Treball de fi de màster

Títol: Hurricane Seasonal Forecast Modelling

Cognoms: Sarmanto

Nom: Natalia

Titulació: Màster en Ciència i Tecnologia de la Sostenibilitat

Director/a: Miquel Sànchez-Marrè

Data de lectura: Octubre 2020

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Institut Universitari de Recerca en Ciència i Tecnologies de la Sostenibilitat

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This research got its start from the support of the company Hurricane Unwinder. The principal idea was to create a hurricane intensity forecast analysis of the company's satellite picture-based prediction. Error analysis of individual storms was conducted based on statistical verification between various variables used in the model. These results were compared with a similar study done on the prediction models of the National Hurricane Centre's intensity forecast. Also, a more comprehensive overall prediction analysis was conducted to identify the specific events where the company's model performs better than other official forecasts.

In addition to the intensity forecast analysis, a second part was intended to include in this master's dissertation. A seasonal forecast model for Hurricanes in the Atlantic basin was planned, and later created.

The research started in parallel by doing both tasks. Over time the complexity of both investigations appeared to be more than expected, and since no direct link could be found between the two studies, one part was decided to be left out from this thesis. No seasonal forecasting had previously been done in the company, therefore the main interest laid in it. Due to this, the initial plan describing both tasks was discarded and only the seasonal forecasting model was included in the thesis. Even though not detailed in this analysis, the intensity forecast error analysis was continued. Some parts of the process of the intensity error analysis can be found in Annex A.5.

ABSTRACT

In this master's dissertation, the aim is to create a seasonal hurricane forecast model based on the best possible random forest technique. Open data on Global Sea Surface Temperature (SST), Lower Stratospheric Temperature, Quasi-biennial Oscillation and the number of hurricanes in a season in the Atlantic basin over the years of 1978-2015 are used and processed. A Data Science approach is followed to understand variable behaviour among the ones studied, using various pre-processing and dimension reduction techniques. An innovative concept in the field of time series is used to introduce additional covariables into the analysis by creating new SST variables according to geographical areas. Two SST databases are compared and tested to find the best data-source for the models built. Machine learning algorithms are used for the creation of different models, which are to be tested. Hyperparameter optimisation is studied for tuning the algorithms in the best possible way, where the results are verified with versatile evaluation metrics. Lastly, the random forest results are tested against models built on other machine learning algorithms. The results show that the tuned random forest algorithm gives the best prediction accuracy when dimensions are reduced using principal component analysis (PCA), whilst some further analysis is suggested to do on gradient boosting in future studies. Results offer a novel approach to understand the relationship between these variables and to improve the ability to predict seasonal hurricanes in the future.

Keywords: Hurricanes, seasonal forecasting, data science, machine-learning, statistics, data pre-processing, hyperparameter model tuning.

RESUM

En aquesta tesi de màster, l'objectiu és crear un model de previsió estacional d'huracans basat en el millor model possible de la tècnica de Random Forest (RF) . Les dades obertes utilitzades són: Temperatura de la superfície marina global (SST), baixa temperatura estratosfèrica, oscil·lacions quasi-biennals, i, el nombre d'huracans en una temporada a la conca Atlàntica entre els anys 1978 i 2015. Seguim un enfocament basat en Data Science per tal de poder entendre el comportament de les variables que hem estudiat, utilitzant diverses tècniques de preprocessament de dades i tècniques de reducció de la dimensionalitat. Utilitzem un concepte innovador pel que fa a sèries temporals que ens permet introduir co-variables addicionals en l'anàlisi creant noves SST variables segons zones geogràfiques. Es comparen i avaluen dues bases de dades de SST són per així trobar la millor font de dades pels models creats. Utilitzem algoritmes d'aprenentatge automàtic per a la creació de diferents models que després seran avaluats. L'optimització d'híper-paràmetres s'utilitza per a modelar els algoritmes de la millor manera possible, on els resultats són verificats amb diferents mètriques d'avaluació . Finalment, s'avaluen els resultats dels models de RF contra els altres models creats amb altres tècniques d'aprenentatge automàtic. Els resultats mostren que els models de RF són els que

prediuen amb més exactitud quan es redueixen dimensions usant tècniques d'anàlisi de components principals (PCA), mentre que deixem una porta oberta a futurs estudis sobre l'ús del model de *gradient boosting*. Els resultats mostren una nova manera d'entendre la relació entre aquestes variables i millora l'habilitat de predir huracans estacionalsen el futur.

Paraules clau: Huracans, pronòstic estacional, data science, aprenentatge automàtic, estadística, pre-processament de dades, ajust d' híper-paràmetres d'un model

RESUMEN

El objetivo de esta tesis de máster es crear un modelo de previsión estacional de huracanes basado el mejor algoritmo posible de la técnica de Random Forest (RF). Los datos abiertos utilizados son: Temperatura de la superficie marina global (SST), baja temperatura estratosférica, oscilaciones casi-bienales y, el número de huracanes en una temporada en la cuenca Atlántica entre los años 1978 y 2015. Seguiremos un enfoque basado en Data Science para así poder entender el comportamiento de las variables estudiadas, utilizando diferentes técnicas de re-procesamiento de datos y técnicas de reducción de la dimensionalidad. Utilizamos un concepto innovador en series temporales que nos permite introducir co-variables adicionales en el análisis creando nuevas variables SST según zonas geográficas.Se comparan y evaluan dos bases de datos de SST para así encontrar la mejor fuente de datos posible para nuestros modelos. Utilizamos algoritmos de aprendizaje automático para la creación de diferentes modelos que después serán evaluados. La optimización de híper-parámetros se utiliza para modelar los algoritmos de la mejor manera posible, donde los resultados son verificados con distintas métricas de evaluación . Finalmente, testeamos los resultados de los modelos de RF contra los otros modelos creados usando técnicas de aprendizaje automático. Los resultados muestran que los modelos de RF son los que predicen con mayor exactitud cuándo se reducen dimensiones utilizando técnicas de análisis de componentes principales (PCA), mientras que dejamos una puerta abierta a futuros estudios sobre el método de *gradient boosting*. Los resultados muestran una nueva manera de entender la relación entre estas variables y mejora la habilidad de predecir huracanes estacionales en el futuro.

Palabras clave: Huracanes, pronóstico estacional, data science, aprendizaje automático, estadística, pre-procesamiento de datos, ajuste de híper-parámetros de un modelo

1. INTRODUCTION

Tropical cyclones (TC) are complex phenomena that affect many lives around the globe. TCs are formed over warm tropical oceans and are large-scale rotary storms that can have an outer circulation that reaches up to 1000 km from the storm centre (Montgomery & Farrell, 1993). The *tropical cyclone* is the generic name for *tropical storms* having a peak wind higher than 17 ms^{-1}. Once the wind increased to more than 33 ms^{-1}, the name of the storms depends on its location. In the Caribbean and the eastern Pacific, these are called "*hurricanes*" while in the western North Pacific the storm of similar intensity is called "*typhoon*". In the North Indian Ocean same kind of storms with surface winds exceeding 33 ms^{-1} are called with the name "*severe tropical storms*" while in the region around Australia they are called "*severe tropical cyclones*" (Meteorology & Program, 2009). With *storm intensity*, it is meant the maximum wind speed average during 1 minute at the height of 10 meters.

The formation of storms has been studied for many years in the literature (Shapiro & Goldenberg, 1998; Montgomery & Farrell, 1993; McBride & Zehr, 1981). Still, it has been shown that it is tough to clarify the exact formation process as there are so many different components affecting. TCs start with local disturbance, and if there are enough external factors favouring the formation of a storm, it might become a self-sustained storm. The process described is called *tropical cyclogenesis*. As explained above, it is a complex system to understand the factors contributing to the formation of tropical cyclogenesis, but some necessary conditions have been identified. Gray (1968) highlighted six features that have been considered the minimum requirements:

1. A minimum sea surface temperature of 26°C to a depth of 60m
2. An increased mid-troposphere relative humidity (700hPa)
3. Limited instability in the atmosphere
4. An increase in the lower troposphere relative vorticity
5. At the genesis site a weakened vertical shear of the horizontal winds
6. Located min 5° latitude away from the equator (mainly a condition for more significant and more intense storms)

The *tropical storms* have in common their structural element. In all tropical storms, an eye, eyewall, boundary layer inflow, rainbands, upper-tropospheric outflow and a cirrus cloud shield are found. A general scheme of a tropical storm is depicted in figure 1.1. The circulation flow of the storm depends on whether it is located in the Northern Hemisphere or the Southern Hemisphere. In the Northern Hemisphere, they rotate counter-clockwise and in the Southern Hemisphere clockwise.

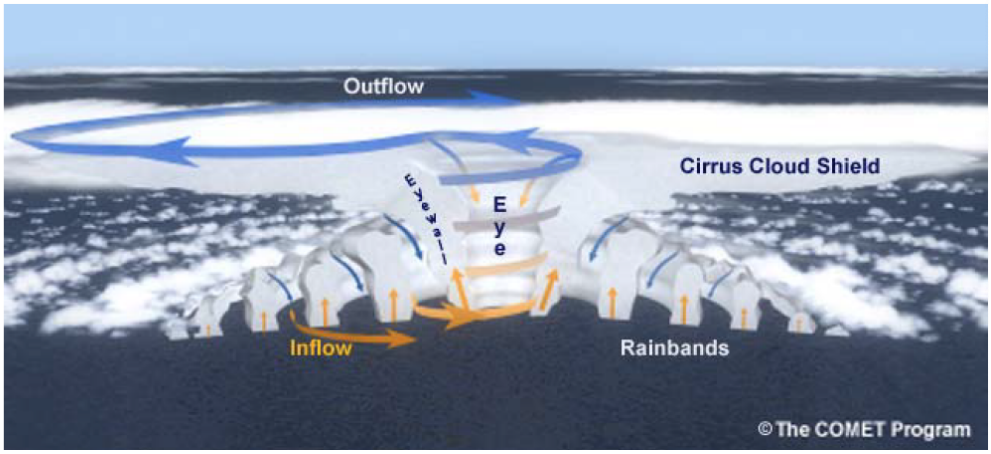


FIGURE 1.1 TROPICAL STORM SCHEME. SOURCE: (COMET, 2009)

The classification method of the cyclones depends on in what basin they take place. In the Atlantic areas and the eastern Pacific, a scheme called *the Saffir-Simpson scale* is used. The scale categorized the storms in 5 classes where hurricanes of category 3-5 are considered major or intense hurricanes. Even if the number of category 3-5 hurricanes is pretty low, the total damage they cause is almost 85% of all damages caused by storms with landfall (Meteorology & Program, 2009). Annex A.1 describes in more detail the different classification categories of the Saffir-Simpson scale. This classification method highlights the importance of the *storm intensity*, and for this reason, a reliable *intensity forecast* enables better preparation at possible landfall. Figure 1.2 shows the tracks and the intensities of the tropical storms between the years 1851 and 2006.

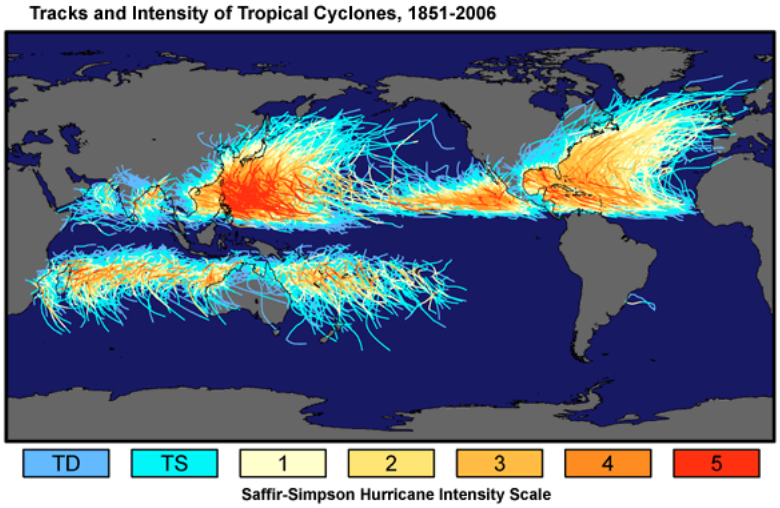


FIGURE 1.2 TRACKS AND THE INTENSITIES OF THE TROPICAL STORMS BETWEEN THE YEARS 1851 AND 2006.

A high-intensity TC can have destructive impacts with increased fatalities and immense capital losses. 2017 remains as one of the most intense hurricane seasons in the Atlantic basin with the most severe

damages until today. Total damages of over 294.92 billion dollars and 3364 fatalities took place, and the storm Maria was the deadliest one of the three major hurricanes with 2900 lives lost (NOAA, 2020a). The damage of a TC correlates strongly with the intensity of the storms, and to minimize the damages, TC forecast has been developed over the years as better preparation could save significant impacts on the damages occurred. The TC *track forecast* is widely developed and can give quite reliable results today, but the *intensity forecast* lags behind with more significant errors in the predictions (DeMaria & Kaplan, 1999).

The Finnish company Hurricane Unwinder recognized this shortcoming in the weather forecasts and has developed a machine learning tool based on high-resolution satellite images to better forecast the TC intensities globally. The different data streams contain information from various sources that the machine learning models continuously processes. The algorithm is built by using a recurrent-neural-network (RNN) architecture, with a convolutional neural network (CNN) processing all time steps in it. Also, the company has realized the lack in *seasonal forecasting* of tropical storms.

Seasonal forecasting of tropical cyclones is a widely researched topic. Every year many areas around the globe are affected by these highly intensive storms, and the cost of done damage is immense. The globe can be divided into seven different basins. Each basin has its monitoring centre, and they are divided on a geographical base. Figure 1.3 demonstrates the tropical cyclone areas.

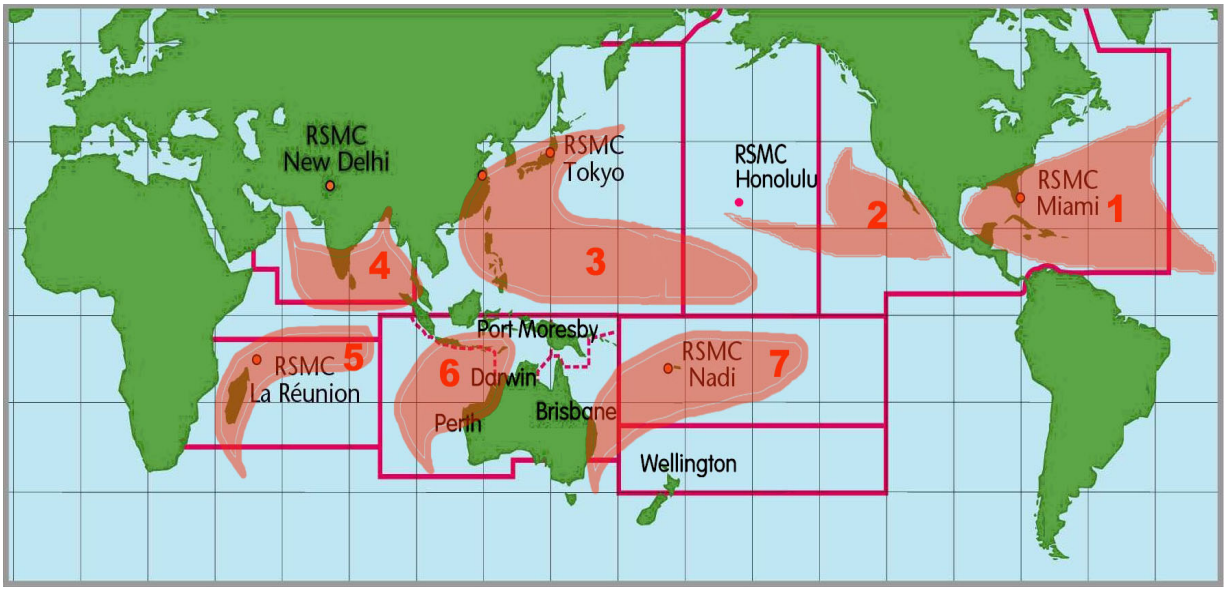


FIGURE 1.3 THE TROPICAL CYCLONE BASINS. SOURCE: HTTPS://ENACADEMIC.COM/DIC.NSF/ENWIKI/5166721

As every basin has its own TC season, when most storms take place, the seasonal forecast has tried to explore the number of cyclones that will take place during the active season.

The active season differs from basin to basin and table 1.1 describes in detail the active TC season of each, its responsible monitoring center, the name used for TC storms and their basin number corresponding to figure 1.3.

TABLE 1.1 THE DIFFERENT TROPICAL CYCLONE BASINS.

Basin	Monitoring center	Active Season	Name	Basin number
Northern Atlantic Ocean	United States National Hurricane Center	1st of June – 30th of November	Hurricane	1
Northeast Pacific Ocean	The Central Pacific Hurricane Center	15th of May – 30th of November	Hurricane	2
North-Western Pacific Ocean	-Joint Typhoon Warning Center -Philippine Atmosphericphysical and Astronomical Services Admin. -RSMC Tokyo-Typhoon Center	Active all year round (min activity in February and March)	Typhoon	3
North Indian Ocean	Indian Meteorological Department	Peaks in April and May, and in October and November	Cyclone	4
South-West Indian Ocean	Météo-France	17th of November to the 20th of April	Cyclone	5
Australian region	Australian Bureau of Meteorology	4th of January to the 23rd of May	Cyclone	6
South Pacific Ocean	Fiji and New Zealand's Meteorological Services	1st of July until the end of the year	Cyclone	7

As table 1.1 shows, the different basins vary quite much when taking into consideration the active times. The activity of each basin is an essential factor when creating a seasonal forecast as it reflects on the months one should use in the research. Since this research is limited to time and resources, the

analysis will focus on the Northern Atlantic basin. The regression will be done on the number of hurricanes that have taken place in the Atlantic basin and hence from now on we centre mainly on hurricanes in that area.

As the *hurricane season* takes place between June and the end of November, our input data will be focus on the months of January to July. In this way a prediction can be generated before the main season has started.

This focus is the same as what Hurricane Unwinder has done in their work in the research on the creation of a predictive model of seasonal forecasts. The company has identified an opportunity of using Random Forest regression as the algorithm in the prediction model, and for that reason, it will be the base for when creating the model in this research.

Not only the basins make a difference, but also other methods have been studied and compared. It has been found that there exists various methods using statistical, statistic-dynamic and dynamic techniques. Many factors impact the number of TCs taking place in a season, and it is a topic that still has a lot of potentials to be improved (P. Klotzbach *et al*., 2019). Factors that most often are taken into account when creating a seasonal forecast in the Atlantic basin are, for example, the current *El Niño*[1] effect and *La Niña*[2] conditions. Other often-used variables are rainfall amounts in Africa, sea surface temperatures, the lower stratosphere winds, and wind tendencies and atmospheric pressure in the Caribbean area. Also, factors as longitude and latitude have significant importance regarding tropical storms and their creation. The majority of the storms are usually formed between 5° and a maximum of 30° poleward due to warmer sea temperatures and favouring winds. Majority of storms (up to 87%) in, for example, the North Atlantic are formed within 20°N of the equator (Vega & Binkley, 1993).

As seen below from figure 1.4, the mean tracks of the storms are located close to the equator, but never crossing it. The storms never cross the equator due to the Coriolis force[3] being zero at the equator. The TCs need the Coriolis force to be able to have the spinning motion and hence if it zero the storm will weaken out rapidly. Due to the force, the storms tend to move towards north from the equator on the Northern Hemisphere and move toward South in the Southern Hemisphere. As demonstrated in figure 1.4, the TC basins are connected, and it should not be forgotten that a possible

[1] *El Niño* phenomenon refers to a large-scale interaction between the ocean and the atmosphere linked to a periodic warming in the sea surface temperatures in the Pacific Ocean (NOAA, 2020b).

[2] *La Niña* phenomenon refers to below-average sea surface temperatures in the Pacific (NOAA, 2020b).

[3] *The Coriolis force* deflects the wind to the right in the northern hemisphere and to the left in the Southern hemisphere. It explains why the winds in low-pressure areas blow anticlockwise in the northern hemisphere and clockwise in the high-pressure areas. Without the force air would move directly from high pressure to low pressure area. This is the force that drives the circulation of TCs (Met Office, 2020)

correlation exists between other basin sea temperatures and the formation of hurricanes in the Atlantic's.

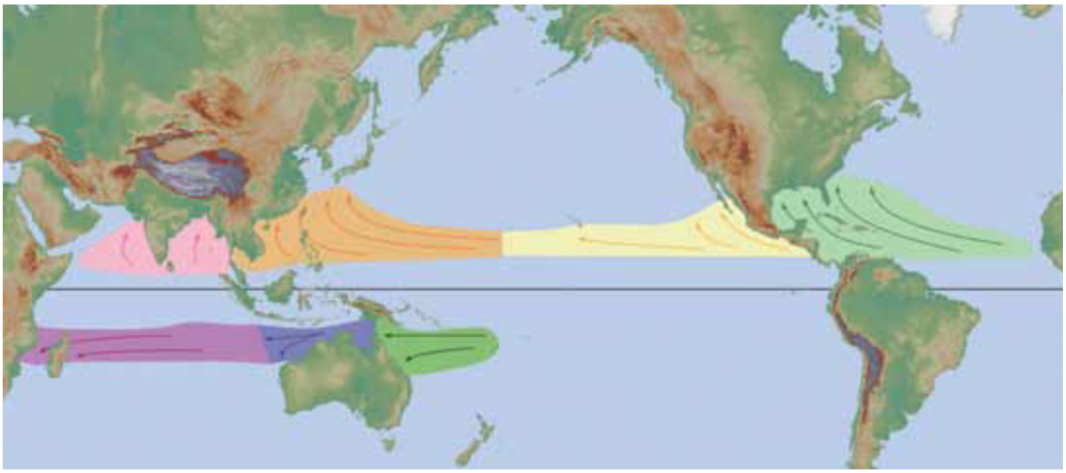


FIGURE 1.4 TROPICAL CYCLONE FORMATION REGIONS WITH MEAN TRACK. SOURCE: (NHC, 2017)

This research will focus on identifying a model that can with as little error as possible predict the number of hurricanes taking place in the Atlantic basin based on different climatological variables.

2. BACKGROUND

In this chapter, previous related works will be analyzed. In addition, the main machine learning and statistical techniques used in the data science modelling process will be explained. Furthermore, the available information to be used, will be detailed.

2.1 PREVIOUS RESEARCH

2.1.1 TROPICAL CYCLONE PREDICTION

Tropical storm track forecasting is the prediction type of TCs that has developed the most over the years. This might be because it has been of high interests over the years, as in this way a prediction of landfall can be made, and the affected regions can take anticipatory actions. Governments face a complicated trade-off between risking people life and a very costly false alarm. This has increased the research in the field of minimizing TC track prediction error. Regnier (2008) describes that using a stochastic model of storm movement created from track history, a model to explore the relationship between lead time and track error could be discovered in the Atlantic basin. The research shows that if the track uncertainty can be reduced to less than 10% of failing, it would be enough for an evacuation to take place. At the current state, 76% of cases might be of false alarm due to high error in hurricane track prediction.

The National Hurricane Center (NHC) has issued TC track forecast since the 1970s. In 2000, McAdie & Lawrence described that the NHC forecast had improved for all 24, 48 and 72 h forecasts. The NHC uses the "best-track" model to state the forecast result, and the expected error describes the forecasting difficulty. Factors like wind, pressure, sea surface temperature, air temperature, ocean current and the Coriolis force (earth rotational force) are all variables affecting the track forecast. Both statistical and dynamical models have been created (Roy & Kovordányi, 2012).

The Met Office started creating TC track forecasts in the late 1980s by analysis of mean sea-level pressure forecasts. The best forecasts were obtained over sea areas, as there are no obstacles to take place (Heming, 2017).

Intensity forecasts of tropical storms have not developed at the same pace as the track forecast of the TCs, and many studies have been conducted to increase the accuracy (Rogers et al., 2006). One of the most common methods is a statistical analysis based on the variable review from the *Statistical Hurricane Intensity Prediction Scheme* (SHIPS). De Maria and Kaplan generated a model based on climatological, persistence and synoptic variables using a multiple regression scheme. Their model is concentrated only on storms taking place over the sea in the Atlantic basin between the years 1982 to 1992. The model showed that by doing a jackknife procedure where the best track information is replaced by initial intensity and operation estimated of track, the results show that average intensity errors are 10-15% smaller than for models with only climatological and persistence variables (DeMaria & Kaplan, 1994). In 1999 the same model was updated and was incorporated into the National

Hurricane Centre (NHC) forecast model (DeMaria & Kaplan, 1999). By using the SHIPS model and including the previous season's forecast results, a statistical model was used for the North Pacific basin. The model is considered a "Statistical-synoptic" according to its creators. Later in 1997, it was modified into its current version, a "statistical-dynamic" model as it includes synoptic predictors from forecast fields. Further modifications have taken place as the land factor was incorporated into the model and therefore it has been treated as a separate model, Decay-SHIPS (DSHP) (DeMaria *et al*., 2014).

2.1.2 SEASONAL FORECASTS

The prediction of *seasonal tropical forecasts* started in early 1980 by the work of Neville Nicholls (Australian Bureau of Meteorology) predicting for the Australian region and William Gray (at Colorado State University) for the Atlantic region. These first forecasts predicted on base of *the statistical relationship* between different climate phenomena, e.g. *El Niño-Southern Oscillation* (ENSO) and tropical cyclones (Klotzbach *et al*., 2019). The ENSO effect is still today widely used, and many models use statistical comparison when creating seasonal forecasts. *Dynamical forecast modelling* is another widely used method as it predicts the amount of TCs in a season based on the interaction between the tropical cyclone and its surrounding. This method has generated reliable forecasts (Camargo *et al*., 2007) but to create these models large and powerful computers are needed. Further, an excellent description of the structure of the TC and its surrounding environment is essential, which in many cases can be challenging. A third model, a *Hybrid model*, that is a mix of statistical and dynamical forecasts is extensively used in seasonal prediction, and as the name indicates, combines features of both methods.

In 1984, Grey describes the correlation between the number of hurricane events in a season and the equatorial Quasi-Biennial Oscillation of stratospheric wind. The research found that there was a negative correlation between the frequency of hurricanes and the presence of equatorial winds at 30 Mb if they are of easterly direction (or are becoming more easterly during the season of storms). The effects of El Niño have been proven and found that usually, the seasonal activity is slightly above normal in non-El Niño years, and clearly above average when the equatorial stratospheric winds blow from a westerly direction (Grey, 1984). Other climatological factors have remarkable importance, such as the Atlantic Multi-decadal Oscillation and the West African monsoon system (Camargo et al., 2007)

A leader in the field of the provision of seasonal forecasts is the Colorado State University. The research group of the university has issued seasonal hurricane forecasts in the North Atlantic basin since 1984. The forecast is published in April, with updates in early June, July and August (P. Klotzbach *et al*., 2019). The 2020 forecast is based on 38 years of data using a statistical methodology together with the output from two statistical/dynamical models from the UK Met Office. Qualitative adjustments are made to the additional process that cannot be done using the algorithms (Klotzbach *et al.*, 2020). The NOAA seasonal forecast has been issued since 1998 and is based on statistical analogue regression. The model operates for the North Atlantic basin but also in the eastern North Pacific.

Other university-based seasonal forecasts are issued by City University of Hong Kong and the UTSR that is based at University College London. Both universities provide forecasts for the Western North Pacific and are of statistical forecast models. Additionally, the UTSR provides Atlantic forecasts too. The model used at the City University of Hong Kong was annual until 2011. Still, it stopped publishing annual forecasts as the number of TCs were decreasing and the forecast was over-predicting the number of storms. The Hong Kong Observatory started issuing experimental forecasts in the early 2000s every year in March. The model is a statistical-dynamic model, and it has been found to be a promising combination for the accuracy of seasonal forecast. Between the years 2009 and 2018, 60% of the forecasts issued turned out to be correct; the error was measured with the mean average error (Klotzbach *et al*., 2019).

Other non-publicly available forecasts are issued, e.g. by the European Centre for Medium-Range Weather Forecast (ECMWF), the Geophysical Fluid Dynamics Laboratory and Japan Meteorological Agency (JMA).

The seasonal forecast has been developing from statistical modelling to statistical-dynamic models, but more accurate models are still to be found. Not just the function used effects, but also the variables included and the length of the forecasting period. Murakami *et al.* (2016), showed that the correlation between observed and the predicted amount of TCs could range from 0.4 to 0.6 by changing the initial month of the forecast.

2.2 Techniques Used

2.2.1 Pearson Correlation Coefficient

Pearson correlation coefficient is a statistical metric that measures the direction and strength for a linear relationship between two variables. It is widely used in data analysis, classification, decision making, etc. The coefficient is measuring the linear dependence between the two random variables. It has been one of the first formal measures for correlation and is still today widely used. It indicates the strength of the variables measured (Zhou *et al.*, 2016). The method attempts to draw a line of best fit through the data of the variables, where the correlation coefficient indicates how far away these points are from the fitted line (Toppr, 2020). In chapter 4 the functionality of the technique will be discussed in detail.

The coefficient was developed by Karl Pearson, where the name has its origin from. However, the idea was introduced in the 1880s by Francis Galton. Pearson was an English mathematician and he found the first statistics department at University College in London, in 1911. Pearson introduced various other mathematical concepts (Yule & Filon, 1936).

2.2.2 Principal Components Analysis

The *principal component analysis* (PCA) was also introduced by Karl Pearson, in 1901. Later it was independently developed and named to its current name by Harold Hotelling in the 1930s. Most often, the Principal Component analysis is used in exploratory data analysis and for predictive models. According to Hotelling, there might be a smaller fundamental set of independent variables which determine the values of the PCA. Before this kind of components has been called *factors*, but Hotelling himself referred to them as *components* as they maximize the successive contribution to the total variance (Bartholomew, 2010). Therefore they are called to the "principal components".

Factor analysis has often been linked to the PCA as the idea behind the two methods are the same, to find the most important factor or components of the data set.

2.2.3 Decision trees and regression trees

Data mining has been of interest over many years and it has been researched both in the field of Statistics and Machine Learning. It is an important development as data mining concentrates on analyzing data from different perspectives. It analyses, modifies and processes the data into a useful form that further can be of high usage of the end-user (Fayyad *et al.,* 1996).

