RESEARCH ARTICLE

Key Points:

- The high skill of Machine Learning (ML) models in predicting tropical cyclone (TC) tracks is consistent with their ability to capture vorticity advection
- ML models accurately predicted the warm-core structure of TCs, although with slightly weaker intensity
- Beyond the known double penalization effect, training data bias also drives ML models' low bias in TC intensity prediction

Author Contributions:

Conceptualization: Sukumaran Sandeep, Hariprasad Kodamana
Data curation: Pankaj Lal Sahu
Formal analysis: Pankaj Lal Sahu, Sukumaran Sandeep, Hariprasad Kodamana
Funding acquisition: Sukumaran Sandeep, Hariprasad Kodamana

Evaluating Global Machine Learning Models for Tropical Cyclone Dynamics and Thermodynamics

Pankaj Lal Sahu[1], Sukumaran Sandeep[1,2], and Hariprasad Kodamana[2,3,4]

[1]Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, New Delhi, India, [2]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi, India, [3]Indian Institute of Technology Delhi-Abu Dhabi, Abu Dhabi, UAE, [4]Department of Chemical Engineering, Indian Institute of Technology Delhi, New Delhi, India

Abstract Machine Learning Weather Prediction (MLWP) models have recently demonstrated remarkable potential to rival physics-based Numerical Weather Prediction (NWP) models, offering global weather forecasts at a fraction of the computational cost. However, thorough evaluations are essential before considering MLWP models as replacements for NWP models. This study presents a comprehensive evaluation of four leading MLWP models—GraphCast, PanguWeather, Aurora, and FourCastNet—against observations and three state-of-the-art NWP models in predicting tropical cyclones (TCs) across all tropical ocean basins. All MLWP models exhibited strong skill in forecasting TC tracks, achieving an average track error of less than 200 km at a 96-hr forecast lead time. However, they consistently underestimated maximum sustained wind speeds compared to NWP models and observations. The low bias in TC intensity forecasts by MLWP models is linked to similar bias in their training data, along with the double penalization effect. MLWP models realistically captured the absolute vorticity patterns and their advection, demonstrating their ability to represent the dynamics underlying TC translation. They also captured the low-level convergence and vertical warm core structure of TCs, although the magnitudes were weaker than observed, highlighting the linkage between dynamical and thermodynamical processes. The consistency in magnitude between various physical fields in the MLWP models suggests that they intuitively learn the interrelationships among different physical fields during the evolution of weather systems, demonstrating their ability to capture complex physical interactions. Among the MLWP models, Aurora showed superior performance, surpassing GraphCast, PanguWeather, and FourCastNet.

Plain Language Summary Machine Learning (ML) models are revolutionizing global weather prediction. Physics-based Numerical Weather Prediction (NWP) models run on high-performance computing systems, requiring hundreds of processors and significant computational resources to generate global weather forecasts. In contrast, trained ML models produce forecasts in seconds on a single graphical processing unit, significantly enhancing the potential to predict extreme weather events like tropical cyclones (TCs). While preliminary studies indicate good skill of ML models in predicting TC tracks, a detailed evaluation of their performance in forecasting TC tracks, intensity, and underlying dynamics globally has been lacking. In this study, we comprehensively analyze the skill of four global ML Weather Prediction (MLWP) models in these aspects and compare them with three NWP models and observations. Our results show that MLWP models achieve smaller track errors than NWP models but predict weaker TC intensities. Among the MLWP models, Aurora, which uses a transformer-based architecture and a variety of training data sets, outperformed models based on a single training data set source. This indicates Aurora's superior ability to represent the dynamics and thermodynamics of TCs. These findings underscore the potential of ML models in weather prediction, offering opportunities to improve the accuracy and efficiency of forecasting extreme weather globally.

1. Introduction

Tropical cyclones (TCs) are among the most devastating natural events, presenting significant hazards to coastal areas due to intense winds, storm surges, and substantial precipitation (Islam et al., 2021; Kantha, 2008; Needham et al., 2015; Seo & Bakkensen, 2017). Precise forecasting of TC tracks and intensities is crucial for mitigating their societal and economic impacts (Fanchiotti et al., 2020; Lenzen et al., 2019; Naguib et al., 2022). Advancements in numerical weather prediction (NWP), data assimilation techniques, and the integration of satellite observations have significantly improved the precision of TC forecasts. Historically, forecasts were initially simplistic, based primarily on empirical techniques and climatological data. With a rise in computational capacity in the mid-20th century, dynamical models began advancement in TC forecasting (Willoughby et al., 2007). Over

Investigation: Pankaj Lal Sahu,
Sukumaran Sandeep,
Hariprasad Kodamana
Methodology: Pankaj Lal Sahu,
Sukumaran Sandeep,
Hariprasad Kodamana
Project administration:
Sukumaran Sandeep,
Hariprasad Kodamana
Resources: Sukumaran Sandeep,
Hariprasad Kodamana
Software: Pankaj Lal Sahu
Supervision: Sukumaran Sandeep,
Hariprasad Kodamana
Validation: Pankaj Lal Sahu,
Sukumaran Sandeep,
Hariprasad Kodamana
Visualization: Pankaj Lal Sahu
Writing – original draft: Pankaj Lal Sahu
Writing – review & editing:
Sukumaran Sandeep,
Hariprasad Kodamana

the past two decades, the enhancement of global and regional models has resulted in significant advancements in track prediction, with errors reduced by about 50% since the 1990s (Cangialosi et al., 2020). The improvement in track predictions is primarily due to improved NWP model initial conditions as a result of wider satellite coverage and advanced ensemble forecasting methods (Goerss, 2009). On the other hand, intensity prediction has lagged behind track forecasts, mainly because of the inadequate representation of many physical processes influencing storm strength, including ocean interactions, atmospheric conditions, and internal storm dynamics (DeMaria et al., 2005). Recent developments in high-resolution models, model physics, data assimilation, and the addition of air-sea interactions have started to enhance the accuracy of intensity forecasts (Alaka Jr et al., 2024; Tallapragada et al., 2014).

In recent years, Machine learning-based Weather Prediction (MLWP) models such as GraphCast, FourCastNet, and Pangu-Weather have shown significant potential to improve TC and general weather forecasting. Using massive data sets and advanced deep learning techniques, these models have distinct advantages over traditional NWP systems. They enable faster computations, achieve high spatial and temporal resolutions, and often match or outperform conventional models in predicting large-scale global weather phenomena, including TCs (Anirudh et al., 2025; Hakim & Masanam, 2024). For example, GraphCast uses graph neural networks (GNNs) to effectively capture complex spatial and temporal dependencies, resulting in more accurate weather predictions (Lam et al., 2023). Similarly, FourCastNet and Pangu-Weather utilize Fourier-based neural networks and transformer-based architectures to model intricate atmospheric dynamics with impressive accuracy in a few minutes (Bi et al., 2023; Bonev et al., 2023; Pathak et al., 2022). Despite these advances, significant challenges remain, particularly in forecasting high-impact weather events. A recent analysis of Storm Ciarán, a European windstorm, reveals considerable strengths and weaknesses in their application for operational forecasting. Models such as GraphCast, FourCast-Net, and Pangu-Weather correctly predicted large-scale synoptic features. However, for this extratropical cyclone, they failed to capture finer-scale details essential for accurate weather warnings. The failure in these models included underprediction of maximum wind speeds and a lack of consistency in the warm core representation (Charlton-Perez et al., 2024). These results show that MLWP models can match or even outperform traditional NWP systems in predicting large-scale storm structures but cannot resolve small-scale features. This limitation arises from several factors, including the quality, diversity, and size of their training data sets and their inability to incorporate physical constraints into their architectures fully. In addition, their performance in different ocean basins and under changing climate conditions still requires rigorous validation and further refinement (Mu et al., 2025; Thuemmel et al., 2023; Wu et al., 2022).

