METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **23**: 587–593 (2016)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.1580

RMetS
Royal Meteorological Society

A prediction scheme for the frequency of summer tropical cyclone landfalling over China based on data mining methods

Huantong Geng,[a]* Dawei Shi,[a,b] Wei Zhang[a] and Chao Huang[a]
[a] *College of Atmospheric Science, Nanjing University of Information Science and Technology, Nanjing, China*
[b] *Meteorological Observatory, Meteorological Bureau of Lianyungang City, China*

ABSTRACT: This study examines the landfalling tropical cyclones (TCs) over *C*hina using state-of-the-art data mining methods (i.e. Finite Mixture Model (FMM) based cluster algorithm and the Classification and Regression Tree (CART)). Using the 1951–2012 TC best track dataset released by the Shanghai Typhoon Institute of the Chinese Meteorological Administration, the tracks of TCs landfalling over the Chinese coast were classified into three clusters through an FMM. Several climate indices were analysed using the CART algorithm for the three clusters. The prediction model built by CART for summer track frequency was based on a random sampling of the data for 46 years (about 75% of the total years) as the training set with a training accuracy of 100% (Cluster-1), 89.96% (Cluster-2) and 100% (Cluster-3). Data for the remaining 16 years (about 25%) were used for testing with a prediction accuracy of 87.5% (Cluster-1), 62.5% (Cluster-2) and 68.75% (Cluster-3). This study focuses on Cluster-1 of summer TCs landfalling over China for its high frequency, strong intensity, severe impacts and long lifespan. Furthermore, it suggests that the FMM algorithm is effective for track classification of TCs landing over China. In addition, the CART algorithm, which was used to build the prediction model of Cluster-1 for the classification of track frequency, showed high accuracy and its results can be explained and understood easily. It provides a novel framework for forecasting the frequency of TCs landfalling over China.

KEY WORDS Cluster-1 TC track; FMM algorithm; CART algorithm

Received 26 September 2015; Revised 7 March 2016; Accepted 7 March 2016

1. Introduction

A tropical cyclone (TC) is one of the most destructive causes of natural disasters (Zhu *et al.*, 2012; Zhang *et al.*, 2015). A landfalling TC can bring devastating impacts, both social and economic, to coastal and even inland areas (Zhu *et al.*, 2012). China has more TC landfalls than other countries. Every summer (June, July and August; referred to as JJA), an average of six TCs move across the Chinese coastline. Therefore, studies of summer TCs landfalling over China have both theoretical and practical significance in improving the accuracy of forecasting TCs.

The influence of a TC is associated almost entirely with its track. TC activity is affected heavily by large-scale circulation, as well as by sea surface temperature in local and remote areas (Wang and Chan, 2002). Wang and Chan (2002) found that a TC is more likely to recurve and could reach higher latitudes in El Niño years; the lifespan and intensity of a TC tends to be longer and stronger in El Niño years than in La Niña years; but the total frequency of TCs landfalling is not significantly related to the El Niño-Southern Oscillation (ENSO). Tao *et al.* (2013) pointed out that the seasonal prediction was impeded by strong TCs, as well as instability of the ENSO; He Pengcheng also discovered that the Pacific Decadal Oscillation (PDO) modulates the influence of the ENSO on TCs over the Western North Pacific (WNP) (He and Jing, 2011). Liu and Chan suggested that the interdecadal variation of TC track is possibly correlated with PDO (Liu and Chan, 2008). Li *et al.* (2011) suggested that the frequency of the northward TCs in East Asia has an out-of-phase relationship with the PDO. When the subtropical high becomes stronger or extends westward, the TCs formed over the South China Sea and the WNP are more likely to move northwestwards and recurve (Ho *et al.*, 2004; Goh and Chan, 2010) and to make landfall over the Chinese coast. A pioneering study on the influence of the quasi-biennial Oscillation (QBO) on TC frequency was conducted by Gray (1984a, 1984b). Chan (1995) summarized that, in non El Niño years, the QBO played a major role in modulating TC frequency over the WNP. The modulation of the QBO on TC tracks was also identified by Ho *et al.* (2009). As stated by Tao *et al.* (2012), weak TCs that form over the WNP are affected mainly by the warming of the Indian Ocean and have a negative correlation with the warming of the eastern Indian Ocean (Zhan *et al.*, 2011). The above factors can be taken as potential predictors for TC frequency.

