Supporting Information for "Can AI weather models predict out-of-distribution gray swan tropical cyclones?"

Y. Qiang Sun*

Pedram Hassanzadeh†
Mohsen Zand,
Ashesh Chattopadhyay,
Jonathan Weare
Dorian S. Abbot

February 20, 2025

Content of this file

1. Figures S1 to S5
2. Table S1

*Corresponding author: qiangsun@uchicago.edu
†Corresponding author: pedramh@uchicago.edu

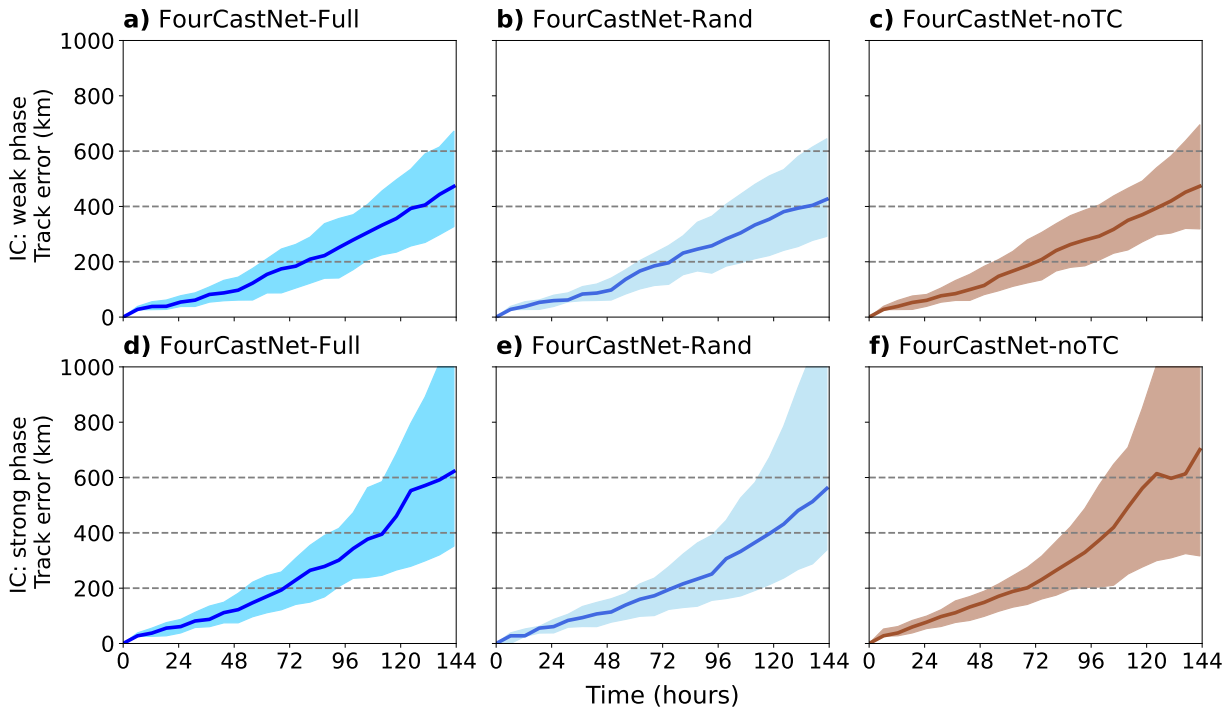

Figure S1: All versions of FourCastNet show similar performance in predicting TC tracks. The forecast errors of the tracks for Category 5 TCs in the test data are shown with shading representing the 25th and 75th percentiles across all forecasts. All forecast data are initialized as in Figure 2. Forecasts with each lead time is generated using 5 realizations and 51 initial conditions (ICs) on all 20 TCs, as described in the main text. Therefore, for each lead time, we have $5 \times 51 \times 20 = 5100$ values. TC track error is defined as the distance between the location of the TC center (minimum mslp) in each forecast and the actual location of the TC in the ERA5 data.

