THE UNIVERSITY OF HONG KONG

DEPARTMENT OF COMPUTER SCIENCE

COMP4802 Extended Final Year Project

CAES9542 Technical English for Computer Science

Progress Report 1

Student Name: FONG Kwan Ching

UID: 3035564902

Project title: A Study on Forecasting Tropical Cyclone Properties

Supervisor: Dr Beta C. L. Yip

Date Submitted: 26 October 2021

Summary

There is an absence of statistical-dynamical forecasting products that assess the probabilities of Hong Kong being affected by tropical cyclones and evaluate the corresponding impact, despite their reliability and intuitiveness to the general public. This project envisages an ensemble forecast, in which decision trees and Monte Carlo simulations are employed to analyze historical warning signal records, tropical cyclone best track data and synoptic meteorological analysis data archives. This progress report details the design choices and initial data preparation tasks done, such as the processing of the Hong Kong Observatory warning issuance database. The upcoming months will see the processing of the meteorological and best track datasets and the development of a baseline model. This project may open new opportunities for Hong Kong-centric probabilistic statistical-dynamical forecasts.

Acknowledgements

I would like to take this opportunity to thank Dr Beta Yip and Dr Ping-Wah Li of the Hong Kong Observatory for guiding me during the project’s inception.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full-Form
CLIPER	Climatology and Persistence
GFS	Global Forecasting System
HKO	Hong Kong Observatory
IBTrACS	International Best Track Archive for Climatological Stewardship
JTWC	Joint Typhoon Warning Center
NWP	Numerical weather prediction
TC	Tropical Cyclone

Chapter 1: Introduction

This chapter summarizes the background of this final year project ("this project"), describes the project objectives and outlines the current progress.

1.1 Background

The term tropical cyclone (TC) refers to storms forming over tropical seas, the stronger categories of which are called typhoons in Eastern Asia [1]. On average, six TCs affect Hong Kong each year [2, pp.34-35]. The adverse weather TCs bring about disrupt Hong Kong citizens' daily activities and result in casualties [3]. Thus, an accurate TC forecasting method and a reliable TC warning system are necessary, to inform the general public of an incoming TC's threat.

Countless techniques were developed in the past several decades to forecast TCs. There are statistical methods that identify patterns in historical data, such as Climatology and Persistence (CLIPER) [4], and dynamical methods that simulate the entire atmosphere using physics and up-to-date observations. The latter, also called numerical weather prediction (NWP), is currently the state-of-the-art means [5]. In recent years, interest arose in combining statistical and dynamical methods to thoroughly harness the advantages of both approaches. An example is [6]. Machine learning techniques also greatly progressed and became a viable statistical means for TC forecasting [7]. By taking the inherent stochasticity of TCs into account, the resultant probabilistic forecasts can be more valuable. They also provide an intuitive way for the general public to assess threats [8].

Probabilistic forecasts designed for Hong Kong exist already. An unofficial organization provides a TC warning signal probabilities forecast [9], which uses a statistical model solely relying on the trajectories of historical TCs ("best track data"). However, there is an absence of statistical-dynamical models that evaluate the probabilities of certain TC warning signals being hoisted.

1.2 Project Objectives

This project focuses on building a statistical-dynamical TC forecasting tool to assess TC strike probabilities for Hong Kong. The tool predicts four values corresponding to minimal, limited, substantial damage and direct TC strike. The first three predictands correspond to TC warning signals number 1, 3, and 8-10 respectively; the last for TCs passing through a 100km radius of Hong Kong. The predictions are valid for up to three days.

This project intends to fill the research gap, such that other statistical-dynamic models could be devised to forecast TC impacts, e.g. by also considering storm surges. This project is exploratory, and the analyses and methodology iteratively deepen.

1.3 Current Status

Thus far, the project has made limited progress. The HKO TC warning signals database was obtained and cleaned, whereas the processing of TC best track statistics is in progress. The acquisition and treatment of the dynamical data have not yet started.

The remainder of this report is organized as follows: Chapter 2 explains the methodology and discusses the particulars of the models and the data used. Chapter 3 describes the current progress of the project, compares it to the project timeline, and outlines the future tasks. Chapter 4 summarizes the report.

