
# THE UNIVERSITY OF HONG KONG

# DEPARTMENT OF COMPUTER SCIENCE

## COMP4802 Extended Final Year Project

## CAES9542 Technical English for Computer Science

## Progress Report 2

Student Name: FONG Kwan Ching

UID: 3035564902

Project title: A Study on Forecasting Tropical Cyclone Properties

Supervisor: Dr Beta C. L. Yip

Date Submitted: 1st December 2021

# Summary

Every year, several tropical cyclones affect Hong Kong and bring about adverse weather, making accurate and informative forecasts necessary. 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 informativeness to the general public. This project builds 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 describes the tasks done so far, namely the processing of best track data and warning signals database and the development of a baseline model. The progress has been satisfactory, and the upcoming months will see the processing of the dynamical meteorological data and the development of the first ensemble member. If successful, 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. I would also like to thank Ms Grace Chang of the CAES for correcting the previous progress report.

# Table of Contents

# List of Figures



# List of Tables



# List of Abbreviations

| Abbreviation | Full-Form |
|---|---|
| **DT** | Decision tree |
| **GFS** | Global Forecasting System |
| **HKO** | Hong Kong Observatory |
| **IBTrACS** | International Best Track Archive for Climatological Stewardship |
| **MLP** | Multilayer perceptron |
| **NWP** | Numerical weather prediction |
| **RF** | Random forest |
| **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 disrupts 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 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 [4]. Recently, interest arose in combining both methods to thoroughly harness their respective advantages. Furthermore, by taking the inherent stochasticity of TCs into account, probabilistic forecasts help the general public intuitively assess threats [5].

Probabilistic forecasts designed for Hong Kong exist already. An unofficial organization provides a TC warning signal probabilities forecast [6], 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 intends to fill the research gap by 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 levels of 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.

## 1.3 Current Status

Thus far, the project has made acceptable progress. The HKO TC warning signals database and TC best track statistics have been acquired and treated; a baseline model has also been developed. The time-consuming process of obtaining and processing the dynamical data have not yet started, which may cause delays to the schedule.

## 1.4 Report Outline

The remainder of this report is organized as follows: Chapter 2 explains the design and structure of the forecasting tool, the selection of data sources, and briefly introduces the planned modelling techniques. Chapter 3 describes the tasks done, such as dataset preparation and baseline model development, compares the progress to the project schedule, and outlines the future tasks. Chapter 4 concludes 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

A 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 allows direct modelling of the atmospheric situation, leading to more realistic forecasts.

Two statistical-dynamical models will be built to forecast as an ensemble. It makes the final forecast more robust and comprehensive by considering each ensemble member's output in a weighted sum. To justify the benefit of including dynamical data, a baseline model will first be built without such data to act as the experiment control. Section 2.3 details the proposed methodologies to experiment building ensemble members with.

The development workflow is simple: Firstly, the data sources (see the next section) are identified. Then, the data are obtained and preprocessed to construct operational datasets on which the models rely. Next, the models are built using the datasets as training data. Finally, the models are individually evaluated.

## 2.2 Data Source Selection and Data Preparation

To build the abovementioned models, three data sources are needed: One for TC signal issuance records, which is available online at the HKO website [7]; 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.

Data from the chosen data sources will be cleaned and transformed, then an operational dataset for the models' use will be constructed. During the transformation process, complex NWP data must be simplified into only a few numbers to permit statistical analysis. This is because there will be unnecessarily many data points in the NWP data. Processing NWP data is the primary difficulty of the whole process (see section 3.3.1).

## 2.3 Proposed Methods for the Forecasting Models

The baseline model and the first ensemble member will be random forest (RF) classifiers. These two models share the same methodology so that the advantages of using dynamical data can be immediately apparent.

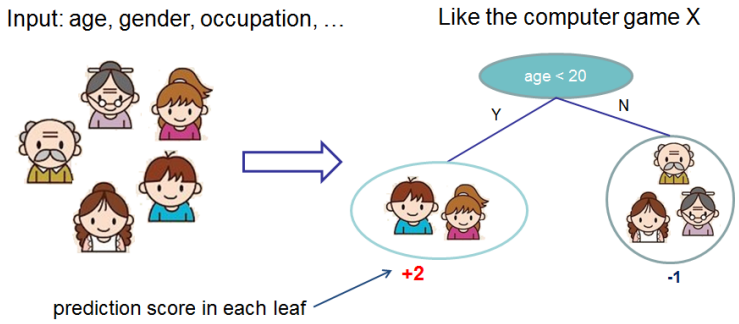



*Figure 2.1: An example of a decision tree (DT).[8]*

Random forest (RF) classifiers contain numerous decision trees, which individually use binary rules to classify data [8, 9]. For example, Figure 2.1 shows that age to reason about whether a person will like some computer game. By combining hundreds of DTs to consider more variables, a comprehensive RF model can be built.

