Hindawi Publishing Corporation
Advances in Meteorology
Volume 2014, Article ID 735491, 13 pages
http://dx.doi.org/10.1155/2014/735491

Research Article

Prediction of Tropical Cyclones' Characteristic Factors on Hainan Island Using Data Mining Technology

Ruixu Zhou,[1] Wensheng Gao,[1] Bowen Zhang,[2] Xianggan Fu,[3] Qinzhu Chen,[3] Song Huang,[3] and Yafeng Liang[3]

[1] *Department of Electrical Engineering, Tsinghua University, Beijing 100084, China*
[2] *China Electric Power Research Institute, Beijing 100192, China*
[3] *Hainan Power Grid Corporation, Haikou, Hainan 570203, China*

Correspondence should be addressed to Ruixu Zhou; zrx13@mails.tsinghua.edu.cn

Received 18 August 2014; Revised 20 October 2014; Accepted 28 October 2014; Published 20 November 2014

Academic Editor: Luis Gimeno

A new methodology combining data mining technology with statistical methods is proposed for the prediction of tropical cyclones' characteristic factors which contain latitude, longitude, the lowest center pressure, and wind speed. In the proposed method, the best track datasets in the years 1949~2012 are used for prediction. Using the method, effective criterions are formed to judge whether tropical cyclones land on Hainan Island or not. The highest probability of accurate judgment can reach above 79%. With regard to TCs which are judged to land on Hainan Island, related prediction equations are established to effectively predict their characteristic factors. Results show that the average distance error is improved compared with the National Meteorological Centre of China.

1. Introduction

Typhoon is a kind of tropical cyclones (TCs), the center-sustained wind speed of which arrives at level 12 to level 13 (typhoon is not distinguished from TC in this paper unless specially emphasized). Hainan Island (108°37′E~111°05′E, 18°10′N~20°10′N) in China is well known as "typhoon corridor." According to the historical data analysis of TCs landing on Hainan Island, the yearly and the monthly statistical results are shown in Figures 1 and 2, respectively. (Note: in this paper the condition to determine whether a typhoon lands or not is that the minimum distance between the typhoon center and Hainan Island is no more than the preset influencing radius, which is 300 km herein). Thus the frequency of TCs landing on Hainan Island is very high. Besides, typhoon ranks at the top among all kinds of disasters on Hainan Island. Taking the typhoon "Damrey" as an example, in 2005, it destroyed 18 cities of Hainan and affected up to 6.305 million people among whom 21 persons were killed. The direct economic loss reached 12.1 billion RMB [1]. Therefore, the timely and accurate forecast of TCs is very important for disaster prevention on Hainan Island. It can also effectively reduce the damage and loss caused by the TC when it happens.

The main methods for traditional TC forecast contain statistical methods and dynamic methods, most of which are along with complicated processes or lower precision. The statistical methods use the historical TCs' positions, intensity, and so on to predict TC's characteristic factors, such as fuzzy multicriteria decision support model [2], conditional nonlinear optimal perturbation, first singular vector, ensemble transform Kalman filter [3], back propagation-neural network [4], adaptive neural network classifier using a two-layer feature selector [5], and a support vector machine using data reduction methods [6]. Dynamic methods are mainly based on numerical forecast, such as a simplified dynamical system based on a logistic growth equation (LGE) [7], a regional coupled atmosphere-ocean model [8], the PSU-NCAR Mesoscale Model version 5 [9], and the GFDL 25-km-Resolution Global Atmospheric Model [10]. Taking three main prediction centers, for example, the average distance error of 24/48 hours' forecast by the USA National Hurricane Center (NHC) is 106/187 km, which is 125/243 km for Japan Meteorological

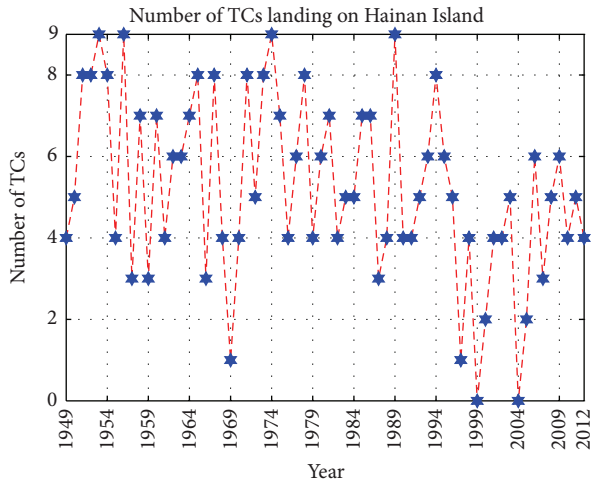


Figure 1: The yearly statistical result.

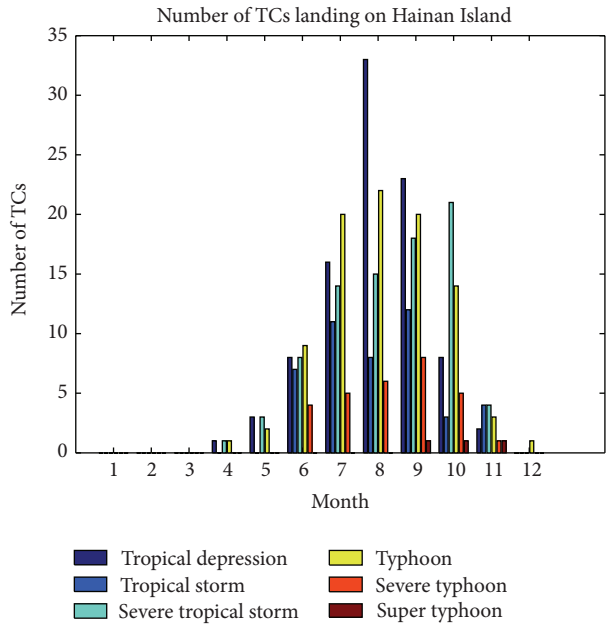


Figure 2: The monthly statistical result.

Agency (JMA) and 120/215 km for National Meteorological Centre of China (NMCC) [11]. Zhang et al. compared the monitoring data from HY-2 and QuikSCAT's satellite scatterometers with the actual typhoon data from ground observation. The result shows that the deviations of typhoon path and intensity are large and their standard deviations are also very big [12]. Therefore, although there are many typhoon forecast methods in use at present, their precision still cannot meet the need for real-time typhoon warning.

By using data mining technology in combination with statistical methods, a new TC forecast method based on the historical data is proposed in this paper. Firstly, the region where typhoon centers were located 48 (or 72 as a comparative experiment) hours before landing on Hainan Island is divided into five (or a number of $\{1, 3, 7\}$ as a comparative experiment) areas using K-means clustering algorithm. Then the TC landing criterion of each area is formed by classification and regression trees (CART). Further, prediction sum of squares (PRESS) algorithm and its progressive optimal algorithm are applied to optimize forecast factor sets. Finally, part of the historical data is used to establish prediction equations by multiple linear regression model (MLRM) and the accuracy of these equations is examined by the remaining historical data. All results show that this methodology is more accurate compared with present existing forecast methods.

2. Data and Methodology

2.1. Data. Data used in this research is based on TCs' best track datasets of the years 1949~2012 in the northwestern Pacific waters (including the South China Sea, northern of the equator, and western of 180°E) [13], which are derived from the TC information center of China Meteorological Administration (CMA) (http://tcdata.typhoon.gov.cn/). CMA best track datasets contain 2172 TCs, which in total have 62663 observation points. Every observation point may provide information as follows: the observation time, strength grade, latitude, longitude, the lowest center pressure (hereinafter referred to as air pressure), 2-minute-average-near-center-maximum wind speed (hereinafter referred to as wind speed), and average wind speed in 2 minutes. Because the average wind speed in 2 minutes of most observation points cannot be obtained, strength grade (SG), latitude (LAT), longitude (LON), air pressure (AR), wind speed (WS), latitude migration velocity (LATMV), and longitude migration velocity (LONMV) are selected as seven predictors (hereinafter referred to as observation point information). Current LATMV and LONMV can be calculated using the following method.

