國立臺灣大學理學院大氣科學研究所

碩士論文

Department of Atmospheric Sciences

College of Science

National Taiwan University

Master Thesis

以機器學習方法預測颱風生成及其 SHAP 詮釋

Prediction of Tropical Cyclogenesis Based on Machine Learning Methods and its SHAP interpretation

呂智樂

Chi-Lok Loi

指導教授：吳俊傑 博士、梁禹喬 博士

Advisor: Chun-Chieh Wu, Ph.D., Yu-Chiao Liang, Ph.D.

中華民國 112 年 7 月

July 2023

doi:10.6342/NTU202303630

Acknowledgements

想不到兩年的碩士班生涯很快就要結束了，好像做了一個五味紛陳的夢：在台灣遇到了很多事情，有甜美的也有苦澀的。但就算如此，我還是認為從香港遠渡而來到台大唸書是沒有錯的選擇。其一是因為終於完成了一個出外闖蕩的願望：一直想要離家自立，趁年少時多遊歷。我認為，人本身就像一艘帆船，雖然停在港口無風無雨很安全，但它本身被創造的目的卻是揚帆冒險航向未知的大海。在這段自己當船長掌舵的旅程，更重要的，讓我覺得自己是個超級幸運兒的，是一群一直都在陪伴我，在背後支持我的朋友和長輩。總是有一種我何德何能的心情，居然能吸引這麼多願意幫助我、聽我講的同伴，實在相當感恩。

首先要感謝的是兩位指導教授吳俊傑老師和梁禹喬老師的悉心教導，在研究上指點迷津，提出可行的想法讓我嘗試。也感激他們很包容我這個不肖學生製造出來的麻煩。同時也多謝曾開治老師的知遇之恩，讓我當應數的助教，體驗教學的樂趣。在此謹以把院長獎獻給你們三位教授。

然後要多謝的是我的實驗室好夥伴：先嶸、禹安、浩廷、峻崑、明翰、俊宇、昱丞，跟隔壁實驗室的尚恩、毓琇。謝謝你們在研究上以及其餘事情上的建言、提點和鼓勵。TDRC 的氣氛很融洽、友善，對我來說是在台灣真正的家。

其他不在台灣的好朋友，像去了美國的玉來，以及香港的老朋友 King、Paul + Felix、Tonybill，也感謝你們縱使身隔千里，依然願意不計時差陪我聽我吐苦水。在相當困難、快要倒下的時候，成為了我站起來繼續前進的力量。另外也相當感謝讓我來台的爸媽，和願意接受我教學、熱情的學弟妹們。

「夢に向かえ まだ不器用でも」這是我最喜歡的幪面超人（台灣：假面騎士）龍騎主題曲《Alive a life》的一句歌詞。藉此勉勵自己就算現在還不夠格，但仍要努力向夢想邁進，大氣（器）晚成也！

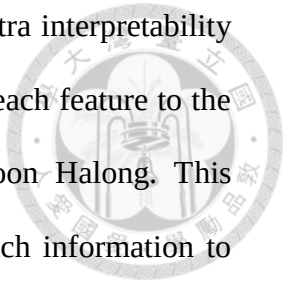
doi:10.6342/NTU202303630

Abstract

Predicting Tropical Cyclone Genesis (TCG) events has been a challenging research topic due to a lack of conclusive theory which unifies different hypotheses about TCG mechanisms. In practice, dynamical models are used to forecast TCG occurrence, but given some of its limitations in recent years machine learning has been proposed as an alternative low-cost approach that can utilize the abundance of reanalysis data. In this study, we attempt to train three machine learning models with varying complexity: Random Forest, Support Vector Machine, and Artificial Neural Network, by feeding various atmospheric and oceanic, dynamic and thermodynamic variables extracted from reanalysis data, to predict cyclogenesis at a forecast lead time of 24 hours for candidate tropical disturbances, identified by an optimized Kalman Filter algorithm. The overall performance is competent in terms of the f1-scores (~0.8) compared to previous researches of the same kind, with recalls (~0.9) generally higher than precisions (~0.7). Operational analysis data is used to further verify the practicality of the models.

An assessment by SHapley Additive exPlanations (SHAP) values reveals that mid-level (500 hPa) vorticity is the most influential factor in deriving the genesis probability at the lead time of 24 hours. Wind shear and tilting are found to possess a considerable level of importance as well. A sensitivity test is done to reaffirm the role of mid-level vorticity and tilting compared to the lower-level ones. These results encourage further experiments that use physical models to explore the dynamical, mid-level pathway to TCG. Nevertheless, some of the thermodynamic variables are also influential, with outer core humidity becoming significant when the forecast lead time is changed to 48

doi:10.6342/NTU202303630



hours. Another usage of SHAP values in this work is providing extra interpretability for the machine learning models, by listing out the contribution of each feature to the output genesis probability, illustrated by a case study of Typhoon Halong. This increases their reliability and forecasters can take advantage of such information to issue tropical cyclone formation warnings more accurately.

Finally, several caveats of current machine learning applications in TCG, including this work, are discussed. One of the main problems is the negligence of presumably negative samples from developing tropical disturbances that only reaches tropical cyclone status long after the required forecast lead time. Several potential improvements for future research are suggested correspondingly.

Keywords: Tropical Cyclones, Tropical Cyclone Genesis, Machine Learning, SHAP values

Disclaimer: The major portion of this research work has been submitted as a paper (Loi et al., 2023) and is currently under review in AGU Journal of Advances in Modeling Earth Systems (JAMES).

doi:10.6342/NTU202303630

摘要

由於缺少統一的理論，預測熱帶氣旋生成一直都是相當困難的研究議題。目前實作主要用動力模式預測熱帶氣旋生成，但機器學習方式最近被提出可作為低成本之替代品，能活用大量再分析資料。這份研究用再分析資料中的大氣及海洋變數，訓練了隨機森林、支持向量機、和神經網絡三個機器學習模型，以預測 24 小時內熱帶擾動生成能否發展為熱帶氣旋。機器學習模型總體表現不俗，f1-分數達 0.8，可比擬前人研究。召回率（約 0.9）普遍比精確率（約 0.7）高。作業用分析資料則進一步用來測試模型實用性。

其後，SHAP 值分析發現中層（500 百帕）渦度是影響熱帶氣旋在 24 小時內生成的最關鍵因素。風切及渦管傾斜也有一定重要性。敏感度測試確認了中層渦度及傾斜比起低層的更重要。此結果鼓勵更多物理模式實驗探討中層動力如何引致熱帶氣旋生成。SHAP 值也增加了機器學習模型的可解釋性。本研究以颱風哈隆為例，展示各變數對其生成預測機率之影響。如此可以增加機器學習模型的可靠度，並提升熱帶氣旋生成預警之準確度。

最後，本論文提出目前以機器學習方式預測熱帶氣旋生成的一些問題。其中之一為：忽略熱帶擾動於預測期間外生成的樣本。同時，亦提出針對各問題未來研究的可改善方向。

doi:10.6342/NTU202303630

Table of Contents

doi:10.6342/NTU202303630

doi:10.6342/NTU202303630

List of Tables

doi:10.6342/NTU202303630

List of Figures

doi:10.6342/NTU202303630

doi:10.6342/NTU202303630

Chapter 1

Introduction

1.1. Review of Current Machine Learning Works on TCG

The difficulties of predicting Tropical Cyclone Genesis (TCG) can be attributed to a variety of reasons, including a lack of surface observational data over oceanic regions, the subjectivity of the Dvorak method (IWTC-9), and the stochastic nature of tropical cyclone (TC) convection (Wang, 2018). One traditional approach to the problem of TCG is exploring dynamical forecast. However, the approach comes with issues such as an incomplete representation of complex physical processes, a resolution too coarse to explicitly resolve processes of sub-grid scales, as well as a substantial computational cost (Thatcher and Pu, 2013; Chen et al., 2020). There are also earlier works of traditional statistical analysis of TCG (Fu et al., 2012; Peng et al., 2012), but they are only diagnostic and hence unable to make predictions. The method of ensemble forecast was suggested to tackle these issues, and has since gained popularity in operational uses, as it can overcome the weaknesses of deterministic forecasts and provide uncertainty information useful for risk assessment (Titley et al., 2019). Nevertheless, an ensemble forecast requires much more computational resources to perform, and thus Machine Learning (ML) methods have been proposed as an alternative that requires low computational cost as compared to other numerical models (Brecht and Bihlo, 2022; Qian et al., 2022). Machine learning models can also be regarded to be a kind of more sophisticated and powerful statistical methods compared to the simpler ones like a logistic regression in the early years.

Recent advances in utilizing ML in the area of TCG have been summarized in

doi:10.6342/NTU202303630

Chen et al. (2020), and some of the relevant works (Zhang et al., 2015; Zhang et al., 2019; Zhang et al., 2022) mentioned there are to be discussed below. In Zhang et al. (2015), the 850 hPa vorticity field was examined to extract tropical disturbances, and a decision tree algorithm was developed based on six classification rules of dynamic and thermodynamic variables that were considered to be potentially influential in TCG. In particular, the vorticity at 800 hPa has to be greater than 4.2×10^{-5} s^{-1} and sea surface temperature (SST) has to be higher than 28.2 °C. They achieved an accuracy of 84.6% in forecasting TCG events. There was a clear visualization by the tree diagram but no feature importance or interpretation was given.

Meanwhile, Zhang et al. (2019) looked for Mesoscale Convective Systems (MCSs) from a satellite dataset (Huang et al. 2018) and included well-established indices like Genesis Potential Index (GPI) and Potential Intensity (PI) in addition to other environmental variables and properties of MCSs (e.g. the lowest/average brightness temperature, and the area coverage). They used a variety of ML algorithms, including the Decision Tree, k-Nearest Neighbors (KNN), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), Random Forest, and ADABoost. Different models were trained for different forecast lead times, including 6, 12, 24 and 48 hours. The ADABoost with a 6-hr lead time model yielded a stunning F1-score of 97.2%. However, the feature importance was expressed by Mean Decrease Impurity Importance (MDI). MDI was dependent on the tree-based nature of algorithms and not all of the models could be explained via MDI. Moreover, MDI only gave each variable a single scalar importance, and hence cannot provide further information about the detailed roles of variables in the process of tropical cyclogenesis.

doi:10.6342/NTU202303630

Finally, Zhang et al. (2022) took ERA5 and Gridded Satellite (GRIDSAT) data to train a Convolutional Neural Network (CNN) model that gave a probabilistic forecast on TCG. The input variables were relative vorticity, relative humidity, wind speed, wind direction, geopotential height, mean sea level pressure (SLP), as well as infrared window (IRWIN) channel brightness temperature, sampled over a box domain of 1000×1000 km centered at the disturbance. They generated probabilistic forecasts at different lead times according to the statistics of output probabilities of the CNN, as compared against the actual fraction of genesis events at those lead times. They also carried out probability bias analysis by constructing composites of different variables in the cases of "best hits" and "worst misses". Some findings were that the "best hits" had a closed geopotential contour and a positive vorticity anomaly near the center in the mid-level, while those were not present in the "worst misses". Finally, they discussed some real-life TC cases and how the model could be applied in these situations. Their probability of detection was 97.1% and the false alarm rate was 20.3%. Nevertheless, using CNN makes the feature patterns highly compressed, thus making the interpretation of the model's result difficult. Table 1.1 summarizes their works. The merits include clear visualization, feature importance information and high performance. A goal of this work is to integrate the advantages and address the shortcomings in these three papers.

1.2. Review of Physical Factors and Pathways Affecting TCG

1.2.1. Dynamical Variables

To create a physically sensible ML model on TCG which is compatible to known governing processes and laws, it is essential to provide as many input variables that

doi:10.6342/NTU202303630

bear a physical relation to TCG as possible. Among dynamical variables, vertical wind shear is perhaps one of the most studied topics in the area of TCG. Tao and Zhang (2014) and Finocchio et al. (2016) discussed the tilting and precession of the vortex tube at the initial stage of TCG under wind shear. Only after realignment of the vortex tube can the disturbance start to intensify. If the wind shear is too strong, the tilting will not be restored and the development cannot begin. Another impact of the vertical wind shear is the ventilation effect proposed by Tang and Emanuel (2010, 2012) where mid-level dry air (i.e. with a low equivalent potential temperature) intrudes the inner core through wind shear and the ensuing evaporation of rain leads to downdrafts of the low-entropy air into the boundary layer. This disturbs the TC energy cycle and hampers its growth.

Another important dynamical factor is the mid-level circulation/vorticity, which is closely linked to the bottom-up theory of TCG that developed by Montgomery et al. (2006). The study of Typhoon Nuri (2008) by Raymond and López Carrillo (2011) revealed that an apparent closed mid-level circulation was seen at the 5 km level as Nuri became a tropical depression. At the same time, the vertical mass flux profile evolved into a more bottom-heavy one (see Fig. 5 in Raymond et al., 2014), marked by a strongly increasing upward mass transport below the mid-level that resulted in the low-level spin-up and the intensification of Nuri. Raymond et al. (2011) proposed that a strong mid-level vortex is the key ingredient to such a bottom-heavy vertical mass flux profile. The first way that such a vertical mass flux profile contributes to TCG is by mass continuity which can cause an intense radial inflow and convergence of vorticity to spin up the low-level warm core vortex later. Another reason is that it reduces vertically integrated lateral export of moist entropy, raising the column-integrated moist static energy and the saturation fraction to the point of criticality, triggering deep

doi:10.6342/NTU202303630

convection. But it was noted that at the beginning, the mid-level circulation is accompanied by a low-level cold core due to the thermal wind relationship, instead of a warm core that is fully developed only after TCG. The emergence of a mid-level vortex with a warm(top)-cold(bottom) core couplet as TCG approaches is further confirmed by the simulations performed by Ge et al. (2013).

Finally, the Okubo-Weiss parameter, being a measure of the degree of rotation relative to deformation, can be derived to assess how well the vortical flow of a disturbance is preserved, as in the paper of Dunkerton et al. (2009), which explores the Marsupial Paradigm. Basically, it is hypothesized that a quasi-closed Lagrangian "pouch" in some "Kelvin Cat's Eye" at the critical layer of easterly wave where the relative flow is effectively zero, would be able to protect the growing disturbance by retaining moisture and resisting environmental wind shear. The Okubo-Weiss parameter in this context, is then an evaluation of the effectiveness of the "pouch" protection against unfavorable factors to TCG. In general, the normalized Okubo-Weiss parameter is calculated as in Raymond et al. (2011).

$$\mathrm{OW}_N = \frac{{\zeta_r}^2 - {\sigma_1}^2 - {\sigma_2}^2}{{\zeta_r}^2 + {\sigma_1}^2 + {\sigma_2}^2}$$

where ζ_r, σ_1, σ_2 are relative vorticity, stretching deformation and shear deformation respectively.

1.2.2. Thermodynamic Variables

On the other hand, some of the thermodynamic variables that may play a role in TCG are column-integrated water vapor and hence convection/precipitation, SST and upper-level warm core. For water vapor, Wang and Hankes (2016) suggested that when

doi:10.6342/NTU202303630

the Saturation Fraction (SF) reaches a critical threshold, exponential precipitation will be triggered. The positive feedback between convection and the low-level moisture convergence would then help sustain the disturbance, in a quasi-equilibrium state with a high precipitation rate and moderately high SF. This chain of events can be captured from multiple variables, such as the cloud brightness temperature that implies the strength of convective clusters. SST is a representative proxy for latent heat flux and ocean heat content, which are the energy source for the Carnot engine cycle proposed by Emanuel (1997), and also a favorable TCG environmental condition in Gray (1979). Last but not least, an upper-level warm core is a fundamental, indicative feature to any TC which continues to build up as the storm intensifies (Wang and Jiang, 2019). Kerns and Chen (2015) demonstrated that TCG occurs when subsidence warming erodes the lower-troposphere cool anomaly and superposes with the pre-existing mid-upper troposphere warm anomaly. This result is quite pertinent to the foregoing issue discussed of warm-cold core couplet, and that an upper-level warm core has to exist beforehand.

In addition, as suggested by Tang and Emanuel (2010, 2012), the ventilation effect arises not only due to the sole existence of wind shear but also the presence of dry, low-entropy air in the mid-level. They utilize a quantity known as the entropy deficit in the calculation of the ventilation index. This form of entropy deficit would be used as an indicator of the thermodynamic aspect of ventilation effect, or simply put, dryness, in the following analysis, independent of the actual, dynamical vertical wind shear.

$$\chi_m = \frac{{s_m}^* - s_m}{{s_{SST}}^* - s_b}$$

χ, s are the non-dimensional entropy deficit and entropy. The subscripts m, b represent

doi:10.6342/NTU202303630

mid-level (600 hPa) and boundary layer. The asterisk * means saturation value. The computation of *s* is based on the pseudo-adiabatic entropy formula from Bryan (2008), which is employed in Tang and Emanuel (2012) as well.