Before jumping to the random forest technique, it is good to check the background of a *decision tree*. A *decision tree* can be used both as a classifier or discriminant method (for predicting a categorical/qualitative variable) or as a regression or predictive method (for predicting a numerical variable), in the same way as a random forest can be used for both kinds of predictions. Figure 2.1 depicts a simple structure of a decision tree.

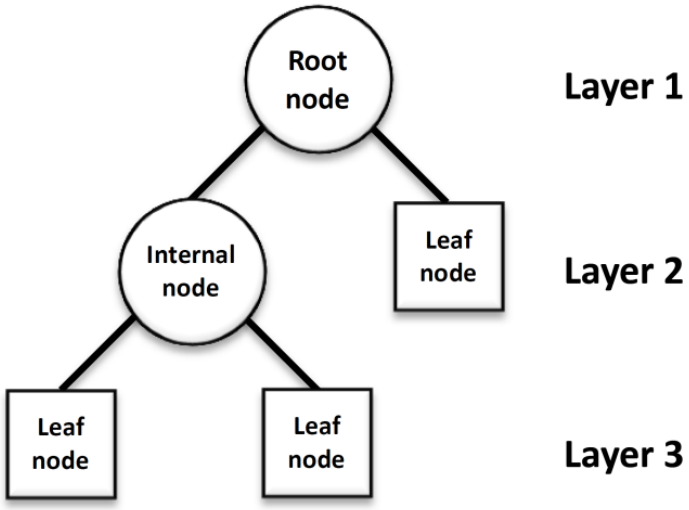


FIGURE 2.1 BASIC DECISION TREE. SOURCE: (CHIU ET AL., 2016)

As this research will be using a random forest as a numerical prediction model, the classification trees are not considered relevant to be discussed at this point. Decision trees with a target variable that can take continuous values are called *regression trees*. A *regression tree* needs a numerical response vector Y, which represents for each observation in the matrix X, the response value. The tree splits up the initial dataset in smaller subsets continuing to the root node of the tree. The splitting in regression trees is based on squared residuals minimization algorithms which tell that the expected sum variance of two nodes should be minimized (Timofeev, 2004).

2.2.4 RANDOM FORESTS

Early models of a *random forest* were started already in 1995 when Ho proposed a method that took into account the problem of limitations that the decision tree classifier had. The problem was that the tree could not grow to arbitrary complexity without sacrificing the accuracy of the unseen data (Ho, 1996). The proposed methodology was to generate classification trees randomly in random subspaces of the initial feature space. Two years later, Amit & Geman, (1997), developed a shape recognition approach that was a combination of shape features and tree classifiers. They used the assumption of an infinite number of features which resulted in the conclusion that there does not exist a classifier that can be based on a full feature set to be evaluated. This is because it was found that it was impossible to find out in advance what features were informative and which ones not (Amit & Geman, 1997). As a result, the standard decision trees were not enough, and an alternative was constructed. A small random sample of features for each node would constrain their complexity to increase with tree depth, hence grow multiple trees. The terminal nodes consisted of estimates of the previous distribution over shape classes.

Later on, Ho proposed a method that is close to the current model of random forest, as the method proposed would solve the dilemma between overfitting and the achievement of the maximum accuracy. The classifier built was based on decision trees that maintained the highest accuracy on training data, whilst it would improve the overall accuracy as it grows in complexity. The idea was that multiple trees were grown at the same time by pseudo-randomly selecting the subset of features that the vector components would be. This method was tested to be better than any other single-tree classifier or other forest construction theories (Ho, 1998).

Leo Breiman developed the most common version of a random forest in 2001. The model includes Breiman's bagging sampling method combining the random selection of features that initially were introduced by Ho and Amit & German to mix decision trees with a variance that is controlled (Fawagreh *et al.*, 2014). Additionally, Breiman introduced the classification and regression trees technique by adding randomness during the construction of the trees. Owing to this, the features selected in each node could be evaluated with the Gini impurity measure, like in CART, or using other measures like Information Gain criterion (Quinlan, 1986) or the Gain Ratio (like in C4.5) (Quinlan, 1993). In the original formulation of Breiman, using the Gini impurity measure, the feature having the highest reduction on the impurity measure is chosen as the split in the node. It is used to measure impurity of data, meaning how uncertain the occurrence of an event is. The general form of calculating the Gini impurity measure is:

$$Gini(t) = 1 - \sum_{j=1}^{L} P\left(C_j|t\right)^2 \quad (1)$$

Where *t* is a condition, *L* represents the number of classes that occur in the data set, and *C* shows the j-*th* class label in the used data set. The Gini index is only used at classification problems, and as we will be using random forest regression (and have no classes in our model), the Gini index is not used in thesis research for evaluation of the model.

The model presented in 2001 by Breiman has later been modified to even better boost its performance. Latinne *et al.* (2001), have been the first ones reporting to improve the model. They used a method based on McNemar's non-parametric test of significance, and the method determines in advance the minimum number of trees in the RF that should be used to get the best overall prediction accuracy. This gives a better prediction and also increases the speed of the operation. Other proposals have been identified like replacement of majority voting with a more sophisticated integration technique (Tsymbal *et al.*, 2006), changing the simple random sampling to pick the eligible subsets of features to split the nodes by weighted random sampling *(Amaratunga et al.*, 2008) and Bader-El-Den's and Gaber's approach to increasing the prediction accuracy by the usage of genetic algorithm (Bader-El-Den & Gaber, 2012).

2.2.4.1 HURRICANE PREDICTION AND RANDOM FORESTS

In tropical cyclone analysis, *machine learning* (ML) tools have already been used. Chen *et al*., (2020) described, in the chart on the figure 2.2, the different ML tools that are recommended to use for the various purposes in TC forecasts.

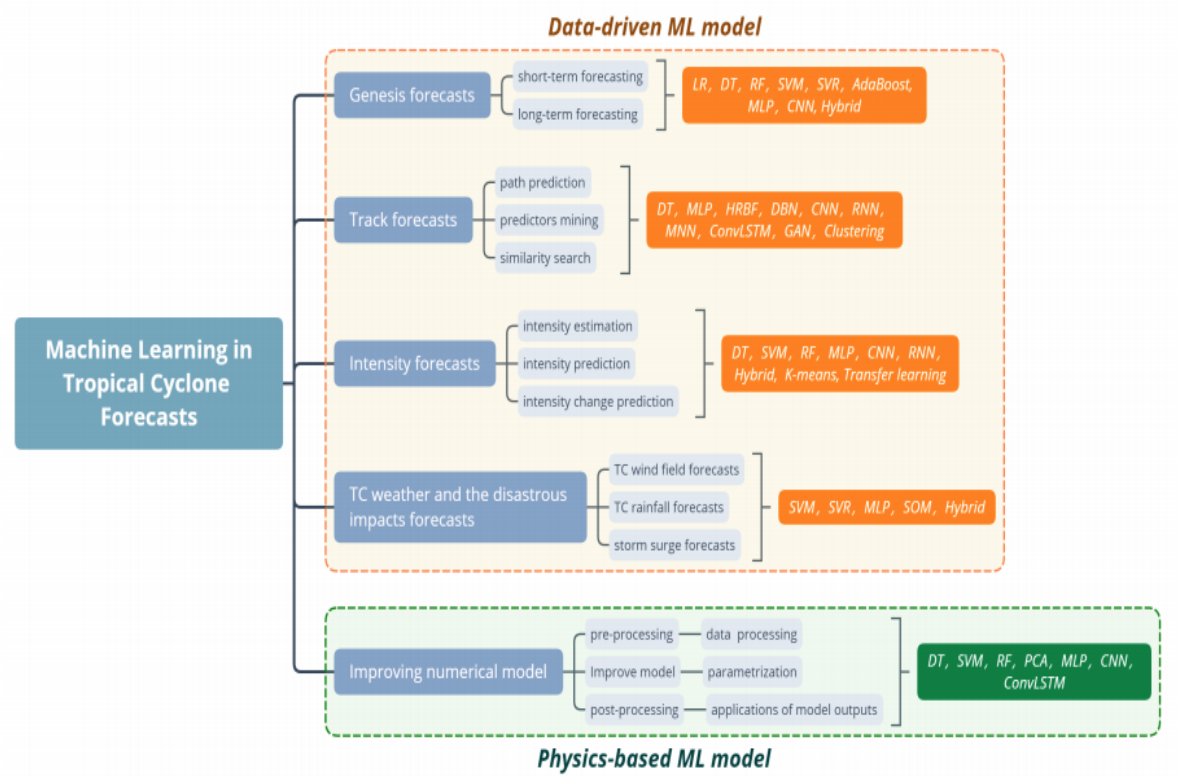


FIGURE 2.2 ML TOOLS FOR TC FORECASTING. SOURCE: (CHEN ET AL., 2020)

Mercer & Grimes (2017) studied the *rapid intensification* (RI) of storms as it still today remains a big forecasting challenge. RI, an increase in the maximum sustained wind of at least 30 knots in a 24-hours period is a phenomenon that not many forecasting models can identify, and it is still unsure what variables cause it. The study used random forest to check for the best artificial intelligence method in RI prediction.

Chen et al. explain that the genesis forecast is not as well defined as track forecast, and in genesis forecast the final goal is to create forecasts for a fixed region in real-time. Until now, ML is only capable of predicting whether the disturbances in the tropical oceans are evolving into tropical cyclones and also their seasonal frequency. For the seasonal forecasts mainly models built of logistic regression (LR), support vector machine (SVM), decision trees (DT) and random forest regressions are done.

Another approach on machine learning tools in the field of tropical cyclones was made by Kim *et al.* in 2019. They used the random forest algorithm in the detection of tropical cyclone formation using satellite data. The model showed to perform better when compared to other models both in the detection and in track forecasting.

Tan *et al.,* (2018), created a prediction scheme of tropical cyclone frequency in the Western North Pacific (WNP) using climatological values, a Lasso variable selection method and finally a random forest prediction model from the years of 1978-2011 as training sample and the years 2012-2016 as validation. The research found that the model is more practical and capable of generating reliable results for tropical cyclone frequency prediction in the WNP.

2.2.5 Linear Regression

Linear regression is a technique used to model the linear relationship between two or more variables. The idea is to fit the observed variables between the variables chosen. Ideally, the observations would be as close as possible to the fitted lines. Linear regression is widely used, and its main applications are within predictions and correlation analysis. Different variations of linear-based regression exist, such as logistic regression.

Linear regression is used in many generally used in biological, behavioural and social sciences to describe the relationship between variables studied. In statistics, linear regression is one of the fundamental learning algorithms due to its simplicity and well-known properties (Yan & Gang Su, 2009). This is one of the reasons it will be included in the research in order to verify the linearity of the results.

2.2.6 Gradient boosting regression

Gradient boosting is a machine learning technique that produces prediction models in the form of an ensemble of weak models, most often decision trees. Jerome H. Friedman developed the gradient boosting regression algorithm simultaneously with other more general function gradient boosting perspectives of Llew Mason, Jonathan Baxter, Peter Bartlett and Marcus Frean (Mason *et al.,* 1999).

Gradient boosting is a generalisation of AdaBoosting, which was the first algorithm to deliver promising boosting results. It improved the approach and introduction of ideas from bootstrap aggregation to further improve the model. Some versions of the gradient boosting have been developed, such as XGBoost and LightBoost (Brownlee, 2020)

2.2.7 Voting Regression

Voting is used in regression and is an ensemble machine learning algorithm. The idea behind the algorithm is to combine the other multiple regression models. It is most often used in the form of voting classifier, but can be applied in regression problems too. Particularly, if there is not one single model that performs well enough, the voting algorithm might yield a better result than any of them (Mangale, 2019). Zhang *et al.,* (2014), discussed better results being found with weighted voting. Optimization was done on the different weights and built on the results found, it was shown that the optimized weighted model generated the most accurate results.

2.3. AVAILABLE INFORMATION

This research is done together with the company Hurricane Unwinder, and as it has been of the company's interest in using certain variables, the research will be focused on these. It is out of the scope of the study to create a more in-depth analysis of the variables that possibly could be used, but during literature review the variables discussed in the following chapters were confirmed. It shall be noted that the databases were carefully chosen and different sources were compared. The sources with the most reliable background were chosen based on expert knowledge (author and company supervisor).

2.3.1 SEA SURFACE TEMPERATURE (SST)

The SST is an important factor when studying hurricane dynamics. SST can be measured over large ocean areas using satellite pictures. The SST varies a lot between seasons and ocean currents, but in general, it is warmer at low (absolute) latitudes and colder at high (absolut) latitudes.

SST and Seasonal forecast

As discussed earlier, TC formation needs warm ocean temperature in order to gain strength, and for this reason, there is a direct correlation with ocean temperature and hurricane formation. If the ocean is found to be warmer than the average temperature for that specific time of the year, it should be expected that above-average hurricane activity will take place. Traditionally SST is used in all kind of hurricane prediction models, and therefore there is widely available data.

In this research, two different SST data sets will be used. Both data sets are extracted from accredited databases and are of high-quality and reliable, that are extensively used in research. They will be compared against each other and the better data set will be used in the final forecast model.

HADISST V1.1

The Hadley Centre Global Sea Ice and Sea Surface Temperature describes globally the monthly SST and sea ice concentrations from 1871 to present. The data is available on the web page of the National Center for Atmospheric Research's (NCAR) data archive. The HADISSTV1.1 only contains the SST temperatures and this is the data that will be used for the thesis. The database gives a monthly time step with a 1° area grid and uses as input data the Met Office Marine Data Banks data that are in situ sea-surface observations and satellite-based estimates (Rayner *et a*l., 2003). Bucket corrections have been done on the gridded fields from 1870 until 1940. The data set of SST is given in Degree Celsius. The monthly mean SST are stored as the counts times 100 given in Celsius and missing data is set to -1e+30. The data array is 360x180 as it corresponds to the latitude and longitude (each grid represents 1°). The dimensions are time, 1791 time steps (each month), latitude, 180 (°), and longitude, 360 (°). It shall be noted that the coordinates cannot be compared to "real" lon and lat coordinates. Item (1,1) represents the value for the 1-deg-area at 179,5W and 89,5N. Item (360, 180) represents the value at 179,5E and 89,5S. The HADISST V1.1 data is restricted to only academic research and cannot be used commercially.

COBE SST2 and Sea-Ice

The COBE SST2 and Sea-Ice data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their Web site at https://psl.noaa.gov/ , is constructed using trends, interannual variations and daily changes. This is done using in situ SST and observations on the consternations of sea ice. The data is reported in monthly files starting from 1850 and is periodically updated. Similarly, as HADISST V1.1, the resolution of the COBE SST2 data is of 1° latitude and 1° of longitude. The coordinates of the global grid also have the spatial coverage of 89.5N to 89.5S, but different from *HADISST V1.1* the longitude starts at 0.5E and goes all the way to 359.5E.

All data is analysed with a theory-based analysis error to secure reliability. The missing grids are presented as values of either 1.e+20 or as -9.96921e+36 ("Climate Data Guide: SST data: COBE: Centennial in situ Observation-Based Estimates | NCAR," n.d.). Figure 2.3 presents the temperatures worldwide for the COBE SST2 dataset for August in 2012.

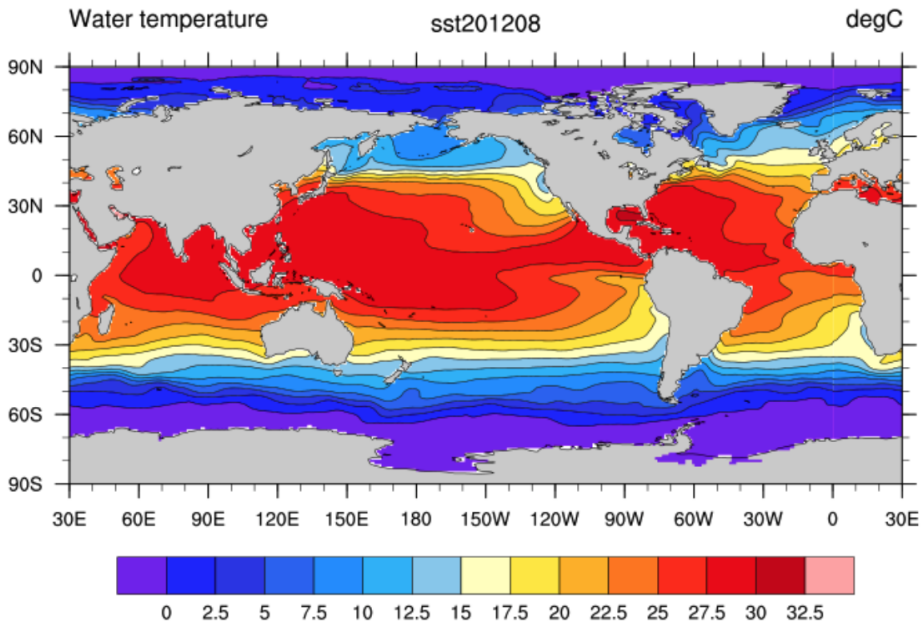


FIGURE 2.3 GLOBAL SST (COBE SST2). SOURCE: (NATIONAL CENTER FOR ATMOSPHERIC RESEARCH STAFF, 2020)

2.3.2 QUASI-BIENNIAL OSCILLATION (QBO)

QBO controls the variability of the equatorial stratosphere, which lays at about 16 km to 50 km height. For understanding the QBO, one should understand the Brewer-Dobson Circulation (BD Circulation). It is a circulation that takes place in the stratosphere where large-scale circulation occurs. Air rises above the equator and is spread to the north and the southern hemisphere, where it starts dropping down at 60° of both sides of the equator. Figure 2.4 demonstrates well the circulation flow.

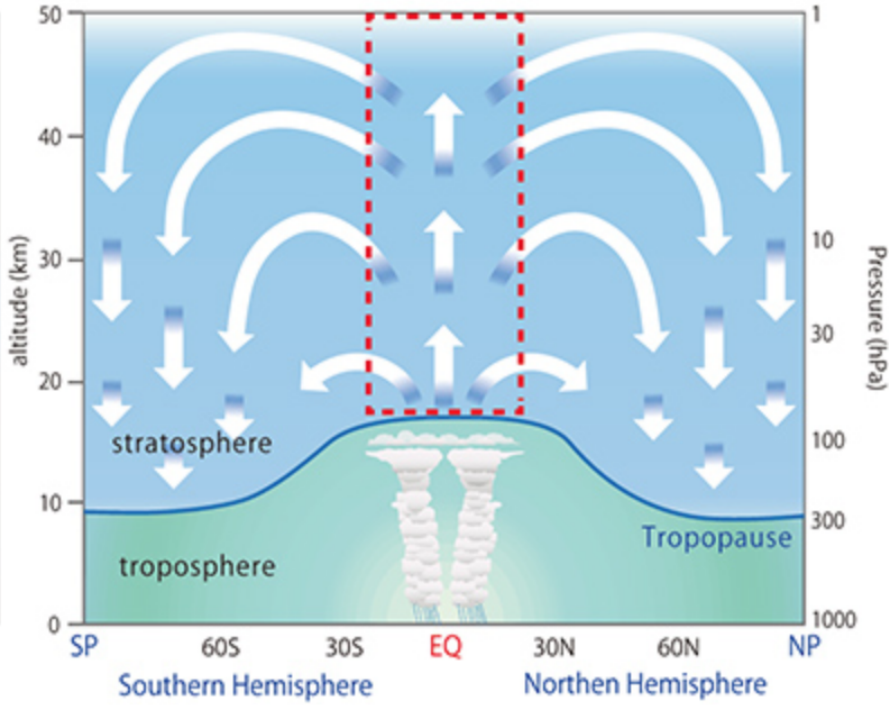


FIGURE 2.4 AIR CIRCULATION FLOW. SOURCE: (JAMSTEC, 2013)

This flow is crucial as it transports ozone, water vapour and other chemicals in the air globally.

The QBO takes place around the equator (around the red box in figure 8). It is seen as downward-moving easterly (blowing from east) and westerly (blowing from west) winds that occur with a period of averaging around 28 months, see figure 2.5. The QBO is an oscillation mean flow that is driven by downward-moving waves with different periods that are not related to the resulting oscillation (Baldwin et al., 2001). These alternating winds are first to develop at the top of the lower stratosphere and later propagate downward at a pace of 1km per month, until they disappear at the tropical tropopause. Figure 2.6 demonstrates well the descending zonal winds. The bigger arrow indicates stronger winds and vice versa. Usually, these winds are located 15° on both sides of the equator. The QBO is an important variable as it regulates the meteorology in the lower stratosphere (Wallace, 1973). The influence of it is extremely strong in the northern hemisphere during winter as the polar stratosphere is more sensitive to the downward moving winds and the breaking of the massive waves that affect the string westerly zonal mean flow.

FIGURE 2.5 EASTERLIES AND WESTERLIES AT THREE CONSECUTIVE YEARS. SOURCE: (JAMSTEC, 2013)

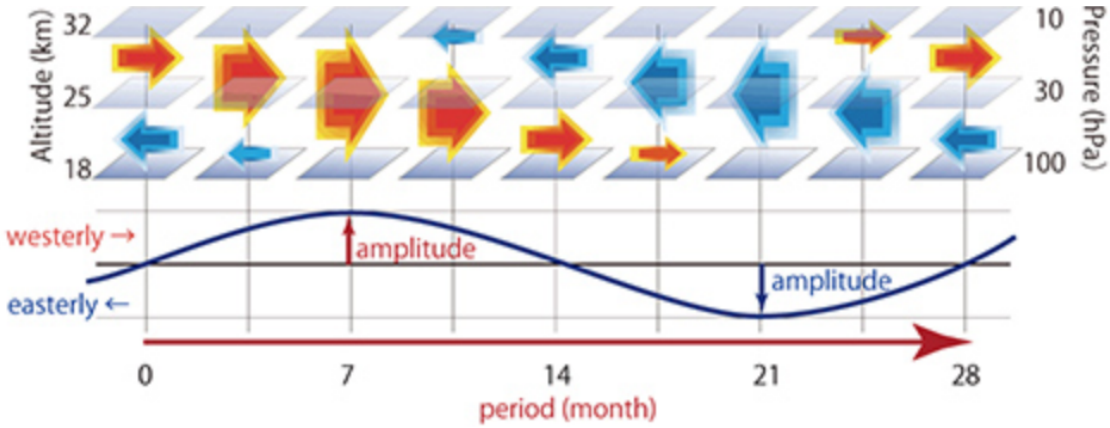


FIGURE 2.6 THE DESCENDING ZONAL WINDS. SOURCE: (JAMSTEC, 2013)

The QBO and Seasonal forecast

The relation between QBO and the Atlantic hurricanes has been hypothesized to relate due to the variation in the vertical wind shear between the upper troposphere and the lower troposphere (as described earlier). Wind shear is in general defined as the change in either vertical or horizontal wind speed or direction. This measure is vital regarding hurricane formation and intensification, especially the vertical formation, starting from the surface going all the way up to the troposphere.

As the changing period in the QBO is quite long (around 28 months) it is easy to conduct long-range predictions of the mean stratospheric zonal winds. An easterly QBO makes strong easterly winds in the lower stratosphere and with this large amount of vertical wind shear is created from the upper troposphere to the lower stratosphere. The west QBO phase usually lasts for 13 to 16 months, and it allows a weaker easterly wind in the tropical North Atlantic stratosphere. Hence, also lower amounts of vertical wind shear. The vertical wind share has been proven to slow down the formation of TCs and for this reason, in the years of west QBO, the number of hurricanes seems to be more active (Camargo et al., 2007).

QBO Calculated at NOAA/ESRL PSL

The NOAA Earth System Research Laboratories pursue research in fields of physical, chemical and biological processes. Their Physical Science Laboratory (PSL) has developed a research that resulted in the QBO data set. The data has been calculated at the PSL and it is continuously updated. It is generated from zonal averages of the 30mb zonal wind as computed from the NCEP Reanalysis Derived data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their webpage at http://psl.noaa.gov/. NCEP and OAR are bodies inside of NOAA conduction environmental prediction and ocean research. The QBO data starts from the year 1948 and goes all the way to 2019. The data is reported as monthly average and the unit given in is ms^{-1}. The westerly QBO is reported as a positive value and an easterly QBO is reported as a negative value. The missing data in the data set is flagged with a value of -9.96921e+36.

2.3.3 Lower Stratospheric Temperature (LST)

The stratospheric temperature has recently become of interest for researchers. It is a critical component when measuring climate change and the interest for it has increased with the trending climate change studied (Houghton *et al*., 2001). The cold temperatures in the Stratosphere overgo changes and in the same way do the different parts of it too. As seen in figure 2.7, the stratosphere extends over many kilometres and is much bigger in proportion to the troposphere. The whole layer is around 40 km high and for this reason, there are differences both in the higher stratosphere as in the lower one. Due to this, we will focus on the lower stratosphere as this is the area that is located closer to the tropical cyclones and other climatological variables affecting the TCs inside of the ozone layer.

The LST shows an important annual cycle and this causes that the two hemispheres move in opposite phases. This can be explained as a consequence of hemispheric asymmetries in the dynamic forcing of the circulation in the stratosphere (Yulaeva *et al*., 1994). The variable has studied to be of a crucial component when studying global change (Randel et al., 2009). Therefore it can be said to be sensitive to changes in the temperature, as it reacts to changes in the presence of ozone.

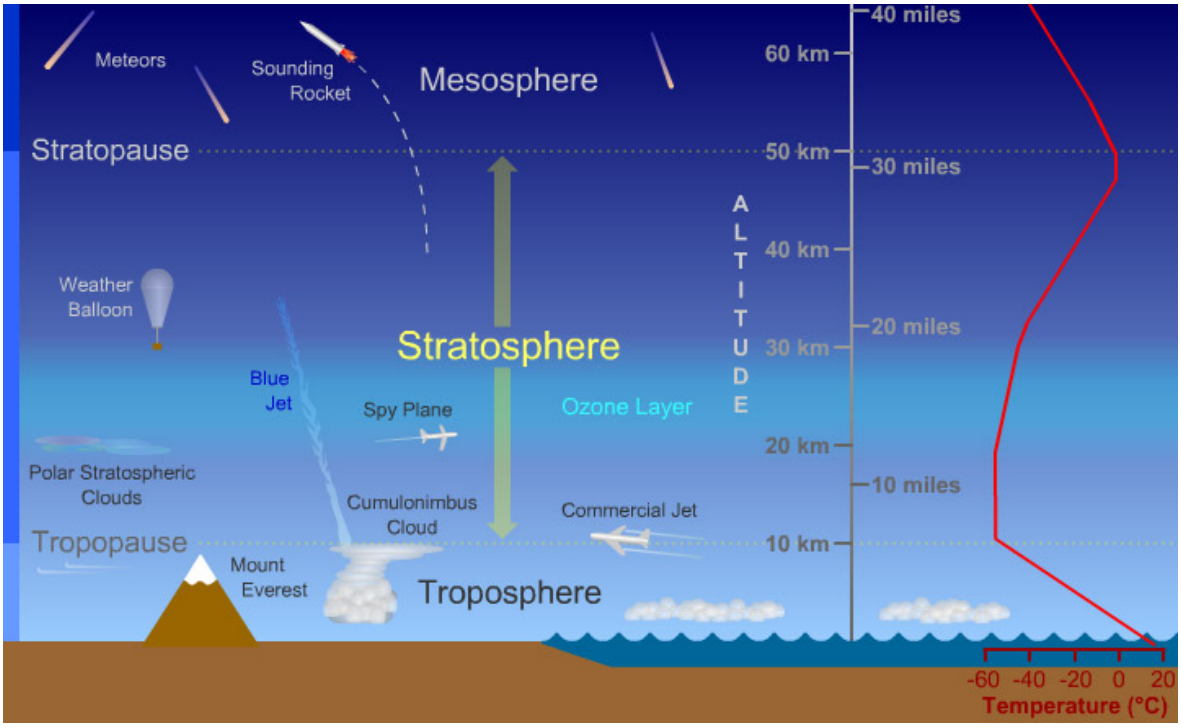


FIGURE 2.7 DESCRIPTION OF THE STRATOSPHERE, TROPOSPHERE AND MESOSPHERE. SOURCE: (UCAR, 2011)

LST and Seasonal Forecast

Little importance has put on research on the relation between the lower stratosphere and TCs. Still, recent studies suggest possible impacts of the lower stratosphere on the TC structure and development (Ferrara *et al.*, 2017). The higher atmospheric temperatures influence the lower troposphere where the hurricanes are built, and hence it is of interest to study the seasonality of the lower stratosphere temperature and study its relation to seasonal hurricane predictability. The upper part of the TCs are reaching the lower stratospheric areas and thus, the area should be taken into consideration when analyzing the storms and their creation.

LST data set from NSSTC

The LST data set was obtained from the official webpage of the National Space Science & Technology Center, operating under the University of Alabama. The reliability of the data has been studied and it has been used in other accredited researches.

The data set contains information from the years from 1978 until 2020 with an updated value for every month. The data is represented in monthly anomalies, a standard representation method in meteorological research. It means it represents the departure from a reference value, or a long term average. For example, a positive anomaly indicates warmer temperature and a negative anomaly indicates colder temperatures than on average. In the dataset used, the most recent updated reference average temperature has been calculated from the years between 1981 and 2010. Missing values are represented as -9999. The dataset reports the daily anomalies in both the southern

hemisphere as in the northern hemisphere, but as we are interested in rather more general patterns, the global anomalies will be used in this study. The latest structural update is from May 2005 and for this reason we can conclude that the data is somewhat up to date. The global coordinates go from -85° latitude to + 85° latitude. A separate analysis could be conducted on polar areas too.

2.3.4 NUMBER OF HURRICANES

In this research, the number of hurricanes is going to be the response value that the regression is done on. Every year the official monitoring centers report the number of TCs that has taken place in the basin. The data used in the analysis is provided by NOAA's Atlantic Oceanographic and Meteorological Laboratory. The hurricane research divisions mission is to predict the hurricanes and other tropical weather. The data is collected by an NOAA aircraft, and the information regarding the hurricanes are reported on their web page[4]. The data set presents *the number of named storms* for every year, the number of *hurricanes*, the number of *major hurricanes* and the *accumulated cyclone energy* (ACE) which measure the activity of the entire tropical cyclone season. ACE of each year is the sum of the ACEs of each storm takin into account the lengths, strengths and number of storms per season. For the research we will be using the variable of *hurricanes* as the storms of lower intensities are not of the main interest. When extracting some graphs, the ACE values are considered to compare different years between each other.

