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.

Two publicly accessible meteorological datasets are utilized in this research: the International Best Track Archive for Climate Stewardship (IBTrACS) global cyclone path dataset and the ERA5 meteorological reanalysis dataset.

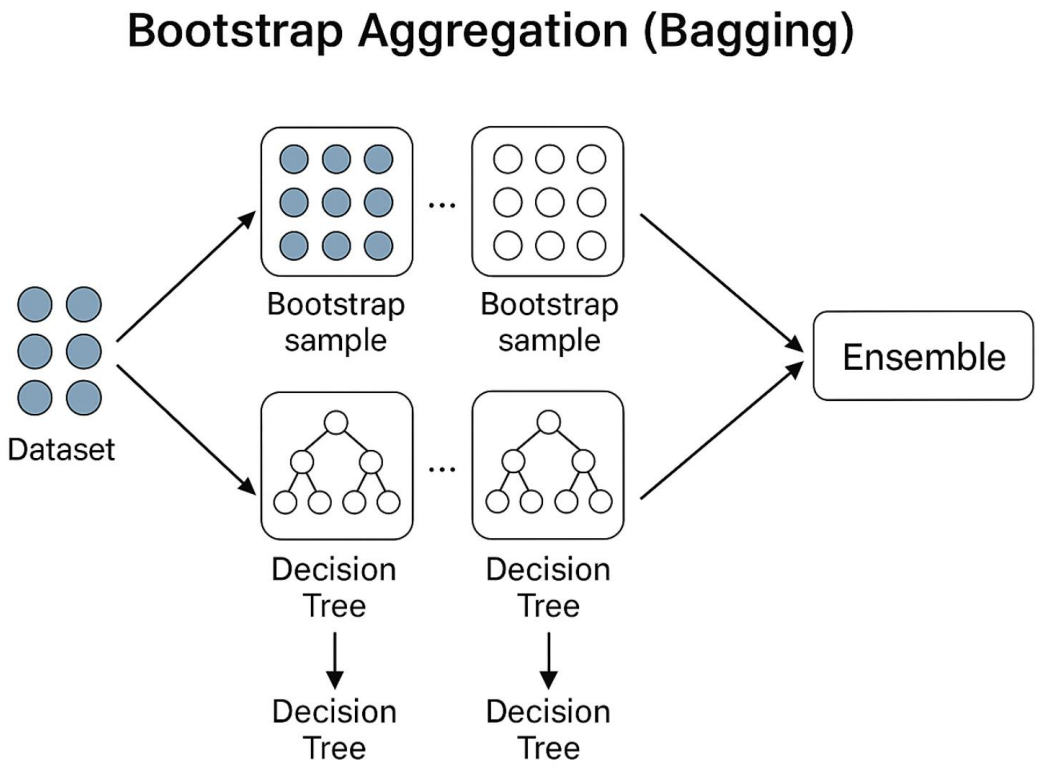
The primary objective of this chapter is to elucidate the process of converting raw meteorological data into model outputs capable of making predictions. Additionally, it will expound on the rationale behind the selection of the Random Forest model. In particular, its proficiency in dealing with non - 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.



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:

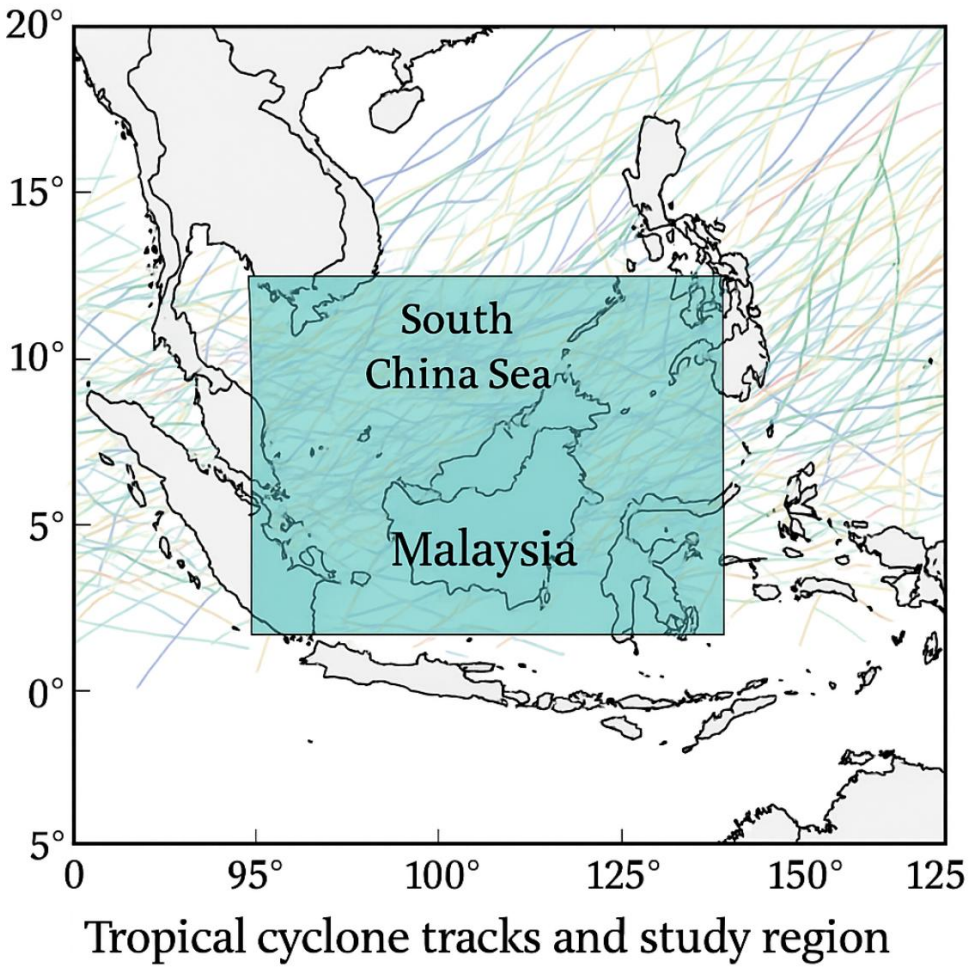
- T denotes the total number of constructed decision trees

- $f_i(x)$ 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 two primary publicly accessible meteorological datasets: the International Best Track Archive for Climate Stewardship (IBTrACS) and the ERA5 reanalysis dataset. Both datasets are extensively employed in global meteorological research and are characterized by their high data quality and extensive spatiotemporal coverage.

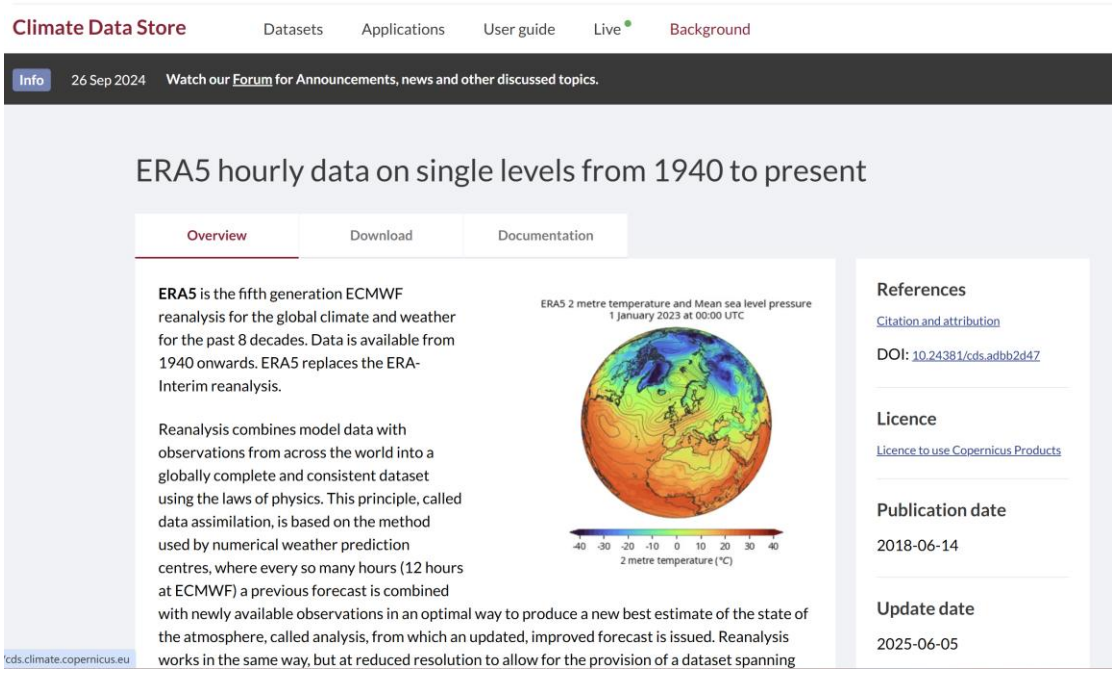


Tropical cyclone tracks and study region

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.

