JOURNAL OF
HYDRO-METEOROLOGY

Research Article

Development of an ensemble dynamic-probabilistic prediction model for tropical storm genesis in the Vietnam East Sea using the Logistic Regression approach

Dao Nguyen-Quynh Hoa[1]*, Tran Tan Tien[1]

[1] VNU University of Science, Vietnam National University; hoadao@vnu.edu.vn; tientt@vnu.edu.vn

***Corresponding author: hoadao@vnu.edu.vn; Tel.: +84–8857758771**

Received: 05 September 2023; Accepted: 23 October 2023; Published: 25 December 2023

Abstract: In this study, a logistic regression model is developed to forecast tropical storm (TS) genesis in the Vietnam East Sea from 2012 to 2019. The model incorporates seven potential predictors including dynamic and thermodynamic parameters at formation time retrieved from the WRF-LETKF outputs. After rigorous testing, six predictors are selected, excluding minimum sea-level pressure. In a broader context, the logistic regression model performs promisingly, generating forecast probabilities that enhance the accuracy of TS genesis predictions, particularly in early forecast cycles. The model's regression coefficients and forecast outcomes align well with test dataset results, affirming its stability and validity. As a result, the forecast probability from this model can be effectively employed as a probabilistic forecast value for predicting TS genesis status.

Keywords: Tropical cyclogenesis; WRF-LETKF; Ensemble prediction system; Logistic regression.

1. Introduction

Tropical storm (TS) genesis forecasting is recognized as one of the most challenging aspects of numerical weather prediction (NWP). The ability to accurately predict the formation of TSs is crucial for effectively managing and mitigating associated risks. With the significant advancements in computational capabilities, NWP models have started to assume a prominent role in forecasting TS genesis and tropical cyclone (TC) as a whole. Consequently, numerous studies have been undertaken to evaluate the reliability and accuracy of predictions regarding the formation and development of TCs to TS intensity as indicated by these models [1–3].

Beyond the utilization of a single-model deterministic forecasts, the adoption of ensemble forecasts for TS genesis prediction has gained popularity, showing superior predictive capabilities. Ensemble prediction systems are designed to address uncertainteis in initial condiitons and imperfections in model formulation, with the goal of providing a range of potential future atmospheric scenarios [4]. Despite the substantial volume of data required for processing, ensemble forecasting has demonstrated its effectiveness in enhancing the accuracy of tropical cyclones [5]. In a study assessing TS genesis from tropical cloud clusters using two global ensemble prediction systems, ECMWF-EPS and UKMO-EPS, conducted over the years 2018 to 2020 in the Northwestern Pacific, Northeastern Pacific, and Northern Atlantic regions. The author [6] reported relatively good forecasting skills. They found that the quality of probabilistic forecasts could be further enhanced by combining predictions from all multi-model ensemble members. Additionally, the authors observed that in cases

where on member did not provide an accurate prediction, favorable conditions for TC genesis were often present. This underscors the utility and value of ensemble forecasting systems in improving the prediction of TS genesis. The Local Ensemble Transform Kalman Filter (LETKF) scheme, pioneered at the University of Maryland [7, 8], has found widespread application in various numerical models, with notable use in the Weather Research and Forecasting (WRF) model [9–11]. Research by [9] highlighed the LETKF's exceptional utility in managing highly diverse data, such as satellite observations. The incorporation of multi-physics ensemble prediction of this scheme has demonstrated ist effectiveness in predicting the genesis of TC Wutip (2013). Building upon their case study, this study delves further into examining the efficacy of the WRF-LETKF when assimialting augmented observations in forecasting the likelihood of 45 TCs that occurred between 2012 and 2019 over the Vietnam East Sea.

Along with dynamical prediction model, numerous studies have delved into the application of statistical models for forecasting the development of TSs [12–15]. A common thread across these diverse researches is the incorporation of seasonal or climatological factors of various scales as predictors in their statistical models. A recent approach is applying these statistical model to the products of pre-existing NWP systems, so called the dynamical-probabilistic forecast models [16]. However, relatively few studies have concentrated on short-term events with lead times of up to 5 days. One of the most frequently employed techniques for probability prediction is the logistic regression model [17]. This model is capable of forecasting and discerning the likelihood of an event occurring, particularly the genesis of TCs [18].

Building upon the foundation of existing scientific knowledge, this study offers a dynamical-probabilistic forecast model to predict TS genesis over the Vietnam East Sea, expanding to ensemble forecast. The construction involves results from the WRF-LETKF forecasts and logistic regression model, with a specific focus on the likelihood of TS genesis from predefined events. This investigation sheds light on vital insights into the predicted TS genesis using products from ensemble prediction system. The subsequent sections of this paper are organized as follows: Section 2 outlines the experimental design. Section 3 provides a brief discussion of the results concerning the direct outputs to predict TC genesis various forecast cycles. Section 4 selects the associated predictors and section 5 constructs the probabilistic prediction model. Finally, Section 6 offers a summary and engages in further discussion.

2. Experiment settings

2.1. Ensemble prediction system

The Local Ensemble Transform Kalman Filter (LETKF) algorithm are applied to the non-hydrostatic version of the WRF model version 3.9.1 to create the ensemble-based data assimilation system (hereafter, WRF-LETKF) with variational data assimilation scheme 3DVAR at cold-start. Additionally, the algorithm's construction and its application in the case study of Wutip (2013) were thoroughly elucidated in the study of [19].

