# Stochastic Simulation of Tropical Cyclones for Risk Assessment at One Go: A Multivariate Functional PCA Approach

Chi Yang[1,1], Jing Xu[2,2], and Jianming Yin[3,3]

[1]College of Global Change and Earth System Science, Beijing Normal University
[2]Chinese Academy of Meteorological Sciences
[3]China Re Catastrophe Risk Management Company LTD.

November 30, 2022

**Abstract**

A multivariate functional principal component analysis (PCA) approach to the full-track simulation of tropical cyclones (TCs) for risk assessment is developed. Elemental variables of TC along the track necessary for risk assessment, such as center coordinates, maximum wind speed, minimum central pressure and ordinal dates, can be simulated simultaneously at one go, using solely the best-track data with no data supplemented from any other sources. The simulation model is optimally determined by means of the ladle estimator. A TC occurrence model using the Conway–Maxwell–Poisson distribution is proposed as well, by which different dispersion features of annual occurrence can be represented in a unified manner. With the occurrence model, TCs can be simulated on an annual basis. The modeling and simulation process is programmed and fully automated such that little manual intervention is required, which greatly improves the modeling efficiency and reduces the turnaround time, especially when newly available TC data are incorporated periodically into the model. Comprehensive evaluation shows that this approach is capable of generating high-performance synthetic TCs in terms of distributional and extreme value features, which can be used in conjunction with wind field and engineering vulnerability models to estimate economic and insurance losses for governments and insurance/reinsurance industry.

# Stochastic Simulation of Tropical Cyclones for Risk Assessment at One Go: A Multivariate Functional PCA Approach

**Chi Yang[1], Jing Xu[2], and Jianming Yin[3]**

[1]College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China.

[2]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing 100081, China.

[3]China Re Catastrophe Risk Management Company LTD., Beijing 100033, China.

Corresponding author: Jing Xu (xujing@cma.gov.cn), Chi Yang (chi@bnu.edu.cn)

**Key Points:**

- Multivariate functional principal component analysis approach to the full-track simulation of tropical cyclones for risk assessment
- Tropical cyclone annual occurrence model using Conway–Maxwell–Poisson distributions
- Fully automated program for the generation of high-performance synthetic tropical cyclones

## Abstract

A multivariate functional principal component analysis (PCA) approach to the full-track simulation of tropical cyclones (TCs) is developed for risk assessment. Elemental variables of TC along the track necessary for risk assessment, such as center coordinates, maximum wind speed, minimum central pressure and ordinal dates, can be simulated simultaneously at one go, using solely the best-track data with no data supplemented from any other sources. The simulation model is optimally determined by means of the ladle estimator. A TC occurrence model using the Conway–Maxwell–Poisson distribution is proposed as well, by which different dispersion features of annual occurrence can be represented in a unified manner. With the occurrence model, TCs can be simulated on an annual basis. The modeling and simulation process is programmed and fully automated such that little manual intervention is required, which greatly improves the modeling efficiency and reduces the turnaround time, especially when newly available TC data are incorporated periodically into the model. Comprehensive evaluation shows that this approach is capable of generating high-performance synthetic TCs in terms of distributional and extreme value features, which can be used in conjunction with wind field and engineering vulnerability models to estimate economic and insurance losses for governments and insurance/reinsurance industry.



## Plain Language Summary

Tropical cyclones (TCs) are one of the biggest threats to life and property around the world. However, the infrequent nature of catastrophic TCs invalidates the standard actuarial loss estimation approaches. TC risk assessment requires estimation of catastrophic TCs having a very low occurrence probability, or equivalently a very long return period spanning up to thousands of years. Since reliable TC data are available only for recently decades, stochastic modeling and simulation turned out to be an effective approach to achieve more stable TC risk estimates for regions where little or no historical TC records exist. Here we present a novel model for the full-track simulation of TCs for risk assessment, via a machine learning approach called multivariate functional principal component analysis (MFPCA). Using this model, high-performance synthetic TCs can be generated in a fully automated manner such that little manual intervention is required, which greatly improves the modeling efficiency and reduces the turnaround time, especially when newly available TC data are incorporated periodically into the model. These synthetic TCs can be used in conjunction with wind field and engineering vulnerability models to estimate economic and insurance losses for governments and insurance/reinsurance industry.


