Decisive Atmospheric Circulation Indices for July–August Precipitation in North China Based on Tree Models

XUAN TONG, ZHONGWEI YAN, AND JIANGJIANG XIA

Institute of Atmospheric Physics, and University of Chinese Academy of Sciences, Beijing, China

XIAO LOU

Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

(Manuscript received 26 February 2019, in final form 21 May 2019)

ABSTRACT

Numerous circulation indices have been applied in practical climate services focused on regional precipitation. It is beneficial to identify the most influential or decisive indices, but this is difficult with conventional correlation analyses because of the underlying nonlinear mechanisms for precipitation. This paper demonstrates a set of the most influential indices for July–August precipitation in North China, based on the recursive random forest (RRF) method. These decisive circulation indices include the Polar–Eurasia teleconnection, North African subtropical high ridge position, India–Burma trough, Antarctic Oscillation, Northern Hemisphere polar vortex central latitude, North Atlantic Oscillation, and western Pacific subtropical high northern boundary position. Some of these factors have been recognized as directly influential to the regional precipitation, for example, those of the northwestern Pacific subtropical high; however, some are not easily understood. Decision tree (DT) models using these indices were developed to facilitate composite analyses to explain the RRF results. Taking one of the most interesting DT rules as an example, when the North African subtropical high ridge position is sufficiently far south, an anomalous anticyclone occurs in the upstream and an anomalous cyclone in the downstream of North China. This is unfavorable for northward moisture transport in eastern China and hence causes less precipitation in North China than climatology. The present results are not only helpful for improving diagnostic models of regional precipitation, but also enlightening for exploring how global climate change could impact a region by modulating large-scale circulation patterns.

1. Introduction

North China, one of the fastest developing regions in China, has been facing a severe water resource shortage for decades (Ma 2007), leading to huge investments in the South-to-North Water Transfer Project by the government. Sharp decreases of summer precipitation in North China, which have resulted from the stepwise weakening of the East Asian summer monsoon (EASM)

system during the past decades, have served as a climatic background (Tu et al. 2010) and increasingly become a matter of concern for scientists and policy-makers (Hao et al. 2010).

Precipitation in North China mainly occurs in July–August, which is the peak stage of the EASM. Previous studies have revealed many influential climate factors on the regional precipitation. The large-scale circulation patterns in association with the atmospheric moisture transport in the monsoon area are the main direct influential factors (Fig. 1; Huang et al. 2007; Lei and Duan 2011; Jin and Guan 2017). Precipitation in North China is also significantly influenced by remote mid-latitude circulations via the Silk Road pattern (SRP) and Pacific–Japan pattern, among others (Lu et al. 2002; Wang and He 2015; Qu et al. 2017). The atmospheric circulation also plays a role in linking the regional precipitation and other remote climate factors such as the

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-19-0045.s1>.

Corresponding author: Zhongwei Yan, yzw@tea.ac.cn

DOI: 10.1175/JHM-D-19-0045.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

FIG. 1. The stations (blue dots) in North China used in this study and the Asian monsoon partition. The arrows indicate major summer monsoon currents in Asia.

polar sea ice and the tropical sea surface temperature anomalies (Wang et al. 2000; Ma 2007; Si and Ding 2016; Yang et al. 2017; Yang and Ma 2017). Therefore, it is beneficial to identify influential circulation factors, not only for improving diagnostic and predictive models of regional precipitation, but also for exploring how large-scale climate change affects a region such as North China.

There are many circulation indices applied in regional climate services. For example, the China Meteorological Administration (CMA) has documented 88 large-scale circulation indices, which are expected to be applied in operational climate diagnosis and prediction (Gao et al. 2017; Gao et al. 2019). In practice, it is difficult to apply all these indices for a given regional problem, even if all the circulation factors are potentially influential. A common way to choose the most influential factors is via an analysis of linear regression or correlation (Wei and Huang 2010). However, linear methods may overlook important relationships in a nonlinear system such as the climate system (Park et al. 2016). The question regarding how to identify the relative importance of the numerous circulation indices for July–August precipitation in North China remains a challenge for researchers.

In recent years, machine learning has rapidly developed under the dramatic advances in the computer sciences. There have been increasing applications of machine learning methods in meteorology and climate studies (Kim et al. 2015; Park et al. 2016; Tao et al. 2018). The random forest model (RF) is one of

the best algorithms at the present because of its advantage in sufficient precision and robustness to outliers (Iverson et al. 2008). In general, the interpretability of RF and decision tree (DT) approaches is better than most of the other machine learning methods (Delerce et al. 2016), which is critical for climate diagnosis on limited data. Park et al. (2016) applied RF and DT in drought prediction based on 16 factors and assessed their relative importance. Delerce et al. (2016) compared linear models with tree models in eight aspects (e.g., nonlinear relationships, handling of missing values, handling mixed data) and concluded that the tree models were substantially better. DT has also been successfully employed in analyses of the recurvature, landfall, and intensity change of tropical cyclones in the western North Pacific (Zhang et al. 2013a,b,c; Yang 2016; Zhang et al. 2015; Gao et al. 2016; Geng et al. 2016). As climatic factors are often interrelated with each other, the collinearity often occurs when there are numerous factors involved in the process of model training. It is necessary to filter out these strongly correlated indices. It is also commonly understood that inclusion of more factors does not necessarily ensure a better performance of a model. To solve these problems, we applied a recursive nonlinear method based on RF itself, as RF has the robustness to counteract the noise, outliers and overfitting problems especially with limited data (Rhee et al. 2014).