Chapter 2: Project methodology

This chapter discusses the high-level design, input data considerations, and modelling methods to build the aforementioned forecasting tool.

2.1 High-Level Design and Approaches

A variant of the statistical-dynamical approach is used. Historical statistics, namely TC warning issuance records and best track data, and synoptic meteorological data obtained from dynamical NWP models are analyzed together. This approach is chosen because it covers both the outcomes (the TC tracks and the signals' issuance) and their underlying factors (the atmospheric situation) so that the forecasts made can be more realistic.

A custom-built control is needed for the experiments because there is no directly comparable prior research to serve as a benchmark. To mimic rigorous research [10], a model similar to CLIPER will be built to act as the baseline for comparison.

Two other models will be built to produce the actual forecasts as an ensemble. It makes the final forecast more robust and comprehensive by considering each ensemble member's output in a weighted sum. Section 2.3 details the proposed methodologies to experiment building ensemble members with.

2.2 Dataset Selection and Data Preparation

This project involves three data sources: One for TC signal issuance records, which is available online at the HKO website [11]; one for best track data, which numerous authorities provide; and one for synoptic meteorological data, which are generated from NWP analyses at various institutions. The selection criteria for best track datasets include accuracy, time coverage, and ease of use. Similarly, the selection criteria for NWP datasets include comprehensiveness, resolution, time coverage, and ease of use.

After confirming the dataset choices, the data will be retrieved and processed to give an operational dataset, which the models will rely on. To prevent the models from analyzing irrelevant information, the source data must be cleaned. Complex NWP data must also be simplified because there will be thousands of data points in the NWP data. These must be summarized into only a few numbers to permit statistical analysis. Processing NWP data is the primary difficulty of this whole process (see section 3.3.3).

2.3 Proposed Methods for the Forecasting Models

The baseline model against which the others are evaluated will be a multilayer perceptron. The multilayer perceptron is a basic machine learning technique, where a simple neural network learns directly on the data [12]. This method is selected because of its simplicity and genericness, such that the effectiveness of more sophisticated and specifically designed models can be justified. To better mimic simplistic CLIPER models, this baseline model will only consider best track data and signal issuance statistics, analogous to how the 1972 CLIPER analyzed only best track data [4].

The first statistical-dynamical model will employ decision trees. This machine learning technique sorts input data into classes using binary rules [13], similar to how the rule “marks < 50” can separate exam scripts into “F-grade” and “not F-grade” groups. Decision trees are favoured because they have proven effective at TC forecasting [7] and can be easily made robust by methods such as pruning [14]. Therefore, decision trees are a valuable method to experiment with.

The second method is to perform Monte Carlo simulations. These simulations are generated based on the probability distribution of the input data [15]. This can be understood as throwing an unfair die hundreds of times to observe the patterns in its behaviour. This method is also known to be suitable for TC forecasts, as illustrated by [16-17]. Furthermore, a probability can be easily computed by counting the number of "hits" among the simulations. As a result, the Monte Carlo method is suitable for this project.