1.3. Objectives of this Work

The main goal of our work is to improve the prediction of TCG events, discerning developing/non-developing tropical disturbances, with ML as a supplement to conventional numerical models. It aims to integrate the merits from and deal with the issues raised by three aforementioned papers on TCG (Zhang et al., 2015; Zhang et al., 2019; Zhang et al., 2021), which is to balance between having a satisfactory performance, and providing a clear visual interpretation of the feature's importance. The forecast lead time of 24 hours is chosen for our machine learning models, because the current ensemble forecast uncertainty becomes increasingly large after 24 hours (Wang et al., 2020) and the Tropical Cyclone Formation Alert (TCFA) by the Joint Typhoon Warning Center (JTWC) is also valid only during the 24 hours after the issuing. We also aim to achieve the following two objectives in this study: first, to understand from a machine learning perspective, which are the most dominant factors for TCG, and compare them to earlier research to verify consistency; second, to explain and illustrate how different physical variables would lead to the bifurcation of development/non-development of disturbances, enhancing the interpretability of the model.

The current chapter has summarized how ML had been applied to the study of TCG and outlined some background knowledge of TCG. Chapter 2 will depict the data used and the methodology. Chapter 3 will present the model results and interpretation.

doi:10.6342/NTU202303630

In Chapter 4, some extensions and further testing of the model will be discussed, along with the inspection of a specific disturbance case (Typhoon Halong, 2014). Discussion and summary will be given in Chapters 5 and 6 respectively.

doi:10.6342/NTU202303630

Chapter 2

Data and Methodology

2.1. Data Used, Spatial Extent and Time Period of Study

The European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5, Hersbach et al., 2018a, 2018b) data with a spatial resolution of 0.25×0.25 degrees are used to identify disturbances and extract the atmospheric variables used in the machine learning training process. Although the ERA5 data are available at an hourly frequency, the sampling is only done four times each day at an interval of six hours (00:00, 06:00, 12:00, 18:00 UTC). Meanwhile, SST, precipitation and the near 11 μm infrared brightness temperature are acquired from the Optimum Interpolation Sea Surface Temperature (OISST, Huang et al., 2020), Tropical Rainfall Measuring Mission (TRMM, Huffman et al., 2016) and NOAA Geostationary Satellite (GRIDSAT, Knapp et al., 2011) datasets, all of which have been re-gridded to the same spatial resolution of 0.25 degrees as the ERA5 data. Operational analysis data of atmospheric variables and precipitation from GFS is later used for further verification to check transferability. Despite the presence of finer 0.25-degree GFS data, we have selected the coarser version with a spatial resolution of 0.5 degrees because of a better accessibility. They are then interpolated to 0.25 degrees and sampled at the same interval of six hours like previously. Due to data availability at the time of writing, only the GFS data from 2021 is used. The regions to be investigated cover the Western North Pacific (WNP) and Central Pacific (CP) (0° - 25°N, 125°E - 165°W, Fig. 2.1). The time period covers June to September from 2003-2015, where the four-month period mostly coincides with the highest seasonal TC activity in WNP (Gao et al., 2020).

doi:10.6342/NTU202303630

2.2. Disturbance Tracking Algorithm by Kalman Filter

This study uses a Kalman filter (Saho, 2017) to detect and track cyclonic centers using the ERA5 850 hPa vorticity field. The vorticity field is smoothed by a Gaussian with a standard deviation of 2 degrees to reduce noise. The Kalman filter tracker is combined with the commonly used area-overlapping method in series (rather than two parallel procedures as in Huang et al., 2018). The detailed implementation of which, K2O Algorithm, is described below. The abbreviation K2O stands for "Kalman-to-Overlapping". At each time step, positive vorticity local maxima that have a vorticity of at least 5×10^{-6} s^{-1} are searched over the gridded data with the constraint that any weaker local maximum is discarded if it is less than 4 degrees away from some stronger local maxima to filter away those vorticity centers being absorbed. Once all such vorticity maxima are identified, they are ranked, arranged and iterated from the highest to lowest. At each iteration, the connected region around the local vorticity maximum where the constituent grids have vorticity values above α% of that maximum is marked and defined as a vorticity cluster. If the determined vorticity cluster spans a previous vorticity cluster in space, then it is discarded. Again, this is to remove weaker systems being heavily influenced by other stronger systems.

After searching all the possible vorticity clusters, at the first time step, we initialize a Kalman Filter for each of them that tracks their vorticity-weighted center position. The Kalman filter has a state vector of

$$\mathbf{x}_\mathrm{t} = (y_t, x_t, v_t, u_t)^T$$

doi:10.6342/NTU202303630

which is composed of the latitude (y_t) and longitude (x_t) of the vorticity-weighted center position, as well as the meridional (v_t) and zonal velocity in degrees (u_t) at the current time. The transition matrix is

$$\mathrm{A} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Where the time step size Δt is 0.25 days. The measurement matrix is

$$\mathrm{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

and the initial process covariance matrix is

$$\mathrm{Q} = \sigma_{\mathrm{a}}^2 \begin{bmatrix} \frac{(\Delta t)^4}{4} & 0 & \frac{(\Delta t)^3}{2} & 0 \\ 0 & \frac{(\Delta t)^4}{4} 1 & 0 & \frac{(\Delta t)^3}{2} \\ \frac{(\Delta t)^3}{2} & 0 & (\Delta t)^2 & 0 \\ 0 & \frac{(\Delta t)^3}{2} & 0 & (\Delta t)^2 \end{bmatrix}$$

with σ_{a} set to 16 degrees/day^2. The initial measurement noise covariance matrix is

$$\mathrm{R} = \begin{bmatrix} \frac{(\Delta t)^2}{2} & 0 \\ 0 & \frac{(\Delta t)^2}{2} \end{bmatrix}$$

Then, at each future time step (t = k+1), each Kalman Filter will predict the movement of the old vorticity cluster center, represented by the blue patch in Fig. 2.2, which will have been tracked by the Kalman Filter at the last time step (t = k). It predicts the state vector by

$$\mathbf{x}_{\mathrm{t+1}} = \mathrm{A}\mathbf{x}_{\mathrm{t}}$$

doi:10.6342/NTU202303630

and forecasts the covariance matrix as

$$P_{t+1} = AP_tA^T + Q$$

All old clusters at t = k are then displaced by their predicted motions, which is illustrated by the green patch in Fig. 2.2, that represents the would-be positions of the old clusters at t = k+1. They are compared to the real new vorticity clusters found in the way mentioned in the last paragraph at t = k+1, demonstrated by the red patch in Fig. 2.2. If there are M old clusters (green) offset by predicted movements and N new clusters (red), then there would be M×N comparisons that can be used to compute the fraction of overlapping by the formula below:

$$\frac{\text{Number of overlapped grids between old (green) and new cluster (red)}}{\text{Number of grids in the old cluster (green)}}$$

Then, we use the Hungarian Algorithm (*scipy.optimize.linear_sum_assignment* in *Python*) that attempts to match each old cluster to a new cluster by maximizing the total overlapping between all old and new clusters. The minimum degree of overlapping is set to be γ%, only above which the pair would be considered as a potential match. If an old cluster (t = k) at the last time step is matched to a new cluster (t = k+1) at the following time step, then they are considered the same disturbance and the Kalman Filter would read the new position to update its internal analysis position, passing the Kalman Filter itself to keep tracking the new vorticity cluster. The analysis step involves the update of the state vector and covariance matrix as follows.

$$S = HPH^T + R$$

$$K = PH^TS^{-1}$$

$$\mathbf{x}_{\text{analysis}} = \mathbf{x}_{\text{forecast}} + K(\mathbf{z} - H\mathbf{x}_{\text{forecast}})$$

$$P_{\text{analysis}} = (I - KH)P_{\text{forecast}}$$

doi:10.6342/NTU202303630

If no corresponding new cluster can be found for an old cluster, it would be considered to have dissipated, and the corresponding Kalman Filter would stop. The entire track of the decayed old cluster is then archived. New Kalman Filters will be initialized for the remaining new clusters that are not linked to any old cluster, and they will enter the tracking procedure starting from the next time step. This process is repeated and continued until the last time step. Only the tracked disturbances satisfying the requirements listed in the next paragraph would be withheld. It seems that setting α=75 and γ=15 is a good empirical choice, which allows most of the historical TCs to be successfully tracked without discontinuity. Or simply put, the single TC entity is not split into two separated tracks or considered as two different systems by the K2O algorithm.

To filter out weak TC seeds, a detected tropical disturbance needs to have a lifetime of at least 3 days. Also, for at least 4 six-hourly time steps, it is required to be sufficiently strong such that its vorticity at 850 hPa is greater than 2×10^{-5} s^{-1}. The vorticity field has been smoothed by a Gaussian filter with a standard deviation of 2 degrees. These requirements are modified from those in Ikehata and Satoh (2021). A final criterion is the mean precipitation in a centered box of 5×5 degrees needs to exceed 0.5 mm/hr for more than 4 time steps as well, in order to remove any dry vortex. The definition of a developing disturbance is simply reaching an intensity of 25 knots or above as recognized by Joint Typhoon Warning Center (JTWC). The time when this first happens is defined as the genesis time. Otherwise, it is non-developing. The International Best Track Archive for Climate Stewardship (IBTrACS), which contains the best track data by JTWC, are used for this purpose. Table 2.1 shows the numbers of

doi:10.6342/NTU202303630

developing and non-developing disturbances, both of which having a balanced sample size coincidentally.

To check the robustness of our tracking method, we also try the well-established Tempest Extremes tracker for reference, and find that our procedure can trace ~85% of the systems identified by Tempest Extremes. While there certainly exists a performance margin, we believe that the Kalman filter method is a simple yet adequately effective method. Also, while it may be surprising that the numbers of developing and non-developing disturbances are similar at first, it is actually due to the stringent conditions placed on the disturbance seeds to be selected. Before applying the thresholds, the non-developing cases are much more (about 5 times) than developing cases as one may have expected.

2.3. Features Selection

The 16 (14+2) variables used to train the ML models and how they are extracted are listed in Table 2.2. Apart from the criterion that they are physically related to TCG, each of them also has to satisfy the condition that the difference of the two probability distributions sampled between the developing and non-developing sets are statistically meaningful. This is confirmed by both the Student's t-test and Kolmogorov-Smirnov test at 99% significance level (see Table 2.3). Two examples of such probability distributions (500 hPa vorticity and 500-850 hPa layer vertical wind shear) are displayed in Fig. 2.3. Nevertheless, two of the variables (denoted by the "+2"), entropy deficit (*Chi*) and instability index (*I*) are incorporated even though these two variables do not pass the 99% significant level. This is because they are viewed as important indicators of TC growth in certain literatures (Raymond et al., 2011; Raymond et al.,

doi:10.6342/NTU202303630

2014), and indeed improves the results considerably, which can be seen in Section 3.1. We further utilize these two statistical tests to refine the vertical levels or horizontal domains where the predictors are extracted through minimizing the computed p-values. The horizontal domains in which the predictors are computed is illustrated in Fig. 2.4. For developing cases, the variables are averaged in time from day -1 (24 hours) to day -1.5 (36 hours) before genesis to reduce possible transient noises. For non-developing cases, a similar time averaging of predictors is done between the moment when the disturbance reaches its maximal 850 hPa vorticity and half a day before the peak.

2.4. Machine Learning Models Used and Training Details

Three Machine Learning models are employed in our study: Random Forest (Liaw and Wiener, 2002), Support Vector Machine (SVM, or SVC, C stands for Classification, Cortes and Vapnik, 1995), and Artificial Neural Network (ANN, Rumelhart et al., 1985). The first two models usually output a definite True or False binary (developing or non-developing in terms of TCG) answer while the last one gives an estimate of genesis probability. Random Forest, being a tree-type algorithm, works simplistically through bisection but suffers (partly alleviated by its bagging nature) when the decision boundaries are supposed to be slanted or curved. SVM can handle non-linearity with the so-called “kernel trick”, often using the radial basis function kernel to map the data points to an infinite-dimensional space. However, the success of the kernel trick is not always guaranteed, and otherwise the dividing hypersurface may not separate the classes cleanly. ANN is the most complex one out of the three selected models. It has the most powerful potential yet quite prone to overfitting and other classes of training problems like dead neurons and vanishing gradient (Géron, 2019).

doi:10.6342/NTU202303630

The time period of training set is chosen as 2003-2013, and the testing set is from the years of 2014-2015 (see Table 2.1 for the counts in the two sets of data). A stratified 5-fold cross-validation for Random Forest and SVM, and a simple 80-20 validation for ANN, are executed in the training set when tuning the hyperparameters/architectures of the three models (see below), and also in a subsequent sensitivity test where predictors are added or removed. For each of the three models, the exactly same splitting of data points is kept throughout the testing (apart from Section 4.3 where the datasets are expanded). Particularly, for Random Forest and SVM, the procedure of grid search is applied. We have to choose two hyperparameters *max_tree_depth* and *ccp_alpha* in random forest, and one hyperparameter *C*, the regularization constant in SVC. Grid search works by cross-validating the model at each possible combination of hyperparameters and return the configurations that have the best score. For example, if we have *k* values of max_depth and *l* of ccp_alpha to select, then the grid search checks all the *k*l* possibilities, which is demonstrated in Fig. 2.5. The final optimal settings of hyperparameters for them are shown in Table 2.4, along with the architecture of the ANN model in Fig. 2.6. To evaluate the performance of the model, metrics including precision, recall and f1-score are calculated. Precision is the fraction of true positives to all positive predictions and recall is the fraction of true positives to all actual positives, while f1-score is the harmonic average (the reciprocal of the average of the reciprocals) of precision and recall. The higher the precision, the less the false alarms. The higher the recall, the less the misses. Since f1-score considers both precision and recall, it is high only if the previous two metrics are high (Zhang et al., 2019). These are demonstrated by the confusion matrix (a 2×2 contingency table split by actual and predicted True or False) in Table 2.5.

2.5. Introduction of SHAP Values

doi:10.6342/NTU202303630

To interpret the model results, SHAP values are utilized (Lundberg and Lee, 2017). SHAP values have been widely employed in the area of TC Rapid Intensification (e.g. Griffin et al. 2022). It represents the average marginal contribution of a feature, i.e. the mean difference in output between all constructed coalitions of variables that have the specific feature versus those without, estimated for a specific data sample. The detailed explanation can be found in Molnar (2022). The key formula of SHAP values is

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} (v(S \cup \{i\}) - v(S))$$

where φ_i is the SHAP value for the *i*-th variable, S is any possible coalition of features, v(S) is the output probability of the coalition S. N represents all variables and $\setminus\{i\}$ means except the *i*-th variable. There is one SHAP value for each feature in each sample, and it indicates how the value of the chosen feature in that sample affects the decision of the model: in the current TCG scenario, it is reflecting whether the contribution to the output genesis prediction of the physical variable, e.g. vorticity or SST, is positive/increasing or negative/decreasing. Refer to Fig. 2.7 for a simple made-up example for the computation of SHAP values. By examining the SHAP value distributions for each predictor, we can rank their importance, as would be seen in the next section. Often, a large SHAP value spread means the variable is crucial for the decision of the model.

doi:10.6342/NTU202303630

Chapter 3

Results and Interpretation

3.1. Models Performance

The performances of the three models are listed in Table 3.1. They all achieve high recalls and medium precisions, i.e. few misses but more false alarms, leading to decent f1-scores of around 77-82%. SVC and ANN have a better recall and hence a higher f1-score due to the two more correct positives. The increased hit rate can ensure better safety precautions for approaching disturbances. Nevertheless, all three models are useful due to the fact that their training processes are independent from each other. It is because through comparing their predictions and the bases on which they perform such predictions, we may obtain useful insights. They are to be seen in the upcoming Section 3.2. The performances when only the 14 variables that passes statistical tests are used, without the entropy deficit and instability index, is alternatively shown in Table 3.2 for reference. By comparison, it can be seen that the inclusion of the two additional variables lead to an increase in f1-scores in all three models by roughly 3-6%, which is not negligible, and support the foregoing decision to designate them as inputs.

3.2. SHAP Values Patterns

3.2.1. Beeswarm Plots

The SHAP values of the three models computed over the test data are displayed in Fig. 3.1 below as the so-called “beeswarm” plots. As mentioned in the last paragraph of Section 2.5, the larger the spread of SHAP values (represented by dots in the plots) a variable has, the greater its mean absolute SHAP value (refer to Fig. 3.2) and the more

doi:10.6342/NTU202303630

important the feature. To further elaborate, the magnitude of any (positive/negative) SHAP value (individual dot) represents the change (increase/decrease) in the predicted genesis probability caused by the feature value of that specific sample. An apparent common finding is that 500 hPa vorticity heavily dominates the prediction of TCG as its importance ranks the first among all three models. It is clear from the SHAP value distribution that when the vorticity at 500 hPa is large, indicated by the reddish dots far to the right, the SHAP values are largely positive and the predicted genesis probability substantially increases. When the mid-level vorticity is small, indicated by the bluish dots much to the left, the predicted chance of TCG decreases markedly.

