PNAS RESEARCH ARTICLE | EARTH, ATMOSPHERIC, AND PLANETARY SCIENCES

Can AI weather models predict out-of-distribution gray swan tropical cyclones?

Y. Qiang Sun[a,1], Pedram Hassanzadeh[a,b,1], Mohsen Zand[c], Ashesh Chattopadhyay[d], Jonathan Weare[e], and Dorian S. Abbot[a]

Affiliations are included on p. 9.

Edited by Isaac Held, Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ; received October 18, 2024; accepted April 8, 2025

Predicting gray swan weather extremes, which are possible but so rare that they are absent from the training dataset, is a major concern for AI weather models and long-term climate emulators. An important open question is whether AI models can extrapolate from weaker weather events present in the training set to stronger, unseen weather extremes. To test this, we train independent versions of the AI weather model FourCastNet on the 1979–2015 ERA5 dataset with all data, or with Category 3–5 tropical cyclones (TCs) removed, either globally or only over the North Atlantic or Western Pacific basin. We then test these versions of FourCastNet on 2018–2023 Category 5 TCs (gray swans). All versions yield similar accuracy for global weather, but the one trained without Category 3–5 TCs cannot accurately forecast Category 5 TCs, indicating that these models cannot extrapolate from weaker storms. The versions trained without Category 3–5 TCs in one basin show some skill forecasting Category 5 TCs in that basin, suggesting that FourCastNet can generalize across tropical basins. This is encouraging and surprising because regional information is implicitly encoded in inputs. Given that current state-of-the-art AI weather and climate models have similar learning strategies, we expect our findings to apply to other models. Other types of weather extremes need to be similarly investigated. Our work demonstrates that novel learning strategies are needed for AI models to reliably provide early warning or estimated statistics for the rarest, most impactful TCs, and, possibly, other weather extremes.

AI weather models | out-of-distribution generalization | gray swan weather extremes

Significance

AI models produce skillful weather forecasts, including for some extreme events. However, forecasting the strongest events that are so rare they did not exist in the training set (the so-called gray swans) remains a major concern for these models' operational use, especially as climate change introduces unprecedented conditions. Here, we train an AI weather model after removing Category 3–5 tropical cyclones from its training set and test it on Category 5 storms. The model could not accurately forecast these unseen cyclones. However, the model shows promise in learning from strong storms in one region and forecasting them in another region. Our work highlights the need for better understanding the limitations of AI weather models and innovations to improve them.

Recent years have seen a rapid emergence of skillful AI weather forecast models such as FourCastNet (1), Pangu (2), GraphCast (3), AIFS (4), Fuxi (5), and Stormer (6). These data-driven models are deep neural networks (NNs) that predict the evolution of the 3D global atmospheric state in six or twelve hour increments after being trained on the ERA5 reanalysis dataset from 1979 to 2015. AI weather models' out-of-sample (2018–2023) forecasts of the global weather, including some aspects of extreme events such as the track of tropical cyclones (TCs), have been shown to outperform predictions from the best numerical models for up to 10 d (2, 3, 7, 8). Aside from increased accuracy, a major advantage of AI weather models is that, once trained, they can be run 10^4 to 10^5 times faster than state-of-the-art numerical models (2, 3, 9), and for example, produce large-ensemble and probabilistic forecasts (10–13). Furthermore, this substantial speed-up has opened a new, rapidly advancing avenue for developing AI models for long-term emulation of the atmosphere, ocean, or the entire climate system (we refer to such models as AI climate emulators, hereafter) (14–22). The speed of such emulators, which can be trained on reanalysis datasets and/or global climate model (GCM) outputs, enables the generation of large ensembles of long runs. It has been suggested that such ensembles could significantly reduce sampling error (i.e., the internal variability uncertainty), a major challenge for climate change projections of extreme weather events, particularly at regional scales (18, 19, 21, 23).

A key question relevant to the fidelity and usefulness of AI weather models and climate emulators is their ability to forecast/emulate the rarest, yet most impactful, extreme events (24–26). The rarity of the most extreme weather events makes them harder to learn for AI models, which is the classic "data imbalance" problem in statistical learning (27–31). AI weather models have often been found to underestimate the peak amplitude of events such as heat waves and TCs (32–34), which may result from data imbalance as well as other problems such as blurring due to spectral bias (23, 35, 36). AI weather models have shown remarkable skill predicting TC tracks (much better than TC intensity) (1–3, 34); however, TC tracks are largely determined by large-scale background winds that are not rare.

Author contributions: P.H., J.W., and D.S.A. designed research; Y.Q.S. and M.Z. performed research; M.Z. and A.C. contributed new reagents/analytic tools; Y.Q.S. analyzed data; and Y.Q.S., P.H., J.W., and D.S.A. wrote the paper.

Competing interest statement: P.H. is a member of the NVIDIA's Advisory Council on Physics ML for Climate Science.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: qiangsun@uchicago.edu or pedramh@uchicago.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2420914122/-/DCSupplemental.

Published May 20, 2025.

Data imbalance reaches its asymptotic limit for so-called "gray swan" weather extremes, first defined by Lin and Emmanuel (37) as rare weather events that are physically possible but have never actually been observed in the historical record. In the context of AI models in climate science, gray swans are physically possible weather events that are rarer and stronger than those in the training set, thus they are "out of distribution." While in general, neural network-based AI algorithms should not be expected to forecast/emulate out-of-distribution events, one might hypothesize that for physical systems such as climate, AI models could learn key physical relationships among variables from weaker (in-distribution) events that would generalize to stronger (out-of-distribution) events. Although this hypothesis has not been rigorously tested for extreme weather events with state-of-the-art AI models, there are recent studies that suggest these AI weather models learn dynamics to some degree, rather than simply memorizing patterns. These include the interesting results of Hakim and Masanam (38), who showed the ability of AI weather models to predict the response of the atmosphere to highly localized (but not extreme) perturbations that are substantially different from any pattern in the training set, and the work of Rackow et al. (39), who demonstrated these models' relatively robust short-term forecasting skill under climate change.

The main objective of this paper is to examine, in highly controlled experiments, whether the AI weather model FourCastNet can forecast gray swan TCs. Our framework is summarized in Fig. 1. Briefly, we have created 4 additional training sets from the original ERA5 1979–2015 training set (Full). In one set (noTC), we have removed any training sample that contained mean sea-level pressure (mslp), a measure of TC strength (the lower mslp, the stronger TC), below the 25th percentile (988 hPa) in the tropics (30°S–30°N); see *Methods and Data*. This is roughly equivalent to removing samples containing TCs of Category 3–5 anywhere in the world. Our next training set (Rand) has the same size and seasonal distribution as noTC, but we have randomly removed 25% of the samples from the full set, while avoiding removing any Category 3-5 TCs (*Methods and Data*). Finally, we have created two additional training sets by removing samples containing tropical mslp below 988 hPa over either the Western Pacific (noWP) or North Atlantic (noNA) basins. We then train five versions of FourCastNet on each training set from five different random realizations of the weights and biases, leading to 25 independently trained models. Next, we quantify

