

PREDICTION AND ANALYSIS OF TROPICAL CYCLONES LANDFALL  
POINTS BASED ON RANDOM FOREST

WANG TONG

UNIVERSITI TEKNOLOGI MALAYSIA



**UNIVERSITI TEKNOLOGI MALAYSIA  
DECLARATION OF THESIS**

Author's full name : WANG TONG

Student's Matric No. : MCS241052 Academic Session : Session 1

Date of Birth : 2001.0212 UTM Email : tong20@graduate.utm.my

Thesis Title : PREDICTION AND ANALYSIS OF TROPICAL CYCLONES LANDFALL POINTS BASED ON RANDOM FOREST

I declare that this **Choose an item.** is classified as:

☒

**OPEN ACCESS**

I agree that my report to be published as a hard copy or made available through online open access.

☐

**RESTRICTED**

Contains restricted information as specified by the organization/institution where research was done.  
(The library will block access for up to three (3) years)

☐

**CONFIDENTIAL**

Contains confidential information as specified in the Official Secret Act 1972)

*(If none of the options are selected, the first option will be chosen by default)*

I acknowledged the intellectual property in the thesis belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this thesis for academic exchange.

Signature of Student:

Signature :

Full Name

Date :

Approved by Supervisor(s)

Signature of Supervisor I:

Signature of Supervisor II

Full Name of Supervisor I

Full Name of Supervisor II

Date :

Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

## ABSTRACT

This research endeavors to establish a prediction model for tropical cyclone landing points grounded in the random forest regression model, with a particular emphasis on cyclone activities within the Northwest Pacific region. The research data is sourced from the tropical cyclone trajectory information in the IBTrACS dataset spanning from 2020 to 2025. The model utilizes wind speed and air pressure as the primary input features to forecast the latitude and longitude of the tropical cyclone landing points.

In the course of this study, exploratory data analysis (EDA) was initially carried out. Subsequently, feature selection, model training, testing, and result evaluation were completed. After data cleaning, a total of 14,908 valid samples were acquired and partitioned into a training set and a testing set at a ratio of 8:2. The experimental findings indicate that the random forest model exhibits excellent performance in addressing the nonlinear relationships and outliers among variables. The average absolute errors of the model in the tasks of latitude and longitude prediction are 5.51 and 10.75, respectively. The stability and predictive capabilities of the model were further corroborated through error distribution plots and scatter plots.

This research not only offers an effective means of enhancing the prediction accuracy of cyclone landing points but also showcases the latent value of data-driven methodologies in the construction of coastal disaster prevention, mitigation, and early warning systems.

*Key words: Random Forest Regression; Meteorological Data Analysis; Machine Learning*

## **TABLE OF CONTENTS**

### **DECLARATION**

### **ABSTRACT**

### **CHAPTER 1 INTRODUCTION**

- 1.1 Research gap
- 1.2 Problem Background
- 1.3 Research Questions
- 1.4 Research Objectives
- 1.5 Research Scope
- 1.6 Research Contribution
- 1.7 Thesis Organization

### **CHAPTER 2 LITERATURE REVIEW**

- 2.1 Overview
- 2.2 Existing model framework
  - 2.2.1 Statistical and Empirical Models
  - 2.2.2 Numerical Weather Prediction Models
  - 2.2.3 Machine Learning Models
  - 2.2.4 Model Comparison Summary
- 2.3 IBTrACS Dataset
- 2.4 Research Gap and Positioning of This Study
  - 2.4.1 Research Gaps
  - 2.4.2 Positioning of This Study

## **CHAPTER 3      RESEARCH METHODOLOGY**

- 3.1    Introduction
- 3.2    Model Overview: Random Forest
- 3.3    Data Sources
- 3.4    Data Preprocessing
- 3.5    Model Structure and Justification
- 3.6    Summary

## **CHAPTER 4      RESULTS AND ANALYSES OF EXPERIMENTS**

- 4.1    Introduction
- 4.2    Exploratory Data Analysis and Feature Analysis
- 4.3    Model Construction and Experimental Design
- 4.4    Prediction Results and Error Analysis
- 4.5    Result visualization and residual analysis
  - 4.5.1    Distribution of prediction errors
  - 4.5.2    True vs Predicted Latitude Fitting Plot
  - 4.5.3    True vs Predicted Longitude Fitting Plot
  - 4.5.4    Conclusion
- 4.6    Summary

## **CHAPTER 5      CONCLUSION AND FUTURE WORK**

- 5.1    Research Conclusion
- 5.2    Limitations of the Study
- 5.3    Future Research Directions
- 5.4    Summary

## **REFERENCES**

# CHAPTER 1

## INTRODUCTION

### 1.1 Research gap

At present, the prediction models concerning the Malaysian region are scarce. Currently, most studies mainly focus on the Northwest Pacific, while the models for the sea areas around Malaysia ( $0^{\circ}$ – $10^{\circ}$ N,  $95^{\circ}$ – $120^{\circ}$ E) are nearly nonexistent, and the research in this region is also limited.

Meanwhile, traditional prediction methods have limitations. Traditional numerical models are computationally complex and more sensitive to initial conditions, failing to achieve real-time early warning.

Finally, there is insufficient modeling. Traditional statistical methods such as linear regression are unable to capture the complex nonlinear relationship between typhoon paths and environmental variables (such as sea temperature gradients and terrain obstruction).

### 1.2 Problem Background

*Tropical cyclones (TCs) are regarded as extreme weather events, along with gales, rainstorms, and storm surges, which can cause huge losses in coastal areas worldwide.( Chen, R., Zhang, W., & Wang, X. 2020)* It will exert a considerable

influence on the residents' housing, people's property, urban construction, road traffic and other economic constructions in coastal areas. In Southeast Asia, particularly in the surrounding sea areas of Malaysia, the generation and trajectory of tropical cyclones are characterized by complexity and uncertainty, which renders predicting the landing point of tropical cyclones an extremely challenging task. *The prediction of the intensity, location and time of the landfall of a tropical cyclone well advance in time and with high accuracy can reduce human and material loss immensely.* (Kumar, S., Biswas, K., & Pandey, A. K. 2021)

Traditional tropical cyclone trajectory prediction primarily depends on conventional numerical models (e.g., WRF, ECMWF), yet their computational complexity is considerable and they are highly sensitive to initial conditions, thereby failing to satisfy the timeliness demands of disaster emergency responses. Moreover, studies on tropical cyclones in the surrounding waters of Malaysia are scarce. The existing models mostly concentrate on the Northwest Pacific or the North Atlantic, leading to insufficient analysis of the crucial factors in the waters around Malaysia and inadequate consideration of the specific geographical and climatic conditions of the Malaysian sea area.

### **1.3 Research Questions**

How to construct a prediction model for the landing points of tropical cyclones in the waters adjacent to Malaysia based on the random forest algorithm and with the utilization of IBTrACS data (outputting longitude and latitude coordinates)?

Which factors affect the prediction of tropical cyclone landfall points in Malaysia?

## **1.4 Research Objectives**

Construct a random forest model: Based on the tropical cyclone trajectories and environmental variables in the IBTrACS dataset, develop a prediction model for tropical cyclone landing points in the surrounding sea areas of Malaysia (with the output of longitude and latitude coordinates).

Analysis of Key Influencing Factors: Through feature importance analysis and nonlinear modeling, quantify the dynamic influence of sea surface temperature, topography and atmospheric conditions on the landing locations of tropical cyclones.

## **1.5 Research Scope**

1、 In terms of the geographical range, the tropical cyclone data from the surrounding sea areas of Malaysia (latitude and longitude range:  $0^{\circ}$  -  $10^{\circ}\text{N}$ ,  $95^{\circ}$  -  $120^{\circ}\text{E}$ ) are employed, with the tropical cyclone data from other regions excluded.

2、 Regarding the definition of landfall, a tropical cyclone is regarded as having landed if the distance between its center and the coastline of Malaysia is no more than 50 kilometers. The potential landfall points on the east coast (the side facing the South China Sea) are of key concern.

3、 Concerning the prediction duration, with the tropical cyclone entering the study area (latitude  $0^{\circ}$ – $20^{\circ}\text{N}$ , longitude  $100^{\circ}$ – $120^{\circ}\text{E}$ ) as the starting point, the landing point within the next 72 hours is predicted.



4、Employing the relevant data in the IBTrACS dataset, a tropical cyclone landing point prediction model is established.

5、The impacts of key factors, such as climate and terrain, on the landing points of tropical cyclones will be analyzed.

<b>Geographical Range</b>	Malaysian waters (0° - 10°N, 95° - 120°E).
<b>Landfall Definition</b>	Tropical cyclones landing within 50 km of the Malaysian coastline
<b>Forecast time</b>	Landfall predictions up to 72 hours in advance
<b>Dataset</b>	IBTrACS dataset <a href="https://www.ncei.noaa.gov/products/international-best-track-archive">https://www.ncei.noaa.gov/products/international-best-track-archive</a>
<b>Tools</b>	Python programming language with Anaconda environment and IDEs like Jupyter Notebook

## 1.6 Research Contribution

**Theoretical Contribution:** It augments the studies on tropical cyclones in the Malaysian context.

**Methodological Contribution:** A novel model of the random forest algorithm for tropical cyclone prediction has been constructed.

**Enhancing disaster early warning capabilities:** The model can assist the Meteorological Department of Malaysia in optimizing the prediction of tropical cyclone landing points, locking high-risk areas 48 hours in advance, and reducing economic losses in coastal communities due to heavy rainfall and tropical cyclone landings.

**Data/Tool Contribution:** It provides a reusable dataset and model framework for future studies.

## 1.7 Thesis Organization

**Chapter1: Introduction.** The background and significance of tropical cyclone landfall points, the methods hypothesized in this article, and the issues to be addressed, etc.

**Chapter2: Literature Review.** An analysis of current prediction methods and their deficiencies, as well as the assistance offered by the literature to this paper.

**Chapter3: Construction of Random Forest Model.** Comprising dataset screening, data selection, data preprocessing, etc., and detailed descriptions of the construction of the random forest model.

**Chapter4: Analysis of Influencing Factors.** An analysis of the influence of various climatic factors on landfall points.

**Chapter5: Results, Discussion and Conclusion.** Perform visualization of the results, expound the insights on the significance and limitations of the obtained results, summarize and present suggestions for subsequent research.

## **CHAPTER 2**

### **LITERATURE REVIEW**

#### **2.1 Overview**

Tropical cyclones (TCs) are regarded as extreme weather events, along with gales, rainstorms, and storm surges, which can cause huge losses in coastal areas worldwide.( Chen, R., Zhang, W., & Wang, X. 2020) It will exert a considerable influence on the residents' housing, people's property, urban construction, road traffic and other economic constructions in coastal areas. In Southeast Asia, particularly in the surrounding sea areas of Malaysia, the generation and trajectory of tropical cyclones are characterized by complexity and uncertainty, which renders predicting the landing point of tropical cyclones an extremely challenging task.

