https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

TC^2 ver. 1.0: An Objective Hybrid Tracker and Classifier for Tropical Cyclones version 1.0

Doo-Sun R. Park[1,2,3], Dasol Kim[4], Hye-Young Ko[3], Hyeong-Seog Kim[5], Dong-Hyun Cha[6], Minhee Chang[7], Seung-Ki Min[8], Minho Kwon[9], and Tae-Won Park[10]

[1]Department of Earth Science Education, Kyungpook National University, Daegu, 41566, Republic of Korea
[2]BK21 Weather Extremes Education & Research Team, Department of Atmospheric Sciences, Kyungpook National University, Daegu, 41566, Republic of Korea
[3]Center for Atmospheric REmote sensing (CARE), Kyungpook National University, Daegu, 41566, Republic of Korea
[4]Department of Environmental Engineering, Seoul National University of Science and Technology, Seoul, 01811, Republic of Korea
[5]Ocean Science and Technology School, Korea Maritime and Ocean University, Busan, 49112, Republic of Korea
[6]Department of Civil, Urban, Earth and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan, 44919, Republic of Korea
[7]Environmental Planning Institute, Seoul National University, Seoul, 08826, Republic of Korea
[8]Division of Environmental Science and Engineering, Pohang University of Science and Technology, Pohang, 37673, South Korea
[9]Ocean Climate Prediction Center, Korea Institute of Ocean Science and Technology, Busan, 49111, Republic of Korea
[10]Department of Earth Science Education, Chonnam National University, Gwangju, 61186, Republic of Korea

Correspondence to: Doo-Sun R. Park (dsrpark@knu.ac.kr) and Dasol Kim (dasol.kim@seoultech.ac.kr)

Abstract. Accurate detection of tropical cyclones (TCs) from gridded climate model data is essential for evaluating model performance and projecting future TC activity. Conventional detection schemes rely on environmental variable thresholds that are frequently tuned to specific basins or models, making them inherently subjective. Conversely, detection schemes based on universal thresholds often fail to capture regional characteristics. While recent machine learning (ML) approaches provide objective data-driven thresholds, they generally require a great number of variables compared to conventional methods and suffer from high false alarm rates (FAR). Here, we introduce an objective hybrid tracker and classifier for tropical cyclones (TC^2), an algorithm combining traditional and ML techniques to establish objective thresholds and minimize FAR. Using an ensemble of six classifiers based on three ML algorithms and two reanalyses, TC^2 avoids dependency on specific ML algorithms or datasets. TC^2 was trained and its hyperparameters were optimized using two reanalysis datasets over 1998-2017 period, while its performance was evaluated on their respective internal test sets and the independent NCEP FNL dataset from 2018 to 2024. Evaluated against the NCEP FNL dataset, TC^2 outperforms the existing algorithms, i.e., TempestExtremes and OWZP, achieving higher F1 score (83.6%) and critical success index (71.8%), while significantly lowering FAR (14.3%) and maintaining a comparable hit rate (81.5%). TC^2 also better reproduces TC count and seasonal variability over each basin. In CMIP6 evaluations, TC^2 successfully captures the overall characteristics of TC activity. Under the SSP2-4.5 scenario, projected spatial changes in TC genesis frequency detected by TC^2 are largely consistent with those of the dynamical genesis potential index, suggesting that TC^2 identifies physically coherent systems governed by large-scale dynamic environments.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

Utilizing only a limited set of commonly available variables, including minimum sea level pressure, low-level relative vorticity, upper- and low-level wind speeds, and an upper-level warm core, TC^2 provides an effective and robust framework for TC detection in gridded climate datasets.

1 Introduction

In projecting changes in tropical cyclone (TC) activity under future warming scenarios, the choice of detection and tracking schemes is one of the critical sources of uncertainty (Horn et al., 2014; Roberts et al., 2020). As highlighted by Horn et al. (2014), while one tracking scheme projected a consistent decrease in global TC frequency across a set of global climate models (GCMs) (Tory et al., 2013c), the application of another algorithm to some of the exact same GCMs resulted in projected increases (Camargo, 2013). Similarly, Horn et al. (2014) showed that individual models exhibited divergent responses in TC genesis frequency solely according to the choice of the tracking scheme under idealized doubled-CO_2 experiments. Furthermore, Roberts et al. (2020) demonstrated that the total number of detected TCs varies substantially between different tracking algorithms, particularly at lower model resolutions. Consequently, alongside the inherent dynamical limitations of GCMs themselves, the selection of a tracking scheme intrinsically adds an extra layer of uncertainty to future TC projections.

One of the main reasons for uncertainty among various trackers is the threshold values, which can be model-, resolution-, and basin- dependent or independent (Zarzycki and Ullrich, 2017; Horn et al., 2014). There is some debate as to which is more appropriate: dependent or independent ones. In dependent schemes, threshold values for variables, such as minimum sea level pressure (MSLP), low-level vorticity, warm-core temperature anomalies, surface wind speeds, and storm duration, are specifically optimized for each model, resolution, and oceanic basin (Camargo and Zebiak, 2002; Walsh et al., 2007). Although this approach has been widely applied since earlier studies, recent research questions its objectivity; such optimizations can artificially conceal the true limitations of the models by conflating model errors with detection errors, thereby hindering objective inter-model comparisons (Tory et al., 2013a; Zarzycki and Ullrich, 2017; Bourdin et al., 2022). To address this issue, independent tracking methods, such as the Okubo-Weiss-Zeta Parameter (OWZP) scheme and the TempestExtremes framework, have been introduced (Tory et al., 2013b; Ullrich and Zarzycki, 2017). These schemes employ universal, resolution-independent threshold criteria across various models without requiring model-, ocean-basin- or resolution-specific tuning (Tory et al., 2013b; Ullrich and Zarzycki, 2017). While the independent methods appear to allow for a more objective evaluation of model performance than the dependent ones, applying uniform criteria could also be questioned. Particularly, fundamental differences in large-scale circulations and background sea surface temperature (SST) across different basins could necessitate basin-dependent criteria (Camargo and Zebiak, 2002).

Recently, Machine Learning (ML) techniques have emerged as a solution, offering a data-driven approach that overcomes the limitations of subjective, user-prescribed thresholds found in traditional physical trackers (Accarino et al., 2023; Vaittinada Ayar et al., 2025; Gardoll and Boucher, 2022; Galea et al., 2024). However, current ML techniques still face notable challenges. Particularly, a high False Alarm Rate (FAR) in most ML-based algorithms emerges because of the severe class imbalance

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

inherent in meteorological datasets, where TC-free atmospheric situations (the "0" samples) vastly outnumber actual TC events (Accarino et al., 2023; Galea et al., 2024). This overwhelming preponderance of negative samples makes it difficult for algorithms to properly learn the decision boundary without becoming biased, often leading ML-based algorithms to misclassify background noise, weak disturbances, and extratropical cyclones as TCs (Accarino et al., 2023; Galea et al., 2024; Wu and Duan, 2023; Gardoll and Boucher, 2022). While Vaittinada Ayar et al. (2025) have recently mitigated this high FAR by effectively handling the data imbalance, structurally, their algorithm predicts only one TC probability per predefined grid box, meaning it fails to distinguish multiple simultaneous TCs occurring within the same area. Furthermore, since the algorithm requires a total of 20 variables, Vaittinada Ayar et al. (2025) attempted to simplify the algorithm by removing less important variables, but this resulted in a sudden spike in FAR, proving that those complex feature sets are strictly necessary to suppress FAR. Finally, these algorithms exhibit a strong dependency on their training basins, lacking the generalizability required to be deployed globally without regional retuning.

To address these limitations, we developed a Tracker and Classifier for Tropical Cyclones (TC^2), a hybrid TC tracker and classifier that integrates traditional vortex tracking with machine learning. This hybrid framework is specifically designed to overcome both the severe class imbalance inherent in meteorological datasets and the reliance on subjective threshold tuning found in conventional approaches. In the first stage, a traditional vortex tracking algorithm identifies potential TC candidates, effectively filtering out many non-TC atmospheric states. This step substantially reduces the class imbalance that typically hinders the direct application of ML to raw atmospheric data. In the second stage, an ML-based classifier evaluates these pre-screened candidates to make a final determination. The classification is based on basin-specific criteria objectively derived from the statistical properties of the training data, thereby eliminating the need for manually tuned thresholds. Detailed methodologies are provided in Section 2.

The remainder of this paper is organized as follows: Section 2 describes the data and methods; Section 3 presents the performance benchmarking of TC^2; Section 4 demonstrates its application to CMIP6; and Section 5 provides the discussion and conclusions.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

2 Data and Methods

2.1 Data

To label the tracked TCs, the official best track datasets for the period of 1998–2024 were utilized. Specifically, the datasets issued by the Joint Typhoon Warning Center (JTWC) were used for the North and South Indian Oceans and the western North and South Pacific, while those from the National Hurricane Center (NHC) were applied for the eastern North Pacific and the North Atlantic. The analysis period was limited from 1998 to 2024, as global TC data became significantly more reliable following the advent of comprehensive geostationary satellite coverage since 1998 (Kossin et al., 2007; Kuleshov et al., 2010). Best-track data for the Southern Hemisphere in the second half of 2024 are not yet available; points in this space–time window are therefore excluded from all analyses. It should be noted that only the systems reaching at least tropical storm (TS) intensity (i.e., maximum sustained wind speeds of at least 34 knots) were defined as actual TCs in this study. The detailed labelling method is described in Section 2.2.1.

To extract environmental features around TCs and evaluate the performance of the TC^2 algorithm, two different reanalysis datasets, the fifth generation of the European Centre for Medium-Range Weather Forecasts Reanalysis (ERA5) and the Japanese Reanalysis for Three Quarters of a Century (JRA-3Q), and one operational analysis dataset, the National Centers for Environmental Prediction (NCEP) Final (FNL) analysis, were utilized. Reanalysis datasets like ERA5 and JRA-3Q, with horizontal resolutions of 0.25°×0.25° and 1.25°×1.25°, respectively, provide a spatially complete and temporally coherent record produced with a frozen data assimilation system, making them suitable for training the classifier with less risk of artificial shifts (Hersbach et al., 2020; Kosaka et al., 2024). To evaluate the TC^2 algorithm, samples within both the ERA5 and JRA-3Q datasets were respectively divided into a training set (1998–2017) and an internal test set (2018–2024). For independent testing, the NCEP FNL was utilized for the same period of internal test set (2018–2024). Unlike reanalyses, the NCEP FNL is an operational analysis that incorporates late-arriving observational data to represent an accurate snapshot in time. Its underlying data assimilation system and forecast model are subject to frequent updates that may introduce temporal uncertainties such as artificial shifts (http://rda.ucar.edu/datasets/ds083.2/). Nevertheless, its resolution of 1.0°×1.0°, which is comparable to the resolution of many GCMs participating in Coupled Model Intercomparison Project phase 6 (CMIP6) (Eyring et al., 2016), makes the NCEP FNL dataset a suitable testbed for this study in terms of the resolution. Furthermore, because the TC^2 algorithm was trained on higher-resolution ERA5 and relatively lower-resolution JRA-3Q, applying it to the NCEP FNL, which has a different origin and resolution, allows us to rigorously evaluate the cross-dataset transferability of the algorithm. Utilizing the NCEP FNL over the 2018–2024 period enables an evaluation of the TC^2 algorithm's performance, as it can be directly verified against the available best-track information. Further details on the application of these datasets are provided in Section 2.2.2.

