Characterizing uncertainty in synthetic tropical cyclone hazard models for U.S. energy infrastructure resilience

ENVIRONMENTAL RESEARCH
CLIMATE

OPEN ACCESS

RECEIVED
5 March 2026

REVISED
18 May 2026

ACCEPTED FOR PUBLICATION
22 May 2026

PUBLISHED
22 June 2026

PAPER

Characterizing uncertainty in synthetic tropical cyclone hazard models for U.S. energy infrastructure resilience

Effy B John[1], **Karthik Balaguru**[1,*], **Sha Feng**[1], **Kerry Emanuel**[2], **Chia-Ying Lee**[3], **Julian Rice**[1], **Nicholas Lalo**[1], **L Ruby Leung**[1], **David Judi**[1] **and Larry Berg**[1]

[1] Pacific Northwest National Laboratory, Richland, WA, United States of America
[2] Lorenz Center, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[3] Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: Karthik.Balaguru@pnnl.gov

Keywords: tropical cyclones, coastal risk, infrastructure resilience

Supplementary material for this article is available online

Abstract

Tropical Cyclones (TCs) are intense storms that pose a persistent and considerable risk to coastal communities and infrastructure in the global tropics and subtropics, including the United States (US). With known limitations associated with observations and high-resolution earth system models, synthetic TC models that capture a wide spectrum of storm possibilities have been developed to robustly quantify TC risk. Here we examine the simulation of various TC features in the North Atlantic relevant for US coastal risk in three synthetic TC models forced with ERA5 reanalysis: Massachusetts Institute of Technology (MIT), Columbia HAZard (CHAZ) and Risk Analysis Framework for Tropical Cyclones (RAFT). While there is a broad agreement among these models in terms of their representation of salient TC characteristics, certain differences do exist. To connect these modeling uncertainties with energy infrastructure resilience, we apply fragility curves that link simulated TC intensities to damage probabilities, demonstrating how uncertainty in storm states may translate into that in coastal impacts. Our study indicates that acknowledging and accounting for inter-model uncertainty leads to more reliable risk assessments, strengthening science-to-action pathways for managing risks associated with TCs.

1. Introduction

Tropical cyclones (TCs) are among the most destructive natural hazards causing widespread damage and disruption in coastal regions around the world (Geiger *et al* 2016, Noy 2016, Cerveny *et al* 2017). They continue to pose a persistent and growing threat to coastal communities, especially in the Atlantic basin, where exposure to high-impact TCs has increased markedly in recent decades (Holland and Bruyère 2014, Klotzbach *et al* 2018, Beven *et al* 2019, Willoughby *et al* 2024). During the past four decades, North Atlantic hurricanes alone have caused cumulative damages exceeding $1.5 trillion in the United States (US), with some individual storms, such as Katrina (2005), Sandy (2012) and Ida (2021) generating economic losses between $75 billion and $170 billion (adjusted to 2023 USD, (NOAA Office for Coastal Management 2025)). In addition to these staggering financial costs, TCs have caused thousands of fatalities and prolonged disruptions in housing, energy infrastructure, and transportation networks (Weinkle *et al* 2012, Muller *et al* 2025). Given that TCs exert their most profound influence at regional-to-local scales, comprehensive risk assessments at these levels are crucial to accurately quantify hazards and guide effective adaptation and mitigation efforts (Peduzzi *et al* 2012, Eberenz *et al* 2020).

Despite the significant threat posed by TCs, landfalling events at a given location are relatively infrequent, resulting in sparse observational records over time (Weinkle *et al* 2012). This scarcity is compounded by historical data gaps, particularly for TCs that occurred before the satellite era (Romero *et al* 2025). As a result, the long-term TC risk assessment based solely on observations are limited in

their ability to capture the full range of TC hazards. On the other hand, high-resolution earth system model simulations that explicitly resolve TC convective processes and produce intense storms tend to be computationally expensive (Murakami and Sugi 2010, Shaevitz *et al* 2014, Wehner *et al* 2018, Bourdin *et al* 2024). Further, the intensity and frequency of TCs are expected to change (Webster *et al* 2005, Elsner *et al* 2008, Knutson *et al* 2008, 2022, Bender *et al* 2010, Holland and Bruyère 2014, Walsh *et al* 2016, Patricola and Wehner 2018, Bhatia *et al* 2019, Kossin *et al* 2020, Murakami *et al* 2020, Balaguru *et al* 2023), and anticipating potential changes in TC risk requires modeling frameworks that can generate realistic long-term TC scenarios in a range of different conditions (Gettelman *et al* 2018). To address this challenge, synthetic TC models have emerged as viable options (Vickery *et al* 2000, Hall and Jewson 2007, Yonekura and Hall 2011, Nederhoff *et al* 2021). These models offer a powerful alternative by simulating thousands of physically plausible TC intensities and tracks that go beyond the limitations of the historical record, even capturing low-probability, high-impact events (Emanuel *et al* 2006, Lee *et al* 2018, Bloemendaal *et al* 2020b, Xu *et al* 2024) at a relatively low computational cost. These models vary in methodology and complexity, ranging from purely statistical methods (Vickery *et al* 2000, Bloemendaal *et al* 2020b, Nederhoff *et al* 2021), hybrid statistical-dynamical systems (Emanuel *et al* 2006, Lee *et al* 2018) to machine-learning based frameworks (Xu *et al* 2024). One of the main advantages of using synthetic TC models is that they can simulate a large number of TCs with relatively few input variables and minimal computational demand, enabling robust probabilistic risk assessment (Hardy *et al* 2003, Haigh *et al* 2014, Xu *et al* 2024).

Synthetic TC models generally consist of three key components: TC genesis, track, and intensity. The genesis component determines when and where TCs are likely to form, often using statistical methods with historic data or indices derived from environmental conditions (Vickery *et al* 2000, Emanuel 2006, Lee *et al* 2018, Xu *et al* 2024). The track component simulates TC motion across space by representing the influence of large-scale atmospheric steering flows, through statistical, dynamical, or hybrid methods (Emanuel 2006, Lee *et al* 2018). The intensity component governs how TCs strengthen or weaken along their tracks, based on statistical models, physics-based models, hybrid methods, or machine learning-based techniques that capture complex relationships between environmental drivers and TC development (Emanuel 2006, Lee *et al* 2018, Bloemendaal *et al* 2020b, Xu *et al* 2021, 2024). Together, these components provide a framework for generating a large number of TCs that are widely used in risk assessment and climate impact studies (Meiler *et al* 2022). Also, the diversity in model design may lead to variations in TC characteristics, which can further inform the sensitivity of simulating TC-related risks to the approach employed (Bloemendaal *et al* 2022, Meiler *et al* 2022, Romero *et al* 2025). By integrating synthetic TC data with exposure and vulnerability data, they inform risk assessments for energy infrastructure supporting resilience design and adaptation planning under current and future environmental conditions (Arkema *et al* 2023).

This manuscript presents an analysis of synthetic TC models using the Massachusetts Institute of Technology (MIT) model, the Columbia HAZard (CHAZ) model and the risk analysis framework for tropical cyclones (RAFTs) to understand uncertainty among models with a focus on the North Atlantic basin. A common forcing dataset, ERA5 reanalysis, is used to force all three TC models, which ensures that all models experience the same large-scale ocean-atmosphere conditions. Using three synthetic TC models that are forced by the same reanalysis, we aim to understand the impact of model formulation on TC risk estimates while minimizing differences in the underlying environmental forcing. This removes a major source of uncertainty and allows differences in results to be attributed primarily to how each model represents TC genesis, track evolution, and intensity. Within this controlled framework, the three models represent three types of statistical dynamical downscaling methods and span a spectrum of methodological complexity. The statistical-dynamical stochastic model (CHAZ) captures physically informed relationships between large-scale climate and cyclone behavior. The statistical-dynamical deterministic model (MIT) reduces stochasticity, providing a more constrained and reproducible mapping from environment to cyclone characteristics. The statistical-dynamical machine learning model (RAFT) introduces flexible, data-driven representations that can capture nonlinear relationships not easily expressed in traditional formulations. By comparing these approaches under identical forcing, the study is able to assess the robustness of hazard estimates to modeling. This framework strengthens confidence in consistent signals across models while highlighting areas where methodological improvements are most needed. Previous studies have conducted inter-comparison of synthetic TC models (Meiler *et al* 2022), whereas this study focuses on methodological uncertainties. While more synthetic TC models could be included for a more robust comparison of different approaches, three models are used in this study to sample not only the diverse methodologies but also to represent models that have been more broadly tested and widely used (MIT and CHAZ) vs. a relatively newer model (RAFT) that incorporates an ML-based method. The synthetic TC datasets span 1980–2018 (39 years), with substantial variation

in the number of simulated tracks across models: ∼900k (CHAZ), ∼80k (RAFT), and ∼100k (MIT). In this study, our analysis is limited to meteorological comparisons, where the performance of synthetic TC hazard models is evaluated against historical observations. We focus on key attributes such as TC frequency, intensity and intensification over the North Atlantic, especially near the US Coast. By benchmarking model outputs against historical records and each other, we aim to identify strengths, limitations and uncertainty within each modeling approach. Such comparative insights are essential for end-users who rely on these models for decision-making in disaster risk reduction and hazard mitigation.

2. Data and methods

2.1. Observation

2.1.1. International best track archive for climate stewardship (IBTrACS)

The observed TC track and intensity data for this study were obtained from the IBTrACS, maintained by NOAA's National Centers for Environmental Information. In the North Atlantic basin, best-track estimates are provided primarily by the National Hurricane Center (Knapp *et al* 2010, Gahtan *et al* 2024). Although the IBTrACS record for the Atlantic dates back to 1851, its temporal consistency and reliability have been substantially improved by key observational advances: routine aircraft reconnaissance beginning in 1946, the advent of satellite imagery in the 1980s, and the deployment of GPS dropsonde measurements in the late 1990s (Hock and Franklin 1999, Neumann 1999, Landsea and Franklin 2013). Consequently, the most reliable information on the best-track archive extends from the 1980s onward (Vecchi and Knutson 2011, Landsea and Franklin 2013). In this study, we focus mainly on TC characteristics in the North Atlantic Ocean and near the US Coast and following (Emanuel 2025), we selected all TCs that attained at least tropical storm (TS) intensity ($\geqslant$40 knots) between 1980 and 2018.