For example:

- In 2017, Typhoon Hato, although it did not make a direct landfall, its peripheral circulation caused floods in many areas along the eastern coast of Malaysia.
- In 2021, Tropical Depression Haitang brought continuous heavy rain to Johor, forcing the evacuation of tens of thousands of people.

The prediction of the intensity, location and time of the landfall of a tropical cyclone well advance in time and with high accuracy can reduce human and material loss immensely.( Kumar, S., Biswas, K., & Pandey, A. K. 2021)

In recent years, machine learning has shown great potential in the field of meteorological prediction. Machine learning, which is data-driven, can capture the complex non-linear relationships between input variables and prediction results (Shah et al., 2020).

In the area of meteorological prediction, machine learning has been extensively applied to the prediction of cyclone intensity, rainfall amount, etc. (Kordmahalleh et al., 2016; Alam et al., 2020). Algorithms such as Random Forest (RF), Support Vector Machine (SVM), and Long Short-Term Memory Network (LSTM) have exhibited relatively high prediction accuracy. Moreover, they feature short training time and low computational cost (Chattopadhyay et al., 2019; Kratzert et al., 2018).

## **2.2 Existing model framework**

### **2.2.1 Statistical and Empirical Models**

Statistical and empirical models are mathematical modeling approaches constructed based on observational data. These models predict future outcomes by relying on the statistical relationships among variables, without the need to consider the underlying physical mechanisms. Such models have found extensive applications in prediction tasks across various disciplines, including economics, biology, and climate science. Common statistical models encompass linear regression, time - series analysis, and classification methods grounded in historical data.

In the domain of tropical cyclone (TC) prediction, statistical and empirical models were among the first to be employed for forecasting the tracks and landfall locations of tropical cyclones. They predominantly depend on historical data and strive to establish predictive associations between historical cyclones and environmental variables. Among them, the most representative model is the CLIPER model (Climatology and Persistence Model) developed by the National Hurricane Center of the United States. This model predicts the future position of cyclones by integrating path persistence and climatological averages through a regression equation (Neumann, 1972).

CLIPER predicts the position at each future time point through polynomial regression:

$$P(t) = a_0 + a_1t + a_2t^2 + \dots + a_nt^n$$

- $P(t)$ : The predicted position (longitude or latitude) after time  $t$
- $a_i$ : Polynomial coefficients obtained by fitting historical path data
- Typically, separate models are constructed for longitude and latitude.

Likewise, linear regression models have been extensively utilized to predict cyclone intensity or landfall locations. The input variables typically consist of sea surface temperature, wind direction, pressure gradient, etc.

This is the most frequently used statistical prediction model, and numerous cyclone track/intensity predictions are predicated on it:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

- $Y$ : predicted value (such as the future longitude, intensity, wind speed, etc. of a cyclone)

- $X_1, X_2, \dots, X_n$ : Input variables (such as SST, air pressure, wind speed, and other environmental variables)

- $\beta_0$ : Intercept term

- $\beta_i$ : Regression coefficient, reflecting the weight of each variable

- $\varepsilon$  : Error term

For example, Dong and Wang (2018) developed a statistical landfall model for the South China Sea region, leveraging historical cyclone track data for model construction. Their study demonstrated that under specific climatic conditions, statistical models can yield relatively promising prediction outcomes. Nevertheless, in coastal areas with complex topography, there remain issues of insufficient prediction of landfall point variations.

Statistical models possess several advantages, such as high computational speed, low requirements for computing resources, and strong interpretability. However, the modeling process of these models is relatively simplistic. As a result, their accuracy is constrained when addressing regions with intricate terrain or substantial weather fluctuations. Most statistical models are based on assumptions of linearity and stability, which are inconsistent with the non - linear and dynamically evolving characteristics of tropical cyclones. In Southeast Asian regions, including the waters of Malaysia, these limitations considerably reduce the reliability of statistical models as standalone prediction tools (Knaff et al., 2003).

### **2.2.2 Numerical Weather Prediction Models**

The Numerical Weather Prediction (NWP) model, a simulation system grounded in physical laws, utilizes a set of mathematical equations depicting processes including atmospheric fluid dynamics, thermodynamics, and radiation transfer to reproduce and prognosticate atmospheric behavior.

Specifically, the NWP model partitions the atmosphere into a three - dimensional grid structure. At each grid point, it computes the temporal variations of variables such as air temperature, humidity, wind speed, and atmospheric pressure. This model serves as the bedrock for modern weather forecasting.

These simulations are principally formulated using three major categories of core equations: the momentum equation, the thermodynamic equation, and the mass conservation equation. Specifically, the simplified form of the momentum equation is presented as follows:

$$\frac{D\vec{v}}{Dt} = -\frac{1}{\rho}\nabla p + \vec{g} + \vec{F}$$

In this equation,  $\vec{v}$  denotes the wind - speed vector,  $\rho$  represents the air density,  $\nabla p$  signifies the pressure gradient,  $\vec{g}$  is the gravitational acceleration, and  $\vec{F}$  represents other external - force terms such as frictional force or Coriolis force.

The thermodynamic equation is employed to compute the temporal evolution of temperature:

$$\frac{DT}{Dt} = \frac{Q}{c_p}$$

In this context,  $T$  denotes the temperature,  $Q$  represents the heat absorbed per unit mass, and  $c_p$  signifies the specific heat capacity at constant pressure.

By numerically solving these systems of equations, the NWP model is capable of simulating the dynamic evolution of the atmosphere over time.

In the prediction of tropical cyclone landfall locations, NWP models have been extensively utilized to simulate the genesis, track, and intensity variations of cyclones. Commonly adopted global models include the ECMWF (European Centre for Medium - Range Weather Forecasts) and the GFS (Global Forecast System), and regional models such as the WRF (Weather Research and Forecasting Model) are also in frequent use. These models are capable of resolving mesoscale structures, accounting for the impact of terrain, and considering the air - sea interaction. For example, the WRF model has been successfully applied to simulate cyclone landfalls and storm surge scenarios in Southeast Asia (Pattnaik et al., 2017).



The primary strength of NWP models lies in the physical interpretability of their results, along with the provision of high - precision spatiotemporal resolution predictions. Instead of relying on past observational data for future predictions, they simulate potential future weather conditions based on physical principles. Nevertheless, these models impose extremely high requirements on computing resources, and their prediction outcomes are highly sensitive to initial input conditions, such as meteorological observation data. Additionally, due to the imperfections of physical schemes and the continuous accumulation of numerical errors, the accuracy of these models in long - term forecasting is constrained. In regions where the meteorological observation network is relatively sparse, such as Malaysia, the performance of NWP models is also somewhat affected (Chattopadhyay & Chandrasekar, 2021).

### **2.2.3 Machine Learning Models**

Machine Learning (ML) encompasses a set of algorithms that endow computers with the ability to discern patterns from data and make predictions, without the need for explicit programming or the construction of physical mechanism models. Distinct from numerical meteorological models or statistical models, ML is completely data - driven. Its objective is to extract patterns from historical data for the purpose of forecasting future states.

In supervised learning tasks, the model is trained using a set of input variables  $X$  (e.g., sea surface temperature, wind speed, humidity, etc.) and a target output variable  $Y$  (landfall point) to learn a predictive function:

$$\hat{Y} = f(X; \theta)$$

- $\hat{Y}$  : The output predicted by the model (e.g., the longitude and latitude coordinates of the landfall point).
- $X$  : The set of input features.
- $\theta$  : Model parameters (acquired through training and learning).
- $f$  : The prediction function, corresponding to different algorithms, such as Random Forest (RF), Support Vector Machine (SVM), Long Short - Term Memory network (LSTM), etc.

In recent years, machine learning models have yielded remarkable achievements in the realm of meteorological prediction.

Within the domain of tropical cyclone research, ML models are capable of addressing intricate nonlinear relationships and have been applied to a variety of tasks, including path prediction, intensity classification, and landfall location regression. For example, Shah et al. (2020) employed a random forest model to perform modeling on IBTrACS and ERA5 data, thereby achieving effective prediction of cyclone landfall points; Kratzert et al. (2018) utilized LSTM networks to model rainfall - runoff time series, highlighting the potential of machine learning in the modeling of climate time - series data; Alam et al. (2020) enhanced the prediction robustness by integrating multiple models.

Figure 2.3 illustrates the general procedure for using machine learning models in the prediction of tropical cyclone landfall points, encompassing the following six steps:

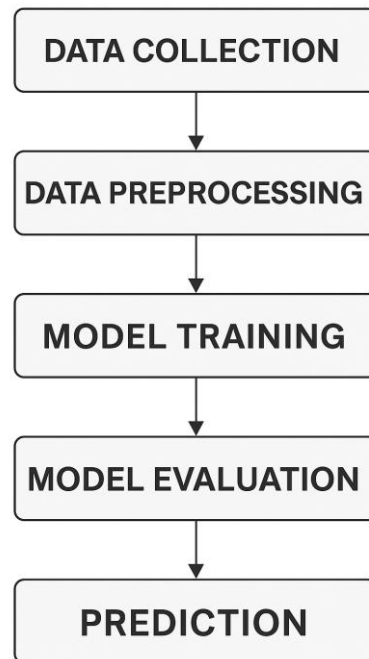


Figure 2.3

- **Data collection** (e.g., IBTrACS and ERA5 datasets)
- **Data preprocessing** (e.g., missing value imputation, coordinate alignment)
- **Feature extraction and construction** (e.g., Sea Surface Temperature (SST), wind shear, terrain elevation)

- **Model training** (e.g., Random Forest (RF), Support Vector Machine (SVM), Long Short - Term Memory (LSTM) models)

- **Model validation** (employing test sets, cross - validation techniques, etc.)

- **Prediction result output** (landfall point coordinates)

This flowchart vividly depicts the complete process from the original data to the generation of prediction results.

Machine learning models are capable of constructing models for nonlinear relationships among high - dimensional and intricate features. These models do not rely on explicit physical assumptions, exhibit rapid training speeds, and possess a certain degree of robustness against noise. Nevertheless, they are highly sensitive to the quality and quantity of training data. There may be issues of overfitting, and moreover, they often suffer from a lack of physical interpretability. In regions such as Malaysia, where the sample data is limited, the model must undergo rigorous validation to ensure its generalization capacity.

#### **2.2.4 Model Comparison Summary**

In the preceding chapters, we have reviewed three major categories of prediction model frameworks: statistical models, Numerical Weather Prediction (NWP) models, and machine learning models.