Meanwhile, other variables show no consistent patterns with regard to importance across the three models. Nevertheless, the ordering of variable importance aligns roughly with the magnitude of the p-values in the two initial statistical tests (see Table 2.3). Namely, a smaller p-value and a greater statistical difference are correlated to a higher SHAP spread, except the additionally included *Chi* and *I* ("+2"). This shows that the SHAP results are reasonable. However, the more interesting findings would come from their difference or similarity in SHAP value patterns of some indicators. First, Tanomaly ranks the second in the random forest model and has an asymmetric SHAP value distribution skewed to the negative side, which is not found in the other two models. Low feature values have more pronounced reduction in SHAP values while high feature values display diminishing increases in SHAP values. It suggests the possibility of Tanomaly having some sort of upper saturation limit, and increasing it further beyond only marginally increases the genesis probability, similar to the concept of diminishing return, and which is like a necessary condition from the perspectives of the Random Forest model. This likely pertains to the fact that random forest builds up through binary separation, and in some top-level nodes, most disturbances that have a

doi:10.6342/NTU202303630

weak warm core are easily discerned and discarded early. This makes physical sense, as an adequate warm core is the backbone of any TC, following Wang and Jiang (2019) that has been included in the introduction. The other two models do not rely on binary separation and this is probably why the importance of Tanomaly falls off in these two models. Entropy deficit (*Chi*) also has a similarly skewed SHAP distribution, where few cases lie to the very left negative side. These cases correspond to the scenario of severe dry air intrusion that hampers the convection. This helps reject some extreme negative cases, hence reducing false positives compared to the alternative experiment without the "+2" addition (*Chi* and *I*), despite them having a low overall SHAP importance. The same situation, with the relative sign of feature value reversed, occurs for *tilt500* in all the three models. Finally, wind shear (*ws*) is also crucial, as it ranks second in SVC and ANN. As a factor impacted by wind shear, the degree of tilting is also quite relevant and will be considered in the upcoming composite analysis in section 3.3.1.

3.2.2. Dependence Plots

Here, we further examine the dependence plots of 500 hPa vorticity and vertical wind shear which have just been found to be essential factors determining TCG according to the SHAP explanation, demonstrated in Fig. 3.3 and 3.4 respectively. The baseline of zero SHAP value is denoted by the orange horizontal line, and the nearby sample points along which the influence of the concerned feature on them is neutral. Note that the SHAP values of both 500 hPa vorticity and vertical wind shear vary quite monotonically, and hence we can loosely define an estimated threshold value for them as where the transition from negative to positive (or reverse) SHAP values happens, which is indicated by the green arrow in the plots. For 500 hPa vorticity (Fig. 3.3), such

doi:10.6342/NTU202303630

transition mainly occurs near 20-25 × 10^{-6} s^{-1}. Notice that the extracted 500 hPa vorticity comes from reanalysis data and has been preprocessed by a Gaussian smoother and may not accurately reflect the real atmospheric conditions, but it gives an approximate scale of magnitude for reference. To the right of this threshold, the models consider the stronger vorticity as producing a positive contribution to the predicted genesis probability. In contrast, to the left, the contribution of the weaker vorticity is negative. Similarly, for wind shear (Fig. 3.4), the models agree that a vertical wind shear of around 4-4.5 m s^{-1} is the turning point where it becomes increasingly detrimental to TC formation. However, it should be noted that these threshold values are close to the average of all data. This possibly suggests that, alongside the monotonic behavior of the computed SHAP values for these two variables, the effects on TCG by these two predictors are fairly linear, and the averages (black dotted vertical lines in Fig. 3.3 and 3.4) are acting as a reference point. Moreover, the threshold may also be sensitive to how the predictors are extracted and calculated.

3.3. Investigation of Dynamical Variables

3.3.1. Shear-coordinate Composites

To shed more light on the above findings, using the prediction of ANN, four shear-coordinate composites of vorticity framed in a 2×2 contingency table are illustrated in Fig. 3.5. It can be seen that true positive cases have the strongest vorticity at both low-level (850 hPa) and mid-level (500 hPa) and they are well-aligned with each other. These two conditions are known to be favorable for the TCG process as discussed in the introduction. But there is significant tilting in false positive samples even if the values of vorticity at the two levels are still somehow large. The vertical wind shear, on average, is also a bit higher in the false positive class. The downshear-left direction of

doi:10.6342/NTU202303630

tilting is consistent with Tao and Zhang (2014) and Finocchio et al. (2016) where the same occurred at the initial times of their simulations. This shows that the model fails to reject negative cases when they have high vorticity that promises some growth potential but the tilting under vertical wind shear is unfavorable, disrupting the vortex tube, impeding convection, and hints otherwise. Lastly, there are only two false negative cases, which are too few to yield any statistically meaningful statement.

We already know that 500 hPa vorticity plays an essential part in the decision-making process of the model. Also, the wind shear and degree of tilting are shown to be relatively important by the SHAP analysis. The composites above may lead to an implication that, either the four more important dynamical variables being inspected for the moment (*vo500*, *vo850*, *tilt500*, and *ws*) are still not sufficiently considered and utilized by the models, or the data points consisted of these four features overlap to a certain extent that prevents discriminative, well-drawn decision boundaries.

3.3.2. Principal Component Analysis (PCA)

To answer the question in the last subsection and confirm the plausibility of the proposed explanation related to the construction of decision boundaries, Principal Component Analysis (PCA, *sklearn.decomposition.PCA* in *Python*) is applied to reduce the four-dimensional feature space to two-dimensional (retaining ~80% of the total variance) phase space for visualization, and some overlapping of points is clearly observed in the middle, as indicated by Fig. 3.6. The four variables have been normalized before computing PCA over the 59×4 testing data matrix and the first two principal components (PC1; PC2) are shown in Table 3.3. The PC1 mainly captures vorticity and tilting which have opposite signs while the PC2 is basically just vertical wind shear. The top-left-bottom-right orientation of the overlapping indicates that the

doi:10.6342/NTU202303630

difficulty of distinguishing between developing and non-developing disturbances mainly lies in the confrontation of high (low) vertical wind shear versus large (small) vorticity and slight (severe) tilting. Hence, it is necessary to introduce other features to give more information and construct a more discerning ML model.

3.4. Geographic Distribution of Cases

We also mark the geographic locations of disturbances, which are separated into the four outcomes classified by the ANN model, 24 hours before genesis (developing) or at the time of reaching 850 hPa vorticity maximum (non-developing) in Fig. 2.1. Many true negative cases are located over the CP near the anti-Meridian (longitude of 180°) which is a gray area with a persistently low TCG frequency (refer to the image hosted in the NASA website by Rohde, 2006), and the model captures this fact quite well. There is a cluster of false positives near the Philippines. It was first speculated that the terrain effect may play a role in producing such results by interfere the growth of some nearby disturbances which may otherwise develop successfully without the disruption of mountains. To investigate this, a follow-up test has been conducted in which a distance-to-land parameter is added. It linearly increases from 0 over land to 1 at sea over 1000 km distance to land, and saturates afterwards. The outcome shows no improvement in all three models. While we won't deny the potential influence from terrain on the actual TCG process (e.g. disrupting the circulation, reducing energy flux. General adverse terrain impacts on TCs are summarized in Petilla et al., 2023), we believe that it is not decisively important in the prediction of TCG by machine learning approach, at least in the current settings.

3.5. Transferability to Operational Analysis Fields

doi:10.6342/NTU202303630

The final section in this chapter is to conduct a test using operational analysis data from GFS to establish the transferability of the reanalysis-trained models to receive operational analysis outputs. The same models trained by the ERA5 datasets are used, without making new ones from GFS. The results are shown in Table 3.4. The performance degrades with a drop of f1-scores ranging from about 3-13% but remains acceptable, in the sense that it is still much better than the naive forecast (guessing all true, with 10/18 = 56% accuracy). Random forest retains the most skill and highest hit, with ANN has one less false alarm case. It is not surprising as random forest has the simplest, straight-forward structure which is less vulnerable to overfitting during extrapolation. There are several possible reasons for the performance drop. First, different physical models and parameterization are used in ERA5 reanalysis and GFS analysis data. Also, the sources and methods in the process of data assimilation between reanalysis versus analysis may vary. Such fundamental differences prevent perfect transfer learning to operational forecast. Finally, fewer cases available for the operational verification may suffers from more noises in the analysis data.

Anyways, in practice, the operational unit can always train their own version of machine learning models (as elaborated in Section 5.2.2 later) so this would not be a serious problem. The more important question is whether there are non-model data unavailable at operational real-time. The 16 variables, while some are obtained from satellite or other sources in our training process in this work, should be all available in any operational center at analysis time. But the real problem is the quality of some variables such as precipitation and brightness temperature which are known to be not so accurate. Solutions include adding a data quality control module for these variables before supplying them to the machine learning models, or reducing the weighting of

doi:10.6342/NTU202303630

these features when training the models. The transferability study in this section, however, exhibits the possibility to use real data in training the machine learning models and apply them in research-use simulation systems like WRF or HiRAM, which can be a rewarding new research frontier.

doi:10.6342/NTU202303630

Chapter 4

Extended Works

4.1. Sensitivity Test of Vorticity and Tilting Variables

This section is an extension study of the results in the previous section, aiming to explore the dependence and relative importance of variables used to train the models. The first one is a sensitivity test to ascertain the importance of the mid-level 500 hPa vorticity in the models. We re-train the same models of Random Forest and SVC (but not ANN due to the considerable stochasticity of the optimizer method) four times: 1) Without 500 hPa vorticity and tilting ("NIL"); 2) With 500 hPa vorticity and tilting ("500"); 3) With 500 hPa vorticity and tilting replaced by those at 700 hPa level ("700"); and 4) With both vorticity and tilting at 500 and 700 hPa levels included. Using the mild randomness in training Random Forest over bootstrapped samples, we train 100 random forests with different random states (that controls the bootstrapping of samples) in each of the four scenarios. The corresponding cross-validation f1-scores of SVC, and those averaged over the multiple random forests (with the first and third quartiles drawn to show the scoring spread) are displayed in Fig. 4.1. It can be seen that the "500" experiment has the best cross-validation f1-score in Random Forest, and f1-scores in both Random Forest and SVC are significantly higher than the "NIL" experiment. If we only use the 700 hPa vorticity and tilting information in the "700" experiment, the performance drops, especially in the random forests. The use of vorticity and tilting information at both levels in the last experiment does not yield a notably better performance than the "500" experiment. From these results, one can conclude that the mid-level vortex has an irreplaceable importance in TCG prediction. This is because, only when a complete vertical vortex structure is present throughout the lower-middle

doi:10.6342/NTU202303630

troposphere would the TC become self-sustainable. If the vorticity data at a lower level is used instead, the mid-level vortex, being a decisive sign of cyclogenesis as explained previously, would not be taken into consideration in the model prediction, and the performance would degrade accordingly. Another point is that the 700 hPa level is too close to the 850 hPa level that has been accounted for in the model prediction, and provides less new information than 500 hPa vorticity.

4.2. Case Study of Typhoon Halong (2014) by Waterfall Plots

Next, we utilize the waterfall plot function to perform a simple case study on the test set, and investigate if the ANN model makes sensible decisions when calculating the predicted genesis probability, to increase its transparency. The ANN model is used to produce the waterfall plot because it returns probability rather than an absolute True or False (1/0) as the other two models usually do. A waterfall plot shows the magnitude of changes in output value due to different factors in one individual sample, arranged in descending order, starting from the biggest increase/decrease. The selected case is Typhoon Halong (WP112014) which is a true positive determined by the ANN model already included in the test set, and the corresponding waterfall plot is shown in Fig. 4.2. It can be seen that the vorticity at 500 hPa (28.5×10^{-6} s^{-1}), above the threshold of 20-25 $\times$ 10^{-6} s^{-1} found in Section 3.2.2, contributes positively to the TCG prediction to a certain extent (~+8%). However, there is a counter effect (~-16%) from the high vertical wind shear (6.5 m s^{-1}) that exceeds the mid-level vorticity contribution. In fact, in the early life time of Halong, its growth was inhibited by the wind shear produced by the large circulation of the nearby monsoon depression Nakri (WP122014), which is clearly reflected by the *ws* term in the waterfall plot. Nevertheless, the next two features *bt* and *vo850* are favorable, both of which being above the mean values in the

doi:10.6342/NTU202303630

developing group (not shown). Together, they lead to a substantial increase (+14%) in the output genesis probability and subsequently a correct positive prediction. As Nakri moved away, its adverse impacts (mainly wind shear) on Halong diminished, and Halong promptly intensified.

This is consistent with the presentations in Fig. 4.2 and 4.3, among which the latter is another waterfall plot but re-drawn 6 hours later than the former (independent of the original test set). It can be clearly seen from their comparison that, as the wind shear decreases, its negative SHAP distribution wanes. Meanwhile, other favorable factors like *bt* and *vo500* are enhanced (+11%/+13%), yielding more positive SHAP values. The net effect is a ~+23% increase in the predicted probability of genesis compared to 6 hours before. This case study shows that the ANN model is able to infer TCG events with physically sound reasons. Based on the waterfall plot output, combined with other known information, the forecaster can make a good judgment on how the disturbance would evolve. In this particular case, the forecasters may have expected that the wind shear, originally with the largest negative SHAP value, would weaken for Halong, potentially based on their experience or the dynamical forecast. They would then be more confident about its formation, even when the waterfall plot in Fig. 4.3, which serves as a verification case, was not yet available.

4.3. Preprocessing Attempt by EOF

To make our model more parsimonious, we attempt to apply Empirical Orthogonal Function (EOF, see Hannachi et al., 2007) to the 16 input variables before training, and take the first few EOF modes that have the most explained variances to re-train the ML models. Theoretically, EOF can achieve the merits of reduced overfitting and increased interpretability. However, the actual cross-validation and testing performance with the

doi:10.6342/NTU202303630

EOF preprocessing is not good as simply using the raw inputs, as shown in Fig. 4.4. Only when most of the EOFs (≥10) are used, then the cross-validation score approaches that of the base models. And even so, the cross-validation and testing score are still worse than the base models. Some possible explanations include the fact that EOF is linear, but the relationship between variables and the models can be non-linear. In addition, EOF forces all the principal eigenvectors to be orthogonal, which produces spurious and physically meaningless weighting combinations across different features.

4.4. Additional Sampling of Negative Cases Backward in Time

A final extension to be made is inclusion of data collected 48 hours before TC formation, as additional negative samples for training. The motivation to apply such a change is the speculation that the growing disturbances might have undergone only a slight or moderate change during the gestation period before their genesis. The features could be very similar in magnitude and the difference might not be recognizable by the model. Such a condition would then likely produce a lot of false positives, in the sense that the model could predict a formation under 24 hours, while in some occasions a disturbance can actually develop after 24 hours. It may be desirable that the model could tell precisely if the disturbance would develop or not, in the next 24 hours exactly, and should return a negative answer if the disturbance would form only at a later time. Previous studies like Zhang et al. (2019) have not considered this situation. While it is also beneficial to predict TC genesis in any future time, i.e. without a fixed time window, this additional consideration would highly complicate our research topic and is out of the scope in this thesis for now.

To achieve our intention, we have added 22 more negative cases with the same set of 16 features sampled 48 hours before the development of disturbance into TCs, and

doi:10.6342/NTU202303630

hence a total of 34+22=56 negative cases, the number of which is more than twice of the positive cases. The performance of the re-trained models is shown in Table 4.1. There is a significant rise in false alarm cases by ~10 when compared to Table 3.1, and the precision falls to ~0.5. This also incurs a chain effect of reduced true positive cases and a drop of f1-score to ~0.6. Surprisingly, many of the false positives are "shuffled" where the previous false positives become true negatives and vice versa, further revealing the high sensitivity of the inclusion of extra negative cases. It strongly implies that the difference within the lifetime of a disturbance ("intra-case") is much minor than that between different disturbances ("inter-case"), and this constitutes an obstacle for the model to delineate decision boundaries effectively. Intuitively, it is illustrated in the schematic diagram of Fig. 4.5, where it is much harder to make a clean decision boundary when the late-developing negative samples are added to the gray area between the positive cases and non-developing negative cases. This shows that it is more difficult than previously anticipated to accurately determine if a disturbance would develop strictly within the specified time window, and to tackle this problem can be one of the major directions in any future ML work on TCG.

doi:10.6342/NTU202303630

Chapter 5

Discussion

5.1. Comparison with Other Studies and Operational Forecast

The overall performance of the three models in Section 3 is comparable to that in Zhang et al. (2019), where they also used Random Forest and SVC along with a variety of other models and achieved a f1-score of 0.790 and 0.657 at the lead time of 24 hours in WNP. However, there are some contrasts between our study and Zhang et al. (2019). First, the performances across models in Zhang et al. (2019) have a much higher variability, with f1-scores ranging from 0.532 to 0.817. In addition, models in Zhang et al. (2019) always yielded greater precision than recall, which is opposite to our observation. It may stem from the fundamental difference that the objects being analyzed by Zhang et al. (2019) were mesoscale convective systems that are basically clusters of clouds inferred from infrared brightness temperature, whereas our study focuses on high low-level vorticity regions. Moreover, the number of samples is much smaller while the classes are more balanced in numbers in our work. Hence, extra cautions should be taken when drawing any conclusion from the two studies.