2.2. Synthetic TC models

A brief description of each of the models used in this study is provided below.

2.2.1. CHAZ model

CHAZ is a statistical-dynamical downscaling TC model that generates tens of thousands of synthetic TCs based on environmental conditions (Lee *et al* 2018). TC genesis is determined based on the tropical cyclone genesis index (TCGI), an index that combines the sea surface temperature (SST), vertical wind shear, vorticity, and environmental moisture (Tippett *et al* 2011, Camargo *et al* 2014, Lee *et al* 2018) to indicate the probability of genesis at a given location. After a weak TC vortex is introduced, the track is subsequently determined using a beta-advection scheme based on large-scale wind fields with stochastic perturbations to represent monthly variability, same as what is used in the MIT model (Emanuel *et al* 2006). Next, the intensity along tracks is estimated using an auto-regressive statistical method that combines deterministic environmental conditions and random stochastic elements (DeMaria *et al* 2005, Lee *et al* 2015, 2016). Note that unlike in MIT and RAFT where all tracks are assigned equal weights, CHAZ tracks are not equally weighted. Each synthetic TC track in CHAZ includes a frequency parameter, in addition to intensity and position. This parameter, which is a new addition to the CHAZ dataset, is designed for post-hoc adjustment to objectively adhere synthetic data to historically observed landfall statistics without compromising the model's reliance on the ambient environment. The frequency information is applied to return period (RP) analysis only.

2.2.2. MIT

The MIT model is a statistical-deterministic downscaling model which generates synthetic TCs based on the ambient storm environment (Emanuel *et al* 2006). In the MIT model, TCs are stochastically initiated ('seeded') at random locations and times throughout the basin. Each seed represents a weak warm-core disturbance. The subsequent track is determined using a beta-advection model, which advects TCs according to the large-scale environmental flow derived from reanalysis or earth system models. Finally, the TC intensity at each track location is simulated using the Coupled Hurricane Intensity Prediction System (CHIPS, (Emanuel *et al* 2004)), a physics-based approach that accounts for both dynamical and thermodynamical processes in the ocean and atmosphere (Emanuel 2006, Emanuel *et al* 2006).

2.2.3. RAFT

RAFT is a hybrid model that integrates statistical, dynamical and machine learning techniques to generate a large number of synthetic TCs for hazard and risk assessment (Xu *et al* 2024). In RAFT, TCs are initiated randomly following a Gaussian spatio-temporal distribution of observed genesis (Emanuel

et al 2006). Once initialized, TC tracks are determined based on a beta-advection scheme (Marks 1992, Emanuel 2006) in which TCs are primarily steered by a weighted mean of 850 hPa and 200 hPa winds (Marks 1992) to which a spatially-varying beta-drift correction term is added (Wu and Wang 2004, Zhao *et al* 2009). Finally, the TC intensity at each track location is computed using a deep neural networks-based approach that predicts TC intensity using the current TC state and the large-scale storm environment (Xu *et al* 2021).

All three TC models considered here—MIT, CHAZ, and RAFT—are forced using environmental fields derived from the ERA5 reanalysis, but they differ significantly in how these fields are used to represent genesis, track, and intensity. For genesis, CHAZ explicitly links TC formation to large-scale environmental conditions derived from ERA5, including low-level absolute vorticity, mid-level relative humidity, SST, and vertical wind shear. These variables are combined into a TCGI, which determines the probability of storm formation as a function of the large-scale environment. In contrast, RAFT generates initial storm locations from a statistical distribution rather than directly from environmental conditions. Specifically, genesis locations are sampled as a random draw from a Gaussian kernel-based probability distribution centered on historically observed TC genesis locations. This approach reproduces the observed spatial climatology of TC genesis, while environmental variables from ERA5 primarily influence subsequent TC evolution, rather than explicitly determining when and where storms form. The MIT model adopts a fundamentally different approach. TCs are initiated by seeding randomly in space and time throughout the domain, independent of environmental favorability. For all three models, only seeds encountering favorable environmental conditions survive and intensify. CHAZ has roughly 70%–80% survival rate while the vast majority of seeds (>99%) in MIT dissipate. These seeds are embedded within a synthetic, time-evolving large-scale environment constructed from ERA5 fields, in which monthly means, variances, covariances, and kinetic energy spectra are preserved using a Fourier-based representation.

All three models use ERA5-derived large-scale winds to drive track evolution. CHAZ, MIT and RAFT combine beta drift with synthetic winds consistent with ERA5 to generate large ensembles of dynamically consistent tracks. For intensity evolution, ERA5-derived thermodynamic and kinematic variables play a central role in all models. The MIT model uses a deterministic, axisymmetric intensity model (CHIPS, (Emanuel *et al* 2004)), which consists of an axisymmetric atmospheric component coupled to a simple one-dimensional ocean model. This model captures key effects of upper-ocean mixing and is forced by potential intensity, along with the local ocean mixed-layer depth and thermal stratification at the TC location. CHAZ employs an autoregressive statistical model in which intensity tendencies depend on predictors such as potential intensity, vertical wind shear, and mid-level humidity. RAFT uses a machine-learning-based model trained on global Statistical Hurricane Intensity Prediction Scheme (SHIPS, (DeMaria and Kaplan 1994, 1999, DeMaria *et al* 2005)) predictors and ERA5-derived variables-including low-level relative humidity, maximum potential intensity and equivalent potential temperature-allowing for nonlinear relationships between environment and intensity change. Despite being forced by the same ERA5 dataset, the models differ in how environmental information is used leading to differences in simulated TC characteristics.

2.3. Evaluation metrics

In this study, we used synthetic TCs generated from the above-mentioned TC models driven by environmental forcings of ERA-5 reanalysis (Hersbach *et al* 2020). Since we analyze various synthetic TC models forced with the same reanalysis data, differences in their simulated TC characteristics reflect mainly the model-specific dynamics and parameterizations, allowing us to understand the uncertainty associated with the outputs, which is the main goal of this study. Despite the existence of multiple approaches in the literature for synthetic TC model intercomparison (Meiler *et al* 2022, Romero *et al* 2025), we restrict our analysis to TC metrics presented in Emanuel (2025). This choice is motivated by the associated statistical testing framework, which provides a rigorous, consistent, and objective basis for evaluating synthetic TC models against observations. In the framework of Emanuel (2025), a set of TC metrics is defined to quantify the key aspects of TC climatology and to enable objective comparison between synthetic model output and historical observations. The metrics span storm frequency, seasonal timing, coastal impacts, and intensity characteristics and are described below:

Intensity histogram: The distribution of lifetime maximum intensities of TCs, focused on North Atlantic, with the total synthetic frequency scaled to match observations; this captures how well a model reproduces the frequency of weak versus strong TCs in the region of interest.

Intensity change histogram: The histogram of 6-hourly intensity changes over each TC's lifetime, expressed as the natural logarithm of one plus the frequency; this metric reflects the model's ability to simulate storm intensification and weakening processes.

Annual cycle: The annual cycle refers to the monthly distribution of TC counts in the North Atlantic Ocean. This metric measures how well a synthetic TC model reproduces the observed seasonal timing of TC activity. In this study, synthetic monthly counts are normalized to match the observed annual frequency.

Gate crossings: The number of TCs crossing predefined coastal gates (Fig S1) along each coastline, with synthetic counts normalized to match observed totals; this metric characterizes TC landfall likelihood for various coastal locations. The coastal gates are defined as an ordered sequence of geospatial points located along the U.S. Gulf and East Coasts, positioned immediately offshore to closely follow the coastal geometry while avoiding inland placement. There are 51 coastal gates that are distributed at approximately uniform intervals of 100 km along the U.S. Gulf and East Coasts. These locations were initially derived from the coastline and subsequently refined through manual inspection to ensure spatial continuity. When connected sequentially, the gates form a continuous coastal transect that functions as a reference boundary for detecting TC coastal-crossing events.

Coastal crossing-to-basin ratio: The ratio of total gate crossings to basin counts for a given coastline; unlike the gate crossings themselves, this ratio is not normalized, providing a direct measure of how frequently TCs from a basin make landfall over a given location.

2.4. Goodness-of-fit tests

These metrics are evaluated through statistical tests (e.g. Poisson log-likelihood and *K*-test) to assess agreement between synthetic and observed distributions, providing a rigorous model intercomparison. All TC metrics are evaluated using a suite of statistical tests to provide a rigorous and objective assessment of model performance. The statistical tests include Poisson log-likelihood(D'), *K*-test, coefficient of determination and χ^2 tests, which together quantify deviations from observations and assess differences in distributions while accounting for sampling variability. A brief description of the statistical tests used in this study is given below.

2.4.1. Poisson-log Likelihood test

Agreement between observed and synthetic counts is quantified using the Poisson deviance, defined as

$$D = \sum_i 2\left[y(i)_{\text{obs}} \ln\left(\frac{y(i)_{\text{obs}}}{y(i)_{\text{syn}}}\right) - \left(y(i)_{\text{obs}} - y(i)_{\text{syn}}\right)\right] \tag{1}$$

where y_{obs} is the observed TC count, y_{syn} is the synthetic TC count and i is the TC metric evaluated. The deviance is normalized and transformed into a skill score with 1 indicating perfect agreement.

