Data Science Research
Predicting Maximum Wind Speed And Presence Or Absence Of Rapid Intensification With Atlantic Hurricanes Data
Project Final Report

Cheuk Hang Ng
a1821087

November 24, 2022

Report submitted for **Data Science Research** at the School of Mathematical Sciences, University of Adelaide

Project Area: **Data Science, Data analysis, Statistical modelling, Machine learning**
Project Supervisor: **Dr John Maclean**

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

Abstract

Meteorologists have been working on improving the accuracy and performance of prediction of hurricanes for decades and varied strategies have been proposed and applied to the real world. Although weather forecasts often able to tell us the approximate path of a hurricane and when would it get stronger, it sometimes fails due to the complexity of hurricanes. This project attempts to predict the maximum wind speed and the presence or absence of rapid intensification using data from storms when they are still weak. In this project, a dataset of 3,700 Atlantic storms occurred during 1979 to 2015 was provided for analysis. The objective of this project is to investigate how good the prediction models will be in predicting the maximum wind speed and the presence or absence of rapid intensification with the data provided.

Three regression models, namely Linear Regression model, Partial Least Squares Regression model and Random Forest Regression model were built to predict the maximum wind speed while two classification models, which were Logistic Regression model and Random Forest Classification model were built to predict the presence or absence of rapid intensification. It was reported that linear regression model performed the best when predicting maximum wind speed by having the lowest mean absolute error of 15.92 and highest percentage variance of 22.93%, compared to the partial least squares model and the random forest regression model. For predicting the presence and absence of rapid intensification, both logistic regression model and random forest classification model failed to correctly identify the true positive cases with the probability threshold set at 0.5, but had reasonable performances with around 18% precision, 62.5% and 67.3% of sensitivity and 80% and 74.7% of specificity respectively when the probability threshold was set at 0.1.

The regression models did not perform well in predicting maximum wind speed from the dataset, and the features of storms in the dataset were not determining features for identifying rapid intensification, which implies there could be better approaches in predicting the maximum wind speed and the presence or absence of rapid intensification of storms during their lifetime.

1 Introduction

This project explores and investigates a dataset provided by WindRisk-Tech, LLC., which consists of data of Atlantic hurricanes observed between 1979 and 2015. The aim of this project is to investigate to what extent the maximum wind speed and the presence or absence of rapid intensification of storms during their lifetime could be predicted with the provided data. The objective is achieved by first performing exploratory data analysis, where the data are cleaned and the structures of the data are revealed before modelling, and unwanted data are filtered out. Next the target variables are modelled with regression models and classification models. Regression models to be implemented are Linear Regression model, Partial Least Squares Regression model and Random Forest Regression model. Classification models to be implemented are Logistic Regression model and Random Forest Classification model.

The following section gives background information on the topics and the data analysis environment of this project. Section 3 provides an overview of the data. Section 4 explains the details of the methods. Section 5 reports the results obtained from the models, and Section 6 discusses the key findings of this project.

2 Background

Hurricane behaviors and processes of formation have always been popular topics in the scope of atmospheric science. Some had done prediction on the frequency of Atlantic hurricane with decadal data [10], some studied structural change over time and modelled the rapid intensification of a single hurricane [5], while some looked at the relationship between rapid intensification and environmental features [15]. People have been studying the factors for the process to happen by analysing hurricane by hurricane [5, 9]. They are curious about the process because most category 4 or 5 hurricanes as defined in the Saffir-Simpson Hurricane Wind Scale [4] has at least one rapid intensification during their lifetime [16], where these hurricanes could be disastrous.

Rapid intensification is a process where a hurricane intensifies in terms of wind speed in a short period of time. In some studies, rapid intensification is suggested to be defined as an increase in the one minute maximum sustained surface wind speed beyond some predefined threshold such as 25, 30 or 35 knots over a 24 hour period [12]. However, for simplicity, this project adopts the definition of rapid intensification as an increase in wind speed of 50 knots or more in 24 hours [3].

The coding environment of this project is RStudio, an integrated development environment for R, which is specialised for data analysis and statistical modelling [13].

3 The Data

The dataset provided consists of:

1. Day
2. Hour
3. Month
4. Year
5. Latitude
6. Longitude
7. Pressure
8. Wind speed
9. Wind shear
10. Potential intensity

According to WindRiskTech, the data were recorded every two hours during a certain time span, and the maximum time span is 800 hours. The duration of each storm is extracted and stored as an extra variable. By investigating the Year data, we found that the dataset consists of 3,700 hurricane events. The first four files contain day, hour, month and the year of the hurricane events respectively. Latitude and Longitude data recorded the geographical location of the centre of hurricanes. Pressure data represents central surface pressure measured in millibar (mb), and Wind speed measured the wind components in knot at 850 hPa level. Wind shear is the data of the difference between 850 hPa and 250 hPa environmental wind speed. As there are no background information for the Potential intensity data from WindRiskTech, the definition for potential intensity is adopted from an article in the Jornal of Cliamte published by American Meteorological Society [7], where it represents the theoretical upper bound of the wind speed that could be attained by the hurricane event at a specific time frame.

4 Methods

An exploratory data analysis is first conducted with the dataset, such as cleaning the data, creating a summary dataframe of the dataset and visualising the relationships between variables. With the summary dataframe and the exploratory data analysis as a guide, the maximum wind speed and the presence or absence of rapid intensification of storms are modelled with varied methods.

4.1 Exploratory Data Analysis

Exploratory data analysis is an essential step when working with any kinds of data. By performing exploratory data analysis, the underlying structure of the data could be revealed, and hence disclosing how the data interact with each others. In general, to perform exploratory data analysis, the data has to be first collected and imported, cleaned and tidied. Then uni-, bi- or multi-variate plots are created, or some models are built to dig into the structures of data, which favors the data modelling afterwards. For instance, the relationships between variables and the shape of data could be disclosed during the exploratory data analysis.

In this project, the exploratory data analysis is started with data cleaning, which includes but not limited to removing missing values and reshaping the data. After that, a summary dataframe is created and the correlation between variables is visualised by a correlation plot. Guided by the correlation plot, the distribution of some major variables are inspected and the relationships between correlated variables are visualised.

4.1.1 Data cleaning

The objective of data cleaning is to ensure the data is useful and functional. To be precise, a useful and functional data is accurate, consistent and complete, has correct format and high readability, and free or errors and redundancy.

In this project, the following steps are executed to clean the data:

- Examine each file: Each file will be read and examined. In this step, the shapes of the data and how the data are presented would be revealed, guiding the remaining steps of the data cleaning process.
- Unify the shapes: The data are reshaped as hurricane events by observations. By unifying the shapes of data, the coding workload could be reduced.
- Missing value handling: Any missing value in the data will be replaced with the logical constant "NA" in R.

- Normality test and transformation: The Shapiro-Wilk test is applied to test the normality of the dataset. As the distribution of the input data may affect the performance of the models, if any non-normality are detected, Box-Cox Transformation will be applied to normalise the data. In this project, the basic formulation of Box-Cox Transformation is used to transform the variables if necessary, where a variable y is transformed with the following equation:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

  defined for $y > 0$ [2].

- Data subset: As the aim of this project is to investigate to what level the maximum wind speed and the presence or absence of rapid intensification of a strong storm could be predicted with the information of storms at early stage, a data subset is created by first filtering out storms with wind speed $<$ 50 knots throughout their lifetimes, and all data before the time frame where wind speed $\geq$ 50 knots are extracted for each storm left in the dataset respectively, whereas the maximum wind speed data are extracted from the original dataset.

4.1.2 Data summarisation

After cleaning the data, a summarisation dataframe is generated for each variable, where the rows correspond to storms, while the columns are the statistical summarisations. The summarisation dataframes are necessary as the original form of the data are time series data. As this project does not aim to attempt to build a time series forecasting model, the statistical summaries are needed to describe the behaviour and characteristics of storms during certain period of time, and at the same time reducing the dimensions of the data.

In the dataset, the Day, Hour, Month and Year data are considered as categorial data and therefore mode is used to summarise these data. The rests are considered as numerical data, hence mean, minimum, maximum, standard deviation and range of value are used to summarise them. These summarisation dataframes are then concatenated to create the summary dataframe.

In addition, a column is created to store ”1” or ”0” to represent whether there is rapid intensification or not in the lifetime of each storm according to the definition of rapid intensification given in section 2.

4.1.3 Data visualisation

To investigate the relations between variables, a correlation plot is created with the mean, minimum and maximum value of latitude, longitude and wind shear data, minimum and maximum of potential intensity data, and the minimum pressure, maximum wind speed and duration of storms. Univariate analysis is first performed to inspect the distribution of maximum wind speed, maximum potential intensity, maximum wind shear, minimum pressure and duration. A bi-variate boxplot of maximum wind speed against year of occurence is plotted to determine whether the Year data should be excluded for training the model or not, since the number of independent variables will increase a lot if Year data is included. Multivariate analysis is then performed to study the relationships between relatively strongly correlated variables.