Along with the increasing availability of meteorological data, data mining that uses an expert system to discover potential and useful information has been applied widely (Han and Kamber, 2006). Several studies have applied this technology to the study of meteorology, which routinely involves large volumes of data (Han and Kamber, 2006; Shi *et al.*, 2015). Camargo *et al.* (2007a) employed the clustering method based on the Finite Mixture Model (FMM) to classify TC tracks over the WNP by using a quadratic polynomial regression function, in which the longitude and latitude time sequences of track points were used as independent variables (Camargo *et al.*, 2007a, 2007b). Zheng *et al.* (2013) also classified the TC tracks over the same ocean using the *k*-mean clustering algorithm, which is not adaptive to track length (Camargo *et al.*, 2007a). Zhang *et al.* (2013a) classified and predicted the changes of TCs (strengthened and weakened) by using a decision tree algorithm. This algorithm also performed

* Correspondence: H. Geng, No. 219, Ningliu Road, Nanjing, Nanjing University of Information Science and Technology, Nanjing 210044, China. E-mail: htgeng@nuist.edu.cn

well when applied to TC landfalls and recurvature (Zhang *et al.*, 2013b, 2013c). As a consequence, the FMM algorithm was used in the present paper to classify the tracks of TCs landfalling over China during 1951–2012. Additionally, it was taken as a binary classification with which to examine whether the TC frequency in three clusters during summer was high or low, and whether the algorithm of the Classification and Regression Tree (CART), proposed by Breiman *et al.* (1984), can be used to build the decision tree. Finally, the prediction model was verified using an independent test data set.

In recent years, climate models have frequently been used to predict TC landfalls. Compared with previous climate models, the present study aimed to build a novel model to predict TC frequency over China by using the data mining approach together with already known variables of TC frequency over the WNP. The new research may enhance the understanding, prediction and management of TC landfalls over China.

The remainder of this study is organized as follows. Section 2 presents data and methodology. Section 3 discusses the experimental results based on the FMM and CART, followed by Section 4 that reaches conclusions.

2. Data and methodology

2.1. Data source

The influence of TCs over China prevails from June to November, and the summer (JJA) is the season with the highest frequency of tropical cyclones making landfall over China. TC data were collected for this study from the best track dataset for 1951–2012, compiled by the Shanghai Typhoon Institute of the Chinese Meteorological Administration, including the latitude and longitude and the maximum sustained wind (MSW) observed at a 6 h intervals. Samples used for the study were TCs landfalling over China from June to November, with a life longer than one day and with MSW $>17.2\ \mathrm{m\ s^{-1}}$. Here, the observation at which TC intensity first reaches $17.2\ \mathrm{m\ s^{-1}}$ is defined as the location of TC genesis. The data also include indices of the ENSO, PDO and QBO collected from the National Oceanic and Atmospheric Administration and climate indices from the Northern Hemisphere summer in the 74 circulation indices of the National Climate Center (see Table 2). The summer values of the indices mentioned above were attained by calculating the average value of those indices during June–August.

2.2. Methodology

2.2.1. Finite mixture model

In previous studies, TC tracks over China have been classified empirically into three major clusters: the westward-moving TCs, the northwestward-moving TCs and the recurving TCs (Zhu *et al.*, 2000). However, such an empirical and shape-based classification fails to describe clearly the exact membership of each TC track. Gaffney *et al.* (2007) designed a CCToolbox of MATLAB that contains the FMM algorithm (download: http://www.datalab.uci.edu/resources/CCT) and applied it to classify TC tracks over the WNP. This algorithm employs mixed polynomial regression models (i.e. curves) to fit the geographical 'shape' of TC tracks and models a TC's longitudinal and latitudinal positions *versus* time (Gaffney *et al.*, 2007). Camargo *et al.* (2007a, 2007b) and Zhang *et al.* (2013d) applied this algorithm recently to classify TC tracks under different conditions and achieved desirable results. (For further details about this clustering method, see the study by Gaffney *et al.* (2007).)

2.2.2. The CART algorithm

The CART is a predictive classification algorithm based on a machine learning method that is commonly used in data mining, and also a statistical approach of nonparametric binary tree that is applicable to the classification of both discrete and continuous variables. This algorithm generates a classification tree if the target is a discrete variable and generates a regression tree if the target is a continuous variable. The classification tree generated by CART was used in this study. In the process of building the classification tree, the attributes of the minimum Gini co-efficient were taken as the testing attributes. A smaller Gini means a smaller sampling heterogeneity, which, in turn, means a better segmentation quality.

Data calculated by CART were arranged in an ascending order and divided into two groups according to the medium value between two adjacent values. The Gini co-efficient was then used to calculate the heterogeneity of output variables of the two sample groups:

$$G(t) = 1 - \sum_{j=1}^{K} p^2(j|t) \tag{1}$$

where t represents the node, K represents the cluster of the output variable and $p(j|t)$ represents the probability that the output variable has value j for node t. When all nodes fall into a single cluster, the heterogeneity of output variables is the smallest and the Gini co-efficient is 0. When the nodes fall equally into all clusters, the heterogeneity is the largest and the Gini co-efficient also reaches its maximum: $1 - 1/k$.

