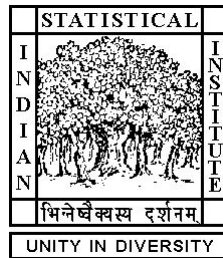


DISCOVERING PREDICTORS FOR RAPID INTENSIFICATION OF CYCLONES IN BAY OF BENGAL

Dissertation Submitted in Partial Fulfillment of the Requirements for the
Degree of
Master of Technology in Computer Science

by
Manideep Aileni
[Roll no: CS2014]

under the supervision of
Dr. Ashish Ghosh
Machine Intelligence Unit

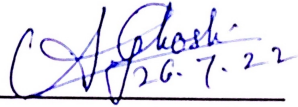


Indian Statistical Institute
Kolkata - 700108, India

July 2022

CERTIFICATE

This is to certify that the thesis titled 'Discovering predictors for Rapid Intensification of Cyclones in Bay of Bengal' submitted by Manideep Aileni, to Indian Statistical Institute, Kolkata, in partial fulfillment for the award of the degree of Master of Technology in Computer Science, has fulfilled all the requirements of the institute for submission.



Dr. Ashish Ghosh
Machine Intelligence Unit,
Indian Statistical Institute,
Kolkata-700108, India.

ABSTRACT

Accurate forecast of cyclone intensities is important for disaster preparedness in the coastal areas. Intensity prediction becomes especially difficult when they undergo rapid intensification (RI). The inability to predict RI comes from the fact that the physical processes that contribute to the cyclonic intensification are not very well understood. Data mining techniques are being used to find the relationship between change in the environmental variable values, and intensification. Though good at identifying precursors to rapid intensification, data mining techniques are computationally expensive. Moreover, very few studies are conducted for the Bay of Bengal region, despite it being a hotbed of cyclonic systems. In our work, we used a computationally cheaper LSH-SNN algorithm to identify precursors to rapid intensification in Bay of Bengal, and showed its effectiveness in identifying precursors.

TABLE OF CONTENTS

Certificate	ii
Table of Contents	iv
List of Figures	v
1 Introduction	1
1.1 Cyclone Forecast	1
1.2 Rapid Intensification	2
1.3 Anomaly	3
2 Literature	4
2.1 Cyclone Forecast	4
2.1.1 RI prediction	4
3 Data	7
4 Method	9
5 Results	12
6 Future work	18
Bibliography	19

LIST OF FIGURES

1.1	Forecast track and intensity along with wind distribution (shaded region) of cyclonic storm 'Asani' on 0600 UTC May 10, 2022, given by Indian Meteorological Department.	2
3.1	Snapshot of reanalysis data downloaded from ECMWF. A circulation can be noticed in the Bay of Bengal, at image coordinates (80, 80).	8
3.2	A 40 x 40 pixel area, zoomed in on the cyclonic system. 1 pixel corresponds to 0.25 degrees latitude / longitude.	8
5.1	SNN output for RI-prior group, 12 h before the start of rapid intensification. A total of 7 clusters are visible. The clusters are concentric, as expected.	13
5.2	LSH - SNN output for RI-prior group, 12 h before the start of rapid intensification. A centroid, and concentric nature of the clusters is noticeable. 9 clusters can be counted.	13
5.3	Region 1, looking like an arc around the cyclonic centre.	14
5.4	Region 2, resembling the centre of the cyclone.	14
5.5	Region 3, also resembling the centre of the cyclone.	14
5.6	In Region 1, the correlation (Pearson's coefficient) between the mean cluster value (sea level temperature), and intensity change, is 0.2778, which is very high.	15
5.7	In Region 2, the correlation (Pearson's coefficient) between the mean cluster value (sea level temperature), and intensity change, is 0.16.	16
5.8	In Region 3, the correlation (Pearson's coefficient) between the mean cluster value (sea level temperature), and intensity change, is -0.15	17

CHAPTER 1

INTRODUCTION

Annually, 4-5 cyclones are formed in the Bay of Bengal, with 1-2 occurring in the pre-monsoon season (March, May), and 2-3 in the post-monsoon season (October, November).

