THE UNIVERSITY OF HONG KONG

DEPARTMENT OF COMPUTER SCIENCE

COMP4802 Extended Final Year Project

Interim Report

Student name: FONG Kwan Ching

UID: 3035569402

Project title: A Study on Forecasting Tropical Cyclone Properties

Supervisor: Dr Beta C.L. Yip

Date submitted: 23 January 2022

Abstract

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 level, despite their reliability and informativeness to the general public. This project builds a statistical-dynamical ensemble forecast, in which decision trees and regression splines 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 completed to date, namely the development of a baseline model using a baseline dataset. This progress report describes the tasks completed so far, namely the development of a baseline dataset and an initial baseline model. The progress has been satisfactory, as the baseline model showed adequate performance, but is slightly behind schedule. The upcoming months will see the preparation of a statistical-dynamical hybrid dataset and the development of experimental models that act as the ensemble. If successful, this project may open new opportunities for probabilistic statistical-dynamical impact level forecasts dedicated to specific locations like Hong Kong.

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 teaching me how to write a proper progress report.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full-Form
CLIPER	Climatology and Persistence
DT	Decision tree
ERA5	ECMWF Reanalysis v5
GFS	Global Forecasting System
HKO	Hong Kong Observatory
HKWW	Hong Kong Weather Watch
IBTrACS	International Best Track Archive for Climatological Stewardship
MLP	Multilayer perceptron
NCEP FNL	NCEP FNL Operational Model Global Tropospheric Analyses
NWP	Numerical weather prediction
PCA	Principal component analysis
sklearn	scikit-learn
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.

As the general public may not have a technical background, a forecasting method that produces intuitive and understandable forecasts should be favoured. Probabilistic forecasts achieve so by truthfully presenting the inherent uncertainty of TC behaviour to the general public [4]. Furthermore, by replacing quantitative technicalities like wind speed, rainfall, TC positions with a simple qualitative "threat index" [4] or "impact level", the general public can quickly grasp the seriousness of the TC one should expect. Therefore, probabilistic impact level forecasts are deemed valuable.

To produce the forecasts in the first place, there have been countless techniques developed in the past decades where one can choose from. In general, these techniques are either statistical methods that identify patterns in historical data or dynamical methods that run simulations of the entire atmosphere using the underlying physical laws on a global scale (“synoptic”). The latter type, also called “numerical weather prediction” (NWP), is currently the state-of-the-art method used by meteorological authorities worldwide [5]. In recent years, interest arose in a “statistical-dynamical” approach that combines both statistical and dynamical methods to harness their respective advantages.

At the moment, there is only one probabilistic TC impact level forecast designed for Hong Kong. Developed by an unofficial organization known as the Hong Kong Weather Watch (HKWW), this forecasting product uses the statistical method of spline regression to produce Hong Kong Observatory (HKO) TC warning signal issuing probabilities, taking TC positions, intensities, and signal status as inputs. This product produces forecasts valid for up to 6 hours, and to extend them to longer periods, the HKWW first forecasts future TC positions and intensities, then feeds them into the forecasting model to obtain further outputs [6].

However, this product does not take synoptic atmospheric factors into account, treating them as neglected hidden variables instead. Because these factors control the future evolution of TCs, it is hypothesized that forecasting models which can consider them alongside past TC trajectories will have superior performance. Unfortunately, there are no such probabilistic statistical-dynamical TC impact level forecasts for Hong Kong yet.

1.2 Project Objectives

This project intends to fill the research gap by building a statistical-dynamical TC forecasting product to assess TC impact level probabilities for Hong Kong. The forecasting product will report the probabilities of a TC leading to minimal impact (TC warning signal no. 1), limited impact (signal no. 3), substantial impact (signals no. 8 to no. 10), and direct strike (TC centre passes through a 100km radius of HKO headquarters), in the upcoming 72 hours.

Since there are no universally accepted quantitative measurements for TC impact or threat severities, the TC warning signals issued by the HKO are assumed to be an appropriate proxy, similar to the HKWW's approach. Signals no. 9 and no. 10 are grouped with no. 8 because of their rarity. The predictand "direct strike" is introduced to compensate for the lack of a warning for the additional threat posed by excessively close TCs, which oftentimes lead to warning signals no. 9 and no. 10.

1.3 Current Status

Thus far, the project has made acceptable progress. The HKO TC warning signals database and TC best track data have been acquired and treated, from both of which a baseline dataset (more in Section 2.1) was built. The first baseline model has also been developed and evaluated, while a second model is being built. The time-consuming process of processing the dynamical data have not yet started, which may cause delays to the schedule and affect the development of experimental models.

1.4 Report Outline

The remainder of this report is organized as follows: Chapter 2 explains the design and structure of the forecasting product, the selection of data sources, the proposed modelling techniques, and the revisions made since the submission of the project plan. Chapter 3 describes the tasks completed, such as dataset preparation and baseline model development, compares the progress to the project schedule, and outlines future tasks. Chapter 4 concludes the report.

Chapter 2: Methodology

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

2.1 Experiment Setting

An important element of the project is to show that the statistical-dynamical models are indeed better than models that do not use dynamical data (see Sections 1.1 and 1.2). Alternatively, this hypothesis necessitates a comparison between a statistical-dynamical dataset ("hybrid dataset") and a control dataset without dynamical data ("baseline dataset"). The modelling methods should therefore be identical except the input data are different. The experiment to test the hypothesis is set up as follows:

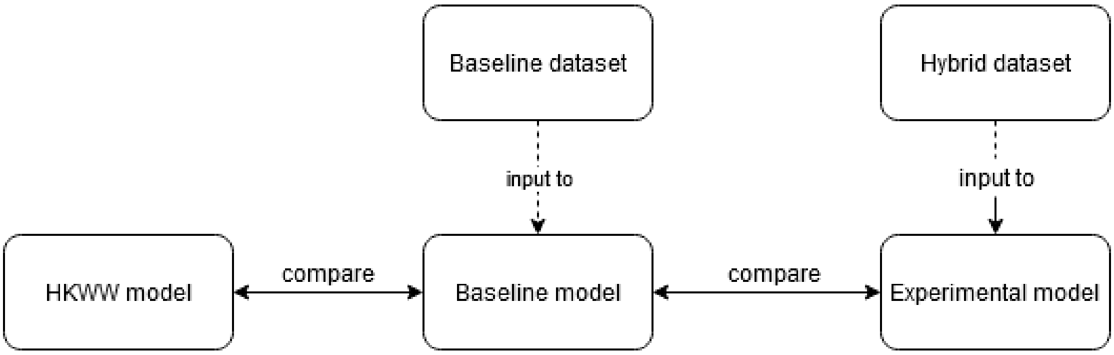


Fig. 2.1 Elements of the experiment setting

As shown in the figure, two models are to be built from the two datasets separately. The one using the baseline dataset is called the "baseline model" and the one using the hybrid dataset “experimental model”. They should share an identical modelling method, such that performance comparisons are possible.

