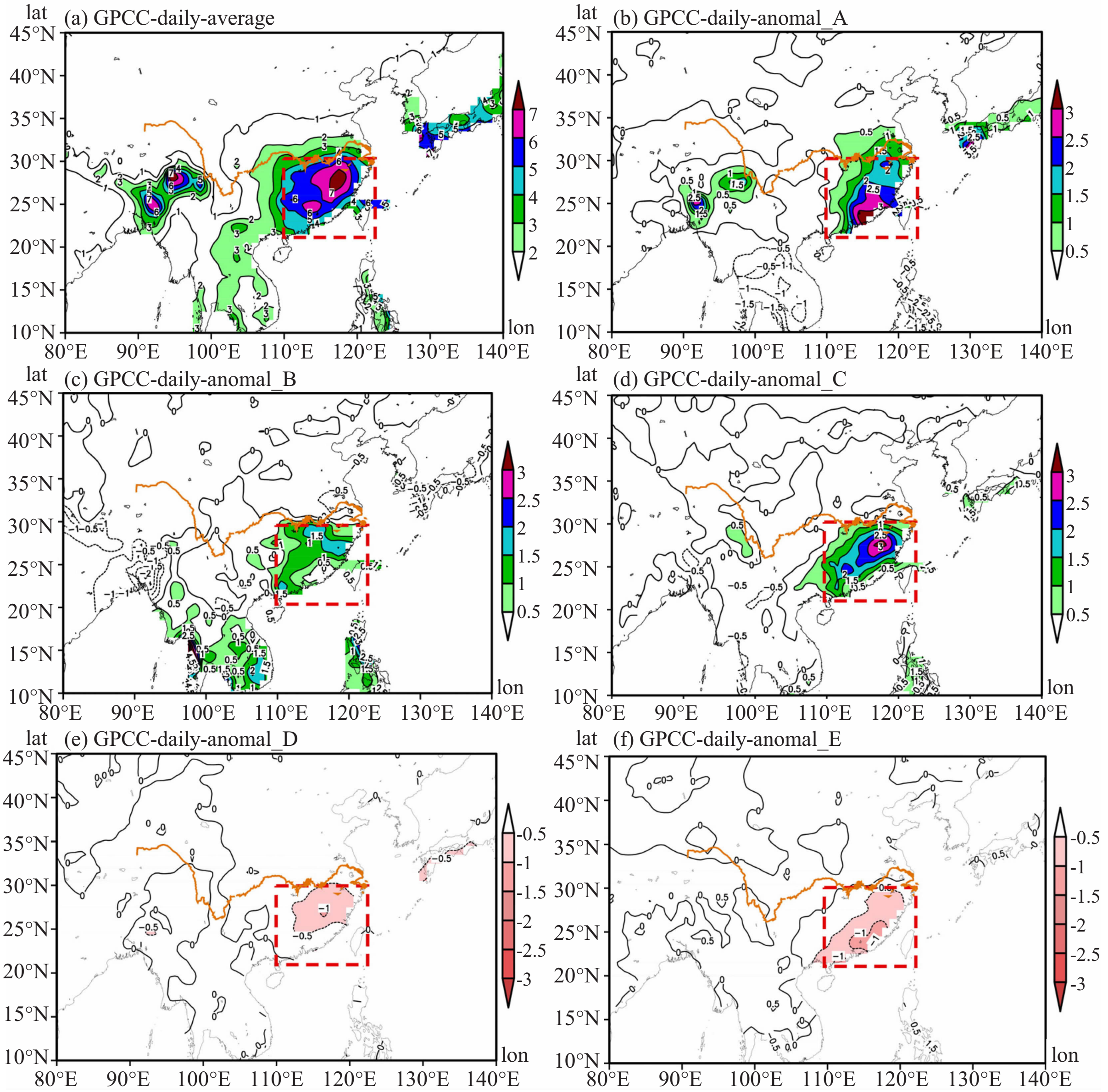

Figure 2. GPCC daily spring rainfall (a. average climate; b. anomalies of type A; c. anomalies of type B; d. anomalies of type C; e. anomalies of type D; f. anomalies of type E). The Chinese mainland in the red rectangle stands for areas subject to spring persistent rains. The orange curve represents the Yangtze River.

