ENVIRONMENTAL RESEARCH
LETTERS

PURPOSE-LED PUBLISHING™ Part of the ENVIRONMENTAL RESEARCH SERIES

LETTER · OPEN ACCESS

Advancing seasonal prediction of tropical cyclone activity with a hybrid AI-physics climate model

To cite this article: Gan Zhang *et al* 2025 *Environ. Res. Lett.* **20** 094031

View the article online for updates and enhancements.

You may also like

- Innovative Ideas in Science 2016
- Spatial and Temporal Variation Characteristics of Northwest Pacific Tropical Cyclone Activity in Global Warming Scenario
  Xinyu Guo, Chenglin Gu, Bei Li et al.
- Northern Hemisphere high latitude climate and environmental change
  Pavel Groisman and Amber Soja

This content was downloaded from IP address 192.17.106.164 on 27/08/2025 at 19:36

IOP Publishing Environ. Res. Lett. 20 (2025) 094031 https://doi.org/10.1088/1748-9326/adf864

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
14 May 2025

REVISED
17 July 2025

ACCEPTED FOR PUBLICATION
6 August 2025

PUBLISHED
15 August 2025

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

LETTER

Advancing seasonal prediction of tropical cyclone activity with a hybrid AI-physics climate model

Gan Zhang[1,*], **Megha Rao**[1], **Janni Yuval**[2] **and Ming Zhao**[3]

[1] Department of Climate, Meteorology, and Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States of America
[2] Google Research, Mountain View, CA 94043, United States of America
[3] Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Princeton, NJ 08540, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: gzhang13@illinois.edu

Keywords: climate prediction, machine learning, tropical cyclone, climate model, model evaluation

Supplementary material for this article is available online

Abstract

Machine learning (ML) models are successful with weather forecasting and have shown progress in climate simulations, yet leveraging them for useful climate predictions needs exploration. Here we show this feasibility using neural general circulation model (NeuralGCM), a hybrid ML-physics atmospheric model developed by Google, for seasonal predictions of large-scale atmospheric variability and Northern Hemisphere tropical cyclone (TC) activity. Inspired by physical model studies, we simplify boundary conditions, assuming sea surface temperature and sea ice follow their climatological cycle but persist anomalies present at the initialization time. With such forcings, NeuralGCM can generate 100 simulation days in ∼8 min with a single graphics processing unit while simulating realistic atmospheric circulation and TC climatology patterns. This configuration yields useful seasonal predictions (July–November) for the tropical atmosphere and various TC activity metrics. Notably, the predicted and observed TC frequency in the North Atlantic and East Pacific basins are significantly correlated during 1990–2023 ($r = \sim 0.7$), suggesting prediction skill comparable to existing physical GCMs. Despite challenges associated with model resolution and simplified boundary forcings, the model-predicted interannual variations demonstrate significant correlations with the observed sub-basin TC tracks ($p < 0.1$) and basin-wide accumulated cyclone energy (ACE) ($p < 0.01$) of the North Atlantic and North Pacific basins. These findings highlight the promise of leveraging ML models with physical insights to model TC risks and deliver seamless weather-climate predictions.

1. Introduction

Machine learning (ML) models recently made breakthroughs in weather forecasting (e.g. Keisler 2022, Pathak *et al* 2022, Bi *et al* 2023, Lam *et al* 2023, Price *et al* 2023, Kochkov *et al* 2024). Trained with atmospheric reanalysis or physical model data, these ML models delivered successful forecasts up to two weeks of lead time with skills comparable to or better than conventional numeric weather forecast (NWP) models. With computational costs at a fraction of NWP models (10^{-3}–10^{-5}), the new ML models unlocked opportunities for improving operational weather service (e.g. Lang *et al* 2024) and advancing fundamental understanding of atmospheric predictability (e.g. Vonich and Hakim 2024). Similar to the early development of NWP and climate models (Phillips 1956), the success of ML models in weather forecasting also inspired researchers to explore their potential applications in climate modeling (Bretherton *et al* 2022, Eyring *et al* 2024). Nonetheless, the feasibility of conducting successful climate simulations and society-relevant climate predictions with the new ML models remains to be explored.

Recent efforts in leveraging the new ML models for climate simulations focused on attaining stable long-term simulations and emulating

© 2025 The Author(s). Published by IOP Publishing Ltd

atmosphere–ocean interactions. For example, Cresswell-Clay *et al* (2024) trained ML emulators for the atmosphere and the surface ocean separately and showed that linking the two emulators can generate stable atmospheric simulations of the current climate for over 1000 years. Other endeavors emphasized achieving stable long-term simulations by incorporating various physical constraints. Bonev *et al* (2023) achieved one-year stable rollouts with the Fourier neural operator (FNO) by replacing an unrealistic flat geometry with spheric geometries. Based on the spheric FNO, Wang *et al* (2024) used gridded atmosphere and ocean data to train separate emulators and link them during roll-outs. With a configuration of lagged ensemble forecasting, their linked emulators achieved skillful seasonal predictions of the El Niño–Southern Oscillation. Watt-Meyer *et al* (2024) introduced mass and moisture constraints to the spheric FNO framework and completed an 80 year historical simulation with realistic atmospheric variability. This set of simulations performed well with in-sample climate forcings but showed unrealistic responses to out-of-sample climate forcings (i.e., zero-shot learning), such as high levels of sea surface temperature (SST) and carbon dioxide. To overcome such limitations, Beucler *et al* (2024) proposed to incorporate the physical knowledge of subgrid processes, which helped ML emulators trained with physical model outputs better generalize across climate regimes.

Distinct from those ML emulators, Kochkov *et al* (2024) developed an ML-physics hybrid model neural general circulation model (NeuralGCM) and achieved multi-decade, stable atmospheric simulations. This model contains an atmospheric dynamical core like the conventional GCMs but replaces the parameterized subgrid physics with ML substitutes. Compared to the existing ML climate emulators, NeuralGCM stands out with its structural similarity to the conventional NWP models and atmospheric GCMs that are grounded on physical principles. NeuralGCM is highly skillful for weather forecasting and can incorporate observed SSTs to simulate climate anomalies, such as simulating realistic tracks and numbers of TCs in the active 2020 Atlantic hurricane season Kochkov *et al* (2024). These traits make NeuralGCM a promising candidate for modeling extreme risks and developing a seamless weather–climate prediction system (Brunet *et al* 2010, Hoskins 2013), provided that NeuralGCM can be configured to deliver skillful climate predictions.

While the current version of NeuralGCM lacks the means to simulate boundary conditions (e.g. ocean and sea ice) and the support for the atmosphere–ocean coupling, previous studies with physical models suggest that simple assumptions of boundary forcings can help establish a performance baseline for atmospheric GCMs in seasonal prediction tasks. Specifically, Zhao *et al* (2010) showed that assuming persistent SST anomalies with a climatological seasonal cycle can help an atmospheric GCM skillfully predict tropical cyclone (TC) activity in the North Atlantic and the Northeastern Pacific. Chen and Lin (2013) suggested that the prediction skill improves when the atmospheric GCM is initialized with observed conditions instead of random conditions from climate simulations. The success of these early studies with physical models builds on the thermal inertia of the tropical ocean and the strong influences of tropical SST on the global atmosphere (e.g. Shukla 1998) and TC activity (e.g. Gray 1984). The exploratory work with atmospheric GCMs served as a stepping stone for the ensuing development of more advanced prediction systems (e.g. Vecchi *et al* 2014, Delworth *et al* 2020).