4.1.4 One-hot encoding month data

To use the summary dataframe as the input of the models, the month data has to be encoded as it is a categorical variable. To avoid creating numerous independent variables, the month are first grouped as seasons with December, January and February being Winter, March, April and May being Spring, and similarly, June, July and August are grouped as Summer and the rests are grouped as Autumn [14]. After that, the seasons are one-hot encoded and the month data is replaced with 4 columns of binary data which indicates the season of occurrence of the storms.

4.2 Modelling Maximum Wind Speed

Three regression models are implemented to model the maximum wind speed of storms, namely Linear Regression model, Partial Least Squares Regression model and Random Forest Regression model. These models are selected due to several reasons. The linear regression model is the best regression model to start with as it is easy to implement, and the model reveals structures of the dataset and the relationships between the independent variables and the response variable. The partial least squares model identifies the major components for relating the response variable and the independent variables, and there is no hyperparameter to tune with in this model. The random forest model is one of the most popular modern machine learning algorithm due to its efficiency in training and fine-tuning and its compatibility to different kinds of data. The data used in this section is a clone of the summary dataframe and the transformed summary dataframe without the summaries of speed data, and the response variable is the maximum speed of each storms.

The data and the response variable are then split into train set and test set with the ratio of 7:3. The train set is used to build the models, and the models will predict the maximum wind speeds from the test set. The results from each model are evaluated and compared.

4.2.1 Linear regression model

Linear regression model is a supervised learning model for modelling continuous outcome. It describes the relationships between independent variables and response variable by fitting a line to the observed data. In this project, multiple linear regression is implemented as more than two independent variables from the dataset are used for predicting the maximum wind speed. The multiple regression model is defined as follow:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

where y is the predicted value, β_0 is the y-intercept of the best fit line, $\beta_1, \beta_2, ..., \beta_n$ are the regression coefficients corresponding to independent variables $X_1, X_2, ..., Xn$ respectively and ϵ is the model error term.

The model is evaluated by the adjusted R-squared value, F-statistic and p-value. The adjusted R-squared value is a number between 0 and 1, which describes how much variance are explained by the independent variables in the model. A R-squared value closer to 1 means the response variable is well explained. A much larger than 1 F-statistic with p-value < 0.05 indicates that at least one of the independent variables in the model is related to the response variable. Similarly, the significance of independent variables are represented by a p-value respectively, with p-value < 0.05 being significant. Apart from evaluating the model with the above numerical values, the diagnostic plots for linear regression analysis are plotted to review whether the model meets the following assumptions:

- Linearity: The relationship between response variable and independent variables should be linear.
- Independence: The independent variables should not be highly correlated with each other.
- Normal distribution: The residuals should be normally distributed.
- Homoscedasticity: The variance of residuals should be consistent across all predicted values.

In this project, the forward stepwise regression method is used to optimize the linear regression model. The method constructs a linear regression model by iteratively adding significant independent variables to the model until the model stops improving. By the forward selection method, the number of independent variables is reduced while the performance of the model can be maintained.

4.2.2 Partial least squares regression model

In contrast to the hard assumptions in linear regression model, partial least squares model is said to be a soft model where it has less restrictions, for example it allows multicollinearity in the independent variables. Although partial least squares model has less strict assumptions about the data, linear relationships between variables and influential outliers should be aware of. The model is able to predict the response variable by identifying the underlying common structure between the response variable and the linear combinations of the independent variables [11]. To predict the response variable, the partial least squares model look for independent variables matrix T which models a set of predictors X and the response variable y simultaneously. The relation between T, X and the predicted response variable Y could be written as:

$$X = TP^T$$

and

$$Y = TBC^T$$

where P and C are the weights for T to model X and to predict Y respectively, and B is a diagonal matrix. By rewriting the above relations, the predicted response variable Y is expressed as a regression model:

$$Y = TBC^T = XB_{PLS}$$

where B_{PLS} being the product of the weights matrices and the diagonal matrix [1].

To obtain the best performing model, the partial least squares model is built with the summary dataframe and the transformed summary dataframe respectively, and the optimal number of components is selected by applying the Elbow Method to the plots of root mean square error of prediction, mean square error of prediction and R-squared value against number of components.

4.2.3 Random forest regression model

Random forest regression model is an ensemble supervised learning model, where multiple decisions trees are grown during training step. A decision tree is a non-parametric supervised learning method which predicts the value of target variable by interpreting decision rules from the independent variables. In random forest regression, the predicted values from all decision trees are aggregated and averaged to cast a prediction. To increase robustness of a random forest, randomness is introduced to the model by growing each tree with a random subset of the independent variables and bootstrap sample of the dataset. The random forest regression model does not require the data to follow any specific distribution, and it handles skewed-data and multi-modal data as well as categorial data.

In this project, the random forest regression model is built with the summary dataframe and the transformed dataframe respectively with the default hyperparameters, and fine-tuned with grid search. The grid search is done by first defining a sequence of hyperparameters with step-sizes for the number of independent variables to be randomly selected, minimum node size when growing each tree and the sample size. After iteratively building models with all combinations of the hyperparameters in the grid, the combination with lowest prediction error is selected to construct the optimal model. The top 20 important variables of fine-tuned models are visualised.

4.2.4 Metrics

In this project, the performance of the regression models are evaluated by mean absolute error (MAE) instead of the more common root mean square error (RMSE). MAE and RMSE are defined as follow:

$$MAE = \sum_{i=1}^{n} \frac{|\hat{y_i} - y_i|}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y_i} - y_i)^2}{n}}$$

where $\hat{y_i}$ are the predicted values from the model and y_i are the observed values in the dataset.

In general, RMSE is the preferred metric in regression problems as it measures the standard deviation of prediction errors, which tells how spread the prediction errors are around the best fit line. However in this project, the maximum wind speed of storms could be large and unpredictable due to their complexity, and the prediction errors from these very strong storms are punished by the nature of RMSE. This is unfavourable as the objective of this part is to investigate to what extent the maximum wind speed could be predicted with the limited information, but not building a high performance model to predict strong storms. Therefore, MAE is selected to evaluate the regression models as it measures the average of unweighted prediction errors.

In addition to the quantitative measurement, the predicted values are plotted versus observed values for each model to observe their tendencies of predictions. The graphs are plotted with a line with 0 intercept and slope 1 to assist in evaluating the predictions.

4.3 Modelling Absence or Presence of Rapid Intensification

Two models are selected to model the absence or presence of rapid intensification of storms. Logistic Regression model and Random Forest Classification model are selected for the binary classification problem. The logistic regression model is a typical binary classification model which is easy to implement and does not require much tuning. The random forest model is implemented with same method as the one used in predicting maximum wind speeds, but the response variable is the absence or presence of rapid intensification instead. In this section, the data used is the summary dataframe and the transformed summary dataframe, and the response variable is the column of binary indicator for rapid intensification contructed in section 4.1.2. The data are first split as storms that rapid intensified and storms that did not rapid intensify. The subsets are further split into train and test set with the ratio 7:3, and then concatenated back, resulting a train set and a test set which consist of both type of storms. The train set is used to build the models and the models will predict results with the test set. The results from each model are evaluated and compared, and the relationships between the important variables inferred by the random forest classification model and the existence of rapid intensification are visualised.

4.3.1 Logistic regression model

Logistic regression model is a statistical model which can be used for binary classification problem. The major feature of a logistic regression model is that it predicts the class of an instance with the logistic function, which is a sigmoid function. The sigmoid function is defined as

$$\sigma(t) = \frac{1}{1 + e^t}$$

which maps any real value t to $(0, 1)$. In our case, the value t is the linear combination of the independent variables, which is:

$$t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

where $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the regression coefficients where $X_1, X_2, ..., Xn$ are the independent variables respectively and ϵ is the error term.

The logistic regression model further maps the output value from the sigmoid function to $\{0, 1\}$ with a probability threshold. For instance, the class of an instance is predicted to be 1 if the output of the sigmoid function is greater than the proabbility threshold. The probability threshold has a default value of 0.5, but the effect of setting it as 0.3 and 0.1 will also be investigated.

There are some assumptions for the logistic regression model. It requires little or no multicollinearity between the independent variables. In addition, the independent variables should have a linear relationship with the logarithm of the likelihood of the instance being the case.

The receiver operating characteristic curve (ROC curve) is plotted to visualise the performance of each model at all probability threshold. The closer the ROC curve is to the diagonal, the worse the model is. In addition, the area under curve (AUC) is also calculated to select the best version of logistic regression model.

4.3.2 Random forest classification model

The random forest classification model implemented in this project has the same structure to that used for predicting maximum wind speed. In general case, the random forest classification model is implemented as outputting the predicted class of the instances. However, the effect of setting different probability threshold for the predictions are investigated and therefore the random forest classification model is implemented as outputting the probabilities of the instances being class 1 or 0 respectively. And if the probability for an instance being classified as class 1 is greater than or equal to the probability threshold, it is classified as 1.