## 1 Introduction

Tropical cyclones (TCs) are one of the biggest threats to life and property around the world. Over the past 50 years, there have been nearly 2,000 disasters linked to tropical cyclones, causing nearly 780,000 deaths and US$ 1,500 billion in economic losses (World Meteorological Organization, 2020). However, the infrequent nature of catastrophic TCs invalidates the standard actuarial loss estimation approaches. Computer models that are able to simulate tens, even hundreds, of thousands of synthetic TC tracks were developed in the past to compensate the scarcity of historical TC loss data, and to achieve more stable TC loss estimates for regions where little to no historical data exist. For insurance and reinsurance companies, it is necessary to evaluate the TC risks as precisely as possible to quantify, manage and mitigate financial losses. TC risk assessment requires estimation of catastrophic TCs having a very low occurrence probability, or equivalently a very long return period (e.g., 1000 years). Since reliable TC data are available only for recently decades, and landfalling TCs are relatively few in nature, stochastic modeling and simulation turned out to be an effective approach to achieve more stable TC risk estimates for regions where little or no historical TC records exist. The common practice consists of two stages (Vickery et al. 2009). The first stage is to fit a basin-wide TC full-track model to TC track data, to generate hundreds of thousands of synthetic TCs that can make up for the sparseness of TC observations while still complying with the statistical characteristics of the observed TCs. The second stage is to couple these synthetic TCs with TC wind field models, either to simulate landfall TC wind fields for wind hazard estimation, or to further drive storm surge models for coastal flood hazard estimation. Therefore, the performance of the synthetic TCs is crucial to the respective risk estimation. The full-track TC data consist of at least the TC center coordinates, maximum wind speed (MWS) as a measure of intensity and/or minimum central pressure (MCP) observed along TC tracks. A full-track model should be able to represent these elements and can be used for simulation.

Vickery et al. (2000) published the first full-track model for the North Atlantic (NA) basin within the regression framework. The track heading, speed and intensity were determined for each 5°×5° grid over the entire basin individually. This approach was then adapted for the Coral Sea (James and Mason 2005) and for the western North Pacific (WNP) basin (Yin et al. 2009; Li and Hong 2016; Chen and Duan 2018), respectively. Casson and Coles (2000) generated TC tracks for the NA basin simply by sampling the historical tracks and then translating by a normally distributed random displacement with the standard deviation less than 100 nm (1 nm = 1.852 km) and used a simple empirical model to simulate the central pressure depth with land effects. Emanuel et al. (2006) presented two different track models for the NA basin: a stochastic Markov chain model and a deterministic beta and advection model. The former propagates tracks by sampling a transition matrix that relates prior track speed and direction to the new speed and direction; the latter determines the TC motion by the weighted average of TC-ambient flow at 850 and 250 hPa plus a beta-drift correction. The TC intensity along tracks was obtained by coupling each synthetic track to a numeric model developed by Emanuel et al. (2004). Following this work, several Markovian-type TC track models were developed, e.g., Hall and Jewson (2007), Rumpf et al. (2007, 2009), Yonekura and Hall (2011), Kriesche et al. (2014) and Nakamura et al. (2015), for the NA or WNP basin or both. Emanuel et al. (2008) further developed a statistical-deterministic model for downscaling TC climatology from global analyses, using a random seeding method to initiate the storm, and a beta and advection model to propagate the storm. Following this approach, Lee et al. (2018) and Jing and Lin (2020) developed similar TC hazard models, either of which is comprised of three component models for TC genesis, track and intensity, respectively, dependent upon local environmental conditions.

In recent two decades, functional data analysis (FDA, Ramsay and Silverman, 2005) achieved rapid development. The object of FDA is a sample of random functions generated from an underlying process, rather than a sequence of individual points as analyzed by traditional approaches. Statistical models for random variables, either by supervised learning (e.g., regression models) or by unsupervised learning (e.g., principal component analysis (PCA)), can also be generalized to apply to random functions. All the elements of a TC can be viewed as functions of time during the TC life cycle. Therefore, TCs from a basin are naturally a sample for FDA. Rekabdarkolaee et al. (2019) proposed a functional analogue of the CLImatology and PERsistence (CLIPER) model (Aberson, 1998), which has long been used to forecast TC tracks in the NA basin. TC center location and intensity along the track were jointly modelled using multivariate functional linear regression with spatially varying coefficients, highlighting the representation of complex spatial-temporal dependency of TC tracks.