The model employs an ensemble size of 21 members, generating perturbations based on the atmospheric state from the previous cycle. These perturbations are integrated into the global deterministic analysis using, maintain uniform scaling with an inflation factor of 1,1 for assimilated variables to enhance the influence of ensemble noise. The multi-physic approach optimizes spread without requiring a larger number of members and combines various parameterization schemes. These schemes include 2 convective schemes, 3 boundary-layer schemes, 3 microphysics schemes, and 2 shortwave-longwave radiation schemes, resulting in a total of 36 potential combinations. From these, 21 combinations were selected for operational

purposes (refer to Table 1 [19]), considering model stability and minimizing internal conflicts during implementation.

The WRF model was configured with a domain that is large enough to cover the entire Vietnam East Sea and the surrounding waters of the Northwestern Pacific, an area with coordinates [95°E - 145°E; 0° - 30°N] (Figure 1). The domain has a spatial resolution of 27 km and includes 31 vertical levels and the forecast lead time is 5 days (120-hr forecasts) within a hourly interval.

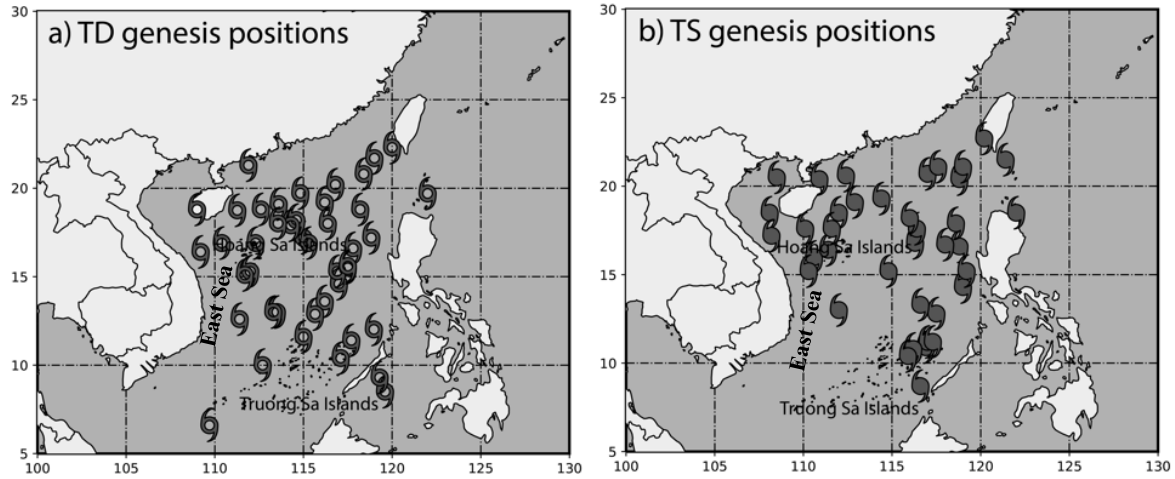

Figure 1. Illustration of the positions where 45 TCs formed (a) and developed into tropical storms (b) in the Vietnam East Sea during the period from 2012 to 2019 retrieved from IBTraCS best-track.

2.2. Data collection and TC tracking

The study relies on the International Best Track Archive for Climate Stewardship (IBTraCS) [20] to serve as the benchmark for TD formation ensemble prediction and for statistically verifying subsequent TS development. The investigation focuses exclusively on TCs in their early stages within the Vietnam East Sea. If TC maximum 10-m sustained wind speed reaches 20 kt (TD intensity), the location and timing of the cyclone in best-track is recorded as TD genesis (Figure 1a). Similarly, when the TC's intensity reaches or exceeds 34 kt (TS), it is considered as a development case, indicating that it has transitioned into a tropical storm (Figure 1b). The initial and boundary conditions for the numerical experiments were derived from the NCEP GFS analysis provided every 3-hr at 0.5° × 0.5° horizontal resolutions.

In this study, the forecasted timing and location of each TCs from TD formation to TS stage is important, as we examine the state of the atmosphere at the genesis of TD and its subsequent potential to reach TS intensity. We employ a straightforward yet essential tracking algorithm, with a specific emphasis on their early development stages over the ocean. The tracking algorithm follows the work made by [19]. The criteria used for selecting cyclones that develop into TSs involve local minimum sea-level pressure (P_{min}) having adjacent local maximum low-level vorticity (ζ_{low}) and minimum geopotential height at 700 hPa while scanning every grid point for each lead time of every forecast. All the extrema are required to exhibit at least 2 closed contours of 2 hPa; 10^{-5} s^{-1}; and 4 dams, respectively, to avoid noise. When TC's maximum 10-meter wind speed (V_{max}) nearest to the P_{min} location exceeds 20 kt, it is recorded as a TD formation. The same implementation has been used to select and verify TC development to TS stage, when V_{max} surpasses a threshold of 34 kt for the first time. In some cases, the phases of TC formation (reaching TD) and development (reaching TS) may overlap. Our definitions for successful TD and TS formation forecasts also consider the location predicted by the model in comparison to real storm's best-track. We do so by selecting TC centers that fall inside the Vietnam East Sea region, within 5° radius proximity to the best-track recorded within a 120-hr lead time.