While MLWP models have demonstrated substantial progress in TC prediction, several research gaps remain unaddressed. First, most existing studies have evaluated these models using a very limited number of cyclone cases, which restricts drawing broader conclusions from the findings. For instance, Bouallègue et al. (2024) assessed PanguWeather for TC tracks and intensity and limited their TC analysis to 2018 data, raising concerns about temporal generalizability. TCs exhibit diverse structures and behaviors depending on their genesis location, seasonality, and environmental conditions, necessitating a broader range of case studies, including basinwide analyses, for robust assessment (Bhatia et al., 2019; Elsner et al., 2008). Each basin has unique environmental and climatological factors influencing cyclone behavior, and a lack of such studies limits the models' applicability in diverse conditions (Emanuel, 2013; Roy & Kovordányi, 2012). Second, the warm-core thermal structure, characterized by higher temperatures at the center of the cyclone, is vital for maintaining low pressure and sustaining the storm's energy (Brueske & Velden, 2003). The development of this warm core is closely linked to diabatic heating, primarily from latent heat release during convection (Sahu et al., 2024; Vigh & Schubert, 2009) and the conservation and amplification of vorticity are essential for the intensification of cyclones. Diabatic processes, such as heating, contribute to changes in potential vorticity, creating conditions favorable for tropical cyclogenesis (Raymond, 2012). The interaction between the thermal structure and vorticity is central to maintaining thermal wind balance, a state necessary for cyclone stability and growth. The temperature adjustments and distribution of vorticity within the vortex core influence the storm's wind structure and intensity evolution (Peng & Fang, 2021). These essential parameters have been largely overlooked in current MLWP model evaluations.

This study aims to address critical research gaps in ML-based TC prediction by focusing on underexplored areas essential for enhancing model reliability and operational applicability. Recognizing the limitations of existing studies that often rely on a narrow set of cyclone cases, this research emphasizes the need for basin-wide analyses to capture the diverse structures and behaviors of TCs influenced by unique environmental and climatological

factors across different genesis locations. Furthermore, it seeks to integrate key physical parameters, including the warm-core thermal structure and vorticity. By addressing these gaps, the study aims to help advance the development of robust and versatile AI-driven forecasting models.

2. Data and Methods

This study utilizes four MLWP models on $0.25° \times 0.25°$ horizontal resolution, namely, FourCastNetv2, GraphCast operational, Aurora pretrained, and Pangu-Weather 6-hr model. All of these models were trained on extensive reanalysis data sets to generate high-resolution weather forecasts. Specifically, FourCastNetv2 was trained on data from 1979 to 2015 and validated using data from 2016 to 2017 (Bonev et al., 2023). GraphCast-operational was pre-trained on data spanning from 1979 to 2017 and fine-tuned with HRES data from 2016 to 2021 (Lam et al., 2023). Pangu-Weather was trained on data from 1979 to 2017 and validated and tested using 2018–2021 data (Bi et al., 2023), and Aurora was pre-trained using 1979 to 2020 data (Bodnar et al., 2024). We generated forecasts of 50 TCs across five ocean basins using all three MLWP models. The details of the TC cases and the forecast durations are provided in Tables S1 to S5 in Supporting Information S1.

All four models rely on ERA5 reanalysis data with a $0.25° \times 0.25°$ resolution, producing outputs for various meteorological variables. This study focuses on a subset of these variables, including mean sea level pressure (MSLP), 10-m wind components ($u10$, $v10$), wind components at 500–200 hPa (u, v), geopotential at 850–200 hPa (z) and temperatures at 1000–200 hPa. ERA5 reanalysis hourly data (Hersbach et al., 2020), with the same resolution, was also used for the comparison of absolute vorticity and temperature anomalies across the models. The models are computationally efficient and capable of generating 96-hr forecasts in a few minutes on a single 40 GB NVIDIA A100 GPU.

The TC best-track observational data used in this study were obtained from the IBTrACS v4 data set (Gahtan et al., 2024; Knapp et al., 2010). This comprehensive data set includes information such as the latitude and longitude of the TC center, maximum sustained near-surface wind speeds, and central pressure values. For evaluation purposes, data from 2022 were used, as some models were tested up to 2021. Tracks shorter than four days were excluded from the analysis, and 10 TC cases were selected per basin, these tracks are randomly selected for all basins except for the North Indian (NI) basin, which had only 10 available cases. To evaluate the accuracy of the MLWP models compared to operational TC forecasts, three global NWP models from the THORPEX Interactive Grand Global Ensemble (TIGGE) project (Bougeault et al., 2010) were chosen. These models, developed by the National Centers for Environmental Prediction, Global Forecast System (GFS), ECMWF Integrated Forecasting System (IFS), and the United Kingdom Meteorological Office, Unified Model (UM), taken at horizontal resolution of $0.25° \times 0.25°$ to fairly compare with MLWP models. In TIGGIE, GFS data is archived in $0.5° \times 0.5°$, IFS data in octahedral reduced Gaussian grid (O640) and UM data in 20 km resolution. The forecast data from these models are provided at vertical pressure levels of 1,000, 925, 850, 700, 500, 300, 250, and 200 hPa.

The TC tracks are detected following the methodology by Bourdin et al. (2022). The tracking technique is briefly described below.

1. Candidate Detection
   - *Local Minima Identification*: Potential cyclone centers are identified as local minima in the MSLP field by comparing each grid point to its eight neighbors.
2. Tracking
   - *Track Initialization*: Tracks are initiated from valid local minima detected in the first timestep, storing attributes like time, pressure, latitude, and longitude.
   - *Track Propagation*: In subsequent timesteps, new local minima within a 250 km radius of the last known position are added to active tracks.
   - *Track Termination*: Tracks without updates in later timesteps are finalized and stored.

The tracks of the TCs obtained by applying the above algorithm on the ERA5 reanalysis, NWP, and MLWP forecasts are shown in Figures 1 and 2. Further, we computed divergence ($\nabla \cdot \mathbf{V}$) at 850 hPa, absolute vorticity (η) and absolute vorticity advection (A_η) in the mid-troposphere (500–200 hPa) to understand the dynamics of TCs (Chan, 2005).
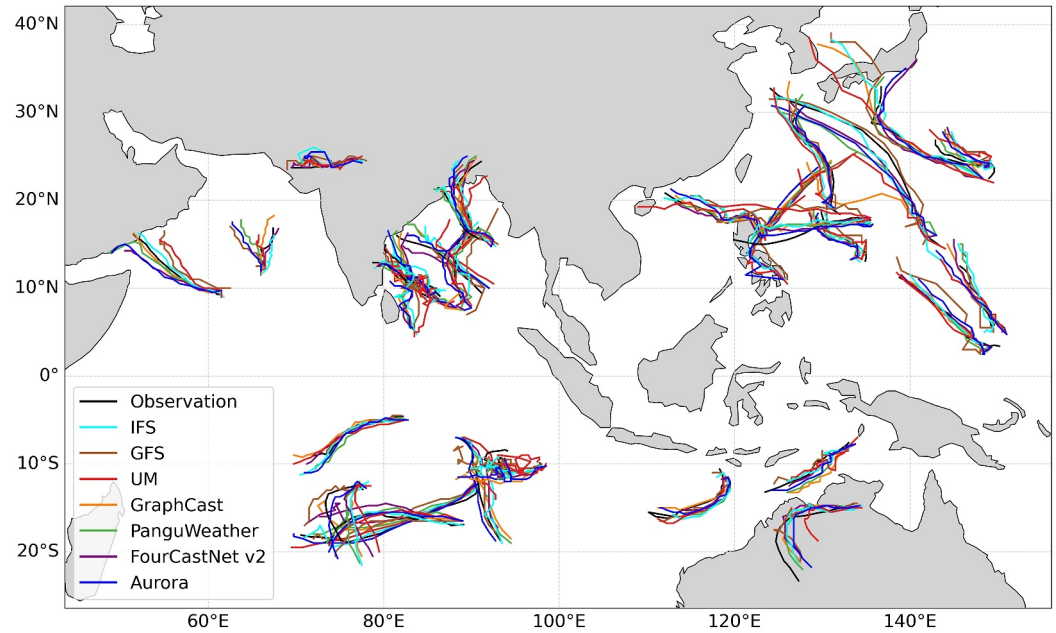
Figure 1. Observed (black) and model-predicted (colored) tropical cyclone tracks in the Northern Indian Ocean, Southern Indian Ocean, and Western North Pacific.

$$\nabla \cdot \mathbf{V} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \tag{1}$$

$$\eta = \zeta + f = \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right) + 2\Omega \sin(\phi) \tag{2}$$

$$A_\eta = -\left(u \frac{\partial \eta}{\partial x} + v \frac{\partial \eta}{\partial y}\right) \tag{3}$$

Where ζ is the relative vorticity, f is the Coriolis parameter, u and v are the wind components in the zonal and meridional directions, Ω is Earth's angular velocity, ϕ is the latitude.