The latest cyclone to originate in the Bay of Bengal is Cyclone Asani, formed on May 5, 2022, at 9.6° North longitude and 91.3° East longitude, near the Andaman and Nicobar Islands. It reached Andhra Pradesh coast on May 11, as a 'cyclonic storm', destroying crops in 30,000 hectares. Three casualties were reported. The cyclone dissipated within the next 24 h.

A cyclonic system starts as a depression, evolves into a deep depression, and then intensifies into a cyclonic storm. Eventually, the cyclone dissipates. When severe cyclonic storms hit coastal areas, they bring heavy rains, and wind, and cause destruction to property and life. If cyclone intensities can be predicted with good accuracy, the coastal areas could be better prepared for receiving the intense cyclones.

1.1 Cyclone Forecast

The Indian Meteorological Department (IMD) keeps a track of the cyclones in Bay of Bengal, and issues forecasts and warnings.

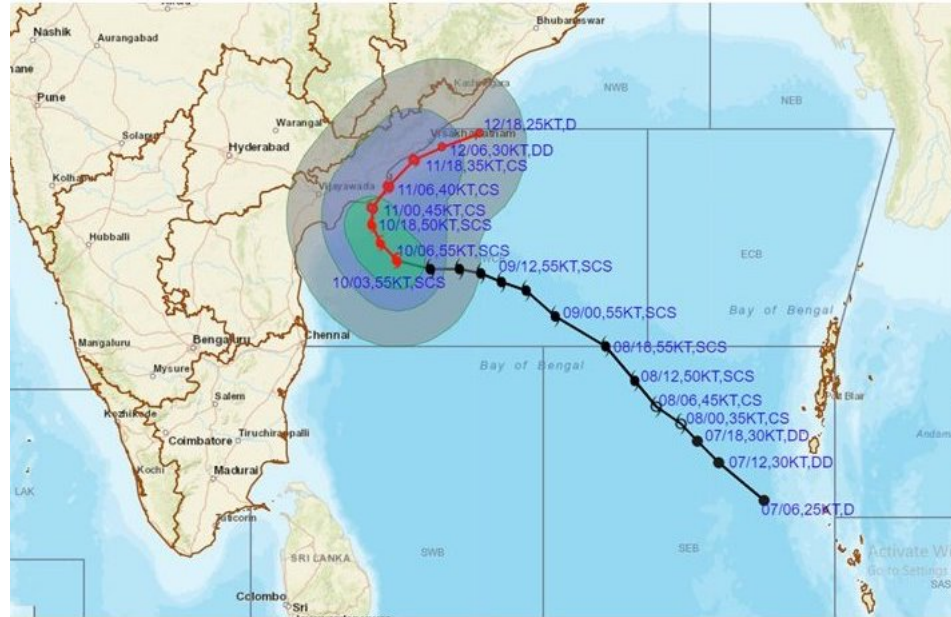


Figure 1.1: Forecast track and intensity along with wind distribution (shaded region) of cyclonic storm 'Asani' on 0600 UTC May 10, 2022, given by Indian Meteorological Department.

A phenomenon that makes the cyclone intensity unexpectedly deviate from the estimated forecast is 'rapid intensification'.

1.2 Rapid Intensification

During their evolution, cyclones are prone to rapid intensification (RI), which may be quantitatively identified as - an increase of sustained wind speed at the cyclonic center by 30 knots in 24 h.

When such sudden increase in the intensity of the cyclone occurs, the preparations made at the coastal regions for receiving the cyclone based its earlier intensity levels, are rendered inadequate. About 1 in 5 cyclones in the Bay of Bengal experience this rapid intensification.