Over the 2012-2019 period, we have collected a dataset comprising a total of 45 recorded TDs, while 35 among them continue to develop to TS stage and the rest are dissipated over the Vietnam East Sea.

2.3. Logistic regression model predicting the tropical storm genesis

2.3.1. Predictors

The purpose is to forecast the probability of TC development reaching TS intensity based on a pre-existing TC formation predicted in ensemble forecasts. The study employs a logistic regression model with the dependent variable as a quantitative variable predicting a probability with values within the range [0; 1]. The initial objective of this study is to establish the predictors for modeling the occurrence probability of TS genesis in the Vietnam East Sea region (refer to Table 1). As outlined in Table 1, these predictors consist of both dynamic parameters (minimum sea-level pressure, low-level relative vorticity, mid-level vertical velocity, and vertical wind shear) and thermodynamic parameters (moist static energy, surface latent heat flux, and low-level horizontal moisture convergence). Research conducted by several authors [21–23] indicates that these critical meteorological parameters can effectively differentiate between developing and non-developing disturbances, particularly in the context of TS formation. The local environment refers to atmospheric state centered on the location of a TC within 5° radius and standardized due to differences in dimension and scale.

Table 1. Descriptions of logistic regression model variables.

Categorization	Variable	Descriptions
Dynamic	P_{min}	Minimum sea-level pressure
	ζ_{low}	Average mid-to-low-level vertical vorticity $\zeta_{low} = \int_{850}^{500} \zeta \left(\frac{dp}{g}\right)$
	ω_{mid}	Average vertical velocity in 700 – 500hPa
	V_{sh}	Vertical shear between 200 and 850 hPa $V_{sh} = \sqrt{(u_{200} - u_{850})^2 + (v_{200} - v_{850})^2}$
Thermodynamic	MSE	Column-integrated moist static energy normalized by C_p $MSE = \frac{\int_{p_s}^{0} C_p T + gz + L_v q_v \left(\frac{dp}{g}\right)}{C_{pd}}$
	SLHF	Surface latent heat flux
	HMC_{low}	Low-level horizontal moisture convergence $HMC_{low} = -\int_{p_s}^{700hPa} \frac{\Delta \overline{uq_v}}{\Delta x} + \frac{\Delta \overline{vq_v}}{\Delta y} \left(\frac{dp}{g}\right)$

Student t-test is conducted on the probability distribution function of each predictor, based on 2 datasets categorized as having TC development reaching TS intensity (DEV) and not having such development (NON-DEV). Null hypothesis H_0: $\bar{x}_{i,1} = \bar{x}_{i,0}$; Alternative hypothesis H_1: $\bar{x}_{i,1} \neq \bar{x}_{i,0}$.

where $\bar{x}_{i,1}$ is the mean value of the i-th predictor in DEV group, and $\bar{x}_{i,0}$ is the corresponding mean value in the NON-DEV group. If the null hypothesis is rejected, and the alternative hypothesis is accepted with a 95% confidence level, meaning that there is a significant difference in the parameter's value between the two groups, then the parameter is selected as a predictor to be included in the logistic regression model.

2.3.2. Logistic regression model configurations

The logistic regression model is based on the concept of linear regression for classification problems. Starting with the output of a linear regression function, the logistic regression model uses the sigma function to find the probability distribution of data within

the range [0; 1]. Assume that we have a regression function with a set of n independent variables $x = (1, x_1, x_2, \dots, x_n)$:

$$\hat{y} = g(x) = w_0 + w_1 x_1 + \cdots + w_n x_n = w^T x \quad (1)$$

Here, represents the regression coefficients. To transform this equation using the sigma function to predict probabilities and introduce non-linearity into the regression model, we have the following expression:

$$p(y = 1|x; w) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}} \quad (2)$$

$p(y = 1|x; w)$ is the conditional probability of the event y = 1 occurring based on the independent variables x and the regression coefficients w.

To assess the regression model's validity through cross-validation analysis, we conduct cross-validation of the regression model on a dataset consisting of 45 cases of TC genesis during the period 2012-2019, with 35 cases of TCs developing to TS and 10 of no development. This is done by dividing the dataset into 5 subsets (5-fold cross-validation). To ensure that the train and test datasets have a sufficient number of cases for both TC development and non-development, the number of TCs in each subset is established as follows:

- Train data: 28 developing TCs and 8 non-developing TCs.
- Test data: 7 developing TCs and 2 non-developing TCs.

In addition, to test the utility of the predictive equation and derive the final forecasting equations, the study employs the Wald test to assess the influence of the predictors on the regression model. The study uses $\alpha = 5\%$ as statistical significance threshold for the regression coefficients. The predicted results from ensemble dynamic-statistical model are determined from the average probability values of the ensemble members:

Here, M = 21 is the total number of ensemble members, and $p_{i,j}$ represents the probability of predicting TC development from the j-th ensemble member for the i-th forecast. $p_{i,j}$ is equal to 0 by default if the ensemble member does not predict TD formation in the Vietnam East Sea within the 120-hr forecast period based on the dynamical model's output.

