PREDICTION OF TROPICAL CYCLONE ACTIVITY OVER THE BAY OF BENGAL USING MACHINE LEARNING MODELS

A THESIS

Submitted by

RISHBHA JAIN

for the award of the degree

of

MASTERS OF TECHNOLOGY

DEPARTMENT OF MECHANICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

CHENNAI-600036

JUNE 2021

THESIS CERTIFICATE

This is to certify that the thesis entitled **"Prediction of tropical cyclone activity over the Bay of Bengal using machine learning models"** submitted by **Rishbha Jain** to the Indian Institute of Technology, Madras for the award of the degree of **Masters in Technology** is a bona fide record of research work carried out by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. C. Balaji
TT Narendran Chair Professor
Department of Mechanical Engineering
Indian Institute of Technology Madras
Chennai – 600 036

Dr. S. Balaji
Associate Professor
Department of Mechanical Engineering
Indian Institute of Technology Madras
Chennai – 600 036.

Place: Chennai

Date: June 2021

ACKNOWLEDGEMENTS

It gives me a sense of fulfilment to write few words to acknowledge Professor C Balaji and Professor S Balaji, who supported me throughout the project. I would also like to acknowledge my mentors Kiran and Gaurav, who were there to help me understand the project details and answered my every woe. I would like to thank Vivek, whose expertise in machine learning got me through some tough times. I am grateful for my family and friends who responded swiftly to any cry for help and sent me valuable resources to do the project well.

I'm grateful to the National Centres for Environmental Protection for providing the datasets used in this work and encouraging environmental studies. I'd also like to acknowledge IIT Madras for forever ensuring and encouraging the pursuit of knowledge, come what may.

ABSTRACT

Keywords: Tropical Cyclones, logistic regression, neural network, long short-term memory, time series, Bay of Bengal

Classical numerical methods of climate studies suffer from computationally expensive and time-consuming simulations. The most popular models are downscaling models, relying on sizeable regional climate data sets or low-resolution models. This study proposes reducing the downscaling time by exploiting machine-learning techniques. Based on either logistic regression or long short-term memory-based time-series prediction, surrogate models can supplement the capabilities of the classical weather models. The ultimate aim of this study is to predict the occurrence of tropical cyclones over the Bay of Bengal in a given month. The test results establish the superiority of the Long short-term memory model over logistic regression for the time series. However, the logistic regression model is easier to implement and provides acceptable results.

Furthermore, this study takes into account many environmental variables that can affect the climate. One such variable is carbon emission. Therefore, future cyclone occurrences are addressed as a final contribution, from 2021 to 2100, based on many possible carbon emission scenarios.

TABLE OF CONTENTS

ABSTRACT ii

TABLE OF CONTENTS iii

LIST OF FIGURES iv

LIST OF TABLES v

LIST OF ABBREVIATIONS vi

CHAPTER 1 ***7***

INTRODUCTION **7**

1.1. Literature Review 7

1.2. Objectives 9

CHAPTER 2 ***10***

DATA AND METHODOLOGY **10**

2.1. Identifying Predictors 10

2.2. Data Preparation 14

2.3. Modified Logistic regression Model 15

2.4. Long Short-Term Memory Algorithm 18

CHAPTER 3 ***22***

RESULTS AND DISCUSSION **22**

3.1. Logistic Regression 23

3.2. Long Short-Term Memory 24

3.3. Prediction 26

3.5. Theoretical Reasoning 28

3.4. Model Validation 29

CHAPTER 4 ***31***

SUMMARY AND CONCLUSIONS **31**

RECOMMENDATIONS FOR THE FUTURE WORK **32**

CHAPTER 5 ***33***

REFERENCES **33**

DATA SOURCES **36**

LIST OF FIGURES

Figure 2.1. Frequency of different tropical cyclone grades in the dataset from 1982-2020 11
Figure 2.2. Monthly frequency of tropical cyclone occurrence from 1982-2020 15
Figure 2.3. Serial Correlation of Mean values of Geopotential and Mean Sea Level Pressure with Tropical Cyclone Occurrence 16
Figure 2.4. Model of an artificial neural network 18
Figure 2.5. Model of a recurrent neural network. 19
Figure 2.6. Model of the LSTM network 19
Figure 3.1. ROC curve of modified logistic regression 24
Figure 3.2. LSTM model loss mapped against the number of epochs run 24
Figure 3.3. ROC curve of modified LSTM 25
Figure 3.4. Carbon dioxide concentrations for different RCP scenarios (Van Vuuren, 2011) 26
Figure 3.5. Moving average of cyclone-prone months per year for different RCP scenarios . 28
Figure 3.6. Average number of TC ridden months from 2090-2100 in different RCP scenarios, compared to the present 29

LIST OF TABLES

Table 2.1. Tropical cyclone classification 10

Table 2.2. Correlation analysis of predictors 12

Table 2.3. Area-wise Mean parameter values for tropical cyclone occurrences in the pre-monsoon period.. 13

Table 2.4. Area-wise mean parameter values for tropical cyclone occurrences in the post-monsoon period.. 14

Table 2.5. Grid validation study for modified logistic regression parameters......................... 17

Table 2.6. Grid validation study for modified LSTM parameters 21

Table 3.1. Confusion matrix of modified logistic regression 23

Table 3.2. Confusion matrix of modified lstm.. 25

Table 3.3. Comparison of performance metrics of algorithms 26

Table 3.4. Simulation data reduction 27

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AS	Arabian Sea
AUC	Area Under Curve
BOB	Bay of Bengal
CMIP5	Coupled Model Intercomparison Project Phase 5
CS	Cyclonic Storm
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GCM	Global Circulation Model
GP	Geopotential
IMD	Indian Meteorological Department
LSTM	Long Short-Term Memory
MSL	Mean Sea Level Pressure
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
RCP	Representation Concentration Pathways
RH	Relative Humidity
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic Curve
SST	Sea Surface Temperature
TC	Tropical Cyclones
TCF	Tropical Cyclone Frequency
TCGF	Tropical Cyclone Genesis Frequency
TN	True Negative
TP	True Positive
TPR	True Positive Rate
WRF	Weather Research and Forecasting

CHAPTER 1
INTRODUCTION

1.1. Literature Review

A tropical cyclone causes extensive casualties and financial losses that directly and indirectly affect the agricultural yield and livelihood of many farmers. Therefore, tropical cyclones are the most dangerous natural hazards around the Bay of Bengal. The cyclone season has a bimodal structure with two peaks; one in April–May and the second in October-December. On average, five storms occur per year. Based on the past 300 years, around 6–7% of global tropical cyclones originate over the North Indian Ocean. However, 75% of storms with a death toll of more than 6,000 occur in this region (Neumann, 1993). This implies that high-intensity storms occur in this region that countries cannot handle. For example, in 2008, tropical cyclone Nargis caused ~1,38,500 human causalities.

In order to combat this, scientists predict when and where a tropical cyclone will occur, along with understanding the path it follows and how intense it will be. Tropical cyclone genesis and intensity depend on many thermodynamic and dynamic factors (DeMaria, et al., 2001). The contributing factors and how much they contribute are under a microscope and seem to differ between different ocean bodies and regions within the same ocean body (Murakami et al., 2016).

So far, most of the work is on high-resolution data, obtained by downscaling low-resolution global circulation model (GCM) parameters and learning on the high-resolution output data. It may be a relatively accurate process; the maximum time error is 13 hours (Raju, et al., 2012); however, it is also highly cumbersome and requires months of waiting for the model to complete downscaling. Data-driven computer modelling systems can reduce the computational requirement of numerical weather prediction systems. There are studies involving various machine learning models, some of them being Poisson regression (Kim H. S., 2010), projection pursuit regression (Chan J. C. L., 2001) and Bayesian regression and classification models (Zhao, 2007). One study conducted by (Nath, et al., 2015) predicted the tropical cyclone frequency using multiple linear regression and neural networks. It produced desirable results, with the neural network faring better than the former traditional machine learning model. Neural networks (ANN) are highly skilled for solving non-linear phenomena due to their adaptive nature and learning capabilities. Advanced neural net models can have a comparable prediction outcome with the highly complex WRF simulations (Hewage, et al., 2021).