To test the TC^2 algorithm's performance when applied to GCMs, we utilized 6-hourly outputs from 13 different GCMs participating in the CMIP6 (Table 1). The selected models were chosen mainly because the necessary 6-hourly output data were publicly accessible when the study was performed. To project future changes in TC activity, we utilized a historical

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

period from 1985 to 2014, and a future period from 2070 to 2099 under the Shared Socioeconomic Pathway 2-4.5 (SSP2-4.5) scenario. Since this application is intended solely to test of the physical reliability of TC^2, we evaluated the far future only under SSP2-4.5, a scenario recently argued to be more realistic than SSP5-8.5 (Chen et al., 2021). Notably, for its application to the CMIP6 datasets, the TC^2 algorithm was trained over the entire 1998–2024 period. It should be noted that MIROC-ES2H was excluded from the calculation of the projected future changes in the dynamic genesis potential index (DGPI) and genesis density shown in Figure 6, because its monthly output to calculate DGPI is not available.

Table 1. GCMs used and their horizontal resolutions (https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html).

GCMs used	Grid dimension (lon × lat)	Nominal resolution (lon × lat)
ACCESS-CM2	192 x 144	1.875° x 1.25°
BCC-CSM2-MR	320 x 160	1.125° x 1.125°
EC-Earth3	512 x 256	0.703125° x 0.703125°
KIOST-ESM	192 x 96	1.875° x 1.875°
MIROC-ES2H	256 x 128	1.40625° x 1.40625°
MIROC6	256 x 128	1.40625° x 1.40625°
MPI-ESM1-2-HR	384 x 192	0.9375° x 0.9375°
MPI-ESM1-2-LR	192 x 96	1.875° x 1.875°
MRI-ESM2-0	320 x 160	1.125° x 1.125°
NESM3	192 x 96	1.875° x 1.875°
NorESM2-LM	144 x 96	2.5° x 1.875°
NorESM2-MM	288 x 192	1.25° x 0.9375°
TaiESM1	288 x 192	1.25° x 0.9375°

2.2 The TC^2 algorithm

The TC^2 algorithm was developed through two main phases (Figure 1): (1) the detection, tracking, environmental feature extraction, and labelling of TC candidates, and (2) the classification of these candidates into true TCs. In the first phase, the detection and tracking of TC candidates followed a procedure similar to conventional threshold-based tracking algorithms, but relied on a vorticity criterion to identify potential candidate locations, with MSLP used to retain and locate the center of each candidate. Following detection and tracking, environmental features were computed at each time step. Lastly, the TC candidates were labelled as either actual TCs or non-TCs by comparing them against best-track datasets. In the second phase, the ML models were trained and their hyperparameters were optimized. Finally, the optimal consensus threshold for the ensemble and the duration criteria were determined by comparing the performance of various combinations. Ultimately, the

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

finalized ensemble voting system classifies each candidate as an actual TC or a non-TC. Detailed descriptions are provided in Sections 2.2.1 and 2.2.2. For user convenience, the released version of TC^2 provides a ready-to-use pipeline that includes the detection, tracking, and environmental feature extraction, as well as the final ensemble voting system with pretrained ML classifiers.

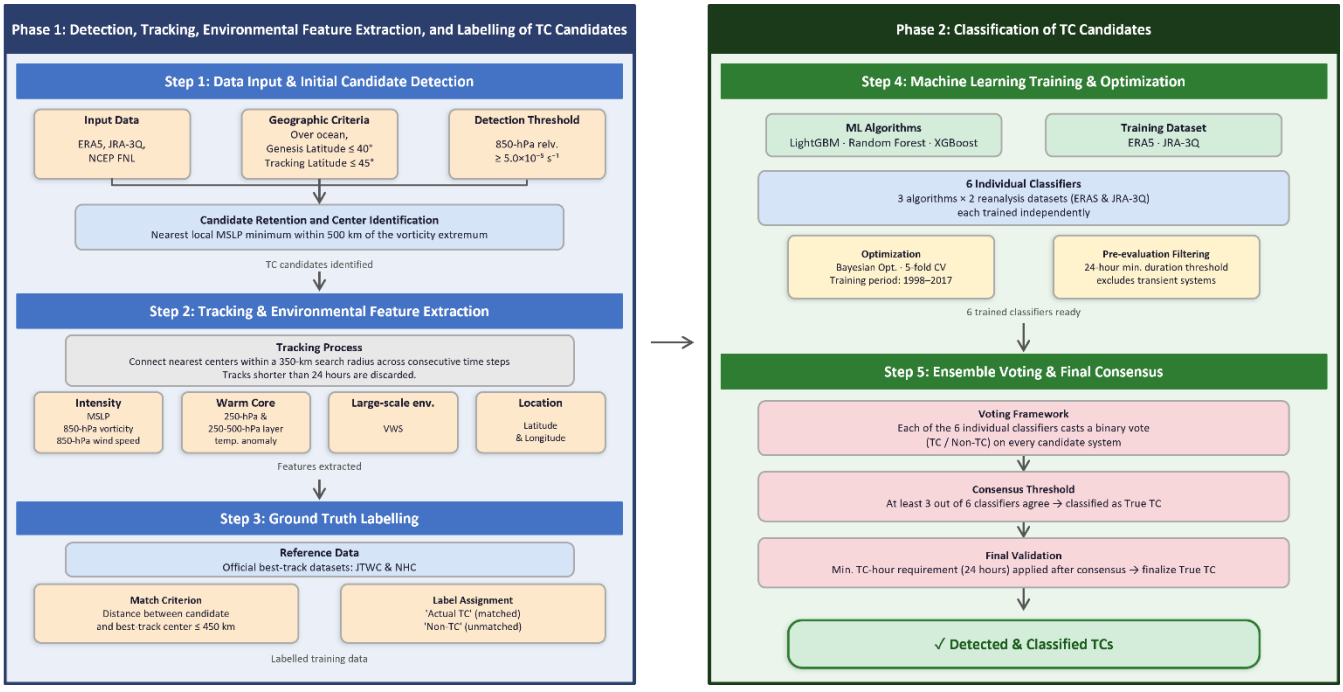

Figure 1. Schematics of the TC^2 algorithm.

2.2.1 Detection, Tracking, Environmental Feature Extraction, and Labelling of TC Candidates

To objectively identify and track TC candidates from the gridded atmospheric data (ERA5, JRA-3Q, and NCEP FNL), a detection and tracking algorithm, which is similar to conventional ones, was employed. For the detection procedure, the locations of the systems were restricted geographically. Genesis location was limited to latitudes at or below 40°, while the tracking location was restricted to within 45° latitude. A land-sea mask was also applied to differentiate the genesis and tracking criteria. The initial genesis of a system was strictly limited to locations over the ocean. However, during the tracking phase, systems moving over land were still detected as long as at least one oceanic grid point was present within a rectangular box of ±4M grid points surrounding the candidate grid point, where $M = \max\left(d/\Delta x, 1\right), d = 2.5°$, and Δx is the grid spacing. This corresponds to approximately ±10° for ERA5 and JRA-3Q, and ±8° for FNL. Unlike conventional tracking algorithms, which usually require multiple intensity-related criteria to be satisfied simultaneously, such as MSLP, surface wind speed, and warm-core temperature anomalies (Horn et al., 2014; Walsh et al., 2007; Ullrich et al., 2021; Camargo and Zebiak, 2002), our initial screening step used only the 850-hPa relative vorticity (magnitude of at least $5.0 \times 10^{-5}\ s^{-1}$) to comprehensively capture all potential TC candidates, both organized and unorganized, by excluding other variables at this stage. After a vorticity extremum

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

exceeded this threshold, the candidate was retained only when a corresponding local minimum in MSLP was found within 500 km of the vorticity extremum, which was then regarded as the center of the system. For the tracking procedure, the identified centers were connected across consecutive time steps. A continuous track was formed by linking the nearest center within a maximum search radius of 350 km. When multiple tracks competed for the same center, the conflict was resolved by proximity, with ties broken by track length and then by MSLP intensity. To eliminate transient or spurious systems, only tracks that persisted for a minimum duration of 24 hours were retained. The vorticity threshold ($5.0 \times 10^{-5}\ s^{-1}$) in this study was empirically determined to balance the detection rate of actual TCs against the FAR. If the threshold is set too high, the algorithm risks missing a substantial number of actual TCs. Conversely, if the threshold is too low, the candidate pool becomes overwhelmingly populated by spurious systems, which severely degrades the performance of the subsequent classification phase. With our chosen threshold, the tracker misses only 1.7%, 1.8%, and 1.2% of best-track TC time steps for ERA5 (1998–2024), JRA-3Q (1998–2024), and FNL (2018–2024), respectively, where a candidate is considered to match a best-track point if the distance between their centers is 450 km or less. At the event level, only 3.1%, 2.6%, and 3.2% of TCs are missed, where a TC is considered captured if a single candidate track covers at least 60% of its best-track time steps.

To assess the background conditions and structural characteristics of the tracked candidates, and to prepare for the subsequent classification, relevant environmental features were calculated at each 6-hourly time step. The intensity of a candidate is represented by the MSLP averaged within a 200-km radius, the 850-hPa relative vorticity averaged within a 350-km radius, and the maximum 850-hPa wind speed among 1-degree radial bin averages within a 500-km radius from the storm center. To identify the upper-level warm core structure, the algorithm computes the maximum 250-hPa temperature and the maximum vertically averaged temperature of the 250–500-hPa layer, both within a 278-km radius from the storm center. The strength of the warm core is then quantified as the anomaly of these maxima relative to the background environment, where the background temperature is defined as the azimuthal mean over a 400–800-km annular region from the storm center. Finally, to evaluate the large-scale environment, the vertical wind shear (VWS) between 850 and 250 hPa is computed as the azimuthal mean over a 400–800-km annular region from the storm center. To avoid weighting bias from the larger number of grid cells at greater radii, all azimuthal means were obtained by first averaging within 1-degree radial bins and then averaging across bins. How these features are aggregated and fed into the classifier is described in Section 2.2.2. Furthermore, to train the subsequent classifier, the detected candidates are labelled using the official best-track datasets provided by the JTWC and the NHC. A detected candidate is considered identical to a best-track TC if the distance between their centers is 450 km or less. Upon a successful match, the candidate is assigned a label of 'actual TC', whereas unmatched candidates are labeled as 'non-TCs'. This labelling process provides the ground-truth labels required to train and evaluate the classifier's ability to distinguish true TCs from spurious disturbances.

2.2.2 Classification of TC Candidates and Classifier Performance Evaluation

Three different ML algorithms were utilized to classify TC candidates into true TCs: Random Forest, XGBoost, and LightGBM (Ke et al., 2017; Breiman, 2001; Chen and Guestrin, 2016). The classifiers were trained using the following features:

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

195 latitude, longitude, maximum 850-hPa wind speed, large-scale VWS, 850-hPa vorticity, MSLP, and 250-hPa and 250–500-hPa warm core anomalies. To find the optimal hyperparameters for each model, we employed Bayesian Optimization (Nogueira, 2014). The optimization was designed to maximize the mean F1 score of a year-based 5-fold grouped cross-validation using only the training set (1998–2017). All algorithms were optimized with 5 initial random explorations and 40 sequential iterations, except for XGBoost on JRA-3Q which used 50 iterations as the initial 40 did not achieve convergence.
200 The search bounds for each algorithm and dataset are summarized in Table 2, along with the final optimized hyperparameters.