According to the Annual Tropical Cyclone Report 2020 of JTWC (Francis and Strahl, 2021), its TCFA has a recall of 94% and precision of 80% (Table 1-4 in their report), achieving a f1-score of ~86%. It is true that the machine learning models developed in this work have not yet reached the same performance of the current operational forecast, in particular, suffering from a lower precision. However, we believe that with the first two suggestions proposed in Section 6.2, the gap may be closed. Another consideration is that, in the foreseeable future, machine learning

doi:10.6342/NTU202303630

techniques would still be a heavily researched topic and likely become more advanced, and the performance would improve accordingly. Moreover, our machine learning approach in the area of TCG is still useful, in the sense that it can potentially uncover new insights and relationships that can in turn improve the operational assessment of TCG events.

5.2. Variable Importance and Caveats

5.2.1. Mid-level Vorticity

The importance of the mid-level vorticity found may be explained by the mechanism proposed in Raymond et al. (2011), which has been described in the introduction. The building up of the mid-level vortex shortly before TCG facilitates a transition from a top-heavy to bottom-heavy vertical mass flux profile, which in turn promotes low-level mass and vorticity convergence and eventually leads to the formation of a complete TC vortex. However, it does not rule out the possibility that thermodynamic or other unused/unknown variables can also participate in the process, maybe to a lesser degree. Actually, the two thermodynamic features SST and Tanomaly are mostly ranked as intermediate by importance. In addition, some thermodynamic variables such as SST (needs to be greater than 26.5°C, see Dare and McBride, 2011) and convection act more like a trigger and afterwards their effects on genesis wear off (hypothesized for Tanomaly in Section 3.2.1), and are not reflected in the prediction of approaching TCG events for pre-existing disturbances. For instance, most of disturbance cases have an SST of at least 27°C, which is well above the minimal requirement.

5.2.2. Caveats and Feasibility to Deploy Operationally

doi:10.6342/NTU202303630

Another issue is that the resolution of the ERA5 reanalysis data is so coarse that small-scale structures of vertical pressure velocity and divergence/convergence cannot be properly resolved. It disables their usage in the training of machine learning models as the comparison of these variables between the developing and non-developing sets displays unreasonable patterns, such as the average upward pressure velocity being greater in the non-developing set than the developing set. So, more delicate phenomena like Vortical Hot Tower (VHT) mergers in Montgomery et al. (2006) cannot be inferred from the reanalysis data, and it is not possible to design the machine learning architecture in a way so as to verify the paradigm. Nevertheless, mid-level vorticity, which usually manifests as a large-scale structure, can be used as a proxy to these unresolved features, enabled by the supposed establishment of a mid-level vortex after the VHT formation and coalescence according to those studies. In fact, the effectiveness of feeding 500 hPa vorticity into model training has been confirmed by the preceding SHAP and composite analysis in Section 3, as well as the sensitivity test in Section 4.1.

When fed with the analysis fields of an operational weather system, the training of the ML models here only requires the area-averaged variables (as computed according to Table 2.2) and hence yield the same results regardless of resolutions. However, different assimilation techniques as well as different physical process representation would have led to varying values of variables. Although the deviations in feature magnitude are possibly limited, they may produce some fluctuations in the machine learning model results, and requiring re-training when deployed for different operational analysis using the corresponding analysis data. Nevertheless, in general, our work provides a machine learning framework that, when necessary, can be adopted to various modeling systems as separately trained models.

doi:10.6342/NTU202303630

Chapter 6

Summary

6.1. Overall Findings

In this study, developing and non-developing disturbances in the WNP and CP regions over a 13-year period are identified and three ML models, Random Forest, SVC, and ANN, are trained to recognize TCG events at the lead time of 24 hours. The performance is quite satisfactory and comparable to the existing literature with a f1-score close to 0.8. The novelty of our work lies in the use of SHAP value to explain predictions generated by the models, which is its first trial in the field of TCG. It is shown that SHAP analysis can assist forecasters in determining TCG, through quantifying how much the various factors contribute to the predicted genesis probability, and would help make the models more transparent. This is an improvement in relation to previous works of the same topic. According to the SHAP analysis, mid-level vorticity is the most crucial indicator in cyclogenesis prediction. Vertical wind shear, along with tilting, ranks second in importance in SVC and ANN. Shear-coordinate composites are produced and a sensitivity experiment is conducted to verify the findings. From these, we conclude that those dynamical variables inducing the establishment of a complete vertical vortex are the main drivers in the TCG process as recognized by our ML approach. These are consistent with previous research, and invite further investigation into the physical mechanisms on how these factors interplay in TCG. Specifically, a reflection on theories associated with the mid-level vortex pathway to TCG may worth further study (e.g. Raymond et al., 2011; Ge et al., 2013).

6.2. Future Works

doi:10.6342/NTU202303630

Some future directions for improvement on TCG forecast by ML may include: 1) Use a ML model that considers the entire time series of the disturbance before genesis, such as a recurrent neural network (RNN), so that the change of feature values in time can be explored, and continuous predictions that are temporally coherent can be made; 2) Add more potentially useful indicators. Although it is found that the difference in entropy deficit between the developing and non-developing sets does not pass the two statistical tests used in our study at 99% significance level (see Table 2.3), it is worthwhile to include them in the training process, as seen by comparing Tables 3.1 and 3.2. Even when the objective criterion aforementioned is not fulfilled, one can always attempt to include more variables subjectively that are empirically known to be essential in the process of TCG. The selection can be optimized using a greedy iterative algorithm, such as one similar to the forward selection method outlined in Section 7.4.2 of Wilks (2021), or the commonly used elbow method; 3) Include negative samples representing developing disturbances that form TCs only after the lead time similar to what is done in Section 4.4, and construct a ML model that is more specific to the lead time. This will help reduce false positives that behave as premature warnings; 4) Standardize the definition of disturbances across different studies. Low-level vorticity and brightness temperature are two frequent choices, and comparison can be meaningfully performed if future researchers adhere to only one definition; 5) Employ more powerful interpretation tools when available, to enhance the transparency of the decision-making process of models, through which researchers can obtain more insights about the physics of TCG. This is equally important for operational tasks, so that forecasters can rely on a more trustworthy ML prediction.

doi:10.6342/NTU202303630

References

Brecht, R., & Bihlo, A. (2022). Computing the ensemble spread from deterministic weather predictions using conditional generative adversarial networks. *arXiv*, release version. https://doi.org/10.48550/arXiv.2205.09182

Chen, R., Zhang, W., & Wang, X. (2020). Machine learning in tropical cyclone forecast modeling: a review. *Atmosphere*, 11(7), 676. https://doi.org/10.3390/atmos11070676

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning.* 20 (3), 273–297. https://doi.org/10.1007/BF00994018

Dare, R. A., & McBride, J. L. (2011). The threshold sea surface temperature condition for tropical cyclogenesis. *Journal of Climate*. 24(17), 4570-4576. http://dx.doi.org/10.1175/JCLI-D-10-05006.1

Emanuel, K. A. (1997). Some aspects of hurricane inner-core dynamics and energetics. *Journal of the Atmospheric Sciences*. 54(8), 1014-1026. https://doi.org/10.1175/1520-0469(1997)054<1014:SAOHIC>2.0.CO;2

Finocchio, P. M., Majumdar, S. J., Nolan, D. S., & Iskandarani, M. (2016). Idealized tropical cyclone responses to the height and depth of environmental vertical wind shear. *Monthly Weather Review*. 144(6), 2155-2175. https://doi.org/10.1175/MWR-D-15-0320.1

doi:10.6342/NTU202303630

Francis, A. S. & Strahl, B. R. (2021). Annual Tropical Cyclone Reports 2020. *Joint Typhoon Warning Center*. https://www.metoc.navy.mil/jtwc/products/atcr/2020atcr.pdf

Fu, B., Peng, M. S., Li, T., & Stevens, D. E. (2012). Developing versus nondeveloping disturbances for tropical cyclone formation. Part II: Western North Pacific. *Monthly Weather Review*. 140(4), 1067-1080. https://doi.org/10.1175/2011MWR3618.1

Gao, S., Zhu, L., Zhang, W., and Shen, X. (2020). Western North Pacific tropical cyclone activity in 2018: A season of extremes. *Scientific Reports*. 10, 5610. https://doi.org/10.1038/s41598-020-62632-5

Ge, X., Li, T., & Peng, M. S. (2013). Tropical cyclone genesis efficiency: Mid-level versus bottom vortex. *Journal of Tropical Meteorology*. 19(3), 197-213.

Géron, A. (2019). Hands-on machine learning with Scikit-learn, Keras, and Tensorflow. Sebastopol, CA: O' Reilly.

Griffin, S. M., Wimmers, A., & Velden, C. S. (2022). Predicting rapid intensification in North Atlantic and eastern North Pacific tropical cyclones using a convolutional neural network. *Weather and Forecasting*, 37(8), 1333-1355. https://doi.org/10.1175/WAF-D-21-0194.1

Hannachi, A., Jolliffe, I.T., & Stephenson, D.B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of*

doi:10.6342/NTU202303630

Climatology. 27(9), 1119-1152. https://doi.org/10.1002/joc.1499

Hersbach, H. et al. (2018a): ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.adbb2d47

Hersbach, H. et al. (2018b). ERA5 hourly data on pressure levels from 1959 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://doi.org/10.24381/cds.bd0915c6

Huang, B. et al. (2020): Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1. *Journal of Climate*. 34, 2923-2939. https://doi.org/10.1175/JCLI-D-20-0166.1

Huang, X., Hu, C., Huang, X., Chu, Y., Tseng, Y., Zhang, G. J., & Lin, Y. (2018). A long-term tropical mesoscale convective systems dataset based on a novel objective automatic tracking algorithm. *Climate Dynamics*. 51, 3145-3159. https://doi.org/10.1007/s00382-018-4071-0

Huffman, G. J., Bolvin, D. T., Nelkin E. J., & Adler, R. F. (2016). TRMM (TMPA) Precipitation L3 1 day 0.25 degree × 0.25 degree V7, edited by Andrey Savtchenko, *Goddard Earth Sciences Data and Information Services Center (GES DISC)*. Accessed: 2022/08/05, https://doi.org/10.5067/TRMM/TMPA/DAY/7

doi:10.6342/NTU202303630

Ikehata, K., & Satoh, M. (2021). Climatology of tropical cyclone seed frequency and survival rate in tropical cyclones. *Geophysical Research Letters*. 48(18), e2021GL093626. https://doi.org/10.1029/2021GL093626

Kerns, B. W., & Chen, S. S. (2015). Subsidence warming as an underappreciated ingredient in tropical cyclogenesis. Part I: Aircraft observations. *Journal of the Atmospheric Sciences*. 72(11), 4237-4260. https://doi.org/10.1175/JAS-D-14-0366.1

Knapp, K. R. et al. (2011). Globally gridded satellite (GridSat) observations for climate studies. *Bulletin of the American Meteorological Society*. 92, 893-907. https://doi.org/10.1175/2011BAMS3039.1

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*. 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *arXiv*, v2. https://doi.org/10.48550/arXiv.1705.07874

Molnar, C. (2022). Shapley Values. In *Interpretable machine learning: A guide for making black box models explainable* (Ch 9.5). Online. https://christophm.github.io/interpretable-ml-book/shapley.html

Montgomery, M. T., Nicholls, M. E., Cram, T. A., & Saunders, A. B. (2006). A vortical hot tower route to tropical cyclogenesis. *Journal of the Atmospheric Sciences*. 63(1), 355-386. https://doi.org/10.1175/JAS3604.1

doi:10.6342/NTU202303630

Peng, M. S., Fu, B., Li, T., & Stevens, D. E. (2012). Developing versus nondeveloping disturbances for tropical cyclone formation. Part I: North Atlantic. *Monthly Weather Review*. 140(4), 1047-1066. https://doi.org/10.1175/2011MWR3617.1

Petilla, C. E. R., Tonga, L. P. S., & Olaguera, L. M. P. et al. (2023). Changes in intensity and tracks of tropical cyclones crossing the central and southern Philippines from 1979 to 2020: an observational study. *Progress in Earth and Planetary Science*. 10, 32. https://doi.org/10.1186/s40645-023-00563-1

Qian, Q., Jia, X., & Lin, Y. (2022). Reduced tropical cyclone genesis in the future as predicted by a machine learning model. *Earth's Future*, 10(2), e2021EF002455. https://doi.org/10.1029/2021EF002455

Raymond, D., Gjorgjievska, S., Sessions, S., & Fuchs, Ž. (2014). Tropical cyclogenesis and mid-level vorticity. *Australian Meteorological and Oceanographic Journal*. 64, 11-25. https://doi.org/10.22499/2.6401.003.

Raymond, D. J., & López Carrillo, C. (2008). The vorticity budget of developing typhoon Nuri (2008). *Atmospheric Chemistry and Physics*. 11(1), 147–163. https://doi.org/10.5194/acp-11-147-2011

Raymond, D. J., Sessions, S. L., & López Carrillo, C. (2011). Thermodynamics of tropical cyclogenesis in the northwest Pacific. *Journal of Geophysical Research: Atmospheres*. 116, D18101. https://doi.org/10.1029/2011JD015624

doi:10.6342/NTU202303630

Rohde, R. A. (2006). *Historic Tropical Cyclone Tracks*. Global Warming Art. https://earthobservatory.nasa.gov/images/7079/historic-tropical-cyclone-tracks

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. University of California, San Diego, CA: Institute for Cognitive Science.

Saho, K. (2017). Kalman Filter for moving object tracking: Performance analysis and filter design. In G. L. de Oliveira Serra (Ed.), *Kalman Filters - Theory for advanced applications* (Ch. 12). IntechOpen. http://dx.doi.org/10.5772/intechopen.71731

Tang, B., & Emanuel, K. (2010). Midlevel ventilation's constraint on tropical cyclone intensity. *Journal of the Atmospheric Sciences*. 67(6), 1817-1830. https://doi.org/10.1175/2010JAS3318.1

Tang, B., & Emanuel, K. (2012). A ventilation index for tropical cyclones. *Bulletin of the American Meteorological Society*. 93(12), 1901-1912. https://doi.org/10.1175/BAMS-D-11-00165.1

Tao, D., & Zhang, F. (2014). Effect of environmental shear, sea-surface temperature, and ambient moisture on the formation and predictability of tropical cyclones: An ensemble-mean perspective. *Journal of Advances in Modeling Earth Systems*. 6(2), 384-404. https://doi.org/10.1002/2014MS000314

doi:10.6342/NTU202303630

Thatcher, L., & Pu, Z. (2013). Evaluation of tropical cyclone genesis precursors with relative operating characteristics (ROC) in high resolution ensemble forecasts: Hurricane Ernesto. *Tropical Cyclone Research and Review*, 2(3), 131-148. https://doi.org/10.6057/2013TCRR03.01

Titley, H. A., Yamaguchi, M., & Magnusson L. (2019). Current and potential use of ensemble forecasts in operational TC forecasting: Results from a global forecaster survey. *Tropical Cyclone Research and Review*, 8(3), 166-180. https://doi.org/10.1016/j.tcrr.2019.10.005

Wang, C., Zeng, Z. & Ying, M (2020). Uncertainty in tropical cyclone intensity predictions due to uncertainty in initial conditions. *Advances in Atmospheric Sciences*. 37, 278-290. https://doi.org/10.1007/s00376-019-9126-6

Wang, X., & Jiang, H. (2019). A 13-Year global climatology of tropical cyclone warm-core structures from AIRS data. *Monthly Weather Review*. 147(3), 773-790. https://doi.org/10.1175/MWR-D-18-0276.1

Wang, Z. (2018). What is the key feature of convection leading up to tropical cyclone formation? *Journal of the Atmospheric Sciences*, 75(5), 1609-1629. https://doi.org/10.1175/JAS-D-17-0131.1

Wang, Z., & Hankes, I. (2016). Moisture and precipitation evolution during tropical cyclone formation as revealed by the SSM/I–SSMIS retrievals. *Journal of the Atmospheric Sciences*. 73(7), 2773-2781. https://doi.org/10.1175/JAS-D-15-0306.1

doi:10.6342/NTU202303630

Wilks, D. (2011). Screening Predictors. In *Statistical Methods in the Atmospheric Sciences* (Vol. 100). Academic Press. ISBN: 978-0-12-385022-5

Wingo, M. T., & D. J. Cecil (2010). Effects of Vertical Wind Shear on Tropical Cyclone Precipitation. *Monthly Weather Review*. 138(3), 645–662. https://doi.org/10.1175/2009MWR2921.1

Zhang, R., Liu, Q., Hang, R., & Liu, G. (2022). Predicting tropical cyclogenesis using a deep learning method from gridded satellite and ERA5 reanalysis data in the western North Pacific basin. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-10. https://doi.org/10.1109/TGRS.2021.3069217

Zhang, T., Lin, W., Lin, Y., Zhang, M., Yu, H., Cao, K., & Xue, W. (2019). Prediction of tropical cyclone genesis from mesoscale convective systems using machine learning. *Weather and Forecasting*, 34(4), 1035-1049. https://doi.org/10.1175/WAF-D-18-0201.1

Zhang, W., Fu, B., Peng, M. S., & Li, T. (2015). Discriminating developing versus nondeveloping tropical disturbances in the western North Pacific through decision tree analysis. *Weather and Forecasting*, 30(2), 446-454. https://doi.org/10.1175/WAF-D-14-00023.1

doi:10.6342/NTU202303630

Machine Learning Models Used	Decision Tree [Zhang et al. (2015)]	Random Forest and AdaBoost [Zhang et al. (2019)]	Convolutional Neural Network [Zhang et al. (2022)]
Pros	It is easy to visualize the decision of model.	MDI (Mean Decrease Impurity) can be computed for feature importance.	It has the most powerful performance among the three works.
Cons	There is no feature importance calculation.	The MDI importance requires the machine learning model being a tree-based algorithm. It only outputs a single scalar value, without detailed analysis.	It is very difficult to interpret and visualize compressed features.