Further, the application of neural networks in climate change forecasting has been wildly successful in studies across the globe (Mitra A. K., 2010) (Bourras D., 2003). For example, the temperature in Nevada was predicted with 97%+ accuracy using a neural network with stacked auto-encoders with higher accuracy than traditional neural networks (Hossain M, 2015). This further encourages the use of complex learning models.

Based on the review of the above literature, it is clear that the prediction of tropical cyclones is an essential and challenging task. However, most studies are concentrated over the Atlantic or Pacific oceans, while the Indian Ocean is just scraping at the tip of the iceberg. These studies investigate different cyclonic parameters using the WRF or any such downscaling models. However, studies that attempt to make researching tropical cyclones easier and more accessible to everyone are scarce. Though there is some work on neural networks over the north Indian Ocean, the application of more robust deep learning models such as recurrent neural networks are lacking. In consideration of the above, the objectives of the current study are listed in section 1.2.

1.2. Objectives

The objectives of the study:

- Predicting the months in a year when a tropical cyclone will occur, isolating the essential data to be entered into mesoscale weather forecasting models. A mesoscale model like the Weather Research and Forecasting (WRF) will downscale relevant parameters for cyclone intensity and track prediction on finding the time windows for cyclone occurrence. This project will, however, be limited to finding the time windows.
- Predicting the tropical cyclone occurrence for different Representative Concentration Pathways (RCP) scenarios.

The structure of this paper is as follows: data and methodology are in chapter 2 that encompasses the method of choosing model predictors, and explains the models themselves. Then, chapter 3 has the results of the models, along with the prediction of the future scenarios and the model validation. Finally, chapter 4 ends with a conclusion and a summary of the future work.

CHAPTER 2
DATA AND METHODOLOGY

2.1. Identifying Predictors

The month-wise tropical cyclone occurrence series over the Bay of Bengal region from 1982 to 2020 is obtained from the cyclone archives of the Indian Meteorological Department (IMD), which is a reputable Regional Specialized Meteorological Center by the World Meteorological Organization.

The number of closed isobars within a system and maximum sustained wind speed are vital for classifying low-pressure systems over the Indian Ocean. If the cyclone is over the land, the pressure criteria is used, and if it is over the sea, the wind criteria is used. Tropical cyclones are classified into low-pressure areas, depressions, deep depressions, cyclonic storms, severe cyclonic storms, very severe cyclonic storms, and super cyclones based on the maximum sustained wind speeds attained by the disturbance (Table 2.1).

Table 2.1. Tropical cyclone classification

Type	Maximum sustained wind speed (knots)	Grade assigned
Low pressure area	<17	1
Depression	17 - 27	2
Deep depression	28 - 33	3
Cyclonic storm	34 - 47	4
Severe cyclonic storm	48 - 63	5
Very severe cyclonic storm	64-119	6
Super Cyclone	≥120	7

A preliminary analysis of the data tells us that there are no immediate visual clusters of tropical cyclone occurrence. The cyclonic grades are also well distributed spatially. On looking at the frequency of the different types of disturbances (Fig 2.1), it is observed that depressions are most common, while severe cyclones are very rare.

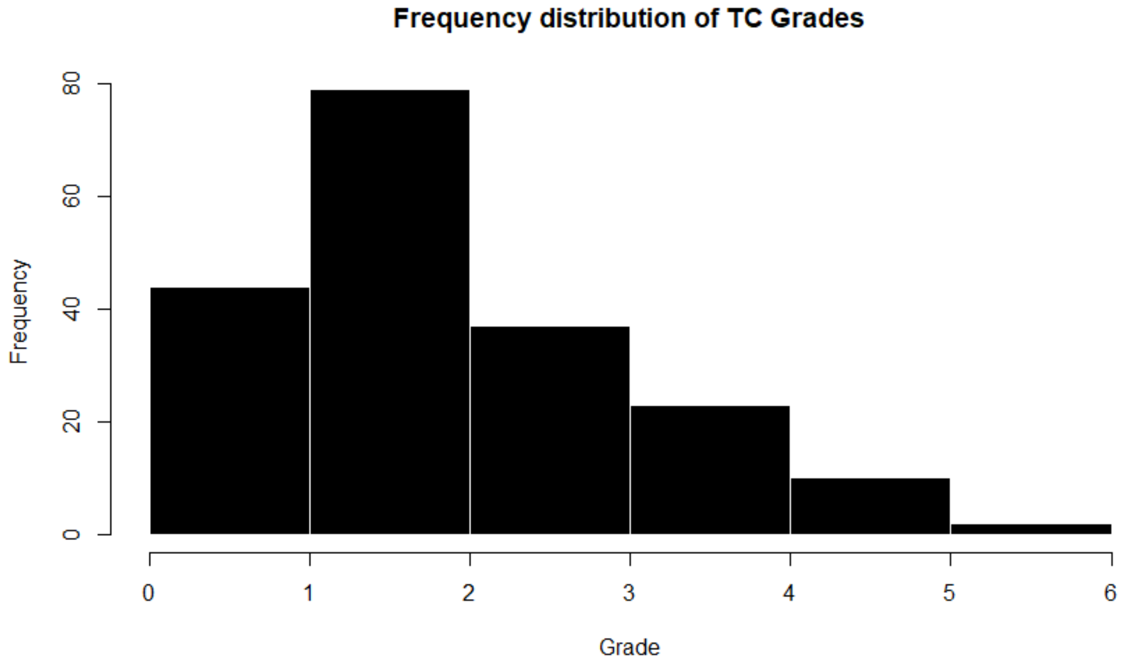

Figure 2.1. Frequency of different tropical cyclone grades in the dataset from 1982-2020

The reanalysed dataset for monthly sea level pressure, geopotential at 500 hPa, zonal wind at 700hPa and relative humidity at 500 hPa (Nath, et al., 2015) are obtained from the National Centers for Environmental Prediction (**NCEP**) and the National Center for Atmospheric Research (**NCAR**) datasets for developing the model. A grid independence study is conducted, and a 5º×4º grid size is used. The existence of correlation bullseyes in NCEP/NCAR analysis datasets (Klotzbach, 2008) supports the claim to use a large grid size. About the parameters:

- Sea Level Pressure: It is the atmospheric pressure at mean sea level. It exhibits a negative correlation with TC occurrence. This behaviour is because the sea surface temperature (SST) increases, and consequently, mean sea level pressure decreases in an area of tropical cyclone genesis.
- Geopotential: This is the potential of the gravity field of the earth. There is a negative correlation between TC occurrence and geopotential at 500 hPa. Geopotential contours can calculate the wind, which is faster where the contours are more closely spaced and is tangential to the geopotential contours.
- Zonal Wind (Zhou, 2017): Zonal winds are winds circulating at the same latitude, parallel to the equator. There are two forms of wind shear: vertical and horizontal.

Vertical wind shear is the change in the speed and direction of winds at increasing heights in the atmosphere. It is more influential than horizontal wind shear when it comes to TCs. It is positively correlated with TC occurrence.

- Relative Humidity: It is the ratio of the amount of water vapour present in the air to the highest amount possible at the same temperature. High relative humidity (RH) is necessary to attain the maximum intensity of a TC (Kaplan, 2003). Moist air also has a positive influence on TC intensification as it promotes the formation of warm air drafts from low levels to the sub cloud layer. Dry air is less buoyant than moist air at the same temperature and thus limits ascending movement.

In this study, Pearson correlation analysis between monthly TC occurrence and the mean of the atmospheric parameters over the region (78 º E - 95º E, 5º N -25º N) is used to identify relevant predictors. Pearson correlation is a measure of linear correlation between two sets of data. It lies between -1 and 1. The correlation coefficient is represented by Equation 2.1.

$$\rho_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} \tag{2.1}$$

The results (Table 2.2) indicate that the most relevant parameters to the study are geopotential at 500hPa and mean sea level pressure.