Table 2. The optimal hyperparameters of each ML-based classifier

Algorithm	Parameter (Search range)	ERA5	JRA-3Q
Random Forest	n_estimators (1–301)	95	245
	max_depth (1–100 for ERA5 &1–61 for JRA-3Q)	21	41
	min_samples_leaf (1–21)	1	1
XGBoost	n_estimators (1–201)	99	147
	max_depth (1–21)	14	16
	learning_rate (0.01–1.0)	0.058	0.099
	subsample (0.1–1.0)	0.466	0.946
	gamma (0.0–5.0)	0.199	4.666
Light GBM	n_estimators (1–200)	160	158
	max_depth (1–20)	3	11
	min_data_in_leaf (1–20)	16	1
	learning_rate (0.01–1.0)	0.390	0.044

Note that an additional duration threshold was applied after classification to exclude borderline systems. While the detection
205 phase required a minimum track duration of 24 hours for all candidates (Section 2.2.1), a further 24-hour persistence criterion

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

was imposed at the classification stage: a candidate was labeled as a TC only if the classifier identified it as such for at least 24 hours (4 time steps, not necessarily consecutive). This minimum TC-hour requirement was empirically determined to maximize the detection skill of TC^2 against the observations in Table 5: the 24-hour threshold yields the highest F1 score and critical success index (CSI) in the evaluation against the best-track over the NCEP FNL (2018–2024). A shorter threshold admits more short-lived spurious systems and thus more false alarms, whereas a longer threshold removes genuine TCs and thus more misses; the 24-hour value best balances the two.

In the internal test sets (2018–2024), the classifiers trained and evaluated on the ERA5 and JRA-3Q datasets exhibited strong classification capabilities (Table 3). The classifiers based on ERA5 achieved F1 scores of 79.5%–80.6%, accuracies of 98.0–98.1%, hit rates of 70.7%–73.5%, and CSIs of 65.9%–67.5%, with FARs of 9.3%–10.8%. The classifiers based on JRA-3Q showed higher hit rates of 75.2%–77.9%, higher CSIs of 71.8%–73.5%, and higher F1 scores of 83.6%–84.7%, with lower FARs of 6.0%–7.2%, compared to those based on ERA5. Overall, the JRA-3Q-based classifiers outperformed the ERA5-based ones across most metrics.

Table 3. The performance of individual classifiers for the internal test sets.

ML	Trained dataset	Accuracy (%)	Hit rate (%)	F1 score (%)	FAR (%)	CSI (%)
Random Forest	EAR5	98.0	70.7	79.5	9.3	65.9
XGBoost		98.1	73.5	80.6	10.8	67.5
Light GBM		98.1	72.0	79.8	10.5	66.4
Random Forest	JRA3Q	96.4	75.2	83.6	6.0	71.8
XGBoost		96.6	77.9	84.7	7.2	73.5
Light GBM		96.5	77.2	84.3	7.2	72.9

To evaluate cross-dataset transferability, all individual classifiers were applied to the NCEP FNL data (2018–2024), which differs in resolution and origin from ERA5 and JRA-3Q. Overall, all classifiers maintained strong performance on this independent dataset, with accuracies of 97.9%–98.1% and F1 scores of 83.2%–83.9% (Table 4). The ERA5-based classifiers achieved F1 scores of 83.2%–83.9%, hit rates of 77.2%–79.5%, and CSIs of 71.2%–72.2%, with FARs of 9.8%–11.3%. The JRA-3Q-based classifiers achieved comparable F1 scores of 83.4%–83.7% and higher hit rates of 81.3%–83.4%, but at the cost of notably higher FARs of 14.2%–16.6%. This contrasts with the internal test, where JRA-3Q-based classifiers exhibited lower FARs than ERA5-based ones, suggesting that the JRA-3Q-based classifiers become more aggressive when applied to a different dataset. Nevertheless, the consistent F1 scores and accuracies across all classifiers support the cross-dataset transferability of the approach.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

Table 4. The performance of individual classifiers for the independent test set, the NCEP FNL.

ML	Trained dataset	Accuracy (%)	Hit rate (%)	F1 score (%)	FAR (%)	CSI (%)
Random Forest	ERA5	98.0	77.2	83.2	9.8	71.2
XGBoost		98.1	79.5	83.9	11.2	72.2
Light GBM		98.0	78.7	83.4	11.3	71.5
Random Forest	JRA-3Q	98.0	81.3	83.5	14.2	71.7
XGBoost		97.9	83.4	83.4	16.6	71.5
Light GBM		98.0	83.2	83.7	15.8	72.0

Our final classifier is based on an ensemble voting system. Because the three different ML algorithms were independently trained on two different reanalysis datasets (ERA5 and JRA-3Q), a total of six individual classifiers were generated. In this ensemble approach, each of the six classifiers casts a vote, and a classified system is ultimately regarded as a true TC based on the consensus of these votes. In this study, a consensus is defined as an agreement among three or more classifiers; the rationale for this specific threshold is provided in Table 5.

Table 5. The performance of the final classifier according to the least number of classifiers in agreement.

Number of classifiers in agreement	Accuracy (%)	Hit rate (%)	F1 score (%)	FAR (%)	CSI (%)
1	97.8	87.1	83.4	20.0	71.5
2	98.0	84.9	84.1	16.6	72.6
3	98.1	82.9	84.3	14.2	72.9
4	98.1	79.7	84.3	10.6	72.8
5	98.0	76.4	83.1	8.9	71.1
6	97.9	72.3	81.3	7.0	68.6

This ensemble configuration (threshold=3) achieves an accuracy of 98.1%, an F1 score of 84.3%, and a CSI of 72.9% on the NCEP FNL dataset, slightly outperforming the best individual classifier in F1 (83.9%) and CSI (72.2%). The ensemble hit rate of 82.9% is marginally lower than the best individual hit rate (83.4%), while the FAR of 14.2% is higher than the lowest individual FAR (9.8%). However, combining classifiers trained on different reanalysis datasets mitigates the risk of overfitting to the specific biases of a single dataset, ensuring higher stability and generalizability when applied to independent datasets

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

such as CMIP models. It should be noted that the minimum TC-hour requirement (24 hours) was applied only after all individual votes were collected and the final consensus was reached.

2.3 Performance Benchmarking and Physical Robustness Assessment of TC^2

To evaluate TC^2's performance relative to conventional methods, we applied two different tracking frameworks exclusively to the NCEP FNL dataset which serves as the independent test set in this study: TempestExtremes and the Okubo-Weiss-Zeta Parameter (OWZP) (Tory et al., 2013b; Ullrich and Zarzycki, 2017). A key rationale for selecting these two schemes is their resolution-insensitive nature; they are designed to avoid resolution-dependent threshold issues. The OWZP scheme was implemented using its universal, resolution-independent thresholds based on large-scale environmental variables. Unlike traditional trackers that attempt to directly identify TC circulations using grid-dependent extrema, the OWZP scheme is designed to identify the large-scale environmental conditions favorable for TC formation (Tory et al., 2013b). Specifically, by evaluating the OWZP to isolate regions of low-deformation vorticity and near solid-body rotation, along with relative humidity and VWS, all of which are well-resolved even in coarse GCMs, the scheme successfully avoids the need for subjective, resolution-dependent threshold adjustments (Tory et al., 2013b). TempestExtremes achieves resolution independence by employing a closed-contour criterion evaluated over a physical great-circle distance, thereby circumventing the resolution-dependency associated with traditional grid-based metrics (Ullrich and Zarzycki, 2017). The detailed detection and tracking criteria can be found in their respective references.

It is important to clarify a methodological difference in the construction of the confusion matrices between the standalone evaluation (Tables 3–5) and the comparative assessment against OWZP and TempestExtremes (Table 6). In the standalone TC^2 evaluation, the confusion matrix is constructed at the point (time-step) level from the candidates identified during Phase 1 (Fig. 1), where true negatives are naturally defined as candidate points not classified as TCs. Actual TCs missed in Phase 1 are implicitly excluded from this matrix, but the number of such cases is negligibly small (less than 2% at the time-step level). When comparing TC^2 with other algorithms, however, the existing algorithms output only detected TC events, so true negatives cannot be defined in the same way. The comparative evaluation in Table 6 therefore uses only TP, FP, and FN, together with metrics that do not require TN (F1 score, hit rate, FAR, and CSI). At each 6-hourly time step, algorithm output points and best-track points of at least TS intensity are matched one-to-one by greedy nearest-neighbor assignment within a 450 km radius: an algorithm point matched to such a best-track point is counted as TP, an unmatched algorithm point as FP, and a best-track point of at least TS intensity not matched by any algorithm output as FN. These methodological differences account for the small discrepancies in TC^2's performance metrics between Table 5 and Table 6.

To investigate future spatial changes in TC genesis frequency extracted by the TC^2 algorithm from CMIP6 models, the spatial distribution of TC genesis frequency, defined as TC genesis density, was computed. To construct TC genesis density, the annual number of genesis events was accumulated within 500 km radius of each grid point on a 2.5° × 2.5° resolution grid based on the detected genesis locations (latitude and longitude). For each model, an annual genesis density map is generated and then averaged over the historical period (1985–2014) and the future SSP2-4.5 scenario period (2070–2099). Finally, for

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

both periods, the multi-model ensemble (MME) mean is computed by averaging the results of the 12 models (i.e., all models except MIROC-ES2H). The projected future changes are subsequently calculated as the difference between the future and historical MME means (i.e., SSP2-4.5 minus historical).

To determine whether the future spatial changes in TC genesis frequency can be physically explained by changes in large-scale environmental conditions, we employed the DGPI (Wang and Murakami, 2020). Recently, the DGPI has been widely utilized to diagnose future changes in TC activity under global warming (Liu et al., 2024; Wang et al., 2024; Murakami and Wang, 2022). By emphasizing robust large-scale dynamical constraints, the DGPI effectively minimizes the potential biases associated with thermodynamic parameters, which often overestimate TC genesis in warming scenarios, thereby showing better agreement with the TC changes directly detected from GCMs (Murakami and Wang, 2022; Liu et al., 2024). Therefore, a correspondence between the TC activity extracted by the TC^2 algorithm and the DGPI serves as robust evidence confirming the physical validity and reliability of the TC^2 algorithm. DGPI is defined in this study as follows:

$$DGPI = (2.0 + 0.1 \times V_s)^{-1.7} \left(5.5 - \frac{du_{500}}{dy} \times 10^5\right)^{2.3} (5.0 - 20 \times \omega_{500})^{3.4} (5.5 + |\zeta_{a850} \times 10^5|)^{2.4} e^{-11.8} - 1.0 \quad (1)$$

where V_s, $\frac{du_{500}}{dy}$, ω_{500}, and ζ_{a850} are the magnitude of vertical wind shear between 200 and 850 hPa ($m\ s^{-1}$), the meridional shear vorticity at 500 hPa (s^{-1}), the omega at 500 hPa ($Pa\ s^{-1}$), and the absolute vorticity at 850 hPa (s^{-1}), respectively (Murakami and Wang, 2022).

3 Verification and benchmark performance evaluation of the TC^2 algorithm

We verify that the algorithm is implemented correctly and behaves reproducibly, separately from its skill against observations. The tracker is deterministic by design: candidate centres are the grid maxima of the 850 hPa relative vorticity above a fixed threshold, tracks are formed by deterministic nearest-candidate linking within the mindist limit, and the classifier applies fixed, pre-trained models. Identical input therefore yields identical output, which we confirmed by running the full detection twice on the same input: the two runs produced byte-identical track files verified by checksum. We further checked that every storm in the released output satisfies, by construction, the four thresholds that define a TC^2 track — a 6-hourly time cadence, a 6-hourly centre displacement within mindist = 350 km, a genesis latitude within critical genesis latitude = 40°, and a duration of at least mindhr = 24 h. Across the entire released data set, no track violated any of these constraints, confirming that the released tracks are a correct and reproducible realization of the TC^2 v1.0 algorithm.