2.4.2. K test

The *K*-test is a non dimensional goodness-of-fit metric that assigns perfect skill when observed counts fall within the Poisson sampling uncertainty bounds of the synthetic predictions, thereby explicitly accounting for observational sampling error. *K*-test is defined as,

$$K = \begin{cases} \sum_i \dfrac{\max\left(y(i)_{\text{obs}} - y(i)_{\text{upper}}, 0\right)}{y(i)_{\text{syn}}}, & y(i)_{\text{obs}} \geqslant y(i)_{\text{syn}} \\ \sum_i \dfrac{\max\left(y(i)_{\text{lower}} - y(i)_{\text{obs}}, 0\right)}{y(i)_{\text{syn}}}, & y(i)_{\text{obs}} < y(i)_{\text{syn}}. \end{cases} \tag{2}$$

Deviations accrue only when observations lie outside this uncertainty envelope, with the magnitude of the departure determining the reduction in skill, which is then transformed into a skill score ($K' = \max(1-K,0)$) yielding a score that increases toward one as model-observation agreement increases.

2.4.3. Coefficient of determination

In this study, we also compute the coefficient of determination (R^2) with values closer to one indicating stronger agreement between synthetic model and observations.

2.4.4. χ^2

χ^2 is used as a goodness-of-fit measure that quantifies the root mean squared differences between observed and synthetic counts across bins, thereby emphasizing discrepancies that exceed expected Poisson sampling variability. This is transformed into a skill score ($\chi^{2\prime} = \max(1-0.5\ \chi^2,0)$) with values closer to 1 indicating better agreement between observations and model.

Further details on the formulation, implementation, and interpretation of these statistical tests are provided in Emanuel (2025). In our study, we computed RP (how often TCs of a given intensity are expected to occur) in order to derive exceedance probabilities and for the assessment of risk to energy infrastructure. For each 1^0x1^0 grid cell, we ranked all TC wind-speeds by intensity for both observations and each synthetic model. We then used an average annual TC rate of 12.73 storms per yr (1950–2023) to compute the wind speed corresponding to specific RPs (50 and 100 year). RPs are calculated as the inverse of the annual probability that an event of at least that specified wind speed will occur. Finally, damage probabilities for various energy infrastructure (distribution lines and distribution towers) are calculated using fragility curves provided by Bennett *et al* (2021). Fragility curves are functions that estimate the extent of infrastructure damage in response to a specified hazard. Thus, estimating the extent of TC associated damage depends on the fragility of infrastructure. By plugging a RP wind speed into the fragility curve, we directly obtain the probability of damage that a given energy infrastructure may likely experience over that period.

3. Results

3.1. Comparison of TC metrics

At the basin scale, a key challenge in characterizing TC activity is capturing the full climatology, as both storm frequency and intensity distributions shape the overall contribution to regional TC risk. At landfall, TCs present their greatest hazard, with storm intensity and the rate of intensification determining the potential magnitude of impacts along affected coastlines. To address both aspects, we evaluate the performance of synthetic TC models in representing TC activity at both the basin scale and at landfall. This ensures that both the overall climatology of TCs and the subset of TCs impacting coastlines are assessed, capturing differences in intensity, frequency, and intensification behavior across scales. As mentioned in the methods section, we follow the framework of Emanuel (2025) for evaluating the performance of synthetic TC models. Applying this framework to various synthetic TC models enables a systematic comparison with observations, and the results presented below demonstrate how well the models reproduce key features of basin-scale storm climatology and landfalling TC behavior. Intensity histograms are constructed with the annual exceedance threshold set to 12 for the basin and 3 for landfall, capturing the distribution of TC intensities and are shown in figure 1. The thresholds of 12 and 3 are derived from the average annual frequencies of North Atlantic synthetic TCs (12 yr^{-1}) and landfalling TCs (3 yr^{-1}) Intensity change histograms are also computed for basin and landfall events to characterize the statistical distribution of TC intensification and weakening over their lifetimes.

Figure 1(a) shows the basin-scale intensity histograms from observations and models. Overall, the models reproduce the observed distribution of North Atlantic TC intensities with reasonable fidelity (figure 1(a)). All three models recognize TS as the dominant category, followed by hurricanes and major hurricanes. RAFT appears to slightly overestimate, while CHAZ underestimates the frequency of mid-range intensity (approximately 70–120 knots) TCs. In contrast, MIT closely follows the observed frequency of TCs in all ranges, indicating better consistency with the available data. Across the three models, goodness-of-fit test scores are uniformly high. The χ^2 values are essentially unity (χ^2 : 0.99 -1), indicating that the range of intensities simulated in each model lies within the expected sampling uncertainty. The explained variance is likewise high (0.97–0.99), demonstrating that the modeled intensity distributions closely follow the observed histogram across various bins. *K*-test scores near unity (0.96–1) and Poisson log-likelihood test values ranges from 0.90–0.97 indicate excellent overall agreement in the representation of basin-scale TC intensity distribution. Next, we consider the intensity change histogram (figure 1(b)). At the basin-scale, the intensity change histogram of all three models agrees well with observations and exhibits broadly similar skill, with only modest quantitative differences (figure 1(b)). The χ^2 values are close to 1 and the variance explained spans 0.96–0.99, with the MIT model capturing the largest fraction of variability. Consistent with this, the *K*-test scores are 0.86,0.92 and 0.93, while the Poisson log-likelihood scores are 0.83, 0.91 and 0.91, respectively, for RAFT, MIT and CHAZ. The various statistical test scores for each TC metric are shown in table 1. Overall, these results indicate that all three synthetic TC models reproduce key features of intensity change reasonably well, although slight differences in performance are evident. These broad agreements with observations likely point to the

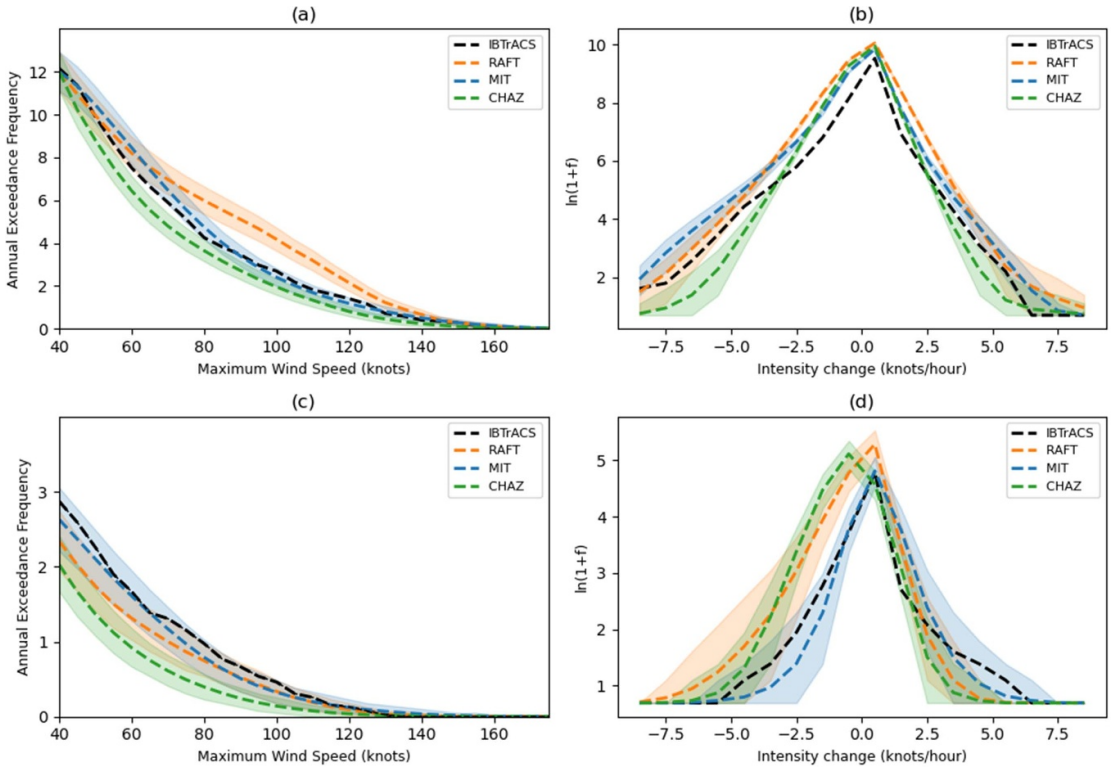

Figure 1. Annual exceedance frequency by lifetime maximum intensity (a) and the natural logarithm of the doublesided cumulative frequency of intensity change (b) for the North Atlantic Ocean. The latter accumulates the bin frequency going from left to right starting at the frequency peak, and separately from right to left starting at the peak. The shading shows the span between the 5th and 95th quantiles of a Poisson distribution based on the synthetic track count scaled to the historical track count. (c) and (d) same as (a) and (b) but for US landfalling TCs.

capacity of the models to replicate the distribution of TC intensities in the North Atlantic, including the dominant but weaker TSs as well as the rarer high-intensity storms-where societal and economic impacts are disproportionately concentrated, highlighting their value for risk and resilience planning.

One of the main challenges in characterizing the TC risk is the uncertainty surrounding how storms behave as they approach land, since they pose their greatest threat during landfall when both their intensity and the rate of intensification determine the scale of possible impacts. Therefore, we begin by examining the annual exceedance histograms of landfalling TCs along the US Coast in observations and synthetic TC datasets (figure 1(c)). In this study, following (Emanuel 2025), TCs that cross a coastline multiple times, only the first sea-to-land crossing is considered; land-to-sea crossings are ignored. This choice is made to avoid double-counting individual storms and to preserve independence across events, and is applied consistently to both the synthetic tracks and the IBTrACS dataset. However, this approach necessarily excludes subsequent landfalls (e.g. storms that re-emerge over water and make additional landfalls), and may therefore underestimate the total number of landfall occurrences and their cumulative impacts.

The χ^2 values (0.94-RAFT , 0.98-MIT, 0.95-CHAZ) indicate no significant deviation from observations and the high coefficients of determination (r^2: 0.85-RAFT, 0.94-MIT, 0.83-CHAZ) demonstrate that the synthetic models capture nearly all observed variance. The *K*-test scores show 0.92,0.98 and 0.88 and Poisson log-likelihood scores of 0.80, 0.93 and 0.79 for RAFT, MIT and CHAZ respectively. Similar results are obtained for the 6 h intensity change of US landfalling TCs (figure 1(d), table 1). Note that the 6 h window for landfalling TCs refers to the period just prior to landfall.