In the CART algorithm, the decrease of the heterogeneity can be described by the decrease of the Gini co-efficient:

$$\Delta G(t) = G(t) - \frac{Nr}{N} G\left(t_{\mathrm{r}}\right) - \frac{Nl}{N} G\left(t_{\mathrm{l}}\right), \tag{2}$$

where $G(t)$, $G(t_{\mathrm{r}})$ and $G(t_{\mathrm{l}})$ represent the Gini co-efficients of the ungrouped output variable, the right subtree and the left subtree, respectively; N, N_{r} and N_{l} stand for their respective sample sizes.

The segmentation point where the heterogeneity decreases the most can be obtained through repeated calculations, the best segmentation point corresponding to the largest $\Delta G(t)$.

The values of climate indices (e.g. ENSO, PDO, QBO etc.) are continuous; each value can be divided into split points, which are the mean of two consecutive values. In the present study, t is the number of samples (62 years) and j stands for the number of years in which the TC frequency is high or not, in the case of different split points for the climate indices.

In this study, the excellent data mining software of SPSS Clementine 12.0 was used to build the CART decision tree model.

3. Cluster analysis and statistics of landfalling TC tracks

In China, TC tracks are classified empirically into three clusters (Zhu *et al.*, 2000). This approach was also used in the present paper for comparative analysis of TC tracks. The FMM algorithm was used to cluster the TC tracks landfalling in China during 1951–2012 based on the similarity of their shapes, lengths and locations, as shown in Figure 1.

Figure 1(a) shows the distribution of all TC tracks over China during 1951–2012; the three clusters (Cluster-1, Cluster-2 and Cluster-3) are shown in Figures 1(b), (c) and (d), respectively. The black bold curves in Figures 1(b)–(d) represent the mean curves of the clusters. As can be seen from Figure 1, Cluster-1 tracks have both the longest and the widest influence, covering

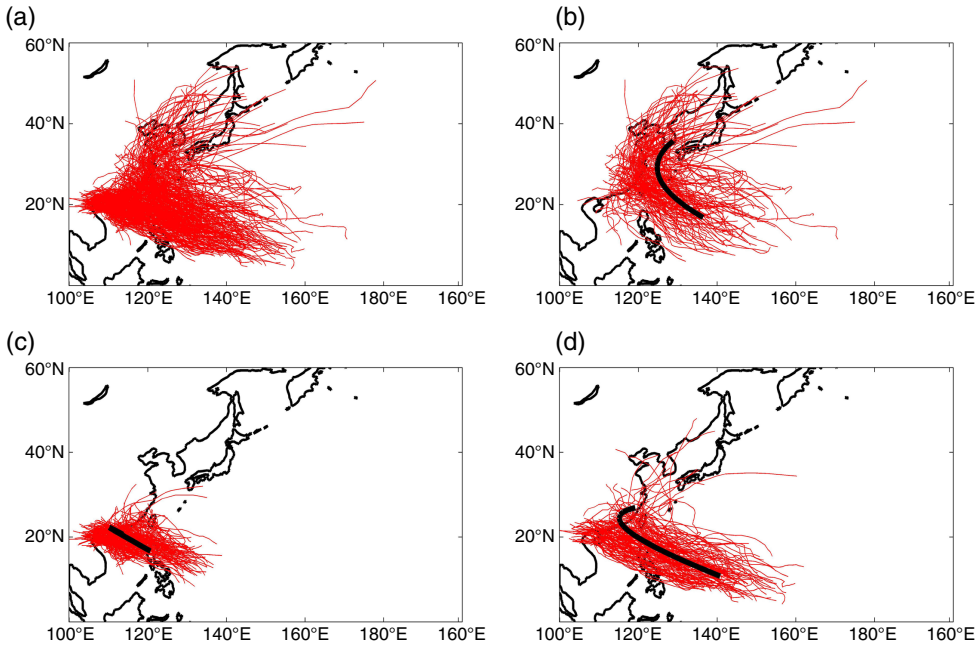

Figure 1. Classification of TC tracks in 1951–2012 over China based on the FMM algorithm. (a) The distribution of all TC tracks over China during 1951–2012; (b) Cluster-1, (c) Cluster-2 and (d) Cluster-3 TC tracks. The black bold curves in Figures 1(b)–(d) represent the average paths.

Table 1. Statistics for summer location, intensity, frequency and lifespan for each TC track class.