The present study is focused on identifying the most important factors among the 78 large-scale circulation

indices for July–August precipitation in North China via the recursive random forest (RRF) method. The underlying mechanisms linking the large-scale circulation factors and the regional precipitation are demonstrated by using composite analyses of the tree models. The rest of the paper is organized as follows. The data, study period, and region are described in section 2. Section 3 introduces the machine learning methods used. The results are demonstrated and compared with some linear analyses in section 4. Conclusions and discussion are presented in section 5.

2. Data

This study used the following datasets:

- 1) The monthly precipitation records for 160 stations in China during 1951–2017, available from the CMA (<http://cmdp.ncc-cma.net/cn/index.htm>). This dataset has been extensively used for studying climate variability in the region (e.g., Ding et al. 2008; Si and Ding 2016; Yang et al. 2017; Yang and Ma 2017).
- 2) The monthly series of 88 atmospheric circulation indices during 1951–2017, available from http://cmdp.ncc-cma.net/Monitoring/cn_index_130.php, as listed in the supplemental material. Due to numerous missing values, 10 of the 88 indices (numbered 4, 9, 15, 20, 37, 42, 76, 79, 87, and 88 in the list in the supplemental material) were excluded. Consequently, only 78 indices were used in the following modeling analysis.
- 3) The JRA-55 global atmosphere reanalysis data since 1958, with geographical resolution of $1.25^\circ \times 1.25^\circ$ (Kobayashi et al. 2015), available from https://jra.kishou.go.jp/JRA-55/index_en.html#download. The variables used in this study include monthly geopotential height, and zonal and meridional winds.

Although the EASM rainy season typically extends from May to September, North China mainly receives rain during July–August, with climatological features different from those in the other East Asian regions (Xing et al. 2017). Therefore, this paper focuses on the monthly rainfall in July and August. There are different definitions of North China in different studies. This paper follows Huang et al. (1999), who divided China into seven regions according to geographical and climatic characteristics. Based on their definition of North China, 19 stations (out of the 160 stations) were used for this study (Fig. 1). These 19 stations have very similar seasonal cycles of precipitation, with a common peak in July–August (Fig. S1).

3. Methods

a. Decision tree

The DT method is commonly used in data mining. It applies tree-shaped rules to make decisions. A DT model is an inverted tree-shaped structure, of which each inner node represents a feature (factor), each rule refers to the test result and each leaf node denotes a class label. The top node in a tree is its root, which represents the most significant feature. In this study, we used the classification and regression tree (CART) algorithm (Breiman et al. 1984) to build a DT model, which partitions indices based on the Gini index. The DT classifier can perform automatic feature selection and complexity reduction. The tree structure provides easily interpretable information with regard to the predictive and generalization capacity of the classification (Sagheb et al. 2014).

b. Random forest

The RF method is based on CART DT (Breiman 2001) and produces numerous independent trees to reach a final decision through random bagging resampling (Prasad et al. 2006) to the selection of a random subset of training samples and a random subset of candidate variables at each node of a tree. Due to the randomization, RF has the robustness to counteract the noise, outliers and overfitting potentially suffered by a DT model (Rhee et al. 2014). Theoretically, the probability that a sample is never picked up in m times sampling is $\lim_{m \rightarrow \infty} [1 - (1/m)]^m$, which is about 36.8%. These data are called out-of-bag (OOB) data. With OOB data from random selections, RF provides internal cross validation and relative importance of a factor variable when samples are OOB (Long et al. 2013; Kim et al. 2014). RF provides mean decrease accuracy or mean decrease Gini index in classification when a variable is permuted. The greater this decrease in accuracy, the more important the variable.

c. Recursive RF

For the present study, we developed RRF for the first time. The main idea of RRF is to divide all indices into a few classes based on intercorrelation and then to construct RF for each of these classes. Finally, a chosen number of effective indices from the 78 indices are applied to construct the final RF model (see the appendix). To accomplish RRF in this study, six models were developed as follows: 1) control trial 1 (CT1), constructed by the common RF procedure, designed to contrast the advantage of RRF; 2) control trial 2 (CT2), constructed by the logistic regression, designed

FIG. 2. The generalization capacity of the six models designed for RRF. The blue bar represents the precision value, and the red bar represents the AUC.