A custom baseline is needed because the modelling methods can vary. To build a performant forecasting product, which is the ultimate objective of this project, the modelling method should be alternated to find an optimal one. The model by HKWW [6] is used to benchmark the performance of a baseline model and ascertain that it correctly realizes the baseline dataset’s potential. The HKWW model is not used directly as a baseline because the experimental models need not use the same modelling method.

2.2 Evaluation Metrics

The success of each model is measured in several ways.

Firstly, the deterministic forecasts made from the models should have a good F1 score. The F1 score is a balanced measurement between precision and recall [7], which respectively assess false alarm rates (high precision implies few false positives [8]) and the ability to warn impending TC threats (high recall means relevant items are better identified [8]). As the forecasts are probabilistic in nature, decision thresholds should be found first, such that the probabilities can be converted to deterministic “threat/no threat” values.

Secondly, the probabilities should be reasonably well-calibrated. That is, the probabilities must also be good confidence scores [9], such that among n predictions with the same probability p there are np items that are indeed positives. This can be qualitatively measured using calibration plots or quantitatively with Brier scores [10] (the lower, the better the calibration is).

It must be noted that the HKWW provides neither Brier scores nor decision thresholds in their documentation, and they do not predict direct strikes either [6]. Thus, it is assumed that if the F1 score of a model is close to that by the HKWW, then the direct strike performance and calibration quality are both within an acceptable range.

2.3 High-Level Design

This project intends to explore multiple modelling methods and iteratively deepen the methodology. As a result, multiple experimental models will be created, the best of which are to be grouped as an ensemble to produce forecasts, similar to the approach taken by major NWP institutions [11]. This makes the final forecast more robust and comprehensive by considering each ensemble member's output in a weighted sum.

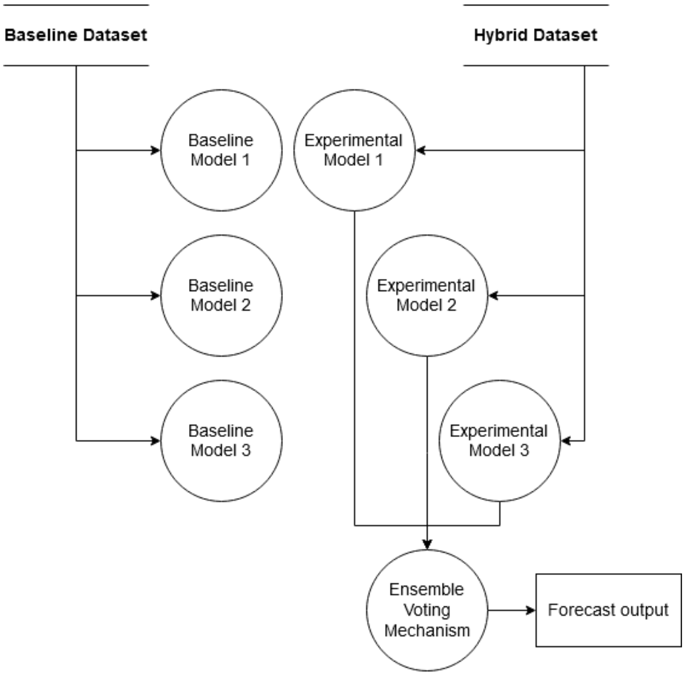


Fig. 2.2: Data flow diagram from datasets to models and outputs

In Figure 2.2, the data flow and structure of the forecasting product are shown. Multiple baseline model – experimental model pairs are present, one for each proposed modelling method and only the statistical-dynamical experimental models participate in the ensemble. The voting mechanism weighs the experimental models' outputs and produces the final output. The number of models in the ensemble may vary as new modelling options are discovered over the course of this project.

The development workflow of the project is as follows: Firstly, the data sources (see the next section) are identified. Then, the data are obtained and preprocessed to construct datasets on which the models rely. Next, the models are built using the datasets as training data. Finally, the models are evaluated, baselines against HKWW and experimental models against the baselines.

A revision has been made since the submission of the Project Plan in October 2021, such that multiple baselines instead of only one are devised. The new approach was adopted to allow for more rigorous and convincing comparisons.

2.4 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 [12]; 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 data quality, time coverage, and ease of use. Similarly, the selection criteria for NWP datasets include comprehensiveness, spatial resolution, time coverage, and ease of use.

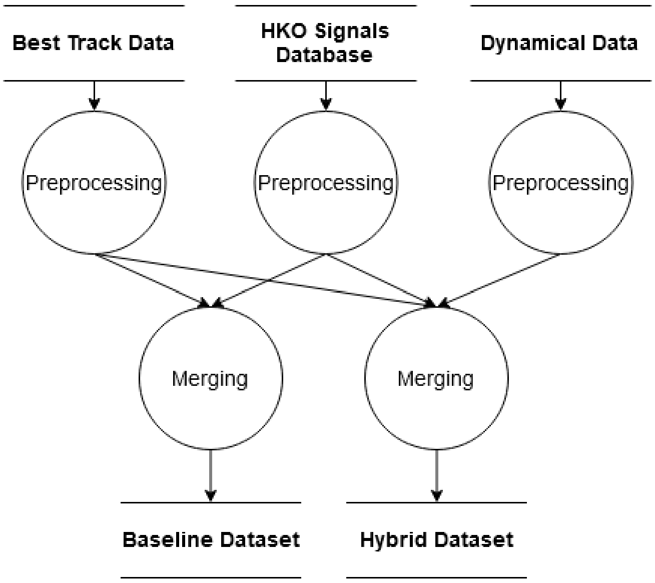


Fig. 2.3: Data preparation process

The diagram above summarizes the data preparation process. The data acquired from the selected data sources are separately preprocessed before they are merged to build the baseline and hybrid datasets.