To evaluate the performance of the TC^2 algorithm with TempestExtremes and the OWZP scheme, we first assessed the performance of each scheme based on the confusion matrix defined in Section 2.3 (Table 6). The results clearly indicate that the TC^2 algorithm exhibits the most outstanding and balanced detection capabilities globally and across individual basins. Specifically, in terms of global metrics, TC^2 achieves the highest F1 score of 83.6%, and CSI of 71.8%, substantially outperforming both TempestExtremes (74.7% and 59.6%, respectively) and the OWZP scheme (67.0% and 50.3%, respectively). Notably, TempestExtremes records the highest global hit rate of 83.4%, which is slightly higher than the 81.5%

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

achieved by TC^2. However, the critical advantage of TC^2 lies in its exceptional ability to suppress false alarms while maintaining high sensitivity. While TempestExtremes suffers from a significantly high FAR of 32.4%, TC^2 maintains a remarkably lower FAR of 14.3%.

This superior and balanced performance of TC^2 is consistently observed across all individual ocean basins; TC^2 achieves the highest F1 scores and CSIs in all six major basins: Western North Pacific (WNP), Eastern North Pacific (ENP), North Atlantic (NA), North Indian Ocean (NI), South Indian Ocean (SI), and South Pacific (SP). This signifies that TC^2 successfully detects actual TCs while effectively minimizing the misclassification of non-TC systems regardless of the region. It is particularly noteworthy to evaluate model performance in traditionally challenging basins such as the NA, NI, and SP. As demonstrated in Table 6, the existing algorithms (TempestExtremes and OWZP) severely struggle in these regions. Although TC^2 is not entirely immune to these regional difficulties, exhibiting its lower F1 score and CSI in the NA, NI, and SP compared to the other basins, its overall performance remains remarkably robust. For instance, in the NI and SP, the conventional methods suffer from exceptionally high FAR (>46%) and significant F1 score degradation (dropping to the 46.8–64.9% range). In stark contrast, TC^2 successfully mitigates these issues, achieving a high F1 score of 80.0% in the NI and effectively suppressing false alarms to maintain a stable 76.8% even in the highly problematic SP. Furthermore, in the NA, TC^2 secures a robust F1 score of 81.7%. Ultimately, these results prove that TC^2's true strength lies in its consistent detection capability and false alarm suppression, successfully overcoming the regional limitations that cause conventional methods to fail.

Secondly, we assessed the methods' ability to capture TC tracks (Figure 2). TC^2 reproduces realistic TC tracks across all basins, performing comparably to or better than the existing methods in most respects. In particular, TC^2 effectively limits TC tracking into high latitudes, closely matching the observations, whereas TempestExtremes and OWZP track a substantial number of TCs propagating to higher latitudes. In the Northern Hemisphere, TC^2 tracks reach at most ~45°N, near the observed maximum (~49°N), while TempestExtremes and OWZP extend to ~70° and ~58°N, respectively. This high-latitude restriction is partly imposed by our filtering of TC candidates that move poleward of 45°, but it does not appear to be solely an artifact of that constraint. A clear indication comes from the Southern Hemisphere, where TC tracks in both the observations and TC^2 are largely confined equatorward of 40°S (with poleward maxima of ~39°S and ~33°S, respectively), whereas TempestExtremes and OWZP track TCs well beyond 40°S (to ~60° and ~48°S, respectively). Because this confinement occurs well equatorward of the 45° threshold, it cannot be attributed to that filter; rather, it indicates that TC^2 intrinsically captures the latitudinal limits of TC propagation.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

Table 6. Comparison of the performance of TC^2, TempestExtremes, and OWZP.

Region	Algorithms	Hit rate (%)	F1 score (%)	FAR (%)	CSI (%)
Globe	TC^2	81.5	83.6	14.3	71.8
	TempestExtremes	83.4	74.7	32.4	59.6
	OWZP	75,6	67.0	39.9	50.3
WNP	TC^2	87.5	85.5	16.3	74.7
	TempestExtremes	89.7	80.0	27.8	66.7
	OWZP	82.0	70.7	37.9	54.7
ENP	TC^2	83.9	85.7	12.5	74.9
	TempestExtremes	78.4	78.4	21.6	64.5
	OWZP	70.8	69.4	31.9	53.2
NA	TC^2	75.6	81.7	11.1	69.1
	TempestExtremes	74.2	72.9	28.4	57.3
	OWZP	62.8	66.0	30.4	49.3
NI	TC^2	79.8	80.0	19.9	66.6
	TempestExtremes	82.5	64.9	46.5	48.1
	OWZP	79.5	46.8	66.9	30.5
SP	TC^2	75.3	76.8	21.7	62.3
	TempestExtremes	84.7	60.5	52.9	43.4
	OWZP	80.7	63.6	47.6	46.6
SI	TC^2	80.9	85.0	10.5	73.9
	TempestExtremes	89.1	78.5	29.9	64.5
	OWZP	81.8	70.6	37.9	54.6

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

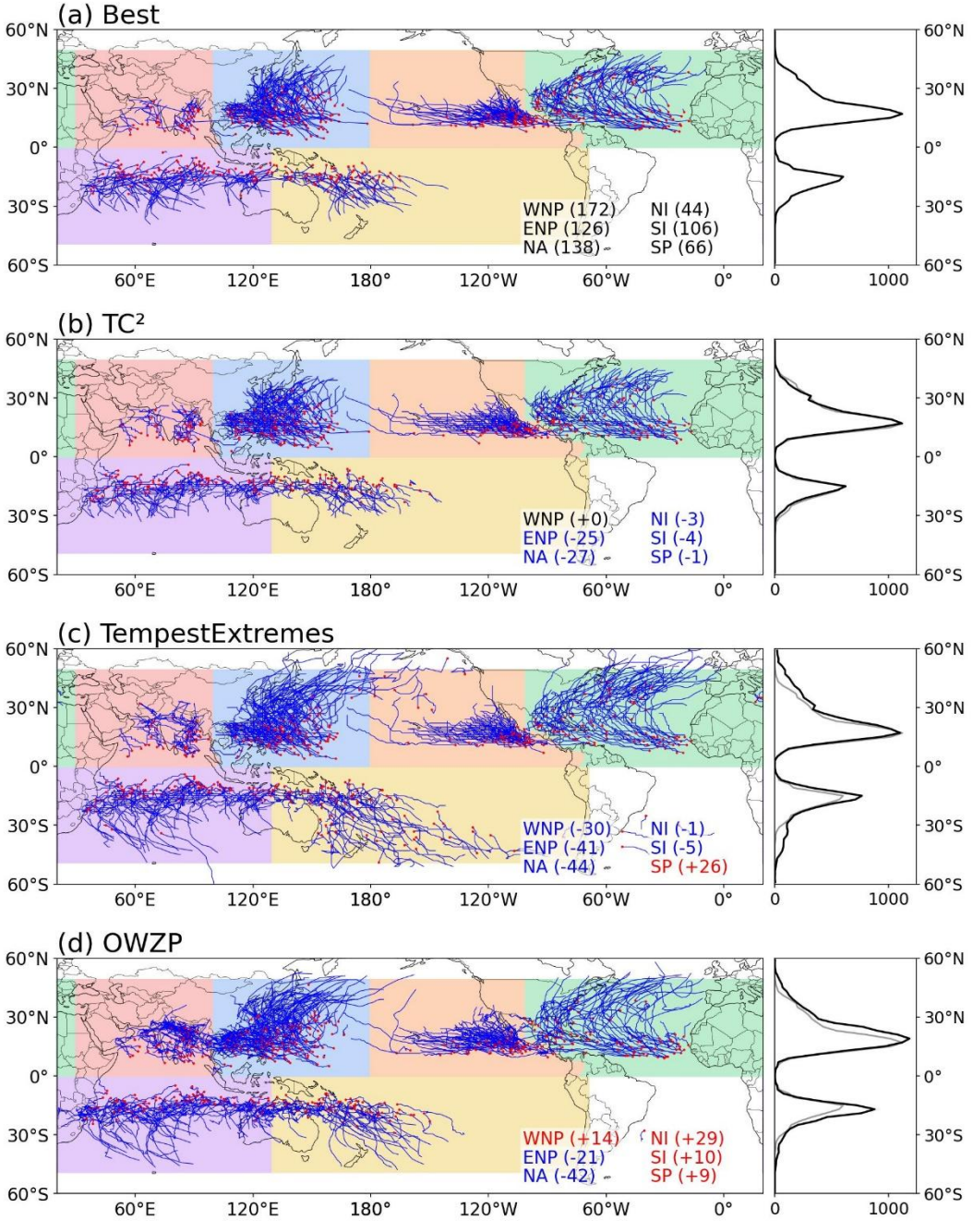

Figure 2. TC tracks over 2018–2024 from the best-track data (a) and as detected on FNL by (b) TC², (c) TempestExtremes, and (d) OWZP. Blue lines show the TC tracks and red dots their genesis points. Shaded areas denote the genesis basins: blue (WNP), orange (ENP), green (NA), red (NI), purple (SI), and yellow (SP). The number in parentheses next to each basin label represents the number of TCs generated in that basin in (a), and the difference from this value in (b) and (c), with blue and red indicating negative and positive differences, respectively. The curve to the right of each panel shows the latitudinal distribution of all track points (counts per 2° latitude band, on a common linear axis); in (b)–(d) the gray curve repeats the best-track distribution for reference.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

Based on the per-basin TC counts (Figure 2), TC^2 generally matches or outperforms TempestExtremes and OWZP, tracking the observed count most closely in four of the six basins (WNP, NA, SI, SP). In the WNP (172 TCs), TC^2 exactly reproduces the observed count, while OWZP slightly overestimates it (+14) and TempestExtremes substantially underestimates it (−30). All three schemes under-detect TCs in the ENP (126) and NA (138); TC^2 shows the smallest deficit in NA (−27, versus −44 for TempestExtremes and −42 for OWZP), while in ENP its deficit (−25) is comparable to OWZP (−21) and well below TempestExtremes (−41). In the NI (44 TCs), TempestExtremes (−1) and TC^2 (−3) closely match the observations, whereas OWZP markedly overestimates the frequency (+29). In the SP (66 TCs), TC^2 is nearly exact (−1), whereas TempestExtremes (+26) and OWZP (+9) overestimate genesis. In the SI (106 TCs), TC^2 (−4) and TempestExtremes (−5) agree closely with the observations, while OWZP overestimates (+10). Globally (excluding the South Atlantic), the observed total of 652 is reproduced almost exactly by OWZP (651) but only because compensating per-basin biases cancel, rather than through skill in any single basin; TC^2 (592) and TempestExtremes (557) both underestimate the global total.