[4] https://www.aoml.noaa.gov/hrd/data_sub/hurr.html

3. DEFINITION OF OBJECTIVE

The main objective of this thesis is to create a data-driven seasonal forecast model, for prediction of hurricanes in a season based on an investigation of data inputs from an interval of months. By using SST data from all over the world for the months chosen, data for quasi-biennial oscillation (QBO) and lower strophic temperature (LST), a model for prediction of hurricanes in the coming season will be aimed for. Two important sources of SST will be compared, and based on their correlation to the number of hurricanes taken place in a season, the better data set will be used in the creation of the "best" model.

The random forest will be the main algorithm used in the model development due to its common high predictive accuracy in many situations. Four different random forest models are created, based on different pre-processing steps. This will help us find the "best possible" random forest model. In addition, three other algorithms are going to be used for validation of the models established. By training all models with the right time period, correct areas and variables, a base for a prediction model that could be used annually is expected to be developed.

To only focus on Random Forest was a decision based on the fact that the model generally performs very well with the use of many tree regressions. Further, it was of interest of the company to explore this kind of a model and for this reason the objective is to find the best random forest model possible.

The outcome is a comparison of the predictions done on the built models with the actual number of hurricanes taken place, that aims to obtain a mean error as close to zero as possible. As the objective of the modeling is to find the best possible random forest prediction model for seasonal hurricanes, the three additional models will not be elaborated as much in detail as the random forest models.

A second part of the research was conducted, but it will not be part of the thesis report. An analysis of the intensity forecasting error was made both for forecasts done by Hurricane Unwinder (HU) and for the National Hurricane Center. The aim of it was to find variables that could explain the errors occurred in the HU prediction and find events where the HU prediction model is more accurate than the official NHC one. Detailed comparisons between forecast errors and other variables were made, and an analysis of correlations between the prediction error and the climatological variables were done in order to find the reasons behind the errors.

Some of these results can be found in the annex A.5. Initially, it was planned to include this part to the master's thesis but as the investigation went on, it was decided to leave this part out of the study as the dissertation would have been too broad with difficulties to focus the work and quality of results.

.

4. METHODOLOGY

4.1 RESEARCH DESIGN

The thesis follows an organised research design as of many components and variables are included in the analysis and a clear order has to be remained (See figure 4.1).

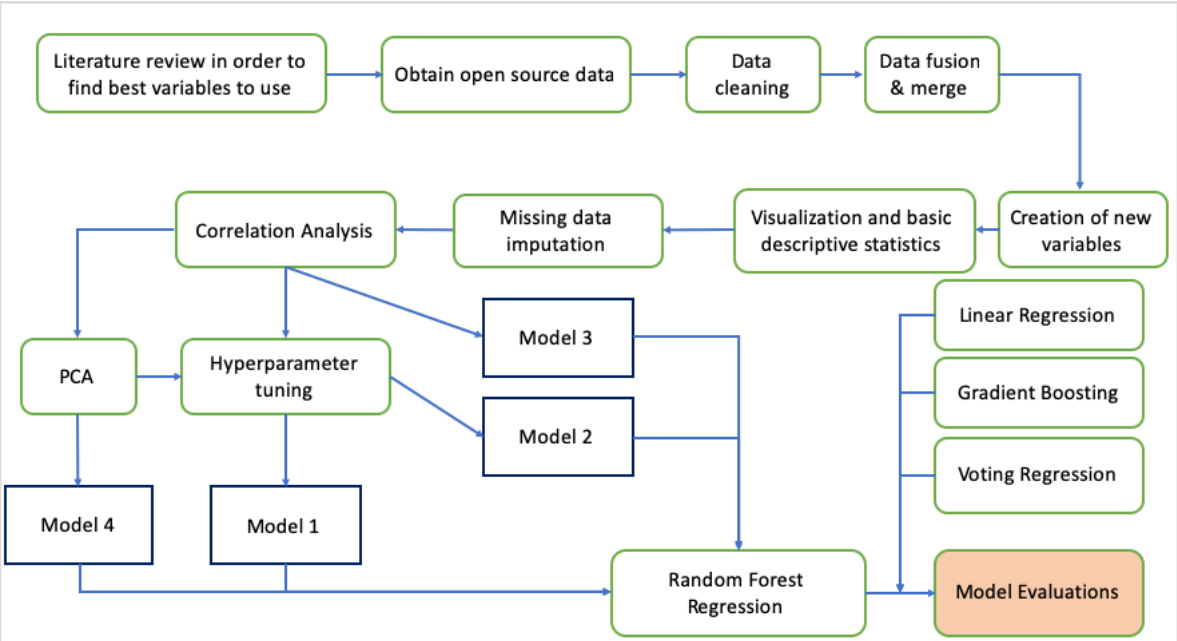


FIGURE 4.1 RESEARCH DESIGN

Before starting the creation of a seasonal forecast, a detailed literature review was done. Different methodologies and variables were studied, and it was decided to build the model on a statistical base as it fits best into the nature of a master thesis dissertation. The literature review showed that much importance had been given to the ENSO effect, whereof it has been studied in various researches. The combination of ENSO effect with Quasi-biennial oscillation has also been studied, and for this reason, it was decided to research if the Quasi-biennial oscillation alone could affect the seasonal predictions of hurricanes. Other variables chosen for the analysis were the lower stratospheric temperature and SST.

Then, The different Random Forest models (model 1 to model 4) were induced from the final data matrix, and compared against the other models.

4.2 IMPLEMENTATION

All analysis and processing of data have been done using the Python programming languages and several Python packages. Table 4.1 shows the used packages for the research;

TABLE 4.1 PYTHON PACKAGES USED FOR THE MASTER'S DISSERTATION

Python 3.0 Packages	**Provided Functionality**
Pandas	Data manipulation
Numpy	Manipulation and computation with arrays
Matplotlib	Data visualization
Seaborn	Data visualization
Scipy	Computing and algebra
Xarray	Tool for working with multi-dimensional arrays
netCDF4	Tool to read data in Python
xlrd	Extraction of data from excel
Sklearn	Machine learning
graphviz	Graph drawing
StringIO	Read string files

In Annex A.4, samples of the Python scripts generated can be found. As the script is very exhaustive, only a few snippets of the code was inserted. Some code parts for pre-processing, hyperparameter tuning, and the random forest regression are attached.

4.3 PRE-PROCESSING

4.3.1 INITIAL PRE-PROCESSING

Before starting the work with the data, some previous steps were needed in order to have a good quality of data. In the pre-processing, some steps of the proposal for pre-processing process presented by Gibert *et al*., (2016) (figure 4.2) were undertaken.

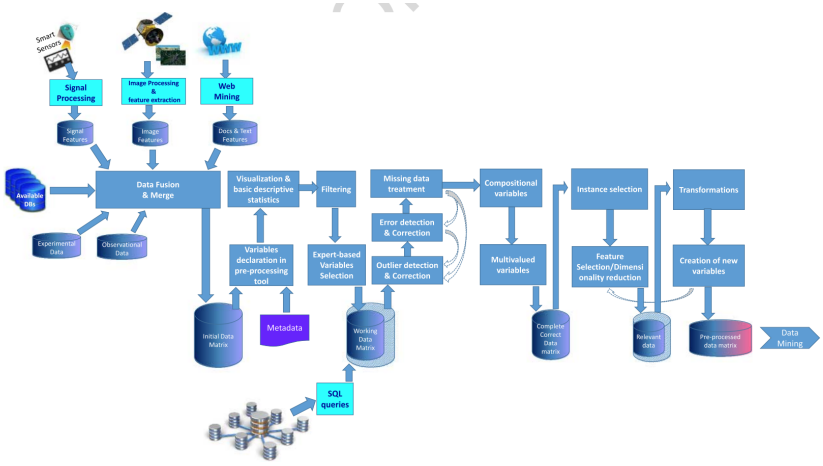


TABLE 4.2 PROPOSAL OF STEPS IN THE PRE-PROCESS. SOURCE: (GIBERT ET AL., 2016)

The following steps were done in order to process the data:

1.1 Variable identification

A careful literature review (Chapter 2) was conducted to identify the previous studies regarding variables and models used. After carefully comparing the studies, the variables of interest could be finalized.

1.2 Identification of suitable databases

Open-source data was used and the sources were studied well in order to check the reliability. For each variable, a different source was needed.

1.3 Data cleaning

The data cleaning aims to unify the data sets obtained and match the time span in order to use them in the analysis. Encoding of the data sets and re-formatting was needed in order to avoid errors and other mistakes. Unnecessary columns of datasets were deleted at this point, only to include the desired variables.

1.4 Data fusion & merge

The creation of a mutual data frame for the seasonal forecast required combining and processing of different databases.

1.5 Creation of new variables

New variables were created and added to the data set for the SST variables. A logic way to break down the variable is to create a smaller geographically-based division. This was done according to coordinates (longitude and latitude). Based on expert knowledge, this new "areas" were created and later each of them was tested statistically. As the coordinates of the data sets do not correspond to real longitude and latitude coordinates the "new areas" has to be built according to the coordinates used in the data sets.

1.6 Visualization and basic descriptive statistics

This step gives the reader a better picture of the behaviours of each variable used and possible further steps for data processing can be identified better. Basic statistical parameters were studied and graphs were plotted to see the variables behaviours better. The following statistical parameters were calculated:

- Mean value: the sum of the ensemble of all the values in the set divided by the number of integers. This was done for each year.

$$M = \sum_{i=1}^{n} x_i \quad (2)$$

Where x_i is each value taken by the variable x and n the total number of observations.

- Median value: The median value separates the higher half from the lower half. It can be seen as the middle value.

- Standard deviation: shows us how much dispersion takes place of the variable.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x)\hat{}2}{n-1}} \quad (3)$$

Where n is the total number of observations, xi representing each value presented by a variable and x representing the mean M.

- Maximum and minimum values were calculated to understand the range of the data better
- The number of observations were shown so that it could be detected if missing values were present.

For all variables boxplots were represented and depending on each variable time-series plots are shown.

1.7 Missing data imputation

The data sets worked with are quite complete, and no significant missing data takes place. For this reason, only the SST data set had to be modified as the other data sets were found to be complete. Based on the database description, the missing data was identified, but as it does not have a significant impact on the research the missing data was replaced with Nan. The NaN values in the SST dataset can be assumed to be of land area and hence it does not affect this research as we only include sea areas of the SST variable. A quick test was done to prove it. Three areas with no land area were studied for missing values, and respectively three areas with little and a lot of land area were analysed. Thus, if the areas with a high amount of land area were the ones with the highest amount of Nans we could trust the assumption that missing values are for land areas and hence do not affect our analysis.

4.3.2 DIMENSIONALITY REDUCTION

4.3.2.1 CORRELATION ANALYSIS

In order to identify the SST data set to use, and also to identify the better ones subset created in the pre-processing, the *Pearson correlation* was calculated. The Pearson correlation coefficient is used to determine the relationship between variables and their strength, and it measures the magnitude and direction of the change in the variables (Schober & Schwarte, 2018). The Pearson correlation was calculated for the months from January to July.

As the variable SST can be divided into endless areas, it is of interest to study what geographical SST area brings useful information to the research. The areas were divided by using detailed expert knowledge, and due to this, it is essential to study and statistically verify which of the areas to include in the analysis. The correlation coefficient was studied for areas and the regions with very low correlations were left out. Many scholars have published different categories for how to classify the correlation with varying cut-off points, but as our case is particular and in general relatively low correlations were found, the areas with two correlation values higher than 0.2 were chosen for the analysis.

Pearson's correlation coefficient, also called *Pearson's r*, is useful when comparing the strength of a linear association between two variables. Equation 4 describes the calculation behind Pearson's Correlation coefficient.

$$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}} \quad (4)$$

The correlation coefficient *r,* can take any value between -1 and 1, with 0 indicating no correlation and a positive value indicating a positive correlation and a negative indicating a negative correlation between the variables. The stronger the association of the two variables are the closer the *r* will be to -1 or 1.

In python, the scipy.stats package is used to calculate the Pearson r. It calculates the correlation coefficient but also a 2-tailed p-value. The p-value is the probability that the absolute value of *r* of a random sample, *x* and *y,* drawn from a population with zero correlation, would be bigger (or equal) than the absolute value of r. In other words, we can assume the our null-hypothesis (H0) to be the following;

H0= the variables are uncorrelated *(5)*

The p-values measure the probability that data would be the same if the H0 were true. Saying this, we can conclude that a low p-value indicates that the H0 can be rejected and a higher p-value indicates of one not being able to reject the H0. Generally, researchers use limit values like 0.01 or 0.05. The test was used to clarify the right geographical areas to include in the research but also with the aim to try to see how many months bring valuable information to the model.

Once the Pearson correlation coefficient was done, an analysis was conducted to compare the two SST datasets used. The correlation coefficient is considered to be of relatively good fit in order to decide on what data set to use in the final analysis.

4.3.2.2 PRINCIPAL COMPONENT ANALYSIS (PCA)

Before starting with the predictive model induction, the algorithm needs a few pre-processing steps just before the execution of the model. Once the dataset used was built, the following steps took place:

1- Train-test split: To validate the model, the data set hasdto be divided into the training set that trains the model, and a separate test set that is used to validate the model built. The aim is to have as much training data as possible but as the data set worked with is relatively small to train and test on, a split of 40% gives a fair split for both sets.
2- Scale data: Before modelling, the data was “centered” and standardized. This was needed as the working data had different variables and were measured on various scales. In this way all variables have the same importance when running the regression.
3- A fit to baseline random forest model was calculated at this point to be able to compare the behaviour of the output of the model once the PCA and other operations were done. This was done using the Mean Absolute Percentage Error (MAPE). The accuracy was calculated by subtracting MAPE from 100% as it is a statistical measure presenting the accuracy of a forecast model. It was calculated by

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Yt-\hat{Y}t}{Yt}\right| \quad (6)$$

, where Y is the actual value vector and $\hat{Y}$ is the predicted value vector.
And further

$$\text{Accuracy} = 100\text{-MAPE} \quad (7)$$

After this step, the PCA was conducted.

As this analysis aims to find the relationship between climatological variables and the number of hurricanes taken place in a season, our interest lays in finding the best predictor variables. Hence, also reducing the dimensionality of our data is of interest as the aim is to get an as precise prediction model as possible. The idea behind PCA is that a rotation generates a new coordinate system for the original dataset, where new factorial axes are linear transformations of the initial variables. The PCA process builds a covariance matrix, a symmetric matrix that is created of its own vectors (Shlens, 2014). Further, PCA reveals underlying trends, influences, and other relationships within the dataset which might not be seen without axis rotation.

The coordinate transformation results in a matrix including the values of the principal diagonal. The values in the diagonal capture information captured by each principal component in terms of variance. The higher the variance is, the more information the component maintains. Based on this the components containing the most information, were chosen for further analysis. The advantage of PCA is that it maintains the characteristics and variables of the dataset that provides the most to the variance.

Before moving on to the hyperparameter tuning the accuracy of the model was calculated again with the new PCA dataset in order to see if any improvement had been obtained.

4.4 Model building

The seven models induced from our data was described in this section. However, first, a hyper-parameter tuning task of the model parameters was detailed, before the building of the different models were done.

4.4.1 Hyper-parameter tuning

Hyperparameter tuning is an operation that adjusts the model to perform the best possible way. In this way, a prediction performance is tried to maximize. As every dataset is different, each model has to be modified according to its needs. Hyperparameter tuning is used in machine learning in order to find the set of hyperparameters that controls the learning path of the algorithm. As these hyperparameters of the models cannot be trained, the model has to be tuned to get the best output. In this analysis, this was done in two steps. The first step was the *Randomized Search Cross Validation,* used in the sklearn libraries. This function finds the best hyperparameters to use in the model from

parameters that are suggested by the user. These are the number of decision trees in the forest, and the number of features used by each tree when the node splitting is done. This function runs multiple *K-Fold Cross Validations* inside of a range that the author has decided itself. The K-fold CV is a method used for cross validation, and it divides the data training set into K number of subsets. The model is then fit K times by training it each time on K-1 of the subsets and evaluating the Kth subset (validation data here). The Randomized Search CV does this function inside of the chosen ranges given. It is done on random sampling running K-Fold CV on each combination of values. The hyperparameters analysed were:

n_estimators – number of trees in the forest

max_fetaures – maximum number of features considered for splitting of a node

max_depth – maximum number of leaves in each decision tree

min_sample_split – minimum number of data point that are placed in a node before it can be split

min_samples_leaf – minimum number of data point allowed in a leaf

bootstrap – a method of sampling the data points either with replacement or without it

Once these hyperparameters have been tested and found, a range of best hyperparameters could be identified. Bar plots were presented in order to show the mean scores of the model at each value and the performance of each hypermeter. The Mean Average Error (MAE) Scores were presented as negative in order to have a maximization problem. This means the bigger value the better, MAE equal to zero would be a perfect model.

With the results obtained in Randomized Search CV the next step, *GridSearchCV* could be done. The best results found from Randomized Search CV were used in the grid search. The *GridSearchCV* is a more refined search for the best hyperparameters. In this step, every single combination of hyperparameter values are tested, which takes a lot of time. Due to this, the RandomSearchCV works as a pre-process for the GridSearch as it frames the used range. By plugging in the average best performing ranges, the test is done in a smaller range. Then, when running the test, the best hyperparameters were identified, which could be used in our final RF model.

4.4.2 RANDOM FOREST MODEL

The random forest regression (RF) model is a classification and regression method, which consists of tree predictors where each tree is generated by utilizing a random vector sampled independently from the input vector. The error of the tree classifier depends on the strength of the individual trees used in the forest and correlation they have amongst them. As the algorithm uses a random selector of

features when splitting the node, the error rates are better than when comparing to Adaboost, but regarding noise, they are more robust. The same ideas apply to regression models.

The main idea behind random forest regression is to combine many predictions made by different decision trees into one single model. Each of the individual trees used in the model brings information as they are considered a random subset of features in the whole, and even have access to a random set of training data points. The model takes the average of all the individual estimates when forming the final prediction.

The RF uses bootstrapping as default when sampling, which means that when the sample is chosen out of the set, they are replaced. The number of samples of the RF is chosen according to how many trees the algorithm has been chosen to use, n_{tree}. Each sample has a probability of (1/n) of being chosen. For this reason it can be assumed that we expect each element to show up at least:

$$(1\text{-}1/n)^n \approx 1/e \approx 0{,}368 \text{ times} \quad (8)$$

This means that approximately 36,8% of the elements are not shown at all in the analysis. The idea behind the bootstrapping is that each tree will be trained on different samples and even if each tree has a high variance, overall the whole forest as a total will have a lower variance together without increasing the bias.

From these bootstrap samples, an unpruned regression tree was created. Pruning is a technique that reduces the size of the decision tree by removing parts of the tree that do not give any power to the regression. The pruning is often used if there is a risk of overfitting (Hoare, 2020). Overfitting takes place when a model is very flexible and memorizes the training data. When using regression trees this problem is taken into account by adding bagging into the model as it reduces variance and increases the bias by combining many trees into one single model. For this reason, the random forest regression model is usually unpruned. The trees are generated in a way that each node, instead of choosing the best split of all the predictors, randomly samples *m* tries of the predictors. From those variables it selects the best split. What makes the random forest so special is that it uses bagging, which is a method to generate a training set by randomly resampling the original dataset with replacement of N examples, where N is the size of the training set used originally (Rodriguez-Galiano *et al.*, 2015). In this way the data is more stable and higher accuracy is achieved. The RF grows the trees until the best result is obtained by using the best split point within a subset of important features that have been chosen randomly from the overall set of inputs.

The final prediction is found once it combines the predictors of the separate decision trees that were only somewhat predictive while the predictions were not correlated to each other, hence, it is called a Bootstrap Aggregation. The average of these predictions are calculated and form the random forest prediction values. The part of the samples that were not included in the n-th tree in the bagging process, are used as part of the subset that is called out-of-bag (oob). These elements might be used for all trees to evaluate performance. For this reason, an error can be computed without having to compare with an external data subset. Figure 4.2 and 4.3 shows how the random forest works.

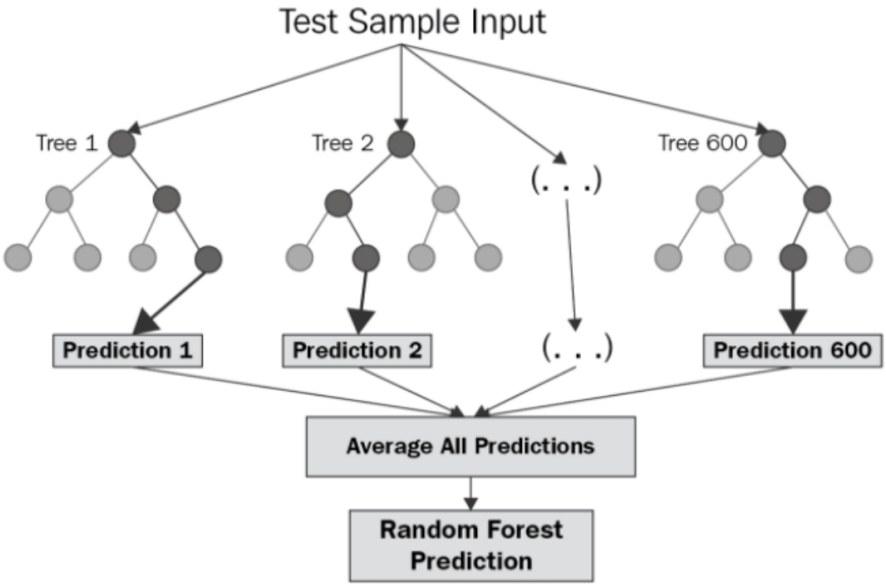


FIGURE 4.2 RANFOM FOREST THEORY MODEL. SOURCE: (BAKSHI, 2020)

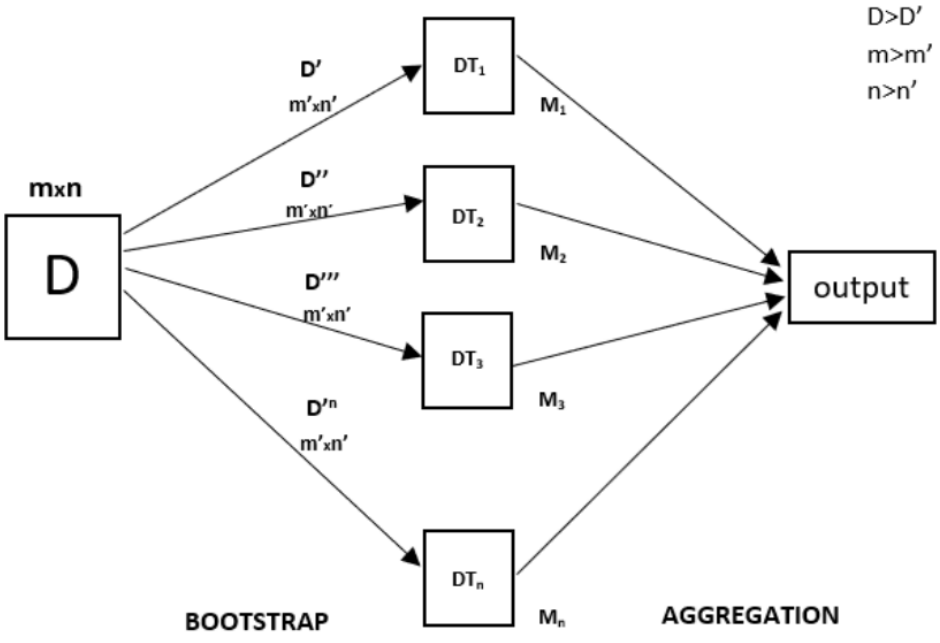


FIGURE 4.3 VISUALISATION OF BOOTSTRAP AGGREGATION THEORY. SOURCE: (DUTTA, 2020)

4.4.2 LINEAR REGRESSION MODEL

As discussed in chapter 2, the simple linear regression model is one of the basic models of machine learning. In this research a very simple model was used. The general equation form of a linear regression model is:

$$Y = a + bX \quad (9)$$

, where Y is the dependent variable, X is the explanatory variable, a the intercept and b the slope of the line.

As the model that was used in this analysis had several explanatory variables, the function contains various X variables (Multiple Linear Regression model).

The dependent variable, number of hurricanes, was predicted with the test set on the trained model. The output of the model is compared to results of the other models generated.

4.4.3 GRADIENT REGRESSION MODEL

The gradient boosting is a method of creating an ensemble. The process starts by fitting an initial tree to the data. After this, a second model is built specializing on the predicting the cases where the first model performed poorly. The same processes is continued many times and the model is expected to perform better at each step the boosting is repeated. When running a gradient regression, the algorithm optimizes on base of Mean Square Errors. Figure 4.4 demonstrates the basic idea behind gradient boosting.

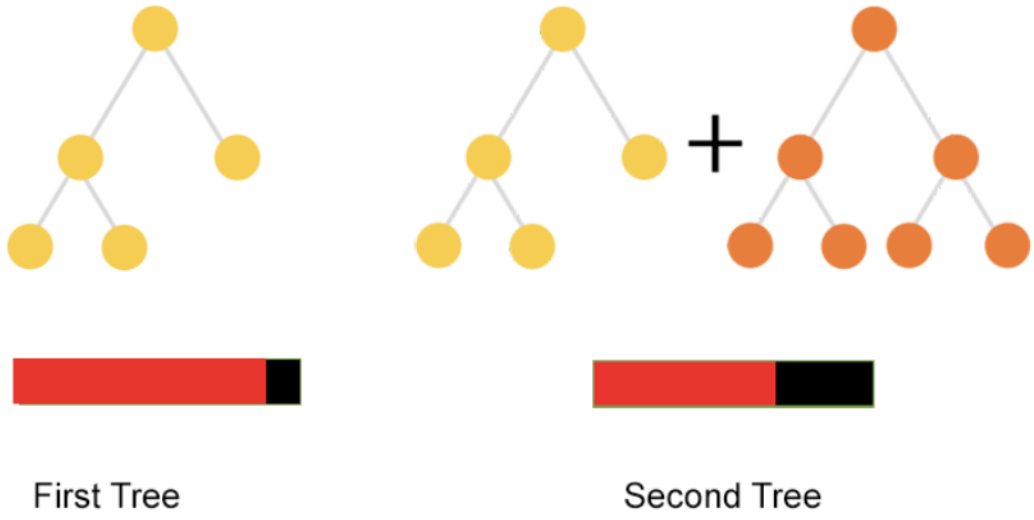


FIGURE 4.4 GRADIENT BOOSTING VISUALIZED. SOURCE: (ERSHOV, 2018)

The first tree works as the basic tree and from there the model builds a second tree trying to optimize the part that is performing the worst. At the same time it can be noted that the errors (red-black) bar diminishes when the new trees are built.

We used the basic hyperparameters given by the algorithm in Python and did not do tuning on the model. The non-tuned algorithm was used, as in this way the general performance of the algorithm could better be understood. It enabled an analysis on finding out which was the most accurate to use in this case. The results from the model were compared against results found from the other models.

4.4.4 VOTING REGRESSION MODEL

As discussed earlier, voting regression combines the average of a set of predictions made by other models. In our analysis the voting regression was done based on the linear regression, the gradient boost regression and a basic non-tuned random forest. This was done to equally compare and implement the different models. A limitation in the voting ensemble is that it treats all the models uniformly, resulting in all contributing equally.

The models used for voting were fit on the data and combined into one voting ensemble. This new design was fit to the data and a prediction was done in the same way as in other prediction models. The voting regression averages the individual predictions and forms a final prediction based on these.

4.5 MODELS EVALUATION

For the evaluation of the models, various verification methods have been used. The minimum error and the maximum error were calculated for *y_pred1* and *y_pred. y_pred1* stands for the regression prediction done on the training set and *y_pred* represents the regression prediction done on the test set. *Y_test* and *y_train* are the true values for each dataset. The maximum and minimum errors are metrics that capture the best and the worst-case scenarios between the error of the predicted value and the true value. In a perfectly fitting model the error would be zero. The errors were calculated using the following formula:

$$Max\ Error(y, \hat{y}) = max\sqrt{((y}_{pred} - y\_test)\text{ˆ}2) \quad (10)$$

$$Max\ Error(y, \hat{y}) = max\sqrt{((y}_{pred1} - y\_train)\text{ˆ}2) \quad (11)$$

$$Min\ Error(y, \hat{y}) = min\sqrt{((y}_{pred} - y\_test)\text{ˆ}2) \quad (12)$$

$$Min\ Error(y, \hat{y}) = min\sqrt{((y}_{pred1} - y\_train)\text{ˆ}2) \quad (13)$$

The mean absolute error (MAE) measures the average magnitude of the errors without taking into consideration of their direction. MAE fails to punish significant errors as it treats all errors the same way. It is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (14)$$

, where *n* represents the number of observations and ŷ the *i-th* estimation. *Y* stand for the true *y* values.

The R^2 represents the proportion of variance of y that can be explained by the independent variables in the model. It provides the goodness of fit and describes how well the samples are probable to be predicted in the model generated. The R2 values lay between +1 and -1. If $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the correct value in a set of n samples, the estimated R^2 can be defined following:

$$\mathrm{R}^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (15)$$

Where,

$$\bar{y} = \frac{1}{n}\sum_{n=1}^{n} y_i \text{ and } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(\epsilon_i^2), \quad (16),(17)$$

ϵ stands for the errors in the model.

The R^2 was calculated for both the training and test set.

The mean squared error computes the average of the squares of the errors. MSE measures the quality of an estimator and values closer to zero are better. The model punishes bigger error (different from MAE) as the bigger errors gets a bigger value once they are squared. The formula is:

$$\mathrm{MSE}\ (\mathrm{y}, \hat{\mathrm{y}}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1}(y_i - \hat{y}_i)^2 \quad (18)$$

The MSE metric is good as it punishes bigger errors, but as it squares all units of data it is not comparable with other metrics. To solve this problem, the root mean square error RMSE was used. It measures the standard deviation of the residuals. It shows how spread out the residuals are, which indicates how concentrated the data is around the lines of best fit. It was calculated by taking the square root of the MSE.

$$\mathrm{RMSE}(\mathrm{y}, \hat{\mathrm{y}}) = \sqrt{MSE} \quad (19)$$

These scored were evaluated when the models were built and an attempt to find the best possible models was done by evaluating these metrics. The metrics don't tell alone much of information but once the results of many different regressions were compared they are of good usage.