Inspired by the recent NeuralGCM development and the previous physical studies, this study explores the feasibility of leveraging NeuralGCM to deliver skillful seasonal climate prediction. We emphasize TC activity since these storms are a leading contributor to life losses and economic damages (World Meteorological Organization 2021) and often remain challenging for physical GCMs to simulate (Roberts *et al* 2020). This TC focus also helps us leverage proven concepts and knowledge in physical model development (Zhao *et al* 2010, Chen and Lin 2013). Overall, this effort establishes a performance baseline for future model development that seeks to extend our climate modeling capability and deliver societally valuable predictions (Emanuel *et al* 2012, Lemoine and Kapnick 2024).

2. Data and methods

2.1. Observational data

The fifth-generation ECMWF atmospheric reanalysis (ERA5) (Hersbach *et al* 2020) serves as the primary data for the training, configuration, and validation of NeuralGCM simulations. The gridded ERA5 is generated by an NWP model that follows physical laws and ingests multi-sourced observational data (e.g. weather station and satellite data). The original grid spacing of ERA5 is approximately 0.25° and contains variables at pressure levels and the surface (e.g. SST and sea ice coverage). The ERA5 data from 1979–2017 and 1979–2019 is used to train the deterministic and the stochastic NeuralGCM, respectively (Kochkov *et al* 2024). Since the training is based on narrow time windows ($\leqslant$5 d), NeuralGCM does not directly learn the seasonal evolution trajectories. To facilitate the configuration and validation of retrospective prediction experiments, we regrid the ERA5 data to match the grid of NeuralGCM. While many recent ML studies use TC tracks extracted from the ERA5 for model evaluation, we evaluate TC predictions using the International Best Track Archive for Climate Stewardship (Knapp *et al* 2010). This dataset

includes a collection of hurricane information based on multi-sourced observations and expert quality control. The best track dataset is widely used in TC research and real-world risk modeling and is generally considered more trustworthy than the reanalysis datasets (e.g., ERA5) that struggle with representing intense hurricanes (Dulac *et al* 2024).

2.2. NeuralGCM and hindcast experiments

NeuralGCM includes a differentiable dynamical core that solves the governing equations of atmospheric dynamics and a neural network that parameterizes unresolved processes of atmospheric columns (Kochkov *et al* 2024). We use the pre-trained, 1.4° version of NeuralGCM to balance the need to conduct ensemble predictions and simulate realistic TC activity. At this resolution, NeuralGCM has two models: a deterministic model and a stochastic model (Kochkov *et al* 2024). The deterministic model was extensively evaluated and showed promise in simulating realistic TC activity in the test of the year 2020 (Kochkov *et al* 2024). The stochastic configuration uses random seeds to generate space-time correlated Gaussian random fields for perturbing initial conditions and insert stochasticity into the neural network parameterization. These random fields are independent of each other and conceptually resemble the NWP techniques of perturbing the initial fields and the parameterized model physics (Kochkov *et al* 2024). Since we use the NeuralGCM versions trained by Kochkov *et al* (2024) without modifying the model architecture or parameterized physics, we provide a high-level technical description in supplementary materials and encourage interested readers to consult Kochkov *et al* (2024) for more details.

We conduct hindcast experiments using both deterministic and stochastic configurations to assess the potential sensitivity of TC activity to the learned model physics parameters. To generate ensemble predictions with the deterministic model, we introduce perturbations to the initial conditions using a Gaussian random field. This field, initialized from a random seed, is applied to the learned correction within the NeuralGCM's encoder. Specifically, the encoder interpolates ERA5 initial conditions to sigma levels and subsequently learns a correction to this interpolation. We then perturb this correction by multiplying it by a factor of (1 + random_field_value), where random_field_value represents the value from the generated Gaussian random field with correlation length of 1000 km for the deterministic model. We use this perturbation strategy to initialize twenty-member ensemble simulations at 0 UTC on 1 July for each year from 1990 to 2023. We also generated additional simulations (e.g. 1979–1989) to facilitate comparisons with previous TC studies that used physical models (supplementary materials).

Inspired by the seasonal prediction experiments by Zhao *et al* (2010) and Chen and Lin (2013), we use the climatological seasonal cycle and persistent anomalies of SST and sea ice to drive the NeuralGCM. Based on the autocorrelation of SST and sea ice, this configuration can approximate the evolution of tropical SST (Chen and Lin 2013) and sea ice (Bushuk *et al* 2022) during July–November. When calculating the anomalies of SST and sea ice at the initialized time, we use the daily climatology of 1991–2020 that is resampled using the monthly data. To ensure the consistency among variables and the configurations described by Kochkov *et al* (2024), the initial states of the SST, sea ice, and atmosphere are acquired from the ERA5. At later steps of the prediction experiment, we force NeuralGCM with the pre-calculated SST and sea ice fields, namely the sum of their daily climate values and anomalies at the initialization time. Therefore, all the information needed for long-range predictions is available near the initialization time. We run the predictions for approximately five months to cover much of the TC season of the Northern Hemisphere. We acknowledge the assumption of persistent anomalies has limitations and consider the prediction skill of our experiments as a lower bound on the attainable skills. The 1.4° versions of NeuralGCM with the simplified boundary forcings can finish 100 simulation days in ∼8 min with a single graphics processing unit (GPU) (supplementary table 1).

2.3. Post-processing and evaluation

The combination of the stochastic NeuralGCM and modified boundary conditions yields stable multi-month predictions in most cases. While the hindcasts with the deterministic physics are generally stable (∼98.5%) in the tropics, about 10% of the simulations with the stochastic physics configuration show spurious small-scale waves (supplementary figures 1 and 2) associated with unrealistic convection and stratosphere features. These waves mostly appear in the tropics and violate the weak gradient constraint of the real-world atmosphere (Charney 1963, Sobel and Bretherton 2000). While fixes are being explored, this study proceeds by labeling the simulations with spurious waves using a check of tropical variability. Specifically, we calculate the standard deviations of 500 hPa geopotential height in the zonal direction and compare the metric between the initial and later prediction steps. If spurious small-scale waves develop, they will substantially increase the zonal variability and thus the instability metric. If this metric exceeds two times the initial metric values at any latitudes during the roll-outs, we flag the corresponding simulation as unstable and assign all the fields to climate values. The flagging is robust to small changes

in the threshold as the spurious waves usually amplify quickly once appearing in the rollouts. As the supplementary materials will show, the skills of the deterministic and the stochastic hindcasts in predicting TC activity are comparable. Unless otherwise specified, the analyses and discussion in the main text focus on the hindcasts with the more stable configuration with deterministic model physics.