ERA5 is a high-precision global reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). It offers hourly estimates of various atmospheric and oceanic variables. Key variables relevant to this study include sea surface temperature (SST), surface pressure, 10-meter wind speed and direction, and mid-level relative humidity. These variables serve as essential factors influencing the path and intensity changes of tropical cyclones.



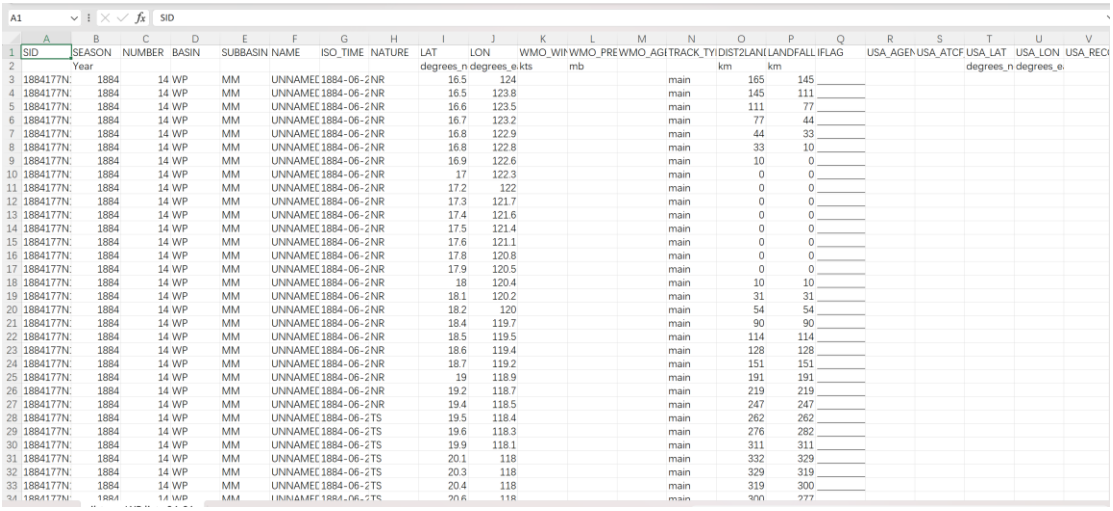
By temporally aligning the cyclone trajectory information from IBTrACS with the environmental variables in ERA5, a supervised learning training dataset can be constructed. In this dataset, cyclone positions are treated as output variables, while the corresponding environmental features at each time step are used as input variables. The study focuses on the South China Sea and surrounding waters of Malaysia, covering a geographical range approximately between 0°–20°N latitude and 95°–125°E longitude.

3.4 Data Preprocessing

Before model training, a series of preprocessing steps must be performed on the IBTrACS and ERA5 datasets 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.

3.4.1 Downloading and Filtering of the IBTrACS Dataset

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



	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_WIN	WMO_PRE	WMO_AGE	TRACK_TY	DIST2LAN	LANDFALL	IFLAG	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	UNNAMED	1884-06-2	NR	16.5	124				main	165	145						
4	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	16.5	123.8				main	145	111						
5	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	16.6	123.5				main	111	77						
6	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	16.7	123.2				main	77	44						
7	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	16.8	122.9				main	44	33						
8	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	16.8	122.8				main	33	10						
9	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	16.9	122.6				main	10	0						
10	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	17	122.3				main	0	0						
11	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	17.2	122				main	0	0						
12	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	17.3	121.7				main	0	0						
13	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	17.4	121.6				main	0	0						
14	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	17.5	121.4				main	0	0						
15	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	17.6	121.1				main	0	0						
16	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	17.8	120.8				main	0	0						
17	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	17.9	120.5				main	0	0						
18	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	18	120.4				main	10	10						
19	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	18.1	120.2				main	31	31						
20	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	18.2	120				main	54	54						
21	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	18.4	119.7				main	90	90						
22	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	18.5	119.5				main	114	114						
23	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	18.6	119.4				main	128	128						
24	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	18.7	119.2				main	151	151						
25	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	19	118.9				main	191	191						
26	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	19.2	118.7				main	219	219						
27	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	NR	19.4	118.5				main	247	247						
28	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	TS	19.5	118.4				main	262	262						
29	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	TS	19.6	118.3				main	276	282						
30	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	TS	19.9	118.1				main	311	311						
31	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	TS	20.1	118				main	332	329						
32	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	TS	20.3	118				main	329	319						
33	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	TS	20.4	118				main	319	300						
34	1884177N	1884	14	WP	MM	UNNAMED	1884-06-2	TS	20.6	118				main	300	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 1980 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)