Although the full-track modeling of TCs has made much progress in the past decades, there are still some deficiencies in the current models. Most of all, they have been becoming more and more complicated. A full-track model usually consists of several components including TC genesis, track, MWS and/or MCP, and lysis, respectively. Some models may even have additional ones for the temporal/spatial clustering of TC tracks and TC behavior at and after landfall. Since model parameters are estimated at grid level as in most methods, model maintenance and update through periodical incorporation of newly available TC data could be quite cumbersome and thus time-consuming as a result. Moreover, these components adopt different methods suitable for their own tasks and are almost irrelevant with each other. Correlations existing between elemental variables of TC are hardly captured and as a result the synthetic TCs may exhibit characteristics inconsistent with those observed in the TC best tracks. In addition, many models use the TC-environmental factors such as sea surface temperature (SST) and ambient flow at 850 and 250 hPa from reanalysis data as predictors. While these data may bring in additional information in modeling and simulation, the additional data, , inevitably bring about extra uncertainties and potential biases into the already complicated and burdensome models. On the other hand, the information contained in the TC track data themselves has yet been far from fully exploited. For TC risk assessment, what is required from the full-track model is the statistical characteristics of historical TCs, which can be fully mined from the data themselves. Based on these considerations, we present in this work a flexible and extensible one-for-all model via the multivariate functional principal component analysis (MFPCA) approach which utilizes solely the TC best-track data to accommodate as many variables as needed by risk assessment. We try to establish a working procedure from modeling to simulation as objective as possible, with minimal subjective intervention. The entire modeling and simulation process is easy to implement in the R environment for  statistical computing (R Core Team, 2021), and is operable on a moderate desktop computer with tolerable simulation time.

This paper is organized as follows. Section 2 describes the data used for modeling. Section 3 introduces the MFPCA method, the simulation model we developed and the model selection criteria. In section 4 we apply the model to simulate elemental variables of TC for the NA and WNP basin, respectively, and evaluate the performance of the synthetic TCs. We summarize our work with discussions in section 5.

## 2 Data

The only raw material we use to construct the simulation model is the historical best-track (reanalyzed) data of TCs. The data sets for the NA and WNP basin were derived from the Atlantic hurricane database (HURDAT) and Joint Typhoon Warning Center (JTWC), respectively, and were redistributed through the International Best Track Archive for Climate Stewardship (IBTrACS, Knapp et al., 2010). The period since 1980 is generally considered as the modern era when geostationary satellite coverage has been nearly global and polar orbiting satellite data has been more widely available than the prior years. Therefore, we take data from 1980 till the recent year available, which is 2019 for the NA basin and 2018 for the WNP basin, respectively. The TC information in the best-track data includes storm type, date and time, center coordinates longitude/latitude (LON/LAT), MWS, MCP, and average translation speed and direction inferred from center coordinates, recorded every 6 hours. For the two USA agencies, MWS is defined as the maximum 1-minute sustained wind speed at 10 m above the surface. For the WNP basin, MCP is available only since 2001. Only those TCs with their lifetime maximum intensity (LMI) reaching the tropical storm (TS) level (34 kt or 17.5 m $s^{-1}$) or above are chosen as sample observations for modeling. As a result, sample data for the NA and WNP basin are comprised of 513 and 1035 TCs, respectively.

In the following discussion, the zonal and meridional components of a translation velocity (denoted as VX and VY, respectively), derived from the translation speed and direction, are used to describe the TC movement. The seasonality of TC activity can be represented by the annual phase angles of the recorded dates during TC life cycles, in the form of pairs of sine and cosine functions of the phase angles (denoted as SIN and COS, respectively). The ordinal dates of TCs in a year can be retrieved from such pairs of trigonometric functions inversely with simple calculation. The relative lasting time (RLT) of a TC, i.e. the time lapse from the TC genesis divided by the TC lifetime, is used to indicate at which stage of life cycle the TC is. With all the above recorded and derived variables, the spatial-temporal evolution of TCs can be fully described.

