

THE UNIVERSITY OF HONG KONG

DEPARTMENT OF COMPUTER SCIENCE

COMP4802 Extended Final Year Project

Final 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: 18 April 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, multilayer perceptrons and generalized additive models are employed to analyze historical warning signal records, tropical cyclone best track data and synoptic meteorological analysis data archives. This final report details the methodology and project outcomes, describes the characteristics of the datasets employed and analyzes the results obtained with different modelling methods. The findings showed that a statistical-dynamical ensemble forecaster was able to outperform baselines that excluded dynamical data, but the exact performance depended on the modelling method used. The satisfactory results of 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. The help and advice offered by great people online, namely Miguel Trejo and hpaulj on Stack Overflow and Ryoumiya, Avi and ∇ (lyfe) in Avi's Discord server, have also proven vital to the project and I would like to express my gratitude to them.

I would also like to, in the most earnest yet civilized manner, condemn the accused reliably of the HKU VPN and the HKU CS GPU Farm in the period between 10 January and 23 January 2021. The network connections between my personal computer and the GPU farm over the HKU VPN and that between the running compute resource instance and the GPU farm gateway have been, to put it mildly, as stable as a bucking horse at a rodeo. I am most grateful to have been given a chance to practice my anger management by whatever was causing the problem, which was never identified.

Table of Contents

Abstract.....	i
Acknowledgements.....	ii
List of Figures.....	vi
List of Tables.....	vii
List of Abbreviations.....	ix
Chapter 1: Introduction.....	1
1.1 Background.....	1
1.2 Project Objectives.....	2
1.3 Current Status.....	3
1.4 Report Outline.....	4
Chapter 2: Literature Review.....	4
2.1 Climatology and Persistence Statistical Models.....	4
2.2 Numerical Weather Prediction and Ensembles.....	5
2.3 Machine Learning in TC Forecasting.....	5
2.4 Strike Probability Research.....	6
2.5 Miscellaneous Relevant Research.....	6
Chapter 3: Methodology.....	7
3.1 Experiment Settings.....	7
3.2 Evaluation Metrics.....	9
3.3 High-Level Design.....	10
3.4 Data Source Selection and Dataset Preparation.....	12
3.5 Data Modelling Techniques Employed.....	13
3.5.1 Tree-Based Models.....	14
3.5.2 Linear Classification and Regression Models.....	14

3.5.3 Direct Time Series Modelling	15
3.5.4 Multilayer Perceptrons	16
3.5.5 Gaussian Processes	16
3.5.6 Ensemble Design	17
3.6 Limitations and Exclusions	17
Chapter 4: Results and Analysis	18
4.1 Dataset Construction	18
4.1.1 Data Source Selection and Data Preprocessing	18
4.1.2 Baseline Dataset Details	19
4.1.3 Experimental Dataset Details	23
4.1.4 Notable Dataset Idiosyncrasies	28
4.2 Results with Tree-Based Models	31
4.2.1 Baseline Models	31
4.2.2 Experimental Models	36
4.2.3 Discussion	39
4.3 Results with Linear Models	40
4.3.1 Baseline Models	40
4.3.2 Experimental Models	42
4.3.3 Discussion	43
4.4 Results with Direct Time Series Modelling	47
4.4.1 Baseline Models	47
4.4.2 Experimental Models	50
4.4.3 Discussion	53
4.5 Results with Multilayer Perceptrons	54
4.5.1 Baseline Models	54

4.5.2 Experimental Models	57
4.5.3 Discussion	59
4.6 Results with Gaussian Processes	60
4.7 Final Ensemble Model	60
4.8 Summary and Overall Comments	67
Chapter 5: Conclusion.....	69
Appendices.....	71
Appendix 1: Feature Importance Reports	71
1.1 Top 50 features for each predictand, baseline tree-based regressor.....	71
1.2 Top 50 polynomial features for each predictand, mutual information and baseline data.....	76
1.3 Top 50 features for each predictand, experimental tree-based classifier	81
1.4 Top 50 polynomial features for each predictand, mutual information and experimental data	86
Appendix 2: References	94

List of Figures

Figure	Title	Page No.
3.1	Elements of the experiment setting	8
3.2	Data flow diagram from datasets to models and outputs	11
3.3	Data preparation process	12
4.1	Count of positives of each predictand in comparison	29
4.2	Scatter plot of 00LAT against LOW_IMPACT	30
4.3	Calibration plots of the (a) uncalibrated and (b) calibrated baseline XGBoost classifiers	33
4.4	Calibration plot of the uncalibrated baseline Extra Trees classifier	34
4.5	Calibration plots of the (a) baseline XGBoost regressor and (b) baseline Extra Trees regressor	35
4.6	Calibration plots of the (a) uncalibrated experimental XGBoost classifier and (b) its calibrated counterpart	37
4.7	Calibration plot of the uncalibrated experimental Extra Trees classifier	39
4.8	The behaviour of (a) GAM and (b) Nystroem approximation on nonlinear sample data	43
4.9	Calibration plots of the baseline GAM with sigmoid and clipping adjustments	45
4.10	Calibration plot of the experimental GAM	46
4.11	Calibration plots of the (a) uncalibrated time series classifier and (b) time series regressor	49

4.12	Calibration plot of the calibrated time series classifier	50
4.13	Calibration plots of the experimental time series classifier (a) before and (b) after calibration	52
4.14	Calibration plot of the experimental time series regressor	53
4.15	The calibration plots of the uncalibrated baseline MLP classifiers trained on (a) all TSNV data and (b) downsampled TSNV data	57
4.16	Calibration plot of the experimental MLP classifier	59
4.17	Calibration plots of the (a) default and (b) regularized ensemble models	64
4.18	Distribution of predicted probabilities of the ensemble	66

List of Tables

Table	Title	Page No.
4.1	Summary of the features present in the original baseline dataset	20
4.2	Summary of the features present in the TSNV variant of the baseline dataset	21
4.3	Proposed dynamical predictors	24
4.4	Summary of the features present in the experimental dataset	26
4.5	Performance of the baseline tree-based models	31
4.6	Performances of the experimental tree-based models in	36

comparison to the baselines

4.7	Performance of the baseline linear models	40
4.8	Performance of the experimental linear models	42
4.9	Performance of the baseline time series models	48
4.10	Performances of the experimental time series models	50
4.11	Performances of the baseline MLPs	55
4.12	Performances of the experimental MLPs	58
4.13	List of ensemble members	61
4.14	Performances of different ensemble voting mechanisms	62
4.15	Comparison between the default and regularized ridge regression ensemble mechanisms	63
4.16	Detailed performance measurements of the ensemble	65
6.1	Top 50 features for minimal impact prediction with baseline data	71
6.2	Top 50 features for limited impact prediction with baseline data	72
6.3	Top 50 features for substantial impact prediction with baseline data	73
6.4	Top 50 features for direct strike prediction with baseline data	74
6.5	Top 50 polynomial features for minimal impact prediction with baseline data	76
6.6	Top 50 polynomial features for limited impact prediction with baseline data	77
6.7	Top 50 polynomial features for substantial impact	79

	prediction with baseline data	
6.8	Top 50 polynomial features for direct strike prediction with baseline data	80
6.9	Top 50 features for minimal impact prediction with experimental data	81
6.10	Top 50 features for limited impact prediction with experimental data	82
6.11	Top 50 features for substantial impact prediction with experimental data	84
6.12	Top 50 features for direct strike prediction with experimental data	85
6.13	Top 50 polynomial features for minimal impact prediction with experimental data	86
6.14	Top 50 polynomial features for limited impact prediction with experimental data	88
6.15	Top 50 polynomial features for substantial impact prediction with experimental data	90
6.16	Top 50 polynomial features for direct strike prediction with experimental data	91

List of Abbreviations

Abbreviation	Full-Form
CLIPER	Climatology and Persistence
DT	Decision tree

ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis v5
GAM	Generalized additive model
GFS	Global Forecasting System
GP	Gaussian process
HKO	Hong Kong Observatory
HKWW	Hong Kong Weather Watch
hPa	Hectopascal
IBTrACS	International Best Track Archive for Climatological Stewardship
JTWC	Joint Typhoon Warning Center
MARS	Multivariate adaptive regression splines
mbar	Millibar
MI	Mutual information
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
TSNV	Time Series, New Variable layout

Chapter 1: Introduction

This chapter summarizes the background of this final year project (“this project”), describes the project objectives and states the current status of the project.

1.1 Background

N.B. A more comprehensive literature review is available in Chapter 2.

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 results 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, and 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. 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 reports 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 were assumed to be an appropriate proxy, similar to the HKWW's approach. Signals no. 9 and no. 10 were grouped with no. 8 because of their rarity. The predictand "direct strike" was 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.

The success of the product depends strongly on **whether a statistical-dynamical forecast is superior to a traditional forecast without dynamical data.** This is the second objective of the project. This aspect is arguably more important because research on similar topics may be encouraged by this project's success.

1.3 Current Status

A rudimentary statistical-dynamical ensemble forecaster has been successfully developed, containing the forecasting models and no auxiliary functions to ease usage. The experimentations have also justified the superiority of the statistical-dynamical approach proposed over approaches without dynamical data.

1.4 Report Outline

The remainder of this report is organized as follows: Chapter 2 is a literature review of the current research on statistical-dynamical TC forecasting and other relevant matters. Chapter 3 explains the design and structure of the forecasting product, the selection of data sources and the employed modelling techniques. Chapter 4 describes the construction and contents of the dataset and analyzes the modelling results. Chapter 5 concludes the report.

Chapter 2: Literature Review

This chapter looks into the research on TC forecasting over the last few decades, in particular research about probabilistic forecasts and notable literature on various forecasting techniques.

2.1 Climatology and Persistence Statistical Models

The traditional approach to statistical TC models is called Climatology and Persistence (CLIPER), first invented by Neumann in 1972 to predict TC movements. His strategy involves a linear regression model and painstaking manual selection among 164 potential predictors, most of which are polynomial interaction terms between best track factors [7]. Nowadays, CLIPER is no longer operational [8, pp.17-18], having been superseded by NWP. It nonetheless serves as a baseline of comparison against other models [8, pp.18] and the basis of other forecasting products, such as TC intensity [9], TC wind field radii [10-11], rainfall [12] and aerosol forecasts [13]. An example is Knaff and Sampson's research, wherein a CLIPER was first built as a control [14]

before a statistical-dynamical regression model was built and evaluated against the CLIPER [15]. The main lesson learnt is that linear models may have a place in this project and a baseline model acting as control is necessary.

2.2 Numerical Weather Prediction and Ensembles

TC track probability products generated by NWP models, which simulate the whole atmosphere following its physical rules, are quite common. For instance, the European Centre for Medium-Range Weather Forecasts (ECMWF), one of the leading institutions in NWP, generates such forecasts [16] using an ensemble [17], i.e. an NWP model is run multiple times and the outcomes are collectively considered to produce the output. A recent study demonstrated that although ensemble systems may have different strengths in different ocean basins, combining them can give better performance than using any one alone, showing that ensemble forecasts are preferable [18]. Ensembles also assist probabilistic forecasts according to several papers focusing on TC track forecasts, all of which rely on NWP models at the core [19-21]. This concludes that ensemble probability forecasts involving NWP-generated data are feasible for TC-related tasks.

2.3 Machine Learning in TC Forecasting

Machine learning as a statistical method received ever-growing attention in the past decade. Machine learning techniques have been successfully employed to evaluate TC track and landfall behaviour patterns [22-23], replace traditional statistical models to produce track forecasts [24], compute TC wind field distribution probabilities [25], analyze satellite images to assess TC damages [26], among many other applications

[27]. Decision trees and regressional models are commonplace, so is the employment of dynamical data in the analyses. For instance, Zhang et al. combined best track data with meteorological variables retrieved from NWP analysis archives in their study, in which 18 rules governing TC recurvature were identified using decision trees [22]. That indicates one possibility for this project: using decision trees and other tree-based methods to analyze dynamical data and achieve a statistical-dynamical analysis.

2.4 Strike Probability Research

Research on strike probabilities or damage estimates is also abundant. These are primarily season-wide probability forecasts, such as year-round TC landfall threat assessments [28] and annual TC activity forecasts for Fiji [29-30]. Notable research includes Wang et al.'s quantitative TC damage estimates for Hong Kong, calculated in terms of Hong Kong Dollars [31], Chin's 1977 statistics-based strike probability values which do not concern individual TCs [32], and the HKWW's TC strike probability models [6]. HKWW based their models on regression splines that are fitted to best track factors, but failed to provide means of verification (e.g. probability decision thresholds, probability calibration scores, etc.), nor are the models statistical-dynamical. As such, it can be seen that a statistical-dynamical TC strike probability forecast will be able to plug the research gap.

2.5 Miscellaneous Relevant Research

DeMaria et al. published several papers, verified by fellow experts [33-34], on the topic of harnessing randomness via Monte Carlo methods to predict TC wind speed probabilities [35-36]. Their research and [28] suggest that Monte Carlo methods may

be feasible for this project, but due to practical difficulties in specifying a suitable problem in terms of Monte Carlo methods, this possibility never came to fruition.¹ On the other hand, Li et al. considered TC positions in polar coordinates [37] instead of Cartesian (i.e. in longitude and latitude), which is appealing because the radial distance is a key factor affecting TC warning signal issuances, at least for signal no. 1 [38].

Chapter 3: Methodology

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

3.1 Experiment Settings

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). This necessitated a comparison between a statistical-dynamical dataset (“experimental dataset”²) and a control dataset without dynamical data (“baseline dataset”). The modelling methods should thus be identical whereas the input data were different. The experiment to test the hypothesis is set up as follows:

¹ Monte Carlo methods were originally promised in the Project Plan, but were abandoned by January 2022.

² The nomenclature in the Interim Report was “hybrid dataset”. It was changed to ensure coherence and minimize confusion.

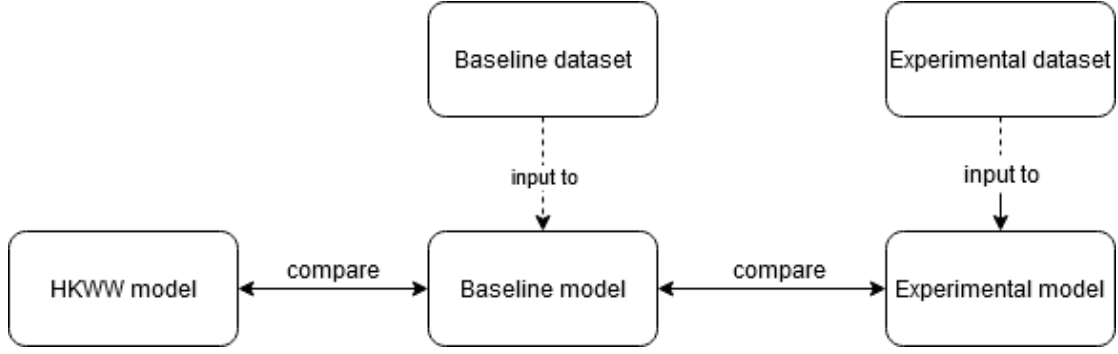


Fig. 3.1: Elements of the experiment setting

As shown in the figure, two models³ were built from the two datasets separately. The one using the baseline dataset was called the “baseline model” and the one using the experimental dataset “experimental model”. They shared an identical modelling method, such that performance comparisons were possible.

A custom baseline was needed because the modelling methods varied (see Section 3.5). To build a performant forecasting product, which is the ultimate objective of this project, the modelling method was alternated to find an optimal one. The model by HKWW [6] was used to benchmark the performance of baseline models and ascertain that they maximally realized the baseline dataset’s potential. The HKWW model was not used directly as a baseline because the experimental models did not use the same modelling method.

The models solved either a simple regression problem, where the models (regressors) produced the probability directly, or a multi-label classification problem [39], where the models (classifiers) assigned to each sample mutually non-exclusive labels that

³ To be precise, two *sets* of models (see Section 3.3).

identify them as positives or negatives. In the latter case, it was imperative that the models also produce confidence scores or class/label probabilities, which are the desired targets of this project. Both approaches were considered in this project with emphasis on the classification approach because the regressors returned values outside of normal bounds⁴. These values require treatment such as clipping or scaling so they could be converted into proper probabilities, but this transformation process would destroy potentially meaningful information carried in the out-of-bounds values (see also Section 4.3.3). In contrast, class probabilities are guaranteed to fall within the correct ranges and were deemed more viable.

3.2 Evaluation Metrics

The models, baseline and experimental, were evaluated using the following metrics:

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 [40], which respectively assess false alarm rates (high precision implies few false positives [41]) and the ability to warn impending TC threats (high recall means relevant items are better identified [41]). To summarize the performances across the four predictands, the arithmetic mean of the respective F1 scores (“macro-averaging [40]”) was taken, so that an overall comparison between models was possible. As the forecasts were probabilistic in nature, decision thresholds were found first, such that the probabilities could be converted to deterministic “threat/no threat” values, depending on whether they crossed the thresholds or not.

⁴ Probabilities should always be within the range [0, 1].

Secondly, the probabilities should be well-calibrated. That is, the probabilities must also be good confidence scores [42], such that among n predictions with the same probability p there were np items that are indeed positives. This was qualitatively measured using calibration plots and quantitatively with Brier scores [43] (the lower, the better the calibration is). Good calibration was necessary to ensure that the ensembles had credible inputs.

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 was assumed that if the F1 score of a model was close to that of the HKWW, then the direct strike performance and calibration quality were both within an acceptable range.

3.3 High-Level Design

This project explored multiple modelling methods and iteratively deepened the methodology. As a result, multiple experimental models were created, the best of which grouped as an ensemble to produce forecasts, similar to the approach taken by major NWP institutions [17] to make the final forecast more robust and comprehensive.

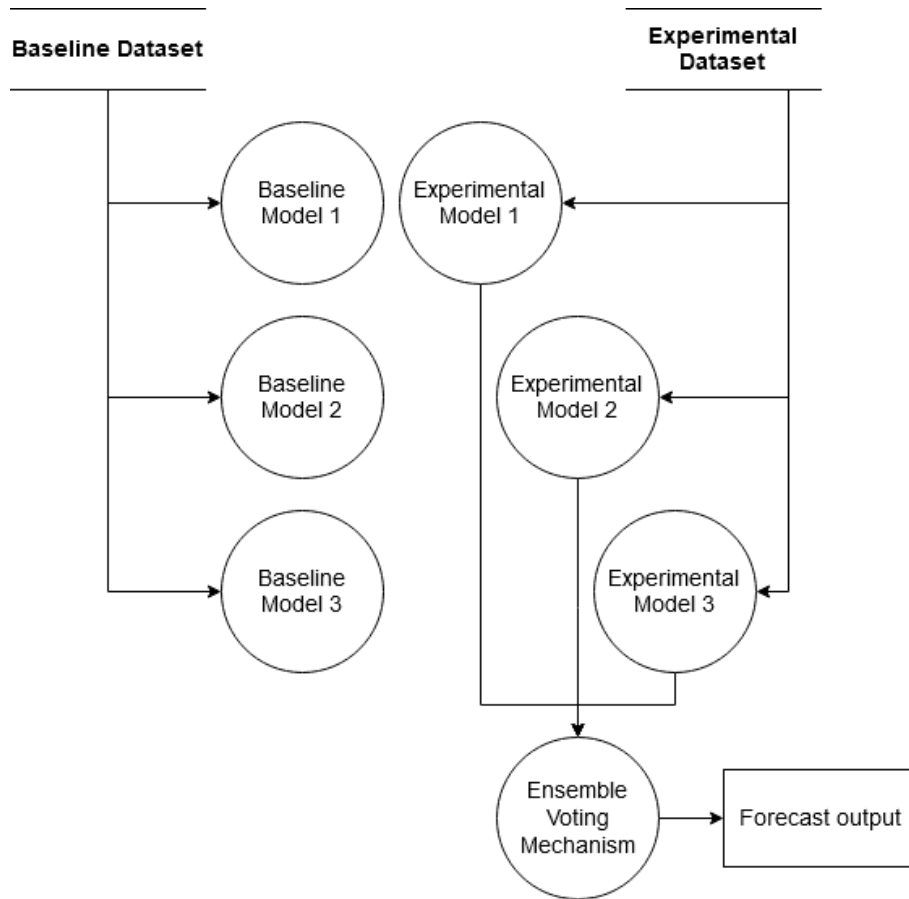


Fig. 3.2: Data flow diagram from datasets to models and outputs

In Figure 3.2, the data flow and structure of the forecasting product are shown. Multiple baseline model – experimental model pairs are present⁵, one for each proposed modelling method and only the statistical-dynamical experimental models participated in the ensemble. The voting mechanism weighed the experimental models' outputs and produced the final output. The diagram shows only three experimental models for demonstration purposes, but the exact number is different (details are in Section 3.5.6 and Section 4.7).