to show the advantage of RRF beyond the linear method; 3), 4), and 5) intermediate trials 1–3 (IT1–3) of RRF; and finally 6) the optimal RF model (ORF). The intermediate RF models (IT1–3) were constructed recursively as the latter model (IT3) uses the results of the former models (IT1–2). More details are provided in the [appendix](#).

d. Performance evaluation

In this study, the precision value (PV) and the receiver operating characteristics (ROC; [Zou et al. 2007](#)) were used to assess the generalization of a model. The PV represents how well the model outcome matches the observation. The ROC is a graphical method for evaluating, organizing and selecting classifiers based on their performances, especially for the two-class model. The area under the ROC curve (AUC) is an indicator of the performance of the model. The closer this AUC value is to 1, the better the classifier's performance. The AUC value can be categorized as poor (0.5–0.6), average (0.6–0.7), good (0.7–0.8), very good (0.8–0.9), or excellent (0.9–1) ([Chen et al. 2018](#)). More details of the ROC can be found in [Fawcett \(2006\)](#) and [Flach et al. \(2003\)](#).

4. Results

a. General model performance

The PV and AUC of the ORF were 0.73 and 0.75, respectively, which was the best performance among the six models analyzed ([Fig. 2](#)). The AUC of the ORF was

larger than that of CT1 by 10%, suggesting that collinearity harms the performance of RF (CT1).

The generalization capacity of the linear model (CT2) was much lower than that of the ORF, and also less than that of the conventional RF model (CT1) when involving all 78 indices. This indicates that RF identifies the most influential or decisive factors more effectively than the linear method. Of the 78 circulation indices, 9 were finally applied to construct the ORF (as listed in [Fig. 4](#)). Note that the intermediate model IT3 applied 26 indices as input [details in section c(2) of the [appendix](#)], which were weakly correlated with each other, but its performance was still weaker than that of the ORF. The AUC of the ORF was 4% larger than that of IT3. The weakened performance of RF with more factors was probably associated with insufficient observation samples, although bagging resampling eased this problem to a certain extent. In addition, we used the chosen nine circulation indices used in the ORF to construct a DT model (e.g., CART) for comparing the performance of RF and DT in classifying the rainfall in North China. The generalization ability of this DT model is 0.68 based on fivefold cross validation and that of RRF is 0.75. This indicates a better performance of RRF than DT due to the typical randomization in RF.

To illustrate how well the ORF model matched the observed monthly rainfall records in North China, we selected two of the most important circulation indices derived by the ORF, the Polar–Eurasia pattern (POL) and the North African subtropical high ridge position

(NASHRP), to form a 2D space and plot the distribution of the monthly rainfall records in this space. The monthly rainfall records were categorized into two states, P and N, representing more and less than the mean climatology during 1951–2017, respectively. As shown in Fig. 3, the 2D space can be divided into four phases. Most of the P months were gathered in and around the second phase (mainly in the blue oval in Fig. 3), suggesting that when the NASHRP is anomalously northward and the POL is weak, there is more rainfall in North China. In contrast, a large part of the N months gathered in the fourth phase, suggesting that when the NASHRP is anomalously southward and the POL is strong, there is less rainfall than climatology in North China. A comparison between Fig. 3a and Fig. 3b clearly shows that the simulated distribution of the peak-summer monthly rainfall states in North China from the ORF matched the observations very well.

b. The relative importance of circulation indices

The relative importance of the 78 indices was analyzed via CT1, which ranked the subtropical high, POL, and polar vortex as the top three circulation factors influencing the peak summer precipitation over North China.

During the RRF process, the total 78 indices were divided into three classes based on intercorrelation [see section c(1) of the appendix] and the relative importance of indices in class 1 and class 2 were calculated. The outcomes reveal that the most effective index in class 1 was the Western Pacific Subtropical High Intensity. The other indices in class 1 were abandoned due to collinearity. In class 2, there were 33 indices, and the top 5 [floor($\sqrt{p}$), $p = 33$; see section c(2) of the appendix] indices were the NASHRP, western Pacific subtropical high western ridge point, POL, India–Burma trough intensity and central Pacific 850-mb trade wind (1 mb = 1 hPa). The other indices of class 2 were also abandoned because of the intercorrelation problem.