The second ensemble member will use Monte Carlo simulations. These simulations are generated based on the probability distribution of the input data [10]. This can be understood as throwing an unfair die hundreds of times to observe its behaviour. Monte Carlo simulations are proven suitable for TC forecasting [11] and allow computation of probabilities. Therefore, Monte Carlo simulations are appropriate 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 Processing of TC Signal Records and Best Track Data

The HKO TC warning signal issuance records were retrieved and processed in early October 2021. The records were downloaded from the HKO website [7] as a webpage and a script was written to extract the records and convert them from text into correct data types. As the HKO uses names to identify TCs, this greatly limited the choice of best track data sources because many do not include TC names.

The *International Best Track Archive for Climatological Stewardship* (IBTrACS) dataset [12] was chosen as the best track data source because it contains both TC names and the well-validated Joint Typhoon Warning Center best track dataset [13]. When the IBTrACS dataset was processed in late October, the main difficulty encountered was to keep track of the numerous variables present, most of which are unnecessary. Documentations had to be made and updated continuously to help identify variables.

The two datasets were then combined in early November to give a “baseline dataset” the baseline model uses. This process was challenging because unhandled null values may crash the model training programs later on. This was overcome by searching for null values in the baseline dataset after its construction so that these irregularities could be revealed and fixed.

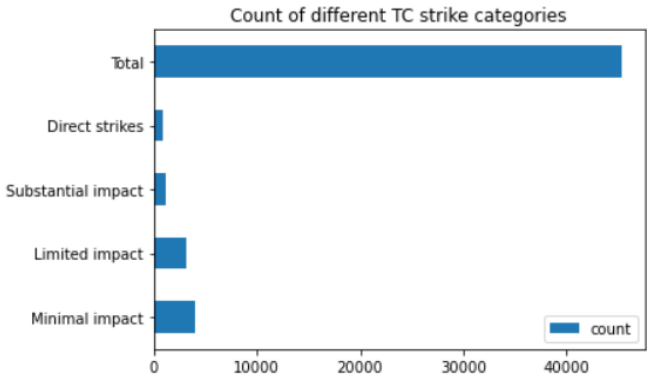



*Figure 3.1: Count of different TC impact categories in the baseline dataset*

In Figure 3.1, the number of data records in the baseline dataset that correspond to each predictand category is shown, alongside the total data record count. It is immediately evident that there are few positives (TC led to some impact) when compared to the huge number of negatives (no impact). As a result, whether the models can correctly predict true positives (see section 3.1.3) will be very important, because this is what the general public should be accurately informed about.

### 3.1.2 Selection of Dynamical Data Source and Preliminary Analyses

When comparing the *ECMWF Reanalysis v5 (ERA5)* dataset [14] to the NCEP FNL dataset [15], it is noted that the latter is built using 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, making it unable to support operational forecasts. Therefore, the NCEP FNL dataset is deemed more useful and was thus selected in October.

Currently, the documentation of the NCEP FNL dataset is being studied to understand its data formats and exact contents. It was found that selective downloads are supported and the whole 600GB dataset need not be downloaded. Therefore, the processing and storage of the required subset of the dataset are simpler than previously expected.

However, further analyses of the data content and literature reviews about summarization methods (see also section 2.2) have not yet started.

### 3.1.3 Baseline Model Development

The baseline model specified in section 2.1 was developed in late November. This was completed in advance because the baseline model does not depend on dynamical data. The initial plan was to employ multilayer perceptrons (MLPs), a simple and robust neural network [16], however, the model could not be fit to the data.

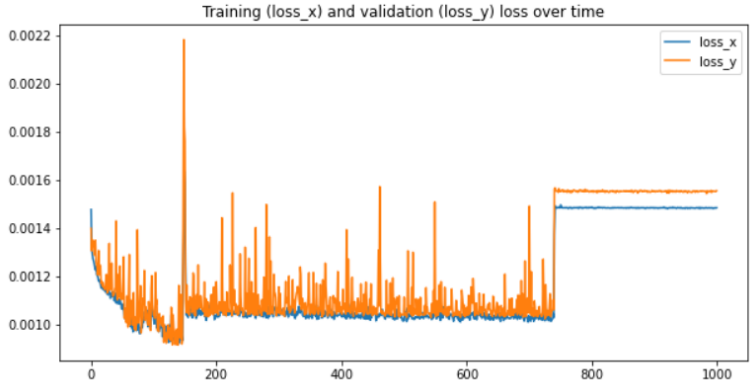
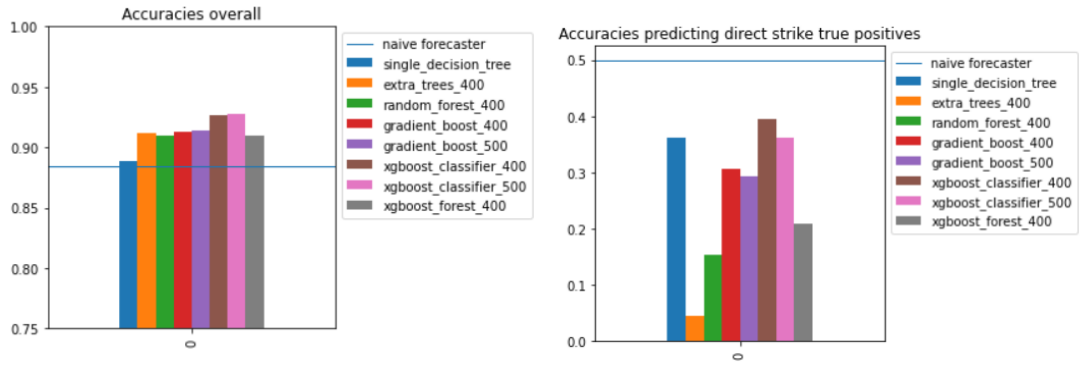



