Response to Anonymous Referee 2

Dear Referee,

We sincerely thank you for your time and effort in reviewing our manuscript: *SPIN (v1.0): A Spontaneous Synthetic Tropical Cyclone Model Empowered by NeuralGCM for Hazard Assessment*. We are grateful for your supportive assessment of our work and for the helpful comments you provided. We have carefully addressed each of your points and revised the manuscript where appropriate. Please find our detailed responses below.

General Comments:

1. The term "downscaling" is used throughout the introduction without explicit definition, which may create ambiguity for readers. In mainstream climatology, "downscaling" typically refers to the spatial refinement of gridded meteorological fields (e.g., temperature, precipitation, wind) from coarse to fine resolution, either through dynamical methods (regional climate models) or statistical approaches (bias correction, regression-based methods). In the tropical cyclone hazard community, however, "downscaling" — or more precisely "statistical-dynamical downscaling" — carries a different and more specific meaning: the generation of synthetic TC tracks and intensities from large-scale environmental fields, without necessarily producing spatially explicit high-resolution wind fields. These two usages are sufficiently different that conflating them risks misleading readers outside the TC hazard community. More broadly, it would be helpful for the authors to better acknowledge the diversity of approaches that exist for generating synthetic TC tracks.

Response:

Thank you very much for pointing out the potential ambiguity in the term "downscaling." We agree that the term is used differently in mainstream climatology and in the TC hazard literature. To avoid confusion, we have revised the Introduction to explicitly define our use of "downscaling" in the TC hazard context. We have also expanded the discussion of existing statistical downscaling approaches to better acknowledge the diversity of these methods.

Changes in the manuscript:

We have therefore revised the manuscript as follows (Lines 24–40):

> "… The high risks associated with TCs have motivated the development of TC hazard downscaling frameworks. Here, "downscaling" denotes approaches that generate event sets large enough for risk assessment, rather than the spatial refinement of gridded meteorological fields. Over recent decades, two major approaches have been used for TC hazard downscaling: statistical and dynamical downscaling. Statistical downscaling models efficiently generate large ensembles of synthetic TCs from large-scale environmental conditions without requiring spatially explicit high-resolution wind fields, enabling estimates of extreme-event likelihoods at low computational cost. Statistical approaches include resampling methods based purely on historical observations (Bloemendaal et al., 2020; Vickery et al., 2000). Although useful for present-day risk assessment, such models are limited in their ability to represent non-stationary climates because they do not explicitly account for changes in the large-scale environment. To address this limitation, environment-dependent synthetic TC models have been developed to allow TC activity to vary

with changes in climate conditions. These models differ in how they represent genesis, track and intensity evolution. Examples include statistical–deterministic models incorporating random-seeding frameworks (Emanuel et al., 2008; Lin et al., 2023) and purely statistical frameworks such as CHAZ (Lee et al., 2018) and PepC (Jing and Lin, 2020).

These models have several constraints when applied to compound TC hazard assessment. First, because they are driven by monthly-mean environmental fields, they are not designed to resolve intra-month variability relevant to TC genesis. Second, synthetic TCs are treated as independent events. Third, they do not provide time-evolving ambient environmental fields associated with each storm. These limitations make it difficult for traditional statistical downscaling models to move beyond statistical co-occurrence and represent compound TC hazards as dynamically organized events ...”

2. The description of NeuralGCM's performance is not fully demonstrated with respect to TC representation (as it is the case for ECMWF models for instance). Several important questions are left unaddressed that are directly relevant to the reliability of the SPIN framework: How long is the model rollout stable? Are there any known biases? Are all simulated synthetic tracks physically plausible, or can NeuralGCM produce spurious vortex-like features (hallucinations) that may be incorrectly identified as TCs by TempestExtremes? What do extreme TCs look like in NeuralGCM outputs (10m wind speed, sea level pressure structure)? What initial conditions are used, and how are observed SSTs incorporated? These aspects are critical to assess the physical realism of the TC catalog.

Response:

Thank you very much for raising these important questions or concerns.

1) How long is the model rollout stable?

We agree that the numerical stability of the NeuralGCM rollouts was not sufficiently assessed in the original manuscript. We have now added a dedicated stability assessment and updated the processing workflow to handle unstable simulations before TC detection.

Following the recommended NeuralGCM checkpoint modification (https://neuralgcm.readthedocs.io/en/latest/checkpoint_modifications.html), we fixed the global-mean log surface pressure to reduce long-term model drift and enhance numerical stability. Although this adjustment substantially improved the model drift of the rollouts, numerical instability still occurs in approximately 1.75% of simulated member months, defined by non-meteorological outliers followed by missing values across the domain at later output steps. Instability begins after 10–12 months of integration in most cases, accounting for 71.9% of unstable member-years. No instability cases begin within the first five months.

To maintain a consistent ensemble size across target years, we generate three additional backup members using the same mid-October initialization protocol but independent initialization times. For each target year, unstable original members are removed before TC detection and replaced by full-year simulations from backup members for the same target year. Backup members are considered in a fixed order. For each target year, the first backup member-year that is both stable and not previously assigned within that target year is used for replacement. Backup member-years that have already been used within the same target year, or are themselves unstable, are skipped.

We then reran the TC detection and regenerated all figures and statistics using this updated protocol. The numerical values change marginally (e.g. the correlations of interannual variability for TCs and MTCEs improved in key basins), and the conclusions remain unchanged.

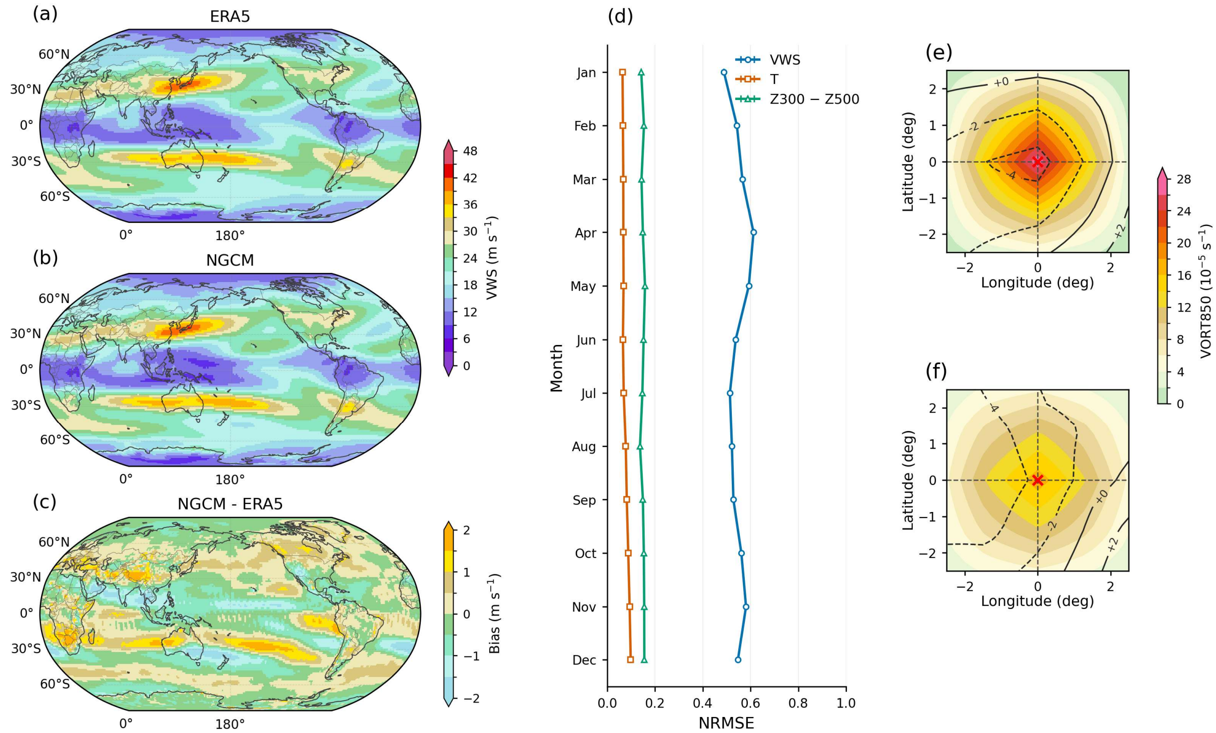
We have therefore included these statements in Lines 124–133.