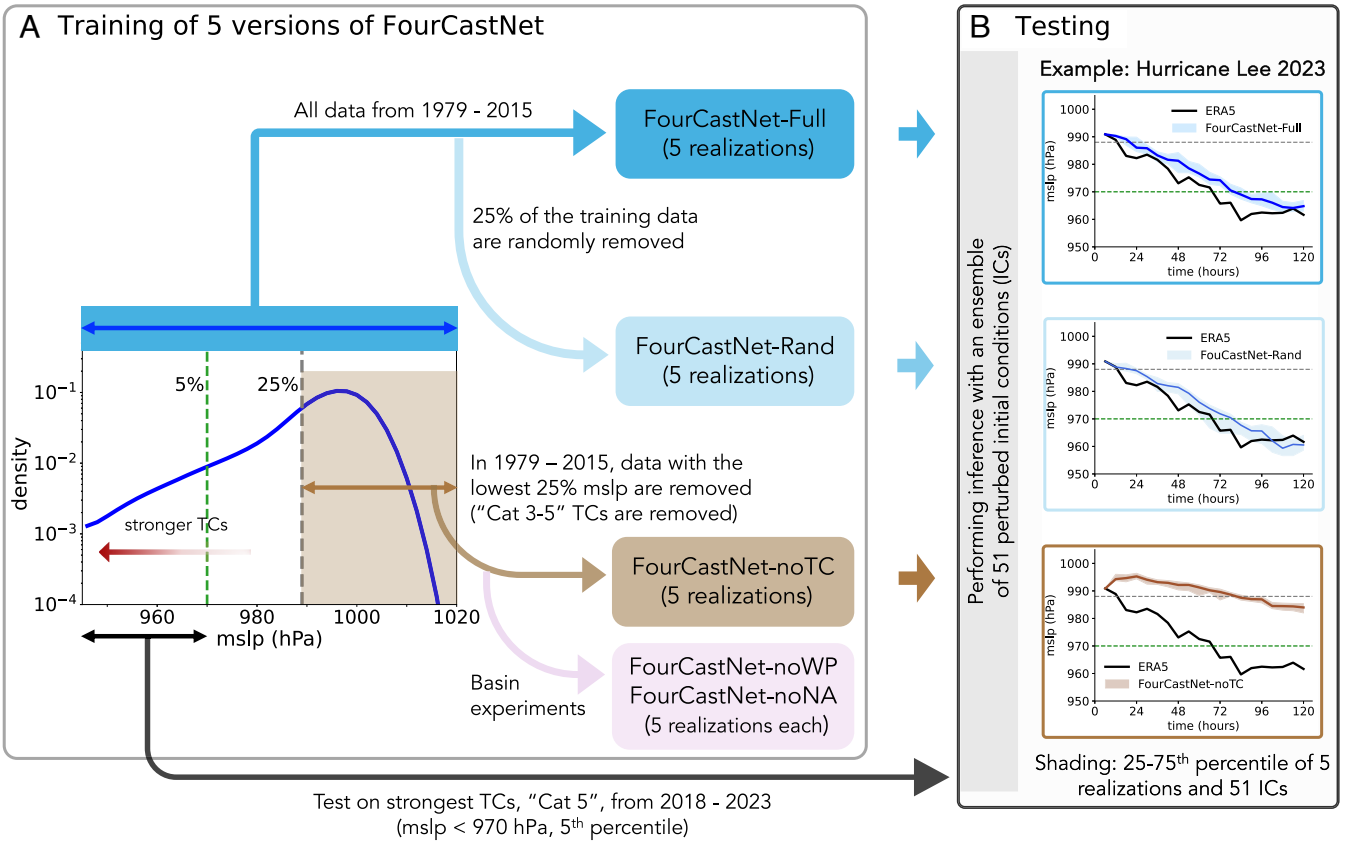

Fig. 1. Schematic overview of this study. (*A*) Training of five versions of FourCastNet. The panel depicts the histogram of minimum mslp in the tropics (30°S–30°N) in the training set (ERA5, 1979–2015). Note that a lower mslp corresponds to a stronger TC. Vertical lines indicate the 5th and 25th percentiles, which are 970 hPa and 988 hPa, respectively. For FourCastNet-Full, the full training dataset is utilized. For FourCastNet-noTC, samples with instances of mslp below 988.0 hPa anywhere in the tropics are removed from the training set. FourCastNet-Rand uses a training set of the same size and seasonal distribution as noTC but with samples removed randomly (while ensuring that samples with mslp < 988.0 hPa are retained). Two additional models are also trained for which samples below 988 hPa only over the tropical Western Pacific (noWP) or tropical North Atlantic (noNA) basin are removed. For each training set, five independent versions (realizations) are trained from different random weight/bias initializations to account for model uncertainty. (*B*) Testing of the five models. The forecast skill of each trained model is evaluated for TCs with mslp below 970 hPa (Category 5) in the test set. The *Right* panels provide an example of the forecast results for Hurricane Lee (2023), a Category 5 TC. Shading represents the 25th to 75th percentile range of forecasts, derived from five model realizations and 51 different initial conditions (ICs) provided by an ensemble of data assimilations (EDA) from ECMWF; See *Methods and Data*.

the forecast skill of each model on the mslp evolution of the strongest (lowest 5th percentile) TCs, roughly corresponding to Category 5, in the test period (2018–2023).

Before presenting results in the next section, we emphasize that while here we focus on one AI model (due to the computational cost of training) and one type of extreme events (TCs), we speculate that the findings and insights of this work are likely to apply to other current state-of-the-art AI weather/climate models, which share the same principal learning process. This point and the potential implications for other kinds of extreme events are discussed further in *Summary and Discussion*.

Results

Failure to Extrapolate to Out-of-Distribution Gray Swan TCs. Fig. 1*B* compares the forecasting skill of FourCastNet-Full, -Rand, and -noTC on Category 5 Hurricane Lee (2023). With its winds increasing by 85 mph (140 km/h) in 24 h, Hurricane Lee underwent rapid intensification before reaching its peak intensity. TCs of similar intensity to Hurricane Lee are present in the training datasets of FourCastNet-Full and FourCastNet-Rand. Both models are able to forecast Lee's rapid intensification fairly well, although their forecasted mslp evolution has an approximately 1-d lag relative to ERA5. Most importantly, almost all 255 ensemble forecasts with these two models reach a minimum mslp below 970 hPa, the threshold for Category 5 TCs in ERA5 data (*Methods and Data*). In contrast, there are no comparably strong TCs in the noTC training dataset (mslp $\geq$ 988 hPa), so Hurricane Lee represents an out-of-distribution gray swan event for FourCastNet-noTC. The performance of FourCastNet-noTC on Hurricane Lee is much worse than that of FourCastNet-Full and FourCastNet-Rand, suggesting difficulty predicting out-of-distribution events. In particular, all members of the FourCastNet-noTC forecast a weakening of Lee (i.e., mslp increases) from the beginning (1b), followed by a very slow reintensification, such that their mslp stays above 980 hPa during the 5-d forecast period. We note here that the mslp threshold for noTC was 988 hPa, which means that FourCastNet-noTC's predictions barely go below the lowest mslp seen in the tropics. Thus, FourCastNet-noTC is not only inaccurate, but its forecasts yield "false negatives," the worst type of error for decision-critical tasks: The model forecasts a moderate Category 3 TC rather than a devastating Category 5 TC, giving no signal that its forecast is inaccurate. As we show below, this is a consistent behavior of FourCastNet-noTC.

Fig. 2 presents similar comparisons but aggregated over all 20 Category 5 TCs in the testing dataset and includes forecasts initialized in both the weak (top row) and strong phases of the TCs—during the transition from Category 4 to 5 (bottom row). The conclusions are the same as those drawn for Hurricane Lee. Both FourCastNet-Full and FourCastNet-Rand perform fairly well in forecasting the evolution of minimum mslp, whether the