To gain a more profound and distinct understanding of their general applicability and disparities in the prediction of tropical cyclone landfall points, in this section, we will summarize the advantages and limitations of each type of model via a comparative analysis. As presented in Table 2.4,

Table 2.4

Comparison Dimension	Statistical Models	NWP Models	ML Models
Theoretical Basis	Statistical regression	Physical laws	Data-driven pattern learning
Physics-based	NO	YES	NO
Model Complexity	Low	High	Medium–High
Computational Cost	Low	Very High	Medium
Predictive Accuracy	Moderate	High (conditional)	High (data-dependent)
Interpretability	High	High	Low–Medium
Suitability for Malaysia	Limited	Medium(with data)	High(data adaptable)

As can be gleaned from the comparison presented in Table 2.4, while statistical models exhibit high computational efficiency and their results are amenable to straightforward interpretation, they encounter issues of insufficient accuracy when addressing regions with nonlinear and dynamically complex characteristics, such as Southeast Asia.

Numerical meteorological models possess physical precision; however, they are highly reliant on computational resources and observational data. On the other hand, machine learning models, despite not being founded on physical mechanisms,

demonstrate excellent adaptability and predictive prowess. This is particularly evident when environmental variables are incorporated.

Given the relatively sparse observational network in Malaysia and the imperative for rapid response in early warnings, it is a more judicious choice to employ machine learning models (e.g., random forest models) as the core models for this research.

### **2.3 IBTrACS Dataset**

IBTrACS (International Best Track Archive for Climate Stewardship), meticulously maintained by the National Oceanic and Atmospheric Administration (NOAA) of the United States, represents a comprehensive global dataset on tropical cyclones. This dataset amalgamates information from various international and regional meteorological organizations, including the China Meteorological Administration (CMA), the Joint Typhoon Warning Center (JTWC), and the Japan Meteorological Agency (JMA). Through this integration, IBTrACS provides a century - spanning record of cyclone activities, distinguished by its exceptional authority and spatio - temporal continuity.

The IBTrACS dataset encompasses the following pivotal variables:

- Cyclone identification number
- Timestamp (year, month, day, hour)

- Geographical coordinates (latitude and longitude) of the cyclone's center
- Maximum sustained wind velocity (knots)
- Central atmospheric pressure (hPa)
- Basin designation or storm classification

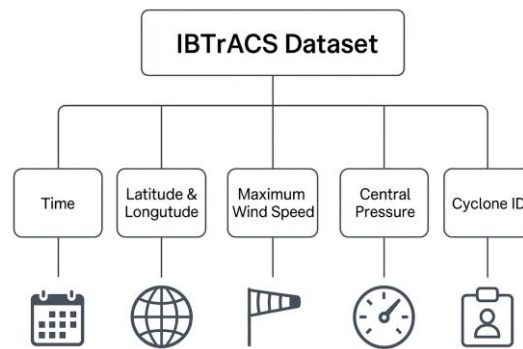


Figure 2.3.1

IBTrACS has garnered widespread application in the realms of global climate change analysis and the modeling of tropical cyclone predictions. Its standardized data structure is robust enough to facilitate multi-dimensional research encompassing cyclone trajectories, landfall locations, frequencies, and intensities. Notably, Knapp et al. (2010) underscored the pivotal role of IBTrACS in harmonizing cyclone records from diverse institutions; Knaff et al. (2014) leveraged this dataset to construct a satellite-based objective cyclone-scale climatology model; and Shah et al. (2020) harnessed IBTrACS data to develop machine learning-based models for predicting

cyclone paths and landfall points, thereby illustrating its multifaceted value in both conventional and contemporary prediction methodologies.

Within the scope of this study, we anticipate selecting cyclone records that have occurred in the South China Sea and in the vicinity of the Malaysian coastline since 1980. This dataset will be employed to construct the target variable, specifically the landfall point, within the context of machine learning models.

## **2.4 Research Gap and Positioning of This Study**

### **2.4.1 Research Gaps**

Despite the abundance of research on tropical cyclone prediction, several significant research gaps persist in Southeast Asia, particularly in Malaysia:

#### **1. Absence of Empirical Studies Focused on Malaysia**

While a plethora of global research exists on tropical cyclone prediction, empirical studies specifically targeting the prediction of landing points in Malaysia or the South China Sea region remain notably scarce.

#### **2. Underutilization of Machine Learning in Landing Prediction Tasks**

Although machine learning techniques have been applied to cyclone path and intensity prediction, their practical implementation in the prediction of precise landing



point coordinates is still limited, particularly in the context of utilizing straightforward models such as Random Forest (RF).

### 3. Exploration of Simple Machine Learning Methods in Data-Scarce Environments

Given that Malaysia is characterized by relatively sparse observational data, there is a notable absence of systematic and effective validation of simple yet interpretable machine learning models in such data-limited scenarios.

#### 2.4.2 Positioning of This Study

This study aims to address the aforementioned research gaps by developing a machine learning-based prediction model for tropical cyclone landfall points in the Malaysian region. Instead of designing an entirely new model structure, this study focuses on adapting and applying existing methods to the local context. The Random Forest algorithm, which is both easy to implement and highly interpretable, has been selected as the core model. By integrating historical cyclone data from the IBTrACS cyclone trajectory dataset with environmental characteristics derived from the ERA5 high-resolution meteorological variable dataset, the model predicts the latitude and longitude coordinates of landfall points. In addition to capturing the historical behavior of cyclones, the model also incorporates the dynamic environmental conditions surrounding each event.

The core algorithm employed in this study is Random Forest, which demonstrates strong capabilities in handling high-dimensional nonlinear relationships while maintaining stability even in scenarios with limited sample sizes.

This study highlights the feasibility and practical significance of leveraging global open-source datasets and interpretable machine learning methods in data-sparse regions such as Malaysia. The constructed prediction model, which outputs estimated landfall latitude and longitude coordinates, provides valuable support for enhancing Malaysia's disaster warning systems and disaster prevention strategies.

## **CHAPTER 3**

### **RESEARCH METHODOLOGY**

#### **3.1 Introduction**

This chapter will elaborate in detail on the methodology for establishing a prediction model of tropical cyclone landfall points. This research is centered on predicting the cyclone paths in the coastal regions of Malaysia. Given the high complexity and non - linearity inherent in the atmospheric system, traditional numerical physical models frequently encounter challenges in accurately predicting cyclone landfall locations.

Consequently, this study employs a data - driven approach grounded in the Random Forest (RF) algorithm for model construction. The overall methodological framework encompasses several crucial steps:

1. Comprehending the essence of the model employed.
2. Obtaining and pre - processing relevant meteorological datasets.

3. Formulating meaningful input features.

4. Training the model and assessing its performance.

One publicly accessible meteorological dataset is utilized in this research: the International Best Track Archive for Climate Stewardship (IBTrACS) global cyclone path dataset.

The primary objective of this chapter is to elucidate the process of converting raw meteorological data into model outputs capable of making predictions. Additionally, it will expound on the rationale behind the selection of the Random Forest model. In particular, its proficiency in dealing with non - linear relationships and its resilience against overfitting render the Random Forest model an optimal choice for spatial meteorological prediction tasks.

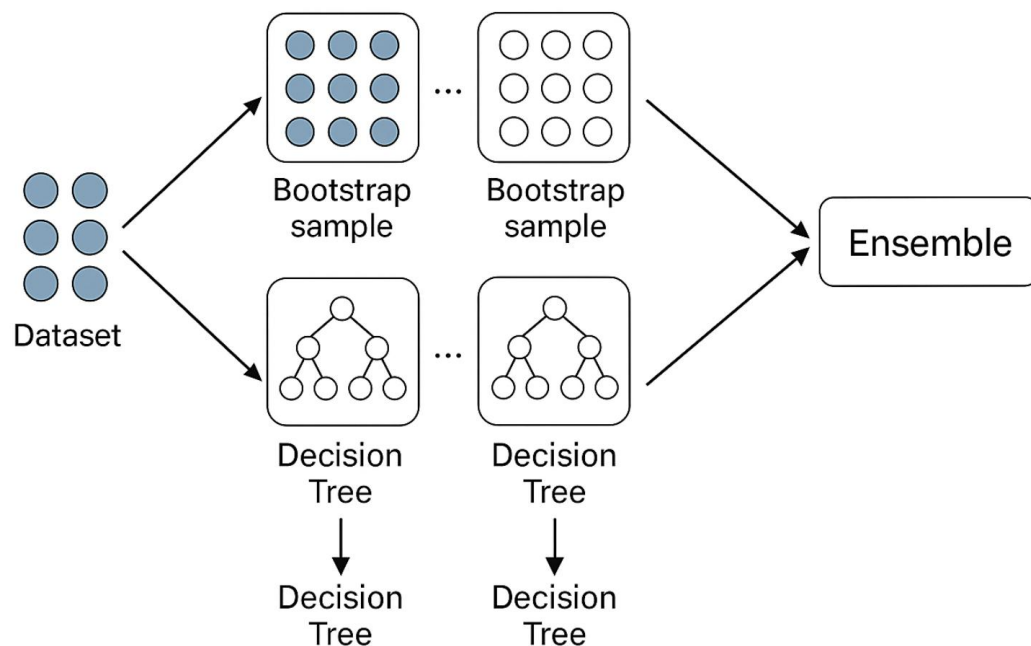
### **3.2 Model Overview: Random Forest**

Random Forest (RF) is a supervised machine learning algorithm and an ensemble learning method that constructs multiple decision trees during the training phase. For prediction, it aggregates the outputs of all sub-models by averaging (for regression tasks) or majority voting (for classification tasks).

In this study, we employ the Random Forest regression model to predict the landing points (latitude and longitude) of tropical cyclones. The Random Forest model demonstrates exceptional performance in handling nonlinear relationships between features, mitigating overfitting, and maintaining high predictive accuracy on medium and small-scale datasets, making it particularly suitable for this research.

This algorithm leverages a mechanism known as Bootstrap Aggregation (Bagging), where each decision tree is trained on a bootstrap sample of the dataset, and only a subset of features is considered at each split. This process enhances diversity among the trees, reduces overall model variance, and improves generalization capability.

