ENVIRONMENTAL RESEARCH
CLIMATE

PURPOSE-LED PUBLISHING™ Part of the ENVIRONMENTAL RESEARCH SERIES

PAPER • OPEN ACCESS

A framework for testing tropical cyclone hazard models

To cite this article: Kerry Emanuel 2025 *Environ. Res.: Climate* **4** 025011

IOP Publishing Environ. Res.: Climate 4 (2025) 025011 https://doi.org/10.1088/2752-5295/add9ba

ENVIRONMENTAL RESEARCH
CLIMATE

OPEN ACCESS

RECEIVED
20 January 2025

REVISED
5 May 2025

ACCEPTED FOR PUBLICATION
16 May 2025

PUBLISHED
27 May 2025

PAPER

A framework for testing tropical cyclone hazard models

Kerry Emanuel

Lorenz Center, Massachusetts Institute of Technology, Cambridge, MA, 02139, United States of America

E-mail: emanuel@mit.edu

Keywords: tropical cyclone, hazards, models

Abstract

The field of tropical cyclone (TC) hazard modeling is in the midst of a rapid transition away from purely statistical techniques based on historical events to more physics-based approaches that can better account for the climate change that has already occurred and that will continue into the future. As such models proliferate, it is important to test them against TC observations and compare them to each other. Here a framework is proposed for comparing hazard model predictions to metrics derived strictly from historical ('best-track') observations, though in the future metrics more directly related to damage should be incorporated. While a specific set of metrics and statistical tests is presented here for illustrative purposes, it is hoped that the TC hazard community along with industry and government interests will convene to agree on a set that will serve as a benchmark for testing and intercomparing TC hazard models. This should allow for more rapid refinement and improvement of such models.

1. Introduction

Much of the damage from weather hazards arises from events that are frequent enough to matter but not so frequent that societies are well adapted to them. These events lie in the tails of the weather hazard probability distributions, but not so far out in the tail that their low frequency makes them largely irrelevant. Hazard modelers are faced with the paradox that the most damaging events are poorly represented in the historical record but of great importance to risk assessment. To make matters worse (and arguably *much* worse), these events are often disproportionately affected by climate change, rendering the sparse historical record of them largely irrelevant to today's weather hazard risk.

It would be difficult to overstate the deleterious effects this circumstance has already had on civilization. Because weather hazards like storms, droughts, heatwaves, and wildfires are not correctly represented in today's risk models, there is little incentive to adapt to the changed risk landscape. Insurance, which in a well-functioning society should communicate risk through pricing, is in most states[1] prevented from doing so by premium rate restrictions imposed by regulators, who were put in place to ensure solvency but are now having the opposite effect, and by the provision of federal flood insurance and disaster relief, which strongly subsidize risk in more dangerous places. These and other policies have arguably caused or exacerbated a strong migration away from less risky locations and into the riskiest places (e.g. Peralta and Scott 2019), leading to a well-documented increase in weather-related damages (Muller *et al* 2025). The collision between this policy-driven migration and the climate change-induced increase in hazardous weather is a recipe for disaster.

To break this cycle, the climate and weather communities must do a better job quantifying weather hazards, both in today's climate (as opposed to the climate of the last ∼100 years) and in future climates. The research community has been stepping up to this challenge and there are now several publicly available weather hazard models for such diverse phenomena as tropical cyclones (TCs) and wildfires (Lee *et al* 2018, Bloemendaal *et al* 2020, Jing and Lin 2020, Oliveira *et al* 2021, Lin *et al* 2023).

[1] Among all the rate filings by reasonably large, U.S. multi-state insurers, fewer than 10% received the rate increases they requested and there is poor correlation between rates and risk in states with the strongest restrictions (Oh *et al* 2021).

As such models develop and proliferate, it is important to be able to compare them to each other and evaluate them against recent historical records of the hazard. There have been a few attempts to do this so far (e.g. Meiler *et al* 2022, 2023) and one hopes there will be further such comparisons. The purpose of this paper is to catalyze a movement toward an agreed-upon set of metrics and statistical tests that TC model developers might routinely use to gauge the performance of their models in comparison to historical records and to other models. This follows one of the recommendations from the 2023 report of the President's Council of Advisors on Science and Technology (PCAST 2023). A preliminary set of metrics and statistical tests is here proposed but should be regarded as 'for instance' and not as a proposal for a final set of tests. We apply it to the output of a single dynamical-statistical TC downscaling for illustrative purposes; in principle, it could be run with any set of TC tracks developed as part of a hazard model.

We confine ourselves to strictly meteorological comparisons of output from a TC hazard model to historical TC observations; we do not undertake to estimate damages. This is a crucial first step in testing a hazard model; it is not plausible that a model that performs poorly against TC observations can predict damage with any fidelity. On the other hand, estimating damage requires knowledge of the vulnerability of the built environment to wind and water, and to the value of the property affected; this adds additional uncertainty to damage prediction. For statistical and statistical-dynamical downscalings, one must also unpack the limited information about size and intensity (e.g. radius and intensity of maximum surface winds) into a fully two-dimensional and time-evolving wind field, adding yet more uncertainty to damage estimates. We therefore postpone consideration of damage metrics to a future paper and here focus on strictly meteorological metrics.

In the following sections we describe the observational and model data sets, the proposed metrics, and a set of statistical tests of goodness-of-fit between model predictions and observations. This is followed by a presentation of the results and a summary with recommendations for further progress.

2. Observational and hazard model data

Historical TC data are provided by the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp *et al* 2010). Here we use data curated by the National Hurricane Center for the North Atlantic, eastern North Pacific and central North Pacific, and data from the Joint Typhoon Warning Center for the rest of the world. We confine our attention to the period 1979–2023, during which satellites are thought to have detected almost all TCs globally.

We use the MIT TC hazard model (Emanuel *et al* 2006, 2008) to generate a large, global set of synthetic TCs. This statistical-dynamical model downscales TCs from global, time-evolving climate states represented by both reanalyses and CMIP6 climate models. In this technique, TCs are initiated by seeding randomly, in time and space, the time-evolving large-scale environment provided by global climate models or reanalyses. The large-scale environment is represented by Fourier series in time with random phase, constrained so that the monthly means of all variables and the monthly mean variances and covariances (based on daily data) among the wind components are identical to those of the gridded data. The kinetic energy spectrum of the synthesized large-scale winds is also constrained to obey geostrophic turbulence scaling. The synthetic time series are then bilinearly interpolated to the storm positions and linearly interpolated to the date and time. We use this procedure, rather than the reanalysis or climate model winds directly, to provide a potentially unlimited number of realizations of the wind.

A TC intensity model (CHIPS; Emanuel and Rappaport 2000) is then run along each of the randomly generated tracks. Owing to the use of an angular momentum radial coordinate, the intensity model has very high spatial resolution in the storm core. It has been shown to produce skillful real-time intensity forecasts (Emanuel and Rappaport 2000). Over 99% of the seeded tracks dissipate and are discarded; the survivors are regarded as constituting the TC climatology of the original reanalysis or climate model. When applied to global reanalysis data, this technique has been shown to simulate with reasonable fidelity most the salient features of the current climatology of TCs (Emanuel *et al* 2008).

The ratio of the number of initial seeds that develop into TCs to the total number of seeds deployed is a measure of the overall frequency of the synthetic TCs. To calibrate these event sets, we multiply this ratio by a single scalar so that the annual, global rate of synthetic TCs averaged over the period 1979–2023 is 85, close to that derived from historical TC data over this period.

This technique has several advantages over conventional downscaling using regional models. The use of angular momentum coordinates allows for increasing spatial resolution of the storm core as its intensity increases; consequently, each storm's intensity is limited by the physical properties of its environment rather than by numerical resolution. Because the TC model is driven by the statistics of the global model or reanalysis, an arbitrarily large number of events can be simulated in a given climate.

Table 1. reanalyses and CMIP6 models downscaled.