⁵ Multiple baselines instead of only one (as in the Project Plan) were devised. The new approach was adopted to allow for more rigorous and convincing comparisons.

The development workflow of the project was as follows: Firstly, the data sources were identified. Then, the data were obtained and preprocessed to construct datasets on which the models relied. Next, the models were built using the datasets. Finally, the models were evaluated, baselines against HKWW and experimental models against the baselines.

3.4 Data Source Selection and Dataset Preparation

To build the abovementioned models, three data sources were needed: One for TC signal issuance records, which are available online at the HKO website [44]; one for best track data, which numerous authorities provide; and one for synoptic meteorological data, generated from NWP analyses at various institutions. The selection criteria for best track datasets included data quality, time coverage, and ease of use. Similarly, the selection criteria for NWP datasets included comprehensiveness, spatial resolution, time coverage, and ease of use.

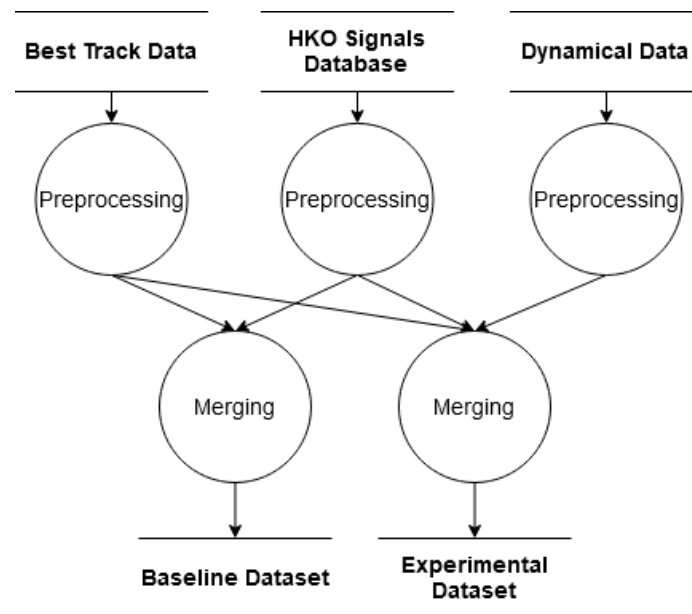


Fig 3.3 Data preparation process

Figure 3.3 summarizes the data preparation process. The data acquired from the selected data sources were separately preprocessed before they were merged to build the baseline and experimental datasets.

Data from the chosen data sources were preprocessed to perform the necessary cleaning and transformations, such as changing units, removing missing data records and computing predictor variables. Complex NWP data were simplified into only a few one-dimensional features to permit efficient statistical analysis because the NWP data are typically two-dimensional grids covering a part of the world map at an instant. The final step of the dataset preparation process was to merge the data sources. Each best track data record was matched with the corresponding dynamical variables (if needed) and impact level labels.

More details regarding data source choice, dataset preparation and dataset contents are available in Section 4.1.

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

3.5 Data Modelling Techniques Employed

Several options can be used to build forecasting models for the ensemble outlined in Section 3.3. Each technique described below is a group of modelling methods categorised together because of their common features. The individual methods of each group were iteratively tested whenever appropriate to identify the best-performing one

for evaluation.

3.5.1 Tree-Based Models

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 [45]. They are known to be useful for TC forecasting and research [22]. 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 [45]. Therefore, DTs and their variants are chosen as the first class of modelling techniques to test.

The results obtained using tree-based models are discussed in Section 4.2.

3.5.2 Linear Classification and Regression Models

This group contains linear classification and regression methods, together with non-linear extensions thereof. These techniques were chosen in reference to CLIPER-like methods (cf. Section 2.1) and their potential was justified by the appealing performance scores provided by HKWW.

Linear models 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 [46].

There were numerous candidate methods, ranging from logistic regression and linear regressions to multivariate adaptive regression splines (MARS) [47] and generalized

additive models (GAMs). MARS and GAM are regression methods similar to penalized spline regressions done by HKWW, whose exact modelling strategy could not be ascertained [6]. These models were implemented using the packages `sklearn`, `py-earth` [48] and `pyGAM` [49].

Two issues undermined the ability of linear models: Firstly, the data contained nonlinearities (see Section 4.1.4) which necessitated nonlinear spline-based models and testing with kernels [50] and kernel approximations to convert the data into linearly separable forms. Secondly, linear models do not naturally consider multiple features at once as DTs do, and thus polynomial interaction terms were to be calculated in advance, whose numerous resultant features were filtered through mutual information (MI) [51] and other techniques. This feature selection step is common to CLIPER-like methods, even for the inaugural CLIPER itself [7].

The results obtained with linear models are reviewed in Section 4.3.

3.5.3 Direct Time Series Modelling

The TC forecasting problem can alternatively be regarded as a time series-related problem since the predictands are effectively each derived from a time series of TC warning signal statuses (the *endogenous* variables) and the model's task is to make sense of the *exogenous* variables that influence them, which can also be presented as time series on their own. That is, given the TC warning signal statuses (endogenous), TC movements and synoptic conditions (both exogenous) of the past 24 hours, the model should evaluate the state of the endogenous variables in the next 72 hours. As the data were indeed in form of time series, it was considered most reasonable to utilize

this potentially helpful nature and introduce dedicated tools to aid the analysis.

The Python library `sktime` is capable of performing time series classification, regression and forecasting tasks [52-53]. It was used to explore possibilities in these three categories to perform multi-label classification, probability regression, and forecasting of the endogenous variables via extrapolation. The results obtained are detailed in Section 4.4.

3.5.4 Multilayer Perceptrons

A multilayer perceptron (MLP) is a simple neural network consisting of input, output layers and at least one intermediate hidden layer. It is the basis of all artificial neural networks and is sufficiently versatile for a variety of tasks, linear or nonlinear alike [54]. While MLPs have not been used for TC forecasting, it has been employed in other fields of weather forecasting, such as handling rainfall time series [55]. This suggests that MLPs are suitable for this project, which concerns both time series data and nonlinear data modelling. The results obtained with MLPs built using `sklearn` and `PyTorch` [56] are in Section 4.5.

3.5.5 Gaussian Processes

Gaussian processes (GPs) are an alternative time series modelling method. GPs allow for reasoning about a posterior distribution (i.e. the endogenous variables) given observations of a related prior distribution (exogenous variables) and can model time series where data mean and variance values change over time [57]. This makes GPs suitable for TC time series modelling and regression as well, even though there have been no antecedents. GPs were implemented using `sklearn` and the third-party package `GPyTorch` [58]. The results obtained are described in Section 4.6.

3.5.6 Ensemble Design

The ensemble consisted of the best four to five experimental models created using the techniques described above. The voting mechanism, which received individual experimental models' probability predictions and returned a unified average value, was implemented using an approach called stacking [59]. Stacking involves fitting an additional estimator – either a classifier or a regressor – that fits the input data to the desired output, thereby saving the effort of manually seeking optimal weights for each input. The ensemble members needed to produce well-calibrated probabilities in the first place, otherwise, the final results would be poorly calibrated and thus provide little predictive value. A few estimator candidates such as linear regression and logistic regression were tested and the best was chosen as the finalized stacking estimator. The results obtained are discussed in Section 4.7.

3.6 Limitations and Exclusions

This project has two primary limitations by nature.

Firstly, the HKO TC warning signal system is assumed to be an appropriate replacement for accurate impact level measurements but 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 [60].

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 4: Results and Analysis

This chapter reviews the results obtained during the execution of the aforementioned project methodology and includes analyses of the outcomes.

4.1 Dataset Construction

4.1.1 Data Source Selection and Data Preprocessing

The HKO warning signal records were retrieved and processed in early October 2021. The records were downloaded from the HKO website [44] 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 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.

There were three candidates for best track data, namely the dataset of the Joint Typhoon Warning Center (JTWC) [61], the *International Best Track Archive for Climatological Stewardship* (IBTrACS) dataset [62-63], and that of the China Meteorological Administration [64-66]. The IBTrACS dataset was chosen because it contains both TC names and the well-validated and well-documented JTWC dataset as its subset, which could be extracted from within the IBTrACS dataset. When the IBTrACS dataset was processed in late October, the main difficulty encountered was keeping track of the numerous variables present, most of which are unnecessary. Documentations had to be made and updated continuously to help identify variables.

There were two main options for the dynamical data source, namely the *ECMWF Reanalysis v5* (ERA5) dataset [67] and the *NCEP FNL Operational Model Global Tropospheric Analyses* (NCEP FNL) dataset [68]. 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.

4.1.2 Baseline Dataset Details

There were three baseline dataset variants.

The first was created in early November 2021 by merging the processed HKO records and IBTrACS records, taking all the available records. 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 original baseline dataset, which had 45293 samples, are as follows, as it was first built in November:

No.	Column name	Description
0-2	MM, DD, HH	The timestamp of the record; corresponds to the last TC position/intensity sample in the time series.
3	LOW_IMPACT	The targets to predict, i.e. the correct labels of the record. The names were changed to shorten the column names.
4	MID_IMPACT	
5	BIG_IMPACT	
6	DIRECT_STRIKE	
7	00LAT	Latitude (degrees N) of the TC centre position, at the time described in columns 0-2.
8	00LON	Longitude (degrees E) of the TC centre position, at the time described in columns 0-2.
9	00WIND	Maximum sustained wind speed (knots) at the TC centre at the time described in columns 0-2.
10-12	06LAT, 06LON, 06WIND	Position and intensity 6 hours before the given timestamp.
13-15	12LAT, 12LON, 12WIND	The extension of the previous 6 columns (two timesteps) until 24 hours back.

16-18	18LAT, 18LON, 18WIND	
19-21	24LAT, 24LON, 24WIND	

Table 4.1: Summary of the features present in the original baseline dataset

As shown in the table, TC trajectories and intensity evolutions were 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) were excluded, assuming that the models could still function well without them. There were two notable intricacies: The time series lasts for only 24 hours, so forecasting newly formed TCs will be possible; the changes in TC warning signal status over the 24 hours were not included, i.e. the target endogenous variables were not present in the input data to the models.

After the first tests were conducted in the subsequent weeks to build baseline models, it was noted that the HKWW includes the endogenous variables in their inputs and the nonlinearity of the data prevented effective linear models from ever being built. A revision was thus made in January 2022 to alter the predictor variable layouts and explicitly encode the time series of both endo- and exogenous variables in the dataset. The resulting dataset was given the short name “TSNV” (Time Series, New Variable layout) and consisted of the following columns:

No.	Column Name	Description
0-4	LOW_IMPACT, MID_IMPACT,	The targets to predict, equivalent to columns 3-6 of the original dataset.

	BIG_IMPACT, DIRECT_STRIKE	
5-7	MM00, DD00, HH00	Timestamp at which the following several columns' values were taken.
8-11	MI_STATUS00, LI_STATUS00, SI_STATUS00, DS_STATUS00	The state of the endogenous variables at the time specified in columns 5-7. The predictand short names were not simplified as in columns 0-4.
12	DIST00	Radial distance (in km) of the TC from Hong Kong.
13	AZM00	The azimuth of the TC from Hong Kong, in true bearing.
14	SPEED00	TC movement speed in knots.
15	DIR00	TC heading bearing, measured in true bearing.
16	VMAX00	Storm intensity in knots, same as column 9 in the original dataset.
17	DVMAX00	Change in TC intensity over the past 6 hours.
18-30	MM06, ..., DVMAX06	The same features, but the values were taken 6 hours before the time specified in columns 5-7.
31-43	MM12, ..., DVMAX12	The same features as columns 5-17, but the values were taken 12 hours ago.
44-56	MM18, ..., DVMAX18	The same features as columns 5-17, but the values were taken 18 hours ago.
57-68	MM24, ..., VMAX24	The same features as columns 5-16, but the values were taken 24 hours ago. DVMAX24 was absent.

Table 4.2: Summary of the features present in the TSNV variant of the baseline dataset

As shown in the table, the features in the original baseline dataset were either rearranged into explicit time series (for the date and intensity parameters) or presented in a new way (position parameters), before new features were added (namely the time series of the predictands, TC movement and intensity change features). Many of these columns were irrelevant (such as rearranging the time information into a series) but were kept nonetheless to maintain a uniform dataset structure.

The TSNV variant proved better than the original dataset. The best baseline tree-based model at the time was a multi-output XGBoost classifier (see also Section 4.2) and the new dataset helped boost the average f1-score from 0.797 to 0.800.

In late March 2022, while experimental models were being built in earnest, it was suspected that the seemingly superior performance of the experimental dataset was due to the dataset being smaller in size. Therefore, the TSNV variant was downsized to match the experimental dataset by only using data dated 2008 or later. The dataset now had 6957 samples only, with the same columns as before. This third variant was called “downsampled TSNV”⁶.

4.1.3 Experimental Dataset Details

The desired dynamical features were first identified through a brief study of related literature in January 2022. The following 19 items were found to be suitable:

⁶ A misnomer because there was no sampling, but rather a filtering.

No.	Predictors	Dynamical Data Needed
1-2	Vertical wind shear magnitudes [15, 24, 69-70], between 200-850 mbar ⁷ levels and between 500-850 mbar [15]	u-, v-components of winds ⁸
3-4	Relative humidity [15, 24, 69-70] at 750-850 mbar and 300-500 mbar levels [15]	Relative humidity
5-6	Temperature closest to surface (1000 mbar) [15, 69-70, 73] and at 200 mbar [15]	Temperature
7-12	Hong Kong surface (1000 mbar) winds, 200 mbar zonal (u) wind [15, 69, 73], 500 mbar winds [24, 74], summer monsoon index (850 mbar zonal) [22, 75-76]	u-, v-components of winds
13	850 mbar relative vorticity [15, 24, 29]	Absolute vorticity
14-17	Westerly index [77]; Western North Pacific subtropical high area, intensity, and westward extension indices [78]	500 mbar geopotential height [22, 24, 74]
18	925 mbar potential temperature [15]	Potential temperature

Table 4.3: Proposed dynamical predictors

As shown in the table, the 18 predictors⁹ were derived from six dynamical data variables. The corresponding data archives of the six variables were downloaded to

⁷ Millibar (mbar) and hectopascal (hPa) are by definition exactly equivalent [71]. Literature tended to use mbar, while NCPE FNL used hPa, hence the mixture of nomenclature in this report.

⁸ The u-component (zonal flow) is parallel to the equator whereas the v-component (meridional flow) is parallel to longitude lines [72].

⁹ The Interim Report states 19, which is incorrect.

compute the predictor values. While some indices were well-defined, the other predictors required custom calculations. This was 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 [15] and a 7-degree average for 200 mbar temperature [69].

The data downloads finished in January. It was deemed most computationally efficient to calculate all predictors values in advance, instead of only opening the data files during the merging step. Therefore, for each 6-hour time step between August 1999 (when NCEP FNL data were first available) and each TC present in the best track data records, the feature values were precomputed by definition and saved to files, so that they could be reused later. The processing faced several difficulties:

Firstly, there had been significant difficulty in opening the downloaded files because they did not exactly follow the standard formats but using the PyNIO package [79] provided by the maintainers of the NCAR Command Language [80], a programming language commonly used in the meteorological community, the files could be opened.

A second difficulty arose as some predictors could not be calculated by definition. The relative vorticity values found were much lower than expected real-life values and absolute vorticity was kept instead; the data for 925 hPa potential temperature were not available, whereas that for sigma level 0.995¹⁰ was available, so it was directly used.

A third difficulty was the irregular sampling in the data source. Most of the dynamical

¹⁰ This refers to the level in the atmosphere where the air pressure is 99.5% that of the surface [81].

data records were taken at times such as 6 a.m., 12 noon, 6 p.m., and midnight, such that they were taken at the same time TC positions and intensities were reported, but a small number of those were taken at 9 a.m. and 3 p.m. instead, which led to missing data problems on paper. The solution was to use interpolation to estimate the values at, for example, noontime using data from 3 hours ago and 3 hours later.

The fourth, final and most vexatious difficulty was encountered when the data were checked for missing values. In theory, there should be 29220 data files for the period between 1998 and 2019, but typically only 18000 were present and in the worst case, only 8422 records could be used. It was found that there were numerous missing records in the data *source*, which could not be compensated for by changing to other data sources (e.g. ERA5) because the data quality and formats were not compatible. As there were no simple solutions to this problem, it was decided that the data processing continue, and a disappointingly small dataset would be tolerated.

The merging step was the same as that of the baseline datasets. Based on the TSNV variant, the dynamical predictors were added to the preexisting columns in the following manner:

No.	Column name	Description
0-3	MINIMAL_IMPACT, ..., DIRECT_STRIKE	The targets to predict, same as columns 0-3 in the baseline TSNV dataset.
4-5	MM00, DD00	The date on which the following several columns of data were taken.
6-9	MI_STATUS00, ...,	State of the endogenous variables. Same as

	DS_STATUS00	columns 8-11 in the baseline TSNV dataset.
10-15	DIST00, ..., DVMAX00	The exogenous variables present in the baseline TSNV dataset (columns 12-17).
16-17	ULVWS00, MLVWS00	Upper-Lower level Vertical Wind Shear and Mid-Lower level Vertical Wind Shear (in ms^{-1}); predictors 1-2 in Table 4.3.
18-19	HI_HUMID00, LO_HUMID00	Upper-level and lower-level relative humidity averages (in %); predictors 3-4.
20-21	STEMP00, UTEMP00	Surface and upper-level temperature (in K); predictors 5-6.
22-23	U_HK00, V_HK00	Surface u- and v-components of winds at Hong Kong (in ms^{-1}); predictors 7-8.
24-26	U20000, U50000, V50000	Winds at different levels (in ms^{-1}); predictors 9-11.
27	EASM00	East Asia Summer Monsoon index (in ms^{-1}); predictor 12.
28	VORT00	850 hPa vorticity average (in s^{-1} , scaled by 10^6 to avoid underflowing); predictor 13.
29	WESTERLY00	Westerly index (in gpm); predictor 14.
30-32	SH_AREA00, SH_INT00, SH_EXT00	Subtropical high area (no unit), intensity (gpm) and westward extension (degree longitude) indices; predictors 15-17.
33	POTT00	Potential temperature average (in K); predictor 18.
34-63	MM06, ..., POTT06	The same as the previous 30 features, but

		from 6 hours earlier.
64-93	MM12, ..., POTT12	The same as columns 4-33, but from 12 hours ago.
94-123	MM18, ..., POTT18	The same as columns 4-33, but from 18 hours ago.
124-152	MM24, ..., POTT24; no DVMAX24	The same as columns 4-33 (with DVMAX omitted), but from 24 hours ago.

Table 4.4: Summary of the features present in the experimental dataset

Several observations can be made from the table. The hour information was removed due to its high irrelevance. The new dynamical features counted 90, so the total number of columns ballooned to 153. This significantly larger column count created practical difficulties for some of the models (see following sections for more details). This combined experimental dataset counted 6955 samples, which was merely 15.3% that of the original baseline dataset. This significantly reduced size would cause issues later on, especially with MLPs (Section 4.5), but data augmentation by synthesizing extra samples was impossible because that would require a generator that correctly understood the complexity and distribution of the TC data, which was an intricate topic on its own.

4.1.4 Notable Dataset Idiosyncrasies

Several peculiarities were identified during the project, and they affect the analyses in the upcoming sections. These particularities are as follows:

Firstly, the dataset was grossly unbalanced, regardless of type and variant. All baseline

dataset variants and the experimental dataset contained only 11% positives, i.e. if all samples that contained at least one positive in the four predictands were counted, they constituted only 11% of the total samples count. This consistent ratio suggests that there may have been little change in TC climatology over the past 60 years.

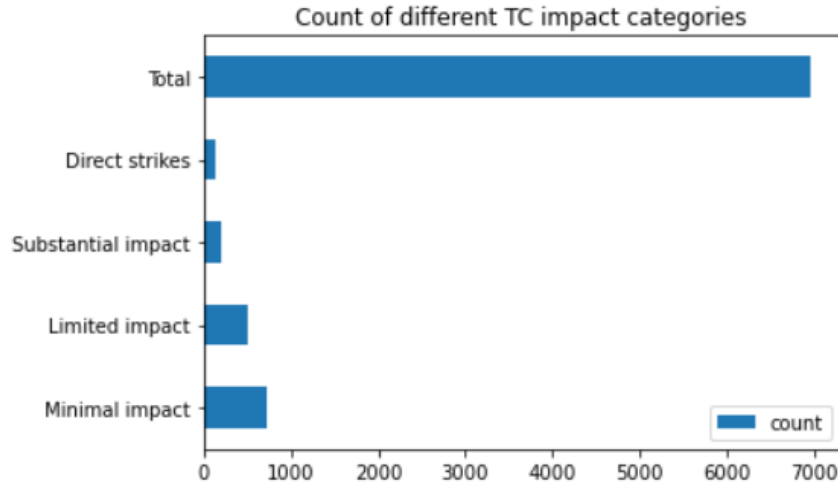


Fig. 4.1: Count of positives of each predictand in comparison

Figure 4.1 shows the imbalance in detail by showing the exact positives counts of each predictand, using data from the experimental dataset. While positives of any kind are scarce, the rarity increases with impact level and direct strikes are the least common. The distribution of positives of each class was also consistent across datasets.