The difficulty in predicting a rapid intensification arises from the fact that it is an anomaly.

1.3 Anomaly

An anomaly may be defined as an occurrence, that deviates significantly from the expectation. Anomalies are characterized by their rare, random, unpredictable occurrences. An example of an anomaly, in the context of atmosphere, is the development of a cyclonic storm itself. Several depressions occur over the ocean / sea surface, but only few develop into a cyclonic storm, in a random and unpredictable fashion.

It is to be noted that, phenomenon that are considered as anomalies today, may eventually be better understood, and predicted, and hence may cease to be considered anomalies in the future.

For example, the solar eclipse, though rare, and unpredictable for early humans, has been now completely predictable, as the celestial motion became well understood. Hence, it ceased to be anomaly.

For the phenomenon of rapid intensification (RI), just like the case of other anomalies, there is no reliable technique at present, that can be used for prediction. The dynamics that contributes to RI is not yet well understood. In this thesis, we attempted to gain insight into the formation of rapid intensification (RI) of cyclones in the Bay of Bengal, using a data mining technique. This would eventually help in understanding of rapid intensification process and prediction.

CHAPTER 2

LITERATURE

2.1 Cyclone Forecast

Numerical simulations can be performed to predict both the track and intensity of a cyclone. But this is computationally expensive. This made the researchers gravitate to using Machine Learning (ML) / Deep Learning (DL) techniques. Even though training a deep learning network is expensive, making predictions using a trained model in real-time is computationally cheaper.

At present, track forecasts have achieved good accuracy, but intensity forecasts are still comparatively far behind. This is owing to the fact that the physical processes involved in the progress of a cyclone are not yet well understood, and also because of the occurrence of anomalous behaviours like rapid intensification in the cyclone.

2.1.1 RI prediction

Geng et al. [4] identified multiple attributes and levels present in a cyclone, and constructed an index based model for predicting the intensity change. They used 100 intensifying and 100 weakening tropical cyclone samples, and determined a set of seven core attributes. The attributes were given a projection weight, and based on a threshold value, they predicted if a cyclone would intensify or weaken with 90% accuracy.

[11] tackled the problem of binary classification of intensity change using a

decision tree method, and maximizing information gain based on attribute values. Their results showed the effectiveness of decision tree model, and thereby the potential for data mining techniques in cyclone forecasting.

Due to the presence of clouds, sea surface temperature cannot be measured by satellites, during a cyclone, though it is the most important factor for intensity prediction. [3] introduced a new attribute, and using it in a decision tree model resulted in a lower false alarm rate, and also higher probability of RI detection.

To classify an instance of the cyclone as an RI or non-RI in 24 h lead time, [7] used SVM, PCA, and feature selection, using the variable such as u , v , temperature, and moisture. Their study concluded that thermodynamic variables are more important than kinetic variables.

Based on the fact that scientific forecasts are only 15% better than climatological estimations, [8] used an ensemble of SVM, ANN, RF methods, and studied the cyclones in Atlantic basin. The results indicated a promising application of ensemble methods.

[2] Examined upper troposphere troughs for RI, non-RI cyclone instances. The troughs were classified into three clusters, using k-means clustering. They studied North Atlantic cyclones from 1989 to 2016. Results showed a strong correlation between cluster classification and rate of RI.

The challenging nature of intensity prediction comes from the fact that the relationships between physical processes in a cyclone are not yet well understood. To find these hidden relationships, [9] used association finding, employing 13 attributes. Favourable situations for intensification were discovered.

Variables that influence cyclone intensity are usually built by human experts. Using an XGBoost algorithm to classify and evaluate variable importance from information gain value, [10] were able to identify new predictors, using data from SHIPS database.

By visualizing reanalysis data of RI and non-RI cyclone instances in vertical planes, [1] were interested in determining the environmental factors that cause RI. They came up with an order of importance for key factors for RI formation.