Figure 2 shows the annual cycle of TCs in the North Atlantic in observations and the three synthetic TC models. All models reproduce the dominant features of the North Atlantic seasonal cycle quite well (figure 2), especially the maximum in September and high activity extending from August to October. The models successfully capture this late-summer to early-fall peak, which is driven by high SSTs, reduced vertical wind shear, and favorable thermodynamic conditions in the North Atlantic. Also, they simulate very low counts in winter and early spring (January–May) and capture the rapid increase in TC activity beginning in June–July. However, small differences persist in the precise month-to-month distribution. For example, CHAZ and MIT slightly overestimate TC counts during June–August while

Table 1. Comparison of statistical metrics across different synthetic TC models. The skill scores (χ^2, R^2, K', and D') for CHAZ, MIT and RAFT are given for various TC metrics.

	D'			K'			R^2			χ^2		
Parameter	CHAZ	MIT	RAFT	CHAZ	MIT	RAFT	CHAZ	MIT	RAFT	CHAZ	MIT	RAFT
Intensity histogram	0.90	0.97	0.95	0.96	1.00	0.98	0.99	0.99	0.97	0.99	1.00	0.99
6 h intensity change histogram	0.91	0.91	0.83	0.93	0.92	0.86	0.96	0.99	0.98	1.00	0.99	0.99
Intensity histogram (Landfall)	0.79	0.93	0.80	0.88	0.98	0.92	0.83	0.94	0.85	0.95	0.98	0.94
6 h intensity change histogram (Landfall)	0.97	0.99	0.94	1.00	1.00	1.00	0.99	0.99	0.97	0.96	1.00	0.96
US East and Gulf gates	0.82	0.87	0.90	0.84	0.86	0.85	0.60	0.73	0.85	0.92	0.92	0.85
Annual cycle	0.90	0.94	0.99	0.94	0.96	1.00	0.95	0.97	1.00	0.97	0.98	1.00

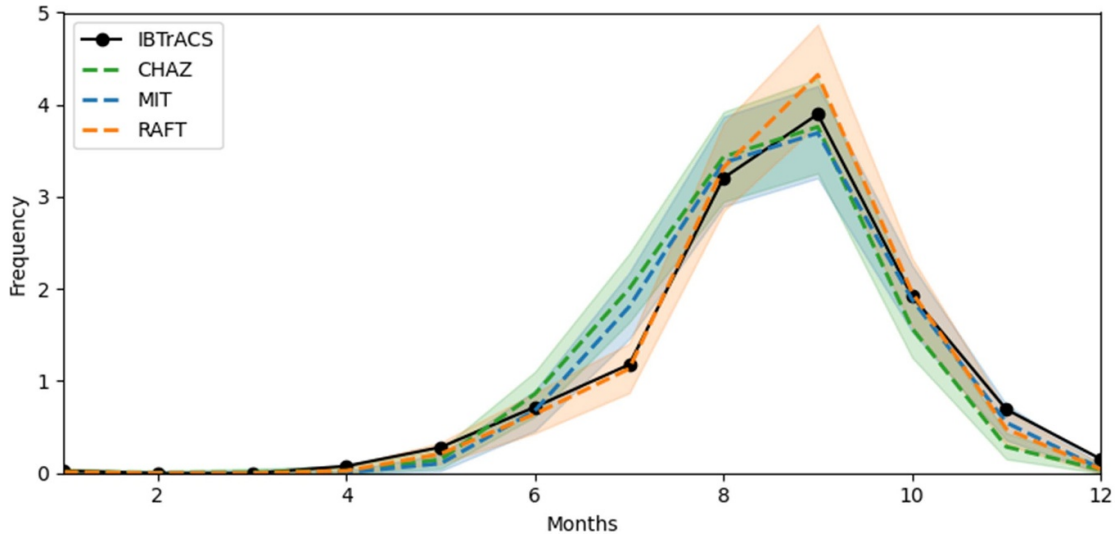

Figure 2. Monthly frequencies of TCs in the North Atlantic in observations, CHAZ, MIT and RAFT. The shading shows an estimate of historical sampling error based on the 5th and 95th percentiles of a Poisson distribution centered on corresponding synthetic TC model monthly counts.

RAFT produces more storms in September (figure 2). These differences are subtle and likely reflect sensitivities to modeled environmental controls such as vertical shear and potential intensity. Overall, the North Atlantic annual cycle results demonstrate that the synthetic TC models capture the primary seasonal forcing mechanisms governing TC activity. The goodness-of-fit statistics indicate strong agreement between the models and observations. The χ^2 values (1.00 for RAFT, 0.98 for MIT, and 0.97 for CHAZ) suggest no statistically significant deviation from the observed distribution. Similarly, the high coefficients of determination (r^2: 1.00 for RAFT, 0.97 for MIT, and 0.95 for CHAZ) demonstrate that the models reproduce nearly all of the observed variance. The K-test scores (1.00, 0.96, and 0.94 for RAFT, MIT, and CHAZ, respectively) further support the consistency between modeled and observed distributions. In addition, the Poisson log-likelihood scores (0.99 for RAFT, 0.94 for MIT, and 0.90 for CHAZ) indicate strong skill in capturing the frequency characteristics of the observed events. This suggests that large-scale environmental controls on TC seasonality are realistically represented in the driving fields, even though fine-scale monthly variability is slightly more difficult to reproduce. Overall, the North Atlantic annual cycle results demonstrate that hazard models capture the primary seasonal forcing mechanisms governing TC activity realistically.

3.2. Gate crossing

Having assessed how TC intensity and intensity change is distributed across the basin and at landfall, we next examine the number of TCs crossing the US coastal gates (figure 3), providing a more direct

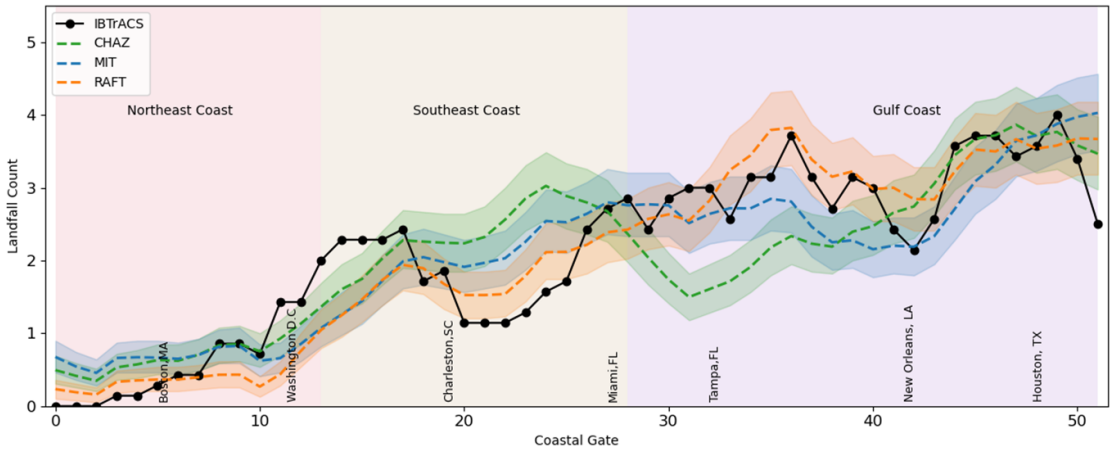

Figure 3. The number of TC crossings for each coastal gate of the US East and Gulf coastline in observations, CHAZ, MIT and RAFT. The shading shows the span between the 5th and 95th quantiles of a Poisson distribution based on the synthetic track count scaled to the historical track count. All distributions have been smoothed with a 7-gate running mean. Major cities-Boston, Washington DC, Charleston, Miami, Tampa, New Orleans and Houston-are marked for reference.

measure of the potential hazard to populated shorelines and regions of socioeconomic exposure. An illustration of the location of various gates is given in figure S1. In our study, Gates 1–12 represent the Northeast coast, gates 13–29 correspond to the Southeast Atlantic coast, and gates 30–52 represent the Gulf Coast. The observed gate-crossing counts vary markedly along the coastline reflecting preferred landfall locations shaped by large-scale steering flow, TC genesis locations and coastal geometry. In observations, it is clearly seen that the Gulf coast has more TC crossings, followed by the Southeast coast and then the Northeast Coast (figure 3). The synthetic models broadly follow this observed spatial pattern, reproducing the locations of high and low land crossings reasonably. All three models exhibit strong consistency with the observed crossing distribution, as indicated by K-test (0.85-RAFT, 0.84-CHAZ, 0.86-MIT) and χ^2 values (0.85-RAFT, 0.92- MIT and 0.92-CHAZ). However, departures from the 5th-95th percentile Poisson envelope occur in certain regions for MIT and CHAZ, especially between coastal gates 30–40, suggesting localized model biases rather than random noise or sampling limitations of the observational record. This underestimation of landfall counts in the Gulf coast likely arise from differences in simulated track curvature or regional steering effects near land. This is reflected in metrics sensitive to spatial variance: the coefficient of determination spans from 0.85 (RAFT) to 0.73 (MIT) and 0.60 (CHAZ). The Poisson log-likelihood tests values are 0.90 (RAFT), 0.87 (MIT) and 0.82 (CHAZ) accounting for TC count uncertainty.

Overall, these results show while all models are broadly consistent with observed landfall distributions at regional scales, their differences reflect how key processes controlling track variations in coastal regions are represented. Unlike basin-wide or cumulative distribution metrics, the gate-crossing metric is tightly linked to coastal exposure and impact making it the most physically relevant indicator of coastal hazard. Therefore, even models that perform similarly under basin scale goodness-of-fit tests can yield slightly different patterns of coastal risk, underscoring the importance of using coastal gate-crossing statistics to quantify and interpret model uncertainty in landfall hazard assessments.