Data from the chosen data sources are preprocessed to perform the necessary cleaning and transformations, 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, which typically are grids covering a part of the world map at an instant. Processing NWP data is the primary difficulty of the whole process.

The final step of the dataset preparation process is to merge the data sources. Each best track data record is matched with the corresponding dynamical variables and impact level labels.

N.B.: In this report, the terms "input variable", "predictor" and "feature" are used interchangeably to refer to input variables that the models consider.

2.5 Proposed Modelling Techniques

Several options can be used to build forecasting models for the ensemble outlined in Section 2.3. Each technique described below is a *group* of modelling methods categorised together because of their common features. The individual methods of each group will be iteratively tested whenever appropriate to identify the best-performing one for evaluation.

2.5.1 Decision Trees and Tree-Based Methods

The first option is to use decision trees (DTs) and variants thereof, such as random forests and gradient boosting DTs. DTs are a machine learning technique that identifies rules to classify data and can be adapted to perform regression [13]. They are known to be useful for TC forecasting [14] and research, such as finding the conditions for TC recurvature [15]. DTs are also easy to develop, using the Python packages scikit-learn (sklearn) and XGBoost, the latter of which further optimizes DTs through gradient boosting [16]. Therefore, DTs and their variants are chosen as the first class of modelling techniques to test.

By treating the forecasting problem as a multi-label classification problem [17], where each TC time series sample should be labelled with mutually non-exclusive classes corresponding to the four predictands, and then the confidence scores of the labels as a marginal probability should be calculated to make a probabilistic forecast.

The choice to use this classification problem approach is made because regression methods and simple line-fitting techniques require more pre- and postprocessing (see next section), while the construction of classifiers is simpler, and they can give reasonable results within a shorter development period.

2.5.2 Line- and Curve-Fitting Methods

This group contains linear classification and regression methods, together with non-linear extensions thereof. These techniques are chosen to mimic Climatology and Persistence (CLIPER) methods, which typically use linear regression to predict TC properties using past data and time series [18].

They differ from DTs because they do not use rules to recursively separate data into smaller splits, but rather, the output is a function of the input features. For example, linear regression requires a line to be fitted to the data, such that outputs can be made as a weighted sum of the inputs [19].

Candidate methods include logistic regression and support vector machines, which are to perform multi-label classification, and regression methods like linear regression. The HKWW model uses penalized splines to perform regression [6], which can be implemented in this project as multivariate adaptive regression splines [20]. These models can be built using sklearn and other third-party Python packages such as py-earth [21].

One important consideration is the relationship between the features and the prediction targets may not necessarily be linear. For example, the probability of a TC strike does not increase with month numbers from 1 through 12, but rather, it tends to peak in the summer months. This will necessitate the use of kernels or kernel approximations to convert features to forms linear models can process [22]. Regression splines have complex shapes that help them adapt to non-linear data, thus they may have better performance while requiring less preprocessing.

On another hand, these models, regardless of linearity, do not naturally consider interactions and relationships between feature variables, e.g. latitude times longitude, and these interaction terms should be calculated as polynomials in advance. The large number of polynomial features generated will necessitate feature selection by ANOVA F-tests or principal component analyses (PCA). This step is common to numerous CLIPER-like methods, such as the inaugural CLIPER itself [18].

2.5.3 Miscellaneous Methods

Several modelling methods have good potential and are worth trying provided there is sufficient time.

The Python package of sktime implements numerous time series classification, regression, and forecasting models [23-24], many of which are novel and have comparatively little related literature. However, if the models are updated to effectively support multivariate inputs (like the datasets of this project) soon, then these inventive models can be a suitable choice because the forecasting of TCs does involve time series processing.

Gaussian processes [25] are an alternative. Gaussian processes allow for reasoning about a posterior distribution given observations of a related prior distribution and can model time series where data mean and variance values change over time [26]. This makes Gaussian processes suitable for TC time series modelling and regression as well.

2.5.4 Revisions since October 2021

Two changes have been made to the list of proposed modelling methods since the submission of the Project Plan in October 2021. Multilayer perceptrons (MLPs) [27] were originally proposed to act as the baseline model, but due to practical difficulties (see Section 3.1.3) this had to be abandoned. Monte Carlo simulations [28] were initially considered viable to model the stochastic behaviour of TCs, but further research showed that this method is more suitable for season-wide climatological analyses instead of individual TC forecasts, therefore Monte Carlo simulations were abandoned as well.

2.6 Limitations

This project has two limitations by nature.

Firstly, the HKO TC warning signal system is assumed to be an appropriate replacement for accurate impact level measurements which there are none. However, the actual amount of damage Hong Kong may suffer under some TC warning signals varies; and there are cases where the signals fail to indicate the threat of TCs. For example, the HKO hoisted signal no. 3 in 2006 during Typhoon Prapiroon which brought about disproportionately severe weather, leading to heavy criticism and subsequent reform of the signal system [29].

Secondly, this project uses statistical methods to analyze historical data under the assumption that past climatological patterns identified remain unchanged in the future. In other words, future changes in general TC behaviour are not accounted for. This limits the forecasting product’s ability to handle extreme cases and new climatological patterns.

As such, the outputs of the forecasting product should only be used for reference regardless of its apparent performance, because of the simplifying assumptions and intricacies overlooked. The authoritative advisories by the HKO should always be considered first.

Chapter 3: Project Progress

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

3.1 Tasks Completed

3.1.1 Construction of the Baseline Dataset

The HKO TC warning signal issuance records and the best track data were retrieved, preprocessed, and combined into the baseline dataset.

The HKO warning signal records were retrieved and processed in early October 2021. The records were downloaded from the HKO website [12] by web scraping and a script was written to extract the records in the obtained webpage 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 primary difficulty encountered during this process was with the web scraping step. The peculiar structure of the obtained webpage made extracting TC warning records from it tremendously difficult. Some manual trial-and-error tests were needed to correctly locate the records and extract them.

The *International Best Track Archive for Climatological Stewardship* (IBTrACS) dataset [30] 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 [31]. 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 the baseline dataset. 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.