Li et al. [6] implemented a shared nearest neighbour algorithm to identify clusters of environmental variables in the data, which can be used as predictors to rapid intensification. When known predictors were used for clustering, they showed clear difference in the RI and non-RI groups, validating the method.

CHAPTER 3

DATA

Atmospheric data comes in three forms,

- Satellite based monitoring data - For example, INSAT-3D is a geostationary satellite launched by ISRO in 2013. The region it monitors includes the region over Bay of Bengal.
- In-situ observations, in which the measurements are performed at the location where the measurements correspond to. For example, drifting buoys can measure sea level pressure / temperature at their respective locations.
- Numerical simulations, in which a known state of the atmosphere is initiated, and forecasting is done based on atmospheric physics.

Reanalysis method is a physics based numerical simulation of the atmosphere, that assimilates all the recorded global in-situ observations. For our work, reanalysis data is downloaded from European Centre for Medium Range Weather Forecasts (ECMWF), which is considered as the world's best global reanalysis data. The cyclonic systems in Bay of Bengal are tracked by Regional Specialized Meteorological Centre (RSMC), New Delhi, from 1982 to 2021. The tracked data includes the cyclones location, estimated central pressure, pressure drop, and maximum sustained surface wind.

If we define an RI-prior as an instance of a cyclone 12 h prior to rapid intensification, then each of these 40×40 arrays, is either an RI-prior or a non-RI-prior. An RI-prior stack, is a 3-D array, constructed by stacking d RI-priors. The 3-D

array can be viewed as a 2-D array of 40×40 cells, with each cell having a depth d . Before stacking, for each 40×40 array, the values in the array are standardized by subtracting the mean and dividing by the standard deviation.

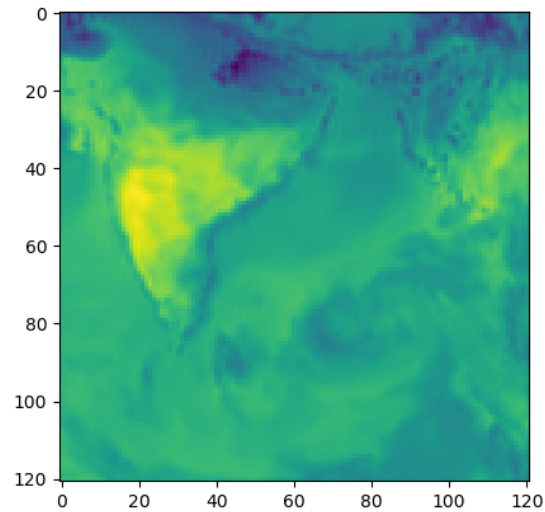


Figure 3.1: Snapshot of reanalysis data downloaded from ECMWF. A circulation can be noticed in the Bay of Bengal, at image coordinates (80, 80).

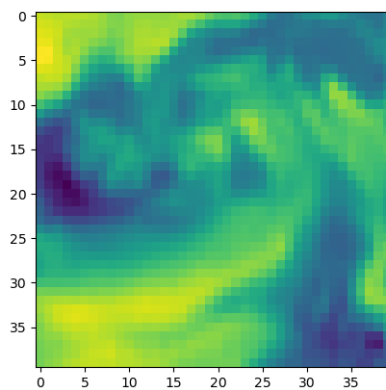


Figure 3.2: A 40×40 pixel area, zoomed in on the cyclonic system. 1 pixel corresponds to 0.25 degrees latitude / longitude.

CHAPTER 4

METHOD

On the 3-D data structure described, an algorithm is implemented for clustering the cells (points) each of depth d .