Following the analysis of coastal gate crossings, we examine the coastal-to-basin ratio to quantify how the proportion of basin-wide storms that translate into coastal exposure varies across regions. The normalized coastal crossing-to-basin ratio (R'_k) is defined as the fraction of the number of gate crossing TCs along the US Coast compared to the total number of TCs generated in the basin, divided by the corresponding observed fraction (Emanuel 2025). It provides a simple measure of the proportion of TCs that make landfall along the coast compared to the observed climatology, which does not require any statistical testing. While a value of 1 indicates excellent agreement with observations, values greater than or less than 1 indicate systematic over- or under representation of coastal landfalls, respectively. In our analysis, all three models represent the coastal-to-basin ratio fairly well with values of 0.93 for RAFT, 1.41 for MIT and 0.86 for CHAZ. Of the three, the MIT model has the highest landfall rate, likely due to the westward bias in TC genesis locations (figure S2). Storms forming farther west in the North Atlantic are initiated closer to continental boundaries, reducing the travel distance required for landfall and increasing the probability of landfall compared to CHAZ and RAFT. Accurate representation

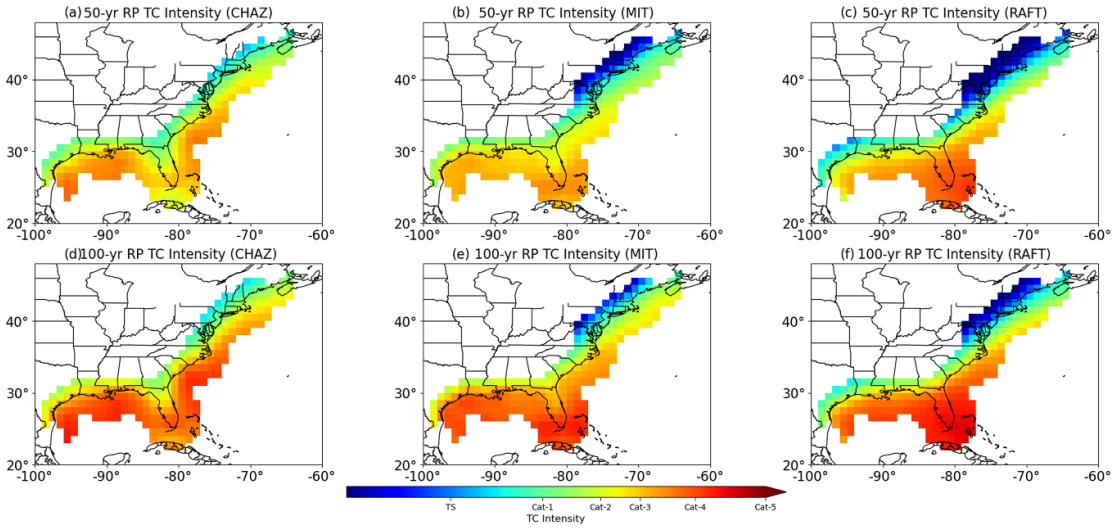

Figure 4. Return period TC intensity computed assuming 12.73 events per year, based on the exceedance probability distribution of tropical cyclones passing through each 1° × 1° grid. Spatial distribution of wind speeds for 50 year (upper panel, (a)–(c)) and 100 year (bottom panel, (d)–(f)) TCs across the four datasets, CHAZ, MIT AND RAFT along US Gulf Coast and East Coast.

of coastal landfalls is critical because these events disproportionately contribute to cumulative hazard, including storm surge, flood and wind damages.

3.3. Return Period

A central challenge in TC impact analysis is determining how often high-intensity TCs can be expected to strike a particular region. Adaptation planning depends on this knowledge, since most protective standards are tied to probabilities of exceedance, which correspond to RPs. Estimating these values requires hazard datasets that cover lifespans longer than the RP itself. However, observational records are generally too short and geographically inconsistent, especially for extreme TC events that occurred before the satellite era (Knapp *et al* 2010, Schreck III *et al* 2014). Synthetic TC tracks overcome this limitation by simulating a sufficiently large number of plausible TCs, allowing one to derive RP maps for any study region (Bloemendaal *et al* 2020a). Spatial maps of the intensity of the return-level TC provides a powerful way to visualize geographical patterns of risk. Because reliable TC observations are limited prior to the satellite era, our 50 year and 100 year RP analysis is restricted to intercomparison of the estimates derived from the three synthetic models. The synthetic TC model data spanning from 1980–2018, encompassing tens of thousands of synthetic TCs were analyzed which allows for a substantially larger sample size compared to the historical TC records, thereby improving the statistical reliability and robustness in the extreme value (EV) analysis. Extreme value analysis is commonly used in observational TC studies to fit a statistical distribution to storm intensity maxima and estimate return levels and RPs of rare events beyond the length of the historical record. This approach is illustrated in global TC hazard datasets such as Bloemendaal *et al* (2020a), where EV methods are applied at basin and coastal scales to derive wind-speed RPs from best-track and synthetic observations. In our study, however, we do not perform observational EV fitting; instead, we focus on comparative analysis within our modeling framework rather than statistical extrapolation from historical TC records.

Figure 4 shows the spatial distribution of 50 year and 100 year RP TC intensities along the US Gulf and East Coasts in the three models in 1° × 1° grid. RP estimates are inherently scale-dependent; when computed at the 1° × 1° grid-scale as in this study, they reflect more localized and frequent extremes, whereas aggregation over larger spatial scales tends to smooth variability and yield less frequent, longer-return-period events. All synthetic TC model simulations show that TCs with a 50 year RP (TC with a 2% annual probability of occurrence) along the US Southeastern and Gulf coasts typically attain intensities corresponding to Categories 2–3 (figures 4(a)–(c)). Along the Gulf Coast, from Texas through the Florida Panhandle, intensities reach major hurricane strength, reflecting the region's high susceptibility to major hurricane landfalls. Similarly, the Florida peninsula and Keys exhibit some of the most EVs, due to their exposure to warm surrounding waters and the convergence of TC tracks (Feng *et al* 2025). Moving northward, the Southeast Coast (Georgia, North and South Carolina) experiences slightly lower intensities, equivalent to Cat. 1–2 TCs (figures 4(a)–(c)). In contrast, the Northeast Coast shows a reduction in 50 year RP TC intensity, usually reaching up to Cat. 1 TCs. For 100 year RP (TCs with

an annual probability of 1% occurrence), a systematic increase in the intensity of TCs is observed in all coastal regions (figures 4(d)–(f)). The Gulf Coast, Florida Peninsula experiences upto Cat. 3–4 range TC intensities, while the Southeast Coast generally reaches up to Cat. 2–3. The Mid-Atlantic TC intensity reduces to approximately Cat. 1, while in the Northeast storms range between TS and Cat. 1 strength. All models agree broadly with this spatial pattern, albeit with some discrepancies in the magnitude of TC intensities (figures 4(a)–(f)). This large-scale consistency suggests a robust representation of coastal TC intensities. Previous studies using the STORM synthetic TC model show a strong coastal gradient in U.S. landfall risk, with higher-intensity TC impacts along the Florida-North Carolina coastline compared to the northeastern U.S., primarily due to prevailing storm tracks and SST-driven intensity differences (Bloemendaal *et al* 2020a). Consistent with these results, the models in this study also reproduce an enhanced risk along the southeastern U.S. relative to the northeastern coastline, indicating a robust spatial pattern across synthetic TC frameworks. However, notable regional discrepancies emerge and differences are evident in the magnitude and spatial extent of peak intensities within localized hotspots such as southern Florida, with RAFT and MIT producing higher intensities relative to CHAZ for both 50 and 100 year RP levels. These regional variations highlight the inter-model differences and uncertainties in representing rare, high-impact events. Again, these results indicate that even when TC hazard models are driven by the same environmental data, differences in their methodologies can lead to markedly different RP estimates. The RP intensity maps based on observations are provided in figure S3 for reference.

Our findings underscore the elevated long-term TC risk for the US Gulf and Southeast Coasts relative to northern regions. They also suggest that while the mid-Atlantic and upper-East Coast regions are relatively less affected by intense TCs, the socio-economic risks from TCs for these areas may not be underestimated. Finally, it is important to note that our observed estimates throughout this study are derived from IBTrACs and are therefore subject to substantial uncertainty due to the limited historical record. Hence, the observational analysis presented here should only be considered as an approximate reference point but not as a strict criterion to evaluate model performance.

3.4. Energy infrastructure risk

Now, we extend the RP analysis of the previous section to estimate the spatially-varying risk of TC-induced damages to energy infrastructure over US coastal regions. Specifically, the TC intensity corresponding to a certain RP was applied directly to fragility curves (figure 5 in Bennett *et al* (2021), figure S4), and thereby used to compute the damage probabilities for the infrastructure of interest, a framework that has been demonstrated previously with RAFT (Lipari *et al* 2024). Note that in this study we use a simplified representative fragility curve from Bennett *et al* (2021) for demonstration purposes, which may not fully capture the variability in structural responses across different energy infrastructure types. However, users may incorporate structure-specific fragility data to refine and improve the accuracy of the risk assessments. Here, the 100-yr RP TC intensity is used with electric power distribution lines and distribution towers along the Gulf and East Coasts. The goal of this analysis is to demonstrate how estimates of TC intensity may be translated into infrastructure risk, and how uncertainty in the former can lead to that in the latter. It is important to note that this method does not explicitly account for the spatial distribution of assets/infrastructure. In other words, the fragility curve—TC intensity relationship was applied under the assumption that distribution lines and towers are present across our entire domain.