2) Are there any known biases?

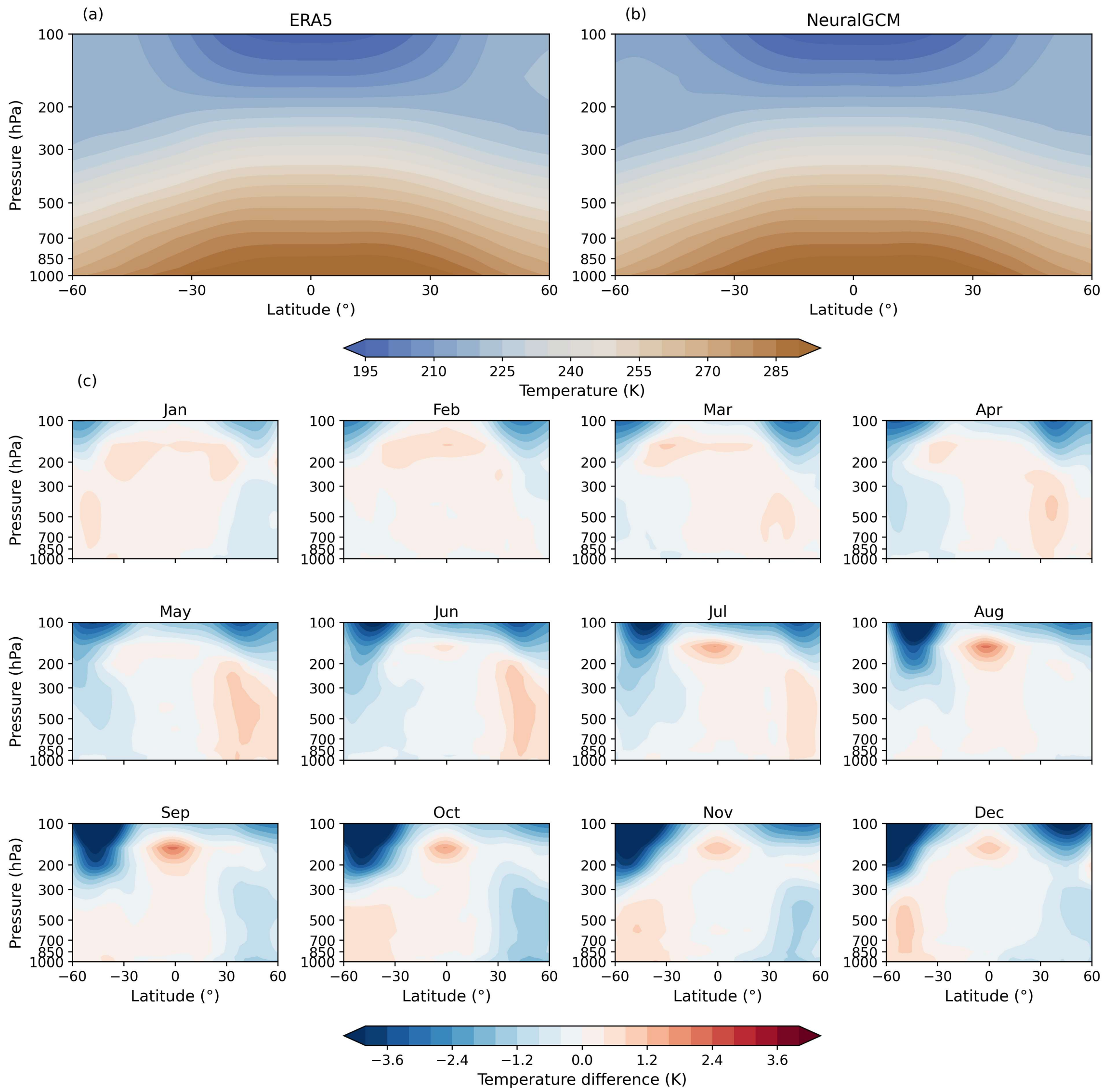
NeuralGCM captures the main climatological structures of the large-scale environments relevant to tropical cyclone activity. The ensemble mean reproduces the observed VWS pattern, including the subtropical maxima evident in ERA5, although regional biases remain (Response Fig. R1a–c). It also captures the broad vertical temperature structure, with warmer conditions in the lower troposphere and colder conditions aloft (Response Fig. R2), and reproduces the expected meridional contrast in Z300-Z500, an upper-level warm-core proxy used in TempestExtremes, with larger values in the tropics and smaller values toward higher latitudes (Response Fig. R3).

Upper-tropospheric temperature biases increase with rollout time (Response Fig. R2). Biases in Z300-Z500 and VWS vary seasonally, with Z300-Z500 tending to be higher than ERA5 during the early-to-peak TC season and lower during the late season, whereas VWS tends to be lower than ERA5 during the early-to-peak season and higher during the late season (Response Fig. R3 and R4). Monthly normalized root-mean-square errors (NRMSEs), normalized by the spatial standard deviation of the ERA5 reference fields, also generally increase with rollout time, indicating gradual bias growth during the simulations (Response Fig. R1d). Despite these biases, the broad agreement with ERA5 supports the use of NeuralGCM to simulate large-scale environmental fields relevant to TC hazards.

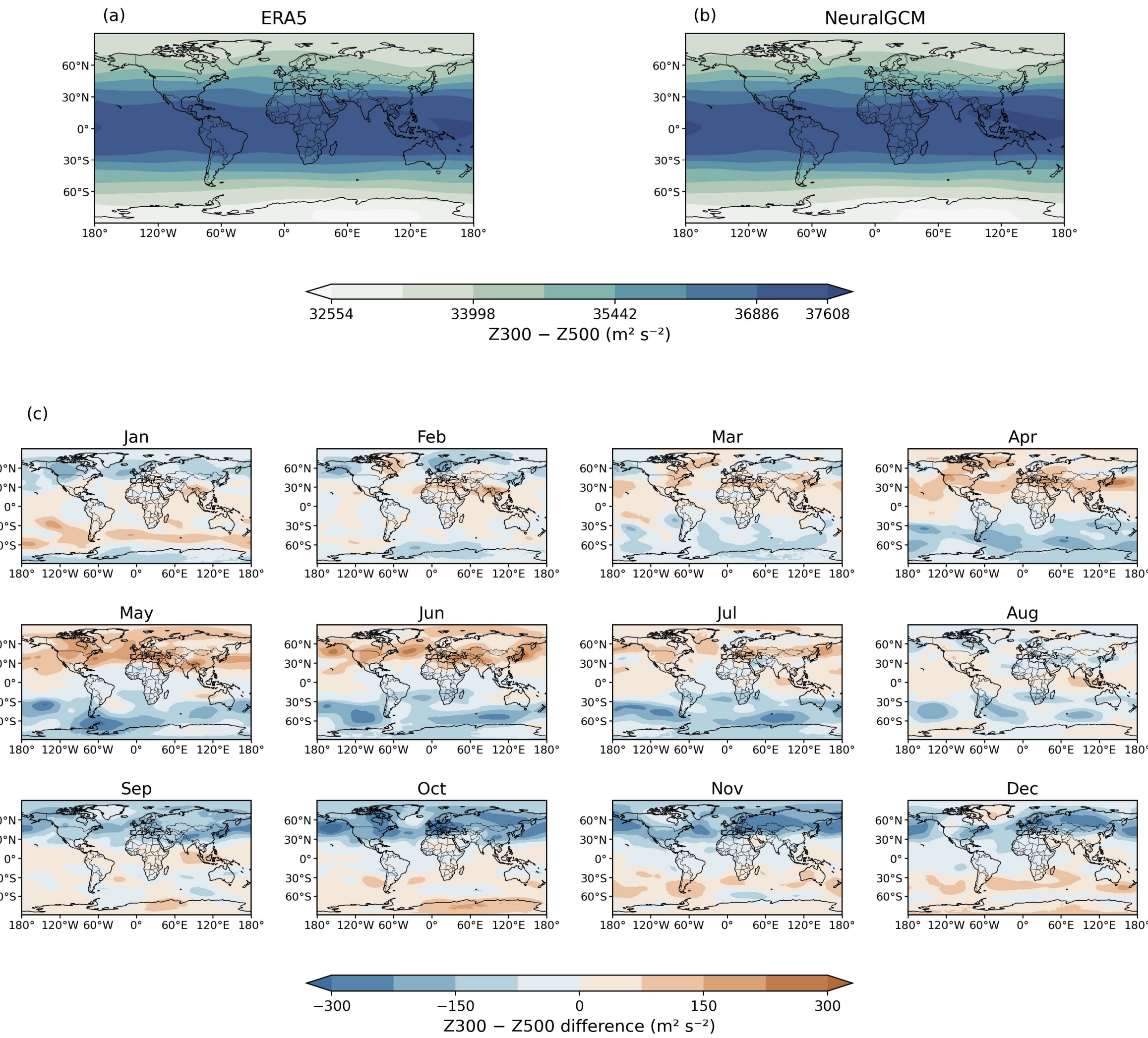
The corresponding text has been added in a new subsection, Section 3.1, “Large-scale atmospheric environment”.