An overview of the Shared Nearest Neighbour (SNN) based clustering algorithm is presented below:

- Given n points v_1, v_2, \dots, v_n belonging to set V , build an $n \times n$ symmetric distance matrix, where each entry at (i, j) is the distance between the two corresponding points (v_i, v_j) .
- Based on the distances calculated above, k nearest neighbours are selected for each point.
- SNN density is initialized as 0 for each point. For each unordered pair of points $\{v_i, v_j\}$, if the number of shared k -nearest neighbours between them is more than eps , then increment the snn density by 1 for each v_i and v_j .
- A point is classified as a core point if its SNN density is atleast $Minpts$.
- Two core points belong to the same cluster if there are more than eps points common to both in their k - nearest neighbours.
- A point that is not a core point, but is an eps -nearest neighbour to a core point, is put in the cluster that the core point belongs to.
- A point that is neither a core point, nor is an eps nearest neighbour to a core point, is classified as noise.

In our case, each point, being of total depth d , has a value of the environment variable, at each level from 1 to d .

To get the distance between two points, cosine similarity is used as the distance metric in our implementation.

The purpose of the above algorithm is to identify clusters such that the points in a cluster have a similar distribution of the environmental variable. An SNN algorithm is suitable for clustering geographical variables, because, geographically, the neighbouring points have similar variable values. Especially, based on cyclone dynamics, we expect concentric clusters. The values of k , eps and $Minpts$ are empirically set to 50, 30 and 25 respectively.

We implemented a variation of SNN algorithm, known as Locality Sensitive Hashing (LSH) based SNN [5]. Using LSH, the n points are first hashed into bins with the expectation that similar valued points fall in the same bin. The distances are calculated only between those points that fall in the same bin. This reduces the complexity of SNN from $O(N^2)$ to $O\left(\frac{N^2}{b}\right)$, where b is the number of bins the values are hashed to.

Once the clusters are identified using the RI-prior stack for a given environmental variable, cluster mean values are calculated in each case of an RI-prior instance and a non-RI-prior instance.

A permutation test is performed, to check if any of these cluster mean values show different distributions in the RI-prior group of instances and non-RI-prior group of instances. If yes, then such a cluster, and the corresponding environment variable, forms a candidate precursor.

A correlation plot between the candidate precursor's mean value and the 24 h intensity change starting 12 h later, is plotted from historical data.

The higher the correlation magnitude, the more suitable is the candidate precursor as a potential precursor.

Mean value in RI-prior group of a potential precursor is taken as the threshold. Probability of occurrence of an RI, when that threshold is exceeded in the test data, is assigned the same value as that of historical probability, based on the correlation plot.

CHAPTER 5

RESULTS

First, the standard SNN output is visualized, to be able to compare the degree of approximation that will take place once an approximation algorithm is employed. The environmental variable used for clustering is sea level temperature.

When LSH is performed, using 4 buckets for hashing the cells, the total of 1600 cells were split into 58, 403, 1138, and 1 cells, thereby reducing the cost of computing the distance matrix by more than 30%. The clusters are shown in Figure 5.2.

For each cluster obtained using the LSH-SNN algorithm, the distribution of environmental variables is checked at in both RI-prior and non-RI-prior groups. Using a permutation test, three regions are found to be different among both the groups, with 0.05 level of significance. The regions are showed in Figure 5.3, 5.4, 5.5.

From historical data, the mean values in each of these clusters at a cyclonic instance, and the corresponding change in intensity starting 12 h later, expressed in terms of change in wind speed in the 24 h period, is plotted in Figures 5.6, 5.7 and 5.8.

The mean value of Region 1 displays strongest correlation with the intensity change. Surprisingly, though the two central clusters are very close by, one has a positive correlation, and the other has a negative correlation.

A threshold is defined as the historically average value of the mean value of

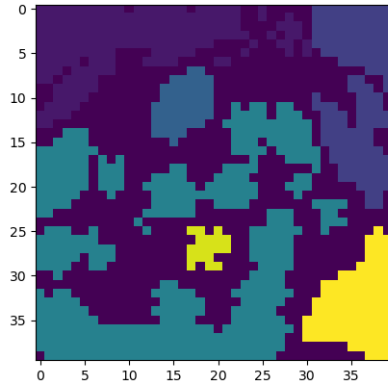


Figure 5.1: SNN output for RI-prior group, 12 h before the start of rapid intensification. A total of 7 clusters are visible. The clusters are concentric, as expected.