4.6 MODELS BUILT

Various models were built and tested in order to find the best one. All models were pre-processed the same way until the usage of the Pearson Correlation Coefficient, meaning all data sets consists of the SST data, LST, QBO and the "Number of Hurricanes" variables. Further all predictions were scaled and a MAPE was calculated for each of them. The following models were tested.

1) PCA + tuned RF
   The data sets dimensionalities were reduced with PCA. The RF algorithm has been tuned in order to get the best model for this specific data set
2) No PCA + tuned RF
   Different from the first model, no PCA was used to reduce the dimensionality of the data but the RF has been tuned for this particular data set.
3) Random RF – no PCA ("*base-line model*")
   In this model the X and Y test and training sets were used directly after pre-processing without any dimension reduction in PCA. Also, the RF model used had a random number of n_estimators and used a random replacement method in bootstrapping.
4) PCA + random RF
   The last RF model was treated with PCA after pre-processing. The components with an explaining variance up to 80% were chosen and these were used in the following steps of the PCA. No tuning was done and the random values of RF were used.
5) Linear Regression model
   The X and Y data sets were used after pre-processing and scaling. No other processing was done on the data before fitting and predicting Y on the test set built.
6) Gradient Regression model
   Similarly to the previous model, no dimension reduction was done with PCA on the model, neither any tuning as the focus was to compare the basic function of the model compared to the rest of the models. The basic hyperparameters were used when running the model.
7) Voting Regression Model.
   Model 3, model 5 and model 6 were combined into model 7. The average performance of the three models was used to do a prediction using the voting regression.

5. RESULTS

As discussed in chapter 4, the results found from the research were generated in the order described. In this chapter, the results obtained are presented with graphs and the corresponding descriptions.

5.1 PRE-PROCESSING

The workflow described in chapter 4 was followed in generation of the results.

5.1.1 INITIAL PRE-PROCESSING

5.1.1.1 VARIABLE IDENTIFICATION

A detailed literature review was done in chapter 2 where it could be found that many variables have been studied together in *seasonal forecasting*. In order to bring some new perspective to the already studied fields, the most common variables such as La Niña and El Niño effects were left out of the research. After consulting with the company experts of Hurricane Unwinder and studying the variables previously used, it was found that the best variables to study were, SST, QBO and LST for prediction of the number of hurricanes that occur in a season in the Atlantic basin.

5.1.1.2 IDENTIFICATION OF SUITABLE DATASET

For each variable a data set was identified, and its properties were studied. Making sense of the data and understanding the variables was a crucial step in order to be able to process the data to a usable form. Data sets were downloaded and opened in Python, where some of the data sources were given in Excel (Number of Hurricane-data set) and some of the data was downloaded in NETCDF4 format (Both SST files). The NETCDF4 data files had to be processed in order to be readable in Python. The LST and QBO data was obtained in csv files, hence they were easily readable as a Python Pandas DataFrame by a transformation from string to floats for the QBO and from object to float for LST using Python.

5.1.1.3 DATA CLEANING

As the variables were obtained from different sources, some cleaning and processing was needed. Notations used in the data were somewhat different in the sense of using "," separations for the same value as for ";" etc., which was needed to be carefully taken into consideration when going through the datasets. This step did not contain a difficult process but was time-consuming and vital as it might

have an impact on the final results if not done carefully as the data could have been read differently depending on the separator in data.

As mentioned, each variable had data from different years and for obtaining the best data, the time span had to be unified to be the same for all data. Even if the sources report that data are continuously updated, the updates in the sources are not immediate. For this reason, table 5.1 presents the years that data was available for each variable.

TABLE 5.1 DATA AVAILABILITY (IN YEARS) FOR EACH VARIABLE

Variable	Data availability (Years)
SST - HADISST	1870 – 2019
SST – COBE SST2	1850 – 2018
QBO	1948 – 2019
LST	1978 – 2019
Number of Hurricanes	1851 – 2015

As it can be noted from table 5.1, there exists variability in the data availability. For both SST variables many years of data was available, but as it is of interest to analyse both QBO and LST, the database has to be limited to years where all variable data were available. The LST variable dataset was available only from 1978 onward and therefore previous years could not be included in the analysis. In the same way, the dataset of the number of hurricanes occurring in a season includes data to 2015, which did put the upper limit to the data used. Hence, this analysis was conducted on data from ***1978-2015.***

Not to forget, only data from the months of January to July were included in the analysis. This means that data from August to December were left out.

Another important part of the data cleaning was to delete the unnecessary columns from the datasets that were not going to be included in the research. Table 5.2 summarizes the variables that were left out from each dataset before creating a common one for our analysis, where NH = Northern Hemisphere, SH=Southern Hemisphere, TRPC= Tropics, NO.DAYS=Number of days. The .1 variables are running means of the previous year and are therefore not of interest in the study.

TABLE 5.2 VARIABLES DELETED FROM CORRESPONDING DATA.

Variable	Variables deleted

SST - HADISST	*Time bands*
SST – COBE SST2	*None*
QBO	*None*
LST	*NH, SH, TRPC, NO.DAYS, GLOBAL.1, NH.1, SH.1, TRPC.1*
Number of Hurricanes	*Named Storms, Major Hurricanes*

5.1.1.4 DATA FUSION AND MERGE

Some of the data was presented in different formats with monthly values on the columns, while other datasets contained annual values in the columns and monthly values in the rows. A common method had to be found in order to align the separate datasets into one. Index resetting and the transposes were created in order to have years and months in an aligned manner. All the variable data frames were merged into one single dataset that was used for the analysis in this thesis. Similarly to the data cleaning, this step was quite time-consuming as it entailed a lot of data processing and coding in Python. The merged data was ordered in a way that the variables were presented on a monthly basis for each year.

5.1.1.5 CREATION OF NEW VARIABLES

New variables were created for the SST variables in order to break down the huge data frame. The *HADISST* data sets use the point 89.5°N 179.5°W as (1,1), hence some modification was needed. The starting coordinate (1,1) was changed to be 89.5°S 179.5W, to in a more straightforward way treat the data used. One shall not confuse these coordinates with a traditional x-y coordinate system as the latitude will always be reported first. The dimensions of our system are (180,360). Figure 5.1 shows the earth coordinates so that a better picture of earth's coordinate system can be built.

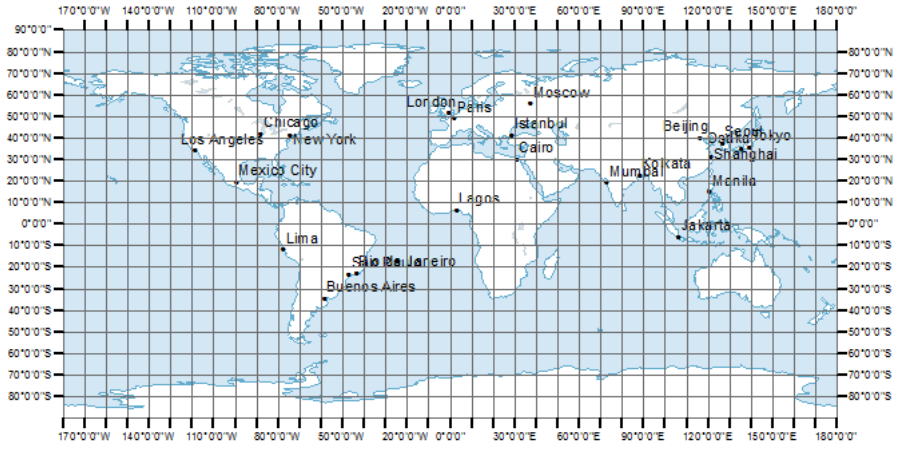


FIGURE 5.1 EARTH COORDINATES. SOURCE: (GIS GEOGRAPHY, 2020)

For example, London, has approximately the coordinates of 51°N 0°W in "real life", while in the *HADISST* dataset London would have the coordinates of:

Latitude: 89.5° + 51° = 140.5°

Longitude: 179.5° + 0° = 179.5°

The coordinates would be (140.5, 179.5).

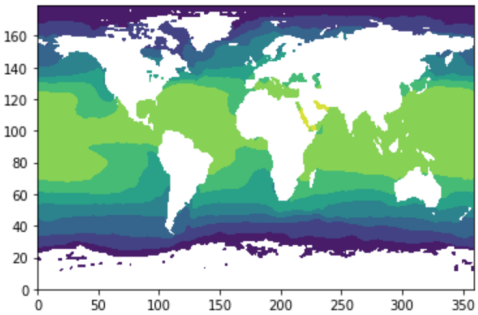
For the *COBE SST2* dataset no changes for the longitude coordinates had to be done rather than after 179.5°E continue the coordinates from 180 to 360. For latitude, the same adjustment as for *HADISST* had to be done. Hence, the coordinates for London would be:

Latitude: 89.5° + 51° = 141°

Longitude: 0°

London in *COBE SST2* (141,0)

The proposed method to "re-center" the maps according to the data was done, and Figure 5.2 presents the results obtained for a visual representation of how the world map is situated on its axis.



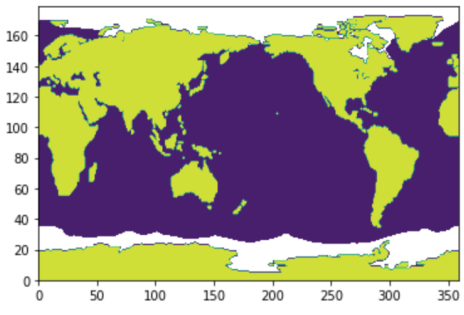


FIGURE 5.2 THE HADISST AND COBE SST2 MAPS CENTERED ACCORDING TO THEIR COORDINATES

Based on this theory, the global SST were divided into smaller sub-parts to study better what oceanic area affects the seasonal forecast of Hurricanes in the Atlantic basin. These areas are not based on any official coordinates as it is not considered necessary for this research. The names and the coordinates were more of estimations and shall not be used as a reference in other researches. In table 5.3, all the new variables that were created can be found. The first column states the new area name (again, not correct names nor coordinates but works as labels in this research), the second column its coordinates given in the HADISST data set with first giving the latitude (lat) starting coordinate to the ending coordinate separated with a ":" sign, and the third column presents the same coordinates for the COBE SST2 for the given area. To make sure that the areas were equal between the two datasets, the areas were cropped at the same "real coordinates" but as the datasets are centered a bit different, the HADISST areas had to be divided in two in the cases of *North Atlantic / Arctic Sea A, Pacific A, Pacific B, Antarctica B and ArcticA.*

TABLE 5.3 TABLE 5 NEW VARIABLES CREATED FOR HADISST AND COBE SST2

Created area name	Coordinates in *HADISST (lat, lon)*		Coordinates in *COBE SST2 (lat, lon)*
West Africa A	(50:95), (180:200)		(50:95), (0:20)
West Africa B	(50:100), (140:180)		(50:100), (320:360)
North Atlantic / Arctic Sea A	North AtlanticA_1	(130:160), (320:360)	(130:160), (140:360)
	(North AtlanticA_2	(130:160), (0:180)	
North Atlantic / Arctic Sea B	(130:160), (180: 250)		(130:160), (0:70)
Atlantic Sea	(100:130), (120:170)		(100:130), (300:350)
West Mexico	(100:130), (50:100)		(100:130), (230:280)
Caribbean	(100:120), (100:120)		(100:120), (280:300)
Pacific A	PacificA_1	100:130), (0:50)	(100:130), (120:230)
	PacificA_2	(100:130), (300:360)	
Pacific B	PacificB_1	(50:100), (300:360)	(50:100), (120:280)
	PacificB_2	(50:100), (0:100)	

Antarctica A	(0:50), (180:330)		(0:50), (0:150)
Antarctica B	AntarcticaB_1	(0:50), (330:360)	(0:50), (150: 360)
	AntarcticaB_2	(0:50),(0:180)	
ArcticA	ArcticA_1	(160:180), (320:360)	(160:180), (140:360)
	ArcticA_2	(160:180), (0:180)	
ArcticB	(160:180), 180:250)		(160:180), (0:70)

In Python these variables were created to run from 1978-2015.

5.1.1.6 VISUALIZATION AND BASIC DESCRIPTIVE STATISTICS

For all variables basic descriptive analysis was represented and depending on each variable, time-series plots or box plots were shown. Plots were only presented if they were considered to bring valuable information to the reader. Box plots were generated for both the variables QBO and LST, and for the number of hurricanes, a time series was plotted in order to visualize the number of hurricanes taking place each year.

Number of hurricanes

As seen from figure 5.3, the number of hurricanes taken place each year varies from year to year without a direct pattern. 2005 seems to have a higher number of hurricanes compared to the rest of the years, with a number of over 15 hurricanes in one season. The years with the lowest number of hurricanes are around 1982 and 2013 where only 2 hurricanes took place. It is interesting to note that the number of hurricanes varies a lot between years and no standard or trend can be identified.

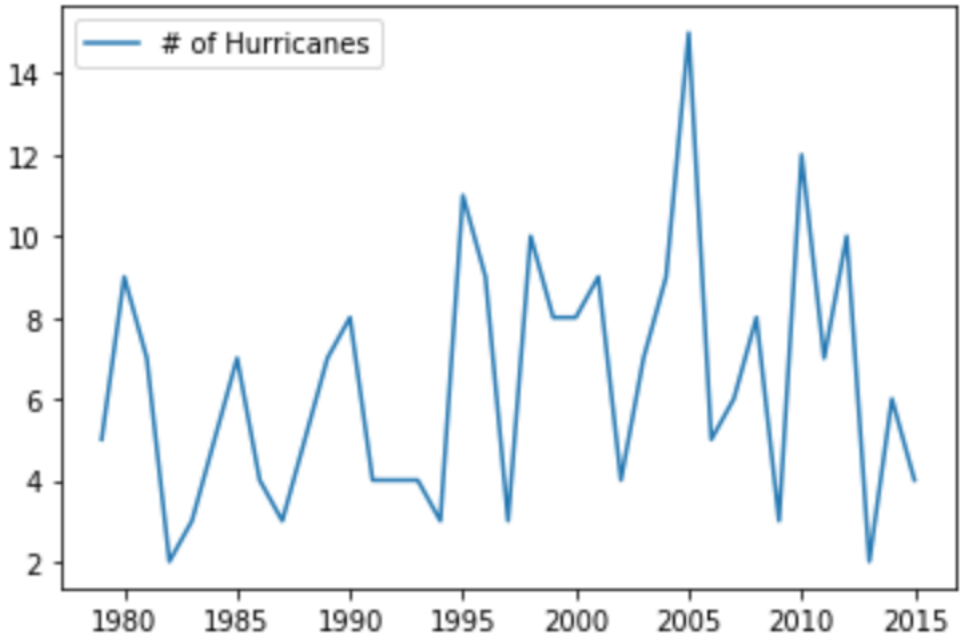


FIGURE 5.3 THE NUMBER OF HURRICANES BETWEEN THE YEARS 1978 AND 2015.

To understand the accumulated cyclonic energy, it was plotted in Fig 5.4. There seems to be a relatively similar pattern as for the number of hurricanes meaning that of the hurricane seasons taken place in the Atlantic basin, all of the demonstrated quite powerful strength. For this reason we shall not underestimate the strength of any of the years in the analysis as each year records hurricanes of strength.

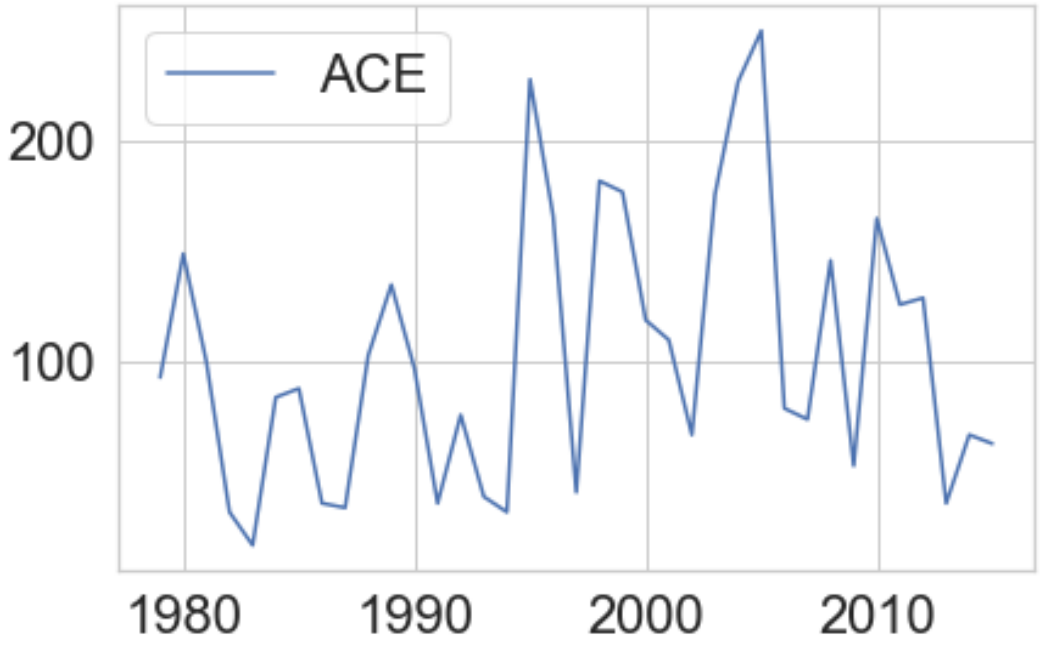


FIGURE 5.4 ACE BETWEEN 1978 AND 2015

Descriptive statistics of the number of hurricanes were summarized in table 5.4, where the results show for both the amount of hurricanes and for the ACE. As seen, the maximum value of the number of hurricanes in a season between the 1978 and 2015 was 15, and the minimum number 2. In most of the cases the number of hurricanes was over 6. In the same way, the ACE shows maximum values of 250 and a minimum of 17. To be noted is that the 2013 storms were in numbers as low as in 1982 but the ACE shows to have been higher in 2013, proving that ACE gives us good guidance for the number of hurricanes in a season but it cannot blindly be trusted as the correlation is not always perfect.

TABLE 5.4 SUMMARY STATISTICS OF VARIABLE "NUMBER OF HURRICANES /YEAR"

	# of Hurricanes	ACE
Mean	6.378378	103.567658
Median	6	93.0
Standard Deviation	3.030872	61.727691
Min	2	17.0
Max	15	250.0

From now on, the ACE will not be included in the analysis as it does not bring additional valuable information. A separate analysis could be done on the ACE if there is interest in rather predicting the accumulated cyclone energy. In theory, the models that were built in this research can be tested on the ACE too.

QBO

For better understanding the QBO variable, a box plot was generated. Figure 5.5 presents the results obtained, with the green line representing the median values of each month. Number 1 represents the month January and 12 December. As seen in the graph, the highest westerly median value could be found in April, and in July the median was at its highest for easterly QBO. Over the other months, the values were quite smoothly divided between the extremes. The negative values in the boxplot represent the easterly QBO values and in this way it can also be compared if there exist any trends on more easterly or westerly oscillation for each month. No considerable trends could be identified.

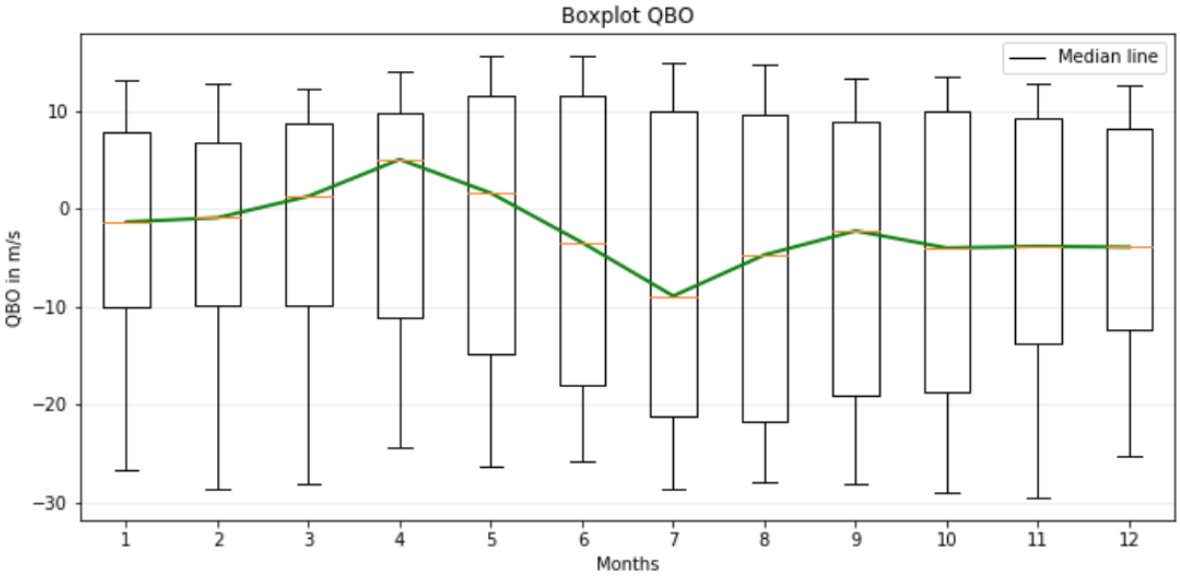


FIGURE 5.5 BOXPLOT ON QBOS

The values are spread between the extreme values in a constant manner and no outliers or extremes outside of the "general boundaries" were found.

As for the number of hurricanes, summary statistics for QBO are found in table 5.5. It is interesting to see that the maximum wind speeds for the easterly QBO have been in general stronger than the westerly ones. The same trend can well be observed in the Boxplot of QBOs (Figure 5.5).

TABLE 5.5 SUMMARY STATISTICS OF QBO

	Mean	Max Easterly	Max Westerly	Standard deviation	Median	Count
Jan	2.51	26.70	13.13	10.09	-1.39	37
Feb	1.44	28.62	12.68	10.17	-0.96	37
Mar	0.75	28.15	12.17	10.86	1.24	37
Apr	0.44	24.38	14.03	12.14	5.0	37
May	1.92	26.28	15.56	13.48	1.56	37
Jun	4.58	25.89	15.62	14.40	-3.49	37
Jul	6.44	28.65	14.85	14.76	-8.95	37
Aug	6.49	27.93	14.66	15.03	-4.75	37
Sep	5.51	28.13	13.21	14.81	-2.30	37
Oct	4.74	29.05	13.38	14.40	-4.04	37
Nov	4.18	29.55	12.79	13.00	-3.88	37
Dec	3.38	25.38	12.55	11.34	-3.96	37

LST

As mentioned, a similar boxplot was generated for LST with a representation of outliers and the median line in green. Figure 5.6 gives a better understanding of the division of the values of LST for each month.

As for the QBO, month 1 represents January month and month 12 represents December. Compared to the QBO boxplot, the LST data set entails more extreme values. The median values seem to be pretty constant with higher outliers between the months of September and February. The mean of the anomalies tend to be negative but close to zero.

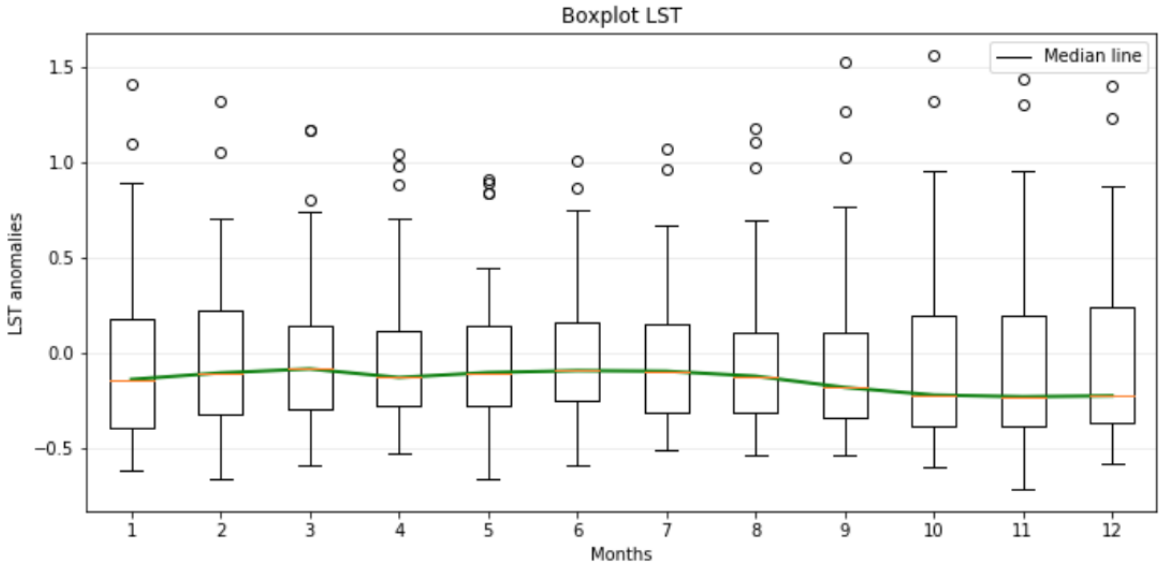


FIGURE 5.6 FIGURE 22 BOXPLOT LST

Descriptive statistics of the variable LST can be found in the table below (Table 5.6).

TABLE 5.6 SUMMARY STATISTICS LST

	Mean	Max	Min	Standard deviation	Median	Count
Jan	-0.001	1.40	-0.62	0.51	-0.14	37
Feb	-0.010	1.32	-0.67	0.48	-0.11	37
Mar	-0.004	1.17	-0.60	0.43	-0.09	37
Apr	-0.017	1.04	-0.54	0.41	-0.13	37
May	-0.019	0.91	-0.66	0.39	-0.11	37
Jun	-0.010	1.00	-0.60	0.39	-0.10	37
Jul	-0.009	1.07	-0.51	0.41	-0.10	37
Aug	-0.017	1.18	-0.54	0.44	-0.13	37
Sep	-0.021	1.52	-0.54	0.51	-0.19	37
Oct	-0.015	1.56	-0.61	0.53	-0.23	37
Nov	-0.011	1.43	-0.72	0.53	-0.23	37
Dec	-0.007	1.40	-0.58	0.51	-0.23	37

As for all the other variables, the two SST variables were described in the table 5.7 and 5.8. The descriptive statistics were taken for each geographical area separately in order to understand their characteristics more in detail and also to detect if outliers or other corrections were needed to be done. Table 9 describes the values for the COBE SST2 dataset and table 10 summarizes the different variables for the HADISST dataset.

SST – COBE SST2

TABLE 5.7 DESCRIPTIVE STATISTICS FOR COBE SST2

	Mean	Max	Min	Standard deviation	Median	Count
Antarctica A	3.5	21.50	-1.95	4.98	2.20	3419708
Antarctica B	3.67	21.23	-2.04	5.44	1.83	4787741
Atlantic	23.47	30.37	13.65	3.28	23.98	684000
Caribbean	27.38	31.08	14.22	2.51	28.28	102000
North Atlantic / Arctic Sea A	9.56	28.61	-2.03	5.71	9.33	1682999
North Atlantic / Arctic Sea B	8.44	23.66	-1.85	4.73	8.40	535498
Pacific A	23.46	34.66	9.74	3.71	24.04	841500
Pacific B	23.99	31.45	7.33	4.51	25.18	204000
West Africa A	21.06	30.35	11.86	4.60	27.27	229500
West Africa B	23.75	30.52	9.74	4.29	20.53	410400
West Mexico	24.39	32.3	9.57	4.48	25.95	684000
ArcticA	-1.19	11.38	-1.87	1.11	-1.62	2006297
ArcticB	-1.33	9.76	-1.87	1.29	-1.80	356998

Some variation does exist in the maximum and minimum values of the SSTs between the different areas. However, most of the areas did perform as predicted. The sizes of the areas are neither align which explains the differences in the count of observations.

HADISST-SST

TABLE 5.8 DESCRIPTIVE STATISTICS FOR HADISST

	Mean	Max	Min	Standard deviation	Median	Count
Antarctica A	4.11	20.64	-1.9	4.95	2.53	3419997
Antarctica B_1	5.40	20.55	-2.1	5.72	5.21	683000

Antarctica B_2	4.40	22.02	-3.0	5.29	3.47	4103999
ArcticA_1	-1.41	12.16	-2.3	1.17	-1.8	364800
ArcticA_2	-1.23	10.84	-1.9	1.39	-1.8	1641591
Arctic_B	0.51	12.31	-2.3	2.84	-1.09	638393
Atlantic	23.58	30.05	13.78	3.27	24.10	684000
Caribbean	27.04	30.59	19.37	1.88	27.45	182400
North Atlantic / Arctic Sea A_1	6.46	25.14	-1.8	4.88	5.72	547199
North Atlantic / Arctic Sea A_2	8.39	26.88	-1.8	5.67	8.40	2462382
North Atlantic / Arctic Sea B	10.55	28.61	-1.8	6.58	9.44	957600
Pacific A_1	22.76	30.05	9.16	4.28	23.88	684000
Pacific A_2	24.39	30.80	-0.51	5.18	26.40	820800
Pacific B_1	26.03	31.57	10.83	4.60	28.37	1368000
Pacific B_2	24.06	30.93	10.73	4.30	25.29	2280000
West Africa A	21.19	30.47	9.91	4.22	20.72	410400
West Africa B	23.74	30.25	10.29	4.04	24.82	912000
West Mexico	24.52	32.80	10.26	4.39	25.99	684000

Similarly as for the COBE SS2 data, HADISST was summarized in order to present the variables that were created on base of the global HADISST data.

5.1.1.7 MISSING DATA TREATMENT

As already discussed, the SST variables were the only ones with missing values. As the data sets included coordinates for land areas it was assumed that the Nan values were of land. To confirm this statement, a small test was conducted to prove it. Based on coordinates from each data set, first areas

with only ocean were picked, then areas on coastline (land and ocean) were tested, and lastly areas with no sea area were tested. Table 5.9 and Table 5.10 summarizes the results found in the tests.