## Bootstrap Aggregation (Bagging)



For regression tasks, the final prediction of the Random Forest is obtained by averaging the outputs of all individual decision trees. The mathematical formulation is presented as follows:

$$\hat{y}^{RF}(x) = \frac{1}{T} \sum_{i=1}^T f_i(x)$$

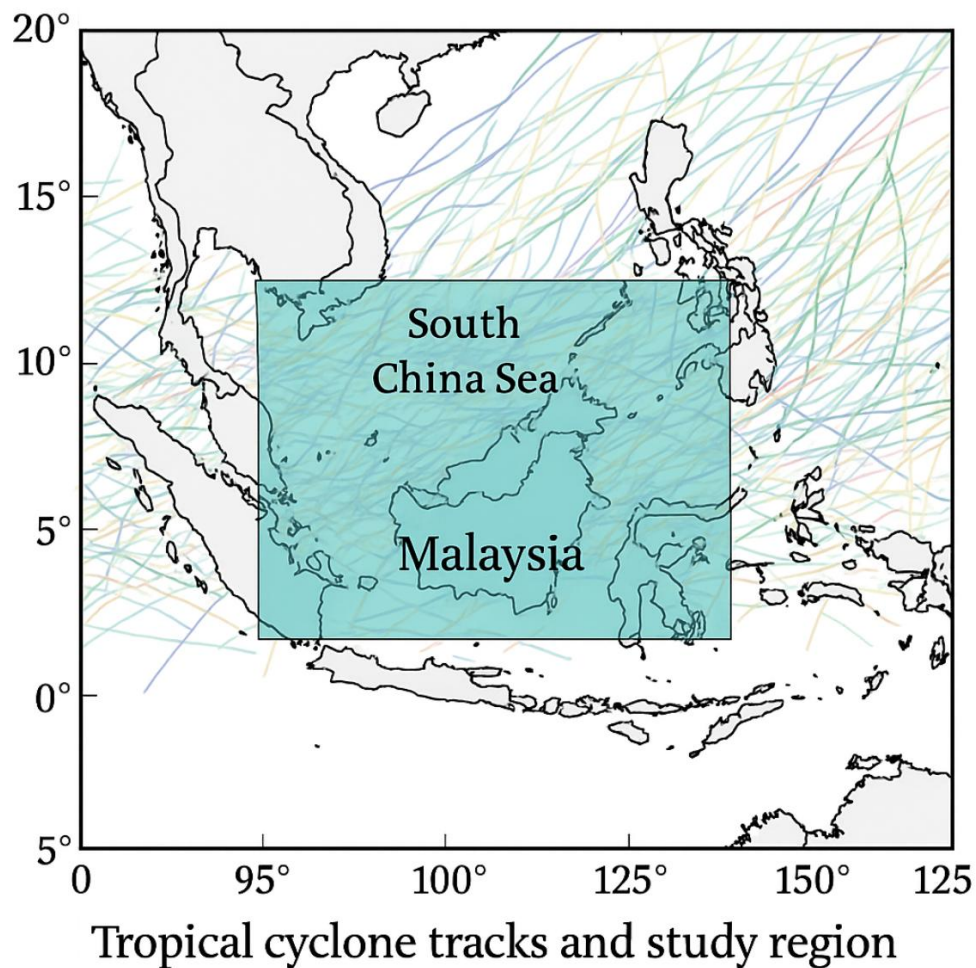
Among them:

- $T$  denotes the total number of constructed decision trees
  
- $f_i(x)$  indicates the prediction outcome of the  $i$  tree for input  $x$

The random forest model was selected because it has been extensively applied in meteorological prediction and geospatial modeling, achieving favorable results. In particular, under conditions where sample sizes are limited, feature dimensions are high, and data contain noise, the random forest model exhibits exceptional stability and robustness. Additionally, this algorithm can provide scores of feature importance, which aids in identifying the environmental variables that most significantly influence the prediction of tropical cyclone landfall locations.

### 3.3 Data Sources

This study utilized one primary publicly accessible meteorological dataset: the International Best Track Archive for Climate Stewardship (IBTrACS). The dataset is extensively employed in global meteorological research and is characterized by their high data quality and extensive spatiotemporal coverage.



IBTrACS is a globally comprehensive dataset of historical tropical cyclone tracks compiled and released by the National Centers for Environmental Information

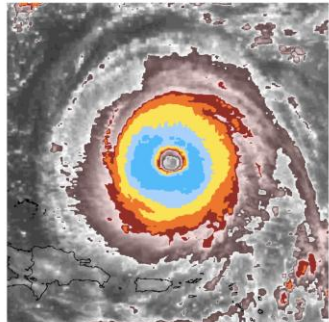
(NCEI), which operates under the U.S. National Oceanic and Atmospheric Administration (NOAA). This dataset includes critical parameters such as the time, latitude and longitude positions, central pressure, and maximum sustained wind speed of cyclones. For this study, we focus on the Northwest Pacific sub-region (ibtracs.WP) to analyze tropical cyclone activities affecting Malaysia and its surrounding areas. The dataset spans from 1980 to the present, with a temporal resolution of six hours.

[Home](#) [Products](#) [International Best Track Archive for Climate Stewardship \(IBTrACS\)](#)

## International Best Track Archive for Climate Stewardship (IBTrACS)

The International Best Track Archive for Climate Stewardship (IBTrACS) project is the most complete global collection of tropical cyclones available. It merges recent and historical tropical cyclone data from multiple agencies to create a unified, publicly available, best-track dataset that improves inter-agency comparisons. IBTrACS was developed collaboratively with all the World Meteorological Organization (WMO) Regional Specialized Meteorological Centres, as well as other organizations and individuals from around the world.

*To help the project receive continued support, updates, and improvement, tell us how you use IBTrACS data by completing our optional User Registration Form.*

[Optional User Registration](#)

### 3.4 Data Preprocessing

Before model training, a series of preprocessing steps must be performed on the IBTrACS dataset to ensure a consistent data structure and complete information, thereby making them suitable for supervised learning tasks. These steps help reduce noise and improve the reliability of the model inputs.



The IBTrACS dataset can be obtained from its official website (<https://www.ncei.noaa.gov/products/international-best-track-archive>).

A1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	SID	SEASON	NUMBER	BASIN	SUBBASIN NAME	ISO_TIME	NATURE	LAT	LON	WMO_WIP	WMO_PRE	WMO_AG	TRACK_TY	DIST2LAN	LANDFALL	I	FLAG	USA_AGEN	USA_ATCF	USA_LAT	USA_LON	USA_REC
2	Year							degrees_n	degrees_e	kts	mb			km	km					degrees_n	degrees_e	
3	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		16.5	124				main	165	145							
4	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		16.5	123.8				main	145	111							
5	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		16.6	123.5				main	111	77							
6	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		16.7	123.2				main	77	44							
7	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		16.8	122.9				main	44	33							
8	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		16.8	122.8				main	33	10							
9	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		16.9	122.6				main	10	0							
10	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		17	122.3				main	0	0							
11	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		17.2	122				main	0	0							
12	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		17.3	121.7				main	0	0							
13	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		17.4	121.6				main	0	0							
14	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		17.5	121.4				main	0	0							
15	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		17.6	121.1				main	0	0							
16	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		17.8	120.8				main	0	0							
17	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		17.9	120.5				main	0	0							
18	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		18	120.4				main	10	10							
19	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		18.1	120.2				main	31	31							
20	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		18.2	120				main	54	54							
21	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		18.4	119.7				main	90	90							
22	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		18.5	119.5				main	114	114							
23	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		18.6	119.4				main	128	128							
24	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		18.7	119.2				main	151	151							
25	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		19	118.9				main	191	191							
26	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		19.2	118.7				main	219	219							
27	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2NR		19.4	118.5				main	247	247							
28	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2TS		19.5	118.4				main	262	262							
29	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2TS		19.6	118.3				main	276	282							
30	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2TS		19.9	118.1				main	311	311							
31	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2TS		20.1	118				main	332	329							
32	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2TS		20.3	118				main	329	319							
33	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2TS		20.4	118				main	319	300							
34	1884177N	1884	14	WP	MM	UNNAMEE1884-06-2TS		20.6	118				main	299	277							

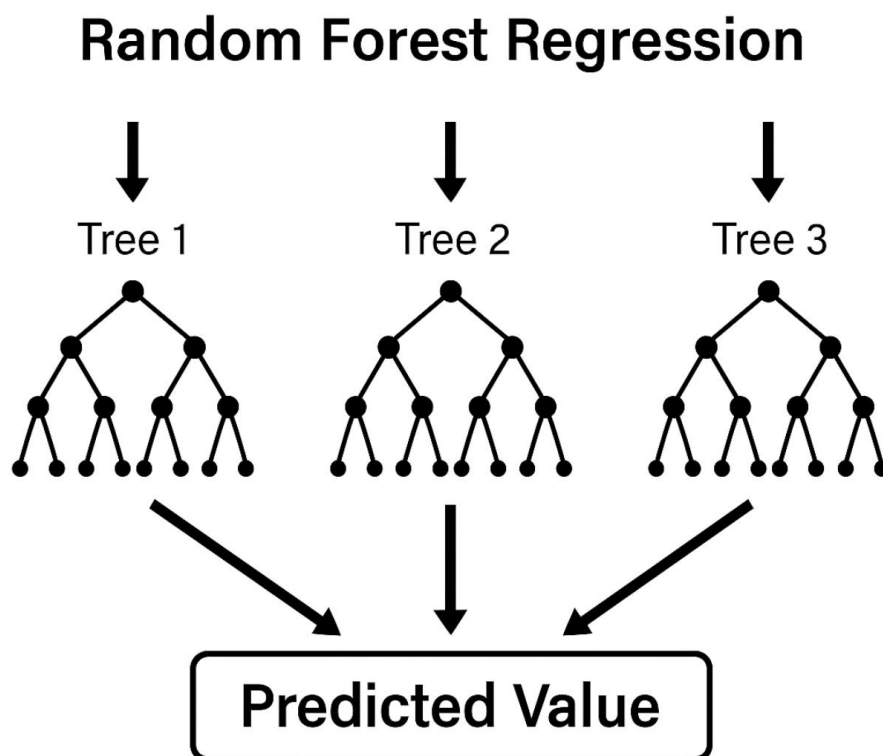
Given the substantial size of the IBTrACS dataset files and the inclusion of unnecessary data for our study, we will conduct a filtering process on the IBTrACS dataset in this step. Our filtering criteria are as follows:

- Time range: from 2000 to the present

- Geographical range: covering Malaysia and its surrounding seas (Longitude: approximately 95°E to 120°E; Latitude: approximately 0°N to 15°N)

### 3.5 Model Structure and Justification

The machine learning model selected for this study is the Random Forest Regression (RFR) algorithm. Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions through averaging, thereby reducing overfitting and enhancing the model's generalization ability. Each decision tree is trained on a random subset of both data samples and features, introducing randomness and diversity into the model. This structure enables Random Forest to effectively capture the complex nonlinear relationships between environmental variables (e.g., sea surface temperature, wind shear, humidity) and the landing locations of tropical cyclones (latitude and longitude).



The Random Forest model constructed in this study incorporates the following key structural parameters:

- Number of trees: Determines the total number of decision trees to be built. A higher number of trees generally improves performance but increases computational cost.

- Maximum tree depth: Controls the maximum depth of each tree. Shallower trees help mitigate overfitting by limiting the complexity of individual trees.

- Number of features to consider at each split: Restricts the number of features evaluated at each split, promoting diversity among trees and reducing correlation between them.

- Bootstrap sampling: Each tree is trained using a dataset sampled with replacement, ensuring variability in the training data for each tree.

Compared with other machine learning methods (e.g., Support Vector Regression, LSTM, Linear Regression), the Random Forest model was chosen for the following reasons:

1. Strong nonlinear modeling capability: The formation and trajectory of tropical cyclones are governed by the nonlinear interactions of various meteorological and oceanic factors. Random Forest excels at capturing such complex relationships due to its inherent flexibility.