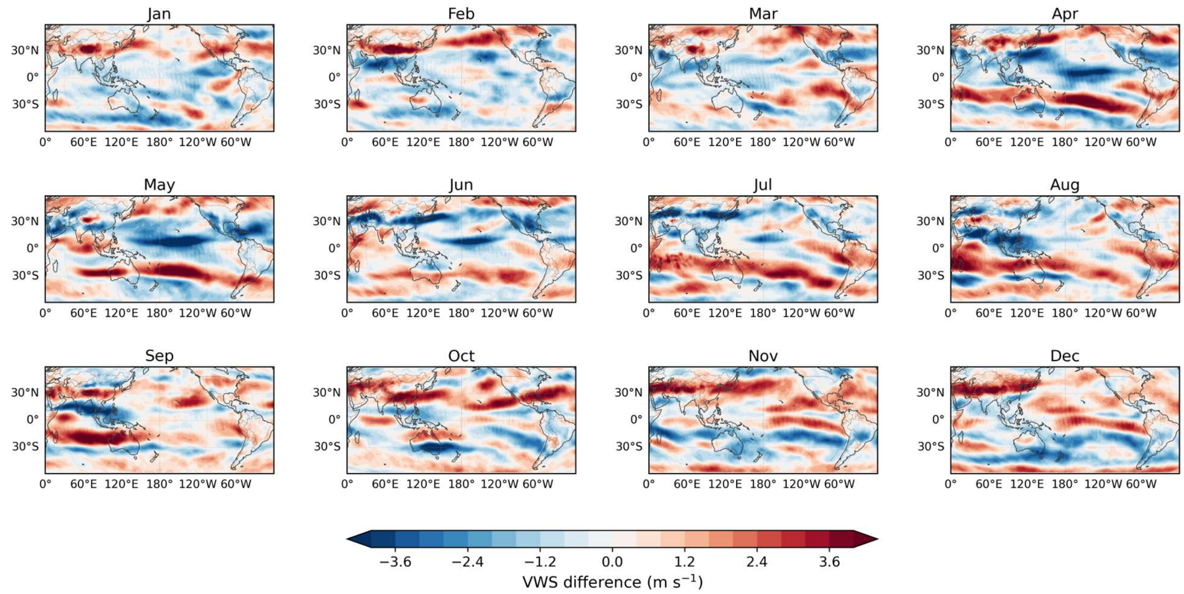
Response Fig. R1. Evaluation of simulated atmospheric environmental fields. Panels show the 1980–2022 mean 200–850-hPa vertical wind shear (VWS; $m\ s^{-1}$) from (a) ERA5, (b) the NeuralGCM ensemble mean, and (c) their difference. Panel (d) shows monthly normalized root-mean-square errors (NRMSEs) between NeuralGCM and ERA5 for VWS, the meridional temperature profile (T; K), and the 300–500-hPa geopotential height difference (Z300-Z500; $m^2 s^{-2}$), averaged over 1980–2022, with error bars indicating the 10th–90th percentile range across ensemble members. Panels (e,f) show storm-centred composites of the surrounding environment for FAST-simulated extreme TCs in the North Atlantic basin during 1980–2022. Composites are shown for the (e) top 10% and (f) lowest 10% of storms ranked by lifetime maximum intensity, evaluated at the time of lifetime maximum intensity. Shading shows 850-hPa relative vorticity ($10^{-5} s^{-1}$), and contours show surface pressure anomalies (hpa) relative to the mean pressure within a storm-centred 500–800-km annulus.



Response Fig. R2. Meridional temperature profiles in ERA5 and NeuralGCM. (a,b) 1980–2022 mean pressure–latitude temperature profiles from (a) ERA5 and (b) the NeuralGCM ensemble mean. (c) Monthly NeuralGCM minus ERA5 temperature differences.



Response Fig. R3. Geopotential difference between 300 hPa and 500 hPa in ERA5 and NeuralGCM. (a,b) The 1980–2022 mean Z300-Z500 from (a) ERA5 and (b) the NeuralGCM ensemble mean. (c) Monthly NeuralGCM minus ERA5 differences in Z300-Z500.



Response Fig. R4. Monthly mean NeuralGCM minus ERA5 differences in 200–850-hPa vertical wind shear (VWS) over 1980–2022.

3) Are all simulated synthetic tracks physically plausible, or can NeuralGCM produce spurious vortex-like features (hallucinations) that may be incorrectly identified as TCs by TempestExtremes?

TempestExtremes (TE) has been applied not only to NeuralGCM output (Kochkov et al., 2024; Zhang et al., 2025), but also to high-resolution climate-model simulations (Moon et al., 2026; Roberts et al., 2020; Tang et al., 2022). Nevertheless, we do not assume that every vortex-like feature identified from raw model output is necessarily a physically plausible TC. In this study, TE is used as an objective candidate-track detection tool, and spurious vortex-like features may in principle be detected if they satisfy the tracking criteria.

Following this comment, we further examined the detected tracks. We found that approximately 0.32% of all detected tracks across ensemble members and years are located entirely over land, and are therefore not physically plausible TCs. We have updated the post-processing procedure to remove these tracks. We now report this diagnostic in Line 140. The relevant figures and statistics throughout the manuscript have been updated accordingly. The conclusions remain unchanged.

We further assessed the physical consistency of the retained tracks using environmental fields extracted along the TC tracks. For a random subset of detected events, the associated 500-hPa geopotential height and surface pressure fields show coherent synoptic-scale circulation patterns with TC-like disturbances. In addition, as shown in our TC-structure analysis (Response Fig. R1e,f), FAST-simulated low-intensity events are associated with weaker 850-hPa vorticity and surface pressure anomalies in NeuralGCM, indicating a consistent relationship between the simulated intensity and the underlying model vortex structure.

Although we cannot manually verify every individual track, these additional checks suggest that the retained TE tracks are broadly physically consistent.

4) What do extreme TCs look like in NeuralGCM outputs (10m wind speed, sea level pressure structure)?

The NeuralGCM outputs do not include 10-m wind speed or sea-level pressure, so we cannot directly show the near-surface wind and sea-level pressure structure of the simulated TCs. Instead, we use the available 850-hPa relative vorticity and surface pressure fields to assess whether the FAST-derived intensity ranking is dynamically consistent with the corresponding NeuralGCM atmospheric fields. Using the NA basin as an example, the top 10% of storms ranked by lifetime maximum intensity show stronger 850-hPa relative vorticity and surface pressure anomalies than the lowest 10% (Response Fig. R1 e,f). This correspondence indicates that the FAST-derived intensity ranking is qualitatively reflected in the NeuralGCM atmospheric fields.

The corresponding text has been added in Lines 258–261.

5) What initial conditions are used, and how are observed SSTs incorporated?

All initial and lower-boundary fields are derived from ERA5. The initial atmospheric states include wind, geopotential, temperature, specific humidity, and specific cloud ice and liquid water content on pressure levels. Monthly mean SST and SIC fields are linearly interpolated to 12-hourly resolution and prescribed as lower-boundary conditions throughout the simulations.

We have revised the relevant text in Lines 120–123 to describe these details more precisely.

3. Not sure to fully understand why SPIN allows "two-way interactions" between synthetic TC and their environment ? FAST computes TC intensity independently along each NeuralGCM-derived track without any feedback into the atmospheric fields. Also, SPIN's improved MTCE representation is attributed to these interactions which likely need more evidences / justifications.