2.3.3. Verification metrics

It is important to note that our study primarily centers around determining whether or not TS genesis has occurred within the Vietnam East Sea within a 120-hr lead time. In this context, the study does not specifically focus on the precise timing and location of the occurrence in comparison to observations. Consequently, the selection of verification metrics in this study is tailored to the specific goal of assessing the accuracy of forecasted probabilities (through Brier score) and the categorization of events versus non-events (through AUC-ROC).

- Brier score: $BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2$ calculates the forecast accuracy by comparing the forecasted probability with the actual observation, indicating if the event occurred or not [24]. N is the total number of observed events. A lower BS indicates a better forecast alignment with reality (BS range: 0-1).

- AUC-ROC assesses the binary classification model performance, measuring a model's ability to distinguish positive and negative classes [25, 26]. A high AUC-ROC value (close to 1) suggests accuracy, close to 0 indicates inverse predictions, and 0.5 signifies poor classification.

3. Genesis forecasting in the WRF-LETKF

Figure 2 represents the verification scores of TD genesis and TS genesis over the Vietnam East Sea for 8 years during 2012-2019 from each ensemble member from WRF-LETKF. For TD genesis forecasts, the WRF-LETKF ensemble system demonstrates reasonably good forecasting skill up to a 5-day lead time, with the lowest BS occurring at

approximately 3.5 days before the formation (equivalent to 84-hr cycle). With the 48-hr forecast cycle, the ensemble system shows a stable increase in the probability of correctly predicted cases.

However, when it comes to forecasting the development of TCs reaching TS stage (TS genesis), the Brier score indicates a significant decrease compared to the TD genesis forecast, with most values exceeding 0.1 and a clear decreasing trend in the near forecast cycles. When combined with the AUC-ROC skill score, it is observed that the WRF-LETKF performs well in classifying the development of TCs to TS intensity in the forecast cycles starting from 96-hr, 84-hr, and 48-hr onward (with AUC-ROC values exceeding 0.6). In the forecast cycle of 120-hr and 108-hr, the model fails to classify the possibility of TD development into storms accurately.

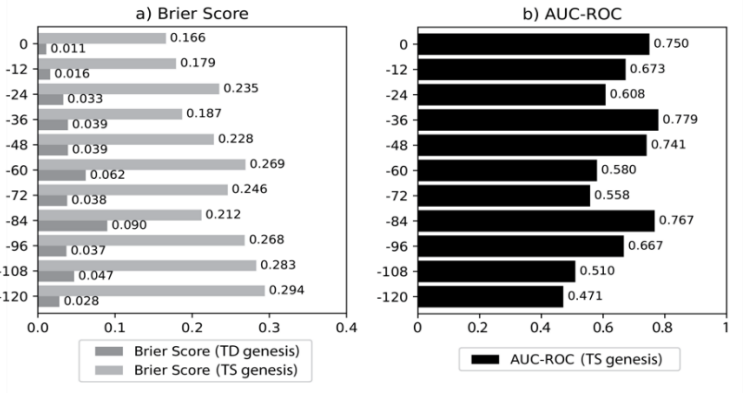

Figure 2. Brier score (a) and AUC-ROC (b) assessing the prediction of TD and TS genesis in the Vietnam East Sea using direct forecast products from WRF-LETKF.

In other words, an accurate forecast for the event of a TD genesis may not necessarily convey more accurate information about the TD's development into a full-fledged TS. It is the motivation of this study to use products in TD forecasted by the WRF-LETKF system to predict the probability of TS in a statistical combination.

4. Selection of predictors

In the context of multivariate regression model, it is crucial to avoid strong correlations among the independent variables. A high correlation suggests that the features of one independent variable closely coincide with those of another, potentially leading to a substantial reduction in the reliability of the model's regression coefficients. Hence, in the current study, we conducted an in-depth assessment aimed at evaluating multicollinearity among the independent variables and explore the relationships between these independent and dependent variables (Figure 3).

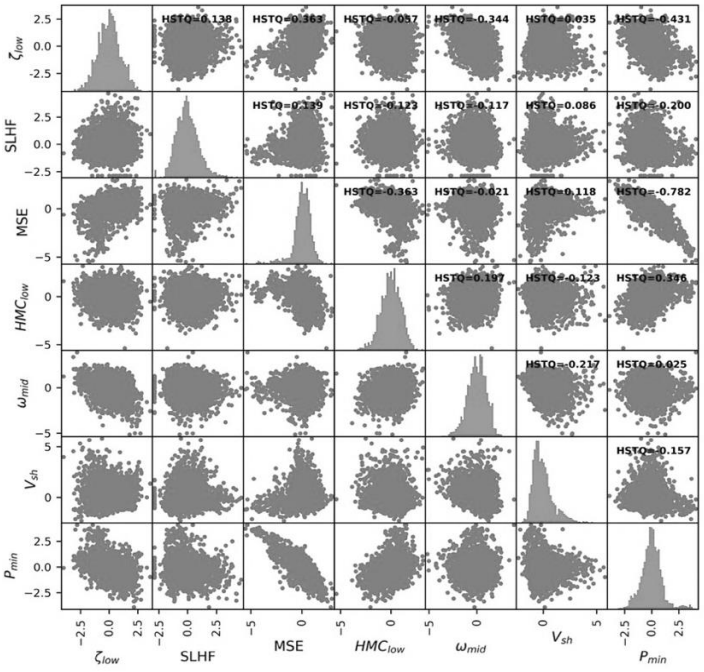

Figure 3. Correlation matrix between pairs of candidate predictors, specific correlation coefficients are highlighted in each subplot. The diagonal line represents probability distribution functions of individual parameters.

