

TCBENCH: A BENCHMARK FOR TROPICAL CYCLONE TRACK AND INTENSITY FORECASTING AT THE GLOBAL SCALE

Anonymous authors

Paper under double-blind review

ABSTRACT

TCBench is a benchmark for evaluating global, short to medium-range (1-5 days) forecasts of tropical cyclone (TC) track and intensity. To allow a fair and model-agnostic comparison, TCBench builds on the IBTrACS observational dataset and formulates TC forecasting as predicting the time evolution of an existing tropical system conditioned on its initial position and intensity. TCBench includes state-of-the-art dynamical (TIGGE) and neural weather models (AIFS, Pangu-Weather, FourCastNet v2, GenCast). If not readily available, baseline tracks are consistently derived from model outputs using the TempestExtremes library. For evaluation, TCBench provides deterministic and probabilistic storm-following metrics. On 2023 test cases, neural weather models skillfully forecast TC tracks, while skillful intensity forecasts require additional steps such as post-processing. Designed for accessibility, TCBench helps AI practitioners tackle domain-relevant TC challenges and equips tropical meteorologists with data-driven tools and workflows to improve prediction and TC process understanding. By lowering barriers to reproducible, process-aware evaluation of extreme events, TCBench aims to democratize data-driven TC forecasting.

1 INTRODUCTION

Tropical cyclones (TCs), also called “hurricanes” or “typhoons” depending on the basin (American Meteorological Society, 2020), are globally devastating weather systems. In the U.S. alone, TCs caused over \$1.5 trillion in damages and over 7,000 deaths between 1980 and 2024 (NOAA National Centers for Environmental Information, 2025). With over half of the global population projected to live in the tropics and low-elevation coastal zones by 2050-2060 (Gu et al., 2021; Neumann et al., 2015), **improving TC forecasting is urgent** for risk mitigation and community resilience.

TC forecasting can be framed as the task of forecasting **two continuous time-series**:

Track: the time-series of storm locations in latitude $\phi \in [-50, 50]^\circ$ and longitude $\lambda \in [0, 360]^\circ$;

Intensity: commonly characterized by maximum sustained wind speed $V_{\max} \in \mathbb{R}_+$ (m s^{-1}) and minimum sea-level pressure $p_{\min} \in \mathbb{R}_+$ (Pa). Both variables are used operationally.

Operational TC forecasts traditionally rely on physics-based models that solve coupled partial differential equations to simulate global weather fields, with nested higher-resolution grids in the vicinity of the TC (Hazelton et al., 2023). However, we require supercomputing infrastructure to run these models at a spatial resolution that marginally resolves TCs (Davis, 2018) and this limits their accessibility. This limitation, combined with the availability of high-quality, open-sourced benchmark datasets (Rasp et al., 2024), has led to an increased interest in data-driven weather forecasting (Ebert-Uphoff & Hilburn, 2023).

Several “neural weather models” now outperform state-of-the-art physics models on the 1–10 day forecasting of select meteorological fields at a 0.25° resolution (Rasp, 2024). Compared to physics-based models, neural weather models—once trained—are computationally efficient and help democratize forecasting. Neural weather models are scalable and ensemble well, which is a desirable property in large forecasts to quantify and reduce uncertainty. Nonetheless, the performance of neural weather models is often summarized in metrics calculated over the global prediction and relatively less focus

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

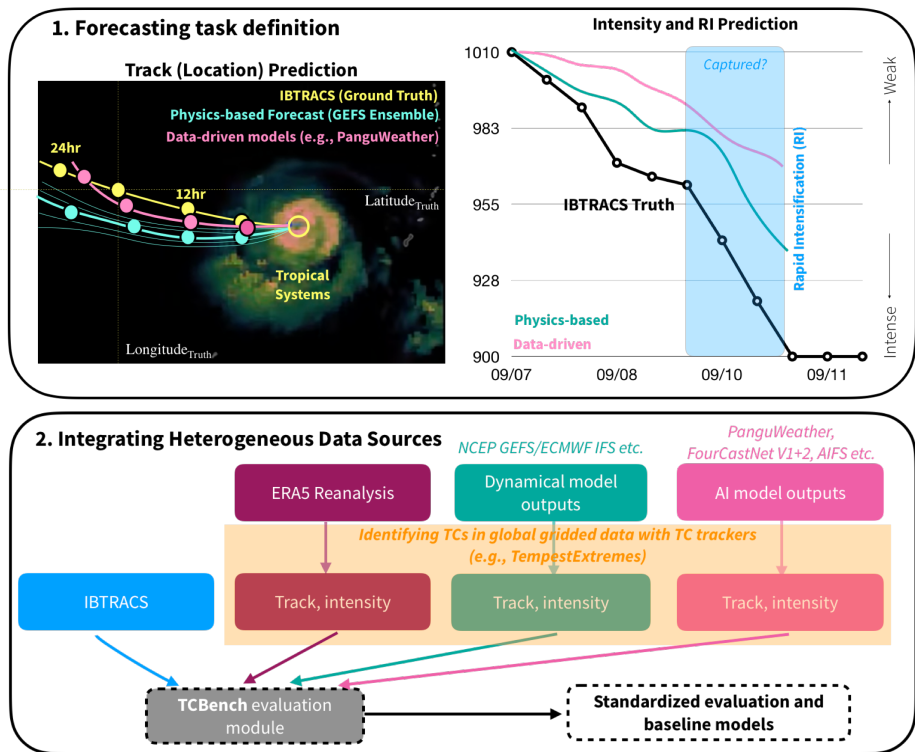


Figure 1: TC Bench defines TC forecasting as predicting time-series of track and intensity knowing the system’s initial state. It integrates heterogeneous data sources (observations, reanalysis, physics/data-driven models) into a unified evaluation framework to standardize model assessment.

on process-based assessments. Divergent approaches in evaluation complicates comparison between models. For example, PanguWeather (Bi et al., 2023) and FourCastNet (Pathak et al., 2022) rely on author-defined TC tracking algorithms, GraphCast (Lam et al., 2023) optimizes tracking for each model, and GenCast (Price et al., 2025) relies on a well-established tracking algorithm (Ullrich et al., 2021) but is forced to adapt it due to the frequency of their model outputs.

While TC track prediction from neural weather models now rivals (DeMaria et al., 2024) and sometimes outperforms physics-based models (Broad, 2024), TC intensity prediction remains a major challenge. The spatial resolution of physics-based global models is often too coarse to represent anomalies in wind speed, wind extremes, in TCs (DeMaria et al., 2014; Baker et al., 2024). Similarly, neural models are often forecast low intensity due to biases in their training data (Dulac et al., 2024). As a result, current data-driven intensity forecasts frequently under-perform simple baselines such as persistence and climatology (DeMaria et al., 2024).

Accurate modeling of TC intensity is of special importance, given that Rapid Intensification (RI) potential—defined as a large increase in intensity over a short period of time, typically around the 95th percentile of intensification in a 24h window—is frequently observed in the most destructive TCs. Yet, the predictive performance of neural weather models for RI has to our knowledge rarely been systematically assessed in previous literature. In this work, we treat RI as a **binary classification task**, which determines whether V_{\max} will increase by at least 30 kt ($\approx 15.4 \text{ m s}^{-1}$) over 24 hours (Kaplan & DeMaria, 2003; Kaplan et al., 2010), emphasizing that this roughly corresponds to the 95th percentile of the rate of intensity increase.

In this work, we present TCBench, a benchmark dataset that **bridges state-of-the-art neural and physics-based weather models with TC observational records to accelerate progress in data-driven forecasting of TCs**. This enables a transparent, reproducible assessment of how TC prediction can be further improved with machine learning. We demonstrate that neural weather models can skillfully forecast TC intensity up to 5 days ahead, particularly when combined with observational

108 data and processed through tools provided in TC Bench. Furthermore, we show that ensemble neural
109 weather models demonstrate TC track prediction skill that rivals traditional, physics-based ensembles.

110 TC Bench frames the problem of TC forecasting under the assumption that there is already a tropical
111 system that exists at an initial time of forecast, thereby focusing on predicting how the system will
112 evolve. TC Bench provides pre-processing pipelines, evaluation protocols, visualization tools, and
113 baselines from both physics- and neural-based models to benefit the atmospheric science and AI
114 communities. Designed for rapid iteration and anticipating fast innovation in data-driven forecasting,
115 TC Bench is extensible to new models, additional predictive targets, and encourages creative use of
116 the data’s spatiotemporal structure.

117 118 2 RELATED WORK

119
120 **Statistical models skillfully forecast TC intensity up to 5 days ahead when combined with**
121 **physics-based global atmospheric models, a strategy known as “statistical-dynamical” predic-**
122 **tion.** A prominent example is the Statistical Hurricane Intensity Prediction Scheme (SHIPS, (DeMaria
123 & Kaplan, 1994; Kaplan & DeMaria, 1999; DeMaria et al., 2005)), which uses a multiple regression
124 model with area-averaged climatological, persistence, satellite observations, and environmental
125 predictors from physics-based models to forecast TC intensity. Statistical-dynamical TC intensity
126 prediction models remain in use operationally due to their skill, particularly at longer forecast lead
127 times (Cangialosi et al., 2020). Statistical-dynamical models are also skillful at predicting RI, unlike
128 purely physics-based models (Torn & DeMaria, 2021). Statistical-dynamical methods for RI forecasts,
129 such as the SHIPS Rapid Intensification Index, or SHIPS-RII (Kaplan et al., 2010; 2015), use linear
130 discriminant analysis (Kaplan et al., 2010; Knaff et al., 2018), Bayesian and logistic regression
131 models (Rozoff & Kossin, 2011; Knaff et al., 2018), binomial logistic regression (DeMaria et al.,
132 2021), as well as consensus models (Kaplan et al., 2015).

133 In recent years, **machine learning models have begun tackling the problem of TC intensity**
134 **prediction.** Machine learning models trained on SHIPS predictors and satellite imagery skillfully
135 predict TC intensity and RI (Su et al., 2020; Griffin et al., 2022), as well as intensity and track errors
136 (Barnes et al., 2023; Fernandez et al., 2025). Convolutional neural networks (CNNs) and k-means
137 clustering better distinguish between environments that are favorable and not favorable to RI (Mercer
138 et al., 2021). Other approaches, such as decision tree-based models trained on SHIPS predictors
139 (Shaiba & Hahsler, 2016), the random Forest-based RI Scheme (FRISA) (Slocum, 2021; Sampson
140 et al., 2023), and the Long Short-Term Memory method (Yang et al., 2020) have also shown skill in RI
141 prediction. The advent of neural weather models for the global atmosphere opens new opportunities
142 for data-driven post-processing, analogous to statistical-dynamical prediction, with the **potential to**
143 **significantly extend the lead time for skillful TC forecasts.**

144 TC Bench targets the challenging problem of global TC forecasting, providing open,
145 real-time-available data for these notoriously difficult-to-predict phenomena in an accessible format
146 designed to accelerate improvement of data-driven models (Table 1). As key contributions from
147 TC Bench to the community, we include routines to deterministically and probabilistically evaluate
148 tracks predicted by models, neural weather model output data associated with our selected test
149 year, and the necessary inputs for the training and validating postprocessing baselines—facilitating
150 evaluation by community members of their own models and approaches with respect to the provided
151 baselines. Furthermore, our forecasting setup provides a straightforward comparison between TC
152 track prediction models to our best observational record, and does not limit comparisons to storms
153 predicted by all models. Finally, we as provide scores for baseline models and thereby streamline
154 comparisons between current and future models.