Additionally, the temperature anomaly was calculated to better understand the storm's intensity, structure, and thermal characteristics (Durden, 2013). The temperature anomaly is computed as follows: a $10° \times 10°$ grid box is defined around the TC center and the average temperature within this box is computed for each pressure level. The mean temperature at each pressure level is then subtracted from the temperature at each grid point within the box to yield the temperature anomaly.
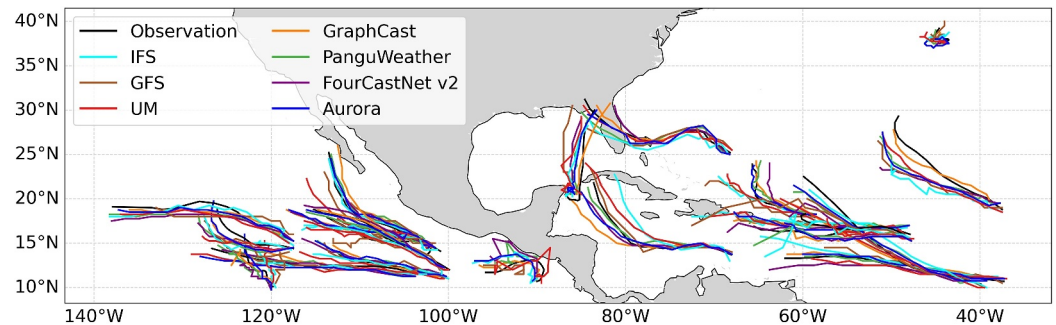

Figure 2. Same as Figure 1 but for the Eastern Pacific and Atlantic basins.
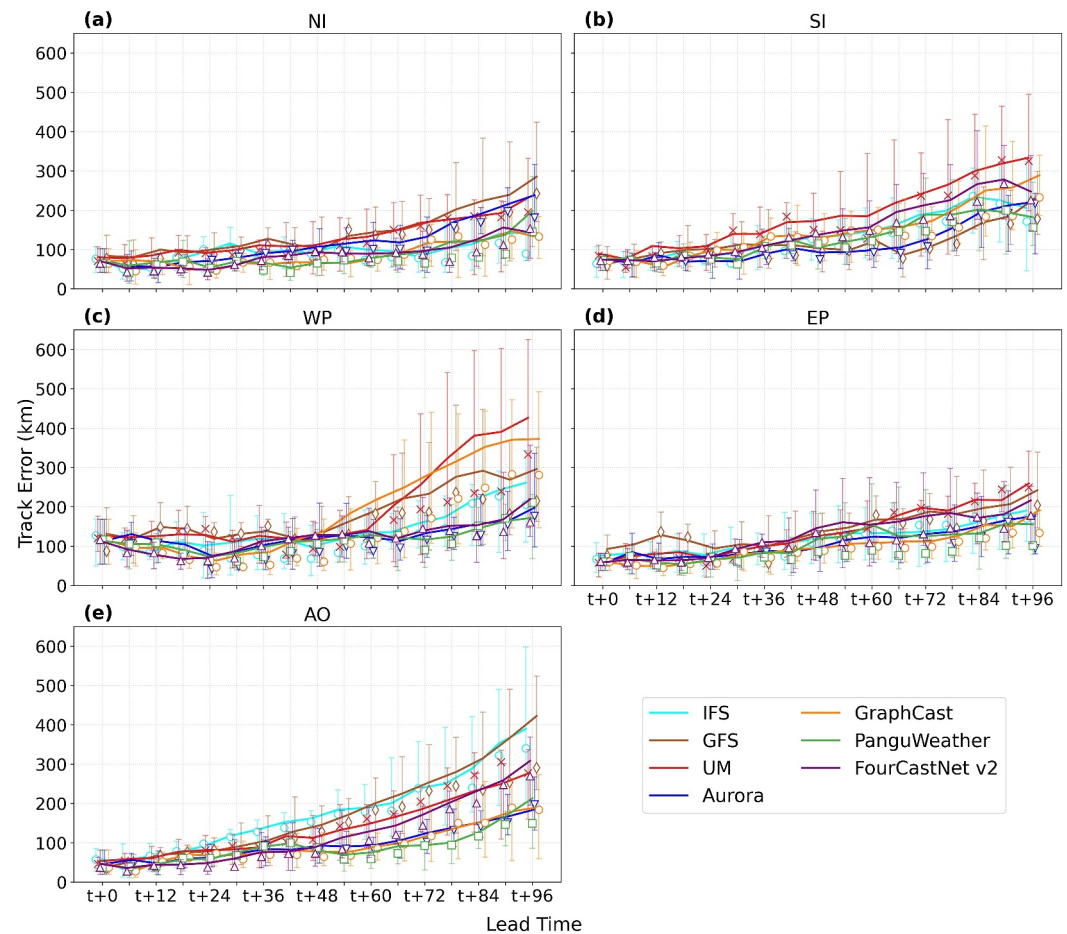
Figure 3. Track error (km) for various models across different ocean basins: (a) North Indian, (b) South Indian, (c) Western North Pacific, (d) Eastern Pacific, and (e) Atlantic Ocean. The x-axis represents lead time ($t + 0$ to $t + 96$ hr), the line plot shows the mean, the marker represents the median, and the error bars indicate the interquartile range (80–20 percentile).

3. Results and Discussion

The mean track error increases with forecast lead time across all ocean basins (Figure 3). Aurora, GraphCast, PanguWeather, and FourCastNet v2 consistently show superior performance, achieving lower errors compared to the physics-based numerical models like GFS, IFS, and UM. In the North Indian (NI), South Indian (SI) and Eastern Pacific (EP) regions, the errors increase steadily with relatively small variability, as indicated by narrower error bars. The Western Pacific (WP) region exhibits the steepest rise in errors and the highest variability, with GFS, IFS, UM, and GraphCast showing particularly poor performance, while Aurora, PanguWeather, and FourCastNet v2 maintain a clear advantage. This performance gap is largely attributed to the failure to properly capture the recurvature of TC Meari, which significantly influenced track prediction. In the EP region, all models perform relatively closely, but Aurora, GraphCast and PanguWeather retain a slight edge, especially at longer lead times, while FourCastNet v2 shows competitive results at shorter lead times. In the Atlantic Ocean (AO) region, a steady increase in error with lead time is observed, with Aurora, PanguWeather and GraphCast again achieving the best performance, and IFS and GFS showing higher errors, particularly at lead times beyond 24 and 48 hr respectively. Overall, these results highlight the advancements of ML-based models, particularly PanguWeather and GraphCast, in reducing mean track error.