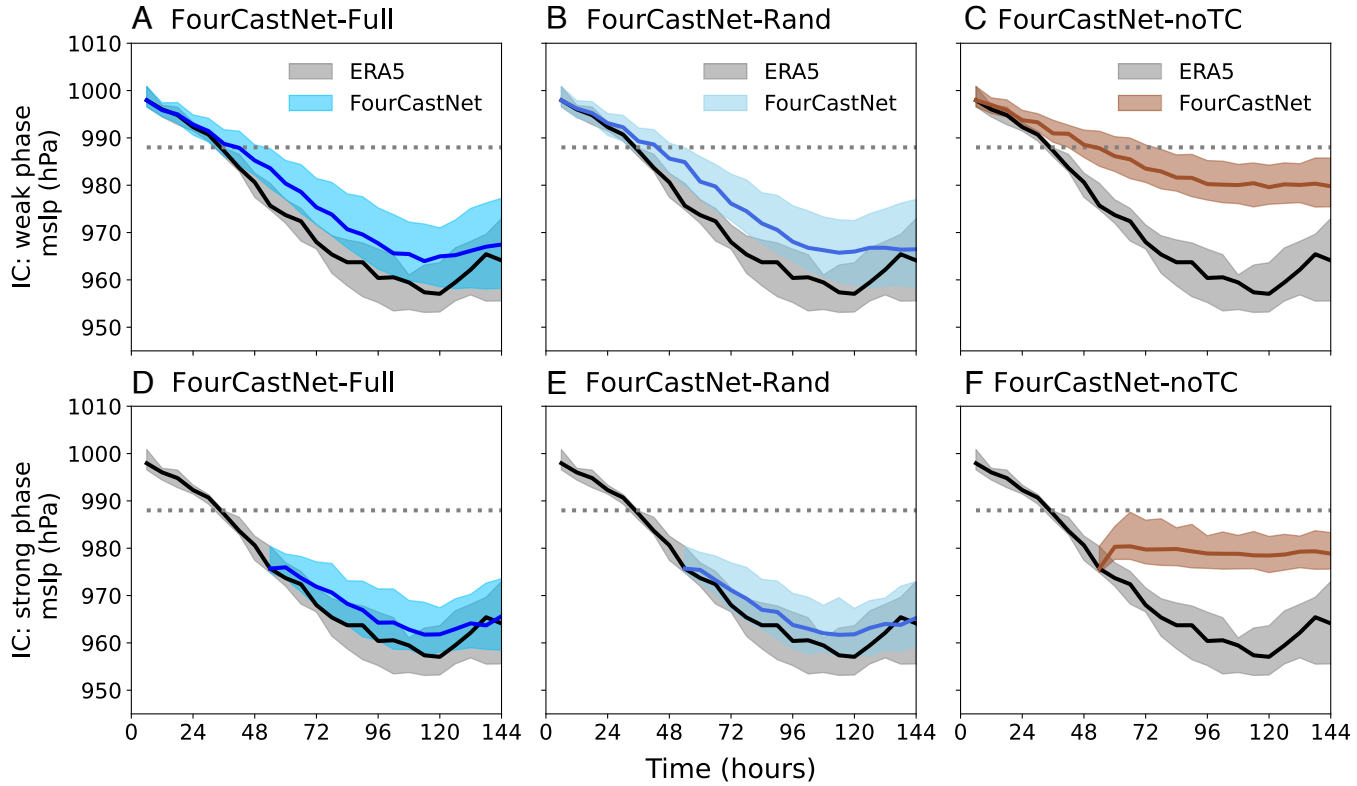

Fig. 2. FourCastNet's difficulty in extrapolating to gray swan TCs. Forecasting of all 20 Category 5 TCs from the test set (2018–2023) by three versions of FourCastNet trained on different datasets: FourCastNet-Full (left column: *A* and *D*), FourCastNet-Rand (middle column: *B* and *E*), and FourCastNet-noTC (right column: *C* and *F*). The dashed line shows the critical threshold for 25th percentile of minimum mslp (roughly Category 3 TC) used in the noTC training set. All panels show the evolution of the median mslp (solid line) and the interquartile range from the 25 to the 75th percentile (shading) over all 20 Category 5 TCs, 5 realizations of each trained model, and 51 perturbed ICs from EDA (5,100 forecasts). Shading for ERA5 is over the 20 TCs. Forecasts are initialized 1 d before each TC reached the critical threshold (weak phase, top row) or 1 d after the TC reached this threshold (strong phase, bottom row). The latter ICs are out-of-distribution. As an additional note, detailed analysis shows that none of the ensemble members in the FoureCastNet-noTC forecasts reached the observed lowest mslp values. Although a few members' mslp reached 970 hPa, this occurred because these members transitioned to an unstable state that eventually led to blow-up, rather than capturing realistic intensification of the storm.

ICs are during the weak or strong phase of the TCs. They do show a consistent bias in underpredicting the strength of TCs, similar to results reported for other state-of-the-art AI weather models for TCs (34) and other extreme events (33). This bias could result from data imbalance or other problems such as blurring due to spectral bias (23, 35).

In contrast, FourCastNet-noTC significantly underpredicts the evolution of mslp (panels *C* and *F*), despite demonstrating comparable performance to FourCastNet-Full in predicting the tracks of TCs (*SI Appendix*, Fig. S1). Although FourCastNet-noTC has some skill on the first day of intensification when initialized in the weak phase, once the ERA5 mslp approaches the critical value of 988 hPa after around 1 d, the intensification in FourCastNet-noTC's forecasts slows dramatically and mslp barely reaches below 980 hPa. When FourCastNet-noTC is initialized in the strong phase, where the ICs are out-of-distribution, FourCastNet-noTC forecasts a weakening of the TC. It appears that the model tries to relax toward the values of intensity it had seen in training (above 988 hPa), instead of intensifying to the lower values of mslp observed in ERA5. As noted above, this behavior would lead to false negative forecasts for destructive storms, which has major societal consequences. Additionally, it is important to note that the similar performance of FourCastNet-Full and -Rand on Category 5 TCs shows that a 25% shorter training set does not impact the forecasting skill, so this cannot be the reason for FourCastNet-noTC's poor performance.

These comparisons consistently show that FourCastNet is unable to learn from weaker TCs (Category 1–2) and extrapolate to stronger, unseen Category 5 TCs. Despite an inability to extrapolate to gray swan TCs, FourCastNet-noTC performs similarly to FourCastNet-Full and FourCastNet-Rand on typical weather that all models see in their training set. Specifically, the three models exhibit similar forecast skill for global and tropical weather [based on anomaly correlation coefficient (ACC) and root mean squared error (RMSE)], as shown in *SI Appendix*, Fig. S2, and for weaker (Category 1–2) TCs in the testing set, as presented in *SI Appendix*, Fig. S3.

Some Success Generalizing Across Tropical Basins. Although the training dataset of FourCastNet-noTC has no TC (or any sample) with mslp below 988 hPa in the tropics, it has plenty of samples with mslp below this value in the midlatitudes, associated with extratropical cyclones. This can be seen in Fig. 3 *A* and *C* and *SI Appendix*, Fig. S4*A*, which show that accounting for the values of mslp globally in the noTC training set, mslp <988 hPa is not unprecedented or out-of-distribution at all.

One might wonder why the presence of strong extratropical cyclones in the training set does not enable FourCastNet-noTC to predict Category-5 TCs with similarly low mslp. We believe that this is due to the difference in the dynamics of TCs and extratropical cyclones. TCs' dynamics are driven by convection and latent heating with azimuthal winds reaching the maximum near the low-pressure center, while extratropical cyclones' dynamics are driven by baroclinic instability and their wind profiles are different. For the AI model, these differences are manifested in the evolution of the entire state vector (input) rather than a single variable (like mslp), and can be seen, for example, in the joint PDFs in Fig. 3. While there is a marked trend of increases in wind speed as mslp decreases below 1,000 hPa in the tropics, in the extratropics, the relationship between mslp and wind speed is much less pronounced, such that it is not uncommon to have low wind speeds when the mslp is low. The distinction between the two phenomena can be also seen in the probability density of 10-m wind for weak (mslp > 988 hPa) and strong (mslp < 988 hPa) cyclones. The different physical mechanisms driving extratropical cyclones and TCs therefore lead to different relationships between physical variables in space and time. As a result, the presence of strong, low-mslp extratropical cyclones (in the training set) does not help FourCastNet-noTC with predicting Category 5 TCs in the tropics during testing.

Given the above discussion, one might next wonder whether FourCastNet can generalize from strong TCs it has seen in the training set in one tropical basin to another. While there are differences between the TCs in the two major basins of activity, the North Atlantic and Western Pacific, mainly due to differences in large-sale circulation, TCs in both regions are fundamentally driven by similar dynamical processes. There is substantial similarity between the mslp and 10-m wind distributions of TCs in these two basins (*SI Appendix*, Fig. S5 *E*–*H*). Both FourCastNet-noWP and FourCastNet-noNA perform significantly better than FourCastNet-noTC when tested on Category 5 TCs in the Western Pacific and North Atlantic basins, respectively (Fig. 4). Specifically, FourCastNet-noWP and FourCastNet-noNA forecast intensification of all tested TCs from both weak and strong ICs and produce minimum mslp well below 970 hPa for many storms (Fig. 4). In fact, FourCastNet-noWP's performance is similar to that of FourCastNet-Full for 3 out of the 13 tested TCs in the WP basin (not shown). The reduced performance of FourCastNet-noWP and FourCastNet-noNA relative to FourCastNet-Full may be due to the smaller number of intense TCs in their training datasets.

Overall, this analysis suggests that FourCastNet can effectively generalize across geographic regions by learning from dynamically similar events. We emphasize the need for dynamical similarity given that FourCastNet-noTC could not generalize from strong extratropical cyclones to unseen strong TCs.

Forecasts Lack Physical Consistency. Adding physical constraints is often cited as an avenue for improving the skill of AI weather/climate models for extreme weather events and out-of-distribution generalization (18, 21, 33, 40). Here, we examine a key physical balance of TCs in all trained versions of FourCastNet to see whether the poor out-of-distribution generalization to gray swans in FourCastNet-noTC could be due to lack of this balance. Above the boundary layer, TCs approximately satisfy the gradient-wind balance between the pressure gradient force and the Coriolis and centrifugal forces in the azimuthal mean (41):

$$g\frac{\partial Z}{\partial r} = \frac{V_g^2}{r} + fV_g. \qquad [1]$$

Here, V_g is the azimuthal gradient wind, Z is the geopotential height (the height of a given pressure surface), g is the gravitational acceleration, r is the radial distance from the center of the cyclone, and f is the Coriolis parameter.