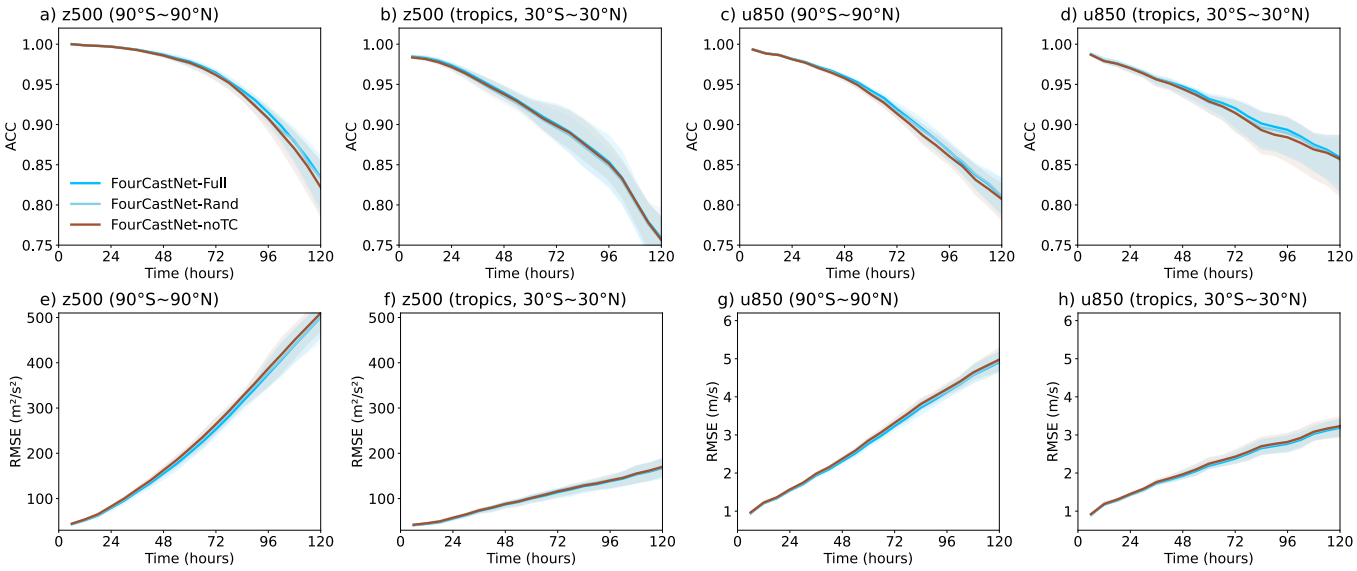

Figure S2: All versions of FourCastNet show similar forecast skill when evaluated with metrics that focus on large-scale weather, such as ACC ((a)-(d)) and RMSE ((e)-(h)). **(a)** ACC for geopotential at 500 hPa (z500) over the globe. **(b)** ACC for z500 over the tropics (30°S - 30°N) only. **(c)** ACC for zonal wind at 850 hPa (u850) over the globe. **(d)** ACC for u850 over the tropics (30°S - 30°N) only. **(e)-(h)** Similar to **(a)-(d)**, but for RMSE instead of ACC. We define the ACC and RMSE as the mean latitude-weighted value over all forecasts, following WeatherBench [1]. The shading represents the 25th and 75th percentiles across all the forecasts. Forecasts are initialized on the 1st and 15th of each month in 2018 for each trained version of FourCastNet. Each forecast is generated using 5 realizations and 51 initial conditions (ICs), as described in the main text.