Set the moments of current observation point and previous two observation points as (E_t, N_t), (E_{t-1}, N_{t-1}), and (E_{t-2}, N_{t-2}), respectively (where E is on behalf of longitude and N is on behalf of latitude). Then the LONMV and LATMV of the current observation time are LONMV_t and LATMV_t, which are calculated as

$$\begin{aligned}\mathrm{LONMV}_t = \frac{1}{2}R \cdot \Big\{ & \cos^{-1}\left[\cos^2 N_t \cdot \cos\left(E_t - E_{t-1}\right) + \sin^2 N_t\right] \\ & \cdot \operatorname{sgn}\left(E_t - E_{t-1}\right) \\ & + \cos^{-1}\left[\cos^2 N_{t-1} \cdot \cos\left(E_{t-1} - E_{t-2}\right) \right. \\ & \left. + \sin^2 N_{t-1}\right] \\ & \cdot \operatorname{sgn}\left(E_{t-1} - E_{t-2}\right)\Big\} \times 6^{-1},\end{aligned} \tag{1}$$

$$\begin{aligned}\mathrm{LATMV}_t &= \frac{1}{2}R \cdot \frac{\left\{\left(N_t - N_{t-1}\right) + \left(N_{t-1} - N_{t-2}\right)\right\}}{6} \\ &= \frac{R \cdot \left(N_t - N_{t-2}\right)}{12},\end{aligned} \tag{2}$$

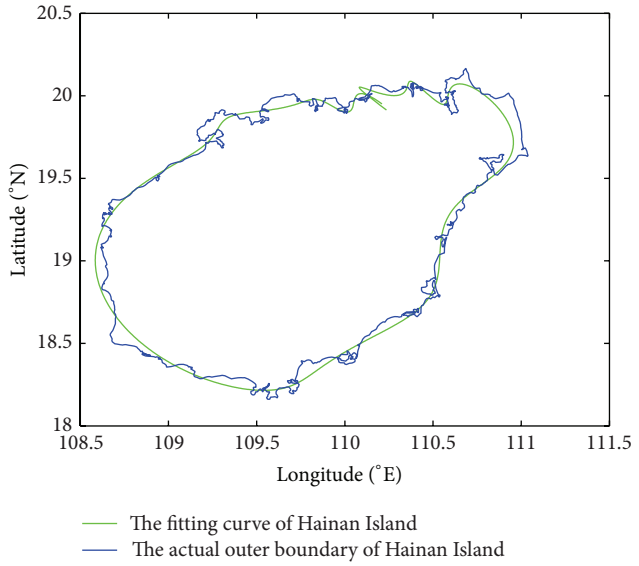


Figure 3: The curve fitting of the outer boundary of Hainan Island.

where R is the mean radius of the Earth with the value of 6370.856 km; $\mathrm{sgn}(x)$ represents the sign function; the unit of LONMV_t and LATMV_t is km/h.

2.2. Methodology. As mentioned in the Introduction, a landing TC is defined as a TC that the minimum distance between the typhoon center and Hainan Island is no more than the preset influencing radius. Hence, in order to distinguish TCs between landing and not landing on Hainan Island, the minimum distance between each TC's track and the outer boundary of Hainan Island needs to be calculated according to CMA best track datasets. Due to the variety of TCs' tracks, applying general curve-fitting directly does not provide a good result. Therefore, polynomial fitting [14] is applied in this paper where an intermediate variable is introduced to conduct curve-fitting with the latitude and longitude, respectively. Taking an arbitrary TC, for example, the specific fitting effects are shown in Figures 3 and 4. Using the fitting polynomials of each TC's and the outer boundary of Hainan Island's latitude and longitude with respect to the corresponding intermediate variables, the distance between any point of each TC's track and any point on the outer boundary of Hainan Island can be calculated, from which the minimum distance can be selected. The great circle distance (GCD) between any two points on the Earth can be calculated using formula (3). The GCD is the shortest distance between any two points on the Earth. Set any two points on the Earth as $d_1(E_1, N_1)$ and $d_2(E_2, N_2)$ and the GCD is

$$\left|d_1 d_2\right| = R \cdot \arccos\left[\cos N_1 \cdot \cos N_2 \cdot \cos\left(E_1 - E_2\right) + \sin N_1 \cdot \sin N_2\right]. \quad (3)$$

The time intervals used to forecast TCs by three main prediction centers (NHC, JMA, and NMCC) are 24, 48, and 72 hours. In order to forecast TCs in a timely manner and compare forecast accuracy among different methods, here

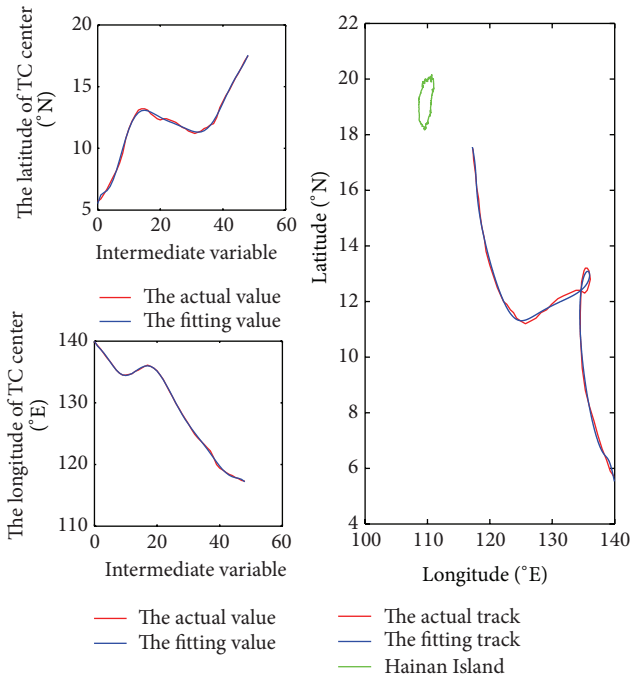


Figure 4: The curve fitting of an arbitrary TC.

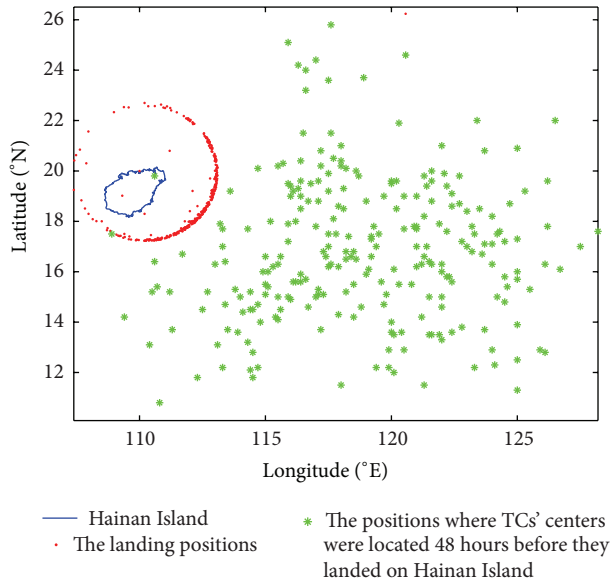


Figure 5: The region selected as research object.

the region where TCs' centers were located 48 hours before they landed on Hainan Island (shown in Figure 5) is selected as research object. In order to narrow the research scope, K-means clustering algorithm [15, 16] is applied to divide the region where TCs' centers were located 48 hours before they landed on Hainan Island into five areas. In this section the situation, in which the region where TCs' centers were located 48 hours before they landed is selected as research object and the research object is divided into five areas, is taken as

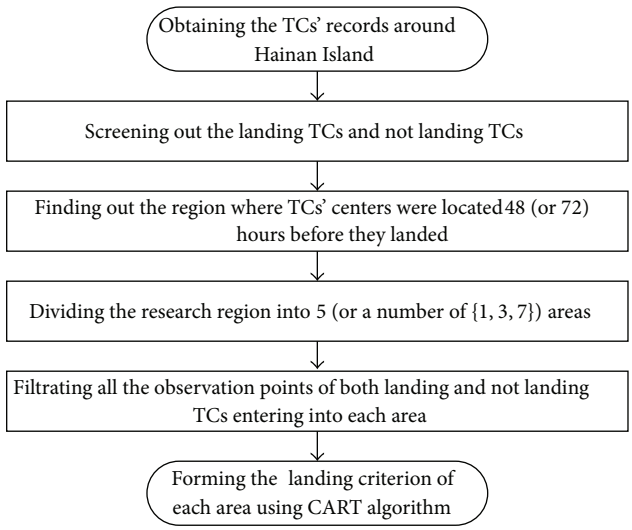


Figure 6: Flow diagram of forming landing criterions.

an example for a convenient statement. Other situations of a comparative experiment are also conducted in Section 3.3.

