CHAPTER 4

RESULTS AND ANALYSES OF MODEL 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

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

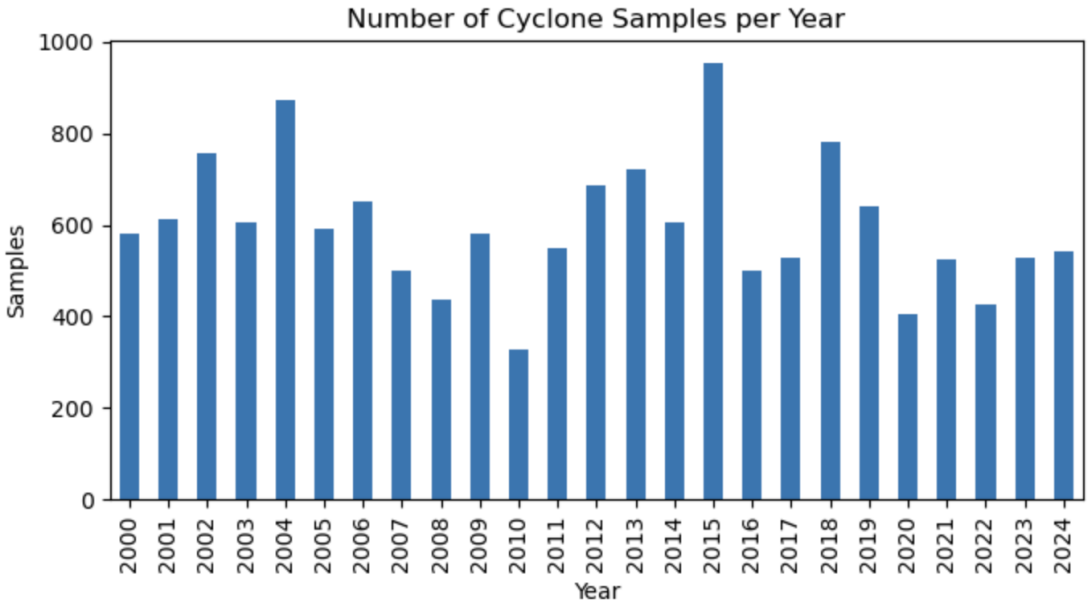

Figure 4-1 Bar Chart of the Number of Tropical Cyclone Samples per Year

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.

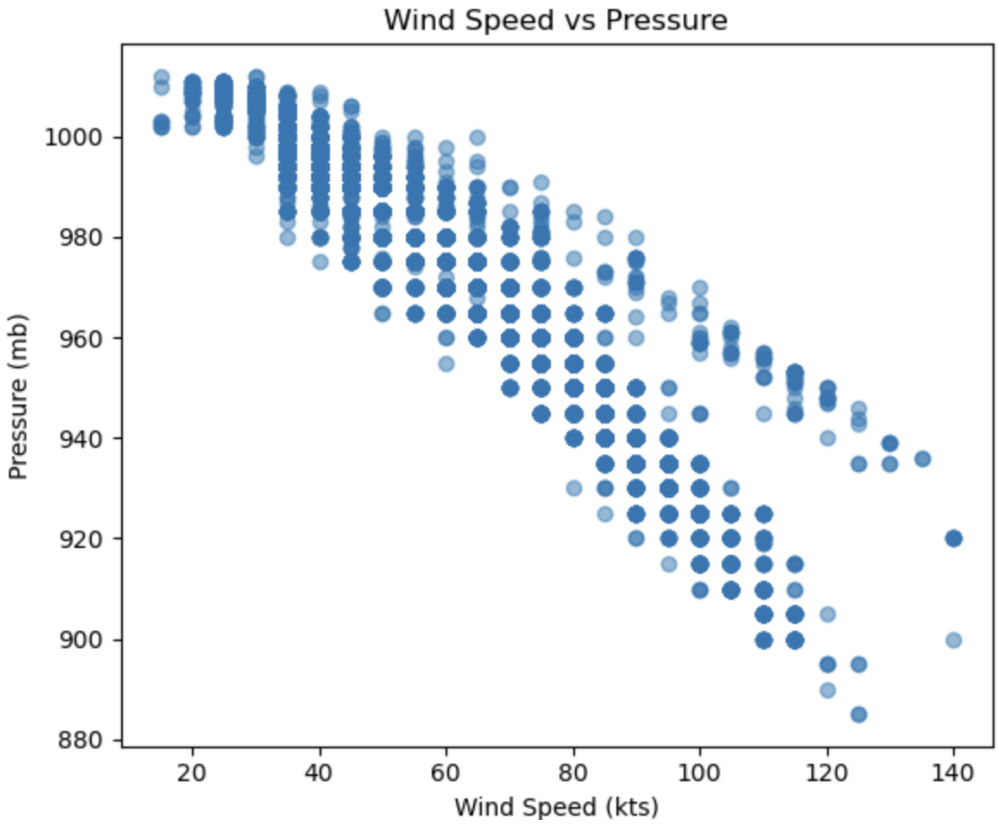

Figure 4-2 Scatter Plot Depicting the Relationship between Wind Speed and Air Pressure

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.