Finally, we assessed the seasonal cycle of TC genesis frequency detected by the three algorithms (Fig. 3), quantified by the Spearman rank correlation and the normalized standard deviation (NSD) of the monthly genesis cycle relative to best-track data. Although all three schemes demonstrate generally strong performance, TC^2 exhibits the most consistent agreement with observations across all individual basins. In terms of seasonal phase, all three schemes capture the timing of the genesis cycle well: the Spearman rank correlations range from 0.84 to 0.98 for TempestExtremes and from 0.65 to 1.00 for OWZP, while those of TC^2 range from 0.89 to 1.00 (WNP: 0.96, ENP: 1.00, NA: 0.92, NI: 0.89, SP: 0.99, and SI: 0.98). Since a correlation coefficient of about 0.587 or higher is statistically significant at the 95% confidence level (n = 12), all three schemes yield statistically significant seasonal correlations in every ocean basin. The unique double-peak characteristic of TC genesis frequency in the NI is captured by all three schemes; however, TC^2 and TempestExtremes reproduce the later (autumn) peak more accurately than OWZP. In terms of seasonal amplitude, the NSDs of TC^2 range from 0.82 to 1.07, closest to the ideal value of 1.0, indicating that TC^2 best reproduces the amplitude of the seasonal cycle. By contrast, TempestExtremes tends to underestimate the amplitude (NSDs of 0.62 to 0.94), while OWZP shows larger and less consistent deviations (NSDs of 0.74 to 1.58), most notably an over-amplification in the NI. Notably, TC^2 is the only scheme that reproduces both the phase and the amplitude of the seasonal cycle in every basin.

Meanwhile, despite the high FAR of TempestExtremes and OWZP (Table 6), their seasonal genesis cycles closely follow the observations because the cycle is defined by genesis alone (Fig. 3). Their high FAR instead stems from track over-extension: TempestExtremes and OWZP track each TC ~44% and ~27% longer than the best-track data (105 h), respectively, whereas TC^2's track duration matches the observations (≈110 h). These excess points generate false alarms both at high latitudes (the poleward/extratropical tails in Fig. 2) and at lower latitudes (the weak phases before and after the stages exceeding TS intensity). Accordingly, properly bounding the TC life cycle, distinguishing these intense stages from weak or extratropical phases, is essential for reducing FAR.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

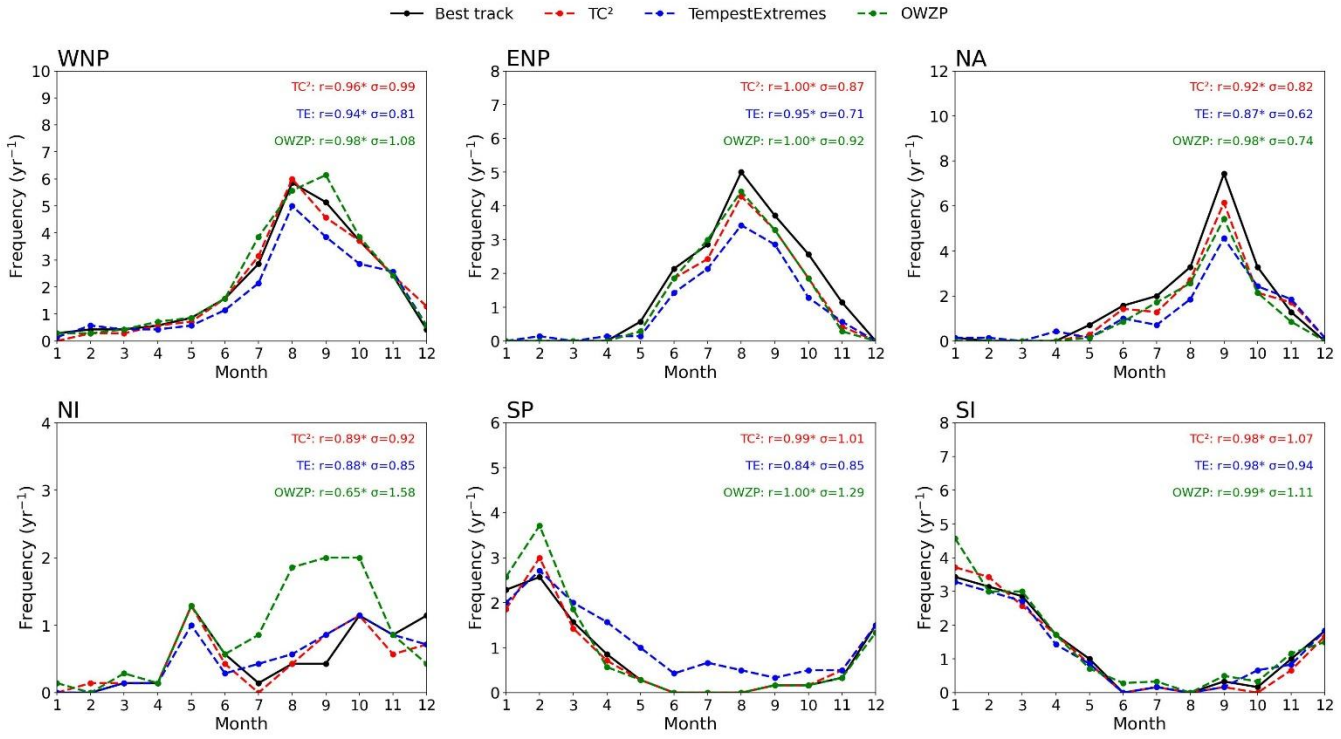

Figure 3. Seasonal variations of TC genesis frequency from the best-track data (black solid lines) and those detected by the TC^2 (red dashed lines), TempestExtremes (blue dashed lines), and OWZP (green dashed lines) schemes across six ocean basins (WNP, ENP, NA, NI, SP, and SI). r and σ denote the correlation coefficients and normalized standard deviations (NSDs), respectively. Asterisks indicate that the correlation coefficients are statistically significant at the 95% confidence level.

The consistent and often superior per-basin performance of TC^2 relative to OWZP and TempestExtremes indicates that data-driven, regionally adaptive criteria are more effective than fixed universal thresholds. Unlike the other two schemes, TC^2 maintains consistently high performance in every ocean basin and the skill to distinguish stages at or above TS intensity from weaker (sub-TS) or extratropical phases. This advantage is rooted in its design: TC^2 classifies candidates with a data-trained model that includes geographic location (latitude and longitude) among its predictors, so its effective detection criteria vary continuously with region, whereas TempestExtremes and OWZP apply the same thresholds everywhere. These results suggest that regionally adaptive, observation-based criteria are important for accurate TC detection, consistent with previous studies reporting that basin-dependent thresholds improve the representation of tracked-TC statistics (Camargo and Zebiak, 2002; Wu and Duan, 2023).

Although TC^2 was trained on ERA5 and JRA-3Q, it generally outperforms TempestExtremes and OWZP when applied to the NCEP FNL, a dataset of different origin and resolution. Two aspects make TC^2 less sensitive to the choice of dataset. First, as described in Section 2.2.1, its classifier predictors are spatially aggregated rather than grid-point quantities, so resolution-dependent extremes are smoothed. Second, as a data-driven ML method trained on multiple datasets, TC^2 is not overfitted to a single dataset, suggesting it generalizes well and is likely applicable to other datasets such as GCM output.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

4 Application to CMIP6

To further demonstrate the robustness and broad applicability of TC^2 to GCMs, we extended our evaluation to historical simulations from ten different CMIP6 models (Table 1). We first assessed the spatial distribution of TC tracks (Figure 4). When applied to the CMIP6 historical simulations, TC^2 effectively captures the overall track patterns. Nevertheless, several well-documented deficiencies in simulation of TC activity are evident, highlighting the inherent limitations of the climate models themselves.

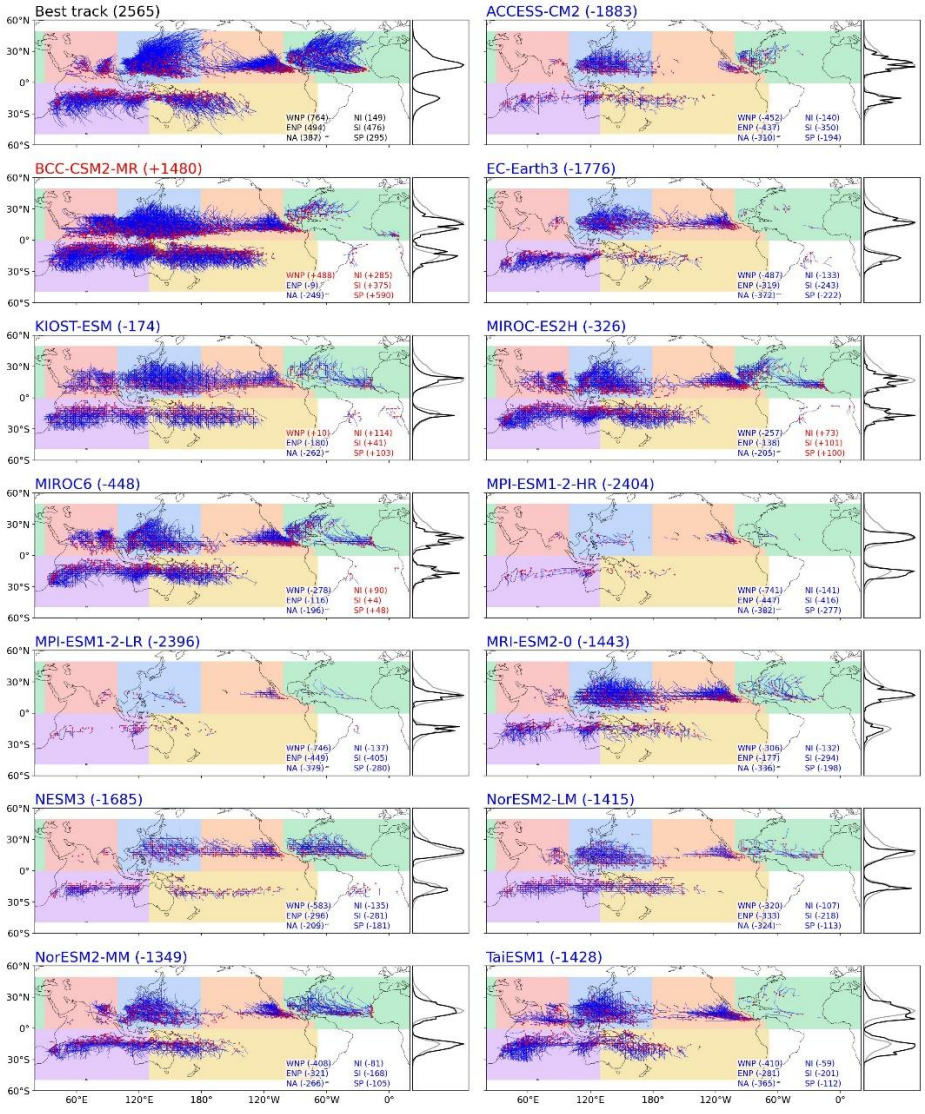

Figure 4. Same as Fig. 2, but for the 1985–2014 period and 13 GCMs. Each panel title indicates the data source and the absolute total TC count for the best-track data, and the difference from this value for the GCMs, with blue and red indicating negative and positive differences, respectively. Total counts are classified by genesis over the six shaded basins (the unshaded South Atlantic is excluded), and the same difference convention applies to the per-basin values at the lower right of each panel.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

Firstly, although the models underestimate TC frequency in most basins, the NA stands out as by far the most poorly captured (Figure 4). Averaged over the 13 models, the detected count relative to best-track data (1985–2014) is only ~23% in the NA, compared with 45–78% in the other basins (ENP 45%, WNP 55%, SI 67%, NI 74%, SP 78%). All 13 models under-detect NA TCs, with deficits of −196 to −382 against the observed 387, i.e., they reproduce only ~1–49% of the observed NA count. This pronounced tendency to underestimate TCs in the NA basin has been widely documented in previous studies (Camargo, 2013; Roberts et al., 2020; Manganello et al., 2012; Camargo et al., 2020). Furthermore, this negative bias has also been reported in tracking results based on reanalysis datasets (Bourdin et al., 2022; Raavi and Walsh, 2020). Our analysis using the NCEP FNL dataset also confirmed a persistent under-detection of TCs across all evaluated tracking schemes, including the TC^2 algorithm, in the NA basin (Figure 2). The difficulty in adequately capturing TCs in this region is attributed to the physical characteristics of NA storms, which tend to be relatively small in spatial scale and have shorter lifespans, as well as the models' poor representation of African easterly waves that serve as primary TC precursors (Tory et al., 2013b; Camargo, 2013; Tory et al., 2013a). Therefore, NA TCs are often poorly detected, particularly in low-resolution models, which is well supported by several studies demonstrating that increasing the horizontal resolution of models generally leads to improved detection and a more realistic representation of NA TCs (Manganello et al., 2012; Strachan et al., 2013; Roberts et al., 2020).