Table 1.1. Pros and cons of the selected machine learning works on TCG [Zhang et al. (2015), Zhang et al. (2019) and Zhang et al. (2022)].

doi:10.6342/NTU202303630

	Developing	Non-Developing	Marginal
Training/ Validation (2003-2013)	102+25=127 (52%)	94+24=118 (48%)	245 (100%)
Testing (2014-2015)	25 (42%)	34 (58%)	59 (100%)
Operational Verification for Transferability (2021)	10 (56%)	8 (44%)	18 (100%)
Total	162 (50%)	160 (50%)	322 (100%)

Table 2.1. The numbers and percentages of developing and non-developing disturbances in training/validation and testing set. “Marginal” means the sum of a row. The numbers to the left/right of plus sign (+) in the training/validation row indicates the counts of training set and validation set respectively. Notice that the data split is not the standard “80%/10%/10%”. Rather, the training/validation and testing sets are first split by the time periods indicated in the parentheses, and then the training/validation set is internally split again by a “80%/20%” ratio.

doi:10.6342/NTU202303630

Predictors (Total: 14+2=16) [abbreviation]	Pre-processing
Vorticity at 850 hPa (low-level) [vo850]	Gaussian smoothing with a standard deviation of 2 degrees, and the location of its maximal value chosen as the low-level disturbance center.
Vorticity at 500 hPa (mid-level) [vo500]	Same smoothing as above. The maximal value closest to the low-level disturbance center defined above is identified, and the location of which is denoted as the mid-level disturbance center.
Tilting of the vortex tube [tilt500]	The distance between the low-level and mid-level disturbance centers assigned as above.
300-1000 hPa column-integrated water vapor (inner) [q_inner]	Averaged inside a circle of 1-degree radius centered at the low-level disturbance center. (Other predictors below are all extracted from inside of some regions that are similarly centered at the low-level disturbance center.)
300-1000 hPa column-integrated water vapor (outer) [q_outer]	Averaged inside an annular region spanning from 2 degrees to 4 degrees.
Fraction of grid points with precipitation greater than 0.25 mm/hr [prec]	Inside a circle having a radius of 6 degrees.
Temperature anomaly between 300-600 hPa [Tanomaly]	The mean inside a circle of 1 degree radius, minus the mean over a box of 10×10 degrees.

doi:10.6342/NTU202303630

Okubo-Weiss parameter at 600 hPa [OW600]	Averaged inside a circle having a radius of 3 degrees. If the vorticity at a pixel is negative, a minus sign is added before the averaging.
Latitude [lat]	--
Vertical Wind Shear between 500 and 850 hPa levels [ws]	Computed as the difference between the mean wind at 500 and 850 hPa level inside an annular ring of 2 degrees to 8 degrees. This annular ring is referenced from literatures like Wingo and Cecil (2010).
Moist Static Energy at 1000 hPa [mse1000]	Averaged inside a circle with a radius of 4.5 degrees.
Sea surface temperature [SST]	Averaged inside a circle with a radius of 3 degrees.
Mean infrared brightness temperature [bt]	Averaged inside a circle with a radius of 6 degrees.
Fraction of grid points with infrared brightness temperature below 218K [btarea]	Counted inside a circle with a radius of 6 degrees.
Entropy Deficit [Chi]	Follow Tang and Emanuel (2012) exactly
Instability Index [I]	Follow Raymond et al. (2014), averaged within a circle with a radius of 1 degree.

Table 2.2. The list of input predictors for model training and their pre-processing procedure.

doi:10.6342/NTU202303630

Predictors (Total: 14+2 = 16)	Student's t	K-S
vo850	1.63×10^{-18}	3.33×10^{-16} *
vo500	1.06×10^{-27}	3.33×10^{-16} *
tilt500	6.05×10^{-17}	9.90×10^{-12}
q_inner	5.69×10^{-5}	3.04×10^{-3}
q_outer	3.43×10^{-12}	1.88×10^{-9}
prec	4.17×10^{-3}	2.39×10^{-3}
Tanomaly	2.88×10^{-17}	3.33×10^{-16} *
OW600	4.27×10^{-11}	5.27×10^{-8}
lat	4.50×10^{-12}	5.31×10^{-14}
ws	3.15×10^{-6}	1.09×10^{-5}
mse1000	1.28×10^{-13}	3.68×10^{-12}
SST	7.61×10^{-6}	4.12×10^{-4}
bt	7.33×10^{-7}	3.95×10^{-5}
btarea	1.51×10^{-6}	5.94×10^{-4}
Chi	0.157	0.533
I	0.067	0.013

Table 2.3. The p-values of Student's t-test and Kolmogorov-Smirnov test for all the 14+2 = 16 variables. The asterisk * indicates the smallest possible p-value of the K-S test computed by the *Python scipy.stats* library.

doi:10.6342/NTU202303630

Machine Learning Models	Hyperparameters
Random Forest	max_depth=15, ccp_alpha=0.005, n_trees=100
SVC	StandardScalar at the top, C=0.25, RBF kernel
ANN	See Fig. 5 for the architecture. Adamax optimizer: learning_rate=0.001, beta_1=0.9, beta_2=0.999; Early stopping based on validation accuracy with a patience of 100 steps; Simple scheduler with a factor of exp(-0.025) applied to every step after 50 iterations; Loss: Binary Cross-entropy

Table 2.4. The values of hyperparameters supplied to the three candidate machine learning models.

doi:10.6342/NTU202303630

		Prediction		
		True	False	
Actual	True	True Positive (Hit)	False Negative (Miss)	*Recall = TP/(TP+FN)*
	False	False Positive (False Alarm)	True Negative (Correct Negative)	
		Precision = TP/(TP+FP)		*F1-Score = 2*(Rec*Prec)/(Rec+Prec)*

Table 2.5. Confusion matrix of four possible outcomes and formula for the three metrics used. Note: Here positive implies developing and negative implies non-developing.

doi:10.6342/NTU202303630

	Random Forest	SVC	ANN
(Cross-) Validation Accuracy	0.847	0.870	0.857
TP (Hits)	21	23	23
FN (Misses)	4	2	2
FP (False Alarms)	8	8	8
TN (Correct Negative)	26	26	26
Test Recall/Precision	0.840/0.724	0.920/0.741	0.920/0.741
Test F1-Score	0.777	0.821	0.821

Table 3.1. Performances of the three machine learning models (Random Forest, SVC, ANN) with the full 14+2 = 16 features.

doi:10.6342/NTU202303630

	Random Forest	SVC	ANN
(Cross-) Validation Accuracy	0.820	0.849	0.857
TP (Hits)	21	22	23
FN (Misses)	4	3	2
FP (False Alarms)	9	11	10
TN (Correct Negative)	25	23	24
Test Recall/Precision	0.840/0.700	0.880/0.667	0.920/0.697
Test F1-Score	0.764	0.759	0.793

Table 3.2. Same as Table 3.1 with only the first 14 variables which pass the statistical tests.

doi:10.6342/NTU202303630

	vo500	*vo850*	*tilt500*	*ws*
PC1	-0.619	-0.604	0.502	-0.010
PC2	-0.072	-0.029	-0.105	0.991

Table 3.3. The four components of the first two EOF (Empirical Orthogonal Function) vectors in the PCA analysis (Section 3.3.2).

doi:10.6342/NTU202303630

	Random Forest	SVC	ANN
TP (Hits)	9	8	8
FN (Misses)	1	2	2
FP (False Alarms)	5	5	4
TN (Correct Negative)	3	3	4
Test Recall/Precision	0.900/0.643	0.800/0.615	0.800/0.666
Test F1-Score	0.750	0.695	0.727

Table 3.4. Same as Table 3.1 but on the operational verification set.

doi:10.6342/NTU202303630

	Random Forest	SVC	ANN
TP (Hits)	19	21	19
FN (Misses)	6	4	6
FP (False Alarms)	19	20	19
TN (Correct Negative)	37	36	37
Test Recall/Precision	0.760/0.500	0.840/0.512	0.760/0.500
Test F1-Score	0.603	0.636	0.603

Table 4.1. Same as Table 3.1 with an addition of 22 negative samples 48 hours before TCG.

doi:10.6342/NTU202303630

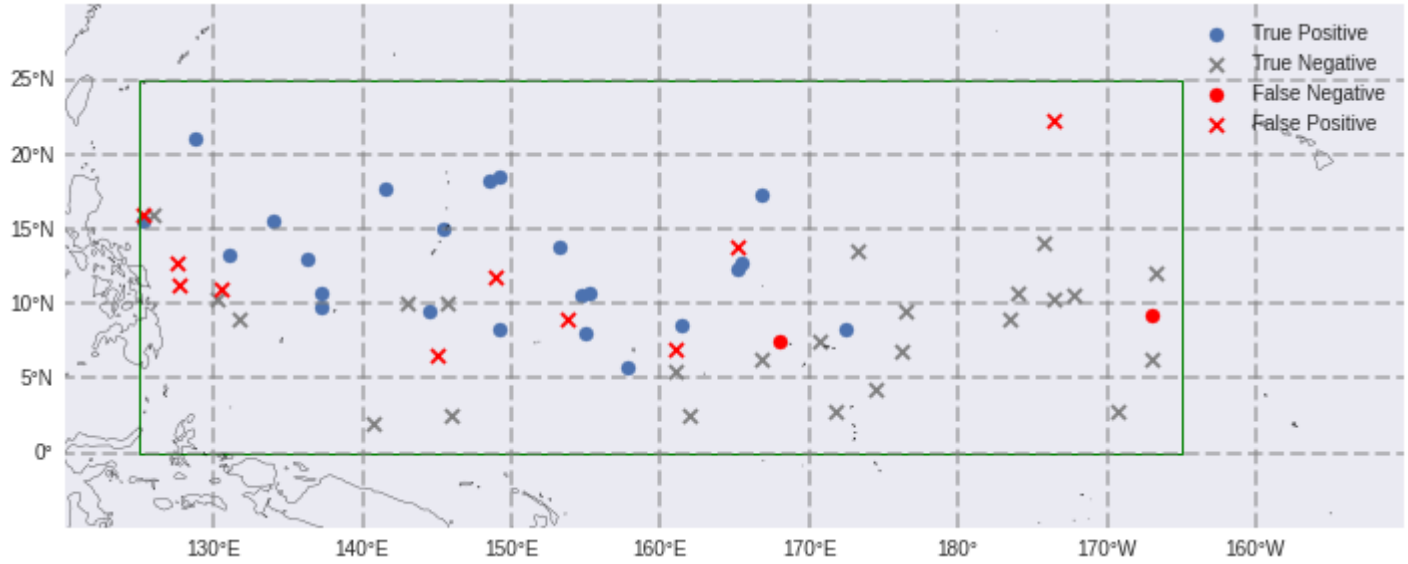


Figure 2.1. The studied regions of Western North Pacific (WNP) and Central Pacific (CP) from 0° to 25°N and 125°E to 165°W indicated by the green box. The symbols of blue dot, gray cross, red dot and red cross represent true positive, true negative, false negative and false positive cases in the test set results generated by the Artificial Neural Network (ANN) model. Their locations represent where the disturbances are, either 24 hours before genesis (developing) or at the moment of reaching 850 hPa vorticity maximum (non-developing).

doi:10.6342/NTU202303630

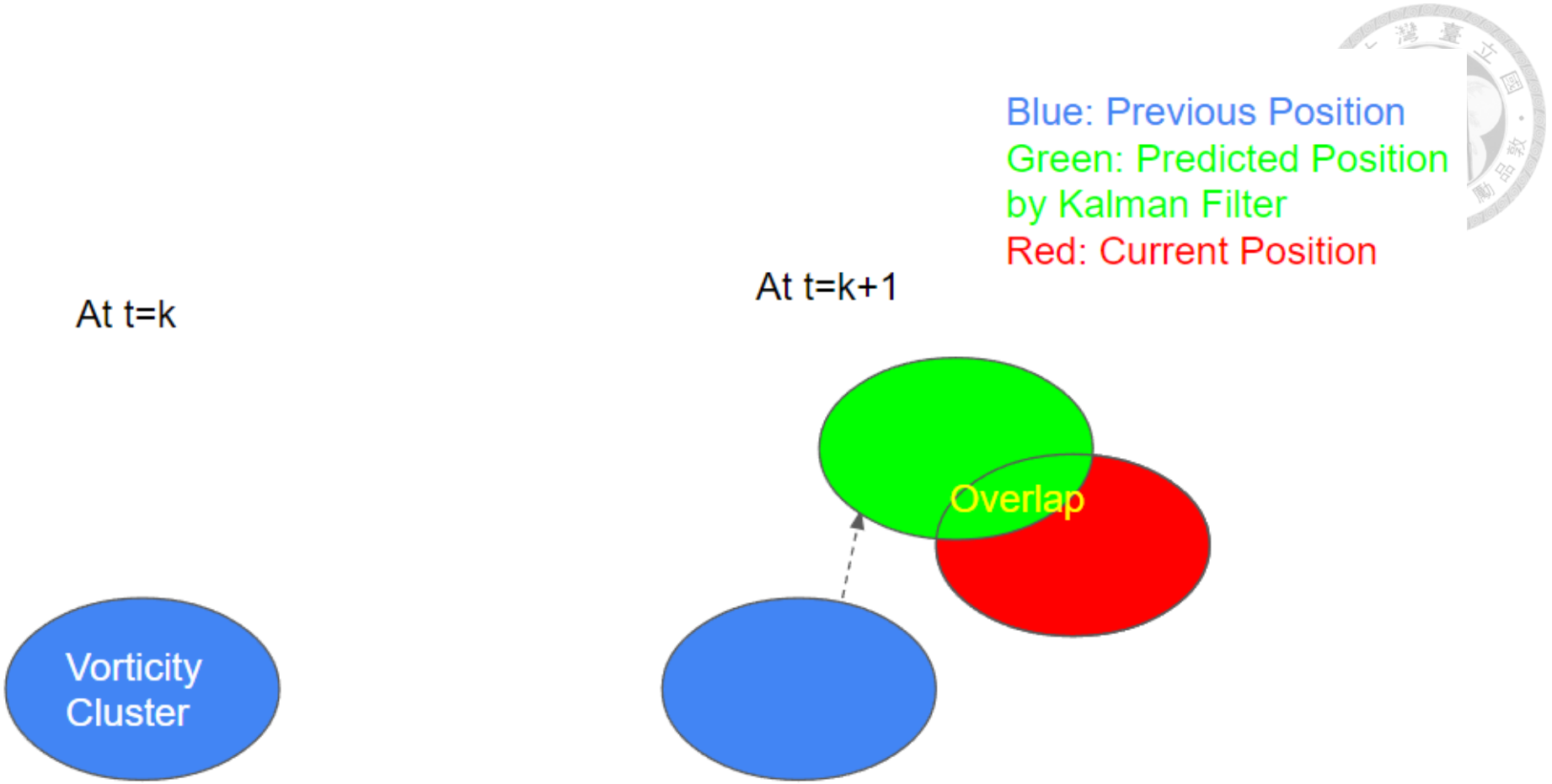


Figure 2.2. Schematic of the Kalman-to-Overlapping procedure. See the text on the figure for meanings of the colors.