## 3 Methods

### 3.1 Multivariate FDA

For a comprehensive introduction to FDA, please refer to Ramsay and Silverman (2005). Here we just briefly review some of the concepts used in this study. A random variable $X = \{X(t), t \in \mathcal{T}\}$ is called functional variable if it takes values in an infinite dimensional space (a functional space), where $\mathcal{T} \subset \mathbb{R}$ is a compact interval. An observation *x* of *X* is called a functional datum. A functional data sample consists of *N* realizations of *X*: $x_1, \cdots, x_N$. Usually, *X* can be viewed as a second order stochastic process in the separable Hilbert space $\mathcal{H}$ of square integrable functions, $L^2(\mathcal{T})$. In practice, functional data are observed discretely, and therefore always come in pairs of the form $(t_{ij}, x_{ij})$ with $x_{ij} = x_i(t_{ij}),\ i = 1, \cdots, N,\ j = 1, \cdots, S_i$. In general, the number and location of $t_{ij} \in \mathcal{T}$ can vary with *i*. Discretized observations have to be transformed into functional data first for subsequent analysis. In most circumstances, interpolation or smoothing methods, e.g. B-splines or smoothing splines, are employed.

Multivariate functional data (MFD) take multiple functions at the same time into account. Each observation unit consists of a fixed number of functions *p*, and is assumed to be a realization of a random process $X = \left(X^{(1)}, \cdots, X^{(p)}\right)$, where $X^{(k)} = \left\{X^{(k)}(t), t \in \mathcal{T}\right\},\ k = 1, \cdots, p$. As only observed discretely, MFD are of the form $\left(t_{ij}^{(k)}, x_{ij}^{(k)}\right),\ i = 1, \cdots, N,\ j = 1, \cdots, S_i,\ k = 1, \cdots, p$. Each element function can be represented separately by its observation points and the observed values. The full MFD sample is a collection of all the *p* element functions.

We assume that a TC in a basin is a realization of the underlying air-sea interactive process responsible for the formation and evolution of the TCs in that basin. Elemental variables of TC along the track, such as the center coordinates LON/LAT, MWS, MCP, etc., recorded at discrete time points during the TC lifetime, constitute the TC MFD. The best-track data are naturally in the form of MFD. Via the multivariate FDA approach, the aspects of a TC throughout its lifetime can be studied as a whole with correlations between them taken into account.

3.2 MFPCA

The TC MFD contain information about not only the TC movement but also the response of TCs to the underlying process. Unlike most existing full-track models that were typically fitted through supervised learning, our innovative model introduces MFPCA method, an unsupervised learning approach that makes full use of the best-track data and requires little to none human intervention. MFPCA is effectively an extension of functional principal component analysis (FPCA) to the multivariate FDA (Ramsay and Silverman, 2005). Here we follow the framework of MFPCA proposed by Happ and Greven (2018). This framework allows for element functions to be defined in different domains possibly with different dimensions. For simplicity we still assume that all the element functions in the model are defined in the same one-dimensional time domain. Like in FPCA, MFPCA aims at a multivariate functional Karhunen-Loève representation of data such that

$$X(t) = \sum_{m=1}^{\infty} \rho_m \psi_m(t)\,, t \in \mathcal{T} \tag{1}$$

where $X(t)$ is multivariate with $\mu(t) = \mathrm{E}[X(t)] = \left(\mathrm{E}\left[X^{(1)}(t)\right], \cdots, \mathrm{E}\left[X^{(p)}(t)\right]\right) = 0$, $\psi_m(t) \in \mathcal{H}$ are complete orthogonal basis of eigenfunctions of covariance operator $\Gamma$ such that

$$\Gamma\psi_m = \nu_m \psi_m \tag{2}$$

where $\nu_m$ are eigenvalues and $\nu_m \to 0$ for $m \to \infty$, and $\rho_m$ are zero mean random variables with $\mathrm{cov}(\rho_m, \rho_n) = \nu_m \delta_{mn}$. Moreover,

$$\mathrm{E}\left[\left\|X(t) - \sum_{m=1}^{M} \rho_m \psi_m(t)\right\|^2\right] \to 0 \ \text{ for } M \to \infty \tag{3}$$

uniformly for $t \in \mathcal{T}$.