The prediction evaluation includes selected environmental variables and metrics of TC activity. The dynamical variables include the 500 hPa geopotential height, which characterizes the steering flow that affects TC tracks, and the 200–850 hPa vertical wind shear, which affects TC genesis and development. While evaluating convection-related variables is important, the NeuralGCM version used here does not include precipitation variables. A new version that can simulate realistic precipitation is under development (Yuval *et al* 2024). We apply the TempestExtreme package (Ullrich *et al* 2021) to track TCs in our retrospective prediction experiments. The tracking uses the vorticity-based method and does not impose any wind speed thresholds. We follow most parameter choices of the TC tracker used by Kochkov *et al* (2024) who tuned the parameters such that the TC counts of the ERA5 at 0.25-degree grid spacing match the values at 1.4-degree grid spacing. To better match the TC counts in the IBTrACS, we lower the vorticity threshold to $4 \times 10^{-5}\,s^{-1}$ and set the storm duration threshold to 54 h.

Following previous studies of TC activity (e.g. Zhao *et al* 2010, Chen and Lin 2013, Zhang *et al* 2021), we mainly evaluate the ensemble mean and examine the metrics of anomaly correlation coefficient and root-mean-squared error. The evaluated variables include detrended environmental variables, regional TC counts, and the ACE. The ACE is defined as the sum of the squares of the maximum wind speed (knots) of all the available track data with a scaling factor of 10^{-4}. We also provide results from other models (e.g. Chen and Lin 2013, Johnson *et al* 2019, Zhang *et al* 2019) for reference. These models, such as the ECMWF seasonal forecasts (SEAS5) (Johnson *et al* 2019), are physical models with higher spatial resolutions of the atmosphere (e.g., 36 km grid spacing). We briefly discuss the performance of models in section 3 and present additional analyses (e.g. SEAS5) and considerations for more comprehensive comparisons in supplementary materials.

3. Results

3.1. Model skill with large-scale atmospheric environment

The NeuralGCM hindcasts with simplified boundary forcings simulate the atmospheric climate and seasonal cycle realistically (figure 1). The July–November means of the 500 hPa geopotential of the NeuralGCM and the ERA5 show consistent climate patterns. Their differences are the smallest in the tropics and the largest in the Arctic region. An inspection of the seasonal evolution of the zonal means of the 500 hPa geopotential suggests the model biases grow over time. The initial biases emerge in the polar region and develop relatively rapidly during the transition season. The biases in the tropics are relatively small and comparable to those in a fully coupled physical prediction system (supplementary figure 3). Preliminary analyses (not shown) suggest that the high-latitude biases are related to the simplified boundary forcings, especially the ice representation in polar regions. The seasonally evolving geopotential biases affect the midlatitude jet streams and may ultimately distort some aspects of the tropical–extratropical teleconnections and TC activity (Zhang *et al* 2016, Wang *et al* 2020). Comparing other atmospheric variables suggests similar evolution between the climate states of the NeuralGCM and the ERA5 (not shown). Similar biases and consistency are present in the hindcasts with the stochastic version of NeuralGCM (not shown). Their overall consistency between the model climate and the observation is notable considering the simplified boundary forcings and the lack of complex atmosphere-land-ocean coupling in the NeuralGCM hindcasts.

The NeuralGCM hindcasts also show skills in predicting year-to-year variability of the monthly mean atmospheric environment. We examine variables including the 500 hPa geopotential, surface pressure, 1000 hPa temperature, and vertical shear of zonal wind (200–850 hPa) and find that the NeuralGCM hindcasts show various skill levels across the examined month leads (supplementary figures 4–7). Preliminary comparisons between the NeuralGCM hindcasts (supplementary figures 4–7) and operational seasonal predictions by a fully coupled physical model (Johnson *et al* 2019; supplementary figures 8–10) suggests the anomaly correlation coefficients with the observation are overall lower for the NeuralGCM hindcasts with simplified boundary forcings. Nonetheless, the anomaly correlation coefficients for NeuralGCM hindcasts show spatial-temporal patterns similar to those of the physical model. The anomaly correlation coefficients of the initial month are the highest and decrease with forecast lead time. While the decay quickly makes extratropical predictions unskillful, the prediction skill persists at much longer forecast lead in the tropics, as suggested by the relatively high correlation coefficients and low prediction errors (supplementary figures 4–7). The relatively high skill in the tropics corresponds to regions where the SST strongly regulates atmospheric variability (Shukla 1998), consistent with similar experiments with physical GCMs (Chen and Lin 2013).

We next focus on the prediction of the atmospheric environment in the main development regions (MDRs). The MDRs, as outlined in

