Strengths and Weaknesses of Global Machine Learning Weather Prediction Models in Forecasting Tropical Cyclones

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

[1] Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, New Delhi 110016, India. Email: asz238477@iitd.ac.in
[2] Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi 110016, India
[3] Indian Institute of Technology, Delhi-Abu Dhabi, Khalifa City, Abu Dhabi, UAE
[4] Department of Chemical Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India

1. Introduction

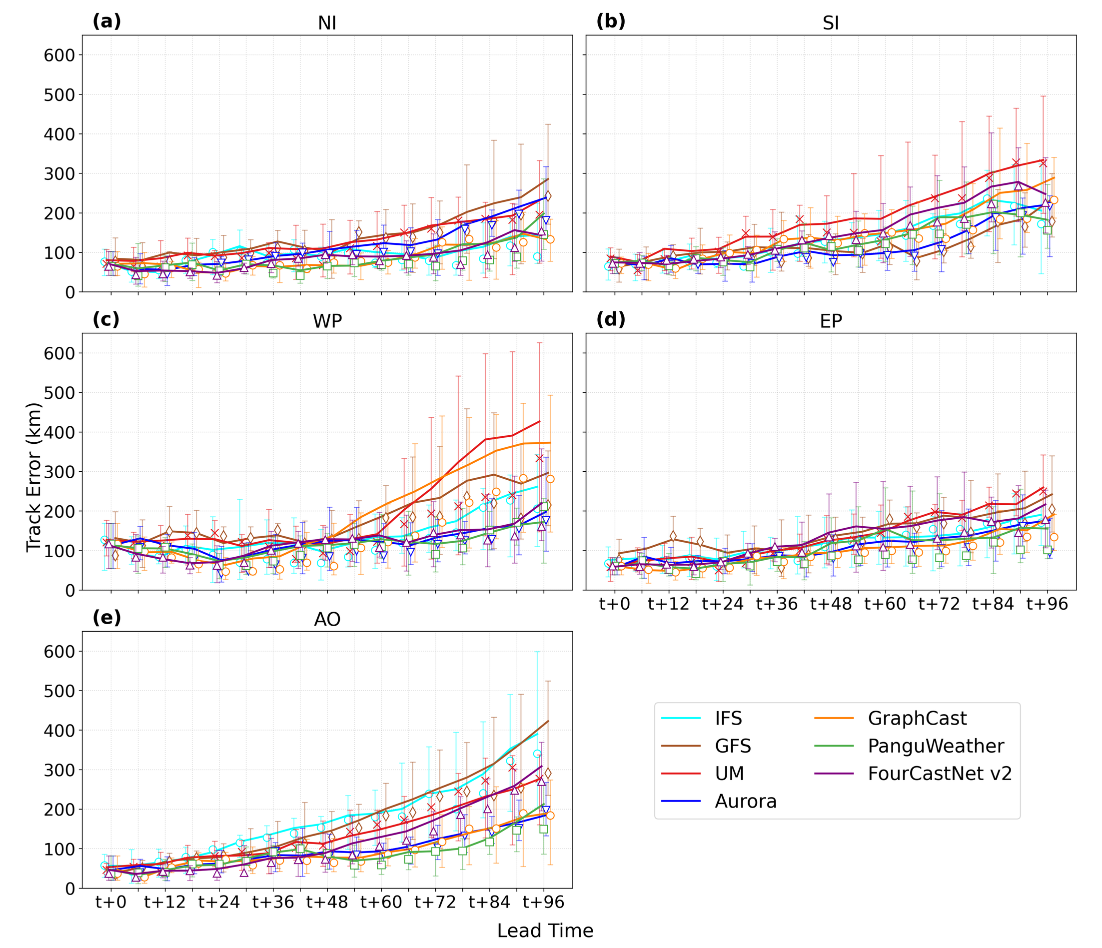
Tropical cyclones (TCs) are major hazards, demanding accurate forecasts for effective risk mitigation. While traditional Numerical Weather Prediction (NWP) models have made significant strides in track forecasting, accurately predicting storm intensity remains a persistent challenge. Machine Learning-based Weather Prediction (MLWP) models—such as GraphCast, Fourcastnet, Aurora, and Pangu-Weather—offer rapid and competitive forecasting capabilities; however, they continue to struggle with small-scale storm dynamics [1]. This study expands on basin-wide analyses to assess the performance of these MLWP models, emphasizing their ability to simulate both the dynamical and thermodynamical parameters that are essential for realistic TC prediction. The evaluation highlights the strengths and weaknesses of MLWP in capturing the physical relationships inherent to tropical cyclones. It underscores the importance of refining these models to enhance their representation of complex storm dynamics.