The random forest classification model is built with default value, and then fine-tuned with grid search. The random forest classification models are compared by the out-of-bag prediction errors, and the model with the lowest error is selected as the optimal model. The top 25 important variables of fine-tuned model are ranked and the relationships between the top few important variables and the existence of rapid intensification are visualised.

4.3.3 Metrics

Three probability thresholds which are 0.1, 0.3 and 0.5 are used to investigate the performances of the models. A probability threshold is a value for the classification models to determine whether the case should be predicted as true or not. That is, the probability threshold convert the soft classification algorithms into hard classifiers. If the probability returned from the model is greater than the threshold, than the model would predict the case as true.

To evaluate and compare the classification models, a confusion matrix is visualised to explore the precision, sensitivity and specificity of the model, where

- Precision is the ratio of true positives to total predicted positives, calculated by:

$$\frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

- Sensitivity is the ratio of true positives to total actual positives in the data, calculated by:

$$\frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

- Specificity is the ratio of true negatives to total actual negatives in the data, calculated by:

$$\frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}$$

5 Results

Following the methods outlined in section 4, an exploratory data analysis was conducted and the results were visualised. The models mentioned in section 4 were built and fine-tuned, and the performances were compared with the metrics described.

5.1 Exploratory Data Analysis

According to the methods described in section 4.1.1 and 4.1.2, the data were cleaned and a data subset consists of 2,145 strong storms was created. A summary dataframe with early phase data was constructed as shown in Table 1a. A column of the rapid intensification indicator was

Modal month	Modal year	Average latitude	Average pressure	Duration
9	1979	11.724	1001.885	88
9	1979	9.273	1000.807	134
7	1979	11.916	1002.096	94
9	1979	12.452	1002.925	102
6	1979	9.432	999.288	90
7	1979	27.140	1001.417	232
9	1979	25.341	1000.967	188
6	1979	16.301	1000.008	118
6	1979	13.059	1001.305	94
10	1979	12.831	1001.654	94

(a) Sample columns from the summary dataframe constructed by the statistical summarisation of data of storms at early stage.

Season 4	Average latitude	Average pressure	Maximum wind shear	Duration
1	5.633	501886.504	1.923	1.553
1	4.674	500806.840	3.598	1.578
0	5.704	502098.506	2.229	1.557
1	5.901	502929.466	2.447	1.562
0	4.739	499287.753	3.400	1.554
0	10.411	501417.573	4.226	1.603
1	9.924	500967.414	3.874	1.594
0	7.229	500007.805	2.986	1.571
0	6.121	501305.458	2.497	1.557
1	6.039	501655.336	3.116	1.557

(b) Sample columns from the summary dataframe after applying Box-Cox Transformation to the summary dataframe and one-hot encoding the months as seasons.

Table 1: Summary dataframe and transformed summary dataframe

added to the summary dataframe and 147 storms were identified as rapid intensified during their lifetimes according to the definition of rapid intensification given in section 2. Some of the variables were plotted in histogram to reveal the distribution of the data. A correlation plot was created and used in multivariate analysis, where the relationships between maximum wind speed and other variables were visualised.

Some major variables such as maximum wind speed, minimum pressure, maximum wind shear, maximum potential intensity and duration were plotted in histogram to reveal the distribution of these data. It was found that most data exhibited right-skewed distribution like Figure 2a, few were left-skewed like Figure 2b and only average and maximum potential intensity were close to normally distributed like Figure 2c.

A correlation plot was created to visualised the correlation between some major variables in the summary dataframe, for example the maximum wind speed, duration, the average, minimum and maximum of wind shear, pressure, potential intensity, longitude and latitude as shown in Figure 1. Focus was placed on the column of maximum wind speed as the relationships between the response variable and the independent variables were investigated before modelling. The correlation plot showed

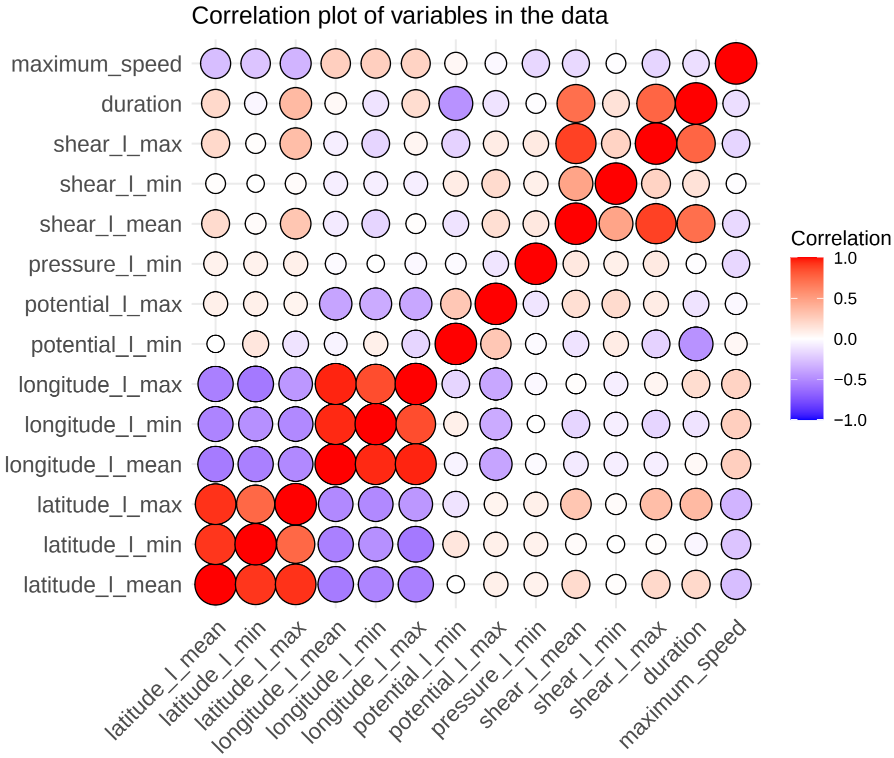
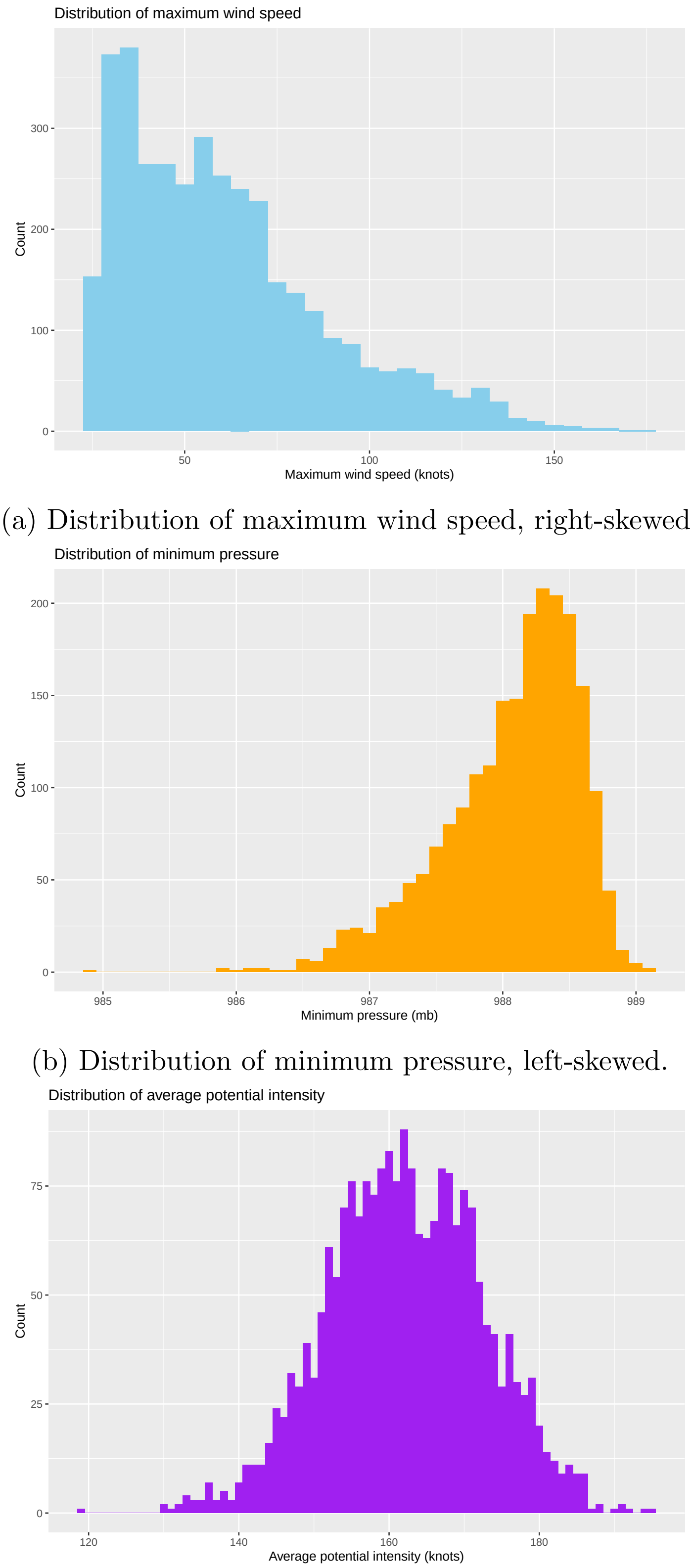


Figure 1: Correlation plot of major variables in the summary dataframe.