The main problem caused by the imbalance was that the margin between under- and overfitting would be very small. That is, with about 88% of the data being negatives, a naïve forecaster predicting all negatives will achieve 88% accuracy and models improve slowly because positives can simply be regarded as noise; on the flip side, models that have high scores are very prone to overfitting, because the small number of positives to model may underrepresent the underlying variance of TC behaviour. As

shown in the rest of this chapter, the majority of the models were able to avoid excessive overfitting.

The second issue was that the relationship between the predictand and the features was not linear. For instance, it is more likely to encounter TC strikes in the summer months than any other, because more TCs may spawn in that period. The following figure shows the relationship between TC position longitude (00LAT) and minimal impact probability (LOW_IMPACT) in the full baseline dataset.

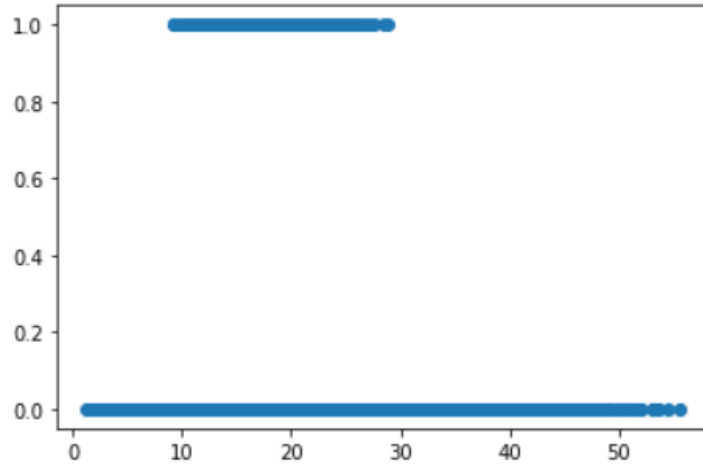


Fig. 4.2: Scatter plot of 00LAT (horizontal axis) against LOW_IMPACT (vertical axis)

Two notable characteristics are present: Firstly, for some values of 00LAT, there are simultaneously two LOW_IMPACT values (0 and 1), indicating that just latitude alone was far from sufficient for a linear model to distinguish between positives (1) and negatives (0). Secondly, if only the positives are considered, then the relationship will have a bell-like shape, which cannot be modelled using linear methods either.

This issue implied that linear models and regression models would have difficulty

fitting to the data. A good regression model must return values strictly within $[0, 1]$ while being well-calibrated, i.e. a model that mostly predicts values close to 0.5 (the mean between 0 and 1) has little utility; while a good linear model must be able to adapt to the nonlinear nature of the data. Sections 4.3 and 4.6 will see the repercussions of these matters manifest.

4.2 Results with Tree-Based Models

Over the course of several months, a large variety of tree-based models were built and tested, including simple single DTs, random forests, highly randomized forests called Extra Trees (can be classifiers or regressors), and XGBoost gradient boosting DTs. It was found that classifiers and regressors had similar performances and XGBoost models and Extra Trees had the best performance in general.

4.2.1 Baseline Models

The following table summarizes the performance of the best baseline tree-based models, compared against the HKWW model scores.

Dataset	Model	Average F1 Score ¹¹
N/A	HKWW (Control)	0.84500 ¹² [6]
Original	XGBoost Classifier	0.78352
Original	XGBoost Regressor	0.65818
Original	Extra Trees Classifier	0.70872
Original	Extra Trees Regressor	0.75880

¹¹ Correct to 5 decimal places unless otherwise specified.

¹² This score does not consider direct strike, which the HKWW model does not predict.

TSNV	XGBoost Classifier	0.81360
TSNV	XGBoost Regressor	0.81553
TSNV	Extra Trees Classifier	0.80852
TSNV	Extra Trees Regressor	0.80149
Downsampled	XGBoost Classifier	0.87635
Downsampled	XGBoost Regressor	0.90432¹³
Downsampled	Extra Trees Classifier	0.88469
Downsampled	Extra Trees Regressor	0.89117

Table 4.5: Performance of the baseline tree-based models

As shown in the table, the performance of the baseline models improved with each dataset revision, with both models using the downsampled TSNV dataset successfully outperforming the HKWW model. The bolded value was the best value obtained and it clearly shows that the baseline tree-based model was capable of fully exploiting the potential of the dataset. Moreover, classifiers and regressors had similar performances at large and oftentimes regressors outperformed classifiers to some extent. It must also be noted, however, that none of these models was deliberately calibrated to obtain better probability calibration because they all had very low Brier scores (typically under 0.03) and calibration tended to disrupt model behaviour.

¹³ Bolded values in these tables indicate the best-scoring model.

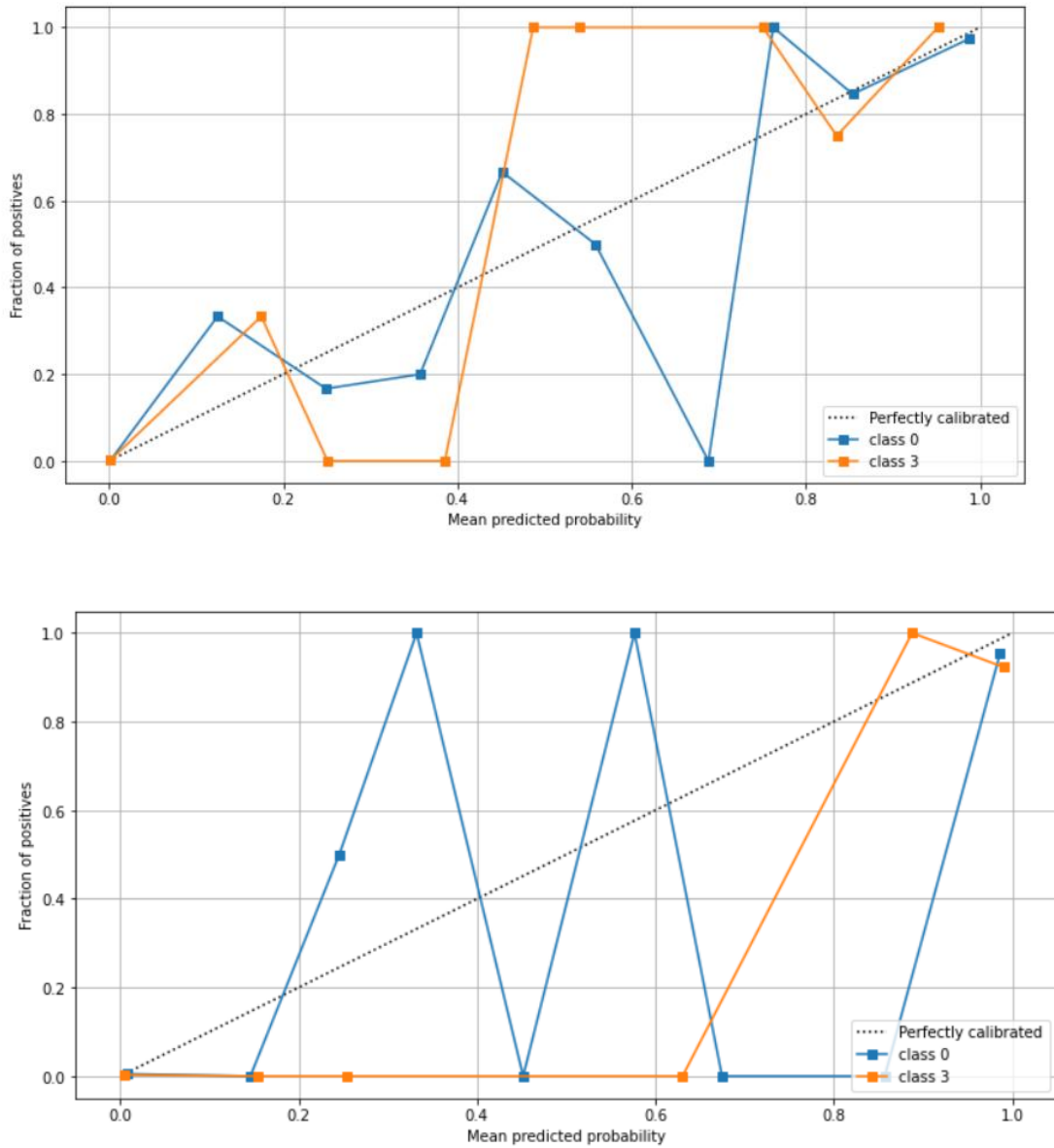


Fig. 4.3: Calibration plots of the (a) uncalibrated (top) and (b) calibrated (bottom) *baseline*¹⁴ XGBoost classifiers

For example, in Figure 4.3a, the calibration plot of the XGBoost classifiers on the downsampled dataset is shown and although the lines are not well-aligned to the diagonal, which indicates perfect probability calibration, they are still less abnormal

¹⁴ Unless otherwise specified, the “baselines” refer to models built using the downsampled baseline dataset.

than those plotted after calibration which are shown in Figure 4.3b. Furthermore, the probability calibration of Extra Trees classifiers tended to be decent without extra postprocessing, as shown in Figure 4.4.

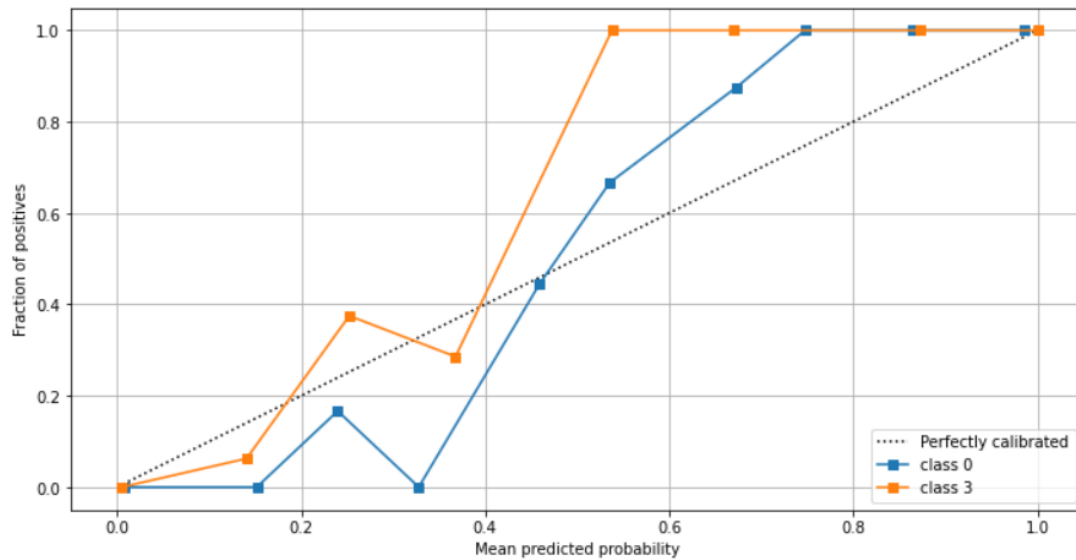


Fig. 4.4: Calibration plot of the uncalibrated baseline Extra Trees classifier

The figure shows the calibration plot of the Extra Trees classifier model on the downsampled dataset. The lines are much closer to the diagonal. Regressors also showed satisfactory calibration:

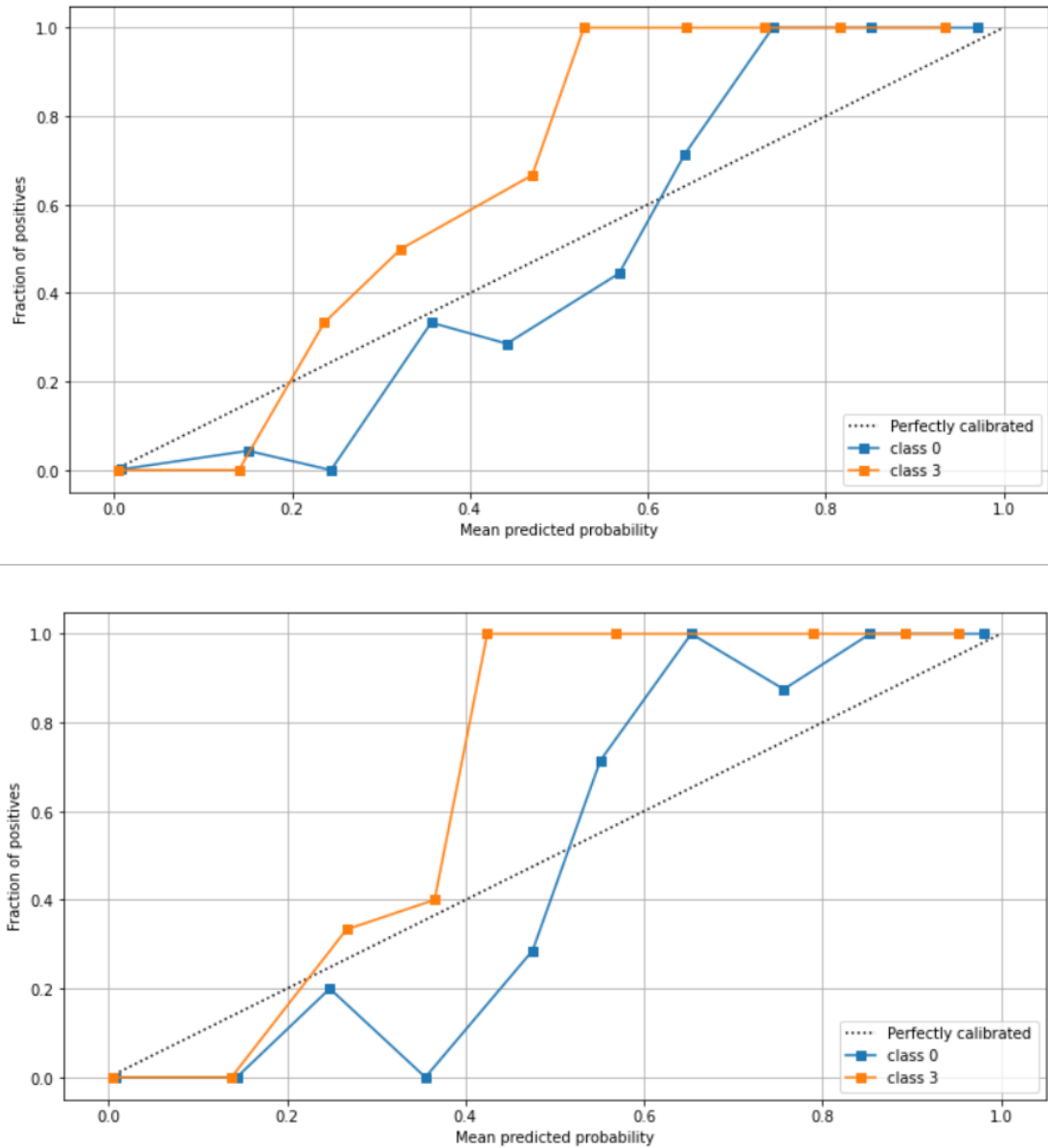


Fig. 4.5: Calibration plots of the (a) baseline XGBoost regressor (top) and (b) baseline Extra Trees regressor (bottom)

As shown in Figure 4.5, the calibration of both regressors was good, especially for the blue lines (indicating minimal impact) that are acceptably close to the diagonal. Comparing the XGBoost models (Figures 4.3a and 4.5a), it is clear that the regressor version was better; whereas for the Extra Trees models (Figures 4.4 and 4.5b), the classifier had marginally better calibration.

All evidence considered, it was decided that tree-based models did not require calibration.

4.2.2 Experimental Models

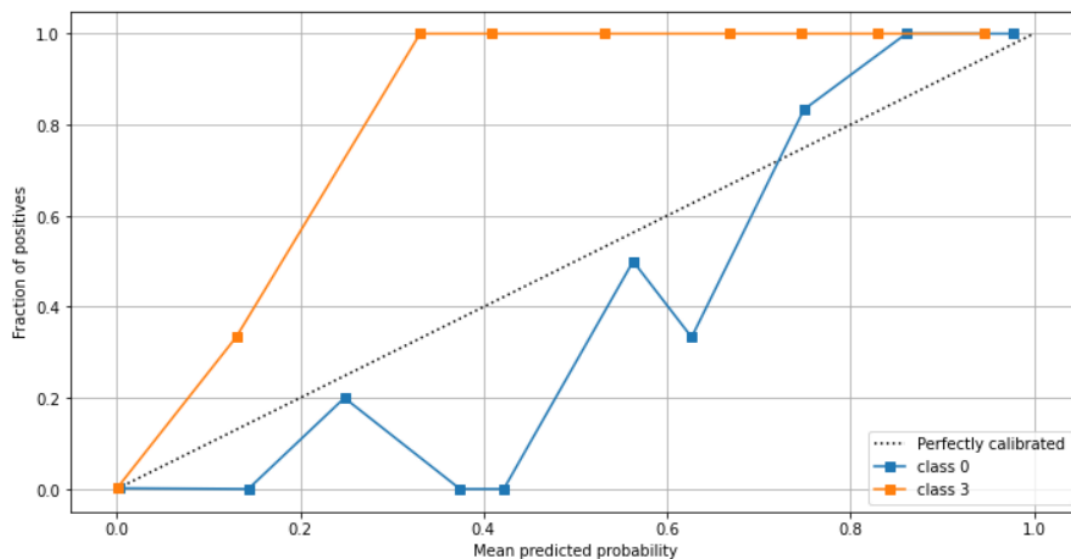
The methods that had worked best for the baseline models were reused for the experimental models and the results are as follows:

Dataset	Model	Average F1 Score
N/A	HKWW	0.84500
Downsampled baseline	XGBoost Classifier	0.87635
Downsampled baseline	XGBoost Regressor	0.90432
Downsampled baseline	Extra Trees Classifier	0.88469
Downsampled baseline	Extra Trees Regressor	0.89117
Experimental	XGBoost Classifier	0.92765
Experimental	XGBoost Regressor	0.87246
Experimental	Extra Trees Classifier	0.89849
Experimental	Extra Trees Regressor	0.89823

*Table 4.6: Performances of the experimental tree-based models in comparison to the
baselines*

In Table 4.6, the performances of the experimental tree-based models (at the bottom) and notable baseline models of the same type (at the top) are listed. From the table, one can notice that the new statistical-dynamical dataset did improve the scores but not by much. For instance, the Extra Trees models had only improved by about 0.01. The two XGBoost models that had more than a 0.9 F1 score could be seen as coincidental outliers as most models had an average score of about 0.88-0.89. In any case, the HKWW model was once again surpassed, which was a desirable outcome. The models built in this project may have slight overfitting, but cross-validation hyperparameter optimization was carried out to reduce the chances of overfitting, so it was not a significant concern.

As with the baseline models, the tree-based experimental models did not require additional calibration, even though the original calibration-level was second-rate.



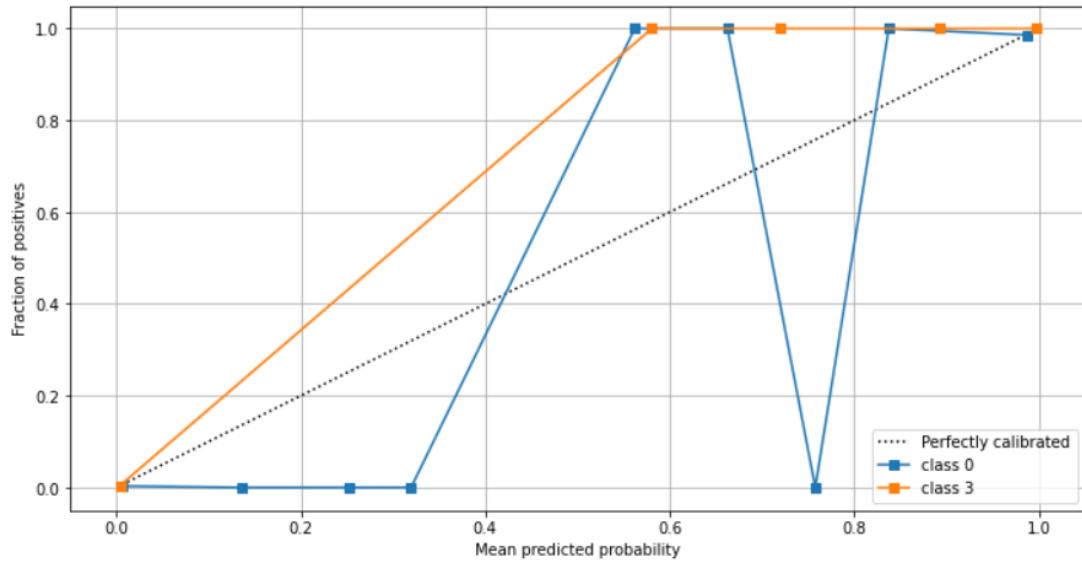


Fig. 4.6: Calibration plots of the (a) uncalibrated experimental XGBoost classifier (last page) and (b) its calibrated counterpart (above)

Figure 4.6 shows that despite the mediocre calibration of the original XGBoost model, where the line plots were not closely aligned to the perfect diagonal (Fig. 4.6a), the calibrated version (Fig. 4.6b) was much less usable because the predictions produced after calibration were aberrant, such that there were fewer points plotted near the diagonal than before. The regressor version had a near-identical calibration performance, but since its predictions were less accurate, the classifier version was favoured.