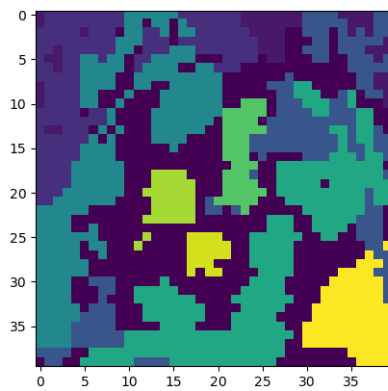


Figure 5.2: LSH - SNN output for RI-prior group, 12 h before the start of rapid intensification. A centroid, and concentric nature of the clusters is noticeable. 9 clusters can be counted.



Figure 5.3: Region 1, looking like an arc around the cyclonic centre.

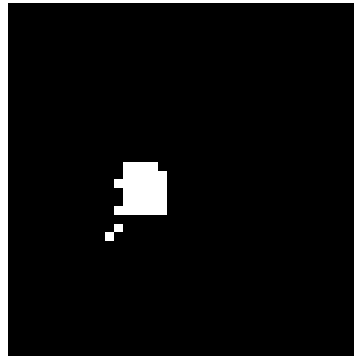


Figure 5.4: Region 2, resembling the centre of the cyclone.

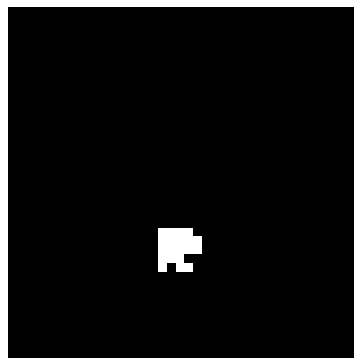


Figure 5.5: Region 3, also resembling the centre of the cyclone.

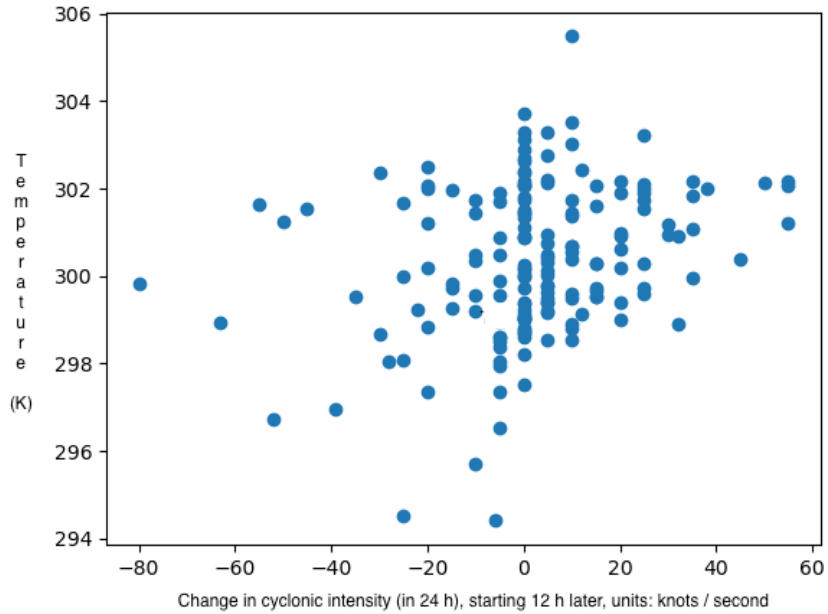


Figure 5.6: In Region 1, the correlation (Pearson’s coefficient) between the mean cluster value (sea level temperature), and intensity change, is 0.2778, which is very high.