JJA	Cluster-1	Cluster-2	Cluster-3
Mean initial longitude	128.94	116.00	126.68
Mean initial latitude	24.61	19.27	17.45
Mean intensity ($m\,s^{-1}$)	51.30	38.35	56.42
Mean lifespan (6 hours)	30.09	18.90	33.01
Frequency	149	169	108
Annual frequency	2.40	2.73	1.74
Frequency ratio (%)	35	40	25

the Yangtze River delta, the Pearl River delta and the Beijing–Tianjin–Hebei Region in China and even parts of the Korean Peninsula and Japan.

Statistics of the formation location, average intensity, average lifespan and frequency of the three clusters of the summer TC tracks are presented in Table 1.

As revealed by the statistics, the TC tracks of Cluster-1 form more eastward and northward and they have stronger intensity and longer lifespan than Cluster-2; the frequency of Cluster-1 is higher than that of Cluster-3 but lower than that of Cluster-2, accounting for about 35% of the total TC frequency.

Table 2 lists the analytical correlation between the TC track frequency and several climate indices. The frequency of Cluster-1 tracks is significantly correlated with Niño1 + 2, Niño3, PDO, Area Indexes of Subtropical High (AISH), the Eurasian Meridional Circulation Index (EMCI), the Western Pacific Subtropical High Ridge Line (WPSHRL), the Area Index for the Subtropical High over the West Pacific (AISHWP), the Strength Index of the Subtropical High in the Western Pacific (SISHWP) and the Asian Meridional Circulation Index (AMCI), whereas the frequencies of the other two clusters are not significantly correlated with the climate indices. Thus, the frequency of Cluster-1 tracks can be predicted with greater accuracy than the frequencies of Cluster-2 and -3 tracks using the indices of ENSO and PDO and the other climate signals, providing good foundations for the prediction model.

4. The CART-based prediction model for TC frequency

The CART algorithm was also used to classify the frequencies of TCs in Cluster-1, Cluster-2 and Cluster-3. The TC samples recorded in 46 years out of the 62 years (about 75%) from 1951 to 2012 for the three clusters landfalling over China were selected randomly as the training set for the model; those samples recorded in the remaining 16 years (about 25%) were taken as the test set to verify the effectiveness of the model. The CART algorithm is a machine learning method and has the ability to filter data through the heterogeneity of attributes and output variables. The indices of ENSO, PDO and QBO and some other climate indices (referred to as 74 circulation indices) in the summer were taken as the learning attributes of the model, to determine 'whether the object variable (namely the frequency) is high or not'. If the frequency of Cluster-1 and Cluster-2 TC tracks in summer is $\geq$3, then the answer is 'Yes' (high frequency); if $<$3, then the answer is 'No' (low frequency). If the frequency of Cluster-3 TC tracks in summer is $\geq$2, then the answer is 'Yes' (high frequency); if $<$2, then the answer is 'No' (low frequency). Of the total 62 years, there were 27 years of Cluster-1, 31 years of Cluster-2 and 34 years of Cluster-3 during which the frequencies of TC tracks were high; frequencies during the remaining years were low. After multiple modelling on the random data, the decision tree of which test set had the highest accuracy was taken as the optimal decision tree model.

By using CART, the decision tree of Cluster-1 involved Niño3, the India-Burma Trough (IBT), QBO, PDO and Niño4; the final decision tree was built as shown in Figure 2. The decision tree of Cluster-2 involved the Index of Asia Polar Vortex Intensity (IAPVI), QBO, EMCI and AISH; the final decision tree was built as shown in Figure 3. The decision tree of Cluster-3 involved EMCI, IBT, AISH and Niño1 + 2; the final decision tree was built as shown in Figure 4. Taking the decision tree of Cluster-1 (as

Table 2. The Pearson correlation co-efficients between the number of TCs in each cluster and several climate indices (the maximum in the three clusters is marked in bold).