Without calibration, the Extra Trees classifier was already quite suitable for use:

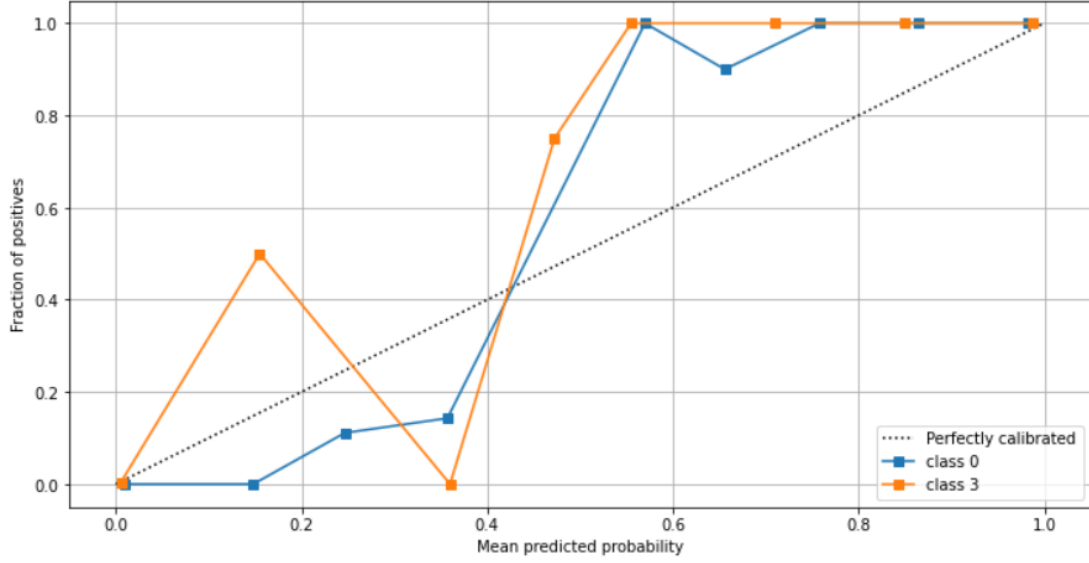


Fig. 4.7: Calibration plot of the uncalibrated experimental Extra Trees classifier

The two lines in the figure were closely aligned to each other, though the probabilities for direct strike (class 3, orange line) tended to be overestimated. The regressor version had comparatively worse calibration and higher Brier scores in two out of four categories, otherwise, the two versions would be indistinguishable.

4.2.3 Discussion

There was relatively little difficulty in the development of tree-based models, and they had promising scores, thanks to DTs' ability to partition input spaces to model nonlinearities. However, the results showed that statistical-dynamical models may not be significantly better than traditional methods. Furthermore, the impossibility of probability calibration implied that the ensemble (Section 4.7) would have to adaptively adjust the mildly biased input probabilities provided by the upstream tree-based classifiers. The two best experimental tree-based models, namely the XGBoost classifier and the Extra Trees classifier, were admitted to the ensemble to utilize their good scores and calibration. Their regressor versions were added to the ensemble later

on to test the ensemble’s upper limits.

4.3 Results with Linear Models

The work on the linear models was far more strenuous than on tree-based models. From January to April 2022, a wide array of methods was tested, including linear regression, logistic regression, MARS, and GAMs for the models; Nystroem kernel approximation and polynomial feature calculation for data transformations; and principal component analysis (PCA), ANOVA F-tests and mutual information (MI) scoring for feature selection. Regardless of dataset choice, the majority of the models struggled to pass an average F1 score of 0.6, presumably owing to the complexity of the data as demonstrated in Section 4.1.4.

4.3.1 Baseline Models

Developing a performant linear model was a difficult challenge. The following table summarizes the performances of the most notable baseline linear models, in comparison with the HKWW model:

Dataset Variant	Model	Average F1 Score
N/A	HKWW (Control)	0.84500
Original	MARS	0.31688
Original	GAM	0.44999
Original	PCA-selected polynomials, Nystroem-approximated polynomial kernel, logistic regression	0.40031
Original	Polynomial kernel on all polynomial	0.43500

	features, logistic regression	
TSNV	Logistic regression	0.54243
TSNV	MARS	0.52897
TSNV	GAM with MI-selected polynomial features	0.60176
Downsampled	GAM with MI-selected polynomial features	0.75851

Table 4.7: Performance of the baseline linear models

From the table, it can be concluded that none of the linear models was able to get close to the HKWW model's performance, even though performance improved with each dataset revision. In fact, the scores reported by HKWW were suspicious in that the details about the spline models used were not disclosed and the scores could not be replicated using a GAM with similar settings. It can also be observed in the table that the model types are inconsistent across dataset variants. This is because modelling techniques that had proven to be ineffective were never retried to save time.

It was also important to bear in mind that not all modelling methods could be reused for a different input dataset. The number of polynomial features generated grew exponentially with the number of base features, so although this project only used degree 3 polynomial features, the size of the polynomial transformed TSNV data grew beyond the total amount of memory available. This made it impossible for PCA, kernel methods and most feature selection methods to function; it was infeasible to fit a GAM on all the polynomial features either. However, since it was found that the GAMs considered data collected at $t=0h$ or $t=-6h$ (i.e. columns 5-30 of the TSNV dataset) most

relevant, all work on feature selection and GAMs as shown in Table 4.7 was done using the polynomial features generated from these features only.

4.3.2 Experimental Models

Only the modelling methods that produced non-trivial results with the baseline datasets were kept and reused for the experimental models. The results found are as follows:

Dataset	Model	Average F1 Score
N/A	HKWW	0.84500
Baseline TSNV	Logistic regression	0.54243
Baseline TSNV	GAM	0.60176
Downsampled baseline TSNV	GAM with MI-selected polynomial features	0.75851
Experimental	Logistic regression	0.56949
Experimental	Linear regression	0.58693
Experimental	GAM with MI-selected polynomial features	0.81090

Table 4.8: Performance of the experimental linear models¹⁵

From the table it can be observed that the experimental models were all superior to the baseline models, once more justifying that a statistical-dynamical forecasting model is better than a model without consideration of dynamical predictors. The scores obtained were still not better than HKWW's, but this is secondary to the baseline-experimental dataset comparison. MARS was not tested again because the software package had a

¹⁵ Unless otherwise specified, all experimental model performance tables henceforth will include notable baselines for comparison.

conflict with sktime and was abandoned, but it was surmised that it would not perform better than GAMs.

4.3.3 Discussion

Despite the apparently underwhelming scores, the best linear model was accepted into the ensemble (Section 4.7) because it added variety to the ensemble members by representing a non-tree-based methodology. Moreover, by the end of this chapter, it will become evident that the experimental GAM also represented the average forecasting model.

Let us now discuss two noteworthy matters that influence linear models' behaviour.

Firstly, as anticipated in Section 4.1.4, the nonlinearity prevalent in the data prevented most linear models from getting good scores. The most performant model, the GAM, overcame this issue using splines, which approximate the complex relationships between inputs and outputs relatively well.

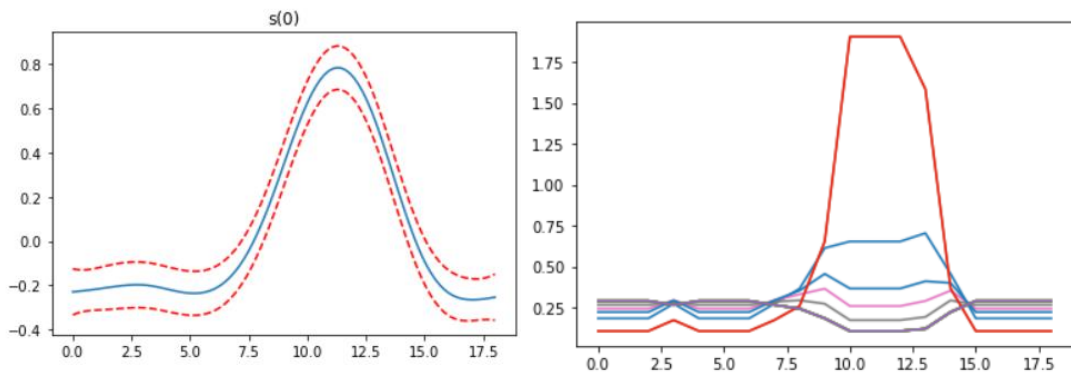


Fig. 4.8: The behaviour of (a) GAM (left) and (b) Nystroem approximation (right) on nonlinear sample data

Figure 4.8 shows how different models handle nonlinear data (corresponding to the blue line in Fig. 4.8a). The GAM shown in Fig. 4.8a can both model the data closely and provide a confidence interval estimate (red dotted lines), which was unused in this project due to the large number of features to track. The data transformed using Nystroem-approximated polynomial kernels are shown in Fig. 4.8b as coloured lines, each corresponding to one component of the kernel. These lines are still not possible to model with a fully linear approach, i.e. without using splines. This test justifies the capability of GAMs in modelling the nonlinear dataset used in this project.

While models like GAM can fit well to individual features, they cannot consider multiple features at once as DTs do. Therefore, polynomial features were introduced so that interaction between features could be represented, even though it did not improve the lack of linear separability in the data. The usage of polynomial features came at the cost of increased memory use and prolonged model fitting times, which were usually tolerable.

The second notable matter is related to postprocessing and calibration. The experimental and baseline GAMs were regressors. Thus, a choice about the handling of out-of-bound values (see Section 3.1) should be made. There were two candidates: to simply clip those values at the nearest bounds (i.e. $adjusted_y_hat = \max(0, \min(1, y_hat))$) as implemented by NumPy [82]) or pass all outputs through a logistic sigmoid function to limit their ranges. While F1 scores did not differ for either choice, its effect on probability calibration was different, especially for linear models.

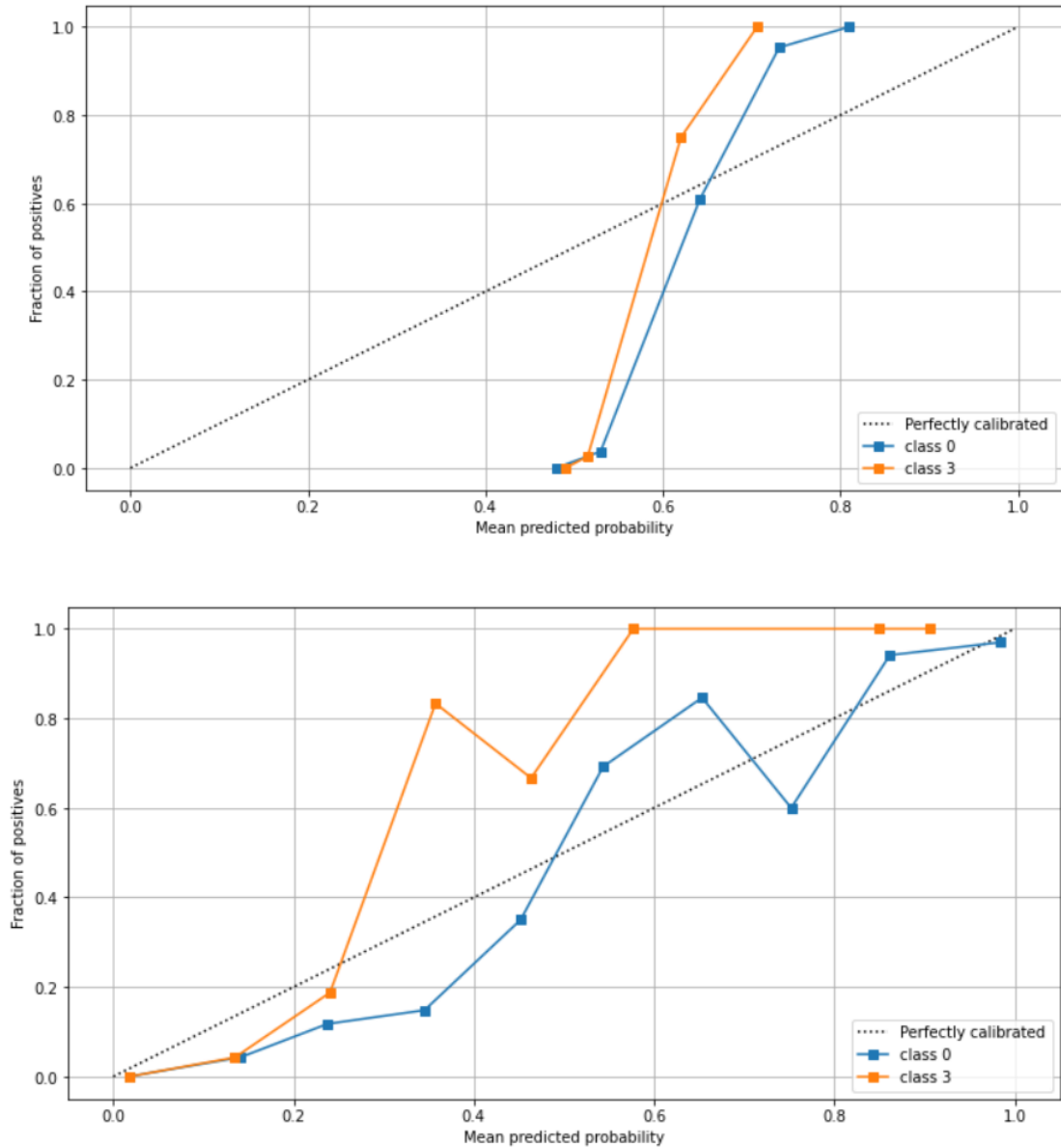


Fig. 4.9: Calibration plots of the baseline GAM with sigmoid (top) and clipping (below) adjustments

As shown in the figures, clipping helped the model obtain a better probability calibration, whereas sigmoid made the model outputs cluster around 0.5-0.8. The superiority of clipping was demonstrated also via the Brier scores, which were around 0.01-0.03 for clipping adjustment but 0.25 for sigmoid adjustment. Clipping helped the experimental GAM obtain a satisfactory calibration too:

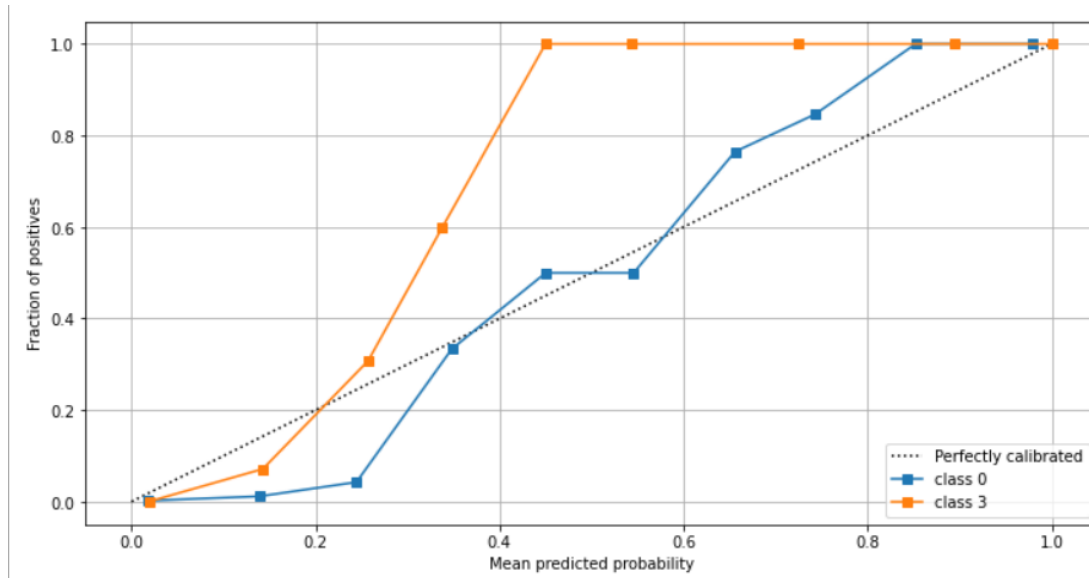


Fig. 4.10: Calibration plot of the experimental GAM

The figure shows the calibration of class 0 (first predictand, i.e. minimal impact) was remarkable because the blue line was very close to the perfect diagonal. The calibration of the fourth predictand (orange line) was still passable, but the room for improvement was obvious. Unfortunately, GAMs (and regressors in general) do not support probability calibration since it is only defined for classifiers [83], so there was no way to better the calibration.

This project suffered a huge crisis when this distinction was discovered in April 2022 because it necessitated the re-testing of numerous models using clipping adjustment. Due to the limited time, only the promising regressors (belonging to any model type) were refitted and re-tested. As mentioned before, it only affected calibration and thus the choice of ensemble members, which should have good calibration (see Section 3.5.6).

4.4 Results with Direct Time Series Modelling

There were three main modelling options under this category as stated in Section 3.5.3, namely time series classification, regression, and forecasting. They were tested from January to March 2022 in that order, and it was found that only one DT-based time series classification model was practical.

4.4.1 Baseline Models

The numerous classification models offered by the software package sktime are categorized based on the algorithm type [84] and at least one model was tested for each type. They typically either took more than several hours to fit or demanded more memory than available and were thus considered impractical. The only practical classifier was a DT-based model called Time Series Forest Classifier. On the other hand, sktime only provides one time series regressor (Time Series Forest Regressor) [85] which had a near-identical score when compared to the classifier version.

Endogenous variable forecasting was found to be unworkable. The forecasters provided by sktime include a naïve forecaster, trend forecasters which could be understood as extrapolations, and exponential smoothing forecasters [86]. They forecast based on a given time series of the endogenous variables and the exogenous variables are optional. The results obtained using forecasters were poor. Three representative examples of their performances are shown in the upcoming table.

The following table summarizes the results obtained with time series modelling:

Dataset	Model	Average F1 Score
N/A	HKWW (control)	0.84500
Original	Time Series Forest Classifier	0.72757
Original	Time Series Forest Regressor	0.70487
Original	Naïve Forecaster	0.36100
Original	Polynomial Trend Forecaster	0.35937
Original	Theta Forecaster ¹⁶	0.35962
TSNV	Time Series Forest Classifier	0.68631
TSNV	Time Series Forest Classifier	0.62575
Downsampled	Time Series Forest Classifier	0.73426
Downsampled	Time Series Forest Regressor	0.73471

Table 4.9: Performance of the baseline time series models

As shown in the table, none of the models performed as good as the HKWW model and the forecasters could not even attain a score above 0.5. It is also interesting that the TSNV variant performed worse than the original baseline dataset unless downsized. This is thought to be because of the numerous additional features that added complexity to the problem, and the tree-based classifier simply functioned better with a smaller dataset.

Once again, the tree-based time series classifiers and regressors had small performance differences, this time even for calibration.

¹⁶ This is a type of exponential smoothing forecaster [87].