However, as emphasized by previous studies, horizontal resolution is not the sole factor; other model configurations, such as the dynamical core (Walsh et al., 2013; Reed and Jablonowski, 2012), the existence of ocean coupling (Li and Sriver, 2018; Zarzycki, 2016), and physical parameterizations (Zhao et al., 2012; Reed and Jablonowski, 2011), also play critical roles. Our results also suggest that this underestimation is not exclusively a resolution-dependent issue. Within the Max Planck Institute (MPI) model family, both versions exhibit a severe underestimation of TC frequencies. Specifically, the MPI-ESM1-2-LR and MPI-ESM1-2-HR models detected only 169 and 161 TCs globally during the 1985–2014 period, respectively, accounting for a mere 6–7% of the 2,565 observations. Notably, increasing the spatial resolution from LR to HR did not improve the detection capability; rather, it further reduced the count. Unlike the HR version, the LR incorporates dynamically computed vegetation and a fully equilibrated land–ocean carbon cycle, and features a longer, daily coupling period between system components, whereas the HR utilizes 1-hourly coupling (Mauritsen et al., 2019). In particular, the high-frequency ocean coupling effect can significantly impact TC simulations (Li and Sriver, 2018; Zarzycki, 2016). Higher coupling frequencies induce more pronounced ocean cooling feedbacks (Scoccimarro et al., 2017), which may contribute to the slightly weaker TC activity in the HR configuration relative to the LR. On the other hand, while there is currently insufficient evidence to conclude that dynamic vegetation significantly impacts TC activity, its potential role cannot be entirely ruled out and needs further investigation. As another example, the KIOST-ESM produced the second-highest global TC count despite having the second-coarsest resolution (Table 1), further demonstrating that a lower resolution does not necessarily result in fewer detected TCs.

Secondly, TCs propagating toward the mid-latitudes are heavily underrepresented in the simulations (Figure 4). As shown in the NCEP FNL evaluation (Figure 2), the TC^2 algorithm effectively captures TCs moving toward the mid-latitudes, which suggests that this underrepresentation in CMIP6 models more likely comes from an inherent limitation of the climate models

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

themselves rather than an algorithmic flaw. This aligns well with previous studies pointing out that most GCMs, regardless of whether they employ low or relatively high resolutions (e.g., 50 km), consistently struggle to simulate recurving tracks and TC propagation toward the mid-latitudes (Camargo and Zebiak, 2002; Bell et al., 2019; Manganello et al., 2012; Zhao et al., 2009). Based on known model behaviours documented in existing literature, we can deduce that this structural limitation likely stems from two factors: the inadequate representation of TC beta gyres under limited resolutions, which severely underestimates the poleward self-propagation (i.e., beta drift) of the simulated storms (Schenkel and Hart, 2012), and systematic biases in large-scale ambient steering flows, such as the position and strength of subtropical highs, which fail to properly guide TCs toward higher latitudes (Camargo, 2013; Yokoi et al., 2013).

Finally, in most models, such as BCC-CSM2-MR, EC-Earth3, KIOST-ESM, MIROC-ES2H, MIROC6, MPI-ESM1-2-LR, NESM3, and NorESM2-LM, TCs were detected in the South Atlantic, in which TCs are historically rare in observations but occasionally emerge. This was also repeatedly reported by several previous studies (Camargo, 2013; Tory et al., 2013a). The potential reasons for this overestimation are positive SST biases that lead to unrealistically high Genesis Potential Index values in the basin (Camargo, 2013), as well as the tendency of detection schemes to misclassify subtropical cyclones associated with the South Atlantic Convergence Zone as regular TCs (Hodges et al., 2017). Meanwhile, because TC^2 is trained only over the observation, this finding ironically alleviates the concern that TC^2 might never or hardly capture TCs in climatologically inactive regions, confirming its objective and robust detection capability across global oceans.

We evaluated the seasonal cycle of TC genesis frequency detected by TC^2 at the basin scale across the 13 GCMs (Figure 4). Although the models generally underestimate amplitude, the temporal correlation coefficients indicate that TC^2 reproduces the seasonal timing well in most basins: the multi-model ensemble mean (MME) attains a Spearman correlation with the best-track of 0.96–0.97 in every basin except the NI (0.50). To quantify the inter-model spread, we counted the models whose seasonal correlation is statistically significant at the 95% confidence level ($r \gtrsim 0.587$). All 13 models are significant in the ENP, NA, SP, and SI, and 12 of 13 in the WNP (only MPI-ESM1-2-HR is not). By contrast, only 5 of the 13 models are significant in the NI, and the MME itself is not ($r = 0.50$), making it the single basin where the seasonal timing is poorly captured. This degradation reflects well-documented limitations of GCMs and detection schemes: in the NI, models chronically struggle with the bimodal monsoon cycle, and detection schemes often make false alarms from monsoon depressions (Camargo et al., 2007; Manganello et al., 2012; Raavi and Walsh, 2020). Overall, apart from the structurally challenging NI basin, TC^2 consistently and accurately captures the regional seasonal timing of TC genesis across the GCMs, even in basins such as the NA, where the seasonal phase is well reproduced despite the strong underestimation of amplitude.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

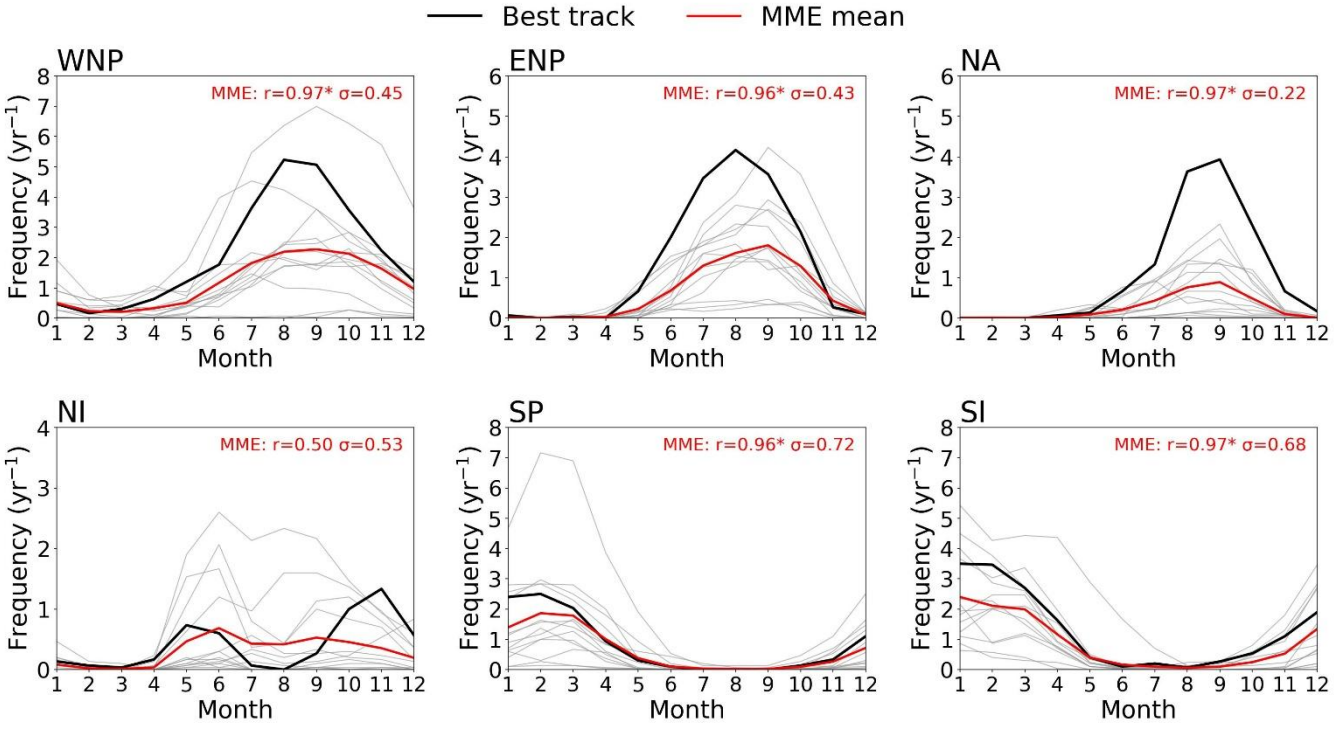

Figure 5. Seasonal variations of TC genesis frequency from the best-track data (black lines) and ensemble mean (red lines) and each ensemble (gray lines) of GCMs across six ocean basins. r and σ denote the correlation coefficients and normalized standard deviations (NSDs), respectively. Asterisks indicate that the correlation coefficients are statistically significant at the 95% confidence level.

As a final assessment, we investigated future spatial changes in TC genesis frequency and compared them with those of the DGPI under the SSP2-4.5 scenario (2070–2099 relative to the historical period 1985–2014; Figure 5). The changes are regionally mixed with a small net increase in the multi-model-mean global count (about +4%). In detail, the primary TC genesis locations exhibit distinct regional shifts, i.e., southwestward in the ENP, northwestward in the NA, and northeastward in the WNP, so that while activity decreases in the core of traditional main development regions, the area over which TCs can form expands to a wider region. The equatorward shift and overall decrease prevail in SP and SI, respectively. Remarkably, this spatial pattern of future change shares a consistent tendency with that of the DGPI: the pattern correlation (Pearson's r) between the genesis-density change and the DGPI change is $r \approx 0.36$ over $|\text{lat}| \leq 40°$ and increases to ≈ 0.41 in the deep tropics ($|\text{lat}| \leq 20°$), statistically significant at the 99% confidence level, even though the exact regional boundaries do not perfectly overlap. This correspondence signifies that the future TC shifts extracted by TC^2 are well explained by actual changes in the dynamic background environment. Consequently, it demonstrates that the TCs detected by TC^2 are not algorithmic artifacts but physically valid systems.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