COBE SST2

TABLE 5.9 NANS ANALYSIS. COBE SST2 DATA

Area name	Coordinates	Shape	# of values totally	# of Nans	% Nans of total
Indian Ocean	(40:80), (70:100)	(456, 40, 30)	547200	0	0%
South Pacific Ocean	(60:80), (200:250)	(456, 40, 50)	912000	0	0%
South Atlantic Ocean	(40:70), (330:359)	(456, 30, 29)	396720	0	0%
East coast of Australia	(40:80), (145:170)	(456, 40, 25)	456000	59736	13,1%
Madagascar	(50:90), (30:55)	(456, 40, 25)	456000	122664	26,9%
Japan area	(110:150), (120:160)	(456, 40, 40)	729600	207024	28,4%
Sahara (North Africa)	(100:120), (0:30)	(456, 20,30)	273600	273600	100%
North America	(125:150), (240:260)	(456, 25, 20)	228000	228000	100%
Northern South America	((60:80), (290:310)	(456, 20, 20)	182400	182400	100%

The results in table 11 confirms the assumption made. Areas from the Indian Ocean, the South Pacific Ocean and the South Atlantic Ocean were picked and for these three areas no missing values were found. The coordinates for the tested areas are found in table above. The test for coastline areas gave some missing values, more specifically 13,1%, 26,9% and 28,4% of Nans. The coastline areas tested were the East coast of Australia, Madagascar and the area around Japan. Based on our assumption, these results make sense as part of the area used in land, meaning an output of NaN. The Sahara, North America and Northern South America were tested as the “all-land-area”, not surprisingly giving an output of only missing values. Based on this, we assume that in COBE SST2 data set the Nans are land area and do not affect our analysis results.

HADISST

The same analysis was conducted on the HADISST dataset using the same test-areas. The outputs were similar.

TABLE 5.10 NANS ANALYSIS. HADISST DATA

Area name	Coordinates	Shape	# of values totally	# of Nans	% Nans of total
Indian Ocean	(40:80), (250:280)	(456, 40, 30)	547200	0	0%
South Pacific Ocean	(60:80), (20:70)	(456, 40, 50)	912000	0	0%
South Atlantic Ocean	(40:70), (150:179)	(456, 30, 29)	396720	0	0%
East coast of Australia	(40:80), (325:350)	(456, 40, 25)	456000	58368	12,8%
Madagascar	(50:90), (210:235)	(456, 40, 25)	456000	118104	25,9%
Japan area	(110:150), (300:340)	(456, 40, 40)	729600	194962	26,7%
Sahara (North Africa)	(100:120), (180:210)	(456, 20,30)	273600	273600	100%
North America	(125:150), (60:80)	(456, 25, 20)	228000	228000	100%
Northern South America	((60:80), (110:130)	(456, 20, 20)	182400	182400	100%

The “only-sea areas” did not contain any Nans whilst land areas are presented 100% of missing values. Similarly to the previous case, areas with both land and sea were tested to support the assumption taken.

These two test supports our assumption of missing values being insignificant for the quality of our research.

5.1.2 Dimensionality Reduction results

In all models some dimensionality reduction was made. The correlation analysis was included in each model, as also the scaling. Further, for only two model the PCA dimensionality reduction was done.

5.1.2.1 Correlation Analysis results

The Pearson Correlation Coefficient was calculated for both SST data sets. The correlation for all areas was calculated to the number of hurricanes appearing for each season and the areas for the months of January to July. The first column describes the correlation coefficient, whilst the second column the calculated p-value. The first row for each area is the month of July, and from there they go back one month at a time until January month. The full tables can be found in Annex A.3. Table 5.11 presents some of the highest and lowest correlations found for the data set of COBE SST2, and table 5.12 represents some of the highest and lowest correlations for HADISST. The closer the correlation coefficient was to the value of 1 or -1, the higher correlation was found, and the smaller the p-value, the more likely it was to be able to reject the H0. For COBE SST2 the highest correlations could be found for the area of West Africa B, with a correlation of -0,37, and Antarctica B with a correlation of 0,39. Other high correlation coefficients could be found in the regions of West Africa A (-0,35) and Arctic A (0,36). The lowest correlation coefficients were found in the areas of Pacific B with a correlation coefficient of 0,009, Arctic A with 0,0008 and the Atlantic ocean of 0,007. These values vary between each month and hence it is hard to know what geographical area to use based on this analysis.

COBE SST2 Results:

TABLE 5.11 PEARSON CORRELATION COEFFICIENT COBE SST2

Correlation coefficient	**p-values**
Pacific B:	
-0.2431285501587313	0.14706025358785063
-0.07510982956568367	0.6586200611926456
0.00901718678704496	0.9577575323648138
-0.005881236292024465	0.9724413197460902
-0.03417294451513	0.8408628471338623
0.006627491169942511	0.9689460641495761
0.026243015316410045	0.8774697681047201

ArcticA:	
0.0008655518958541469	0.995943385634001
0.25939588308745654	0.1210550278999894
0.36239156178086906	0.027509213753614016
0.18592605187639455	0.2705624466404285
0.28435014828195865	0.0880676612929987
0.10120074882121313	0.5511850855286774
0.1155816353584741	0.4957363663414306
West AfrikaA:	
-0.1627415316663698	0.3358598870736353
-0.18684064561871788	0.268170557456723
-0.23676770051857576	0.15826705908414188
-0.27202668026337884	0.10336343839464428
-0.28083179389000046	0.09224436181290868
-0.1650955794991429	0.3288208765484777
-0.34679988103162296	0.0354785936128579
West AfrikaB:	
-0.07806689612273601	0.6460432044657357
-0.223549392943256	0.18350816435910094
-0.2546059406421637	0.1283255249046742
-0.3509167511194734	0.03321098671977361
-0.2807228209159556	0.09237611217408764
-0.3667093525823901	0.025584579426540233
-0.3738297431392875	0.02265459379345043
AntarcticaA:	
-0.17796647805326948	0.29196804783071006

0.11476497044240344	0.4988069834313692
0.2779286686839585	0.09580389350343001
0.2555221612980012	0.1269103090615277
0.29430261692796067	0.07704041723247432
0.2957016130900356	0.07558075207150128
0.15852479781258602	0.3486985128958547
AntarcticaB:	
0.09543239632765621	0.5742157783197654
0.06606672139662201	0.6976460629117315
0.10878833180368386	0.5215695845833715
0.26418691046459863	0.1140956589455952
0.3937391459152525	0.015903785672090047
0.3717846824120584	0.02346605030516122
0.2500423209953372	0.1355497858833186
Atlanten:	
0.049949288774937206	0.7690752421989069
0.012809804345004956	0.9400177062571673
-0.0755419098553711	0.6567764811863346
-0.008019635692686357	0.9624272106472073
0.024841984644668788	0.8839657114018604
0.007760905266646509	0.96363857204888
0.049599000812530644	0.7706511193225437

For HADISST the highest correlation coefficients were found in areas of the Caribbean with a coefficient of 0,47 and the Atlantic ocean with correlation of up to 0,53. Both Pacific B and Pacific B_1 show results of high correlation too. The lowest coefficients for HADISST data was found for the areas of North

Atlantic A and North Atlantic A_1 with a correlation coefficient of 0,006, but also for Arctic A (coefficient 0,01). In general, Mexico had quite low correlation coefficients.

Results HADISST:

TABLE 5.12 PEARSON CORRELATION COEFFICIENT HADISST

Correlation Coefficient	**p-value**
Caribbean:	
0.4659129397316978	0.0036581575708933374
0.4411228131572418	0.0062783102559886336
0.2575903881864634	0.12375843005295807
0.053920440905262865	0.7512749885210119
-0.0014267921122741445	0.9933130474877361
-0.13847499662145718	0.41372900607209306
-0.011015774816135297	0.9484061821982568
Pacific B:	
0.4595232371434644	0.004220590878703153
0.32412556615902965	0.050334948807550524
0.39900940386064004	0.014431485539342946
0.3487234575795102	0.03440383610385555
0.32461178512689826	0.04997098107549703
0.31199918879641264	0.06012030598181714
0.028514607889251767	0.8669543561199637
Pacific B_1:	
0.4595232371434644	0.004220590878703153
0.32412556615902965	0.050334948807550524
0.39900940386064004	0.014431485539342946
0.3487234575795102	0.03440383610385555
0.32461178512689826	0.04997098107549703
0.31199918879641264	0.06012030598181714
0.028514607889251767	0.8669543561199637
ArcticA:	
0.06554515009444067	0.699921618124107
0.06054322484485247	0.7218728298724836
0.016936903043067927	0.9207459288652459
-0.09026735958402618	0.5951997967391751
-0.19338081679923289	0.2514719375198012
-0.15353812734916836	0.36426039611130134
-0.26396591843373446	0.11440988983996976
Atlanten:	
0.4362081506308303	0.0069559144854281964
0.5259180148194696	0.000828666230038813
0.4646766669351435	0.003761587413598012
0.45631001811059035	0.004530724525519118
0.3540257222770984	0.03157831573730598
0.41093525079534815	0.01151903225317533
0.3458903758293768	0.03599621276900224
North AtlanticA:	
0.20197757481919099	0.23059915728007552

-0.117447333707466	0.48875781555914666
-0.2278870456460758	0.1749295965048606
-0.22525266952611095	0.1801047583476557
-0.09104993093152138	0.5919990111706845
-0.006013425371880571	0.9718221431893337
0.08418239026531465	0.6203410093026153
North AtlanticA_1:	
0.20197757481919099	0.23059915728007552
-0.117447333707466	0.48875781555914666
-0.2278870456460758	0.1749295965048606
-0.22525266952611095	0.1801047583476557
-0.09104993093152138	0.5919990111706845
-0.006013425371880571	0.9718221431893337
0.08418239026531465	0.6203410093026153
Mex:	
-0.2246185439159463	0.1813665571119686
-0.17106402489010464	0.3113877420263247
-0.1602927909024915	0.3432796741032282
-0.19740825424470973	0.24154172478409933
-0.10837285726458902	0.5231707312950495
-0.07785573359892341	0.6469381620947653
-0.0863785551278606	0.6112160886413271

At it can be seen from the results presented above, the data set HADISST showed much higher correlations to the hurricane data set. Also, results made better sense which can be concluded based on expert knowledge regarding sea surface temperatures and their behaviour in the different areas. The higher p-values and lower correlations of the COBE SST2 dataset indicated that better prediction can be obtained using the HADISST data. For these reasons, the HADISST was used in the research as it is considered to be of better quality.

As mentioned in Chapter 4, only values with at least two correlation coefficient of higher than 0,2 (or smaller than -0,2) was used in the analysis. Many of the SST variables showed a tendency to have higher correlation coefficient in months closer to the hurricane season, but as the west coast of Africa show opposite tendency we included all months, as the temperatures of the African west coast are important in the formation process of Atlantic hurricanes. The SST variables that have two or more months of a correlation coefficient higher than 0,2 are:

Caribbean, Pacific A_1, Pacific B, Pacific B_1, Arctica A_1, West Africa A, West Africa B, Atlantic, North Atlantic A, North Atlantic A_1, North Atlantic B

These variables were included in the analysis as they are considered to bring information that has an impact on the predicted value.

The variables left out due to low correlations were:

Pacific A, Arctic A, Arctic B, Antarctica A, Antarctica B, Antarctica B_1, Mexico

5.1.2.2 PRINCIPAL COMPONENT ANALYSIS RESULTS

As the data set now used has shown to be HADISST, the variables were put together with the LST and QBO variables to have the predicting variables for our dependent variable Y, Number of Hurricanes. This data was centralized and normalized as described in chapter 4.

The original shape of the X-set was (37, 142), as we have 37 years (rows) in our analysis and 142 variables (columns), but after the Pearson Correlation analysis the lowest correlated variables were removed and the new shape of the X-set was (37, 93). The shape of our Y arrays was (37, 0) as it only consists of one column of 37 years.

By using sklearn libraries the model "train-test-split" was used to create the train and test sets for the analysis. A split of 0.4 was used. The following properties of the test- and training-sets were found (Table 5.13):

TABLE 5.13 TRAINING- AND TEST-SET SHAPES

X-test set	(15,93)
X-train set	(22, 93)
Y-train	(22, 0)
Y-test	(15, 0)

As described in Chapter 4, before running the PCA analysis, the fit to the "Baseline" Random Forest model was done. The scaled data was fitted to the current model and based on this, it can be compared if the generated models bring higher accuracy or not

Baseline Accuracy: 45.19%

The accuracy of the model was calculated to be 45.19% without any kind of data processing. If the results obtained after data processing were higher than the baseline model, it was proven that the processing had been fruitful.

Hereafter the PCA was calculated.

By using the PCA function, the Cumulative Explained Variance was calculated and plotted in figure 5.7. The graph shows a red line which presents the line where added components do not bring useful information to explaining the variance. After 15 components we can see the explained variance slows

down for each added component. In general, 80% of explained variance is a good measure for a model and it can be seen that this 80% can be found already at the 9th component.

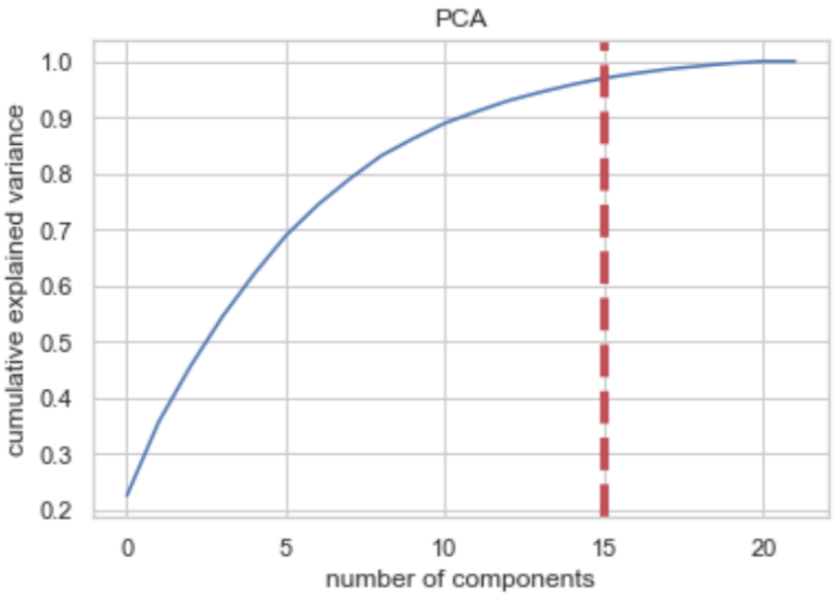


FIGURE 5.7 PCA PLOTTED

Figure 23 shows that the first components explain the best the variance of the data, but once the number of the components increases it can be noted that marginal increase of explained variance for each component decreases.

In order to support figure 5.7, table 5.14 was generated. This table shows the values for the explained variance ratio for each component.

TABLE 5.14 THE CUMULATIVE AND EXPLAINED VARIANCE OF DATA

	Cumulative Variance Ratio	Explained Variance Ratio
0	0.225177	0.225177
1	0.356618	0.131441
2	0.456207	0.099589
3	0.544236	0.088029
4	0.620592	0.076356
5	0.689171	0.068579
6	0.743852	0.054681
7	0.790146	0.046293
8	0.831675	0.041529
9	0.862098	0.030423
10	0.889815	0.027718
11	0.910567	0.020752
12	0.929873	0.019306
13	0.944898	0.015025
14	0.958386	0.013487

The cumulative variance ratio shows how much of the model is explained by the components in total, adding up for each added component in an order where the first component is the one with the highest explained variance ratio and so on. The table above (table 5.14) shows the 15 highest components. From this table it can be decided on how many component will be used when taking into account how much of the components describe the data. The nine first components were saved as its own training set (X_train_pca) and was used in the analysis when constructing the RFs. The nine first components contains over 80% of the explained variance, and for this reason, it was believed to remain enough information for further research. See table 5.15 for the new shape of the new test and train sets obtained.

TABLE 5.15 TRAINING- AND TEST-SET SHAPES AFTER PCA TREATMENT

X-test set	(15,9)
X-train set	(22, 9)
Y-train	(22, 0)
Y-test	(15, 0)

By fitting this new data set consisting of the first nine components to the RF model, a new model accuracy was obtained.

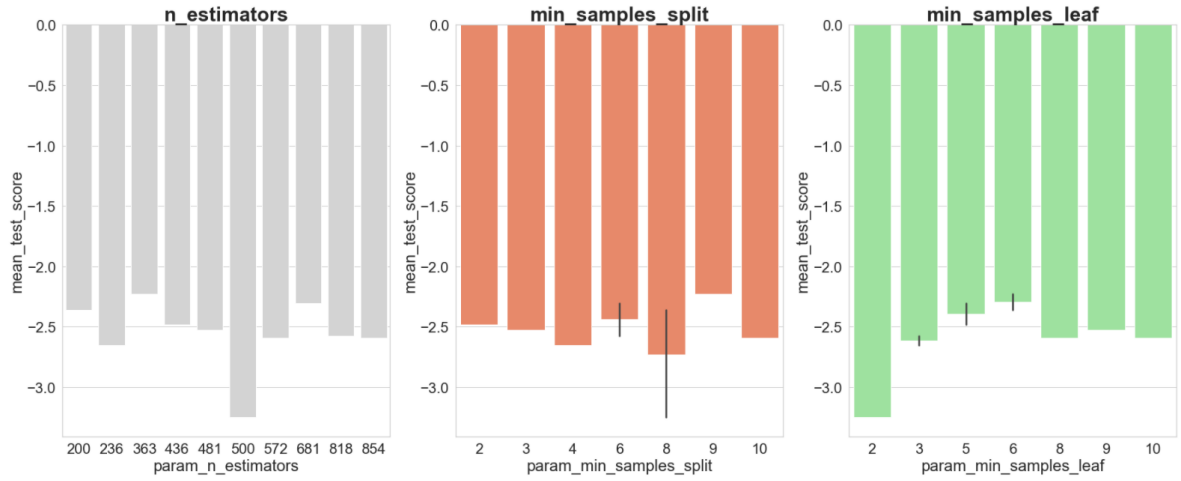
Baseline Accuracy after PCA: 53.87%

The accuracy of the model increased with 8% by reducing the dimensions with PCA. This means that there is a lot of unnecessary variables in the original data set and by reducing the dimensions of it a more accurate data set can be worked with.

5.2 Predictive models results

5.2.1 Hyperparameter tuning

Once the dataset had been reduced with the PCA, the random forest algorithm was tuned into the best possible ones. The Random Search CV hyperparameter tuning was done with sklearn and in order to find different options for the grid CV tuning. Ranges of possible best hyperparameters were proposed into the model and figure 5.8 summarizes the results obtained from the random search CV.



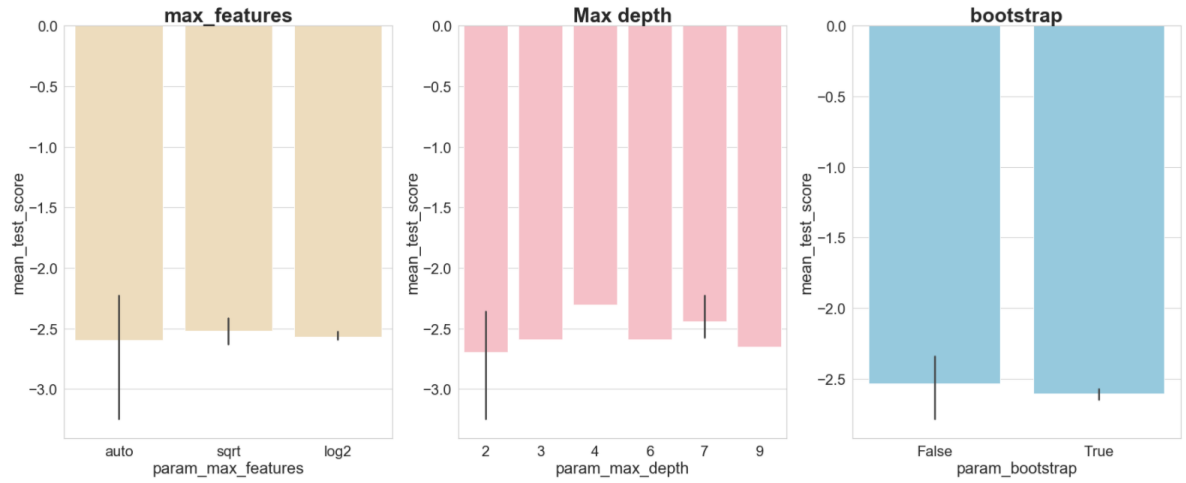


FIGURE 5.8 RANDOM SEARCH CV RESULTS (WITH PCA)

From the above-given graphs, a direction of the best hyperparameters could be concluded and the "best ones" were further imputed to the grid search CV. The hyperparameters that have the smallest Mean absolute score, MAE in our case, are the optimal ones that were aimed to be used. The Random Search CV suggested the following best hyperparameters;

n_estimator: 363

min_sample_split: 9

min_sample_leaf: 6

max_features: auto

max:depth: 7

bootstrap: False

These hyperparameters are one possible solution for a "best" model, but as we do the Grid Search all of the possible combinations were tested and confirmed if the Random Search "best model" actually was the optimal one.

Based on these results, a Grid Search CV could be built. The Grid Search CV tested all possible combination from the parameters given, as the above found "best hyperparameters" were introduced to the model. The values inserted to the Grid Search CV are presented in table 5.16.

TABLE 5.16 VALUES INTRODUCED TO THE GRID SEARCH CV

Hyperparameter	Values inserted in Grid Search CV

n_estimator	200, 363, 436, 481, 681
max_feature	Auto, sqrt
max_depth	2, 3, 4, 7
min_sample_split	2, 3, 6, 9
min_sample_leaf	5, 6, 9, 10
bootstrap	True, False

Running the Grid Search CV, the following results were obtained:

n_estimator: 200

min_sample_split: 3

min_sample_leaf: 6

max_features: sqrt

max_depth: 7

bootstrap: False

```
{'bootstrap': False,
 'max_depth': 7,
 'max_features': 'sqrt',
 'min_samples_leaf': 6,
 'min_samples_split': 3,
 'n_estimators': 200}
```

The results obtained from the Grid Search are the best hypermeters that were suggested be used in the random forest. Interesting to see is that the given hyperparameters are different from the “best” ones obtained from the random search. As the grid search tries all the possible combinations possible, the output of the grid search was used in the tuned model.

NO PCA – tuned model

As we are comparing our results with other models too, a tuned model for *no PCA treated* data was created. The same process was done as in the previous part, just without doing the PCA. This was done as our analysis aims to find the best possible data processing methods and algorithms for the forecast of seasonal hurricanes. Even if the PCA showed to improve the accuracy of the baseline model, a tuned model of RF was built so that more reliable results could be found to support the decision of the best one. The initial scaled data was used in order to do a Random Search CV hyperparameter tuning for the RF algorithm. Figure 5.9 shows the results found.

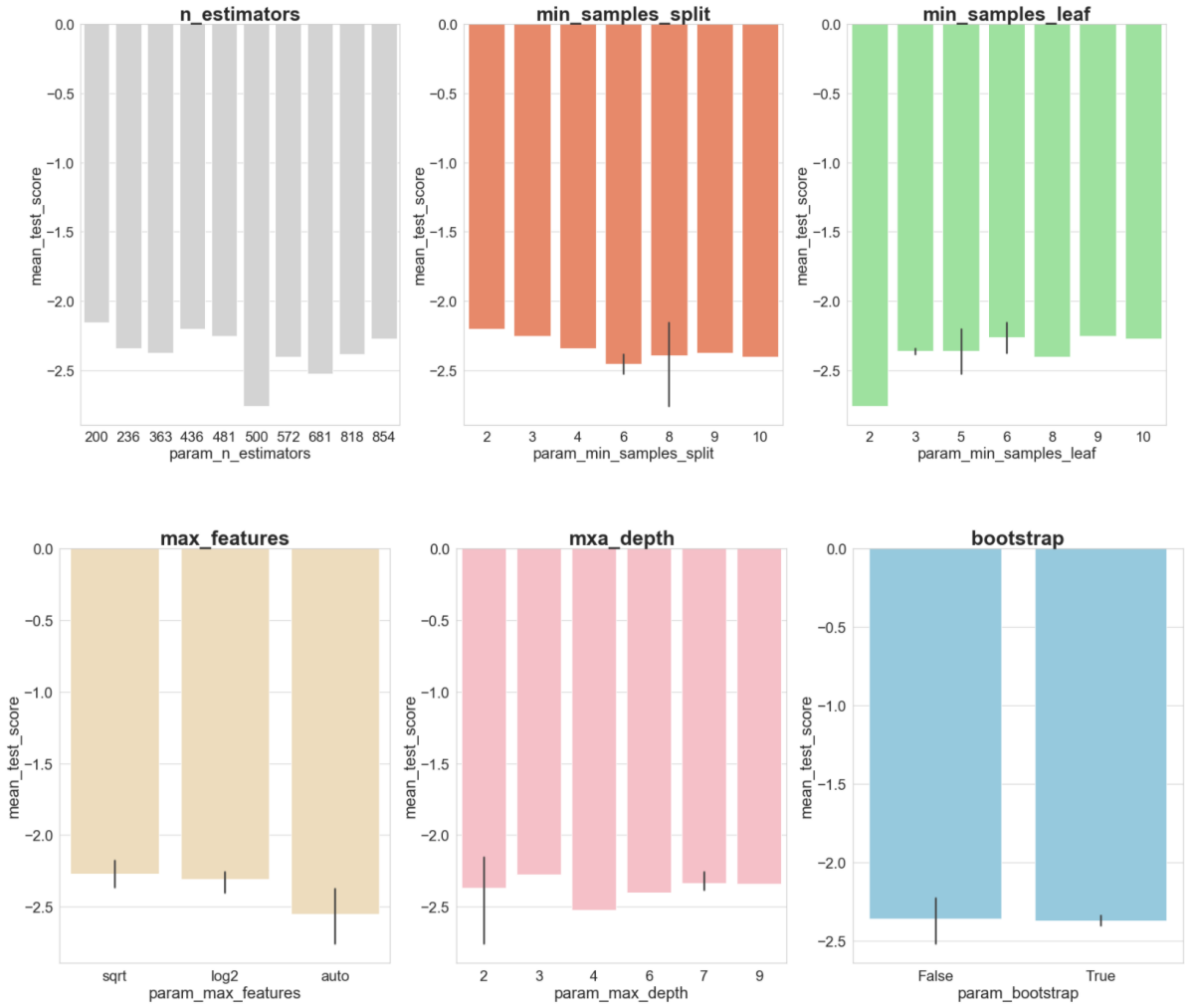


FIGURE 5.9 RANDOM SEARCH CV RESULTS (NO PCA)

From the above-given figure, a direction could be concluded and the best hyperparameters were further taken to the grid search CV. The Random Search CV suggested the following best hyperparameters;

n_estimator: 200

min_sample_split: 8

min_sample_leaf: 6

max_features: sqrt

max_depth: 2

bootstrap: False

Following the same steps as when treating data with PCA, the values from Random Search CV were introduced to the Grid Search CV. Table 5.17 demonstrated the values used for the Grid Search CV with no PCA.

TABLE 5.17 VALUES INTRODUCED TO THE GRID SEARCH CV (WITH NO PCA)

Hyperparameter	Values inserted in Grid Search CV
n_estimator	200, 236, 436, 481, 854
max_feature	Sqrt, log2
max_depth	2, 3, 7, 9
min_sample_split	2, 3, 4, 8
min_sample_leaf	5, 6, 9, 10
bootstrap	False, True

The results obtained from the Grid Search CV were:

n_estimator: 200

min_sample_split: 4

min_sample_leaf: 5

max_features: sqrt

max_depth: 2

bootstrap: False

```
{'bootstrap': False,
 'max_depth': 2,
 'max_features': 'sqrt',
 'min_samples_leaf': 5,
 'min_samples_split': 4,
 'n_estimators': 200}
```

Similarly, as in the test with PCA data, the grid search tuning did not give the same hyperparameters as the random search. The output from the grid search was used for the tuned random forest model for the data with no PCA.

5.2.2 RANDOM FOREST MODELS EVALUATION

Once the best hyperparameters were found, the different RF models could be created. Two different RF regressions were created based on the best hyperparameters found for each data set (no PCA and with PCA). The algorithm was trained with both data sets and they were evaluated separately with the test data. Two more models (model 3 and model 4) were also built so that a better comparison could be done. These two models were not tuned for their hyperparameters.

Model 1. PCA with Tuning

The tuned RF model was trained on the training set that had been processed with PCA. Thereafter the trained model was tested on the test set. Results on the performance of the model were printed in order to be able to compare with other models. Predictions of Y was built on both the training sets and thereafter for the test sets. An error analysis was graphed with the Root Mean Square Error on the y-axis and the number of predictions on the x-axis. In figure 5.10, the orange line demonstrates the errors obtained when creating a regression forecast using the RF algorithm and the X-test data, comparing the results to the Y-test data. The blue line presents the results of a similar procedure done on the train dataset.

FIGURE 5.10 REGRESSION ERROR FOR BOTH TRAINING AND TEST SETS (MODEL 1)

As seen from the graph above, the training set seems to perform better than the test set. The RMSE error is higher for all test set predictions while the training set predictions remain almost all under 2.2. The evaluation metrics are calculated for both the training and test sets to be able to compare more in detail.

Min error for test set = 2.50

Min error for train set= 2.12

The prediction on Y generated from the test set was plotted out in figure 5.11, with the predicted values on the x-axis and the true Y-test values on the y-axis. Figure 5.12 demonstrates a similar graph with the results based on the training set.

FIGURE 5.11 REGRESSION RESULTS MODEL 1 (TEST)

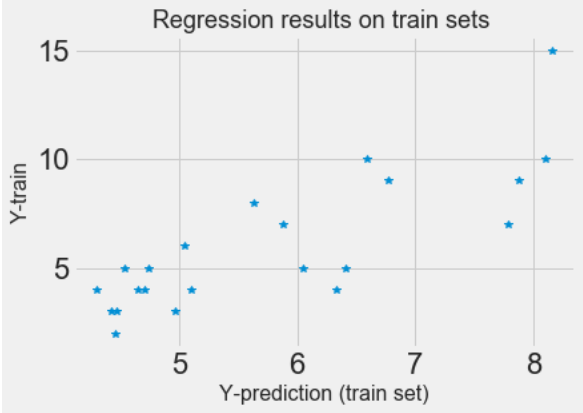


FIGURE 5.12 REGRESSION MODEL 1 (TRAIN)

It can be seen that the prediction on the test set has a more linear pattern than the predictions on the test set. This was also well demonstrated when calculating the evaluation metrics for the predictions above.

To better be able to analyse these results, various evaluation metrics were calculated for both the test and training set predictions. These metrics results have been summarized in table 5.18.