2. Robustness against overfitting: As an ensemble method, Random Forest reduces variance by averaging predictions across multiple trees, enabling it to generalize well even in scenarios with limited data.

3. High fault tolerance: Random Forest is relatively insensitive to noise and missing data, making it well-suited for environmental datasets that often contain incomplete or noisy observations.

4. Interpretability of variable importance: Random Forest provides a mechanism to calculate feature importance, allowing researchers to identify which environmental variables most significantly influence the predicted landing locations of tropical cyclones. This enhances the interpretability and transparency of the model.

In conclusion, Random Forest is particularly well-suited for predicting tropical cyclone landing points in Malaysia and its surrounding regions, where the problem is characterized by high complexity and sparse data.

### **3.6 Summary**

This chapter presents the methodology employed in this study for predicting the landing points of tropical cyclones in the Malaysian region. First, it elaborates on the random forest regression algorithm chosen for this study, emphasizing its strengths in nonlinear modeling, stability, and interpretability for geographical space prediction problems.

Subsequently, the chapter details the primary data source utilized: the IBTrACS cyclone trajectory dataset. It also outlines the data acquisition and preprocessing procedures. Furthermore, the chapter examines the structural composition of the random forest model and provides justification for its selection, highlighting its effectiveness in addressing multi-variable, nonlinear, and incomplete data challenges.

The methodological framework established in this chapter serves as a foundation for the model training and evaluation presented in the subsequent chapter.

## CHAPTER 4

### RESULTS AND ANALYSES OF EXPERIMENTS

#### 4.1 Introduction

This chapter provides a comprehensive introduction to Exploratory Data Analysis (EDA), model training, and model performance evaluation using the International Best Track Archive for Climate Stewardship (IBTrACS) dataset.

Initially, a systematic data exploration and feature statistical analysis of the cyclone observation samples employed in this study are carried out. This endeavor visually depicts the distribution patterns of the samples across spatial, temporal dimensions, and major meteorological variables. Through preliminary descriptive statistics and visualization techniques, a solid data foundation is established for subsequent modeling tasks.

Subsequently, the machine learning modeling strategy adopted in this research is expounded in detail. This includes aspects such as feature selection, dataset partitioning, model type determination, parameter setting, and the overall modeling workflow. Special emphasis is placed on the applicability of the Random Forest Regression model and its advantages within the realm of meteorological forecasting.

During the model training and experimentation phase, the prediction performance of the model on the test set is meticulously analyzed. Error assessments are conducted using metrics such as the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). In addition, the limitations of the model and areas for improvement are discussed in conjunction with representative cases. Through the visualization of errors and residual analysis, a profound understanding of the model's generalization ability and practical application significance is achieved.

Ultimately, the key findings and conclusions derived from the experimental process are synthesized. This serves to provide theoretical and data support for future research endeavors and the optimization of relevant methodologies.

## **4.2 Exploratory Data Analysis and Feature Analysis**

To comprehensively and profoundly comprehend the structure and distribution patterns of tropical cyclone sample data, this section undertakes a systematic statistical analysis and visual exploration of the core characteristic variables.

First and foremost, Table 4-1 is employed to present the descriptive statistical results of the key variables. As can be observed, the total number of samples amounts to 14,908. The mean value of latitude (Latitude) is 21.75, the mean value of longitude (Longitude) is 135.02, the mean value of wind speed (Wind\_kts) is 59.90 knots, and the mean value of air pressure (Pressure\_mb) is 973.82 hectopascals. These data

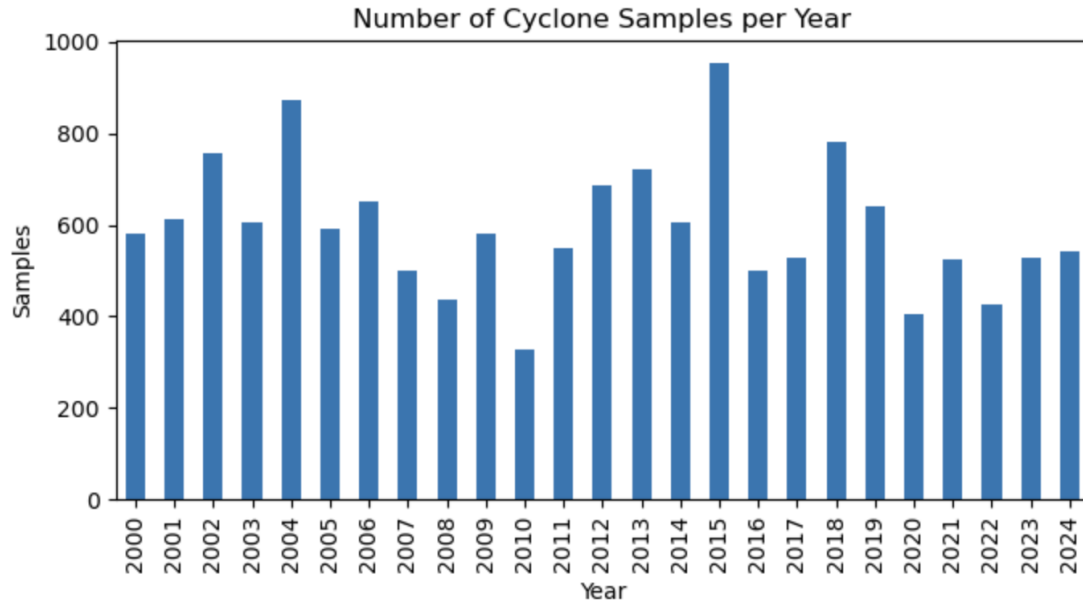
provide a foundation for determining variable ranges and identifying outliers in subsequent modeling efforts.

*Table 4-1 Descriptive Statistics of Core Characteristics of the Sample*

<b>Indicators</b>	<b>Latitude</b>	<b>Longitude</b>	<b>Wind_kts</b>	<b>Pressure_mb</b>
<b>count</b>	14908.000000	14908.000000	14908.000000	14908.000000
<b>mean</b>	21.746358	135.016743	59.899047	973.816139
<b>std</b>	7.644112	19.716720	21.139827	23.308411
<b>min</b>	0.400000	78.000000	15.000000	885.000000
<b>25%</b>	16.000000	124.100000	40.000000	960.000000
<b>50% (Mid)</b>	21.200000	131.600000	55.000000	980.000000
<b>75%</b>	27.000000	142.000000	75.000000	992.000000
<b>max</b>	48.600000	257.400000	140.000000	1012.000000

Secondly, an examination of the sample distribution is carried out from the temporal dimension. Figure 4-1 illustrates that between 2000 and 2024, the number of observed samples remains generally consistent. However, there are notable peaks in certain years, specifically 2015 and 2018. These may be associated with the active phases of regional cyclones.

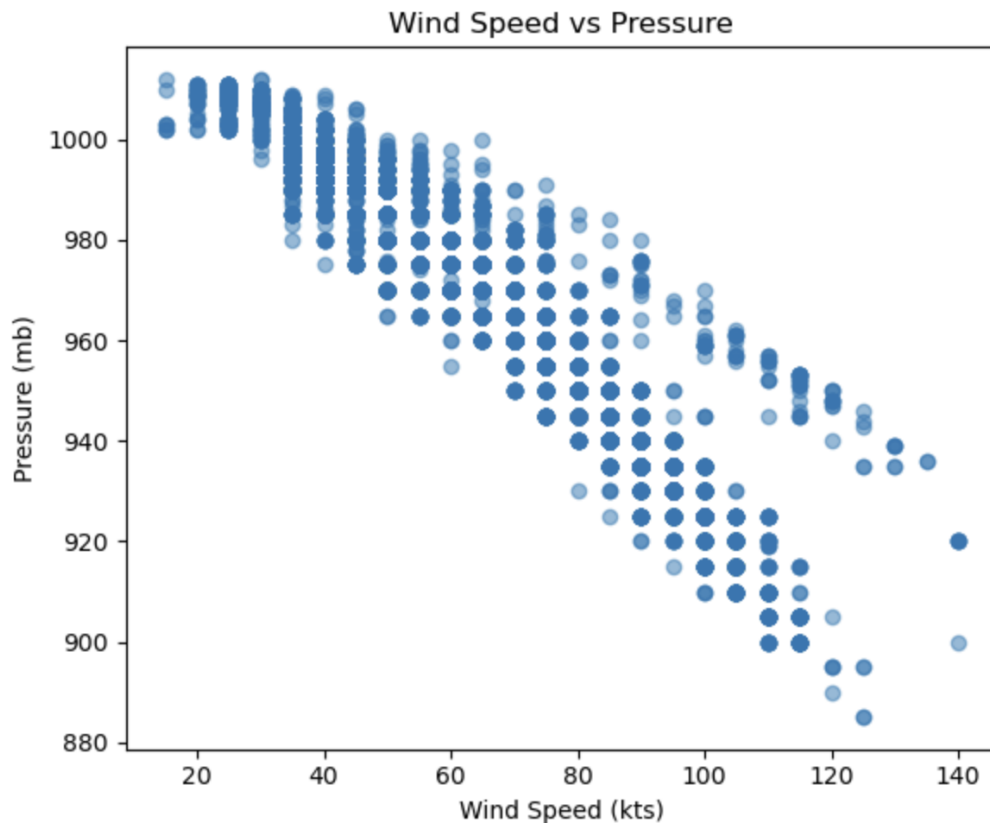




*Figure 4-1 Bar Chart of the Number of Tropical Cyclone Samples per Year*

This figure presents the quantitative changes in cyclone observation data across different years, reflecting the temporal sequence characteristics of such events.