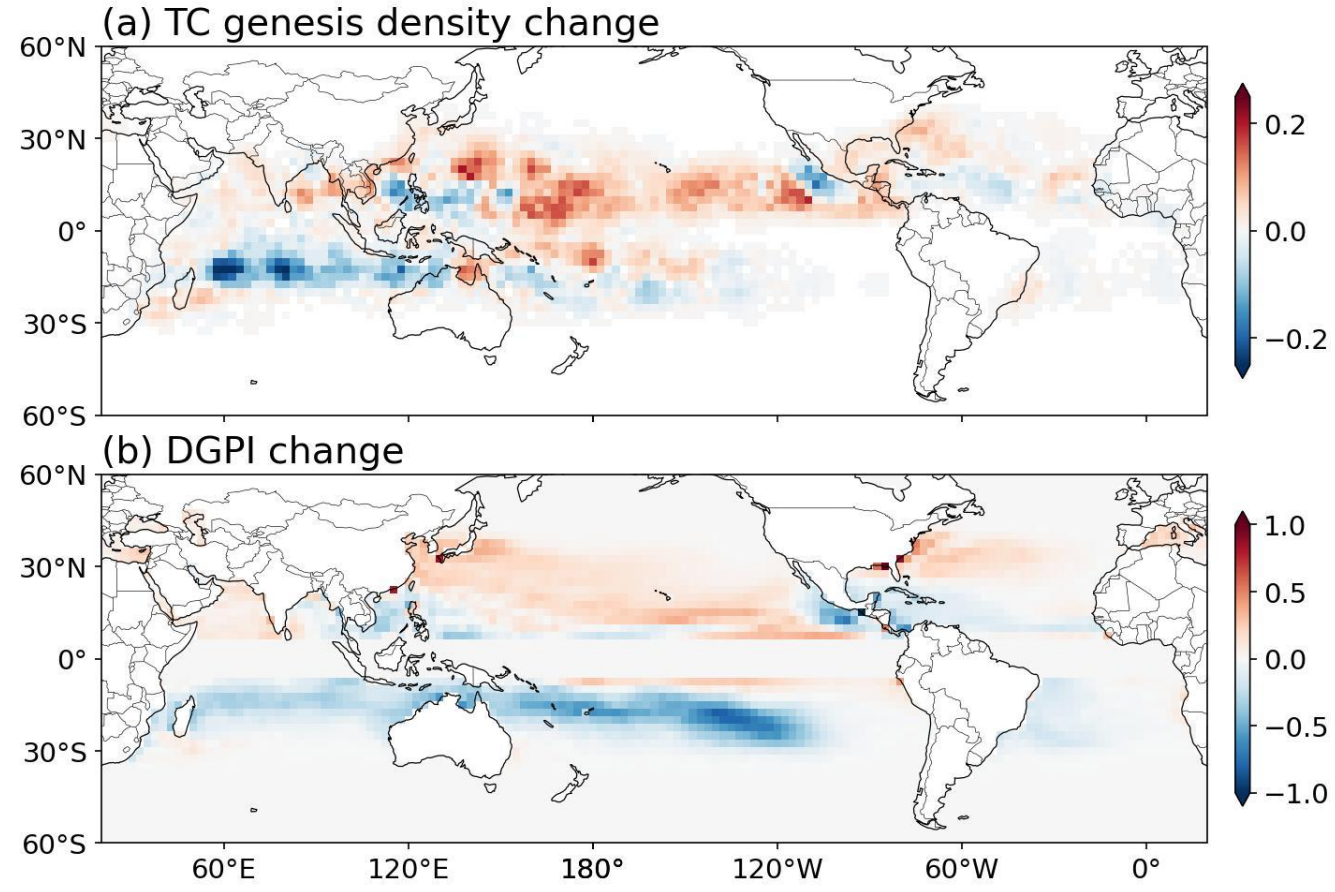

Figure 6. Differences of (a) TC genesis density and (b) DGPI between 2070-2099 and 1985-2014 under the SSP 2-4.5 scenario based on the 12-model MME (all models except MIROC-ES2H).

5 Discussion and Conclusion

In this study, to overcome the limitations of conventional and ML-based TC detection schemes, we developed the TC^2 algorithm, a hybrid TC detection algorithm that integrates a traditional tracking method with ML techniques. The conventional approach defines a TC by applying thresholds to a few representative variables (e.g., warm core anomalies, low-level vorticity, and MSLP). However, these thresholds have traditionally been optimized for specific basins and models. This raises a fundamental question, as the conventional methods used to define these tailored thresholds remain inherently subjective. To address this, basin- and model-independent schemes were introduced; nevertheless, such universal approaches may inadvertently exclude the unique environmental characteristics of individual basins. Most recently, ML-based detection schemes have been proposed to establish objective thresholds, but they often require a wider array of variables than traditional methods or suffer from a relatively high FAR. By integrating a traditional tracking method with ML techniques, TC^2 secures the objective threshold values and successfully reduces non-TC cases, and thereby significantly lowering FAR. TC^2 is an

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

ensemble of six classifiers trained on three different decision tree-based ML algorithms and two types of reanalysis datasets, ensuring that it is not overly dependent on any specific ML method or reanalysis data.

TC^2 shows overall robust performance. When applied to NCEP FNL data, TC^2 demonstrated superior performance compared to existing schemes evaluated in this study, i.e., TempestExtremes and OWZP, achieving higher F1 score (83.6%) and CSI (71.8%), along with a significantly lower FAR (14.3%). The hit rate was 81.5%, which is comparable to that of TempestExtremes. Furthermore, TC^2 captures the TC counts and seasonal variability over each basin more effectively than other tracking schemes. When applied to the CMIP6 historical simulations, TC^2 effectively captures the overall TC characteristics such as track patterns and seasonal cycles. Although TC^2 reflected the general underestimation of TC frequency and the poor representation of TC activity in the South Atlantic and NI basins in some models, these issues are well-known inherent limitations of GCMs, indicating that they are not algorithmic deficiencies of TC^2. Under the SSP2-4.5 scenario, future projections generally exhibited a slight increasing trend in global TC frequency. Regionally, the primary TC genesis locations displayed distinct spatial shifts: southwestward in the ENP, northwestward in the NA, and northeastward in the WNP. Remarkably, this spatial pattern of future changes shares a highly consistent spatial tendency with that of the DGPI, signifying that the future TC shifts extracted by TC^2 are well explained by actual changes in the dynamic background environment.

TC^2 is a highly efficient and reliable tool for TC detection because it requires only fundamental variables, such as MSLP, 850-hPa and 250-hPa zonal and meridional winds, 250- and 500-hPa temperatures, and land-sea mask, all of which are generally provided by CMIP simulations. In contrast, OWZP needs 950 (or 925), 850, 700, 500, and 200-hPa zonal and meridional winds, relative humidity, and temperature, together with topography. TempestExtremes requires MSLP, 300-hPa and 500-hPa geopotential heights, 10-m zonal and meridional winds, and topography. Although TC^2 uses far fewer variables than OWZP and a comparable set to TempestExtremes, TC^2 matches or surpasses TempestExtremes and OWZP, making it a computationally efficient yet highly capable tool for TC detection in both reanalysis and GCM datasets.

Code and data availability

The source code of TC² is available via GitHub at https://github.com/dsrpark/TC2-v1.0 under the MIT License. The specific version used in this study is archived on Zenodo (https://doi.org/10.5281/zenodo.20563573) (Park et al., 2026b). Pretrained model weights are separately archived on Zenodo (https://doi.org/10.5281/zenodo.20563586) (Park et al., 2026c). The TC tracks and verification data that reproduce the results of this study, the TC² final tracks for the FNL and 13 CMIP6 models (historical and SSP2-4.5), the TempestExtremes and OWZP tracks used for inter-comparison, the candidate tracks and script that reproduce the best-track miss-rate statistics, the genesis-density and DGPI grids, and the determinism and internal-consistency verification reports, are archived on Zenodo (https://doi.org/10.5281/zenodo.21185270) (Park et al., 2026a). File formats, a manifest with SHA-256 checksums, and instructions are included in that record.

The input data are third-party products, obtained from their respective archives and not redistributed here. CMIP6 model output was downloaded from the Earth System Grid Federation (https://esgf.llnl.gov/). The 6-hourly data from ERA5 data was

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

downloaded from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=download) (Hersbach et al., 2020). JRA3Q (Kosaka et al., 2024) and FNL (http://rda.ucar.edu/datasets/ds083.2/) data were downloaded from the Geoscience Data Exchange (https://gdex.ucar.edu/datasets/d640000/dataaccess/ and https://gdex.ucar.edu/datasets/d083002/dataaccess/), respectively. The best-track data were download from the JTWC (https://www.metoc.navy.mil/jtwc/jtwc.html?best-tracks) and NHC (https://www.nhc.noaa.gov/data/) websites, respectively.

Author contributions

DSRP conceptualized the study, designed the experiments, and supervised the entire research process. DSRP also prepared the schematics. The initial manuscript was drafted by DSRP and DK, with DK performing the preliminary algorithm testing. HSK developed the tracker code. HYK developed the classifier code, and conducted all calculations for the figures and tables, and generated the final visualizations. All authors contributed to the revision of the initial manuscript.

Competing interests

We declare there is no conflict of interest.

Acknowledgements

The authors would like to thank Dr. Mincheol Moon and Ms. Eunji Kim for their technical guidance on applying the OWZP and TempestExtremes. The authors used AI-assisted tools (Claude code, Codex, and Google Notebook LM) for computational assistance, figure plotting, and English language editing. The authors take full responsibility for the content of this manuscript.

Financial support

This study was supported by the Korea Meteorological Administration Research and Development Program (RS-2022-KM221312 and RS-2025-02309058), and the National Research Foundation of the Korean government (RS-2025-00556009 and RS-2026-25497720).

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

References

Accarino, G., Donno, D., Immorlano, F., Elia, D., and Aloisio, G.: An ensemble machine learning approach for tropical cyclone localization and tracking from ERA5 reanalysis data, Earth Space Sci., 10, e2023EA003106, 10.1029/2023EA003106, 2023.

Bell, S. S., Chand, S. S., Camargo, S. J., Tory, K. J., Turville, C., and Ye, H.: Western North Pacific tropical cyclone tracks in CMIP5 models: Statistical assessment using a model-independent detection and tracking scheme, J. Climate, 32, 7191–7208, 10.1175/JCLI-D-18-0785.1, 2019.

Bourdin, S., Fromang, S., Dulac, W., Cattiaux, J., and Chauvin, F.: Intercomparison of four algorithms for detecting tropical cyclones using ERA5, Geosci. Model Dev., 15, 6759–6786, 10.5194/gmd-15-6759-2022, 2022.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, 10.1023/A:1010933404324, 2001.

Camargo, S., Sobel, A. H., Barnston, A. G., and Emanuel, K. A.: Tropical cyclone genesis potential index in climate models, Tellus A, 59 A, 428–443, 10.1111/j.1600-0870.2007.00238.x, 2007.

Camargo, S. J.: Global and regional aspects of tropical cyclone activity in the CMIP5 models, J. Climate, 26, 9880–9902, 10.1175/jcli-d-12-00549.1, 2013.

Camargo, S. J. and Zebiak, S. E.: Improving the detection and tracking of tropical cyclones in atmospheric general circulation models, Wea. Forecasting, 17, 1152–1162, 10.1175/1520-0434(2002)017<1152:Itdato>2.0.Co;2, 2002.

Camargo, S. J., Giulivi, C. F., Sobel, A. H., Wing, A. A., Kim, D., Moon, Y., Strong, J. D. O., Del Genio, A. D., Kelley, M., Murakami, H., Reed, K. A., Scoccimarro, E., Vecchi, G. A., Wehner, M. F., Zarzycki, C., and Zhao, M.: Characteristics of model tropical cyclone climatology and the large-scale environment, J. Climate, 33, 4463–4487, 10.1175/jcli-d-19-0500.1, 2020.

Chen, D., Rojas, M., Samset, B. H., Cobb, K., Niang, A. D., Edwards, P., Emori, S., Faria, S. H., Hawkins, E., Hope, P., Huybrechts, P., Meinshausen, M., Mustafa, S. K., Plattner, G.-K., and Tréguier, A.-M.: Framing, Context, and Methods, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy,

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 147–286, 10.1017/9781009157896.003, 2021.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, The 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794, 10.1145/2939672.2939785,

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, 10.5194/gmd-9-1937-2016, 2016.

Galea, D., Hodges, K., and Lawrence, B. N.: Investigating differences between tropical cyclone detection systems, Artificial Intelligence for the Earth Systems, 3, e220046, 10.1175/AIES-D-22-0046.1, 2024.