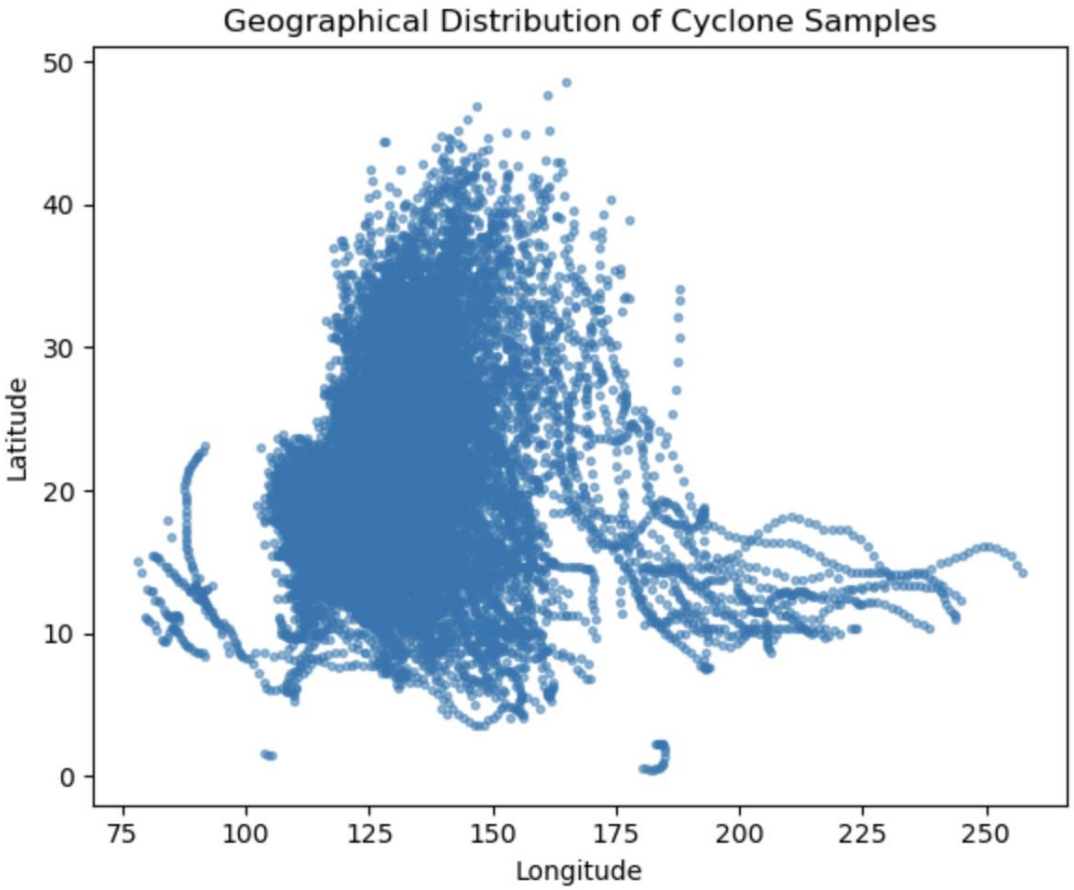

Figure 4-3 Spatial Distribution of Cyclone Samples by Latitude and Longitude

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.