Table 2.2. Correlation analysis of predictors

Parameter	Pearson Correlation Coefficient (95% confidence)	
	Correlation Coefficient	*Correlation type*
Geopotential at 500hPa	0.40	Negative
Mean Sea Level Pressure	0.37	Negative
Relative Humidity at 500hPa	0.12	Positive
Vertical Zonal wind at 700hPa	0.17	Positive

On analysing the relationship between the mean values of the chosen parameters and the tropical cyclone occurrences around the Bay of Bengal (Table 2.3), the pre-monsoon tropical cyclone occurrence (May-July) is heavily dependent on the mean sea level pressure variations. Hence, the geopotential variations do not play much of a role here. Conversely, in the post monsoon season (October to December), the geopotential variations play a far more critical role in tropical cyclone occurrences (Table 2.4).

In the pre-monsoon period, the mean value of the geopotential is 57399.2 m^2/s^2 with a standard deviation of 231.3 m^2/s^2 when there is no cyclone, and that of mean sea level pressure is 100924.4 Pa with a standard deviation of 123.3 Pa. It is clear from Table 2.3 that all of the mean sea level pressure values for the months with tropical cyclone occurrence lie well below 100924 Pa (due to the negative correlation), but the same does not stand for the geopotential. Note that the empty spaces in the graph occur because there are places over the Bay of Bengal where cyclone genesis is less common, and hence the data is not available in the dataset at hand.

Table 2.3. Area-wise Mean parameter values for tropical cyclone occurrences in the pre-monsoon period

Latitudes	**Longitudes**							
	78-82.5		***82.5-87***		***87-91.5***		***91.5-96***	
	MSL	**GP**	**MSL**	**GP**	***MSL***	***GP***	***MSL***	***GP***
5-9	100634.2	57488.2			100478.3	57491.3		
9-13								
13-17			100790	57590	100315	57365.93	100340	
17-21	100321.2	57354.8	100670.5	57458.2	100839.4	57504.41	100580.7	57359.4
21-25	100480.6	57341.9	100491.4	57358.2	100765.4	57486.94		

A grid of the mean parameter values of mean sea level pressure (MSL) in Pa and geopotential (GP) in m^2/s^2 over the ranges of latitudes and longitudes displayed in the pre-monsoon months when a tropical cyclone occurs in the dataset

In the post-monsoon period, the mean value of the geopotential is 57514.1 m^2/s^2 with a standard deviation of 118.2 m^2/s^2 when there is no cyclone, and that of mean sea level pressure is 101253.5 Pa with a standard deviation of 132.5 Pa. It is clear from Table 2.4 that all of the geopotential values for the months with tropical cyclone occurrence lie well below 57514.1 Pa (due to the negative correlation), but the same does not stand for the MSL pressure.

Table 2.4. Area-wise mean parameter values for tropical cyclone occurrences in the post-monsoon period

Latitudes	**Longitudes**							
	78-82.5		***82.5-87***		***87-91.5***		***91.5-96***	
	MSL	**GP**	**MSL**	**GP**	***MSL***	***GP***	***MSL***	***GP***
5-9	101540.9	57246.9	101185.4	57352				
9-13	101028.4	57487.8	101510.8	57302	101484.6	57367.5		
13-17	100984.9	57320.4	101036.1	57270.5	101206.8	57402.1	101365	57321.6
17-21			101011	57326.6	101003.9	57469.7	100897.2	57276.8
21-25					101200.4	57472.7		

A grid of the mean parameter values of mean sea level pressure (MSL) in Pa and geopotential (GP) in m^2/s^2 over the ranges of latitudes and longitudes displayed in the post-monsoon months when a tropical cyclone occurs in the dataset

2.2. Data Preparation

Normalisation is a technique often applied as part of data preparation for machine learning (Sola, 1997). Normalisation changes the values of numeric columns in the dataset to a standard scale without distorting differences in the ranges of values. It is required when features have different ranges. Here, the geopotential height (m) has an order of magnitude of 4, while the mean sea level pressure (Pa) has an order of 5. Therefore, each column of the data is normalised. Normalisation is an excellent technique when the data distribution is unknown or when the distribution is not Gaussian. It is also referred to as a min-max scaling and is represented by Equation 2.2.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2.2}$$

Two more predictors are created. The first is "month", which indicates which month of the year it is. It consists of three buckets: "1" for January, February, March and April; "2" for May, June, July, August and September; and "3" for October, November and December. This is because the likelihood of a cyclone occurring in any month is different, and the predictor can help the model learn what a 'monsoon'. In addition, it can help the cyclone learn the differences between the pre and post-monsoon period, an important distinction as explained in section 2.1. For example, from Figure 2.2, it is clear that there is a significant bias for cyclones occurring during the monsoon season, i.e. May-July or October-December. By providing the month of the year as an input, the model learns that these months are more favourable than others.

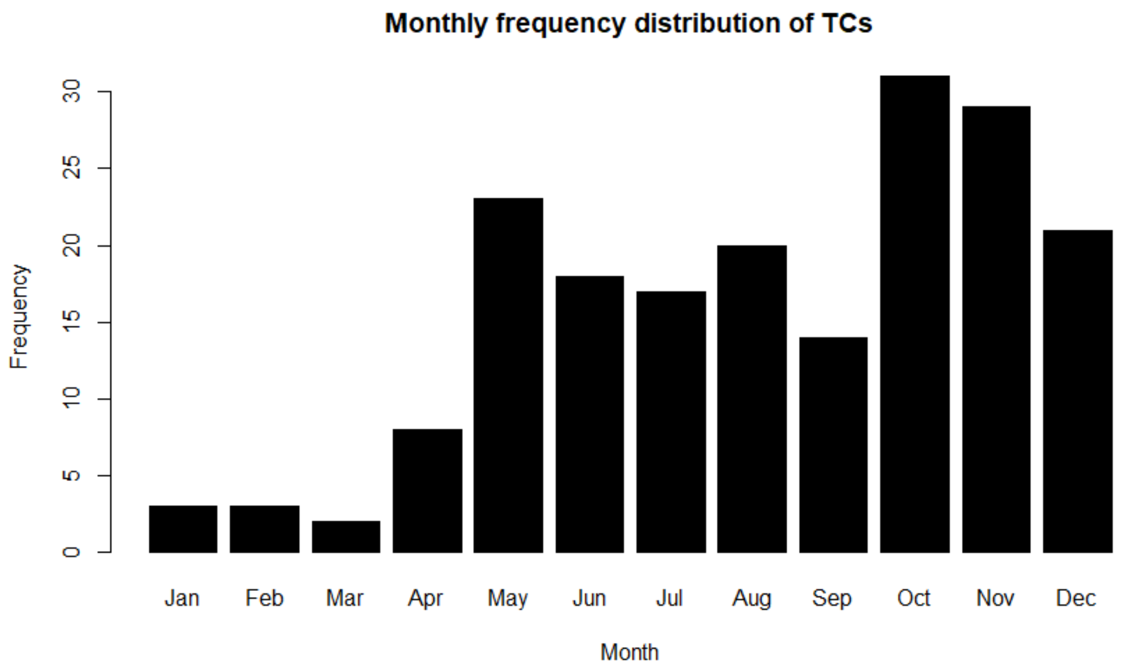

Figure 2.2. Monthly frequency of tropical cyclone occurrence from 1982-2020

The second feature is "Grade", which indicates the intensity of the cyclone. Its values range from 1 to 7, depending on the intensity of the cyclone that month. These are used as row weights to give the model a clearer understanding of cyclone occurrence and give high importance to predicting more severe cyclones.

The data from each geographical grid point is a column in the dataset. Further, the correlation coefficient between all the predictor columns is checked and used to remove redundant columns (correlation>0.95).