Results indicate that out of the total 7 candidate predictors selected to construct the logistic regression

equation from the forecasted product, P_{min} exhibits the highest correlation with the other variables. Specifically, the highest correlation coefficient value is observed between P_{min} and the normalized MSE (-0.782; inverse correlation). The findings also highlight the cause-and-effect relationship between the dynamic and thermodynamic features concerning the development of TCs, as manifested through the sea-level pressure at the cyclone center. Consequently, with the objective of building a logistic regression equation from independent predictors, P_{min} is not utilized as a candidate variable.

The predictor variables are directly chosen from the remaining candidate variables, which include Student t-tests. The selection results, as shown in Table 2, demonstrate that, at a statistical significance level of 5%, all candidate predictor variables meet the criteria and are selected as forecasting variables. Therefore, all dynamic and thermodynamic parameters selected for evaluation at the time of TD genesis exhibit distinct characteristics between developing and non-developing TC groups up to TS intensity.

5. Results

5.1. Ensemble dynamic - probabilistic model detecting TS genesis

Table 2. Regression coefficients and statistical test values in logistic regression model.

Eq.	Variable	Regression coefficient	Wald	P-value	Eq.	Variable	Regression coefficient	Wald	P-value
	Intercept*	1.6243	1343.6146	<0.01		intercept*	1.6661	1350.3482	<0.01
	$\zeta_{\lambda o \omega}$*	0.3576	50.2584	<0.01		$\zeta_{\lambda o \omega}$*	0.4036	61.2068	<0.01
1	*SLHF**	0.4075	78.5723	<0.01	2	*SLHF**	0.4268	82.0777	<0.01
	*MSE**	1.4384	404.3414	<0.01		*MSE**	1.4215	405.1923	<0.01
	HMC_{low}*	0.2898	36.3438	<0.01		HMC_{low}*	0.3384	47.6302	<0.01
	$\omega_{\mu\iota\delta}$	-0.0246	0.2533	0.6147		$\omega_{\mu\iota\delta}$	-0.0162	0.108	0.7425
	V_{sh}	-0.0444	1.1774	0.2779		V_{sh}	-0.076	3.2888	0.0698
	intercept*	1.615	1321.1707	<0.01		intercept*	1.5918	1322.6212	<0.01
	$\zeta_{\lambda o \omega}$*	0.3662	51.6439	<0.01		$\zeta_{\lambda o \omega}$*	0.3658	53.2859	<0.01
3	*SLHF**	0.4947	110.8745	<0.01	4	*SLHF**	0.4133	81.8117	<0.01
	*MSE**	1.4487	415.7788	<0.01		*MSE**	1.3863	400.4556	<0.01
	HMC_{low}*	0.2965	38.7205	<0.01		HMC_{low}*	0.2942	37.8519	<0.01
	$\omega_{\mu\iota\delta}$	-0.0137	0.0786	0.7792		$\omega_{\mu\iota\delta}$	-0.0123	0.0634	0.8012
	V_{sh}	-0.0813	3.7935	0.051		V_{sh}	-0.0415	1.0286	0.3105
	intercept*	1.6016	1318.983	<0.01					
	$\zeta_{\lambda o \omega}$*	0.3677	53.4547	<0.01		*Values are significant at 95% level.			
5	*SLHF**	0.4353	87.701	<0.01					
	*MSE**	1.4676	421.0099	<0.01					
	HMC_{low}*	0.3307	47.9665	<0.01					
	ω_{mid}	-0.0015	0.001	0.9752					
	V_{sh}	-0.0632	2.4137	0.1203					

From the results shown in Table 2, the regression coefficients of most predictors, except for ω_{mid} and V_{sh}, are positive, indicating a positive relationship between thermodynamic factors at the time of formation and the development of TCs to TS intensity. This implies that the likelihood of TCs reaching TS intensity increases with low-level vorticity and enhanced moisture convergence at the time of formation. The results clearly demonstrate the impact of vorticity, humidity, and low-level moisture convergence at formation on the forecast probability of TC development to TS, with statistical significance (p-value < 5%).

Conversely, the estimated regression coefficients for ω_{mid} and V_{sh}, are negative, suggesting that mid-level vorticity and decreased wind shear in the environment contribute

to the potential for TC development. However, the Wald test indicates that these parameters do not have a significant impact (ω_{mid} has a p-value > 0.6 in all prediction equations) because they do not meet the 95% confidence level of the test. With these results, the probabilistic equations for the development of TCs in the Vietnam East Sea from the logistic regression model as follows: $\mathrm{p(TS\ genesis)} = \frac{1}{1+e^{-z}}$ with:

Eq. 1: $z = 1.6243 + 1.4384 \times MSE + 0.4075 \times SLHF + 0.3576 \times \zeta_{low} + 0.2898 \times HMC_{low}$

Eq. 2: $z = 1.6661 + 1.4215 \times MSE + 0.4268 \times SLHF + 0.4036 \times \zeta_{low} + 0.3384 \times HMC_{low}$

