A Study on The Frequency Characteristics of Typhoon Landing in Guangdong, China, Based on Machine Learning Methods

Chenning Pan
Joint International Research Laboratory of Catastrophe Simulation and Systemic Risk Governance, Beijing Normal University, Zhuhai 519087, China;
School of National Safety and Emergency Management, Beijing Normal University, Zhuhai 519087, China
202321099015@mail.bnu.edu.cn

Xiaoyong Ni*
Joint International Research Laboratory of Catastrophe Simulation and Systemic Risk Governance, Beijing Normal University, Zhuhai 519087, China;
School of National Safety and Emergency Management, Beijing Normal University, Zhuhai 519087, China
nixiaoyong@bnu.edu.cn

Ruoxi Lai
Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China
202211079136@mail.bnu.edu.cn

Dafeng Ma
College of Education for the Future, Beijing Normal University, Zhuhai 519087, China
202211038583@mail.bnu.edu.cn

Xifan Shen
Bay Area Internatinoal Business School,
Beijing Normal University, Zhuhai 519087, China
202111059036@mail.bnu.edu.cn

*Correspondence: nixiaoyong@bnu.edu.cn

ABSTRACT

Located in the southern coastal region of China, Guangdong Province is perennially threatened by typhoons originating from the Northwest Pacific Ocean. Compared to studies focusing on typhoon intensity, paths, and catastrophic effects, research on typhoon landing frequency features remains relatively limited. This study systematically constructs a dataset based on the landing typhoon frequency of Guangdong over 71 years (as the target variable) and a set of 88 atmospheric circulation indices as well as 26 sea surface temperature indices (as the potential feature variables). Two machine learning models, including Random Forest Regression (RR) and Support Vector Regression (SVR), are used to predict the frequency of typhoons making landfall in Guangdong, and the results show that both models have good characterisation ability while the RR model has slightly better fitting performance on the training set. This study provides scientific support for understanding the characteristics of typhoons in Guangdong and better responding to typhoon disasters.

Keywords

Tropical cyclone, frequency characteristics, machine learning, vulnerability functions.

INTRODUCTION

According to the definition of the World Meteorological Organization, a tropical cyclone with sustained wind speeds above level 12 (i.e. 32.7 to 41.4 meters per second) at its center is called a typhoon or hurricane, where generally the former is used to refer to tropical cyclones generated in the Pacific Ocean, while the latter is used to refer to those generated in the Atlantic Ocean. According to the traditional Chinese custom of naming tropical cyclones, tropical cyclones are collectively referred to as typhoons in the following. The 2022 Global Natural Disaster Assessment Report points out that there are a total of 66 major storm disasters (typhoons and hurricanes) worldwide in 2022, accounting for approximately 21% of the total frequency of major disasters; 16.93 million people are affected, resulting in direct economic losses of approximately 131 billion US dollars (Academy of Disaster Reduction and Emergency Management et al., 2023). The impact and challenges caused by typhoon disasters cannot be ignored and are a global problem facing humanity. China is located in the northwestern part of the Pacific Ocean, which is a high risk area for typhoon disasters. In the 71 years from 1949 to 2019, there are 2,333 typhoons in the Northwest Pacific, of which 636 have landed in China, accounting for 27.3% of the total (Liu et al., 2012; Su et al., 2021). Guangdong Province is located in the coastal area of southern China and is the main region where typhoon makes landfall in China, making it one of the provinces most severely affected by typhoons (Zhang et al., 2009; Chou et al., 2018). In 2018, Typhoon Mangkhut makes landfall in Guangdong with 14 gusts of wind, affecting 3.066 million people, causing the collapse of 200,000 houses and direct economic losses of 13.29 billion yuan. In early September 2023, Typhoon Sura and Haikui make landfall in Guangdong, causing record-breaking rainfall in various parts of the province. The disasters have caused serious consequences, including road collapses, interruptions in communication and power supply, house collapses, fallen trees, and damage to water conservancy projects, particularly in Shenzhen and Yangjiang.

With the severity of global climate change, the increasing intensity of worldwide natural disasters such as typhoons is causing progressively greater losses globally. Typhoon characteristics prediction has become a major focus of research. In general, the studies on typhoon characteristics prediction include typhoon frequency prediction (Jia et al., 2014; Hai et al., 2019), track prediction (Murakami et al., 2010; Hsan et al., 2021), landfall prediction (Wahiduzzaman et al., 2017; Kumar et al., 2021), wind field strength prediction (Kotal et al., 2020; Chen et al., 2023), rain field strength prediction (Matyas, 2010), and so on. The study of the characteristics of typhoon landing frequency can understand the law and mechanism of typhoon activities and the impact of climate change on typhoon activities, better early warning and risk management, so this paper mainly focuses on the prediction of typhoon landing frequency.

The traditional methods for studying the characteristics of typhoon landing frequency mainly include two categories: statistical analysis methods and numerical analysis methods. Statistical methods do not focus on physical principles but on the exploration of data patterns in historical disasters. Yang et al. find that the landfall typhoons numbers as well as the south-north range of landfall typhoons have decreased by using the data from the Tropical Cyclone Yearbooks (Yang et al., 2009). Numerical analysis methods are built based on principles of atmospheric dynamics and thermodynamics to discretize and simulate the atmosphere for weather forecasting. Nie et al. find that the average intensity and number of typhoons landing in China under three RCPs will increase in varying degrees in the future, based on outputs from the global climate model HadGEM2-ES in CMIP5 (Nie et al., 2023). With the popularization of artificial intelligence technology, machine learning has gradually become a popular method in typhoon characteristics prediction research due to their ability to process large amounts of complex meteorological data and capture their nonlinear and non parametric relationships. Richman et al. implement a Support Vector Regression model based on the May-September Nino index, a novel El Nino, a quasi-biennial oscillation in the stratosphere, and a Southern Hemisphere ring mode to estimate the frequency of typhoons in the Australian typhoon season from December to April (Richman et al., 2012). Tan et al. select environmental fields such as sea surface temperature, sea level pressure, Nino 3.4 index, wind shear, vorticity, subtropical high pressure and sea ice cover, and establish a random forest model based on the prediction of tropical cyclone frequency in the Northwest Pacific Ocean (Tan et al., 2018). Chen chooses six types of forecast elements, including temperature field, altitude field, sea-level pressure, specific humidity field, meridional wind field vertical shear, and sea surface temperature field to perform single machine learning model predictions and integrated predictions (Chen, 2022).

Overall, compared to traditional weather prediction methods such as numerical models, machine learning methods are still in the exploratory stage of typhoon characteristics prediction research. Its application research scope is wide, but it mainly focuses on the study of typhoon intensity, weather and its catastrophic effects such as wind speed and rain intensity (Chen et al., 2020). There is still relatively little research on typhoon landing frequency. Studying the frequency of landfall can help understand the characteristics and trends of typhoon frequency in different regions, help governments and communities better prepare and respond to possible typhoon disasters, conduct more effective meteorological disaster risk assessments, and reduce the impact of disasters on people's

lives and property. Therefore, studying the frequency of typhoons is of great significance for guiding regional disaster prevention and reduction.