For each of the five areas, all the observation points of both landing and not landing TCs which entered into this area are filtrated. With strength grade, latitude, longitude, air pressure, wind speed, latitude migration velocity, and longitude migration velocity as classification properties, the TC landing criterion of each area is formed by using CART algorithm [17, 18]. The flow diagram of forming landing criterions is shown in Figure 6.

For TCs which are judged to be landing on Hainan Island, PRESS and its progressive optimal algorithm and MLRM can be used to forecast TCs' characteristic factors (including latitude, longitude, the lowest center pressure, and wind speed). Forecasts in this paper contain landing prediction pattern and dynamic prediction pattern. Landing prediction pattern is defined as employing the observation point information of those points which first enter into any area to predict the characteristic factors when TC lands. Dynamic prediction pattern is defined as 24 hours' and 48 hours' prediction with respect to the observation point which enters into any area. The flow diagrams of landing forecast pattern and dynamic forecast pattern are shown in Figures 7 and 8, respectively. Here PRESS [19] and its progressive optimal algorithm [20, 21] are used to select the best forecast factor set from seven predictors which will be used to forecast corresponding characteristic factor. MLRM [22] is used to establish corresponding forecast equations. MLRM is expressed as [23]

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i \quad (i = 1, 2, \cdots, n), \tag{4}$$

where y_i is the estimated value, $\beta_0 \sim \beta_m$ is the regression coefficients, ε_i is the random error, and $x_{i1} \sim x_{im}$ are the forecast factors of the observation point.

Table 1: The geometric centers and scopes of five areas.

Area number	Longitude Unit: °E	Latitude Unit: °N	Radius Unit: degree
1	118.4511	15.2845	2.7902
2	113.1435	14.9623	3.1706
3	116.9517	20.8023	3.3048
4	123.9859	14.4951	2.7098
5	122.6830	18.5854	2.8346

Table 2: The number of OPs in five areas.

Area number	Number of landing (Z_1)	Number of not landing (Z_2)	The ratio of landing $Z_1/(Z_1 + Z_2)$
1	537	1031	34.25%
2	855	1389	38.10%
3	788	1514	34.23%
4	378	1070	26.10%
5	340	1419	19.33%

3. Results and Discussions

In this section, the situation, in which the region where TCs' centers were located 48 hours before they landed is selected as research object and the research object is divided into five areas, is firstly researched. Other situations of a comparative experiment, in which the research object may be the region where TCs' centers were located 72 hours before they landed and the number of areas of divided research object may be any number of {1, 3, 5, 7}, are also discussed at the end of the section.

3.1. Dividing the Research Region into Five Areas. K-means clustering algorithm is used to divide our research region into five areas as described in Section 2.2, of which the geometric centers and scopes are shown in Table 1. With respect to each area, all the observation points of both landing and not landing TCs which entered into this area are filtrated. The positions of these observation points are shown in Figure 9 and are used to form TCs' landing criterions. The numbers of these observation points (OPs) for both landing and not landing TCs are shown in Table 2. The division of five areas further narrows the research scope and makes the selection of the observation points more pertinent so as to form the effective landing criterions, which will be illustrated further in Section 3.3.

3.2. The Formation of Landing Criterions in Five Areas. According to the CART algorithm, the landing criterions in five areas are shown in Figure 10 (refer to Section 2.1 for the meaning of seven predictors). The corresponding probability of accurate judgment (P_{AJ}), probability of false alarm (P_{FA}), and probability of false dismissal (P_{FD}) are shown

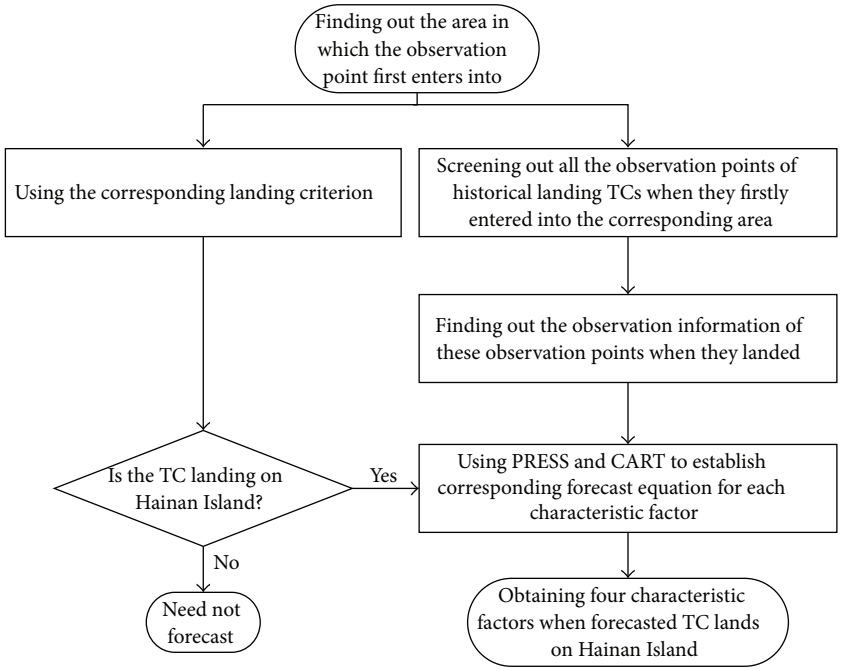


Figure 7: Flow diagram of landing prediction pattern.

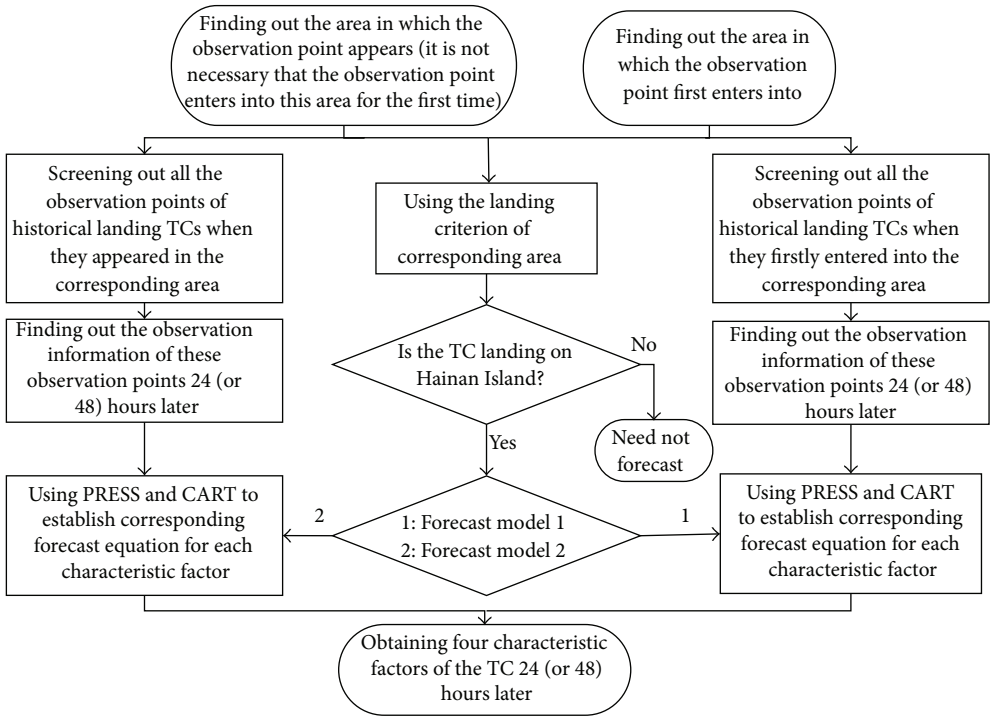


Figure 8: Flow diagram of dynamic prediction pattern.

in Table 3. Set the numbers of OPs for landing and not landing TCs in any area to be Z_1 and Z_2, respectively; the number of OPs which are judged to be landing according to landing criterions when they landed truly is denoted as M_1 and the number of OPs which are judged to be not landing according to landing criterions when they did not land truly is denoted