Model/reanalysis	Institution	Type	Designation in this paper
ERA5	European Center for Medium-Range Forecasts	Reanalysis	ERA5
MERRA2	National Aeronautics and Space Administration	Reanalysis	MERRA2
NCEP	NOAA National Centers for Environmental Prediction	Reanalysis	NCEP
CanESM5	Canadian Center for Climate Modeling and Analysis	CMIP6 GCM	CANESM
CESM-2	National Center for Atmospheric Research	CMIP6 GCM	CESM2
CNRM-CM3	Centre National de Recherches Météorologiques, Météo-France	CMIP6 GCM	CNRM
EC-Earth3	EC-Earth consortium	CMIP6 GCM	ECEARTH
FGOALS-g3	Chinese Academy of Sciences	CMIP6 GCM	FGOALS
IPSL-CM6A-LR	Institute Pierre Simon Laplace	CMIP6 GCM	IPSL
MIROC6	Center for Climate System Research, University of Tokyo; Japan Agency for Marine-Earth Science and Technology; National Institute for Environmental Studies	CMIP6 GCM	MIROC
MPI-ESM1–2-HR	Max Planck Institute	CMIP6 GCM	MPI
MRI CGCM 2.3.2a	Meteorological Research Institute, Japan	CMIP6 GCM	MRI
UKESM1–0-LL	United Kingdom Meteorological Office	CMIP6 GCM	UKMO

For the present purpose, we generate 45 000 synthetic TCs downscaled from each of three global reanalyses and ten CMIP6 global climate models over the period 1979–2023[2]. This works out to a total of 585 000 synthetic TC events. The reanalyses and CMIP6 models used are summarized in table 1. In the case of the CMIP6 models, we use output from the historical simulations for the period 1979–2014 and from the SSP3-7.0 simulations for 2015–2023. The choice of SSP3-7.0 is arbitrary because the emissions pathways of all the SSPs are nearly identical during the years 2015–2023. Casual examination of TC hazard model times series shows no sign of discontinuity between 2014 and 2015.

3. Metrics

Decisions about what metrics to use for testing TC Hazard model datasets are here guided by four key considerations:

1. Robustness in the historical data sets
2. Meteorological fidelity
3. Eventual utility
4. Amenability to standard statistical goodness-of-fit tests

Probably the most robustly observed variable, at least since around 1979, is the number of TCs. While it is possible that some 'midget TCs' escaped detection, the requirement imposed on the synthetic tracks—that wind speeds remain at or above 35 knots for at least two days—probably filters out any midget TCs that might form. A more serious source of error is that the operational determination that a TC exists depends on the criteria used, which vary from place to place, and over time, and always have some subjective component. In the comparisons that follow, we consider only those observed and synthetic TCs whose lifetime maximum intensity exceeds 40 knots. Some of the uncertainty in observational counts arises from inaccuracies in whether this or any intensity threshold has been exceeded.

In assessing the robustness of the observed hurricane record, a large factor in many of the comparisons is the low sample size. For metrics that stem from counts, we assume that the sampling error obeys a Poisson

[2] The MERRA2 reanalysis begins in 1980, thus we use 1980–2023 for thatreanalysis.

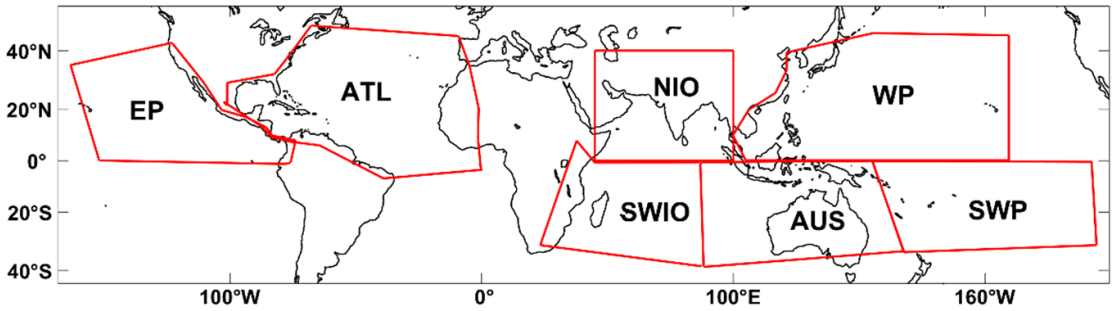

Figure 1. Basin definitions used in this study: Eastern North Pacific (EP), North Atlantic (ATL), southwest Indian Ocean (SWIO), North Indian Ocean (NIO), Australian region (AUS), western North Pacific (WP), southwestern South Pacific (SWP). TCs are assigned basins based on where their intensity first reaches 40 knots.

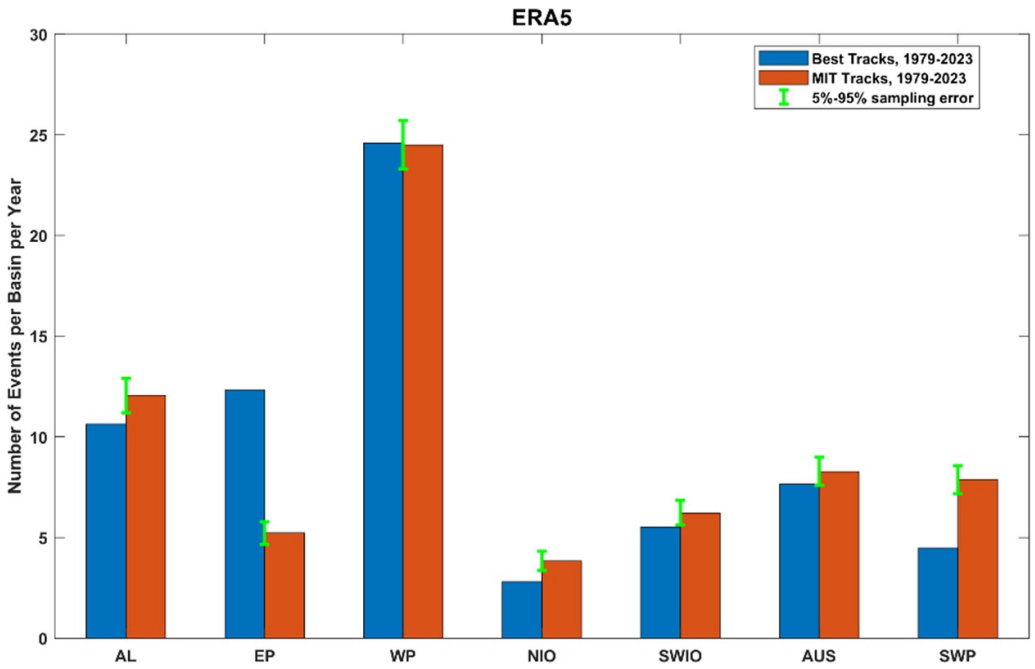

Figure 2. Annual average basin TC counts over the period 1979–2023 from historical ('best-track') observations (blue) and from MIT TC Hazard model synthetic events downscaled from ERA5 (red). The latter have been normalized so that the global total synthetic count matches the corresponding historical count. The green confidence intervals show estimated observational sampling error defined as the 5th and 95th quantiles of a Poisson distribution based on the synthetic basin frequency (see text for details).

distribution. We can use this estimate of sampling error in statistical tests that assume some prior knowledge of the magnitude of such errors. For this reason, we stick to metrics that involve TC counts in this paper.

Global TC annual frequency is not a candidate metric for the MIT downscaling, as we have calibrated the model to yield the correct average over the period 1979–2023 and the year-to-year variability in this number is no greater than what would be expected from a pure Poisson process, thus we cannot reject the hypothesis that the interannual variability of global TC counts is a random process. Therefore, we do not use either time-mean or annual global frequency as a metric.

One metric that a good TC hazard model ought to simulate well is the global spatial distribution of TCs. A simple measure of this is the total number of storms during the years 1979–2023 in each of a set of TC basins. For this purpose, we use the TC basin definitions shown in figure 1. Therefore, our first metric is the set of basin total TC counts; we shall refer to this metric as NB_i, the number of storms in basin i.

An example comparing the synthetic to the historical basin counts is presented in figure 2. Here the 45 000 global synthetic events have been downscaled from the ERA5 reanalysis; recall that the global total count has been constrained to equal that of the historical record. There are far too few events in the eastern North Pacific, and too many in the southwest South Pacific, but the historical and synthetic counts agree well elsewhere.