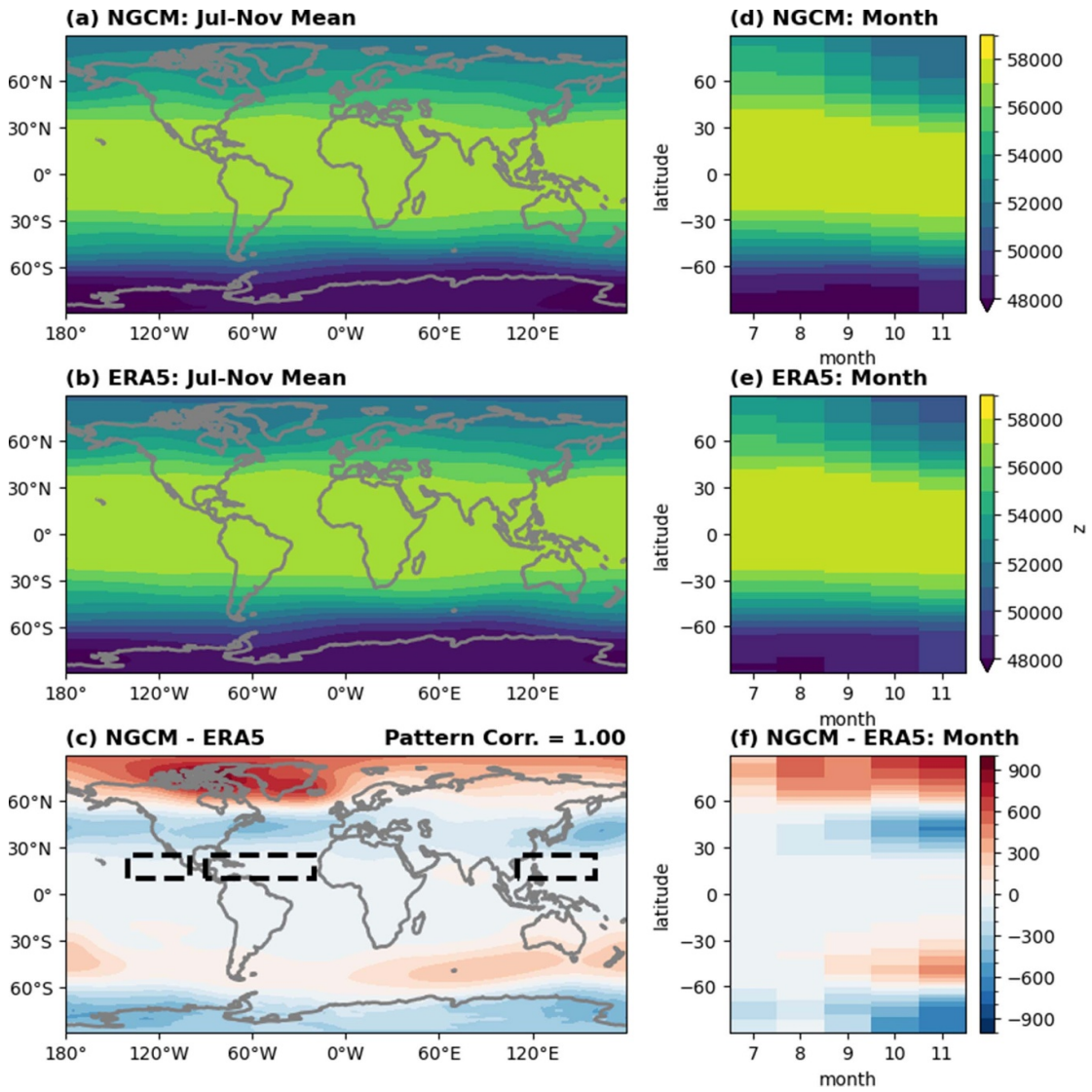

Figure 1. Comparison of the large-scale atmospheric environment in NeuralGCM hindcasts and the ERA5. (a) July–November ensemble mean (1990–2019) of the 500 hPa geopotential (m^2 s^{-2}) in the NeuralGCM hindcasts. (b) Same as (a), but for the ERA5. (c) The difference between (a) the NeuralGCM hindcasts and (b) the ERA5. The pattern correlation (rounded to 2 decimal places) between (a) and (b) is denoted in the top right corner of (c). (d) The monthly evolution of the zonal mean of the 500 hPa geopotential (m^2 s^{-2}) in the NeuralGCM hindcasts. (e) Same as (d), but for the ERA5. (f) The difference between (d) the NeuralGCM hindcasts and (e) the ERA5. The dashed black lines in (c) highlight the main development regions (MDRs) where the environmental prediction skill is evaluated. The definitions of the MDRs roughly follow previous studies but use slightly larger latitudinal ranges (10° N–25° N). The longitudinal ranges for the MDRs of the North Atlantic, the Northeast Pacific, and the Northwest Pacific are 90° W–20° W, 140° W–100° W, and 110° E–160° E.

figure 1(c), span over the tropical North Pacific and North Atlantic, which contribute a majority of TCs that form in the Northern Hemisphere (Goldenberg *et al* 2001, Doi *et al* 2013, Jien *et al* 2015, Zhang and Wang 2015, Feng *et al* 2021). Figure 2 shows the skill of the NeuralGCM with simplified boundary forcings in predicting the MDR atmospheric environment. The predictions of near-surface air temperature and the 500 hPa geopotential show the highest anomaly correlation coefficients with the observation across the three examined MDRs. For the prediction of these two variables, the anomaly correlation coefficients are statistically significant for each calendar month (figures 2(a)–(c)). In comparison, the anomaly correlation coefficients for the surface pressure and the zonal wind shear are much lower but can remain statistically significant in August (lead days = 31–62). The anomaly correlation coefficients generally exceed those of persisting the monthly anomalies of June (supplementary figure 11). The findings thus suggest that the NeuralGCM with simplified boundary forcings can predict some aspects of the large-scale atmospheric variability in the MDRs. Such skills in predicting the large-scale environment are essential for the subseasonal-to-seasonal predictions of TC activity.

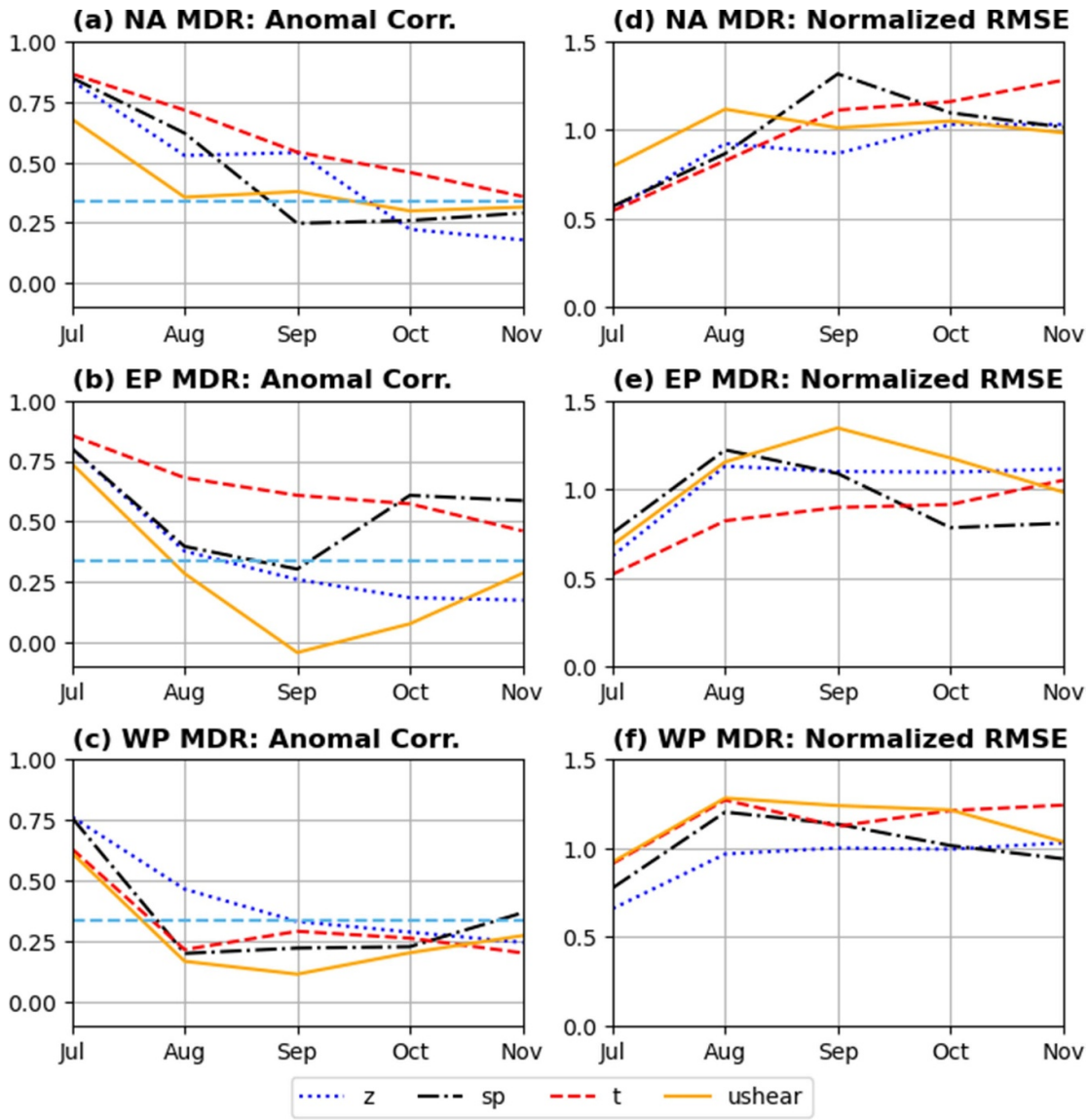

Figure 2. Prediction skills of NeuralGCM for environmental fields in key TC regions. The plot shows selected atmospheric variables, including the 500 hPa geopotential (z, unit: m^2 s^{-2}), surface pressure (sp, unit: hPa), 1000 hPa temperature (t, unit: K), and vertical shear of zonal wind (200–850 hPa) (ushear). (a) The anomaly correlation coefficient of the MDR in the North Atlantic (NA). (b) Same as (a), but for the Northeast Pacific (EP). (c) Same as (a), but for the Northwest Pacific (WP). (d)–(f) Same as (a), (b), (c), but for the normalized root-mean-squared error. The scaling factor is the year-to-year standard deviations of the ERA5 data. The skill metrics are calculated using data from 1990–2023 for each grid point and then averaged over the MDR domains highlighted in figure 1(c). The input variables were detrended using a linear least-squared fitting. The 95% confidence threshold for the anomaly correlation coefficient is decided with t-statistics and denoted with horizontal cyan dashed lines in (a)–(c).