As expected, the gradient-wind balance holds for Category 5 TCs in the ERA5 data across all radial length scales (Fig. 5). This reflects the fact that ERA5 is based on a physical model with relatively small adjustments during data assimilation to better fit with observations, and the dominant balance in the equations of motion for TCs is gradient-wind balance (Eq. **1**). In contrast, the gradient-wind balance is not observed within ∼200 km from the center of TCs in forecasts from both FourCastNet-Full and FourCastNet-noTC. The most significant deviations from balance, i.e., the difference between the full wind and the gradient wind, occur when forecasts are initialized from the strong phase (Fig. 5 *E* and *F*). While the magnitude of the gradient

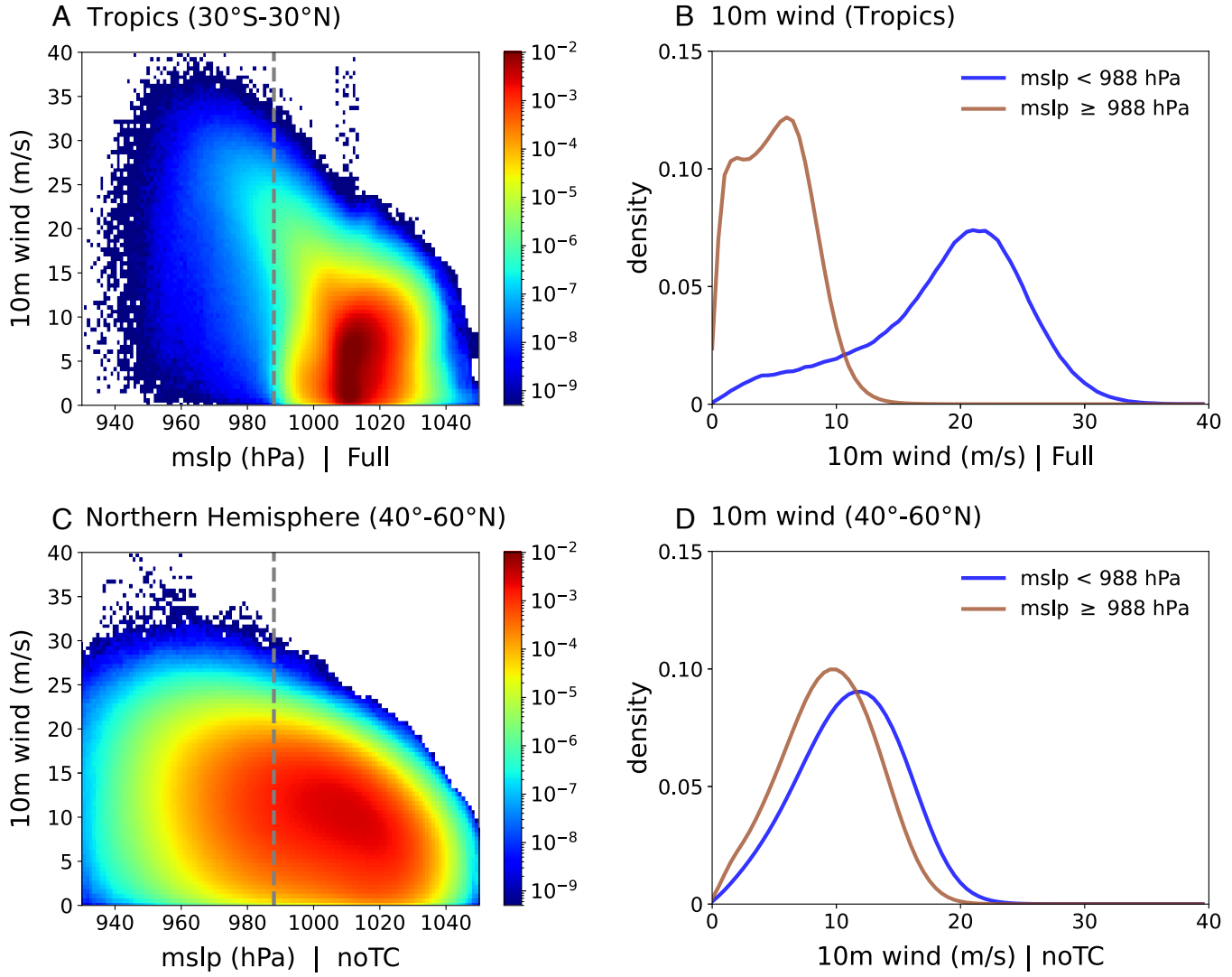

Fig. 3. Extra-TCs and TCs exhibit different dynamical behavior. (*A*) Joint PDF of mslp and 10-m winds in the tropics (30°S–30°N) in the Full training set. (*B*) Probability density of 10-m winds in the tropics in the Full training set, conditioned on the mslp threshold. (*C*) Similar to (*A*) but for the midlatitudes (40°–60°N) of the noTC training set. (*D*) Similar to (*B*) but for the midlatitudes of the noTC training set.

wind is comparable to the full wind, the radius of maximum gradient wind is too large relative to that of the full wind. Although FourCastNet-Full appears to simulate gradient wind balance slightly better than FourCastNet-noTC when forecasts are initialized from the weak phase (Fig. 5 *B* and *C*), the difference is within the uncertainty range. Interestingly, despite the fact that FourCastNet-Full is much more successful at forecasting the evolution of Category 5 TCs than FourCastNet-noTC, the wind and pressure fields are no more physically consistent in FourCastNet-Full than in FourCastNet-noTC.

Lack of physical consistency is not unique to FourCastNet. A few recent studies have also pointed out lack of physical balances in other state-of-the-art AI weather models (33, 35). The implications of these findings for potential avenues for improving the out-of-distribution generalization of AI weather/climate models will be discussed in the next section.

Summary and Discussion

We conduct controlled experiments in which we train the AI weather model FourCastNet after removing Category 3–5 TCs from the training sets, either globally (FourCastNet-noTC) or only over the Northern Hemisphere Atlantic or Pacific basins (FourCastNet-noNA and -noWP). By analyzing the forecasting skill of these different models on Category 5 TCs in the test set we demonstrate the following:

1. **Lack of out-of-distribution generalization (extrapolation) for gray swan TCs:** FourCastNet is unable to learn about unseen strong TCs (Category 5) from the weaker ones (Category 1–2) that were present in the training set (Figs. 1 and 2). This is despite the fact that low mslp extratropical cyclones (at the level of Category 5 TCs) exist in the noTC training set. However, because extratropical cyclones and TCs have different dynamics (as, for example, manifested in the joint PDFs of mslp-wind, Fig. 3 and *SI Appendix*, Fig. S4), the low mslp values in extratropical cyclones could not help FourCastNet-noTC with predicting Category 5 TCs.
2. **Some generalization for dynamically similar storms across ocean basins:** FourCastNet is able to learn something about unseen strong TCs in one ocean basin from strong TCs it has seen in another basin in the training set (Fig. 4). This is somewhat surprising, but encouraging, as location-specific

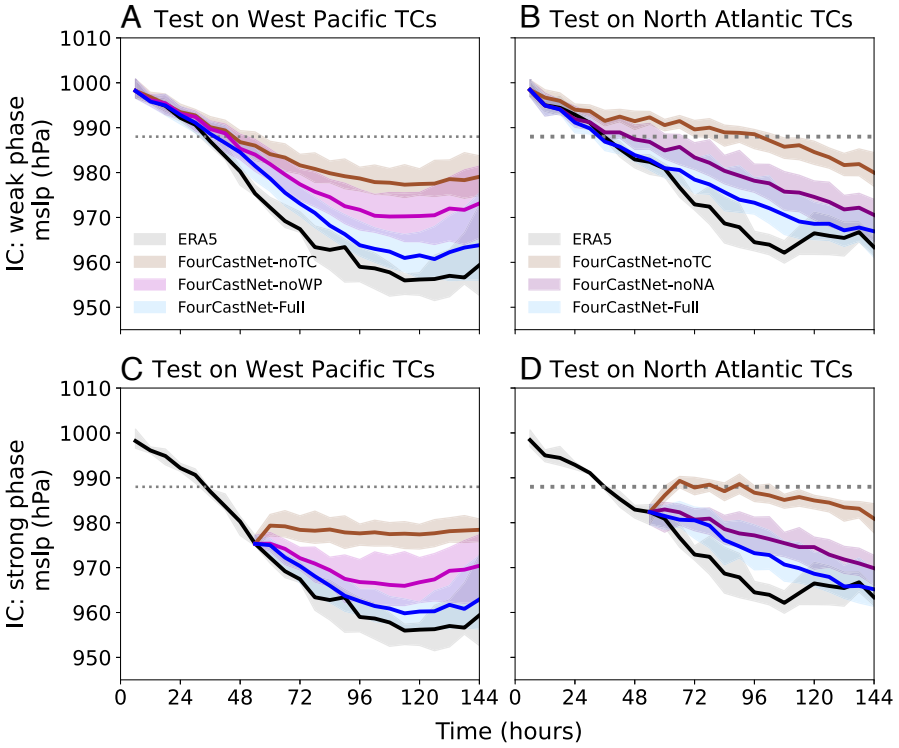

Fig. 4. FourCastNet generalizes across tropical regions for dynamically similar events. (*A*) Comparison of the forecast skill of FourCastNet-noWP against other models for Category 5 TCs (from the test set) in the Western Pacific, initialized at the TC's weak phase. (*B*) As in (*A*) but for TCs in the North Atlantic basin. (*C* and *D*) As in (*A* and *B*) but initialized at the strong phase of the TCs. Solid lines and shading are as in Fig. 2.

information (e.g., latitude, longitude, topography) is implicitly encoded in the inputs.