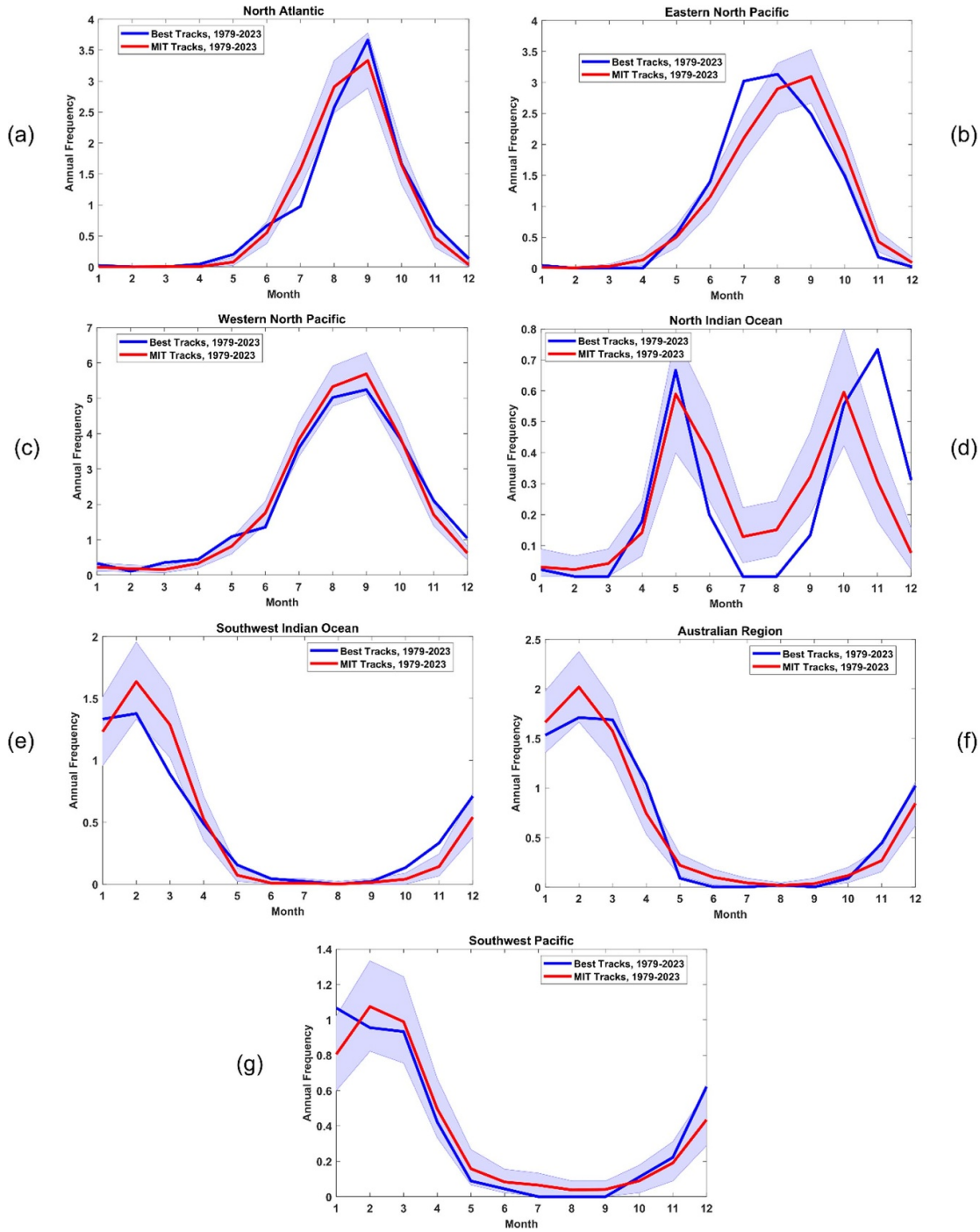

Figure 3. Monthly frequencies of TCs in each of the 7 basins (ATL, EP, WP, NIO, SWIO, AUS, and SWP shown in figure 1) averaged over the period 1979–2023. Counts based on historical observations and MIT downscaling of ERA5 shown by blue and red curves respectively. The blue shading shows an estimate of historical sampling error based on the 5th and 95th percentiles of a Poisson distribution centered on the MIT monthly counts.

The confidence intervals shown in figure 2 were calculated by finding the 5th and 95th quantiles of a Poisson distribution based on the estimated total number of observations in each basin (the synthetic count shown in the graph multiplied by the number of years (45) in the record), and the result was then divided by the number of years in the record. This procedure tests the hypothesis that the observed count was drawn from the same distribution as the (much larger) synthetic count. This same procedure was used to estimate sampling error in all but one of the metrics presented in this paper.

The annual cycle of TCs is a strong signal in all seven of the basins shown in figure 1. A good TC hazard model should be able to replicate these annual cycles and these therefore constitute our next seven metrics. To keep this metric separate from the basin totals metric, the synthetic annual cycle in each basin was normalized so that its sum over the 12 months equals the observed basin total over those months. We refer to

this metric as NM_{ij}, the number of storms in basin i and month j. While this metric is a good test of the meteorological fidelity of a hazard model, it is not clear how useful it is in actual applications. An example of this metric, again using events downscaled from ERA5, is shown in figure 3. While the correspondence between historical observations and synthetic events is generally good, some biases are evident that are nearly ubiquitous across the reanalyses and CMIP6 models downscaled here. In the eastern North Pacific region, the synthetic events peak in September while observed events peak in August. (Recall that the EP basin totals are greatly unpredicted.) And while the hazard model successfully captures the seasonal bimodality of events in the North Indian Ocean, the mid-summer lull is underpredicted while the autumn maximum occurs in October rather than November. These differences would not appear to be due to sampling error.

To be useful, TC hazard models must be able to simulate with good fidelity the frequency and intensity of storms at landfall along populous coastlines. Here we must balance the desire to emphasize the most destructive events against the need to have sufficient historical data to make the comparison meaningful. For this reason, we consider two metrics: Coastal crossings of all events in the dataset, e.g. events whose lifetime maximum wind speed exceeds 40 kts, and events that cross the coastlines as hurricanes, with maximum sustained surface winds of at least 64 kts at the time of coastline crossing. For TCs that cross a coastline from sea to land more than once, we here consider only the first crossing and we do not consider crossings from land to sea.

To capture the all-important distribution of events along coastlines, we divide the coast into a set of connected line segments often referred to as coastal gates; these are widely used in industry catastrophe models. Balancing the desire to include populous coastlines against the need to have accurate and plentiful historical observations, we consider the coastlines and coastal gates displayed in figure 4.

For each of the five sets of coastal gates we consider two metrics: first, the distribution of events among the gates of each set, with the synthetic track total normalized to be equal to that of the historical events. This metric can be expressed as NG_{kl}, where k is the coastline index (1–5) and l is the gate number on that coastline. In what follows, we smooth NG_{kl} with a 7-point running mean along the dimension l. The second metric is designed to measure whether the proportion of TCs in a particular basin that pass through any of the set of coastal gates is consistent with historical records. Unlike the other indices, this is not an integer but rather the ratio of two integers:

$$R_k \equiv \frac{\sum_{l=1}^{N_l} NG_{kl}}{NB_i}, \tag{1}$$

where the basin count that appears in the denominator is understood to be that basin that is relevant to the coastline in question. For example, for the Lesser Antilles, we would use the North Atlantic basin. When evaluating (1) for the synthetic tracks, we do not normalize either the gate counts or the basin count. The intent here is to separately measure the distribution of storms along coastlines and the proportion of basin storms that cross anywhere over that coastline.

Figure 5 shows an example of the first type of coastal impact metric.

The last two metrics proposed here relate to storm intensity. In keeping with the decision to stick with counts, we compare synthetic and historical histograms of intensity. Specifically, we compare the exceedance histograms of lifetime maximum intensities and exceedance histograms of the 6 h intensity change histogram of all points along each track. We labels these NV_n and NDV_m, where n is the index of wind intensity and m is the index of 6 h intensity changes.

For these two intensity metrics, we confine ourselves to the North Atlantic because intensity estimates elsewhere are somewhat less reliable as they are based entirely on satellite data (except for the western North Pacific from 1979 to 1987). Figure 6 shows an example of each of these two metrics.

The set of proposed metrics is summarized in table 2.

4. Goodness-of-fit tests

With the exception of the coastline-to basin ratio, R_k, all of the metrics in table 2 are integer counts. In comparing synthetic to observed counts, it is important to account for sampling error in the observations. (There are enough synthetic events to make their sampling error very small.) We assume that the observed counts are samples of a Poisson distribution centered at the predicted value, and wish to measure how close the observations are to the synthetic values. We do this, rather than the other way around, because the historical sampling error will be much larger than the synthetic sampling error. In essence, we wish to test the hypothesis that the observed counts are drawn from the same Poisson distribution as the synthetic event counts.