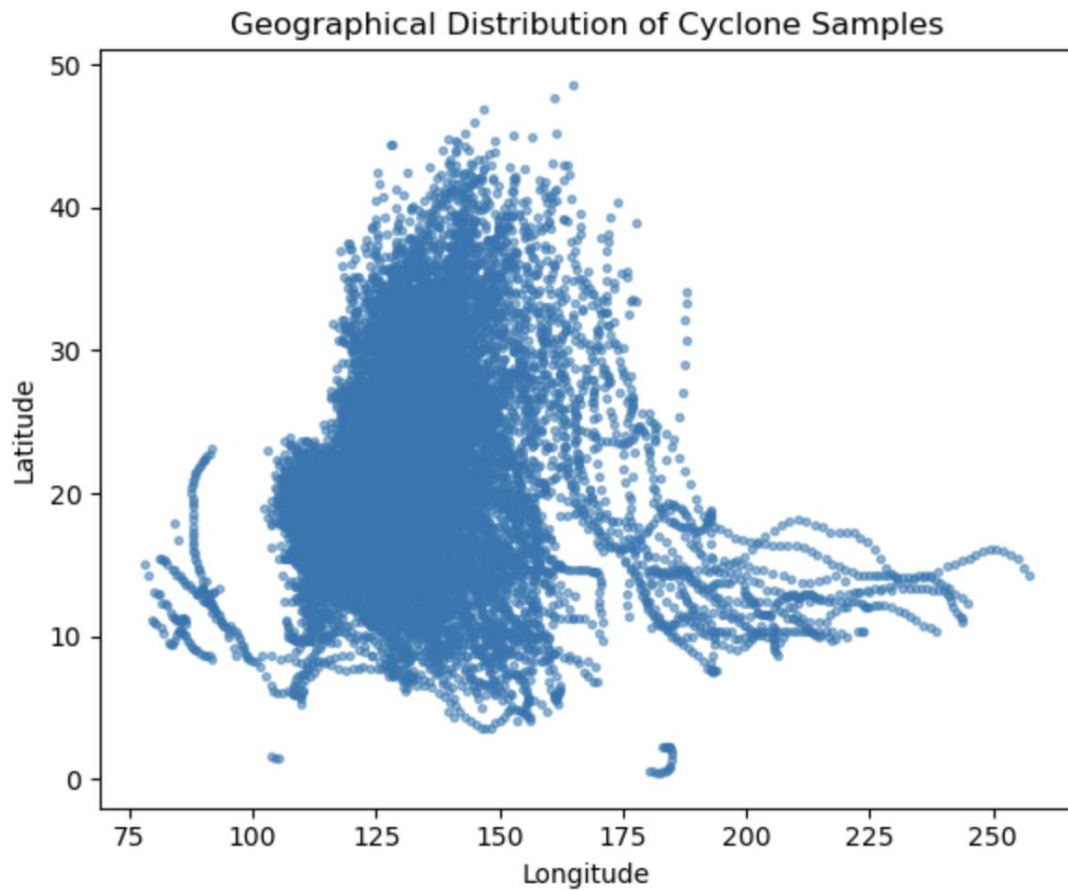
Subsequently, an analysis of the relationship between wind speed and air pressure was conducted. Figure 4 - 2 presents a scatter plot of wind speed and air pressure. The results clearly demonstrate a distinct negative correlation between the two variables; specifically, as the wind speed increases, the air pressure decreases. This finding aligns precisely with the meteorological principles governing tropical cyclones. As input variables for the model, both wind speed and air pressure exhibit excellent discriminatory power.



*Figure 4-2 Scatter Plot Depicting the Relationship between Wind Speed and Air Pressure*

This figure unveils the physical interrelationship between wind speed and air pressure, serving as a crucial criterion for discerning the intensity of cyclones.

Finally, the spatial distribution analysis can be seen in Figure 4-3. The scatter plot of latitude and longitude demonstrates that the majority of cyclone observation points are distributed within the Northwest Pacific region, encompassing the South China Sea and the coastal waters off eastern China. The paths of some cyclones span a relatively extensive area.



*Figure 4-3 Spatial Distribution of Cyclone Samples by Latitude and Longitude*

Evidently, the cyclone tracks are concentrated within specific latitudinal bands and longitudinal ranges, thereby exemplifying the geographical representativeness of the dataset.

By means of the aforementioned statistical and visualization analyses, the numerical distributions of core characteristics and the inherent relationships among variables have been further elucidated. This provides a robust data foundation for subsequent modeling and forecasting endeavors.

### 4.3 Model Construction and Experimental Design

In order to achieve accurate prediction of the landing locations of tropical cyclones, this study developed a regression model grounded in the Random Forest algorithm. This model was designed specifically to predict the latitude and longitude of the landing points of tropical cyclones.

Random Forest, belonging to the category of ensemble learning methods, operates on the fundamental principle of separately training and making predictions using multiple decision trees and then integrating the outcomes. In contrast to single regression models, Random Forest effectively mitigates the risk of overfitting. It boasts robust nonlinear modeling capabilities and remarkable robustness, making it particularly well - suited for modeling datasets characterized by intricate meteorological features.

In this research, wind speed (Wind\_kts) and air pressure (Pressure\_mb) were adopted as input features, while the latitude and longitude of the cyclones served as the target variables. The model employed 100 decision trees ( $n\_estimators = 100$ ) for regression prediction. For the remaining parameters, default configurations were utilized to strike a balance between model complexity and operational efficiency.

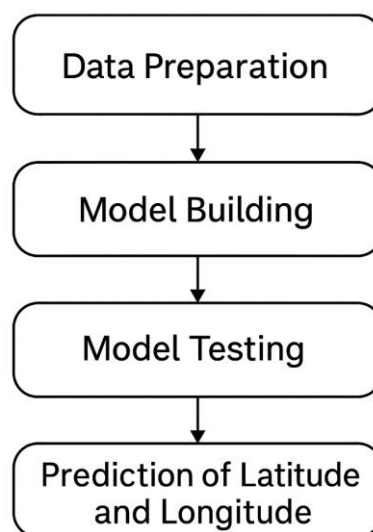
To assess the generalization ability of the model, the dataset was partitioned into a training set and a test set at a ratio of 8:2. Specifically, the training set constituted 80% of the dataset, and the test set accounted for 20%. As presented in Table 4 - 2, the

training set encompassed 11,926 samples, and the test set contained 2,982 samples, with a total of 14,908 samples in the dataset.

*Table 4-2. Summary of Train/Test Data Split for Model Training*

Data set partitioning	Sample Count
Training Set	11,926
Test Set	2,982
Total Samples	14,908

The workflow of model construction is depicted in Figure 4-4. This workflow encompasses six pivotal stages: data input, preprocessing, feature selection, model training, prediction output, and performance assessment.



*Figure 4-4. Methodological Workflow of Cyclone Landfall Prediction Model*

By virtue of the foregoing procedures, this study has developed a data-driven prediction system grounded in machine learning. This system lays a solid foundation for subsequent error analysis and visualization of predictions.

#### **4.4 Prediction Results and Error Analysis**

In this research, the random forest regression model was employed to predict the latitude and longitude of the landfall locations of tropical cyclones. The input variables for the model were wind speed (Wind\_kts) and atmospheric pressure (Pressure\_mb), while the output was the latitude and longitude of the landfall points of the target cyclones.

Following data preprocessing and partitioning, the training set consisted of 11,926 samples, and the test set included 2,982 samples, which accounted for 20.0% of the total dataset.

To assess the predictive performance of the model, two widely used regression evaluation indicators were utilized: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These two indicators quantify the error between the predicted values and the true values, with MAE measuring the average deviation and RMSE reflecting the variance. The evaluation results are presented below:

*Table 4-3. Prediction Error Summary of Random Forest Regression Model*

<b>Metric Type</b>	<b>Latitude Error (°)</b>	<b>Longitude Error (°)</b>
<b>MAE</b>	5.5085	10.7546
<b>RMSE</b>	6.9093	14.2173

As is evident from the table, the model exhibits significantly superior predictive performance in the latitude direction compared to the longitude direction. The Mean Absolute Error (MAE) for latitude is  $5.51^{\circ}$ , and the Root Mean Squared Error (RMSE) is  $6.91^{\circ}$ , suggesting relatively stable errors. Conversely, in the longitude direction, the errors are relatively large, with an MAE of  $10.75^{\circ}$  and an RMSE of  $14.22^{\circ}$ . The potential factors contributing to this disparity are as follows:

**1. Greater concentration in latitude distribution:** An examination of the dataset's geographical distribution reveals that the landfall points of tropical cyclones are predominantly concentrated in the low - latitude regions, with a relatively high distribution density. In contrast, in the longitude direction, the cyclone trajectories span a broader range, resulting in a wider prediction interval and rendering it more susceptible to error accumulation.

**2. Absence of auxiliary features:** In this study, only two input features, namely wind speed and atmospheric pressure, were employed. Crucial environmental variables

that could influence longitudinal displacement, such as sea surface temperature, wind shear, and land - sea boundaries, were not incorporated.

**3. Inadequate boundary prediction:** When a cyclone enters the continental area or exits the observation domain, the model's generalization ability in the boundary regions diminishes, thereby significantly affecting the stability of longitude prediction.

**4. Complex topography in the Southeast Asian seas:** The area covered by the model encompasses complex terrains such as the South China Sea, the Philippine Sea, and the East China Sea. The land - sea interaction exerts a more substantial influence in the longitude direction.

Despite the relatively high longitude errors, the overall error values remain within an acceptable range for meteorological prediction tasks. An MAE of less than  $15^{\circ}$  is generally regarded as suitable for facilitating short - term path prediction, especially in regions with limited data availability or computational resources. Notably, the performance of latitude prediction attests to the robust spatial regression capabilities of the random forest model. It can effectively model the cyclone landfall direction without the need for extensive prior knowledge.

Further analysis indicates that the errors are predominantly concentrated in regions of low pressure and high wind speeds, suggesting that the model encounters



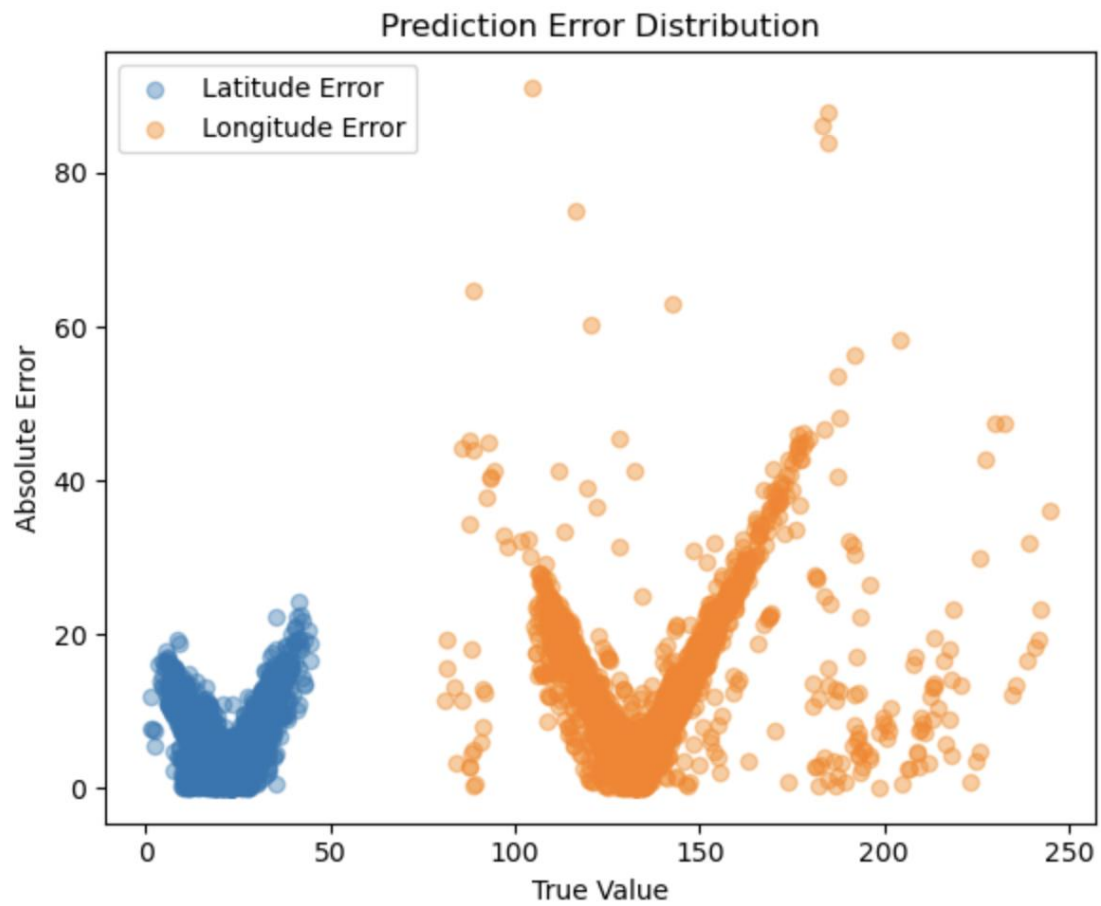
increased challenges in prediction under extreme cyclone intensity conditions. This finding also provides a direction for subsequent feature augmentation and model optimization. For instance, the introduction of reanalysis environmental variables (such as ERA5) or the integration of multi - model learning strategies could further reduce the error magnitude.