3. **Lack of gradient-wind balance:** Whether trained with the Full or noTC dataset, FourCastNet is unable to reproduce gradient-wind balance, a key physical constraint that is obeyed by TCs in the training set (Fig. 5). Enforcing gradient-wind balance in the loss function could potentially help the model learn this physical relationship, though there could be tradeoffs in the accuracy or the representation of other physical properties (e.g., the ageostrophic flow).
4. **Common metrics can obscure poor performance for gray swan TCs:** FourCastNet-Full, FourCastNet-Rand, and FourCastNet-noTC show similar forecast skill as measured by ACC or RMSE calculated both globally and over the tropics, as well as by forecasting skill on in-distribution TCs (Category 1–2 here); see *SI Appendix*, Figs. S1–S3.

As discussed below, (1) and (2), which have been unambiguously tested for an AI weather model here, along with (3) and (4), have major implications for efforts to understand and predict TCs using AI weather models. We speculate that our findings might have implications for the gray swans of other types of extreme events and for AI climate emulators.

Before discussing these implications, potential solutions, and proper tests for learning gray swans, we highlight that here we have focused on only one model, FourCastNet. We did this because it was the first state-of-the-art AI weather model publicly available for training and is much cheaper to train than new models like PanguWeather or GraphCast (2, 3, 42). Still, training 25 independent versions was computationally demanding (note that in this work, each version had to be trained from random initial weights; we cannot fine-tune a pretrained FourCastNet for our experiments). However, given that current state-of-the-art AI weather models and climate emulators share the same principal learning process (i.e., physics-free deterministic or probabilistic evolution of the mapping of $\mathbf{x}(t)$ to $\mathbf{x}(t + \Delta t)$), we expect limitation (1) to apply to them as well. In fact, while newer models often show improved accuracy on global and some regional metrics, recent studies have found similar types of shortcomings, e.g., physical inconsistency like (3) or missing the peak amplitude of extreme events, in different state-of-the-art models (23, 33, 35) (note that ref. 34 found FourCastNet to have forecast skill for TCs comparable to newer models). Even most of the emerging "foundation weather/climate models" (43–45) use the same overall principal learning process. There are recent examples of AI models that use self-supervised learning algorithms, such as pretraining with masked-autoencoders (46–48), or hybrid models such as NeuralGCM (22), although, at least so far, these approaches do not have any component to address data imbalance. Whether they improve (1) and (3) remains to be thoroughly investigated, and should not be assumed without rigorous demonstration (see below).

We have focused here on only one type of extreme weather (TCs), again due to computational cost. Whether (1) and (2) apply to other major types of extreme weather events needs to be thoroughly studied using similar controlled experiments. While stronger TCs could not be learned from the weaker ones, it is possible that some types of extreme events could be learned from weaker examples. Furthermore, extreme weather events often have distinct dynamics (e.g., dry and moist heat waves, atmospheric rivers, and cold snaps), which can hinder generalization among them . However, some of the main physical processes, such as zonal and meridional thermal advection in heat and cold waves, share similarities. To what degree learning one

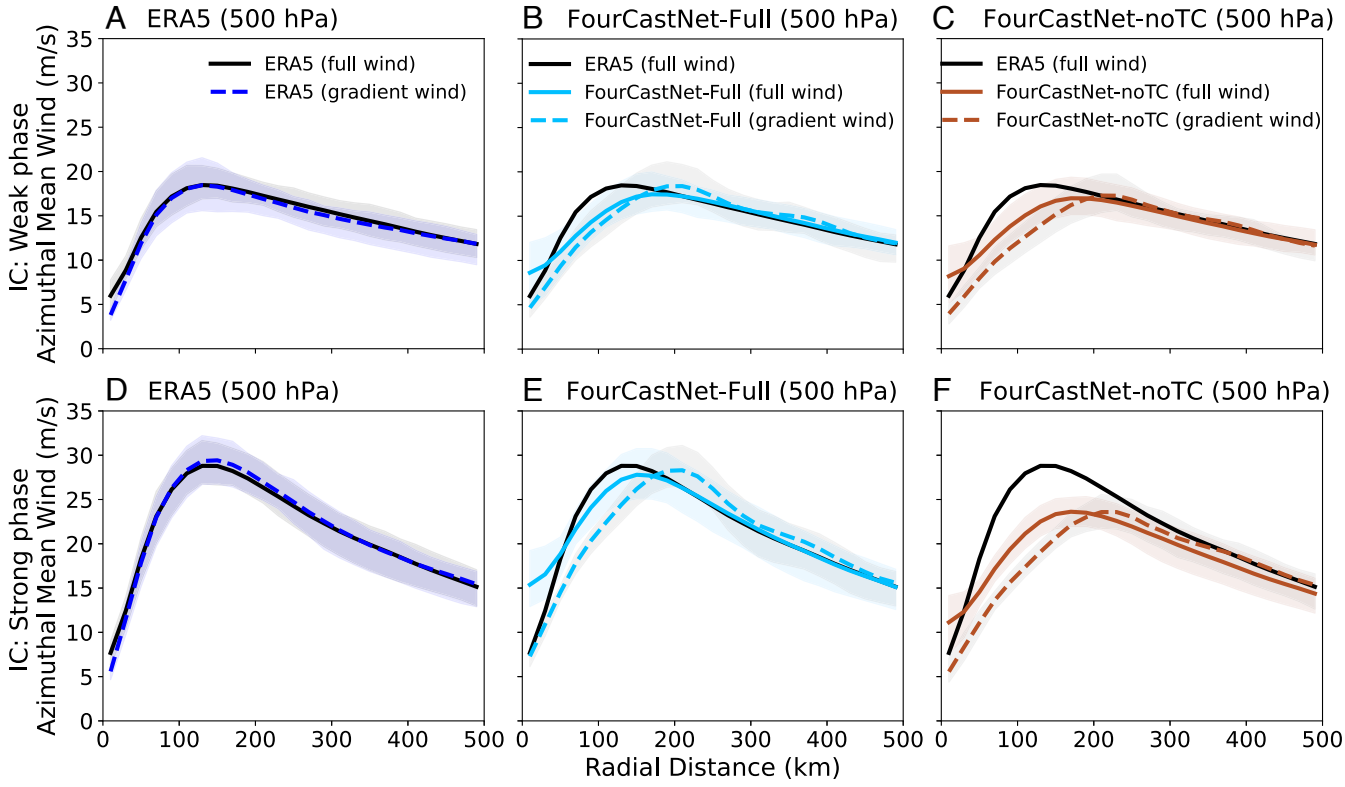

Fig. 5. Lack of physical consistency in the forecasts. (*A*) Gradient-wind balance in ERA5. Radial profiles of azimuthal wind and the gradient-wind derived from Eq. **1** at 500 hPa for all Category 5 TCs in the test set (2018–2023) in their weak phase. (*B*) As in (*A*) but for FourCastNet-Full's forecasts. (*C*) As in (*B*) but for FourCastNet-noTC's forecasts. The bottom row (*D–F*) is the same as the top row (*A–C*) but for the strong phase of Category 5 TCs. The shading indicates the 25th to 75th percentile range across the 20 Category 5 TCs in the test (*Left* panel). In the *Middle* and *Right* panels, the shading is over the 20 TCs, 5 realizations, and 51 perturbed ICs.

type of event can translate into another type remains to be seen. Our current results clearly demonstrate that the presence of other related extreme events (i.e., extratropical cyclones) does not help with gray swan TCs.