The algorithm used in this study starts with a sample of $X$: $x_1, \cdots, x_N$ with its estimated multivariate mean $\hat{\mu}$ subtracted, and consists of four steps:

(1) For each element function $j = 1, \cdots, p$ of $x_i$, create a B-splines representation with $M_j$ basis functions $\hat{\phi}_1^{(j)}, \cdots, \hat{\phi}_{M_j}^{(j)}$ and corresponding coefficients $\hat{\xi}_{i,1}^{(j)}, \cdots, \hat{\xi}_{i,M_j}^{(j)}$. Other choices for function representation can be principal component functions of FPCA or arbitrary basis functions in $L^2(\mathcal{T})$ (Happ and Greven, 2018).

(2) Combine all coefficients into one big matrix $\Xi \in \mathbb{R}^{N \times M_+}$ with $M_+ = M_1 + \cdots + M_p$, the *i*th row of which

$$\Xi_{i,\cdot} = \left(\hat{\xi}_{i,1}^{(1)}, \cdots, \hat{\xi}_{i,M_1}^{(1)}, \cdots, \hat{\xi}_{i,1}^{(p)}, \cdots, \hat{\xi}_{i,M_p}^{(p)}\right) \tag{4}$$

and then estimate the joint covariance matrix $\hat{Z} = \frac{1}{N}\Xi^T\Xi$.

(3) Find eigenvectors $\hat{c}_m$ and eigenvalues $\hat{\nu}_m$ of $\hat{Z}$ for $m = 1, \cdots, M_+$.

(4) The multivariate principal component functions and scores are estimated accordingly by

$$\hat{\psi}_m^{(j)} = \sum_{n=1}^{M_j} [\hat{c}_m]_n^{(j)} \phi_n^{(j)}, \qquad \hat{\rho}_{i,m} = \sum_{j=1}^{p} \sum_{n=1}^{M_j} [\hat{c}_m]_n^{(j)} \hat{\xi}_{i,n}^{(j)} = \Xi_{i,\cdot} \cdot \hat{c}_m, \quad m = 1, \cdots, M_+ \tag{5}$$

respectively.

The multivariate Karhunen-Loève representation of $x_i$ is finally given as

$$x_i = \hat{\mu} + \sum_{m=1}^{M_+} \hat{\rho}_{i,m} \hat{\psi}_m \tag{6}$$

where $\hat{\psi}_m = \left(\hat{\psi}_m^{(1)}, \cdots, \hat{\psi}_m^{(p)}\right)$ having the same multivariate structure of *X*. The R package "MFPCA" (Happ-Kurz, 2020) provides an easy way to implement the above algorithm.

When applied to the best-track data, Step (1) requires that all the TCs have the same lifetime such that they share the same set of B-spline basis functions for each element function of TC MFD. To achieve this, the longest lifetime among all the TCs is set to be the interval $\mathcal{T}$ for the TC MFD. For TCs with lifetime shorter than $\mathcal{T}$, their element functions will be prolonged with constant values after the lysis. Specifically, LON/LAT remains the coordinates of the last observation, MWS is set to be 0 m s$^{-1}$, and MCP is set to be the mean sea-level pressure (MSLP), after the lysis. As a result, all the TCs have exactly the *S* number of 6-hour observation points in the interval $\mathcal{T}$. In addition, for the B-spline representation with an order of 4 (cubic splines, the default choice for most applications), the maximum number of basis functions is $S + 2$. For the *p* element functions of TC MFD, the numbers of basis functions $M_i, i = 1, \cdots, p$ needed are usually less than $S + 2$ and may differ from each other according to their own intrinsic behaviours. However, for the sake of minimal subjective choices, we simply set $M_1 = \cdots = M_p = S + 2$ so that $M_+ = p \times (S + 2)$. For each individual element function, the degree of freedom is obviously redundant with this choice of basis functions and could be optimized. At this stage we keep all the excessiveness for computational simplicity and leave the optimization task to the final order determination stage.