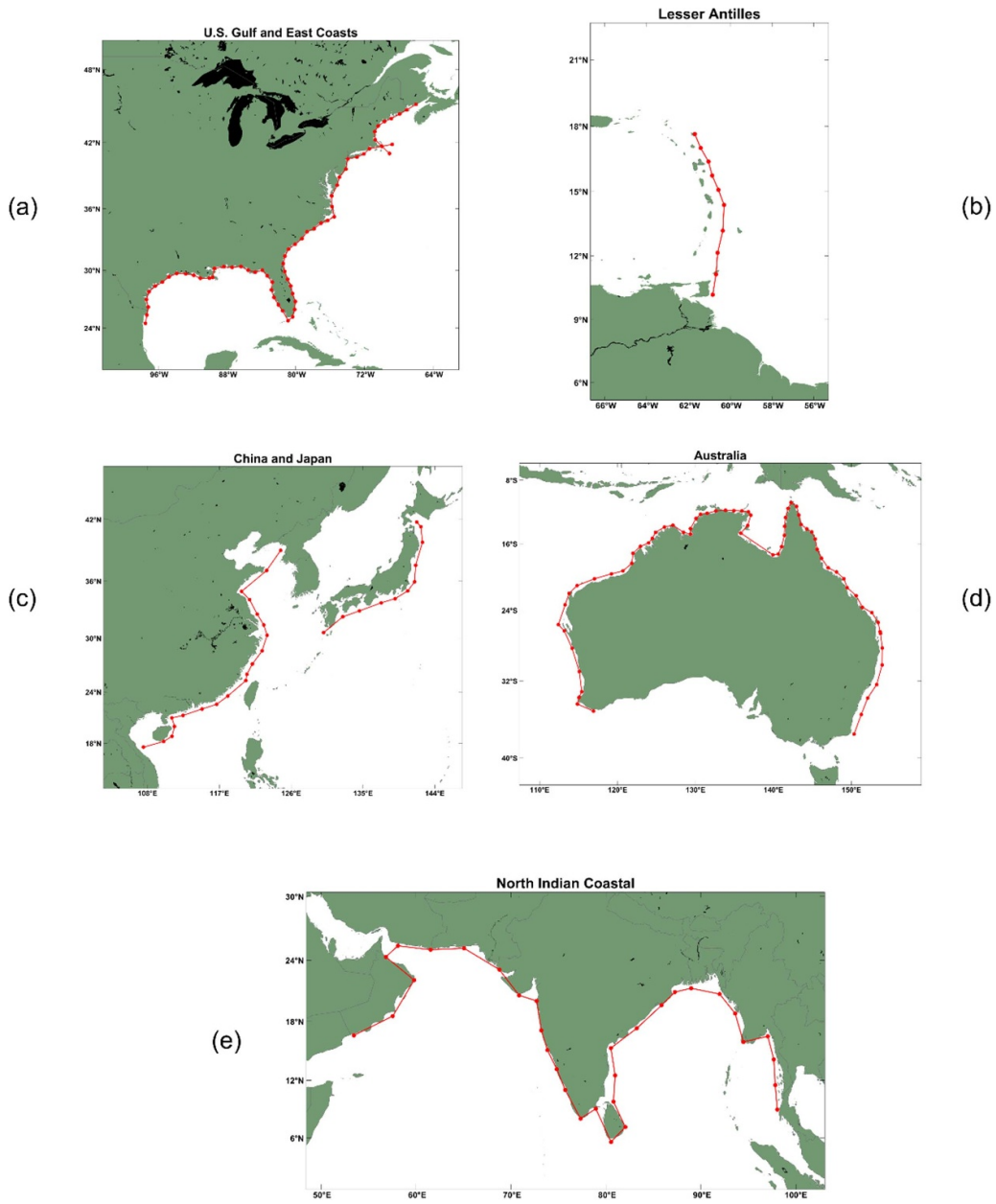

Figure 4. Definitions of 5 sets of coastal gates: (a) U.S. Gulf and East coasts, (b) Lesser Antilles, (c) China and Japan, (d) Australia, and (e) North Indian coastal.

To quantify the magnitude of Poisson random noise, for a given count, λ, we estimate the pth quantile of the inverse of the Poisson cumulative distribution function as

$$y_p = F^{-1}(p;\lambda), \tag{2}$$

where $F(x;\lambda)$ is the CDF of the Poisson distribution with rate parameter λ and $F^{-1}(p;\lambda)$ gives the smallest x such that $F(x;\lambda) \geqslant p$.

Here we chose to characterize the sampling error on the high side of the prediction as

$$y_{\text{upper}} \equiv F^{-1}(0.95;\lambda), \tag{3}$$

and sampling error on the low side as

$$y_{\text{lower}} \equiv F^{-1}(0.05;\lambda), \tag{4}$$

and further define the range of sampling error as

$$y_{\text{spread}} \equiv y_{\text{upper}} - y_{\text{lower}}. \tag{5}$$

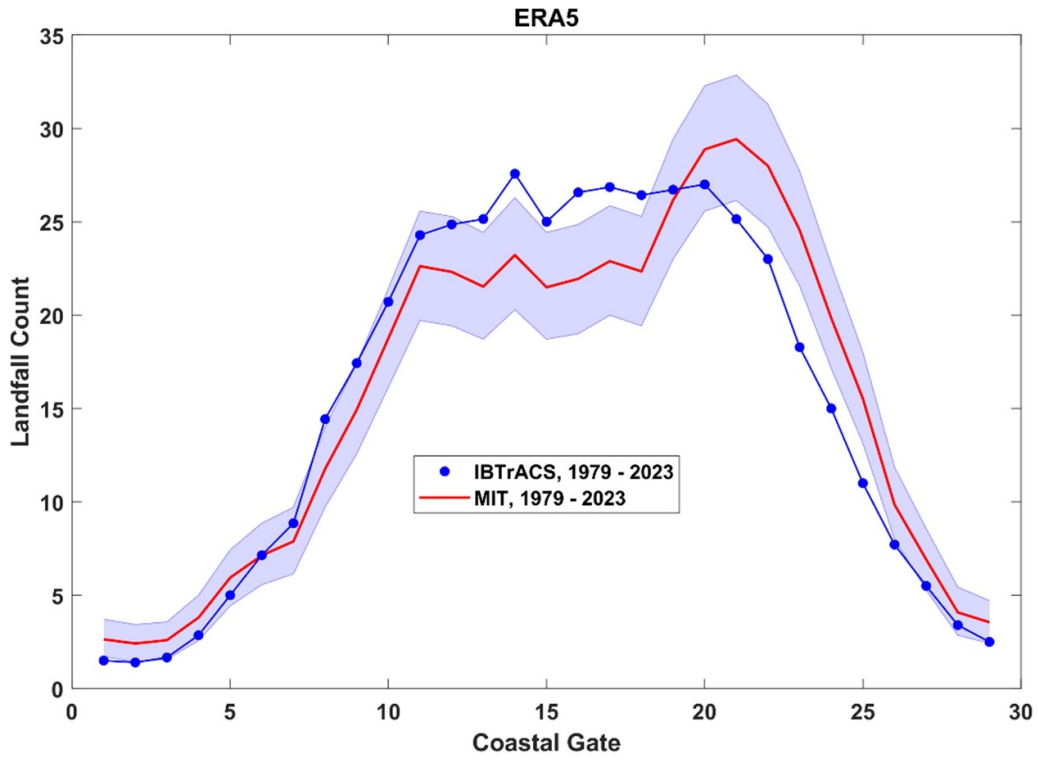

Figure 5. Comparison between number of TC crossings for each coastal gate of the China-Japan coastline shown in figure 4(c). Synthetic track count downscaled from ERA5 shown in red while historical tracks are in blue. The blue shading shows the span between the 5th and 95th quantiles of a Poisson distribution based on the synthetic track count scaled to the historical track count. Both distributions have been smoothed with a 7-gate running mean.