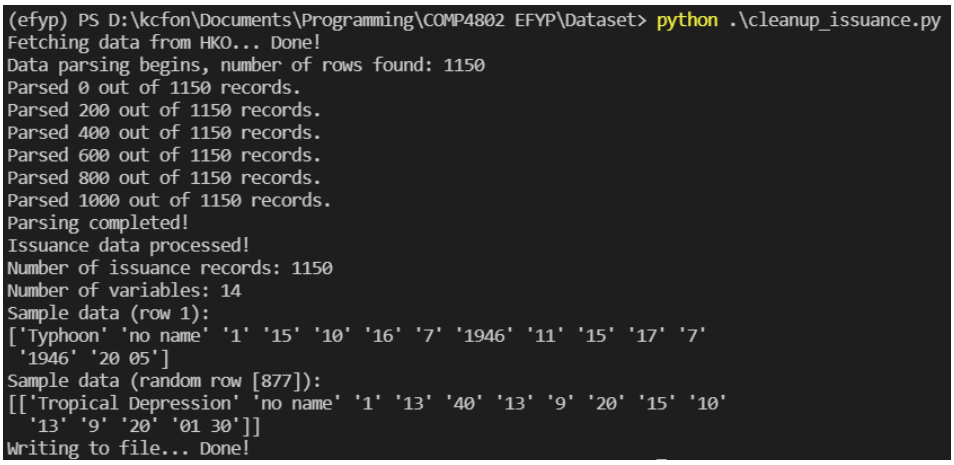
Chapter 3: Project Progress

This chapter reports the current progress of the project, including work completed, tasks in progress, and next steps as illustrated via a timeline.

3.1 Details of Work Completed and in Progress

3.1.1 Conversion of HKO TC Warning Signal Records

The HKO TC warning signal issuance records have been retrieved and grouped into a machine-readable dataset. The records are available on the HKO website [11] as a human-readable webpage instead of being in machine-friendly formats. Therefore, a Python script was written to fetch the webpage and extract the records, before transforming them from text to the correct data types. Finally, the converted data were written to a local file for later reuse.



```
(efyp) PS D:\kcfon\Documents\Programming\COMP4802 EFYP\Dataset> python .\cleanup_issuance.py
Fetching data from HKO... Done!
Data parsing begins, number of rows found: 1150
Parsed 0 out of 1150 records.
Parsed 200 out of 1150 records.
Parsed 400 out of 1150 records.
Parsed 600 out of 1150 records.
Parsed 800 out of 1150 records.
Parsed 1000 out of 1150 records.
Parsing completed!
Issuance data processed!
Number of issuance records: 1150
Number of variables: 14
Sample data (row 1):
['Typhoon' 'no name' '1' '15' '10' '16' '7' '1946' '11' '15' '17' '7'
 '1946' '20 05']
Sample data (random row [877]):
[['Tropical Depression' 'no name' '1' '13' '40' '13' '9' '20' '15' '10'
  '13' '9' '20' '01 30']]
Writing to file... Done!
```

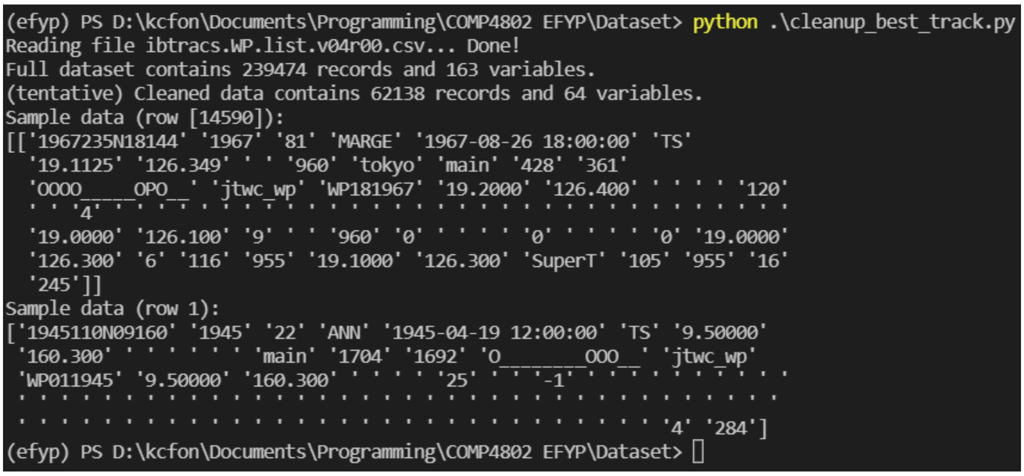
Figure 3.1: The Output of the warning signal issuance records processing program

Figure 3.1 shows the output of the Python script. The data extracted were parsed into the correct data types, and statistics alongside some sample data were displayed for reference before the converted data was saved to a file. As shown in the samples, TCs are identified by names instead of a unique identifier. This had consequences for the best track dataset selection process below.