3.2. Model skill with TC activity

The NeuralGCM hindcast can simulate a realistic spatial-temporal distribution of TC activity (figure 3). The kernel density estimation of the simulated and observed TC activity shows consistent spatial patterns, including the high density of TC tracks in parts of the Northeast Pacific and the Northwest Pacific (figures 3(a)–(c)). The relative track density among the Northern Hemisphere basins is also realistic, free of the common bias of many physical GCMs in severely underestimating TC activity in the Northern Hemisphere (e.g., Roberts *et al* 2020). The seasonal cycles of TC activity in the Neural GCM and the observation are also similar (figures 3(d) and (e)). The similarities include the latitudinal shift towards lower latitudes in the late season and the high concentration (∼75%) of samples in July-September. The comparisons also show subtle biases of the NeuralGCM hindcasts with deterministic physics. For instance, the kernel density of TC tracks is too high near 10–20° N in the Northeast Pacific; the decay of TC activity in the late season is also too fast, with October–November accounting for 17% instead of 25% of tracks. Similar seasonality biases are present in the NeuralGCM hindcasts with stochastic physics (not shown). This suggests these biases might arise from the simplified boundary forcings and biases in simulating the large-scale environment (figure 1(f)),

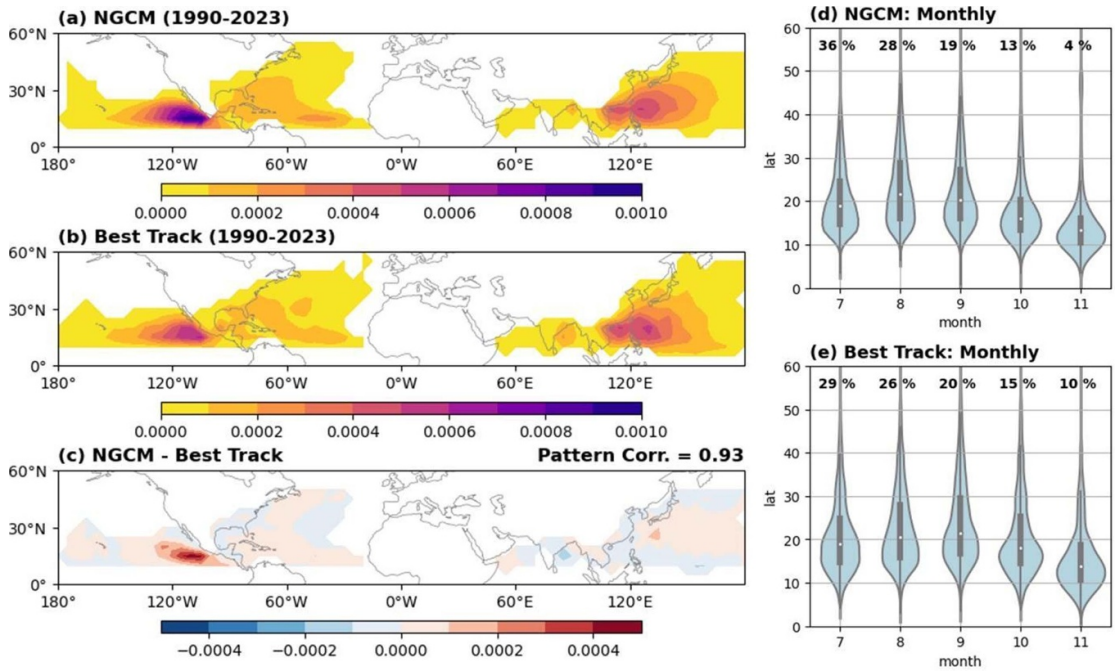

Figure 3. Comparison of July–November TC activity in NeuralGCM hindcasts and the IBTrACS (1990–2023). (a) Kernel density estimates (KDEs) of the TCs identified in the NeuralGCM hindcasts. (b) Same as (a), but for the best track observations. (c) The difference between (a) the NeuralGCM hindcasts and (b) the IBTrACS. The KDEs are calculated on a latitude–longitude grid with 5° spacing, and the unit of KDEs is $1/(25^{\circ 2})$. The pattern correlation between (a) and (b) is denoted in the top right corner of (c). (d) The monthly evolution of the latitudinal distributions of the Northern Hemisphere TCs in the NeuralGCM hindcasts. The miniature box-and-whisker plots show the median, the interquartile range (IQR), and the 1.5× IQR range. The monthly percentage of the track points in the July–November total is denoted in the upper part. (e) Same as (d), but for the best track.

though the parameter choice of tracking algorithms may also be a contributing factor.

The NeuralGCM hindcast also simulates interannual variations of TC activity that are significantly correlated with the observation. The correlations between the seasonal prediction and the observation of the basin-wide TC frequency are statistically significant in the North Atlantic and the Northeast Pacific (figure 4). Despite the much lower model resolution and computational costs, the NeuralGCM hindcasts demonstrate skill comparable to previous physical model simulations with similar simplified boundary forcings (Zhao *et al* 2010, Chen and Lin 2013) or a more realistic representation of boundary forcings (e.g. Zhang *et al* 2019) in those two basins. Table 1 shows a direct, like-for-like comparison of prediction skill for TC frequency. When model performance is ranked for the North Atlantic and the Northeast Pacific, the NeuralGCM hindcasts with deterministic physics are comparable to or better than at least one of the examined physical models. The prediction skill is associated with the relationship between TC activity and the large-scale atmospheric environment; moreover, the NeruralGCM hindcast can also predict at least some aspects of TC activity on the subseasonal scale, in hyperactive seasons, and beyond the model training period (supplementary materials).

The prediction of TC frequency in the Northwest Pacific and the North Indian Ocean is relatively poor (figures 4(c) and (d)) and appears related to the prediction skill of large-scale environment. Consistent with the results of the basin-wide TC frequency, the predicted track density in open-ocean areas is significantly correlated with the observation in parts of the North Atlantic and North Pacific but not the North Indian Ocean (figure 5(a)). The regions with high prediction skills are consistent with physical model simulations and predictability analysis (Zhang *et al* 2019). Comparable basin-wide and regional correlations for the NeuralGCM hindcasts with the stochastic configuration (figure 5(b) and supplementary figure 12). Since the environmental constraint of convective activity is crucial for long-range predictions (e.g. Shukla 1998), low skill in predicting TC activity is likely associated with large-scale environmental variables. For instance, the prediction skill of the atmospheric environment of the MDR of the Northwest Pacific is lower than that of the North Atlantic and the Northeast Pacific (figure 2). In the North Indian Ocean, the prediction skill of local environmental variables is low (e.g. supplementary figure 7), and the TC-environment relationship is weak, making skillful seasonal predictions challenging (Supplementary Materials).