Having excluded those factors with collinearity, 26 out of the 78 indices remained for further analysis [see details in section c(2) of the appendix]. The IT3 model was used to calculate the relative importance of these 26 indices. Figure 4 shows that the top two most important indices were the POL and the NASHRP index, which was consistent with the outcome of CT1. This suggests that the main characteristics of the circulation–rainfall relationship were captured despite the substantial reduction of indices due to intercorrelation. Previous studies have suggested that the POL and NASHRP are likely to play crucial roles in summer rainfall in North China (Lin 2014; Jia et al. 2002). The

highest generalization of the ORF suggests that the relative importance of the chosen indices by RRF is reasonable. As shown in Fig. 4, the most influential factors for peak summer precipitation in North China include the POL, subtropical high ridge (SHR), India–Burma trough, and Antarctic Oscillation (AAO).

c. Major underlying mechanisms

The interpretability of the RF is not straightforward because it is an ensemble of massive DT models. The interpretability of DT is good although its generalization was poor compared with RF. To help understand the RRF results, we used the top nine indices in the ORF to construct an optimal DT model. The process of constructing the optimal DT model was similar to RF. The root node of the optimal DT was the NASHRP index, which was in the second position in RF. According to the DT model, when the NASHRP is farther north, there is more precipitation in North China in general. This conclusion is consistent with the study by Jia et al. (2002), who found a large correlation coefficient between summer precipitation in China and the NASHRP index. The rules of the optimal DT are illustrated in Fig. 5 and detailed in Table 1.

To explore possible mechanisms underlying the DT rules, we took three major rules as examples, that is, rules 1–3 (details in Table 1 and Fig. 5). The PVs of these chosen rules are the highest (larger than 0.8) among all the six rules of the model, which imply robust underlying physical mechanisms. In the composite analyses of the atmospheric circulation, we focused on August partly because some teleconnections such as the SRP involved in the DT rules are relatively stable in August (Thompson et al. 2019).

For rule 1, there are 25 monthly samples, 14 of which occur in August. The composite circulation fields of these 14 months are shown in Fig. 6, to visualize why a very southerly location of the NASHRP ($< -1.04^\circ$) is unfavorable for precipitation in North China. At 500-hPa geopotential height (Fig. 6a), a wide range of positive anomalies occurred over central Africa (mainly to the south of 25°N) and an anomalous negative center occurred over the Mediterranean region. Usually, the North African subtropical high is located around 30°N in August, so the anomalous pattern in Fig. 6a indicates that the North African subtropical high was located farther southward than climatology, as rule 1 shows. Figure 6a also shows that an anomalous positive center or anticyclone existed in the upstream of North China. Furthermore, an anomalous negative center or cyclone occurred in the downstream of North China. Figure 6b shows the composite 250-hPa geopotential height field,

FIG. 3. 2D scatter diagram of (a) the peak-summer monthly rainfall records in North China and (b) the ORF model output. The two axes represent the POL and NASHRP indices, respectively. Blue/red dots represent the monthly precipitation in North China more/less than climatology. Most of the monthly records of more/less precipitation are outlined by the blue/red oval.

FIG. 4. Relative importance of the 26 indices in IT3; the x axis is the relative importance of each index and the y axis is the corresponding serial number of the raw 88 circulation indices. The top 9 indices are used to construct the optimal RF model: Polar–Eurasia pattern (77), North African subtropical high ridge position index (24), India–Burma trough intensity index (67), Antarctic Oscillation (69), Northern Hemisphere polar vortex central latitude index (57), North Atlantic Oscillation (70), western Pacific subtropical high northern boundary position index (38), western Pacific subtropical high western ridge point index (45), and Asia polar vortex intensity index (51).

as done in other studies of relevant circulation patterns (e.g., Thompson et al. 2019). There was an anomalous atmospheric wave pattern extending from Europe to eastern Asia, with anomalous northerly wind over North China (Fig. 6b). This wave pattern has been referred to as the SRP in previous studies. Figure 6b suggests that the SRP plays a role in connecting the anomalous circulation around North Africa and the peak summer rainfall in North China. The anomalous northerly winds in the mid- to upper troposphere over North China are surely unfavorable for the northward transport of air moisture from the lower latitudes into North China and therefore lead to less rainfall in North China than climatology.

For rule 2, there were 18 monthly samples, 11 of which occurred in August. Figure 7 shows the composite anomalous circulation fields of these 11 months. The anomalous southward location of the polar vortex at the 500 hPa can be inferred in Fig. 7a [$-4.42 < \text{Northern Hemisphere polar vortex central latitude (NHPVCL)} \leq 0.58$, as defined in rule 2, $\text{NHPVCL} = \text{the latitude of the center of the polar vortex, the lowest geopotential height nearby the polar}$

region at 500 hPa]. As Fig. 7a shows, the anomalous positive center in the polar region indicates a weaker polar vortex. In this case, North China was located at the bottom of a weak anomalous cyclone, and an anomalous anticyclone occurred over the coast of southeastern China. As a result, there was an anomalous northerly cold current and a southerly warm moist current converging in North China, which produced favorable conditions for higher precipitation in North China than climatology. For rule 2, there was no obvious signal at the upper troposphere (Fig. 7b).