The columns of the baseline dataset are as follows, as it was first built in November:

No.	**Column name**	**Explanation**
0	MM	Date and time of the final record in the time series; each row in the dataset contains one.
1	DD	
2	HH	
3	LOW_IMPACT	These 4 columns are the correct labels, i.e. whether

4	MID_IMPACT	each event will take place in 72 hours.
5	BIG_IMPACT	
6	DIRECT_STRIKE	
7	00LAT	Latitude (in degrees North of the equator) of the TC centre position at time *t* = 0h back.
8	00LON	Longitude (in degrees East of the prime meridian) of the TC centre position at time *t* = 0h back.
9	00WIND	Maximum sustained wind speed (in knots) at the centre of the TC at time *t* = 0h back.
10-12	06LAT, 06LON, 06WIND	Position and intensity of the TC at time *t* = 6h back, i.e. 6 hours **before** the time specified in columns 0-2.
13-15	12LAT, 12LON, 12 WIND	The extension of the previous 6 columns (two timesteps) until 24 hours back.
16-18	18LAT, 18LON, 18WIND	
19-21	24LAT, 24LON, 24WIND	

Table 3.1: Baseline dataset (initial version) summary

As shown in the table, TC trajectories and intensity evolutions are encoded as a time series in the dataset, while interaction terms or inferable terms (e.g. dates of the other records in the time series) are excluded, assuming that the models can still function well without them. Moreover, the time series lasts for only 24 hours, so forecasting newly formed TCs will be possible.

3.1.2 Selection of Dynamical Predictors and Data Source

The data source from which the dynamical data archives are to be acquired was finalized and a brief literature review was conducted to identify useful variables to include in the hybrid dataset as predictors.

When comparing the *ECMWF Reanalysis v5* (ERA5) dataset [33] to the *NCEP FNL Operational Model Global Tropospheric Analyses* (NCEP FNL) dataset [34], 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 2021.

In January 2022, some 11 articles on statistical-dynamical TC forecasts were studied and their choices of dynamical predictor variables were compared. It was concluded that the following predictors are the most common and can be used as a starting point:

No.	Predictors	Dynamical data needed
1-2	Vertical wind shear magnitudes [34-37], between 200-850 mbar levels and between 500-850 mbar [34]	u-, v-components of winds
3-4	Relative humidity [34-37] at 750-850 mbar and 300-500 mbar levels [34]	Relative humidity
5-6	Temperature closest to surface (1000 mbar)	Temperature

	[34, 36-38] and at 200 mbar [34]	
7-12	Hong Kong surface (1000 mbar) winds, 200 mbar zonal (u) wind [34, 37, 38], 500 mbar winds [35, 39], summer monsoon index (850 mbar zonal) [15, 40-41]	u-, v-components of winds
13	850 mbar vorticity [34, 35, 42]	Absolute vorticity
14-18	Westerly index [43]; Western North Pacific subtropical high area, intensity, and westward extension indices [44]	500 mbar geopotential height [15, 35, 39]
19	925 mbar potential temperature [34]	Potential temperature

Table 3.2: Proposed dynamical predictors

As shown in the table, 19 predictors are derived from six dynamical data variables. That means the data archives of the six variables will be downloaded such that the 19 predictors can be computed therefrom. While some indices are well-defined, the other predictors require custom calculations. This is done by averaging the data values in the vicinity of the TC centre, such as taking a 2-degree average around it for potential temperature [34] and a 7-degree average for 200 mbar temperature [37].

3.1.3 Development of the First Baseline Model

The first baseline model was developed in November and December 2021.

The originally proposed baseline modelling technique of MLP proved unsuitable. Various network structures were tested, ranging from small ones with 32-64 neurons in each of the 1-3 hidden layers to large ones with up to 1024 neurons in 10 hidden layers.

Accuracies never grew above zero even though a naïve forecaster can achieve 88% accuracy, while training loss showed erratic behaviour over time.

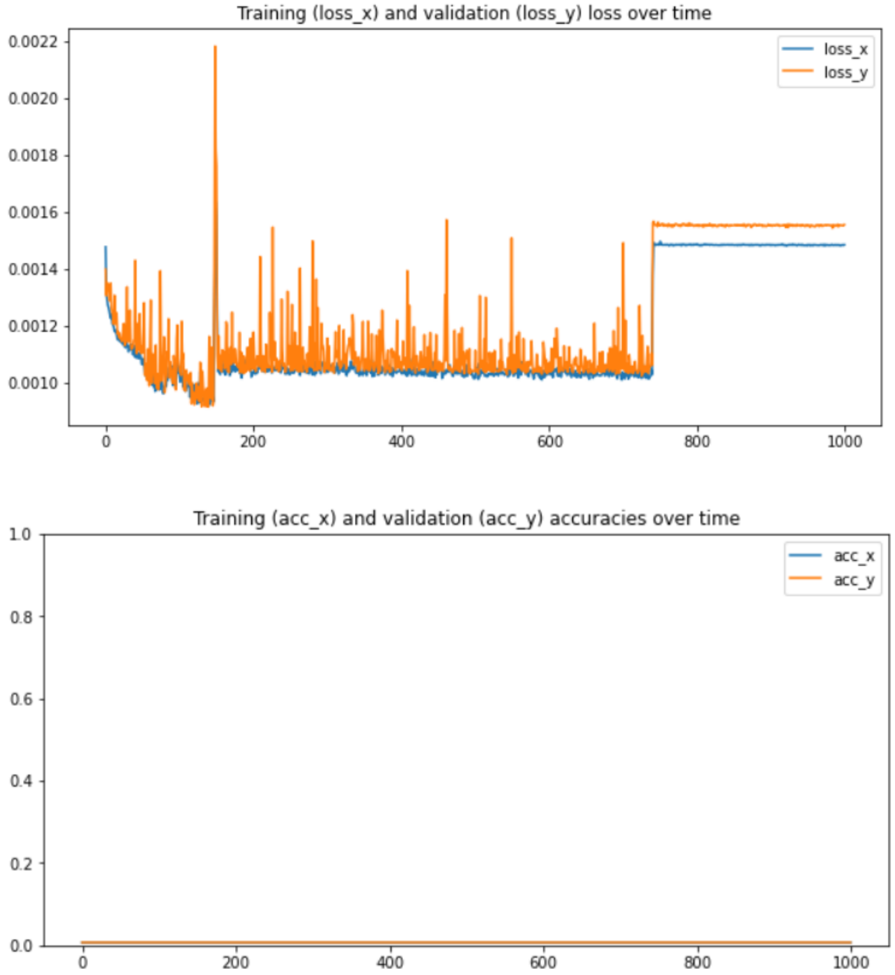
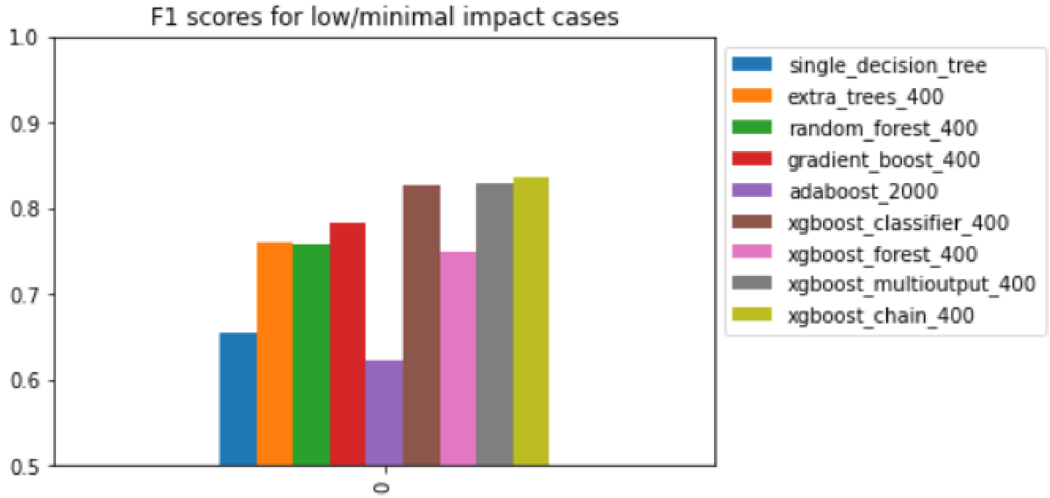