The main potential implications of (1) to (4) for current state-of-the-art AI weather models and climate emulators, are discussed below. We emphasize that these items are of significant importance but are currently speculative and require extensive and careful investigation.

a. **AI weather models might produce unreliable early warning for some types of unprecedented weather events:** AI weather models might fail to accurately forecast weather extremes that are unprecedented globally. However, it is encouraging that if dynamically similar events of comparable or larger amplitudes exist in other parts of the world in the training set, the AI models may show some forecast skill. *The possibility of AI weather models misforecasting extreme weather events, particularly if they produce the false negatives we consistently find here, creates serious societal risk*. This concern is particularly acute as climate change is increasing the likelihood of some types of gray swans (including TCs), just as rapid advances in AI weather forecasting invite more operational reliance on such models over traditional physics-based ones. This suggests that the limitations of the AI weather models for forecasting gray swans must be fully characterized [see (c)]. Additionally, it would be valuable to identify early warning signals that an AI weather model is failing. As highlighted above, the strengths and weaknesses of AI models for gray swans might vary among different types of extreme weather events, requiring comprehensive studies.

b. **AI climate emulators might mischaracterize extreme weather statistics for some types of events:** Our results suggest that current AI climate emulators may not be able to reliably reduce sampling error in gray-swan-event statistics, such as return periods. The possible inability to reduce this type of error, which arises due to internal variability, undermines a major motivation behind developing climate emulators. This is because to do so the emulators would have to learn about gray swan events from the weaker events in the training set, which we find they cannot do for TCs. As a result, the long, large ensembles of synthetic data these emulators generate may not contain reliable information about extreme events stronger and rarer than those that exist in the (typically short) training set, and the accuracy of the estimated statistics will be still limited by the length of the training set. Again, this potential limitation might vary among different types of extreme events.

c. **Proper definitions/metrics for gray swans and unprecedented rare events for AI models is necessary:** Given the partial ability of FourCastNet to generalize learning dynamically similar storms across regions as well as differences in common normalization schemes in climate analysis and AI algorithms, the question of which events qualify as gray swans needs to be carefully examined. First, an event considered unprecedented in one region may already have occurred elsewhere. Second, in climate science, unprecedented events are often defined based on comparing the anomaly of one variable (e.g., near-surface temperature or rainfall) to local

statistics. Current AI weather models often normalize each input variable with respect to its global mean and SD. Our findings suggest that accounting for the presence of dynamically similar events in other regions, examining global rather than local fields, can be important in determining whether an event is truly out-of-distribution for an AI weather model or climate emulator.

There are some potential remedies that could help AI weather and climate models with forecasting and emulating gray swans. Common AI strategies for dealing with data imbalance are using weighted loss functions and resampling. Weighted loss functions can improve the learning of rare events that exist in the training set (see refs. 31, 49, and 50 for applications in climate science). But this approach will not help with gray swans. Up-sampling is a potential remedy, but it requires generating strong synthetic TCs and including them in the noTC training set. Generative AI models (e.g., diffusion models) might help here (21); however, such models have not yet been shown to generate physically realistic extreme events beyond what was in their own training set. In fact, iterative training of generative models on their own output has been shown to lead to model collapse (51). A more promising approach is to use physics-based or theoretical models to generate strong synthetic TCs (52). While such TCs would satisfy physical constraints, how to properly include them in snapshots of more realistic global circulation without causing unphysical predictions needs to be investigated.

Incorporating physical constraints, such as conservation laws and symmetries, is often mentioned as a method of improving the representation of extreme events in AI models (28, 53, 54). While such approaches have shown promise in toy models and idealized systems, their applicability to state-of-the-art models and complex data has not been demonstrated. Furthermore, whether such constraints would help with gray swans remains to be seen. Here, we show that gradient-wind balance is not noticeably worse in the predictions of FourCastNet-noTC than in FourCastNet-Full. This suggests that the primary reason FourCastNet-noTC does not capture Category 5 TCs is not lack of gradient-wind balance but rather the absence of training samples with comparable intensities. Even if this balance were somehow enforced, which would not be a trivial task, the representation of gray swans may not improve.

Finally, an emerging remedy would integrate innovations in mathematical methods for rare events with AI weather/climate models. Applying rare-event sampling algorithms to conventional physics-based geophysical models has already shown promise in producing rare extreme events and even gray swans as well as estimating their statistics, such as return periods (25, 55–59). The resampling of physics-based models, for example a GCM, is guided by a quick best guess of which simulations are progressing toward the rare event of interest. Cheap ensemble forecasts with an AI weather model, or an AI model designed for the specific rare event (60–62), could be used to provide the quick best guess. Stronger, rarer extreme weather events produced by the physics-based model (e.g., a numerical weather/climate model) using rare-event simulation can then be added to the training set of the AI weather/climate model. A related approach that is also promising is ensemble boosting (63).

In addition to remedies, rigorous tests are needed to quantify the fidelity of any novel AI model in predicting/emulating gray swans. The approach used in this paper is unambiguous and also provides a straightforward way to examine the physical consistency of the predicted events. However, this approach is computationally demanding as it requires training the AI model from random weights multiple times. An alternative approach is to use an emulator to generate a synthetic dataset that is much longer than the training set and compare extreme event statistics with those of a proper ground truth. Note that in this approach, the physical consistency of the rare events would need to be fully examined in addition to summary statistics like return periods. A major challenge with this approach is that it requires a ground truth, e.g., a long dataset of at least hundreds of years, which is not available for ERA5, but can be produced with physics-based climate models.

Our work demonstrates the importance of testing and improving the behavior of AI weather/climate models on out-of-distribution gray swan events in addition to tests on typical global weather with measures such as the ACC, or on extreme weather events from distributions similar to that of the training set. We have shown that out-of-distribution extrapolation is nontrivial for extreme weather events and moving forward, the burden of proof is on anyone who claims it. This should be done with carefully designed metrics and reliable ground truths, as weak baselines and reporting biases can lead to overoptimism and misleading results (64). Given the outsized societal impact of extreme weather events, this charts a critical path forward for improving and fully exploiting powerful new AI weather and climate models.

Materials and Methods

ERA5 Data and Its TCs. All training and testing data are derived from the 0.25 degree ECMWF's ERA5 reanalysis (65), available through the Copernicus Climate Data Store (CDS, https://cds.climate.copernicus.eu/). As in ref. 1, data from 1979 to 2015 are used for training. Testing experiments in this study utilize data from 2018 to 2023. Each Category 5 TC test case consists of a control member (IC, directly from ERA5 at 0.25° resolution) and 50 members with perturbed ICs. The 50 perturbations are obtained from the EDA from ECMWF, accessible via the THORPEX Interactive Grand Global Ensemble (TIGGE) Data Retrieval (https://apps.ecmwf.int/datasets/data/tigge/). The EDA data, originally at 0.5° resolution, were regridded to the 0.25° grid to match the control member and be usable as FourCastNet's input. The perturbation fields of the EDA ensemble are first derived as the difference between the EDA members and their mean. We then rescale these fields by a factor of 0.1 and add them to the ERA5 field to generate 50 members of ICs for testing (inference).

TCs are identified within the ERA5 dataset by tracking closed mslp contours, with the minimum pressure serving as the criterion for intensity. The TC center is identified at the minimum mslp location. For the gradient-wind analysis, the TC center is defined as the location of the lowest 500 hPa geopotential height, given a possible vertical tilt in the TC structure.

Training FourCastNet with Five Training Sets. FourCastNet (1) is a recently developed global data-driven deep learning-based weather forecasting model that autoregressively predicts $\mathbf{x}(t + \Delta t)$ from $\mathbf{x}(t)$

$$\mathbf{x}(t + \Delta t) = \mathcal{M}(\mathbf{x}(t), \theta), \quad [2]$$

where $\mathbf{x}(t)$ is the 3D state of the atmosphere consisting of 20 atmospheric variables from ERA5 and θ represents the trainable parameters of the model. FourCastNet uses the Adaptive Fourier Neural Operator (AFNO) framework (66, 67) to efficiently parameterize the attention mechanism in a vision transformer. Both training and testing (inference) use a timestep of $\Delta t = 6$ h. We use FourCastNet's official code* and exact same state variables, architecture, and hyperparameters as in Pathak et al. (1). The one difference between the training procedure in Pathak et al. (1) and the one here is that we add zero-mean Gaussian noise with a variance of 0.3 to the inputs, i.e., $\mathbf{x}(t)$, during training. We find this to be essential for the stability of all generated ensemble forecasts during inference.