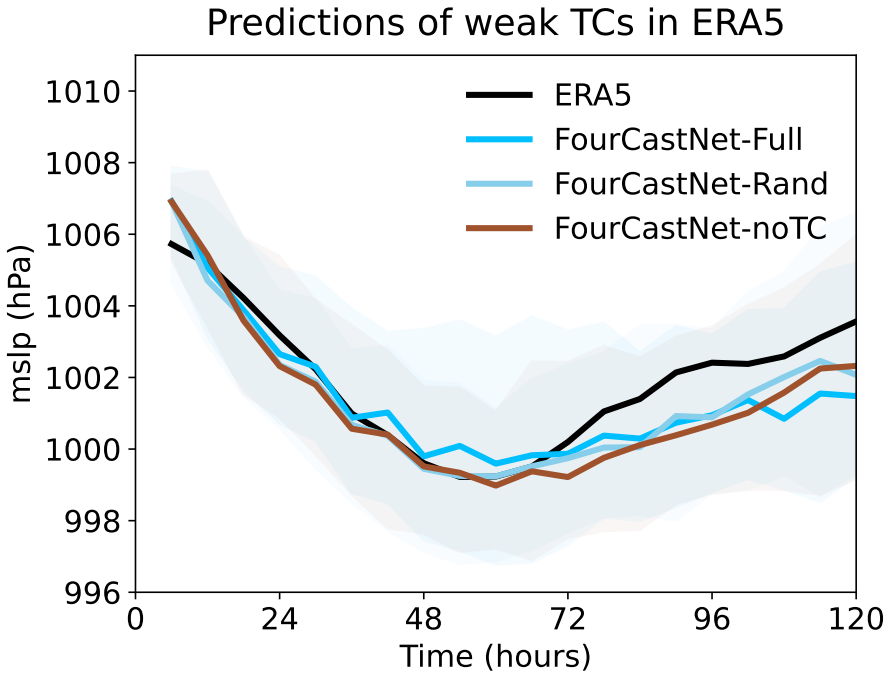

Figure S3: All versions of FourCastNet show similar performance on weaker TCs. Evolutions of mslp at the center of Category 1-2 TCs in ERA5. The lines represent the median values in ERA5 (black line) and different versions of FourCastNet's forecasts (colored lines). Two criteria are used to define weaker TCs: a) TCs are classified as Category 1 or 2 in the IBTrACS data; b) the TC center reached a minimum mslp between 988.0 hPa and 1000.0 hPa in the ERA5 data during their life cycle. All forecasts are initialized 2 days before each TC reaches Category 1 in the IBTrACS data. The shading represents the 25th and 75th percentiles of the predicted values.