Fig. 3.1: The change in (a) losses (top) and (b) accuracies (bottom) during a typical MLP training process

In Figure 3.1b, it is evident that the MLP cannot be fitted to the data. The model was left to run for 1000 epochs, but the accuracy values computed with training and validation datasets both remained close to zero. Figure 3.1a shows that the loss values showed even more inexplicable behaviour, that the losses abruptly grew twice during the training process. At the moment of writing, the root causes of these abnormal phenomena have not been found.

The failure of MLPs necessitated the methodology changes (see Section 2.5.4). As a result, the first proposed experimental modelling technique in the project plan, namely DTs and derived methods (see also Section 2.5.1), was tried next. This was because the debugging of the MLPs might require too much time, thus compressing the time available for other tasks. Eight variants of decision trees were built, and many showed adequate performance.

It was observed that the F1 scores were the best for simple predictands (e.g. minimal impact) and gradually worsened as difficulty increased (e.g. up to "direct strike", the hardest predictand). The following figures show the F1 scores of the models predicting the easiest and hardest predictands.



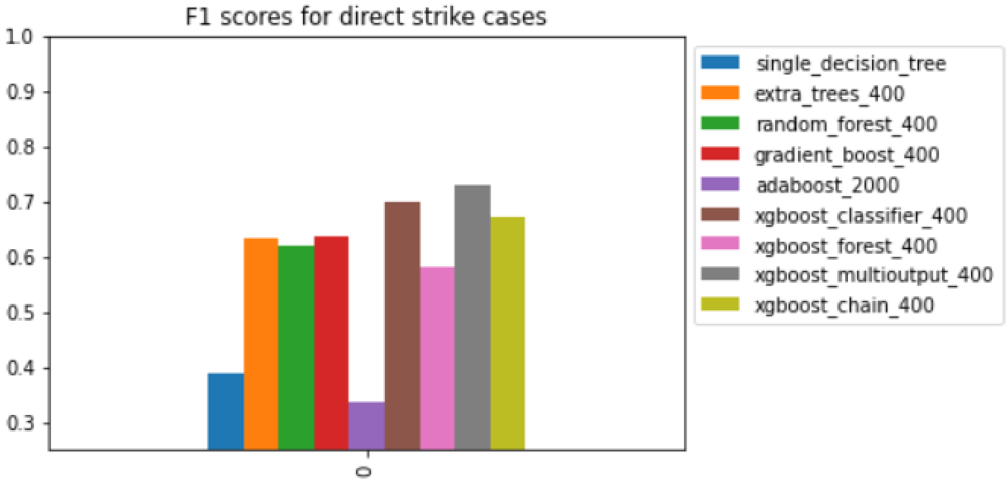


Fig. 3.2: F1 scores for (a) minimal impact predictions (previous page) and (b) direct strike predictions (this page).

As shown in the figure, the “xgboost” models have better performance than the others in general, except “xgboost_forest_400”. Amongst them, the best model was chosen based on the average F1 scores across all categories (also called “macro-average” [7]). The “xgboost_multioutput_400” model had the best average F1 score and was chosen to be the first baseline model.

Probability calibration was checked afterwards. The Brier scores tended to show similar trends like F1 scores and worsened (increased) with difficulty. Calibration did not improve the scores significantly, and in some cases aggravated the probability distributions.