2. Data and Methodology

Four MLWP models—FourCastNetv2, GraphCast, Aurora and Pangu-Weather—trained on ERA5 reanalysis data (0.25°×0.25°) were used to generate high-resolution 96-hour forecasts for 50 tropical cyclones across five ocean basins. The models' outputs were compared with IBTrACS v4 best-track observational data and operational forecasts from NCEP (GFS), ECMWF (IFS) and UKMO (UM). Cyclone tracks were detected using local minima in mean sea level pressure, while key variables—including absolute vorticity, its advection, and temperature anomalies—were computed to assess storm dynamics and thermal characteristics.

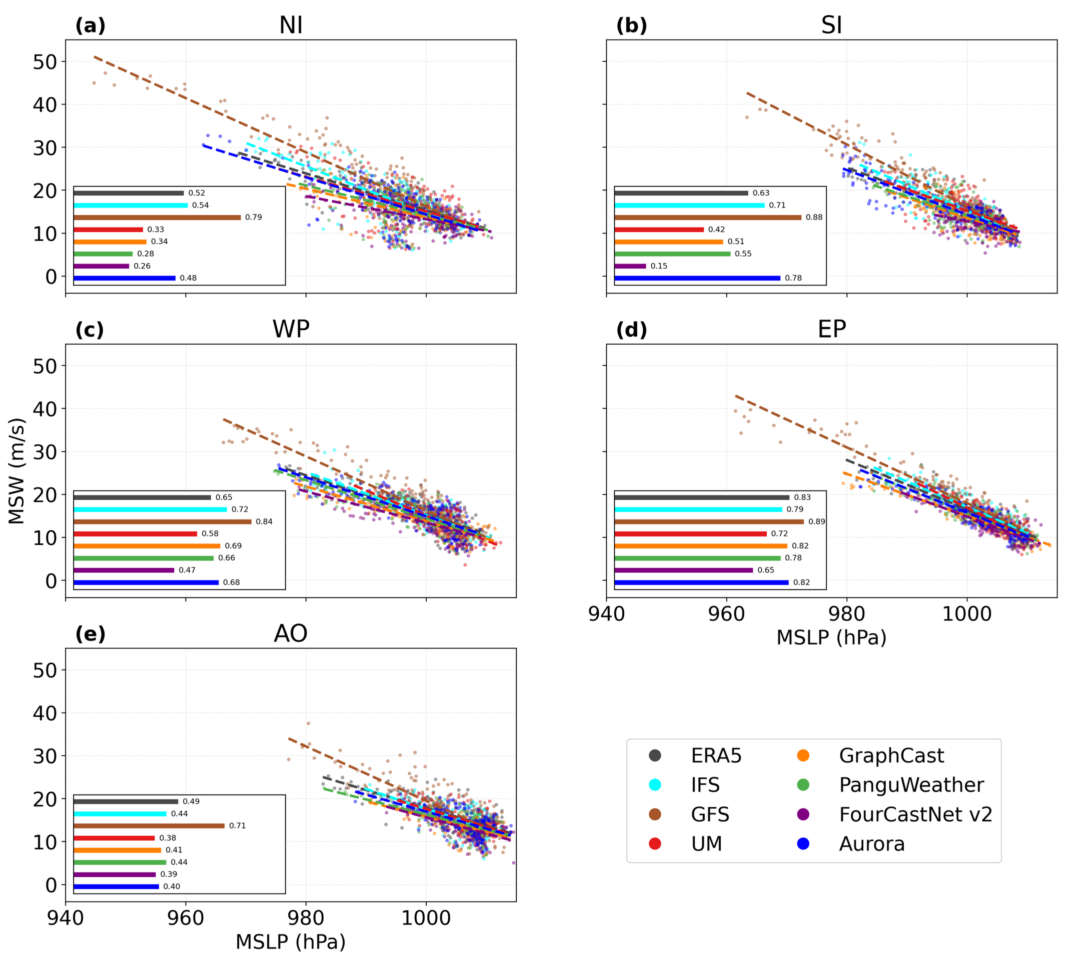
3. Track error