The horizontal movement of TCs is fundamentally governed by vorticity advection, with environmental steering flows exerting significant influence through their modulation of this process (Y. Wang & Holland, 1996; Chan, 2005). Thus, we computed the vorticity advection for ERA5 and all forecast models. Storm-centered composites of absolute vorticity and its advection are constructed for the phase of maximum TC intensity for
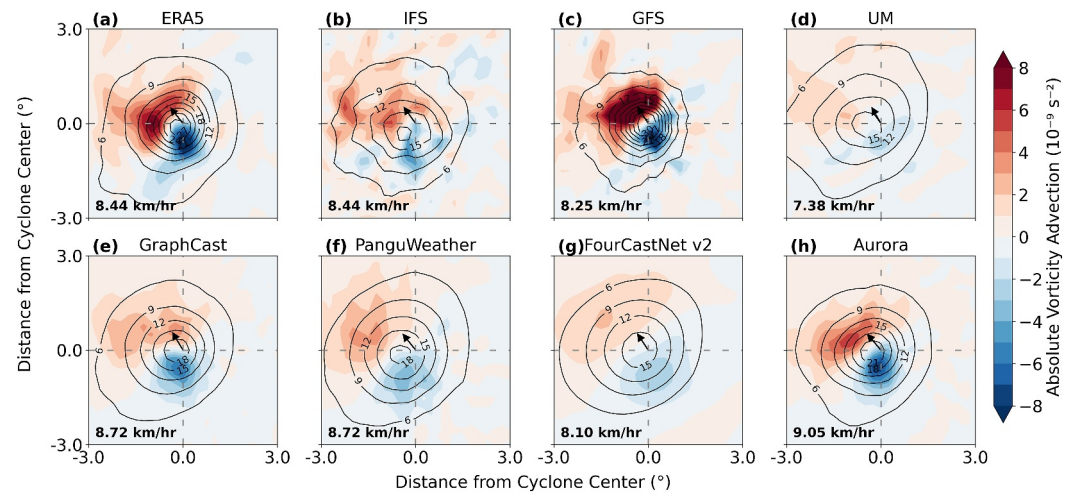
Figure 4. Storm-centered composite plots of absolute vorticity advection (10^{-9} s^{-2}, shading) and absolute vorticity (10^{-5} s^{-1}, contours) averaged in the mid-troposphere (500–200 hPa) over the North Indian Ocean basin for (a) ERA5, (b) Integrated Forecasting System, (c) Global Forecast System, (d) Unified Model, (e) GraphCast, (f) PanguWeather, (g) FourCastNet v2, and (h) Aurora. The arrow at the center shows the mean translational speed and magnitude written in bold.

the NI Ocean basin (Figure 4). ERA5, NWP, and MLWP models show typical vorticity patterns associated with TCs. ERA5 and GFS model forecast show strong vorticity advection, while the IFS and UM model show a weak vorticity advection. The MLWP models captured the pattern of vorticity advection though the magnitudes are weaker than ERA5 and GFS. Among the MLWP models, FourCastNet v2 composite shows weaker vorticity advection while Aurora shows more stronger advection than other three. Similar results are obtained for other ocean basins as well (Figures S1–S4 in Supporting Information S1). This analysis indicates that the purely data-driven MLWP models capture the underpinning physics intuitively.

Figure 5 illustrates the errors in maximum sustained wind speed (MSW) forecasts as a function of forecast lead time for all ocean basins, comparing NWP and MLWP models. The MSW error is calculated as the difference between observed and predicted MSW values, with positive errors indicating underestimation and negative errors indicating overestimation. The MSW and MSLP errors for individual ocean basins as a function of forecast lead time are shown in Figures S5–S9 in Supporting Information S1. Both NWP and MLWP models exhibit increasing MSW errors with longer forecast lead times. Regional variations in error magnitude are apparent, with some regions showing larger forecast discrepancies than others. The GFS model generally demonstrates relatively low errors at shorter lead times. However, a notable observation is that all MLWP models exhibit larger MSW prediction errors compared to NWP models. For instance, while the GFS model might show mean errors of around 5–10 m s^{-1} at $t + 72$ hr in some regions, the MLWP models display errors as high as 15–18 m s^{-1}. The underestimation of TC intensities by MLWP models is consistent across all ocean basins.

This underprediction aligns with similar issues reported in capturing the intensity of heat extremes by MLWP models, as noted by Pasche et al. (2025). One contributing factor to these errors could be inaccuracies in the training data, as evidenced by the MSW errors in ERA5 shown in Figure 5. Another potential factor is the inherent double penalization issue in ML models caused by the minimization of the mean squared error. Further investigation is needed to pinpoint the relative contributions of biases in the training data and double penalization to the large errors in TC intensity predictions by MLWP models.

The spatial structure of the low-level wind speeds associated with TCs predicted by the models provides valuable insights into the errors in MSW. Storm-centered composites of 10 m wind speeds and MSLP were constructed for ERA5 reanalysis, NWP models, and MLWP models for TCs over the NI Ocean basin for all the TCs at their peak intensity (Figure 6). These composites correspond to the time when the TCs achieved maximum intensity. The ERA5 reanalysis and all models captured the closed isobaric structure of the MSLP pattern associated with TCs. The central minimum pressure in ERA5 is approximately 988 hPa, while that of the GFS model is 982 hPa. The mean intensity of the TCs forecasted by the IFS and UM model is notably weaker, with a central minimum MSLP of about 992 and 996 hPa respectively. The central minimum MSLP forecasted by the Aurora, GraphCast and
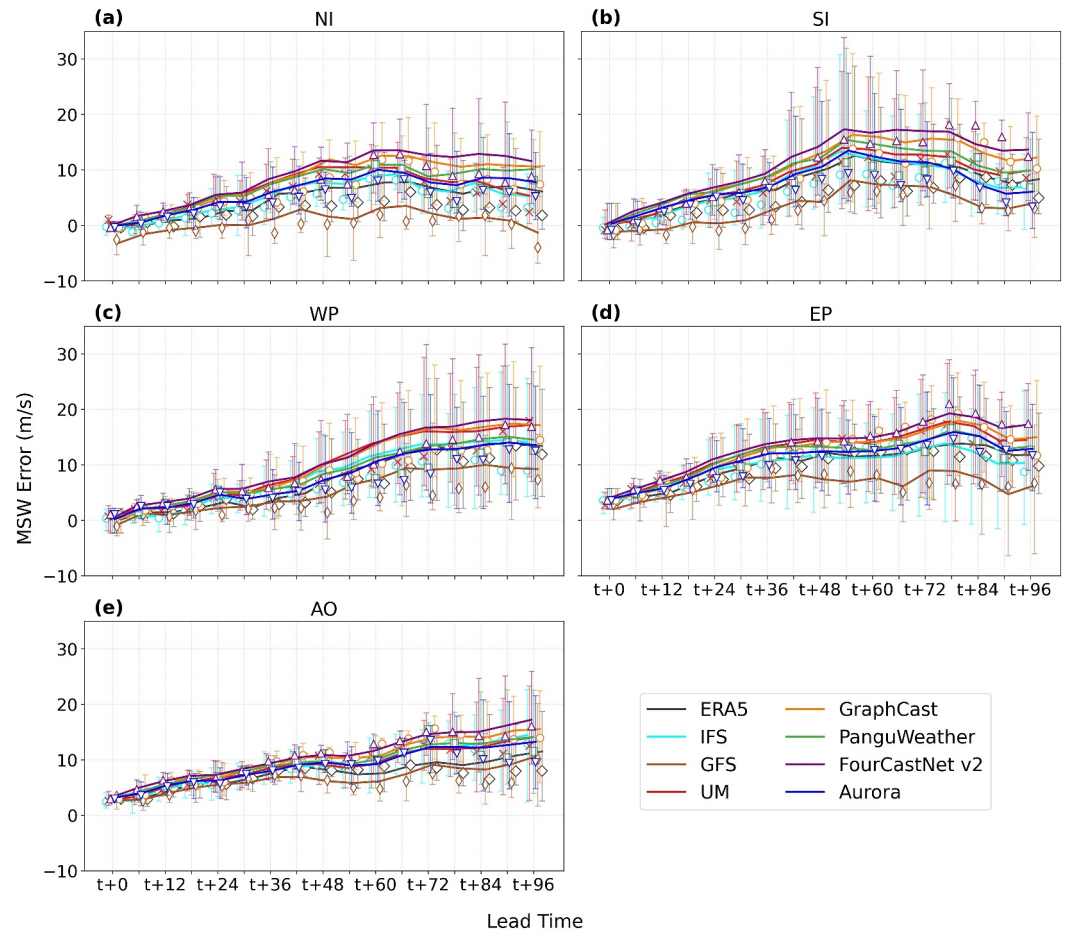
Figure 5. Maximum sustained wind speed error (m s^{-1}) for various models across different ocean basins: (a) North Indian, (b) South Indian, (c) Western North Pacific, (d) Eastern Pacific, and (e) Atlantic Ocean. The *x*-axis represents the forecast lead time ($t + 0$ to $t + 96$ hr), the line plot shows the mean, the marker represents the median, and the error bars indicate the interquartile range (80–20 percentile). Positive errors indicate model underestimation relative to observation, while negative errors indicate overestimation.
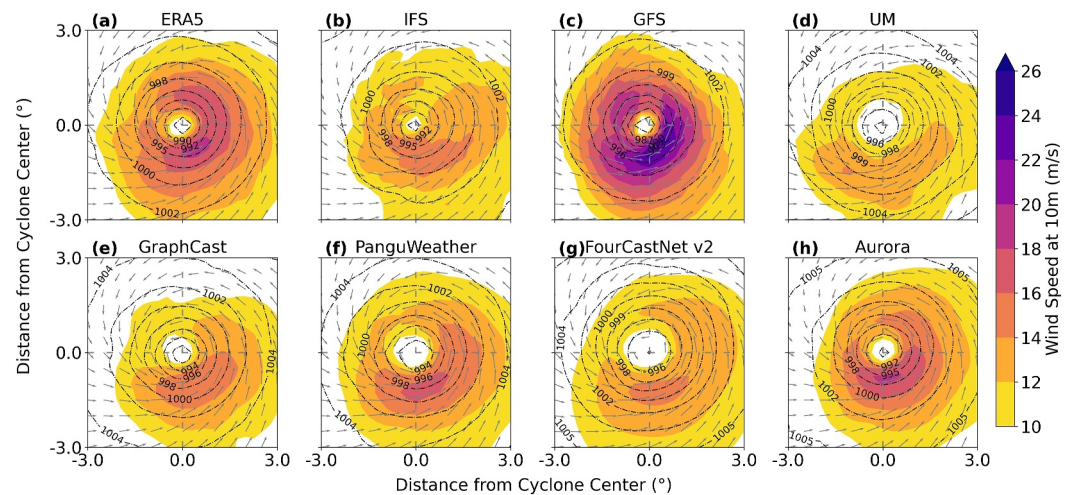