For rule 3, there were 26 monthly samples, 14 of which occurred in August. In this case, the composite signal of the anomalous circulation was weak at 500 hPa, but obvious at 850 hPa. As shown in Fig. 8a, the anomalous negative center in the northern polar region indicates a strengthened polar vortex and hence the anomalously northward location of the lowest geopotential height. This matches with rule 3 ($\text{NHPVCL} > 3.08$). In this case, an anomalous positive center occurred downstream of North China. There was a wide-range anomalous anticyclone over the northwestern Pacific. As a result, the anomalous

FIG. 5. The optimal DT model using the top 9 influential indices in the ORF. The red curve indicates a rule of DT used for a composite analysis. The oval frame represents an inner node, which means the index used for partition. The box represents a leaf node, which means the final classification result for each rule. Arrow connectors represent conditions of the index for partition.

southerly wind prevailed over North China, bringing moisture from the lower latitudes into North China and therefore leading to more rainfall than climatology in North China. Figure 8b shows the SRP at 250 hPa in terms of the composite anomalous meridional winds. This SRP pattern has been regarded as

responsible for unprecedented heat waves in Southeast China (Thompson et al. 2019), because of the anomalous anticyclone over the south of eastern China. Compared with the SRP in rule 1, this SRP had an eastward shift, with a southerly wind over North China. Undoubtedly, different phases of the

TABLE 1. The rules of diagnosing July and August rainfall records in North China based on circulation indices as discovered by the optimal DT.

Decision rule	Decision indices	PV
1. If NASHRP ≤ -1.04 , then less precipitation in North China	NASHRP	21/25 = 84%
2. If NASHRP > -1.04 and $-4.42 < \text{NHPVCL} \leq 0.58$ and $\text{POL} \leq 0.2$, then more precipitation in North China	NASHRP, NHPVCL, and POL	17/18 = 94%
3. If NASHRP > -1.04 and $\text{NHPVCL} > 3.08$ and $\text{POL} \leq 0.2$, then more precipitation in North China	NASHRP, NHPVCL, and POL	23/26 = 88%
4. If NASHRP > -1.04 and $\text{NHPVCL} \leq -4.42$, then less precipitation in North China	NASHRP and NHPVCL	15/21 = 71%
5. If NASHRP > -1.04 and $0.58 < \text{NHPVCL} \leq 3.08$ and $\text{POL} \leq 0.2$, then less precipitation in North China	NASHRP, NHPVCL, and POL	12/20 = 60%
6. If NASHRP > -1.04 and $\text{NHPVCL} > -4.42$ and $\text{POL} > 0.2$, then less precipitation in North China	NASHRP, NHPVCL, and POL	15/24 = 63%

FIG. 6. (a) Composite field of the 500-hPa geopotential height (normalized) and the 850-hPa wind (normalized) of rule 1 in August; the shaded colors represent the 500-hPa geopotential height and the arrows represent the 850-hPa wind. (b) The 250-hPa meridional winds (normalized) in August; the red box outlines North China.

SRP could have different effects on precipitation in North China.

5. Conclusions and discussion

This paper demonstrates the most influential or decisive circulation factors for July–August precipitation in North China from a nonlinear climate perspective among the 88 circulation indices operationally applied at the CMA, based on the RRF method. DT models using these indices were developed to facilitate composite analyses to explain the RRF results. The RRF method was designed partly to solve the collinearity problem, which is found in statistical modeling and climate diagnosis, especially when numerous factors are

involved. The main conclusions of this study were as follows:

- 1) The RRF method had the best generalization capacity of the different models developed in this paper especially better than logistic regression, a traditional linear method. The optimal RF model based on RRF not only captured the main physical relationships derived from RF, but also showed a better generalization capacity.
- 2) The most influential or decisive circulation factors for rainfall in the peak summer months in North China, derived from RRF, included the POL, NASHRP, India–Burma trough, AAO, Northern Hemisphere polar vortex central latitude, North

FIG. 7. As in Fig. 6, but for rule 2.

Atlantic Oscillation, and western Pacific subtropical high northern boundary position. Some of these are well known in literature of climate studies for the region, whereas some need further research.

- 3) DT can help understand the RRF results. For the major rules of the DT model, there are robust underlying physical mechanisms as revealed by the composite analyses of the same-rule samples.

The peak summer precipitation in North China is modulated by combined effects of numerous anomalous circulation patterns. The present results also can help us to understand how large-scale climate change could influence regional precipitation. For example, the North Africa subtropical high not only played an important role in modulating interannual variations of summer rainfall in North China, as derived from the present

RRF analysis, but also showed a long-term trend that was in phase with the decreasing trend of summer rainfall in North China since the late 1970s (Fig. S2). In addition, although the present model is a diagnostic model in order to reveal the decisive circulation factors for precipitation in North China, it is possible to develop a predictive model if some precursor signals of the decisive circulation factors can be incorporated via further statistical analyses and incorporating operational numerical weather/climate predictions. These deserve further studies. It is noted that, due to the limited data, the present RRF modeling cannot capture all of the nonlinear relationships that occurred. Nevertheless, with increasing observations, the nonlinear method has potential for improving modeling of the regional climate, whereas linear methods do not because they are based on stationary statistics. The present results are not

FIG. 8. As in Fig. 6, but for rule 3 and the composite field of geopotential height is at 850 hPa.

only helpful for improving diagnostic models of regional precipitation, but also enlightening for exploring how global climate change could impact a region via responses in large-scale circulation patterns.