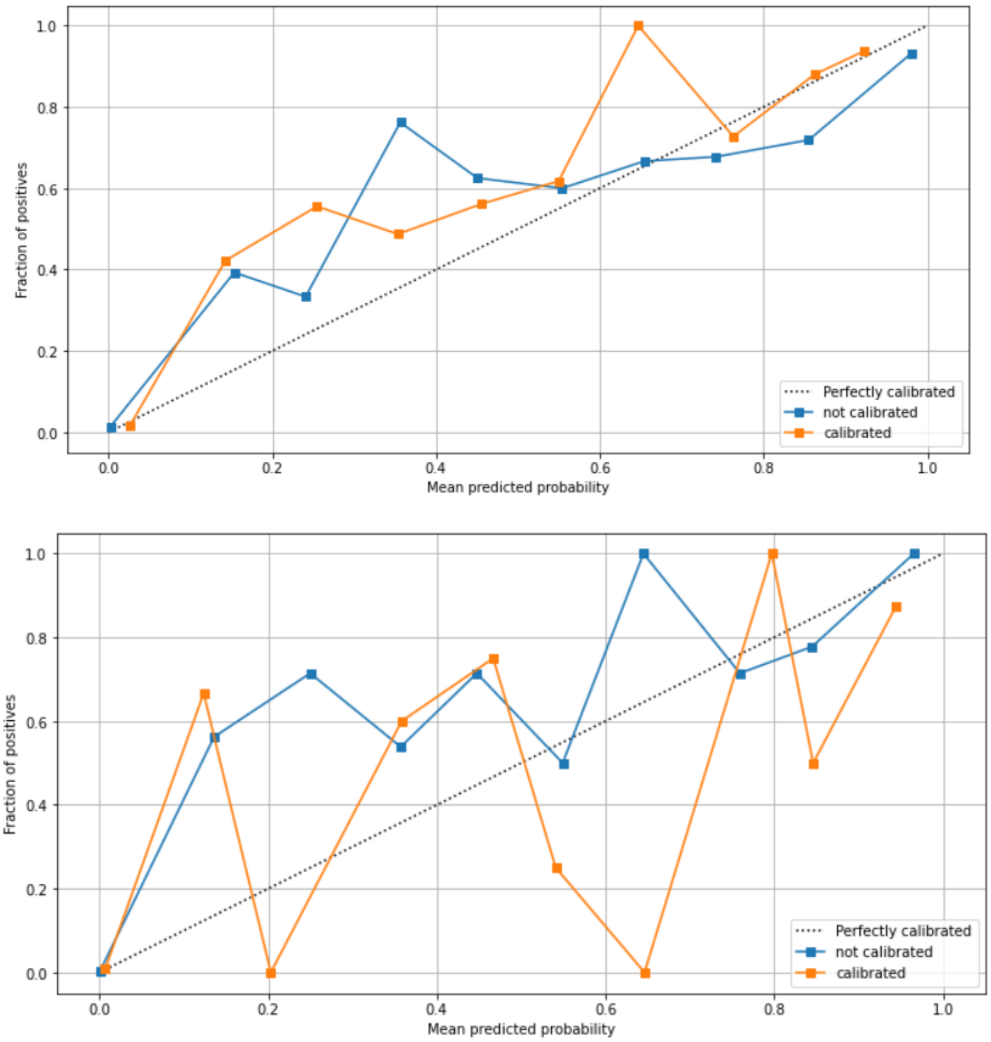


Fig. 3.3: Calibration plots for the XGBoost model predicting (a) minimal impact (top) and (b) direct strike (bottom) cases

An ideal model should have a perfectly diagonal calibration plot like the dotted line in Figures 3.3a and 3.3b. The blue and orange lines are the plots of the original XGBoost model and the calibrated model respectively. In Figure 3.3a, it is apparent that the difference in performance between the two models are near indiscernible, whereas in Figure 3.3b, the calibrated model’s plot shows much more fluctuations than the non-calibrated one, indicating calibration is not helpful. As a result, this project claims ignorance about probability calibration and considers it unnecessary, at least for the XGBoost model.

As the final step, the best decision thresholds were then found. The F1 scores were then finalized, and the following evaluation tasks ensued.

3.1.4 Evaluation of the First Baseline Model

With the completion of the first baseline model, an XGBoost gradient boosting DT classifier, comparison with the HKWW model to verify model performance became possible. The following table summarizes the XGBoost model's performance and the reported performance of the HKWW model [6] predicting warning signals no. 1, no. 3, no. 8-10.

		HKWW model	XGBoost baseline
Minimal impact (signal no. 1)	Precision	**0.95**	0.88
	Recall	**0.85**	0.78
	F1 score	**0.897**	0.828
Limited impact (signal no. 3)	Precision	**0.93**	0.84
	Recall	0.78	**0.83**
	F1 score	**0.848**	0.837
Substantial impact (signals no. 8-10)	Precision	**0.94**	0.76
	Recall	0.68	**0.82**
	F1 score	0.789	**0.799**
Direct strike	Precision	N/A	0.82
	Recall		0.65
	F1 score		0.724

Table 3.3: Comparison between the HKWW model and the XGBoost baseline model.

In the table, bolded values are the superior ones. While the HKWW model generally outperforms the XGBoost baseline, in terms of bold numbers count, it must be noted that the prediction targets are different for both models. It should be expected that the baseline has worse performance because it has to forecast 72 instead of 6 hours ahead. Thus, the baseline model shown above was considered acceptable, as it showed performance close to the HKWW one and its recall scores were oftentimes surprisingly good. An issue is the precision of the baseline model tended to be low. This indicates there is still room for improvement.

In December 2021, the XGBoost baseline model was made to forecast the impending Typhoon Rai. This was to observe the model's behaviour facing statistically rare TCs and to obtain experience building an operational forecast, so the final product may be built similarly. The model consistently failed to predict the ultimate issuing of the TC warning signal no. 1 on 20 December. This was expected because of Rai's rarity; it is surmised that the introduction of dynamical data may help future models identify these cases better.

There are several recommendations suggested based on the performance of the model:

- The dataset can use longer time series for each sample, e.g. 48h instead of 24h. This comes at the cost of about 10-20% of the dataset size and makes it impossible to forecast TCs that have newly spawned. The latter is unfavourable because the ability to forecast newly formed TCs near Hong Kong will be lost.
- Use a different representation of the TC position and intensity values. As the (radial) distance of a TC to Hong Kong is typically one of the criteria behind TC signal issuances [45], the position of a TC can be described in a polar coordinate

system instead. This has the additional advantage of eliminating the nonlinearity between TC impact and TC coordinates (the closer a TC is, the likelier it will affect Hong Kong). Similarly, variables such as the change in intensity, TC movement speed and heading can be introduced.

- Considering how HKWW uses current and previous TC warning signal status as an input feature to their models, the same could be done for the baseline dataset. It can be rearranged into an explicit time series, such that each record in the series contains signal statuses and dates in addition to TC positions and intensities.

A small preliminary test was carried out to check whether the recommendations were helpful. The unverified initial results showed increases in F1 scores by 0.01 to 0.10 for different categories, with the time series approach even raising precision scores to 0.8, which were considered promising.

3.2 Tasks in Progress

3.2.1 Dynamical Data Acquisition and Processing

To build the dynamical features, the corresponding atmospheric variables as described in Table 3.2 need to be downloaded and preprocessed first.

The NCEP FNL dataset supports subset downloads, therefore only about 5GB of compressed data files have to be downloaded. To minimize the download sizes, each variable is separately downloaded in a subset data request. The extracted meteorological data total at about 10000 files per variable. At the moment, files of 500 mbar geopotential height, 850 mbar absolute vorticity, and u- and v-components of winds at 1000 mbar, 850 mbar, 500 mbar and 200 mbar have been downloaded.