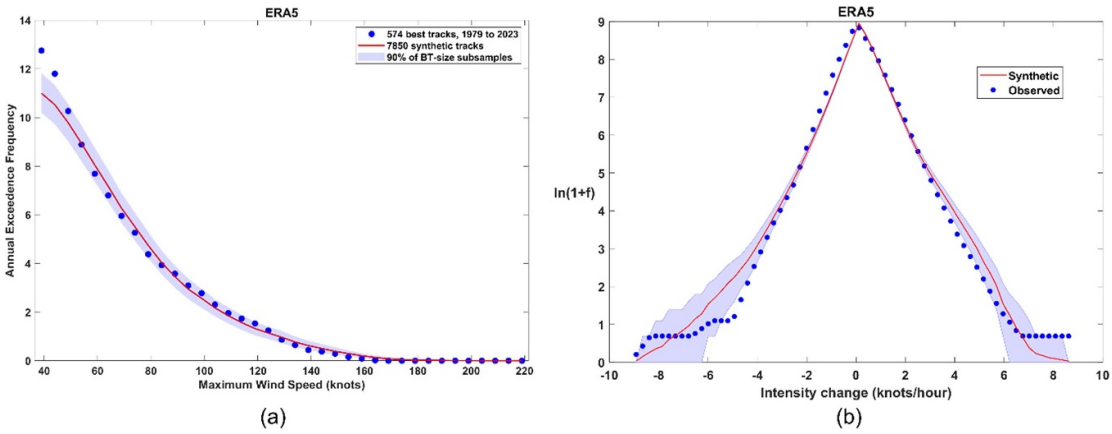

Figure 6. Annual exceedance frequency by lifetime maximum intensity (a) and the natural logarithm of the double-sided cumulative frequency of intensity change (b). The latter accumulates the bin frequency going from left to right starting at the frequency peak, and separately from right to left starting at the peak. ERA5 downscaling shown in red, results based on historical records by blue dots. The blue 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.

a. Poisson log-likelihood test

One measure of goodness-of fit appropriate to counts with Poisson sampling error is the Poisson log-likelihood, or deviance, defined

$$D = \sum_i 2\left[y(i)_{\rm obs} \ln\left(\frac{y(i)_{\rm obs}}{y(i)_{\rm syn}}\right) - \left(y(i)_{\rm obs} - y(i)_{\rm syn}\right)\right], \tag{6}$$

where $y_{\rm obs}$ is the observed count, $y_{\rm syn}$ is the synthetic count, and the index i represents the dimension over which y is varying; for example, the basin index for basin counts, or the month of the year for the annual

Table 2. Summary of TC metrics.

Short Name	Description	Symbol
Basin counts	Number of TCs in each ocean basin $i = 1:7$	NB_i
Annual cycle	Number of storms in each month, $j = 1:12$, performed in each ocean basin, i. Synthetic counts normalized so that the annual total equals the observed annual total.	NM_{ij}
Gate crossings	Number of storms crossing each gate $l = 1:\mathrm{NC}_k$ performed for each coastline, k. NC_k is the total number of gates in coastline k. Synthetic counts normalized so that the coastline total equals the observed coastline total.	NG_{kl}
Coastal crossing-to-basin ratio	The ratio of the total number of gate crossings in coastline k, $\sum_{l=1}^{\mathrm{NC}_k}\mathrm{NG}_{kl}$, to the basin count NB_i relevant to that coastline. No normalizations applied.	R_k
Intensity histogram	Histogram counts of North Atlantic lifetime maximum intensity. Total annual frequency of synthetic events set at 12	NV_n
Intensity change histogram	Natural logarithm of 1 plus the histogram of North Atlantic 6 h intensity changes over whole lifetime of each event. Peak synthetic frequency set equal to peak observed frequency.	NDV_m

cycle metric. Because D is dimensional, having the same units as y, we normalize it by the same quantity calculated using the mean rather than individual values of y_{obs} :

$$D_{\mathrm{mean}} = \sum_i 2\left[y(i)_{\mathrm{obs}} \ln\left(\frac{y(i)_{obs}}{\overline{y_{\mathrm{syn}}}}\right) - \left(y(i)_{\mathrm{obs}} - \overline{y_{\mathrm{syn}}}\right)\right], \tag{7}$$

where the overbar denotes the mean of y_{syn} over the index i.

One drawback of the Poisson log-likelihood is that it blows up if either $y(i)_{\mathrm{obs}}$ or $y(i)_{\mathrm{syn}}$ vanishes. In applying this metric here, we sum only over those index values for which $y(i)_{\mathrm{obs}} > 5$ and $y(i)_{\mathrm{syn}} > 5$. Note also that D is positive definite. Larger differences indicate less agreement between the synthetic and observed counts. To convert this into something that looks more like a skill metric, we transform it by

$$D' \equiv \max\left(1 - 0.4\frac{D}{D_{\mathrm{mean}}}, 0\right). \tag{8}$$

The factor of 0.4 in (8) is designed to bring this test roughly in line with the tests described in the following section.

This test will produce smaller skill if the synthetic counts differ from the observed by a multiplicative factor. (This is also an issue with the K-test and χ^2 tests described in the next subsections.) This problem is largely addressed by normalizing the synthetic counts as described in the previous section.

b. K-test

This test, presented here for the first time, assigns a skill of unity when the observed count falls within the sampling error spread of the synthetic count. Specifically, we first define an error measure K :

$$K \equiv \begin{array}{l} \sum_i \frac{\max\left(y(i)_{\mathrm{obs}} - y(i)_{\mathrm{upper}}, 0\right)}{y(i)_{\mathrm{syn}}} \quad y(i)_{\mathrm{obs}} \geqslant y(i)_{\mathrm{syn}} \\ \sum_i \frac{\max\left(y(i)_{\mathrm{lower}} - y(i)_{\mathrm{obs}}, 0\right)}{y(i)_{\mathrm{syn}}} \quad y(i)_{\mathrm{obs}} < y(i)_{\mathrm{syn}} \end{array}. \tag{9}$$

This measure of error accrues only when the observed count lies outside the sampling error envelope and is proportional to how far outside the envelope it is. The K-test is non-dimensional, but like the D test is

singular when the synthetic count vanishes. Therefore, as with the D test, we sum only over those indices for which $y(i)_{\rm obs} > 5$ and $y(i)_{\rm syn} > 5$. We also divide the result by the number of bins that meet these criteria.

To turn this into something more like a skill metric, we transform K by

$$K' \equiv \max\left(1 - K, 0\right). \tag{10}$$

The K-test effectively gives the benefit of the doubt to the synthetic TCs, where in this case the 'doubt' is due to the often-large sampling error of the observations.

c. Coefficient of determination

We also calculate the conventional coefficient of determination, or r^2. This has the advantage of being invariant to both multiplicative and additive errors.

d. χ^2

The final goodness-of-fit test we apply is the standard χ^2 calculation, measuring the root-mean-square difference between the observed and synthetic data normalized by the square of the predicted values:

$$\chi^2 = \sum_i \frac{\left(y(i)_{\rm obs} - y(i)_{\rm syn}\right)^2}{y(i)^2_{\rm syn}}. \tag{11}$$

In performing the sum, we apply the same exclusions as described in the sections on the Poisson log-likelihood and K tests and divide the result by the number of bins that meet these criteria. To convert to a skill metric, we apply

$$\chi^{2\prime} = \max\left(1 - 0.5\chi^2, 0\right). \tag{12}$$

The coefficient 0.5 is designed to bring this skill metric more or less into the same range of numerical values as the previously described skill measures.

e. A note on the coastline to basin ratio, R_k

There are enough historical events summed over basins and over the coastal gates used here that the sampling error is small. For this reason, we present the raw ratio of synthetic to observed values of R_k for each coastline, k:

$$R_k{}' \equiv \frac{R_{\rm ksyn}}{R_{\rm kobs}}. \tag{13}$$

5. Results

We present the results of these tests, applied to the three reanalyses and ten CMIP6 climate models listed in table 1, all for the period 1979–2023, in figures 7–11. In each case, the samples and their sizes can be inferred from table 2.

The first four of these figures show the skill scores D', K', r^2, and $\chi^{2\prime}$, respectively. In all cases, a score of unity denotes perfect skill while zero is regarded as very low or no skill. The bottom row of these figures shows the arithmetic mean over all the metrics. For ease of comparison, the cells in the tables are color-coded for each skill test S as follows: $S < 0.25$, red; $0.25 \leqslant S < 0.5$, orange; $0.5 \leqslant S < 0.75$, yellow, $S \geqslant 0.75$, green.

Figure 11 shows the ratio given by (12) for all TCs and for hurricane-strength TCs. The bottom row shows the average skill, defined as

$$\overline{R_k{}'} \equiv \overline{\max\left(1 - |1 - R_k{}'|, 0\right)}, \tag{14}$$

where the overbar signifies the average over all the other rows in this table. This penalizes departures from unity in either direction. The color coding is the same as in figures 7–10 but since $R_k{}'$ can be greater than unity, we add the following: $1 \leqslant R_k{}' < 1.25$, green; $1.25 \leqslant R_k{}' < 1.5$, yellow, $1.5 \leqslant R_k{}' < 1.75$, orange; and $R_k{}' \geqslant 1.75$, red.