2.3. Modified Logistic regression Model

Logistic regression is a supervised machine learning algorithm used for binary classification. The central function, the logistic function, is stated in equation 2.3.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

Where e is the base of natural logarithms and x is the value to transform. This function outputs a value between 0 and 1.

Logistic regression begins with a linear transformation, much like linear regression. First, input values (x) are combined with weights and added to an intercept. Next, this expression is input into the logistic function (Equation 2.4). Finally, the output data is classified into "0" or "1" classes (Equation 2.5) by choosing an optimum cutoff value.

$$y = \frac{1}{1 + e^{-(a+bx)}} \tag{2.4}$$

$$Output = \begin{cases} 0, & andy < cutoff \\ 1, & andy \geq cutoff \end{cases} \tag{2.5}$$

The coefficients are optimised using the maximum-likelihood algorithm. The loss function for the same is given by Equation 2.6.

$$\log(loss) = \frac{1}{N}\sum_{i=1}^{n}[-\left(w_0(y_i \log(\hat{y}_i)) + w_1\left((1-y_i)\log(1-\hat{y}_i)\right)\right)] \tag{2.6}$$

Modifications:

In this study, the time-series nature of the data calls for altering the input of the regression model. The predictors of previous rows of the dataset are added to the given row of the dataset as columns. In order to understand the number of previous rows to be considered, a correlation plot (auto-correlation function) has been drawn between tropical cyclone occurrence in a month and the mean values of the predictors of the past months (Fig 2.3). It is observed that around T-12 months can be considered correlated, with a standard threshold of 0.1

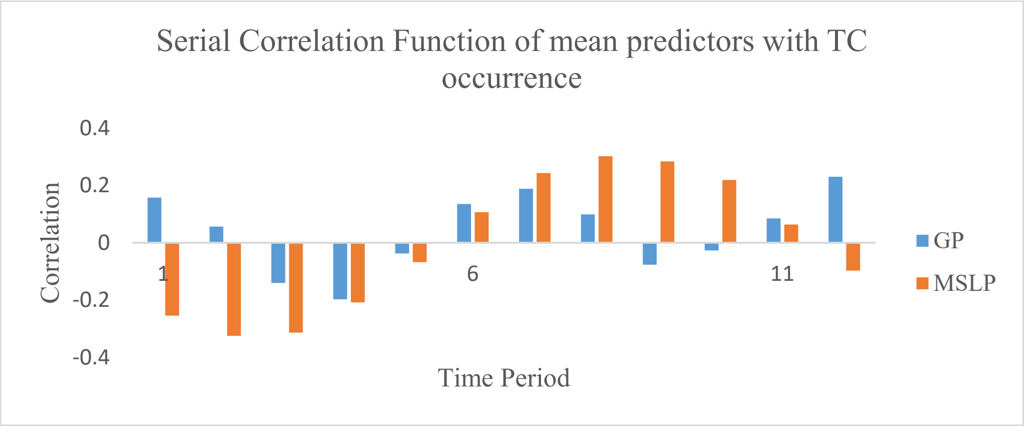

Figure 2.3. Serial Correlation of Mean values of Geopotential and Mean Sea Level Pressure with Tropical Cyclone Occurrence

There is also a class imbalance in the dataset. The number of data points where a cyclone does not occur outnumbers the number where a cyclone occurs by the ratio of 2:1. In order to solve this, the input data is assigned weights in the form of G where “G” is the grade of the cyclone. The logistic regression equation now became equation 2.7.

$$y = \frac{1}{1 + e^{-(G)(a+bx)}} \tag{2.7}$$

Where a, b are parameters and G, x are inputs.

Finally, the model is k-fold cross-validated. Cross-validation is a valuable way to determine the accuracy when there is not enough data. It outputs the average accuracy for different test-train data combinations.

- ‘K’ groups of the test dataset are extracted, ensuring the optimum class balance.
- One unique group is selected. The remaining data acts as the training set.
- The model is then fit on the training set and evaluated on the test dataset.
- This process repeats for each fold, and the average accuracy is output.

10-fold cross-validation has proved most successful for the current dataset.

A grid study is conducted to optimise the hyperparameter (Table 2.5). The data split between training and testing datasets is in the ratio of 0.85:0.15. Furthermore, the cutoff probability for outputting ‘1’ is 0.65.

Table 2.5. Grid validation study for modified logistic regression parameters

Data split (across) **Probability (down)**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**
0.5	0.71	0.74	0.61	0.71	0.68	0.62
0.55	0.68	0.61	0.71	0.71	0.7	0.74
0.6	0.69	0.67	0.68	0.65	0.64	0.75
0.65	0.74	0.69	0.64	0.7	0.67	**0.82**
0.7	0.66	0.7	0.72	0.75	0.67	0.59
0.75	0.68	0.77	0.7	0.62	0.69	0.62
0.8	0.7	0.66	0.73	0.64	0.75	0.68
0.85	0.68	0.69	0.65	0.74	0.63	0.63
0.9	0.7	0.65	0.62	0.8	0.63	0.79
0.95	0.72	0.69	0.57	0.66	0.73	0.71

2.4. Long Short-Term Memory Algorithm

Neural networks are a set of algorithms modelled after the human brain and designed to recognise patterns. Neural networks consist of input, output and hidden layers. The core unit in a neural network is called the neuron or node (Fig 2.4). These units receive input from either the data source or some other units. Next, the input is combined with various weights and biases to generate an output. Then node weights are allotted based on relative importance. Finally, the node output goes through a transfer function.

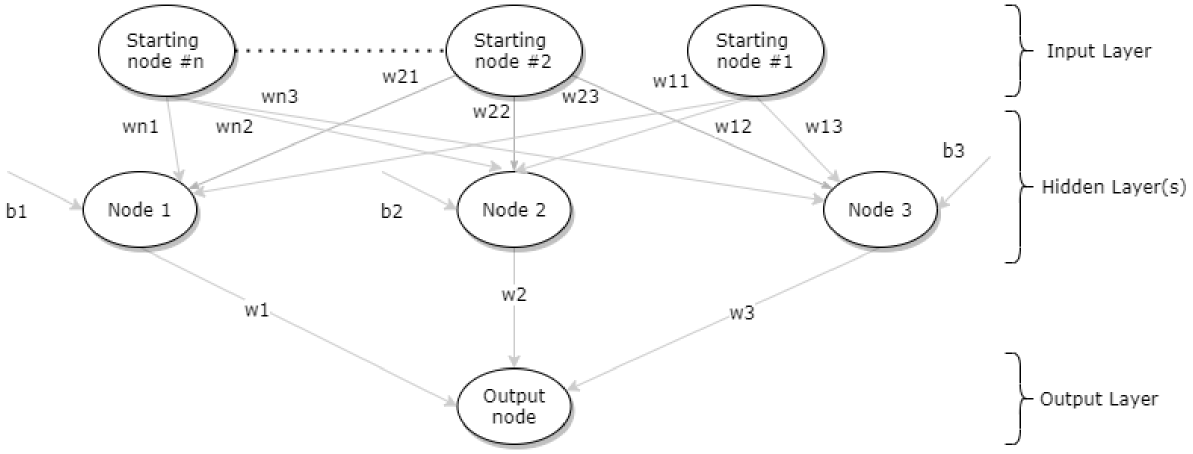

Figure 2.4. Model of an artificial neural network

Classical neural networks have a particular disadvantage for the time series dataset. There is no memory of past observations and outputs. Thus, an input data point to the model produces an output with no input from the preceding data point. Recurrent neural networks (RNNs) solve this. As shown in Fig 2.5, any output from the previous data point is processed and fed into the model again for the next time instant, along with the following data point.

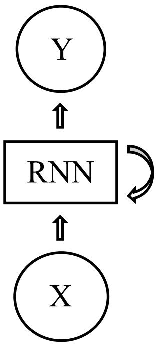

Figure 2.5. Model of a recurrent neural network.