doi:10.6342/NTU202303630

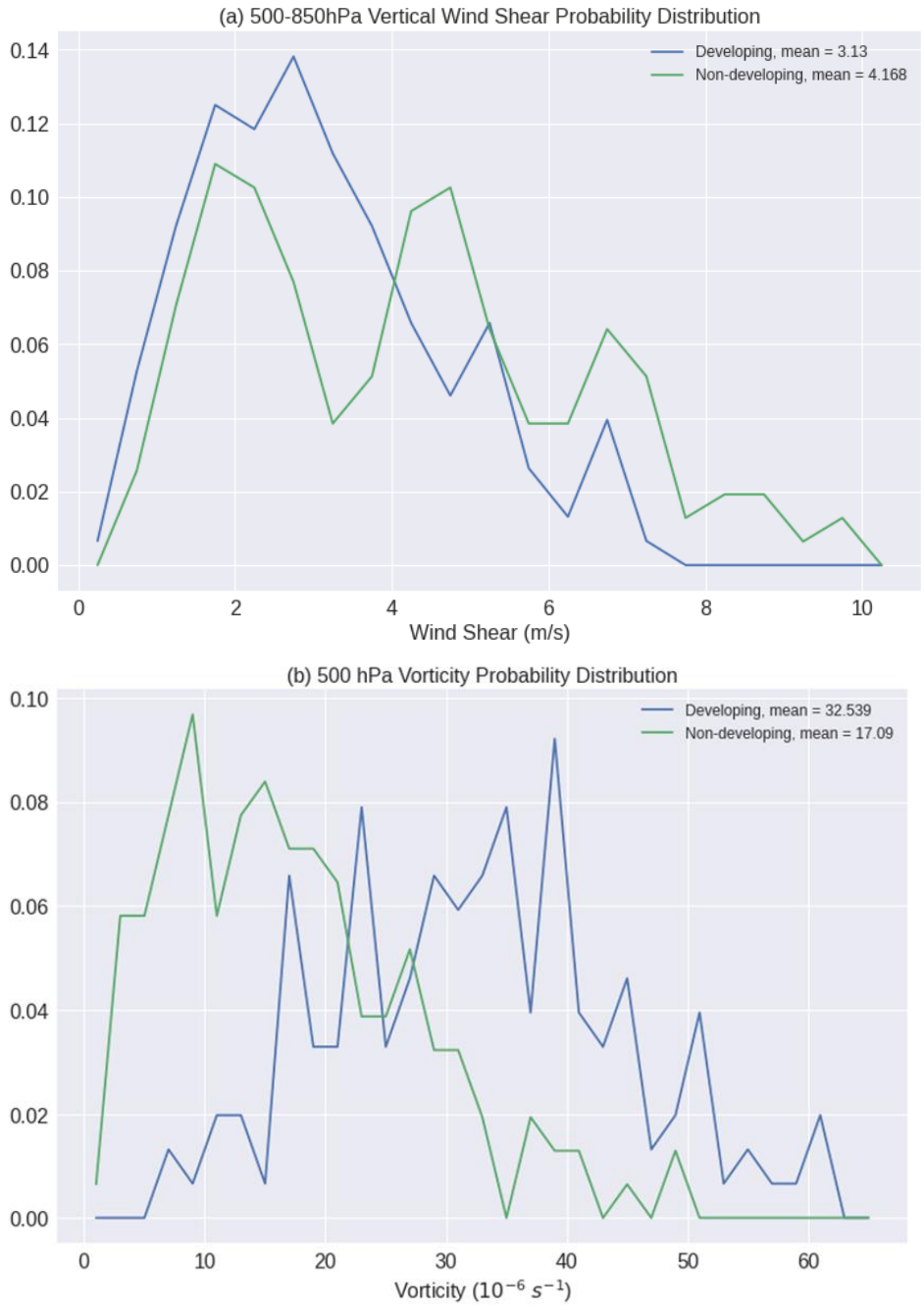


Figure 2.3. The probability distributions of developing (blue) and non-developing (green) sets for the variables: (a) 500-850 hPa vertical wind shear (unit: m s-1), and (b) 500 hPa vorticity (unit: 10^{-6} s^{-1}). The p-values of t-test and K-S test are (a) 3.15×10^{-6} and 1.10×10^{-5}; (b) 1.06×10^{-27} and 3.33×10^{-16} respectively.

doi:10.6342/NTU202303630

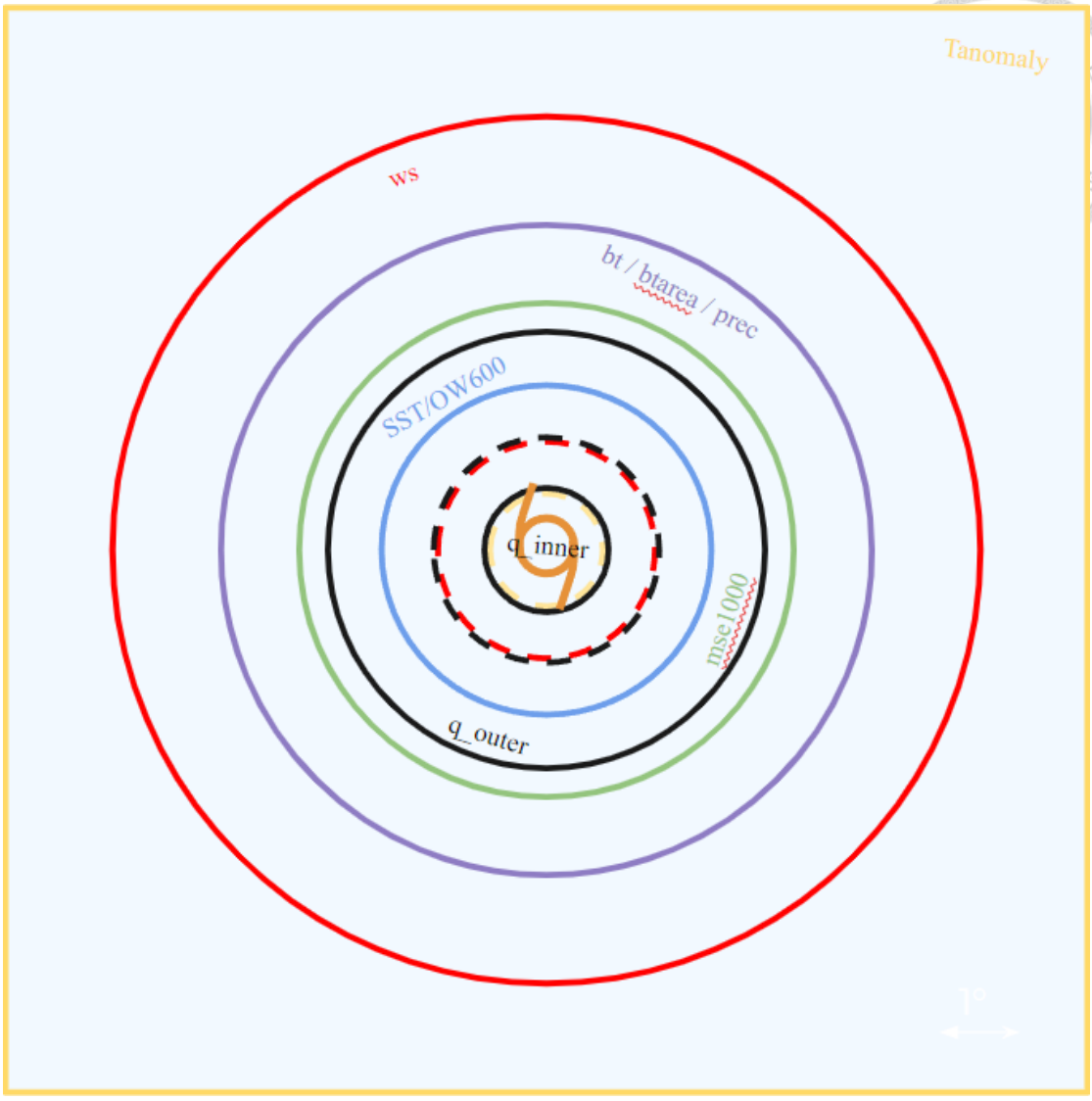


Figure 2.4. A schematic diagram showing the circular/annular regions where different variables are averaged/extracted. Different colors indicate different variables. Solid/dotted lines represent outer/inner boundaries.

doi:10.6342/NTU202303630

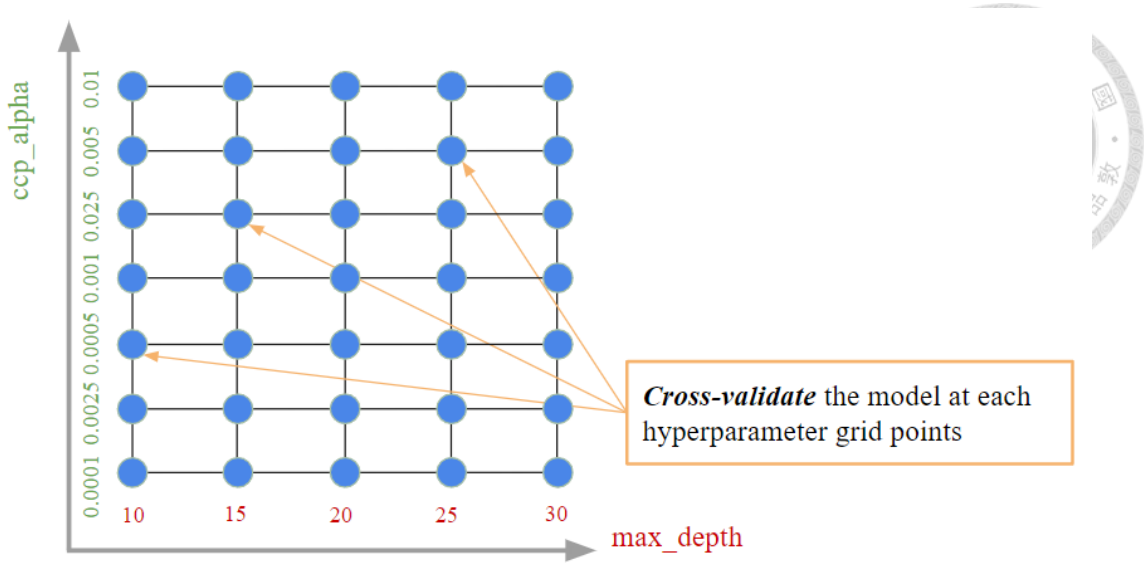


Figure 2.5. A schematic diagram showing how the method of grid search works, by cross-validating different models with every possible hyperparameter configurations arranged in an array.

doi:10.6342/NTU202303630

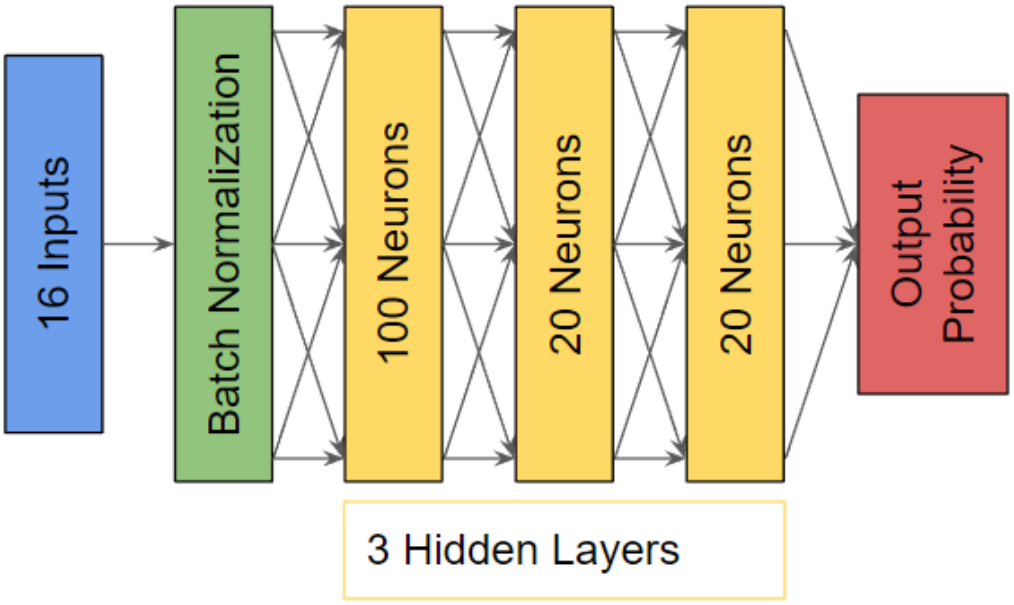


Figure 2.6. The architecture of the ANN used, with one layer of Batch Normalization and three fully-connected hidden layers, yielding the output genesis probability at the end.

doi:10.6342/NTU202303630

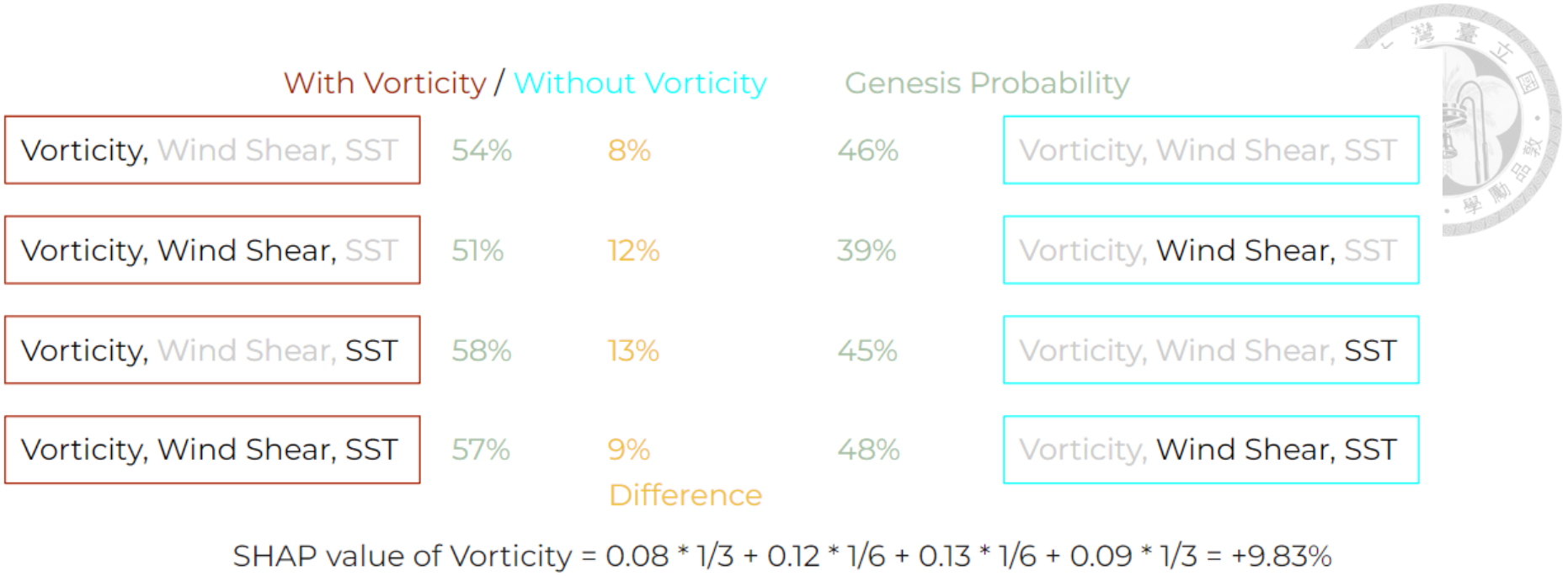


Figure 2.7. A made-up example of SHAP value computation. There are three features: vorticity, wind shear and SST, leading to $2^3 = 8$ possible configurations of elements. To obtain the SHAP value of vorticity, find all coalitions that have vorticity present (black text), paired against those with vorticity absent (gray text). Their difference (yellow text) in output predicted genesis probability (green text) are then calculated pair by pair. For example, the first pair, only vorticity versus none, the difference is 54% - 46% = 8%. A weighted average of these differences is then taken to obtain the SHAP value for vorticity.