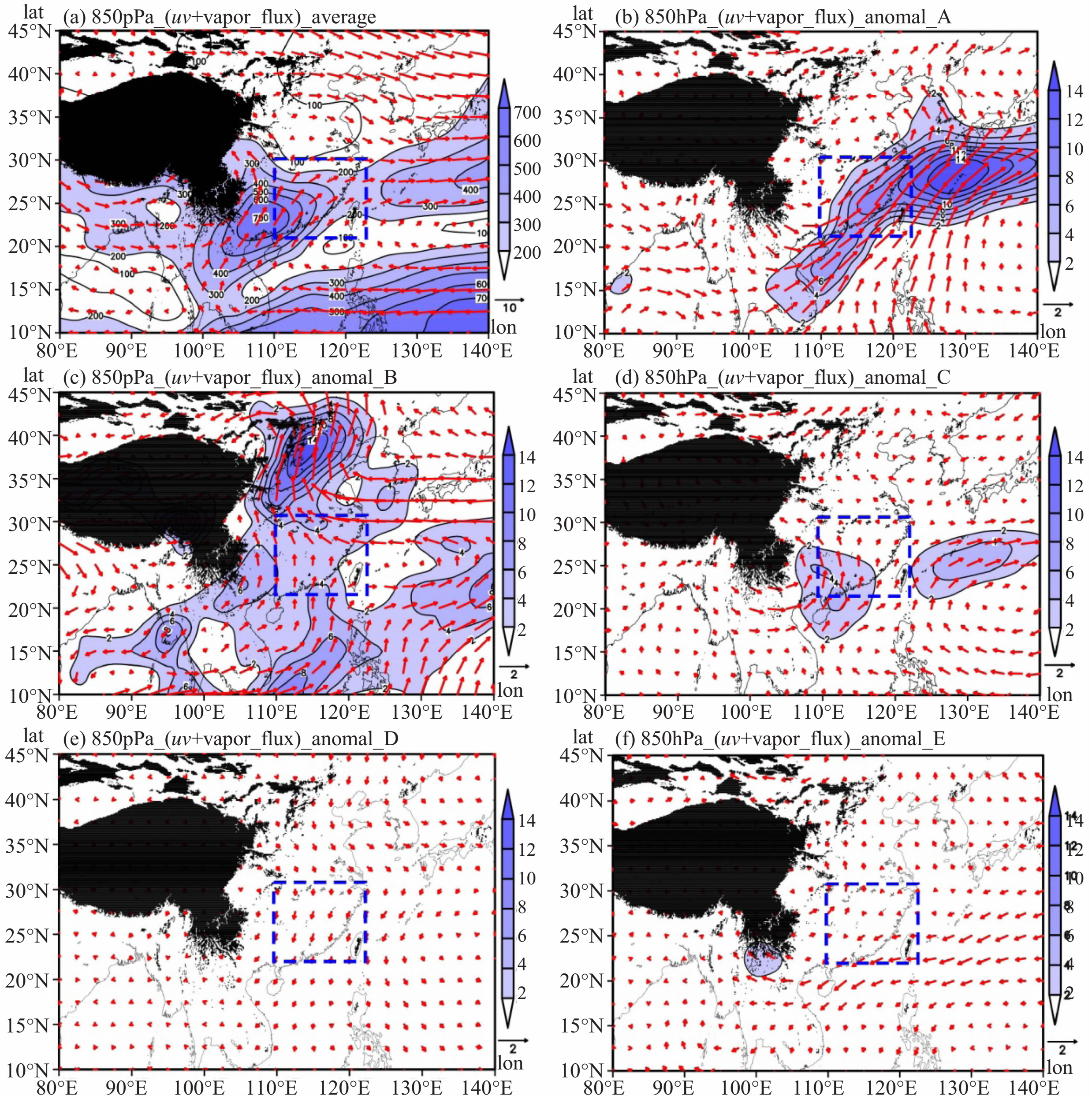
(a) 850pPa_(uv+vapor_flux)_average
(b) 850hPa_(uv+vapor_flux)_anomal_A
(c) 850pPa_(uv+vapor_flux)_anomal_B
(d) 850hPa_(uv+vapor_flux)_anomal_C
(e) 850pPa_(uv+vapor_flux)_anomal_D
(f) 850hPa_(uv+vapor_flux)_anomal_E
lat
lon
45°N
40°N
35°N
30°N
25°N
20°N
15°N
10°N
80°E
90°E
100°E
110°E
120°E
130°E
140°E
700
600
500
400
300
200
10
14
12
10
8
6
4
2
2

Figure 3. The 850 hPa wind field (unit: m s^{-1}) and water-vapor flux (unit: 10^{-4}kg·m^{-1}·s^{-1}·hPa^{-1}) in spring(a. average climate; b. anomalies of type A; c. anomalies of type B; d. anomalies of type C; e. anomalies of type D; f. anomalies of type E. The black shadow area shows the Tibetan plateau topography, the purple shadow area denotes the water-vapor flux, and the Chinese mainland in the blue rectangle stands for areas subject to spring persistent rains).