THE OVERALL STRUCTURE

The general structure of this study is shown in Figure 1. Guangdong Province has been selected as the study area, which is one of the most typhoon-prone provinces in mainland China. Typhoon track data from 1952 to 2022 and a potential set of feature variables have been collected. Through data processing, datasets of both feature variables and target variables have been obtained, used for training two machine learning models. The ability of the two models to characterise and predict the frequency of typhoons making landfall in Guangdong Province is compared. The results of this paper help to understand the characteristics of typhoons in Guangdong Province and provide reference and technical support for subsequent studies on typhoon landing frequency predicion, as well as help to improve the government's capabilities in response to typhoons.

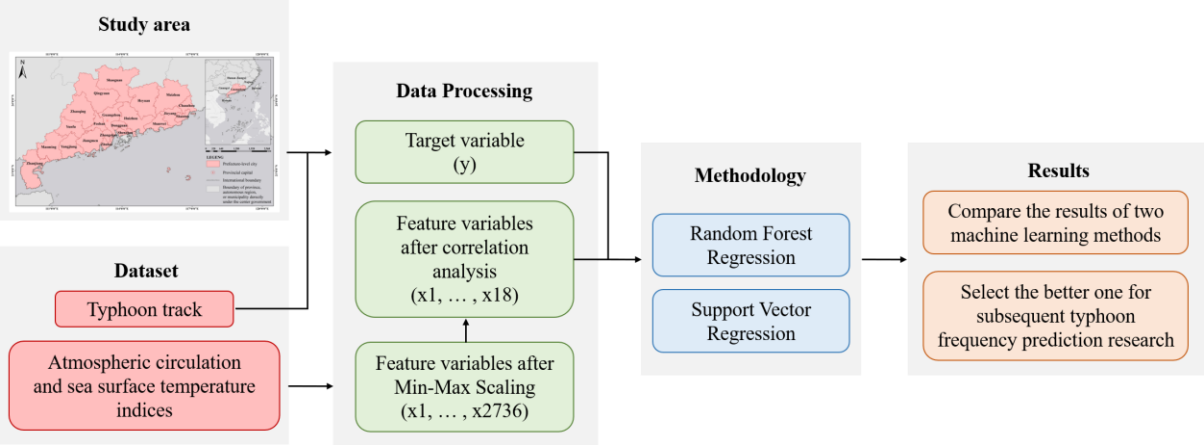


Figure 1. The Overall Structure of This Study

STUDY AREA AND DATASETS

Study Area

Guangdong Province is a provincial-level administrative region in China. It is located along the coast of the South China Sea, bordered by Hong Kong, Macau, Guangxi, Hunan, Jiangxi and Fujian, and across the sea from Hainan Province, with a total land area of 179,800 square kilometers. As of October 2022, Guangdong is composed of 21 prefecture-level cities, 65 districts, 20 county-level cities, 34 counties, and 3 autonomous counties. In 2022, the gross domestic product (GDP) of Guangdong Province reached 12,911.86 billion yuan, with a per capita GDP of 101,905 yuan. At the end of 2022, the permanent population of Guangdong was 126.568 million, with an urbanization rate of 74.79%. Figure 2 illustrates the study area in this paper.

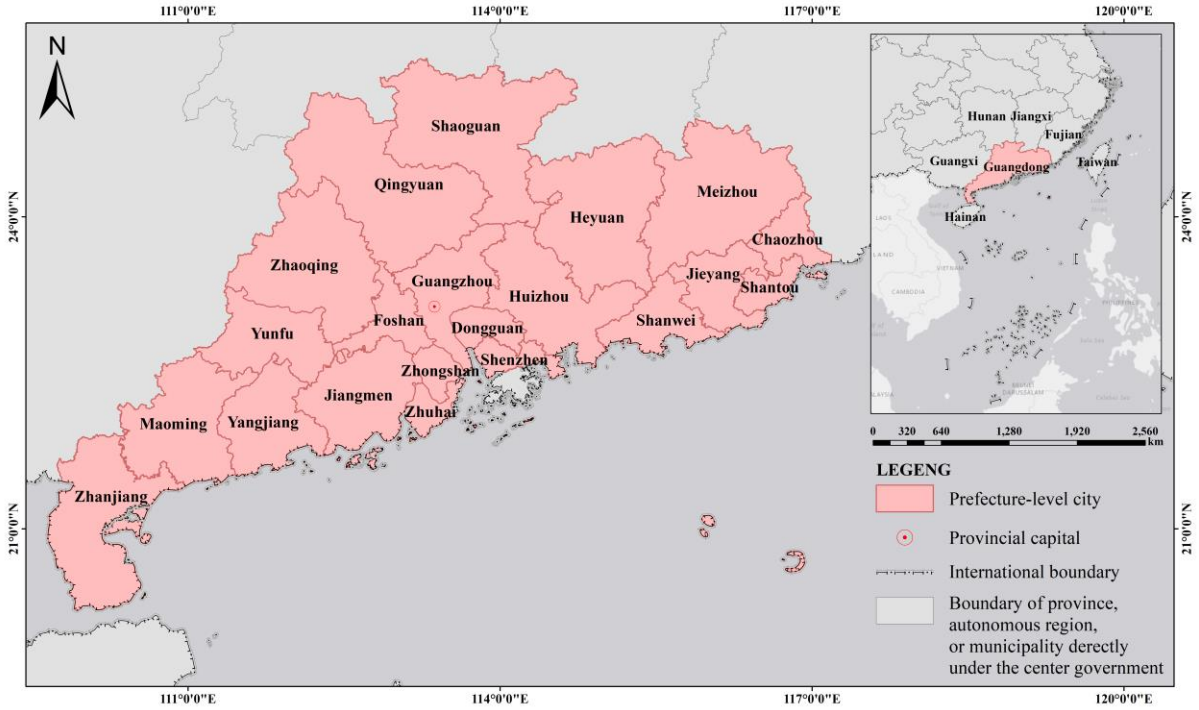


Figure 2. The Study Area in This Study

Data Sources

The typhoon track data is sourced from the Best Track Data provided by the China Meteorological Administration Tropical Cyclone Data Center (CMA, 2023; Ying et al., 2014; Lu et al., 2021). This dataset provides basic information on the position and intensity of tropical cyclones in the northwest Pacific (including the South China Sea, north of the equator, and west of 180°E longitude) every 6 hours since 1949. Starting from 2017, for typhoons that make landfall in China, the time frequency of the best track data has been increased to once every 3 hours within the 24-hour period before their landfall. Since 2018, for typhoons making landfall in China, the time frequency of the best track data is increased to every 3 hours within the 24 hours before their landfall and during their land activities In China.

Regarding the potential feature factors influencing typhoon frequency, the monthly data of 88 atmospheric circulation indices and 26 sea surface temperature indices provided by the National Climate Center from 1951 to 2022 are considered by this study, mainly based on the research of Rong et al. (National Climate Center, 2023; Rong et al., 2023).