Acknowledgments. This study is supported by the National Key R&D Program (2016YFA0600404) and the CAS project (XDA20020201).

APPENDIX

Details of the Recursive Random Forest (RRF) Procedure

a. The purpose of RRF

As climatic factors are often interrelated, the collinearity often occurs when there are many factors involved

in the process of model training. Moreover, it is widely accepted that more factors do not necessarily ensure a better performance of the model. To solve these problems, we applied the RRF method, a nonlinear method based on random forest (RF) itself.

b. Data preprocessing

Before constructing a model, the dataset is processed into $(X_1, X_2, \dots, X_n, Y)$ for training models, where n is the total number of the circulation indices used in a model, X is the set of n anomalous circulation indices, and Y is the labeled data: $Y = 1$ (0) means more (less) precipitation in North China than climatology.

c. Workflow of RRF

The general workflow of RRF is shown in Fig. A1, with details as follows.

relative importance of each index was also calculated ($i = 2$ in Fig. A1).

- 3) The top index (western Pacific subtropical high intensity index) in class 1, determined by IT1 and the top floor($\sqrt{p}$) indices (NASHRP index, western Pacific subtropical high western ridge point index, India–Burma trough intensity index, POL, and central Pacific 850-mb trade wind index), in class 2 (where p is the number of indices in class 2) determined by IT2 were combined with those in class 3, as the input X filtering dataset. This input X filtering dataset was used to construct a RF model to calculate the relative importance of each index in the input X filtering dataset [$i = 3$ in Fig. A1, hereinafter referred to as intermediate trial 3 (IT3)].
- 4) Finally, the descending order indices calculated by IT3 were added into the RF model one by one and the optimal number of indices was determined with the best generalization value. The determined indices were used to construct the optimal RF model (ORF), namely RRF ($i = 4$ in Fig. A1).

Besides the above four models, two additional models were designed as control models:

- 1) control trial 1 (CT1), an RF model constructed with all the 78 indices, was designed to show the advantage of RRF over the common RF method;
- 2) control trial 2 (CT2), constructed by the logistic regression, was designed to show the advantage of RRF over the common linear method.

3) PARAMETER SELECTION

All six models were constructed through relevant parameter selection. The details of parameter selection are listed in the supplemental material (Table S2). More explanations of the parameters can be found in the sklearn toolkit (<https://scikit-learn.org/stable/index.html>).