Figure 5 shows the spatial distribution of damage probabilities for distribution lines and towers based on 100 year RP TC intensities obtained from RAFT, CHAZ and MIT. The 100 year RP is used because it aligns with the typical lifespan of the infrastructure, ensuring resilience against extreme events that are likely to occur at least once during its intended life span. In all models, the Florida peninsula emerges as the region of highest risk for distribution lines and towers, followed by the Northern Gulf Coast, both of which are historically known for their exposure to intense landfalls, as seen previously (figure 4). In comparison, the Southeast Coast experiences a lower, although still notable, likelihood of damages relative to these two regions (figure 5). This agreement in identifying geographic hotspots of TC risk adds value and credibility to the analysis, as it reflects the convergence of multiple independent models in the same high-risk areas. In all models, distribution lines are shown to be more vulnerable to TC-risk than towers, with consistently higher probabilities of damage. This reflects the inherent fragility of overhead lines relative to tower structures and carries important operational implications. Our results show that for Florida peninsula, the magnitude of damage probabilities of distribution lines (distribution towers) ranges from 40% to 100% (30%–70%) depending on the choice of model, reflecting considerable inter-model spread. Following Florida peninsula, the New Orleans region displays a high probability of TC-related damages to distribution lines and towers ranging from 30% to 50%. In this region, MIT and CHAZ indicate a higher probability of damage compared to RAFT (figures 5(c) and (d)). These

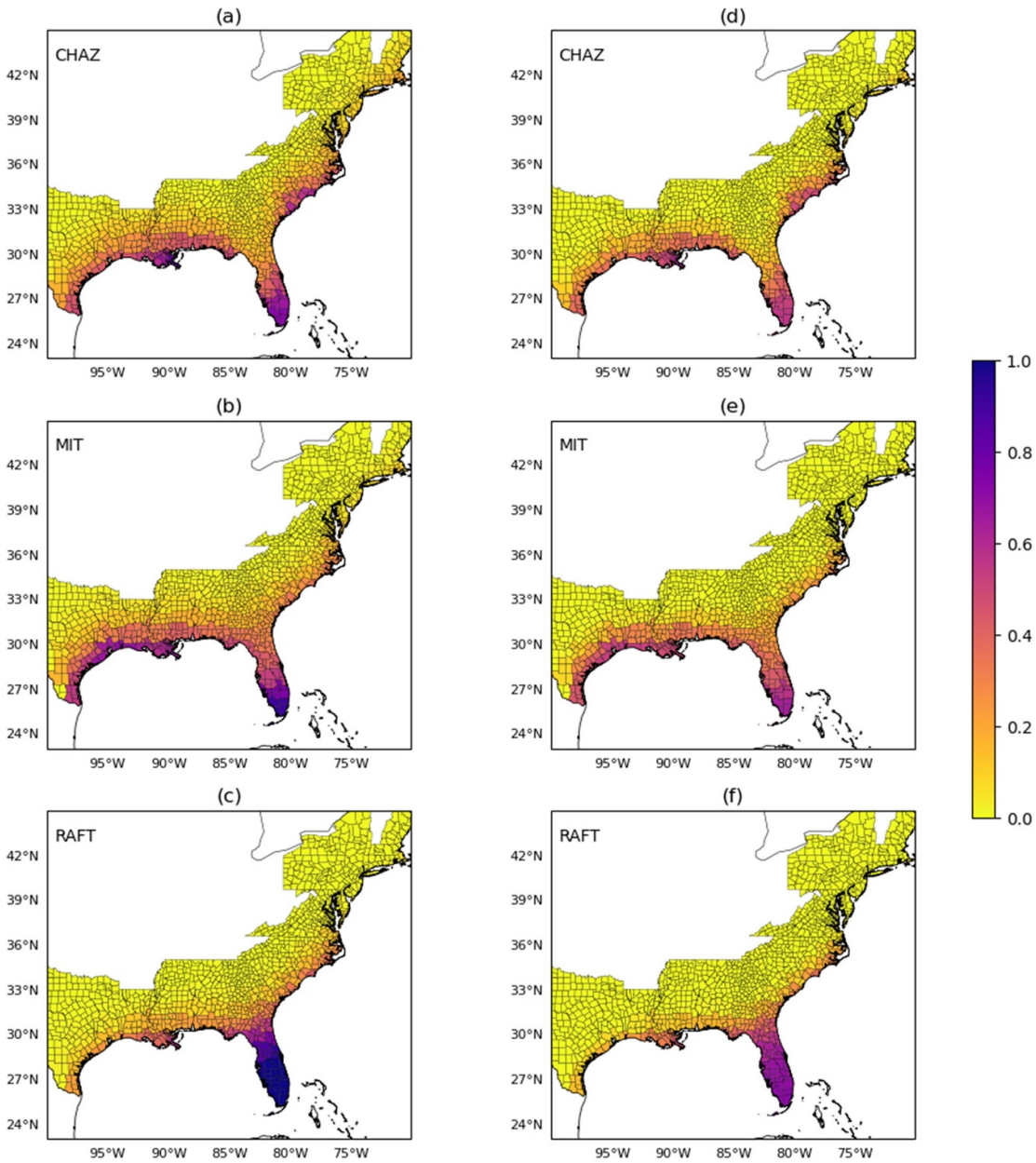

Figure 5. Spatial distribution of damage probability of distribution lines (a)–(c) and towers (d)–(f) based on 100-year RP TC intensity across the models (a), (d) CHAZ, (b), (e) MIT and (c), (f) RAFT along US Gulf Coast and East Coast.

differences indicate the influence of model-specific methodologies in individual models, particularly in the simulation of TC intensity. Similar results are obtained when computing expected decadal failures per asset, representing the spatial distribution of expected average decadal damages to distribution lines and towers from TCs, as shown in figure S5. The three hazard models yield different 100 year RP wind speeds; differences on the order of several knots, which, when passed through the same fragility curve, translate into damage probabilities of varying magnitudes. The fragility curve has a steep slope in the mid-intensity range; therefore, even modest shifts in TC intensity can be amplified into large differences in damage probability (Bennett *et al* 2021, Lipari *et al* 2024). Although this introduces uncertainty, the spread between models should be viewed as a valuable measure of confidence bounds rather than a limitation.

4. Discussion

A useful way to interpret inter-model differences is to relate them directly to the underlying modeling techniques, as these strongly shape how TC behavior is represented. The three modeling frameworks

used in this study—statistical-dynamical stochastic (CHAZ), statistical-dynamical deterministic (MIT), and statistical-dynamical machine learning (RAFT)—represent complementary approaches to estimating TC climatology and extremes. Statistical-dynamical models combine environmental fields (often from reanalysis or climate models) with simplified or empirical representations of TC evolution. This approach retains some physical realism while being computationally efficient, enabling large synthetic event sets for robust RP estimation. Their outcomes, however, depend strongly on the quality of the driving fields and the assumptions embedded in the statistical relationships, which can smooth extremes or constrain variability. Statistical-dynamical-ML models extend this framework by using machine learning to learn complex, and potentially nonlinear, relationships from data. This can improve the representation of processes such as intensity evolution or regional variability, especially where traditional parameterizations are insufficient. At the same time, these models may be sensitive to training data limitations and can exhibit reduced interpretability, which introduces additional uncertainty in extrapolating to rare extremes or changing environmental conditions. In this study, these differences in formulation translate into variations in estimated intensities and RPs. The statistical-dynamical approaches provide stable and consistent estimates across regions due to their constrained structure, while the inclusion of ML can introduce greater flexibility in capturing regional features.

A key place where these methodological differences become evident is in the estimation of RP intensities, since these depend strongly on how each model represents the tail of the intensity distribution. Each model incorporates the same large-scale environmental controls but differs in how that information is utilized to simulate storm genesis and intensity evolution, leading to a range of plausible return-period estimates. For example, in Miami, the 50 year RP intensity increases across the models, from 96.34 knots (CHAZ) to 103.36 knots (MIT) and 113.77 knots (RAFT). In contrast, in New Orleans the pattern reverses, with CHAZ giving 94.47 knots, MIT giving 89.67 knots, and RAFT giving 79.99 knots. This spatially variability in RP estimates shows that differences between models are not systematic but region-dependent. In CHAZ, estimates of RP intensities are obtained from large ensembles of synthetic storms generated using empirically derived relationships. The advantage here is sampling: the very large number of events improves the stability of EV estimates and reduces sampling uncertainty at longer RPs. However, because intensity evolution is governed by simplified parameterizations, the upper tail may be somewhat smoothed or capped by the assumptions in the model. This can lead to slightly lower RP intensities at high RPs, especially in regions where extreme intensification processes are not fully captured. MIT relies on CHIPS, which is a physics-based TC intensity model. CHIPS simulates intensity evolution by resolving key thermodynamic processes in an idealized but dynamically consistent way. Because of this, intensity changes are tightly constrained by environmental conditions such as SST. Since CHIPS does not include stochastic perturbations in intensity evolution, extreme intensities can only arise if the environmental conditions strongly support them. This tends to produce RP estimates that are physically consistent and regionally realistic, however the lack of stochastic variability narrows the distribution. Because each storm's intensity follows a deterministic pathway given the inputs, the spread of intensities is reduced. This leads to smoother and more stable RP curves, but also means that rare, high-end intensification pathways might be under-represented. RAFT model approach introduces an additional layer of flexibility in how intensities are modeled. By learning nonlinear relationships from data, ML components can better capture regional variability and subtle environmental dependencies, which can influence RP estimates. In some regions, this can improve the realism of moderate-to-high RP intensities. However, for very rare events (e.g. long RPs), the estimates depend on how well the extremes are captured by the training data. If extremes are underrepresented, the model may underpredict the highest RP intensities or produce less stable tail behavior.

Overall, the spread in RP intensities highlights a trade-off between physical consistency, sampling robustness, and flexibility in representing extreme events, which is central to interpreting risk estimates. These results indicate that all three models produce physically reasonable and consistent estimates of extreme wind speeds, but differ in how they represent regional variability and the tail of the intensity distribution. Overall, the spread across models should be interpreted as an uncertainty envelope rather than disagreement: the lower values represent baseline risk, while the higher values represent a plausible upper-bound scenario. This range is particularly valuable for risk assessment, as it avoids over-reliance on any single model and instead captures the spectrum of potential outcomes. Importantly, all three models remain physically reasonable and mutually consistent in magnitude, with differences primarily reflecting uncertainty in how extremes are extrapolated rather than disagreement on the underlying climatology. Taken together, they provide a complementary and robust ensemble that captures uncertainty in TC intensity estimates and supports more reliable risk assessment than any single modelling approach alone.