Response:

Thank you very much for pointing this out. We agree with the reviewer that the original statement overstated the degree of two-way coupling between TCs and the environment in SPIN. The statement that SPIN enables two-way TC–environment interaction has therefore been removed. The revised manuscript now clarifies that SPIN's key advantage lies in diagnosing synthetic storms from evolving environmental fields, thereby representing intra-month variability relevant to TC genesis and providing synthetic storms together with their ambient environments. Further evidence on SPIN's improved representation of MTCEs are provided in our responses to the specific comments on Sections 3.5 and 3.6.

Specific Comments

Section 2.2.1

Could the authors elaborate on how the NeuralGCM ensemble simulations were conducted? Why were 14 members chosen specifically? The statement that "the 1.4° version produces reliable TC activity" is vague for a claim that underpins the entire study. Did the authors perform substantial checks to verify that NeuralGCM reliably simulates TCs across the full 1980-2022 period? What is the spread (in tracks representation) between ensemble members for an identical simulation year?

Response:

Thank you very much for raising these important questions.

1) Could the authors elaborate on how the NeuralGCM ensemble simulations were conducted?

For each target year from 1980 to 2022, we initialize a 14-member NeuralGCM ensemble simulation. The ensemble is generated using lagged initial conditions, with members initialized at 6-hour intervals in mid-October of the preceding year and integrated continuously through December of the target year. For example, the simulations used for the 1980 analysis are initialized in mid-October 1979 and integrated through December 1980. The preceding October–December period is treated as spin-up, and the January–December period of the target year is used for TC tracking and subsequent analysis. All initial and lower-boundary fields are derived from ERA5. The initial atmospheric states include wind, geopotential, temperature, specific humidity, and specific cloud ice and liquid water content on pressure levels. Monthly mean SST and SIC fields are linearly interpolated to 12-hourly resolution and prescribed as lower-boundary conditions throughout the simulations.

We have revised the manuscript accordingly (Lines 115–123). In addition, we have included a schematic illustrating the SPIN workflow as Figure 1 in the revised manuscript.

2) Why were 14 members chosen specifically?

The 14-member ensemble is chosen as a balance between statistical robustness and computational cost. Sensitivity tests indicate that modestly increasing or decreasing the ensemble size does not affect the conclusions.

3) The statement that "the 1.4° version produces reliable TC activity" is vague for a claim that underpins the entire study

We agree that the original phrase “producing reliable TC activity” was vague. We have revised it to a more specific statement based on previous studies, noting that the 1.4° NeuralGCM version is computationally efficient and has been shown to reasonably reproduce TC tracks and storm counts (Kochkov et al., 2024; Zhang et al., 2025).

Changes in the manuscript (Lines 113–115):

> “We use the pre-trained 1.4° deterministic version of NeuralGCM to simulate the environmental fields from which TC tracks are extracted. The 1.4° version is about ten times faster in inference than the 0.7° version and has been shown in previous studies to reasonably reproduce trajectories and counts of TCs (Kochkov et al., 2024; Zhang et al., 2025).”

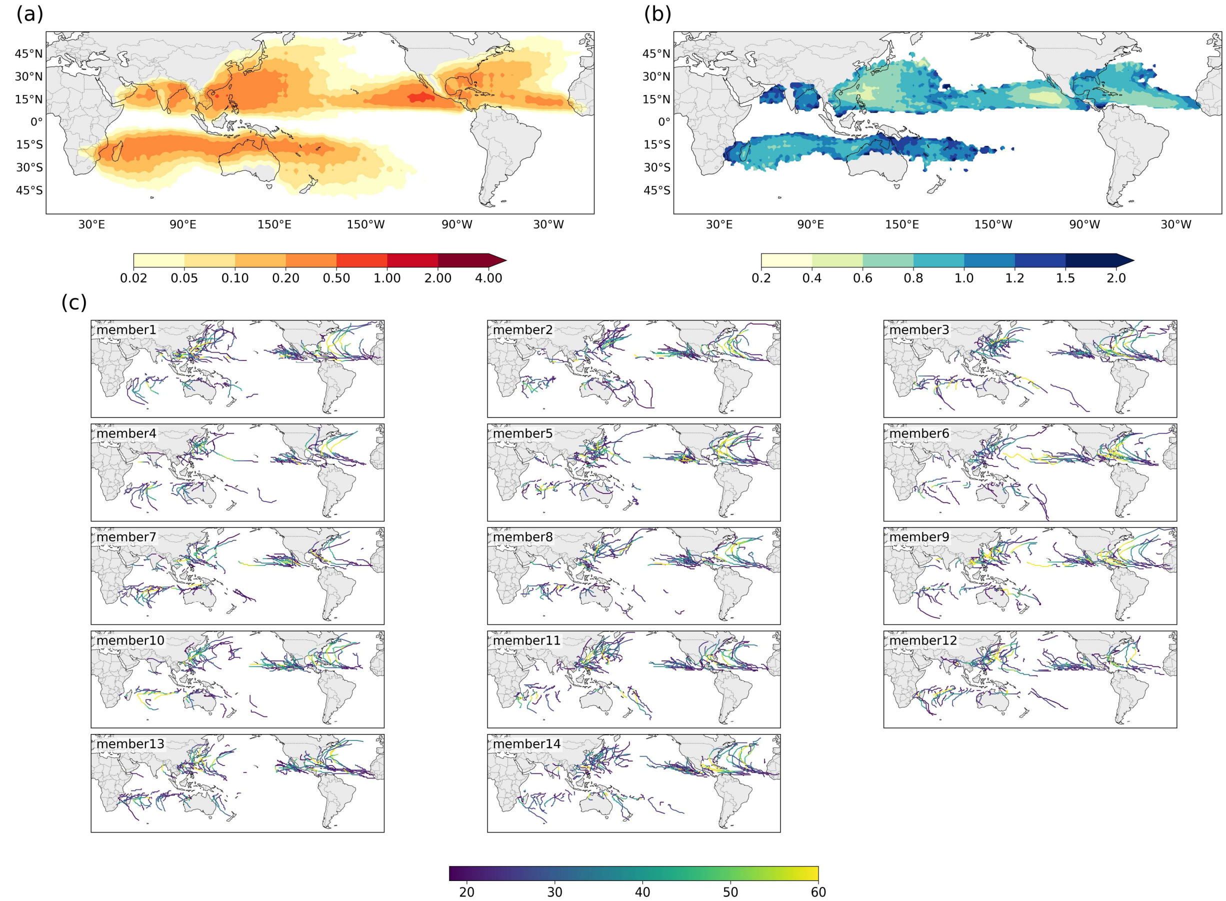
4) Did the authors perform substantial checks to verify that NeuralGCM reliably simulates TCs across the full 1980-2022 period?

Yes. We performed multiple diagnostics to evaluate NeuralGCM-based TC simulations over 1980–2022. As discussed in our response to General Comment 2, subquestion 2 and 4, NeuralGCM reasonably reproduces large-scale TC-relevant atmospheric environments, although some biases remain. The FAST-derived intensity ranking is also qualitatively reflected in the NeuralGCM fields, with stronger events associated with stronger 850-hPa relative vorticity and surface pressure anomalies. The original manuscript evaluates TC frequency, interannual variability, seasonal cycle, genesis distribution, and track patterns. The simulated interannual variability shows correlations with IBTrACS that are comparable to or higher than those from existing statistical hazard models (Figures 3 and 4), and the seasonal cycle, genesis distribution, and tracks are reasonably reproduced (Figures 5, 6, and 8). These diagnostics support the use of NeuralGCM-based SPIN simulations for representing TC activity over the full study period.

5) What is the spread (in tracks representation) between ensemble members for an identical simulation year?

Here, we show the 43-year mean standard deviation of inter-member TC track density in panel a, the coefficient of variation (CV) in panel b, and the TC tracks in 2020 for all 14 ensemble members in panel c. The CV is defined as the inter-member standard deviation divided by the inter-member mean track density.

The standard deviation is generally higher in regions with larger track density, especially in the WP, EP, NA, SI, and AU basins, indicating that the largest absolute inter-member spread occurs in the major TC-active regions. In contrast, the CV is higher mainly along regions characterized by sparse TC occurrence, suggesting larger relative uncertainty there. Panel c shows that the number and intensity of TC tracks vary across ensemble members, while the broad spatial patterns remain consistent with regions climatologically favourable for TC genesis.