Raw Data Processing

Feature Variables Processing

For the feature variables, this study utilizes monthly atmospheric circulation indices and sea surface temperature indices for the years 1951-2022. The entire feature variable dataset comprises 71 entries (an entry per year from 1952 to 2022), with each entry including 88 atmospheric circulation indices such as the Northern Hemisphere Subtropical High Area Index, as well as 26 sea surface temperature indices like the NINO 1+2 SSTA Index, for a total of 24 months corresponding to the target variable's current year and previous year, resulting in a total of 2736 [=(88+26)*24] feature variables per entry.

In the feature variable dataset, the differences in meanings of various variables and units of measurement can affect the performance of machine learning algorithms. Therefore, it is necessary to normalize data from different dimensions, allowing for comparison and interpretation on the same scale. In this study, Min-Max Scaling is employed to normalize the feature variable dataset.

The Min-Max Scaling method scales the original data proportionally to fit within the range [0, 1]. The specific formula for this process is as follows:

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (1)$$

Here, x_{ij} denotes the value of the j-th variable for the i-th year, $\min(x_j)$ and $\max(x_j)$ are the minimum and maximum values of the j-th variable in the whole dataset, and x'_{ij} denotes the normalized value after processing.

Acquisition and Analysis of the Target Variable

Based on the best track data of tropical cyclones, which provides latitude and longitude information at regular intervals for each typhoon, combined with the longitude and latitude range of Guangdong Province, whether the typhoon made landfall in mainland China from Guangdong Province is determined. Here, landing in Guangdong refers to the first landfall location of a typhoon in China being Guangdong, the situation where a typhoon lands in other provinces of China first and then makes a secondary landfall in Guangdong is not considered. Summarizing the typhoons landing in Guangdong each year, the frequency data of typhoons making landfall in Guangdong from 1952 to 2022 (target variable) is obtained. It reveals that a total of 250 typhoons made landfall in Guangdong during this period, with an average of 3.52 typhoons per year and a standard deviation of 1.71. The highest annual typhoon frequency reached seven, while the lowest was one. Figure 3 reflects the annual changes in the frequency of typhoons landing in Guangdong Province. The linear trendline indicates a significant downward trend in the number of typhoons making landfall in Guangdong Province over the 71 years. The 5-year moving average curve shows that the landing typhoon frequency exhibits overall fluctuating characteristics, entering a period of reduced frequency starting from the late 20th century.

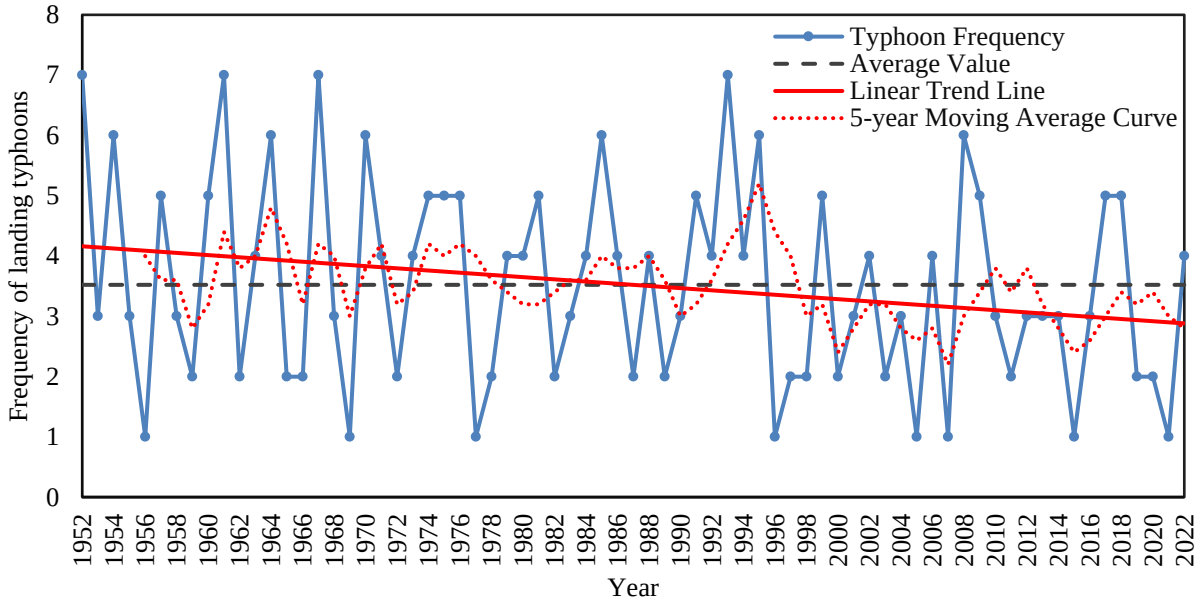


Figure 3. The Annual Distribution of Landing Typhoon Frequency in Guangdong Province

METHODOLOGY

Random Forest Regression

Random Forest Regression (RR) is an important application branch of Random Forest. The RR model builds multiple unrelated decision trees by randomly extracting samples and features, following the principles of Bagging, and obtains predictions in parallel. Each decision tree can produce a prediction result from the extracted samples and features, and the regression prediction result of the whole forest is obtained by averaging the results of all trees. This model improves the Bagging ensemble algorithm by incorporating random attribute selection alongside the bootstrap sampling method during the training of decision trees. This introduces diversity among weak learners, enhancing the model's generalization performance.