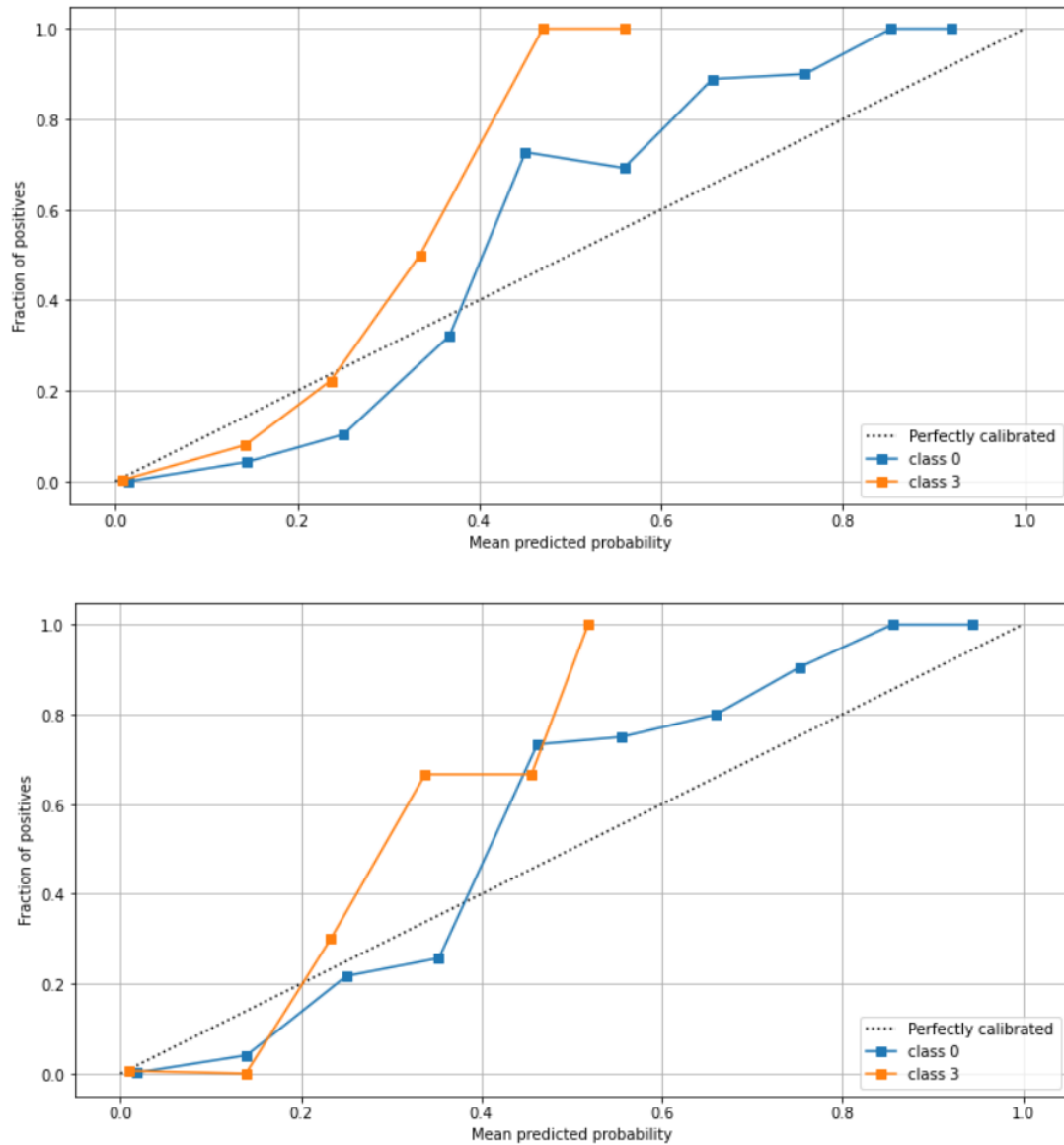


Fig. 4.11: Calibration plots of the (a) uncalibrated time series classifier (top) and (b) time series regressor (bottom)

The figures show that the calibration was already quite decent for the classifier and the regressor did not show any improvement. In fact, the Brier scores suggested that the classifier was better by a small margin of around 0.001. Purposeful calibration was not needed either, as shown in the upcoming figure.

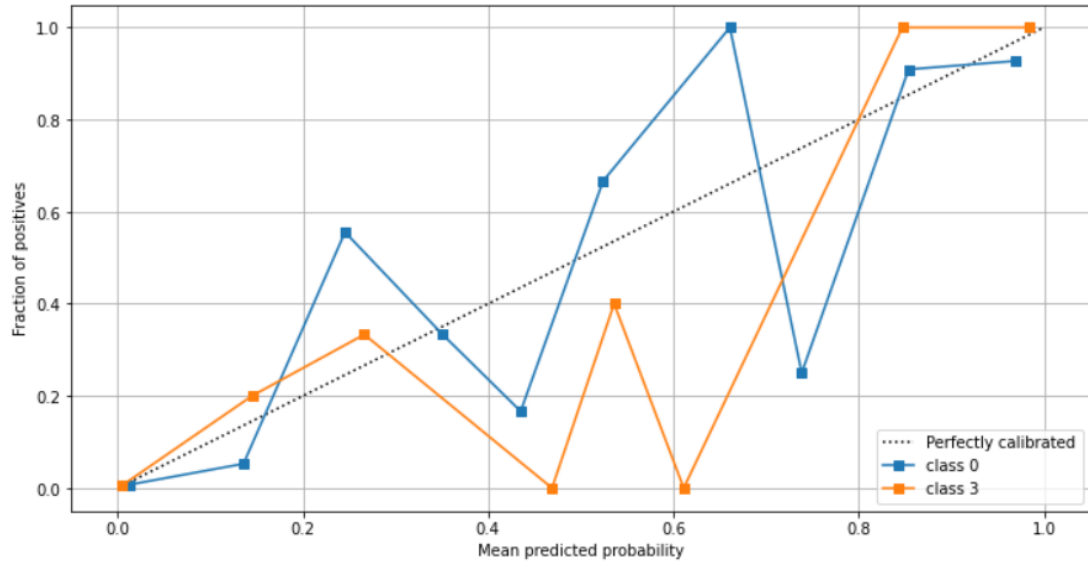


Fig. 4.12: Calibration plot of the calibrated time series classifier

The figure shows that the calibrated probabilities became more erratic after the process. While probability values bigger than 0.6 became common for class 3 (direct strike), the overall peculiarity suggested that it would not be a worthwhile choice.

4.4.2 Experimental Models

The tree-based time series classifier and regressor were reused for the experimental models. The results obtained are as follows:

Dataset	Model	Average F1 Score
N/A	HKWW	0.84500
Downsampled baseline	Time Series Forest Classifier	0.73426
Downsampled baseline	Time Series Forest Regressor	0.73471

Experimental	Time Series Forest Classifier	0.81236
Experimental	Time Series Forest Classifier (calibrated)	0.80568
Experimental	Time Series Forest Regressor	0.72107

Table 4.10: Performances of the experimental time series models

While the experimental classifier was superior to the baseline as expected, there were some unanticipated issues with the experimental time series models in general. Firstly, as shown in the table, the HKWW continued to outperform the experimental time series models just like linear models. If the values given by HKWW were truthful, then this would be worrying because they could be used as counterexamples to demonstrate that statistical-dynamical models might not necessarily be better than other preexisting models. Secondly, the regressor had seemingly overfitted even to validation data (with which a score of 0.86687 was attained) despite cross-validation. Thirdly, calibration did not affect the probabilities' distribution much and the regressor was worse calibrated than the classifier.

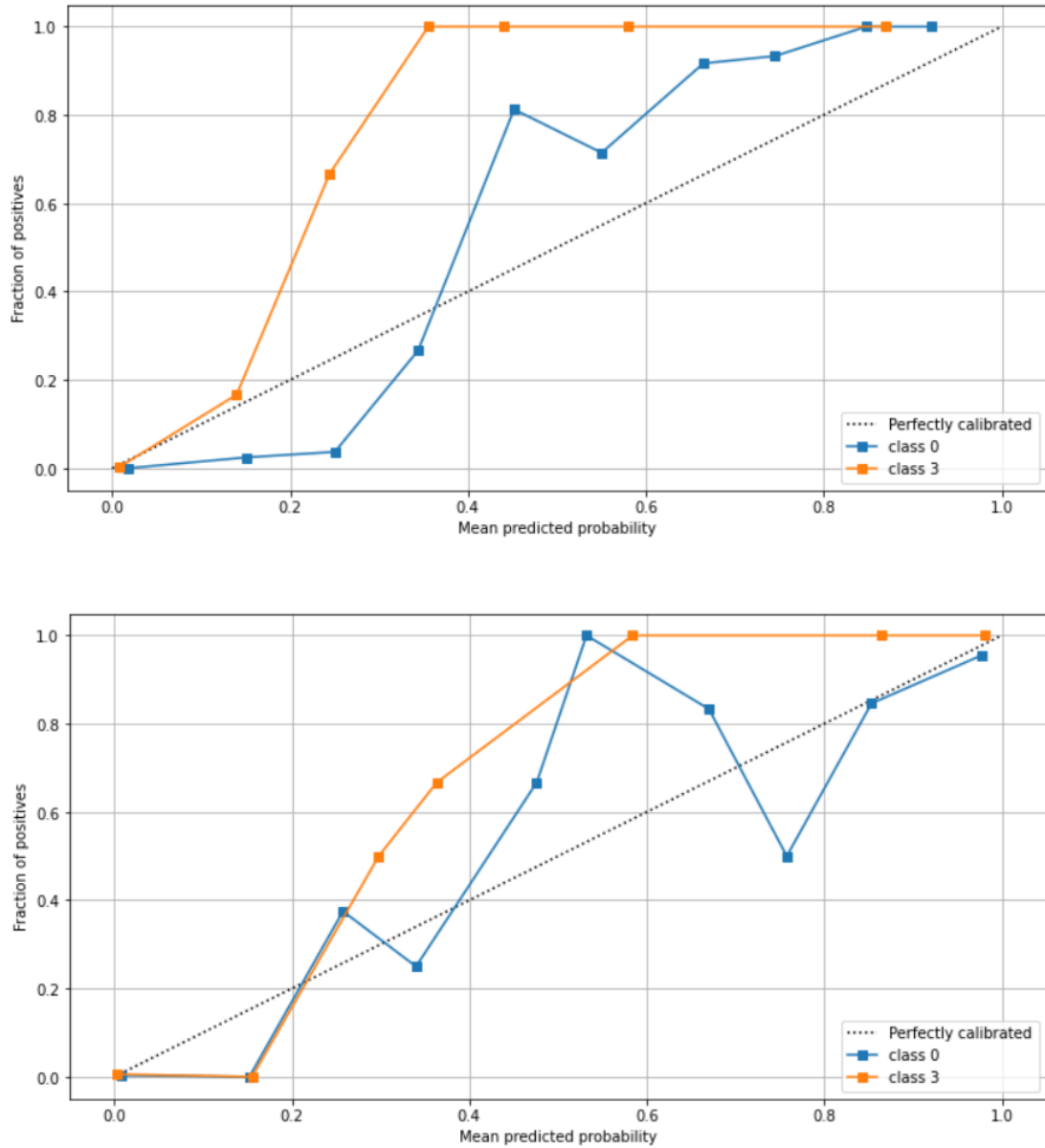


Fig. 4.13: Calibration plots of the experimental time series classifier (a) before (top) and (b) after (below) calibration

As shown in the figures, calibration did not significantly alter the probability distribution to the extent that it became too erratic to be practical. In fact, this was the one time when calibration was indeed useful. Table 4.10 showed that this calibration procedure negatively affected model scores, but the decrease was considered to be inconsequential.

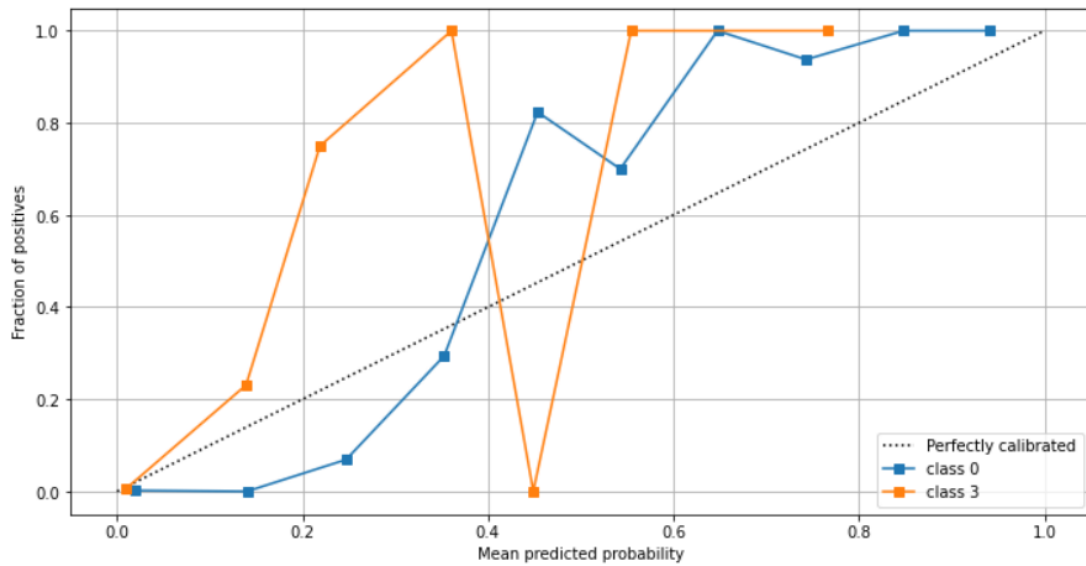


Fig. 4.14: Calibration plot of the experimental time series regressor

According to Figure 4.14, the calibration for class 0 (blue line) was acceptable but not so for class 3 (orange line). Indeed, the Brier scores for the regressor ranged from 0.01 to 0.03, while that for the classifier ranged from 0.007 to 0.02. That indicates the time series regressor was less appropriate than the corresponding classifier, which contradicted the results found with baseline datasets (see the previous section).

4.4.3 Discussion

The performance of the time series models was mediocre when compared to the best DT-based models, even if the best time series model was effectively yet another DT-based model; but with the results obtained with linear models (see Section 4.3) in mind, one must acknowledge that time series models were acceptable, nonetheless. Therefore, the calibrated time series classifier was admitted to the ensemble (see Section 4.7).

Forecasters had been a sizeable disappointment. Because the given endogenous time

series was short (5 observations only) compared to the prediction target (equivalent to 12 observations ahead) and the forecasters might not have used the exogenous variables well, extrapolating and forecasting the future was difficult. Moreover, the forecasters were unable to generalize over multiple samples and they only fitted to one sample at a time, making it meaningless to use forecasters for this project.

The failure of the experimental regressor to maintain a performance comparable to the corresponding classifier could be explained by irregularity in dataset splits, i.e. the test data had a different distribution than the training and validation data, but this remains mere speculation to date.

4.5 Results with Multilayer Perceptrons

Multilayer perceptrons (MLPs) as described in Section 3.5.4 exhibited behaviour that contrasted the other methodologies significantly. Both MLP classifiers and regressors were tested and the following will detail the results obtained and issues encountered.

4.5.1 Baseline Models

The development of baseline MLPs had initially been unsuccessful, as both software packages sklearn and PyTorch could not construct MLPs that gave non-trivial scores. MLPs built with PyTorch had zero accuracies whereas those built with sklearn had insufficient accuracy, e.g. MLPs with 10 hidden layers had an average F1 score below 0.3. It was believed that the wrong evaluation metrics were specified for PyTorch MLPs whereas the sklearn MLP model structure was inappropriate. A breakthrough was found in February and successful models were then built, as it was found that models with large numbers of neurons in the first layers were more suitable.

Dataset	Model	Structure	Average F1 Score
N/A	HKWW model	N/A (spline based)	0.84500
Original	(MLP) Classifier	3072, 1024, 1024, 1024, 256	0.75770
Original	(MLP) Regressor	3072, 1024, 1024, 1024, 256	0.61195
TSNV	Classifier	3072, 1024, 1024, 1024, 256	0.79509
TSNV	Regressor	3072, 1024, 1024, 1024, 256	0.53694
Downsampled	Classifier	3072, 1024, 1024, 1024, 256	0.69481
Downsampled	Regressor	3072, 1024, 1024, 1024, 256	0.58962

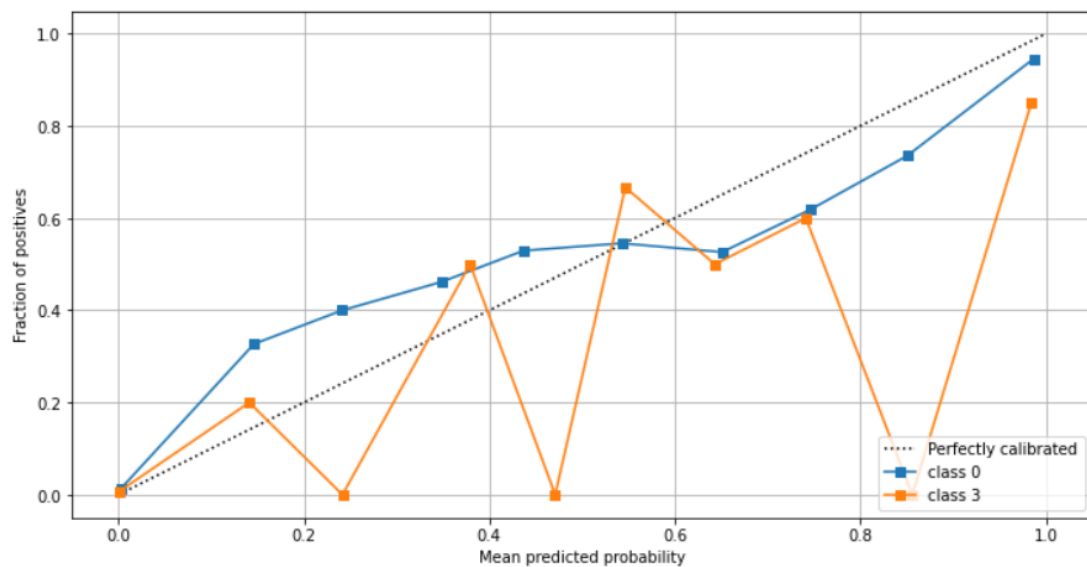
Table 4.11: Performances of the baseline MLPs

As shown in the table, MLPs generally did not perform well when compared to the HKWW model or the other baseline models in this project. This was because the MLPs were typically overfitted. For example, the downsampled baseline MLP classifier had an average score of about 0.9 on the training data but the score decreased to 0.838 on validation data, using which decision thresholds were found. These thresholds were then reused for final scoring with test data and the score was 0.695 at the end. However, increasing regularization and decreasing model structure complexity did not improve the scores by any remarkable degree. Also noteworthy is that as shown in the table,

MLP regressors tended to perform worse than a classifier fitted using the same settings. Due to time constraints, there were no in-depth investigations made to examine this matter, but overfitting was surmised to be the underlying reason.

More importantly, the model performance did not increase with each dataset revision. This was essentially the exact opposite of the expected behaviour as inferred from other modelling techniques. The table shows that the full-sized TSNV dataset was more suitable than the downsized one. This can be explained in terms of overfitting: the smaller dataset was easier for the models to overfit.

The MLP models showed acceptable probability calibration for predictands that had more positives.



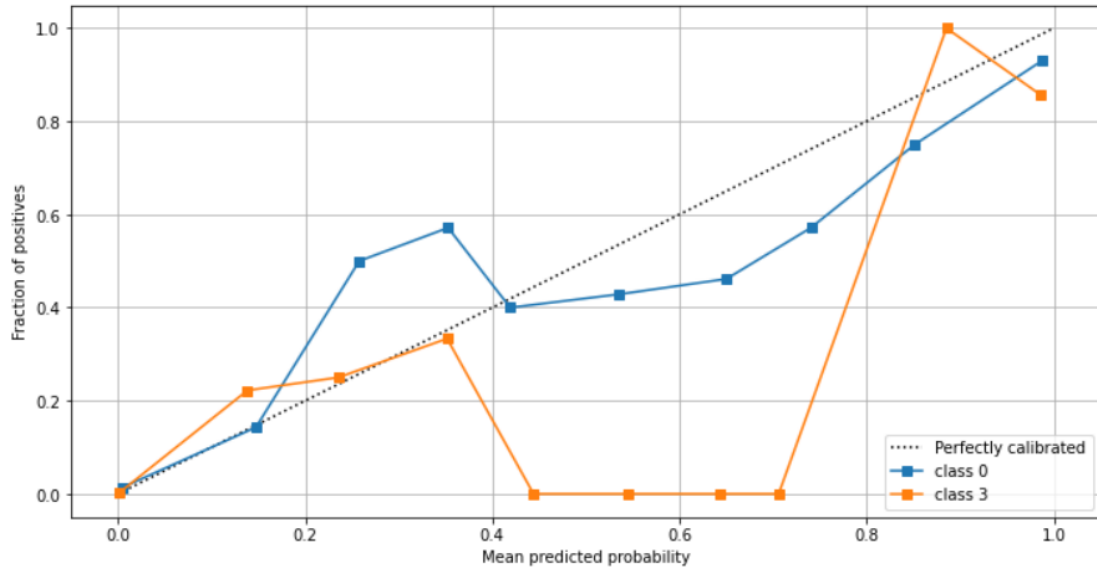


Fig. 4.15: The calibration plots of the uncalibrated baseline MLP classifiers trained on (a) all TSNV data (last page) and (b) downsampled TSNV data (above)

As shown in Figure 4.15, both models showed remarkably good calibration for class 0 (blue lines) which was the class with the most positive samples in the data. While that was much appreciated, the calibration for the rarer classes was not as good. The orange lines (class 3) in both figures were erratic, but the model that had access to more data (Fig. 4.15a) could predict better because it had more points near the diagonal than that in Fig. 4.15b. That meant, with sufficient data, intentional probability calibration would be unnecessary.