3.4.2 Downloading and Filtering of the ERA5 Dataset

The ERA5 dataset can be obtained from its official website (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview).

Given the substantial size of the ERA5 dataset files and the inclusion of unnecessary data for our study, we will conduct a filtering process on the ERA5 dataset in this step. Our filtering criteria are as follows:

- Time range: from 1980 to the present

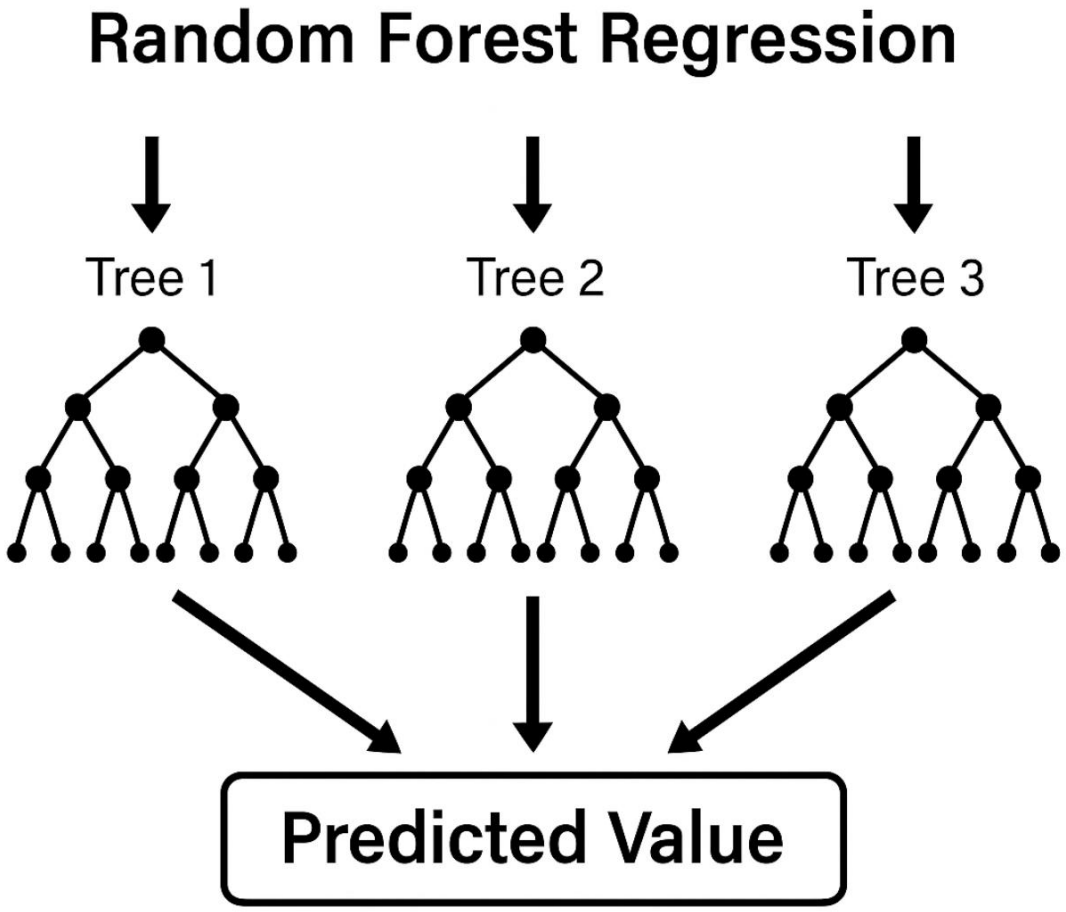
- Temporal resolution: every 6 hours

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

- Environmental variables for each cyclone at each time step: sea surface temperature, wind speed, air pressure, humidity, etc.

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 two primary data sources utilized: the IBTrACS cyclone trajectory dataset and the ERA5 environmental variable 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.