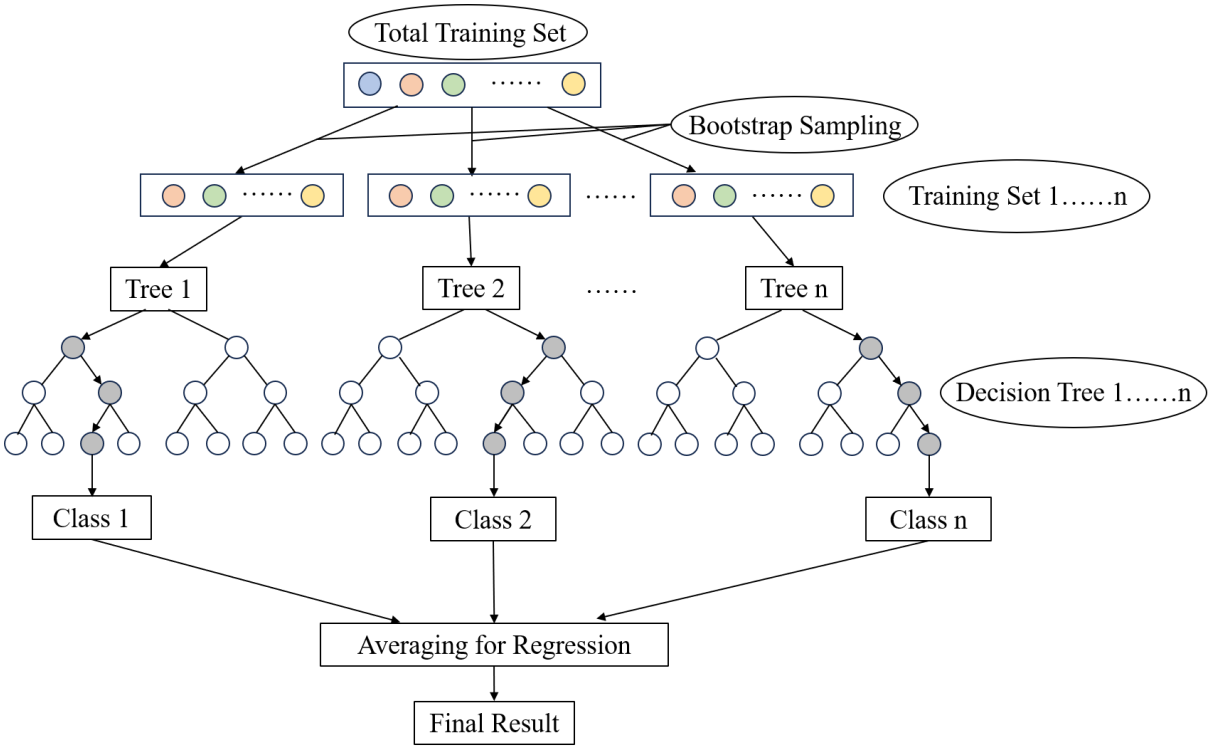


Figure 4. Flowchart of Random Forest Regression Algorithm

Support Vector Regression

Support Vector Regression (SVR) is a supervised learning method that uses Support Vector Machines (SVM) for regression analysis. The goal of SVR is to find the optimal hyperplane, allowing for a certain amount of error, such that the sample points are as close to this hyperplane as possible. SVR achieves this by optimizing the following convex quadratic programming problem:

$$\min_{\omega,b,\xi,\hat{\xi}} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}(\xi_i + \hat{\xi}_i) \tag{2}$$

Here, ω denotes the weight vector of the hyperplane, b denotes the intercept, ξ and $\hat{\xi}$ denote slack variables allowing some data points to be within the margin or error range, respectively, and C denotes the regularization parameter controlling the trade-off between margin size and error.

The objective of SVR is to minimize the loss function mentioned above while satisfying the following constraints:

$$yi - \langle\omega, \phi(x_i)\rangle - b \leq \epsilon + \xi_i \tag{3}$$

$$\langle\omega, \phi(x_i)\rangle + b - yi \leq \epsilon + \hat{\xi}_i \tag{4}$$

$$\xi_i, \hat{\xi}_i \geq 0 \tag{5}$$

These constraints ensure that data points are within the margin or error range. The parameter ϵ controls the size of the margin. SVR finds the optimal hyperplane by minimizing the loss function, enabling it to make predictions.

Metrics for Evaluating Machine Learning Performance

In order to compare the learning effects of the two machine learning models, Mean Squared Error (MSE), Root Mean Squard Error (RMSE) and Mean Absolute Error (MAE) are chosen as the evaluation metrics for the training results. Smaller MSE, MAE and RMSE suggest better fit of the model to the data. The formulas to calculate MSE, RMSE and MAE are as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

In which, n denotes the number of samples, y_i denotes the actual values, and $\hat{y}_i$ denotes the predicted values of the model.

RESULTS

Correlation Analysis

To prevent overfitting and reduce redundant information caused by numerous feature variables, a correlation analysis between the target variable and feature variables is performed, yielding correlation coefficients of each feature variable to the target variable. Based on the correlation coefficients, features with a strong correlation with the target variable have higher priority to be selected thereby simplifying the feature variables dataset, which is recommended by Rong et al. as a simple but effective way of parameter selection (Rong et al., 2023). Given a dataset D = $\{(X_K, Y_K)\}$, where $K \in n$, the correlation coefficient r is defined as:

$$r = \frac{\sum_{K=1}^{n}(X_K - \bar{X})(Y_K - \bar{Y})}{\sqrt{\sum_{K=1}^{n}(X_K - \bar{X})^2}\sqrt{\sum_{K=1}^{n}(Y_K - \bar{Y})^2}} \tag{9}$$

where $\bar{X}$ denotes the mean of X and $\bar{Y}$ denotes the mean of Y.

After calculating the correlation coefficients for each feature variable to the target variable, those with missing values (originally marked as -999 in the raw data) are removed. The remaining features are sorted in descending order based on the absolute values of their correlation coefficients. Subsequently, multiple tests are conducted based on the changing dataset, where every three sequential cumulative increases in the number of feature variables, and the dataset composed of the first 18 feature variables is discovered to have the better prediction effect. Therefore, in this paper, 18 feature variables are set as shown in Table 1, and due to space limitations, the process of selecting the number of feature variables is omitted. These 18 feature variables respectively represent the sea surface temperature indices (x2723, x2734, x1902, x2620, x2724, x1580, x2609, x2718, x2610, x2604, x2506, x1694, x2512) and atmospheric pressure indices (x2575, x2573, x2337, x411, x1605) related to typhoon formation (For detailed explanations of these indices, please refer to the National Climate Center's website). Furthermore, only x411 is a feature variable from the previous year, affecting this year's frequency of typhoon landing. Therefore, it can be observed that sea surface temperature factors have a greater impact on the frequency of typhoon landing in Guangdong, and the index of the same year has more influence compared to the index of the previous year. The dataset of feature variables together with the target variable will serve as the initial dataset for machine learning models. The dataset is split into 80% for training and 20% for testing.

Table 1. The Top Eighteen Feature Variables Ranked by the Absolute Values of Their Correlation Coefficients

Number	Feature variable identifier	Feature variable name	Absolute value of the correlation coefficient
1	x2723	Indian Ocean Warm Pool Area Index in December of the same year	0.4015
2	x2734	Indian Ocean Basin-Wide Index in December of the same year	0.3995
3	x1902	Scandinavia Pattern , SCA in May of the same year	0.3949
4	x2620	Indian Ocean Basin-Wide Index in November of the same year	0.3766
5	x2724	Indian Ocean Warm Pool Strength Index in December of the same year	0.3731

6	x1580	Tropical Northern Atlantic SST Index in February of the same year	0.3659
7	x2609	Indian Ocean Warm Pool Area Index in November of the same year	0.3605
8	x2575	India-Burma Trough Intensity Index in November of the same year	0.3564
9	x2718	NINO B SSTA Index in December of the same year	0.3514
10	x2610	Indian Ocean Warm Pool Strength Index in Novermber of the same year	0.3483
11	x2604	NINO B SSTA Index in Novermber of the same year	0.3385
12	x2573	Tibet Plateau Region 1 Index in Novermber of the same year	0.3376
13	x2337	Northern Hemisphere Polar Vortex Central Latitude Index in September of the same year	0.3373
14	x2506	Indian Ocean Basin-Wide Index in October of the same year	0.3373
15	x411	Antarctic Oscillation, AAO in April of the previous year	0.3334
16	x1605	South China Sea Subtropical High Area Index in March of the same year	0.3332
17	x1694	Tropical Northern Atlantic SST Index in March of the same year	0.3298
18	x2152	Western Hemisphere Warm Pool Index in July of the same year	0.3293

Random Forest Regression Model Prediction

Key parameters affecting the fitting performance of the RR model, namely n_estimators (number of trees), max_depth (maximum depth of the tree), min_samples_split (minimum samples required for a node to split), and min_samples_leaf (minimum samples required for a leaf node), are performed parameter tuning by using a parameter grid which is formed by commonly used numerical values respectively. After multiple rounds of optimization and combining the fitting results, the values 200, 20, 10, and 4 are chosen for the respective parameters.