While being generally highly successful, one more issue arises here for a time series dataset. There is a vanishing gradient problem, or the inputs from past data points rapidly decreasing in importance as the model trains. Long short-term memory networks (LSTMs) solve this. Unlike feedforward neural networks, LSTM has feedback connections. As a result, it can process large sequences of data effectively. This is because of the property of remembering patterns for long durations of time.

A typical LSTM network comprises different memory blocks called cells (middle rectangles in Fig 2.6). Each cell receives 2 data points, the first one being the input data point (x), and the second is input from the previous cell (h or hidden state). Also, each memory block has three major manipulations performed on the input data, and these mechanisms are called gates.

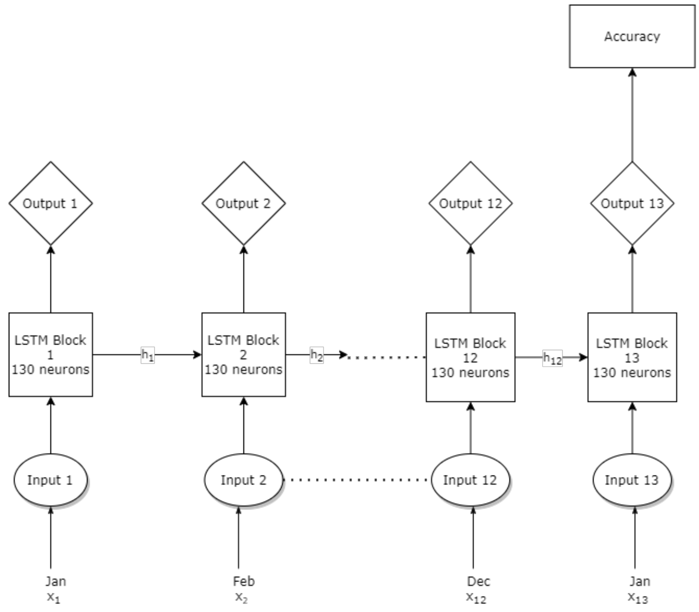

Figure 2.6. Model of the LSTM network

The inputs are multiplied with a weight matrix and added to a bias. The output passes through a sigmoid function in order to normalise it. The sigmoid function practically decides if a number is essential or not. The block output is further manipulated and combined with the output of the previous block via a regulatory filter to create the input to the next block. Due to the normalising filter, the network keeps a memory of all the previous blocks. Therefore, it can effectively decide which data points to remember. This property is crucial for seasonal use cases since the most significant data point can be a whole year before the current data point.

The LSTM network has a single hidden layer (Sankar et al., 2015). A 12-month lookback window is considered in the network, allowing the model to take in the past 12 months of data points and the current data point to predict the output of those months. This structure makes the problem a ‘many to many’ kind of LSTM. That means that the model will be trained to input 13 rows of input (current month and past 12 months) and will output 13 months’ worth binary output. However, only the output of the last month is used to calculate the accuracy of the model.

A dense time distributed layer is the next layer. It allows the use of a single layer for every input, i.e. the reuse of the single-layer when the next time step comes but with different inputs. The layer is “dense”. That means all the data inputs pass through a single hidden layer. This property helps in 2 ways:

- It allows the problem to be modelled as defined, with a single output to a single input.
- It makes the network much simpler by needing much fewer weights and biases. However, the network needs to loop through more epochs.

A grid study is conducted to choose the hyperparameters. These are the number of epochs and the number of neurons in the hidden layer. In this case, 100 epochs with 130 neurons have the highest validation accuracy (Table 2.6). Here, the classification nature of the problem calls for optimising the binary cross-entropy loss.

Table 2.6. Grid validation study for modified LSTM parameters

# of epochs (across) **# of neurons (down)**	**50**	**100**	**200**	**400**	**700**	**1000**
10	0.77	0.81	0.83	0.86	0.82	0.84
20	0.8	0.83	0.87	0.87	0.9	0.85
30	0.81	0.9	0.92	0.88	0.86	0.9
50	0.78	0.91	0.87	0.93	0.91	0.89
100	0.89	0.93	0.88	0.89	0.91	0.8
130	0.9	**0.94**	0.9	0.85	0.89	0.84
150	0.91	0.9	0.89	0.87	0.9	0.83

One may think that the large number of parameters yielded by this model, with our limited dataset, will cause massive overfitting. However, the test accuracy proves otherwise. Besides this, Vinyals et al., 2016 show that a simple neural network with 2n+d parameters can perfectly fit any dataset of n samples of dimension d. However, even though commonly used neural networks have much more than 2n+d parameters, they do not necessarily overfit.

The dataset splits into training, testing and validation data. Training data is the dataset used to train the model (weights, biases and other parameters). The model sees and learns from this data. The validation dataset is the unbiased estimate of fit on the training dataset while tuning model hyperparameters. Finally, the test dataset provides an unbiased estimate of the final model fit.

CHAPTER 3

RESULTS AND DISCUSSION

In this section, the success metrics of both predictive models are evaluated. The following parameters are used to evaluate the performance: The accuracy, recall (or true positive rate, TPR), AUC (Area Under The Curve)- ROC (Receiver Operating Characteristics) curve and f1 score.

The accuracy is the fraction of data points correctly classified.

A true positive (TP) is when the model correctly predicts the positive class. Similarly, a true negative (TN) is when the model correctly predicts the negative class. A false positive (FP) is when the model incorrectly predicts the positive class. A false negative (FN) is when the model incorrectly predicts the negative class. Since the data at hand is devastating cyclones, the number of false negatives are minimised on priority.

The ROC curve is a metric for classification problems, plotted with TPR (Equation 3.1) on the y axis and false positive rate (FPR) (Equation 3.2) on the x-axis. The ROC curve measures the degree of separability at different thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both FPs and TPs. The AUC is the area under the ROC curve. It is an aggregate measure of performance for the model and is threshold invariant. If the AUC is high, the model is generally better at ranking positive examples over negative ones. The best model has an AUC of 1.

$$TPR = Recall = \frac{TP}{TP + FN} \tag{3.1}$$

$$FPR = \frac{FP}{TN + FP} \tag{3.2}$$

The f1 score explains the balance between recall and precision. Precision is the fraction of “useful” instances among the retrieved instances (Equation 3.3). It is the number of true 1s out of all the 1s predicted. The recall (or TPR) is the fraction of relevant instances retrieved or the number of true 1s divided by the total number of 1s in the original dataset. To fully evaluate the effectiveness of a model, both precision and recall are valuable. Unfortunately, precision and recall are often inversely proportional. For example, increasing the classification threshold may increase the precision but reduce the recall.

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.4}$$

3.1. Logistic Regression

In a simple weighted logistic regression, where the class weights are assigned to remove the class imbalance by giving data points where the cyclone occurs (y=1) a class weight of 2, the accuracy on the test dataset is 0.74, and the recall is 0.63.

In the modified logistic regression case, as elaborated in section 2.2, the accuracy on the test dataset is 0.82, recall is 0.81, and precision is 0.76. The recall is better than the precision, and this balance is suitable for this use case because it minimises the number of false negatives. The f1 score is 0.78. It indicates a good balance between the two.

The confusion matrix is shown below in Table 3.1.

Table 3.1. Confusion matrix of modified logistic regression

Predicted (across) **Real (down)**	**0**	**1**
0	32	7
1	5	22

The model has done relatively well at predicting the data. The AUC-ROC is 0.79, which indicates that the model is suitable for classification. A point to note is that the model always correctly predicts high-grade cyclones (Grade>=3 or anything above a deep depression). In addition, the threshold selected via the ROC curve indicates that it is well balanced (black dotted line in Fig 3.1).