Interestingly, the NeuralGCM hindcast and the observation show significant correlations in the ACE (section 2.3). For the experiments with deterministic physics, we identified statistically significant correlations for the North Atlantic ($r = 0.68$), the Northeast

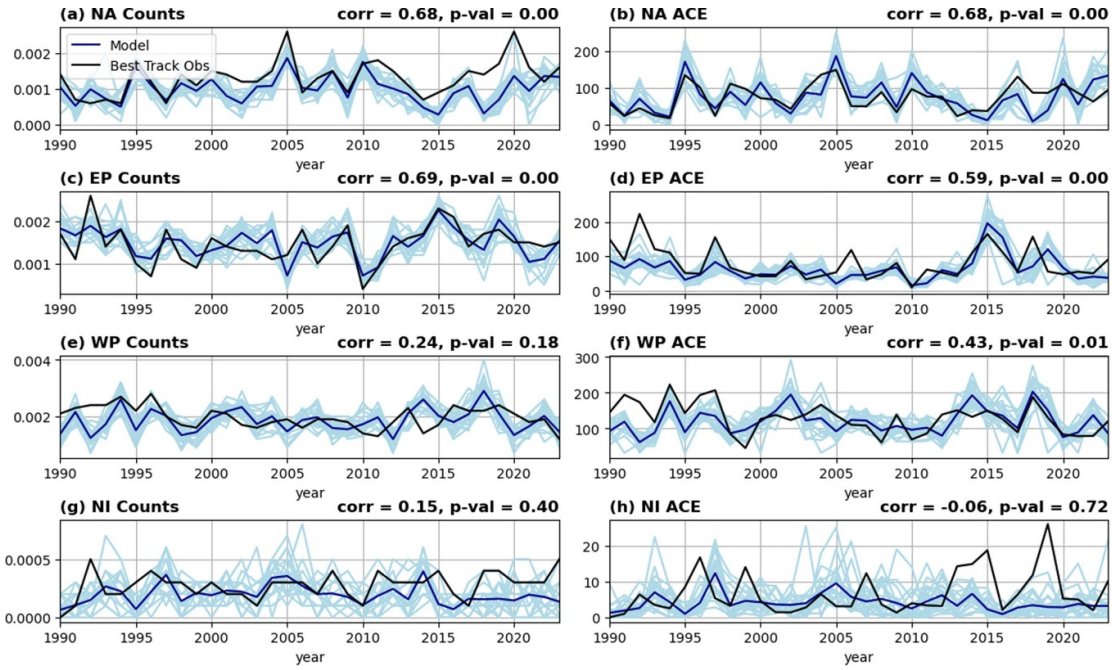

Figure 4. Predictions and observations of the basin-wide TC counts and the accumulated cyclone energy (ACE; unit: 10^4 knot^2) (1990–2023). (a) North Atlantic TC counts, (b) North Atlantic ACE. (c)–(d), (e)–(f), and (g)–(h) Same as (a)–(b), but for Northwest Pacific, Northeast Pacific, and North Indian Ocean, respectively. The black line shows the annual hurricane counts in the observation (July–November). The blue lines show the counts in the NeuralGCM hindcasts. The dark blue line shows the ensemble mean, and the light blue line shows the individual ensemble members. The correlations between the ensemble means and the observation, as well as the associated statistical significance, are denoted in the top right corner of each subplot.

Table 1. Correlation of basin-wide TC frequency between seasonal predictions and the best track observations. The NeuralGCM results contain two rows that represent the hindcasts with the deterministic version (top) and the hindcasts with the stochastic version (bottom), respectively. The value ranges of NeuralGCM indicate 95%-confidence level intervals estimated using resampling with replacement. Except for the smaller number of resampling runs ($N = 1000$), the other settings of the skill estimation are similar to those in Zhang *et al* (2019). The ensemble size and evaluation period are consistent between studies so the comparisons are relatively fair.

Reference and evaluation configuration	Model	N Atlantic	NE Pacific	NW Pacific
Chen and Lin (2013) 5-member ensemble 1990–2010	HiRAM	0.88	0.61	0.34
	NeuralGCM	0.74–0.87 0.69–0.89	0.54–0.73 0.37–0.71	0.19–0.47 0.05–0.37
Zhang *et al* (2019) 12-member ensemble 1981–2014	FLOR	0.60–0.75	0.47–0.60	0.27–0.44
	NeuralGCM	0.67–0.76 0.63–0.78	0.53–0.65 0.41–0.60	0.09–0.23 0.01–0.17

Pacific ($r = 0.59$), and the Northwest Pacific ($r = 0.43$) (figure 4). The hindcast with the stochastic physics shows comparable or better skill in predicting the ACE of the North Atlantic ($r = 0.69$), the Northeast Pacific ($r = 0.52$), and the Northwest Pacific ($r = 0.57$) (supplementary figure 12). We also compared the skill of these hindcast experiments and a physical model with a higher spatial resolution (Vecchi *et al* 2014, Zhang *et al* 2019). During the 1981–2014 period, the NeuralGCM hindcasts and the physical model have comparable correlation coefficients in predicting the ACE in the North Atlantic, but the skill of the NeuralGCM hindcasts is notably lower with the Northeast and Northwest Pacific (not shown). The skill difference in predicting the ACE is possibly attributable to regional model biases and difficulties of the NeuralGCM hindcasts in representing intense TCs (>60 m s^{-1}) (supplementary figure 13). Nonetheless, this intensity-related issue is expected considering the lower resolution of the NeuralGCM (1.4° vs ∼0.5°) used to generate our hindcasts.

4. Summary and discussion

This study conducts experimental seasonal predictions with the newly available NeuralGCM and simplified boundary forcings. Inspired by earlier studies with physical GCMs, the hindcast experiments focus on July to November which account for most TC activity in the Northern Hemisphere. The