3.1.2 Selection and Processing of Best Track Dataset

Compared to other datasets, the Joint Typhoon Warning Center (JTWC) best track dataset [18] provides much more information about the caveats of the dataset, such as error ranges and verification techniques used. The crucial flaw of the JTWC dataset is that TC names are not included, making it difficult to match a TC with the corresponding signals issued. Fortunately, the *International Best Track Archive for Climatological Stewardship* (IBTrACS) dataset [19] contains the entirety of the JTWC dataset and additional variables including TC names. Because of its comprehensiveness and easier usage, the IBTrACS dataset was then chosen.

Another Python script was written to perform data type conversions and cleanups. This was necessary because some text-based raw data must be converted into numbers and the dataset had more variables than needed. This cleanup is currently in progress.



```
(efyp) PS D:\kcfon\Documents\Programming\COMP4802 EFYP\Dataset> python .\cleanup_best_track.py
Reading file ibtracs.WP.list.v04r00.csv... Done!
Full dataset contains 239474 records and 163 variables.
(tentative) Cleaned data contains 62138 records and 64 variables.
Sample data (row [14590]):
[['1967235N18144' '1967' '81' 'MARGE' '1967-08-26 18:00:00' 'TS'
  '19.1125' '126.349' ' ' '960' 'tokyo' 'main' '428' '361'
  'OOOO_____OPO__' 'jtwc_wp' 'WP181967' '19.2000' '126.400' ' ' ' ' '120'
  ' ' '4' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' '
  '19.0000' '126.100' '9' ' ' '960' '0' ' ' ' ' '0' ' ' ' ' '0' '19.0000'
  '126.300' '6' '116' '955' '19.1000' '126.300' 'SuperT' '105' '955' '16'
  '245']]
Sample data (row 1):
['1945110N09160' '1945' '22' 'ANN' '1945-04-19 12:00:00' 'TS' '9.50000'
 '160.300' ' ' ' ' ' ' 'main' '1704' '1692' 'O_______OOO__' 'jtwc_wp'
 'WP011945' '9.50000' '160.300' ' ' ' ' '25' ' ' '-1' ' ' ' ' ' ' ' ' ' ' '
 ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' '
 ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' '4' '284']
(efyp) PS D:\kcfon\Documents\Programming\COMP4802 EFYP\Dataset> 
```

Figure 2.3: The Output of the provisional best track data cleanup program

Figure 2.3 shows the statistics of the dataset after basic cleaning which eliminated some irrelevant variables. A mask was applied to screen out the JTWC data, identified by the text “jtwc_wp”, and 99 unneeded variables were removed. Note that the TC names (‘MARGE’ and ‘ANN’) are present in the two sample records shown.

3.1.3 Selection of Dynamical Data Source

When comparing the *ECMWF Reanalysis v5 (ERA5)* dataset [20] to the NCEP FNL dataset [21], it is noted that the NCEP FNL dataset is built using the analyses of the Global Forecasting System (GFS), the NWP system used by US authorities. As a result, forecasts can be made for new TCs using the latest GFS analyses. In contrast, the ERA5 dataset is only updated after long delays and it is impossible to obtain the latest analyses for operational forecasts. Therefore, the NCEP FNL dataset is deemed more useful and is thus selected.

3.2 Project Timeline

The following table is the project schedule.

Date	Task	Status
October 2021	High-level design, Dataset acquisition, Data processing	Design: Finished Dataset acquisition: In progress Data processing: in progress
November 2021	Construction and evaluation of the baseline model	To do
December 2021 - January 2022	Building and evaluating ensemble member 1, Interim report	To do
February 2022	Construction and evaluation of ensemble member 2	To do
March 2022 – April 2022	Building and evaluating the ensemble forecast, Final report	To do

Table 3.1: Project Schedule and Current Status

As shown in the table, data processing should be finished by the end of October 2021. However, section 3.1 shows that the work on the datasets is behind schedule because the NWP data have not been processed yet, which is unfavourable.

3.3 Future Work and Expected Difficulties

3.3.1 Processing Best Track Data and Dataset Construction

As mentioned in section 3.1.2, the processing of the IBTrACS dataset is still in progress. More variables may be removed, and the data types have not been converted yet. After that, the combined dataset containing both the best track records and the HKO issuance records will be prepared for the baseline model's use, by matching records using dates and TC names.