As one may have anticipated, the skill with which the synthetic events mimic observed events depends on both the metric and the statistical test. Nevertheless, there is some commonality among the skill scores. Some general observations include

D'

	ERA5	MERRA2	NCEP	CANSEM	CESM2	CNRM	ECEARTH	FGOALS	IPSL	MIROC	MPI	MRI	UKMO
Basin Counts	0.90	0.90	0.84	0.96	0.85	0.99	0.86	0.80	0.84	0.76	0.90	0.93	0.90
Atlantic AC	0.96	0.95	0.99	0.75	0.96	0.91	0.92	0.95	0.73	0.93	0.96	0.89	0.93
East Pacific AC	0.94	0.92	0.86	0.90	0.80	0.80	0.62	0.75	0.77	0.91	0.92	0.82	0.86
West Pacific AC	0.98	0.98	0.99	0.80	0.96	0.79	0.88	0.97	0.75	0.94	0.87	0.86	0.92
North Indian AC	0.73	0.61	0.76	0.46	0.52	0.28	0.54	0.61	0.00	0.67	0.67	0.00	0.67
Southwest Indian AC	0.85	0.89	0.92	0.79	0.65	0.91	0.75	0.91	0.80	0.82	0.79	0.00	0.67
Australian AC	0.93	0.92	0.97	0.84	0.56	0.54	0.62	0.95	0.92	0.93	0.87	0.22	0.73
Southwest Pacific AC	0.93	0.91	0.90	0.28	0.52	0.00	0.00	0.65	0.23	0.93	0.65	0.28	0.67
U.S. Gulf and East Gates	0.80	0.84	0.84	0.71	0.72	0.88	0.89	0.76	0.56	0.70	0.00	0.69	0.93
U.S. Hurricane Gates	0.63	0.70	0.61	0.63	0.77	0.61	0.76	0.68	0.30	0.29	0.66	0.57	0.76
Lesser Antilles Gates	0.85	0.03	0.09	0.07	0.00	0.22	0.68	0.00	0.42	0.94	0.66	0.28	0.17
Lesser Antilles Hurricane Gates	0.00	0.59	0.44	0.00	0.71	0.27	0.69	0.65	0.49	0.00	0.08	0.12	0.70
China and Japan Gates	0.97	0.98	0.96	0.93	0.94	0.94	0.94	0.96	0.88	0.96	0.96	0.93	0.93
China and Japan Hurricane Gates	0.89	0.89	0.83	0.68	0.81	0.76	0.81	0.87	0.57	0.75	0.93	0.73	0.71
North Indian Gates	0.81	0.76	0.75	0.87	0.89	0.76	0.84	0.62	0.78	0.82	0.84	0.87	0.81
North Indian Hurricane Gates	0.85	0.89	0.83	0.17	0.81	0.74	0.88	0.63	0.81	0.76	0.80	0.94	0.40
Australia Gates	0.86	0.85	0.81	0.86	0.79	0.75	0.56	0.55	0.63	0.89	0.80	0.86	0.88
Australia Hurricane Gates	0.84	0.82	0.87	0.92	0.83	0.83	0.16	0.53	0.17	0.77	0.77	0.88	0.92
North Atlantic LMI Histogram	1.00	0.99	0.99	0.85	1.00	0.98	0.97	0.99	0.81	0.92	0.95	0.92	0.96
North Atlantic log Intensification	0.98	0.97	0.99	0.89	0.97	0.97	0.92	0.98	0.91	0.96	0.93	0.94	0.93
Mean Skill Score	0.83	0.82	0.81	0.67	0.75	0.69	0.71	0.74	0.62	0.78	0.75	0.64	0.77

Figure 7. The skill score D' for three reanalyses and ten CMIP6 climate models (columns), across 20 metrics and the arithmetic mean of them (rows), color-coded for each skill test S as follows: $S < 0.25$, red; $0.25 \leqslant S < 0.5$, orange; $0.5 \leqslant S < 0.75$, yellow, $S \geqslant 0.75$, green.

K'

	ERA5	MERRA2	NCEP	CANSEM	CESM2	CNRM	ECEARTH	FGOALS	IPSL	MIROC	MPI	MRI	UKMO
Basin Counts	0.76	0.74	0.64	0.86	0.47	0.72	0.54	0.55	0.64	0.64	0.74	0.67	0.73
Atlantic AC	0.96	0.96	1.00	0.48	0.97	0.81	0.74	0.91	0.17	0.85	0.90	0.65	0.86
East Pacific AC	0.92	0.88	0.79	0.85	0.67	0.45	0.48	0.57	0.56	0.86	0.88	0.37	0.81
West Pacific AC	0.90	0.92	0.99	0.34	0.79	0.36	0.65	0.93	0.47	0.73	0.33	0.54	0.64
North Indian AC	0.79	0.66	0.83	0.57	0.82	0.49	0.68	0.79	0.27	0.76	0.74	0.00	0.78
Southwest Indian AC	0.88	0.89	0.91	0.96	0.70	0.96	0.78	0.97	0.85	0.80	0.95	0.32	0.88
Australian AC	0.95	0.90	1.00	0.78	0.56	0.89	0.93	0.96	0.96	0.97	0.90	0.81	0.65
Southwest Pacific AC	0.98	0.98	0.97	0.75	0.61	0.50	0.52	0.76	0.73	1.00	0.77	0.77	0.79
U.S. Gulf and East Gates	0.98	0.99	0.98	0.91	0.94	0.98	0.99	0.95	0.77	0.91	0.79	0.89	1.00
U.S. Hurricane Gates	0.97	0.99	0.96	0.99	0.99	0.97	0.99	0.95	0.96	0.95	1.00	0.98	1.00
Lesser Antilles Gates	1.00	0.97	0.97	0.97	0.96	0.98	1.00	0.92	0.99	1.00	1.00	0.98	0.98
Lesser Antilles Hurricane Gates	0.95	1.00	1.00	0.92	1.00	1.00	1.00	1.00	1.00	0.87	0.99	0.99	1.00
China and Japan Gates	0.97	0.97	0.92	0.91	0.91	0.92	0.91	0.93	0.82	0.90	0.95	0.90	0.89
China and Japan Hurricane Gates	0.91	0.91	0.84	0.74	0.83	0.77	0.83	0.90	0.62	0.59	0.95	0.74	0.76
North Indian Gates	0.69	0.81	0.66	0.79	0.86	0.64	0.79	0.53	0.63	0.73	0.75	0.83	0.74
North Indian Hurricane Gates	0.97	0.98	0.97	0.89	0.92	0.97	1.00	0.76	0.94	0.88	0.98	0.99	0.93
Australia Gates	0.93	0.93	0.89	0.95	0.87	0.86	0.69	0.67	0.77	0.96	0.89	0.92	0.94
Australia Hurricane Gates	0.95	0.96	0.99	1.00	0.93	0.97	0.57	0.79	0.65	0.88	0.91	1.00	1.00
North Atlantic LMI Histogram	0.99	0.97	0.96	0.00	1.00	0.86	0.78	0.97	0.00	0.33	0.61	0.26	0.63
North Atlantic log Intensification	0.84	0.48	0.81	0.00	0.59	0.28	0.00	0.77	0.00	0.32	0.00	0.00	0.00
Mean Skill Score	0.91	0.88	0.90	0.73	0.82	0.77	0.74	0.83	0.64	0.80	0.80	0.68	0.80

Figure 8. The skill score K′ for three reanalyses and ten CMIP6 climate models (columns), across 20 metrics and the arithmetic mean of them (rows), color-coded for each skill test S as follows: $S < 0.25$, red; $0.25 \leqslant S < 0.5$, orange; $0.5 \leqslant S < 0.75$, yellow, $S \geqslant 0.75$, green.