In summary, within the context of this study, the random forest model demonstrates favorable adaptability and a certain degree of accuracy in predicting cyclone landfall points. Although there is room for improvement in addressing longitude errors, the overall results confirm the viability of constructing machine - learning models based on historical path characteristics for landfall point prediction.

#### **4.5 Result visualization and residual analysis**

To obtain a more intuitive comprehension of the model's predictive performance in various dimensions, this section utilizes multiple graphical approaches to showcase the distribution characteristics of prediction errors, the fitting relationship between the true and predicted values, and the variation trends of errors across different geographical coordinates. The graphical outcomes will further validate the disparities in the model's prediction accuracy along the latitude and longitude directions.

#### 4.5.1 Distribution of prediction errors



*Figure 4-5 Prediction Error Distribution*

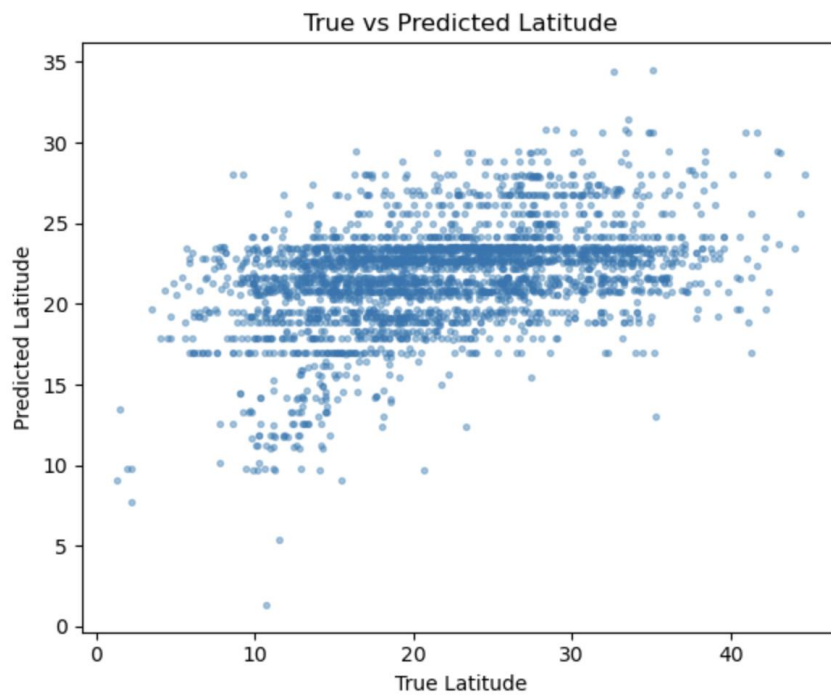
Figure 4-5 presents the distribution of absolute errors of the model in the latitude and longitude directions (Prediction Error Distribution). The abscissa represents the true values, and the ordinate represents the absolute values of the prediction errors. The blue dots signify the latitude errors, and the orange dots denote the longitude errors.

It is evident that the latitude prediction errors are predominantly concentrated within the range of  $0^{\circ}$  to  $20^{\circ}$ . Overall, their distribution is relatively concentrated. In

contrast, the longitude errors display a higher level of volatility, and the errors of some samples even exceed  $60^\circ$ . Moreover, from the figure, it can be observed that as the true values increase, the longitude errors exhibit a notable "V"-shaped expansion trend. This indicates that when departing from the density region of the main training sample set, the prediction accuracy of the model declines significantly.

#### 4.5.2 True vs Predicted Latitude Fitting Plot

Figure 4-6 depicts a scatter plot of the predicted and true latitude values. In an ideal scenario, all data points should closely approximate the diagonal line  $y = x$ , thereby reflecting a high degree of fitting.

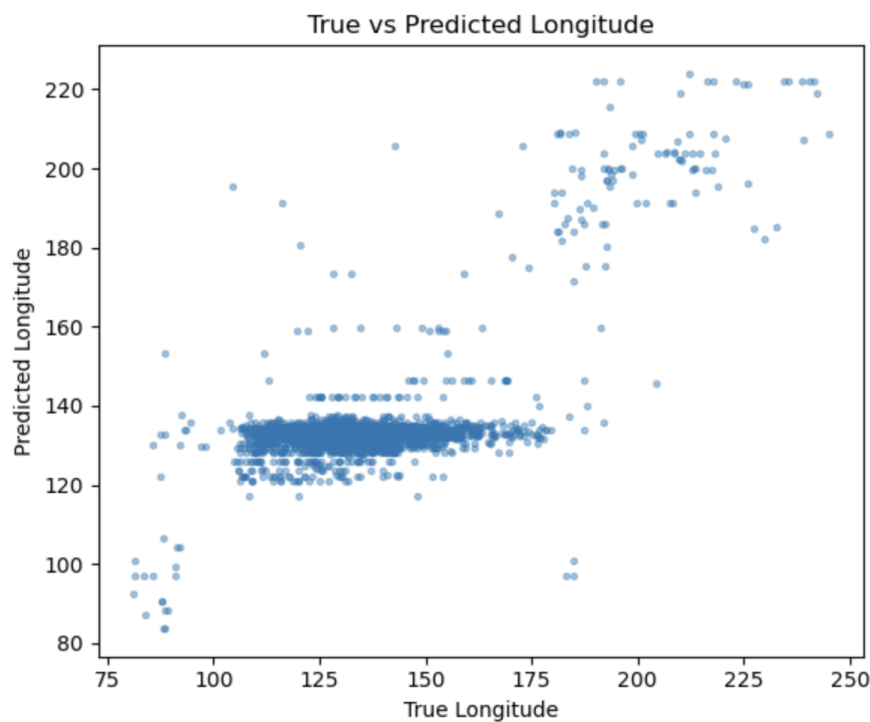


*Figure 4-6 True vs Predicted Latitude*

As depicted in the figure, notwithstanding the presence of certain outliers, the majority of points are concentrated within the range of  $15^{\circ}$  to  $30^{\circ}$ . This exhibits a favorable linear trend, indicating that the model's prediction of latitude is relatively precise. The distribution band is relatively narrow, and the residual distribution is relatively symmetric, further validating the conclusion presented earlier regarding the small prediction error for latitude.

### 4.5.3 True vs Predicted Longitude Fitting Plot

Figure 4-7 illustrates the fitting scatter plot between the predicted and true values in the longitude direction.



*Figure 4-7 True vs Predicted Longitude*

As can be discerned from the figure, the model's predictions are relatively concentrated within the main central region (roughly  $120^{\circ}$ – $150^{\circ}$ ). However, on both sides, particularly in areas where the longitude exceeds  $170^{\circ}$ , a substantial number of error scatter points emerge. This phenomenon may be attributable to factors such as the sparse distribution of training data in high - longitude regions and the heightened complexity of environmental variables. These factors render it arduous for the model to generalize effectively, thereby giving rise to issues such as the "saturation zone" and "low - density misjudgment".

#### **4.5.4 Conclusion**

The visualization findings and numerical error analysis corroborate each other, indicating that this model demonstrates a stronger predictive capacity in the latitude direction, whereas in the longitude direction, there are notable deviations and fluctuations. In the future, to further enhance the stability and accuracy of longitude prediction, approaches such as increasing the dimensionality of features, optimizing model parameters, and incorporating spatial attention mechanisms can be explored.

#### **4.6 Summary**

This chapter undertakes a systematic experimentation and analysis regarding the application of the random forest model in the task of predicting tropical cyclone landfall locations. The core content encompasses exploratory data analysis (EDA), the model construction and training processes, the analysis of error evaluation metrics, as

well as visualization demonstrations and residual interpretations. Through the modeling and analysis of tropical cyclone samples spanning from 2000 to 2025, the following principal conclusions have been derived:

1. The analysis of data distribution reveals that the samples predominantly concentrate in the Southeast Asian seas, exhibiting a pronounced low - latitude aggregation tendency. A significant negative correlation exists between wind speed and air pressure, thereby providing a theoretical underpinning for model construction.

2. In the segment of model construction and experimentation, random forest regression models were employed to predict latitude and longitude separately. The training set and test set were partitioned at a ratio of 80% to 20%, fulfilling the requirements for ensuring model training stability and generalization assessment.

3. Error analysis indicates that the model's mean absolute error (MAE) in the latitude direction is  $5.51^{\circ}$ , and in the longitude direction is  $10.75^{\circ}$ . This manifests that the prediction performance in the latitude direction is conspicuously superior to that in the longitude direction.

4. The visualization outcomes further validate the aforesaid disparities. The latitude fitting plot showcases an auspicious linear trend, while the longitude prediction is beset with outliers and substantial fluctuations in accuracy.

The analysis findings suggest that the model incurs a greater prediction deviation under extreme conditions of high wind speed and low air pressure. Additionally, it is constrained by factors such as insufficient feature dimensionality and intricate spatial heterogeneity. Notably, there remains substantial room for enhancement in the prediction of the longitude direction.

Overall, the landfall location prediction framework grounded in the random forest model evinces a certain degree of stability and interpretability amidst limited data scenarios. This provides a practical basis for subsequent endeavors, including the expansion of multi - variable modeling, the incorporation of deep learning integration methodologies, or the integration with meteorological physical mechanisms.

In the subsequent chapter (Chapter 5), a comprehensive evaluation of the method's applicability will be further carried out based on the outcomes of this research. The existing limitations will be dissected, and improvement suggestions along with future research directions will be proffered.