doi:10.6342/NTU202303630

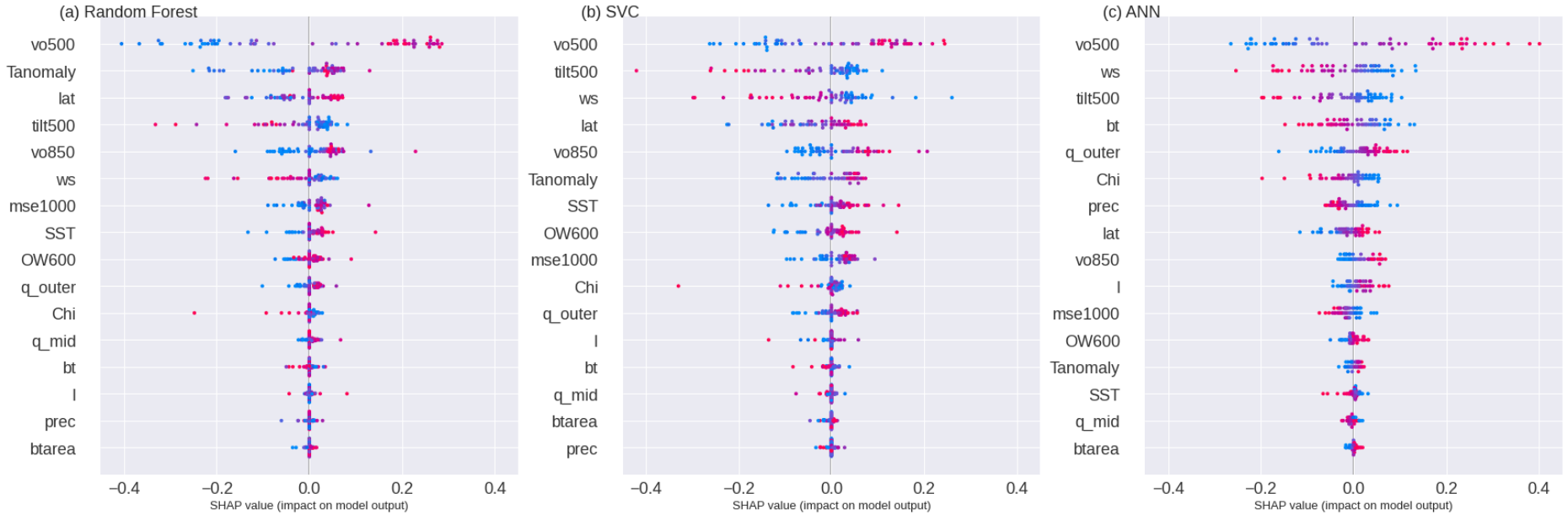


Figure 3.1. Beeswarm plots showing the SHAP values for each feature in each test sample as colored dots for the model of (a) Random Forest, (b) SVC, (c) ANN. X-axis is SHAP value and y-axis represents different variables. Cooler (Warmer) color represents a relatively lower (higher) value of the variable. The features are ranked in terms of mean absolute SHAP values (as an indicator of importance, see Figure 3.2.) from the top (more important) to the bottom (less important).

doi:10.6342/NTU202303630

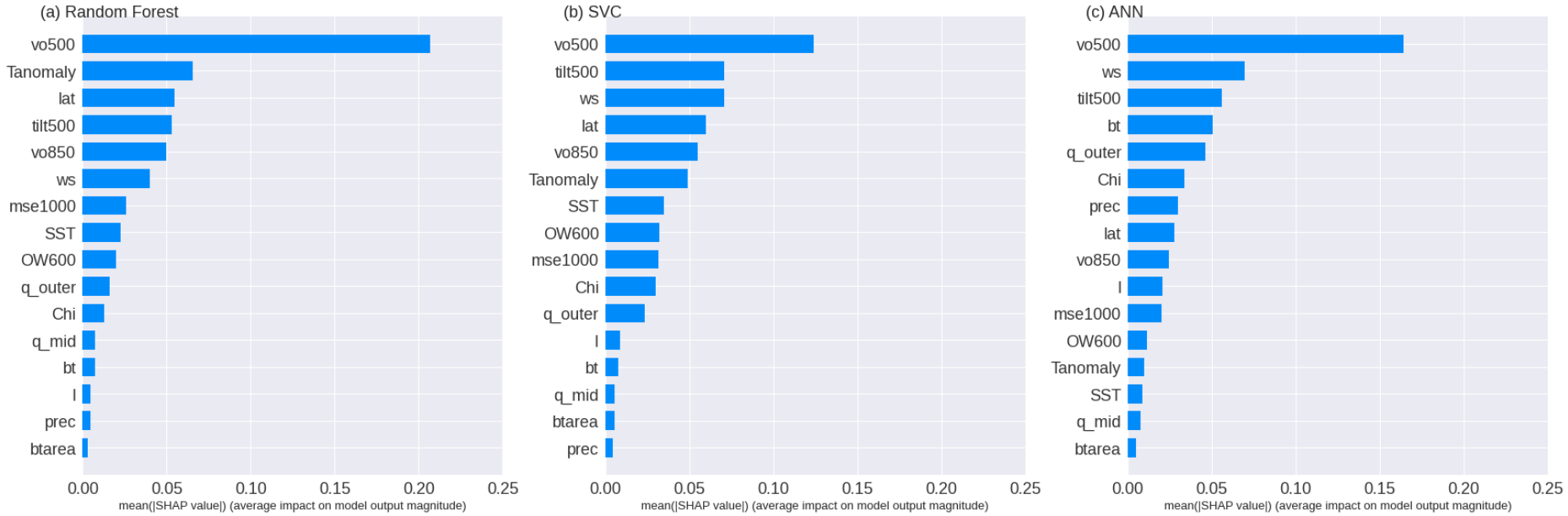


Figure 3.2. Barplots showing the mean absolute SHAP values of (a) Random Forest, (b) SVC, (c) ANN.

doi:10.6342/NTU202303630

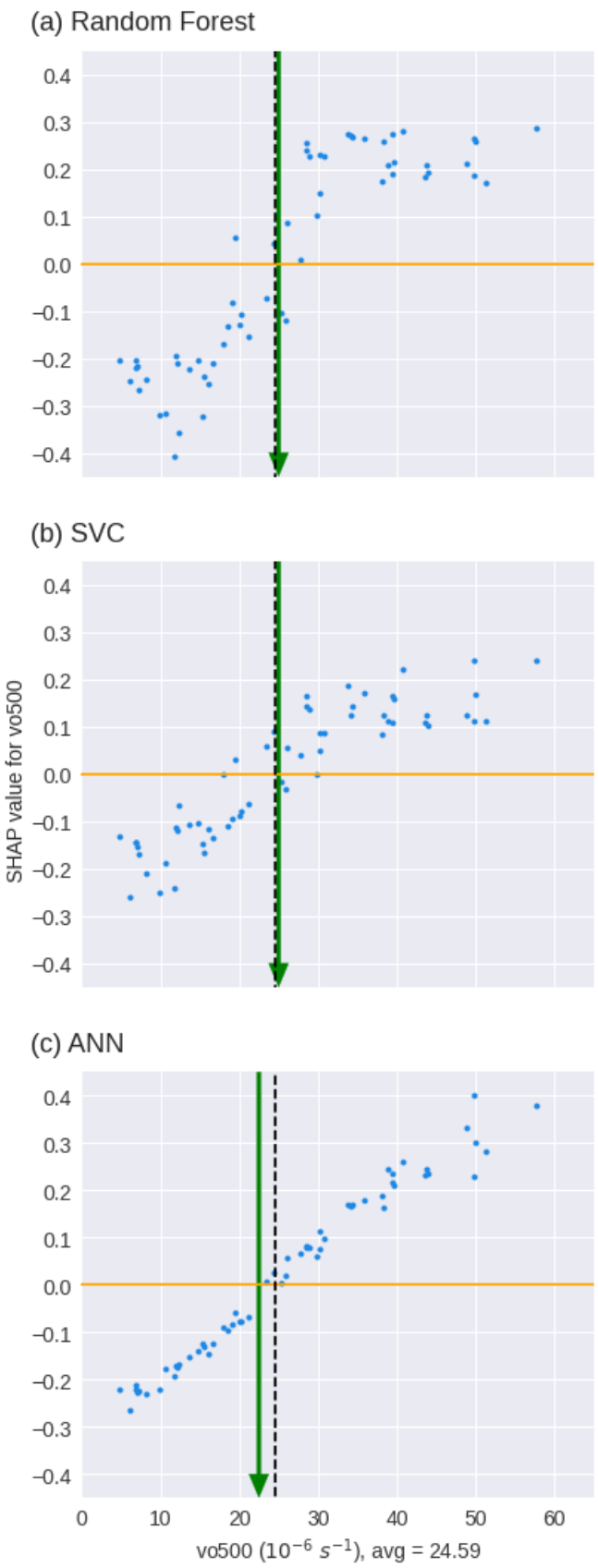


Figure 3.3. Dependence plots of 500 hPa vorticity for (a) Random Forest, (b) SVC, (c) ANN. X-axis is vorticity (unit: 10^{-6} s^{-1}) and y-axis is the SHAP value corresponding to 500 hPa vorticity. Each dot represents one test sample. The orange horizontal line shows the baseline where SHAP value is zero, and the green vertical arrow approximately identifies the location where the sign of SHAP changes. The black vertical line denotes the average of the variable value, shown at the bottom of the plot.

doi:10.6342/NTU202303630

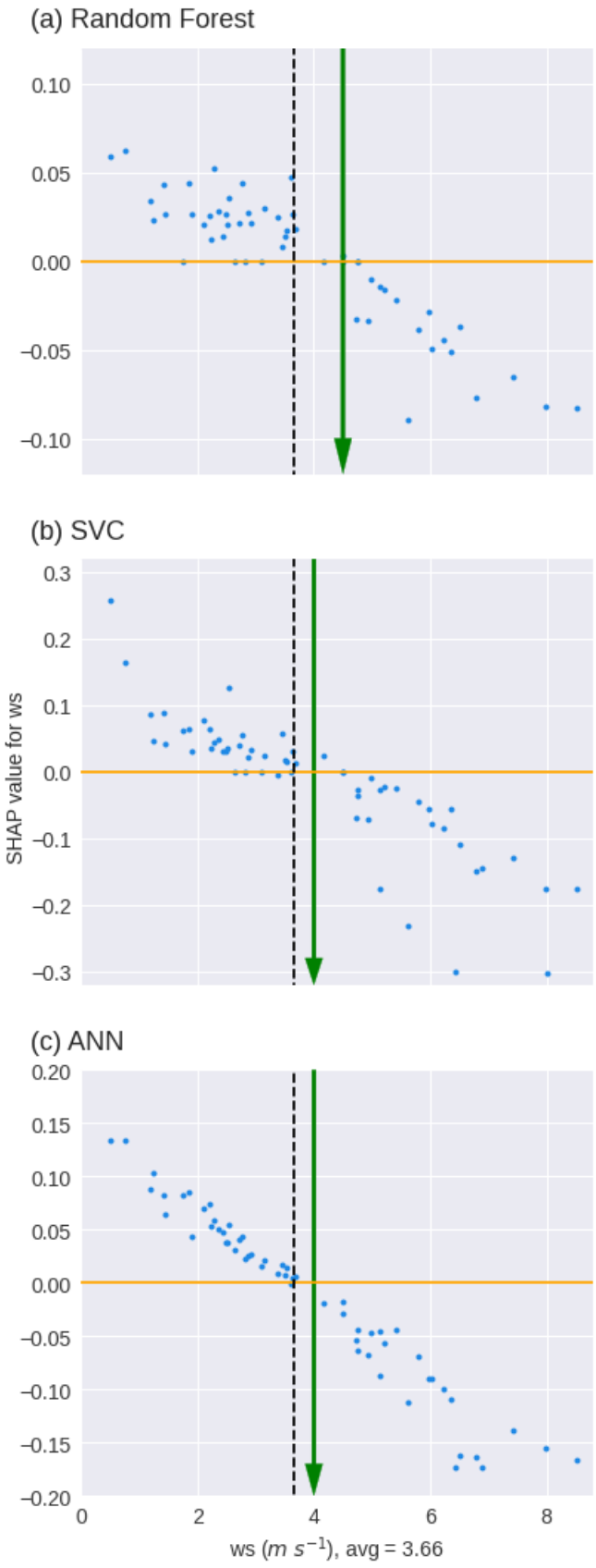


Figure 3.4. Same as Figure 3.3, but for wind shear. (unit: m s^{-1}).

doi:10.6342/NTU202303630

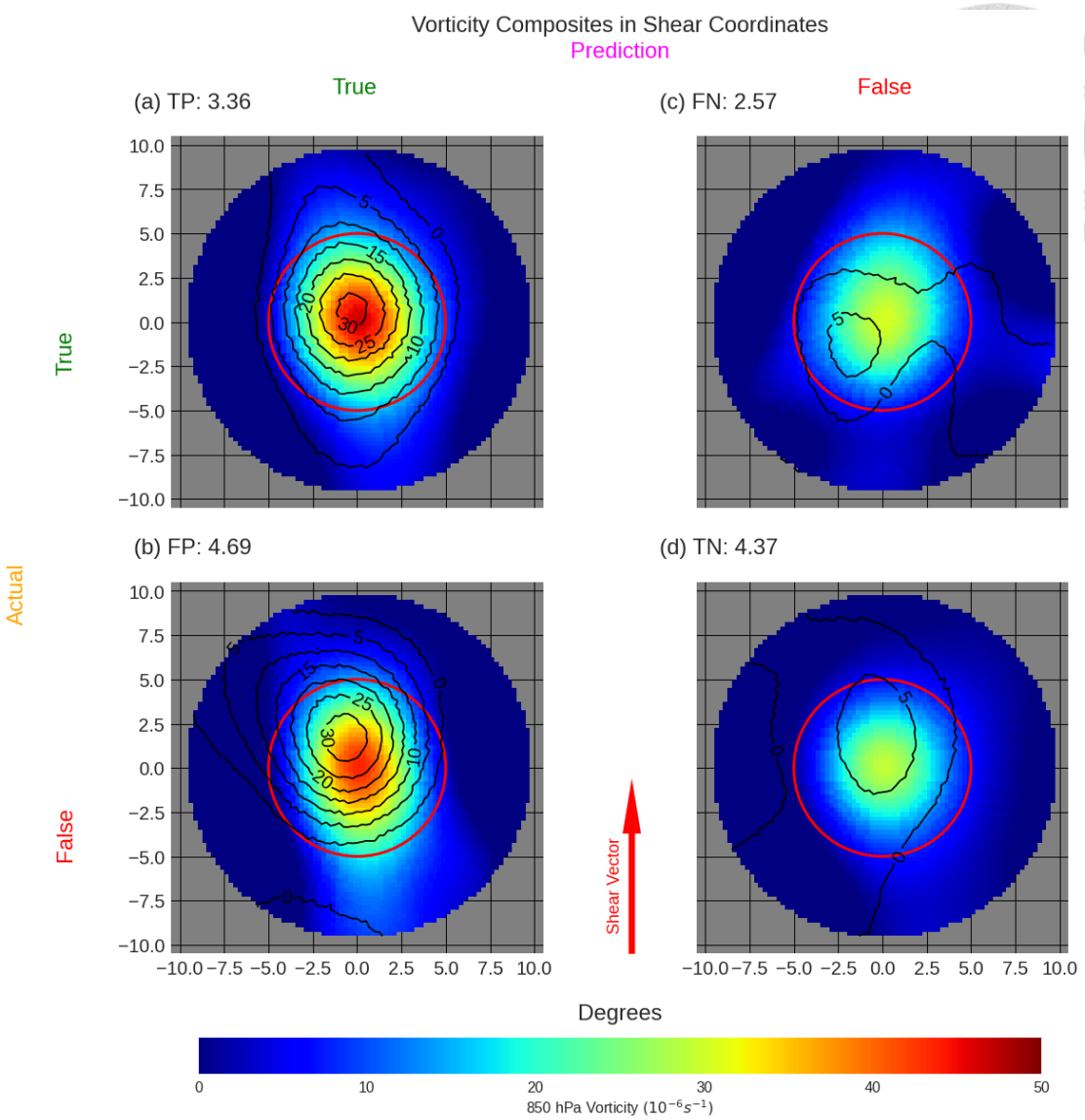


Figure 3.5. Shear-coordinate composite of vorticity at 500 hPa (contour, unit: 10^{-6} s^{-1}) and 850 hPa (shading, unit: 10^{-6} s^{-1}) for (a) True Positive, (b) False Positive, (c) False Negative, (d) True Negative cases. The numbers in brackets are the average wind shear magnitude (unit: m s-1) for each of the four classes respectively. The red arrow stands for the shear vector.