(a) Distribution of maximum wind speed, right-skewed.

(b) Distribution of minimum pressure, left-skewed.

(c) Distribution of average potential intensity, close to normal distribution.

Figure 2: Three types of distribution found in the data.

that none of the independent variables had strong correlation with the maximum wind speed, while some of the independent variables were relatively strongly correlated. Strong correaltions were found between duration and average wind shear, potential intensity and longitude, longitude and latitude, and different statistical summaries from same data.

A boxplot of maximum wind speed grouped by year of occurrence was visualised to determine whether the year data should be one-hot encoded and included or not. The graph is shown in Figure 3 where no trend or relationships were observed in the graph. Although the outliers in each year showed some kinds of wave pattern, the bodies of the boxplots stay at a similar level throughout all years. As a result, the year data were not included in models training.

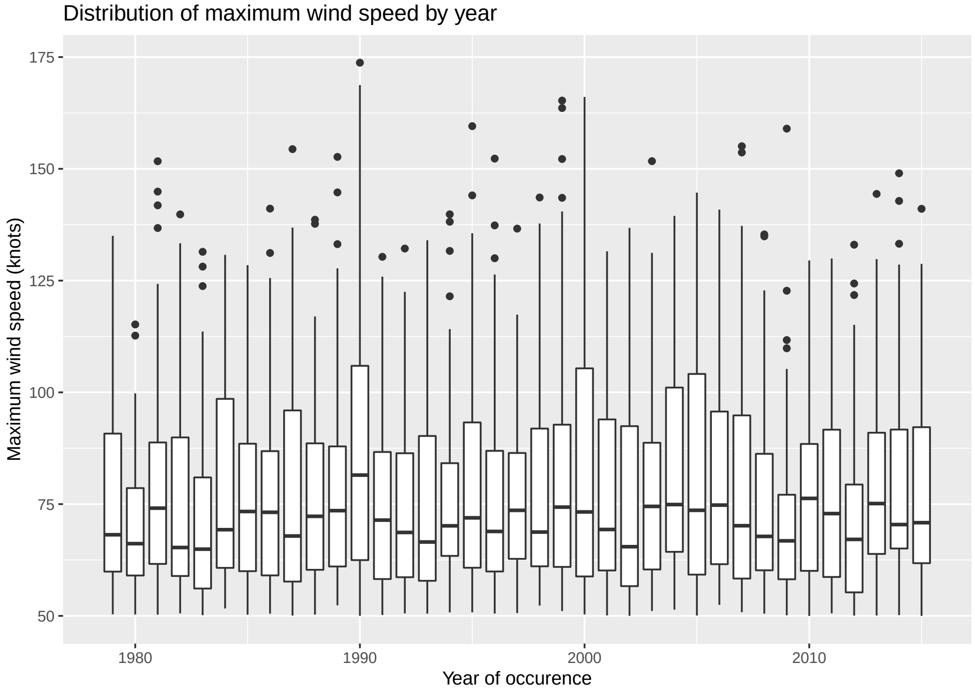
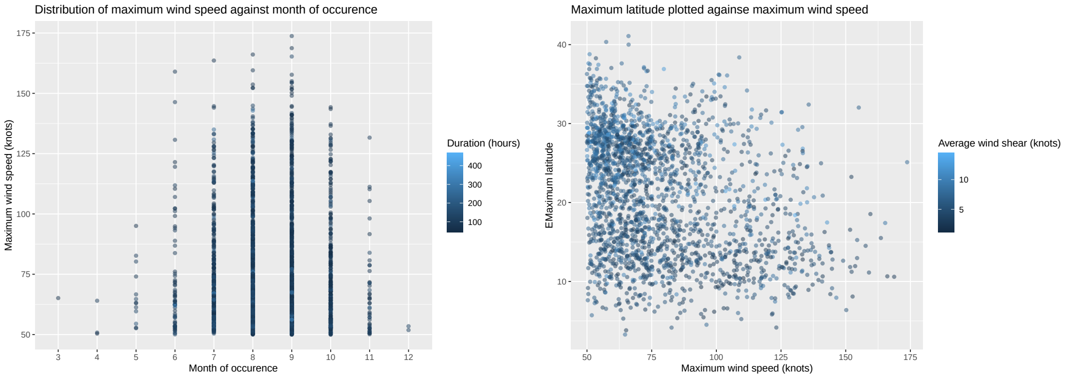


Figure 3: Boxplot of maximum wind speed grouped by year. No trend was observed.

In addition, scatter plots of independent variables versus maximum wind speed were plotted to observe their relationships. From Figure 9a, it was found that most of the storms occurred during July and October, which showed the existence of relationship between maximum wind speed and months. Although the other independent variables were proved to be correlated with maximum wind speed, most of them exhibited weak linear relationships with maximum wind speed, similar to the relationship between maximum latitude and maximum wind speed as demonstrated in Figure 9b.



(a) Distribution of maximum wind speed by occurrence month. Most storms were occurred between July and October.

(b) Relationship between maximum latitude and maximum wind speed, colored by magnitude of average wind shear. Very weak relationship observed.

Figure 4: Multivariate plots visualised

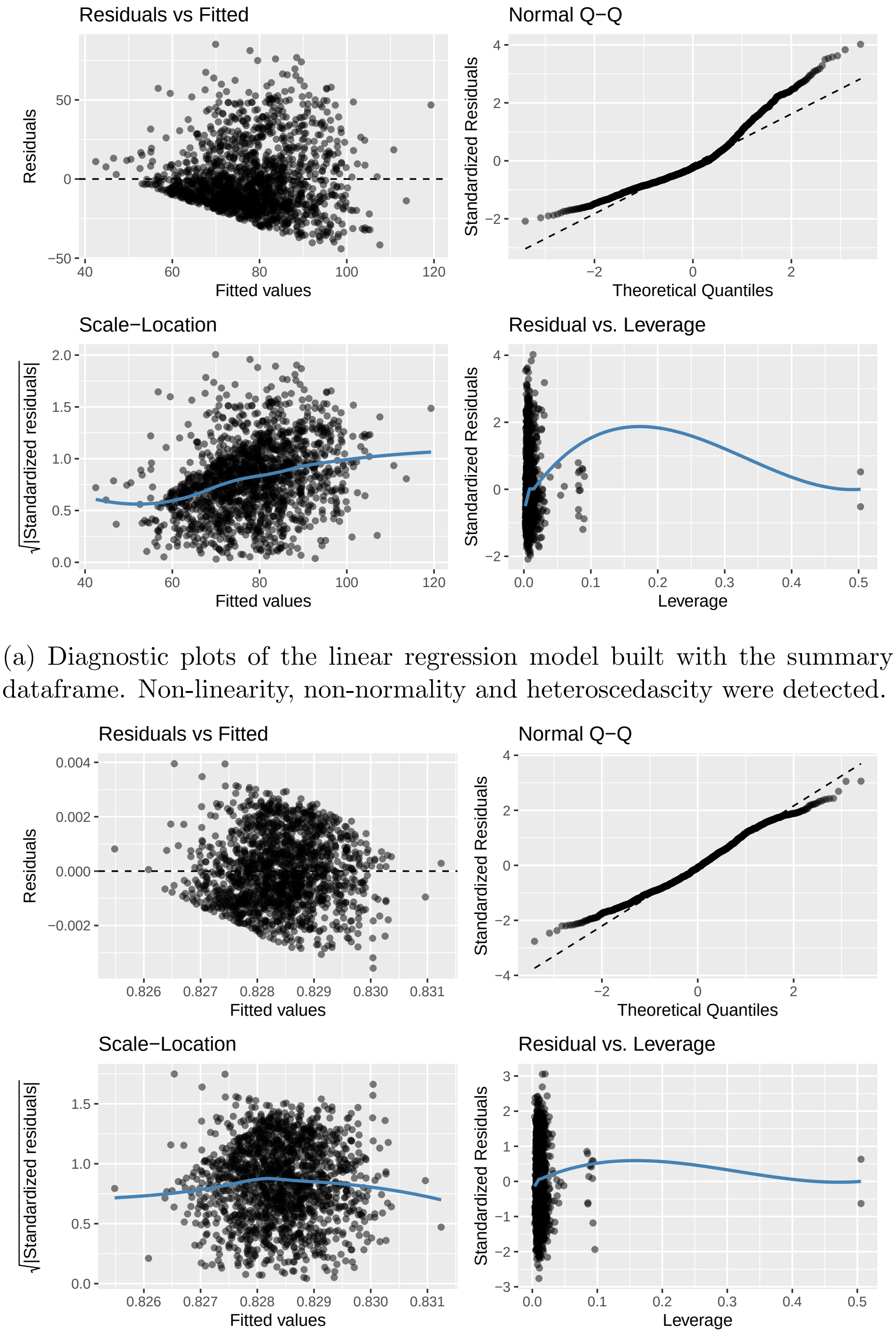
From the histograms for data distribution, non-normality were observed and hence the Shapiro-Wilk test was applied to the summary dataframe. With the threshold for p-value set as 0.05, it was reported that non-normality was detected in all independent variables except average potential intensity. For convenience, the whole dataframe except the season variables was cloned and transformed with Box-Cox Transformation as outlined in section 4.1.1. The transformed summary dataframe is shown in Table 1b.