As shown in Fig. 4 (a), in spring with average climate, 700hPa in the SPR area is covered by a negative vertical speed area, with the intensity of the negative center reaching −0.8Pa s^{-1}, corresponding to a strong ascending motion. The vertical speed distribution of 700hPa has a good correspondence with the 850hPa water-vapor flux and wind field distribution (Fig. 3a). As the negative center of the vertical speed, the SPR area corresponds to the great value area of water-vapor flux and the southwest jet zone. As indicated by the anomalous distribution of 700hPa vertical speed of Type A, B and C in spring, when the SPR is more than the normal, 700hPa is the anomalously negative vertical speed in the SPR area, implying an abnormal ascending motion. On the contrary, when it comes to Type D and E, 700hPa is the anomalously positive vertical speed in the SPR area, indicating an abnormal descending motion. It is worth noting that as for Type A, B and C in spring, the 700hPa vertical speed distribution and 850hPa water-vapor flux and wind field (Fig. 3a) enjoy a good configuration: Type A, B, and C have their own abnormal 850hPa strong southerly wind fields and water-vapor flux distributions in spring, and there is an abnormal ascending motion at 700hPa, providing favorable conditions for the establishment and maintenance of the SPR rain bands.

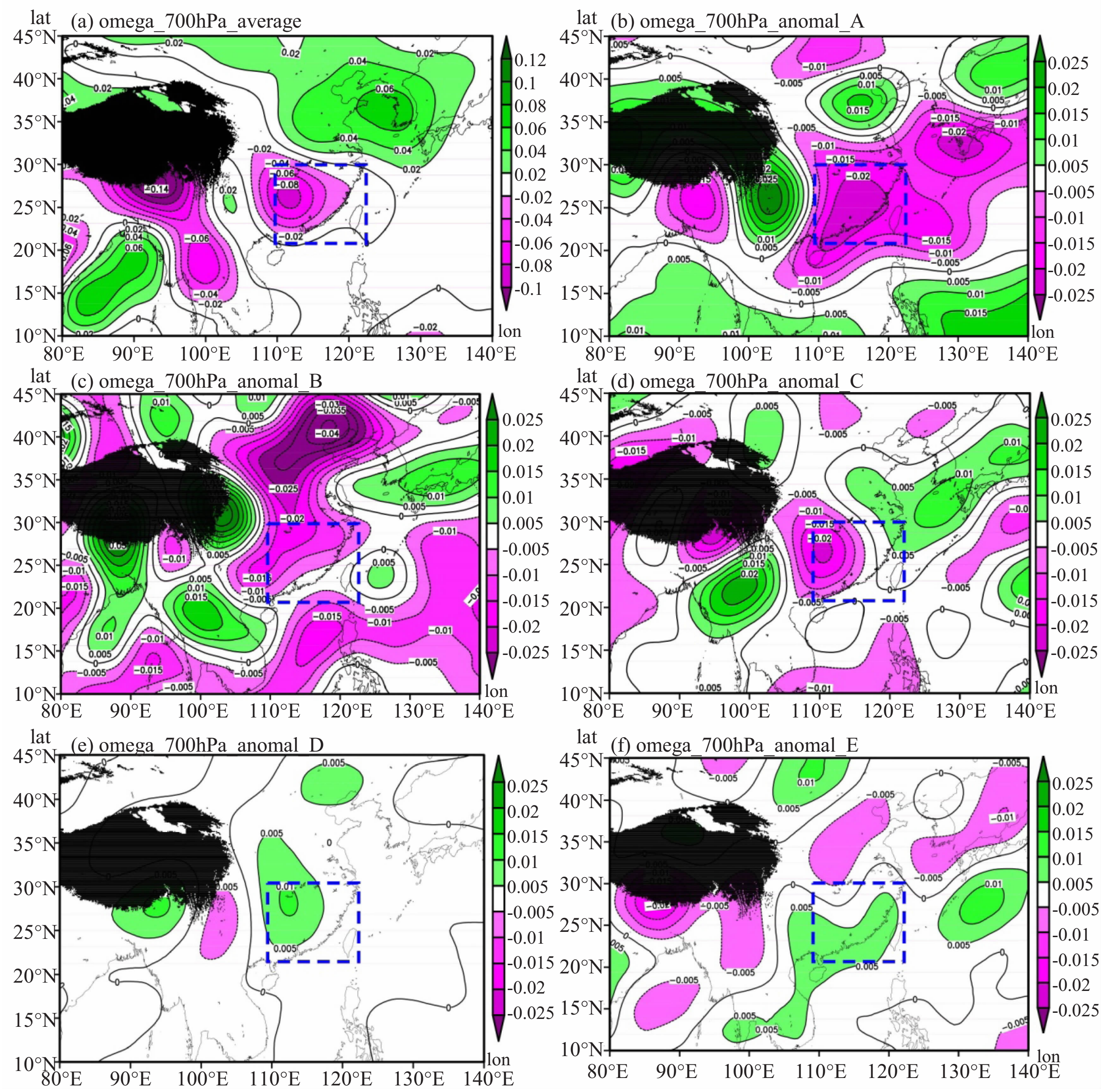

Figure 4. The 700 hPa vertical speed in spring (unit: Pa s^{-1}) (a. average climate; b. anomalies of type A; c. anomalies of type B; d. anomalies of type C; e. anomalies of type D; f. anomalies of type E). The black shadow area shows the Tibetan Plateau topography, the blue shadow area denotes the vertical speed, and the Chinese mainland in the blue rectangle stands for areas subject to spring persistent rains.

5 SUMMARY AND DISCUSSION

SPR is the main rain band of the SPR area in East Asia in spring, but it was until the end of the last century that it was proposed by Tian et al as a climatic event [8]. Most of the previous researches studied SPR from the perspectives of external factors such as the dynamic and thermodynamic effects of the Tibetan Plateau, sea surface temperature, etc. As a classic data mining method, the C4. 