The processing step has not been started yet, because a Python package that can handle the data files has yet been found. The NCAR Command Language [46] is a reputable choice to serve as an alternative to Python, but as it is migrating to Python [47], the corresponding package PyNIO [48] will be tried next.

3.2.2 Development of a Second Baseline Model

The second baseline model is being developed in parallel to the dynamical data preparation. This baseline model uses the methods outlined in Section 2.5.2, i.e. line-fitting and curve-fitting methods. The extra steps to generate interaction terms by computing polynomial features and introducing kernels to handle non-linearity have been done. A dimensionality reduction procedure using PCA has also been added to remove irrelevant generated features to reduce computation.

At the moment, there are only preliminary results. Logistic regression, support vector machines and linear regression have been tested and it is found that polynomial features and kernels improved model performance as expected, while dimensionality reduction brought about minor performance degradations instead of improvements. While as few as 30 principal components can explain 98.75% of the variance, 450 components were needed to make effective predictions. Alternative feature selection methods such as ANOVA F-tests have been tested and none came to avail.

The performance of the current best line-fitting model, a logistic regressor classifier that takes in PCA-transformed 450-dimensional data as input (“PCA classifier”), is shown in the following table, where it is compared with the XGBoost model of Section 3.1.3.

		XGBoost baseline	PCA classifier
Minimal impact (signal no. 1)	Precision	0.88	0.48
	Recall	0.78	0.74
	F1 score	0.828	0.579
Limited impact (signal no. 3)	Precision	0.84	0.45
	Recall	0.83	0.76
	F1 score	0.837	0.563
Substantial impact (signals no. 8-10)	Precision	0.76	0.37
	Recall	0.82	0.56
	F1 score	0.799	0.449
Direct strike	Precision	0.82	0.19
	Recall	0.65	0.50
	F1 score	0.724	0.280

Table 3.4: Comparison between the XGBoost baseline model and the PCA classifier model

It is immediately evident in the table that the XGBoost model is superior in all aspects. Similar to the XGBoost model, the PCA classifier model also suffers from a discrepancy between precision and recall. It is further noted that the best performing model at the moment is a classifier but not a regressor, which echoes Section 2.5 that classifiers show results faster than regressors. The planned work on regression splines (Section 2.5.2) has not yet been started.

3.3 Project Timeline

The following table outlines the schedule of this project.

Date	**Tasks**	**Status**
October – November 2021	High-level design, Baseline dataset preparation, Baseline model 1 development	All completed
December 2021 – January 2022	Hybrid dataset preparation, Baseline model 2 development	All in progress
February 2022	Experimental model 1 development and evaluation	To-do
March 2022	Experimental model 2 development and evaluation	To-do
April 2022	Ensemble forecast, Any remaining tasks	To-do

Table 3.5: Project timeline

With the baseline dataset already prepared and the first model built, the tasks related to baseline models as listed in the table have made appropriate progress. The development of the first baseline model was a longer task than initially expected, with fine-tuning and experimentations with the recommendations (Section 3.1.4) continuing well into January 2022. The revision of methodologies (see Sections 2.1 and 2.5.4) introduced new tasks such as developing another baseline model, which has commenced in January 2022. The tasks related to dynamical data were originally planned for 2021 in the Project Plan but have only started recently. To conclude, a satisfactory number of tasks have been completed, but the current state of the project is better described to be behind schedule, owing to the slow progress in processing dynamical data.

3.4 Future Tasks

3.4.1 Construction of the Hybrid Dataset

The construction of the hybrid dataset involves the computation of predictor variables and the merging of these variables to the baseline dataset (with data before 1st August 1999 excluded, because the NCEP FNL dataset does not have records for them). The recommendations regarding the baseline dataset (see Section 3.1.4) will also be considered.

The incorporation of a new data source to the baseline dataset is expected to be a major difficulty because there will be considerable overhead to open raw dynamical data files and calculate these values, especially if the predictor values are calculated during the merging process. A solution is to precompute all dynamical predictors first, using TC positions supplied by the best track archives, and then save them to file so that a smaller number of files will have to be opened during the data source merging process.

3.4.2 Development of the First Experimental Model

After the construction of the hybrid dataset, the first experimental model can then be built. This model will be DT-based, and it is predicted that the XGBoost gradient boosting models will once again be the best-performing.

The development of this model will likely encounter few problems because the baseline model development code can be reused. The main difficulties that may arise have been solved and the only change will be the dataset provided. Nonetheless, the first batch of models built may underperform because the input dataset has different dimensions than before, but model hyperparameters have not been adjusted accordingly. The solution will involve techniques such as grid search [49] that automatically optimizes model hyperparameters.

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. DTs and line-fitting methods are the primary modelling techniques employed, while a few cutting-edge time series modelling techniques are also being considered.

The project has made acceptable progress thus far. The HKO warning records and the best track data have been processed and combined to give a baseline dataset, and a baseline model has been successfully built despite initial setbacks. This initial baseline model has shown reasonable performance and possible means of improvement have been identified. The project is still on a tight schedule, however, because the NWP data processing tasks are making slow progress, the development of the ensemble members may be delayed.

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 and thus prevent the usage of the forecasting product elsewhere. 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] 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.

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

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

[7] "sklearn.metrics.f1_score." scikit-learn 1.0.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (accessed Jan. 11, 2022).

[8] "Classification: Precision and Recall." Google Developers Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall (accessed Jan. 11, 2022).

[9] "1.16. Probability calibration." scikit-learn 1.0.2 documentation. https://scikit-

learn.org/stable/modules/calibration.html (accessed Jan. 23, 2022)

[10] A. P. Dawid and M. Musio, "Theory and Applications of Proper Scoring Rules," *METRON*, vol. 72, no. 2, pp. 169-183, Jan. 2014. DOI:10.1007/s40300-014-0039-y.

[11] H. A. Titley, R. L. Bowyer and H. L. Cloke, "A global evaluation of multi-model ensemble tropical cyclone track probability forecasts," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 726, pp. 531-545, Jan. 2020. DOI: 10.1002/qj.3712.

[12] "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 (accessed Oct. 25, 2021).

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

[14] 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.

[15] W. Zhang, Y. Leung and J. C. L. Chan, "The Analysis of Tropical Cyclone Tracks in the Western North Pacific through Data Mining. Part I: Tropical Cyclone Recurvature," *Journal of Applied Meteorology and Climatology,* vol. 52, no. 6, pp. 1394-1416, Jun. 1, 2013. DOI: 10.1175/JAMC-D-12-045.1.