Response Fig. R5. Inter-member spread in simulated TC track density. (a) Standard deviation and (b) coefficient of variation of annual TC track density are first calculated across the 14 ensemble members for each year and then averaged over 1980–2022. (c) TC tracks from all 14 ensemble members in 2020, with track colour indicating TC intensity ($m\ s^{-1}$).

Section 2.2.2

The authors state that TempestExtremes parameters are fine-tuned separately for each basin to align simulated annual average track counts and storm lifetimes with IBTrACS. Could the authors elaborate on this procedure? Please also specify the total number of simulated years and clarify the link with Figure 14.

Response:

Thank you very much for the questions. We first used the TempestExtremes parameters from Kochkov et al. (2024) as a baseline. This setting gives a reasonable global TC count, but when applied uniformly to all basins, it underestimates TC activity in some basins and overestimates it in others. We therefore ran the tracking separately for each basin and used basin specific detection thresholds. The thresholds were selected to make the historical period mean TC counts and storm lifetimes in each basin broadly consistent with IBTrACS. To avoid overfitting, this tuning was done once for each basin over the full historical period, not separately for individual years or ensemble members. The specific parameter settings for each basin are provided in our open-source code.

The NeuralGCM simulations consist of 14 ensemble members, each covering 1980–2022, corresponding to 14 × 43 = 602 simulated member years. Because FAST uses a random draw to initialize the maximum azimuthal wind speed, its intensity simulation is not deterministic for a given track and environmental forcing. We therefore perform ten FAST intensity simulations for each detected NeuralGCM track, using the same track and environmental forcing but different initial intensity states. This yields 14 × 43 × 10 = 6,020 intensity-realization years, as noted in Lines 159–162 of the revised manuscript. In other words, NeuralGCM determines the TC genesis, tracks, and environmental conditions, while the repeated FAST simulations sample intensity uncertainty conditional on those tracks and environments. These 6020 intensity realization years are used for the return period analysis shown in the original Figure 14, now Figure 17 in the revised manuscript.

Section 3.1

The authors note that interannual correlations could be improved by initializing NeuralGCM in April rather than mid-October, yet this configuration was not adopted. The rationale for choosing October initialization should be explicitly justified.

Response:

Thank you very much for the questions. We use January–December simulations here to maintain comparability with previous hazard-model evaluations. For application-oriented experiments, we suggest initializing closer to the target TC season, which can shorten the forecast lead time and reduce the accumulation of model drift.

The approximately 2.5-month spin-up period follows the experimental setup of Kochkov et al. (2024) and Cheng et al. (2022). This choice is appropriate for our full January–December evaluation, as initializing later would reduce the spin-up before January and could affect early-year TC activity, particularly in Southern Hemisphere basins where the TC season is active during January–March.

Changes in the manuscript:

We have therefore revised the manuscript as follows to make this clearer (Lines 276–280):

> "We use January–December simulations here to maintain comparability with previous hazard-model evaluations (Lee et al., 2018; Lin et al., 2023). For application-oriented experiments, however, initializing closer to the target TC season can shorten the forecast lead time and reduces the accumulation of model drift. Consistent with this interpretation, April initialization simulations followed by evaluation over the Northern Hemisphere hurricane season yields higher basin-scale TC-frequency correlations (Figure 4)."

Section 3.1

The proposed explanation for SPIN's improved interannual skill in mixed-ENSO-signal regions is physically plausible but remains qualitative and unverified. The authors should support this claim more explicitly, for instance by showing the spatial distribution of ENSO influence on TC genesis, and by demonstrating that SPIN's improvement over JL23 is indeed concentrated in regions with mixed-sign ENSO responses.

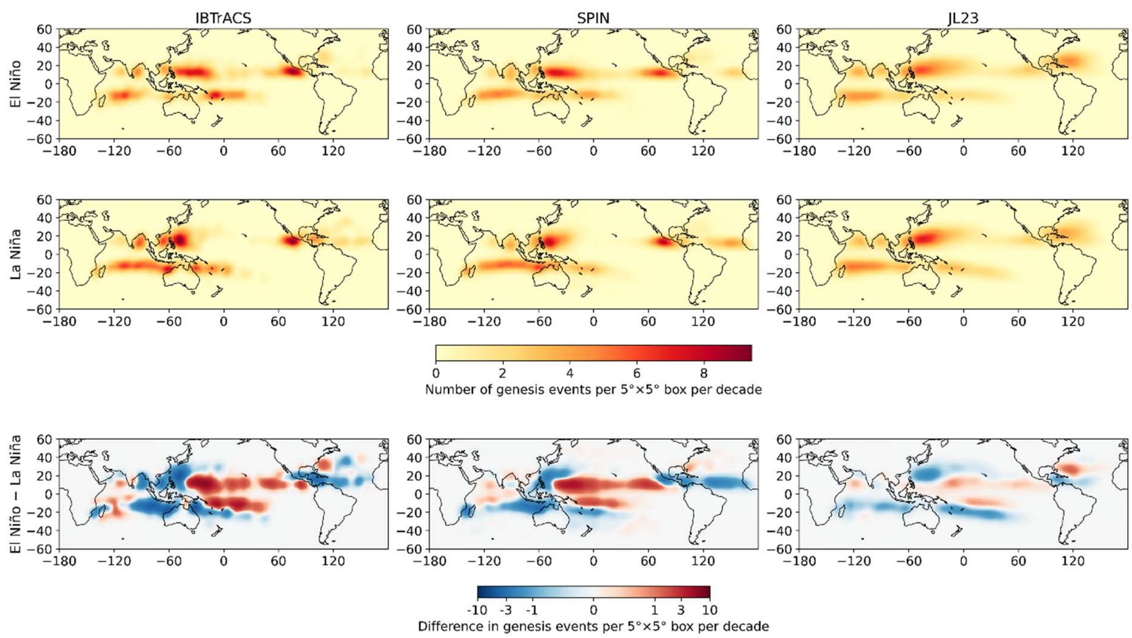
Response:

Thank you very much for raising this question. According to JL23's own manuscript, they denoted that their model's difficulty in simulating annual TC frequency in these basins may be due to:

> "El Niño-Southern Oscillation, the main source of tropical interannual variability, has single-signed signals in TC genesis in the Atlantic and Eastern Pacific basins, but mixed-signed signals in the Western Pacific, South Pacific, and South Indian basins. It is likely that the interannual signal in genesis gets averaged out in basins such as the Western Pacific."

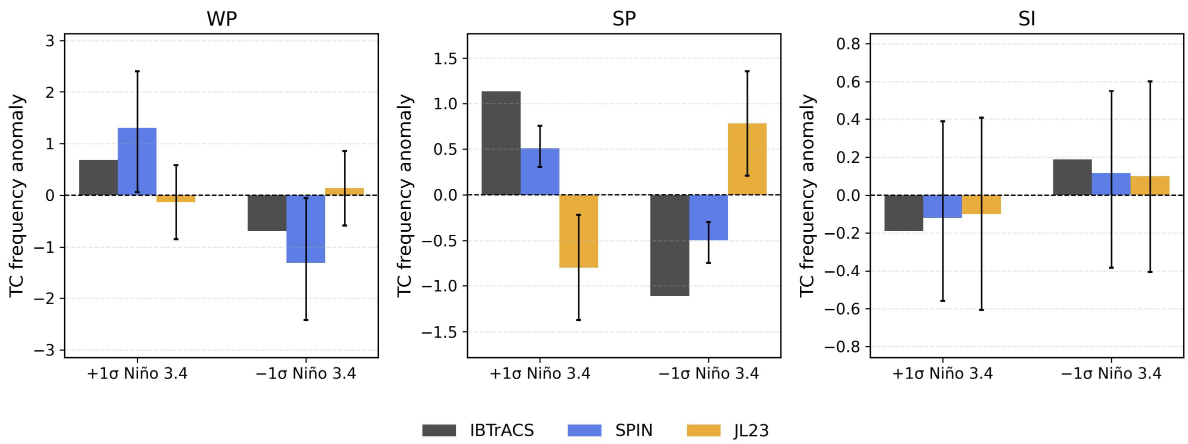
Here we show the maps of the El Niño minus La Niña difference in TC genesis density. Both SPIN and JL23 reproduce the sign of the multi-year mean ENSO-related genesis signal. Notably, in the WP, SP, and SI basins, positive and negative genesis anomalies coexist within the same basin. However, SPIN shows the magnitude of genesis density difference between El Niño and La Niña conditions that are closer to IBTrACS, whereas JL23 produces weaker and smoother anomalies (Response Fig. R6).