TABLE 5.18 EVALUATION METRICS RF MODEL (WITH PCA AND TUNING)

Metrics	Test Set	Train Set
Max Error	4.36	6.93
MAE	2.19	1.62
MAPE	57.17%	
MSE	6.45	4.65
RMSE	2.54	2.15
R^2	0.19	0.50

These metrics do not tell much alone and without any other model it is hard to know how well the model performs. In general, it can be said that the R^2 being 0.19 is relatively low, especially when the same value for the train set is 0.5. After calculating the same metrics for the other models, a proper comparison could be made. In general, it can be concluded that the performance of the test set

predictions were lower than for the train set, except of the max error. The max error is lower for the test set predictions than for the train set ones.

Model 2. No PCA with Tuning

To be able to compare if the PCA increased the accuracy of the regression model, the RF model tuned for no PCA was evaluated in the same way as the model above.

Figure 5.13 represents the regression errors for the model output of Random Forest without PCA treated data.

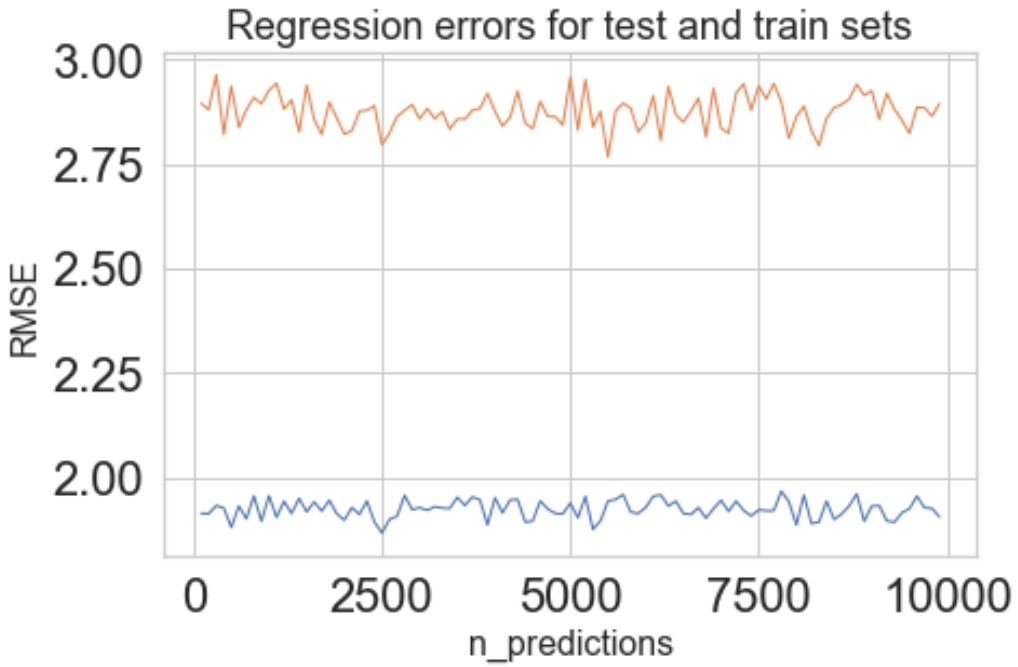
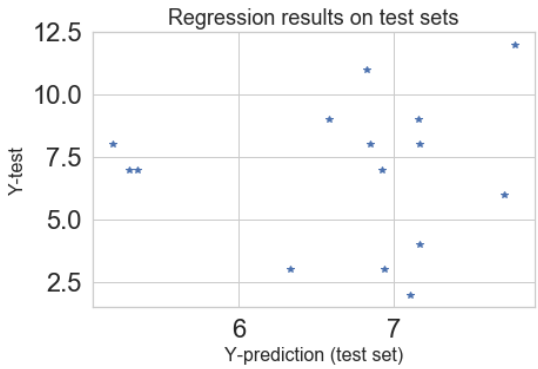


FIGURE 5.13 REGRESSION ERROR FOR BOTH TRAINING AND TEST SETS (MODEL 2)

The results show similarly pattern as for the PCA processed data. The prediction errors for the test-data set (orange) and the errors for the training data (blue) are considerably different with the regression error made on the test set, over 2.75 meanwhile the train set regression error remains under 2.0. Now, the minimum error found for the test set was 2.77, and the minimum error for the train set was 1.87.

Figure 5.14 and figure 5.15 plots the predictions against the true values for both the test and train sets.



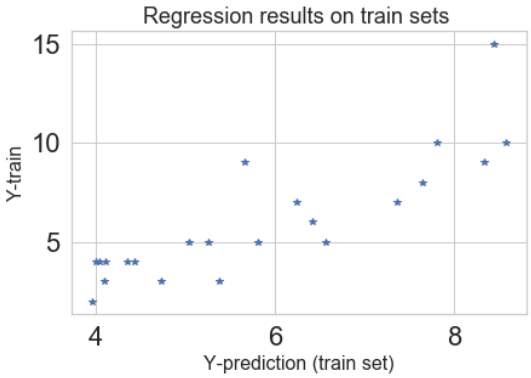


FIGURE 5.14 REGRESSION RESULTS MODEL 2 (TEST)

FIGURE 5.15 REGRESSION RESULTS MODEL 2 (TRAIN)

The regression results on the test set were more disperse and spread out compared to model 1. It is hard to identify a pattern for the regression results which can be supported by the evaluation metrics shown in table 5.19 below. The regression on the train set shows less deviation but still the bigger the prediction value is, the more dispersion there seemed to take place.

TABLE 5.19 EVALUATION METRICS RF MODEL (WITH TUNING BUT NO PCA)

Metrics	Test Set	Train Set
Max Error	5.20	6.94
MAE	2.54	1.24
MAPE	44.35	
MSE	8.38	3.63
RMSE	2.89	1.90
R^2	-0.06	0.61

The MAPE was lower than for the model that has reduced the dimensions of the data with PCA. As expected, the R^2 was very low for the regression prediction done on the test set, turning out to be -0,06. This means that the goodness of fit is negative which in this analysis is not desirable. The maximum error was also bigger for the regression done on the train set than for the test set, similarly as in the case of model 1. Regarding the other metrics, the prediction on the train set performs better than the one done on the test set. The MSE is very big for the test set prediction compared to the one for the train set.

Model 3. No PCA and no tuning

The third RF regression model was already created. This time no tuning was done to the model, as part of the research was to study if hyperparameter tuning increases the quality of our research. The Random Forest algorithm was set to its default hyperparameters with a random number of n_estimators and random replacing in bootstrapping. The data used will neither be processed with PCA for dimensions reduction. This model is most modest one of the models tested, in fact it is the "base-line" model created in the beginning of the research. The interesting point of this model is to compare it to the model that is tuned with data not processed with PCA, Model 2. The regression errors retrieved from model 3 were presented in figure 5.16.

FIGURE 5.16 REGRESSION ERROR FOR BOTH TRAINING AND TEST SETS (MODEL 3)

Different from the two previous models, much more variability of the errors could be found. The regression errors goes all the way up to 4 for the test set and later smoothened out to vary just below 3. On the other hand the regression error appears to be lower for the training set regression varying between 2 at its maximum and values below 1.

The same metrics were calculated as for the previous two models.

Min error for test set = 2.49

Min error for train set= 0.79

Figure 5.17 and figure 5.18 plots the regression results for the test and train set predictions.

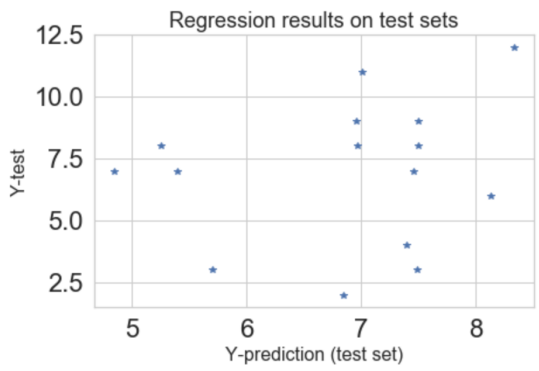


FIGURE 5.17 REGRESSION RESULTS MODEL 3 (TEST)

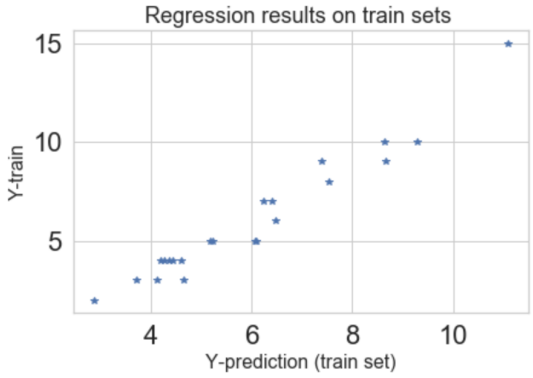


FIGURE 5.18 REGRESSION RESULTS MODEL 3 (TRAIN)

Now, a more evident difference could be seen between the test set prediction and the prediction done on the training set. Looking at figure 5.18, an almost linear perfect looking fit is shown for the training set while the test set based predictions show a collection of widely spread out prediction values. When the difference between the two predictions is this big, there are quite obvious signs of some error in the model. Further, table 5.20 presents the metrics for the model and as we can see the difference in R^2 is significant between the train and test set predictions.

The evaluation metrics were calculated for the model to be able to compare its performance with the other models.

TABLE 5.20 EVALUATION METRICS RF MODEL (WITH NO TUNING AND NO PCA)

Metrics	Test Set	Train Set
Max Error	4.86	3.91
MAE	2.48	0.87
MAPE	45.19%	
MSE	7.94	1.38
RMSE	2.82	1.17
R^2	-0.00	0.85

As said before, the evaluation metrics alone do not tell the reader much about the model but when the metrics differ significantly between the test and train set some conclusions can be done. As in this model, especially the MSE and the R^2 values were significantly different between the two sets. If the difference is very big between the two data sets it often is due to overfitting of the training model. The R^2 for the training set is as high as 0.85, with a corresponding value of zero for the test set alarms of a

model that is not able to deal with new data introduced. A model that is not capable of implementing the model trained on new data is not going to be a suitable one for future predictions.

Model 4. PCA and no tuning

The fourth RF regression model was created based on PCA and the Random Forest algorithm without tuning (as in model 3). This model works as a final comparison for the three other ones when searching for the best RF model for seasonal hurricane analysis. The dimensions of the data was reduced with PCA and thereafter the data was introduced to the random forest with j+1 random threes and a total random sampling of the model. The errors retrieved from the regressions are presented in figure 5.19.

FIGURE 5.19 REGRESSION ERROR FOR BOTH TRAINING AND TEST SETS (MODEL 4)

Similar to model 3, model 4 experiences more variability in the errors than the models that has been tuned. The errors for the regression on the test and train set were pretty close to each other for the first predictions but once the model was tested further, the error decreases for the training set more than for the predictions made on the test set. The regression errors on the training set were lower than for the ones that are tuned but the errors on the test set are quite similar.

As for the three previous models the minimum errors were calculated for both predictions;

Min error for test set = 2.30

Min error for train set= 1.08

As for Model 3, the predictions were plotted for both sets and as expected, the training set predictions seems to have high goodness of fit whilst the test set was not performing as well. Figure 5.20 and Figure 5.21 visualized the prediction results below.

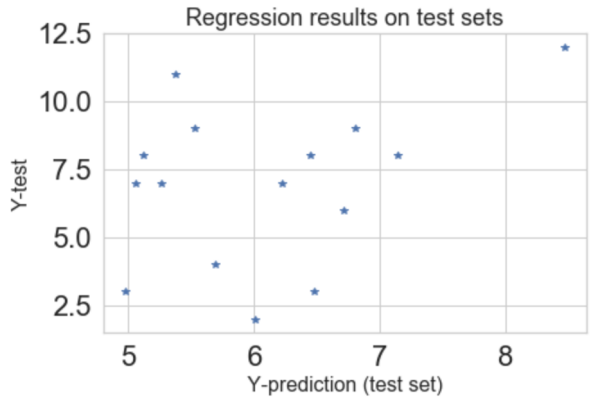


FIGURE 5.20 REGRESSION RESULTS MODEL 4 (TEST

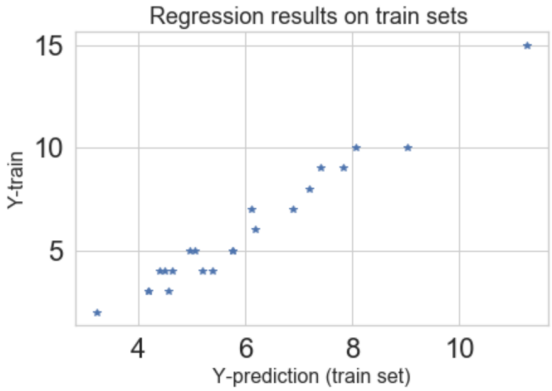


FIGURE 5.21 REGRESSION RESULTS MODEL 4 (TRAIN)

A close to perfect fit is obtained for the regression results on the train set. The results on the test sets perform worse and do not display a similar pattern as the model done on training data.

Again, the evaluation metrics were calculated for the model to be able to compare its performance with the other models (table 5.21).

TABLE 5.21 EVALUATION METRICS RF MODEL (WITH NO TUNING AND WITH PCA)

Metrics	Test Set	Train Set
Max Error	5.63	3.75
MAE	2.42	1.01
MAPE	52.68%	
MSE	7.68	1.64
RMSE	2.77	1.28
R^2	0.03	0.82

Model 4 seems to overfit the training data in the same way as model 3, but its accuracy remains higher than for model 3, where dimensions has not been reduced with PCA. R^2 on the train set is not as high as for model 3, but a goodness of fit of 0.82 is still extremely high. The test increased its performance a bit regarding the R^2 as it now results to be 0.03, not zero anymore. However, the difference between the two sets are certainly big and it is indicating a not so good model that is over-fitting based on the train set. As mentioned, the maximum errors are quite similar but the MSE shows big difference in this model too. As the differences of MSE and R^2 are as high, the model cannot be reliable for future predictions. Interesting to note however, is that the MAPE is the second largest for all of the studied models so far.

The four models generated show quite different results on the metrics used to evaluate and therefore summarizing figures 5.22 to 5.28 were plotted.

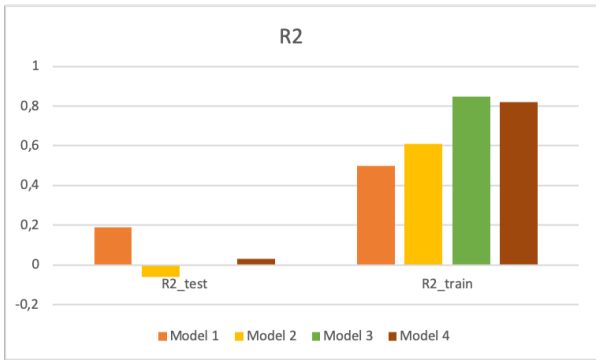


FIGURE 5.22 SUMMARY OF R2

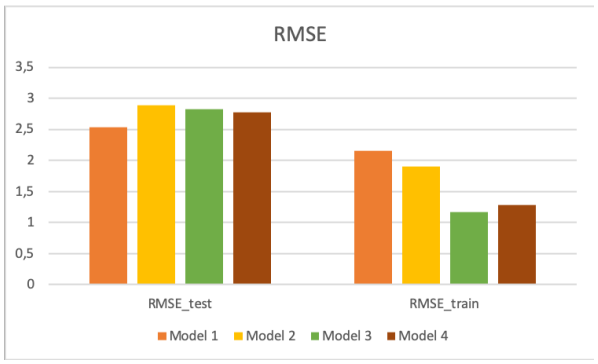


FIGURE 5.23 SUMMARY OF RMSE

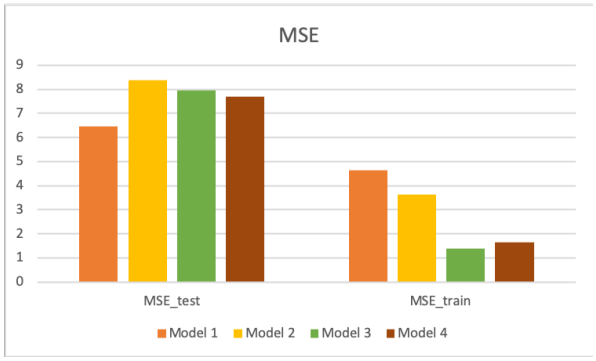


FIGURE 5.24 SUMMARY OF MSE

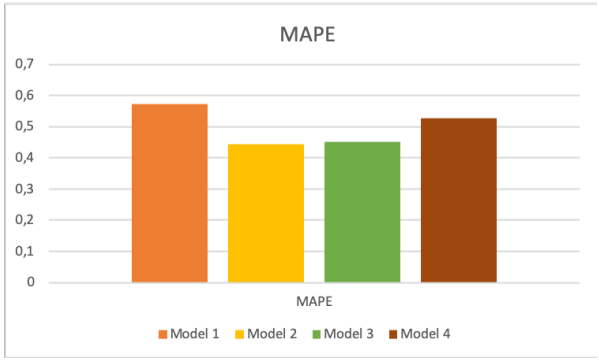


FIGURE 5.25 SUMMARY OF MAPE

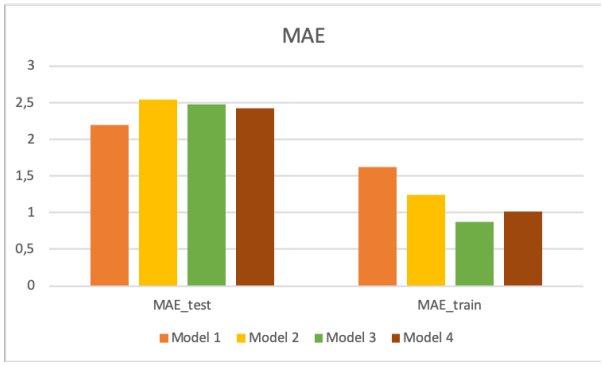


FIGURE 5.26 SUMMARY OF MAE

FIGURE 5.27 SUMMARY OF MAX ERROR

FIGURE 5.28 SUMMARY OF MIN ERROR

As we have seven different metrics to evaluate our models, some variations on the model performances can be found. The evaluation metrics estimate different models to perform the best, and hence, not all metrics could be taken into account when identifying the best model for a seasonal forecast of hurricanes.

As briefly discussed earlier, Model 3 and 4 seem to over-fit as the difference between the R^2 for the train and test sets was considerable. One shall note that the difference was big for Model 1 and 2 also, but not in the same magnitude as for the two latter ones. The goodness of fit for Model 2 resulted negative, which can be interpreted as a negative relationship between the predictions and real values. Based on the goodness of fit, Model 1 seemed to perform the best.

Comparing the RMSE and MSE of the models, Model 1 performs clearly the best compared to the three other ones. Model 1 had the lowest MSE and lowest RMSE. If comparing the metrics for the training set it could at first glance look like the models without tuning were doing better, but again, the excellent performance of the two last ones on the training set based predictions could be believed to be based on over-fitting of the models. This being said, Model 1 was considered to be the best one, when taking these metrics into account, as it performed the lowest error on the test model predictions and did not seem to overfit either.

The MAPE was discussed already earlier, but as represented, the accuracy seemed to increase with the dimensions reduction done with PCA. That can be seen in figure 5.25, which presents that the accuracy is highest for the two models which have been reduced. Model 1 displayed the highest accuracy, while model 2 presented the lowest accuracy of them all.

The same trend could be shown for the MAE values. Model 1 illustrated the lowest error when assessing the predictions on the test set, but relatively high error when comparing to the other models for predictions on the training set. Still, the test set prediction results were the important ones, and it is a good sign that the difference between these two values was not too big.

It was curious to note the difference in performance for maximum and minimum values. Model 4 seemed to perform the best regarding the minimum errors, whilst model 1 displays errors similar to

model 3. With reference to the maximum error, model 1 showed the lowest values, but one should notice the high maximum error the model presented for the training set predictions.

When analysing these metrics, it was relatively straightforward that model 1 performed the best of all four models. This to say, PCA dimensions reduction and hyperparameter optimization was proven to increase the performance of the forecast results, and can be considered to be the best random forest model in this dissertation.

5.2.3 LINEAR REGRESSION MODEL EVALUATION

Model 5, the linear regression model was built after scaling the test and train data. The model was fit into the training set and a regression was done afterwards on the test set. As we do not compare as much in detail the linear regression as the random forest models, only the test prediction results are presented. This is only due to the fact that this (and the two following models) were mainly used as verification models to support the final results.

The results for the linear regression are summarized in table 5.22.

TABLE 5.22 LINEAR REGRESSION RESULTS

Metrics	Test Set
Min Error	2,3
Max Error	6.63
MAE	2.54
MAPE	49.02%
MSE	8.63
RMSE	2.94
R^2	-0.08

The model presents quite similar results as the other models, slightly worse. The goodness of fit is worse than the other models and the MSE is higher. However, the MAPE, 49.02%, shows better accuracy than the two models that were not processed with PCA and used the RF algorithm.

5.2.4 GRADIENT BOOSTING REGRESSION MODEL EVALUATION

Similarly as for the linear regression, the gradient boost model was built on the data after scaling the data set. As said, no tuning was done and the following results could be found (table 5.23).

TABLE 5.23 GRADIENT BOOSTING REGRESSION RESULTS

Metrics	Test Set
Min Error	2.37
Max Error	4.56
MAE	2.29
MAPE	50.10%
MSE	7.45
RMSE	2.73
R^2	0.06

The MAPE increased to over 50%, which is better than the majority of the models. The R^2 was now positive, even though it was still very low. The MAE was found to be 2.29, being better than for the linear regression model, similarly as the MSE and RMSE too. The maximum error was pretty much lower than for the linear regression but on the other hand the minimum error was higher for the gradient boost regression than for the linear model.

5.2.5 VOTING REGRESSION MODEL EVALUATION

The last model in this analysis, model 7, the voting regression model is a combination of 3 models used. The average performance was taken from the linear regression, the gradient boost and the random forest model. Predictions results from the model are to be found I table 5.24.

TABLE 5.24 VOTING REGRESSION RESULTS

Metrics	Test Set
Min Error	2.49
Max Error	4.15
MAE	2.36
MAPE	49.24%

MSE	7.06
RMSE	2.66
R^2	0.11

Comparing the min and max errors, model 7 is performing as one of the best models, with the maximum error being the lowest of all models. The MAE can be considered good when comparing to the other models, and for the MSE, only model 1 was scored better. The same trend could be found for the RMSE and R^2 . Model 1 is the only model that could outperform the voting regression.

5.3 COMPARATIVE ANALYSIS OF MODELS

As discussed earlier, the main focus of the dissertation was to find the optimal random forest model. However, it was decided to include three additional algorithms to get a wider understanding of the performance of the predictions when compared to other methods. The five main evaluation metrics were summarized for each prediction model, see table 5.25.

TABLE 5.25 PREDICTION RESULTS SUMMARY

	Model 1	Model 2	Model 3	Model 4	Model 5 - LR	Model 6 - GB	Model 7 - Vo
MAE	2,19	2,54	2,48	2,42	2,54	2,29	2,36
MAPE	0,5717	0,4435	0,4519	0,5268	0,4902	0,501	0,4924
MSE	6,45	8,38	7,94	7,68	8,63	7,45	7,06
RMSE	2,54	2,89	2,82	2,77	2,94	2,73	2,66
R2	0,19	-0,06	0	0,03	-0,08	0,06	0,11

Respecting the Random Forest-based predictions, Model 1 represents the best scores of all four. Model 5, the Linear Regression prediction has worse scores than Model 1 in all five evaluations. However,

interesting to note is that when comparing the Linear regression results to the basic random forest (model 3), MAPE is higher for linear regression. As all other metrics are better for Model 1, we still consider it to be a better overall model than the linear regression.

The Gradient Boost regression performs overall better than the linear regression, but still not as well as Model 1. The general RF model was outperformed by the gradient boost, indicating that perhaps the random forest algorithm is not the only good one for seasonal prediction. Due to this, the last model, the voting regression model, performs rather well too. Model 3 is outperformed by the voting regression, but as observed, Model 1 remains the best model yet.

Interesting to note is the bad performance of model 2. Tuning was done on data that was not reduced with PCA, displaying results that perform the worst compared to all models. Only the linear regression model shows higher values on the MSE, RMSE and R^2 metrics. This is a compelling remark as it highlights the importance of pre-processing of the data. Comparing model 2 and 3, it is shown that the results are worse for model 2 than the base-line prediction, meaning one should be very careful when doing tuning of models as it can harm the model negatively.

Figure 5.29 plots the predictions made on the linear regression, the gradient boost, the random forest model and the actual regression points (blue line). As the graph shows, possibilities to increase accuracy do exist.

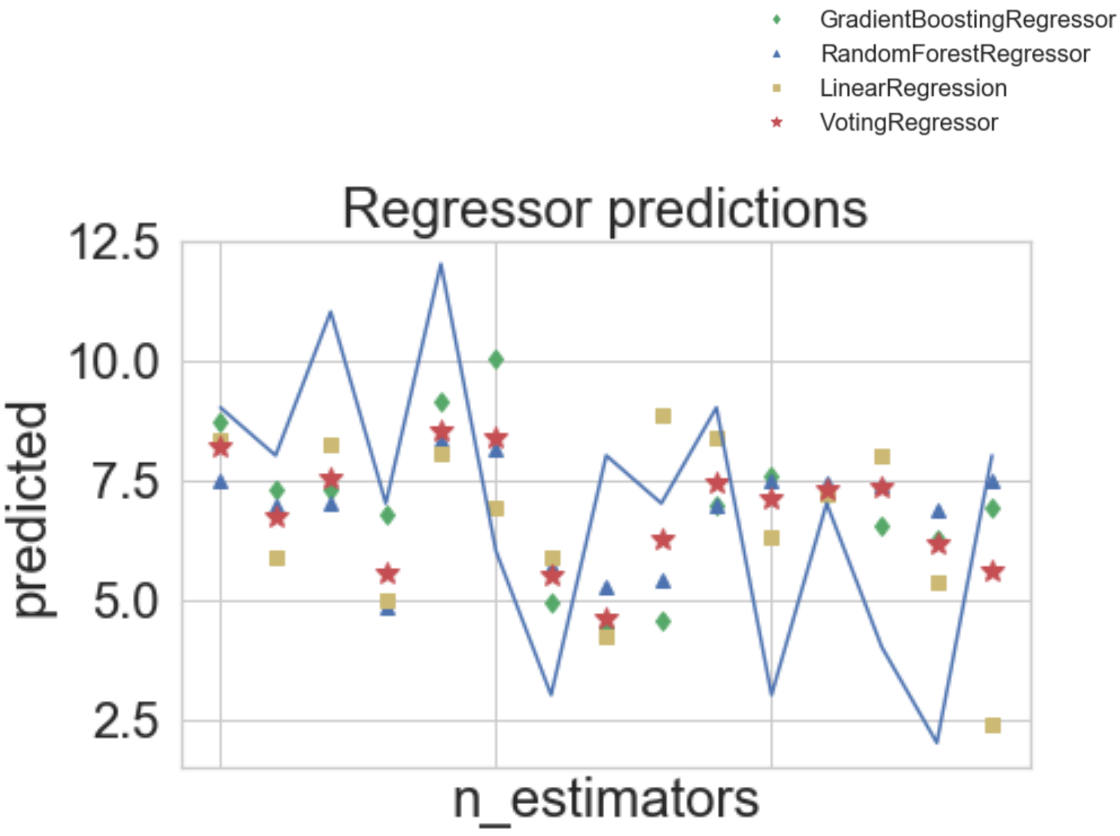


FIGURE 5.29 REGRESSION SUMMARY OF COMPARISON REGRESSIONS

A lot of variation in prediction accuracy can be found between the models. Interesting to note is that there exist some events where none of the algorithms has been able to predict well, and also on events where just one or a few models are well forecasted. Based on the graph, it is impossible to conclude the best model available, which means there are still possibilities to increase the performance of some of the models.

6. ANALYSIS OF RESULTS

6.1 DISCUSSION OF RESULTS

The analysis has generated various curious results. With the help of different evaluation metrics, a detailed study was conducted, and it was found that model 1 was the model predicting the best. As the dimensions of the model were first reduced with the help of PCA and thereafter tuned for the hyperparameters of the algorithm, it was shown that the steps taken in the research did increase the quality and accuracy of the prediction model.

It was not known if the PCA would increase accuracy, but by testing it could be found that a decrease of the errors for both the tuned and the non-tuned models took place. This being said, the original model contained data that did not support the model with valuable information, in fact, it even worsened the model. Hyperparameter tuning was shown to reduce overfitting and to increase the prediction results. No detailed testing was done on the magnitude of the impact of PCA, compared to the effect of model tuning. It is hard to say if one of these two has a more significant influence on prediction performance or not. Yet, when reviewing the prediction results of the models, it can be noted that tuning a model that has not been processed with PCA, results worse than the base-line model. This indicates that tuning is an adequate tool when the data processed describes well enough the predicted variable.

As of general performance, a predicting accuracy of 57% is a good start, but for implementation in "real-life" forecasts in commercial usage, it should be more accurate. Also, the comparison between the "base-line" random forest and gradient boost regression showed that without further development of the models, the gradient boosting model performed better prediction results than the random forest. However, we do not know how PCA and tuning would affect the model. Perhaps an increased performance could take place, but for this, a comprehensive research would have to be conducted, following similar steps as done in this research.

The result generated in this research could not be compared to other prediction result done in the field. Other forecast models do exist, but due to the usage of other variables and other data sets, it was considered to be of no benefit to predict the results to other previous researches.

6.2 LIMITATIONS

Limitations of the research could be identified along the process. As there only existed data for LST from 1978 forward, the years included in the research got limited. As discussed, over 35 years of data

is enough for conducting a study like this, but perhaps more data could have provided a possibility to identify overall patterns and trends.

Another limiting factor is the fact that both the LST and QBO data used were of global average values. More area-specific data could have contributed with more precise information, and in this way also improved the predictions. Nonetheless, a more in-depth detailed knowledge on what atmospheric areas impact on the hurricanes seasons, and in what way, would have been needed. This was unfortunately not possible to obtain within the time span of the master's thesis period.

Lack of expert knowledge on the SST areas could be considered a limitation. Smaller and more precise areas could have been created, taking into account factors such as ocean currents etc., that could affect the sea surface temperatures.