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

[17] "1.12. Multiclass and multioutput algorithms." scikit-learn 1.0.2 documentation. https://scikit-learn.org/stable/modules/multiclass.html#multiclass-and-multioutput-algorithms (accessed Jan. 11, 2022).

[18] 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.

[19] “1.1 Linear models.” scikit-learn 1.0.2 documentation. https://scikit-learn.org/stable/modules/linear_model.html#linear-model (accessed Jan. 23, 2022).

[20] J. Friedman, “Multivariate adaptive regression splines,” *The Annals of Statistics*, vol. 19, no. 1, pp. 1-67, Mar. 1991. DOI: 10.1214/aos/1176347963

[21] *Py-earth*. (2017). Accessed Jan. 23, 2022. [Online] Available: https://github.com/scikit-learn-contrib/py-earth.

[22] M. I. Jordan. The Kernel Trick [Online]. Available: https://people.eecs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf.

[23] M. Löning, A. Bagnall, S. Ganesh, V. Kazakov, J. Lines and F. Király, “sktime: A Unified Interface for Machine Learning with Time Series,” 2019. [Online] Available: https://arxiv.org/abs/1909.07872.

[24] *alan-turing-institute/sktime*. 2021. Zenodo. Accessed: Jan. 11, 2022. [Online] Available: https://zenodo.org/record/5610006#.Yd1erVl-VEY. DOI: 10.5281/zenodo.3749000.

[25] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger and A. G. Wilson, “GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration,” *Advances in Neural Information Processing Systems*, 2018.

[26] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson and S. Aigrain, “Gaussian processes for time-series modelling,” *Philosophical Transactions of the Royal Society A*, vol. 371, no. 1984, Feb. 13, 2013. DOI: 10.1098/rsta.2011.0550.

[27] “Multilayer Perceptron.” DeepAI. https://deepai.org/machine-learning-glossary-

and-terms/multilayer-perceptron (accessed Oct. 3, 2021).

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

[29] HKO, “Review of the Tropical Cyclone Warning System in 2006 and New Measures in 2007”, Feb. 26, 2007. Accessed: Jan. 23, 2021. [Online] Available: https://www.hko.gov.hk/en/wxinfo/currwx/files/tc_review_rpt.pdf.

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

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

[32] *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.

[33] *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.

[34] 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.

[35] J. Tan, S. Chen and J. Wang, “Western North Pacific tropical cyclone track forecasts by a machine learning model,” *Stochastic Environmental Research and Risk Assessment*, vol. 35, pp. 1113-1126, Jun. 2021. DOI: 10.1007/s00477-020-01930-w.

[36] J. A. Knaff, C. R. Sampson and G. Chirokova, “A Global Statistical–Dynamical

Tropical Cyclone Wind Radii Forecast Scheme," *Weather and Forecasting*, vol. 31, no. 2, pp. 629-644, Apr. 2017. DOI: 10.1175/WAF-D-16-0168.1.

[37] S. Kim, I. Moon and P. Chu, "Statistical–Dynamical Typhoon Intensity Predictions in the Western North Pacific Using Track Pattern Clustering and Ocean Coupling Predictors," *Weather and Forecasting*, vol. 33, no. 1, pp. 347-365, Feb. 2018. DOI: 10.1175/WAF-D-17-0082.1.

[38] H. Kim, C. Ho, J. Kim and P. Chu, "Track-Pattern-Based Model for Seasonal Prediction of Tropical Cyclone Activity in the Western North Pacific," *Journal of Climate*, vol. 25, no. 13, pp. 4660-4678, Jul. 2012. DOI: 10.1175/JCLI-D-11-00236.1.

[39] S. Giffard-Roisin, M. Yang, G. Charpiat, C. K. Bonfanti, B. Kégl and C. Monteleoni, "Tropical Cyclone Track Forecasting Using Fused Deep Learning From Aligned Reanalysis Data," *Frontiers in Big Data*, vol. 3, 2020. DOI: 10.3389/fdata.2020.00001.

[40] B. Wang and Z. Fan, "Choice of South Asian Summer Monsoon Indices," *Bulletin of the American Meteorological Society*, vol. 80, no. 4, pp. 629-638, Apr. 1999. DOI: 10.1175/1520-0477(1999)080<0629:COSASM>2.0.CO;2.

[41] B. Wang, Z. Wu, J. Li, J. Liu, C. Chang, Y. Ding and G. Wu, "How to Measure the Strength of the East Asian Summer Monsoon", *Journal of Climate*, vol. 21, no. 17, pp. 4449-4463, Sep. 2008. DOI: 10.1175/2008JCLI2183.1.

[42] S. S. Chand and K. J. E. Walsh, "Modeling Seasonal Tropical Cyclone Activity in the Fiji Region as a Binary Classification Problem," *Journal of Climate*, vol. 25, no. 14, pp. 5057-5071, Jul. 2012. DOI: 10.1175/JCLI-D-11-00507.1.

[43] C. Rossby, "Relation between variations in the intensity of the zonal circulation of the atmosphere and the displacements of the semi-permanent centers of action," *Journal of Marine Research*, vol. 2, pp. 38-55, 1939.

[44] 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.

[45] HKO. "Hong Kong's Tropical Cyclone Warning Signals." https://www.hko.gov.hk/en/publica/gen_pub/files/tcws.pdf (accessed Jan. 23, 2022).

[46] *The NCAR Command Language (NCL).* (2019), Computational and Information Systems Lab, National Center for Atmospheric Research. Accessed: Jan. 23, 2022. [Online] Available: https://www.ncl.ucar.edu/. DOI: 10.5065/D6WD3XH5.

[47] "The Future of NCL and the Pivot to Python." NCAR Command Language. https://www.ncl.ucar.edu/Document/Pivot_to_Python/ (accessed Jan. 23, 2022).

[48] *PyNIO*. (2018), Computational and Information Systems Lab, National Center for Atmospheric Research. Accessed: Jan. 23, 2022. [Online] Available: https://www.pyngl.ucar.edu/Nio.shtml.

[49] "3.2 Tuning the hyper-parameters of an estimator." scikit-learn 1.0.2 documentation. https://scikit-learn.org/stable/modules/grid_search.html#exhaustive-grid-search (accessed Jan. 23, 2022).