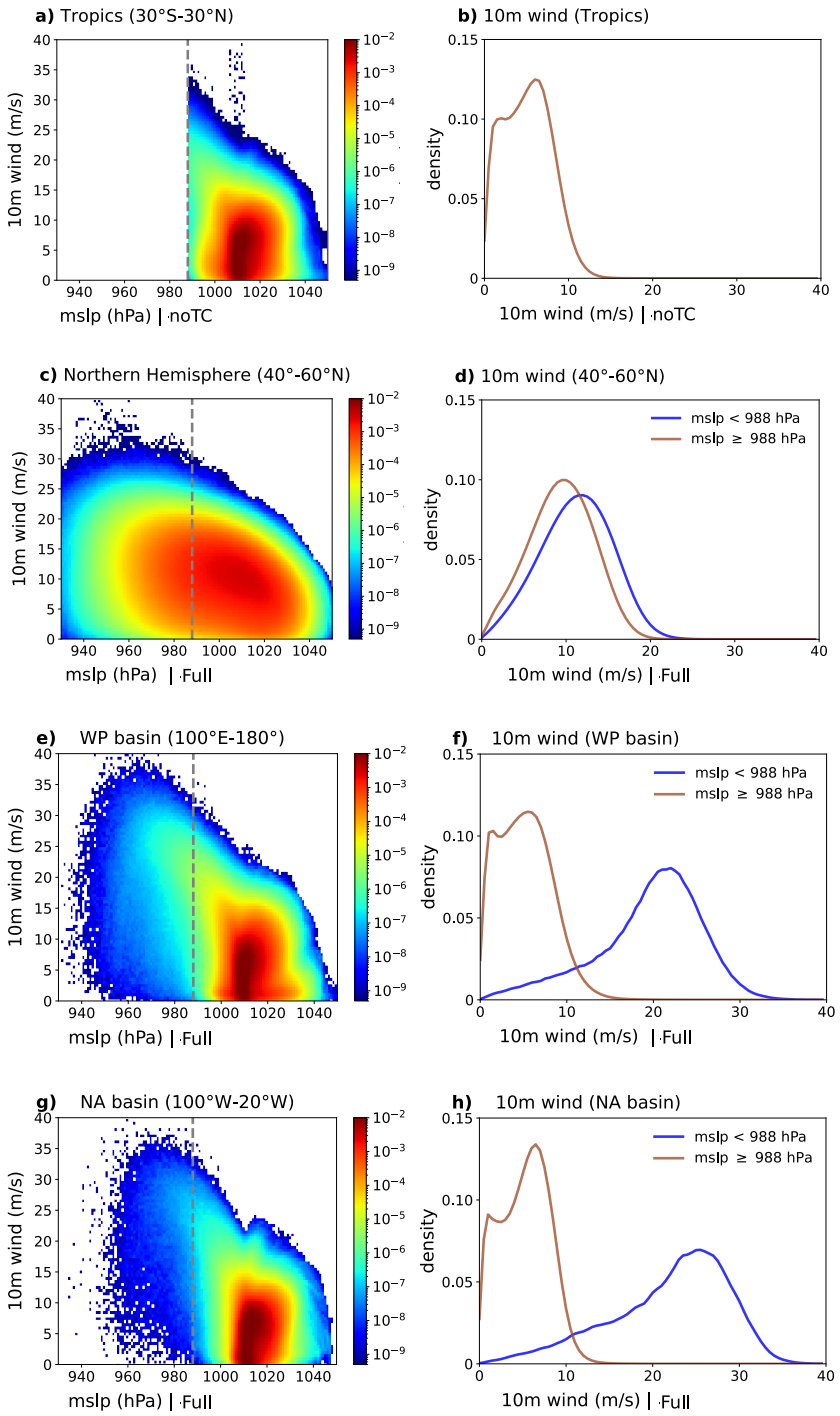

Figure S4: Dynamical similarity between the North Atlantic (NA) and Western Pacific (WP) tropical basins. a) Similar to Figure 3(a) in the main text, except that here we show the joint PDF between 10-m winds and mslp in the noTC dataset. As expected, mslp is always larger than 988.0 hPa. **b)** Similar to Figure 3(b) in the main text, but for the noTC dataset. **c)** Similar to Figure 3(c) in the main text, except that here we show the joint PDF between 10-m winds and mslp of the mid-latitudes in the Full dataset. **b)** Similar to Figure 3(d) in the main text, but for the Full dataset. **e)** Similar to Figure 3(a) in the main text, but for the WP basin (box region between $0° - 30°$ N and $100°$ E $- 180°$). **f)** Similar to Figure 3(b) in the main text, but for the WP basin. **g)** Similar to Figure 3(c) in the main text, but for the NA basin (box region between $0° - 30°$ N and $100°$ W $- 20°$ W). **h)** Similar to Figure 3(d) in the main text, but for the NA basin.