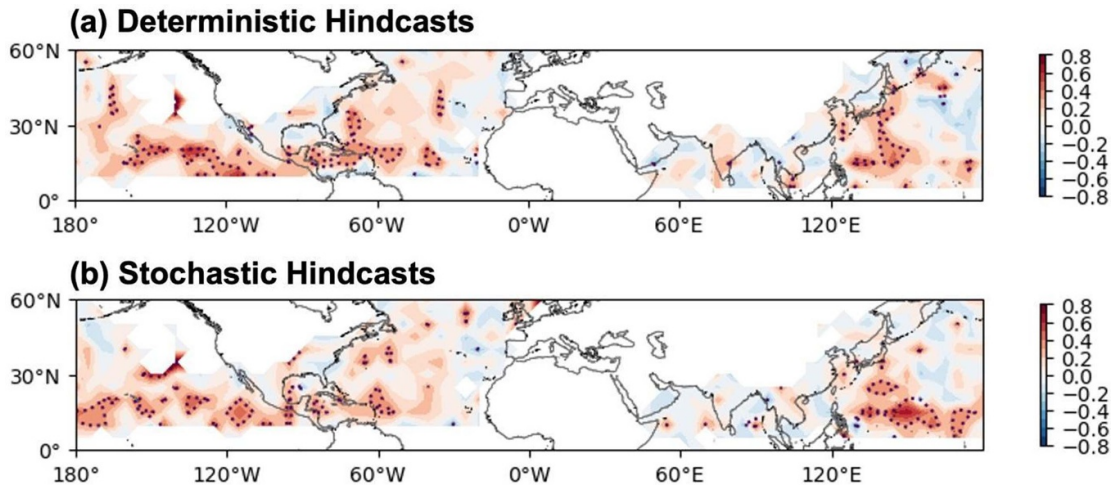

Figure 5. The correlations between the cell counts of TC tracks in the NeuralGCM hindcast and the best track observation. (a) Hindcasts with deterministic physics. (b) Hindcasts with stochastic physics. The annual counts (July–November) are calculated based on a 5° grid. The dashed contours highlight the correlation coefficients at the 90% confidence level.

NeuralGCM hindcasts of 1990–2023 can simulate realistic climate states of the atmosphere environment and TC activity. When predicting atmospheric variability, the NeuralGCM hindcast shows statistically significant anomaly correlation coefficients with the ERA5 reference at various forecast leads. Consistent with experiments conducted with physical GCMs, the skills are the highest for the environmental variables in the tropics. The hindcast also achieves relatively high skills in predicting seasonal metrics of TC activity, notably in the North Atlantic and the Northeast Pacific ($r = \sim 0.7$). For instance, the prediction skill of TC activity metrics such as basin-wide TC frequency is comparable to physical models with much higher spatial resolution (e.g. Chen and Lin 2013) or more complex coupled processes (e.g. Zhang *et al* 2019) (table 1). Despite challenges associated with intense TCs and some aspects of regional activity (e.g., the North Indian Ocean), the model-predicted interannual variations show significant correlations with the observation, including the sub-basin TC tracks ($p < 0.1$) (figure 5) and basin-wide ACE ($p < 0.01$) of the North Atlantic and North Pacific basins (figure 4). The skill with this physics-ML model is encouraging considering the simplified nature of boundary forcings and the low computational costs (supplementary table 1).

This study has several caveats related to the comparison with operational prediction models and the simplified boundary forcings. Since the TC data of most operational climate prediction models are not publicly accessible, we were unable to comprehensively compare our results with state-of-the-art models and evaluate potential differences in prediction skills and computation costs (supplementary materials). The simplified boundary forcings used in this study rely on the persistence of anomalies and can be less reasonable for other initialization time. We speculate that more realistic representations of boundary forcings or the inclusion of coupled climate processes (e.g. land-atmosphere coupling) may help NeuralGCM to accomplish more skillful predictions of TC activity (e.g. Zhang *et al* 2021) and other aspects of the Earth system (e.g. Yeager *et al* 2022). Such development can be accomplished by coupling NeuralGCM with statistical models, ML emulators, or other hybrid models of the ocean and other Earth system components.

Contributing to the rapidly evolving field of the ML-based climate modeling, this study demonstrates a practical application of the NeuralGCM and provides valuable insights for future model development. Our hindcast experiments with simplified boundary forcings shows that the NeuralGCM can represent the atmospheric responses to boundary forcings (e.g. SST) that are critical for the subseasonal-to-seasonal prediction. These experiments establish a performance baseline against which future model iterations can be benchmarked. Furthermore, our results suggest that the NeuralGCM holds significant potential as a foundation for developing a computationally affordable system for seamless subseasonal-to-seasonal prediction. Nevertheless, the evaluation also underscores some challenges of applying the current ML atmospheric models, including limitations inherited from training datasets and the lack of coupling among key climate system components. Recognizing that physical GCMs required decades of refinement to achieve milestones like simulating realistic TC activity (Manabe *et al* 1970, Zhao *et al* 2009, Roberts *et al* 2020), patience and continued effort are warranted despite recent breakthroughs in ML modeling efforts. We expect intensified collaboration among ML and physical science communities to

alleviate many of the identified issues, ultimately accelerating the transformation of climate model development and applications.

Data availability statement

The ERA5 dataset (Hersbach *et al* 2020) is accessible via the Climate Data Store of the Copernicus Climate Change Service 2019 (https://doi.org/10.24381/cds.6860a573). The IBTrACS dataset (Knapp *et al* 2010) is archived by the U.S. National Centers for Environmental Information (www.ncei.noaa.gov/products/international-best-track-archive). The NeuralGCM code (Kochkov *et al* 2024) is available at GitHub (https://github.com/google-research/neuralgcm) under the Apache 2.0 license. The code used to generate the plots in this study is available the Zenodo repository (https://doi.org/10.5281/zenodo.15319941).

Acknowledgments

G Z is supported by the faculty development fund of the University of Illinois at Urbana-Champaign, and the faculty fellowship of the Office of Risk Management and Insurance Research (ORMIR) of Gies Business School, and the U.S. National Science Foundation Awards AGS-2327959 and RISE-2530555. The authors thank Sarah Henry and Drs. Zhuo Wang, Stephan Hoyer, and Dmitrii Kochkov for stimulating discussions about evaluating and refining the NeuralGCM simulations. GZ thanks the organizers of the Rossbypalooza workshop at the University of Chicago for providing a welcoming environment for attendees.

Conflict of interest

J Y is an employee of Google. J Y have filed international patent application PCT/US2023/035420 in the name of Google LLC, currently pending, relating to the NeuralGCM (Kochkov *et al* 2024).

ORCID iD

Gan Zhang 0000-0002-7323-3409

References

Beucler T *et al* 2024 Climate-invariant machine learning *Sci. Adv.* **10** eadj7250

Bi K, Xie L, Zhang H, Chen X, Gu X and Tian Q 2023 Accurate medium-range global weather forecasting with 3D neural networks *Nature* **619** 533–8

Bonev B, Kurth T, Hundt C, Pathak J, Baust M, Kashinath K and Anandkumar A 2023 Spherical Fourier neural operators: learning stable dynamics on the sphere (arXiv: 2306.03838v1)

Bretherton C S, Henn B, Kwa A, Brenowitz N D, Watt-Meyer O, McGibbon J, Perkins W A, Clark S K and Harris L 2022 Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations *J. Adv. Model Earth Syst.* **14** e2021MS002794

Brunet G *et al* 2010 Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction *Bull. Am. Meteorol. Soc.* **91** 1397–406

Bushuk M *et al* 2022 Mechanisms of regional arctic sea ice predictability in two dynamical seasonal forecast systems *J. Clim.* **35** 4207–31

Charney J G 1963 A note on large-scale motions in the tropics *J. Atmos. Sci.* **20** 607–9

Chen J-H and Lin S-J 2013 Seasonal predictions of tropical cyclones using a 25-km-resolution general circulation model *J. Clim.* **26** 380–98