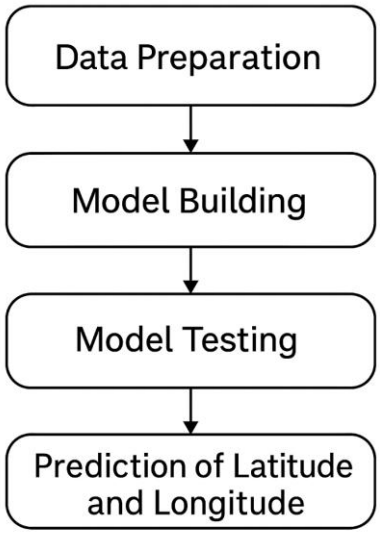

Figure 4-4. Methodological Workflow of Cyclone Landfall Prediction Model

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

Metric Type	Latitude Error (°)	Longitude Error (°)
MAE	5.5085	10.7546
RMSE	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°, and the Root Mean Squared Error (RMSE) is 6.91°, suggesting relatively stable errors. Conversely, in the longitude direction, the errors are relatively large, with an MAE of 10.75° and an RMSE of 14.22°. 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° 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

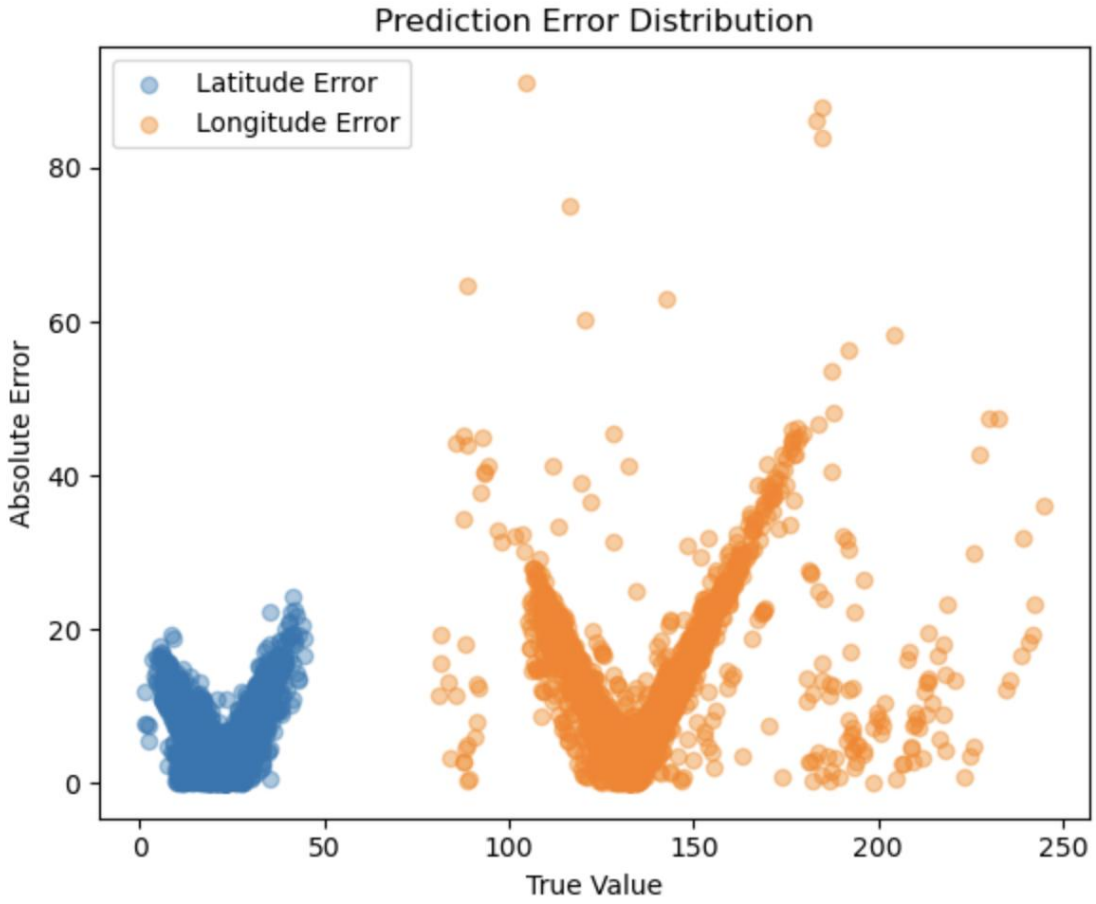

Figure 4-5 Prediction Error Distribution

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° to 20°. 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°. 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.