Due to the random features inherent in the RR model, the results obtained from each training iteration are not entirely consistent. Therefore, Table 2 represents the average training results after 20 repetitions. The model exhibites good performance on the training set, where the prediction deviation for typhoon landing frequency is around 1. However, the performance on the test set is relatively poor, with the values of three metrics significantly higher than on the training set. Comparison of the actual and predicted values for the training and test sets from the 20th training iteration is shown in Figure 5. The RR model demonstrates a consistent trend with the actual values in the training set, showing relatively high concordance with peak and trough values in certain years. On the other hand, the results for the test set, highlighted with a deep red background in the Figure 5, exhibites suboptimal trends and extreme value predictions. The maximum prediction error for the number of landing typhoons reaches 2.4.

Table 2. The Results of RR Training Evaluation Metrics

Type of dataset	Evaluation indicators	Values
Training set	MSE	1.0857 ± 0.0152
	RMSE	1.0419 ± 0.0073

	MAE	0.8758 ± 0.0096
	MSE	1.8042 ± 0.0511
Test set	RMSE	1.3431 ± 0.0192
	MAE	1.1669 ± 0.0178

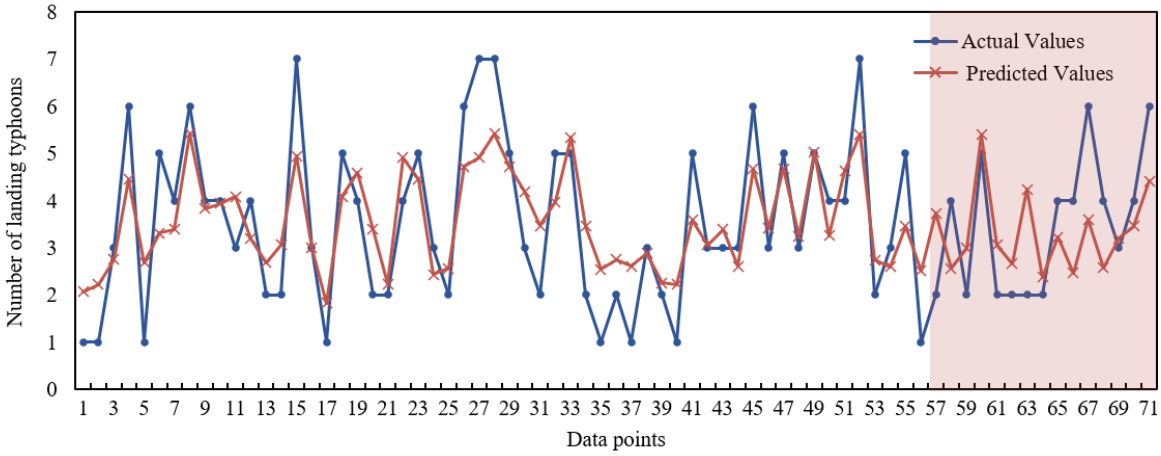


Figure 5. RR Training Results Plot for Training and Test Sets (Highlighted in Deep Red Background for Test Set Training Results)

The average feature importance values from the 20 training iterations are illustrated in Figure 6. Among them, the four feature variables, Scandinavia Pattern, SCA in May of the same year (x1902), Antarctic Oscillation, AAO in April of the previous year (x411), Tropical Northern Atlantic SST Index in February of the same year (x1580), and Tibet Plateau Region 1 Index in Novermber of the same year (x2573), demonstrate higher importance in influencing the overall training performance of the model, with importance values of 0.2097, 0.1159, 0.1154, and 0.0899, respectively. The relevant results show that time coefficients of the ninth mode from empirical orthogonal function analysis (EOF) of the normalised 500hPa height field in the region 20°N-90°N, 0-360° in May as well as regional average of sea surface temperature distance levels in the region 5.5°N-23.5°N, 57.5°W-15°W in February of the same year (x1902, x1580) and normalised series of time coefficients of the first mode from EOF of the anomalous field at 700hPa height in the region 20°-90°S, 0-360° in April of the previous year (x411) will have a relatively important impact on the number of typhoon landing in Guangdong Province in a year.

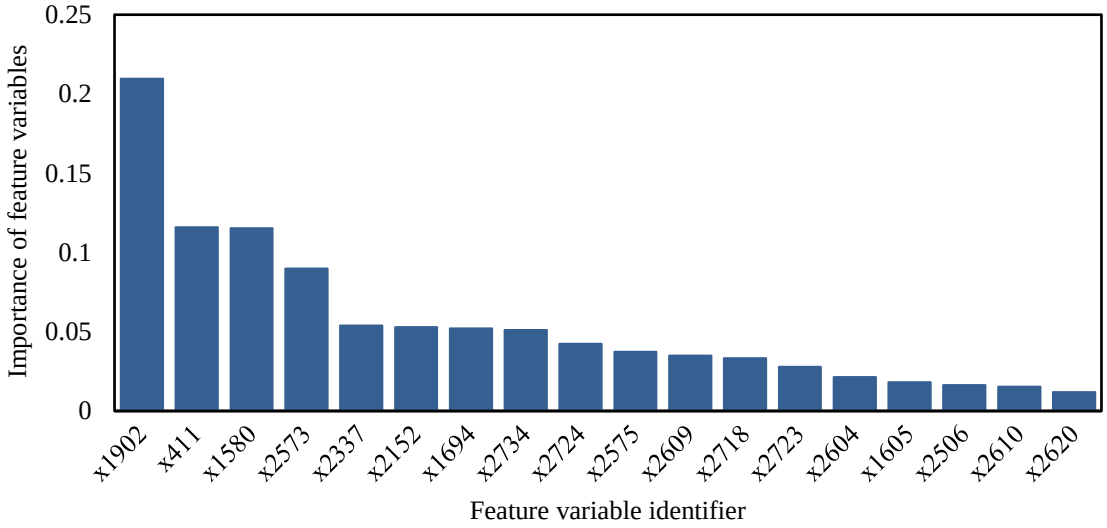


Figure 6. Importance Level of Eighteen Feature Variables

Support Vector Regression Model Prediction

In SVR model, the parameters affecting training performance include kernel (kernel function), C (penalty term), epsilon (tolerance), and gamma (affecting the width of the kernel function). Similar to Random Forest Regression, common values for these parameters are selected to form a parameter grid, picking the best parameter combination through exhaustive methods. SVR does not involve the concept of "random" features like Random Forest, so the optimal parameter set is obtained through one training: C is set to 10, epsilon is set to 1, gamma is set to 0.1, and the kernel is set to sigmoid.