### 3.3 Order determination

Underlying Eq. (6) is a general noisy model for PCA (Jolliffe, 2002, p. 151)

$$X = Z + \epsilon \tag{7}$$

where *Z* and $\epsilon$ are independent *p*-dimensional random vectors for signal and noise, respectively, $\Sigma = \mathrm{var}(Z)$ is a singular matrix with rank $d < p$, and $\mathrm{var}(\epsilon) = \sigma^2 I_p$ where $I_p$ is the identity matrix. The principal components are the projections of *X* onto the first *d* leading eigenvectors of Σ. Here, the order determination problem is to estimate *d*, the rank of Σ. In the context of our MFPCA model, the problem is to estimate an optimal truncation lag $M \leq M_+$ such that Eq. (6) can be approximated by the signal part of *X*:

$$x_i \approx \hat{\mu} + \sum_{m=1}^{M} \hat{\rho}_{i,m} \hat{\psi}_m \tag{8}$$

Here we use the ladle estimator (Luo and Li, 2016) to determine $d$. This estimator combines both the eigenvalues and the bootstrap eigenvector variability of $\hat{\Sigma}$. The idea behind it is based on the fact that when the eigenvalues of a random matrix are far apart, the bootstrap variability of the corresponding eigenvectors tends to be small. On the other hand, this bootstrap variability tends to be large when the eigenvalues are close together. The ladle estimator of the rank $d$ is achieved by minimizing the objective function

$$g_n(k) = f_n(k) + \phi_n(k) \tag{9}$$

where $f_n(k)$ and $\phi_n(k)$ represent the bootstrap eigenvector variability and sample eigenvalues, respectively, $n$ is the number of bootstrap samples (half the number of data by default), $k = 0, \cdots, p-1$. Refer to Eqs. (4) and (5) in Luo and Li (2016) for the mathematical forms of the two terms. The eigenvalue term $\phi_n(k)$ is large when $k < d$; the eigenvector term $f_n(k)$ is large when $k > d$; but both are small when $k = d$. Therefore, $g_n(k)$ is expected to reach its minimum approximately at $d$. The function curve of $g_n(k)$ resembles a ladle, hence the name. The R code provided in the Supplementary material of Luo and Li (2016) can be adapted to estimate the optimal truncation lag $M$ in the MFPCA context.

### 3.4 Full-track simulation

#### 3.4.1 Simulation model

Once the order $M$ is determined by the ladle estimator, the multivariate functional representation of TC data can be written as

$$x_i = \hat{\mu} + \sum_{m=1}^{M} \hat{\rho}_{i,m}\hat{\psi}_m + \sum_{m=M+1}^{M_+} \hat{\rho}_{i,m}\hat{\psi}_m \tag{10}$$

which is a mixed model by analogy: the first two terms on the right-hand side are of fixed effect, the last term is of random effect that can be utilized for simulation. The simulation procedure starts with randomly choosing a historical observation $x_i$, draws a sample of multivariate normal $\left(\rho_{i,M+1}, \cdots, \rho_{i,M_+}\right)$ with zero means and $\mathrm{cov}\left(\rho_{i,m}, \rho_{i,n}\right) = \nu_m \delta_{mn}$ where $m, n = M+1, \cdots, M_+$, and substitutes the sample for the estimated $\hat{\rho}_{i,m}, m = M+1, \cdots, M_+$ in the last term to finally synthesize a full-track TC. Unlike regression-based simulations in most previous works, this approach still relies on historical TCs to serve as "seeds" to grow more analogues, somewhat similar to the random perturbation method in Casson and Coles (2000), but is much more comprehensive and exhaustive in data utilization and information extraction.

For TC risk assessment, it is often desirable that the synthetic TCs are generated on an annual basis so that the return periods of extreme events can be estimated. To achieve this, we first sample the number of TCs in a year using a fitted TC occurrence model (see below) in advance, and then randomly draw that number of TCs from the whole historical data as the candidates for applying the above procedure to simulate TCs for that year. This step is repeated to simulate a series of annual TCs until the desired length of simulation period is reached.

#### 3.4.2 Occurrence model

For count data like the annual number of TCs, Poisson distribution is usually the preferred model in which the expected value stands for the annual rate of occurrence. Poisson distribution has the equidispersion property, i.e., its mean is equal to its variance. In real data, however, such equidispersion is rarely satisfied. In most situations, the variance is greater than the mean, a phenomenon known as overdispersion and otherwise known as underdispersion. Interestingly, the annual TC occurrence in the NA basin is overdispersed, whereas that in the WNP basin is underdispersed (section 4.2). There are various alternative models for overdispersed count data, such as the negative binomial distribution, but much fewer models for underdispersed count data. Vickery et al. (2000) used the negative binomial distribution to sample the annual number of TCs in the NA basin. For the WNP basin, however, Poisson and negative binomial distributions are actually not applicable; they may well overestimate the annual variation of TC occurrence.