Gardoll, S. and Boucher, O.: Classification of tropical cyclone containing images using a convolutional neural network: performance and sensitivity to the learning dataset, Geosci. Model Dev., 15, 7051–7073, 10.5194/gmd-15-7051-2022, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146, 1999–2049, 10.1002/qj.3803, 2020.

Hodges, K., Cobb, A., and Vidale, P. L.: How well are tropical cyclones represented in reanalysis datasets?, J. Climate, 30, 5243–5264, 10.1175/jcli-d-16-0557.1, 2017.

Horn, M., Walsh, K., Zhao, M., Camargo, S. J., Scoccimarro, E., Murakami, H., Wang, H., Ballinger, A., Kumar, A., Shaevitz, D. A., Jonas, J. A., and Oouchi, K.: Tracking scheme dependence of simulated tropical cyclone response to idealized climate simulations, J. Climate, 27, 9197–9213, 10.1175/JCLI-D-14-00200.1, 2014.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: a highly efficient gradient boosting decision tree, The 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 3149–3157,

Kosaka, Y., Kobayashi, S., Harada, Y., Kobayashi, C., Naoe, H., Yoshimoto, K., Harada, M., Goto, N., Chiba, J., Miyaoka, K., Sekiguchi, R., Deushi, M., Kamahori, H., Nakaegawa, T., Tanaka, T. Y., Tokuhiro, T., Sato, Y., Matsushita, Y., and Onogi, K.: The JRA-3Q Reanalysis, Journal of the Meteorological Society of Japan. Ser. II, 102, 49–109, 10.2151/jmsj.2024-004, 2024.

Kossin, J. P., Knapp, K. R., Vimont, D. J., Murnane, R. J., and Harper, B. A.: A globally consistent reanalysis of hurricane variability and trends, Geophys. Res. Lett., 34, L04815, 10.1029/2006gl028836, 2007.

Kuleshov, Y., Fawcett, R., Qi, L., Trewin, B., Jones, D., McBride, J., and Ramsay, H.: Trends in tropical cyclones in the South Indian Ocean and the South Pacific Ocean, J. Geophys. Res.-Atmos., 115, https://doi.org/10.1029/2009JD012372, 2010.

Li, H. and Sriver, R. L.: Tropical Cyclone Activity in the High-Resolution Community Earth System Model and the Impact of Ocean Coupling, J. Adv. Model. Earth Syst., 10, 165–186, https://doi.org/10.1002/2017MS001199, 2018.

Liu, C., An, S.-I., Zhao, J., Son, S.-W., Jin, F.-F., and Zhan, R.: Hemispheric asymmetric response of tropical cyclones to CO_2 emission reduction, npj Clim. Atmos. Sci., 7, 83, 10.1038/s41612-024-00632-2, 2024.

Manganello, J. V., Hodges, K. I., Kinter, J. L., Cash, B. A., Marx, L., Jung, T., Achuthavarier, D., Adams, J. M., Altshuler, E. L., Huang, B., Jin, E. K., Stan, C., Towers, P., and Wedi, N.: Tropical Cyclone Climatology in a 10-km Global Atmospheric GCM: Toward Weather-Resolving Climate Modeling, J. Climate, 25, 3867–3893, https://doi.org/10.1175/JCLI-D-11-00346.1, 2012.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-de-la-Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S.-S., Paulsen, H., Peters, K.,

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J.-S., Tian, F., Voigt, A., Vrese, P., Wieners, K.-H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2, J. Adv. Model. Earth Syst., 11, 998–1038, https://doi.org/10.1029/2018MS001400, 2019.

Murakami, H. and Wang, B.: Patterns and frequency of projected future tropical cyclone genesis are governed by dynamic effects, Commun. Earth Enviro., 3, 77, 10.1038/s43247-022-00410-z, 2022.

Nogueira: Bayesian optimization: Open source constrained global optimization tool for Python, GitHub repository, https://github. com/fmfn/BayesianOptimization, 2014.

Park, D.-S., Kim, D., Ko, H.-Y., Kim, H.-S., Cha, D.-H., Chang, M., Min, S.-K., and Park, T.-W.: Tropical cyclone tracks and verification data for TC² (Tropical Cyclone Tracker and Classifier) (v.1.0), Zenodo [code], 10.5281/zenodo.21185270, 2026a.

Park, D.-S., Kim, D., Ko, H.-Y., Kim, H.-S., Cha, D.-H., Chang, M., Min, S.-K., and Park, T.-W.: TC² (Tropical Cyclone Tracker and Classifier) (v1.0), Zenodo [code], 10.5281/zenodo.20563573, 2026b.

Park, D.-S., Kim, D., Ko, H.-Y., Kim, H.-S., Cha, D.-H., Chang, M., Min, S.-K., and Park, T.-W.: Pretrained model weights for TC² (Tropical Cyclone Tracker and Classifier) (v1.0), Zenodo [code], 10.5281/zenodo.20563587, 2026c.

Raavi, P. H. and Walsh, K. J. E.: Sensitivity of tropical cyclone formation to resolution-dependent and independent tracking schemes in high-resolution climate model simulations, Earth Space Sci., 7, e2019EA000906, 10.1029/2019EA000906, 2020.

Reed, K. A. and Jablonowski, C.: Impact of physical parameterizations on idealized tropical cyclones in the Community Atmosphere Model, Geophys. Res. Lett., 38, https://doi.org/10.1029/2010GL046297, 2011.

Reed, K. A. and Jablonowski, C.: Idealized tropical cyclone simulations of intermediate complexity: A test case for AGCMs, J. Adv. Model. Earth Syst., 4, https://doi.org/10.1029/2011MS000099, 2012.

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vannière, B., Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L.-P., Chauvin, F., Terray, L., Valcke, S., Moine, M.-P., Putrasahan, D., Roberts, C. D., Senan,

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

R., Zarzycki, C., Ullrich, P., Yamada, Y., Mizuta, R., Kodama, C., Fu, D., Zhang, Q., Danabasoglu, G., Rosenbloom, N., Wang, H., and Wu, L.: Projected future changes in tropical cyclones Using the CMIP6 HighResMIP multimodel ensemble, Geophys. Res. Lett., 47, e2020GL088662, 10.1029/2020GL088662, 2020.

Schenkel, B. A. and Hart, R. E.: An examination of tropical cyclone position, intensity, and intensity life cycle within atmospheric reanalysis datasets, J. Climate, 25, 3453–3475, 10.1175/2011JCLI4208.1, 2012.

Scoccimarro, E., Fogli, P. G., Reed, K. A., Gualdi, S., Masina, S., and Navarra, A.: Tropical Cyclone Interaction with the Ocean: The Role of High-Frequency (Subdaily) Coupled Processes, J. Climate, 30, 145–162, https://doi.org/10.1175/JCLI-D-16-0292.1, 2017.

Strachan, J., Vidale, P. L., Hodges, K., Roberts, M., and Demory, M.-E.: Investigating global tropical cyclone activity with a hierarchy of AGCMs: The role of model resolution, J. Climate, 26, 133–152, 10.1175/JCLI-D-12-00012.1, 2013.

Tory, K. J., Chand, S. S., Dare, R. A., and McBride, J. L.: An assessment of a model-, grid-, and basin-independent tropical cyclone detection scheme in selected CMIP3 global climate models, J. Climate, 26, 5508–5522, 10.1175/jcli-d-12-00511.1, 2013a.

Tory, K. J., Chand, S. S., Dare, R. A., and McBride, J. L.: The development and assessment of a model-, grid-, and basin-independent tropical cyclone detection scheme, J. Climate, 26, 5493–5507, 10.1175/jcli-d-12-00510.1, 2013b.

Tory, K. J., Chand, S. S., McBride, J. L., Ye, H., and Dare, R. A.: Projected Changes in Late-Twenty-First-Century Tropical Cyclone Frequency in 13 Coupled Climate Models from Phase 5 of the Coupled Model Intercomparison Project, J. Climate, 26, 9946–9959, https://doi.org/10.1175/JCLI-D-13-00010.1, 2013c.

Ullrich, P. A. and Zarzycki, C. M.: TempestExtremes: a framework for scale-insensitive pointwise feature tracking on unstructured grids, Geosci. Model Dev., 10, 1069–1090, 10.5194/gmd-10-1069-2017, 2017.

Ullrich, P. A., Zarzycki, C. M., McClenny, E. E., Pinheiro, M. C., Stansfield, A. M., and Reed, K. A.: TempestExtremes v2.1: a community framework for feature detection, tracking, and analysis in large datasets, Geosci. Model Dev., 14, 5023–5048, 10.5194/gmd-14-5023-2021, 2021.

Vaittinada Ayar, P., Bourdin, S., Faranda, D., and Vrac, M.: Ensemble random forest for tropical cyclone tracking, Nat. Hazards Earth Syst. Sci., 25, 4655–4672, 10.5194/nhess-25-4655-2025, 2025.

https://doi.org/10.5194/egusphere-2026-3285
Preprint. Discussion started: 9 July 2026

Walsh, K., Lavender, S., Scoccimarro, E., and Murakami, H.: Resolution dependence of tropical cyclone formation in CMIP3 and finer resolution models, Clim. Dyn., 40, 585–599, 10.1007/s00382-012-1298-z, 2013.

Walsh, K. J. E., Fiorino, M., Landsea, C. W., and McInnes, K. L.: Objectively determined resolution-dependent threshold criteria for the detection of tropical cyclones in climate models and reanalyses, J. Climate, 20, 2307–2314, 10.1175/jcli4074.1, 2007.

Wang, B. and Murakami, H.: Dynamic genesis potential index for diagnosing present-day and future global tropical cyclone genesis, Environ. Res. Lett., 15, 114008, 10.1088/1748-9326/abbb01, 2020.

Wang, P., Wang, C., Wu, L., Cao, J., and Zhao, H.: A dynamical downscaling framework for tropical cyclone activity over the western North Pacific, J. Geophys. Res.-Atmos., 129, e2024JD041946, 10.1029/2024JD041946, 2024.

Wu, T. and Duan, Z.: A new and efficient method for tropical cyclone detection and tracking in gridded datasets, Weather Clim. Extremes, 42, 100626, 10.1016/j.wace.2023.100626, 2023.

Yokoi, S., Takayabu, Y. N., and Murakami, H.: Attribution of projected future changes in tropical cyclone passage frequency over the western North Pacific, J. Climate, 26, 4096–4111, 10.1175/jcli-d-12-00218.1, 2013.

Zarzycki, C. M.: Tropical Cyclone Intensity Errors Associated with Lack of Two-Way Ocean Coupling in High-Resolution Global Simulations, J. Climate, 29, 8589–8610, https://doi.org/10.1175/JCLI-D-16-0273.1, 2016.

Zarzycki, C. M. and Ullrich, P. A.: Assessing sensitivities in algorithmic detection of tropical cyclones in climate data, Geophys. Res. Lett., 44, 1141–1149, 10.1002/2016GL071606, 2017.

Zhao, M., Held, I. M., and Lin, S.-J.: Some Counterintuitive Dependencies of Tropical Cyclone Frequency on Parameters in a GCM, J. Atmos. Sci., 69, 2272–2283, https://doi.org/10.1175/JAS-D-11-0238.1, 2012.

Zhao, M., Held, I. M., Lin, S.-J., and Vecchi, G. A.: Simulations of Global Hurricane Climatology, Interannual Variability, and Response to Global Warming Using a 50-km Resolution GCM, J. Climate, 22, 6653–6678, https://doi.org/10.1175/2009JCLI3049.1, 2009.