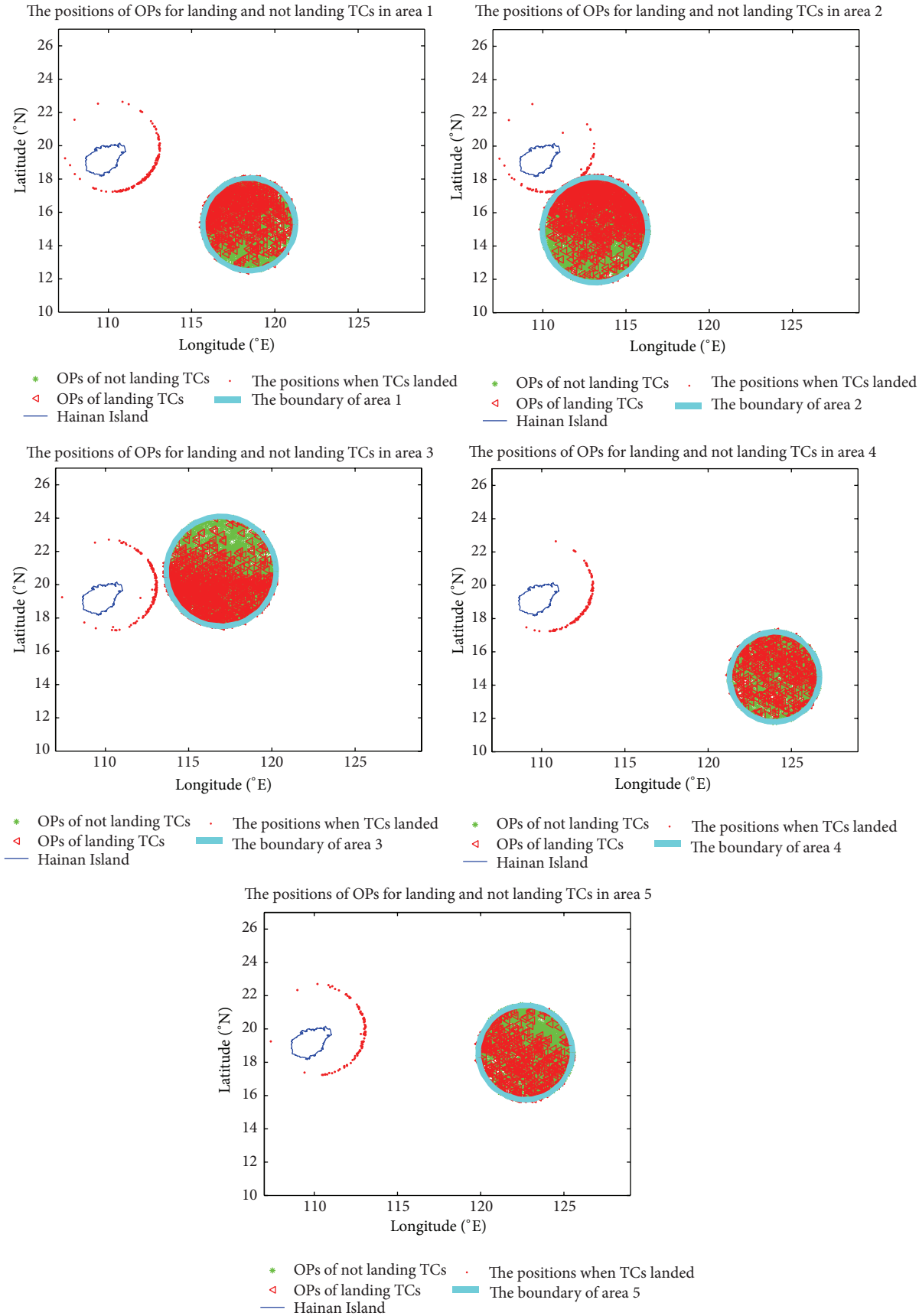


Figure 9: The positions of OPs for landing and not landing TCs which entered into each area.

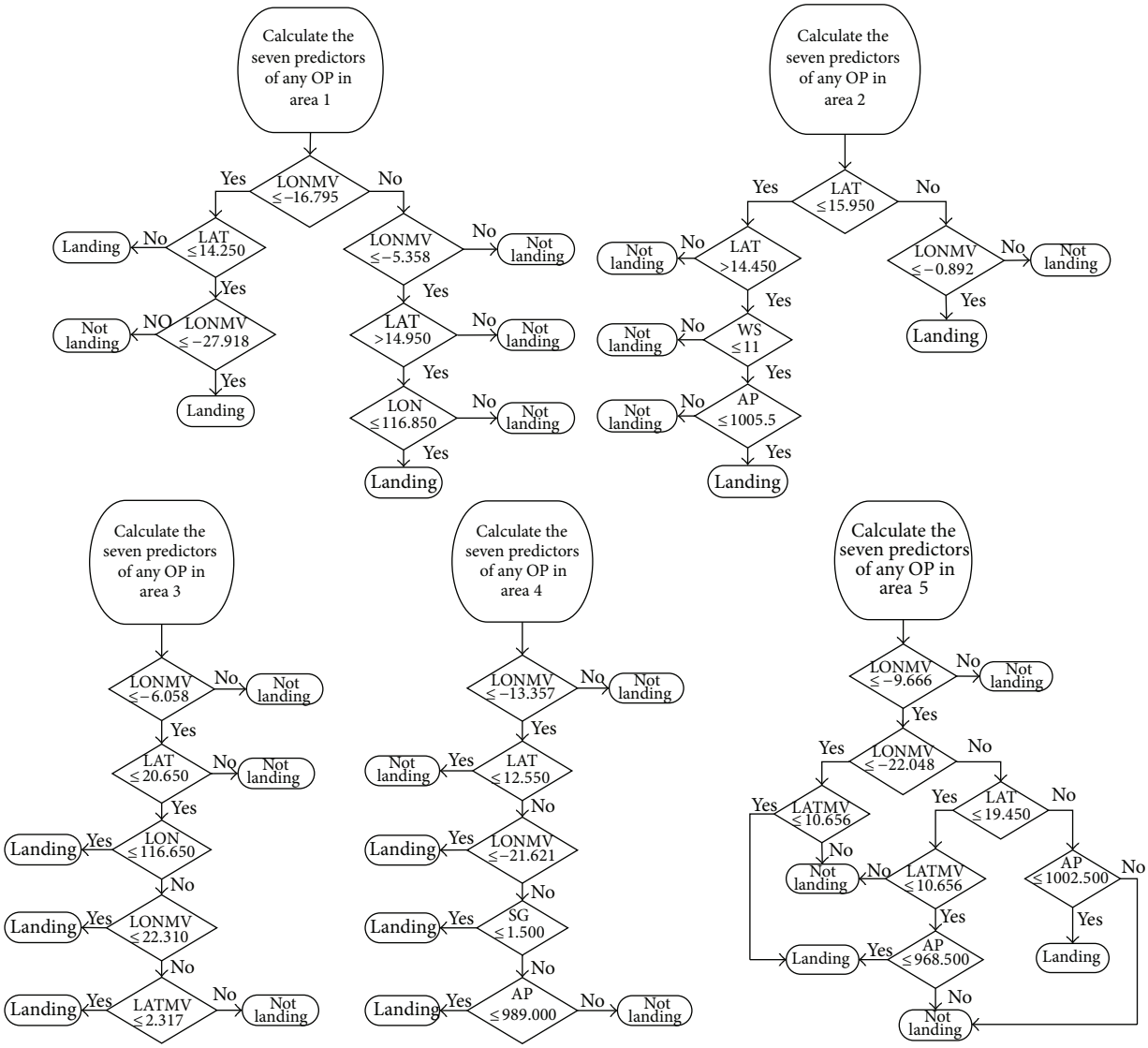


Figure 10: The landing criterions in five areas.

Table 3: P_{AJ}, P_{FA}, and P_{FD} of each criterion.

Area number	P_{AJ}	P_{FA}	P_{FD}
1	74.68%	17.94%	39.48%
2	75.31%	15.55%	39.53%
3	77.50%	11.76%	43.15%
4	66.23%	36.54%	25.93%
5	79.76%	16.49%	35.88%

as M_2. Then P_{AJ}, P_{FA}, and P_{FD} of this area are calculated as follows:

$$P_{AJ} = \frac{M_1 + M_2}{Z_1 + Z_2}, \qquad P_{FA} = \frac{Z_2 - M_2}{Z_2}, \qquad P_{FD} = \frac{Z_1 - M_1}{Z_1}. \tag{5}$$

Table 4: The labels of different situations.