doi:10.6342/NTU202303630

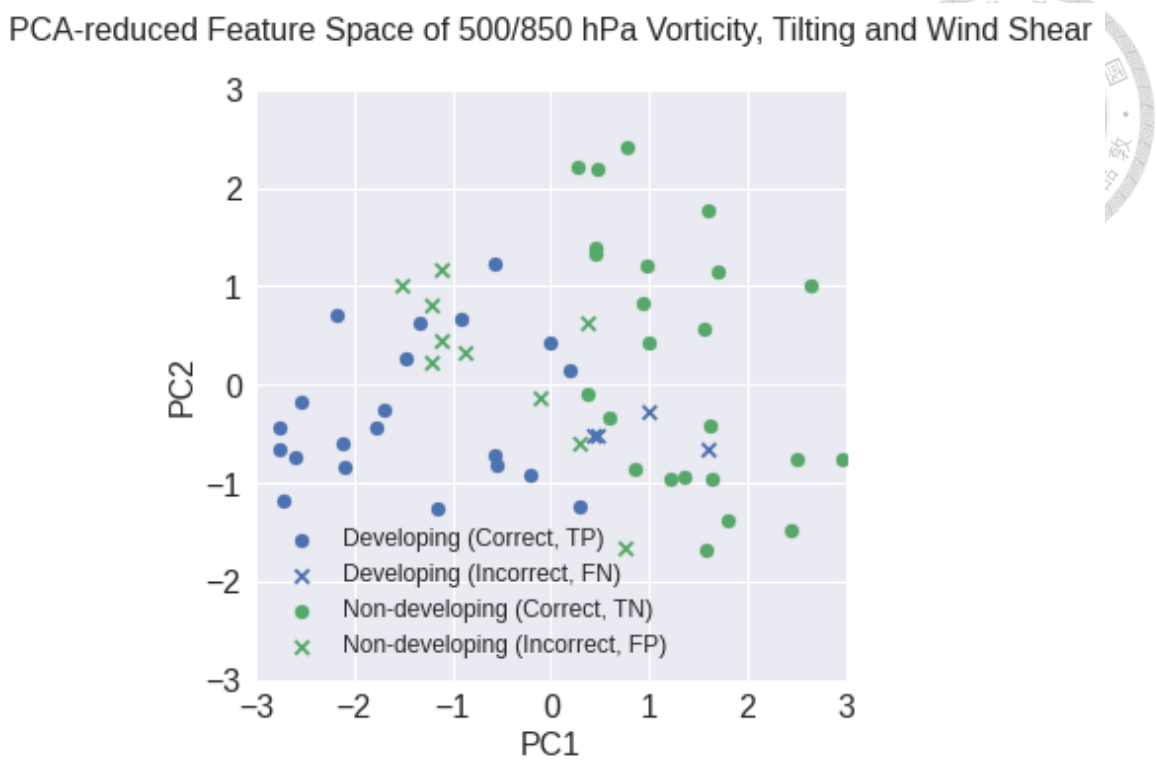


Figure 3.6. Principal Component Analysis applied to the (standardized) four-dimensional feature space created by the variables *vo500*, *vo850*, *tilt500*, and *ws*. The two leading PCA modes are retained and the distribution of data over the reduced two-dimensional PCA space is displayed. Blue: developing; Green: non-developing; Circles: correct for all the three models; Crosses: incorrect for at least one model. PC1/PC2 explains ~55%/25% of the total variance respectively.

doi:10.6342/NTU202303630

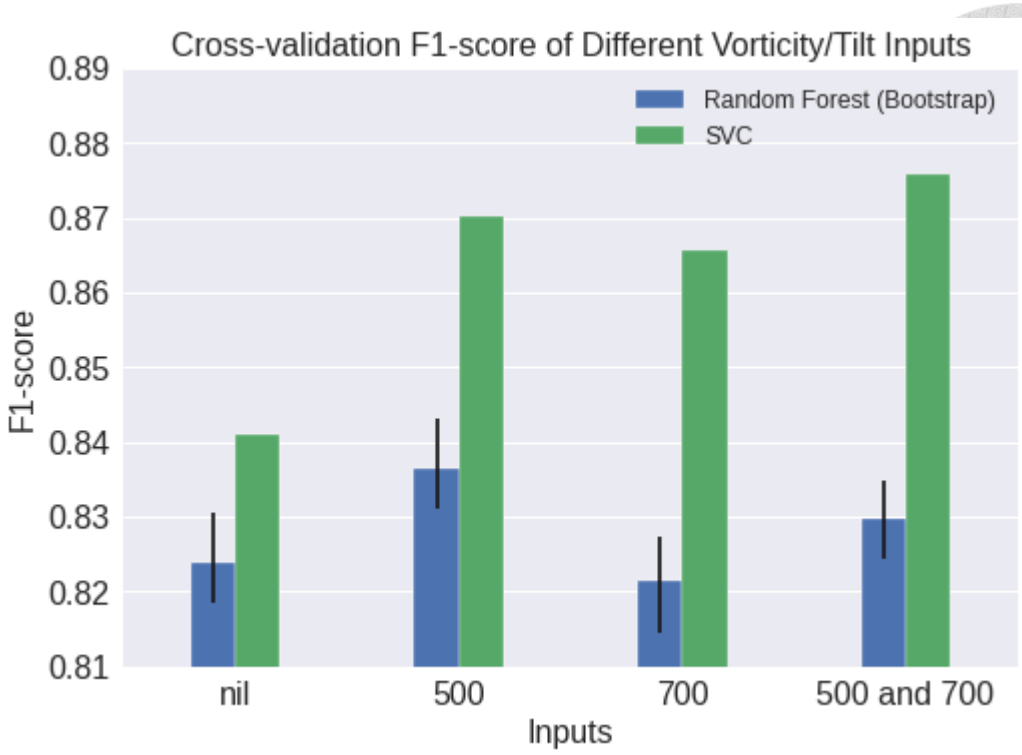


Figure 4.1. (Mean) Cross-validation f1-scores of 100 Random Forests (blue) and SVC (green) in the four scenarios of the sensitivity experiment. (See text for the meanings of x-axis labels) The black vertical bar denotes the range between the first and third quartiles of f1-scores within the 100 random forests. The difference in f1-scores of the random forests in any pair of experiments passes t-test at 99% significant level (except the second-fourth pair).

doi:10.6342/NTU202303630

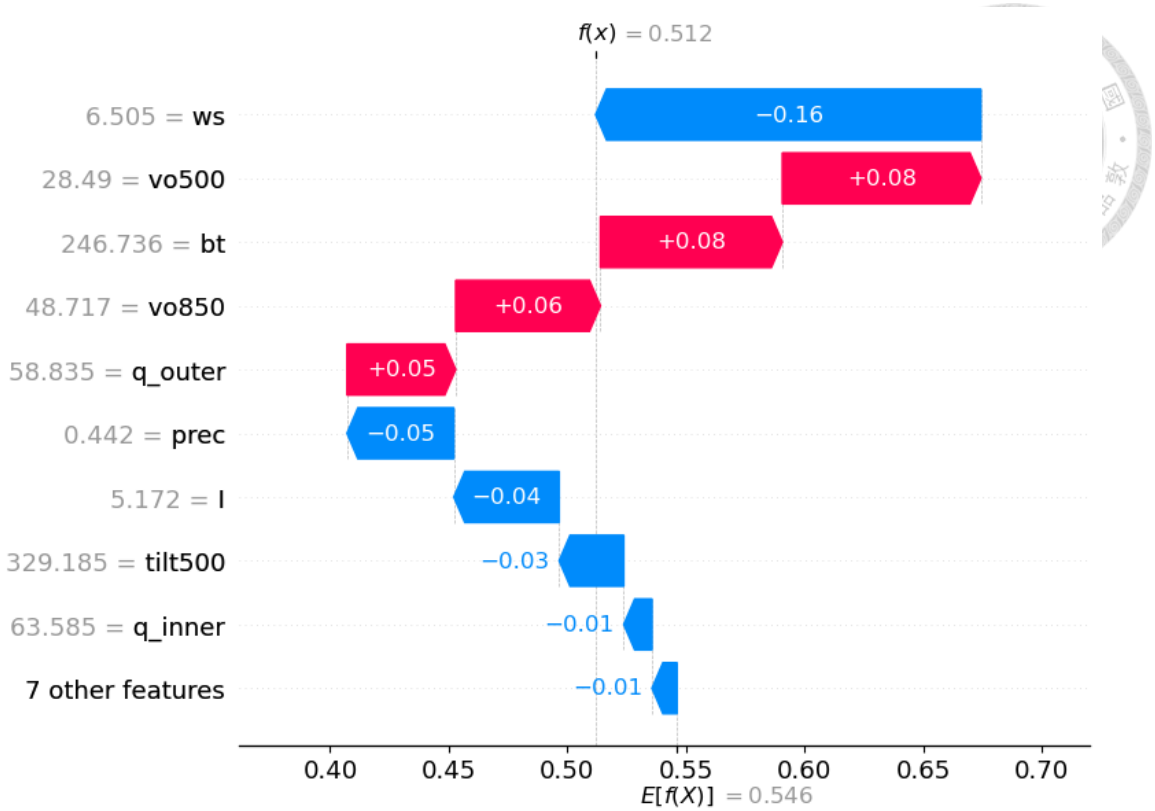


Figure 4.2. Waterfall plot for the development prediction of Typhoon Halong (WP112014) by the ANN model 24 hours before (2014-07-27 06:00 UTC) its genesis. Arrow in each row shows how much the variable increases or decreases the output predicted genesis probability. The E[f(X)] is the base expected probability and f(x) is the output probability. E[f(X)] and all the arrows add up to f(x).

doi:10.6342/NTU202303630

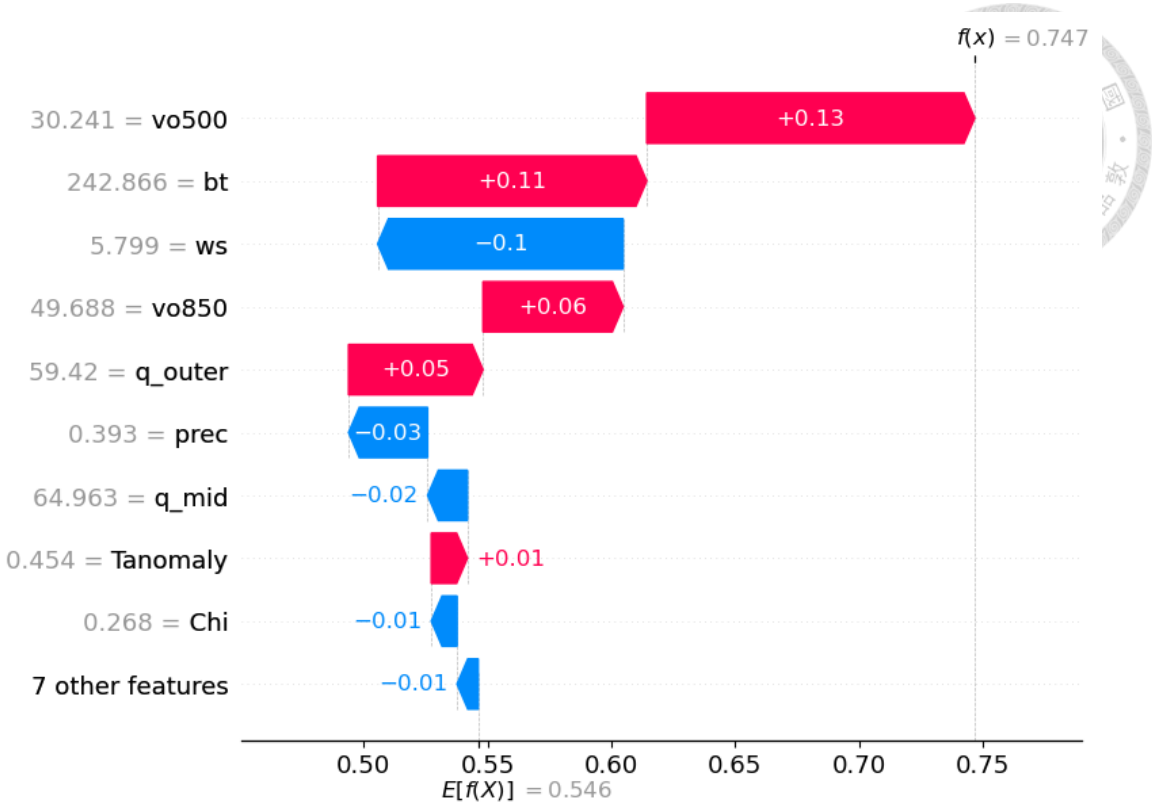


Figure 4.3. Same as Figure 4.2, but re-drawn for the same set of variables extracted 6 hours later, or 18 hours before the genesis of Halong (2014-07-27 12:00 UTC).

doi:10.6342/NTU202303630

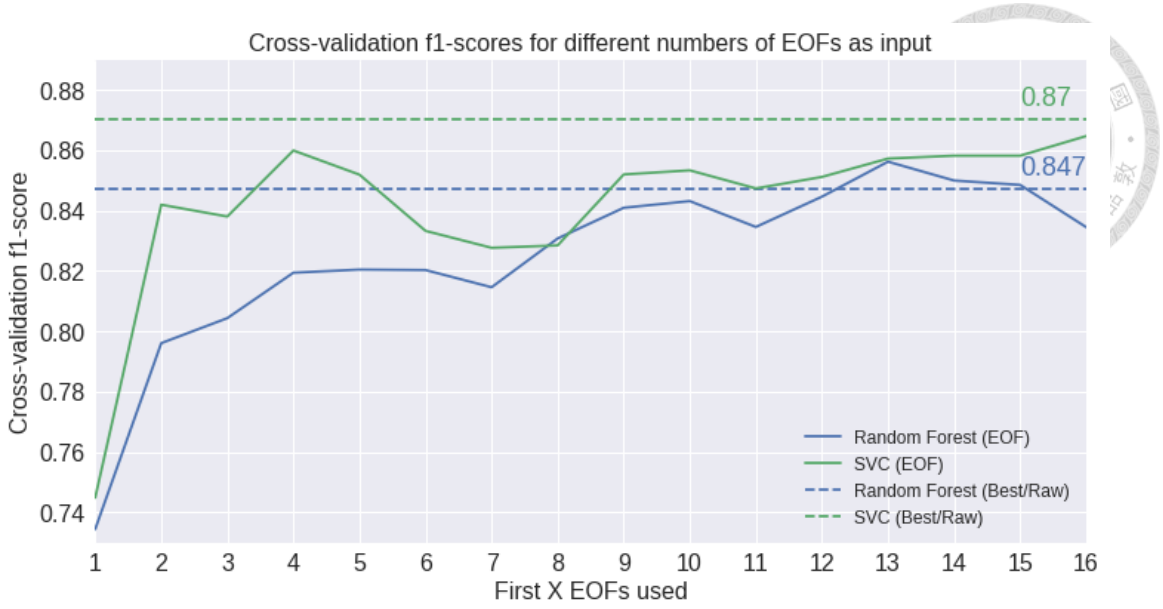


Figure 4.4. The cross-validation f1-scores with EOF preprocessing for Random Forest (blue) and SVC (green). The solid curves represent the cross-validation performance against how many the largest EOFs are used. The dotted horizontal lines are the reference cross-validation f1-scores when the raw variables are used in the original machine learning models.

doi:10.6342/NTU202303630

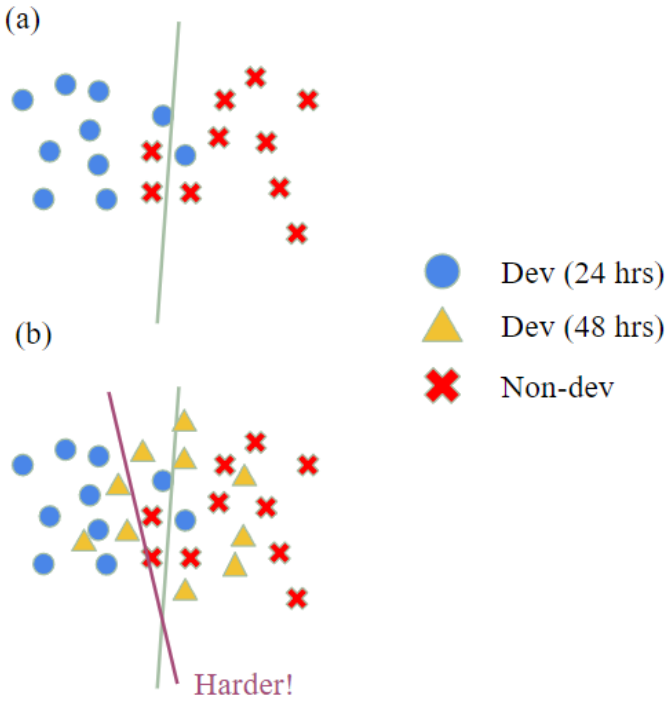


Figure 4.5. A schematic showing hypothetical distributions of developing (blue circles, 24 hours before genesis) and non-developing (red crosses) cases reduced to some phase space, (a) before (b) after the addition of the negative cases of developing disturbances 48 hours before genesis (yellow triangles). The light green and magenta lines represent the old and new decision boundaries, respectively.

doi:10.6342/NTU202303630