*https://github.com/NVlabs/FourCastNet.

For each training set (e.g., Full, noTC, etc.), we train five models, each starting from different random initializations of the trainable parameters to account for the uncertainty in the generalization error of the model. Each model is first trained for 80 epochs with a cosine learning-rate schedule at a starting learning rate of $\ell_1 = 5 \times 10^{-4}$ and then it is fine-tuned for an additional 50 epochs using a cosine learning-rate schedule with a lower learning rate of $\ell_2 = 10^{-4}$. The full training of one model takes roughly 7 d on 4 A100 GPUs.

Preparing the Five Training Sets. The histogram of the minimum of mslp in the tropics (30°S–30°N) for each sample in the training data is shown in Fig. 1. The plot identifies critical thresholds: the 5th percentile (∼970.0 hPa) and the 25th percentile (989.0 hPa). The fine-tuning step of FourCastNet requires two future timesteps, thus, in addition to $\mathbf{x}(t)$ (sample for which mslp falls below 25th in tropics), we must remove two additional samples ($\mathbf{x}(t - 2\Delta t)$ and $\mathbf{x}(t - \Delta t)$) from the training set. We therefore set the critical mslp threshold to 988.0 hPa to ensure that just 25% of the training data is excluded. The resulting training set (noTC) is used to train FourCastNet-noTC. Note that the 25th percentile threshold of mslp (<988.0 hPa) roughly corresponds to pressure at the center of major Category 3–5 TCs within the ERA5 dataset. Based on the IBTrACs data, approximately 24.8% of storms become major TCs (Category 3–5), and the median value of mslp in ERA5 for Category 3 TCs is 987.0 hPa, fairly close to the 988.0 hPa threshold we used (*SI Appendix*, Fig. S5). We define Category 5 in ERA5 as TCs that have mslp less than 970.0 hPa, sustained for 12 h or longer. Note that 970.0 hPa is lower than the average mslp values for Category 5 TCs of IBTraACs in the ERA5. There are in fact more than 20 Category 5 TCs in the real world during 2018–2023 based on the Saffir–Simpson hurricane wind scale (the average number is 5 to 7 per year). In this study, we train and test the AI model within the world of the ERA5 dataset. We use the Category 1–5 terminology mainly to facilitate communication.

FourCastNet-Rand is trained with the Rand dataset. Rand incorporates all the excluded samples in the noTC dataset. To match the size of the noTC training set, we randomly remove samples that do not include Category 3–5 TCs. Thus, Rand includes all major TCs and has the same total number of training samples and the annual cycle distribution as noTC.

The noWP (noNA) training set excludes samples when there are mslp values below 988.0 hPa in the tropical Western Pacific (North Atlantic) basin. There are many more strong TCs in the Western Pacific compared to the North Atlantic basin. This holds true for both the ERA5 data and observations. In our study, ∼51% of TCs removed in noTC are due to cyclones in the Western Pacific, while only ∼9% of the removed samples are due to TCs in the North Atlantic.

Data, Materials, and Software Availability. Some study data are available. Due to size limitations, some data cannot be uploaded to the repository. But we will share all our codes and provide all the details required for the readers to reproduce all the data used in this study themselves. Here are more details: We use the original FourCastNet with modifications for our customized training sets. These codes are publicly available at https://github.com/envfluids/FourCastNet (68). The necessary data to reproduce the results, including the weights of the 25 trained models and indices of dates that are removed in each training dataset, can be found on Zenodo at https://zenodo.org/uploads/13835657 (69) and https://zenodo.org/uploads/13834149 (70).

ACKNOWLEDGMENTS. We thank the editor and two anonymous reviewers for insightful comments and suggestions. This work was supported by ONR Award N000142012722 (to P.H.), ARO Grant W911NF-22-2- 0124 (to D.S.A. and J.W.), and NSF Grant AGS-2046309 (to P.H.). Computational resources were provided by NSF ACCESS (allocation ATM170020), NCAR's CISL (allocation URIC0009), and the University of Chicago Research Computing Center.

Author affiliations: [a]Department of the Geophysical Sciences, University of Chicago, Chicago, IL 60637; [b]Committee on Computational and Applied Mathematics, Division of the Physical Sciences, University of Chicago, Chicago, IL 60637; [c]Research Computing Center, The Office of the Provost, University of Chicago, Chicago, IL 60637; [d]Department of Applied Mathematics, University of California, Santa Cruz, CA 95064; and [e]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012

1. J. Pathak *et al.*, FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv [Preprint] (2022). https://arxiv.org/abs/2202.11214 (Accessed 1 November 2022).
2. K. Bi *et al.*, Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
3. R. Lam *et al.*, Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
4. S. Lang *et al.*, AIFS-ECMWF's data-driven forecasting system. arXiv [Preprint] (2024). https://arxiv.org/abs/2406.01465 (Accessed 12 August 2024).
5. L. Chen *et al.*, A cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **6**, 190 (2023).
6. T. Nguyen *et al.*, "Scaling transformers for skillful and reliable medium-range weather forecasting" in *ICLR 2024 Workshop on AI4 Differential Equations in Science* (2024).
7. S. Rasp *et al.*, WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.* **16**, e2023MS004019 (2024).
8. Z. Ben Bouallègue *et al.*, The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bull. Am. Meteorol. Soc.* **105**, E864–E883 (2024).
9. T. Kurth *et al.*, "FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators" in *Proceedings of the Platform for Advanced Scientific Computing Conference* (2023), pp. 1–11.
10. I. Price *et al.*, Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025). https://doi.org/10.1038/s41586-024-08252-9.
11. L. Li, R. Carver, I. Lopez-Gomez, F. Sha, J. Anderson, Generative emulation of weather forecast ensembles with diffusion models. *Sci. Adv.* **10**, eadk4489 (2024).
12. A. Mahesh *et al.*, Huge ensembles. Part I: Design of ensemble weather forecasts using spherical Fourier neural operators. arXiv [Preprint] (2024). https://arxiv.org/abs/2408.03100 (Accessed 15 August 2024).
13. A. Mahesh *et al.*, Huge ensembles. Part II: Properties of a huge ensemble of hindcasts generated with spherical Fourier neural operators. arXiv [Preprint] (2024). https://arxiv.org/abs/2408.01581 (Accessed 15 August 2024).
14. O. Watt-Meyer *et al.*, ACE: A fast, skillful learned global atmospheric model for climate prediction. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.02074 (Accessed 12 December 2023).
15. S. R. Cachay, B. Henn, O. Watt-Meyer, C. S. Bretherton, R. Yu, Probabilistic emulation of a global climate model with spherical dyffusion. arXiv [Preprint] (2024). https://arxiv.org/abs/2406.14798 (Accessed 1 September 2024).
16. J. P. Duncan *et al.*, Application of the AI2 climate emulator to E3SMV2'S global atmosphere model, with a focus on precipitation fidelity. *J. Geophys. Res.: Mach. Learn. Comput.* **1**, e2024JH000136 (2024).
17. H. Guan, T. Arcomano, A. Chattopadhyay, R. Maulik, Lucie: A lightweight uncoupled climate emulator with long-term stability and physical consistency for o (1000)-member ensembles. arXiv [Preprint] (2024). https://arxiv.org/abs/2405.16297 (Accessed 1 June 2024).
18. C. Y. Lai *et al.*, Machine learning for climate physics and simulations. arXiv [Preprint] (2024). https://arxiv.org/abs/2404.13227 (Accessed 22 August 2024).
19. N. Cresswell-Clay *et al.*, Deep learning earth system model for stable and efficient simulation of the current climate. arXiv [Preprint] (2024). https://arxiv.org/abs/2409.16247 (Accessed 1 October 2024).
20. S. Dheeshjith *et al.*, Transfer learning for emulating ocean climate variability across CO_2 forcing. arXiv [Preprint] (2024). https://arxiv.org/abs/2405.18585 (Accessed 5 August 2024).
21. A. Bracco *et al.*, Machine learning for the physics of climate. *Nat. Rev. Phys.* **7**, 6–20 (2024).
22. D. Kochkov *et al.*, Neural general circulation models for weather and climate. *Nature* **632**, 1060–1066 (2024).
23. A. Chattopadhyay, P. Hassanzadeh, Long-term instabilities of deep learning-based digital twins of the climate system: The cause and a solution. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.07029 (Accessed 7 December 2024).
24. C. B. Field, V. Barros, T. F. Stocker, Q. Dahe, *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2012).
25. F. Ragone, J. Wouters, F. Bouchet, Computation of extreme heat waves in climate models using a large deviation algorithm. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 24–29 (2018).
26. V. Eyring *et al.*, Pushing the frontiers in climate modelling and analysis with machine learning. *Nat. Clim. Change* **14**, 916–928 (2024).
27. B. Krawczyk, Learning from imbalanced data: Open challenges and future directions. *Progr. Artif. Intell.* **5**, 221–232 (2016).
28. A. Chattopadhyay, E. Nabizadeh, P. Hassanzadeh, Analog forecasting of extreme-causing weather patterns using deep learning. *J. Adv. Model. Earth Syst.* **12**, e2019MS001958 (2020).
29. J. E. Walsh *et al.*, Extreme weather and climate events in northern areas: A review. *Earth-Sci. Rev.* **209**, 103324 (2020).
30. I. Ebert-Uphoff, K. Hilburn, Evaluation, tuning and interpretation of neural networks for working with images in meteorological applications. *Bull. Am. Meteorol. Soc.* **101**, E2149–E2170 (2020).
31. G. Miloshevich, B. Cozian, P. Abry, P. Borgnat, F. Bouchet, Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *Phys. Rev. Fluids* **8**, 040501 (2023).
32. O. C. Pasche, J. Wider, Z. Zhang, J. Zscheischler, S. Engelke, Validating Deep Learning Weather Forecast Models on Recent High-Impact Extreme Events. *Artif. I. earth syst.* **4**, e240033 (2025).
33. A. J. Charlton-Perez *et al.*, Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of storm Ciarán. *npj Clim. Atmos. Sci.* **7**, 93 (2024).
34. M. DeMaria *et al.*, Evaluation of tropical cyclone track and intensity forecasts from artificial intelligence weather prediction (AIWP) models. arXiv [Preprint] (2024). https://arxiv.org/abs/2409.06735 (Accessed 15 September 2024).
35. M. Bonavita, On some limitations of current machine learning weather prediction models. *Geophys. Res. Lett.* **51**, e2023GL107377 (2024).
36. T. Selz, G. C. Craig, Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophys. Res. Lett.* **50**, e2023GL105747 (2023).