4.5.2 Experimental Models

The results obtained with experimental MLPs are as follows:

Dataset	Model	Model Structure	Average F1 Score
N/A	HKWW model	N/A (spline-based)	0.84500
TSNV	(MLP) Classifier	3072, 1024, 1024, 1024, 256	0.79509
Downsampled baseline	Classifier	3072, 1024, 1024, 1024, 256	0.69481
Downsampled baseline	(MLP) Regressor	3072, 1024, 1024, 1024, 256	0.58962
Experimental	Classifier	3072, 2048, 1024, 256	0.74856
Experimental	Regressor	3072, 2048, 1024, 256	0.14684

Table 4.12: Performances of the experimental MLPs

Several observations can be made from the table. Firstly, a great discrepancy between the performances of MLP classifiers and regressors remained, even though the dataset was changed. Secondly, the experimental MLP classifier worked better than its counterpart built with the downsampled baseline dataset, which was welcomed, but the performance could not match that of the full TSNV dataset version. This once again suggested that MLPs preferred larger datasets instead of small ones.

The calibration of the MLP classifier was somewhat unfavourable.

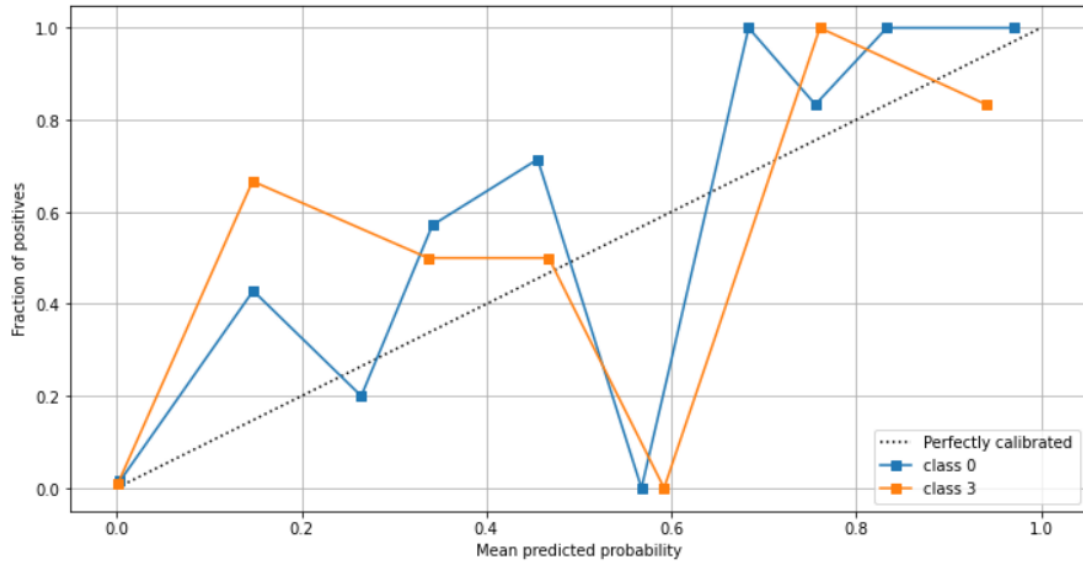


Fig. 4.16: Calibration plot of the experimental MLP classifier

Both lines in the diagram are improperly aligned to the diagonal, indicating that the probability calibration was unsatisfactory, despite the Brier scores remaining low (0.008-0.029). This was believed to be caused by the lack of data. The more positive samples there are, the more likely that a more comprehensive plot can be made, as there will be more predictions to infer from.

4.5.3 Discussion

While MLPs had shown promising performances on the full-size baseline dataset, the results obtained with smaller datasets such as the experimental one were disappointing. Linear models such as GAMs were initially thought to be less appropriate, but they managed to surpass MLPs, who had then come to be the worst models.

As it has been reiterated throughout the previous sections, the reduced dataset size was the most probable culprit. While DTs were able to generalize over the small experimental dataset well and attain respectable scores on held-out data, MLPs had

simply failed to do so and were overfitted.

Nonetheless, the experimental MLP classifier was admitted to the ensemble to promote heterogeneity in ensemble member types.

4.6 Results with Gaussian Processes

Gaussian Processes (GPs) as described in Section 3.5.5 were outright inappropriate. GPs demanded much memory and for the full baseline dataset, only an approximate GP built with GPyTorch was possible. Approximate GPs base their inference on a subset of the training data and thus accuracy would be lost to some extent. Regardless of model design, no GP was able to attain a non-zero score. This was not because of an inappropriate loss function and evaluation metric choice (compare MLPs), as the behaviour repeated when the smaller experimental dataset was used to build exact GPs. It was found that the model easily overfitted on training data and claimed to have a 100% accuracy, but all probabilistic predictions made on unseen data were 0.5, which was a phenomenon model design could not affect. It was suspected that GPs were unable to effectively distinguish between noise and relevant data, and simply returned the mean between positives and negatives as the prediction. Therefore, all GPs were considered unsuitable for this project.

4.7 Final Ensemble Model

The final ensemble model consists of the best individual models from Sections 4.2 to 4.5 (GPs excluded due to impracticality) and an ensemble voting mechanism, which

was a further estimator¹⁷ according to Section 3.5.6. The ensemble members selected are as follows:

No.	Model type	Model	Average F1 Score	Brier Score Range ¹⁸
1	Tree-based	XGBoost Classifier	0.92765	0.001-0.008
2	Tree-based	Extra Trees Classifier	0.89849	0.005-0.010
3	MLP	MLP Classifier	0.74856	0.008-0.028
4	Time series	calibrated Time Series Forest Classifier	0.80568	0.007-0.020
5	Linear	GAM	0.81090	0.012-0.038
6	Tree-based	XGBoost Regressor	0.87246	0.007-0.017
7	Tree-based	Extra Trees Regressor	0.89823	0.007-0.011

Table 4.13: List of ensemble members

The two regressors were added as an afterthought to observe whether the addition of more varied models could further improve the ensemble. From these models, several configurations were tested, each consisting of some input models and a voting mechanism (the estimator). The voting mechanism was fitted using the validation data

¹⁷ A regressor or a classifier.

¹⁸ Correct to 3 decimal places.

which the input models have never encountered and then evaluated using the remaining held-out test data. The results found are as follows:

Voting Method	Input Models	Average F1 Score	Brier Score Range ¹⁹
Linear regression	No. 1-4	0.96364	0.0011-0.0076
Logistic regression	No. 1-4	0.93069	0.0023-0.0084
Ridge regression ²⁰	No. 1-4	0.95534	0.0013-0.0076
Linear regression	No. 1-5	0.96874	0.0011-0.0077
Ridge regression	No. 1-5	0.95401	0.0014-0.0077
Linear regression	No. 1, 2, 4, 5	0.94782	0.0011-0.0072
Ridge regression	No. 1, 2, 4, 5	0.95234	0.0015-0.0077
Linear regression	No. 1-7	0.96874	0.0012-0.0077
Ridge regression	No. 1-7	0.95978	0.0016-0.0077

Table 4.14: Performances of different ensemble voting mechanisms

There were four batches of tests conducted, as shown in the differences in input models in Table 4.14. The first batch excluded the GAM, the second excluded tree-based regressors, and the third excluded the worst-performing input model (MLP), but it was found that in terms of deterministic forecasting performance measured in F1 scores, the best choice was to include all the selected models. This echoes the concept mentioned in Section 3.3 that ensembles can make the final output more robust and comprehensive.

Amongst the voting mechanism (stacked estimator) options, logistic regression was

¹⁹ Correct to 4 decimal places

²⁰ Linear regressor with regularization applied [88].

quickly found to be the worst and was abandoned. Between linear and ridge regression, there was a trade-off to be made: ridge regression allowed fine-tuning of calibrations via regularization whereas linear regression simply maximized the prediction scores. There was also doubt regarding the F1 scores found because the validation and test datasets had few positive samples and could easily have been overfitted. The decision was ultimately made to select the second most performant ensemble model, i.e. the ridge regressor taking all seven models as input, and then apply additional regularization to control overfitting and decision thresholds.

Regularization	Testing F1 Score	Brier Scores ²¹
Default	0.95978	0.0061, 0.0077, 0.0016, 0.0051
Applied	0.93773	0.0068, 0.0079, 0.0025, 0.0054

Table 4.15: Comparison between the default and regularized ridge regression ensemble mechanisms

The table shows that with regularization, the testing F1 score and Brier scores of the model worsened. This was however acceptable because the F1 score of the ensemble remained higher than any individual ensemble member and the calibration plots suggested that the regularized model was still favourable.

²¹ Correct to 4 decimal places and listed in the order of the predictands (minimal impact to direct strike)

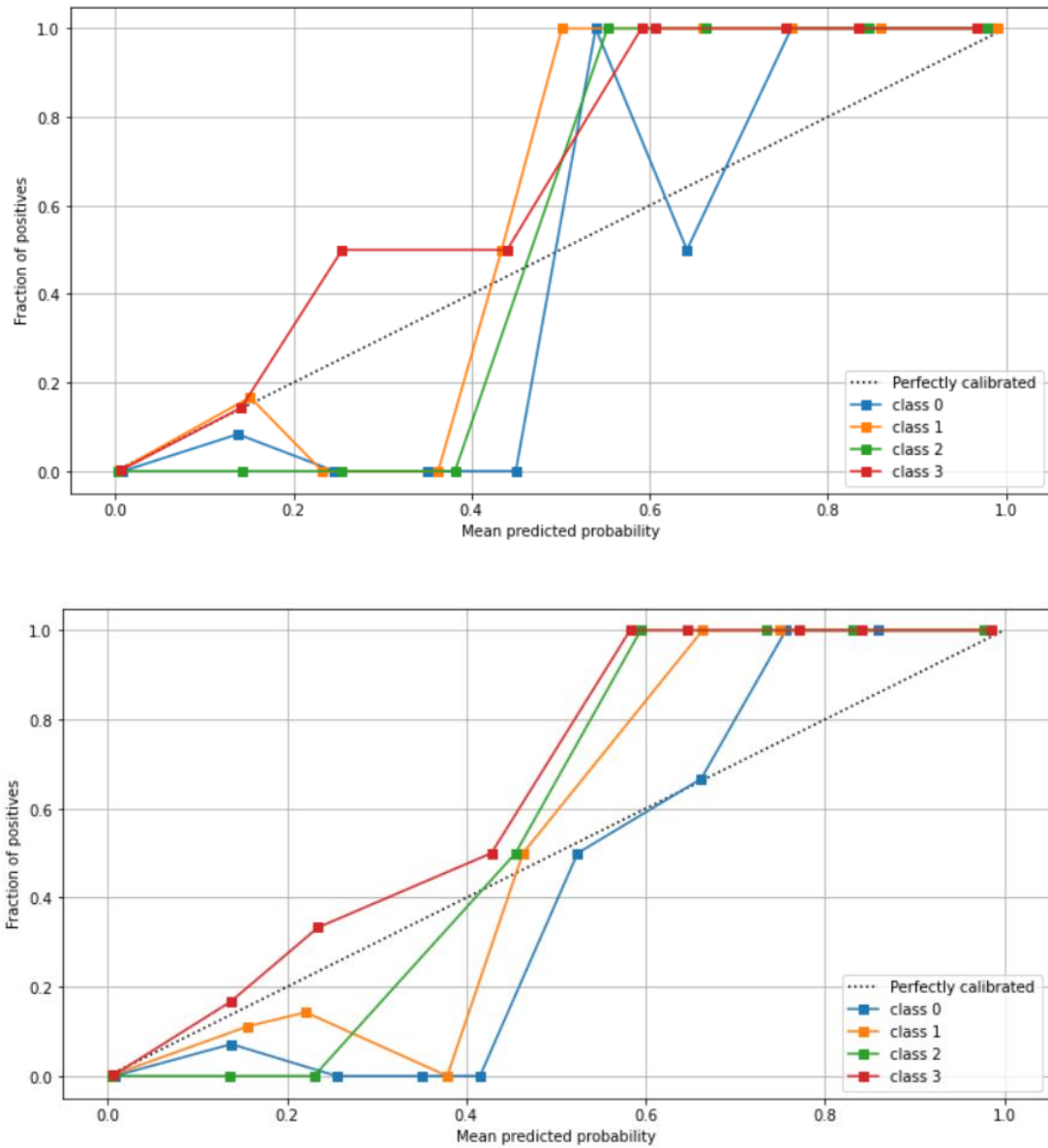


Fig. 4.17: Calibration plots of the (a) default (top) and (b) regularized (bottom) ensemble models

As shown in the figures, the lines in the regularized version were closer to the perfect diagonal than without regularization. This way, the predicted probabilities were more credible and over-/underestimations less problematic. It was believed that the decrease in F1 scores was a necessary price to pay.

The following chart shows in detail the behaviour of the ensemble when performing deterministic predictions, i.e. converting the probabilities to binary “yes/no” forecasts.

Predictand	Metric	Value ²²
Minimal Impact (75 positives in the testing data)	Precision	0.96104
	Recall	0.98667
	F1 Score	0.97638
	Decision threshold	0.40494
	Brier score	0.00678
Limited Impact (57 positive samples)	Precision	0.93103
	Recall	0.94737
	F1 Score	0.93913
	Decision threshold	0.37428
	Brier score	0.00789
Substantial Impact (17 positive samples)	Precision	0.94444
	Recall	1.00000
	F1 Score	0.97143
	Decision threshold	0.39512
	Brier score	0.00246
Direct Strike (16 positives)	Precision	0.92857
	Recall	0.81250
	F1 Score	0.86667
	Decision threshold	0.27891

²² Correct to 5 decimal places.

	Brier score	0.00536
Macro-Average (80 total positives among 696 testing samples)	Precision	0.94127
	Recall	0.93663
	F1 Score	0.93773
	Brier score	0.00561

Table 4.16: Detailed performance measurements of the ensemble

From the table, it can be observed that the decision thresholds deviated further away from the ideal value of 0.5 the rarer the predictand was. That was a prevalent issue, however, as most models (baseline and experimental alike) had difficulty making unbiased predictions about the less common targets, due to the rarity of positives.

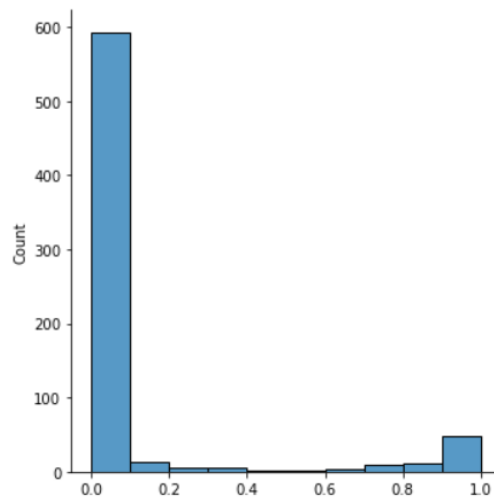


Fig. 4.18: Distribution of predicted probabilities of the ensemble

Figure 4.18 shows that the ensemble tended to predict values close to 0 and 1. That implied that the model would over- or underestimate often, even though an ideal model would only correctly predict 0 or 1.

4.8 Summary and Overall Comments

This project had two main objectives (as stated in Section 1.2): To develop a statistical-dynamical probabilistic forecast to assess TC impact level probabilities in Hong Kong and to ascertain whether a statistical-dynamical approach is better than not considering dynamical factors. The results showed that both objectives have been generally achieved.

Firstly, the forecasting *product* mentioned in Section 1.2 was realized as a prototype at best, since only the forecasting models at its core were complete. To complete a forecasting product, auxiliary functions would have to be developed to accept user input of TC position and intensity features, automate the retrieval and preprocessing of dynamical data, and present the forecasts to the user. Nonetheless, it was still possible to manually carry out these procedures and conduct operational forecasts using the models built throughout this project.

Secondly, the superiority of the statistical-dynamical approach was justified successfully. With dataset size and all other features fixed, a dataset containing dynamical predictors can be used to develop better forecasts (compare the downsampled baseline and experimental datasets), even though the exact performance scores depend on the modelling method. For instance, the average F1 scores of the best tree-based model and the MLP model differ by 0.1791. The margin between the two datasets might not be significant either, e.g. the best tree-based baseline was only worse than the best experimental model by 0.02.

The top 50 most relevant features found by baseline XGBoost models for each

predictand were listed in Appendix 1.1. From the lists, it was found that the DTs tended to regard endogenous variables' statuses, dates (especially the month), and TC position information as highly important. For direct strikes, the recent azimuths and TC intensities were given more emphasis than for other predictands. These choices were understandable, e.g. the endogenous variables have strong persistence and seasonal factors indeed influence TC strike risks. The 50 most significant polynomial features for each predictand according to mutual information were in Appendix 1.2. A different trend was found, where distance, month, and intensity were common factors among the selected polynomials while endogenous variables ranked much lower. This was presumably because the endogenous variables mostly had a value of 0, making their polynomials relatively unusable.

The usage of dynamical features by the experimental models was apparent. Appendix 1.3 lists out feature importance ranks according to experimental XGBoost models and different predictands preferred different dynamical features, e.g. for minimal impact surface temperature was highly important but to predict substantial impact the knowledge of 500hPa u-winds was more important. The general trend to use endogenous variables, months and distances was also present, but dynamical factors such as surface temperature, summer monsoon index and winds were also commonplace. The most relevant polynomials, as shown in Appendix 1.4, continued similar trends where distance and month were dominant factors, but the factors they paired up with were often dynamical predictors such as temperature, potential temperature, westerly index, and summer monsoon index. The relationship between the favoured dynamical predictors with the predictands was rather obscure and hard to analyse, e.g. potential temperature being one of the key factors influencing direct strike.

These findings may require meteorology professionals to further scrutinize. In any case, these proved that the models utilized the new dynamical data well.

It was also found that treating the forecasting problem as a classification or regression problem had little difference, with the former being slightly more effective in general but the latter more appropriate for the final ensemble forecaster, which succeeded in balancing the individual forecasting models to obtain even better results.

However, the datasets used were not large and were inherently unsuitable for many modelling methods. This may have had the effect of exaggerating the scores due to overfitting, even if cross-validation was occasionally carried out. That is, the models may have overfitted the entireties of the datasets, which underrepresented the real behaviour of TCs.

Chapter 5: Conclusion

This project explored the means to develop statistical-dynamical forecasting models that evaluate the likelihood of TCs affecting Hong Kong. The models worked together as an ensemble and produce probabilistic forecasts in categories corresponding to different levels of impact. DTs, GAMs, time series modelling, MLPs and ridge regression were the main modelling methods employed. The usefulness of dynamical data in TC strike probability forecasts had also been studied.

The results described in the previous chapter show that statistical-dynamical TC impact probability forecasts are indeed desirable. All the modelling techniques applied, namely

DTs, GAMs, MLPs and time series modelling, gave experimental models that outperformed their corresponding traditional baselines and, in some cases, even the current best TC impact probability forecaster by HKWW. A further ensemble consisting of the best seven experimental models showed capability in predicting with enhanced accuracy and comprehensiveness, albeit running a risk of overfitting. At the moment, the ensemble forecaster can produce operational forecasts, but the usability is low, as manual input and preprocessing of data are still needed. Nonetheless, this project is considered a success on the whole.

It is believed that the results obtained in this project demonstrate the potential of probabilistic statistical-dynamical TC forecasts designed for Hong Kong while serving as an additional means to evaluate TC threats for the general public. Future weather forecasts with similar concepts may also be feasible.

However, this project has used vague definitions for TC impact levels, which were only loosely tied to TC warning signals and thus prevent the usage of the forecasting product elsewhere. In addition, the selection of dynamical predictors by the models has not been well understood, necessitating further studies. 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. Furthermore, the forecasting product developed in this project is unwieldy and requires much manual work to produce operational forecasts. The development of supporting and automation functions will be beneficial as they make the models more usable.

Appendices

Appendix 1: Feature Importance Reports

The exact importance values are available in the source code.