The primary difficulty will be to keep track of the numerous variables present in the best track dataset. It is expected that about 30 variables, each identified by a column number instead of a name, will remain after processing, thus identifying variables for removal, conversion or matching will be difficult. This may be overcome by writing helper programs to manage variable names and column numbers.

3.3.2 Development of the Baseline Model

The natural next step is to develop the baseline model. As this involves machine learning (see section 2.3), it will be most appropriate to employ the department's GPU farm as the computing resource. Therefore, securing a new resource allocation will be necessary (see also next section). The model itself will be built using the PyTorch package because it supports GPU acceleration well [22] and is easy to use. It is expected that the development will be a smooth process.

3.3.3 NWP Data Processing

The NCEP FNL dataset needs to be retrieved and processed. However, it is more than 600GB in size and will have to be stored on the department's GPU farm. Therefore, the new computing resource must have sufficient storage space.

After that, the correct predictors need to be selected from the dataset. Not all variables will be necessary, and those who are must be compressed into only a few values (see section 2.2). The compression techniques will have to be specifically chosen for the variable in question. For instance, geopotential height data can be summarized using indices [23]. To tackle this issue, the literature on the topic will be studied.

Chapter 4: Conclusion

This project explores the means to develop forecasting models that evaluate the likelihood of TCs affecting Hong Kong. The models work together as an ensemble and produce probabilistic forecasts in categories corresponding to different levels of impact.

At the moment, there is limited progress, as the input data is still being processed. The TC warning signal issuance records of the HKO have been obtained and the conversion and cleanup of the IBTrACS best track dataset are still in progress. The NWP data from the NCEP FNL dataset has yet been handled because of a lack of computing resources. It is observed that the data collected are complex with large volumes and numerous variables to filter out. Therefore, thoroughly studying the documentation and relevant literature will be necessary.

This project, if successful, may open new grounds for probabilistic statistical-dynamical TC forecasts designed for Hong Kong, while serving as an additional means to evaluate TC threat for the general public. However, this project uses vague definitions for TC impact levels, which are only loosely tied to TC warning signals. The forecasts could be more valuable if other factors such as storm surge, rainfall and exact wind speeds are considered to better characterize the levels of damage.

References

[1] "Hurricanes Frequently Asked Questions." NOAA's Atlantic Oceanographic and Meteorological Laboratory. https://www.aoml.noaa.gov/hrd-faq/#what-is-a-hurricane (accessed Sep. 30, 2021).

[2] "Tropical Cyclones in 2020," HKO, Hong Kong, Jul. 2021. Accessed Sep. 30, 2021. [Online] Available: https://www.hko.gov.hk/en/publica/tc/files/TC2020.pdf.

[3] "Social and Economic Impact of Tropical Cyclones." HKO. https://www.hko.gov.hk/en/informtc/economice.htm (accessed Sep. 30, 2021).

[4] C. J. Neumann, "An Alternative to the HURRAN (Hurricane Analog) Tropical Cyclone Forecast System," National Hurricane Center, Miami, FL, USA, NOAA Tech. Memo. NWS SR-62, Jan. 1972. [Online] Available: https://repository.library.noaa.gov/view/noaa/3605.

[5] R. L. Elsberry, "Advances in research and forecasting of tropical cyclones from 1963–2013," *Journal of the Korean Meteorological Society* (한국기상학회지한국기상학회지), vol. 50, no. 1, pp. 3-16, 2014. DOI: 10.1007/s13143-014-0001-1.

[6] J. A. Knaff and C. R. Sampson, "Southern hemisphere tropical cyclone intensity forecast methods used at the Joint Typhoon Warning Center, Part II: statistical-dynamical forecasts," *Australian Meteorological and Oceanographic Journal*, vol. 58, no.1, pp. 9-18, Mar. 2009. DOI: 10.22499/2.5801.002.