5.2 Modelling Maximum Wind Speed

To model the maximum wind speed of storms in their life times with the statistical summaries of data in early stage, linear regression model, partial least squares regression model and random forest regression model were built with both the summary dataframe and the transformed summary dataframe respectively. The best performing models were compared and discussed.

5.2.1 Linear regression model

The linear regression model with forward stepwise selection was built with the summary dataframe and revised with the transformed summary dataframe. 13 independent variables were selected in the first model while 18 independent variables were selected in the second model.



(a) Diagnostic plots of the linear regression model built with the summary dataframe. Non-linearity, non-normality and heteroscedascity were detected.

(b) Diagnostic plots of the linear regression model built with the transformed summary dataframe. Non-linearity, non-normality and heteroscedascity were slightly improved.

Figure 5: Diagnostic plots for linear regression model were plotted to evaluate the models.

Common significance independent variables selected were average and maximum latitude, minimum and standard deviation of pressure. Duration was selected in the first model only, while the standard deviation and range of wind shear minimum longitude appeared in the second model only. Both models had large F-statistics with p-value smaller than 0.05, which indicate that they were significant model. But the adjusted R-squared values for both model were only 0.1905 and 0.2293 respectively, which means they explained only around 20% of the variations in the training data.

The diagnostic plots were plotted to reveal how well the models were. By observing Figure 5a, the residuals vs fitted plot suggested non-linear relationship between response variable and one or more independent variables as the instances were not spread around the horizontal dashed line evenly and equally. A curve was observed in the normal Q-Q plot instead of a straight line, indicating non-normal distribution in residuals. The shape in scale-location plot showed that the residuals were spread unequally, which means it violates the assumption of homoscedasticity.

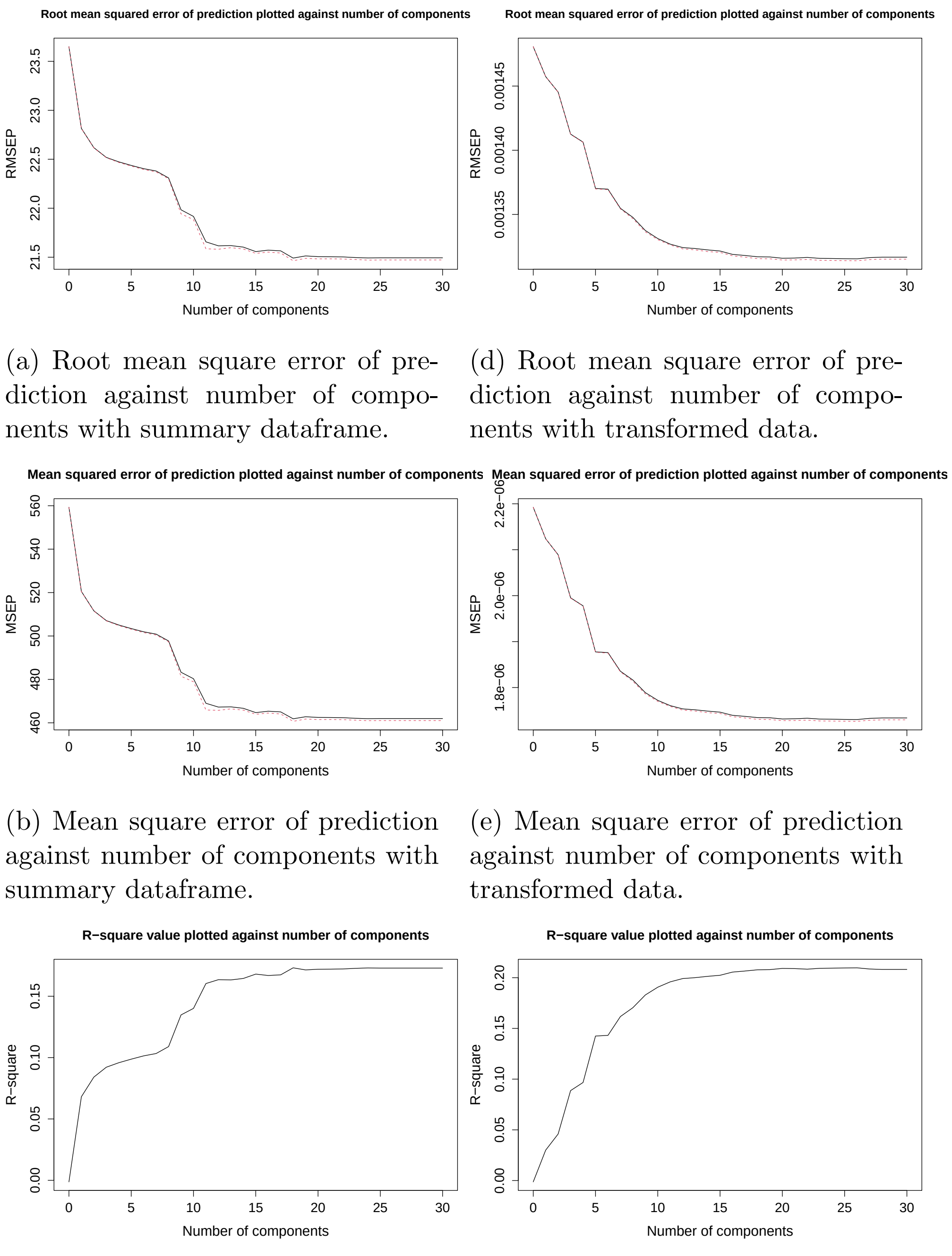
In contrast, with the transformed summary dataframe, the residuals follow the assumptions for linear regression model better as demonstrated in Figure 5b, with equally spread residuals in the residuals vs fitted plot and the scale-location plot, and a straighter line in the normal Q-Q plot. Hence it was selected as the optimal linear regression model obtained.

The test set from the transformed summary dataframe was passed to the optimal model and the predicted values were visualised in Figure 8a against observed values. To summarise, the model failed to predict most of the high maximum wind speed (≥ 100 knots) storms, and was better at predicting low maximum wind speed (< 100 knots) storm. The mean absolute error measured was 15.92 knots.

5.2.2 Partial least squares regression model

As described in section 4.2.2, the partial least squares regression model has softer assumptions than the linear regression model, hence similar performances were expected in the models built with the summary dataframe and the transformed summary dataframe respectively.

The model summaries showed that both models attained optimal results with 11 components, in terms of the decreasing rate of root mean square error of prediction, and the increasing rate of percentage variance explained. The optimal number of components was selected with the Elbow method by observing the metrics plots as shown in Figure 6. In particular, it was observed that the curves were smoother in the metrics plots from the model built with transformed data, compared to the staircase looking curve from that with the summary dataframe.



(a) Root mean square error of prediction against number of components with summary dataframe.

(d) Root mean square error of prediction against number of components with transformed data.

(b) Mean square error of prediction against number of components with summary dataframe.

(e) Mean square error of prediction against number of components with transformed data.

(c) R-squared value against number of components with summary dataframe.

(f) R-squared value against number of components with transformed data.

Figure 6: Metrics plots of partial least squares regression model built with the summary dataframe ((a)-(c)) and the transformed summary dataframe ((d)-(e)) respectively.

The partial least square regression model built with the transformed summary dataframe performed slightly better than the other one with a mean absolute error of 16.20 knots. The predicted values were plotted against observed values in Figure 8b. The pattern was found to be similar to that from the linear regression model.

5.2.3 Random forest regression model

Baseline random forest regression models was built with the summary dataframe and the transformed summary dataframe with default hyperparameters, where 500 trees were grown with 10 randomly selected independent variables each, and the minimum node size which controls the depth of each tree was 5. After fine-tuning the model, it was found that the fine-tuned random forest regression model built with the transformed dataframe performed the best among all random forest regression models.