Eq. 3: $z = 1.6150 + 1.4487 \times MSE + 0.4947 \times SLHF + 0.3662 \times \zeta_{low} + 0.2965 \times HMC_{low}$

Eq. 4: $z = 1.5918 + 1.3863 \times MSE + 0.4133 \times SLHF + 0.3658 \times \zeta_{low} + 0.2942 \times HMC_{low}$

Eq. 5: $z = 1.6016 + 1.4676 \times MSE + 0.4353 \times SLHF + 0.3677 \times \zeta_{low} + 0.3307 \times HMC_{low}$

5.2. Validation of the model predictability

The predictive outcomes of the logistic regression model concerning the potential development of TCs reaching TS stage in the Vietnam East Sea are assessed across test data reflecting variations between different forecasting cycles (Table 3). In general, BS values are relatively low, mostly below 0.2, in all forecasting cycles and prediction cases, demonstrating the model's high predictive accuracy. Specifically, BS is at its lowest value in the 60-hr forecast cycle (~2,5 days) when the logistic regression model is applied, especially in Eq. 3, with a value of 0.0893.

Table 3. Forecasting skill scores evaluating the probabilistic model on test data.

Fcst. cycles (hrs)	Ensemble dynamic-probabilistic model on test data									
	Eq. 1		Eq. 2		Eq. 3		Eq. 4		Eq.5	
	AUC-ROC	BS	AUC-ROC	BS	AUC-ROC	BS	AUC-ROC	BS	AUC-ROC	BS
-120	0.6538	0.1482	0.7083	0.1515	0.6731	0.1582	0.8846	0.1495	0.8077	0.1355
-108	0.7885	0.1369	0.8654	0.1391	0.8077	0.1344	0.8462	0.1337	0.5417	0.0972
-96	0.6667	0.1681	0.5417	0.1723	0.5833	0.1668	0.4792	0.1718	0.6875	0.1556
-84	0.7556	0.1073	0.6667	0.1145	0.6667	0.1248	0.6222	0.1170	0.7333	0.1033
-72	0.6667	0.1672	0.7308	0.1538	0.6731	0.1598	0.7692	0.1397	0.6346	0.1650
-60	0.7949	0.0893	0.7949	0.1063	0.7885	0.1353	0.8974	0.1008	0.8077	0.1377
-48	0.7115	0.1522	0.7292	0.1443	0.7436	0.1218	0.7179	0.1165	0.8205	0.1041
-36	0.7708	0.1504	0.7727	0.1590	0.6250	0.1539	0.8333	0.1466	0.7576	0.1360
-24	0.8974	0.0977	0.7500	0.1497	0.5385	0.1595	0.7500	0.1537	0.6923	0.1514
-12	0.8393	0.1317	0.6538	0.1472	0.7857	0.1412	0.7500	0.1522	0.7857	0.1393
0	0.7308	0.1527	0.7115	0.1593	0.8205	0.1211	0.8077	0.1435	0.6875	0.1630

Furthermore, AUC-ROC are relatively high, ranging between 0.8 to 0.9 from all equations. This indicates that the regression models effectively differentiate between cases of TC development and non-development based on the environmental variables described. It also suggests that the regression model performs most effectively during the 60-hr forecast cycle compared to other cycles. In contrast, during the 96-hr cycle, all the prediction equations yield BS values exceeding 0.16, and AUC-ROC hovers around 0.5.

An overall assessment reveals that among the proposed prediction equations, Eq. 1 demonstrates the highest predictive capability across the test dataset, with relatively high AUC-ROC (the lowest being 0.6538 at the 120-hr cycle) and BS values below 0.1 in the 60-hr and 24-hr cycles before the formation.

Table 4. Forecasting skill scores evaluating the probabilistic model on test data.

Fcst. cycles (hrs)	Ensemble dynamic-probabilistic model in 2012-2019 period									
	Eq. 1		Eq. 2		Eq. 3		Eq. 4		Eq. 5	
	AUC-ROC	BS	AUC-ROC	BS	AUC-ROC	BS	AUC-ROC	BS	AUC-ROC	BS
-120	0.7885	0.1507	0.7692	0.1504	0.8077	0.1497	0.8077	0.1517	0.8077	0.1508
-108	0.8846	0.1198	0.8846	0.1192	0.8846	0.1171	0.8846	0.1207	0.8846	0.1192
-96	0.7500	0.1725	0.7500	0.1732	0.7500	0.1719	0.7500	0.1737	0.7500	0.1729
-84	0.8333	0.1256	0.8333	0.1256	0.8333	0.1253	0.8333	0.1271	0.8333	0.1263
-72	0.6731	0.1926	0.6731	0.1921	0.6731	0.1918	0.6731	0.1926	0.6731	0.1918
-60	0.8462	0.1459	0.8462	0.1454	0.8462	0.1442	0.8462	0.1463	0.8462	0.1453
-48	0.7500	0.1464	0.7500	0.1464	0.7500	0.1447	0.7500	0.1464	0.7500	0.1457
-36	0.7292	0.1602	0.7083	0.1603	0.7083	0.1601	0.7292	0.1609	0.7083	0.1598
-24	0.6923	0.1483	0.6923	0.1489	0.6923	0.1481	0.6923	0.1490	0.6923	0.1483
-12	0.7143	0.1964	0.7143	0.1977	0.7143	0.1971	0.7143	0.1969	0.7143	0.1970
0	0.6923	0.1987	0.6731	0.1998	0.6923	0.1979	0.6923	0.1988	0.6923	0.1984