Figure 6. Storm-centered composite plots of 10 m wind speed (m/s, shading) and mean sea level pressure (hPa, contours) over the North Indian Ocean basin for (a) ERA5, (b) Integrated Forecasting System, (c) Global Forecast System, (d) Unified Model, (e) GraphCast, (f) PanguWeather, (g) FourCastNet v2 and (h) Aurora.

Figure 7. Storm-centered composite plots of divergence (10^{-5} s^{-1}, shading) and wind speed (>15 m/s, quiver) at 850 hPa over the North Indian Ocean basin for (a) ERA5, (b) Integrated Forecasting System, (c) Global Forecast System, (d) Unified Model, (e) GraphCast, (f) PanguWeather, (g) FourCastNet v2 and (h) Aurora. The blue shading shows convergence.
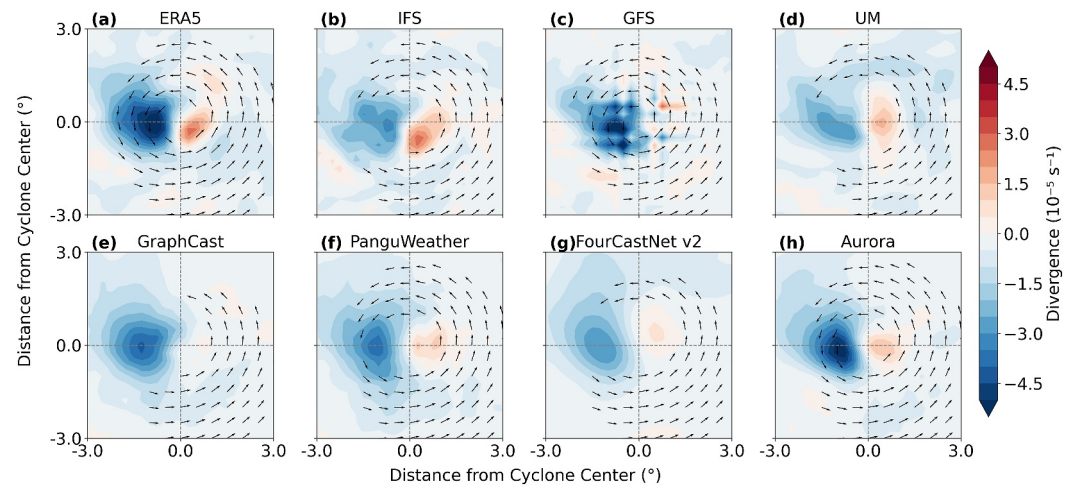
PanguWeather models is around 992 hPa, whereas the FourCastNetv2 model predicts a slightly weaker intensity with a central minimum MSLP of 996 hPa.

The spatial patterns of 10 m wind speeds were consistent with the MSLP patterns, with the GFS model forecasting stronger wind speeds compared to other models and ERA5 reanalysis. A comparison of the forecasted MSW and minimum MSLP with corresponding observations suggests that ERA5 underestimated the intensity of TCs over the NI Ocean (Figure S5 in Supporting Information S1). MSLP and 10 m wind speed composites for TCs in other ocean basins also indicate that the GFS model outperformed ERA5 and all other forecast models, except in the Atlantic Ocean, where ERA5 showed slightly stronger TCs (Figures S10–S13 in Supporting Information S1). Further insight into the model performance can be obtained by directly comparing the geopotential with the MSLP as the total air column pressure derived from the hydrostatic equation provides a robust validation of surface pressure estimates (H.-Y. Liu et al., 2021). In Figures S14–18 in Supporting Information S1, geopotential averaged at 850–200 hPa and MSLP are evaluated together for all the ocean basins. The underlying logic is that the geopotential height is directly related to the mass distribution in the atmospheric column, and in a well-resolved atmosphere, this should be consistent with the surface pressure and in all MLWP models there was a close correspondence between the predicted MSLP and the spatial variations in geopotential. This consistency indicates that the pressure fields, are accurately represented in the models, while they have underestimation issues.

All forecast models, including the MLWP models, exhibit a clear negative slope in the MSLP–MSW scatter plots, indicating that lower central pressures generally coincide with stronger surface winds (Figure S19 in Supporting Information S1). This robust inverse relationship between MSLP and MSW is consistent across different ocean basins (Bai et al., 2019; Carstens et al., 2024). Notably, the GFS forecasts tend to produce stronger wind speeds than both ERA5 reanalysis and the other forecast models, while ERA5 consistently underestimates TC intensity in most basins as shown in Figure 5. This shows the MLWP models reproduce the fundamental pressure–wind coupling with high fidelity. Reanalysis data, often considered the ground truth in the absence of direct observations, serves as the training data set for MLWP models. These analyses strongly suggest that the large errors in TC intensity predicted by the MLWP models stem from the underestimation of TC intensities in the training data.