Region	Probability of RI
1	8% vs 14%
2	6.5% vs 10.14%
3	5.4% vs 6%

Table 5.1: Probability of RI depending on whether the mean value in the region not exceeded vs exceeded the threshold value.

the environment variable in all the RI-prior cyclone instances. For Regions 1, 2, and 3, the thresholds are 301 K, 300 K, and 300 K respectively.

The RI probability is estimated from the correlation plot, when threshold is not exceeded vs when it is exceeded, and shown in Table 5.1.

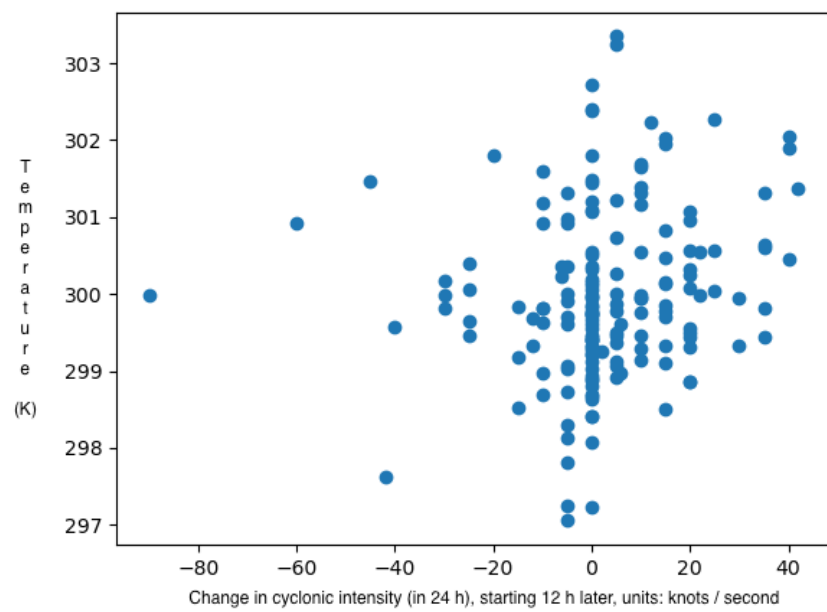


Figure 5.7: In Region 2, the correlation (Pearson's coefficient) between the mean cluster value (sea level temperature), and intensity change, is 0.16.

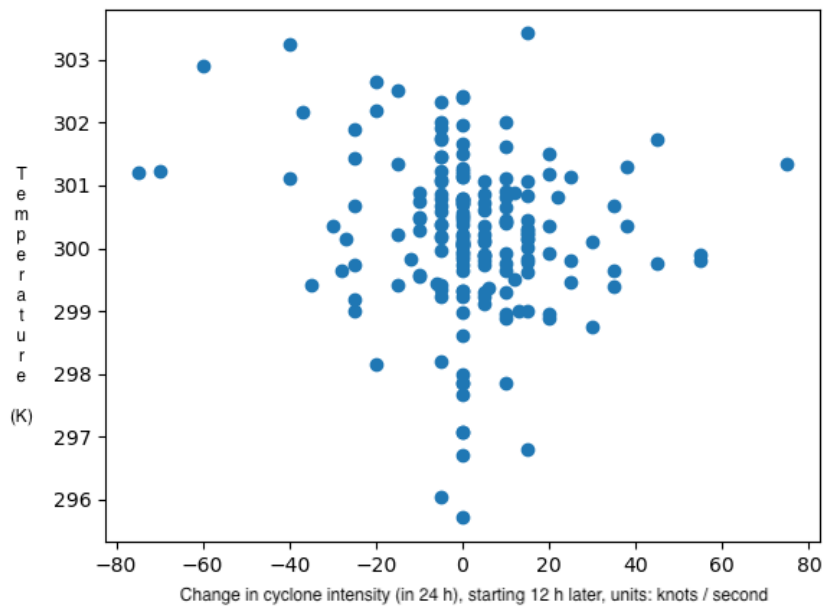


Figure 5.8: In Region 3, the correlation (Pearson's coefficient) between the mean cluster value (sea level temperature), and intensity change, is -0.15

CHAPTER 6

FUTURE WORK

The algorithm described in our method can be used by an atmospheric scientist, to check if and how any given environment variable can affect the development of a cyclone. The cluster regions that strongly correlate with the development phenomena can be classified as potential precursors.