A possible reason is the difference in the treatment of seed disturbances. Previous diagnostic frameworks express TC genesis rates as the product of seed frequency and a transition probability controlled by the large-scale environment (Emanuel, 2022; Hsieh et al., 2020). In JL23, the seed supply is prescribed randomly and does not vary dynamically with ENSO related circulation changes. In contrast, SPIN identifies candidate storms directly from the evolving circulation in NeuralGCM, allowing both environmental favourability and precursor disturbance supply to vary from year to year. This may help SPIN retain regional ENSO responses that are partly smoothed in JL23. In other words, random seeding in JL23 may introduce disturbances in regions where activity should be suppressed, or miss regions where activity should be enhanced. We therefore present this mechanism as a plausible explanation, rather than a definitive attribution.



Response Fig. R6. TC genesis counts in 5° × 5° grid boxes per decade from IBTrACS, SPIN and JL23 during 1980-2022. JL23 genesis densities are shown after applying a constant global scaling factor so that the total JL23 genesis count over 1980–2022 matches that of IBTrACS. Top and middle rows show genesis counts during El Niño-like and La Niña-like months, respectively, defined using the Niño 3.4 SST index with thresholds of $+1\sigma$ and -1σ. The bottom row shows the El Niño minus La Niña difference in genesis counts per decade. The colour scale is linear between −1 and 1 and logarithmic for absolute differences greater than 1.

In addition, we examine ENSO conditioned TC frequency anomalies in regions with mixed-sign ENSO responses (Response Fig. R7). SPIN shows closer agreement with IBTrACS than JL23. In the WP, SPIN and IBTrACS both show positive anomalies under El Niño-like conditions and negative anomalies under La Niña-like conditions, whereas JL23 remains close to zero. In the SP, SPIN is also consistent with IBTrACS in sign, while JL23 shows the opposite tendency. In the SI, the ENSO-related anomalies are weaker and more uncertain; nevertheless, SPIN remains closer to the IBTrACS response than JL23. These results suggest that SPIN better captures ENSO modulation of TC frequency in these regions with mixed-sign ENSO responses.



Response Fig. R7. ENSO modulation of TC frequency in the WP, SP and SI basins. Bars show the TC frequency anomalies expected under +1σ (El Niño-like) and −1σ (La Niña-like) Niño 3.4 conditions, estimated from the linear relationship between TC frequency anomalies and the Niño 3.4 index over 1980–2022. TC frequency and Niño 3.4 index are calculated over May–December for the WP basin and December–April for the SP and SI basins. TC anomalies are expressed relative to each dataset's own climatology. Error bars show the 10th–90th percentile range across SPIN ensemble members and the 95% OLS confidence interval for JL23. JL23 counts are scaled separately within each basin so that the total JL23 TC count in each basin matches that of IBTrACS.

Changes in the manuscript:

To make this point clearer in the manuscript, we have revised the text in Lines 292–305 as follows:

> "SPIN also shows markedly improved agreement with observations relative to JL23 in the WP, SP, and SI basins. Lin et al. (2023) suggested that lower predictability in these basins may reflect the mixed-sign influence of El Niño–Southern Oscillation (ENSO) on TC genesis across subregions (Figure S6, Camargo et al., 2007; Lin et al., 2023), which can cause the basin-integrated interannual genesis signal to be averaged out. In contrast, SPIN better retains ENSO modulation in these basins, with ENSO conditioned TC frequency anomalies closer to IBTrACS than JL23 (Figure S7). A possible reason is the difference in the treatment of seed disturbances. Previous diagnostic frameworks express TC genesis rates as the product of seed frequency and a transition probability controlled by the large-scale environment (Emanuel, 2022; Hsieh et al., 2020). JL23 uses random seeding (Emanuel et al., 2008; Lin et al., 2023), in which seeds are placed randomly in space and time and then allowed to evolve within the large-scale environment. Because this seed supply is not dynamically coupled to ENSO related circulation changes, it may artificially increase genesis probability in low-activity regions while decreasing it in high-activity regions. In contrast, SPIN diagnoses candidate storms from the evolving circulation fields generated by

NeuralGCM. This allows SPIN to retain year-to-year changes in both environmental favorability and precursor disturbance supply. We regard this as a plausible mechanism rather than a definitive attribution. Overall, these results highlight the potential of SPIN to capture interannual variability in TC activities across most basins.”

Section 3.1

The tendency of SPIN to overestimate TC genesis in the early-to-peak season and underestimate it toward the end of the season is acknowledged but not investigated. Given that this bias is attributed to NeuralGCM, the authors should provide a more thorough discussion of its likely physical origin — whether related to biases in large-scale circulation timing, smoothed monthly SST forcing, or misrepresentation of TC precursors.

Response:

Thank you very much for this comment. We consider the seasonal genesis bias appears to be most closely related to biases in the NeuralGCM simulated environment. Compared with ERA5, NeuralGCM tends to produce lower VWS during the early-to-peak TC season, but higher VWS during the late season (Response Fig. R4). These biases favour TC genesis earlier in the season but suppress it later, consistent with the simulated seasonal genesis bias. Smoothed monthly SST forcing and misrepresentation of synoptic-scale TC precursors may contribute to individual genesis errors, but they are less likely to explain the systematic early-season overestimation and late-season underestimation.

Changes in the manuscript:

We have therefore further discussed this in the updated manuscript (Lines 312–316):

> “A similar bias was noted by Zhang et al. (2025, supplementary) that evaluates the performance of NeuralGCM. This seasonal bias is likely linked to the large-scale environmental biases in NeuralGCM. As shown in Section 3.1, NeuralGCM tends to predict lower VWS than ERA5 during the early-to-peak TC season, but higher VWS during the late season (Figure S5). These biases favour TC genesis earlier in the season but suppress it later, consistent with the simulated seasonal genesis bias.”

Section 3.2

The paragraph describing the track density results (lines 234-238) is repeated at lines 244-249.

Response:

Thank you for pointing this out. The duplicated paragraph has been removed.

Section 3.3

The tuning procedure for Ck is not described. The authors should specify which dataset was used and how Ck was calibrated, to ensure the subsequent intensity evaluation in Sections 3.3 and 4 is not a possible artifact of parameter fitting.

Response:

Thank you very much for pointing this out. In the previous manuscript, different values of Ck were used for different basins and regions to improve agreement with observations. We agree that this made the intensity evaluation less straightforward and could make it difficult to distinguish the baseline performance of SPIN from improvements introduced by basin- or region-specific parameter tuning.

In the revised manuscript, we have removed all basin- and region-specific tuning of Ck. Instead, all SPIN results are produced using a single fixed value, Ck = 1.05. This value is used consistently throughout the manuscript. The intensity evaluation in Sections 3.3 and 4 therefore reflects the performance of SPIN under one fixed FAST configuration, rather than an optimized fit to individual regions.

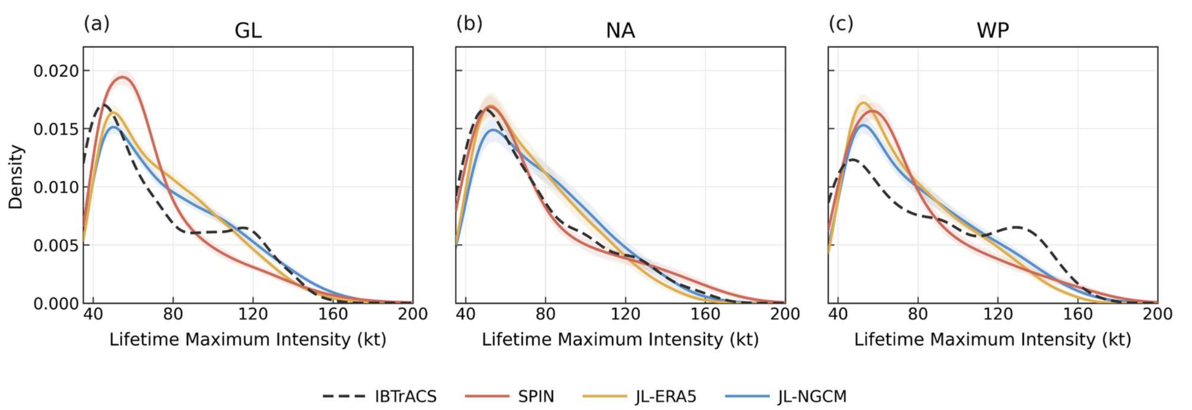
However, this value is not intended to be an optimal calibration for every basin or regional intensity distribution. As noted in the JL23 documentation, model parameters are provided through the namelist for tuning.