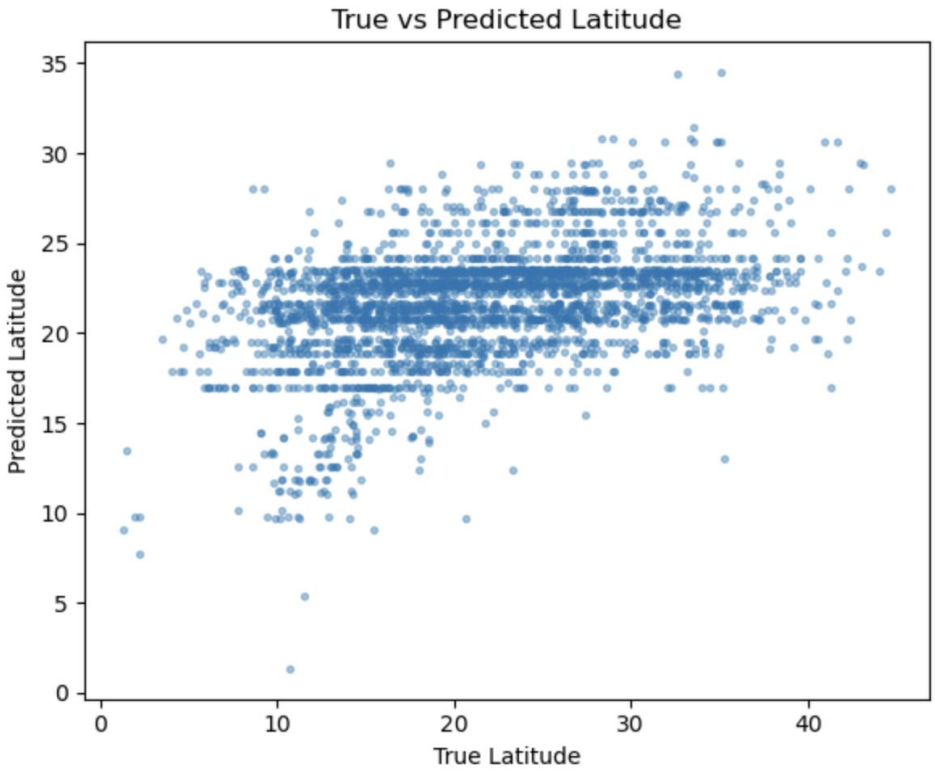

Figure 4-6 True vs Predicted Latitude

As depicted in the figure, notwithstanding the presence of certain outliers, the majority of points are concentrated within the range of 15° to 30°. 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.

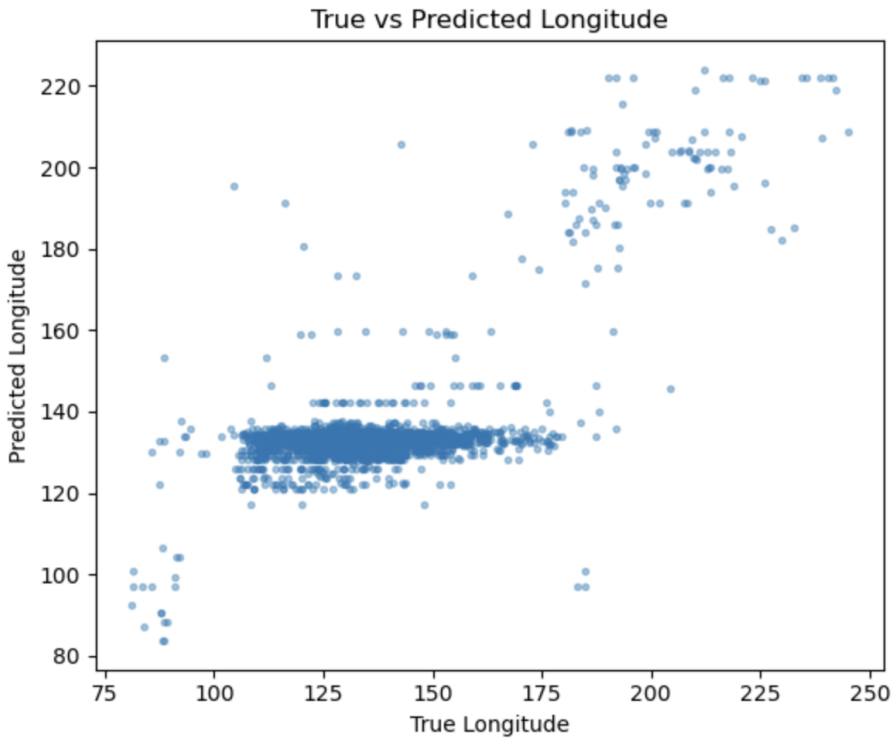

Figure 4-7 True vs Predicted Longitude

As can be discerned from the figure, the model's predictions are relatively concentrated within the main central region (roughly 120°–150°). However, on both sides, particularly in areas where the longitude exceeds 170°, 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°, and in the longitude direction is 10.75°. 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.