By examining BS and AUC-ROC, it is evident that the dynamical-statistical hybrid model has seen a significant improvement compared to the forecast results analyzed in Figure 2, especially in the early forecast cycles before actual formation time. In these cycles, all prediction equations exhibit notably reduced BS values, ranging from over 0.23-0.3 (Figure 2) down to 0.11-0.19 (Table 4). This reduction indicates that the forecast model provides a more accurate description of whether or not a TC will develop to TS intensity. The lowest BS values are achieved in the 120-hr and 108-hr forecast cycles preceding the actual formation, with BS decreasing from 0.294 to an average of 0.15 and from 0.283 to an average of 0.12, respectively. Although the results do not significantly differ between the various forecast cycles, the analysis shows that BS increases in the 12-hr preceding the formation compared to the original ensemble forecasting products, suggesting that the potential development of at TD to TS intensity can be better identified when the initial conditions incorporate basic information about atmospheric circulation and the nearby TC structure at the time of formation.

Observing BS and AUC-ROC, no significant differences are found among the logistic regression equations, indicating similarity in performance across the cross-validated train and test datasets. The verification metrics for each prediction equation on the cross-validated test dataset (Table 3) and the entire dataset (Table 4) yield similar results, ensuring the model's consistent forecasting capabilities. Therefore, the optimized logistic regression model, based on the proposed ensemble forecasting results, can be considered an effective tool for forecasting the likelihood of TC development to TS intensity in the Vietnam East Sea.

6. Summary and conclusion

While the general performance of the assimilated multi-physics WRF-LETKF in short-term forecasting, regarding tropical disturbances evolving into TD, is quite accurate when assessing probabilities and errors. However, the pure dynamical model's forecast for the likelihood of further development into TS exhibits lower accuracy. In the study, we tried to develop a simple statistical model from these products and discussed the capability of forecasting the TS genesis of TCs in the SCS using the logistic regression model. We develop an ensemble dynamic-probabilistic forecast model aiming at forecasting a probability of the higher or lower level of TS genesis frequency of TC in each of the 45 TC formation events during the period 2012-2019.

The forecast model incorporated a total of seven candidate predictors, which encompassed both dynamic parameters (minimum sea-level pressure, relative vorticity of the

lower-level, vertical velocity of the middle-level, and vertical wind shear) and thermodynamic parameters (moist static energy, surface latent heat flux, accumulated moisture convergence of the lower-level). Out of these seven variables, six were selected as predictors for the logistic regression model after multicollinearity verification and a significance test, except for minimum sea-level pressure. We developed a combination of equations for each train-test datasets in cross-validation procedure and concluded that vertical velocity at lower level and vertical wind shear have a minor impact to the probabilistic forecast, therefore excluding them from our newly developed logistic regression model.

In a broader context, when considering the performance of the logistic regression equations derived from the ensemble prediction products, the results are promising. The model generates a forecast probability between 0 and 1 for each member forecast. The combination of probabilities between ensemble members for a forecast lead time determines the overall likelihood of TS genesis. The forecasting skill using this approach has improved compared to direct forecast products, especially in the early forecast cycles before the actual formation occurs. Additionally, the results of the regression coefficients and forecasts produced by the model align closely with the outcomes from the utilization of test dataset, thus indicating the stability and validity of the newly developed model as an accurate statistical tool. Consequently, the forecast probability generated by this model can be effectively utilized as a probabilistic forecast value in predicting the status of TS genesis.

The logistic regression model offers the advantage of providing rapid forecasting information while accommodating the non-linear nature of variables. In the future, our focus will be on extending the forecast time range of the model and incorporating a more comprehensive set of variables for a predictive model of TS genesis.

Author contributions: Conceptualization: T.T.T., D.N.Q.H.; Methodology: D.N.Q.H.; Data curation: D.N.Q.H.; Writing-original draft: D.N.Q.H.; Writing-editing: T.T.T.

Acknowledgement: The authors are grateful for the time and effort given by the anonymous reviewers whose contributions greatly strengthened this manuscript.

Statement: The authors collectively declare that this article represents our research work, has not been previously published elsewhere, and is not copied from any previous studies. There are no conflicting interests among the group of authors.

References

1. Halperin, D.J.; Fuelberg, H.E.; Hart, R.E.; Cossuth, J.H. Verification of tropical cyclone genesis forecasts from global numerical models: Comparisons between the North Atlantic and eastern North Pacific basins. *Wea. Forecasting* **2016**, *31*, 947–955.
2. Liang, M.; Chan, J.C.L.; Xu, J.; Yamaguchi, M. Numerical prediction of tropical cyclogenesis Part I: Evaluation of model performance. *Quart. J. Roy. Meteor. Soc.* **2021**, *147*, 1626–1641.
3. Jaiswal, N.; Kishtawal, C.M.; Bhomia, S.; Pal, P.K. Multi-model ensemble-based probabilistic prediction of tropical cyclogenesis using TIGGE model forecasts. *Meteor. Atmos. Phys.* **2016**, *128*, 601–611.
4. Pedlosky, J. Geophysical fluid dynamics. 1979.
5. Zhang, X.; Yu, H. A probabilistic tropical cyclone track forecast scheme based on the selective consensus of ensemble prediction systems. *Wea. Forecasting* **2017**, *32*, 2143–2157.
6. Zhang, X.; Fang, J.; Yu, Z. The forecast skill of tropical cyclone genesis in two global ensembles. *Wea. Forecasting* **2023**, *38(1)*, 83–97. https://doi.org/10.1175/WAF-D-22-0145.1.