5. Summary

In this study, we compared three synthetic TC models to assess the inter-model uncertainty in representing TC parameters relevant for risk assessment. While there are some similarities in the methodology used in the three models, there are also certain differences among the models, notably in the approach used to generate TC intensities. CHAZ relies on the stochastic auto regressive intensity model, which links large-scale environmental parameters to the development of TCs (Lee *et al* 2018). MIT uses CHIPS, a physics-based coupled model that accounts for air–sea interactions (Emanuel *et al* 2006). RAFT leverages deep neural networks, allowing it to learn complex relationships between environmental conditions and TC intensification from training data (Xu *et al* 2021, 2024). These contrasting approaches highlight the diversity in TC intensity simulations, each with its own strengths and limitations with respect to risk assessment. All three models are able to reproduce the observed distribution of TC intensities, including weak and major hurricane events. Our study shows that the frequency of rapid intensification events is also well simulated in all models (figure 1). This agreement among models provides confidence in their application for risk assessments, even though uncertainties remain at the tails of the distribution where the rarest and most extreme events occur.

In observations, there is a substantial degree of uncertainty (Meiler *et al* 2022, Romero *et al* 2025), making it difficult to draw definitive conclusions when based on them alone. On the other hand, the comparison between synthetic TC datasets and IBTrACS reflects the complexities of simulating TC characteristics and limitations of both data sources. All models produce similar intensities corresponding to the 50 year and 100 year RPs along Gulf and Florida Coasts, leading to confidence in their projections of extreme TCs.

The main goal of our study is to understand the similarities and differences among various synthetic TC models rather than evaluating their performance on certain TC metrics. These results reflect the intricate nature of interactions between storms and their environment in each modeling approach, making it difficult for any one model to capture every aspect of TC risk. Rather, our study underscores the idea that each modeling approach has its own distinct strengths and limitations. The inter-model comparison highlights that no single modeling framework is universally optimal; rather, model suitability is strongly application-dependent. Within the suite of approaches considered here-CHAZ, MIT and RAFT-each offers distinct advantages and limitations that make them complementary rather than interchangeable. These models provide a useful balance between physical realism and computational efficiency, making them particularly well suited for large-ensemble probabilistic risk assessments and climate sensitivity analyses. Both CHAZ and MIT frameworks offer interpretability and consistency in process representation, making it valuable for controlled experiments and diagnostic studies where isolating physical mechanisms is important. In contrast, the RAFT model enhances flexibility in capturing nonlinear relationships in environmental predictors and storm behavior, but its performance depends strongly on training data coverage and generalization outside observations.

Across model components, all models follow similar methodology (beta-advection) and use the same environmental forcing (ERA5) to simulate TC motion, which provides a common dynamical framework for track evolution. Because all three models use ERA5-derived large-scale winds within a beta-advection framework for track evolution, the dynamical basis for storm motion is shared across models. A quantitative comparison of track climatology—including mean track density, recurvature latitude, and landfall angle distributions—is beyond the scope of the present study and is identified as a priority for future work. For the purposes of this intercomparison, inter-model differences in genesis location and intensity are treated as the primary sources of divergence in risk estimates, consistent with the analysis presented here. As a result, we presume that track behavior may appear similar across the three models, although we do not quantify this agreement here. In contrast, large inter-model spread emerges in TC genesis locations (figure S2), where seeding strategies and environmental conditions can lead to notable discrepancies in where TCs form. For example, in the MIT model, enhanced TC genesis occurs in the Caribbean and US Gulf compared to observations, which may lead to more frequent landfall events as mentioned earlier. Even for TC intensity, additional discrepancies may arise because the intensity components differ substantially across models. Variations in how air–sea interaction processes, inner-core processes and thermodynamic limits are represented can significantly contribute to the differences in TC intensity. Addressing these uncertainties will require additional sensitivity experiments to better isolate the contribution of each component to the overall inter-model differences. Taken together, this suggests that targeted improvements in genesis and intensity modules would yield the greatest gains in model reliability. Overall, a multi-model strategy that leverages statistical-dynamical stochastic, deterministic, and machine learning frameworks in a complementary manner offers the most robust pathway for risk assessment, while mitigating the individual limitations of each approach.

Current analyses focus exclusively on wind-driven impacts, despite the fact that storm surges and rainfall-induced inland flooding are frequently responsible for catastrophic losses (Rappaport 2014). Since the TC-risk metrics we considered in our study do not necessarily reflect the surge and rainfall effects, they tend to understate the losses associated with flood-dominated events that have relatively modest TC intensity. Incorporating explicit storm-surge and rainfall-related metrics in future synthetic TC model assessment will be essential to portray a more accurate and complete picture of TC risk.

The uncertainties discussed in this study have direct implications for future TC risk and adaptation assessments. For instance, biases in TC intensity and frequency can distort hazard projections and infrastructure design standards. Underestimated intensities may result in under-engineered assets vulnerable to extreme TCs, while overestimation may cause unnecessary economic burden. Spatial inconsistencies may also misidentify high-risk zones, leading to inaccurate planning and insurance pricing. Given these challenges, ensemble modeling approaches that combine multiple models to balance individual weaknesses offer a more reliable pathway for hazard estimation.

In the context of energy infrastructure risk, the scarcity of detailed fragility assessments combined with the uncertainty in future TC characteristics underscores the urgent need to address these knowledge gaps and implement robust mitigation strategies. Protecting critical energy assets, from power generation facilities to transmission networks and substations, against the increasing threat of extreme TCs requires measures such as rigorous site selection protocols, stronger structural design standards, and resilient-driven operational planning. These steps are essential to maintain reliable power delivery, reduce system vulnerabilities, and ensure the long-term sustainability of the energy sector in a non-stationary environment.

Acknowledgments

This research was supported by the U.S. DOE's Office of Critical Minerals and Energy Innovation (CMEI) Integrated Energy System Office (IESO). The research was also funded by the U.S. Department of Energy (DOE) Office of Science Biological and Environmental Research (BER) as part of the RGMA and MultiSector Dynamics (MSD) program areas through the collaborative, multiprogram Integrated Coastal Modeling (ICoM) project. The research used computational resources from the National Energy Research Scientific Computing Center (NERSC), a U.S. DOE User Facility supported by the Office of Science under Contract DE-AC02-05CH11231. The Pacific Northwest National Laboratory is operated for U.S. DOE by Battelle Memorial Institute under Contract DE-AC05-76RL01830.

Data availability statement

RAFT-simulated TCs forced by ERA5 reanalysis are freely accessible at https://zenodo.org/records/18879988. TC data based on the MIT model is available for non-profit research only upon signing a data license. For further details, please reach out to info@windrisktech.com. Similarly, TC data from CHAZ is available for research purposes upon request to co-author Lee.

The data that support the findings of this study are openly available at the following URL/DOI:. https://zenodo.org/records/18879988 (John *et al* 2026).

Supplementary Information available at: https://doi.org/10.1088/2752-5295/ae7202/data1.

ORCID iDs

Karthik Balaguru 0000-0003-0181-2687
Sha Feng 0000-0002-2376-0868
Kerry Emanuel 0000-0002-2066-2082
Chia-Ying Lee 0000-0002-1644-375X
Julian Rice 0000-0003-3253-7833
L Ruby Leung 0000-0002-3221-9467
Larry Berg 0000-0002-3362-9492

References

Arkema K, Bailey A, Compeán R G, Fernandez P M and Reguero B 2023 Modeling tropical cyclone risk while accounting for climate change and natural infrastructure in the Caribbean (https://doi.org/10.18235/0004966)

Balaguru K, Xu W, Chang C-C, Leung L R, Judi D R, Hagos S M, Wehner M F, Kossin J P and Ting M 2023 Increased us coastal hurricane risk under climate change *Sci. Adv.* **9** eadf0259

Bender M A, Knutson T R, Tuleya R E, Sirutis J J, Vecchi G A, Garner S T and Held I M 2010 Modeled impact of anthropogenic warming on the frequency of intense Atlantic hurricanes *Science* **327** 454–8

Bennett J A, Trevisan C N, DeCarolis J F, Ortiz-García C, Pérez-Lugo M, Etienne B T and Clarens A F 2021 Extending energy system modelling to include extreme weather risks and application to hurricane events in Puerto Rico *Nat. Energy* **6** 240–9

Beven J L I, Berg R and Hagen A 2019 Hurricane Michael (al 14/2018) *Tropical Cyclone Report* (National Hurricane Center) (available at: www.nhc.noaa.gov/data/tcr)

Bhatia K T, Vecchi G A, Knutson T R, Murakami H, Kossin J, Dixon K W and Whitlock C E 2019 Recent increases in tropical cyclone intensification rates *Nat. Commun.* **10** 635

Bloemendaal N *et al* 2022 A globally consistent local-scale assessment of future tropical cyclone risk *Sci. Adv.* **8** eabm8438

Bloemendaal N, de Moel H, Muis S, Haigh I D and Aerts J C 2020 Estimation of global tropical cyclone wind speed probabilities using the storm dataset *Sci. Data* **7** 377

Bloemendaal N, Haigh I, deMoel H, Muis S, Haarsma R and Aerts J 2020 Generation of a global synthetic tropical cyclone hazard dataset using storm *Sci. Data* **7** 40

Bourdin S, Fromang S, Caubel A, Ghattas J, Meurdesoif Y and Dubos T 2024 Tropical cyclones in global high-resolution simulations using the IPSL model *Clim. Dyn.* **62** 4343–68

Camargo S J, Tippett M K, Sobel A H, Vecchi G A and Zhao M 2014 Testing the performance of tropical cyclone genesis indices in future climates using the HiRAM model *J. Clim.* **27** 9171–96

Cerveny R S *et al* 2017 WMO assessment of weather and climate mortality extremes: lightning, tropical cyclones, tornadoes and hail *Weather Clim. Soc.* **9** 487–97