Ti	Nu	Label of corresponding situation
48	1	FE1
48	3	FE3
48	5	FE5
48	7	FE7
72	1	ST1
72	3	ST3
72	5	ST5
72	7	ST7

3.3. Other Situations as a Comparative Experiment. In Sections 3.1 and 3.2, the region where TCs' centers were located 48 hours before they landed is selected as research object and the research object is divided into five areas. In this section other situations as a comparative experiment are researched

Table 5: P_{AJ}, P_{FA}, and P_{FD} for each of the remaining seven situations.

(a) P_{AJ}, P_{FA}, and P_{FD} of each area for FE1

Area number	P_{AJ}	P_{FA}	P_{FD}
1	0.7884	0.0696	0.5279

(b) P_{AJ}, P_{FA}, and P_{FD} of each area for FE3

Area number	P_{AJ}	P_{FA}	P_{FD}
1	0.7454	0.1139	0.5191
2	0.7636	0.1875	0.3054
3	0.7829	0.0446	0.7814

(c) P_{AJ}, P_{FA}, and P_{FD} of each area for FE7

Area number	P_{AJ}	P_{FA}	P_{FD}
1	0.8522	0.0154	0.8414
2	0.7486	0.1991	0.3959
3	0.7302	0.1808	0.4176
4	0.7719	0.1165	0.4398
5	0.7236	0.0740	0.6735
6	0.7834	0.1512	0.2975
7	0.7631	0	1

(d) P_{AJ}, P_{FA}, and P_{FD} of each area for ST1

Area number	P_{AJ}	P_{FA}	P_{FD}
1	0.8164	0.0412	0.6998

(e) P_{AJ}, P_{FA}, and P_{FD} of each area for ST3

Area number	P_{AJ}	P_{FA}	P_{FD}
1	0.8320	0.0677	0.5877
2	0.7396	0.2062	0.3462
3	0.8270	0	1

(f) P_{AJ}, P_{FA}, and P_{FD} of each area for ST5

Area number	P_{AJ}	P_{FA}	P_{FD}
1	0.8284	0	1
2	0.7613	0	1
3	0.7566	0.1590	0.3564
4	0.8596	0	1
5	0.8281	0.0678	0.5813

(g) P_{AJ}, P_{FA}, and P_{FD} of each area for ST7

Area number	P_{AJ}	P_{FA}	P_{FD}
1	0.7874	0	1
2	0.7020	0.1664	0.4723
3	0.8799	0.0118	0.8777
4	0.8358	0	1
5	0.8537	0	1
6	0.7761	0.0955	0.5047
7	0.7407	0.1028	0.6720

and compared with each other. Finally, we select the situation which produces the best result.

In order to distinguish different situations, the labels of them are denoted in Table 4, where the parameter Ti is used to illustrate the research object is the region where TCs' centers are located. Ti hours before they landed on Hainan Island; Nu denotes the number of areas of divided research object.

It can be seen from Table 4 that FE5 is the situation which has been researched in Sections 3.1 and 3.2. The remaining seven situations are researched as follows using the methods which are identical with FE5.

The P_{AJ}, P_{FA}, and P_{FD} of each area for each of the remaining seven situations can be calculated according to formula (5), the results of which are shown in Table 5.

In order to select the best of these eight situations in Table 4, an evaluation method is introduced, with which the Index of each situation is calculated, where Index is defined according to formula (6). For any situation, N denotes the number of areas of divided research object and P_{AJ_i}, P_{FA_i}, and P_{FD_i} denote the P_{AJ}, P_{FA}, and P_{FD} of the ith area, respectively. Consider the following:

$$\text{Index} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left[\frac{1}{10}P_{AJ\_i}^2 + \frac{3}{10}\left(1 - P_{FA\_i}\right)^2 + \frac{3}{5}\left(1 - P_{FD\_i}\right)^2\right]}. \tag{6}$$

It is obvious that the higher the Index is, the better the result of the landing criterions on the whole is. The Index for each of eight situations is shown in Table 6. The situation FE5 shows the best result, which also illustrates that the research scheme in Sections 3.1 and 3.2 is better compared with other situations. Finally, situation FE5 is selected to form the landing criterions.

Table 6: The Index for each of eight situations.

Label of corresponding situation	Index
FE1	0.6750
FE3	0.6659
FE5	0.7026
FE7	0.6625
ST1	0.6297
ST3	0.6572
ST5	0.6391
ST7	0.6231

3.4. Forecast of TCs' Characteristic Factors

3.4.1. Landing Prediction Pattern. The landing forecast pattern is defined as follows: obtaining the observation point information (seven predictors) when landing TCs' centers first enter into any area, which can be used to forecast the characteristic factors (LAT, LON, AP, and WS) when TCs land on Hainan Island. The flow diagram is shown in Figure 6. Taking area 1, for example, the OPs of historical landing TCs' centers when they first entered into area 1 are shown in Figure 11 and the tracks of historical landing TCs which passed through area 1 are shown in Figure 12.

Dividing the historical landing TCs passing through each area into two groups with the same number, one group of TCs is used to establish prediction equations and the other group

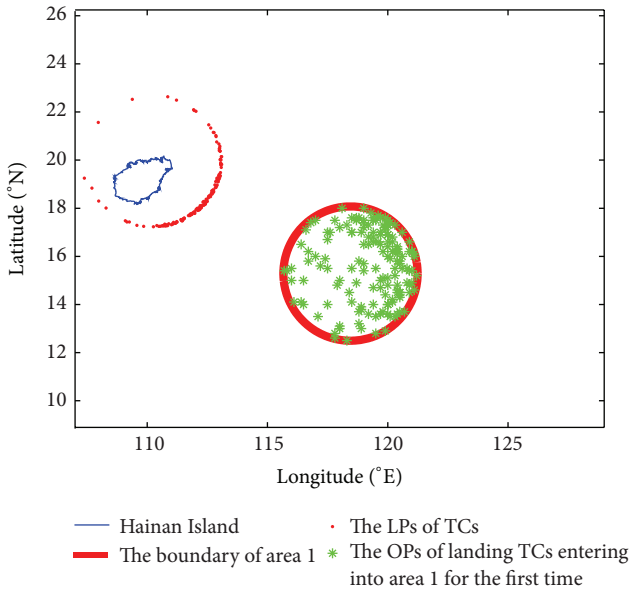


Figure 11: The OPs of historical landing TCs' centers when they firstly entered into area 1 (LP: landing position).

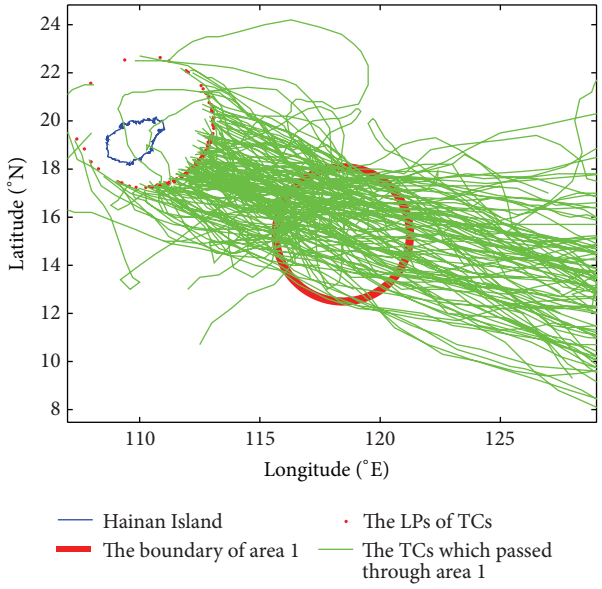


Figure 12: The tracks of historical landing TCs which passed through area 1.

is used to test the accuracy of these equations. The results of testing of these prediction equations for TCs which passed through each area are shown in Table 7. Making use of the actual and predicted longitude and latitude of TCs' centers, in combination with the formula (3), the calculated mean and standard deviation of GCD in the landing prediction pattern are shown in Figures 13 and 14, respectively. Averaging the results of five areas, it can be obtained that the average of the mean/standard deviation (SD) of GCD is 144.6382/97.8740 km. In [24], Yu et al. analyze the average GCD error of 48 hours' forecast in the South China Sea, which is 222.6 km. Therefore, the landing prediction pattern proposed in this

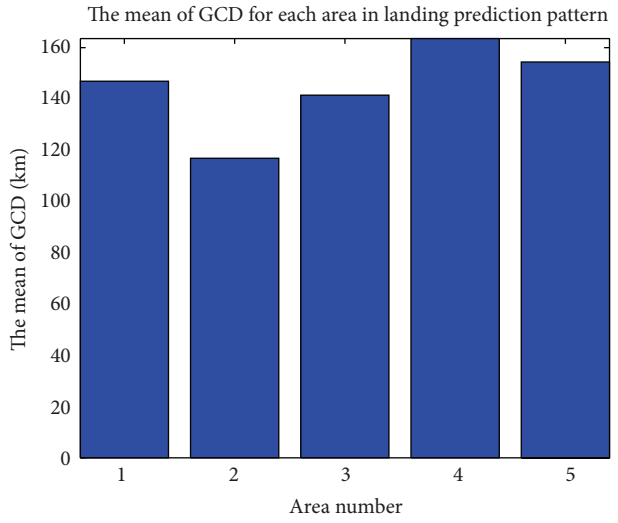


Figure 13: The mean of GCD for each area in landing prediction pattern.

