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:

$$P(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n$$

- $P(t)$: 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)

- $X1, X2, \ldots, Xn$: Input variables (such as SST, air pressure, wind speed, and other environmental variables)

- β_0: Intercept term

- β_i: Regression coefficient, reflecting the weight of each variable

- ε : 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, ρ represents the air density, ∇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.

- θ : 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:

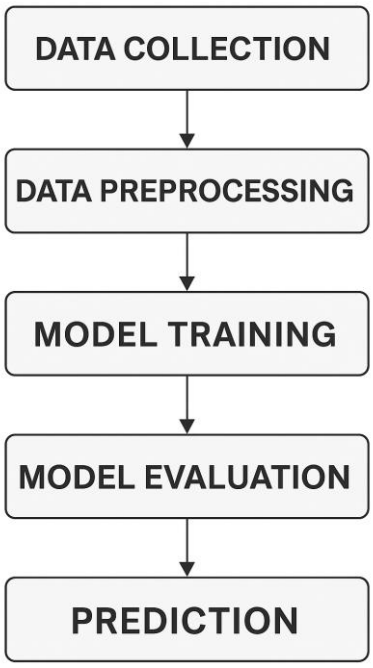

Figure 2.3

- **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 Review of Relevant Datasets

2.3.1 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

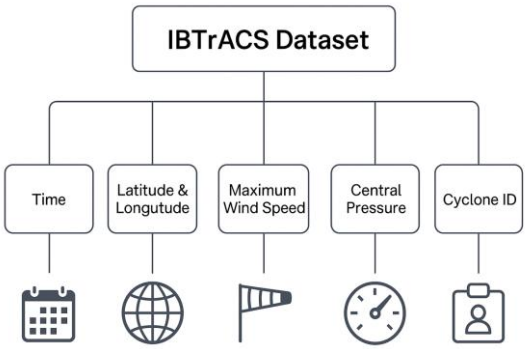

Figure 2.3.1

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.

2.3.2 ERA5 Dataset

ERA5, developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), represents the fifth generation of global reanalysis data, serving as an advanced and replacement version of the earlier ERA-Interim. This dataset achieves its comprehensive coverage by merging contemporary numerical weather prediction models with global observational data through the application of data assimilation techniques. Consequently, ERA5 provides hourly estimates for a wide array of atmospheric, terrestrial, and oceanic variables. Notably, it boasts a spatiotemporal resolution of $0.25° \times 0.25°$, with data availability dating back to 1979 (Hersbach et al., 2020).

In the context of tropical cyclone prediction endeavors, a suite of ERA5 variables are frequently employed, encompassing:

- Sea Surface Temperature (SST)

- Surface Atmospheric Pressure

- Wind Speed Components at Multiple Altitudes (U, V)
- Vertical Wind Shear
- Relative Humidity
- Total Precipitation
- Geopotential Height

These variables are instrumental in reflecting the environmental conditions surrounding the cyclone, and they play a pivotal role in the construction of feature variables within machine learning models (Bhatia et al., 2018).

ERA5 data, distinguished by its high accuracy and spatiotemporal consistency, has garnered widespread application in various research domains, including climate change analysis, storm surge modeling, and the prediction of tropical cyclone intensity. Notably, Bhatia et al. (2018) utilized ERA5 to construct a rapid intensification prediction index for tropical cyclones; Tao et al. (2022) combined wind shear and SST features, leveraging neural network models to enhance the accuracy of tropical cyclone intensity classification. In the present study, we intend to temporally synchronize the environmental variables extracted from ERA5 with the cyclone track data from IBTrACS, with the aim of utilizing these integrated data as input features for training the random forest model.

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. Limited Instances of Joint Utilization of Reanalysis and Trajectory Data

ERA5 and IBTrACS are pivotal meteorological and cyclone trajectory datasets accessible for research purposes. However, the integration of these datasets for the purpose of predicting landing points in Southeast Asia is underrepresented in the literature, warranting further investigation.

4. 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.

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

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., & Molenar, 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*