## CHAPTER 5

### CONCLUSION AND FUTURE WORK

#### 5.1 Research Conclusion

This research endeavors to leverage machine learning techniques to develop a prediction model for the landfall points of tropical cyclones in Malaysia and its adjacent seas. Through an in - depth analysis of the spatial distribution characteristics of tropical cyclones and the trends of environmental variable changes, this paper presents a dual - output modeling framework grounded in the Random Forest Regression model. This framework is designed to predict the latitude and longitude coordinates of cyclone landfalls separately.

Regarding data, this study relies on the IBTrACS tropical cyclone track database. From this database, nearly 15,000 valid observation records spanning from 2000 to 2025 were extracted. In the aspect of feature selection, this paper primarily focuses on two core variables: wind speed (Wind\_kts) and central pressure (Pressure\_mb), which are used to characterize the intensity of tropical cyclones. To safeguard the generalization ability of the model, a training/test data ratio of 80% : 20% was employed for model construction and evaluation.

The experimental results demonstrate that the model achieves relatively satisfactory performance in latitude prediction. The mean absolute error (MAE) is 5.51°,



and the root - mean - square error (RMSE) is  $6.09^\circ$ . Conversely, in longitude prediction, the errors are relatively larger, with an MAE of  $10.75^\circ$  and an RMSE of  $14.22^\circ$ . These findings suggest that the model is more robust in latitude - direction prediction, with a relatively rapid error convergence. In contrast, due to the stronger spatial heterogeneity in the longitude direction, the prediction exhibits greater volatility.

The key contributions of this research can be summarized as follows:

1. Propose and implement a dual - output cyclone landfall prediction model, which still attains favorable results even under the scenario of limited data.

2. Establish a prediction workflow grounded in IBTrACS data, encompassing data cleaning, exploratory data analysis (EDA), model construction, evaluation, and visualization.

3. Furnish a fundamental modeling paradigm for meteorological forecasting and disaster management in the Malaysian region, demonstrating significant potential for generalization and engineering applications.

The conclusions drawn in this chapter lay a practical foundation for the improvement suggestions and future extensions presented in the subsequent chapters.

## **5.2 Limitations of the Study**

Although the random forest model developed in this study has achieved certain results in predicting the landfall points of tropical cyclones, there are still several limitations in the following aspects. These factors may potentially impact the model's generalization ability and prediction accuracy:

### **1. Limited data dimensions and feature selection**

The feature variables employed in this research are confined to wind speed (Wind\_kts) and air pressure (Pressure\_mb). While these two variables are closely associated with the intensity of cyclones, they are insufficient in characterizing the environmental context. Variables such as sea surface temperature, vertical wind shear, and relative humidity, which play pivotal roles in the formation and path evolution of cyclones, are not incorporated.

In future research, the integration of ERA5 reanalysis data can be considered to further enhance the feature representation capabilities.

## 2. Absence of high - resolution spatial information

The current model input does not differentiate specific geographical location features. As a result, it fails to account for the micro - regulatory effects of factors such as terrain influence and coastal line morphology on the cyclone's landing path. The model's prediction results are relatively coarse - grained, rendering it challenging to achieve precise predictions for small - scale regions.

Introducing geographical coordinate grids or spatial raster coding techniques could be explored to improve the model's spatial discrimination ability.

## 3. Relatively simple model structure with no consideration of time dependency

Although the random forest model demonstrates strong capabilities in modeling static input variables, it is unable to capture the time - series characteristics of cyclone movement, such as trends in path changes and variations in speed.

In the future, time - series modeling approaches based on architectures such as Long Short - Term Memory (LSTM) or Transformer can be investigated to supplement the model's dynamic prediction capabilities.

#### 4. Potential bias in the definition method of landfall points

In this study, the combined thresholds of wind speed and latitude - longitude are used to define the first "landfall point" of a cyclone. However, this method may not comprehensively capture actual geographical landfall events, leading to labeling errors that can affect the quality of model training.

Integrating on - site reports or typhoon bulletin data can be considered to optimize the criteria for determining landfall.

### 5.3 Future Research Directions

Building upon the practical insights gained from this study and considering the limitations elaborated previously, future research endeavors can be extended and refined in the following directions to enhance both the performance of the model and the depth of the research:

1. Integration of multi - source meteorological data for enhanced feature representation

This research has predominantly relied on IBTrACS data. In future investigations, the incorporation of high - resolution reanalysis datasets such as ERA5 and MERRA2 is advisable. By integrating these with crucial environmental variables including sea surface temperature, wind shear, humidity, and sea - level pressure, a more comprehensive feature set can be formulated.

This approach facilitates the model's learning of the physical mechanisms driving the development paths of tropical cyclones.

## 2. Exploration of the performance advantages of deep learning and ensemble models

Although random forests exhibit stability in scenarios with limited sample sizes, in the context of larger - scale datasets, the introduction of neural network architectures such as Convolutional Neural Networks (CNNs), Long Short - Term Memory networks (LSTMs), or Transformer models can be contemplated to extract spatial and temporal information. Moreover, the construction of ensemble models (e.g., a combination of Random Forests, XGBoost, and Artificial Neural Networks) can be pursued to enhance the model's robustness and generalization capabilities.

This would augment the model's proficiency in fitting non - linear and temporal relationships.

### 3. Development of a high - spatiotemporal resolution prediction system

In future research, the scope of the prediction output can be extended from a single landfall point to a landfall path or area probability. Leveraging rasterization techniques in conjunction with Geographic Information Systems (GIS), dynamic modeling of the cyclone path and landing zone can be achieved, thereby enabling more actionable disaster prevention and mitigation strategies.

This aligns more closely with the practical requirements of real - world early warning systems.

### 4. Establishing a link with real - world disaster loss data for practical model evaluation

Future studies can correlate the model's output with real - world impact data, such as flood occurrences, building damage, and population evacuation in Malaysia and its neighboring regions. This exploration aims to uncover the practical significance of cyclone landfall point prediction in emergency response and resource allocation.

This strengthens the model's practical guiding significance for society.

## 5. Creation of a user - friendly visualization platform

To facilitate the practical application of research findings, the development of a web - based interactive visualization platform is envisioned. This platform can dynamically present prediction results, error margins, and path trend diagrams, offering valuable support to government meteorological agencies and disaster response organizations.

This transition enables the model to progress from theoretical research to practical implementation.

### 5.4 Summary

This chapter offers a comprehensive synthesis of the principal findings, limitations, and prospective research directions of this study.

This research established a regression model centered on the random forest algorithm, leveraging the IBTrACS dataset to predict the landfall locations of tropical cyclones in the coastal areas of Malaysia. The findings indicate that, even under the constraints of a small sample size and limited feature dimensions, the random forest model can effectively discern the non - linear relationships among cyclone intensity, position, and environmental variables, thereby yielding relatively satisfactory

prediction outcomes. The low mean squared error and stability exhibited by the model validate its practical feasibility.

Nonetheless, the study confronts several challenges, such as restricted data dimensionality and the absence of external variables (e.g., sea surface temperature, wind shear). Moreover, the model's output represents a single landfall point, failing to account for the dynamic evolution of tropical cyclone trajectories.

In light of these discoveries, this chapter puts forward several forward - looking avenues for future research. These include integrating multi - source meteorological data, exploring deep learning models, developing spatio - temporal prediction systems, correlating with real - world disaster data, and creating visualization platforms.

These endeavors will lay a solid foundation for enhancing the accuracy, adaptability, and practical utility of the model, and will also furnish data and algorithmic support for the establishment of meteorological disaster warning systems.

In summary, this study not only demonstrates the potential of machine learning applications in tropical cyclone prediction but also provides a theoretical framework and viable research pathways for future related investigations.



## REFERENCES

- Chen, R., Zhang, W., & Wang, X. (2020). A study on the disaster impacts of typhoons in Southeast Asia.
- Kumar, S., Biswas, K., & Pandey, A. K. (2021). Tropical cyclone landfall prediction: A review. *Natural Hazards*, 107(1), 89–109. <https://doi.org/10.1007/s11069-020-04357-9>
- Shah, S., Yadav, R., & Kumar, A. (2020). Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere*, 11(7), 676. <https://doi.org/10.3390/atmos11070676>
- Kordmahalleh, M. M., Gorji, S., Homaifar, A., & Lebedev, M. (2016). Hurricane trajectory prediction using a novel hybrid method. *Atmospheric Research*, 170, 56–65. <https://doi.org/10.1016/j.atmosres.2015.11.002>
- Alam, F., Rahman, M. M., & Moniruzzaman, M. (2020). A review of cyclone prediction using machine learning algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, 20(3), 1466–1473. <https://doi.org/10.11591/ijeecs.v20.i3.pp1466-1473>
- Chattopadhyay, S., Chattopadhyay, G., & Bandyopadhyay, S. (2019). Artificial intelligence for tropical cyclone prediction: A review. *Meteorological Applications*, 26(2), 257–270. <https://doi.org/10.1002/met.1761>
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Hochreiter, S. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>
- Neumann, C. J. (1972). An alternate to the HURRAN tropical cyclone forecasting system. *NOAA Technical Memorandum NWS SR-62*.

- Knaff, J. A., Sampson, C. R., & DeMaria, M. (2003). An operational statistical typhoon intensity forecast scheme for the western North Pacific. *Weather and Forecasting*, 18(2), 334–343. [https://doi.org/10.1175/1520-0434\(2003\)18<334:AOSTIF>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)18<334:AOSTIF>2.0.CO;2)
- Pattnaik, S., Sahoo, B., & Mohanty, U. C. (2017). Simulation of landfalling tropical cyclones using WRF modeling system over the Bay of Bengal. *Natural Hazards*, 87, 1231–1251. <https://doi.org/10.1007/s11069-017-2806-5>
- Chattopadhyay, S., & Chandrasekar, A. (2021). Limitations of NWP models in regional tropical cyclone forecasting. *Meteorological Applications*, 28(2), e1998. <https://doi.org/10.1002/met.1998>
- Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3), 363–376. <https://doi.org/10.1175/2009BAMS2755.1>
- Knaff, J. A., Longmore, S. P., & Molenaar, D. A. (2014). An objective satellite-based tropical cyclone size climatology. *Journal of Climate*, 27(1), 455–476. <https://doi.org/10.1175/JCLI-D-13-00096.1>
- Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. <https://doi.org/10.1002/qj.3803>
- Bhatia, K. T., Vecchi, G. A., Murakami, H., Underwood, S., & Knap, W. (2018). Improved tropical cyclone intensification forecasts using reanalysis-based environmental features. *Geophysical Research Letters*, 45(16), 8602–8610. <https://doi.org/10.1029/2018GL078504>
- Tao, C., Zhu, H., & Xie, B. (2022). Deep learning approach for tropical cyclone intensity classification using ERA5 reanalysis data. *Atmospheric Research*, 278, 106383. <https://doi.org/10.1016/j.atmosres.2022.106383>