The training results for the training set and test set are shown in Table 3. It can be observed that the training performance of SVR on the training set composed of these 18 feature variables is not very ideal compared to RR, with an average of 1.18 typhoon landing frequency prediction errors according to MAE, which is higher than RR. However, the test set fitting effect is good, with the values of three metrics lower than RR. Combining Figure 7, SVR's prediction values on the training set align well with the overall trend of the actual values, but the fitting performance for the peak and trough values is relatively poor, even showing completely opposite predicted values to the actual values in some years. The test set results highlighted with a deep green background show suboptimal trend and extreme value performance, with a maximum prediction error of 2.29 for the number of landing typhoons.

Table 3. The Results of SVR Training Evaluation Metrics

Type of dataset	Evaluation indicators	Values
Training set	MSE	1.9894
	RMSE	1.4105
	MAE	1.1840
Test set	MSE	1.5054
	RMSE	1.2269
	MAE	1.0259

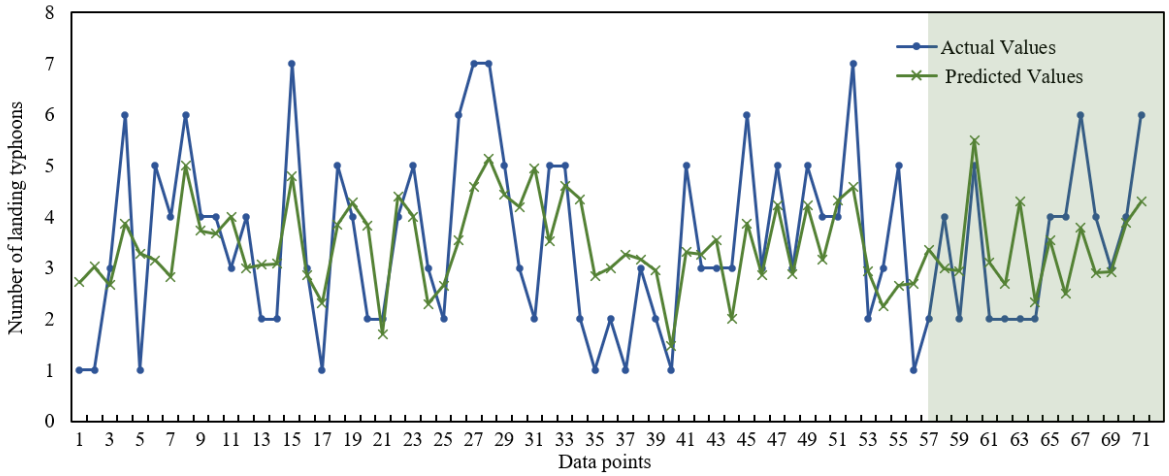


Figure 7. SVR Training Results Plot for Training and Test Sets (Highlighted in Deep Green Background for Test Set Training Results)

Comparison of Model Performance after Simple Secondary Selection

Usually, multicollinearity among feature variables can lead to model instability and excessive sensitivity to input variables. It is necessary to adopt methods to select optimal feature variables from a pre-selected set to enhance model performance and interpretability. Based on the idea of Rong et al., the author's team proposes a new approach: conducting preliminary selection of feature variable dataset by correlation analysis and secondary selection based on importance obtained through the RR method mentioned earlier to simplify the number of feature variables in the dataset, and then performing regression training based on two machine learning models. Table 4 below shows the 8 feature variables with importance values exceeding 0.05, which will constitute a new feature variable dataset for training both models.

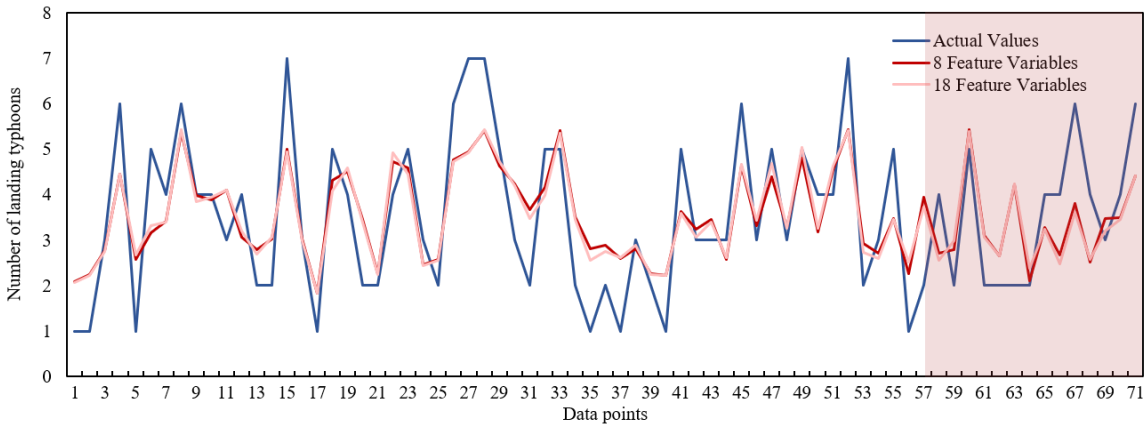
Based on this idea, the author's team does a preliminary exploration The training results of the two models are shown in Table 5. Compared to the dataset with 18 feature variables, the values of MSE, RMSE, and MAE on the test set all decrease, indicating that the dataset with 8 feature variables shows some improvement in the RR model's fitting performance on the test set. However, in the training set, the simplified dataset has slightly worse model fitting performance than the original dataset. For the SVR model, the simplified dataset does not improve its fitting performance. Comparing the actual and predicted values of the two models on the training and test sets with different feature variable datasets, the simplified dataset does not play much role in two model fitting performances.

Table 4. The Eight Feature Variables with Importance Values Exceeding 0.05

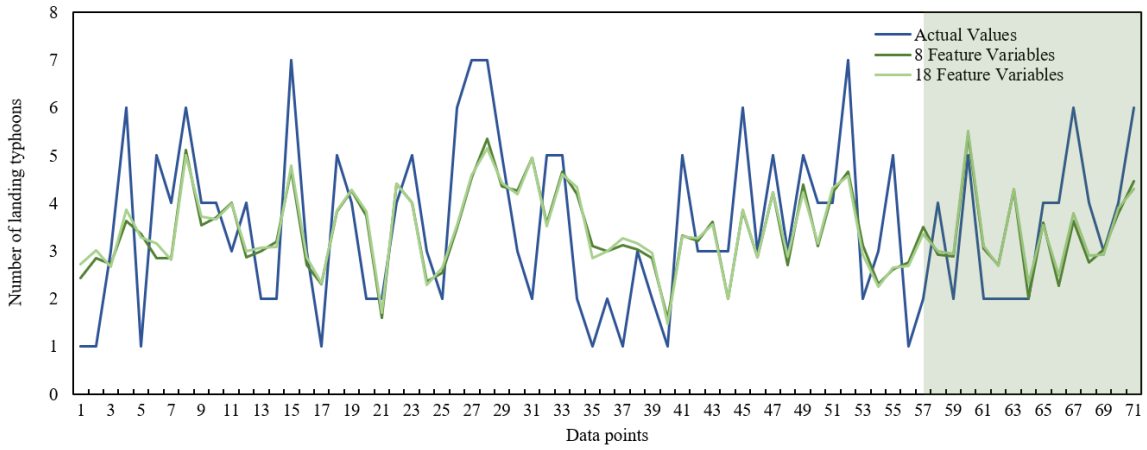
Number	Feature variable identifier	Feature variable name	Importance value of the characteristic indicator
1	x1902	Scandinavia Pattern , SCA in May of the same year	0.2097
2	x411	Antarctic Oscillation, AAO in April of the previous year	0.1159
3	x1580	Tropical Northern Atlantic SST Index in February of the same year	0.1154
4	x2573	Tibet Plateau Region 1 Index in Novermber of the same year	0.0899
5	x2337	Northern Hemisphere Polar Vortex Central Latitude Index in September of the same year	0.0540
6	x2152	Western Hemisphere Warm Pool Index in July of the same year	0.0530
7	x1694	Tropical Northern Atlantic SST Index in March of the same year	0.0521
8	x2734	Indian Ocean Basin-Wide Index in December of the same year	0.0511