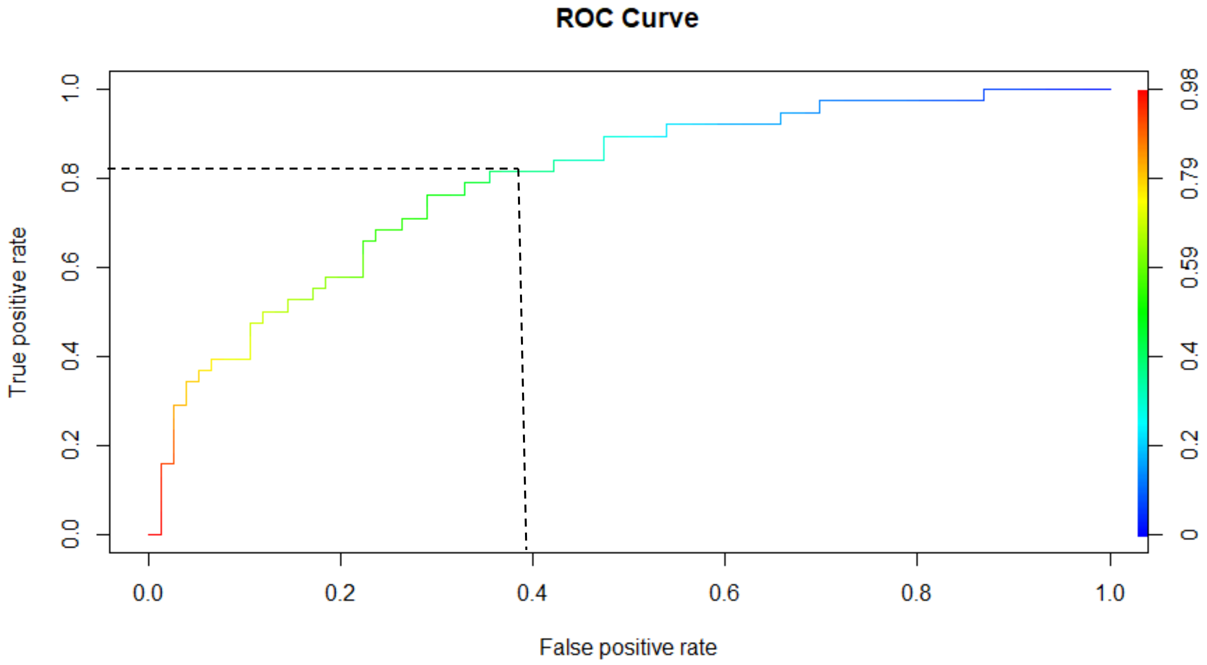

Figure 3.1. ROC curve of modified logistic regression

3.2. Long Short-Term Memory

In the case of simple artificial neural networks, the accuracy was a meagre 0.66. However, the long short-term memory algorithm yielded a test accuracy of 0.93 and an f1 score of 0.92. The epoch loss is plotted with many epochs, and the stopping point is selected as the minima of the validation loss curve (Fig 3.2). Any stopping point beyond this will overfit the model.

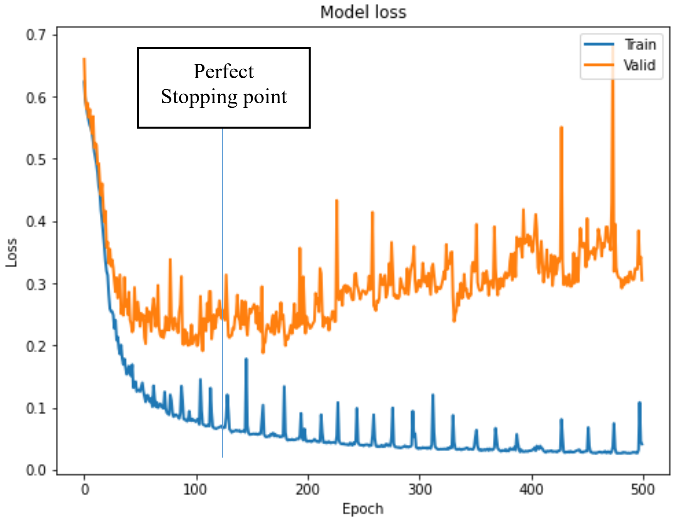

Figure 3.2. LSTM model loss mapped against the number of epochs run

The confusion matrix is shown below in Table 3.2. The recall is 0.92, which is more than adequate. There is a meagre number of false negatives.

Table 3.2. Confusion matrix of modified lstm

Predicted (across) **Real (down)**	**0**	**1**
1	2	22
0	28	2

Only low-grade cyclones are classified incorrectly, i.e., low-pressure areas or depressions. Anything above depression is consistently classified correctly.

The ROC curve has a near-perfect shape (Fig 3.3), with an AUC-ROC of 0.93. The sharp elbow is because of the limited dataset size.

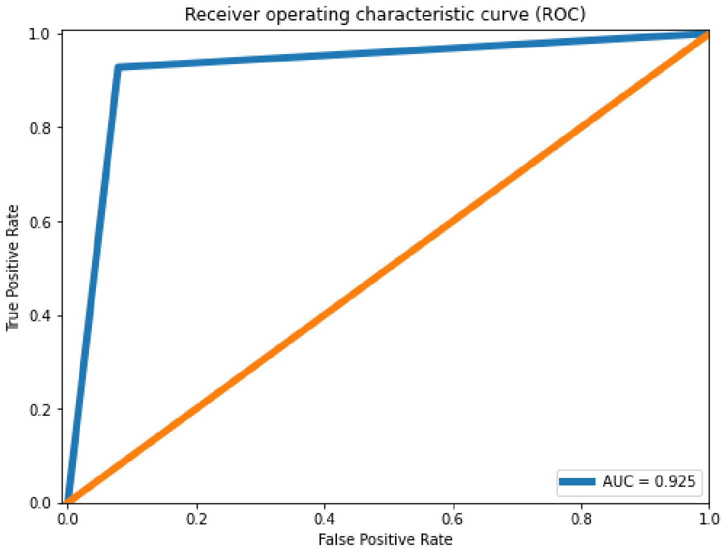

Figure 3.3. ROC curve of modified LSTM

On comparing the logistic regression and neural network model (Table 9), it is clear that the LSTM model is superior by all counts when it comes to performance metrics.

Table 3.3. Comparison of performance metrics of algorithms

Parameter	**Algorithm**	
	Modified Logistic regression	***LSTM neural network***
Test Accuracy	0.82	0.90
Recall	0.81	0.90
F1 score	0.78	0.91
AUC-ROC	0.79	0.93

However, when we look at the amount of time it takes the model to run, we see that the logistic regression is far quicker. This factor can be significant for larger datasets.

3.3. Prediction

Representative Concentration Pathways (RCPs) are scenarios that convey different trajectories for carbon emissions and the resulting atmospheric carbon dioxide concentration from 2000 to 2100. Thus, they encompass an extensive range of possible climate policy outcomes for the 21st century. The RCPs describe four different scenarios (Fig 3.4, derived from [4]) based on different assumptions about economic growth, population, energy consumption and land use.

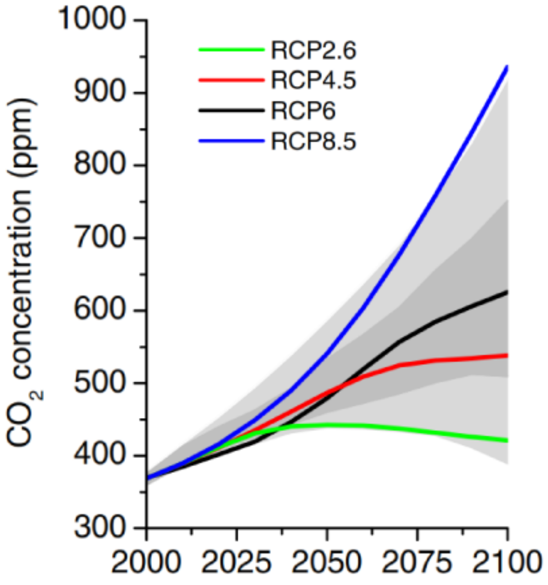

Figure 3.4. Carbon dioxide concentrations for different RCP scenarios (Van Vuuren, 2011)

The LSTM model predicts future TC occurrences for different RCP scenarios. The input data is from the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor, 2012). It is a project of the World Climate Research Programme (WCRP) for providing time-projected environmental variables. The work represents the unification of climate modelling teams across the globe. A large repository of output physical variables and their time series was generated.