155 3 TC BENCH DATA TOOLBOX

156
157 To provide a clear test set and baselines ensuring fair inter-model comparison, TC Bench integrates a
158 diverse set of climate data with physics-based and neural weather model forecasts. A summary of
159 all data sources used in TC Bench is provided in Appendix C. These sources were selected based
160 on availability, spatial resolution, historical record length, and their relevance to forecasting or post-
161 processing TC intensity of tracks (Section 3.2). Figure 2 provides an example of some of the various
data sources included in TC Bench, including observed TC track and intensity data (Figure 2a,b),

Dataset	Forecast evaluation	Global coverage	Open source	TC-specific	Environ. predictors	AI-ready format
WeatherBench ² ¹	✓	✓	✓	✗	✓	✓
ChaosBench ²	✓	✓	✓	✗	✓	✓
SHIPS ³	✗	✗	✗	✓	✓	✗
IBTrACS ⁴ (“ground truth”)	✗	✓	✓	✓	✗	✗
TC-PRIMED ⁵	✗	✓	✓	✓	✓	✓
TropiCycloneNet ⁶	✗	✓	✓	✓	✓	✓
TCBench	✓	✓	✓	✓	✓	✓

Table 1: Comparison of datasets and benchmarks across key criteria: forecast evaluation support, coverage, openness, TC specificity, environmental predictors, and ML readiness.

¹ Rasp et al. (2024); ² Nathaniel et al. (2024); ³ DeMaria et al. (2022); ⁴ Knapp et al. (2010); ⁵ Razin et al. (2023); ⁶ Huang et al. (2025).

reanalysis (Figure 2c), and neural weather model forecasts (Figure 2d). In addition to the data we provide, we set forth guidelines for community members to submit data/and or models for their evaluation and listing on the leaderboard.

3.1 DATA PROVIDED WITH TCBENCH

The International Best Track Archive for Climate Stewardship (IBTrACS) serves as the “ground truth” for model evaluation. IBTrACS is the most complete observational archive of global TCs that is currently available (Knapp et al., 2010; Gahtan et al., 2024). IBTrACS provides global coverage of TC tracks, including various parameters, such as location, intensity, and size. For consistency, we retain only the 6-hourly time stamps (00:00, 06:00, 12:00, 18:00 UTC). We note that the definition of TC intensity is inconsistent across different meteorological agencies (Schreck III et al., 2014). In TCBench, we use the U.S. agencies’ definition: minimum sea-level pressure and 1-minute maximum

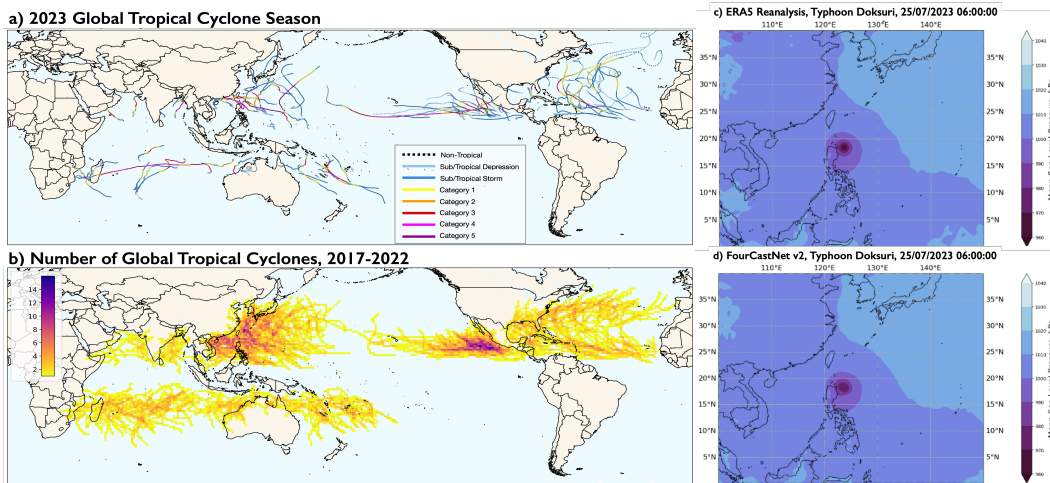


Figure 2: (a) 2023 tropical cyclones in the TCBench test year, from IBTrACS. The lines represent the position of each tropical cyclone over time, with the line color representing the storm’s intensity at that position. (b) IBTrACS estimate of tropical cyclone numbers from 2017-2022 (corresponding to TCBench’s training and validation years). Tropical cyclone counts binned into a 1° latitude by 1° longitude grid. (c) Mean sea level pressure from ERA5 reanalysis during Typhoon Doksuri, 25/07/2023 06:00:00 UTC. Represents the ground truth targeted by neural weather models, as opposed to the best record of TC intensity (IBTrACS) (d) 6-hour forecast of sea level pressure from the FourCastNet-v2 neural weather model during Typhoon Doksuri, valid at 25/07/2023 06:00:00 UTC. Panels a) and b) were made using the troPYcal software package (Burg & Lillo, 2019).

sustained wind speed at an altitude of 10 meters, both of which are provided in IBTrACS. Another source of uncertainty and inconsistency across agencies and time periods is the first data point included in each TC track in the IBTrACS dataset, e.g. there is no specific intensity for when the agencies start the track of a specific TC. We provide tools for processing the IBTrACS file provided by NOAA and provide the subsection of data of interest per identified TC.

ERA5 (Hersbach et al., 2020) is the 5th generation reanalysis product provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 provides dozens of meteorological variables at high spatial, vertical, and temporal resolutions, and models or assimilates a large amount of historical data. Though ERA5 is used to provide initial conditions to physical and neural weather models, we do not provide this data as it is made available by the ECMWF through its Copernicus programme.

Physics-based models predict the weather by solving partial differential equations, such as the laws of fluid dynamics, thermodynamics, radiative transfer, and atmospheric chemistry. We include hindcasts from several of these models via The International Grand Global Ensemble (TIGGE) product, including ensemble hindcasts of the Global Ensemble Forecasting System (the GEFS) and the International Forecast System (the IFS)(Bougeault & Coauthors, 2010). In contrast, **neural weather models** make weather forecasts using entirely data driven methods. These methods vary by model but include graph neural networks (Lam et al., 2023), Fourier Neural Operators (Pathak et al., 2022), and vision transformers (Bi et al., 2023). We include hindcasts from several neural weather models, including NVIDIA’s FourCastNetv2 (Bonev et al., 2023), Huawei’s Pangu-Weather (Bi et al., 2023), Google DeepMind’s GenCast (Price et al., 2025), and the ECMWF’s single-member AIFS v1.0 (Lang et al., 2024).

With regards to **data processing**, all datasets were reformatted into a unified structure based on IBTrACS for consistency across sources, using the IBTrACS storm identifier for identifying TCs and following unit conventions in IBTrACS. For the physics based models, track data (position, intensity) from international weather prediction centers was extracted from .xml files and standardized to the format. For neural weather model forecasts, forecast tracks and intensities were parsed from model outputs and matched to IBTrACS using TempestExtremes (Ullrich et al., 2021) and HuracanPy (Bourdin & Saffin, 2025) and then standardized to the format. The postprocessing models we provide predict intensification via a parametric distribution, and we thus sample in order to follow the standardized. Additional details for the postprocessing models are provided in Appendix D.4.

3.2 DATA INCLUSION CRITERIA

To ensure consistency and usefulness across all experiments, each forecast or model dataset included in TCbench had to meet the following requirements: a) They must provide at least 2 forecast initializations per day (00z, 12z); b) They must provide forecasts at least through a 5-day lead time, or forecasts until the IBTrACS record for that TC ends; c) the forecasts must be available as 6-hourly forecast data; d) the forecasts must provide at least the latitude, longitude, minimum pressure, and maximum wind speed track variables; e) models must provide any accompanying environmental fields (e.g., surface wind speed, surface pressure, geopotential height) used for tracking and intensity estimation; f) if tracking storms in environmental fields, a documented tracking algorithm and reproducible code examples; and g) if not directly forecasting track and intensity, the provided fields must have a minimum resolution of 0.5° , preferably 0.25° . These criteria are guidelines for both the data we provide, and data provided by community members in the future.

3.3 GUIDELINES FOR MODEL SUBMISSION

Dataset Split Recommendation: IBTrACS storms are divided based on the year they occurred to prevent leakage, noting that we divide based on calendar year and not TC season. We suggest at least 4 years be used for training and at least 2 years be used for validation, noting that we assign 2017-2020 to training and 2021-2022 to validation. **We require that the year 2023 be left for testing**, given that the models we include as baselines are not trained or tuned on this year.

Model Submission Guidelines: TCbench submissions must include: a) a list of open data sources used for model training, including any relevant links; b) any additional data needed to run the model on the test set; c) the minimal data sources needed to train the baselines presented for the training

and validation years, d) code and hyperparameters to train and run the model and generate the predictions, e) the model and/or trained weights in a standard format (e.g., a pickled file – .pkl), f) and an environment file to recreate the computing environment (or a container with which to run the model). All data will be added to the GitHub and HuggingFace repositories as appropriate.

4 TC BENCH BENCHMARKING TOOLBOX

TCBench provides a standardized set of metrics to provide objective comparisons for TC forecasting.

Intensity: Intensity forecasts are evaluated using the root mean square error (RMSE) and the mean absolute error (MAE).

Track Error Metrics: To evaluate the quality of predicted storm tracks, a set of deterministic error metrics is computed following the methodology described by Heming (Heming, 2017). These include:

Direct Positional Error (DPE): The straight-line distance between the forecast and observed storm positions at the same verification time.

Cross-Track Error (CTE): The component of DPE that lies perpendicular to the observed storm motion, indicating lateral displacement.

Along-Track Error (ATE): The component of DPE that lies along the observed storm motion, indicating whether the forecast storm is ahead or behind in time.

A complete description of the formulae and computational procedures is provided in Appendix E.

Probabilistic Evaluation: To quantify uncertainty in ensemble forecasts, we report the fair variant of the continuous ranked probability score (CRPS). Given an ensemble of N forecast values $\{x_i\}_{i=1}^N$ and an observation y , the fair CRPS provides an unbiased estimation of the discrepancy between the forecast distribution and the observed outcome:

$$\text{CRPS}(\{x_i\}, y) = \frac{1}{N} \sum_{i=1}^N |x_i - y| - \frac{1}{2N(N-1)} \sum_{i=1}^N \sum_{j=1}^N |x_i - x_j|. \quad (1)$$

The first term captures ensemble accuracy (mean absolute error), while the second reflects ensemble sharpness (spread). Lower CRPS indicates better-calibrated forecasts centered on the observation. We define **Track CRPS** for TC tracks by replacing absolute differences with Haversine distances between predicted and observed storm center positions.

Rapid Intensification Evaluation: Though Rapid intensification (RI) is a rare event, it is not exceedingly so. We define it as approximately the 95th percentile of intensity change over a 24-hour window. Thus, we evaluate model performance on RI forecasts using metrics more tailored for rare event forecasts. We note that while most of TCBench is configured as a **regression** problem, we have formulated RI as a **binary classification problem**. Thus, rapid intensification models are tasked with making a simple “yes/no” prediction for the occurrence of rapid intensification (i.e., the presence of an intensification of 30 knots per 24h window, where the window is rolled over the 120h forecasts). For these metrics, TP refers to true positives (hits); TN refers to true negatives, or correct negatives; FP refers to false positives, or false alarms; and FN refers to false negatives, or misses.