Table 5. The Training Evaluation Results of the Two Models

Type of dataset	Evaluation indicators	RR		SVR	
		18 feature variables	8 feature variables	18 feature variables	8 feature variables
Training set	MSE	1.0857 ± 0.0152	1.1033 ± 0.0163	1.9894	2.0137
	RMSE	1.0419 ± 0.0073	1.0504 ± 0.0078	1.4105	1.4191
	MAE	0.8758 ± 0.0096	0.8832 ± 0.0060	1.1840	1.1913
Test set	MSE	1.8042 ± 0.0511	1.7534 ± 0.0382	1.5054	1.5912
	RMSE	1.3431 ± 0.0192	1.3241 ± 0.0144	1.2269	1.2614
	MAE	1.1669 ± 0.0178	1.1426 ± 0.0139	1.0259	1.0301



(a) RR



(b) SVR

Figure 8. The Training Results of the Two Models on the Training Set and Test Set Based on Two Datasets (Highlighted in Background for Test Set Training Results)

The results show that the simplification method described above applied to the feature variable dataset does slightly improve the model fitting performance, but the improvement is limited. This suggests that the combined indicator selection approach of initial screening based on relevance and secondary selection based on importance has some potential, but its utility needs further exploration.

LIMITATIONS AND FUTURE WORK

In the selection of machine learning prediction models, due to time and space constraints, this paper only utilizes common machine learning regression models-Random Forest Regression and Support Vector Regression. However, there are many different regression models in machine learning, and for a more comprehensive and refined application to train and predict the landing typhoon frequency in Guangdong Province, other machine learning methods and even deep learning or large model learning methods can be considered as future research methods. By comparing the training results, the optimal method for predicting the landing typhoon frequency in Guangdong Province may be identified. After continuous optimization of the model, further practice can be considered - translating research findings into an information system. This system would display the predicted typhoon landing frequency for each region based on input atmospheric and sea temperature indices. We briefly conceptualized the initial interface and functionality of the system as shown in the figure below (the image results displayed are not accurate). The development of this visualization system can help for typhoon crisis management.

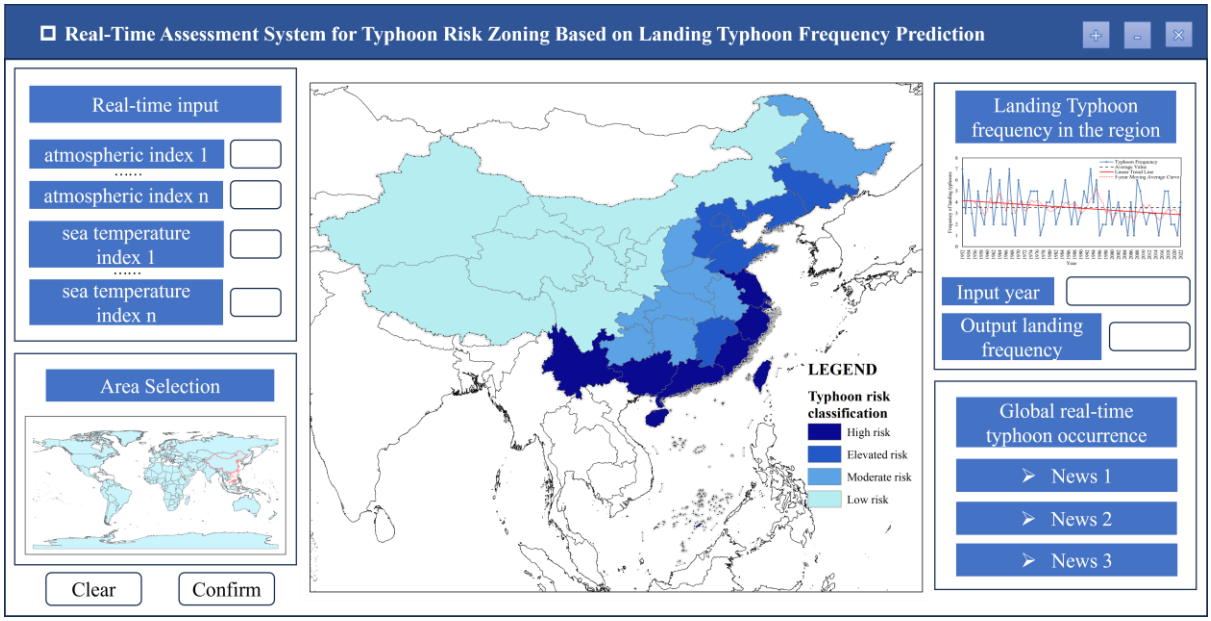


Figure 9. The Initial Concepts for the System Interface

CONCLUSION

Tropical cyclones are one of the destructive and globally impactful natural disasters. With the increasingly severe global climate change, the frequency of extreme weather events is rising, and tropical cyclone disasters have become a significant global concern. This study is based on collected data of the best tracks of typhoons from 1952 to 2022 and monthly atmospheric circulation and sea surface temperature indices from 1951 to 2022, constructing target variable and feature variable datasets separately. These datasets are used for training Random Forest Regression (RR) and Support Vector Regression (SVR) models, allowing a comparison of the fitting performance of the two machine learning models.

The research findings show that the frequency of typhoons making landfall in Guangdong Province has shown a decreasing trend from 1952 to 2022, particularly entering an interdecadal period of scarcity began at the end of the 20th century. Based on the training results of RR and SVR models on the dataset with 18 feature variables, RR demonstrates better fitting performance on the training set compared to SVR, while the opposite is observed on the test set. Both models exhibit good fitting trends on the training set, but RR shows higher sensitivity to peaks and valleys than SVR. On the test set, both models demonstrate relatively less-than-ideal fitting in terms of trends and extreme values. Generally speaking, the values of MSE, RMSE and MAE are small, so it can be seen that the prediction of the number of landing typhoons in this paper is accurate, basically able to predict the exact number of landing typhoon, with a possible error of no more than 2 typhoons. Additionally, analysis of the importance of the feature variables by RR model shows that SCA in May and Tropical Northern Atlantic SST Index in February of the same year and AAO in April of the previous year have a relatively important impact on the number of typhoon landing in Guangdong Province in a year. After the second screening and simplification of the dataset using the RR method, there is not a significant improvement in the fitting performance of the two models. The RR model shows some improvement, but it has the opposite effect on the SVR model, leading to a decrease in performance. Therefore, the combined indicator selection approach of initial screening based on relevance and secondary selection based on importance has some potential, which needs further exploration.