JJA	Cluster-1	Cluster-2	Cluster-3
Niño 1 + 2	**−0.342****	−0.022	0.166
Niño 3	**−0.273***	0.014	0.199
Niño 4	**−0.147**	0.139	0.139
Niño 3.4	−0.169	0.107	**0.185**
Multivariate ENSO Index (MEI)	**0.175**	0.162	−0.011
Pacific Decadal Oscillation (PDO)	**−0.303***	−0.045	0.088
Quasi-biennial Oscillation (QBO)	−0.08	**−0.175**	0.037
North Atlantic Oscillation (NAO)	**0.088**	−0.023	−0.024
Northern Boundary of the Northern Hemisphere Subtropical High	0.12	**0.21**	−0.147
Area Indexes of Subtropical High (AISH)	**−0.270***	0.084	0.034
Southern Oscillation Index	0.016	0.068	**−0.183**
Northern Boundary of the South China Sea Subtropical High	−0.106	**0.238**	0.021
South China Sea Subtropical Ridge Line	−0.088	**0.219**	0.013
Eurasian Meridional Circulation Index (EMCI)	**−0.258***	−0.117	0.211
Eurasian Zonal Circulation Index	**0.118**	−0.015	−0.012
Pacific Ocean Subtropical High Ridge	**0.145**	0.075	−0.055
Area Index for the Subtropical High over the Pacific	**−0.229**	0.092	0.008
Pacific Polar Vortex Intensity Index	0.088	**0.125**	−0.065
Northern Boundary of the West Pacific Subtropical High Index (110° E–150° E)	**0.199**	0.058	0.021
Western Pacific Subtropical High Ridge Line (WPSHRL) (110° E–150° E)	**0.372****	0.106	−0.03
Area Index for the Subtropical High over the West Pacific (AISHWP) (110° E–180° E)	**−0.252***	0.045	0.072
Strength Index of the Subtropical High in the Western Pacific (SISHWP) (110° E–180° E)	**−0.27***	0.039	0.063
Asian Meridional Circulation Index (AMCI)	**−0.258***	−0.139	0.186
Index of Asia Polar Vortex Intensity (IAPVI)	**0.144**	0.114	−0.075
Index of Asian Zonal Circulation	**0.145**	−0.024	−0.106
India–Burma Trough (IBT)	**−0.285***	0.070	−0.004

Level of significance: *$P = 0.05$; **$P = 0.01$.

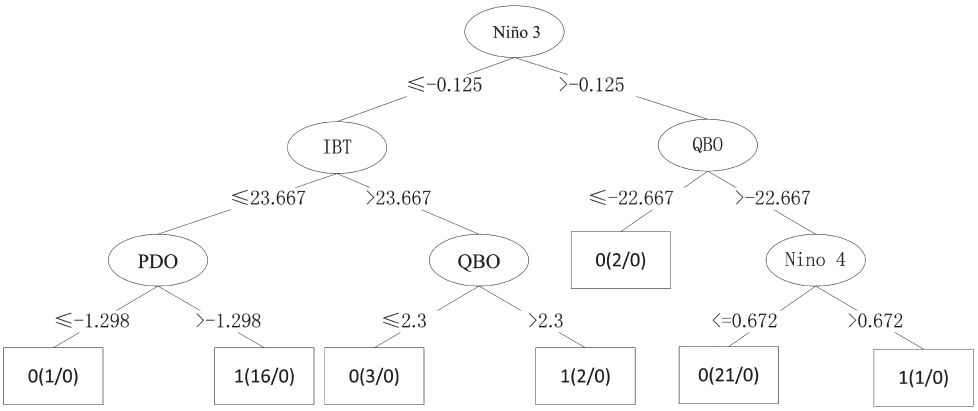

Figure 2. Decision tree model of Cluster-1 track frequency prediction based on the CART algorithm.

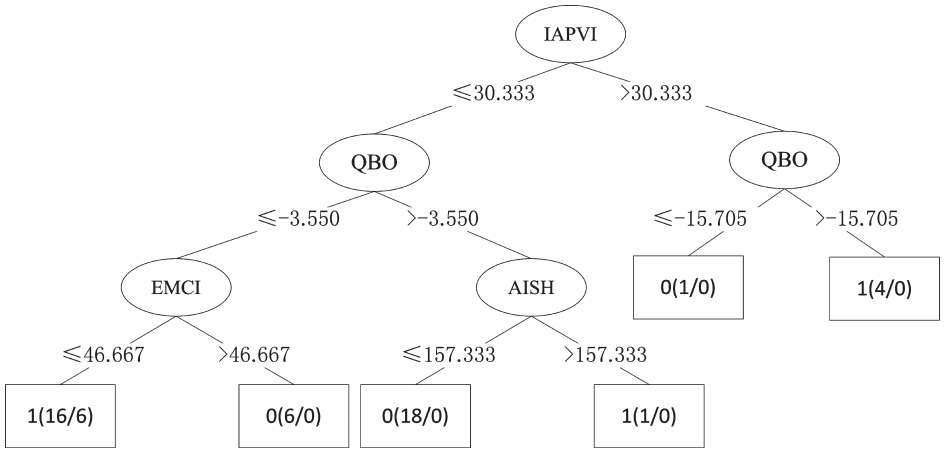

Figure 3. Decision tree model of Cluster-2 track frequency prediction based on the CART algorithm.