DeMaria M and Kaplan J 1994 A statistical hurricane intensity prediction scheme (ships) for the Atlantic basin *Weather and Forecasting* **9** 209–20

DeMaria M and Kaplan J 1999 An updated statistical hurricane intensity prediction scheme (ships) for the Atlantic and eastern north pacific basins *Weather and Forecasting* **14** 326–37

DeMaria M, Mainelli M, Shay L K, Knaff J A and Kaplan J 2005 Further improvements to the statistical hurricane intensity prediction scheme (ships) *Weather Forecast.* **20** 531–43

Eberenz S, Lüthi S and Bresch D N 2020 Regional tropical cyclone impact functions for globally consistent risk assessments *Nat. Haz. Earth Syst. Sci. Discuss.* **21** 1–29

Elsner J B, Kossin J P and Jagger T H 2008 The increasing intensity of the strongest tropical cyclones *Nature* **455** 92–95

Emanuel K 2006 Climate and tropical cyclone activity: a new model downscaling approach *J. Clim.* **19** 4797–802

Emanuel K 2025 A framework for testing tropical cyclone hazard models *Environ. Res. Clim.* **4** 025011

Emanuel K, DesAutels C, Holloway C and Korty R 2004 Environmental control of tropical cyclone intensity *J. Atmos. Sci.* **61** 843–58

Emanuel K, Ravela S, Vivant E and Risi C 2006 A statistical deterministic approach to hurricane risk assessment *Bull. Am. Meteorol. Soc.* **87** 299–314

Gahtan J, Knapp K R, Schreck C J, Diamond H J, Kossin J P and Kruk M C 2024 International best track archive for climate stewardship (IBTrACS) project, version 4r01 (NOAA National Centers for Environmental Information)

Geiger T, Frieler K and Levermann A 2016 High-income does not protect against hurricane losses *Environ. Res. Lett.* **11** 084012

Gettelman A, Bresch D N, Chen C C, Truesdale J E and Bacmeister J T 2018 Projections of future tropical cyclone damage with a high-resolution global climate model *Clim. Change* **146** 575–85

Haigh I D, MacPherson L R, Mason M S, Wijeratne E, Pattiaratchi C B, Crompton R P and George S 2014 Estimating present day extreme water level exceedance probabilities around the coastline of Australia: tropical cyclone-induced storm surges *Clim. Dyn.* **42** 139–57

Hall T M and Jewson S 2007 Statistical modelling of north Atlantic tropical cyclone tracks *Tellus A* **59** 486–98

Hardy T A, McConochie J D and Mason L B 2003 Modeling tropical cyclone wave population of the great barrier reef *J. Waterway Port Coast. Ocean Eng.* **129** 104–13

Hersbach H *et al* 2020 The era5 global reanalysis *Q. J. R. Meteorol. Soc.* **146** 1999–2049

Hock T F and Franklin J L 1999 The NCAR GPS dropwindsonde *Bull. Am. Meteorol. Soc.* **80** 407–20

Holland G and Bruyère C L 2014 Recent intense hurricane response to global climate change *Clim. Dyn.* **42** 617–27

John E B*et al* 2026 Environmental Research: Climate [dataset] *Zenodo* (available at: https://zenodo.org/records/18879988)

Klotzbach P J, Bowen S G, Pielke R and Bell M 2018 Continental us hurricane landfall frequency and associated damage: observations and future risks *Bull. Am. Meteorol. Soc.* **99** 1359–76

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

Knutson T R, Sirutis J J, Bender M A, Tuleya R E and Schenkel B A 2022 Dynamical downscaling projections of late twenty-first-century us landfalling hurricane activity *Clim. Change* **171** 28

Knutson T R, Sirutis J J, Garner S T, Vecchi G A and Held I M 2008 Simulated reduction in Atlantic hurricane frequency under twenty-first-century warming conditions *Nat. Geosci.* **1** 359–64

Kossin J P, Knapp K R, Olander T L and Velden C S 2020 Global increase in major tropical cyclone exceedance probability over the past four decades *Proc. Natl Acad. Sci.* **117** 11975–80

Landsea C W and Franklin J L 2013 Atlantic hurricane database uncertainty and presentation of a new database format *Mon. Weather Rev.* **141** 3576–92

Lee C-Y, Tippett M K, Camargo S J and Sobel A H 2015 Probabilistic multiple linear regression modeling for tropical cyclone intensity *Mon. Weather Rev.* **143** 933–54

Lee C-Y, Tippett M K, Sobel A H and Camargo S J 2016 Autoregressive modeling for tropical cyclone intensity climatology *J. Clim.* **29** 7815–30

Lee C-Y, Tippett M K, Sobel A H and Camargo S J 2018 An environmentally forced tropical cyclone hazard model *J. Adv. Modeling Earth Syst.* **10** 223–41

Lipari S, Balaguru K, Rice J, Feng S, Xu W, Berg L K and Judi D 2024 Amplified threat of tropical cyclones to us offshore wind energy in a changing climate *Commun. Earth Environ.* **5** 755
Marks D G 1992 The beta and advection model for hurricane track forecasting
Meiler S, Vogt T, Bloemendaal N, Ciullo A, Lee C-Y, Camargo S J, Emanuel K and Bresch D N 2022 Intercomparison of regional loss estimates from global synthetic tropical cyclone models *Nat. Commun.* **13** 6156
Muller J, Mooney K, Bowen S G, Klotzbach P J, Martin T, Philp T J, Dhruvkumar B, Dixon R S and Girimurugan S B 2025 Normalized hurricane damage in the united states: 1900–2022 *Bull. Am. Meteorol. Soc.* **106** E51–67
Murakami H, Delworth T L, Cooke W F, Zhao M, Xiang B and Hsu P-C 2020 Detected climatic change in global distribution of tropical cyclones *Proc. Natl Acad. Sci.* **117** 10706–14
Murakami H and Sugi M 2010 Effect of model resolution on tropical cyclone climate projections *Sola* **6** 73–76
Nederhoff K, Hoek J, Leijnse T, van Ormondt M, Caires S and Giardino A 2021 Simulating synthetic tropical cyclone tracks for statistically reliable wind and pressure estimations *Nat. Haz. Earth Syst. Sci.* **21** 861–78
Neumann C J 1999 Tropical cyclones of the North Atlantic Ocean 1871–998
NOAA Office for Coastal Management 2025 Hurricane costs (available at: https://coast.noaa.gov/states/fast-facts/hurricane-costs.html) (Accessed 22 September 2025)
Noy I 2016 The socio-economics of cyclones *Nat. Clim. Change* **6** 343–5
Patricola C M and Wehner M F 2018 Anthropogenic influences on major tropical cyclone events *Nature* **563** 339–46
Peduzzi P, Chatenoux B, Dao H, De Bono A, Herold C, Kossin J, Mouton F and Nordbeck O 2012 Global trends in tropical cyclone risk *Nat. Clim. Change* **2** 289–94
Rappaport E N 2014 Fatalities in the united states from Atlantic Tropical cyclones: new data and interpretation *Bull. Am. Meteorol. Soc.* **95** 341–6
Romero D, Appendini C M, Emanuel K, Lee C-Y, Nederhoff K, Bloemendaal N, Ruiz-Salcines P, Vigh J and Domínguez C 2025 Assessment of synthetic tropical cyclones in the north Atlantic basin *Atmos. Res.* **327** 108404
Schreck C J, Knapp K R and Kossin J P 2014 The impact of best track discrepancies on global tropical cyclone climatologies using IBTrACS *Mon. Weather Rev.* **142** 3881–99
Shaevitz D A *et al* 2014 Characteristics of tropical cyclones in high-resolution models in the present climate *J. Adv. Modeling Earth Syst.* **6** 1154–72
Tippett M K, Camargo S J and Sobel A H 2011 A Poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis *J. Clim.* **24** 2335–57
Vecchi G A and Knutson T R 2011 Estimating annual numbers of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density *J. Clim.* **24** 1736–46
Vickery P J, Skerlj P and Twisdale L A 2000 Simulation of hurricane risk in the US using empirical track model *J. Struct. Eng.* **126** 1222–37
Walsh K J *et al* 2016 Tropical cyclones and climate change *Wiley Interdiscip. Rev.- Clim. Change* **7** 65–89
Webster P J, Holland G J, Curry J A and Chang H-R 2005 Changes in tropical cyclone number, duration and intensity in a warming environment *Science* **309** 1844–6
Wehner M F, Reed K A, Loring B, Stone D and Krishnan H 2018 Changes in tropical cyclones under stabilized 1.5 and 2.0 c global warming scenarios as simulated by the community atmospheric model under the HAPPI protocols *Earth Syst. Dyn.* **9** 187–95
Weinkle J, Maue R and Pielke R 2012 Historical global tropical cyclone landfalls *J. Clim.* **25** 4729–35
Willoughby H, Hernandez J and Pinnock A 2024 Trends in us Atlantic tropical cyclone damage, 1900–2022 *J. Appl. Meteorol. Climatol.* **63** 1499–510
Wu L and Wang B 2004 Assessing impacts of global warming on tropical cyclone tracks *J. Clim.* **17** 1686–98
Xu W, Balaguru K, August A, Lalo N, Hodas N, DeMaria M and Judi D 2021 Deep learning experiments for tropical cyclone intensity forecasts *Weather Forecast.* **36** 1453–70
Xu W, Balaguru K, Judi D R, Rice J, Leung L R and Lipari S 2024 A north Atlantic synthetic tropical cyclone track, intensity and rainfall dataset *Sci. Data* **11** 130
Yonekura E and Hall T M 2011 A statistical model of tropical cyclone tracks in the western north pacific with enso-dependent cyclogenesis *J. Appl. Meteorol. Climatol.* **50** 1725–39
Zhao H, Wu L and Zhou W 2009 Observational relationship of climatologic beta drift with large-scale environmental flows *Geophys. Res. Lett.* **36** L18809