5 algorithm was used to study SPR, which can provide a new research idea for analyzing the nonlinear climatic characteristics and climatic prediction of SPR. Therefore, with 26 circulation indexes of each spring and the previous winter issued by NOAA and the 74 circulation indexes issued by NCC, the C4.5 algorithm was used to establish the decision tree model to predict the SPR from 1961 to 2015, achieving a good prediction effect. Based on the corresponding years of the three types of SPR more than the normal (Type A, B and C), the abnormal characteristics of rainfall and circulation of the three types were analyzed. The main conclusions obtained are as follows: (1) With the 100 circulation indexes as the input variable, the C4.5 algorithm was used to establish the SPR prediction model and relevant prediction rule sets. The accuracy rate of the test set, namely “whether the SPR is more than the normal”, is 98.18% . After evaluating the model by conducting ten 10-fold cross validations to take the average value, the test accuracy rate gained is 84% . (2) Type A, B and C are all characterized by a rainfall with the amount being more than the normal in the SPR area during spring, but there are differences between the three types when it comes to intensity and distribution, which is worthy of attention in the prediction of SPR. (3) Type A, B and C have their own abnormally strong southerly wind fields and water-vapor flux distributions and ascending motions in spring, which indicates that the differences in intensity and distribution of SPR being more than the normal correspond to the differences in water-vapor flux distribution and circulation pattern, providing reference for predicting whether the SPR is more than the normal and the types different SPRs belong to.

With the advent of the big data era, data mining technologies have been applied in many fields. The accumulation of rainfall data and multiple circulation indexes lays a solid foundation for data mining technologies to be applied in rainfall prediction, providing a new idea for statistical forecasting. The study lays emphasis on the climatic prediction of the SPR that is “more than the normal”. How to use data mining technologies to gain a more refined prediction of SPR from the perspective of time and space remains to be a question that needs to be further explored and discussed.

REFERENCES:

[1] ZHU K Z. The enigma of southeast monsoon in China [J]. Acta Geographica Sinica, 1934, 1(1): 1-27 (in Chinese).

[2] GAO Y X, XU S Y, GUO Q Y. Some problems of east Asian monsoon [M]. Beijing: Science Press, 1962: 49-63 (in Chinese.)

[3] CHEN L T, WU R G. Climatic division of precipitation in eastern China and drought-flood variation in various regions [J]. Chin J Atmos Sci, 1994, 18 (5): 586-595 (in Chinese).