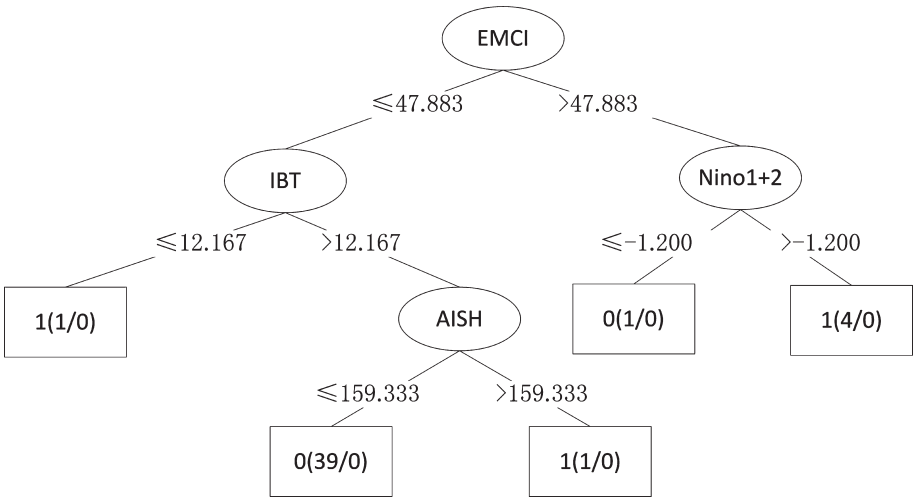

Figure 4. Decision tree model of Cluster-3 track frequency prediction based on the CART algorithm.

Table 3. Rules on predicting the frequency of Cluster-1 TC tracks discovered by the CART algorithm.

Rules	Decision attributes	Learning accuracy
If (Niño3 ≤ −0.125 and IBT ≤ 23.667 and PDO ≤ −1.298), then the frequency is not high	Niño3, IBT, PDO	1/1 = 100%
If (Niño3 ≤ −0.125 and IBT ≤ 23.667 and PDO > −1.298), then the frequency is high	Niño3, IBT, PDO	16/16 = 100%
If (Niño3 ≤ −0.125 and IBT > 23.667 and QBO ≤ 2.3), then the frequency is not high	Niño3, IBT, QBO	3/3 = 100%
If (Niño3 ≤ −0.125 and IBT > 23.667 and QBO > 2.3), then the frequency is high	Niño3, IBT, QBO	2/2 = 100%
If (Niño3 > −0.125 and QBO ≤ −22.667), then the frequency is high	Niño3, QBO	2/2 = 100%
If (Niño3 > −0.125 and QBO > −22.667 and Niño4 ≤ 0.672), then the frequency is not high	Niño3, QBO, Niño4	21/21 = 100%
If (Niño3 > −0.125 and QBO > −22.667 and Niño4 > 0.672), then the frequency is high	Niño3, QBO, Niño4	1/1 = 100%

shown in Figure 2) as an example, each route starting from the root node to the leaf node represents a rule predicting whether the frequency of the Cluster-1 TC tracks in summer is high or not. Take the leaf node '1(16/0)', for example, the number 1 before the parentheses means the frequency of Cluster-1 summer TC track is high; numbers 16 and 0 mean that the node contains 16 samples of which there are $16-0=16$ samples in which the frequency has been classified correctly and no (0) samples with the wrong classification. The accuracy of the self-learning by the model was 100%. The test set was substituted into the decision tree for verification and the classification result showed that the prediction accuracy was 87.5%. In general, a strategy may be used to prune the branch(es) of the decision tree in order to prevent the model from overfitting. However, this pruning strategy was not used in the current study because, with a small sample size, it had little impact on the experiment. Comparison results show that the accuracy of the model reached its highest level without the application of this pruning strategy. The construction method of the CART model for Cluster-2 and Cluster-3 was identical to that for Cluster-1. Figures 3 and 4 are the decision trees of Cluster-2 and Cluster-3, respectively. The accuracy of the self-learning by the CART model was 86.96% (Cluster-2) and 100% (Cluster-3) and the testing accuracy was 62.5% (Cluster-2) and 68.75% (Cluster-3).

With reference to Figure 2, a set of seven rules for predicting whether the frequency of the summer Cluster-1 TC tracks over China is high or not was extracted and is shown in Table 3: this was the qualitative prediction model. In the first column, each rule in the qualitative prediction model is described by the term 'If-then'. The second column describes the attribute variables for making the rules. The learning accuracy of each rule is listed in the third column. The accuracy is calculated as the ratio of the correctly classified samples to the total samples in the leaf node. In a similar way, Tables 4 and 5 contain the rule sets generated according to Figures 3 and 4.

The above classification results indicated that the nonparametric statistical approach CART can produce satisfactory accuracy in predicting Cluster-1 TC tracks in summer and a simple, scientific and easy rule set for prediction can be obtained. Using CART to study TC frequency also provides a novel framework for investigating the nonlinear statistics of TC track frequency.