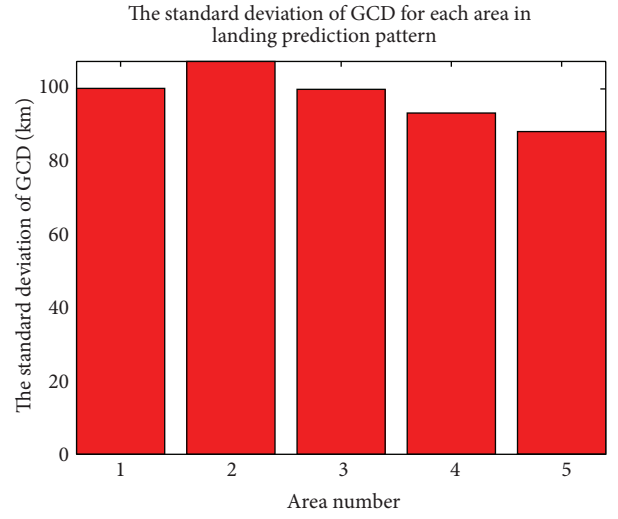
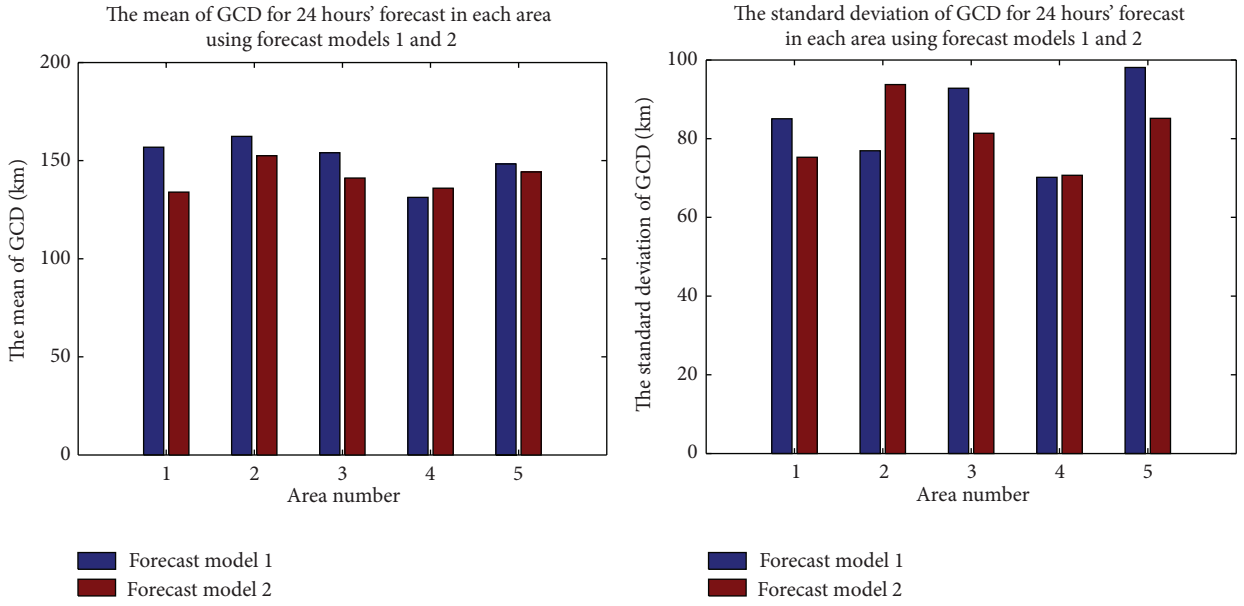


Figure 14: The standard deviation of GCD for each area in landing prediction pattern.

paper shows good prediction accuracy. For TCs which are judged to be landing on Hainan Island, as long as the observation point information when their centers first enter into any area are obtained, the corresponding forecast equations can be used to predict characteristic factors when they land.

3.4.2. Dynamic Prediction Pattern. The dynamic prediction pattern is using the current observation point information to conduct 24 hours' and 48 hours' forecast, which is also illustrated in Figure 8. There are two different forecast models in dynamic prediction pattern that are described as follows.

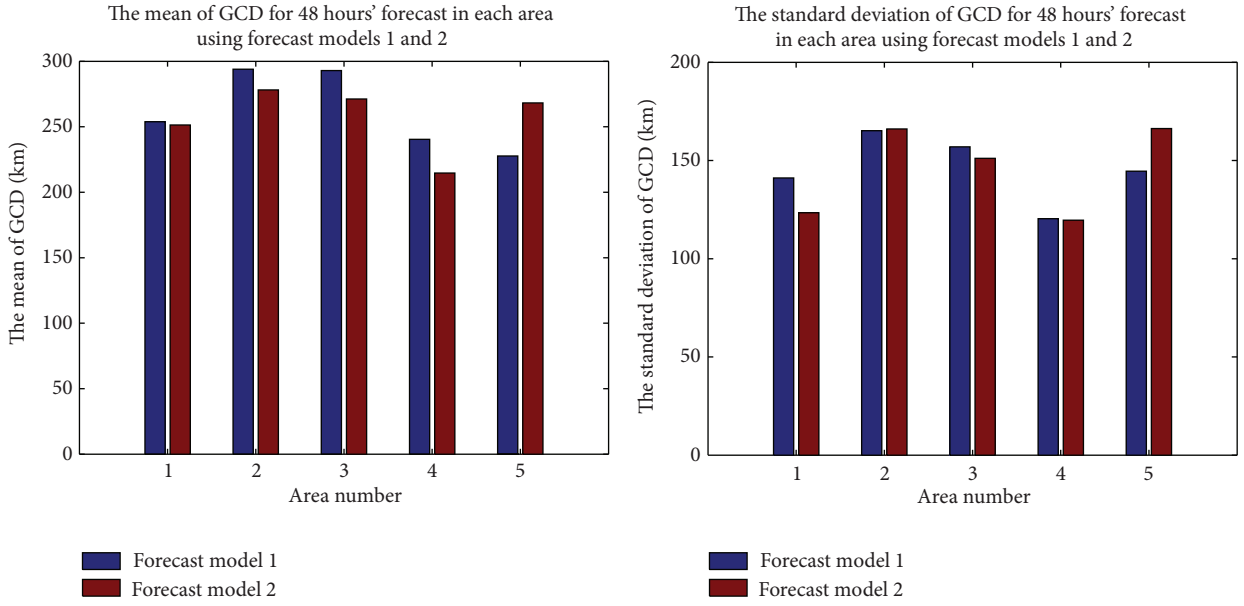
Forecast Model 1. It is to obtain the current observation point information (seven predictors) when landing TCs' centers enter into any area for the first time, making use of which to conduct 24 hours' and 48 hours' forecast.



(a) The mean of GCD for 24 hours' forecast of two forecast models

(b) The standard deviation of GCD for 24 hours' forecast of two forecast models

Figure 15: The mean/standard deviation of GCD for 24 hours' forecast of two forecast models.



(a) The mean of GCD for 48 hours' forecast of two forecast models

(b) The standard deviation of GCD for 48 hours' forecast of two forecast models

Figure 16: The mean/standard deviation of GCD for 48 hours' forecast of two forecast models.

Forecast Model 2. It is to obtain the current observation point information (seven predictors) when landing TCs' centers are in any area (not necessarily enter into any area for the first time), making use of which to conduct 24 hours' and 48 hours' forecast.

In the process of actual prediction, for TCs, which are judged to be landing on Hainan Island, the observation point information when their centers enter into any area for the first time and the established equations in forecast model 1 are used to conduct dynamic prediction. Furthermore, the observation point information when TCs' centers are in area (it is not necessary that TCs' centers enters into this area for the first time) any area and the established equations in forecast model 2 can be used to conduct dynamic prediction.