[4] YEH T C, DAO S Y, LI M T. The abrupt change of circulation over the northern hemisphere during June and October. In: Bolin B (ed) The atmosphere the and sea in motion [M]. New York: The Rockefeller Institute Press, 1959: 249-267.

[5] WU B J, PENG Z B. Progress in the study on spring persistent rains in Lingbei, south of the Yangtze river [J]. Sci Bull, 1996, 12: 65-70.

[6] WANG Q Q, CHEN X D. SVD analysis of the relationship between Jiangnan rainy season precipitation and sea surface temperature in the tropical oceans [J]. Arid Meteor, 2004, 22: 11-16 (in Chinese).

[7] YANG F, LAU KM. Trend and variability of China precipitation in spring and summer: linkage to sea-surface temperatures [J]. Int J Climatol, 2004, 24: 1625-1644.

[8] TIAN S F, YASUNARI T. Climatological aspects and mechanism of Spring Persistent Rains over central China [J]. J Meteor Soc Jpn, 1998, 76(1): 57-71.

[9] WAN R J, WU G X. Temporal and spatial distribution of the spring persistent rains over southeast China [J]. Acta Meteor Sinica, 2009, 23(5): 598-608.

[10] WAN R J, WU G X. A study on the formation mechanism of the spring persistent rains climate [J]. Sci China Ser D-Earth Sci, 2006, 36: 936-950.

[11] ZHANG B, ZHONG S S, ZHAO B, et al. The influence of the subtropical sea surface temperature over the western Pacific on spring persistent rains [J]. J Appl Meteor Sci, 2011, 22(1): 57-65 (in Chinese).

[12] SHANG K, HE J H, ZHU Z W, et al. Comparison between correlations of heat content and sea surface temperature over western Pacific warm pool with spring persistent rain [J]. Sci Geogr Sinica, 2013, 33 (8): 986-992 (in Chinese).

[13] ZHANG W, LEUNG Y, CHAN JCL. The analysis of tropical cyclone tracks in the western North Pacific through data mining, Part I: Tropical cyclone recurvature [J]. J Appl Meteor Climatol, 2013a, 52(6): 1394-1416.

[14] ZHANG W, LEUNG Y, CHAN J C L. The analysis of tropical cyclone tracks in the western North Pacific through data mining, Part II: Tropical cyclone landfall [J]. J Appl Meteor Climatol, 2013, 52 (6): 1417-1432.

[15] GENG H T, SHI D W, ZHANG W, et al. A prediction scheme for the frequency of summer tropical cyclone landfalling over China based on data mining methods [J]. Meteorol Appl, 2016, 23(4): 587-593.

[16] DAVID A, JAMES O, JOHN K, MATTHIAS S. Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique [J]. Wea Forcasting, 2016, doi: 10.1175/WAF-D-15-0113.1.

[17] BAO Ming. Comparison of the effects of anomalous convective activities in the tropical western pacific on two persistent heavy rain events in south China [J]. J Trop Meteor, 2008, 14(1): 28-32.

[18] ZI Y, XU Y L, FU Y F. Climatological comparison studies between GPCP and rain gauges precipitations in China [J]. Acta Meteor Sinica, 2007, 65 (1): 63-74 (in Chinese).

[19] QUINLAN J R. Decision trees as probabilistic classifiers [A]. In: Proceedings of the 4th international workshop on machine learning[C]//American Association for Artificial Intelligence, Irvine, CA, 1987: 31-37.

[20] HAN J, KAMBER M. Data mining: Concepts and techniques [M]. Morgan Kaufmann, San Francisco, CA, 2006.

[21] HUANG R, WU Y. The Influence of ENSO on the summer climate change in China and its mechanism [J]. Advance in Atmospheric Science, 1989, 6(1): 21-32.

[22] ZHAO Z. Impact of El Niño events on atmospheric circulations in the Northern Hemisphere and precipitation in China [J]. Scientia Atmospherica Sinica, 1996, 20(4): 422-428 (in Chinese).

[23] LI C, XU H M, ZHU S X, et al. Numerical analysis on formation mechanism of spring persistent rain [J]. Plateau Meteor, 2010, 29(1): 99-108 (in Chinese).

Citation: LI Chao, SHI Da-wei, CHEN Yu-tian, et al. A prediction scheme for the precipitation of SPR based on the data mining algorithm and circulation analysis [J]. J Trop Meteor, 2019, 25(4): 519-527.