The model was fine-tuned with the following grid search with 100 combinations in total:

- Number of independent variables was searched in the range from 4 to 28, with a stepsize of 4
- Minimum node size was searched in the range from 2 to 10, with a stepsize of 2
- Sample size is searched by 4 values, 0.4, 0.55, 0.7 and 0.85

The optimal hyperparameters from grid search were found to be 28 independent variables, minimum node size of 6 with 0.85 sample size. The top 20 important variables are shown in Figure 7, where maximum latitude was inferred to be the most important variable, followed by standard deviation, minimum and average of pressure and average of latitude. The mean absolute error of the prediction from the optimal model was 17.00 knots. The predicted values were plotted against observed values in Figure 8c. It was observed that the model failed to predict storms with high maximum speed ($\geq$ 100 knots), while having similar predictions to the previous two regression models for storms with maximum wind speed $<$ 100 knots.

5.2.4 Model comparison

Table 2 was created to compare the optimal models side-by-side. The attributes that used to compare the models were the mean absolute error of predictions, the numbers of predictors used in building the models and

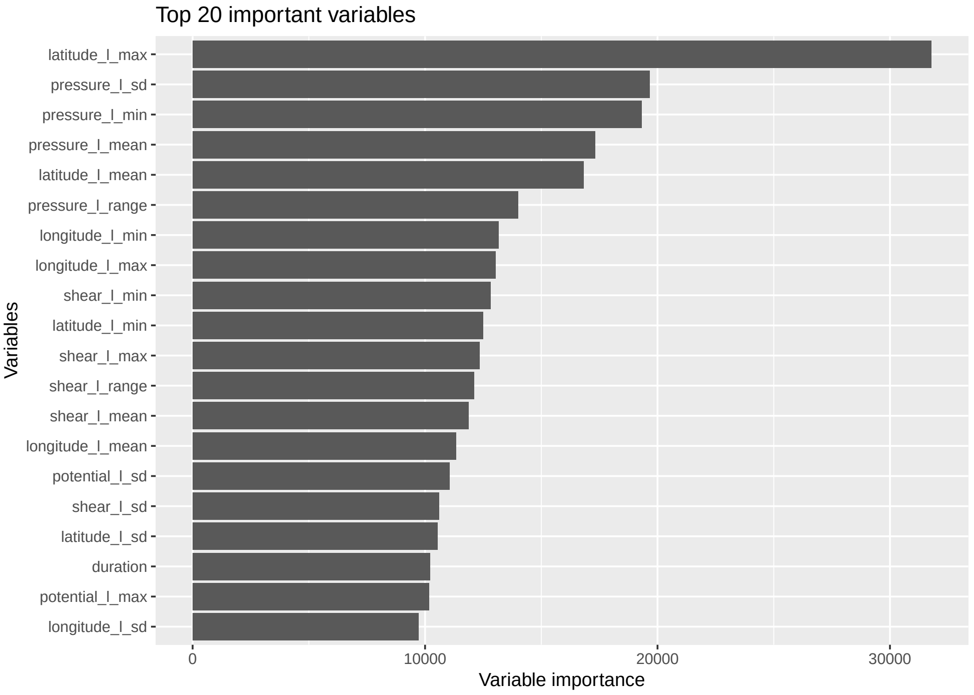


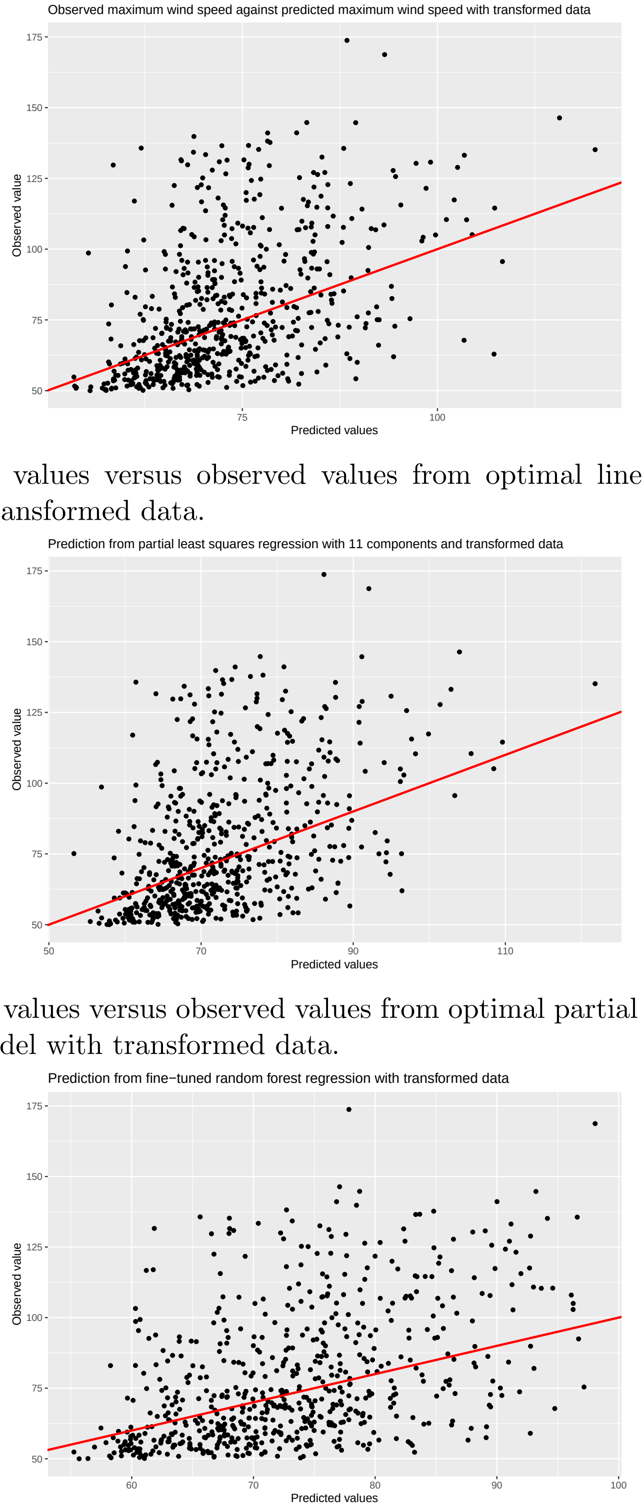
Figure 7: Top 20 important variables inferred by the fine-tuned random forest model with the transformed dataframe.

the percentage variance explained by the models. These attributes were selected to compare the models as they showed the accuracy and dimensions of the models and how much variation the model could explain with the independent variables.

With the transformed summary datframe, the linear regression model

Models			
Type of model or attributes	Linear regression model	Partial least squares regression model	Random forest regression model
Mean absolute error	15.92	16.20	17.00
Number of predictors	18	11	28 (each tree)
Percentage variance explained	22.93%	21.67%	19.63%

Table 2: A side-by-side comparison table of the optimal regression models. It was found that the linear regression model performed the best among three.



(a) Predicted values versus observed values from optimal linear regression model with transformed data.

(b) Predicted values versus observed values from optimal partial least squares regression model with transformed data.

(c) Predicted values versus observed values from optimal random forest regression model with transformed data.

Figure 8: Predicted values versus observed values from the optimal models. None of them was able to predict high maximum wind speed.

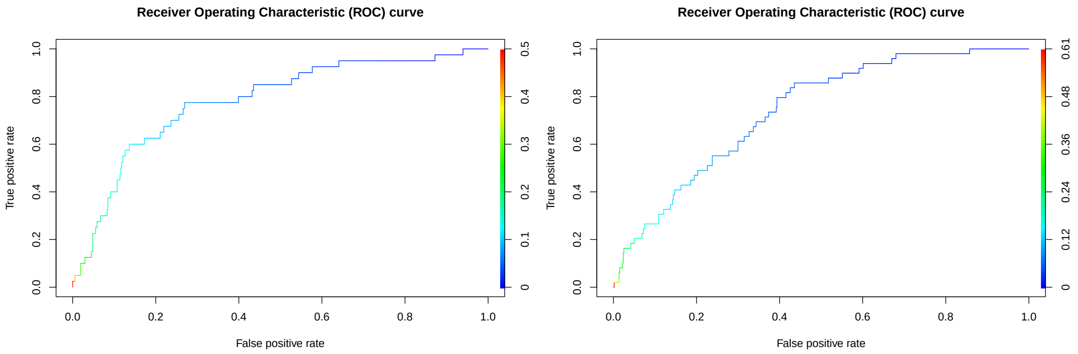
outperformed the other two models in terms of mean absolute error and percentage variance explained, while the partial least squares regression model was the best among three models with regard to the number of predictors in the model. The random forest regression model ranked the last with this particular dataset.