Since most of the MLWP models do not provide vertical velocity explicitly, looking at low-level divergence/convergence patterns can serve as a useful proxy for the strength of the vertical motion. Stronger inflow and convergence in the boundary layer often imply more vigorous updrafts and, thus, a potentially deeper warm core aloft (Willoughby, 1990). Figure 7 highlights intermodel differences in TC dynamics through horizontal wind vectors and divergence within a 3° radius of the center. MLWP models, like Aurora, PanguWeather, and Fourcastnet, resolve the convergence and divergence pattern in lower levels but underestimate the magnitudes. At the same time, GraphCast shows the divergence pattern as very weak compared to convergence, and the GFS
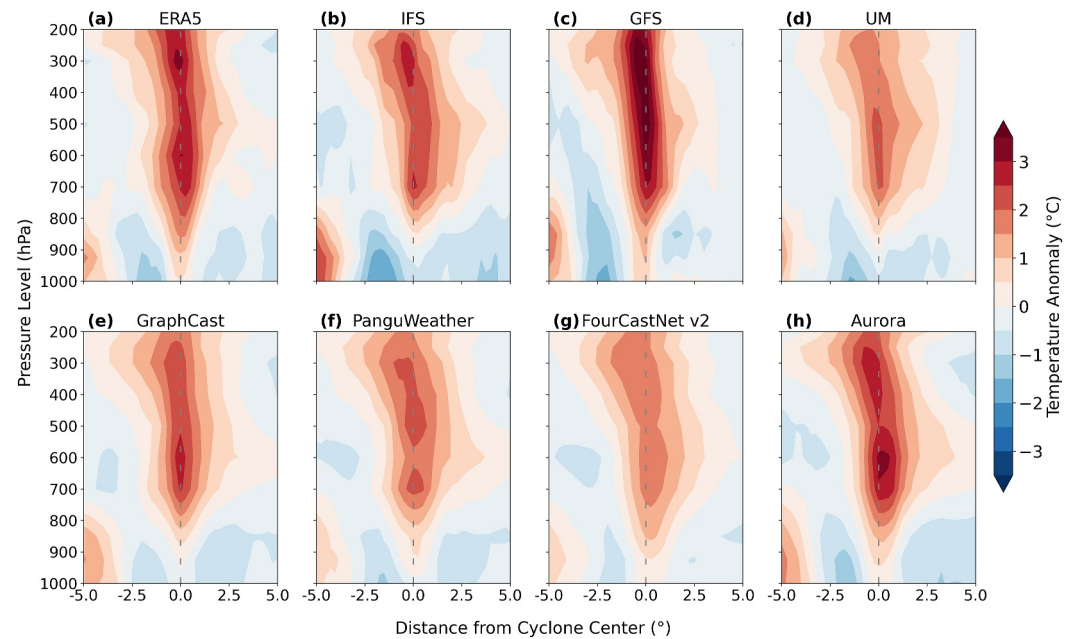
Figure 8. Vertical cross-sections of the temperature anomalies (°C) over the North Indian Ocean basin for (a) ERA5, (b) Integrated Forecasting System, (c) Global Forecast System, (d) Unified Model, (e) GraphCast, (f) PanguWeather, (g) FourCastNet v2 and (h) Aurora. Shading represents composite temperature deviations, with red indicating positive anomalies and blue indicating negative anomalies.

model shows strong convergence/divergence magnitude, but its pattern was not consistent across all ocean basins (Figures S20–23 in Supporting Information S1). These discrepancies directly impact MSW and MSLP spread; concentrated low-level convergence near the core corresponds to stronger updrafts, lower MSLP, and elevated MSW, aligning with TC intensification. Thermal wind balance links warm-core intensity to convergence strength. Weak convergence in models reduces warm-core magnitude, propagating errors to TC intensity.

Several recent studies have examined the skill of standalone ML models and global MLWP models in predicting TC intensities. Yang et al. (2024) utilized a deep learning model trained on satellite images to predict extreme rainfall associated with TCs, achieving an accuracy of 87% for TC rainfall intensity predictions within a 0–120 min window. Errors in TC intensity within ERA5 reduced when it was downscaled to high resolution using a generative super-resolution model (Lockwood et al., 2024), indicating that ERA5's resolution may be inadequate to capture the finer features of TCs. An evaluation of global MLWP models for the Eastern and Western North Pacific basins also showed substantial underestimation of TC intensity when initialized with ERA5 data (C. Liu et al., 2024). Also, the double penalization effect inherent in the RMSE loss function of this MLWP models plays a critical role in weak intensity biases (Subich et al., 2025). However, a hybrid ML-physics model, such as the framework proposed by H.-Y. Liu et al. (2024), may further improve TC intensity prediction. This approach integrates the Pangu-Weather model for large-scale circulation and TC track forecasting with the high-resolution WRF model for simulating core physical processes. It also incorporates spectral nudging to refine large-scale circulation and an ocean mixed-layer model to update sea surface temperatures. Such a hybrid framework has been shown to enhance long-range TC forecasts by reducing track and intensity errors compared to standalone MLWP models or traditional physics-based NWP models.

A key feature of TCs is their vertical thermal structure. It is important to examine whether the MLWP models can capture the thermodynamic structure of TCs, as they did with TC dynamics. To this end, we computed the storm-centered horizontal temperature anomalies from the temperature fields forecasted by the models and compared them with those computed from ERA5 reanalysis for the NI Ocean (Figure 8) and other ocean basins (Figures S24–S27 in Supporting Information S1), following the method described in Section 2. All models successfully captured the warm core structure, with positive temperature anomalies concentrated in the mid-troposphere (roughly between 300 and 700 hPa). The pattern and magnitude of these warm core anomalies are consistent

with the satellite-derived climatology of TC warm core anomalies (X. Wang & Jiang, 2019). However, substantial inter-model variability in the magnitude and vertical extent of the warm core anomalies was observed.

The GFS model exhibited the strongest warm core, with temperature anomalies exceeding 4°C around 200–500 hPa. ERA5 also showed pronounced warm cores, with anomalies reaching approximately 3–4°C in the same altitude range. The MLWP models, while displaying a distinct warm core, exhibited weaker temperature anomalies of 2–3°C. This weaker warm core in the MLWP models aligns with the previous observation of lower maximum wind speeds and higher central pressures, suggesting a less intense storm overall.

The vertical extent of the warm core also varied slightly among the models. In GFS forecasts, the warm anomaly extended relatively high into the troposphere, reaching up to around 200–300 hPa. In contrast, the warm core in the other models was more confined to the mid-troposphere. Additionally, a slight cold anomaly near the surface (below 850 hPa) was evident. This surface cooling can be attributed to evaporative cooling from rainfall. The more pronounced vorticity advection pattern in the GFS model indicates more vigorous vertical motions, leading to a stronger warm core in that model. However, previous studies (Bonavita, 2024) have shown that MLWP models tend to smooth out fine-scale atmospheric structures, leading to discrepancies in the representation of sub-synoptic and mesoscale weather phenomena. Their forecasts exhibit notably different energy spectra compared to both their training reanalysis fields and traditional NWP models, which can result in overly smooth predictions. This smoothing effect may contribute to weaker intensity biases observed in MLWP models, as it dampens the sharp temperature gradients and strong vorticity advection that drive cyclone intensification. Although the MLWP models are not explicitly designed to capture the physics associated with weather systems, they demonstrated an ability to do so, as evidenced by the consistency between their relatively weaker vorticity advection and the thermodynamic structure of TCs.

The performance disparities in predicting dynamical and thermodynamical features across MLWP models stem from differences in architectural design and training data heterogeneity. While GraphCast, PanguWeather, and FourCastNetV2 rely solely on ERA5 reanalysis data, Aurora employs a multimodal pretraining strategy on six diverse weather and climate data sets: ERA5 reanalysis, CMCC climate simulations, IFS-HR high-resolution forecasts, HRES operational forecasts, GFS analysis, and GFS forecasts. This composite data set exposes Aurora to a broader spectrum of atmospheric regimes, temporal scales, and operational biases, enabling it to learn generalizable representations of atmospheric dynamics (Bodnar et al., 2024). GraphCast's GNN excels at multi-scale interactions critical for TC genesis and steering flows. However, it's fine-tuning on high-resolution HRES operational forecasts may enhance its ability to capture small-scale atmospheric processes, particularly in high-impact weather events, potentially explaining its improved performance compared to PanguWeather and Four-CastNetV2 (Lam et al., 2023). PanguWeather's hierarchical transformer prioritizes medium-range stability through temporal aggregation, while FourCastNetV2's Fourier Neural Operators (FNOs) enable rapid global modeling but lack finer scale precision (Bi et al., 2023; Pathak et al., 2022). In contrast, Aurora's 3D Swin Transformer architecture is pre-trained on standardized meteorological variables across heterogeneous data sets, resolutions, and pressure levels, minimizing a composite loss function that harmonizes discrepancies between forecast/analysis pairs and climate simulations. By assimilating operational forecast errors (e.g., GFS/HRES) alongside reanalysis ground truths (ERA5), Aurora likely internalizes correction mechanisms for convective processes and thermodynamic couplings, explaining its superior skill in rapid intensification scenarios. Further investigation is needed to disentangle how pretraining data diversity, particularly the inclusion of climate simulations (CMCC) and operational forecasts, complements architectural innovations to enhance robustness across weather regimes.