Changes in the manuscript:

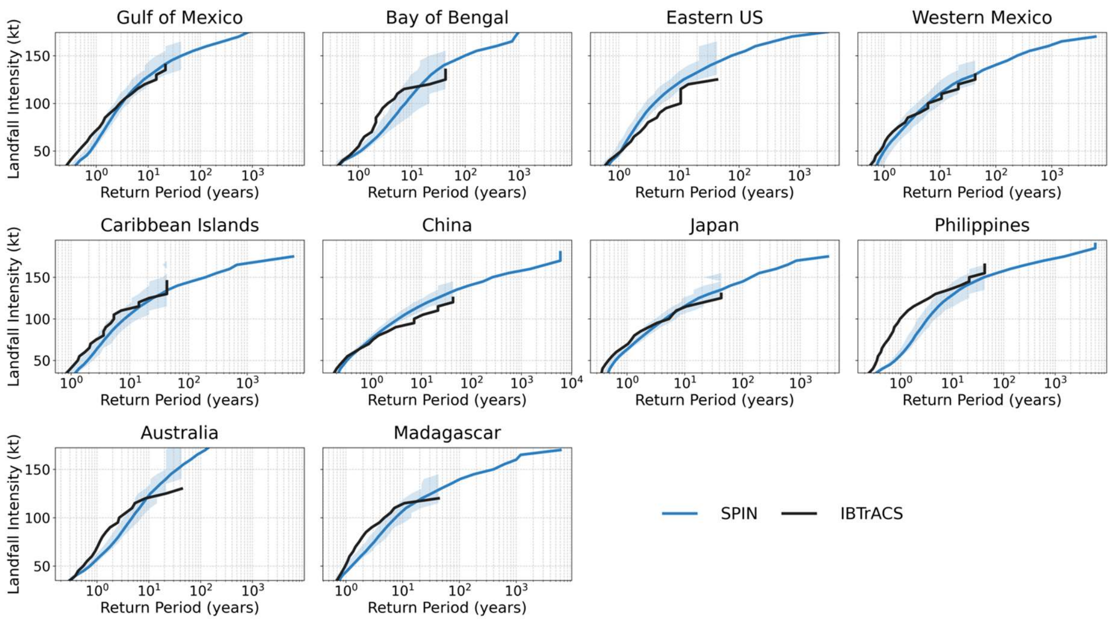
We have added the following clarification in Section 2.2.3:

> "All model parameters are adopted from Lin et al. (2023), except for the surface enthalpy coefficient (Ck). We use Ck = 1.05 as the default setting for all SPIN results presented in this manuscript. This value is slightly smaller than the Ck = 1.2 used in JL23 default settings, as larger values produced an overly heavy upper tail in the simulated LMI distribution under the 6-hourly forced SPIN configuration. No basin- or region-specific tuning is applied. However, this value is not intended to be an optimal calibration for every basin or regional intensity distribution. As noted in the JL23 documentation, model parameters are provided through the namelist for tuning. Users may therefore explore parameter space for application-specific regional hazard assessments."

All affected figures have been regenerated using this fixed Ck setting, including the LMI distributions and regional landfall intensity return period curves shown in Response Figs. R8 and R9. Although some regional biases change quantitatively, the main conclusions remain unchanged. We have also revised Sections 3.3 and 4.1 to discuss the possible origins of the remaining biases.



Response Fig. R8. Distribution of LMI of TCs during 1980–2022. (a) shows the global distribution, (b) the NA basin, and (c) the WP basin distributions. Distributions are estimated using Gaussian kernel density fitting. Results of IBTrACS are shown as black dashed lines, SPIN simulations in red, JL-NGCM (using a single ensemble member) in blue, and JL-ERA5 in yellow. For JL-NGCM and JL-ERA5, downscaled events are randomly subsampled to match the number of observed events. For SPIN, shading denotes the 10th–90th percentile range across ensemble members; for the JL-downscaled results, shading denotes the 10th–90th percentile range from repeated subsampling to the observed sample size.



Response Fig. R9. Return periods of landfall intensity for the following regions: Gulf of Mexico, Bay of Bengal, Eastern United States, Western Mexico, Caribbean Islands, China, Japan, Philippines, Australia, and Madagascar. Results from IBTrACS are shown in black, the solid blue line shows the return period curve computed from all ensemble members of SPIN combined (i.e., stacking all 14 members, 43 years, and 10 runs, yielding a total of 6,020 years). The shaded region represents the 10th–90th percentile spread across all individual ensemble members.

Section 3.5

The authors do not discuss the physical mechanisms behind SPIN's improved representation of inter-genesis times.

Response:

Thank you very much for pointing this out. Uniformly sampling genesis dates within each month assumes independence among individual TCs, which is problematic for MTCEs because TC genesis can cluster under persistent favorable large-scale conditions (Gao and Li, 2011; Ito and Yamauchi, 2026), and pre-existing TCs may also influence the genesis of subsequent storms (Ritchie and Holland, 1999; Yoshida and Ishikawa, 2013). The improved performance of SPIN in representing inter-genesis times may be linked to the demonstrated skill of the NeuralGCM component in representing intra-month-scale environmental evolution, including MJO teleconnections (Peings et al., 2026). This may help SPIN capture TC inter-genesis time distributions without imposing independence among storms.

We have added the corresponding discussion to Section 3.6 Lines 414–418.

Section 3.6

Line 360: "Figure 3" in place of "Figure 11."

Response:

Thank you for pointing this out. The incorrect figure reference has been corrected.

Section 3.6

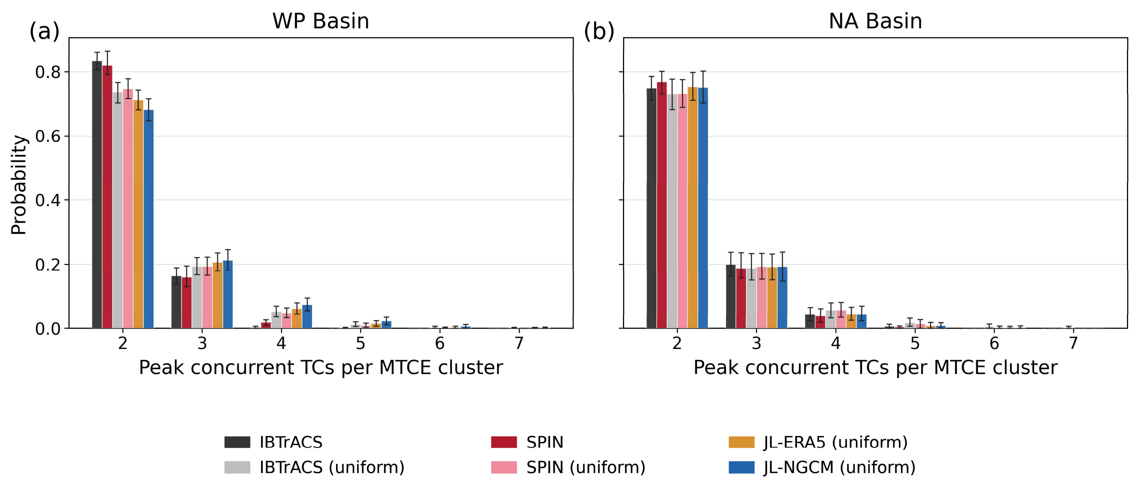
The authors attribute SPIN's improved MTCE representation to two-way TC-environment interactions. Would a "SPIN-uniform" experiment, in which NeuralGCM tracks are retained but genesis dates are resampled uniformly within each month, help isolate the contribution of physical TC-environment interactions from that of temporal resolution alone? The aspect is quite important as it is presented as one of the major improvements of SPIN model compared to existing TCs tracks models.