7. CONCLUSIONS AND FUTURE WORK

7.1 CONCLUSIONS

A comprehensive study to find the best random forest prediction model was done for the seasonal forecasts. According to literature, some research had been done, but a lack of well-performing models was identified. Many previous models have used similar variables in the prediction models, and for this reason, a slightly different variable approach was tried. A combination of SST, LST and QBO variables to predict the number of hurricanes was selected. Two SST datasets were compared against each other, and based on a correlation test, it was shown that the data set HADISST performed better in terms of correlation with the number of hurricanes. Various pre-processing steps were done, and four random forest models were developed. From these, it was tested that models with PCA performed in general better than the ones without PCA. The poor results of tuning with no dimension reduction, showed in this case, that the technique can be potent when used right, but if the dataset is not optimized for the prediction, the output of the tuned model can be lousy.

The tuned PCA model performed the best among the random forest models, and in comparison with other machine learning algorithms, some further analysis was done. For not enlarging the thesis too much, the general base-line random forest model was compared against a linear regression, gradient boosting and voting regression models. None of the models was tuned to keep the research as simple as possible, as this step only served as an indicative test. Results show that of the general models, the voting regression and the gradient boost regressions achieved the best results. This might require additional studies on the topic. Considering this, we cannot claim that model 1 found in this research is generally the best model to use with these variables for the prediction, as a more profound analysis of the other algorithms should be done. Nevertheless, model 1 performs better than any of the algorithms used, and for that reason, it is believed to be the best random forest model, but also a worthy model outperforming the non-tuned gradient boost and voting regression models.

7.2 RECOMMENDATIONS FOR FUTURE WORK

As the focus was mainly on exploring a prediction model on the random forest algorithm, it is of high recommendation to do a similar study exploring the opportunities of the gradient boosting model in seasonal hurricane forecast. A research on the best machine-learning algorithm for seasonal prediction of hurricanes is proposed.

Also, an exploration of other variables could bring a useful addition to the prediction. Many of the climatological factors depend on each other, which would be essential to understand. For this reason, a careful study of the different variables would be recommended and from there study the impact each variable has on the seasonal hurricanes.

Feature selection could have a positive impact on the model. By doing it, the most important features could be identified, which would help to determine the variables that are the most important ones in

seasonal hurricane prediction. It is recommended to investigate its effect it would have on the prediction model combined with PCA.

As the topic of seasonal forecasts of TCs, in general, is still new there exist a lot of opportunity for further development of the best prediction model. The model created in this analysis could be developed into other active tropical storm areas with some needed adjustments. As every area is unique, the models would have to be adapted according to the climatological factors impacting the focus basin.

BIBLIOGRAPHY

Amaratunga, D., Cabrera, J., & Lee, Y.-S. (2008). Enriched random forests. *Bioinformatics*, *24*(18), 2010–2014. https://doi.org/10.1093/bioinformatics/btn356

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, *9*(7), 1545–1588.

Bader-El-Den, M., & Gaber, M. T. (2012). Towards self-optimised random forests. In *19th International Conference, ICONIP 2012 Doha, Qatar, Proceedings, Part II, Lecture Notes in Computer Science (pp. 506–515). Berlin: Springer*.

Bakshi, C. (2020). Random Forest Regression. Random Forest Regression is a… | by Chaya Bakshi | Level Up Coding. Retrieved September 27, 2020, from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., … Takahashi, M. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, *39*(2), 179–229. https://doi.org/10.1029/1999RG000073

Bartholomew, D. J. (2010). Principal components analysis. *International Encyclopedia of Education*, 374–377. https://doi.org/10.1016/B978-0-08-044894-7.01358-0

Brownlee, J. (2020). How to Develop a Gradient Boosting Machine Ensemble in Python. Retrieved September 26, 2020, from https://machinelearningmastery.com/gradient-boosting-machine-ensemble-in-python/

Camargo, S. J., Barnston, A. G., Klotzbach, P. J., & Landsea, C. W. (2007). Seasonal tropical cyclone forecasts, *56*(October), 297–309.

Chen, R., Zhang, W., & Wang, X. (2020). Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere*, *11*(7), 1–29. https://doi.org/10.3390/atmos11070676

Chiu, M.-H., Yu, Y.-R., Liaw, H., & Chun-Hao, L. (2016). The Use of Facial Micro-Expression State and Tree-Forest Model for Predicting Conceptual-Conflict Based Conceptual Change, (January).

Climate Data Guide: SST data: COBE: Centennial in situ Observation-Based Estimates | NCAR. (n.d.). Retrieved August 25, 2020, from https://climatedataguide.ucar.edu/climate-data/sst-data-cobe-centennial-situ-observation-based-estimates

COMET. (2009). *Introduction to Tropical Meteorology, Version 1.3 The COMET ® Program*.

DeMaria, M., & Kaplan, J. (1994). A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic Basin. *Hurricane Research Division, NOAA/AOML*, 209–220. https://doi.org/10.1016/B978-0-12-382225-3.00154-7

DeMaria, M., & Kaplan, J. (1999). An updated Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic and eastern North Pacific basins. *Weather and Forecasting*, *14*(3), 326–337. https://doi.org/10.1175/1520-0434(1999)014<0326:AUSHIP>2.0.CO;2

DeMaria, M., Sampson, C. R., Knaff, J. A., & Musgrave, K. D. (2014). Is tropical cyclone intensity guidance improving? *Bulletin of the American Meteorological Society*, *95*(3), 387–398. https://doi.org/10.1175/BAMS-D-12-00240.1

Dutta, A. (2020). Random Forest Regression in Python - GeeksforGeeks. Retrieved September 27, 2020, from https://www.geeksforgeeks.org/random-forest-regression-in-python/

Ershov, V. (2018). CatBoost - open-source gradient boosting library. Retrieved September 26, 2020, from https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus

Fawagreh, K., Medhat Gaber, M., Elyan, E., & Gaber, M. M. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, *2*(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17(3)*(37). https://doi.org/https://doi.org/10.1609/aimag.v17i3.1230

Ferrara, M., Groff, F., Moon, Z., Keshavamurthy, K., Robeson, S. M., & Kieu, C. (2017). Large-scale control of the lower stratosphere on variability of tropical cyclone intensity. *Geophysical Research Letters*, *44*(9), 4313–4323. https://doi.org/10.1002/2017GL073327

Gibert, K., Sànchez-Marrè, M., & Izquierdo, J. (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. *AI Communications*, *29*, 627–663. https://doi.org/10.3233/AIC-160710

GIS Geography. (2020, March 5). Latitude, Longitude and Coordinate System Grids - GIS Geography. Retrieved September 27, 2020, from https://gisgeography.com/latitude-longitude-coordinates/

GRAY, W. M. (1984). Atlantic Seasonal Hurricane Frequanecy. Part 1: El Nino and 30 mb Quasi-Biennial Oscillation Influences. *Department of Athmospheric Science*, *112*(1), 6–8. https://doi.org/10.16309/j.cnki.issn.1007-1776.2003.03.004

Heming, J. T. (2017). Tropical cyclone tracking and verification techniques for Met Office numerical weather prediction models. *Meteorological Applications*, *24*(1), 1–8. https://doi.org/10.1002/met.1599

Ho, T. . (1996). Random Decision Forest. In *Proceeding of the 3rd International Conference on Document analysis and Recognition* (pp. 278–282).

Ho, T. K. (1998). The random subspace method for constructing decision forests. *Intelligence, IEEE Transactions on Pattern Analysis and Machine*, *20*(8), 832–844.

Hoare, J. (2020). Machine Learning: Pruning Decision Trees | Displayr. Retrieved August 23, 2020, from https://www.displayr.com/machine-learning-pruning-decision-trees/

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., Van der Linden, P. J., Dai, X., … Johnson, C.A., E. (2001). *IPCC Climate Change. The Scientific Basis. Contribution of Working Group I to the Third of the Intergovernmental Panel on Climate Change Assessment*.

Japan Agency for Marine-Earth Science and Technology (JAMSTEC). (2013). *Discovery of Weakening Trend for Equatorial Quasi-Biennial Oscillation with Global Warming -New Observational Findings Verify Changes in Global-Scale*. Retrieved from https://www.jamstec.go.jp/e/about/press_release/20130523/#z1

Kim, M., Park, M. S., Im, J., Park, S., & Lee, M. I. (2019). Machine learning approaches for detecting tropical cyclone formation using satellite data. *Remote Sensing*, *11*(10), 1–19. https://doi.org/10.3390/rs11101195

Klotzbach, P., Blake, E., Camp, J., Caron, L.-P., Chan, J. C. L., Kang, N.-Y., … Zhan, R. (2019). Seasonal Tropical Cyclone Forecasting. *Tropical Cyclone Research and Review*, *8*(3), 134–149. https://doi.org/10.1016/j.tcrr.2019.10.003

Klotzbach, P. J., Bell, M. M., & Jones, J. (2020). Extended Range Forecast of Atlantic Seasonal Hurricane activity and lanfdall strike probability for 2020, 1–38.

Latinne, P., Debeir, O., & Decaestecker, C. (2001). Limiting the number of trees in random forests. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 2096, pp. 178–187). Springer Verlag. https://doi.org/10.1007/3-540-48219-9_18

MANGALE, S. (2019). Voting Classifier. *Medium*.

Mason, L., Baxter, J., Frean, M., & Bartlett, P. (1999). Boosting Algorithms as Gradient Descent. *Advances in Neural Information Processing Systems 12*, *91*(7), 512–518. https://doi.org/10.1103/PhysRevD.91.072004

McAdie, C. J., & Lawrence, M. B. (2000). Improvements in Tropical Cyclone Track Forecasting in the Atlantic Basin, 1970–98. *Bulletin of the American Meteorological Society*, *81*(5), 989–998. https://doi.org/10.1175/1520-0477(2000)081<0989:IITCTF>2.3.CO;2

McBride, J. L., & Zehr, R. (1981). *Observational Analysis of Tropical Cyclone Formation. Part II: Comparison of Non-Developing versus Developing Systems*. *Journal of the Atmospheric Sciences* (Vol. 38). American Meteorological Society. https://doi.org/10.1175/1520-0469(1981)038<1132:OAOTCF>2.0.CO;2

Mercer, A., & Grimes, A. (2017). Atlantic Tropical Cyclone Rapid Intensification Probabilistic Forecasts from an Ensemble of Machine Learning Methods. In *Procedia Computer Science* (Vol. 114, pp. 333–340). Elsevier B.V. https://doi.org/10.1016/j.procs.2017.09.036

Met Office. (2020). Coriolis effect - Met Office. Retrieved September 24, 2020, from https://www.metoffice.gov.uk/weather/learn-about/weather/how-weather-works/coriolis-effect

Meteorology, T., & Program, T. C. (2009). CHAPTER 10, (March), 1–208.

Montgomery, M. T., & Farrell, B. F. (1993). Tropical cyclone formation. *Journal of the Atmospheric Sciences*, *50*(2), 285–310. https://doi.org/10.1175/1520-0469(1993)050<0285:TCF>2.0.CO;2

National Center for Atmospheric Research Staff. (2020). The Climate Data Guide: SST data: COBE: Centennial in situ Observation-Based Estimates.

NHC. (2017). NATIONAL HURRICANE CENTER and CENTRAL PACIFIC HURRICANE CENTER. Retrieved from https://www.nhc.noaa.gov/CLIMO/

NOAA. (2020a). Hurricane Costs. Retrieved June 14, 2020, from https://coast.noaa.gov/states/fast-facts/hurricane-costs.html

NOAA. (2020b). What are El Niño and La Niña? Retrieved September 24, 2020, from https://oceanservice.noaa.gov/facts/ninonina.html

Randel, W. J., Shine, K. P., Austin, J., Barnett, J., Claud, C., Gillett, N. P., … Yoden, S. (2009). An update of observed stratospheric temperature trends. *Journal of Geophysical Research Atmospheres*, *114*(2). https://doi.org/10.1029/2008JD010421

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., … Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, *108*(14). https://doi.org/10.1029/2002jd002670

Regnier, E. (2008). Public evacuation decisions and hurricane track uncertainty. *Management Science*, *54*(1), 16–28. https://doi.org/10.1287/mnsc.1070.0764

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rogers, R., Aberson, S., Black, M., Black, P., Cione, J., & Dodge, P. (2006). The Intensity Forecasting Experiment: A NOAA Multiyear Field Program for Improving Tropical Cyclone Intensity Forecasts. *Bulletin of the American Meteorological Socie*, *87*(11), 1523–1537.

Roy, C., & Kovordányi, R. (2012, February 1). Tropical cyclone track forecasting techniques - A review. *Atmospheric Research*. Elsevier. https://doi.org/10.1016/j.atmosres.2011.09.012

Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Shapiro, L. J., & Goldenberg, S. B. (1998). *Atlantic Sea Surface Temperatures and Tropical Cyclone Formation*. *Journal of Climate* (Vol. 11). American Meteorological Society. https://doi.org/10.1175/1520-0442(1998)011<0578:ASSTAT>2.0.CO;2

Shlens, J. (2014). A Tutorial on Principal Component Analysis. Retrieved from http://arxiv.org/abs/1404.1100

Tan, J., Liu, H., Li, M., & Wang, J. (2018). A prediction scheme of tropical cyclone frequency based on lasso and random forest. *Theoretical and Applied Climatology*, *133*(3–4), 973–983. https://doi.org/10.1007/s00704-017-2233-3

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications*.

Toppr. (2020). Karl Pearson's Correlation Coefficient: Formula, Property, Video, Example. Retrieved September 26, 2020, from https://www.toppr.com/guides/business-mathematics-and-statistics/correlation-and-regression/karl-pearsons-coefficient-correlation/

Tsymbal, A., Pechenizkiy, M., & Cunningham, P. (2006). Dynamic integration with random forests. In *In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), Machine Learning ECML 2006 : 17th European Conference on Machine Learning Berlin, Germany* (pp. 801–808).

UCAR. (2011). The Stratosphere - overview | UCAR Center for Science Education. Retrieved September 27, 2020, from https://scied.ucar.edu/shortcontent/stratosphere-overview

Vega, A. J., & Binkley, M. S. (1993). Tropical cyclone formation in the North Atlantic Basin, 1960-1989. *Climate Research*, *3*(3), 221–232. https://doi.org/10.3354/cr003221

Wallace, J. M. (1973, May 1). General circulation of the tropical lower stratosphere. *Reviews of Geophysics*. John Wiley & Sons, Ltd. https://doi.org/10.1029/RG011i002p00191

Yan, X., & Gang Su, X. (2009). Linear Regression Analysis: Theory and Computing. *World Scientific*.

Yulaeva, E., Holton, J. R., & Wallace, J. M. (1994). On the cause of the annual cycle in tropical lower-stratospheric temperatures. *Journal of the Atmospheric Sciences*, *51*(2), 169–174. https://doi.org/10.1175/1520-0469(1994)051<0169:OTCOTA>2.0.CO;2

Yule, G. U., & Filon, L. N. G. (1936). Karl Pearson. 1857–1936. *Obituary Notices of Fellows of the Royal Society*, *2*(5), 72–110.

Zhang, Y., Zhang, H., Cai, J., & Yang, B. (2014). A weighted voting classifier based on differential evolution. *Abstract and Applied Analysis*, *2014*. https://doi.org/10.1155/2014/376950

Zhou, H., Deng, Z., Xia, Y., & Fu, M. (2016). A new sampling method in particle filter based on Pearson correlation coefficient. *Neurocomputing*, *216*, 208–215. https://doi.org/10.1016/j.neucom.2016.07.036

ANNEX A

A.1 Hurricane Category Classification

FIGURE .0.1 HURRICANE CLASSIFICATION.

Saffir-Simpson Category	Maximum Sustained Wind Speed (V_{MAX}; 1-minute average)[b]			Minimum Central Pressure (p_{MIN})
	m s^{-1}	km h^{-1}	mph	hPa
1	33-42	119-153	74-95	> 980
2	43-49	154-177	96-110	979-965
3	50-58	178-209	111-130	964-945
4	59-69	210-249	131-155	944-920
5	70+	250+	156+	< 920

Source : (COMET, 2009)

A.2 SST Area maps from *Missing Values-test*

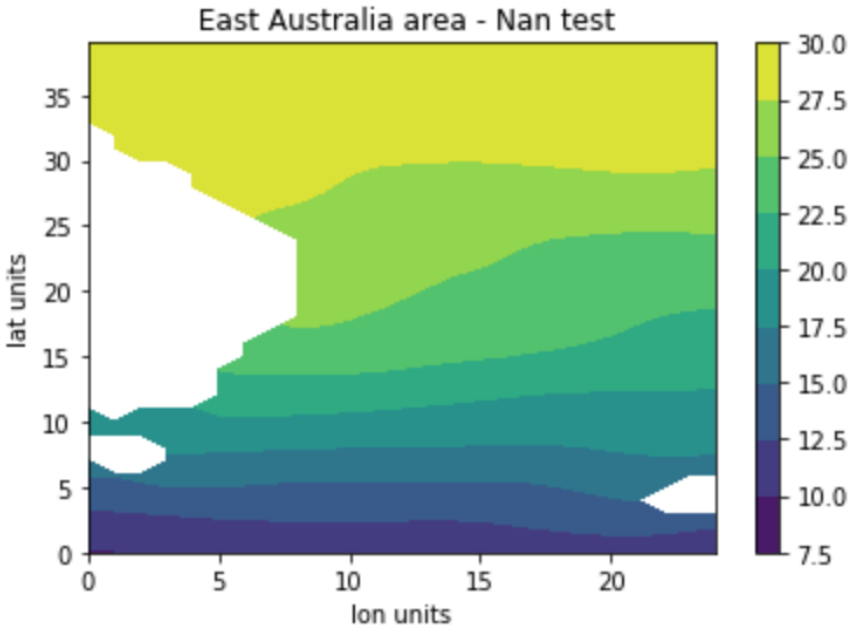


FIGURE 0.2 EAST AUSTRALIA

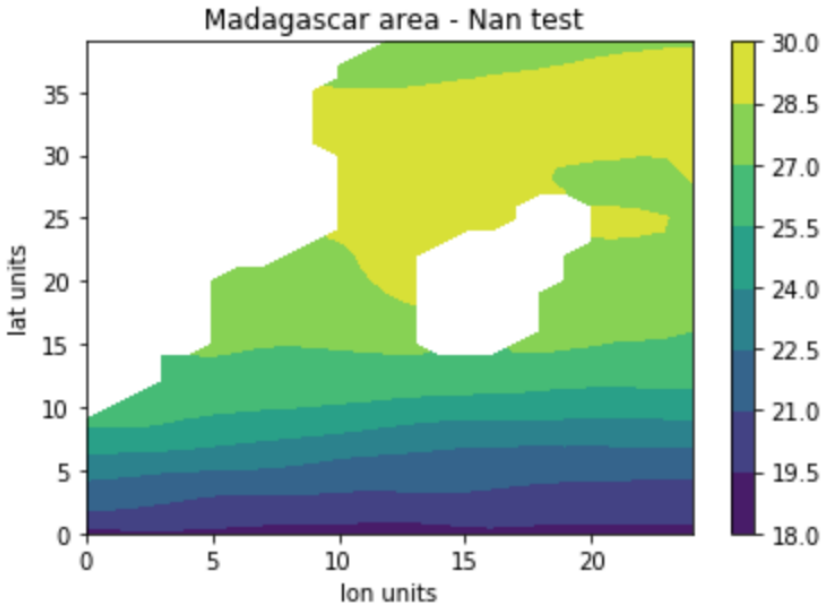


FIGURE 0.3 EAST COAST OF MOZAMBIQUE/MADAGASCAR

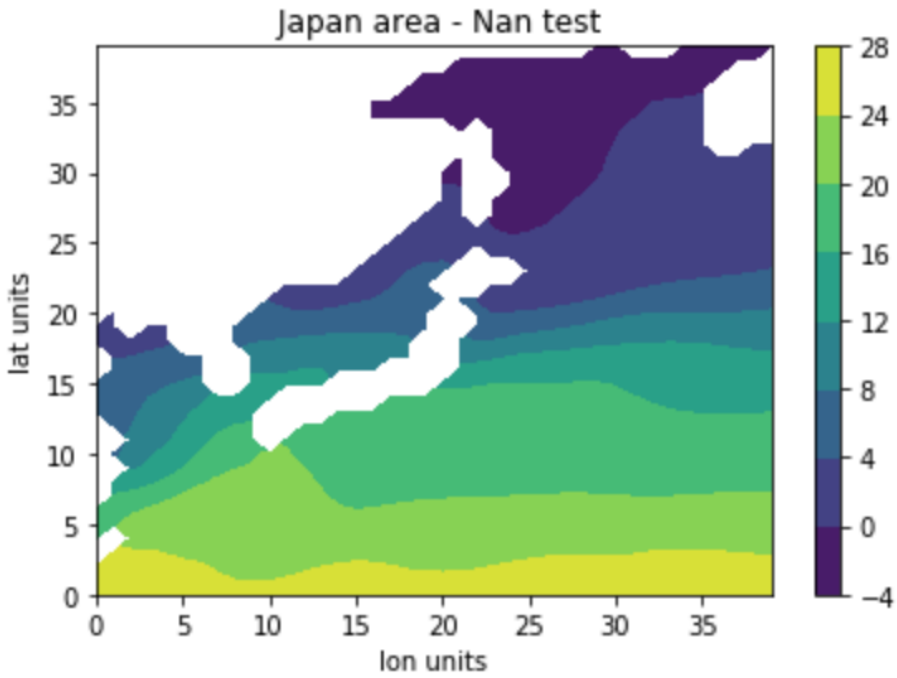


FIGURE 0.4 FIGURE 0.4 EAST COAST OF MOZAMBIQUE/MADAGASCAR

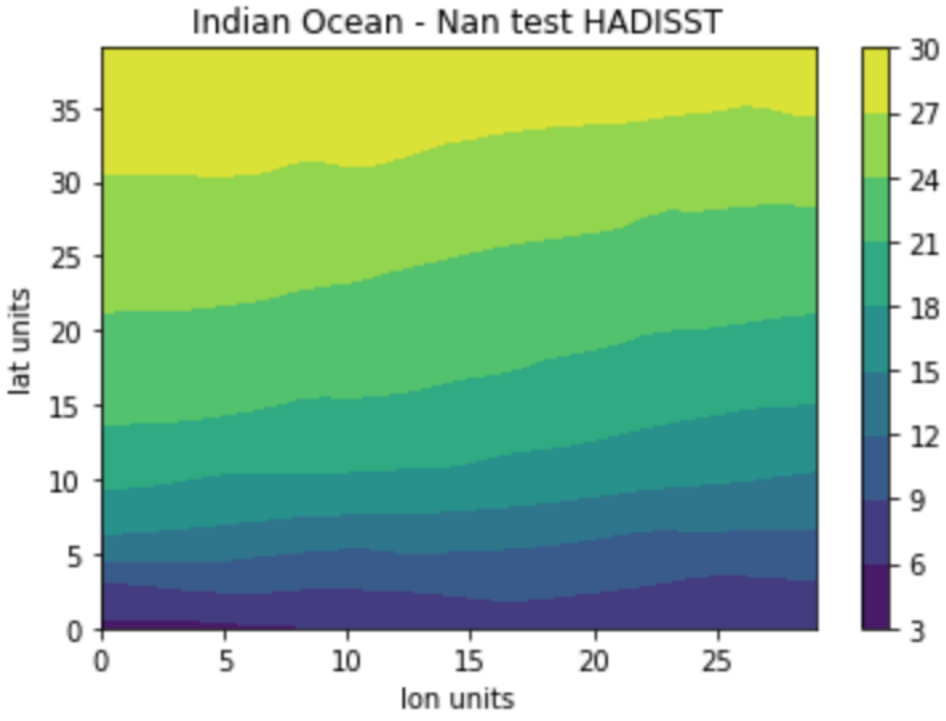


FIGURE 0.5 INDIAN OCEAN - HADISST

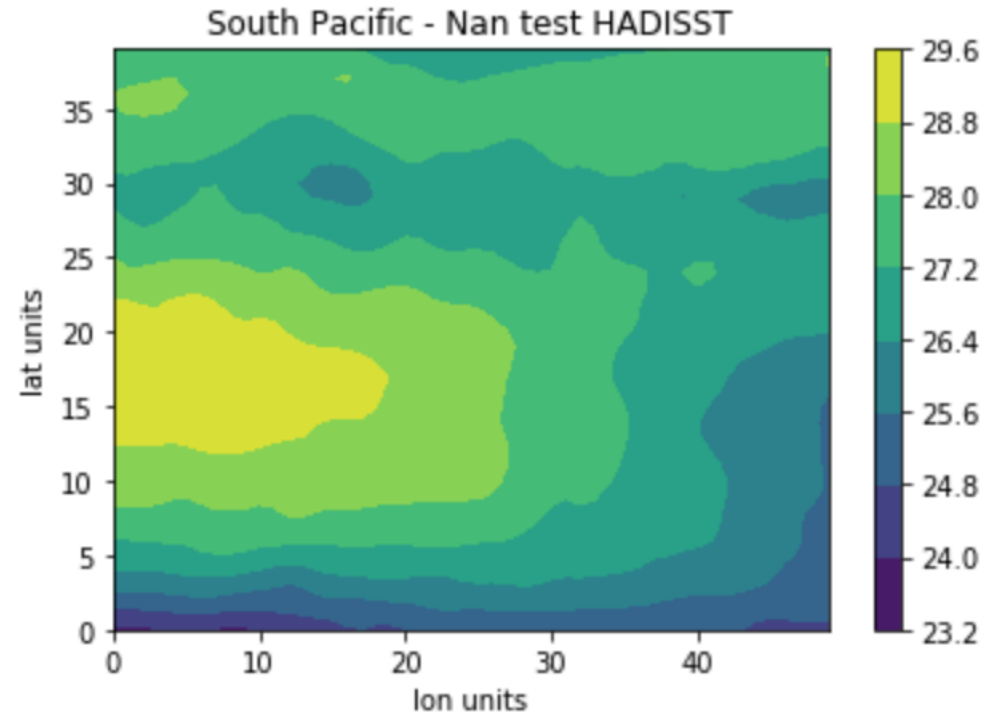


FIGURE 0.6 SOUTH PACIFIC – HADISST

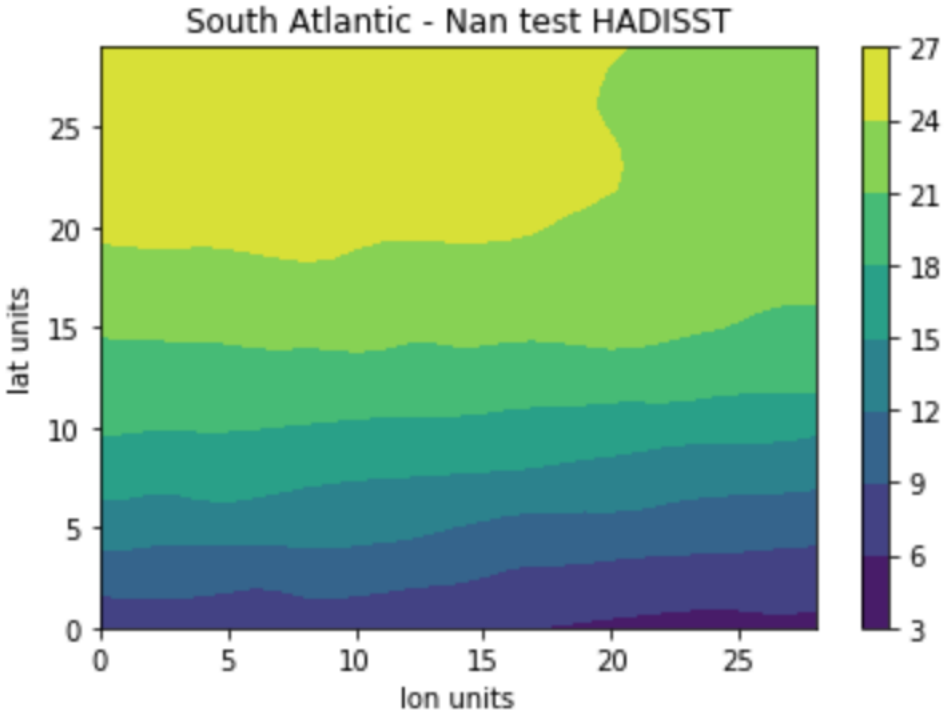


FIGURE 0.7 SOUTH ATLANTIC - HADISST

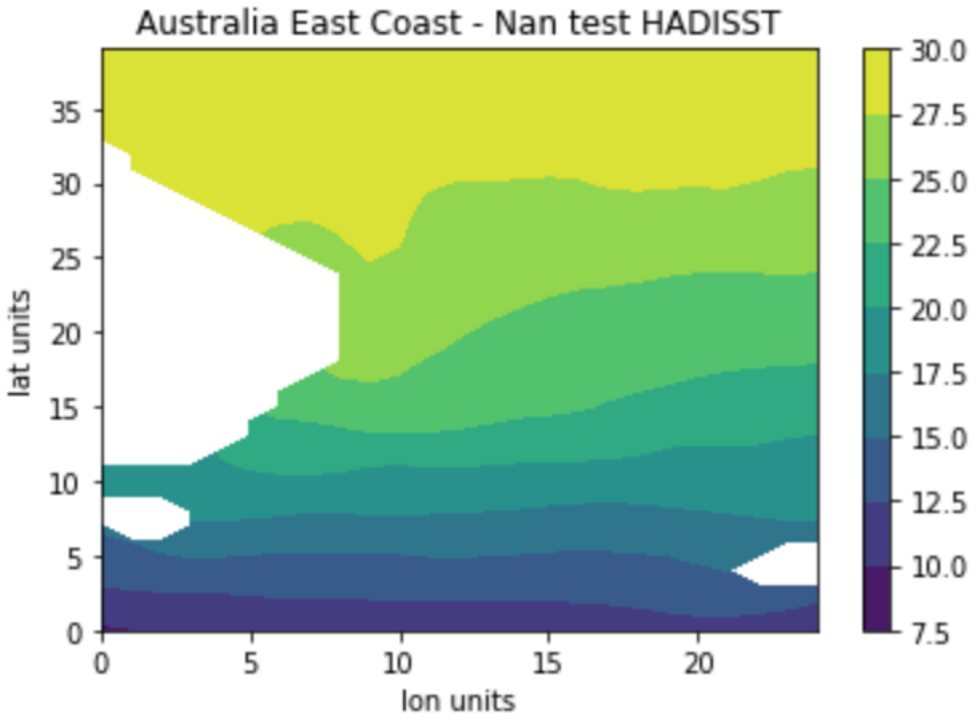


FIGURE 0.8 AUSTRALIA EAST COAST - HADISST

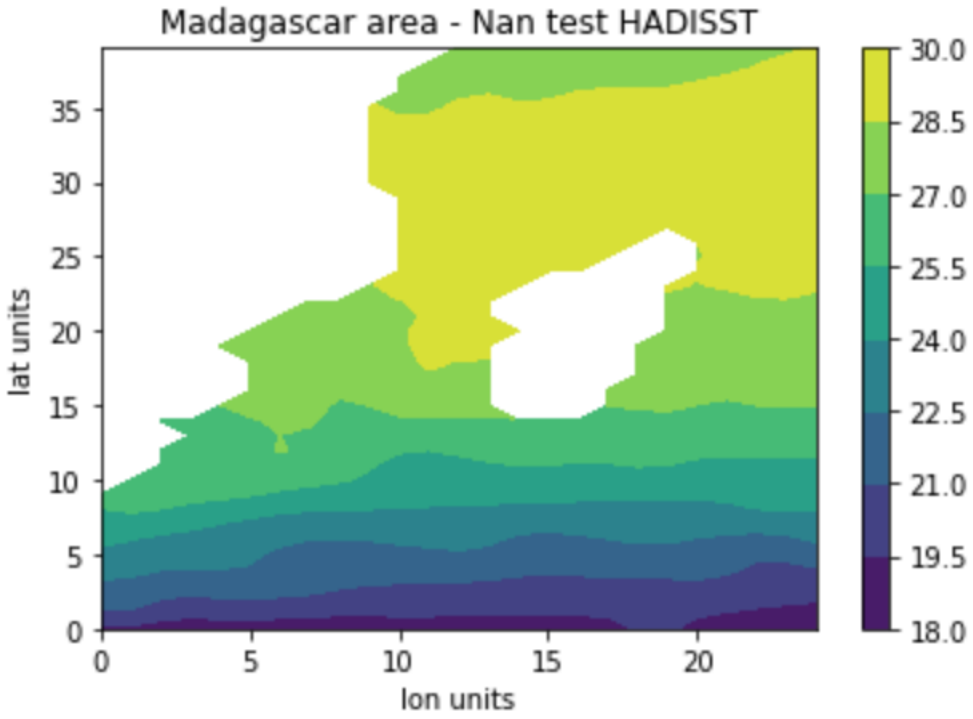


FIGURE 0.9 MOZAMBIQUE (EAST) / MADAGASCAR - HADISST

A.3 Pearson Correlation Coefficient results

COBE SST2 Results:

TABLE 0.1 COBE SST2 CORRELATION COEFFICIENTS

Caribbean:	
-0.005346938739471771	0.974944139415656
0.14199661640505676	0.401836223207704
0.22019898536867596	0.190335220292752
-0.031242708471536942	0.854355579077904
-0.09427623387014633	0.578883717742897
-0.13180280866761518	0.436799283518642
-0.12677125467169628	0.454654972129698
Pacific A:	
-0.08852330685676697	0.6023601499352593
-0.1406292259789903	0.4064305188555085
-0.15666173578447806	0.3544646282011579
-0.11856386263048359	0.4846058497629803
-0.043755124405689315	0.7970706279055108
-0.11325422029653268	0.504512702773148
-0.2247361362216814	0.1811320964741569
Pacific B:	
-0.2431285501587313	0.14706025358785063
-0.07510982956568367	0.6586200611926456
0.00901718678704496	0.9577575323648138
-0.005881236292024465	0.9724413197460902
-0.03417294451513	0.8408628471338623
0.006627491169942511	0.9689460641495761
0.026243015316410045	0.8774697681047201
ArcticA:	
0.0008655518958541469	0.995943385634001
0.25939588308745654	0.1210550278999894
0.36239156178086906	0.027509213753614016
0.18592605187639455	0.2705624466404285
0.28435014828195865	0.0880676612929987
0.10120074882121313	0.5511850855286774
0.1155816353584741	0.4957363663414306
ArcticB:	
0.2702634689929866	0.1057075440250521
0.15861976208708195	0.3484061368362962
0.16773186767310366	0.3210473658327999
0.11635962578513429	0.49282016495235065
0.23609992028966903	0.1594784825154851
0.14651104708089974	0.386881367180002

0.24626821416696132	0.1417476600772113

West AfrikaA:

-0.1627415316663698	0.3358598870736353
-0.18684064561871788	0.268170557456723
-0.23676770051857576	0.15826705908414188
-0.27202668026337884	0.10336343839464428
-0.28083179389000046	0.09224436181290868
-0.1650955794991429	0.3288208765484777
-0.34679988103162296	0.0354785936128579

West AfrikaB:

-0.07806689612273601	0.6460432044657357
-0.223549392943256	0.18350816435910094
-0.2546059406421637	0.1283255249046742
-0.3509167511194734	0.03321098671977361
-0.2807228209159556	0.09237611217408764
-0.3667093525823901	0.025584579426540233
-0.3738297431392875	0.02265459379345043

AntarcticaA:

-0.17796647805326948	0.29196804783071006
0.11476497044240344	0.4988069834313692
0.2779286686839585	0.09580389350343001
0.2555221612980012	0.1269103090615277
0.29430261692796067	0.07704041723247432
0.2957016130900356	0.07558075207150128
0.15852479781258602	0.3486985128958547

AntarcticaB:

0.09543239632765621	0.5742157783197654
0.06606672139662201	0.6976460629117315
0.10878833180368386	0.5215695845833715
0.26418691046459863	0.1140956589455952
0.3937391459152525	0.015903785672090047
0.3717846824120584	0.02346605030516122
0.2500423209953372	0.1355497858833186

Atlanten:

0.049949288774937206	0.7690752421989069
0.012809804345004956	0.9400177062571673
-0.0755419098553711	0.6567764811863346
-0.008019635692686357	0.9624272106472073
0.024841984644668788	0.8839657114018604
0.007760905266646509	0.96363857204888
0.049599000812530644	0.7706511193225437

North AtlanticA:

-0.2042926305241648	0.2251857585607359
-0.30057668975737517	0.07066230772641739
-0.3010090042135738	0.0702385638263961

-0.2870810729303365	0.08492747161327113
-0.19068013091845792	0.25828129158006813
-0.21883101801239874	0.19317347894165054
-0.14215803137287691	0.4012958590706758
North Atlantic B:	
-0.0721011572239812	0.6715118821090331
0.01168775650095144	0.9452634592342332
-0.12238561221415323	0.4705336752599071
-0.16766300706800108	0.32124893957493983
-0.10435245827200476	0.5387880236853009
-0.0317700177575105	0.8519243777130044
0.09889031229539565	0.5603574016928089
Mex:	
0.05515025687824293	0.7457875717991778
0.05336194023407942	0.7537710043195973
-0.0944906425242873	0.5780167688039786
-0.11882780180561728	0.4836270346638432
-0.1535784031641655	0.3641330690539067
-0.17333053977926358	0.30492323340754374
-0.1423609515828586	0.4006171413514414

Results HADISST:

TABLE 0.2 HADISST CORRELATION COEFFICIENTS

Caribbean	
0.4659129397316978	0.0036581575708933374
0.4411228131572418	0.0062783102559886336
0.2575903881864634	0.12375843005295807
0.053920440905262865	0.7512749885210119
-0.0014267921122741445	0.9933130474877361
-0.13847499662145718	0.41372900607209306
-0.011015774816135297	0.9484061821982568
Pacific A:	
-0.21464012142820402	0.20205344294395972
-0.097161634700309	0.5672660770889242
-0.10812078943009376	0.5241433193946443
-0.1821205474579214	0.28066456649483196
-0.07571144190589144	0.6560536744698784
-0.015224945350444635	0.9287352594177876
0.12215035715759588	0.47139363242998955
Pacific A_1:	
0.2973028141901668	0.07393669994616359
0.14660332085465289	0.3865791228895379
0.13966530563063412	0.4096871327851945
0.15194337952416961	0.3693233670049442
0.1037434549875266	0.5411730406982528

0.13859534170023555	0.41331933360559103
0.22017689306580812	0.190380822781064
Pacific B:	
0.4595232371434644	0.004220590878703153
0.32412556615902965	0.050334948807550524
0.39900940386064004	0.014431485539342946
0.3487234575795102	0.03440383610385555
0.32461178512689826	0.04997098107549703
0.31199918879641264	0.06012030598181714
0.028514607889251767	0.8669543561199637
Pacific B_1:	
0.4595232371434644	0.004220590878703153
0.32412556615902965	0.050334948807550524
0.39900940386064004	0.014431485539342946
0.3487234575795102	0.03440383610385555
0.32461178512689826	0.04997098107549703
0.31199918879641264	0.06012030598181714
0.028514607889251767	0.8669543561199637
ArcticA:	
0.06554515009444067	0.699921618124107
0.06054322484485247	0.7218728298724836
0.016936903043067927	0.9207459288652459
-0.09026735958402618	0.5951997967391751
-0.19338081679923289	0.2514719375198012
-0.15353812734916836	0.36426039611130134
-0.26396591843373446	0.11440988983996976
ArcticA_1:	
0.17802929347050447	0.2917949759541868
0.21582466712948617	0.19951521069117983
-0.05559696394764818	0.7437974065241619
0.23084686210759736	0.16924273881650867
0.13523229741621612	0.42485387597685875
0.2337581863026959	0.16377962788935502
0.14547791960595302	0.3902747976358229
ArcticB:	
-0.06272835284358641	0.7122551120159085
0.16919402197934935	0.3167859041922313
-0.1267488537298434	0.45473533617224904
0.20063187280534817	0.23378619801063713
0.15325592109525774	0.36515330307125
0.15295860022364816	0.366095445806919
0.06901644335834221	0.6848261660464692
West AfrikaA:	
0.16156215777118735	0.3394210264363824
0.14610061176691452	0.3882274245405052
-0.06686882663543703	0.6941516455999889
0.1589189521424048	0.3474859649391992
0.35347100689437233	0.031864698407732005
0.3913665851577361	0.016606681822147962

0.23176627475086004	0.16750352563330306
West AfrikaB:	
0.19850489073875296	0.23888419515727946
0.1874436754941941	0.26660110862095376
0.2765189842291284	0.0975697897804214
0.2955447435186984	0.07574334186231171
0.2769611064412055	0.0970132915497336
0.12687698127360456	0.454275778134292
0.06706050379640732	0.6933175095968713
AntarcticaA:	
0.12610819206465357	0.45703696564121304
0.10590374637483205	0.5327356387558946
0.046767653482900566	0.7834215417099668
0.10308744737713649	0.5437477680345688
0.19873100410196096	0.23833870735240306
0.08653672718045521	0.6105610955876918
0.22953834919393432	0.1717402753723522
AntarcticaB:	
0.06691478837940128	0.693951598060169
0.058672902661232665	0.7301385454214938
0.26612523393612897	0.1113673161607439
0.18530560572236282	0.272193004385467
0.07345235254399532	0.6657104515826895
0.11147661874957873	0.5112681420391536
-0.13080860705287012	0.4402964894407384
AntarcticaB_1:	
0.06691478837940128	0.693951598060169
0.058672902661232665	0.7301385454214938
0.26612523393612897	0.1113673161607439
0.18530560572236282	0.272193004385467
0.07345235254399532	0.6657104515826895
0.11147661874957873	0.5112681420391536
-0.13080860705287012	0.4402964894407384
Atlanten:	
0.4362081506308303	0.0069559144854281964
0.5259180148194696	0.000828666230038813
0.4646766669351435	0.003761587413598012
0.45631001811059035	0.004530724525519118
0.3540257222770984	0.03157831573730598
0.41093525079534815	0.01151903225317533
0.3458903758293768	0.03599621276900224
North AtlanticA:	
0.20197757481919099	0.23059915728007552
-0.117447333707466	0.48875781555914666
-0.22788704564607 58	0.1749295965048606
-0.22525266952611095	0.1801047583476557
-0.09104993093152138	0.5919990111706845
-0.006013425371880571	0.9718221431893337
0.08418239026531465	0.6203410093026153

North AtlanticA_1:	
0.20197757481919099	0.23059915728007552
-0.117447333707466	0.48875781555914666
-0.2278870456460758	0.1749295965048606
-0.22525266952611095	0.1801047583476557
-0.09104993093152138	0.5919990111706845
-0.006013425371880571	0.9718221431893337
0.08418239026531465	0.6203410093026153
North Atlantic B:	
-0.1679440192527154	0.32042683727371696
-0.057506313493835365	0.735309470307819
0.04354994827373041	0.7980024464102448
0.12712482445238246	0.45338754428441913
0.10915678153996994	0.5201516812968804
0.22526102681604576	0.18008817056000007
0.2923328865257282	0.07913269573869416
Mex:	
-0.2246185439159463	0.1813665571119686
-0.17106402489010464	0.3113877420263247
-0.1602927909024915	0.3432796741032282
-0.19740825424470973	0.24154172478409933
-0.10837285726458902	0.5231707312950495
-0.07785573359892341	0.6469381620947653
-0.0863785551278606	0.6112160886413271

A.4 Python Scripts

In [668]:
```
x = df_use_QBO['Jan']
x1= df_use_QBO['Feb']
x2= df_use_QBO['Mar']
x3= df_use_QBO['Apr']
x4= df_use_QBO['May']
x5= df_use_QBO['Jun']
x6= df_use_QBO['Jul']
x7= df_use_QBO['Aug']
x8= df_use_QBO['Sep']
x9= df_use_QBO['Oct']
x10= df_use_QBO['Nov']
x11= df_use_QBO['Dec']


data = (x, x1, x2, x3, x4, x5, x6, x7, x8,x9, x10, x11 )

fig, ax1 = plt.subplots(figsize=(10, 6))
fig.canvas.set_window_title('A Boxplot Example')
fig.subplots_adjust(left=0.075, right=0.95, top=0.9, bottom=0.25)
ax = plt.boxplot([x, x1, x2, x3, x4, x5, x6, x7, x8,x9, x10, x11 ])
plt.legend( ['QBO']);
ax1.yaxis.grid(True, linestyle='-', which='major', color='lightgrey',
               alpha=0.5)
ax1.set_axisbelow(True)
ax1.set_title('Boxplot QBO')
ax1.set_xlabel('Months')
ax1.set_ylabel('QBO in m/s')

mins = [d.min() for d in data]
median = [d.median() for d in data]

#plt.plot([1,2,3,4,5,6,7,8,9,10,11,12], mins, c="r", lw=2)
plt.plot([1,2,3,4,5,6,7,8,9,10,11,12], median, c="g", lw=2)
plt.legend( ['Median line'])
```

FIGURE 0.10 PYTHON SCRIPT (QBO-BOXPLOT)

In [1667]:

```
## MODEL WITH PCA & TUNING
rangen=np.arange(100,10000,100)
errors_train=np.zeros(rangen.shape[0])
errors_test=np.zeros(rangen.shape[0])

for j in range(rangen.shape[0]):

    rf_1 = RandomForestRegressor(n_estimators=200,criterion="mae",max_depth=7,
                                 min_samples_split=3,min_samples_leaf=6,
                                 max_features="sqrt",
                                 bootstrap=False)
    rf_1.fit(X_train_scaled_pca,Y_train)
    #display(rf.score(X_train, Y_train))
    #display(rf.score(X_train_scaled_pca, Y_train))



    Y_predict=rf_1.predict(X_test_scaled_pca)
    Y_predict1=rf_1.predict(X_train_scaled_pca)
    errors_test[j]=np.sqrt(np.mean((Y_predict-Y_test)**2))
    errors_train[j]=np.sqrt(np.mean((Y_predict1-Y_train)**2))



plt.plot(rangen,errors_train, linewidth=1)
plt.plot(rangen,errors_test, linewidth=1)
plt.xlabel('n_predictions', fontsize=16)
plt.ylabel('RMSE', fontsize=16)
plt.title('Regression errors for test and train sets', fontsize=18)
plt.show()
print("Min error = ",min(errors_test))

from sklearn.metrics import max_error
```

FIGURE 0.11 PYTHON SCRIPT (TUNED RF WITH PCA)

In [1523]:

```
from sklearn.model_selection import GridSearchCV
n_estimators = [363,681,200, 436, 481]
max_features = ['auto','sqrt']
max_depth = [3,4,7,2]
min_samples_split = [9,3,6,2]
min_samples_leaf = [9,6, 5,10]
bootstrap = [False, True]
param_grid = {'n_estimators': n_estimators,
              'max_features': max_features,
              'max_depth': max_depth,
              'min_samples_split': min_samples_split,
              'min_samples_leaf': min_samples_leaf,
              'bootstrap': bootstrap}
gs = GridSearchCV(rf, param_grid, cv = 3, verbose = 1, n_jobs=-1)
gs.fit(X_train_scaled_pca, Y_train)
rfc_3 = gs.best_estimator_
gs.best_params_
# {'bootstrap': False,
# 'max_depth': 7,
# 'max_features': 'sqrt',
# 'min_samples_leaf': 3,
# 'min_samples_split': 2,
# 'n_estimators': 500}
```

```
Fitting 3 folds for each of 1280 candidates, totalling 3840 fits
```

FIGURE 0.12 PYTHON SCRIPT (HYPERPARAMETER TUNING MODEL 1)

A.5 Hurricane intensity error analysis results

Here, just a few samples of what was done with the intensity error analysis. Figures represent the error analysis with some variables, for both individual storms but also overall data generated for over the years.

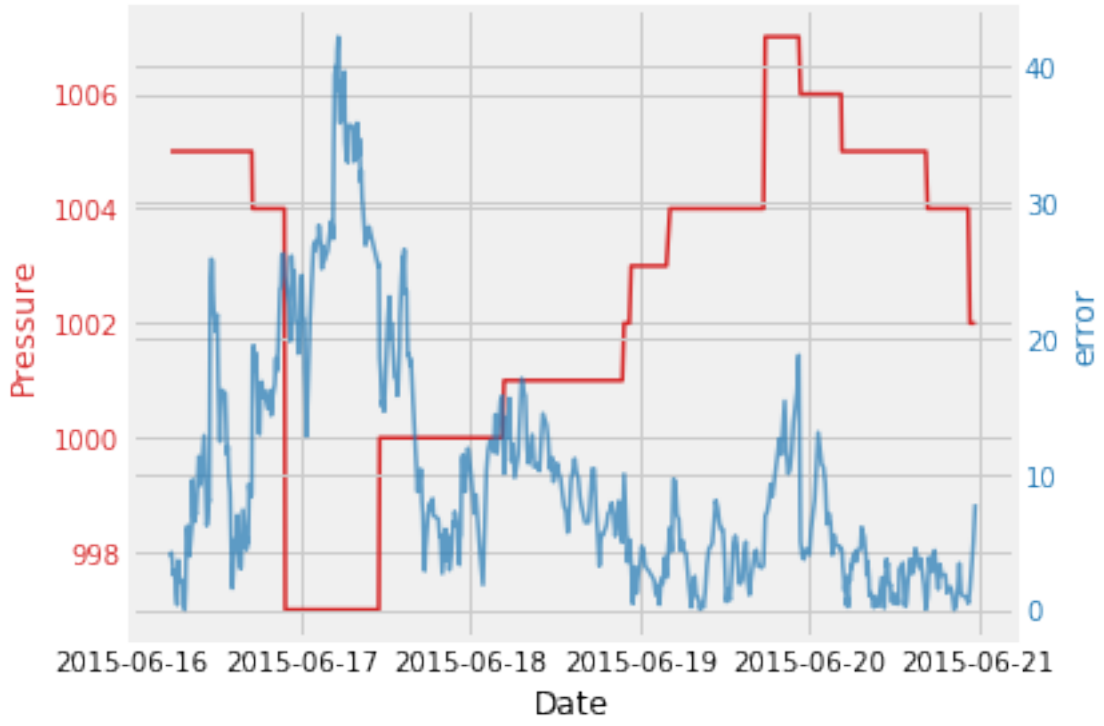


FIGURE 0.13 ERROR ANALYSIS WITH PRESSURE-HURRICANE BILL

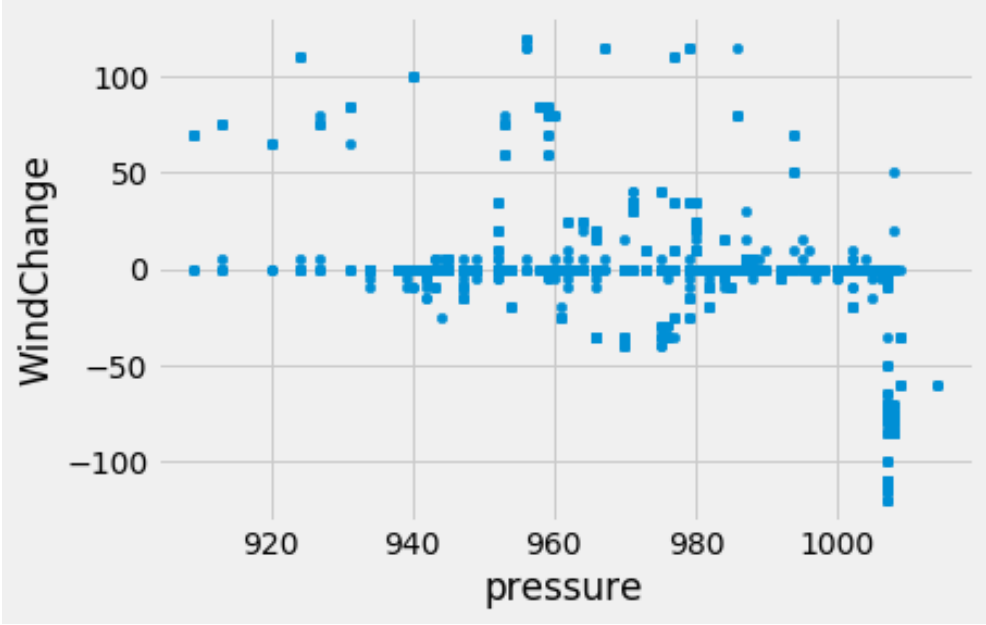


FIGURE 0.14 VARIABLE ANALYSIS ON PRESSURE AND WINDCHANGE – ALL STORMS

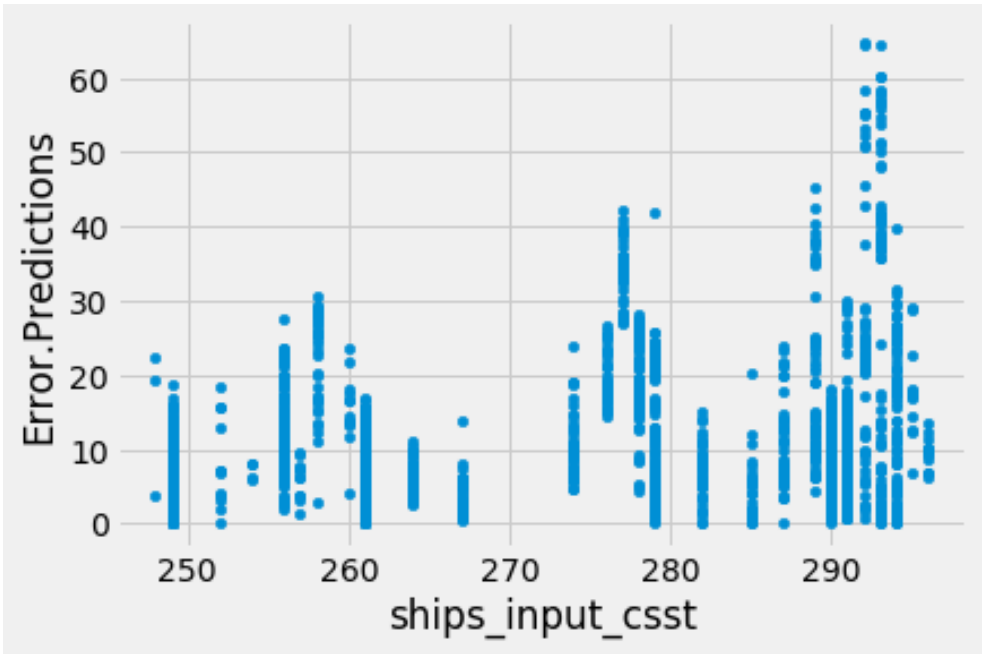


FIGURE 0.15 ERROR ANALYSIS ON SST VARIABLE – ALL STORMS

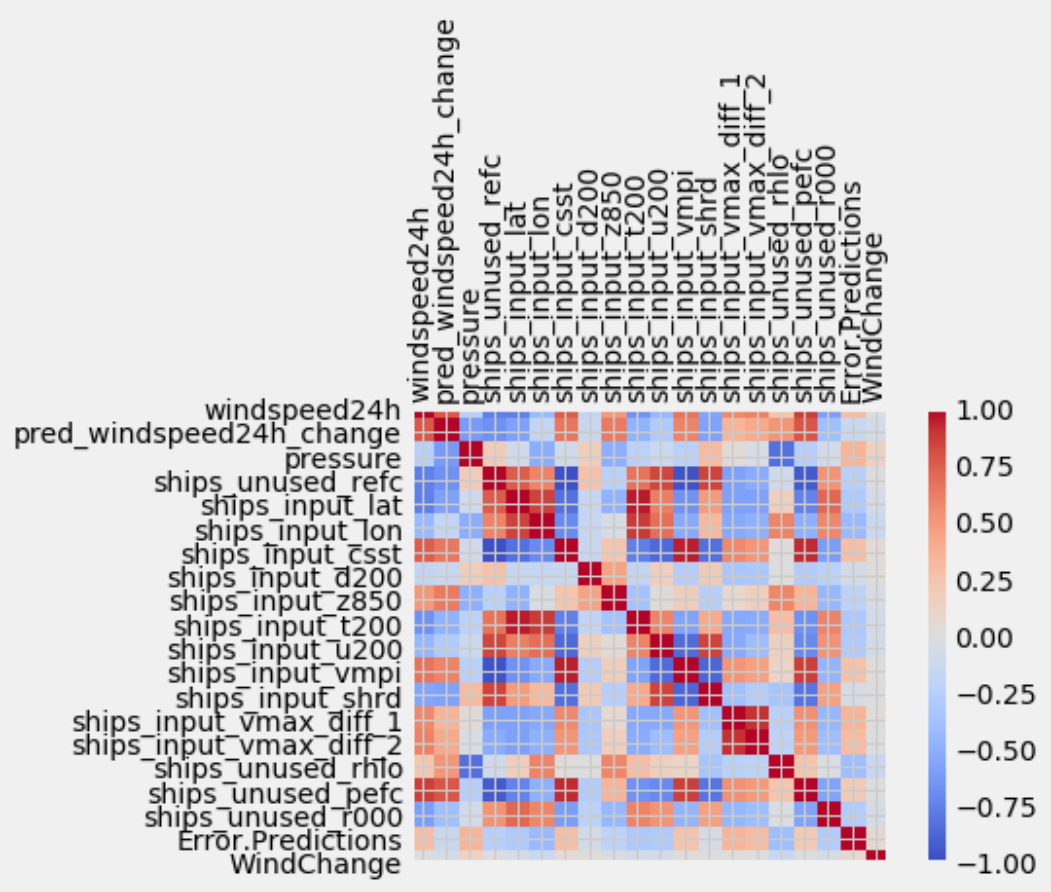


FIGURE 0.16 HEAT MATRIX ON ALL VARIABLES – ALL STORMS

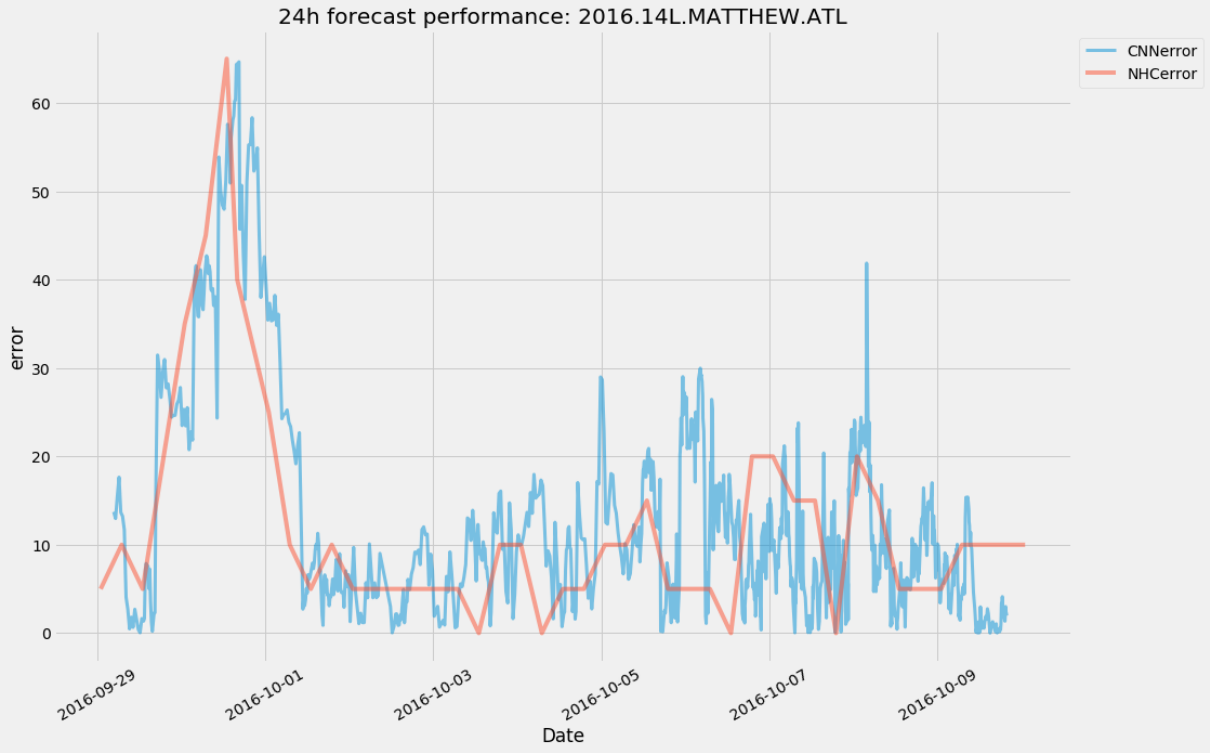


FIGURE 0.17 ERROR COMPARISON BETWEEN NHC'S AND HU'S PREDICTIONS