4. Conclusions

Recent advancements in global MLWP models have demonstrated comparable or better forecast accuracy relative to traditional physics-based NWP models. However, these MLWP models have not been rigorously tested for their skill in predicting extreme weather phenomena, such as TCs, across global ocean basins. Furthermore, questions remain about whether MLWP models capture the underlying physics of weather systems, as they are not explicitly designed to do so.

In this study, we evaluated the skill of four MLWP models—GraphCast, PanguWeather, Aurora, and Four-CastNet v2—against three NWP models, GFS, IFS and UM, in forecasting TCs globally. Our analysis revealed that the MLWP models achieved high skill in predicting TC tracks, as evidenced by low track errors. However, all

MLWP models exhibited a pronounced low-intensity bias in both maximum sustained wind speeds and minimum MSLP. A comparison of TC intensities in ERA5 reanalysis with observations revealed a similar low bias, suggesting that errors in the training data likely contributed to the MLWP models' TC intensity prediction errors. Addressing these data quality issues could substantially enhance the accuracy of TC intensity forecasts by MLWP models. This study did not explore the impact of double penalization, a common issue in ML models, on TC intensity forecast errors, presenting an avenue for future investigation.

We also investigated whether the MLWP models capture the physics of TCs. Despite being data-driven and not explicitly designed to model physical processes, these models were able to reasonably predict the horizontal structure of sea level pressure, low-level convergence, absolute vorticity, and vorticity advection. Additionally, they captured the vertical structure of warm-core anomalies associated with TCs. The consistency in magnitude between various physical fields in the MLWP models suggests that they intuitively learned the interrelationships among different physical fields during the evolution of weather systems. Among MLWP models, Aurora outperformed other three MLWP models in capturing TC physics. This advantage likely stems from the Aurora's training on variety of data sets.

Our results highlight the promise of MLWP models in forecasting extreme weather events, such as TCs, and underscore the need for improved training data sets. Incorporating higher-quality training data or coupling physics-based principles with deep learning models could further enhance the predictive capabilities of MLWP models, particularly for extreme weather events.

Data Availability Statement

The IBTrACS-WMO v4 data set is available from Gahtan et al. (2024). The TIGGE data set is available from Bougeault et al. (2010). ERA5 hourly data are available at Hersbach et al. (2023). The MLWP forecasts are produced using the ai-models toolbox developed by ECMWF (Raoult et al., 2024). All plots are produced using Matplotlib (Hunter, 2007). All TC track data used in this study are available publicly at Sahu and Sandeep (2025).

References

Alaka, G. J., Jr., Sippel, J. A., Zhang, Z., Kim, H.-S., Marks, F. D., Tallapragada, V., et al. (2024). Lifetime performance of the operational hurricane weather research and forecasting model (hwrf) for North Atlantic tropical cyclones. *Bulletin of the American Meteorological Society*, *105*(6), E932–E961. https://doi.org/10.1175/bams-d-23-0139.1

Anirudh, K. M., Raj, P., Sandeep, S., Kodamana, H., & Sabeerali, C. T. (2025). A skillful prediction of monsoon intraseasonal oscillation using deep learning. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(2). https://doi.org/10.1029/2024JH000504

Bai, L., Yu, H., Black, P. G., Xu, Y., Ying, M., Tang, J., & Guo, R. (2019). Reexamination of the tropical cyclone wind–pressure relationship based on pre-1987 aircraft data in the western North Pacific. *Weather and Forecasting*, *34*(6), 1939–1954. https://doi.org/10.1175/waf-d-18-0002.1

Bhatia, K. T., Vecchi, G. A., Knutson, T. R., Murakami, H., Kossin, J., Dixon, K. W., & Whitlock, C. E. (2019). Recent increases in tropical cyclone intensification rates. *Nature Communications*, *10*(1), 1–9. https://doi.org/10.1038/s41467-019-08471-z

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., & others (2024). Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*.

Bonavita, M. (2024). On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, *51*(12), e2023GL107377. https://doi.org/10.1029/2023gl107377

Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., & Anandkumar, A. (2023). Spherical Fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning* (pp. 2806–2823).

Bouallègue, Z. B., Clare, M. C., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., et al. (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, *105*(6), E864–E883. https://doi.org/10.1175/bams-d-23-0162.1

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., et al. (2010). The thorpex interactive grand global ensemble. *Bulletin of the American Meteorological Society*, *91*(8), 1059–1072. https://doi.org/10.1175/2010bams2853.1

Bourdin, S., Fromang, S., Dulac, W., Cattiaux, J., & Chauvin, F. (2022). Intercomparison of four algorithms for detecting tropical cyclones using ERA5. *Geoscientific Model Development*, *15*(17), 6759–6786. https://doi.org/10.5194/gmd-15-6759-2022

Brueske, K. F., & Velden, C. S. (2003). Satellite-based tropical cyclone intensity estimation using the NOAA-KLM series advanced microwave sounding unit (AMSU). *Monthly Weather Review*, *131*(4), 687–697. https://doi.org/10.1175/1520-0493(2003)131⟨0687:SBTCIE⟩2.0

Cangialosi, J. P., Blake, E., DeMaria, M., Penny, A., Latto, A., Rappaport, E., & Tallapragada, V. (2020). Recent progress in tropical cyclone intensity forecasting at the national hurricane center. *Weather and Forecasting*, *35*(5), 1913–1922. https://doi.org/10.1175/waf-d-20-0059.1

Carstens, J. D., Didlake, A. C., Jr., & Zarzycki, C. M. (2024). Tropical cyclone wind shear-relative asymmetry in reanalyses. *Journal of Climate*, *37*(22), 5793–5816. https://doi.org/10.1175/jcli-d-23-0628.1

Chan, J. C. (2005). The physics of tropical cyclone motion. *Annual Review of Fluid Mechanics*, *37*(1), 99–128. https://doi.org/10.1146/annurev.fluid.37.061903.175702

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., et al. (2024). Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of storm Ciarán. *npj Climate and Atmospheric Science*, *7*(1), 93. https://doi.org/10.1038/s41612-024-00638-w

DeMaria, M., Mainelli, M., Shay, L. K., Knaff, J. A., & Kaplan, J. (2005). Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). *Weather and Forecasting*, *20*(4), 531–543. https://doi.org/10.1175/waf862.1

Durden, S. L. (2013). Observed tropical cyclone eye thermal anomaly profiles extending above 300 hPa. *Monthly Weather Review*, *141*(12), 4256–4268. https://doi.org/10.1175/mwr-d-13-00021.1

Elsner, J. B., Kossin, J. P., & Jagger, T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature*, *455*(7209), 92–95. https://doi.org/10.1038/nature07234

Emanuel, K. A. (2013). Downscaling CMIP5 climate models shows increased tropical cyclone activity over the 21st century. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(30), 12219–12224. https://doi.org/10.1073/pnas.1301293110

Fanchiotti, M., Dash, J., Tompkins, E. L., & Hutton, C. W. (2020). The 1999 super cyclone in Odisha, India: A systematic review of documented losses. *International Journal of Disaster Risk Reduction*, *51*, 101790. https://doi.org/10.1016/j.ijdrr.2020.101790

Gahtan, J., Knapp, K. R., Schreck, C. J., Diamond, H. J., Kossin, J. P., & Kruk, M. C. (2024). International best track archive for climate stewardship (IBTrACS) project version 4r01 [Dataset]. https://doi.org/10.25921/82ty-9e16