r^2

	ERA5	MERRA2	NCEP	CANSEM	CESM2	CNRM	ECEARTH	FGOALS	IPSL	MIROC	MPI	MRI	UKMO
Basin Counts	0.83	0.83	0.77	0.91	0.75	0.91	0.63	0.58	0.70	0.28	0.80	0.91	0.84
Atlantic AC	0.96	0.94	0.97	0.76	0.95	0.92	0.98	0.90	0.84	0.89	0.98	0.95	0.96
East Pacific AC	0.91	0.90	0.80	0.85	0.77	0.82	0.66	0.67	0.73	0.87	0.89	0.84	0.79
West Pacific AC	0.99	0.98	0.99	0.92	0.96	0.95	0.97	0.93	0.94	0.89	0.96	0.97	0.97
North Indian AC	0.56	0.39	0.66	0.40	0.68	0.28	0.47	0.55	0.24	0.58	0.45	0.22	0.54
Southwest Indian AC	0.93	0.96	0.93	0.89	0.67	0.90	0.78	0.90	0.80	0.82	0.90	0.69	0.86
Australian AC	0.95	0.93	0.96	0.96	0.63	0.90	0.91	0.91	0.93	0.96	0.95	0.89	0.92
Southwest Pacific AC	0.93	0.91	0.90	0.79	0.54	0.67	0.70	0.71	0.77	0.93	0.77	0.75	0.76
U.S. Gulf and East Gates	0.73	0.79	0.74	0.83	0.65	0.80	0.83	0.58	0.65	0.84	0.61	0.79	0.93
U.S. Hurricane Gates	0.67	0.71	0.67	0.56	0.71	0.53	0.69	0.30	0.26	0.48	0.56	0.42	0.67
Lesser Antilles Gates	0.77	0.39	0.29	0.22	0.46	0.31	0.19	0.61	0.19	0.85	0.15	0.06	0.25
Lesser Antilles Hurricane Gates	0.72	0.01	0.19	0.68	0.41	0.52	0.31	0.16	0.26	0.65	0.32	0.33	0.29
China and Japan Gates	0.90	0.90	0.88	0.76	0.83	0.81	0.80	0.86	0.59	0.95	0.88	0.73	0.73
China and Japan Hurricane Gates	0.79	0.78	0.67	0.55	0.63	0.65	0.66	0.75	0.41	0.89	0.86	0.55	0.55
North Indian Gates	0.62	0.54	0.51	0.70	0.88	0.43	0.59	0.12	0.64	0.59	0.65	0.67	0.58
North Indian Hurricane Gates	0.80	0.81	0.73	0.73	0.68	0.68	0.86	0.16	0.66	0.55	0.74	0.84	0.68
Australia Gates	0.66	0.68	0.60	0.65	0.57	0.45	0.23	0.26	0.39	0.80	0.51	0.70	0.74
Australia Hurricane Gates	0.83	0.83	0.87	0.84	0.81	0.81	0.30	0.47	0.35	0.74	0.75	0.83	0.89
North Atlantic LMI Histogram	0.99	0.99	0.99	0.96	1.00	0.99	0.99	0.99	0.96	0.98	0.98	0.98	0.99
North Atlantic log Intensification	0.95	0.95	0.96	0.87	0.95	0.95	0.90	0.97	0.94	0.93	0.88	0.93	0.90
Mean Skill Score	0.82	0.76	0.75	0.74	0.73	0.71	0.67	0.62	0.61	0.77	0.73	0.70	0.74

Figure 9. The coefficient of determination, r^2, for three reanalyses and ten CMIP6 climate models (columns), across 20 metrics and the arithmetic mean of them (rows), color-coded for each skill test S as follows: $S < 0.25$, red; $0.25 \leqslant S < 0.5$, orange; $0.5 \leqslant S < 0.75$, yellow, $S \geqslant 0.75$, green.

χ'^2

	ERA5	MERRA2	NCEP	CANSEM	CESM2	CNRM	ECEARTH	FGOALS	IPSL	MIROC	MPI	MRI	UKMO
Basin Counts	0.86	0.87	0.68	0.97	0.65	0.90	0.77	0.71	0.81	0.95	0.88	0.83	0.86
Atlantic AC	0.97	0.96	0.99	0.34	0.98	0.85	0.73	0.94	0.00	0.89	0.91	0.61	0.89
East Pacific AC	0.95	0.93	0.88	0.93	0.79	0.32	0.63	0.71	0.66	0.93	0.94	0.00	0.89
West Pacific AC	0.89	0.92	0.98	0.00	0.74	0.06	0.65	0.93	0.32	0.77	0.00	0.44	0.51
North Indian AC	0.78	0.58	0.74	0.32	0.83	0.35	0.60	0.79	0.00	0.72	0.63	0.00	0.76
Southwest Indian AC	0.83	0.85	0.88	0.97	0.74	0.96	0.80	0.95	0.89	0.73	0.96	0.60	0.92
Australian AC	0.94	0.92	0.97	0.66	0.60	0.92	0.95	0.96	0.95	0.97	0.90	0.84	0.51
Southwest Pacific AC	0.97	0.96	0.96	0.72	0.56	0.37	0.35	0.77	0.70	0.98	0.75	0.80	0.80
U.S. Gulf and East Gates	0.96	0.96	0.96	0.86	0.93	0.96	0.97	0.94	0.66	0.88	0.73	0.85	0.98
U.S. Hurricane Gates	0.90	0.93	0.90	0.95	0.95	0.88	0.94	0.87	0.89	0.90	0.96	0.95	0.94
Lesser Antilles Gates	1.00	0.97	0.98	0.98	0.97	0.98	0.99	0.95	0.99	1.00	0.99	0.98	0.98
Lesser Antilles Hurricane Gates	0.90	0.98	0.97	0.87	0.99	0.96	0.98	0.98	0.97	0.78	0.95	0.95	0.98
China and Japan Gates	0.98	0.98	0.96	0.96	0.96	0.97	0.96	0.97	0.92	0.92	0.97	0.96	0.95
China and Japan Hurricane Gates	0.95	0.95	0.92	0.86	0.91	0.86	0.92	0.95	0.77	0.17	0.97	0.86	0.87
North Indian Gates	0.73	0.65	0.78	0.81	0.89	0.68	0.84	0.68	0.58	0.75	0.78	0.86	0.78
North Indian Hurricane Gates	0.91	0.95	0.93	0.84	0.87	0.91	0.98	0.65	0.87	0.82	0.94	0.99	0.87
Australia Gates	0.92	0.91	0.86	0.91	0.85	0.86	0.66	0.61	0.76	0.94	0.88	0.90	0.92
Australia Hurricane Gates	0.91	0.91	0.94	0.98	0.88	0.92	0.39	0.69	0.54	0.84	0.87	0.94	0.96
North Atlantic LMI Histogram	0.99	0.97	0.98	0.00	0.99	0.90	0.85	0.98	0.00	0.33	0.72	0.18	0.71
North Atlantic log Intensification	0.93	0.66	0.93	0.00	0.78	0.41	0.00	0.92	0.00	0.50	0.00	0.00	0.00
Mean Skill Score	0.91	0.89	0.91	0.70	0.84	0.75	0.75	0.85	0.61	0.78	0.79	0.65	0.80

Figure 10. The quantity χ'^2 for three reanalyses and ten CMIP6 climate models (columns), across 20 metrics and the arithmetic mean of them (rows), color-coded for each skill test S as follows: $S < 0.25$, red; $0.25 \leqslant S < 0.5$, orange; $0.5 \leqslant S < 0.75$, yellow, $S \geqslant 0.75$, green.

R_k'

	ERA5	MERRA2	NCEP	CANSEM	CESM2	CNRM	ECEARTH	FGOALS	IPSL	MIROC	MPI	MRI	UKMO
U.S. Gulf and East	1.17	1.11	1.08	1.48	1.21	0.91	0.93	0.86	0.67	1.27	1.56	1.22	1.07
U.S. Gulf and East Hurricanes	1.60	1.44	1.39	1.28	1.52	1.08	0.76	0.88	0.52	1.16	1.49	1.23	1.07
Lesser Antilles	0.68	0.59	0.49	0.43	0.61	0.52	0.22	0.22	0.16	0.42	0.79	0.64	0.24
Lesser Antilles Hurricanes	1.01	1.02	0.87	0.26	0.84	0.78	0.06	0.29	0.04	0.37	0.69	0.28	0.14
China and Japan	0.86	0.81	0.76	0.63	0.70	0.62	0.79	0.81	0.72	0.83	0.96	0.87	0.70
China and Japan Hurricanes	0.83	0.76	0.64	0.36	0.52	0.60	0.52	0.93	0.49	0.60	0.54	0.45	0.45
North Indian Coastal	1.05	1.01	0.91	1.02	1.02	1.20	1.10	0.95	1.06	1.09	0.90	1.11	1.13
North Indian Coastal Hurricanes	1.02	0.92	0.91	0.79	0.78	1.02	0.65	1.30	0.50	1.01	0.40	0.68	0.75
Australia Coastal	1.66	1.52	1.62	1.65	1.43	1.50	1.82	1.35	1.68	1.80	1.94	2.04	1.65
Australia Coastal Hurricanes	2.20	1.90	1.88	1.93	1.29	1.83	1.78	1.84	1.81	2.16	1.82	2.00	1.90
Mean Skill Score	0.69	0.71	0.66	0.51	0.70	0.68	0.52	0.61	0.46	0.59	0.55	0.52	0.55

Figure 11. The ratio of the total number of synthetic tracks passing through the indicated coastal gates to the total number of synthetic tracks in the relevant basin, normalized by the equivalent ratio for observed TCs. These are color coded according to $R_k' < 0.25$, red; $0.25 \leqslant R_k' < 0.5$, orange; $0.5 \leqslant R_k' < 0.75$, yellow, $0.75 \leqslant R_k' < 1.25$, green; $1.25 \leqslant R_k' < 1.5$, yellow, $1.5 \leqslant R_k' < 1.75$, orange; and $R_k' \geq 1.75$, red.