Copernicus Climate Change Service 2019 ERA5 monthly averaged data on pressure levels from 1940 to present (available at: https://cds.climate.copernicus.eu/doi/10.24381/cds.6860a573)

Cresswell-Clay N, Liu B, Durran D, Liu A, Espinosa Z I, Moreno R and Karlbauer M 2024 A deep learning earth system model for stable and efficient simulation of the current climate (arXiv: 2409.16247v2)

Delworth T L *et al* 2020 SPEAR: the next generation GFDL modeling system for seasonal to multidecadal prediction and projection *J. Adv. Model. Earth Syst.* **12** e2019MS001895

Doi T, Vecchi G A, Rosati A J and Delworth T L 2013 Response to CO_2 doubling of the Atlantic hurricane main development region in a high-resolution climate model *J. Clim.* **26** 4322–34

Dulac W, Cattiaux J, Chauvin F, Bourdin S and Fromang S 2024 Assessing the representation of tropical cyclones in ERA5 with the CNRM tracker *Clim. Dyn.* **62** 223–38

Emanuel K, Fondriest F and Kossin J 2012 Potential economic value of seasonal hurricane forecasts *Weather Clim. Soc.* **4** 110–7

Eyring V *et al* 2024 Pushing the frontiers in climate modelling and analysis with machine learning *Nat. Clim. Chang* **14** 916–28

Feng X, Klingaman N P and Hodges K I 2021 Poleward migration of western North Pacific tropical cyclones related to changes in cyclone seasonality *Nat. Commun.* **12** 6210

Goldenberg S B, Landsea C W, Mestas-Nuñez A M and Gray W M 2001 The recent increase in Atlantic hurricane activity: causes and implications *Science* **293** 474–9

Hersbach H *et al* 2020 The ERA5 global reanalysis *Q. J. R. Meteorol. Soc.* **146** 1999–2049

Hoskins B 2013 The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science *Q. J. R. Meteorol. Soc.* **139** 573–84

Jien J Y, Gough W A and Butler K 2015 The influence of El Niño–Southern oscillation on tropical cyclone activity in the Eastern North Pacific basin *J. Clim.* **28** 2459–74

Johnson S J *et al* 2019 SEAS5: the new ECMWF seasonal forecast system *Geosci. Model Dev.* **12** 1087–117

Keisler R 2022 Forecasting global weather with graph neural networks (arXiv: 2202.07575)

Knapp K R, Kruk M C, Levinson D H, Diamond H J and Neumann C J 2010 The international best track archive for climate stewardship (IBTrACS): unifying tropical cyclone data *Bull. Am. Meteorol. Soc.* **91** 363–76

Kochkov D *et al* 2024 Neural general circulation models for weather and climate *Nature* **632** 1060–6

Lam R *et al* 2023 Learning skillful medium-range global weather forecasting *Science* **382** 1416–21

Lang S *et al* 2024 AIFS—ECMWF's data-driven forecasting system (arXiv: 2406.01465v2)

Lemoine D and Kapnick S 2024 Financial markets value skillful forecasts of seasonal climate *Nat. Commun.* **15** 4059

M G W 1984 Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences *Mon. Weather Rev.* **112** 1649–68

Manabe S, Holloway J L and Stone H M 1970 Tropical circulation in a time-integration of a global model of the atmosphere *J. Atmos. Sci.* **27** 580–613

Pathak J *et al* 2022 FourCastNet: a global data-driven high-resolution weather model using adaptive Fourier neural operators (arXiv: 2202.11214v1)
Phillips N A 1956 The general circulation of the atmosphere: a numerical experiment *Q. J. R. Meteorol. Soc.* **82** 123–64
Price I, Sanchez-Gonzalez A, Alet F, Ewalds T, El-Kadi A, Stott J, Mohamed S, Battaglia P, Lam R and Willson M 2023 GenCast: diffusion-based ensemble forecasting for medium-range weather (arXiv: 2312.15796)
Roberts M J *et al* 2020 Projected future changes in tropical cyclones using the CMIP6 HighResMIP multimodel ensemble *Geophys. Res. Lett.* **47** e2020GL088662
Shukla J 1998 Predictability in the midst of chaos: a scientific basis for climate forecasting *Science* **282** 728–31
Sobel A H and Bretherton C S 2000 Modeling tropical precipitation in a single column *J. Clim.* **13** 4378–92
Ullrich P A, Zarzycki C M, McClenny E E, Pinheiro M C, Stansfield A M and Reed K A 2021 TempestExtremes v2.1: a community framework for feature detection, tracking, and analysis in large datasets *Geosci. Model Dev.* **14** 5023–48
Vecchi G A *et al* 2014 On the seasonal forecasting of regional tropical cyclone activity *J. Clim.* **27** 7994–8016
Vonich P T and Hakim G J 2024 Predictability limit of the 2021 Pacific Northwest heatwave from deep-learning sensitivity analysis *Geophys. Res. Lett.* **51** e2024GL110651
Wang C, Pritchard M S, Brenowitz N, Cohen Y, Bonev B, Kurth T, Durran D and Pathak J 2024 Coupled ocean-atmosphere dynamics in a machine learning earth system model (arXiv: 2406.08632v1)
Wang Z, Zhang G, Dunkerton T J and Jin F-F 2020 Summertime stationary waves integrate tropical and extratropical impacts on tropical cyclone activity *Proc. Natl Acad. Sci. USA* **117** 22720–6
Watt-Meyer O, Henn B, McGibbon J, Clark S K, Kwa A, Perkins W A, Wu E, Harris L and Bretherton C S 2024 ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses (arXiv: 2411.11268v1)
World Meteorological Organization 2021 *WMO Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (1970–2019) (WMO-No. 1267)* (WMO)
Yeager S G *et al* 2022 The seasonal-to-multiyear large ensemble (SMYLE) prediction system using the Community Earth System Model version 2 *Geosci. Model Dev.* **15** 6451–93
Yuval J, Langmore I, Kochkov D and Hoyer S 2024 Neural general circulation models optimized to predict satellite-based precipitation observations (arXiv: 2412.11973v1)
Zhang G, Murakami H, Gudgel R and Yang X 2019 Dynamical seasonal prediction of tropical cyclone activity: robust assessment of prediction skill and predictability *Geophys. Res. Lett.* **46** 5506–15
Zhang G, Murakami H, Yang X, Findell K L, Wittenberg A T and Jia L 2021 Dynamical seasonal predictions of tropical cyclone activity: roles of sea surface temperature errors and atmosphere–land initialization *J. Clim.* **34** 1743–66
Zhang G and Wang Z 2015 Interannual variability of tropical cyclone activity and regional Hadley circulation over the Northeastern Pacific *Geophys. Res. Lett.* **42** 2473–81
Zhang G, Wang Z, Dunkerton T J, Peng M S and Magnusdottir G 2016 Extratropical impacts on Atlantic tropical cyclone activity *J. Atmos. Sci.* **73** 1401–18
Zhao M, Held I M, Lin S-J and Vecchi G A 2009 Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50-km resolution GCM *J. Clim.* **22** 6653–78
Zhao M, Held I M and Vecchi G A 2010 Retrospective forecasts of the hurricane season using a global atmospheric model assuming persistence of SST anomalies *Mon. Weather Rev.* **138** 3858–68