Critical Success Index (CSI): Also known as the Threat Score, measures the ratio of correctly predicted positive observations to the sum of all predicted positives, actual positives, and minus true positives. CSI can be used both for probabilistic track evaluation and RI.

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}},$$

5 BASELINES

Following standard evaluation procedures (Knaff et al., 2003), we first compare TC forecasts to the Mean Tendency by Lead & Basin (MT-LB) and persistence baselines. MT-LB samples the empirical

Closed Contour Command Parameters				Stitch Nodes Parameters	
Variable	Delta	Dist. (°)	Min/Max Dist. (°)	Parameter	Value
MSLP	200 Pa	6.5	0	Minimum Track Duration	12h
$\Delta z_{300,500}$	$-58.8 \text{ m}^2\text{s}^{-2}$	5.5	1.0	V_{\max} Threshold	10 m/s
				Latitude (ϕ)	$ \phi \leq 50^\circ$

Table 2: TempestExtremes Parameters. Distances are expressed in great circle degrees.

distribution of IBTrACS targets (intensity and position changes relative to initial conditions) from 1980 to 2022, conditioned on basin and lead time. For Persistence, at each lead time L the forecast equals the initial state at t_0 (i.e., we predict neither motion nor a change in intensity)

Baselines for Track and Intensity Predictions: We provide scores for tracks obtained directly from DeepMind’s Weather Lab website for GenCast and Weatherlab FNv3, and for the ECMWF IFS ensemble and the NCEP GEFS models using tracks from the TIGGE dataset. For PanguWeather, FourcastNetv2, and the single-member version of AIFS, we process global outputs to obtain tracks using TempestExtremes (Ullrich et al., 2021). The parameters we used are listed in Table 2, and are the same as in Ullrich et al. (2021), with the exception of the minimum track duration, which is set to 12h. While this would normally lead to overly sensitive tracking (i.e., resulting in less filtering of short-lived systems that are not true cyclones), we frame the problem of TC prediction with the assumption that there is a prior system of interest known to exist at an initial time and we are interested in the evolution of said tropical system. We thus eliminate all spurious tracks that could not be matched to the systems of interest. We generate two sets of tracks to evaluate - one in which there is a fall back to the persistence baseline whenever the model does not produce a forecast for a storm (e.g., because the storm does not exist in the model outputs) and one that includes only the storms predicted by the models. Our main results rely on the former, while plots associated with the latter can be found in Appendix F.

We provide additional baselines that post-process data-driven weather model forecasts to predict tropical cyclone intensity, using a pipeline that follows (Gomez et al., 2025). These baselines provide only an intensity forecast, and a more thorough description of these baselines is given in Appendix D.4. Finally, we use all baselines for intensity prediction to evaluate for rapid intensification (RI) by thresholding the wind speed change—i.e., if the 24 hour change in wind speed is at least 30 kts, RI has occurred.

6 RESULTS

In this section we present examples of using TC Bench to evaluate the deterministic and probabilistic skills of neural weather models (AI Models) and physics-based models on TC track and intensity predictions, covering lead times of up to five days. The results are summarized in Figures 3 and 4.

Some AI Models perform well for tracks: As shown in Figure 3, track prediction skills deteriorate with increasing lead time. All evaluated models exhibit track errors that are smaller than the persistence error at all lead times, indicating that all models produce useful track prediction. This is especially apparent when looking at the error for the predicted tracks without a fallback to persistence, shown by Figure 6 in Appendix F. While the track error for most AI models and physics-based models are very similar across lead times, we see a clear indication of the AIFS model to outperform other models in deterministic metrics beyond 24 hours. We largely attribute this to AIFS producing TC track predictions for a greater number of time steps (see Figure 7 in Appendix F), which greatly reduces the number of samples for which the evaluation falls back to the persistence baseline. Regarding the quality of model track ensembles, the physics-based GEFS model still outperforms the AI model alternatives we evaluated with regards to track displacement CRPS. These results highlight the potential of some existing data-driven forecasting models to provide **reliable deterministic** TC track forecasts, but suggest that they are yet to be as good as physics-based ensemble models in probabilistic track predictions.

Postprocessing AI models yields skillful intensity predictions: The coarse spatial resolution of the physics-based global weather prediction models and the underlying training data for the AI models mean that neither type of model is likely to perform well for intensity predictions. In fact, all models produce forecasts of V_{\max} and p_{\min} that are worse than the persistence baseline for shorter lead times but that outperform it for longer lead times (fig. 3). We see that the lead time at which the intensity predictions become useful varies between 24 hours and 48 hours, and that the physics-based GEFS model tends to perform best for both V_{\max} and p_{\min} . However, the difference in deterministic and probabilistic skills between the best raw AI model and GEFS is slight. Furthermore, we do not discount the possibility that a high spatial resolution, non-global physics-based model can be dramatically better than both GEFS and the AI models at predicting TC intensity. Finally, a well-designed post-processing algorithm of raw AI predictions can improve upon an originally inaccurate AI model (e.g., PanguWeather) to the extent of producing predictions on par with or better than the GEFS model.

Only the postprocessed AI models capture RI: The challenges of predicting TC RI events with existing global AI models and physics-based models can clearly be observed in Fig. 4, where most models show little to no ability to forecast such events across different lead times. Of the models we evaluated, only the post-processed PanguWeather model shows any skill at RI prediction, with some success in detecting RI for lead times between 48 and 96 hours. We note that neural weather models have generally been trained on ERA5 reanalysis, which is known to have a negative bias with regards to TC intensity. This issue, combined with the fact that many models have been trained to reduced a mean error (e.g., MAE, RMSE) and the relatively coarse resolution of the predicted fields, makes predictions of rapid intensification by the models a significantly more challenging task.

Overall, our evaluation reveals that:

- Some AI models outperform GEFS in deterministic track prediction, but not probabilistic ones.
- GEFS provides slightly better deterministic and probabilistic intensity forecasts than AI models.
- Intensity forecasts from AI models can be improved with a postprocessing pipeline.

These findings emphasize the complementary strengths of neural and physics-based systems, and point toward hybrid approaches as a promising future direction in TC forecasting. These findings also agree with (DeMaria et al., 2024; Sahu et al., 2025), who found that neural weather models could make skillful TC track predictions, but lacked skill in intensity forecasting. The results from the post-processed neural models are encouraging and point the way towards making data-driven TC intensity forecasts more reliable.

7 DISCUSSION AND CONCLUSION

Other Potential Applications: In addition to the simple deterministic and probabilistic TC track and intensity predictions presented above, the TCBenchmark dataset can be applied to other scientific applications related to different aspects of tropical cyclones. These include assessments of wind-related risks posed by tropical cyclones, tropical cyclone wind and precipitation nowcasting, wind field reconstruction, and data aggregation for physics discovery. Additional description of these potential applications is provided in Appendix G.

Limitations: There are clear limitations regarding this work, including the variance and evolution of uncertainty in the data, a limited selection of sources for tractability, and TC track inconsistencies across agencies. We acknowledge these and expand on them further in Appendix H.

Conclusion: By emphasizing flexibility in data experimentation, TCBenchmark provides a data foundation that can be extended with advancements in TC observations and modeling. TCBenchmark’s goal is to help the research community improve TC forecasts, our understanding of the physical processes governing TC behavior, ultimately mitigating TC impacts on local communities and society at large. We additionally provide a common point of comparison between TC forecasting models that is centered around the prediction of TC behavior once systems have formed. We show that for this task there is ample space for further research and improvement beyond our current data-driven forecasting capabilities, and provide tools for helping researchers in this field.

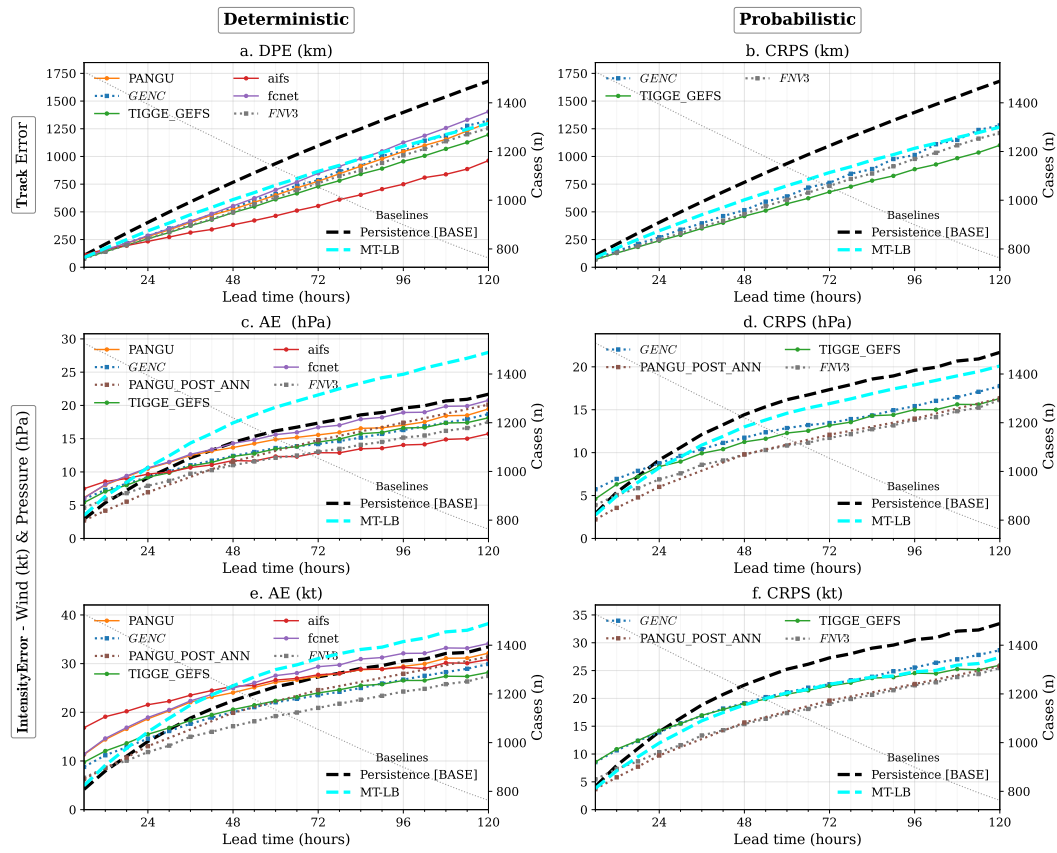


Figure 3: **FAIR per-lead comparison on TC Bench-2023.** Deterministic (left) and probabilistic (right) scores from 6–120 h for: (a) DPE, (b) CRPS–track, (c) AE–pressure, (d) CRPS–pressure, (e) AE–wind, (f) CRPS–wind. Means are computed on IBTrACS verification keys (00/12Z inits), with missing entries filled via persistence for fair comparison. Baselines: persistence (black dashed) and **MT-LB** (cyan dashed; mean tendency by lead & basin from IBTrACS, 1980–2022). For CRPS, persistence equals the MAE delta forecast. *GENC* and *FNV3* are italicized because we use the providers’ tracks (not re-derived). The post-processing model (dotted; e.g., *PANGU_POST_ANN*) deviates from our protocol (non-6 h lead grid; trained/validated on different years). Right axes show IBTrACS case counts. See SI for the corresponding raw (non-filled) comparison.