Mean track error increases with lead time across all basins. MLWP models consistently outperform physics-based models (GFS, IFS, UM), with particularly strong performance in the NI, SI, and EP regions. WP shows the steepest error rise and highest variability, while AO maintains a steady error increase.



4. Wind-Pressure relationship

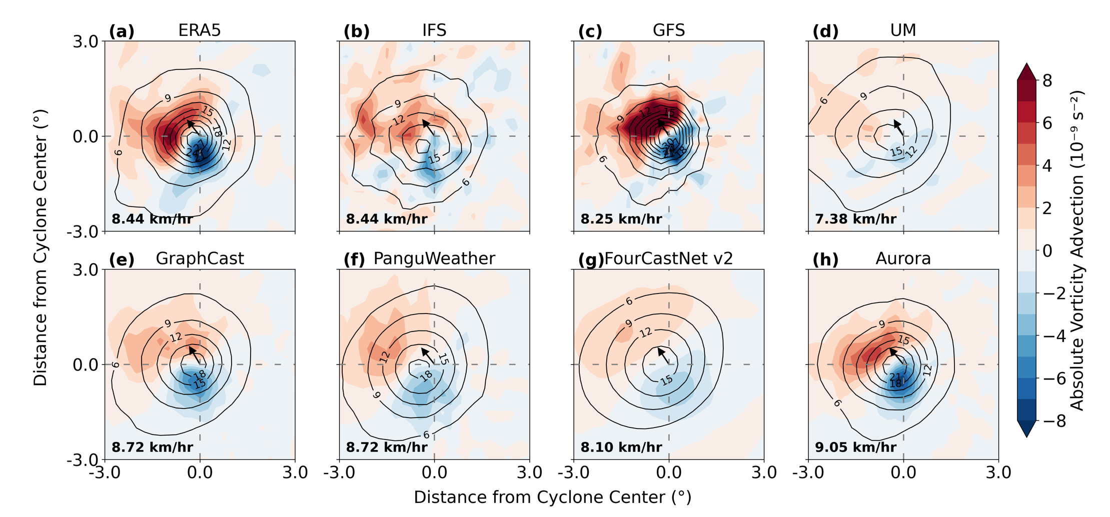
Each model successfully captures the inverse relationship between maximum sustained wind speed (MSW) at 10 m height and central minimum pressure, with steeper best-fit curves reflecting a more accurate representation of this fundamental balance. Notably, the GFS model demonstrated the highest performance in every ocean basin. Meanwhile, the MLWP models consistently maintained the physical balance between wind speed and pressure, effectively capturing the essential dynamics of tropical cyclones.