The output number of months predicted when a tropical cyclone would occur for the different scenarios is in Table 10. The simulation time is at least 50% lower for all scenarios. For context, it takes a 64 core aqua cluster around a week, producing 100GB of data to produce a 10 km resolution of cyclone parameters for three months. By cutting between 490 and 680 months from our simulation time requirements, we potentially save 16,300-22,700 GB of server space, plenty of money and many months of simulations. Thus, predictive models are an effective method that can be applied to narrow down our downscaling window to perform TC predictions faster and more efficiently.

Table 3.4. Simulation data reduction

Scenario	Number of months predicted	Reduction in simulation time
RCP 4.5	280	70.83%
RCP 6.0	470	51.04%
RCP 8.5	377	60.72%

Fig 3.5 shows the average number of cyclonic months per year from 1982 to 2100. RCP 4.5 seems to have the shortest average cyclone period every year, while RCP 6.0 has the largest. The average cyclonic period per year is mapped using the moving average method (number of years = 10) and is smoothened. The yearly average period of tropical cyclone occurrence by 2100 is 4.2 months for RCP 8.5, 6.1 months for RCP 6.0 and 3.0 months for RCP 4.5. Note that this is not a direct comment on the actual number of tropical cyclone occurrences or the intensity of the cyclones.

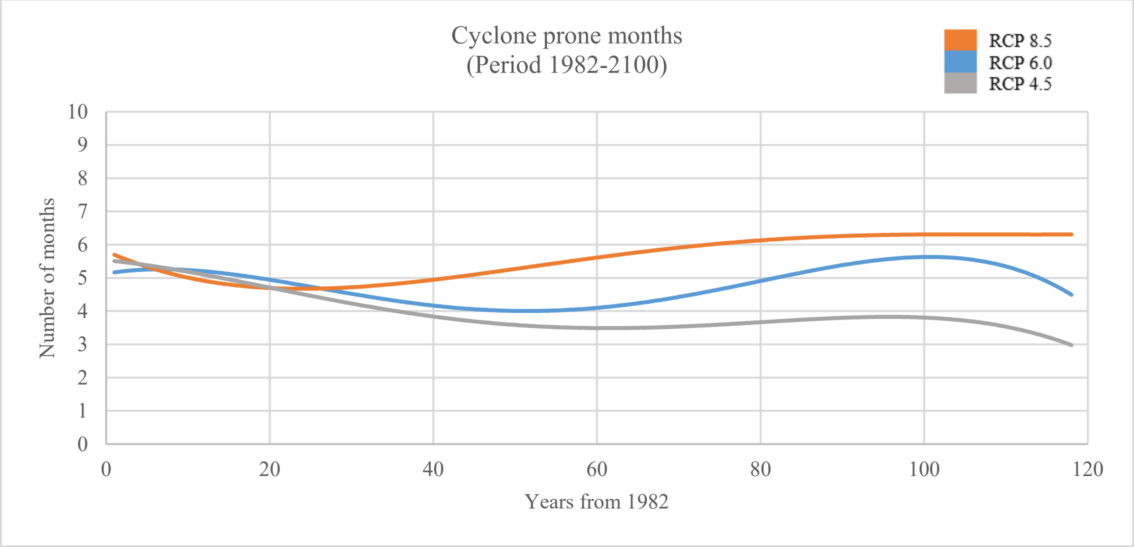

Figure 3.5. Moving average of cyclone-prone months per year for different RCP scenarios

3.5. Theoretical Reasoning

There are many reasons why tropical cyclone occurrence changes. According to the predictions, the tropical cyclone occurrence will reduce over BOB by 2100 (in RCP 8.5 and 6.0 scenarios). That is mainly due to the vertical wind shear and vorticity (Murakami, 2013). Conversely, TC occurrence is predicted to increase over the Arabian Sea (AS). The TC frequency (TCF) can be defined as Equation 3.5.

$$TCF(A) = \iint g(A_0)t(A, A_0)dA_0 \tag{3.5}$$

Where TCF(A) is the TCF in a specific region, $g(A_0)$ is TC genesis frequency (TCGF) in the region A_0, and $t(A, A_0)$ is the probability that a TC generated in region A_0 travels to region A.

Tropical cyclone change can be defined as Equation 3.6.

$$\delta TCF = \delta G \times T + G \times \delta T + \delta G \times \delta T \tag{3.6}$$

Where δ is future change. Tropical cyclone frequency is broken down into three factors: future change due to (a) TC genesis (first term), (b) TC tracks (second term), and (c) a complex term (third term). This can further be broken down by season.

In the pre-monsoon season of April–May-June, TCGF is predicted to decrease in the BOB. Vertical ascent (omega), relative humidity and vertical wind shear appear to be the significant factors influencing this change. In contrast, TCGF is projected to increase in the peak-monsoon season of October-November-December. An increase in the relative humidity is the most significant factor contributing to this increase, followed by a change in vertical ascent (Bell, 2020). This is consistent with the results obtained in this study (Fig 3.6), assuming that the number of TC ridden months is a good proxy for the actual number of TCs per month.

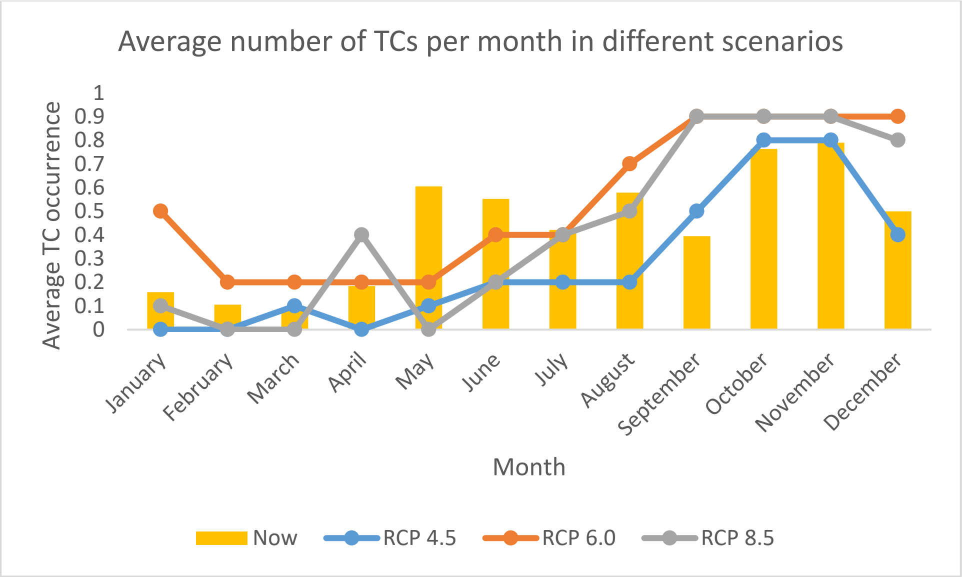

Figure 3.6. Average number of TC ridden months from 2090-2100 in different RCP scenarios, compared to the present

Although the TC track factor is of less importance for the TCF changes, there is predicted to be a westward shift of TC tracks. Thus, overall, the TCF will decrease over the BOB and increase over the AS.

The increasing sea surface temperature (SST) is widely regarded as a reason for an increase in the frequency of TCs. However, it has been proven to be a threshold of around 26-27 degrees (Tory, 2015). Further increase of SST will not necessarily lead to increasing cyclones, and many other factors come into play. It is also hypothesised that the frequency of cyclones will decrease, but the intensity will increase (Bell, 2020)

3.4. Model Validation

In order to validate the numerical result of the present study, the simulation results are compared to the results from the literature. There is limited research in applying machine learning models in the prediction of tropical cyclones around the Bay of Bengal region. One paper, Nath et al., 2015, has compared the number of tropical cyclones occurring per year using different machine learning models. Their neural network model is also superior in predicting tropical cyclone occurrences using large scale climate variables. However, since the task involves regression, not classification, the metrics cannot be compared.