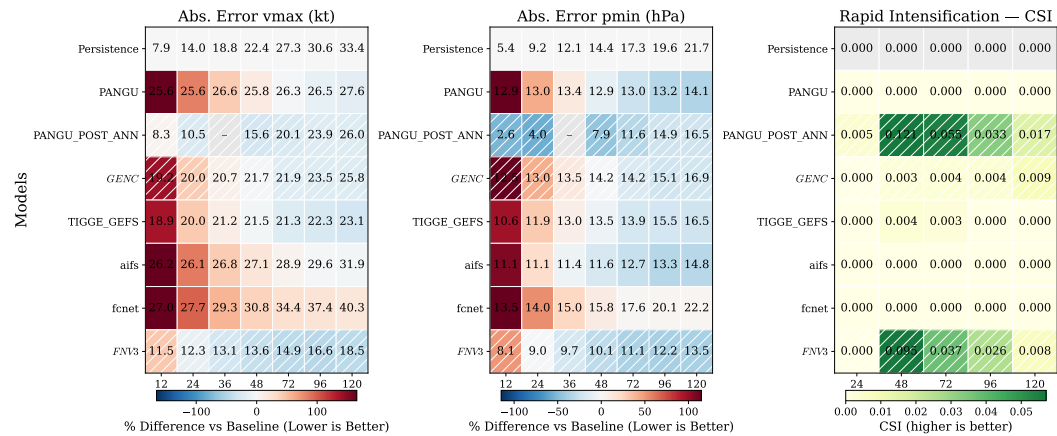


Figure 4: TC Bench deterministic scorecard (2023, baseline = Persistence). Cells show mean error; colors show % difference vs baseline at the same lead (12–120h; columns) for each model (rows).

8 LLM USAGE DISCLAIMER

Conversational AI tools (e.g., ChatGPT and Copilot) were used for the following tasks during the manuscript’s preparation: text editing (specifically, in making certain sections more succinct), latex formatting, and as an aid in code development when writing the benchmark framework.

REFERENCES

- Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. Machine learning for precipitation nowcasting from radar images. *arXiv preprint arXiv:1912.12132*, 2019.
- Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, et al. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025.
- American Meteorological Society. Tropical cyclone. Glossary of Meteorology, 2020. http://glossary.ametsoc.org/wiki/Tropical_cyclone.
- Alexander J. Baker, Benoît Vannière, and Pier Luigi Vidale. On the realism of tropical cyclone intensification in global storm-resolving climate models. *Geophysical Research Letters*, 51(17): e2024GL109841, 2024. doi: <https://doi.org/10.1029/2024GL109841>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024GL109841>. e2024GL109841
- E.A. Barnes, R.J. Barnes, and M. DeMaria. Sinh-arcsinh-normal distributions to add un- certainty to neural network regression tasks: Applications to tropical cyclone intensity forecasts. *Environmental Data Sci.*, 2:e15, 2023. doi: 0.1017/eds.2023.7.
- Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06185-3.
- Marja Bister and Kerry A Emanuel. Low frequency variability of tropical cyclone potential intensity 1. interannual to interdecadal variability. *Journal of Geophysical Research: Atmospheres*, 107 (D24):ACL–26, 2002.
- Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier neural operators: Learning stable dynamics on the sphere, 2023. URL <http://arxiv.org/abs/2306.03838>.
- P. Bougeault and Coauthors. The THORPEX interactive grand global ensemble. *Bull. Amer. Meteor. Soc.*, 91:1059–1072, 2010. doi: 10.1175/2010BAMS2853.1.
- Stella Bourdin and Leo Saffin. Huracanpy: A python package for reading and analysing cyclone tracks. *Journal of Open Source Software*, 10(111):8174, 2025. doi: 10.21105/joss.08174. URL <https://doi.org/10.21105/joss.08174>.
- Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- William J. Broad. Artificial intelligence gives weather forecasters a new edge, July 2024. URL <https://www.nytimes.com/interactive/2024/07/29/science/ai-weather-forecast-hurricane.html>. The New York Times, Published July 29, 2024. Accessed May 14, 2025.
- Tomer Burg and Sam Lillo. TroPYcal, 2019. URL <https://github.com/tropycal/tropycal>. Github repository.
- J.P Cangialosi, E. Blake, M. DeMaria, A. Penny, A. Latta, E. Rappaport, and V. Tallapragada. Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Weather and Forecasting*, 35:1913–1922, 2020. doi: 10.1175/WAF-D-20-0059.1.

- 540 Daniel R Chavas, Ning Lin, and Kerry Emanuel. A model for the complete radial structure of the
541 tropical cyclone wind field. part i: Comparison with observed structure. *Journal of the Atmospheric*
542 *Sciences*, 72(9):3647–3662, 2015.
- 543
- 544 CA Davis. Resolving tropical cyclone intensity in models. *Geophysical Research Letters*, 45(4):
545 2082–2087, 2018.
- 546 M. DeMaria and J. Kaplan. A statistical hurricane intensity prediction scheme (SHIPS) for the
547 Atlantic basin. *Wea. Forecasting*, 9:209–220, 1994. doi: 10.1175/1520-0434(1994)009<0209:
548 ASHIPS>2.0.CO;2.
- 549
- 550 M. DeMaria, C. R. Sampson, J. A. Knaff, and K. D. Musgrave. Is tropical cyclone intensity guidance
551 improving? *Bull. Amer. Meteor. Soc.*, 95:387–398, 2014. doi: 10.1175/BAMS-D-12-00240.1.
- 552
- 553 M. DeMaria, J.L. Franklin, M.J. Onderlinde, and J. Kaplan. Operational forecasting of tropical
554 cyclone rapid intensification at the National Hurricane Center. *Atmosphere*, 12:683, 2021. doi:
555 10.3390/atmos12060683.
- 556 Mark DeMaria, Michelle Mainelli, Lynn K. Shay, John A. Knaff, and John Kaplan. Further improve-
557 ments to the statistical hurricane intensity prediction scheme (ships). *Weather and Forecasting*, 20
558 (4):531 – 543, 2005.
- 559
- 560 Mark DeMaria, James L. Franklin, Rachel Zelinsky, David A. Zelinsky, Matthew J. Onder-
561 linde, John A. Knaff, Stephanie N. Stevenson, John Kaplan, Kate D. Musgrave, Galina
562 Chirokova, and Charles R. Sampson. The national hurricane center tropical cyclone
563 model guidance suite. *Weather and Forecasting*, 37(11):2141 – 2159, 2022. doi: 10.
564 1175/WAF-D-22-0039.1. URL [https://journals.ametsoc.org/view/journals/
565 wefo/37/11/WAF-D-22-0039.1.xml](https://journals.ametsoc.org/view/journals/wefo/37/11/WAF-D-22-0039.1.xml).
- 566 Mark DeMaria, James L. Franklin, Galina Chirokova, Jacob Radford, Robert DeMaria, Kate D.
567 Musgrave, and Imme Ebert-Uphoff. Evaluation of tropical cyclone track and intensity forecasts
568 from artificial intelligence weather prediction (aiwp) models, 2024. URL [https://arxiv.
569 org/abs/2409.06735](https://arxiv.org/abs/2409.06735).
- 570
- 571 William Dulac, Julien Cattiaux, Fabrice Chauvin, Stella Bourdin, and Sébastien Fromang. Assessing
572 the representation of tropical cyclones in era5 with the cnrm tracker. *Climate Dynamics*, 62(1):
573 223–238, 2024.
- 574 Imme Ebert-Uphoff and Kyle Hilburn. The outlook for ai weather prediction, 2023.
- 575
- 576 M.A. Fernandez, E.A. Barnes, R.J. Barnes, M. DeMaria, M. McGraw, G. Chirokova, and L. Lu.
577 Predicting tropical cyclone track forecast errors using a probabilistic neural network. *Artificial*
578 *Intelligence for the Earth Systems*, 4, 2025. doi: 10.1175/AIES-D-24-0066.1.
- 579
- 580 Jennifer Gahtan, Kenneth R. Knapp, Carl J. III Schreck, Howard J. Diamond, James P. Kossin, and
581 Michael C. Kruk. International best track archive for climate stewardship (ibtracs) project, version
582 4.01, 2024. URL <https://doi.org/10.25921/82ty-9e16>. [indicate subset used].
- 583
- 584 Tilmann Gneiting and Adrian E Raftery and. Strictly proper scoring rules, prediction, and es-
585 timation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi:
586 10.1198/016214506000001437.
- 587
- 588 Milton Gomez, Louis Poulain-Auzéau, Alexis Berne, and Tom Beucler. Global forecasting of tropical
589 cyclone intensity using neural weather models. *arXiv preprint arXiv:2508.17903*, 2025.
- 590
- 591 S.M. Griffin, A. Wimmers, and C.S. Velden. Predicting rapid intensification in North Atlantic
592 and eastern North Pacific tropical cyclones using a convolutional neural network. *Weather and*
593 *Forecasting*, 37:1333–1355, 2022. doi: 10.1175/WAF-D-21-0194.1.
- 594
- 595 Danan Gu, Kirill Andreev, and Matthew E Dupre. Major trends in population growth around the
596 world. *China CDC weekly*, 3(28):604, 2021.

- 594 Andrew Hazelton, Ghassan J Alaka Jr, Lew Gramer, William Ramstrom, Sarah Ditchek, Xiaomin
595 Chen, Bin Liu, Zhan Zhang, Lin Zhu, Weiguo Wang, et al. 2022 real-time hurricane forecasts from
596 an experimental version of the hurricane analysis and forecast system (hafsv0. 3s). *Frontiers in*
597 *Earth Science*, 11:1264969, 2023.
- 598
599 JT Heming. Tropical cyclone tracking and verification techniques for met office numerical weather
600 prediction models. *Meteorological Applications*, 24(1):1–8, 2017.
- 601 Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater,
602 Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis.
603 *Quart. J. Roy. Meteorol. Soc.*, 146(730):1999–2049, 2020.
- 604
605 Cheng Huang, Pan Mu, Jinglin Zhang, Sixian Chan, Shiqi Zhang, Hanting Yan, Shengyong Chen, and
606 Cong Bai. Benchmark dataset and deep learning method for global tropical cyclone forecasting.
607 *Nature Communications*, 16(1):5923, 2025.
- 608
609 J. Kaplan and M. DeMaria. An updated statistical hurricane intensity prediction scheme (SHIPS) for
610 the Atlantic and eastern North Pacific basins. *Weather and Forecasting*, 14:326–337, 1999.
- 611
612 J. Kaplan, M. DeMaria, and J.A. Knaff. A revised tropical cyclone rapid intensification index for the
613 Atlantic and eastern North Pacific basins. *Weather and Forecasting*, 25:220–241, 2010.
- 614
615 J. Kaplan, C.M. Rozoff, M. DeMaria, C.R. Sampson, J.P. Kossin, C.S. Velden, J.J. Cione, J.P. Dunion,
616 J.A. Knaff, J.A. Zhang, J.F. Dostalek, J.D. Hawkins, T.F. Lee, and J.E. Solbrig. Evaluating
617 environmental impacts on tropical cyclone rapid intensification. *Weather and Forecasting*, 30:
1374–1396, 2015. doi: 10.1175/WAF-D-15-0032.1.
- 618
619 John Kaplan and Mark DeMaria. Large-scale characteristics of rapidly intensifying
620 tropical cyclones in the north atlantic basin. *Weather and Forecasting*, 18(6):1093
621 – 1108, 2003. doi: 10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2. URL
622 [https://journals.ametsoc.org/view/journals/wefo/18/6/1520-0434_](https://journals.ametsoc.org/view/journals/wefo/18/6/1520-0434_2003_018_1093_lcorit_2_0_co_2.xml)
623 [2003_018_1093_lcorit_2_0_co_2.xml](https://journals.ametsoc.org/view/journals/wefo/18/6/1520-0434_2003_018_1093_lcorit_2_0_co_2.xml).
- 624
625 J. A. Knaff, M. DeMaria, C. R. Sampson, and J. M. Gross. Statistical, 5-day tropical cyclone intensity
626 forecasts derived from climatology and persistence. *Wea. Forecasting*, 18:80–92, 2003. doi:
10.1175/1520-0434(2003)018<0080:SDTCIF>2.0.CO;2.
- 627
628 J.A. Knaff, C.R. Sampson, and K.D. Musgrave. An operational rapid intensification prediction
629 aid for the western North Pacific. *Weather and Forecasting*, 33:799–811, 2018. doi: 10.1175/
630 WAF-D-18-0012.1.
- 631
632 Kenneth R Knapp, Michael C Kruk, David H Levinson, Howard J Diamond, and Charles J Neumann.
633 The international best track archive for climate stewardship (ibtracs) unifying tropical cyclone data.
Bulletin of the American Meteorological Society, 91(3):363–376, 2010.
- 634
635 Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran
636 Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan
637 Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and
638 Peter Battaglia. Learning skillful medium-range global weather forecasting, 2023.
- 639
640 Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult,
641 Mariana CA Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, et al. Aifs–ecmwf’s
data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024.
- 642
643 Joseph W Lockwood, Avantika Gori, and Pierre Gentine. A generative super-resolution model for
644 enhancing tropical cyclone wind field intensity and resolution. *Journal of Geophysical Research:*
645 *Machine Learning and Computation*, 1(4):e2024JH000375, 2024.
- 646
647 A.E. Mercer, A.D. Grimes, and K.M. Wood. Application of unsupervised learning techniques to
identify Atlantic tropical cyclone rapid intensification environments. *J. App. Meteorol. and Climat.*,
60:119–138, 2021. doi: 10.1175/JAMC-D-20-0105.1.