The research process and related results provide some references and guidance for further research and prediction of landing typhoon frequency and other characteristics of typhoons. This paper also helps the government departments to assess the risk of typhoons in the region and to improve the capabilities of emergency response management to disaster risks.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (Grant No. 72304038), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022A1515110049), Young Innovative Talents Project from Department of Education of Guangdong Province (Grant No. 2022KQNCX157).

DATA AVAILABILITY STATEMENT

All the data used in this article are provided with references and links at the manuscript.

REFERENCES

Academy of Disaster Reduction and Emergency Management, National Disaster Reduction Centre of China, International Federation of Red Cross and Red Crescent Societies (IFRC), Beijing Normal University. (2023). 2022 Global natural disaster assessment report. https://www.preventionweb.net/publication/2022-global-natural-disaster-assessment-report

Liu, H., & Liu, R. (2012). Tropical Cyclones Activities in the Western North Pacific from 1990 through 2009. Resources Science, *34*(2), 242-247.

Su, H., Yuan, L., Wang, M., Dong, G., & Fei, X. (2021). Characteristics and disaster analysis of tropical cyclones landing in China in 1949-2019. *Journal of Applied Oceanography*, *40*(3), 382-387.

Zhang, L., Liu, M., Quan, R., Lu, M., Wang, J., & Xu, S. (2009). Characteristics and disaster valuation of the tropical cyclones in southeast coastal areas of China. *Journal of East China Normal University (Natural Science)*, *2*, 41-49.

Chou, J. M., Ban, J. H., Dong, W. J., Hu, C., & Dai, R. (2018). Characteristics analysis and assessment of economic damages caused by tropical cyclones in Guangdong Province. *Chinese Journal of Atmospheric Sciences*, *42*(2), 357-366.

Jia, X., Chen, L., & Luo, J. (2014). Climate Prediction Experiment for Tropical Cyclone Genesis Frequency Using the Large-Scale Circulation Forecast by a Coupled Global Circulation Model. *Journal of Tropical Meteorology*, *20*(2), 103–111.

Hai, Y., & Chen, G. H. (2019). Prediction of frequency of tropical cyclones forming over the Western North Pacific using an artificial neural network model. *Clim Environ Res*, *24*(3), 324-332.

Murakami, H., & Wang, B. (2010). Future change of North Atlantic tropical cyclone tracks: Projection by a 20-km-mesh global atmospheric model. *Journal of Climate*, *23*(10), 2699-2721. https://doi.org/10.1175/2010JCLI3338.1

Hsan, T. Z., & Sein, M. M. (2021). Combining Support Vector Machine and Polynomial Regressing to Predict Tropical Cyclone Track. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)* (pp. 220-221). IEEE. https://doi.org/10.1109/LifeTech52111.2021.9391780

Wahiduzzaman, M., Oliver, E. C., Wotherspoon, S. J., & Holbrook, N. J. (2017). A climatological model of North Indian Ocean tropical cyclone genesis, tracks and landfall. *Climate Dynamics*, *49*, 2585-2603. https://doi.org/10.1007/s00382-016-3461-4

Kumar, S., Biswas, K., & Pandey, A. K. (2021). Predicting landfall's location and time of a tropical cyclone using reanalysis data. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part IV 30* (pp. 372-383). Springer International Publishing. https://doi.org/10.1007/978-3-030-86380-7_30

Kotal, S. D., & Bhattacharya, S. K. (2020). Improvement of wind field forecasts for tropical cyclones over the North Indian Ocean. *Tropical Cyclone Research and Review*, *9*(1), 53-66. https://doi.org/10.1016/j.tcrr.2020.03.004

Chen, J., & Chavas, D. R. (2023). A Model for the Tropical Cyclone Wind Field Response to Idealized Landfall. *Journal of the Atmospheric Sciences*, *80*(4), 1163-1176. https://doi.org/10.1175/JAS-D-22-0156.1

Matyas, C. J. (2010). Associations between the size of hurricane rain fields at landfall and their surrounding environments. *Meteorology and atmospheric physics*, *106*, 135-148. https://doi.org/10.1007/s00703-009-0056-1

Yang, Y. H., Ying, M., & Chen, B. D. (2009). The climatic changes of landfall tropical cyclones in China over the past 58 years. *Acta Meteorol Sin*, *67*(5), 689-696.

Nie, X., Tan, H., Cai, R., & Gao X. (2023). Projection of the tropical cyclones landing in China in the future based on regional climate model. *Advances in Climate Change Research*, *19*(1), 23. https://doi.org/10.12006/j.issn.1673-1719.2022.064

Richman, M. B., & Leslie, L. M. (2012). Adaptive machine learning approaches to seasonal prediction of tropical cyclones. *Procedia Computer Science*, *12*, 276-281. https://doi.org/10.1016/j.procs.2012.09.069

Tan, J., Liu, H., Li, M., & Wang, J. (2018). A prediction scheme of tropical cyclone frequency based on lasso and random forest. *Theoretical and Applied Climatology*, *133*, 973-983. https://doi.org/10.1007/s00704-017-2233-3

Chen, Y. (2022). Research on Prediction of Typhoon Frequency in theNorthwest Pacific Based on Multi-model Integration (Master's thesis). Central South University. https://doi.org/10.27661/d.cnki.gzhnu.2022.001080

Chen, R., Zhang, W., & Wang, X. (2020). Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere*, *11*(7), 676. https://doi.org/10.3390/atmos11070676

China Meteorological Administration Tropical Cyclone Data Center. 1952-2022. CMA Best Track Dataset. https://tcdata.typhoon.org.cn/zjljsjj.html

Ying, M., Zhang, W., Yu, H., Lu, X., Feng, J., Fan, Y., Zhu, Y., & Chen, D. (2014). An overview of the China Meteorological Administration tropical cyclone database. *Journal of Atmospheric and Oceanic Technology*, *31*(2), 287-301. https://doi.org/10.1175/JTECH-D-12-00119.1

Lu, X., Yu, H., Ying, M., Zhao, B., Zhang, S., Lin, L., Bai, L., & Wan, R. (2021). Western North Pacific tropical cyclone database created by the China Meteorological Administration. *Advances in Atmospheric Sciences*, *38*, 690-699. https://doi.org/10.1007/s00376-020-0211-7

National Climate Center. 1951-2022. A Collection of 100 Climate System Indices. http://cmdp.ncc-cma.net/Monitoring/cn_index_130.php

Rong, X., Qin, W., & Wei, W. (2023). Typhoon Number Prediction Based on Three Machine Learning Algorithms. *Marine Forecasts*, *40*(5), 1-9.