5. Conclusions and discussion

Cluster analysis has been used widely to unravel the different types of tropical cyclone (TC) tracks. However, less attention has been paid to the cluster analysis of landfalling TCs, especially over the Chinese coast. This study attempted to classify historical landfalling TC tracks over the Chinese coast using Finite Mixture Model (FMM) based cluster analysis and to apply the Classification and Regression Tree (CART) to build a prediction scheme for landfalling TC frequency. Both the FMM and CART performed encouragingly in extracting useful knowledge from the historical TC archive. The research findings of this study can be summarized as follows.

1. TC tracks over China during 1951–2012 were classified into three clusters using the FMM algorithm. It was found that the tracks of the Cluster-1 TCs formed more eastwards and northwards in the study area and had stronger intensity and longer lifespans. They accounted for a larger proportion of the total frequency and exerted influence on many developed regions.

Table 4. Rules on predicting the frequency of Cluster-2 TC tracks discovered by the CART algorithm.

Rules	Decision attributes	Learning accuracy
If (IAPVI ≤ 30.333 and QBO ≤ −3.550 and EMCI ≤ 46.667), then the frequency is high	IAPVI, QBO, EMCI	12/16 = 75%
If (IAPVI ≤ 30.333 and QBO ≤ −3.550 and EMCI > 46.667), then the frequency is not high	IAPVI, QBO, EMCI	6/6 = 100%
If (IAPVI ≤ 30.333 and QBO > −3.550 and AISH ≤ 157.333), then the frequency is not high	IAPVI, QBO, AISH	18/18 = 100%
If (IAPVI ≤ 30.333 and QBO > −3.550 and AISH > 157.333), then the frequency is high	IAPVI, QBO, AISH	1/1 = 100%
If (IAPVI > 30.333 and QBO ≤ −15.705), then the frequency is not high	IAPVI, QBO	1/1 = 100%
If (IAPVI > 30.333 and QBO > −15.705), then the frequency is high	IAPVI, QBO	4/4 = 100%

Table 5. Rules on predicting the frequency of Cluster-3 TC tracks discovered by the CART algorithm.

Rules	Decision attributes	Learning accuracy
If (EMCI ≤ 47.883 and IBT ≤ 12.167), then the frequency is high	EMCI, IBT	1/1 = 100%
If (EMCI ≤ 47.883 and IBT > 12.167 and AISH ≤ 159.333), then the frequency is not high	EMCI, IBT, AISH	39/39 = 100%
If (EMCI ≤ 47.883 and IBT > 12.167 and AISH > 159.333), then the frequency is high	EMCI, IBT, AISH	1/1 = 100%
If (EMCI > 47.883 and Niño1 + 2 ≤ −1.200), then the frequency is not high	EMCI, Niño1 + 2	1/1 = 100%
If (EMCI > 47.883 and Niño1 + 2 > −1.200), then the frequency is high	EMCI, Niño1 + 2	4/4 = 100%

2. The frequency of the Cluster-1 TCs was significantly correlated with the El Niño-Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO) and several other climate indices which form the basis of building the seasonal prediction model.
3. The CART algorithm was used to derive the Cluster-1 decision tree containing seven leaf nodes, in which ENSO, PDO, the quasi-biennial Oscillation (QBO) and several other climate indices were taken as predictors. Meanwhile, seven rules were established for predicting tracks. The self-learning accuracy of the model was 100% and the prediction accuracy was 87.5%, showing the reliability of the model.

In order to formulate a more accurate prediction model for the frequencies of Cluster-2 and Cluster-3 TC tracks, it will be necessary to find a more suitable dataset for the prediction: the data and the output attribute should contain higher heterogeneity.

The TC data volume is increasing with constantly improved observational methods and skills. Thus, the technology of data mining, which is a cutting-edge tool in the age of large amounts of data, will play an increasingly important role both in improving our understanding of TC landfall variability and in shedding light on the prediction and analysis of TC landfall over China. This study has provided a practical point of view from which to analyse TC landfall data and predict TC landfall over China. Dynamic models will be integrated to complement data mining methods in TC-related studies in the future.

Acknowledgements

This work was sponsored jointly by the National Natural Science Foundation of China (grant nos 41201045, 41430427) and the Natural Science Foundation of Jiangsu Province (grant no. BK20151458).

References

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*. Wadsworth and Brooks: Belmont, CA.

Camargo SJ, Robertson AW, Gaffney SJ, Smyth P, Ghil M. 2007a. Cluster analysis of typhoon tracks. Part I: General properties. *J. Clim.* **20**(14): 3635–3653.

Camargo SJ, Robertson AW, Gaffney SJ, Smyth P, Ghil M. 2007b. Cluster analysis of typhoon tracks. Part II: Large-scale circulation and ENSO. *J. Clim.* **20**(14): 3654–3676.