- 648 Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover,
649 and Pierre Gentine. Chaosbench: A multi-channel, physics-based benchmark for
650 subseasonal-to-seasonal climate prediction. In A. Globerson, L. Mackey, D. Belgrave,
651 A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Informa-*
652 *tion Processing Systems*, volume 37, pp. 43715–43729. Curran Associates, Inc., 2024.
653 URL [https://proceedings.neurips.cc/paper_files/paper/2024/file/](https://proceedings.neurips.cc/paper_files/paper/2024/file/4d3684dd7926754b48bc6cd99a840232-Paper-Datasets_and_Benchmarks_Track.pdf)
654 [4d3684dd7926754b48bc6cd99a840232-Paper-Datasets_and_Benchmarks_](https://proceedings.neurips.cc/paper_files/paper/2024/file/4d3684dd7926754b48bc6cd99a840232-Paper-Datasets_and_Benchmarks_Track.pdf)
655 [Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/4d3684dd7926754b48bc6cd99a840232-Paper-Datasets_and_Benchmarks_Track.pdf).
- 656 Barbara Neumann, Athanasios T Vafeidis, Juliane Zimmermann, and Robert J Nicholls. Future
657 coastal population growth and exposure to sea-level rise and coastal flooding-a global assessment.
658 *PLoS one*, 10(3):e0118571, 2015.
- 659 NOAA National Centers for Environmental Information. U.S. Billion-Dollar Weather and Climate
660 Disasters (2025). <https://www.ncei.noaa.gov/access/billions/>, 2025. Accessed: 09/05/2025.
- 661 Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay,
662 Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram
663 Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A global data-
664 driven high-resolution weather model using adaptive fourier neural operators, 2022. URL [http:](http://arxiv.org/abs/2202.11214)
665 [://arxiv.org/abs/2202.11214](http://arxiv.org/abs/2202.11214).
- 666 Henri Rossi Pinheiro, Kevin Ivan Hodges, and Manoel Alonso Gan. An intercomparison of subtropical
667 cut-off lows in the southern hemisphere using recent reanalyses: Era-interim, ncep-cfrs, merra-2,
668 jra-55, and jra-25. *Climate Dynamics*, 54(1):777–792, 2020.
- 669 Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic
670 Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic
671 weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.
- 672 Stephan Rasp. AI-Weather SotA vs Time. 12 2024. doi: 10.6084/m9.figshare.28083515.
673 v1. URL [https://figshare.com/articles/dataset/AI-Weather_SotA_vs_](https://figshare.com/articles/dataset/AI-Weather_SotA_vs_Time/28083515)
674 [Time/28083515](https://figshare.com/articles/dataset/AI-Weather_SotA_vs_Time/28083515).
- 675 Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell,
676 Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A
677 benchmark for the next generation of data-driven global weather models. *Journal of Advances in*
678 *Modeling Earth Systems*, 16(6):e2023MS004019, 2024.
- 679 Muhammad Naufal Razin, Christopher J. Slocum, John A. Knaff, Paula J. Brown, and Michael M.
680 Bell. Tropical cyclone precipitation, infrared, microwave, and environmental dataset (tc
681 primed). *Bulletin of the American Meteorological Society*, 104(11):E1980 – E1998, 2023.
682 doi: 10.1175/BAMS-D-21-0052.1. URL [https://journals.ametsoc.org/view/](https://journals.ametsoc.org/view/journals/bams/104/11/BAMS-D-21-0052.1.xml)
683 [journals/bams/104/11/BAMS-D-21-0052.1.xml](https://journals.ametsoc.org/view/journals/bams/104/11/BAMS-D-21-0052.1.xml).
- 684 C.M. Rozoff and J.P. Kossin. New probabilistic forecast models for the prediction of tropical cyclone
685 rapid intensification. *Weather and Forecasting*, 26:677–689, 2011. doi: 10.1175/WAF-D-10-05059.
686 1.
- 687 Pankaj Lal Sahu, Sukumaran Sandeep, and Hariprasad Kodamana. Evaluating global machine
688 learning models for tropical cyclone dynamics and thermodynamics. *Journal of Geophysical*
689 *Research: Machine Learning and Computation*, 2(2):e2025JH000594, 2025. doi: [https://doi.org/](https://doi.org/10.1029/2025JH000594)
690 [10.1029/2025JH000594](https://doi.org/10.1029/2025JH000594). URL [https://agupubs.onlinelibrary.wiley.com/doi/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025JH000594)
691 [abs/10.1029/2025JH000594](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025JH000594). e2025JH000594 2025JH000594.
- 692 C.R. Sampson, J.A. Knaff, C.J. Slocum, M.J. Onderlinde, A. Brammer, M. Frost, and B. Strahl.
693 Deterministic rapid intensity forecast guidance for the Joint Typhoon 2 Warning Center’s Area of
694 Responsibility. *Weather and Forecasting*, 2023. doi: 10.1175/WAF-D-23-0084.1.
- 695 Carl J Schreck III, Kenneth R Knapp, and James P Kossin. The impact of best track discrepancies on
696 global tropical cyclone climatologies using ibtracs. *Monthly Weather Review*, 142(10):3881–3899,
697 2014.

- 702 H. Shaiba and M. Hahsler. Applying machine learning methods for predicting tropical cyclone rapid
703 intensification events. *Research Journal of Applied Sciences, Engineering and Technology*, 8:
704 638–651, 2016. doi: 10.19026/rjaset.13.3050.
- 705
706 C.J. Slocum. What can we learn from random forest in the context of the tropical cyclone rapid
707 intensification problem? 2nd NOAA Workshop on Leveraging AI in the Environmental Sciences,
708 2021. URL [https://www.star.nesdis.noaa.gov/star/documents/meetings/
709 2020AI/presentations/202101/20210128_Slocum.pptx](https://www.star.nesdis.noaa.gov/star/documents/meetings/2020AI/presentations/202101/20210128_Slocum.pptx).
- 710 H. Su, L. Wu, J.H. Jiang, R. Pai, A. Liu, A.J. Zhai, P. Tavallali, and M. DeMaria. Applying satellite
711 observations of tropical cyclone internal structures to rapid intensification forecast with machine
712 learning. *Geophys. Res. Lett.*, 47:L089102, 2020. doi: 10.1029/2020GL089102.
- 713
714 R.D. Torn and M. DeMaria. Validation of ensemble-based probabilistic tropical cyclone intensity
715 change. *Atmosphere*, 12:373, 2021. doi: 10.3390/atmos12030373.
- 716
717 P. A. Ullrich, C. M. Zarzycki, E. E. McClenny, M. C. Pinheiro, A. M. Stansfield, and K. A.
718 Reed. Tempestextremes v2.1: a community framework for feature detection, tracking, and
719 analysis in large datasets. *Geoscientific Model Development*, 14(8):5023–5048, 2021. doi:
720 10.5194/gmd-14-5023-2021. URL [https://gmd.copernicus.org/articles/14/
721 5023/2021/](https://gmd.copernicus.org/articles/14/5023/2021/).
- 722 Qidong Yang, Chia-Ying Lee, and Michael K. Tippett. A long short-term memory model for global
723 rapid intensification prediction. *Weather and Forecasting*, 35(4):1203 – 1220, 2020.
- 724 Qidong Yang, Chia-Ying Lee, Michael K Tippett, Daniel R Chavas, and Thomas R Knutson. Machine
725 learning–based hurricane wind reconstruction. *Weather and Forecasting*, 37(4):477–493, 2022.
- 726
727 Michaël Zamo and Philippe Naveau. Estimation of the continuous ranked probability score with
728 limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*,
729 50(2):209–234, 2018.

731 APPENDICES

732 733 A ACCOUNTABILITY AND REPRODUCIBILITY

734
735 TC Bench is released under the open-source GNU General Public License. Continued development,
736 including updates discussed in the limitations section, will be managed on the official TC Bench
737 GitHub page. Maintenance includes community support, issue tracking, and ongoing curation of
738 benchmark content. All assets are hosted on GitHub and HuggingFace to ensure stability and
739 accessibility.

740 741 Resources:

- 742 • **Dataset:** Available at *redacted for anonymity*
- 743 • **Model Checkpoints:** *redacted for anonymity*
- 744 • **Code:** *redacted for anonymity*
- 745 • **Documentation:** *redacted for anonymity*
- 746

747 An anonymized subset of the code and data has been prepared and uploaded to ProtonDrive:

748 <https://drive.proton.me/urls/SKT1FGC3KG#1IIOFkFcL4xOU>

749
750 Please note that though any reference to usernames and paths that might be used to identify the
751 developers were removed, metadata scrubs were not conducted. We request that reviewers refrain
752 from viewing file metadata in case any identifying information be contained within.

753 754 A.1 HUGGINGFACE REPOSITORY STRUCTURE

755 We organize the TC Bench HuggingFace repository as follows:

- 756 • `texttt2023_IBTrACS.csv`: The subset of IBTrACS with the storms and timestamps that we use
757 to evaluate models.
- 758 • `matched_tracks/`: A collection of CSVs with the tracks we evaluate as baselines, including
759 AIFS, FourCastNetv2, Weatherlab FNv3, GenCast, Panguweather, TIGGE GEFS, and TIGGE
760 IFS.
- 761 • `neural_weather_models/`: Raw, global neural weather model outputs for up to 5 days
762 lead time, initialized at each timestamp associated with a storm in `2023_IBTrACS.csv` (i.e., the
763 ground truth).
- 764 • `postprocessing_models/`: Postprocessing model weights and the data required to run
765 them on the test set.