[7] R. Chen, W. Zhang and X. Wang, "Machine Learning in Tropical Cyclone Forecast Modeling: A Review," *Atmosphere*, vol. 11, no. 7, pp. 676, Jun. 27, 2020. DOI: 10.3390/atmos11070676.

[8] J. Jarrell and S. Brand, "Tropical Cyclone Strike and Wind Probability Applications," *Bulletin of the American Meteorological Society*, vol. 64, no. 9, pp.

1050-1056, Sep. 1983. [Online] Available: https://www.jstor.org/stable/26223426.

[9] "Model-Based TC Signal Probabilities – Methodology." Hong Kong Weather Watch. http://www.hkww.org/weather/signalprob/method.html (accessed Oct. 24, 2021).

[10] D. A. Zelinsky and R. J. Pasch, "Tropical Cyclone Track Prediction." Severe Weather. http://severeweather.wmo.int/TCFW/RAIV_Workshop2021/20_TC-TrackForecasting_DaveZelinsky_Richardpasch.pdf (retrieved Sep. 27, 2021).

[11] "HKO Warnings and Signals Database." HKO. https://www.hko.gov.hk/cgi-bin/hko/warndb_e1.pl?opt=1&sgnl=1.or.higher&start_ym=194601&end_ym=202110 (retrieved Oct. 25, 2021).

[12] "Multilayer Perceptron." DeepAI. https://deepai.org/machine-learning-glossary-and-terms/multilayer-perceptron (retrieved Oct. 3, 2021).

[13] "Decision Tree." DeepAI. https://deepai.org/machine-learning-glossary-and-terms/decision-tree (retrieved Oct. 3, 2021).

[14] S. Kumar, "3 Techniques to Avoid Overfitting of Decision Trees." Towards Data Science. https://towardsdatascience.com/3-techniques-to-avoid-overfitting-of-decision-trees-1e7d3d985a09 (retrieved Oct. 3, 2021).

[15] "Monte Carlo Simulation." IBM Cloud Learn Hub. https://www.ibm.com/cloud/learn/monte-carlo-simulation (retrieved Oct. 3, 2021).

[16] M. DeMaria, J. A. Knaff, R. Knabb, C. Lauer, C. R. Sampson and R. T. DeMaria, "A New Method for Estimating Tropical Cyclone Wind Speed Probabilities," *Weather and Forecasting*, vol. 24, no. 6, pp. 1573-1591, Dec. 1, 2009. DOI: 10.1775/2009WAF2222286.1.

[17] X. Xie, B. Xie, J. Cheng, Q. Chu and T. Dooling, "A simple Monte Carlo method for estimating the chance of a cyclone impact," *Natural Hazards*, vol. 107, pp. 2573-2582, Jan. 13, 2021. DOI: 10.1007/s11069-021-04505-2.

[18] *Western North Pacific Ocean Best Track Data*, JTWC, n.d. [Online]. Available: https://www.metoc.navy.mil/jtwc/jtwc.html?western-pacific.

[19] *International Best Track Archive for Climate Stewardship* (IBTrACS), World Data Center for Meteorology, Asheville, n.d. doi:10.25921/82ty-9e16.

[20] *ECMWF Reanalysis v5 (ERA5)*, European Center for Medium-Range Weather Forecasts, n.d. [Online] Available: https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5.

[21] *NCEP FNL Operational Model Global Tropospheric Analyses*, continuing from July 1999, National Centers for Environmental Prediction, National Weather Service, National Oceanographic and Atmospheric Administration, U.S. Department of Commerce, 2000, DOI: 10.5065/D6M043C6.

[22] "pytorch/pytorch: Tensors and Dynamic neural networks in Python with strong GPU acceleration." GitHub. https://github.com/pytorch/pytorch (retrieved Oct. 25, 2021).

[23] J. Nie, P. Liu and C. Zhao, "Research on Relationship between Various Indexes of the Western North Pacific Subtropical High and Summer Precipitation in Eastern China," *Chinese Journal of Atmospheric Sciences* (in Chinese), vol. 45, no. 4, pp. 833-850, Jul. 2021. DOI: 10.3878/j.issn.1006-9895.2009.20160.