REFERENCES

- Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, <https://doi.org/10.1023/A:1010933404324>.
- , J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984: *Classification and Regression Trees*. Wadsworth and Brooks, 385 pp.
- Chen, W., H. R. Pourghasemi, and S. A. Naghibi, 2018: A comparative study of landslide susceptibility maps produced using support vector machine with different kernel functions and entropy data mining models in China. *Bull. Eng. Geol. Environ.*, **77**, 647–664, <https://doi.org/10.1007/s10064-017-1010-y>.
- Delerce, S., H. Dorado, A. Grillon, M. C. Rebolledo, S. D. Prager, V. H. Patino, G. G. Varon, and D. Jimenez, 2016: Assessing weather–yield relationships in rice at local scale using data mining approaches. *PLOS ONE*, **11**, e0161620, <https://doi.org/10.1371/journal.pone.0161620>.
- Ding, Y. H., Z. Y. Wang, and Y. Sun, 2008: Inter-decadal variation of the summer precipitation in East China and its association with decreasing Asian summer monsoon. Part I: Observed evidences. *Int. J. Climatol.*, **28**, 1139–1161, <https://doi.org/10.1002/joc.1615>.
- Fawcett, T., 2006: An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874, <https://doi.org/10.1016/j.patrec.2005.10.010>.
- Flach, P., H. Blockeel, C. Ferri, J. Hernández-Orallo, and J. Struyf, 2003: Decision support for data mining: Introduction to ROC analysis and its applications. *Data Mining and Decision Support: Integration and Collaboration*, D. Mladenović et al., Eds., Kluwer Academic Publishers, 81–90, https://doi.org/10.1007/978-1-4615-0286-9_7.
- Gao, L. H., F. Y. Wei, Z. W. Yan, J. Ma, and J. J. Xia, 2019: A study of objective prediction for summer precipitation patterns over eastern China based on multinomial logistic regression model. *Atmosphere*, **10**, 213, <https://doi.org/10.3390/atmos10040213>.
- Gao, S., W. Zhang, I. I. Lin, and L. S. Chiu, 2016: Improvements in typhoon intensity change classification by incorporating an ocean coupling potential intensity index into decision trees. *Wea. Forecasting*, **31**, 95–106, <https://doi.org/10.1175/WAF-D-15-0062.1>.
- Gao, T., J. Y. Yu, and H. Paek, 2017: Impacts of four northern-hemisphere teleconnection patterns on atmospheric circulations over Eurasia and the Pacific. *Theor. Appl. Climatol.*, **129**, 815–831, <https://doi.org/10.1007/s00704-016-1801-2>.
- Geng, H. T., D. W. Shi, W. Zhang, and C. Huang, 2016: A prediction scheme for the frequency of summer tropical cyclone landfalling over China based on data mining methods. *Meteor. Appl.*, **23**, 587–593, <https://doi.org/10.1002/met.1580>.
- Hao, L. S., J. Z. Min, Y. H. Ding, and J. Wang, 2010: Relationship between reduction of summer precipitation in North China and atmospheric circulation anomalies. *J. Water Resource Prot.*, **2**, 569–576, <https://doi.org/10.4236/jwarp.2010.26065>.
- Huang, R. H., Y. H. Xu, and L. T. Zhou, 1999: The interdecadal variation of summer precipitation in China and the drought trend in North China (in Chinese). *Plateau Meteor.*, **18**, 465–476.
- , J. L. Chen, and G. Huang, 2007: Characteristics and variations of the East Asian monsoon system and its impacts on climate disasters in China. *Adv. Atmos. Sci.*, **24**, 993–1023, <https://doi.org/10.1007/s00376-007-0993-x>.
- Iverson, L. R., A. M. Prasad, S. N. Matthews, and M. Peters, 2008: Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *For. Ecol. Manage.*, **254**, 390–406, <https://doi.org/10.1016/j.foreco.2007.07.023>.
- Jia, L., S. W. Zhou, and F. Ding, 2002: The climatic character of North Africa subtropical high and its effect on summer precipitation in China (in Chinese). *Trans. Atmos. Sci.*, **25**, 816–822.
- Jin, D., and Z. Guan, 2017: Summer rainfall seesaw between Hetao and the middle and lower reaches of the Yangtze River and its relationship with the North Atlantic Oscillation. *J. Climate*, **30**, <https://doi.org/10.1175/JCLI-D-16-0760.1>.
- Kim, M., J. Im, H. Han, J. Kim, S. Lee, M. Shin, and H. S. F. Kim, 2015: Landfast sea ice monitoring using multisensor fusion in the Antarctic. *GLSci. Remote Sens.*, **52**, 239–256, <https://doi.org/10.1080/15481603.2015.1026050>.
- Kim, Y. H., J. Im, H. K. Ha, J.-K. Choi, and S. Ha, 2014: Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GLSci. Remote Sens.*, **51**, 158–174, <https://doi.org/10.1080/15481603.2014.900983>.