767 A.2 GITHUB REPOSITORY STRUCTURE

768 We organize the TC Bench Github repository as follows:

- 770 • `track_processing/`: The TempestExtremes routines needed to generate the tracks associ-
771 ated with the neural weather models, and the python routines that rely on HuracanPY to match the
772 tracks found with TempestExtremes to the TCs in IBTrACS.
- 773 • `utils/`: Collection of utilities used to process IBTrACS data, to process neural weather model
774 data for post-processing, and to calculate metrics on the tracks.
- 775 • `scriptnames.py`: various python scripts used to generate the track baseline predictions
776 (e.g., `baselines.py`, `climatology_maker.py`, `compute_persistence.py`), to evaluate a set of tracks
777 (e.g., `evaluate_tracks.py`)

779 B GETTING STARTED

780 We provide a detailed description of how to prepare the necessary data, perform training, and
781 benchmark your own model on our website. Please visit the following URL to see the guide: *redacted*
782 *for anonymity*. We also ask that if you encounter any issues to please feel free to contact us or raise
783 an issue on GitHub.

784 A getting started notebook for the anonymized review was prepared and is available at:
785 <https://drive.proton.me/urls/SKT1FGC3KG#1IOFkFcL4xOU>
786 under the filename `Getting_Started.ipynb`

789 C TC BENCH DATA ORGANIZATION AND WORKFLOW

791 C.1 DATA SOURCES

792 TC Bench integrates multiple observational, reanalysis, forecast model, and AI-based data sources
793 relevant to tropical cyclone (TC) analysis and prediction. Table S1 summarizes datasets currently
794 included, and points to additional datasets of relevance.

795 **Atmospheric Field Data:** We use atmospheric field data from a variety of sources, including:

- 796 • Climate reanalysis and forecast datasets including ERA-5, TIGGE, IFS, and GFS.
- 797 • Outputs from AI models, including PanguWeather, FourcastNet, and the single member version
798 of AIFS 1.0.

802 C.2 DATA WORKFLOW

803 The TC Bench workflow standardizes diverse data sources into a common format for evaluation:

- 804 1. Forecast and reanalysis fields are subset in space (storm-centered region) and time (forecast
805 lead times) using storm initialization from IBTrACS.
- 806 2. Model-specific track outputs (e.g., TIGGE XML files, neural weather model forecasts in
807 netCDF format) are converted into a uniform CSV format that includes the storm identifier
808 (taken from IBTrACS), position, maximum sustained wind, and minimum sea-level pressure.
809

Data Source	Description	Website	Provided
Reanalysis			
ERA5	European Re-Analysis 5	https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview	No
Ground Truth			
IBTrACS	Best Track Archive	https://www.ncdc.noaa.gov/ibtracs/	Yes
Extended Tracks	Best- TC tracks with size	https://rammb2.cira.colostate.edu/research/tropical-cyclones/tc_extended_best_track_dataset/	No
TC PRIMED	ML Ready Benchmark dataset	https://rammb-data.cira.colostate.edu/tcprimed/	No
Statistical Models			
SHIPS	Statistical Hurricane Intensity Prediction Scheme (developmental dataset)	https://rammb2.cira.colostate.edu/research/tropical-cyclones/ships/development_data/	No
Models			
NCEP GFS	Global Forecast System	https://www.nco.ncep.noaa.gov/pmb/products/gfs/	No
TIGGE GFS	NCEP GEFS ensemble model tracks dataset	https://rda.ucar.edu/datasets/d330003/dataaccess/#	Yes [†]
TIGGE IFS	ECMWF IFS ensemble model tracks dataset	https://rda.ucar.edu/datasets/d330003/dataaccess/#	Yes [†]
PanguWeather	AI global model outputs	https://github.com/198808xc/Pangu-Weather	Yes
FourCastNet	AI model outputs	https://github.com/NVlabs/FourCastNet	No
FourCastNetv2	AI model outputs	https://github.com/NVIDIA/torch-harmonics	Yes
AIFS	AI global forecasting	https://arxiv.org/abs/2406.01465	Yes

Table S1: Data Sources Summary. [†] non-gridded data only

3. A `TCTrack` object is created for each storm, encapsulating observed and predicted values across time steps.

4. Gridded predictors (e.g., neural weather model forecasts) are stored as multidimensional arrays `[samples, time, lat, lon, variables]` for use in ML-based experiments.

C.3 DATA PROCESSING

Each dataset undergoes preprocessing to ensure comparability:

- **TIGGE ensembles:** Extract ensemble mean and member tracks, harmonize to IBTrACS temporal resolution.
- **AI models:** Post-process global forecasts to storm-centered tracks; derive intensity and MSLP comparable to IBTrACS.

- **Post-processing Models:** Extract 50 member ensemble from parametric distribution through sampling, clip predictions to physical values ($V_{max} \geq 0 \text{ kt}$).

C.4 MODEL EVALUATION AND METRICS

We evaluate forecast performance using the following metrics, standard in tropical cyclone verification:

- **Track error:** Great-circle distance (km) between forecast and observed positions.
- **Intensity error:** Mean absolute error (MAE) of maximum sustained winds (kt) and minimum sea-level pressure (hPa).
- **Skill scores:** Relative to climatology, persistence, and SHIPS baselines.
- **Probabilistic metrics:** Brier score and reliability diagrams for ensemble/AI probability forecasts.
- **Partitioning:** Cross-validation by year ensures temporal independence between training and evaluation sets.

C.5 BIASES AND LIMITATIONS

- The dataset represents a subset of all possible TC tracks; some regions and seasons may be underrepresented, especially where fewer storms materialize.
- Forecast evaluation may omit rare teleconnection patterns or unusual environmental scenarios, limiting generalizability.
- Definitions of storm intensity (e.g., 1-min vs. 10-min sustained winds) vary by agency and observation method, introducing potential inconsistencies across basins. We control for this by relying on USA reported values.
- IBTrACS quality varies by basin, with lower reliability in the South Indian Ocean and in the pre-satellite era.
- Reanalysis products such as ERA5 contain uncertainties in poorly observed regions and for weak or short-lived systems.
- AI models are rapidly updated (e.g., FourCastNet v1 vs. v2 vs. v3); benchmark results reflect the versions available to the authors at the time of dataset creation.
- Storm size and structural parameters are inconsistently available across sources, limiting evaluation beyond track and intensity.

D BASELINE MODELS

TCBench provides a set of baseline models including traditional statistical approaches, numerical weather prediction (NWP) models, and simple machine learning (ML) approaches. These baselines serve as reference points for track, intensity, and rapid intensification forecasts.

D.1 PERSISTENCE AND CLIMATOLOGY BASELINES

We rely on persistence and climatology as so called "naïve" baselines to provide context for the performance of the evaluated models against straightforward, transparent prediction methods.

- **Persistence:** Assumes that storm position, intensity, and structure remain unchanged from the current state. Provides a naive forecast useful for short-lead comparisons.
- **Climatology / Mean Track – Long Baseline (MT-LB):** Uses historical averages of storm positions and intensities within each basin and season to generate forecasts. Captures long-term trends and typical seasonal behavior.

D.2 NUMERICAL WEATHER PREDICTION (NWP) BASELINES

We rely on physics-based Numerical Weather Prediction models as baselines for global predictions of TCs. These models rely on physics and numerical solutions of known equations to predict the state of the atmosphere based on initial conditions. We use the following models in TCBench:

- **NCEP GFS / GEFS:** Global Forecast System provides deterministic forecasts, while GEFS provides ensemble reforecasts from TIGGE. Baselines include position, maximum wind, and minimum sea-level pressure.
- **ECMWF IFS / TIGGE IFS:** Ensemble forecasts from ECMWF’s Integrated Forecasting System. Used for both track and intensity benchmarks.

We take the data provided by TIGGE as it is openly available and due to computational limitations associated with running ensembles of physics-based NWP models.

D.3 NEURAL WEATHER MODEL BASELINES

We rely on neural weather models available at the time of writing to provide baselines for global weather prediction using artificial intelligence. A brief description of the models is given below:

- **FourCastNetV2** - neural weather model that relies on spherical harmonics based neural operators to learn a grid-invariant evolution of the state of the atmosphere.
- **PanguWeather** - neural weather model that relies on an earth-specific transformer architecture and multiple, lead time specific models to predict the evolution of the atmosphere.
- **AIFS Single 1.0** - neural weather model whose architecture relies on a graph-based latent representation of the earth and a sliding window transformer processor in order to predict the evolution of the atmosphere.
- **Weathernext Gen (GenCast)[†]** - neural weather model that relies on a conditional diffusion architecture to predict a possible evolution of the state of the atmosphere.
- **FNv3[†]** - neural weather model that is not currently open, but is purported to be a functional generative network (FGN) (Alet et al., 2025) fine-tuned on IBTrACS to better predict tropical cyclones.

Track and intensity forecasts for the models marked with † are taken from Google WeatherLab (<https://deepmind.google.com/science/weatherlab>), due to either computational limitations for running the models or to the models not being open at the time of writing.

D.4 POST PROCESSING BASELINES

We provide a set of postprocessing baseline models following Gomez et al. (2025). Each postprocessing model uses square, clipped neural weather model fields centered on the position of the tropical cyclone at the initial time of forecast, with an extent of $\pm 30^\circ$ latitude and longitude. These models further use an embedding of the position, maximum wind speed, and minimum sea level pressure at the initial time of forecast for their prediction. These models are set up for distributional regression of the rate of the intensification, and output the mean and standard deviation for a gaussian distribution associated with forecast. Of particular note, the postprocessing models only predict intensification, and do not predict the motion of the storm (i.e., the evolution of the position of the TC).

We provide three postprocessing models with varying levels of computational complexity:

- **Multiple Linear Regression:** the neural weather model forecasts are reduced (e.g., by calculating extrema) and a pair of MLRs are trained to predict the mean and standard deviation of the rate of intensification.
- **Multilayer Perceptron (ANN):** the neural weather model forecasts are reduced in a manner consistent with the reduction used for the MLR, but the regression algorithm is instead a MLP.
- **UNet:** the neural weather model forecasts and an embedding of the scalar quantities associated with the forecast (e.g., the position and intensity at the initial time of forecast) are used in the regression task.

We use the details provided for these architectures and associated hyperparameters provided by Gomez et al. (2025) in training the models.

D.5 TRAINING SETUP

All baselines are trained and evaluated consistently:

- **Data Splits:** Storms are split into folds by year to minimize temporal leakage, ensuring that models are evaluated on unseen storms.
- **Input Variables:** Depending on the baseline, inputs include:
 - Current storm state: latitude, longitude, maximum wind, minimum pressure
 - Environmental predictors: derived from ERA5 or TIGGE
 - Historical storm statistics (for climatology/MT-LB)
- **Ensemble Strategy:** For probabilistic evaluation, ML ensembles produce multiple realizations; NWP ensembles (e.g., GEFS, IFS) are used directly as provided; Post-processing model output distributions are sampled to produce a 50-member ensemble.

Baseline forecasts are evaluated using the same deterministic, probabilistic, and rare event metrics described in Section 4.