Reanalysis modelling takes time to accommodate the observations, and generate simulations. Hence, rather than using reanalysis data, the proposed method can be implemented on satellite data, to be able to make quick real time predictions, more accurately.

In our study, only a single time instance (12 h prior to rapid intensification) is checked. Other time instances that are smaller (6 h) or higher (18 h) need to be tested, to see if better precursors could be identified. The method can also be applied for mining predictors for other weather anomalies, such as floods, or drought.

Only one hash function is used in our implementation of the approximation algorithm, to separate the cells into bins. Multiple hash functions can be used, and the best k nearest neighbours for a cell can then be selected from these multiple bins, resulting in a more accurate clustering, with only a marginal increase in computational requirement.

BIBLIOGRAPHY

- [1] Yao Chen, Si Gao, Xun Li, and Xinyong Shen. Key environmental factors for rapid intensification of the south china sea tropical cyclones. *Frontiers in Earth Science*, 8, 2021.
- [2] Michael S. Fischer, Brian H. Tang, and Kristen L. Corbosiero. A climatological analysis of tropical cyclone rapid intensification in environments of upper-tropospheric troughs. *Monthly Weather Review*, 147(10):3693 – 3719, 2019.
- [3] Si Gao, Wei Zhang, Jia Liu, I.-I. Lin, Long S. Chiu, and Kai Cao. Improvements in typhoon intensity change classification by incorporating an ocean coupling potential intensity index into decision trees. *Weather and Forecasting*, 31(1):95 – 106, 2016.
- [4] Huantong Geng, Jiaqing Sun, Wei Zhang, and Chao Huang. Study on index model of tropical cyclone intensity change based on projection pursuit and evolution strategy. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 145–150, 2015.
- [5] Sawsan Kanj, Thomas Bröls, and Stéphane Gazut. Shared nearest neighbor clustering in a locality sensitive hashing framework. *Journal of Computational Biology*, 25(2):236–250, 2018. PMID: 28953425.
- [6] Yun Li, Ruixin Yang, Hui Su, and Chaowei Yang. Discovering precursors to tropical cyclone rapid intensification in the atlantic basin using spatiotemporal data mining. *Atmosphere*, 13(6), 2022.
- [7] Andrew Mercer and Alexandria Grimes. Diagnosing tropical cyclone rapid intensification using kernel methods and reanalysis datasets. *Procedia Computer Science*, 61:422–427, 2015. Complex Adaptive Systems San Jose, CA November 2-4, 2015.
- [8] Andrew Mercer and Alexandria Grimes. Atlantic tropical cyclone rapid intensification probabilistic forecasts from an ensemble of machine learning methods. *Procedia Computer Science*, 114:333–340, 2017. Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA.
- [9] Jiang Tang, Ruixin Yang, and Menas Kafatos. Data mining for tropical cyclone intensity prediction. *Sixth Conference on Coastal Atmospheric and Oceanic Prediction and Processes*, 01 2005.

- [10] Yijun Wei and Ruixin Yang. An advanced artificial intelligence system for investigating tropical cyclone rapid intensification with the ships database. *Atmosphere*, 12(4), 2021.
- [11] Wei Zhang, Si Gao, Bin Chen, and Kai Cao. The application of decision tree to intensity change classification of tropical cyclones in western North Pacific. *Geophysical Research Letters*, 40(9):1883–1887, May 2013.