Response:

Thank you very much for this important suggestion. First of all, the wording in the original manuscript may have overstated the role of two-way TC–environment interactions, because the FAST intensity component is applied offline and does not feed back onto the simulated environment. We have therefore avoided the statement that SPIN allows two-way TC–environment interactions from the revised manuscript. Instead, we interpret the improvement in MTCE representation as mainly arising from NeuralGCM's ability to represent dynamically evolving intra-month environmental conditions from which TCs can emerge directly.

Following the reviewer's suggestion, we added a "SPIN-uniform" experiment. This experiment tests whether SPIN captures physically organized TC genesis timing associated with intra-month processes, rather than merely producing random genesis timing at a finer temporal resolution.

The SPIN-uniform experiment generally overestimates the frequency of larger TC clusters compared with the original SPIN simulation. This indicates that the original SPIN genesis timing is not equivalent to random intra-month allocation, but may be organized by the evolving environmental conditions represented by NeuralGCM. The larger difference between the uniform and non-uniform timing experiments in WP basin may reflect the stronger influence of intra-month to subseasonal variability on TC genesis in this basin, such as MJO- and BSISO-related variability, but a full mechanistic attribution of basin-dependent MTCE formation is beyond the scope of this study.



Response Fig. R10. Histogram of MTCE cluster size for the (a) WP and (b) NA basins during 1980–2022. Cluster size is defined as the peak number of concurrent TCs within each MTCE cluster. The x-axis shows the peak concurrent TC count, and the y-axis shows the probability. Results are shown for IBTrACS (black), IBTrACS with uniformly sampled genesis dates (grey), the SPIN ensemble mean (red), SPIN with uniformly sampled genesis dates (pink), JL23 downscaling driven by ERA5 environmental fields with uniformly sampled genesis dates (yellow), and JL23 downscaling driven by NeuralGCM environmental fields with uniformly sampled genesis dates (blue). Error bars indicate the 10th–90th percentile range. For IBTrACS, the range is estimated by year bootstrapping over 1980–2022. For SPIN, it represents the spread across ensemble members. For the "uniform" experiments, the range is estimated across 1000 Monte Carlo realizations.

Changes in the manuscript:

To make this point clearer in the manuscript, we have revised the text in Lines 455–466 as follows:

> "Next, we evaluate the distribution of MTCE cluster sizes, defined as the peak number of concurrent TCs within a cluster (see section 2.5; Figure 13). SPIN closely reproduces the observed distribution, with MTCEs most frequently exhibiting two peak concurrent TCs. The probability decreases with increasing numbers of concurrent TCs per MTCE cluster, and clusters with more than five concurrent TCs are rare. In contrast, uniform-date sampling methods, whether applied to IBTrACS, SPIN, or synthetic TCs generated by JL23 using NeuralGCM or ERA5 environmental inputs, generally overestimate the frequency of larger clusters, especially in the WP basin. This bias is consistent with the overrepresentation of short TC inter-genesis intervals within the 0–5-day bin under uniform-date sampling (Figure 11), which artificially increases the likelihood of multiple TCs being included within the same cluster. The stronger contrast between the original and uniform-timing experiments in the WP basin than in the NA basin may reflect a greater role of intra-month to subseasonal variability in organizing TC genesis in the WP, although the underlying dynamical controls remain to be clarified. These results suggest that SPIN captures physically organized intra-month TC genesis timing, providing a more realistic basis for estimating MTCE cluster sizes than uniform-date sampling."

References:

Bloemendaal, N., Haigh, I. D., Moel, H. de, Muis, S., Haarsma, R. J., and Aerts, J. C. J. H.: Generation of a global synthetic tropical cyclone hazard dataset using STORM, Scientific Data, 7, 40, https://doi.org/10.1038/s41597-020-0381-2, 2020.

Camargo, S. J., Emanuel, K. A., and Sobel, A. H.: Use of a Genesis Potential Index to Diagnose ENSO Effects on Tropical Cyclone Genesis, https://doi.org/10.1175/JCLI4282.1, 2007.

Cheng, K.-Y., Harris, L., Bretherton, C., Merlis, T. M., Bolot, M., Zhou, L., Kaltenbaugh, A., Clark, S., and Fueglistaler, S.: Impact of Warmer Sea Surface Temperature on the Global Pattern of Intense Convection: Insights From a Global Storm Resolving Model, Geophysical Research Letters, 49, e2022GL099796, https://doi.org/10.1029/2022GL099796, 2022.

Emanuel, K.: A fast intensity simulator for tropical cyclone risk analysis, Nat Hazards, 88, 779–796, https://doi.org/10.1007/s11069-017-2890-7, 2017.

Emanuel, K.: Tropical Cyclone Seeds, Transition Probabilities, and Genesis, https://doi.org/10.1175/JCLI-D-21-0922.1, 2022.

Emanuel, K., Sundararajan, R., and Williams, J.: Hurricanes and Global Warming: Results from Downscaling IPCC AR4 Simulations, Bull. Amer. Meteor. Soc., 89, 347–368, https://doi.org/10.1175/BAMS-89-3-347, 2008.

Hsieh, T.-L., Vecchi, G. A., Yang, W., Held, I. M., and Garner, S. T.: Large-scale control on the frequency of tropical cyclones and seeds: a consistent relationship across a hierarchy of global atmospheric models, Clim Dyn, 55, 3177–3196, https://doi.org/10.1007/s00382-020-05446-5, 2020.

Jing, R. and Lin, N.: An Environment-Dependent Probabilistic Tropical Cyclone Model, Journal of Advances in Modeling Earth Systems, 12, e2019MS001975, https://doi.org/10.1029/2019MS001975, 2020.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., and Hoyer, S.: Neural General Circulation Models for Weather and Climate, Nature, 632, 1060–1066, https://doi.org/10.1038/s41586-024-07744-y, 2024.

Krichene, H., Vogt, T., Piontek, F., Geiger, T., Schötz, C., and Otto, C.: The social costs of tropical cyclones, Nat Commun, 14, 7294, https://doi.org/10.1038/s41467-023-43114-4, 2023.

Lee, C., Tippett, M. K., Sobel, A. H., and Camargo, S. J.: An Environmentally Forced Tropical Cyclone Hazard Model, J Adv Model Earth Syst, 10, 223–241, https://doi.org/10.1002/2017MS001186, 2018.

Lin, J., Rousseau-Rizzi, R., Lee, C.-Y., and Sobel, A.: An Open-Source, Physics-Based, Tropical Cyclone Downscaling Model With Intensity-Dependent Steering, Journal of Advances in Modeling Earth Systems, 15, e2023MS003686, https://doi.org/10.1029/2023MS003686, 2023.

Moon, J., Kim, D., Balaguru, K., Leung, L. R., and Seo, E.: Joint Influence of Supportive and Unsupportive Environmental Conditions on Tropical Cyclone Rapid Intensification in Two HighResMIP Simulations, Geophysical Research Letters, 53, e2025GL119986, https://doi.org/10.1029/2025GL119986, 2026.

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vannière, B., Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L.-P., Chauvin, F., Terray, L., Valcke, S., Moine, M.-P., Putrasahan, D., Roberts, C. D., Senan, R., Zarzycki, C., Ullrich, P., Yamada, Y., Mizuta, R., Kodama, C., Fu, D., Zhang, Q., Danabasoglu, G., Rosenbloom, N., Wang, H., and Wu, L.: Projected Future Changes in Tropical Cyclones Using the CMIP6 HighResMIP Multimodel Ensemble, Geophysical Research Letters, 47, e2020GL088662, https://doi.org/10.1029/2020GL088662, 2020.

Tang, Y., Huangfu, J., Huang, R., and Chen, W.: Simulation and Projection of Tropical Cyclone Activities over the Western North Pacific by CMIP6 HighResMIP, Journal of Climate, 35, 7771–7794, https://doi.org/10.1175/JCLI-D-21-0760.1, 2022.

Vickery, P. J., Skerlj, P. F., and Twisdale, L. A.: Simulation of Hurricane Risk in the U.S. Using Empirical Track Model, Journal of Structural Engineering, 126, 1222–1237, https://doi.org/10.1061/(ASCE)0733-9445(2000)126:10(1222), 2000.

Zhang, G., Rao, M., Yuval, J., and Zhao, M.: Advancing seasonal prediction of tropical cyclone activity with a hybrid AI-physics climate model, Environ. Res. Lett., 20, 094031, https://doi.org/10.1088/1748-9326/adf864, 2025.