D.6 BASELINE MODELS SUMMARY TABLE

Baseline	Type	Inputs	Outputs / Target
Persistence	Deterministic	Current storm state	Track, intensity
Climatology / MT-LB	Statistical	Historical storm averages	Track, intensity
NCEP GFS / GEFS	NWP / Ensemble	Model initial conditions	Track, intensity
TIGGE IFS	NWP / Ensemble	Model initial conditions	Track, intensity
FourCastNetv2	AI	Model Initial Conditions	Track, intensity
PanguWeather	AI	Model Initial Conditions	Track, intensity
AIFS Single	AI	Model Initial Conditions	Track, intensity
WeatherNext Gen	AI	N/A*	Track, intensity
FNv3	AI	N/A*	Track, intensity
Pangu+MLR	PostProcessing AI	AI Model Forecasts, Current Storm state	Intensity
Pangu+ANN	PostProcessing AI	AI Model Forecasts, Current Storm state	Intensity
Pangu+UNet	PostProcessing AI	AI Model Forecasts, Current Storm state	Intensity

Table 2: Summary of baseline models included in TCBench. Each baseline serves as a reference for model evaluation.*Forecasts taken directly from Google WeatherLab

E EVALUATION METRICS

To test the forecast performance, TCBench use both deterministic and probabilistic metrics. Deterministic evaluation focuses on the accuracy of predicted intensity and track positions. Standard regression metrics include the root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). Track errors are quantified using direct positional error (DPE), cross-track error (CTE), and along-track error (ATE), which decompose the forecast position error along and across the observed storm motion.

For probabilistic forecasts, we evaluate ensemble predictions using the Continuous Ranked Probability Score (CRPS), which measures the discrepancy between the forecast distribution and the observed outcome, capturing both ensemble accuracy and spread. For track forecasts, CRPS is generalized by replacing absolute differences with Haversine distances between forecast and observed storm centers. Additional probabilistic metrics include Brier Skill Score (BSS), Continuous ranked probability score (CRPS), that provide complementary assessments of calibration and uncertainty representation. All details of metrics and computational procedures for all metrics are provided in this section.

E.1 DETERMINISTIC METRICS

- **Root Mean Square Error (RMSE):** Measures the square root of the average squared differences between predicted and actual values, indicating the model’s prediction accuracy.

- Mean Absolute Error (MAE): Represents the average absolute differences between predicted and actual values, reflecting the magnitude of errors in predictions.
- R-squared Score (R^2): Quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables, indicating the model’s explanatory power.

Direct Positional Error (DPE) is the most basic measure of the positional accuracy of a tropical cyclone (TC) forecast. It is defined as the distance between the forecast and observed positions of the storm at the same *verification time (VT)*.

There are two commonly used approaches for calculating DPE:

- **Great Circle Distance (GCD)-based DPE:** This version computes the shortest distance over the Earth’s surface between the forecast and observed positions, treating Earth as a sphere. The *Haversine formula* is commonly used:

$$d = 2r \arcsin \left(\sqrt{\sin^2 \left(\frac{\Delta\phi}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\Delta\lambda}{2} \right)} \right)$$

where:

- ϕ_1, ϕ_2 are the latitudes of the observed and forecast points (in radians),
- $\Delta\phi = \phi_2 - \phi_1, \Delta\lambda = \lambda_2 - \lambda_1$ are the latitude and longitude differences,
- r is the Earth’s radius (typically ≈ 6371 km).

This method accounts for Earth’s curvature and is the standard in operational TC verification.

- **Cartesian Projection-based DPE:** An alternative approach involves projecting both observed and forecast positions onto a local tangent plane (e.g., using an azimuthal equidistant or equiarectangular projection). The DPE is then computed using standard Euclidean distance:

$$d = \sqrt{(x_{\text{fcst}} - x_{\text{obs}})^2 + (y_{\text{fcst}} - y_{\text{obs}})^2}$$

where (x, y) are the projected coordinates. This method may be preferred for regional studies or error decomposition in zonal/meridional directions.

Note: DPE provides a scalar magnitude of error but does not convey directionality (e.g., north/south or ahead/behind). Directional errors are captured by the metrics below.

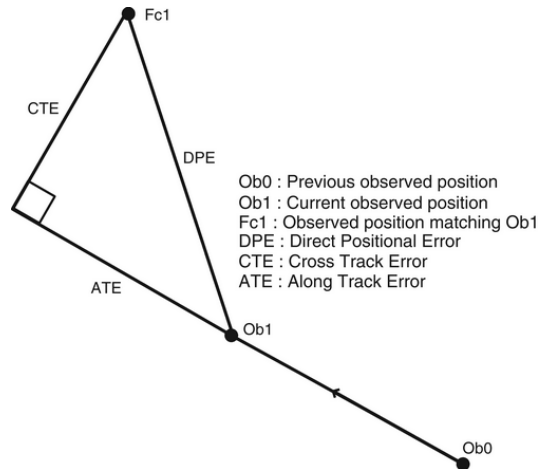


Figure 5: Illustration of track error metrics adapted from (Heming, 2017)

Cross-Track Error (CTE) measures the component of the track forecast error *perpendicular* to the observed direction of storm motion.

1. Define the observed motion vector from the storm position 12 hours prior to VT to the observed position at VT.

2. Draw a perpendicular line from the forecast position to this vector.
3. Find the intersection point of this perpendicular with the observed vector.
4. CTE is the great circle distance between the forecast position and this intersection point.
 - In the **Northern Hemisphere**, a positive CTE indicates the forecast is to the **right** of the observed path.
 - In the **Southern Hemisphere**, the interpretation is reversed.

Note: CTE is undefined at the first valid time (due to lack of a 12-hour prior observed position).

Along-Track Error (ATE) ATE measures the component of the error *along* the direction of storm motion and reflects timing accuracy.

- Using the same intersection point from the CTE calculation, compute the great circle distance between the observed position at VT and the intersection point.
- Positive ATE indicates the forecast is **ahead** of the observed position (i.e., storm is moving too fast).
- Negative ATE indicates the forecast is **behind** (i.e., storm is too slow).

E.2 PROBABILISTIC METRICS

Here, we consider an ensemble of N forecasts as a set denoted $\{x_i | i = 1, \dots, N\}$, and $\{o_i\}_N$ for short.

The **Brier skill score (BSS)** (Brier, 1950) is a relative measure that evaluates the improvement of a probabilistic forecast over a reference forecast or climatology. BSS ranges $-\infty$ to 1, where 1 indicates a perfect skill and 0 indicates no improvement over the reference. A negative BSS suggests that the forecast is less accurate than the reference. The BSS is based on the Brier Score (BS), a metric used to measure the accuracy of probabilistic predictions for binary outcomes, defined as

$$\text{BS} = \frac{1}{N} \sum_{i=1}^N (x_i - o_i)^2$$

where x_i is a forecast probability, o_i is the actual outcome, and N is the number of forecasts. Then BSS is defined as

$$\text{BSS} = 1 - \frac{\text{BS}_{\text{forecast}}}{\text{BS}_{\text{reference}}}$$

The **Continuous Ranked Probability Score (CRPS)** is widely used to assess the skill of ensemble forecasts by measuring the discrepancy between the predicted and observed cumulative distribution functions (CDFs). Unlike the BS, which applies to binary events, the CRPS evaluates the accuracy of probabilistic forecasts for continuous variables. Formally, for a forecast CDF \hat{F} and a realized value y , the CRPS is

$$\text{CRPS} = \int_{\mathbb{R}} \left[\hat{F}(x) - \mathcal{H}(x - y) \right]^2 dx, \quad (2)$$

where \mathcal{H} denotes the Heaviside step function, which represents the degenerate CDF of a single observation. For discrete ensembles, we use the kernel representation of the CRPS (Gneiting & and, 2007):

$$\text{CRPS}(\{x_i\}_N, y) = \frac{1}{N} \sum_{i=1}^N |x_i - y| - \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N |x_i - x_j|. \quad (3)$$

The first term is the mean absolute error of the ensemble members, while the second term measures the ensemble spread via the mean absolute pairwise difference. To mitigate finite-ensemble bias (inflated scores due to underestimation of the second term), we adopt the fair CRPS (Zamo & Naveau, 2018):

$$\text{fCRPS}(\{x_i\}_N, y) = \frac{1}{N} \sum_{i=1}^N |x_i - y| - \frac{1}{2N(N-1)} \sum_{i=1}^N \sum_{j=1}^N |x_i - x_j|. \quad (4)$$

E.3 METRICS FOR EXTREMES

Rapid Intensification: Rapid intensification is a rare event, but not an exceedingly rare event; by definition, it is the 95th percentile of intensity change. Thus, we evaluate model performance on rapid intensification forecasts using metrics more tailored for rare event forecasts. We note that while most of TC Bench is configured as a **regression** problem, we have formulated rapid intensification as a **binary classification problem**. This means rapid intensification models will be tasked with making a simple “yes/no” prediction for the occurrence of rapid intensification.

- **Critical Success Index (CSI):** Also known as the Threat Score, measures the ratio of correctly predicted positive observations to the sum of all predicted positives, actual positives, and minus true positives. Can be used both for probabilistic track evaluation and RI.
- **Peirce skill score (PSS):** The Peirce skill score (also known as the Hansen and Kuipers discriminant, or the true skill statistic) is an estimate of how well the forecast separates “yes” events from “no” events. It ranges from -1 to 1, with +1 being a perfect score and 0 indicating no skill. PSS is generally considered to be better for rare events, though for extremely rare events it tends to 0.

$$PSS = TPR - FPR \quad \text{where } TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}. \quad (5)$$

F EXTENDED RESULTS

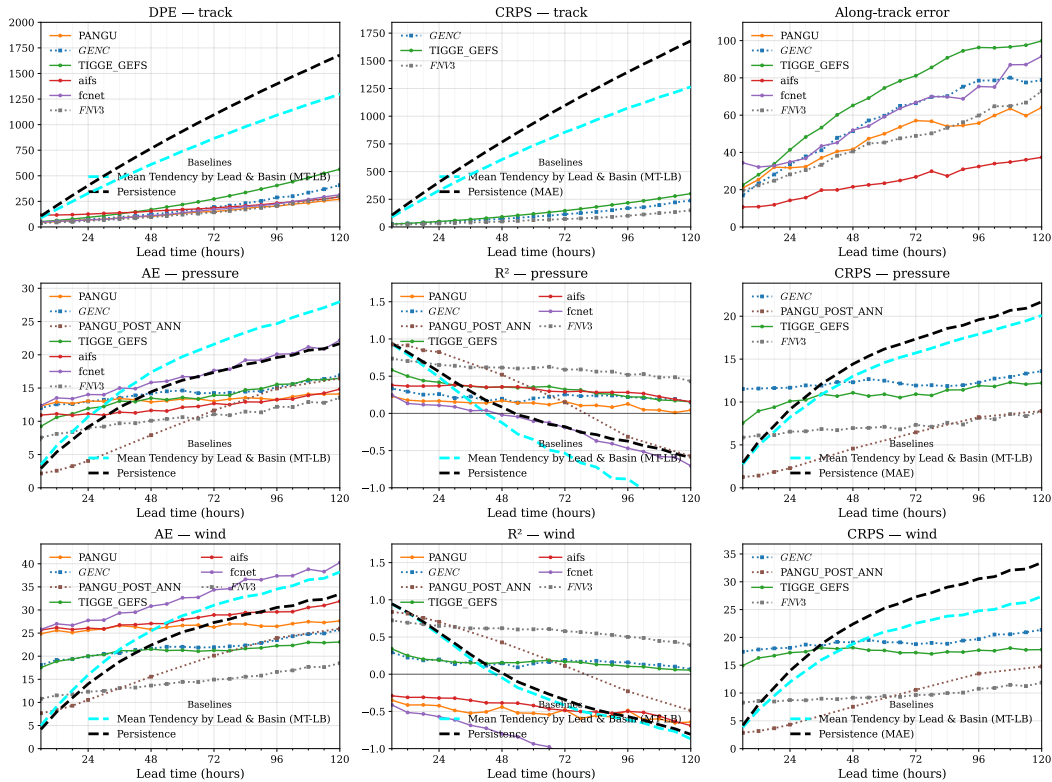
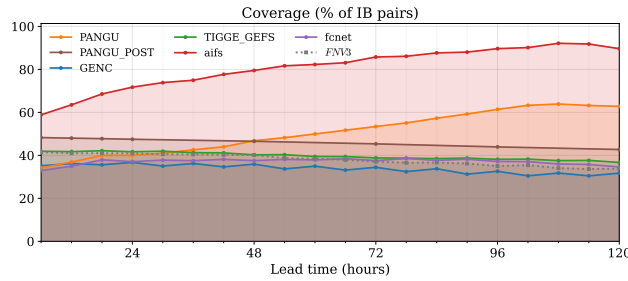