5.3 Modelling Absence or Presence of Rapid Intensification

To model the absence or presence of rapid intensification in the lifetime of storms with the statistical summaries of data in early stage, logistic regression model and random forest classification model were built with both the summary dataframe and the transformed summary dataframe respectively. The models were evaluated with the receiver operating characteristic curve (ROC curve) and confusion matrix. The best performing models were compared and discussed.

5.3.1 Logistic regression model

The logistic regression model with forward stepwise selection was built with the summary dataframe and revised with the transformed summary dataframe. 10 independent variables were selected in the first model while 11 independent variables were selected in the second model.



(a) ROC curve of logistic model built with the summary dataframe, the AUC is 0.7818.

(b) ROC curve of logistic model built with the transformed summary dataframe, the AUC is 0.7415.

Figure 9: ROC curves plotted from the logistic models. The model built with the summary dataframe performed slightly better at lower probability thresholds.

Similar to the linear regression models, the logistic regression models shared some common independent variables such as maximum latitude, maximum and standard deviation of pressure and the average wind shear. Duration appeared in the model with the summar dataframe only, while maximum speed was selected in the second model only. The AUC was computed and the results showed that the logistic regression model built with the summary dataframe performed better, with an AUC of 0.7818 compared to the AUC of 0.7415 from the model built with the transformed summary dataframe.

The ROC curves are visualised in Figure 9, where the curves from the model with transformed data appeared to be closer to the diagonal line, and was verified by the AUC. It was observed that for probability thresholds > 0.2 or < 0.1, the performances were similar but the first model was slightly better for probability thresholds between those values.

5.3.2 Random forest classification model

Baseline random forest classification models was built with the summary dataframe and the transformed summary dataframe with default hyperparameters, where 500 trees were grown with 11 randomly selected independent variables each, and the minimum node size which controls the depth of each tree was 10. After fine-tuning the model, it was found that the fine-tuned random forest classification model built with the transformed dataframe was the optimal model, which had an out-of-bag prediction error of 0.0599, lower than the out-of-bag error of 0.0643 from the model built with the summary dataframe.

The model was fine-tuned with the following grid search with 160 combinations in total:

- Number of independent variables was searched in the range from 4 to 32, with a stepsize of 4
- Minimum node size was searched in the range from 2 to 10, with a stepsize of 2
- Sample size is searched by 4 values, 0.4, 0.55, 0.7 and 0.85

The optimal hyperparameters from grid search were found to be 4 independent variables, minimum node size of 10 with 0.4 sample size. The top 25 important variables are shown in Figure 10, where minimum pressure and average wind speed were ranked as first 2 important variables had almost the same importance, followed by average and range of pressure, maximum wind speed and maximum latitude. From the visualised relationships between first two important variables, minimum pressure and average wind speed and the existence of rapid intensification in Figure 11, it was observed that most class 1 instances, which were the instances with presence of rapid intensification had the same value as the class 0 instances, those without rapid intensification during their lifetimes, in these two independent variables. Even with the most important variable inferred from the random forest classification model, it is hard to separate class 1 instances from class 0 instances with a line or a S-curve, which explained the performance of the models. And the results showed that none of the independent variables had an exceptional high importance, which implies none of them is a determining feature for rapid intensification to happen showed .

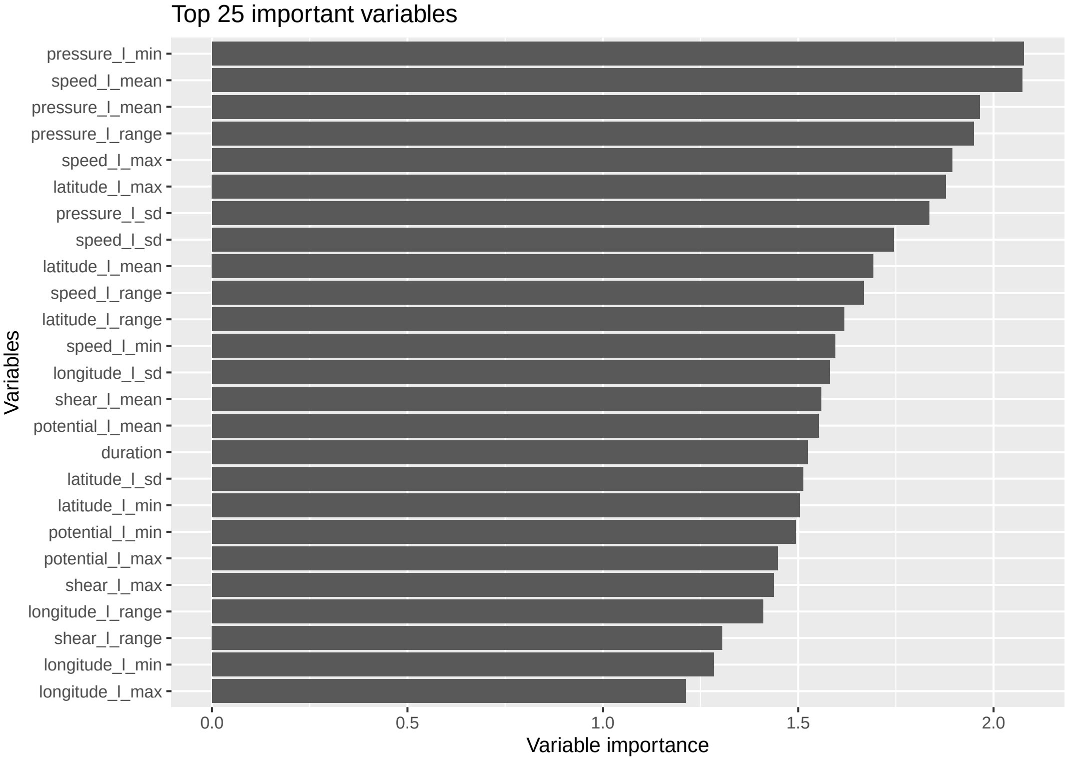
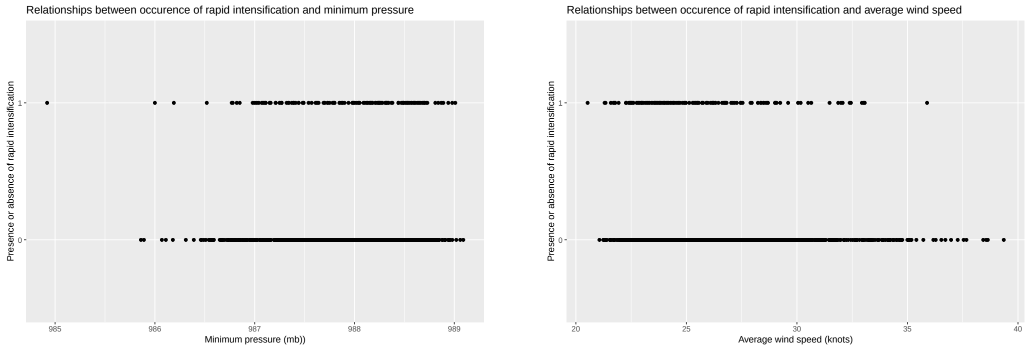


Figure 10: Top 25 important variables inferred from the optimal random forest classification model with transformed dataframe. None of the independent variables had exceptional high importance.

5.3.3 Model comparison

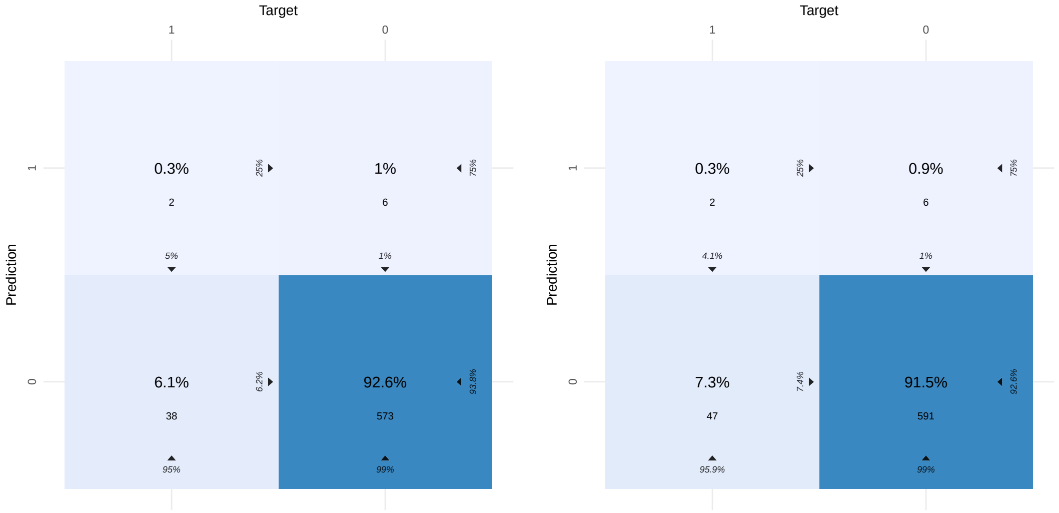
The models were compared with the confusion matrix, together with some metrics for classification models, which were precision, sensitivity and specificity. The following tables showed the performances of the models at probability threshold 0.1, 0.3 and 0.5.