Chan JCL. 1995. Tropical cyclone activity in the Western North Pacific in relation to the stratospheric quasi-biennial oscillation. *Mon. Weather Rev.* **123**(8): 2567–2571.

Gaffney SJ, Robertson AW, Smyth P, Camargo SJ, Ghil M. 2007. Probabilistic clustering of extratropical cyclones using regression mixture models. *Clim. Dyn.* **29**(4): 423–440.

Goh AZ, Chan JCL. 2010. An improved statistical scheme for the prediction of tropical cyclones making landfall in South China. *Weather Forecasting* **25**(2): 587–593.

Gray WM. 1984a. Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences. *Mon. Weather Rev.* **112**(9): 1649–1668.

Gray WM. 1984b. Atlantic seasonal hurricane frequency. Part II: Forecasting its variability. *Mon. Weather Rev.* **112**(9): 1669–1683.

Han J, Kamber M. 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann: San Francisco, CA.

He P, Jing J. 2011. Effect of PDO on the relationships between large scale circulation and tropical cyclone activity over the Western North Pacific. *J. Meteorol. Sci.* **31**(3): 266–273.

Ho C, Baik J, Kim J, Gong D, Sui D. 2004. Interdecadal changes in summertime typhoon tracks. *J. Clim.* **17**(9): 1767–1776.

Ho C, Kim J, Jeong J, Son S. 2009. Influence of stratospheric quasi-biennial oscillation on tropical cyclone tracks in the western North Pacific. *Geophys. Res. Lett.* **36**(6): 141–153.

Li J, Li F, Lin R, Zhang H. 2011. Interdecadal variations of midsummer Northward-Going Typhoons over East Asia and the relationships with Pacific decadal oscillation. *J. Trop. Meteorol.* **27**(5): 731–737.

Liu KS, Chan JCL. 2008. Interdecadal variability of Western North Pacific tropical cyclone tracks. *J. Clim.* **21**(17): 4464–4476.

Shi D, Geng H, Ji C, Huang C. 2015. Construction and application of road icing forecast model based on C4.5 decision tree algorithm. *J. Meteorol. Sci.* **35**(2): 204–209.

Tao L, Jin T, Pu M-J, Xia Y. 2013. Review of the researches on climatological variations of tropical cyclones over Western North Pacific. *Trans. Atmos. Sci.* **36**(4): 504–512.

Tao L, Wu LG, Wang YQ. 2012. Influence of tropical Indian Ocean warming and ENSO on tropical cyclone activity over the western North Pacific. *J. Meteorol. Soc. Jpn.* **90**(1): 127–144.

Wang B, Chan JCL. 2002. How strong ENSO events affect tropical storm activity over the western North Pacific. *J. Clim.* **15**(13): 1643–1658.

Zhan R, Wang Y, Lei X. 2011. Contributions of ENSO and East Indian Ocean SSTA to the interannual variability of Northwest Pacific tropical cyclone frequency. *J. Clim.* **24**(2): 509–521.

Zhang W, Gao S, Chen B, Cao K. 2013a. The application of decision tree to intensity change classification of tropical cyclones in western North Pacific. *Geophys. Res. Lett.* **40**(9): 1883–1887.

Zhang W, Leung Y, Chan JCL. 2013b. The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part I:

Tropical cyclone recurvature. *J. Appl. Meteorol. Climatol.* **52**: 1394–1416.

Zhang W, Leung Y, Chan JCL. 2013c. The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part II: Tropical cyclone landfall. *J. Appl. Meteorol. Climatol.* **52**: 1417–1432.

Zhang W, Leung Y, Wang Y. 2013d. Cluster analysis of post-landfall tracks of landfalling tropical cyclones over China. *Clim. Dyn.* **40**(5–6): 1237–1255.

Zhang M, Qiu H, Fang X, Lu N. 2015. Study on the multivariate statistical estimation of tropical cyclone intensity using FY-3 MWRI brightness temperature data. *J. Trop. Meteorol.* **31**(1): 87–94 (in Chinese).

Zheng Y, Yu J, Wu Q, Lin J, Gong Z. 2013. K-means clustering method for classification of the Northwestern Pacific Tropical cyclone tracks. *J. Trop. Meteorol.* **29**(4): 607–615.

Zhu Q, Lin J, Shou S, Tang D. 2000. *Meteorology Principles and Methods*, 4th edn. China Meteorological Press: Beijing; 521–523.

Zhu Z, Yin Y, Ye D. 2012. Analysis on the character of tropical cyclones making landfall on different regions of Chinese continent. *J. Trop. Meteorol.* **28**(1): 41–49 (in Chinese).