In order to validate the predictions (Section 3.3), a study by Bell et al., 2020 is consulted. The study projects the number of tropical cyclones in the future for the RCP 8.5 scenario using different models. The output shows a decrease in tropical cyclones in 2070-2100 compared to 1990-2020. Furthermore, the average number of cyclones per year by 2100 is between 3 and 4 (depending on the model). These conclusions are comparable with those concluded in this study, with the average number of cyclonic months being around 4.2. Unfortunately, data for the other RCP scenarios was not available. However, since the same model is used to predict the other RCP outcomes and the model itself has been validated, it can be stipulated that the predictions are valid.

CHAPTER 4
SUMMARY AND CONCLUSIONS

WRF simulations are time-consuming and computationally expensive, leading to high monetary cost. The dependence on WRF is reduced by predicting tropical cyclone occurrence via machine learning models. It also paves the way for more democratised research into the devastating tropical cyclones over the North Indian Ocean. The summary and conclusions:

1. Using correlation analysis, mean values of geopotential at 500 hPa and mean sea level pressure are chosen as predictors, along with the month of occurrence. Finally, cyclonic intensity is chosen as a weight. The output is a list of months from 1982 to 2020 and whether a tropical cyclone occurred in that month.
2. The mean sea level pressure is far more critical during the pre-monsoon period than the geopotential for cyclonic activity detection. However, in the post-monsoon period, the converse is true.
3. A modified version of logistic regression and long short-term memory recurrent neural network models that capture the time-series nature of the data is designed. Accuracy, recall, f1 score and AUC-ROC are the performance parameters are chosen to evaluate the models.
4. The results of the training period are consistent with those of the testing period, and the performance parameters are satisfactory. LSTM RNN is, however, far better than logistic regression in terms of performance. Furthermore, both models can predict high-grade cyclones with a 100 per cent accuracy, i.e., all the false negatives pertain to lower grade cyclones.
5. RCP 4.5,6 and 8.5 scenarios are considered for future tropical cyclone prediction (2022-2100). The data required to downscale via WRF is successfully shortened by 50-70% by predicting the months when a tropical cyclone will most likely occur in advance and only downscaling in those months. RCP 6.0 has the greatest projected cyclonic period per year in 2100, followed by RCP 8.5 and RCP 4.5.

RECOMMENDATIONS FOR THE FUTURE WORK

There is much work that needs to be done over the north Indian ocean with regards to reducing the dependence on WRF simulations:

- This study had reduced the window of downscaling on a monthly basis. The same is to be done with a shorter period to reduce the window further and save more time.
- More advanced deep learning models should be applied for different use cases since they yield a better accuracy
- Larger datasets should be used with more environmental variables in order to produce more widely accepted outputs.

CHAPTER 5

REFERENCES

Bell, S. &. C. S. &. T. K. &. Y. H. &. T. C., 2020. North Indian Ocean tropical cyclone activity in CMIP5 experiments: Future projections using a model-independent detection and tracking scheme. *International Journal of Climatology. 40.*

Bourras D., W. T. L. a. E. W. T., 2003. Evaluation of latent heat flux fields from satellites. *J. Appl. Meteorol 42,* pp. 227-239.

Chan J. C. L., J. S. a. K. S. L., 2001. Improvements in the seasonal forecasting of tropical cyclones over the western North Pacific. *Weather Forecast. 16,* pp. 491-498.

Davidovic, D. &. S. K. &. B. D. &. P. M., 2009. Grid implementation of the weather research and forecasting model. *Earth Science Informatics,* pp. 199-208.

DeMaria, M., Knaff, J. A. & Connell, B. H., 2001. A Tropical Cyclone Genesis Parameter for the Tropical Atlantic. *Weather and Forecasting,* pp. 219-233.

Hewage, P., Trovati, M. & Pereira, E., 2021. Deep learning-based effective fine-grained weather forecasting model. *Pattern Anal Applic 24,* p. 343–366 .

Hossain M, R. B. L. S. D. S., 2015. Forecast-ing the weather of Nevada: a deep learning approach. *2015 international joint conference on neural networks (IJCNN), July 2015,* pp. 1-6.

Kannaiyan, M. &. G. K. &. J. G., 2019. Prediction of specific wear rate for LM25/ZrO2 composites using Levenberg–Marquardt backpropagation algorithm. *Journal of Materials Research and Technology. 9.*

Kaplan, J. a. M. D., 2003. Large-scale characteristics of rapidlyintensifying tropical cyclones in the North Atlantic basin. *WeatherForecast., 18(6),* p. 1093–1108.

Kim H. S., H. .. C. P. S. C. a. J. H. K., 2010. Seasonal prediction of summertime tropical cyclone activity over the East China Sea using the least absolute deviation regression and poisson regression. *Int. J. Climatol, 30,* pp. 210-219.

Klotzbach, P., 2008. Refinements to Atlantic basin seasonal hurricane prediction from 1 December. *Journal of Geophysical Research. 113.*

Mitra A. K., P. K. K. A. K. S. a. S. K. R., 2010. A neural network approach for temperature retrieval from AMSU-A measurements onboard NOAA-15 and NOAA-16 satellites and a case study during Gonu cyclone. *Atmosfera 23,* pp. 225-239.

Murakami, H. S. M. &. K. A., 2013. Future changes in tropical cyclone activity in the North Indian Ocean projected by high-resolution MRI-AGCMs. *Clim Dyn ,* Volume 40, p. 1949–1968.

Murakami, H. et al., 2016. Statistical-Dynamical Seasonal Forecast of North Atlantic and U.S. Landfalling Tropical Cyclones using the High-Resolution GFDL FLOR Coupled Model. *Monthly Weather Review.*

Nath, S., Kotal, S. D. & Kundu, P. K., 2015. Seasonal prediction of tropical cyclone activity over the north Indian Ocean. *Atmósfera 28(4),* pp. 271-281.

Neumann, 1993. *Global Guide to Tropical Cyclone Forecasting,* s.l.: s.n.

Raju, P., Potty, J. & Mohanty, U. C., 2012. Prediction of severe tropical cyclones over the Bay of Bengal during 2007–2010 using high-resolution mesoscale model. *Natural Hazards, Volume 63.*

Sola, J. &. S. J. (., 1997. Importance of input data normalization for the application of neural networks to complex industrial problems.. *Nuclear Science, IEEE Transactions on. 44.,* pp. 1464-1468.

Taylor, K. E. S. R. J. &. M. G. A., 2012. An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society, 93(4),* pp. 485-498.

Tory, K. J. a. D. R. A., 2015. Sea Surface Temperature Thresholds for Tropical Cyclone Formation. *Journal of Climate ,* 28(20), pp. 8171-8183.

Van Vuuren, D. E. J. K. M. e. a., 2011. The representative concentration pathways: an overview. *Climatic Change,* Volume 109.

Vinyals, C. Z. a. S. B. a. M. H. a. B. R. a. O., 2016. Understanding deep learning requires rethinking generalization. *CoRR, 1611.03530.*

Zhao, C. P. S. a. X., 2007. A Bayesian regression approach for predicting tropical cyclone activity over the central North Pacific. *J. Climate. 20,* pp. 4002-4013.

Zhou, B. X. Y., 2017. How the “best” CMIP5 models project relations of Asian–Pacific Oscillation to circulation backgrounds favorable for tropical cyclone genesis over the western North Pacific. *J Meteorol Res 31,* p. 107–116 .

DATA SOURCES

[1] TC occurrence 1982-2020: www.imd.gov.in

[2] Environment variables 1982-2020: ERA5 | ECMWF

[3] RCP scenario environment variables 2021-2100: RCP prediction

[4] Figure. 10: The representative concentration pathways: an overview | SpringerLink