*Figure 3.2: The change in loss over time during the MLP model training.*

Loss is a measurement of error and is used to track a model's training performance. Figure 3.2 shows the change in training and validation losses of the MLP. They decreased initially, but then sharply increased to plateau at a high value while the accuracy remained at zero. This indicated that MLPs are inept at modelling the TC data. As an alternative, DTs and RFs were tried.



*Figures 3.3a (left) and 3.3b (right): Accuracies of DTs and RFs*

Usable models were successfully built using DTs and several variations of RFs. In Figure 3.3a, the overall forecasting accuracies of these models were shown. Compared to a naïve forecaster that always predicts negatives, all models proved superior. Figure 3.3b demonstrates that none of them warns about true positives in the rarest category better than a naïve forecaster that always makes a random decision. However, in both cases, the "xgboost_classifier" models have the best performance. This RF variant will likely also perform well as an ensemble member with dynamical data.

## 3.2 Project Timeline

The latest progress necessitates an update to the project schedule because it was found that model development took less time than expected whereas dataset preparation was time-consuming. The revised project schedule is as follows:

| **Date** | **Task** | **Status** |
|---|---|---|
| October – November 2021 | High-level design,<br>Baseline dataset preparation,<br>Baseline model development | High-level design: completed,<br>Baseline dataset: completed,<br>Baseline model: completed |
| December 2021 – January 2022 | NWP data analysis,<br>NWP data processing,<br>Combined dataset preparation | NWP data analysis: in progress,<br>NWP data processing: to do,<br>Combined dataset: to do |
| February 2022 | First ensemble member development and evaluation | To do |
| March 2022 | Second ensemble member development and evaluation | To do |

| | | |
|---|---|---|
| April 2022 | Ensemble forecast development,<br>Final report | To do |

*Table 3.1: Revised Project Schedule and Current Status*

As shown in the table, the project has made satisfactory progress because the baseline model has been completed and the upcoming NWP data-related tasks have commenced. However, because further literature reviews are needed to guide data processing (see section 3.3.1), the completion of the combined dataset may still be delayed. This will be unfavourable because it will compress the time available for model development.

## 3.3 Future Work and Expected Difficulties

### 3.3.1 NWP Data Processing

The NCEP FNL dataset needs to be retrieved and processed to select the correct predictors, which will then be added to the baseline dataset to give the combined dataset that the ensemble relies upon. The desirable variables identified must be compressed into only a few values (see section 2.2). The compression techniques will have to be specifically chosen for the variables. For instance, geopotential height data are best summarized as indices [17]. To tackle this issue, the literature on the topic will be studied.

### 3.3.2 Development of the First Ensemble Member

After the completion of a combined statistical-dynamical dataset, the ensemble member models are to be built next. Based on the experience gained developing the baseline model, RFs will be tested first, because they are straightforward to build, and the results can immediately be compared to the baseline (see section 2.3). It will be difficult to optimize the model because of the exceedingly many possibilities. Techniques such as

grid search may be used to automatically explore those opportunities.

# 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. RFs and Monte Carlo simulations are the primary modelling techniques employed.

The project has made acceptable progress thus far. The HKO warning records and the best track data have been processed and combined, and a baseline model has been successfully built despite initial setbacks. The project is still on a tight schedule, however, because the forthcoming NWP data processing tasks will be time-consuming, potentially delaying the development of the ensemble members.

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. Further research taking other factors such as storm surge, rainfall and exact wind speeds into consideration may produce more valuable forecasts by better characterizing damage levels.

# 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] 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.

[5] 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.

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

[7] "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).

[8] "Introduction to Boosted Trees." xgboost 1.5.1 documentation. https://xgboost.readthedocs.io/en/stable/tutorials/model.html (retrieved Nov. 30, 2021).

[9] "Decision Tree." DeepAI. https://deepai.org/machine-learning-glossary-and-

terms/decision-tree (retrieved Oct. 3, 2021).

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

[11] 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.

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

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

[14] *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.

[15] *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.

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

[17] 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.