1. The scores or generally higher for the three downscaled reanalyses than for the ten downscaled CMIP6 models. This is unsurprising, given that the reanalyses are constrained by observations.
2. The skill scores vary substantially among the metrics used here.
3. There is considerable variation from one climate model to the next. Some of this is owing to the different resolutions, numerics, and physics of each model, and some may be owing to different phases of oscillations such as ENSO and the Pacific Decadal Oscillation.
4. The basin counts are generally fair to good, though in the case of four CMIP6 models (ECEARTH, FGOALS, IPSL, and MIROC) the correlations are not strong.
5. The Poisson log-likelihood scores are quite good for the basin accounts and the annual cycles, except for the North Indian and southwest South Pacific basins. The U.S hurricane gates and the Lesser Antilles gates score relatively poorly in this test. Recall that these metrics measure the skill of variation along the coastlines; the ratio of coastline total crossings to basin events is shown separately in figure 11.
6. The K-test scores (figure 8) are generally higher, as they do not penalize predictions that are within the sampling error of the historical data. Somewhat surprisingly, the scores using this test are quite small for the intensity and intensification metrics for quite a few of the CMIP6 downscalings. Examination of the intensity histograms for some of the lowest scores shows that they greatly underestimate the frequencies of storms of mid-range intensity (roughly 70–120 kts). This is likely because many of the CMIP6 model-simulated potential intensities in the North Atlantic are too small. This is a common problem in GCMs, perhaps related to insufficient ocean heat transport into the North Atlantic.
7. The coefficient of determination, r^2 (figure 9), show good correlations between predicted and observed basin counts, except in the case of a few CMIP6 models, but tend to be low for variations along coastal gates. The correlations are high for the cumulative intensity and intensification histograms simply because they are dominated by a monotonic downward slope toward high intensities/intensification rates. This demonstrates a weakness of this skill measure.

8. In the χ^2 skill test (figure 10), the downscaled reanalyses score quite well for the most part, but the CMIP6 downscaling show weakness in the annual cycle tests and intensity and intensification cumulative histograms. This is probably for the same reason discussed in Point 6 above. There is some similarity with the results of the K-test discussed in Point 6, but the latter, being proportional to linear differences rather than differences in squares, is less sensitive to outliers and does not penalize downscaled results that lie within the envelope of historical sampling error.
9. Of particular interest are the skills of the ratios of coastal gate totals to basin totals, shown in figure 11. Downscalings from the three reanalyses show reasonable results for all U.S. Gulf and East coast downscalings, compared to Atlantic basin totals, but somewhat surprisingly, show too many hurricane crossings relative to basin TCs. It is possible that some of this has to do with fact that the raw data for the synthetic tracks is stored in two-hour intervals, compared to the six-hour intervals in the historical data, as it affects records of TCs of marginal hurricane strength near the time of landfall. For purposes of comparison, the six-hour historical wind speeds are linearly interpolated to two-hour time resolution. Both synthetic and linearly interpolated historical winds are then linearly interpolated to the time of gate crossing. By these means, the historical winds at gate crossing can be more severely contaminated by peaks winds after landfall than can the synthetic TCs. This effect should be further investigated.
10. The ratio of gate-crossing Australian hurricanes to basin counts is much too large across the whole sets of reanalyses and models (figure 11), even though the synthetic counts for Australian basin storms (not shown here except for ERA5 in figure 2) are reasonable. Some of the storms that cross the Australian east coast originate not in the Australian basin but in the southwest South Pacific, according to the basin definitions shown in figure 1. There tends to be a high bias in South Pacific synthetic storms, and this is evident in figure 2 in the case of the ERA5 downscalings. This could explain the excess Australian gate crossings. It is possible that some historical storms that would have been identified as having developed in the southwest South Pacific with better observations end up being classified as originating in the Australian basin.
11. The number of synthetic TCs passing through the Lesser Antilles, downscaled from the three reanalyses, is too small relative to the Atlantic basin counts, when compared to historical data. Yet the number of hurricanes passing through this region, in these three reanalysis downscalings, is about right (figure 11). On the other hand, both the number of TCs and the number of hurricanes passing through the Lesser Antilles are too small compared to basin counts for all of the CMIP6 downscalings. There is a strong north-south gradient of TC track density across these islands, and small biases in the track directions and/or genesis latitudes could greatly change the total number of tracks in this region.

6. Summary

As more TC hazard models are developed, it is important to create a uniform framework for comparing their performance to each other and to historical TC records. Here we have proposed both a set of metrics and a group of statistical goodness-of-fit tests to evaluate the performance of hazard models against historical TC records, accounting as much as possible for the low sample size of the latter. These metrics and statistical tests should be regarded as provisional and are meant to catalyze a discussion in the TC hazard community aimed toward the establishment of a uniform set of metrics and tests.

To illustrate these provisional metrics and tests, we applied them to large sets of synthetic TCs generated by the MIT TC hazard model applied to three reanalyses and ten CMIP6 models. One can anticipate that the application of these tests and metrics to TC hazard models more closely based on historical TC records (e.g. Vickery *et al* 2000, Bloemendaal *et al* 2020) will lead to better scores. This must be weighed against the fact that such methods will be pertinent to an historical period that may no longer accurately reflect today's current TC climatology, let alone that of the future, given the pace of climate change. By contrast, the TC hazard model tested here is completely independent of any historical TC data except insofar as it has been used to test the model. The biases revealed by tests such as these must be accounted for in applications of the hazard model and can also be used to drive improvements in the model. One interesting and large challenge is to account for biases in the climate models used to drive downscalings.

Of necessity, the tests presented here were performed by the author, who is an advocate for the hazard model being tested. This is far from ideal, and it would better for unbiased, independent groups to apply community-developed tests to a range of TC hazard models.

Data availability statement

The data that support the findings of this study are availableupon reasonable request from the authors.

Acknowledgment

This research is part of the MIT Climate Grand Challenge on Weather and Climate Extremes. Support was provided by Schmidt Sciences, LLC.

ORCID iD

Kerry Emanuel https://orcid.org/0000-0002-2066-2082

References

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

Emanuel K A, Ravela S, Vivant E and Risi C 2006 A statistical-deterministic approach to hurricane risk assessment *Bull. Am. Meteorol. Soc.* **19** 299–314

Emanuel K and Rappaport E 2000 Forecast skill of a simplified hurricane intensity prediction model *Preprints of the 24th Conf. Hurricanes and Tropical Meteorology, Ft.* (*Lauderdale, FL*) (American Meteorological Society) pp 236–7

Emanuel K, Sundararajan R and Williams J 2008 Hurricanes and global warming: results from downscaling IPCC AR4 simulations *Bull. Am. Meteorol. Soc.* **89** 347–67

Jing R and Lin N 2020 An environment-dependent probabilistic tropical cyclone model *J. Adv. Model. Earth Syst.* **12** e2019MS001975

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 best track data *Bull. Am. Meteorol. Soc.* **91** 363–76

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

Lin J, Rousseau-Rizzi R, Lee C-Y and Sobel A 2023 An open-source, physics-based, tropical cyclone downscaling model with intensity-dependent steering *J. Adv. Model. Earth Syst.* **15** e2023MS003686

Meiler A C, Kropf C M, Emanuel K and Bresch D N 2023 Uncertainties and sensitivities in the quantification of future tropical cyclone risk *Commun. Earth Environ.* **4** 371

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–E67

Oh S, Sen I and Tenekedjieva A-M 2021 Pricing of climate risk insurance: regulation and cross-subsidies *J. Finance* (https://doi.org/10.2139/ssrn.3762235)

Oliveira S, Rocha J and Sá A 2021 Wildfire risk modeling *Curr. Opin. Environ. Sci. Health* **23** 100274

PCAST 2023 Extreme weather risk in a changing climate: enhancing prediction and protecting communities (available at: https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/04/PCAST_Extreme-Weather-Report_April2023.pdf)

Peralta A and Scott J B 2019 Moving to flood plains: the unintended consequences of the national flood insurance program on population flows *Proc. Environmental Risk, Justice and Amenities in Housing Markets* vol 19

Vickery P J, Skerlj P F and Twisdale L A 2000 Simulation of hurricane risk in the U. S. using empirical track model *J. Struct. Eng.* **126** 1222–37