(a) Minimum pressure was the most important variable interpreted by the random forest classification model. It is hard to differentiate the classes with the value of minimum pressure by a line or curve.

(b) Average wind speed was the second most important variable interpreted by the random forest classification model. Most class 1 instances were stacked on class 0 instances.

Figure 11: Visualied relationships of first two important variables from the optimal random forest classification model and the existence of rapid intensification.



(a) Confusion matrix of the logistic model built with the summary dataframe at probability threshold = 0.3.

(b) Confusion matrix of the random forest model built with the transformed summary dataframe at probability threshold = 0.3.

Figure 12: Multivariate plots visualised

At threshold = 0.1, both models had similar precision, the logistic model had lower sensitivity but higher specificity while the random forest model was the other way round, which means the logistic model was better at founding out the true negative (absence of rapid intensification) cases and the random forest model was better at classifying the true positive (presence of rapid intensification) cases. At threshold = 0.3, although both model had precision of 25%, but the sensitivities were very low, indicating that the models classified very few cases as positive and identified very few true positive cases. It was supported by the high specificities resulted from classifying most cases as negative as visualised in Figure 12. At threshold = 0.5, both models classified almost all cases as negative. The logistic model had a 100% precision simply due to no

Probability threshold = 0.1			
Type of model or metrics	Precision (%)	Sensitivity (%)	Specificity (%)
Logistic model	17.7	62.5	80.0
Random forest classification model	17.9	67.3	74.7

(a) Performances of models with probability threshold = 0.1.

Probability threshold = 0.3			
Type of model or metrics	Precision (%)	Sensitivity (%)	Specificity (%)
Logistic model	25.0	5.00	99.0
Random forest classification model	25.0	4.1	99.0

(b) Performances of models with probability threshold = 0.3.

Probability threshold = 0.5			
Type of model or metrics	Precision (%)	Sensitivity (%)	Specificity (%)
Logistic model	100.0	2.5	100.0
Random forest classification model	0.0	0.0	100.0

(c) Performances of models with probability threshold = 0.5.

Table 3: A comparison table for the performances of the classification models measured with precision, sensitivity and specificity.

false positive case was identified, which was observed from the very low sensitivity. No case was identified as true in the random forest model at this threshold.

By analysing the confusion matrix with different probability thresholds, it was concluded that the classification models were not performing well in identifying the storms that have rapid intensification during their lifetimes with the data provided.

6 Discussion

6.1 Key findings

The key results from this project are highlighted as follows. The strategy of using a summary dataframe to compress the time series data was successful, where the dimensions of data were greatly reduced and only columns of data were concerned instead of 10 dataframes with hundreds of columns.

From modelling the maximum wind speed, a key takeaway is the importance of a normally distributed dataset, as better performances were observed from the models built with the transformed dataset. A shortcoming of the prediction models are that they failed to predict the high maximum wind speeds. In the classification models, it was good to see the models had certain predictive power if the probability threshold was set low.

One possible explanation for the failure of the models is that the features provided in the summary dataframe were weakly related to the response variables and were not the determining features.

6.2 Future works

To extend this research, a few aspects could be reviewed. The first aspect is the dataset. For statistical modelling or machine learning, it is always better to have more observations. More observations means that there will be more variations in the data, and the pattern of behaviours would usually be more obvious. Apart from the number of storms to be included in the dataset, the number of storms that experience rapid intensification should also be concerned. In this dataset, only 147 storms were with rapid intensification. To allow the models to learn the characteristics of storm that will rapid intensify, more instances are needed.

The second aspect is to predict the response variable with other approaches. As the data provided were time series data, time series model such as autoregressive integrated moving average model (ARIMA model) would be an appropriate starting point. A more machine learning-wise attempt would be the long short-term memory model or hybrid models, which are proven to be a good approach and are already applied in predicting rapid intensification of storms or hurricane forecasting [17, 6].

The third aspect is to extend the dataset. For example, apart from the features in the dataset such as wind speed, pressure, geographic location and wind shear, some studies suggested that the sea surface temperature is one the conditions for rapid intensification to happen [8], which this dataset lacks.

7 Conclusion

This project investigated to what extent the maximum wind speed and the presence or absence of rapid intensification of storms during their lifetime could be predicted with statistical models and machine learning methods, using the statistical summaries of data of storms at early stage. Linear regression models, partial least squares regression models and random forest regression models were implemented to predict the maximum wind speed, while logistic regression models and random forest classification models were implemented in predicting presence or absence of rapid intensification.

It was reported that linear regression model performed the best when predicting maximum wind speed by having the lowest mean absolute error of 15.92 and highest percentage variance explained of 22.93%, compared to the partial least squares regression model and the random forest regression model. For predicting the presence and absence of rapid intensification, both logistic regression model and random forest classification model failed to correctly identify the true positive cases with the probability threshold being set at 0.5, but had reasonable performances with around 18% precision, 62.5% and 67.3% of sensitivity and 80% and 74.7% of specificity respectively when the probability threshold was set at 0.1.

The results indicated that with the dataset provided, the approaches used in this project were not the best approaches in predicting the maximum wind speed and the presence or absence of rapid intensification of storms during their lifetime. It was observed that the regression models only had average performance in predicting maximum wind speed from the dataset, and the features of storms in the dataset were not the determine features for the classification models to identify rapid intensification.

References

[1] Hervé Abdi and Lynne J. Williams. Partial least squares methods: Partial least squares correlation and partial least square regression. *Computational Toxicology*, pages 549–579, 2012.

[2] Manuele Bicego and Sisto Baldo. Properties of the box–cox transformation for pattern classification. *Neurocomputing (Amsterdam)*, 218:390–400, 2016.

[3] Samson Brand. Rapid intensification and low-latitude weakening of tropical cyclones of the western north pacific ocean. *Journal of Applied Meteorology (1962-1982)*, pages 94–103, 1973.

[4] National Hurricane Center and Central Pacific Hurricane Center. Saffir-simpson hurricane wind scale.

[5] Hua Chen, Da-Lin Zhang, James Carton, and Robert Atlas. On the rapid intensification of hurricane wilma (2005). part i: Model prediction and structural changes. *Weather and forecasting*, 26(6):885–901, 2011.

[6] Rui Chen, Xiang Wang, Weimin Zhang, Xiaoyu Zhu, Aiping Li, and Chao Yang. A hybrid cnn-lstm model for typhoon formation forecasting. *GeoInformatica*, 23(3):375–396, 2019.

[7] Melissa Free, Marja Bister, and Kerry Emanuel. Potential intensity of tropical cyclones: Comparison of results from radiosonde and reanalysis data. *Journal of Climate*, 17(8):1722 – 1727, 2004.

[8] Charles R. Holliday and Aylmer H. Thompson. Climatological characteristics of rapidly intensifying typhoons. *Monthly Weather Review*, 107(8):1022 – 1034, 1979.

[9] Jyothi Lingala, Sudheer Joseph, and Suneetha P. Role of environmental factors in rapid intensification and weakening of cyclone ockhi (2017). *Earth and Space Science Open Archive ESSOAr*, 2020.

[10] James M Murphy, Doug M Smith, Rosie Eade, David Fereday, Holger Pohlmann, Nick J Dunstone, and Adam A Scaife. Skilful multi-year predictions of atlantic hurricane frequency. *Nature geoscience*, 3(12):846–849, 2010.

[11] Dante Pirouz. An overview of partial least squares. *SSRN Electronic Journal*, 10 2006.

[12] Christopher M. Rozoff and James P. Kossin. New probabilistic forecast models for the prediction of tropical cyclone rapid intensification. *Weather and forecasting*, 26(5):677–689, 2011.

[13] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.

[14] Kim Rutledge, Melissa McDaniel, Santani, Hilary Hall, Tara Ramroop, Erin Sprout, Jeff Hunt, Diane Boudreau, and Hilary Costa. Season, May 2022.

[15] Joshua B. Wadler, Jun A. Zhang, Robert F. Rogers, Benjamin Jaimes, and Lynn K. Shay. The Rapid Intensification of Hurricane Michael (2018): Storm Structure and the Relationship to Environmental and Air–Sea Interactions. *Monthly Weather Review*, 149(1):245–267, January 2021.

[16] B Wang and X Zhou. Climate variation and prediction of rapid intensification in tropical cyclones in the western north pacific. *Meteorology and atmospheric physics*, 99(1-2):1–16, 2007.

[17] Qidong Yang, Chia-Ying Lee, and Michael K. Tippett. A long short-term memory model for global rapid intensification prediction. *Weather and forecasting*, 35(4):1203–1220, 2020.