Similar to Section 3.4.1, the historical observation points that meet the corresponding requirements in corresponding

Table 7: Prediction results in landing prediction pattern.

Area number	The mean/SD of LAT deviation Unit: ˚N	The mean/SD of LAT deviation Unit: ˚E	The mean/SD of AP deviation Unit: hpa	The mean/SD of WS deviation Unit: m/s
1	0.9849/0.8671	0.7264/0.6404	9.2889/6.7882	5.9540/4.3839
2	0.5268/0.6009	0.8677/0.8973	7.9946/6.5339	5.3036/3.8997
3	0.9710/0.7805	0.6340/0.7711	10.6941/8.3758	7.0915/5.2637
4	1.1537/0.8640	0.7150/0.6162	10.3463/6.1761	6.8151/4.7498
5	1.1152/0.6941	0.6728/0.7103	12.1138/8.5744	8.0403/5.3759

Table 8: The results of testing forecast model 1 and forecast model 2.

Forecast hour	Area number	Forecast model (1 or 2)	The mean/SD of LAT deviation Unit: ˚N	The mean/SD of LON deviation Unit: ˚E	The mean/SD of AP deviation Unit: hpa	The mean/SD of WS deviation Unit: m/s
24 h	1	1	0.8128/0.6515	1.0295/0.7600	4.9278/3.9733	3.7787/2.8858
		2	0.7004/0.5153	0.8802/0.7011	5.4114/4.3187	3.9637/2.9843
	2	1	0.9131/0.5832	0.9864/0.7794	6.3412/4.4247	3.8268/2.7857
		2	0.7538/0.6538	1.0050/0.8670	4.9175/4.9673	3.8487/3.0261
	3	1	0.7555/0.5945	1.0528/0.9093	7.8149/6.9316	4.7342/4.1782
		2	0.7727/0.6183	0.8949/0.7255	7.3489/6.6512	5.1718/4.3368
	4	1	0.6582/0.5365	0.8797/0.6251	6.6373/5.5333	3.6128/2.6486
		2	0.6404/0.4911	0.9500/0.6746	6.2280/5.6617	3.7521/2.5645
	5	1	0.6899/0.4989	0.8618/0.6124	6.2218/6.3820	2.9503/2.5823
		2	0.6996/0.5634	1.0017/0.7915	6.3203/6.1131	4.1867/3.2785
48 h	1	1	1.2198/0.8343	1.7735/1.4304	9.7782/5.7246	6.8402/4.0289
		2	1.2598/0.8949	1.6941/1.2488	9.4557/7.2439	6.5030/4.4361
	2	1	1.2250/0.9604	2.1960/1.6746	10.9775/8.4764	7.9032/5.9307
		2	1.2857/1.0131	2.0330/1.5552	8.1434/6.3813	6.5293/4.3273
	3	1	1.2407/0.7253	2.1655/1.7786	9.6580/6.9204	7.5541/5.2989
		2	1.3250/0.9565	1.8414/1.5666	8.1024/7.2309	6.2452/5.3772
	4	1	0.9162/0.8120	1.7618/1.3047	7.9009/5.5968	5.7899/4.5371
		2	0.9896/0.7848	1.5020/1.1732	7.0978/4.8208	5.2826/3.7708
	5	1	0.9650/0.7833	1.6960/1.4327	9.1107/7.4706	8.7411/6.4006
		2	1.1315/1.0845	1.9731/1.5516	9.2449/6.8041	6.8057/4.8502

forecast model (1 or 2) are divided into two groups with the same number of observation points. One group of observation points is used to establish prediction equations and the other group of points is used to test the accuracies of these equations. The results of testing these prediction equations in forecast model 1 and forecast model 2 are shown in Table 8. In combination with formula (3), the calculated mean and standard deviation of GCD in two forecast models are shown in Figures 15 and 16, respectively. Averaging the results of five areas, it can be obtained that the averages of the mean/standard deviation of GCD under forecast model 1 and forecast model 2 for 24 hours' forecast fare 150.5192/84.6156 km and 141.5464/81.2509 km, respectively. For 48 hours' forecast, the averages of the mean/standard deviation of GCD under two different forecast models are 261.7517/145.6345 km and 256.7109/145.2903 km, respectively. Even though the mean of GCD in dynamic prediction pattern is no less than three main prediction centers (NHC, JMA, and NMCC), it is much less than the numerical prediction model in [25], the means of which are 186.3/319.5 km based on System T106 and 161.8/295.8 km based on T213 both for 24/48 hours' forecast. Besides, forecast models 1 and 2 are all more accurate than the forecast using satellite scatterometer's monitoring data in the sense of the standard deviation of GCD and the mean of weed speed error, which are 149.6002 km and 11.9618 m/s in [12]. It can be seen from Table 8 and Figures 15 and 16 that the accuracies of forecast model 1 and forecast model 2 vary from different areas and different characteristic factors. The more accurate forecast model can be selected from forecast models 1 and 2 according to actual conditions. The results of statistical significance tests for each equation used to forecast corresponding characteristic factor in forecast model 1 and forecast model 2 are shown in Table 9, which show that P value is much less than

Table 9: The results of statistical significance tests for each equation used to forecast corresponding characteristic factor in forecast model 1 and forecast model 2.

Forecast hour	Area number	Forecast model (1 or 2)	R-square/P value for LAT	R-square/P value for LON	R-square/P value for AP	R-square/P value for WS
24 h	1	1	$0.5883/9.52 \times 10^{-11}$	$0.6011/7.91 \times 10^{-9}$	$0.6222/4.46 \times 10^{-10}$	$0.7615/5.47 \times 10^{-15}$
		2	$0.4493/2.26 \times 10^{-32}$	$0.5809/5.93 \times 10^{-50}$	$0.7030/1.31 \times 10^{-68}$	$0.7731/1.01 \times 10^{-81}$
	2	1	$0.7177/1.86 \times 10^{-14}$	$0.4391/3.66 \times 10^{-5}$	$0.7556/5.03 \times 10^{-16}$	$0.7995/3.57 \times 10^{-18}$
		2	$0.6808/5.17 \times 10^{-101}$	$0.6753/1.77 \times 10^{-99}$	$0.7481/2.13 \times 10^{-122}$	$0.7277/3.71 \times 10^{-114}$
	3	1	$0.6471/6.36 \times 10^{-11}$	$0.6559/3.48 \times 10^{-11}$	$0.6310/1.83 \times 10^{-10}$	$0.6495/3.24 \times 10^{-10}$
		2	$0.6802/6.46 \times 10^{-94}$	$0.7506/1.19 \times 10^{-114}$	$0.5509/1.40 \times 10^{-62}$	$0.5573/9.84 \times 10^{-65}$
	4	1	$0.7572/1.05 \times 10^{-12}$	$0.5348/3.32 \times 10^{-8}$	$0.7594/1.15 \times 10^{-13}$	$0.8174/2.47 \times 10^{-15}$
		2	$0.6405/4.83 \times 10^{-42}$	$0.6266/1.64 \times 10^{-40}$	$0.6903/1.01 \times 10^{-44}$	$0.7652/9.25 \times 10^{-57}$
	5	1	$0.5031/2.40 \times 10^{-6}$	$0.5154/1.51 \times 10^{-6}$	$0.8732/2.73 \times 10^{-14}$	$0.8524/5.03 \times 10^{-15}$
		2	$0.6239/2.84 \times 10^{-35}$	$0.6464/1.67 \times 10^{-37}$	$0.7835/6.98 \times 10^{-52}$	$0.7550/1.52 \times 10^{-48}$
48 h	1	1	0.1337/0.0240	$0.3943/3.91 \times 10^{-5}$	$0.3208/1.77 \times 10^{-4}$	$0.4725/5.34 \times 10^{-6}$
		2	0.2177/0.0049	0.0399/0.3398	0.2532/0.0016	$0.4275/7.92 \times 10^{-6}$
	2	1	$0.4515/9.05 \times 10^{-6}$	0.2668/0.0218	$0.6720/1.72 \times 10^{-10}$	$0.4622/5.99 \times 10^{-6}$
		2	$0.4022/1.64 \times 10^{-36}$	$0.3095/6.51 \times 10^{-26}$	$0.2960/5.05 \times 10^{-23}$	$0.3332/6.27 \times 10^{-27}$
	3	1	0.2251/0.0218	$0.3935/4.57 \times 10^{-5}$	0.0026/0.7270	0.1362/0.1596
		2	$0.3073/6.18 \times 10^{-25}$	$0.4369/2.13 \times 10^{-39}$	$0.1768/2.51 \times 10^{-12}$	$0.2479/8.91 \times 10^{-18}$
	4	1	$0.4992/9.60 \times 10^{-7}$	$0.3055/2.74 \times 10^{-4}$	$0.4303/1.53 \times 10^{-5}$	$0.6249/1.84 \times 10^{-9}$
		2	$0.3323/3.77 \times 10^{-16}$	$0.3270/1.02 \times 10^{-16}$	$0.2471/1.70 \times 10^{-9}$	$0.3580/3.65 \times 10^{-16}$
	5	1	$0.4546/6.19 \times 10^{-5}$	0.2554/0.0043	$0.3705/1.91 \times 10^{-4}$	0.3535/0.0012
		2	$0.4219/1.09 \times 10^{-18}$	$0.3825/3.44 \times 10^{-17}$	$0.2120/2.18 \times 10^{-7}$	$0.3322/2.46 \times 10^{-12}$