- Kobayashi, S., and Coauthors, 2015: The JRA-55 Reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan Ser. II*, **93**, 5–48, <https://doi.org/10.2151/jmsj.2015-001>.
- Lei, Y. H., and A. M. Duan, 2011: Prolonged dry episodes and drought over China. *Int. J. Climatol.*, **31**, 1831–1840, <https://doi.org/10.1002/joc.2197>.
- Lin, Z. D., 2014: Intercomparison of the impacts of four summer teleconnections over Eurasia on East Asian rainfall. *Adv. Atmos. Sci.*, **31**, 1366–1376, <https://doi.org/10.1007/s00376-014-3171-y>.
- Long, J. A., R. L. Lawrence, M. C. Greenwood, L. Marshall, and P. R. Miller, 2013: Object-oriented crop classification using multitemporal ETM+ SLC-off imagery and random forest. *GLSci. Remote Sens.*, **50**, 418–436, <https://doi.org/10.1080/15481603.2013.817150>.
- Lu, R.-Y., J.-H. Oh, and B.-J. Kim, 2002: A teleconnection pattern in upper-level meridional wind over the North African and Eurasian continent in summer. *Tellus*, **54A**, 44–55, <https://doi.org/10.3402/tellusa.v54i1.12122>.
- Ma, Z. G., 2007: The interdecadal trend and shift of dry/wet over the central part of North China and their relationship to the Pacific Decadal Oscillation (PDO). *Chin. Sci. Bull.*, **52**, 2130–2139, <https://doi.org/10.1007/s11434-007-0284-z>.
- Park, S., J. Im, E. Jang, and J. Rhee, 2016: Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. *Agric. For. Meteorol.*, **216**, 157–169, <https://doi.org/10.1016/j.agrformet.2015.10.011>.
- Prasad, A. M., L. R. Iverson, and A. Liaw, 2006: Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, **9**, 181–199, <https://doi.org/10.1007/s10021-005-0054-1>.
- Qu, J. X., D. Y. Gong, M. Rui, J. Yang, and S. Li, 2017: Possible influence of Arctic oscillation on precipitation along the East Asian rain belt during boreal spring. *Theor. Appl. Climatol.*, **130**, 487–495, <https://doi.org/10.1007/s00704-016-1900-0>.
- Rhee, J., S. Park, and Z. Lu, 2014: Relationship between land cover patterns and surface temperature in urban areas. *GLSci. Remote Sens.*, **51**, 521–536, <https://doi.org/10.1080/15481603.2014.964455>.
- Saghebian, S. M., M. T. Sattari, R. Mirabbasi, and M. Pal, 2014: Ground water quality classification by decision tree method in Ardebil region, Iran. *Arab. J. Geosci.*, **7**, 4767–4777, <https://doi.org/10.1007/s12517-013-1042-y>.
- Si, D., and Y. H. Ding, 2016: Oceanic forcings of the interdecadal variability in East Asian summer rainfall. *J. Climate*, **29**, 7633–7649, <https://doi.org/10.1175/JCLI-D-15-0792.1>.
- Tao, Y., K. Hsu, A. Ihler, X. Gao, and S. Sorooshian, 2018: A two-stage deep neural network framework for precipitation estimation from bispectral satellite information. *J. Hydrometeorol.*, **19**, <https://doi.org/10.1175/JHM-D-17-0077.1>.
- Thompson, V., N. J. Dunstone, A. A. Scaife, D. M. Smith, S. C. Hardiman, H. L. Ren, B. Lu, and S. E. Belcher, 2019: Risk and dynamics of unprecedented hot months in South East China. *Climate Dyn.*, **52**, 2585–2596, <https://doi.org/10.1007/s00382-018-4281-5>.
- Tu, K., Z. W. Yan, and W. J. Dong, 2010: Climatic jumps in precipitation and extremes in drying North China during 1954–2006. *J. Meteor. Soc. Japan*, **88**, 29–42, <https://doi.org/10.2151/jmsj.2010-103>.
- Wang, B., R. G. Wu, and X. H. Fu, 2000: Pacific-East Asian teleconnection: How does ENSO affect East Asian climate? *J. Climate*, **13**, 1517–1536, [https://doi.org/10.1175/1520-0442\(2000\)013<1517:PEATHD>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<1517:PEATHD>2.0.CO;2).
- Wang, H. J., and S. P. He, 2015: The North China/Northeastern Asia severe summer drought in 2014. *J. Climate*, **28**, 6667–6681, <https://doi.org/10.1175/JCLI-D-15-0202.1>.
- Wei, F. Y., and J. Y. Huang, 2010: A study of downscaling factors atmospheric circulation in the prediction in the prediction model of summer precipitation in Eastern China (in Chinese). *Chin. J. Atmos. Sci.*, **34**, 202–212.
- Xing, W., B. Wang, S. Yim, and K. Ha, 2017: Predictable patterns of the May–June rainfall anomaly over East Asia. *J. Geophys. Res. Atmos.*, **122**, 2203–2217, <https://doi.org/10.1002/2016JD025856>.
- Yang, R., 2016: A systematic classification investigation of rapid intensification of Atlantic tropical cyclones with the SHIPS database. *Wea. Forecasting*, **31**, 495–513, <https://doi.org/10.1175/WAF-D-15-0029.1>.
- Yang, Q., and Z. G. Ma, 2017: Decadal modulation of precipitation patterns over eastern China by sea surface temperature anomalies. *J. Climate*, **30**, 7017–7033, <https://doi.org/10.1175/JCLI-D-16-0793.1>.
- , —, and B. L. Xu, 2017: Modulation of monthly precipitation patterns over East China by the Pacific Decadal Oscillation. *Climatic Change*, **144**, 405–417, <https://doi.org/10.1007/s10584-016-1662-9>.
- Zhang, W., Y. Leung, and J. C. L. Chan, 2013a: The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part I: Tropical cyclone recurvature. *J. Appl. Meteor. Climatol.*, **52**, 1394–1416, <https://doi.org/10.1175/JAMC-D-12-045.1>.
- , —, and —, 2013b: The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part II: Tropical cyclone landfall. *J. Appl. Meteor. Climatol.*, **52**, 1417–1432, <https://doi.org/10.1175/JAMC-D-12-046.1>.
- , S. Gao, B. Chen, and K. Cao, 2013c: The application of decision tree to intensity change classification of tropical cyclones in western North Pacific. *Geophys. Res. Lett.*, **40**, 1883–1887, <https://doi.org/10.1002/grl.50280>.
- , B. Fu, M. S. Peng, and T. Li, 2015: Discriminating developing versus nondeveloping tropical disturbances in the western North Pacific through decision tree analysis. *Wea. Forecasting*, **30**, 446–454, <https://doi.org/10.1175/WAF-D-14-00023.1>.
- Zou, K. H., A. J. O'Malley, and L. Mauri, 2007: Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, **115**, 654–657, <https://doi.org/10.1161/CIRCULATIONAHA.105.594929>.