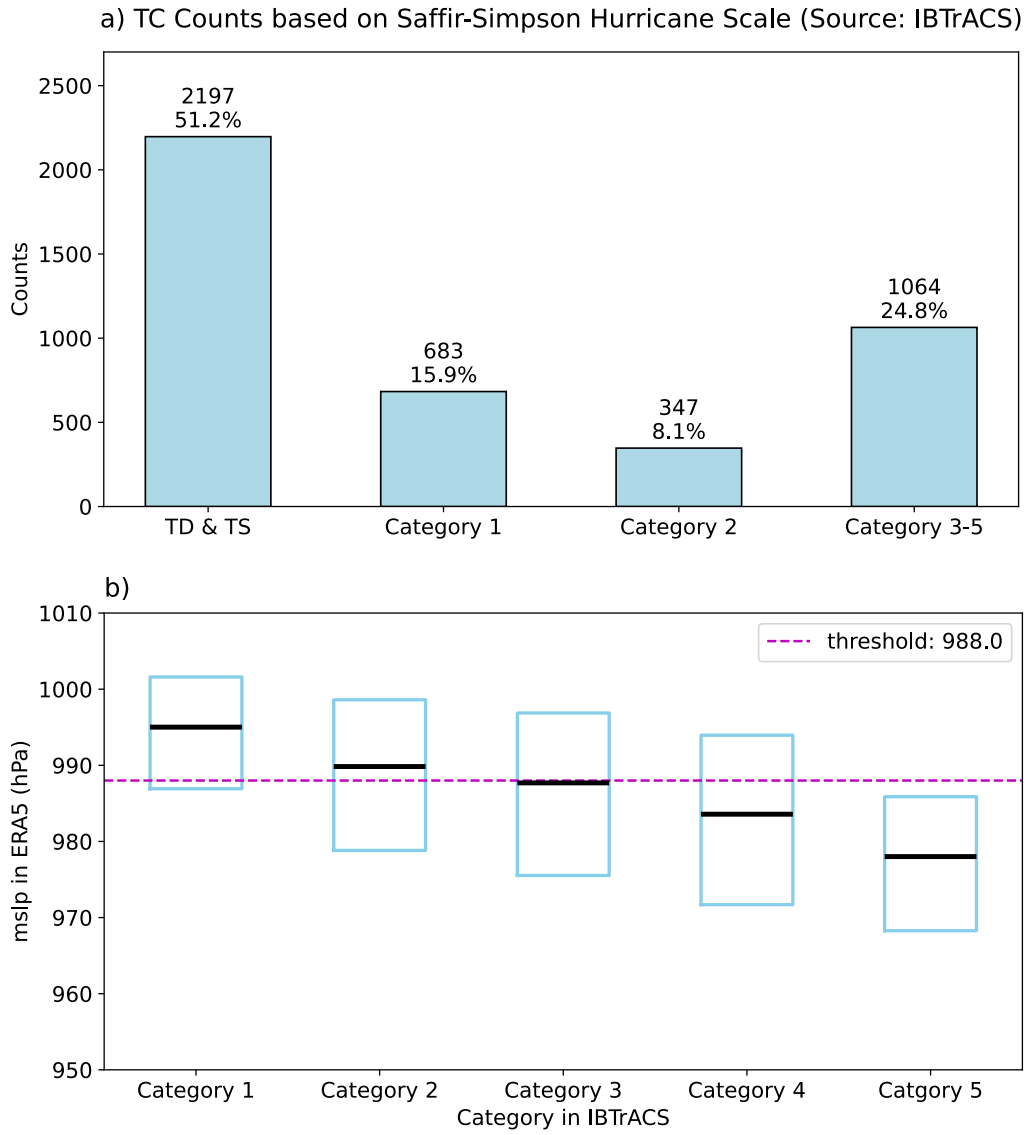

Figure S5: TC categories based on IBTrACS data and their corresponding mslp values in ERA5. a) TC counts for each category based on IBTrACS data, showing that approximately 24.8% of storms develop into major TCs (Category 3–5). **b)** The corresponding ERA5 mslp values for each TC category (based on IBTrACS data). The black line represents the median value for each category, with the top and bottom of the box indicating the 25th and 75th percentiles of the ERA5 mslps for that category. For each TC, a time series of ERA5 mslp values at the storm's center is tracked throughout its life cycle. Note that all real-world Category 1–5 TCs passed through the Category 1 stage, so their ERA5 mslp values at that stage are included in the Category 1 bar plot. The same applies to the bar plots for Categories 2–5.

TC name	Basin	yyyy/mm/dd
1. AMPHAN	Northern Indian Ocean	2020-05-15
2. DUMAZILE	Southern Indian Ocean	2018-03-03
3. HAGIBIS	WP	2019-10-05
4. HAISHEN	WP	2020-09-01
5. JEBI	WP	2018-08-29
6. KHANUN	WP	2023-07-28
7. KONGREY	WP	2018-09-29
8. LARRY	NA	2021-09-03
9. LEE	NA	2023-09-07
10. LORENZO	NA	2019-09-25
11. MANGKHUT	WP	2018-09-09
12. MAWAR	WP	2023-05-22
13. MAYSAK	WP	2020-08-28
14. MINDULLE	WP	2021-09-24
15. NANMADOL	WP	2022-09-13
16. SURIGAE	WP	2021-04-14
17. TEDDY	NA	2020-09-16
18. TRAMI	WP	2018-09-22
19. YUTU	WP	2018-10-22
20. ERA5	Southern Hemisphere	2018-03-28

Table S1: List of all 20 Category-5 TCs in the ERA5 test dataset from 2018 to 2023, and their corrsponding names from IBTrack. Of the 20, 13 occurred in the Western Pacific (WP), 4 in the North Atlantic (NA), 1 in the Northern Indian Ocean, and 2 in the Southern Indian Ocean. The last TC exists only in the ERA5 dataset, having been generated by the IFS model (model error) and lacks a corresponding observed TC name.

References

[1] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.