7. Hunt, B.R.; Kostelich, E.J.; Szunyogh, I. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena* **2007**, *230*, 112–126.
8. Miyoshi, T.; Kunii, M. Using AIRS retrievals in the WRF-LETKF system to improve regional numerical weather prediction. *Tellus A* **2012**, *64(1)*, 18408. https://doi.org/10.3402/tellusa.v64i0.18408.
9. Kieu, C.Q.; Truong, N.M.; Mai, H.T.; Ngo-Duc, T. Sensitivity of the track and intensity forecasts of Typhoon Megi (2010) to satellite-derived atmospheric motion vectors with the ensemble Kalman filter. *J. Atmos. Oceanic Technol.* **2012**, *29(12)*, 1794–1810. https://doi.org/10.1175/jtech-d-12-00020.1.
10. Liu, J.; Fertig, E.; Li, H.; Kalnay, E.; Hunt, B.; Kostelich, E.; Szunyogh, I.; Todling, R. Comparison between local ensemble transform Kalman filter and PSAS in the NASA finite volume GCM - Perfect model experiments. *Nonlinear Processes Geophys.* **2007**, *15*, 645–659. https://doi.org/10.5194/npg-15-645-2008.
11. Kwon, H.; Lee, W.; Won, S.H.; Cha, E.J. Statistical ensemble prediction of the tropical cyclone activity over the western North Pacific. *Geophys. Res. Lett.* **2007**, *34*, 24805. https://doi.org/10.1029/2007GL032308.
12. Leroy, A.; Wheeler, M. Statistical prediction of weekly tropical cyclone activity in the Southern Hemisphere. *Mon. Weather Rev.* **2008**, *136*, 3637–3654. https://doi.org/10.1175/2008MWR2426.1.
13. Mestre, O.; Hallegatte, S. Predictors of Tropical Cyclone Numbers and Extreme Hurricane Intensities over the North Atlantic Using Generalized Additive and Linear Models. J. Clim. **2009**, *22*, 633–648. https://doi.org/10.1175/2008JCLI2318.1.
14. Chan, J.; Shi, J.E.; Lam, C.M. Seasonal forecasting of tropical cyclone activity over the western North Pacific and the South China Sea. *Wea. Forecasting* **1998**, *13*, 997–1004. https://doi.org/10.1175/1520-0434(1998)013<0997:SFOTCA>2.0.CO;2.
15. Wijnands, J.; Qian, G.; Kuleshov, Y. Variable selection for tropical cyclogenesis predictive modeling. *Mon. Weather Rev.* **2016**, *144*, 4605–4619. https://doi.org/10.1175/MWR-D-16-0166.1.
16. Wilks, D.S. Statistical methods in the atmospheric sciences. *Int. Geophys. Series* **2006**, 59, xi. https://doi.org/10.1016/S0074-6142(06)80036-7.
17. Choi, J.W.; Kang, K.R.; Kim, D.W.; Kim, T.R. Development of a probability prediction model for tropical cyclone genesis in the northwestern pacific using the logistic regression method. *J. Korean Earth Sci. Soc.* **2010**, *31(5)*, 454–464. https://doi.org/10.5467/JKESS.2010.31.5.454.
18. Tien, T.T.; Hoa, D.N.Q.; Thanh, C.; Kieu, C. Assessing the impacts of augmented observations on the forecast of Typhoon Wutip (2013)'s formation using the ensemble Kalman filter. *Wea. Forecasting* **2020**, *35(4)*, 1483–1503. https://doi.org/10.1175/waf-d-20-0001.1.
19. Knapp, K.R.; Kruk, M.C.; Levinson, D.H.; Diamond, H.J.; Neumann, C.J. The international best track archive for climate stewardship (IBTrACS): Unifying tropical cyclone data. *Bull. Am. Meteorol. Soc.* **2010**, *91(3)*, 363–376.
20. Fu, B.; Peng, M.S.; Li, T.; Stevens, D.E. Developing versus nondeveloping disturbances for tropical cyclone formation. Part II: Western North Pacific. *Mon. Weather Rev.* **2012**, *140(4)*, 1067–1080. https://doi.org/10.1175/2011MWR3618.1.
21. Kerns, B.W.; Chen, S.S. Cloud clusters and tropical cyclogenesis: Developing and nondeveloping systems and their large-scale environment. *Mon. Weather Rev.* **2013**, *141(1)*, 192–210. https://doi.org/10.1175/MWR-D-11-00239.1.
22. Gray, W. Environmental influences on tropical cyclones. *Aust. Meteorol. Mag.* **1988**, *36*, 127–139.

23. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78(1)*, 1–3.
24. Swets, J.A. The relative operating characteristic in psychology. *Science* **1973**, *182*, 990–1000.
25. Buizza, R.; Miller, M.; Palmer, T.N. Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. 1999, pp. 2887-2908.