Goerss, J. S. (2009). Impact of satellite observations on the tropical cyclone track forecasts of the navy operational global atmospheric prediction system. *Monthly Weather Review*, *137*(1), 41–50. https://doi.org/10.1175/2008mwr2601.1

Hakim, G. J., & Masanam, S. (2024). Dynamical tests of a deep-learning weather prediction model. *Artificial Intelligence for the Earth Systems*, *3*(3). https://doi.org/10.1175/aies-d-23-0090.1

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., & Thépaut, J.-N. (2023). ERA5 hourly data on single levels from 1940 to present [Dataset]. *CDS*. https://doi.org/10.24381/cds.adbb2d47

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *IEEE COMPUTER SOC*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Islam, M. T., Charlesworth, M., Aurangojeb, M., Hemstock, S., Sikder, S. K., Hassan, M. S., et al. (2021). Revisiting disaster preparedness in coastal communities since 1970s in Bangladesh with an emphasis on the case of tropical cyclone Amphan in May 2020. *International Journal of Disaster Risk Reduction*, *58*, 102175. https://doi.org/10.1016/j.ijdrr.2021.102175

Kantha, L. (2008). Tropical cyclone destructive potential by integrated kinetic energy. *Bulletin of the American Meteorological Society*, *89*(2), 219–221.

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, *91*(3), 363–376. https://doi.org/10.1175/2009bams2755.1

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421. https://doi.org/10.1126/science.adi2336

Lenzen, M., Malik, A., Kenway, S., Daniels, P., Lam, K. L., & Geschke, A. (2019). Economic damage and spillovers from a tropical cyclone. *Natural Hazards and Earth System Sciences*, *19*(1), 137–151. https://doi.org/10.5194/nhess-19-137-2019

Liu, C., Hsu, K., Peng, M. S., Chen, D.-S., Chang, P.-L., Hsiao, L.-F., et al. (2024). Evaluation of five global AI models for predicting weather in eastern Asia and western Pacific. *npj Climate and Atmospheric Science*, *7*(1), 221. https://doi.org/10.1038/s41612-024-00769-0

Liu, H.-Y., Tan, Z.-M., Wang, Y., Tang, J., Satoh, M., Lei, L., et al. (2024). A hybrid machine learning/physics-based modeling framework for 2-week extended prediction of tropical cyclones. *Journal of Geophysical Research: Machine Learning and Computation*, *1*(3), e2024JH000207. https://doi.org/10.1029/2024JH000207

Liu, H.-Y., Wang, Y., & Gu, J.-F. (2021). Intensity change of binary tropical cyclones (TCS) in idealized numerical simulations: Two initially identical mature TCS. *Journal of the Atmospheric Sciences*, *78*(4), 1001–1020. https://doi.org/10.1175/jas-d-20-0116.1

Lockwood, J. W., Gori, A., & Gentine, P. (2024). A generative super-resolution model for enhancing tropical cyclone wind field intensity and resolution. *Journal of Geophysical Research: Machine Learning and Computation*, *1*(4), e2024JH000375. https://doi.org/10.1029/2024JH000375

Mu, M., Qin, B., & Dai, G. (2025). Predictability study of weather and climate events related to artificial intelligence models. *Advances in Atmospheric Sciences*, *42*(1), 1–8. https://doi.org/10.1007/s00376-024-4372-7

Naguib, C., Pelli, M., Poirier, D., & Tschopp, J. (2022). The impact of cyclones on local economic growth: Evidence from local projections. *Economics Letters*, *220*, 110871. https://doi.org/10.1016/j.econlet.2022.110871

Needham, H. F., Keim, B. D., & Sathiaraj, D. (2015). A review of tropical cyclone-generated storm surges: Global data sources, observations, and impacts. *Reviews of Geophysics*, *53*(2), 545–591. https://doi.org/10.1002/2014RG000477

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., & Engelke, S. (2025). Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, *4*(1), e240033. https://doi.org/10.1175/AIES-D-24-0033.1

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., & others (2022). *Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators*. arXiv preprint arXiv:2202.11214.

Peng, K., & Fang, J. (2021). Effect of the initial vortex vertical structure on early development of an axisymmetric tropical cyclone. *Journal of Geophysical Research: Atmospheres*, *126*(4), e2020JD033697. https://doi.org/10.1029/2020jd033697

Raoult, B., Pinault, F., Mertes, G., Dramsch, J. S., Cook, H., & Chantry, M. (2024). AI-models (Apache-2.0 License) [Software]. *ECMWF*. https://github.com/ecmwf-lab/ai-models

Raymond, D. J. (2012). Balanced thermal structure of an intensifying tropical cyclone. *Tellus A: Dynamic Meteorology and Oceanography*, *64*(1), 19181. https://doi.org/10.3402/tellusa.v64i0.19181

Roy, C., & Kovordányi, R. (2012). Tropical cyclone track forecasting techniques—A review. *Atmospheric Research*, *104*, 40–69.

Sahu, P. L., Pattnaik, S., & Rath, P. (2024). Factors driving intensification of pre-monsoon tropical cyclones over the Bay of Bengal: A comparative study of cyclones Fani and Yaas. *Journal of the Indian Society of Remote Sensing*, *52*(10), 2191–2205. https://doi.org/10.1007/s12524-024-01930-1

Sahu, P. L., & Sandeep, S. (2025). Data for "evaluating global machine learning models for tropical cyclone dynamics and thermodynamics" [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.15036304

Seo, S. N., & Bakkensen, L. A. (2017). Is tropical cyclone surge, not intensity, what kills so many people in South Asia? *Weather, climate, and society*, *9*(2), 171–181. https://doi.org/10.1175/wcas-d-16-0059.1

Subich, C., Husain, S. Z., Separovic, L., & Yang, J. (2025). Fixing the double penalty in data-driven weather forecasting through a modified spherical harmonic loss function. *arXiv preprint arXiv:2501.19374*.

Tallapragada, V., Bernardet, L., Biswas, M. K., Gopalakrishnan, S., Kwon, Y., Liu, Q., & others (2014). Hurricane weather research and forecasting (HWRF) model: 2013 Scientific documentation. *HWRF Development Testbed Center Tech. Rep*, *99*.

Thuemmel, J., Karlbauer, M., Otte, S., Zarfl, C., Martius, G., Ludwig, N., & others (2023). *Inductive biases in deep learning models for weather prediction*. arXiv preprint arXiv:2304.04664.

Vigh, J. L., & Schubert, W. H. (2009). Rapid development of the tropical cyclone warm core. *Journal of the Atmospheric Sciences*, *66*(11), 3335–3350. https://doi.org/10.1175/2009jas3092.1

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

Wang, Y., & Holland, G. J. (1996). The beta drift of baroclinic vortices. Part I: Adiabatic vortices. *Journal of the Atmospheric Sciences*, *53*(3), 411–427. https://doi.org/10.1175/1520-0469(1996)053⟨0411:TBDOBV⟩2.0.CO

Willoughby, H. E. (1990). Gradient balance in tropical cyclones. *Journal of the Atmospheric Sciences*, *47*(2), 265–274. https://doi.org/10.1175/1520-0469(1990)047<0265:gbitc>2.0.co;2

Willoughby, H. E., Rappaport, E., & Marks, F. (2007). Hurricane forecasting: The state of the art. *Natural Hazards Review*, *8*(3), 45–49. https://doi.org/10.1061/(asce)1527-6988(2007)8:3(45)

Wu, L., Zhao, H., Wang, C., Cao, J., & Liang, J. (2022). Understanding of the effect of climate change on tropical cyclone intensity: A review. *Advances in Atmospheric Sciences*, *39*(2), 205–221. https://doi.org/10.1007/s00376-021-1026-x

Yang, N., Wang, C., & Li, X. (2024). Improving tropical cyclone precipitation forecasting with deep learning and satellite image sequencing. *Journal of Geophysical Research: Machine Learning and Computation*, *1*(2), e2024JH000175. https://doi.org/10.1029/2024JH000175