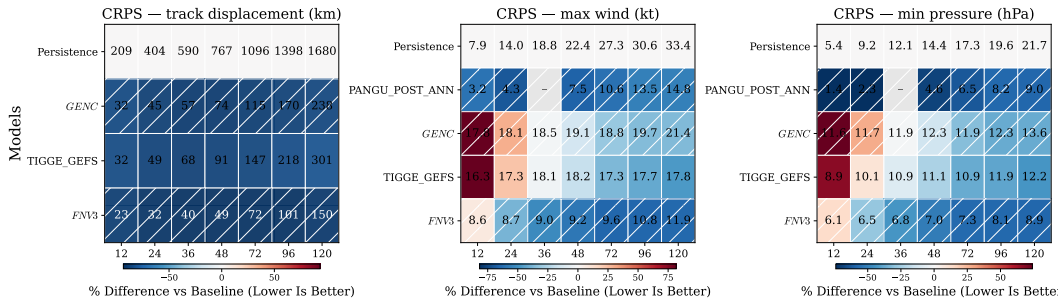
Figure 6: **Per-lead verification on TC Bench-2023 (6–120 h), non-filled.** Rows: track, pressure, wind; columns: AE, R^2 , CRPS (track uses DPE, CRPS-track, along-track). Curves use *raw model coverage*—no persistence filling—so each mean is over the forecast–verification pairs available for that model and lead. Models are colored (WeatherLab FNv3 in gray). Baselines: persistence (black dashed) and climatology/MT-LB (cyan dashed). $R^2 = 1 - \text{MSE}/\text{Var}$ vs. IBTrACS

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197



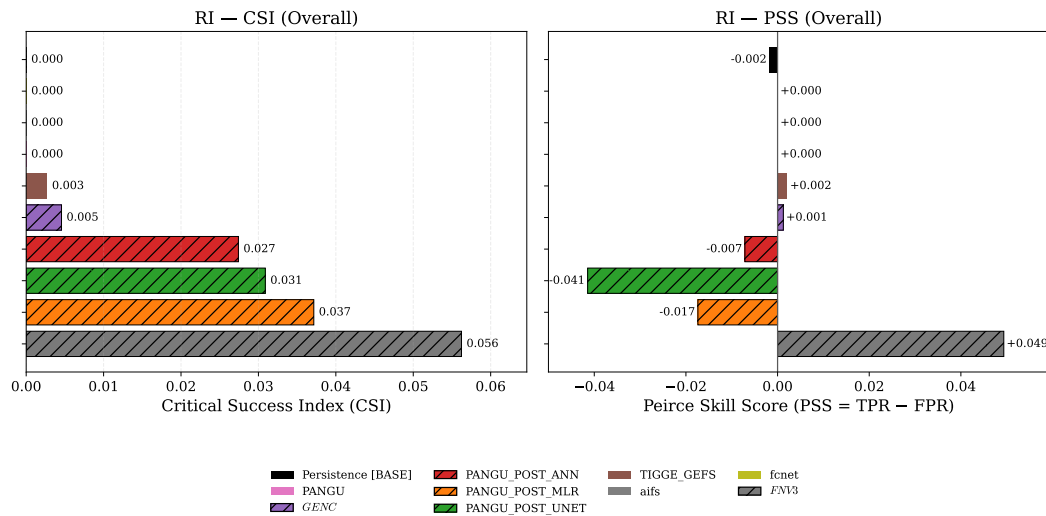
1198 **Figure 7: Coverage on the 2023 test set (% of IBTrACS verification pairs) by lead.** An IBTrACS
1199 pair is a unique (SID, t_0, t_0+L) observed on the 6-hour grid with $t_0 \in \{00, 12\}$ UTC. A model
1200 covers a pair if it outputs any row for that key. Shaded regions indicate the fraction covered at each
1201 lead. Large coverage disparities motivate the FAIR comparison (persistence-filled on the same IB
1202 grid) reported elsewhere in the paper.

1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214



1215 **Figure 8: CRPS scorecards (probabilistic).** Track displacement (km), max wind (kt), and min
1216 pressure (hPa) on TC Bench-2023 for 6–120,h leads. Each cell is the percent difference in CRPS
1217 relative to the Persistence baseline at the same lead (lower is better); the Persistence row reports
1218 absolute CRPS (units in panel titles). Results use *non-filled* data (raw model coverage); dashes denote
1219 no verification pairs at that lead. Hatched rows mark externally provided products (*GENC*, *FNv3*);
1220 *PANGU_POST_ANN* is a learned post-processor.

1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235



1236
1237
1238
1239
1240
1241

1239 **Figure 9: Rapid Intensification (RI) skill—overall by model.** Bars show overall Critical Success
1240 Index (CSI; left) and Peirce Skill Score (PSS = TPR–FPR; right) computed against the IBTrACS RI
1241 ground truth for 2023. Scores use the common (SID, t_0, t_0+L) key set on the 6 h IBTrACS grid.

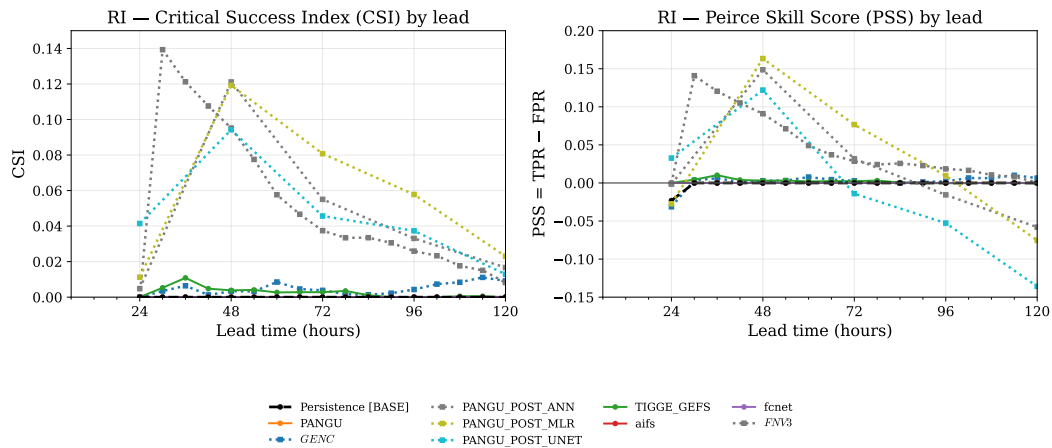


Figure 10: **Rapid Intensification (RI) skill by lead time.** CSI (left) and PSS (right) as a function of lead time (6–120 h, 6 h steps), evaluated vs. the IBTrACS 2023 RI ground truth on the 6 h grid. Axes are capped for readability.

G POTENTIAL APPLICATIONS

Accurate assessments of the wind-related risks posed by tropical cyclones rely heavily on knowledge of their full wind structures (2D images stacked in the vertical). It is challenging to obtain such wind fields because satellite observations often only provide partial coverage of TCs with limited temporal sampling. Various methods have been proposed to address this challenge, ranging from analytical parametric models (Chavas et al., 2015) to modern machine learning techniques (Yang et al., 2022), which attempt to infer spatial wind fields from easily observed scalar quantities such as minimum sea-level pressure or the radius of maximum surface winds. By offering IBTrACS in an AI-ready format, TC Bench facilitates the design and training of models to **predict key scalars needed for wind reconstruction**. Moreover, the standardized aggregation of outputs from multiple AI models allows users to assess and compare tropical cyclone size predictions across models, thereby helping to identify and quantify TC size biases in neural weather models.

By providing different routines to process TC observations in different formats, the TC Bench dataset is ideal for **developing very short-term TC winds and precipitation predictions**. Data-driven nowcasting models (Agrawal et al., 2019) are more computationally competitive than explicitly simulating the TCs with high-resolution numerical models. The ability of probabilistic data-driven models to create hundreds of possible predictions for a given initial condition within minutes and adjust based on new satellite image inputs makes them useful additions to the existing weather forecasting pipeline.

The examples provided in the previous sections only evaluate the intensity at a given point in the TCs. Similar to the wind field reconstruction tasks described earlier, we can **develop wind downscaling models** (Lockwood et al., 2024) to add additional spatial details to the TCs in the neural weather model outputs in coarse resolution. Another extension to the prediction tasks shown in the text is to provide **global prediction of TC activity across timescales**. The neural weather model outputs and simple tracking algorithm in TC Bench facilitate consistent evaluation of the skill of different AI models in developing the TCs at the right time and right place. Based on the comparison of TC activity patterns of different neural weather models, basin-specific refinements can be made to reduce potential spatial biases in the existing trained neural weather models.

Finally, the aggregation of satellite images, environmental fields, and intensity observations in an easy-to-access dataset enables **identification of critical predictive environmental properties for TC intensity** (Bister & Emanuel, 2002), thereby forming the basis for novel discovery on tropical cyclone physics.

1296 H LIMITATIONS
1297

1298 First, uncertainties associated with each product vary, which pose a challenge in estimating and
1299 providing accurate error margins for reanalyses and observations. An ambitious goal is to include ob-
1300 servational error estimates (e.g., intensity, track location) that vary over time, reflecting improvements
1301 in satellites and technology.

1302 Although using multiple reanalyses and tracking algorithms is ideal due to the variation in data
1303 quality and sensitivity of cyclone detection schemes, initial efforts may need to rely on a single
1304 reanalysis dataset for tractability, while clearly acknowledging its limitations (Pineiro et al., 2020).
1305 Furthermore, creating a single dataset that fits all purposes is impossible, necessitating trade-offs
1306 such as neglecting some teleconnections or using datasets only available for the past approximately
1307 15 years. Additionally, the definition of TC intensity varies by agency, primarily in the number of
1308 minutes used to calculate maximum sustained wind speeds, which introduces inconsistency.

1309 The first position of TCs across agencies and across time is also subject to large uncertainty and
1310 inconsistency. Therefore, in subseasonal to seasonal time-scales and climate studies, it is common
1311 to consider the first time the TC reaches tropical storm intensity (35 kt) in the track to improve
1312 consistency. Moreover, creating a seamless list of predictors for the entire tropics is challenging due
1313 to the variation in relevant variables across different basins or seasons.

1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349