1.1 Top 50 features for each predictand, baseline tree-based regressor

Minimal Impact:

Rank	Feature	Rank	Feature
1	'MI_STATUS00'	2	'MI_STATUS06'
3	'MI_STATUS12'	4	'MM18'
5	'DIST06'	6	'DIST00'
7	'MM24'	8	'AZM06'
9	'LI_STATUS00'	10	'MM00'
11	'MM06'	12	'MM12'
13	'AZM18'	14	'AZM00'
15	'LI_STATUS06'	16	'SI_STATUS06'
17	'LI_STATUS18'	18	'DS_STATUS00'
19	'DIST18'	20	'VMAX18'
21	'DD18'	22	'DD06'
23	'DIR06'	24	'DD00'
25	'VMAX06'	26	'LI_STATUS24'
27	'AZM12'	28	'DS_STATUS12'
29	'DIR12'	30	'DD24'
31	'DIST24'	32	'VMAX24'
33	'VMAX00'	34	'DD12'

35	'LI_STATUS12'	36	'DS_STATUS06'
37	'DIR00'	38	'SPEED00'
39	'VMAX12'	40	'AZM24'
41	'SPEED18'	42	'SPEED06'
43	'DIST12'	44	'SI_STATUS00'
45	'SI_STATUS12'	46	'DVMAX12'
47	'DIR24'	48	'SPEED12'
49	'DVMAX00'	50	'SPEED24'

Table 6.1: Top 50 features for minimal impact prediction with baseline data

Limited impact:

Rank	Feature	Rank	Feature
1	'SI_STATUS24'	2	'LI_STATUS00'
3	'MI_STATUS00'	4	'SI_STATUS00'
5	'MI_STATUS06'	6	'LI_STATUS06'
7	'DIST00'	8	'MI_STATUS12'
9	'MM18'	10	'DIST06'
11	'SI_STATUS06'	12	'MM06'
13	'MM00'	14	'AZM00'
15	'AZM18'	16	'AZM06'
17	'LI_STATUS18'	18	'MM12'
19	'DIST18'	20	'SI_STATUS12'
21	'MM24'	22	'DD24'
23	'VMAX06'	24	'DD18'

25	'DIR00'	26	'DIR06'
27	'SPEED18'	28	'DD00'
29	'VMAX18'	30	'DD06'
31	'VMAX00'	32	'DIST24'
33	'DIR12'	34	'SPEED00'
35	'VMAX24'	36	'MI_STATUS18'
37	'VMAX12'	38	'DIST12'
39	'DD12'	40	'AZM24'
41	'DS_STATUS18'	42	'SPEED06'
43	'AZM12'	44	'LI_STATUS12'
45	'DVMAX18'	46	'DVMAX00'
47	'DS_STATUS00'	48	'SPEED24'
49	'DIR24'	50	'SPEED12'

Table 6.2: Top 50 features for limited impact prediction with baseline data

Substantial impact:

Rank	Feature	Rank	Feature
1	'SI_STATUS00'	2	'MI_STATUS00'
3	'DS_STATUS00'	4	'SI_STATUS06'
5	'DIST06'	6	'DIST00'
7	'MM18'	8	'MI_STATUS06'
9	'AZM18'	10	'AZM06'
11	'MM12'	12	'AZM00'
13	'DS_STATUS06'	14	'MM00'

15	'DIR12'	16	'DIR06'
17	'MM06'	18	'DIST18'
19	'DD00'	20	'MI_STATUS12'
21	'DD18'	22	'DIST12'
23	'SPEED00'	24	'VMAX00'
25	'DD12'	26	'DIR00'
27	'VMAX06'	28	'VMAX18'
29	'DD06'	30	'DIST24'
31	'DD24'	32	'SPEED12'
33	'HH06'	34	'DS_STATUS12'
35	'MM24'	36	'DIR24'
37	'MI_STATUS18'	38	'SPEED18'
39	'DIR18'	40	'AZM24'
41	'LI_STATUS06'	42	'VMAX24'
43	'VMAX12'	44	'SPEED24'
45	'AZM12'	46	'SPEED06'
47	'LI_STATUS00'	48	'DVMAX00'
49	'SI_STATUS12'	50	'DVMAX18'

Table 6.3: Top 50 features for substantial impact prediction with baseline data

Direct strike:

Rank	Feature	Rank	Feature
1	'LI_STATUS24'	2	'SI_STATUS18'
3	'SI_STATUS24'	4	'DS_STATUS18'

5	'DS_STATUS00'	6	'MI_STATUS00'
7	'MI_STATUS06'	8	'AZM00'
9	'DIST18'	10	'DD18'
11	'AZM06'	12	'DIST00'
13	'MM18'	14	'VMAX06'
15	'DIR12'	16	'DD06'
17	'AZM18'	18	'DIR06'
19	'DD12'	20	'DIR00'
21	'SPEED18'	22	'MM06'
23	'VMAX18'	24	'DIST12'
25	'DIR18'	26	'VMAX00'
27	'DD24'	28	'MM00'
29	'DIST06'	30	'DD00'
31	'SPEED06'	32	'DIR24'
33	'MI_STATUS18'	34	'SPEED12'
35	'AZM12'	36	'DIST24'
37	'SPEED00'	38	'SPEED24'
39	'MM24'	40	'VMAX24'
41	'AZM24'	42	'MM12'
43	'VMAX12'	44	'DS_STATUS06'
45	'HH12'	46	'DVMAX18'
47	'DVMAX00'	48	'DVMAX12'
49	'SI_STATUS00'	50	'HH00'

Table 6.4: Top 50 features for direct strike prediction with baseline data

1.2 Top 50 polynomial features for each predictand, mutual information and baseline data

Minimal impact:

Rank	Feature	Rank	Feature
1	MM00 DIST00^2	2	DIST00
3	DIST00^3	4	DIST00^2
5	DIST00^2 VMAX00	6	MM00 DIST00
7	DIST00^2 SPEED00	8	DIST00^2 AZM00
9	MM00^2 DIST00	10	MI_STATUS00^2 SPEED00
11	MM00 DD00^2	12	MI_STATUS00 VMAX00^2
13	MM00 DD00 MI_STATUS00	14	MI_STATUS00 DIST00
15	MI_STATUS00 VMAX00	16	DIST00 VMAX00
17	MI_STATUS00 SPEED00 VMAX00	18	MI_STATUS00^2 DIST00
19	DIST00^2 DVMAX00	20	MM00 MI_STATUS00 AZM00
21	MI_STATUS00 SPEED00	22	MM00 MI_STATUS00 VMAX00
23	MI_STATUS00 DIR00	24	HH00 DIST00^2
25	MI_STATUS00 DIR00 VMAX00	26	DD00^2 MI_STATUS00
27	MM00 DIST00 VMAX00	28	MI_STATUS00 DIST00 AZM00
29	MM00^2 MI_STATUS00	30	MI_STATUS00 SPEED00^2
31	MI_STATUS00^2 AZM00	32	MM00 MI_STATUS00 DIST00
33	DD00 MI_STATUS00	34	DD00 DIST00^2

	DIST00		
35	DD00 MI_STATUS00^2	36	MI_STATUS00^2
37	MI_STATUS00 AZM00 SPEED00	38	MI_STATUS00 AZM00^2
39	AZM00	40	DD00 MI_STATUS00
41	MM00 MI_STATUS00^2	42	AZM00^3
43	AZM00^2	44	DD00 MI_STATUS00 AZM00
45	MI_STATUS00 AZM00	46	MI_STATUS00^3
47	DD00 MI_STATUS00 VMAX00	48	MI_STATUS00 AZM00 DIR00
49	MM00 MI_STATUS00	50	MM00^2 DD00

Table 6.5: Top 50 polynomial features for minimal impact prediction using baseline data

Limited impact:

Rank	Feature	Rank	Feature
1	DIST00	2	DIST00^3
3	MM00 DIST00^2	4	DIST00^2
5	MM00 DIST00	6	DIST00^2 VMAX00
7	DIST00^2 AZM00	8	DIST00^2 SPEED00
9	MM00^2 DIST00	10	DD00 DIST00^2
11	HH00 DIST00^2	12	AZM00^3
13	AZM00^2	14	AZM00
15	MM00 DD00^2	16	DIST00 VMAX00
17	DIST00^2 DIR00	18	DIST00 DIR00^2

19	DIST00 SPEED00	20	DIST00 AZM00
21	MM00 DIST00 VMAX00	22	MM00 DIST00 SPEED00
23	DIST00^2 DVMAX00	24	MM00^2 VMAX00
25	MM00^2 DD00	26	MM00 DIST00 AZM00
27	MM00 VMAX00^2	28	LI_STATUS00^2 SPEED00
29	HH00 DIST00	30	DIST00 DVMAX00
31	MM00^3	32	MM00 AZM00^2
33	MM00 SPEED00^2	34	MM00 LI_STATUS00 VMAX00
35	MM00^2 SPEED00	36	MM00 LI_STATUS00 DIST00
37	LI_STATUS00 DIST00 DIR00	38	DIST00 AZM00 DIR00
39	MM00 HH00 DIST00	40	MM00
41	DD00 LI_STATUS00 VMAX00	42	LI_STATUS00^3
43	DD00 LI_STATUS00^2	44	AZM00^2 DIR00
45	MM00^2 LI_STATUS00	46	MM00 DD00
47	LI_STATUS00 SPEED00 VMAX00	48	DD00 DIST00
49	LI_STATUS00 DIR00 VMAX00	50	LI_STATUS00^2

Table 6.6: Top 50 polynomial features for limited impact prediction using baseline data

Substantial impact:

Rank	Feature	Rank	Feature
1	DIST00	2	DIST00^2
3	DIST00^3	4	MM00 DIST00^2
5	DIST00^2 SPEED00	6	DIST00^2 VMAX00
7	MM00 DIST00	8	DIST00^2 AZM00
9	MM00^2 DIST00	10	HH00 DIST00^2
11	DD00 DIST00^2	12	MM00 DIST00 AZM00
13	DIST00^2 DVMAX00	14	MM00^2 DIR00
15	DIST00 AZM00	16	MM00 DIST00 VMAX00
17	DIST00 VMAX00	18	MM00 DD00^2
19	HH00 DIST00	20	DIST00 DVMAX00
21	DIST00^2 DIR00	22	AZM00 DIR00
23	DIST00 DIR00^2	24	MM00 DIST00 SPEED00
25	MM00 DIR00	26	MM00 HH00 DIST00
27	MM00^2 DD00	28	MM00^2
29	MM00	30	MM00^3
31	AZM00^2 DIR00	32	MM00 DD00
33	MM00 SPEED00^2	34	MM00 DIST00 DVMAX00
35	AZM00^3	36	DIST00 SPEED00
37	AZM00^2	38	AZM00
39	SI_STATUS00 DIR00^2	40	SI_STATUS00^2 VMAX00
41	DIST00 AZM00 VMAX00	42	DIST00 VMAX00 DVMAX00
43	HH00 DIST00 AZM00	44	SI_STATUS00 DIST00

45	SI_STATUS00 SPEED00 VMAX00	46	MM00^2 SPEED00
47	MM00^2 VMAX00	48	SI_STATUS00 DIST00^2
49	MM00 DD00 DIST00	50	SI_STATUS00 DIR00

Table 6.7: Top 50 polynomial features for substantial impact prediction using baseline data

Direct strike:

Rank	Feature	Rank	Feature
1	MM00 DIST00	2	MM00 DIST00^2
3	DIST00^2	4	DIST00
5	DIST00^3	6	DIST00^2 VMAX00
7	MM00^2 DIST00	8	DIST00 VMAX00
9	MM00 DD00^2	10	DIST00^2 AZM00
11	MM00 DIST00 VMAX00	12	DIST00^2 SPEED00
13	MM00^2 DIR00	14	MM00 VMAX00^2
15	MM00^2 VMAX00	16	HH00 DIST00^2
17	MM00^3	18	MM00^2
19	MM00 DIST00 AZM00	20	DD00 DIST00^2
21	MM00^2 DD00	22	DIST00 AZM00
23	MM00 SPEED00^2	24	MM00 DIR00
25	HH00 DIST00 VMAX00	26	MM00 HH00 DIST00
27	MM00 DIST00 SPEED00	28	MM00^2 SPEED00
29	MM00 MI_STATUS00 AZM00	30	DIST00^2 DIR00

31	DIST00 DVMAX00	32	MM00^2 HH00
33	MM00	34	MM00 DIST00 DIR00
35	MM00 DD00	36	HH00 DIST00
37	DD00^2 DS_STATUS00	38	DIST00 SPEED00
39	MM00 DIST00 DVMAX00	40	DIST00 SPEED00 VMAX00
41	MM00 DS_STATUS00^2	42	DIST00 AZM00 VMAX00
43	MM00 SPEED00	44	DIST00^2 DVMAX00
45	DS_STATUS00^2 VMAX00	46	DS_STATUS00^3
47	DIST00 VMAX00 DVMAX00	48	MM00 MI_STATUS00 SPEED00
49	MM00 DS_STATUS00 VMAX00	50	DS_STATUS00 AZM00 SPEED00

Table 6.8: Top 50 polynomial features for direct strike prediction using baseline data

1.3 Top 50 features for each predictand, experimental tree-based classifier

Minimal impact:

Rank	Feature	Rank	Feature
1	'MI_STATUS00'	2	'DIST00'
3	'STEMP00'	4	'AZM00'
5	'MM06'	6	'U50018'
7	'MM00'	8	'AZM12'
9	'MM12'	10	'AZM06'
11	'U_HK06'	12	'U_HK00'
13	'LO_HUMID12'	14	'LI_STATUS06'

15	'U_HK18'	16	'DIST06'
17	'MM18'	18	'STEMP06'
19	'AZM24'	20	'WESTERLY06'
21	'POTT00'	22	'MM24'
23	'DD06'	24	'WESTERLY24'
25	'AZM18'	26	'DD24'
27	'U20012'	28	'DIR00'
29	'UTEMP00'	30	'U20006'
31	'U_HK12'	32	'DIST24'
33	'DIST18'	34	'DIST12'
35	'DIR06'	36	'POTT06'
37	'DD00'	38	'LI_STATUS00'
39	'MI_STATUS12'	40	'HI_HUMID06'
41	'LI_STATUS12'	42	'U50006'
43	'EASM12'	44	'MI_STATUS06'
45	'POTT12'	46	'U50000'
47	'VORT12'	48	'VMAX00'
49	'SPEED00'	50	'UTEMP06'

Table 6.9: Top 50 features for minimal impact prediction using experimental data

Limited impact:

Rank	Feature	Rank	Feature
1	'DIST00'	2	'LI_STATUS06'
3	'LI_STATUS00'	4	'MM12'

5	'SH_EXT00'	6	'U_HK18'
7	'MI_STATUS00'	8	'U_HK00'
9	'STEMP00'	10	'MM06'
11	'AZM00'	12	'DIST06'
13	'MM00'	14	'POTT18'
15	'U50018'	16	'U_HK06'
17	'AZM12'	18	'SI_STATUS00'
19	'AZM06'	20	'U_HK12'
21	'MM18'	22	'AZM24'
23	'SH_INT18'	24	'AZM18'
25	'DIST24'	26	'WESTERLY24'
27	'WESTERLY06'	28	'DIR00'
29	'DIR06'	30	'DIST12'
31	'V_HK06'	32	'POTT06'
33	'VMAX18'	34	'MI_STATUS12'
35	'DIST18'	36	'MI_STATUS06'
37	'SPEED00'	38	'V_HK00'
39	'MM24'	40	'VMAX06'
41	'VMAX00'	42	'U50006'
43	'DVMAX06'	44	'U20012'
45	'V50018'	46	'POTT00'
47	'DD06'	48	'EASM06'
49	'WESTERLY12'	50	'LO_HUMID06'

Table 6.10: Top 50 features for limited impact prediction using experimental data

Substantial impact:

Rank	Feature	Rank	Feature
1	'SI_STATUS00'	2	'MI_STATUS00'
3	'DIST00'	4	'U50006'
5	'DS_STATUS00'	6	'EASM12'
7	'DIST06'	8	'MM00'
9	'U20006'	10	'DIR00'
11	'AZM18'	12	'AZM12'
13	'DIR06'	14	'MM06'
15	'AZM24'	16	'U_HK00'
17	'WESTERLY18'	18	'STEMP00'
19	'UTEMP18'	20	'V_HK06'
21	'U_HK18'	22	'EASM06'
23	'UTEMP24'	24	'DIST12'
25	'STEMP06'	26	'EASM00'
27	'AZM00'	28	'V_HK24'
29	'WESTERLY06'	30	'DIR18'
31	'DD06'	32	'V50012'
33	'POTT06'	34	'AZM06'
35	'V50000'	36	'EASM24'
37	'STEMP12'	38	'VMAX24'
39	'DD12'	40	'LI_STATUS00'
41	'DIST18'	42	'DIR24'
43	'DD00'	44	'DD18'

45	'WESTERLY24'	46	'U_HK06'
47	'U20012'	48	'V50006'
49	'U20000'	50	'POTT18'

Table 6.11: Top 50 features for substantial impact prediction using experimental data

Direct strike:

Rank	Feature	Rank	Feature
1	'DS_STATUS00'	2	'POTT24'
3	'V50018'	4	'EASM24'
5	'MM00'	6	'AZM12'
7	'MI_STATUS00'	8	'DIR06'
9	'DIR18'	10	'DIST12'
11	'DIST00'	12	'EASM18'
13	'EASM12'	14	'DIST06'
15	'DIR12'	16	'EASM06'
17	'POTT18'	18	'AZM06'
19	'DIR00'	20	'U50006'
21	'MM06'	22	'V50012'
23	'UTEMP18'	24	'AZM00'
25	'U50000'	26	'DIST24'
27	'LO_HUMID00'	28	'U20006'
29	'HI_HUMID00'	30	'WESTERLY24'
31	'WESTERLY06'	32	'UTEMP00'
33	'MI_STATUS24'	34	'MM24'

35	'UTEMP24'	36	'WESTERLY12'
37	'U_HK06'	38	'UTEMP12'
39	'STEMP00'	40	'V50000'
41	'AZM18'	42	'MM12'
43	'DIST18'	44	'EASM00'
45	'WESTERLY18'	46	'POTT06'
47	'V50006'	48	'POTT00'
49	'AZM24'	50	'STEMP12'

Table 6.12: Top 50 features for direct strike prediction using experimental data

1.4 Top 50 polynomial features for each predictand, mutual information and experimental data

Minimal impact:

Rank	Feature	Rank	Feature
1	DIST00 ² STEMP00	2	DIST00 UTEMP00
3	DIST00 UTEMP00 ²	4	MM00 DIST00 ²
5	DIST00 ² POTT00	6	DIST00 ² WESTERLY00
7	DIST00	8	DIST00 POTT00
9	DIST00 ²	10	DIST00 ³
11	DIST00 STEMP00 POTT00	12	DIST00 STEMP00
13	DIST00 UTEMP00 POTT00	14	DIST00 POTT00 ²
15	DIST00 ² UTEMP00	16	DIST00 STEMP00 ²
17	DIST00 STEMP00	18	DIST00 STEMP00

	UTEMP00		WESTERLY00
19	DIST00^2 VORT00	20	MM00 DIST00 POTT00
21	DIST00^2 HI_HUMID00	22	DIST00^2 LO_HUMID00
23	DIST00 WESTERLY00	24	DIST00 WESTERLY00 POTT00
25	DIST00^2 EASM00	26	DIST00^2 VMAX00
27	DIST00 UTEMP00 WESTERLY00	28	MM00 DIST00 STEMP00
29	DIST00^2 SPEED00	30	MM00^2 STEMP00
31	MM00 DIST00 UTEMP00	32	MM00 DIST00
33	MM00^2 UTEMP00	34	MM00 UTEMP00
35	MM00 UTEMP00^2	36	DIST00 LO_HUMID00 WESTERLY00
37	DIST00 HI_HUMID00 WESTERLY00	38	MM00 STEMP00
39	DIST00^2 U20000	40	DIST00^2 U50000
41	DIST00^2 AZM00	42	MM00 DIST00 WESTERLY00
43	DIST00^2 U_HK00	44	DIST00 STEMP00 EASM00
45	MM00^2 DIST00	46	DIST00^2 SH_INT00
47	DIST00 EASM00	48	DIST00^2 SH_EXT00
49	DIST00 EASM00 POTT00	50	DIST00 STEMP00 VORT00

Table 6.13: Top 50 polynomial features for minimal impact prediction using experimental data

Limited impact:

Rank	Feature	Rank	Feature
1	DIST00 STEMP00 POTT00	2	DIST00 UTEMP00 POTT00
3	DIST00 STEMP00	4	DIST00 POTT00
5	DIST00 STEMP00 UTEMP00	6	DIST00 UTEMP00
7	DIST00	8	DIST00 STEMP00 WESTERLY00
9	DIST00 UTEMP00 WESTERLY00	10	DIST00 WESTERLY00 POTT00
11	MM00 DIST00	12	DIST00 LO_HUMID00 WESTERLY00
13	DIST00 HI_HUMID00 WESTERLY00	14	MM00 DIST00 STEMP00
15	MM00 DIST00 POTT00	16	DIST00 WESTERLY00
17	MM00 DIST00 UTEMP00	18	DIST00 EASM00
19	DIST00 EASM00 POTT00	20	DIST00 HI_HUMID00 STEMP00
21	DIST00 LO_HUMID00 STEMP00	22	DIST00 STEMP00 EASM00
23	DIST00 HI_HUMID00 UTEMP00	24	DIST00 LO_HUMID00 UTEMP00
25	DIST00 UTEMP00 EASM00	26	MM00 DIST00 WESTERLY00
27	DIST00 STEMP00 VORT00	28	MM00 STEMP00

29	DIST00 SH_INT00 POTT00	30	DIST00 UTEMP00 SH_INT00
31	MM00 DIST00 EASM00	32	DIST00 STEMP00 SH_INT00
33	MM00 UTEMP00	34	DIST00 STEMP00 SH_EXT00
35	DIST00 UTEMP00 SH_EXT00	36	MM00 POTT00
37	DIST00 SH_EXT00 POTT00	38	DIST00 HI_HUMID00 POTT00
39	DIST00 LO_HUMID00 POTT00	40	MM00 DD00 UTEMP00
41	AZM00 UTEMP00 POTT00	42	MM00 UTEMP00 POTT00
43	DIST00 VORT00 POTT00	44	DIST00 SH_INT00
45	MM00 DIST00 LO_HUMID00	46	MM00 DIST00 HI_HUMID00
47	MM00 DIST00 VORT00	48	AZM00 UTEMP00
49	DIST00 UTEMP00 VORT00	50	DIST00 LO_HUMID00

Table 6.14: Top 50 polynomial features for limited impact prediction using experimental data

Substantial impact:

Rank	Feature	Rank	Feature
1	DIST00^2 WESTERLY00	2	DIST00^2 STEMP00
3	DIST00^2 POTT00	4	DIST00 UTEMP00^2
5	DIST00 STEMP00	6	DIST00 UTEMP00 POTT00
7	DIST00 POTT00^2	8	DIST00 UTEMP00
9	DIST00 STEMP00 UTEMP00	10	DIST00^3
11	DIST00^2 EASM00	12	DIST00
13	DIST00^2	14	DIST00^2 UTEMP00
15	DIST00 UTEMP00 WESTERLY00	16	DIST00 STEMP00 POTT00
17	DIST00 STEMP00 WESTERLY00	18	DIST00 POTT00
19	DIST00 WESTERLY00	20	MM00 DIST00^2
21	DIST00 STEMP00^2	22	DIST00 WESTERLY00 POTT00
23	DIST00^2 VORT00	24	MM00 DIST00 STEMP00
25	DIST00^2 AZM00	26	MM00 DIST00 UTEMP00
27	DIST00^2 U50000	28	DIST00^2 U20000
29	DIST00^2 HI_HUMID00	30	DIST00^2 LO_HUMID00
31	DIST00 HI_HUMID00 WESTERLY00	32	DIST00 LO_HUMID00 WESTERLY00
33	MM00 DIST00 POTT00	34	DIST00^2 VMAX00
35	DIST00^2 U_HK00	36	MM00 DIST00

37	DIST00 STEMP00 SH_EXT00	38	DIST00 EASM00 POTT00
39	DIST00^2 SPEED00	40	DIST00^2 V50000
41	DIST00^2 SH_EXT00	42	DIST00^2 SH_INT00
43	DIST00 SH_EXT00 POTT00	44	DIST00 UTEMP00 EASM00
45	DIST00 EASM00	46	MM00 UTEMP00 POTT00
47	MM00^2 STEMP00	48	DIST00 STEMP00 EASM00
49	DIST00 UTEMP00 SH_EXT00	50	DIST00 STEMP00 SH_INT00

Table 6.15: Top 50 polynomial features for substantial impact prediction using experimental data

Direct strike:

Rank	Feature	Rank	Feature
1	DIST00 UTEMP00 WESTERLY00	2	DIST00 STEMP00 WESTERLY00
3	DIST00 WESTERLY00 POTT00	4	DIST00 WESTERLY00
5	DIST00 LO_HUMID00 WESTERLY00	6	DIST00 HI_HUMID00 WESTERLY00
7	MM00 UTEMP00 POTT00	8	MM00 UTEMP00
9	MM00 DIST00	10	MM00 DIST00 UTEMP00
11	MM00 DIST00 WESTERLY00	12	MM00 DIST00 POTT00
13	DIST00 STEMP00	14	MM00 DIST00 STEMP00

15	MM00 DD00 UTEMP00	16	DIST00 UTEMP00 POTT00
17	DIST00 STEMP00 UTEMP00	18	DIST00 UTEMP00
19	DIST00 POTT00	20	DIST00 STEMP00 POTT00
21	DIST00	22	DIST00 VMAX00 STEMP00
23	DIST00 VMAX00 WESTERLY00	24	DIST00 SH_EXT00
25	DIST00 VMAX00 UTEMP00	26	DIST00 VMAX00
27	MM00 STEMP00	28	DIST00 HI_HUMID00 UTEMP00
29	DIST00 LO_HUMID00 UTEMP00	30	DIST00 STEMP00 SH_EXT00
31	MM00 DIST00 VORT00	32	MM00 VMAX00 UTEMP00
33	MM00 POTT00	34	DIST00 UTEMP00 VORT00
35	DIST00 LO_HUMID00 POTT00	36	DIST00 HI_HUMID00 POTT00
37	DIST00 STEMP00 SH_INT00	38	DIST00 WESTERLY00 SH_INT00
39	DIST00 VMAX00 POTT00	40	DIST00 VORT00 WESTERLY00
41	DIST00 HI_HUMID00	42	DIST00 LO_HUMID00
43	DIST00 SH_EXT00 POTT00	44	DIST00 WESTERLY00 SH_EXT00
45	MM00 STEMP00 POTT00	46	DIST00 UTEMP00

			SH_EXT00
47	DIST00 EASM00 POTT00	48	DIST00 VMAX00 SH_EXT00
49	MM00 STEMP00 UTEMP00	50	DIST00 LO_HUMID00 STEMP00

Table 6.16: Top 50 polynomial features for direct strike prediction using experimental data

Appendix 2: 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* (한국기상학회지), 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] 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.
- [8] 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).
- [9] J. A. Knaff, M. DeMaria, C. R. Sampson and J. M. Gross, “Statistical, 5-Day Tropical Cyclone Intensity Forecasts Derived from Climatology and Persistence,”

- Weather and Forecasting*, vol. 18, no. 1, pp. 80-92, Feb. 1, 2003. DOI: 10.1175/1520-0434(2003)018<0080:SDTCIF>2.0.CO;2.
- [10] J. A. Knaff, C. R. Sampson, M. DeMaria, T. P. Marchok, J. M. Gross and C. J. McAdie, “Statistical Tropical Cyclone Wind Radii Prediction Using Climatology and Persistence,” *Weather and Forecasting*, vol. 22, no. 4, pp. 781-791, Aug. 1, 2007. DOI: 10.1175/WAF1026.1.
- [11] J. A. Knaff, C. R. Sampson and K. D. Musgrave, “Statistical Tropical Cyclone Wind Radii Prediction Using Climatology and Persistence: Updates for the Western North Pacific,” *Weather and Forecasting*, vol. 33, no. 4, pp. 1093-1098, Aug. 1, 2018. DOI: 10.1175/WAF-D-18-0027.1.
- [12] F. Marks, G. Kappler and M. DeMaria. “Development of a Tropical Cyclone Rainfall Climatology and Persistence (R-CLIPER) Model,” National Oceanic and Atmospheric Administration (NOAA) Hurricane Research Division, Miami, FL, Dec. 31, 2003. Accessed: Sep. 27, 2021. [Online] Available: https://origin.www.nhc.noaa.gov/jht/final_rep/R-CLIPER_final.pdf.
- [13] J. Reid, J. Zhang, B. Sampson, J. Hansen and W. Sessions. “Climatology-Persistence Models (CLIPERs).” ICAP-Ensemble Meeting. http://icap.atmos.und.edu/EnsembleForecastsDataAssimilation/MeetingPDFs/May12/06_Reid-Cliper.pdf (retrieved Sep. 27, 2021).
- [14] J. A. Knaff and C. R. Sampson, “Southern hemisphere tropical cyclone intensity forecast methods used at the Joint Typhoon Warning Center, Part I: control forecasts based on climatology and persistence,” *Australian Meteorological and Oceanographic Journal*, vol. 58, no.1, pp. 1-7, Mar. 2009. DOI: 10.22499/2.5801.001.
- [15] 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.
- [16] “Tropical cyclone strike probability.” ECMWF.
https://www.ecmwf.int/en/forecasts/charts/tcyclone/tc_strike_probability
 (retrieved Sep. 26, 2021).
- [17] “TC Show Guide.” ECMWF.
http://www.ecmwf.int/sites/default/files/TC_ShowGuide.pdf (retrieved Sep. 26, 2021).
- [18] 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.
- [19] S. J. Majumdar and P. M. Finocchio, “On the Ability of Global Ensemble Prediction Systems to Predict Tropical Cyclone Track Probabilities,” *Weather and Forecasting*, vol. 25, no. 2, pp. 659-680, Apr. 1, 2010. DOI: 10.1175/2009WAF2222327.1.
- [20] L. Qi, H. Yu and P. Chen, “Selective ensemble-mean technique for tropical cyclone track forecast by using ensemble prediction systems,” *Quarterly Journal for the Royal Meteorological Society*, vol. 140, no. 680, pp. 805-813, Apr. 2014. DOI: 10.1002/qj.2196.
- [21] X. Zhang and H. Yu, “A Probabilistic Tropical Cyclone Track Forecast Scheme Based on the Selective Consensus of Ensemble Prediction Systems,” *Weather and Forecasting*, vol. 32, no. 6, pp. 2143-2157, Dec. 2017. DOI: 10.1175/WAF-D-17-0071.1.

- [22] 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.
- [23] W. Zhang, Y. Leung and J. C. L. Chan, “The Analysis of Tropical Cyclone Tracks in the Western North Pacific through Data Mining. Part II: Tropical Cyclone Landfall,” *Journal of Applied Meteorology and Climatology*, vol. 52, no. 6, pp. 1417-1432, Jun. 1, 2013. DOI: 10.1175/JAMC-D-12-046.1.
- [24] 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, Nov. 13, 2020. DOI: 10.1007/s00477-020-01930-w.
- [25] T. Loridan, R. P. Crompton and E. Dubossarsky, “A Machine Learning Approach to Modeling Tropical Cyclone Wind Field Uncertainty,” *Monthly Weather Review*, vol. 145, no. 8, pp. 3203-3221, Aug. 1, 2017. DOI: 10.1175/MWR-D-16-0429.1.
- [26] J. Devaraj, S. Ganesan, R. M. Elavarasan and U. Subramaniam, “A Novel Deep Learning-Based Model for Tropical Intensity Estimation and Post-Disaster Management of Hurricanes,” *Applied Sciences*, vol. 11, no. 9, pp. 4129, Jan. 1, 2021. DOI: 10.3390/app11094129.
- [27] 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.
- [28] 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.

- [29] 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. 15, 2012. DOI: 10.1175/JCLI-D-11-00507.1.
- [30] S. S. Chand, K. J. E. Walsh and J. C. L. Chan, “A Bayesian Regression Approach to Seasonal Prediction of Tropical Cyclones Affecting the Fiji Region,” *Journal of Climate*, vol. 23, no. 13, pp. 3425-3445, Jul. 1, 2010. DOI: 10.1175/2010JCL3521.1.
- [31] C. Wang and H. Zhang, “Probability-based estimate of tropical cyclone damage: An explicit approach and application to Hong Kong, China,” *Engineering Structures*, vol. 167, pp. 471–480, Jul. 15, 2018. DOI: 10.1016/j.engstruct.2018.04.064.
- [32] P. C. Chin, “Tropical Cyclone Strike Probability Values for Ten Target Locations in Southeast Asia,” Royal Observatory, Hong Kong, Occasional Paper No. 39, Nov. 1977.
- [33] M. E. Splitt, J. A. Shafer, S. M. Lazarus and W. P. Roeder, “Evaluation of the National Hurricane Center’s Tropical Cyclone Wind Speed Probability Forecast Product,” *Weather and Forecasting*, vol. 25, no. 2, pp. 511-525, Apr. 1, 2010. DOI: 10.1175/2009WAF2222279.1.
- [34] M. E. Splitt, S. M. Lazarus, S. Collins, D. N. Botambekov and W. P. Roeder, “Probability Distributions and Threshold Selection for Monte Carlo–Type Tropical Cyclone Wind Speed Forecasts,” *Weather and Forecasting*, vol. 29, no. 5, pp. 1155-1168, Oct. 1, 2014. DOI: 10.1175/WAF-D-13-00100.1.
- [35] 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.

- [36] M. DeMaria, J. A. Knaff, M. J. Brennan, D. Brown, R. D. Knabb, R. T. DeMaria, A. Schumacher, C. A. Lauer, D. P. Roberts, C. R. Sampson, P. Santos, D. Sharp and K. A. Winters, “Improvements to the Operational Tropical Cyclone Wind Speed Probability Model,” *Weather and Forecasting*, vol. 28, no. 3, pp. 586-602, Jun. 1, 2013. DOI: 10.1175/WAF-D-12-00116.1.
- [37] Q. Li, P. Xu, X. Wang, H. Lan, C. Cao, G. Li, L. Zhang and L. Sun, “An Operational Statistical Scheme for Tropical Cyclone Induced Wind Gust Forecasts,” *Weather and Forecasting*, vol. 31, no. 6, pp. 1817-1832, Dec. 1, 2016. DOI: 10.1175/WAF-D-16-0015-1.
- [38] HKO. “Hong Kong’s Tropical Cyclone Warning Signals.” https://www.hko.gov.hk/en/publica/gen_pub/files/tcws.pdf (accessed Jan. 23, 2022).
- [39] “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).
- [40] “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).
- [41] “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).
- [42] “1.16. Probability calibration.” scikit-learn 1.0.2 documentation. <https://scikit-learn.org/stable/modules/calibration.html> (accessed Jan. 23, 2022).
- [43] 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.
- [44] “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).
- [45] “Introduction to Boosted Trees.” xgboost 1.5.1 documentation. <https://xgboost.readthedocs.io/en/stable/tutorials/model.html> (accessed Nov. 30, 2021).
- [46] “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).
- [47] J. Friedman, “Multivariate adaptive regression splines,” *The Annals of Statistics*, vol. 19, no. 1, pp. 1-67, Mar. 1991. DOI: 10.1214/aos/1176347963
- [48] *py-earth*. (2017). Accessed Jan. 23, 2022. [Online] Available: <https://github.com/scikit-learn-contrib/py-earth>.
- [49] *pyGAM*. (2018), D. Servén, C. Brummitt. Accessed Apr. 18, 2022. [Online] Available: <https://github.com/dswah/pyGAM>. DOI: 10.5281/zenodo.1208723.
- [50] M. I. Jordan. The Kernel Trick [Online]. Available: <https://people.eecs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf>.
- [51] “sklearn.feature_selection.mutual_info_classif.” scikit-learn 1.0.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html (accessed Apr. 18, 2022).
- [52] 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>.
- [53] *alan-turing-institute/sktime*. (2021). Zenodo. Accessed: Jan. 11, 2022. [Online] Available: <https://zenodo.org/record/5610006#.Yd1erVI-VEY>. DOI: 10.5281/zenodo.3749000.
- [54] “Multilayer Perceptron.” DeepAI. <https://deepai.org/machine-learning-glossary-and-terms/multilayer-perceptron> (retrieved Oct. 3, 2021).
- [55] L. C. P. Velasco, R. P. Serquiña, M. S. A. A. Zamad, B. F. Juanico and J. C. Lomosco. “Week-ahead Rainfall Forecasting Using Multilayer Perceptron Neural Network,” *Procedia Computer Science*, vol. 161, pp. 386-397. 2019. DOI: 10.1016/j.procs.2019.11.137.
- [56] *PyTorch*. (2019). Accessed Apr. 18, 2022. [Online] Available: <https://github.com/pytorch/pytorch>.
- [57] 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. 2013. DOI: 10.1098/rsta.2011.0550.
- [58] 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.
- [59] “1.11 Ensemble methods.” scikit-learn 1.0.2 documentation. <https://scikit-learn.org/stable/modules/ensemble.html#stacking> (Accessed Apr. 18, 2022).
- [60] 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.
- [61] *Western North Pacific Ocean Best Track Data*, JTWC, n.d. [Online]. Available: <https://www.metoc.navy.mil/jtwc/jtwc.html?western-pacific>.

- [62] *International Best Track Archive for Climate Stewardship (IBTrACS)*, World Data Center for Meteorology, Asheville, n.d. doi:10.25921/82ty-9e16.
- [63] K. R. Knapp, M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, “The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data,” *Bulletin of the American Meteorological Society*, vol. 91, pp. 363-376. 2010. DOI:10.1175/2009BAMS2755.1.
- [64] *CMA Tropical Cyclone Best Track Dataset*, China Meteorological Administration, 2021. [Online]. Available: https://tcdata.typhoon.org.cn/en/ziljsjj_sm.html.
- [65] M. Ying, W. Zhang, H. Yu, X. Lu, J. Feng, Y. Fan, Y. Zhu and D. Chen, “An overview of the China Meteorological Administration tropical cyclone database,” *J. Atmos. Oceanic Technol.*, vol. 31, pp. 287-301, 2014. DOI: 10.1175/JTECH-D-12-00119.1.
- [66] X. Q. Lu, H. Yu, M. Ying, B. K. Zhao, S. Zhang, L. M. Lin, L. N. Bai and R. J. Wan, “Western North Pacific tropical cyclone database created by the China Meteorological Administration,” *Adv. Atmos. Sci.*, vol. 38, no. 4, pp. 690–699, 2021. DOI: 10.1007/s00376-020-0211-7.
- [67] *ECMWF Reanalysis v5 (ERA5)*, ECMWF, n.d. [Online] Available: <https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5>.
- [68] *NCEP FNL Operational Model Global Tropospheric Analyses*, continuing from July 1999, National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, 2000, DOI: 10.5065/D6M043C6.
- [69] 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.
- [70] 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.
- [71] “Millibar.” Merriam-Webster.com Dictionary. <https://www.merriam-webster.com/dictionary/millibar> (Accessed Apr. 18, 2022).
- [72] H. Blanchonett. “What is the direction convention for the U and V components of winds?” ECMWF Confluence Wiki. <https://confluence.ecmwf.int/pages/viewpage.action?pageId=111155337> (Accessed Apr. 18, 2022).
- [73] 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.
- [74] 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.
- [75] 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.
- [76] 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.
- [77] 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.
- [78] 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.
- [79] *PyNIO*. (2019). NCAR. Accessed Apr. 18, 2022. [Online] Available: <https://github.com/NCAR/pynio>
- [80] *NCAR Command Language 6.6.2*. (2019). NCAR. Accessed Apr. 18, 2022. [Online] Available: 10.5065/D6WD3XH5.
- [81] “FAQ on Reading PSL netCDF files”. NOAA Physical Sciences Laboratory. <https://psl.noaa.gov/data/gridded/faq.html#15> (Accessed Apr. 18, 2022).
- [82] “numpy.clip.” NumPy v1.22 Manual. <https://numpy.org/doc/stable/reference/generated/numpy.clip.html> (Accessed Apr. 18, 2022).
- [83] “1.16 Probability calibration.” scikit-learn 1.0.2 documentation. <https://scikit-learn.org/stable/modules/calibration.html#calibration> (Accessed Apr. 18, 2022)
- [84] “Time series classification.” sktime documentation. https://www.sktime.org/en/stable/api_reference/classification.html (Accessed Apr. 18, 2022).
- [85] “Time series regression.” sktime documentation. https://www.sktime.org/en/stable/api_reference/regression.html (Accessed Apr. 18, 2022).
- [86] “Forecasting.” sktime documentation. https://www.sktime.org/en/stable/api_reference/forecasting.html (Accessed Apr. 18, 2022).

18, 2022).

[87] “ThetaForecaster.” sktime documentation.

https://www.sktime.org/en/stable/api_reference/auto_generated/sktime.forecasting.theta.ThetaForecaster.html#sktime.forecasting.theta.ThetaForecaster (Accessed Apr. 18, 2022).

[88] “sklearn.linear_model.Ridge.” scikit-learn 1.0.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge (Accessed Apr. 18, 2022).