0.05 in almost every case and prove that the corresponding prediction equation is significant.

4. Summary

In this paper, the CMA best track datasets from 1949 to 2012 are used, in combination with data mining technology and statistical methods, to put forward a new methodology to forecast TCs' characteristic factors. This methodology can accurately judge whether TCs land on Hainan Island or not and forecast their characteristic factors (including longitude, latitude, the lowest center pressure, and wind speed). The average of the probabilities of accurate judgment for landing criterions is 74.70% and the highest accuracy can reach 79.76%. For the forecast of landing TCs' characteristic factors, landing prediction pattern and dynamic prediction pattern are proposed, which not only can accurately forecast the characteristic factors when TCs land but also realize dynamically 24 hours' and 48 hours' forecast. The effect of the landing prediction pattern is better, of which the mean of GCD is 144.6382 km, compared with the current 48 hours' forecast in the South China Sea, which is 222.6 km. Even though the mean of GCD in dynamic prediction pattern is no less than three main prediction centers (NHC, JMA, and NMCC), it is much less than the numerical prediction model in [25] and the method using satellite scatterometer's monitoring data in [12]. The forecast methodology proposed in this paper provides a new method for typhoon warning on Hainan Island without getting too much knowledge of meteorology involved and thus simplifies the implementation of the prediction process and meanwhile guarantees the accuracy of prediction.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is in part supported by the National Science-Technology Support Plan in the domain of advanced energy technology. Hainan Power Grid Corporation provided support for this research through "Integrated Demonstration Project of Regional Smart Grid" no. 2013BAA01B03. Dr. Xinhong Huang, a Research Engineer in the Department of Electrical & Computer Engineering from the University of Western Ontario, also puts forward valuable revision comments on this paper.

References

[1] National Climate Center of China, *Typhoon "Damrey" Has Caused Serious Damages to Hainan Province*, 2005, http://ncc.cma.gov.cn/Website/index.php?NewsID=1454.

[2] J. S. Pedro, F. Burstein, and A. Sharp, "A case-based fuzzy multicriteria decision support model for tropical cyclone forecasting," *European Journal of Operational Research*, vol. 160, no. 2, pp. 308–324, 2005.

[3] X. Qin and M. Mu, "A Study on the reduction of forecast error variance by three adaptive observation approaches for tropical cyclone prediction," *Monthly Weather Review*, vol. 139, no. 7, pp. 2218–2232, 2011.

[4] Y. Wang, W. Zhang, and W. Fu, "Back Propogation(BP)-neural network for tropical cyclone track forecast," in *Proceedings of the 19th International Conference on Geoinformatics*, pp. 1–4, IEEE, Shanghai, China, June 2011.

[5] B. Feng and J. N. K. Liu, "An adaptive neural network classifier for tropical cyclone prediction using a two-layer feature selector," in *Advances in Neural Networks—ISNN 2005*, vol. 3497 of *Lecture Notes in Computer Science*, pp. 399–404, Springer, Berlin, Germany, 2005.

[6] H.-J. Song, S.-H. Huh, J.-H. Kim, C.-H. Ho, and S.-K. Park, "Typhoon track prediction by a support vector machine using data reduction methods," in *Computational Intelligence and Security*, vol. 3801 of *Lecture Notes in Computer Science*, pp. 503–511, Springer, Berlin, Germany, 2005.

[7] M. DeMaria, "A simplified dynamical system for tropical cyclone intensity prediction," *Monthly Weather Review*, vol. 137, no. 1, pp. 68–82, 2009.

[8] H. R. Winterbottom, E. W. Uhlhorn, and E. P. Chassignet, "A design and an application of a regional coupled atmosphere-ocean model for tropical cyclone prediction," *Journal of Advances in Modeling Earth Systems*, vol. 4, no. 10, Article ID M10002, 2012.

[9] L. M. Ma and Z. M. Tan, "Improving the behavior of the cumulus parameterization for tropical cyclone prediction: convection trigger," *Atmospheric Research*, vol. 92, no. 2, pp. 190–211, 2009.

[10] J. S. Gall, I. Ginis, S.-J. Lin, T. P. Marchok, and J.-H. Chen, "Experimental tropical cyclone prediction using the GFDL 25-km-resolution global atmospheric model," *Weather and Forecasting*, vol. 26, no. 6, pp. 1008–1019, 2011.

[11] G. Lv, "The Northwestern Pacific typhoon track forecast in 2005," in *Proceedings of the Meteorological Science and Technology Symposium on Taiwan Strait*, p. 7, Chinese Meteorological Society, 2006, (Chinese).

[12] D. Zhang, Y. Zhang, T. Hu, B. Xie, and J. Xu, "A comparison of HY-2 and QuikSCAT vector wind products for tropical cyclone track and intensity development monitoring," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 8, pp. 1365–1369, 2014.

[13] M. Ying, W. Zhang, H. Yu et al., "An overview of the China meteorological administration tropical cyclone database," *Journal of Atmospheric and Oceanic Technology*, vol. 31, no. 2, pp. 287–301, 2014.

[14] G. E. Forsythe, "Generation and use of orthogonal polynomials for data-fitting with a digital computer," *Journal of the Society for Industrial & Applied Mathematics*, vol. 5, no. 2, pp. 74–88, 1957.

[15] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained k-means clustering with background knowledge," in *Proceedings of the 18th International Conference on Machine Learning*, vol. 1, pp. 577–584, 2001.

[16] Y. Ye, "Neighborhood density method for selecting initial cluster centers in K-mean clustering," in *Proceedings of the Workshop on Data Mining for Biomedical Applications (PAKDD '06)*, pp. 189–198, 2006.

[17] P. Xiong, *Data Mining Algorithm and Clementine Practice*, Tsinghua University Press, Beijing, China, 2011 (Chinese).

[18] S. L. Crawford, "Extensions to the CART algorithm," *International Journal of Man-Machine Studies*, vol. 31, no. 2, pp. 197–217, 1989.

[19] D. M. Allen, "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, vol. 13, pp. 469–475, 1971.

[20] S. Yu and J. Shen, "Forward and backward algorithms for selecting predictors on the basis of the criterion from prediction sum of squares and their application," *Acta Meteorologica Sinica*, vol. 1, pp. 83–90, 1988.

[21] D. Yao and S. Yu, "The stepwise algorithm of selecting forecast factors based on PRESS rule," *Journal of Atmospheric Sciences*, vol. 2, pp. 129–135, 1992 (Chinese).

[22] Y. Lu, *Mathematical Statistics Methods*, East China University of Science and Technology Press, Shanghai, China, 2005, (Chinese).

[23] K. Xie and B. Liu, "An ENSO-forecast independent statistical model for the prediction of annual Atlantic tropical cyclone frequency in April," *Advances in Meteorology*, vol. 2014, Article ID 248148, 11 pages, 2014.

[24] J. Yu, J. Tang, Y. Dai, and B. Yu, "The error and cause analysis of China's typhoon path prediction," *Journal of Weather*, vol. 6, pp. 695–700, 2012.

[25] S. Ma, A. Qu, and Z. Yu, "The parallelization of typhoon numerical prediction model of and track forecast error analysis," *Journal of Applied Meteorology*, vol. 3, pp. 322–328, 2004 (Chinese).