37. N. Lin, K. Emanuel, Grey swan tropical cyclones. *Nat. Clim. Change* **6**, 106–111 (2016).
38. G. J. Hakim, S. Masanam, "Dynamical tests of a deep-learning weather prediction model" in *Artificial Intelligence for the Earth Systems* (2024).
39. T. Rackow *et al.*, Robustness of AI-based weather forecasts in a changing climate. arXiv [Preprint] (2024). https://arxiv.org/abs/2409.18529 (Accessed 1 October 2024).
40. T. Beucler *et al.*, Climate-invariant machine learning. *Sci. Adv.* **10**, eadj7250 (2024).
41. H. E. Willoughby, Gradient balance in tropical cyclones. *J. Atmos. Sci.* **47**, 265–274 (1990).
42. E. Guo *et al.*, FourCastNext: Improving FourCastNet training with limited compute. arXiv [Preprint] (2024). https://arxiv.org/abs/2401.05584 (Accessed 1 April 2024).
43. T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, A. Grover, Climax: A foundation model for weather and climate. arXiv [Preprint] (2023). https://arxiv.org/abs/2301.10343 (Accessed 20 December 2023).
44. J. Schmude *et al.*, Prithvi WxC: Foundation model for weather and climate. arXiv [Preprint] (2024). https://arxiv.org/abs/2409.13598 (Accessed 25 September 2024).
45. X. Wang *et al.*, Orbit: Oak ridge base foundation model for earth system predictability. arXiv [Preprint] (2024). https://arxiv.org/abs/2404.14712 (Accessed 24 August 2024).
46. X. Man, C. Zhang, J. Feng, C. Li, J. Shao, W-MAE: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.08754 (Accessed 20 December 2023).
47. C. Lessig *et al.*, AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning. arXiv [Preprint] (2023). https://arxiv.org/abs/2308.13280 (Accessed 12 September 2023).
48. A. McNally *et al.*, Data driven weather forecasts trained and initialised directly from observations. arXiv [Preprint] (2024). https://arxiv.org/abs/2407.15586 (Accessed 27 August 2024).
49. Y. Q. Sun, P. Hassanzadeh, M. J. Alexander, C. G. Kruse, Quantifying 3D gravity wave drag in a library of tropical convection-permitting simulations for data-driven parameterizations. *J. Adv. Model. Earth Syst.* **15**, e2022MS003585 (2023).
50. L. M. Yang, E. P. Gerber, Overcoming set imbalance in data driven parameterization: A case study of gravity wave momentum transport. arXiv [Preprint] (2024). https://arxiv.org/abs/2402.18030 (Accessed 2 March 2024).
51. I. Shumailov *et al.*, AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
52. J. L. Willson *et al.*, DCMIP 2016: The tropical cyclone test case. *Geosci. Model Dev.* **17**, 2493–2507 (2024).
53. A. Chattopadhyay, M. Mustafa, P. Hassanzadeh, E. Bach, K. Kashinath, Towards physics-inspired data-driven weather forecasting: Integrating data assimilation with a deep spatial-transformer-based U-NET in a case study with ERA5. *Geosci. Model. Dev.* **15**, 2221–2237 (2022).
54. R. Wang, R. Walters, R. Yu, Data augmentation vs. equivariant networks: A theory of generalization on dynamics forecasting. arXiv [Preprint] (2022). https://arxiv.org/abs/2206.09450 (Accessed 1 February 2023).
55. R. J. Webber, D. A. Plotkin, M. E. O'Neill, D. S. Abbot, J. Weare, Practical rare event sampling for extreme mesoscale weather. *Chaos* **29**, 053109 (2019).
56. F. Ragone, F. Bouchet, Rare event algorithm study of extreme warm summers and heatwaves over Europe. *Geophys. Res. Lett.* **48**, e2020GL091197 (2021).
57. D. S. Abbot, R. J. Webber, S. Hadden, D. Seligman, J. Weare, Rare event sampling improves mercury instability statistics. *Astrophys. J.* **923**, 236 (2021).
58. J. Finkel, E. P. Gerber, D. S. Abbot, J. Weare, Revealing the statistics of extreme events hidden in short weather forecast data. *AGU Adv.* **4**, e2023AV000881 (2023).
59. J. Finkel, P. A. O'Gorman, Bringing statistics to storylines: Rare event sampling for sudden, transient extreme events. *J. Adv. Model. Earth Syst.* **16**, e2024MS004264 (2024).
60. V. Jacques-Dumas, F. Ragone, P. Borgnat, P. Abry, F. Bouchet, Deep learning-based extreme heatwave forecast. *Front. Clim.* **4**, 789641 (2022).
61. H. Zhang, J. Finkel, D. S. Abbot, E. P. Gerber, J. Weare, Using explainable AI and transfer learning to understand and predict the maintenance of Atlantic blocking with limited observational data. arXiv [Preprint] (2024). https://arxiv.org/abs/2404.08613 (Accessed 16 April 2024).
62. D. S. Abbot, J. Laurence-Chasen, R. J. Webber, D. M. Hernandez, J. Weare, AI can identify solar system instability billions of years in advance. *Res. Notes AAS* **8**, 3 (2024).
63. E. M. Fischer *et al.*, Storylines for unprecedented heatwaves based on ensemble boosting. *Nat. Commun.* **14**, 4643 (2023).
64. N. McGreivy, A. Hakim, Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nat. Mach. Intell.* **6**, 1256–1269 (2024).
65. H. Hersbach *et al.*, The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
66. Z. Li *et al.*, Fourier neural operator for parametric partial differential equations. arXiv [Preprint] (2020). https://arxiv.org/abs/2010.08895 (Accessed 1 February 2023).
67. J. Guibas *et al.*, Adaptive Fourier neural operators: Efficient token mixers for transformers. arXiv [Preprint] (2021). https://arxiv.org/abs/2111.13587 (Accessed 12 February 2023).
68. Y. Q. Sun *et al.*, Gray Swan Tropical Cyclone Forecasting with AI? Github. https://github.com/envfluids/FourCastNet. Deposited 13 October 2024.
69. Y. Q. Sun *et al.*, UChicago Hurricane AI Project - Results 1. Zenodo. https://zenodo.org/records/13835657. Deposited 24 September 2024.
70. Y. Q. Sun *et al.*, UChicago Hurricane AI Project - Model Weights. Zenodo. https://zenodo.org/records/13834149. Deposited 24 September 2024.