5. Absolute Vorticity

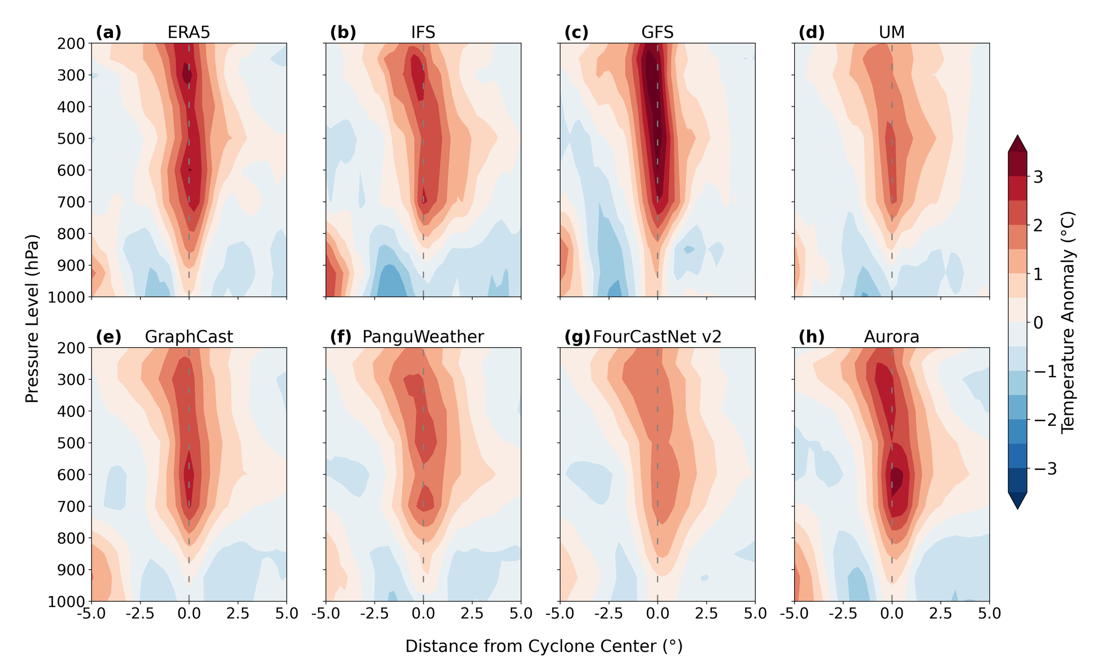
The horizontal motion of TCs is driven by environmental steering and vorticity advection. Storm-centered composites for the North Indian Ocean reveal that all models capture typical TC vorticity patterns. ERA5 and GFS show strong advection, while IFS and UM exhibit weaker signals.

Among ML models, Aurora displays stronger advection, with GraphCast and PanguWeather similar, and FourCastNet v2 slightly weaker. This suggests that ML models intuitively capture the underlying physics, supporting the potential of hybrid ML-physics approaches for improved intensity prediction.



6. Thermal structure

All models captured the TC warm core, and the pattern and magnitude of these warm core anomalies are consistent with the satellite-derived climatology of TC warm core anomalies [2], with positive temperature anomalies in the mid-troposphere (300–700 hPa). However, ML models showed weaker anomalies (2–3°C) compared to GFS (4°C) and ERA5 (3–3.5°C), indicating generally less intense storms. Variations in the vertical extent of the warm core were also observed, with GFS and IFS extending higher into the troposphere.



7. Conclusions

1. ML-based weather prediction models achieve low track errors in forecasting tropical cyclones, matching or surpassing traditional NWP models.
2. All MLWP models tend to underpredict TC intensity.
3. Despite not being explicitly designed for physical modeling, these MLWP models effectively capture the horizontal and vertical structures of tropical cyclones, including sea level pressure patterns, vorticity fields, and warm-core anomalies.

8. References

[1] Andrew J Charlton-Perez, Helen F Dacre, Simon Driscoll, Suzanne L Gray, Ben Harvey, Natalie J Harvey, Kieran MR Hunt, Robert W Lee, Ranjini Swaminathan, Remy Vandaele, et al. 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, 2024.

[2] Xiang Wang and Haiyan Jiang. A 13-year global climatology of tropical cyclone warm-core structures from airs data. *Monthly Weather Review*, 147(3):773 – 790, 2019.

9. Acknowledgment

The authors thank the IIT Delhi HPC facility for computational resources.