Fortunately, there are flexible generalizations of the Poisson distribution called Conway–Maxwell–Poisson (CMP) distributions for modeling overdispersed or underdispersed count data (Shmueli et al., 2005), of which Poisson process is a special case. The probability mass function of the CMP distribution with rate $\lambda$ and dispersion $\nu$ takes the form

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu}\frac{1}{Z(\lambda, \nu)}, \qquad y = 0, 1, 2, \cdots \tag{11}$$

where $\lambda > 0$, $\nu \geq 0$, $Z(\lambda, \nu) = \sum_{y=0}^{\infty} \frac{\lambda^y}{(y!)^\nu}$ is a normalizing constant. $\nu < 1$, $\nu = 1$ and $\nu > 1$ lead to overdispersion, equidispersion (Poisson distribution) and underdispersion, respectively. Huang (2017) suggested a reparameterization of CMP distributions with mean $\mu$ and dispersion $\nu$, which is more suitable for fitting Generalized Linear Models. As a result, the variance of the CMP distribution is a function of $\mu$ and $\nu$, or $V(\mu, \nu)$. In this work, we fit the CMP distribution in the $\mu$-$\nu$ form to the annual TC number sequence as the occurrence model with the help of the R package "mpcmp" (Fung et.al, 2020), the R implementation of Huang (2017).

## 4 Results

### 4.1 Pre-processing of best-track data and post-processing of simulations

Prior to the MFPCA, the best-track data with lifetime shorter than $\mathcal{T}$ are patched with proper values in a manner described in section 3.2. For the NA basin, MCP values after the lysis are set to be 1021.36 hPa, the MSLP estimated using the Dvorak wind-pressure relationship (WPR): $\text{MSLP} = 1021.36 - 0.36 \times \text{MWS} - (\text{MWS}/20.16)^2$ (Knaff and Zehr, 2007). For the WNP basin, MCP is not included in the MFD for modeling. If the MCP simulations are also desired, they can be derived from the MWS simulations using an appropriate WPR. In doing so, however, the complexity of WPRs from various agencies should be aware of (Kueh, 2012; Knapp et al., 2013).

For TC risk assessment, variables to be simulated are center coordinates LON/LAT, MCP or MWS, SIN and COS. The last two are used to retrieve the ordinal dates of TCs. If the translation speed and heading direction are needed, they can be derived from the LON/LAT simulations. Due to the randomness in simulations, ordinal dates retrieved from the SIN/COS simulations may not be strictly regular step functions as recorded dates of observations (see example below). However, in TC risk assessment, the impact of the seasonality on TC activity is typically measured on monthly or even quarterly basis, for which the simulated dates are accurate enough to use. As such, all simulated ordinal dates remain as-is without further adjustment, and all simulation years are treated as non-leap years in the simulation model.

The MWS values in the best-track data were estimated in multiples of 5 kt (1 kt ≈ 0.514 m $s^{-1}$), with the minimal MWS estimate of 10 kt. The freshly simulated MWS is, however, continuous and includes MWS values below 10 kt. To ensure that the synthetic TCs are formally consistent with the best-track data, we round the simulated raw MWS into multiples of 5 kt and then remove the track point at which the rounded MWS is equal to or less than 10 kt. As a result, a freshly simulated track that is intermitted with very low MWS can be split into a few shorter track segments. Another restriction that the LMI must reach the TS level or above is applied subsequently to remove storms with LMI strength of tropical depression (TD) or weaker, as TDs rarely cause statistically meaningful economic and/or insurance losses.

Figure 1 illustrates the above pre- and post-processing procedures, using the hurricane Irma (2017242N16333) as an example. The time span for observation is 00Z 30 August to 12Z 13 September 2017, a period of 348 hours. In order to prepare the TC MFD for modeling, all the variable records are extended with constant values to 570 hours, the interval $\mathcal{T}$ on which the MFD are defined for the NA basin. The simulation is randomly generated by using Irma as the "seed". Simulated track points with MWS equal to or less than 10 kt are removed, resulting in two track segments. The one with LMI less than 34 kt is also discarded. The remaining one finally becomes a synthetic TC. Note that the simulated dates are not a strictly regular step function as recorded dates for observation (Fig. 1e).

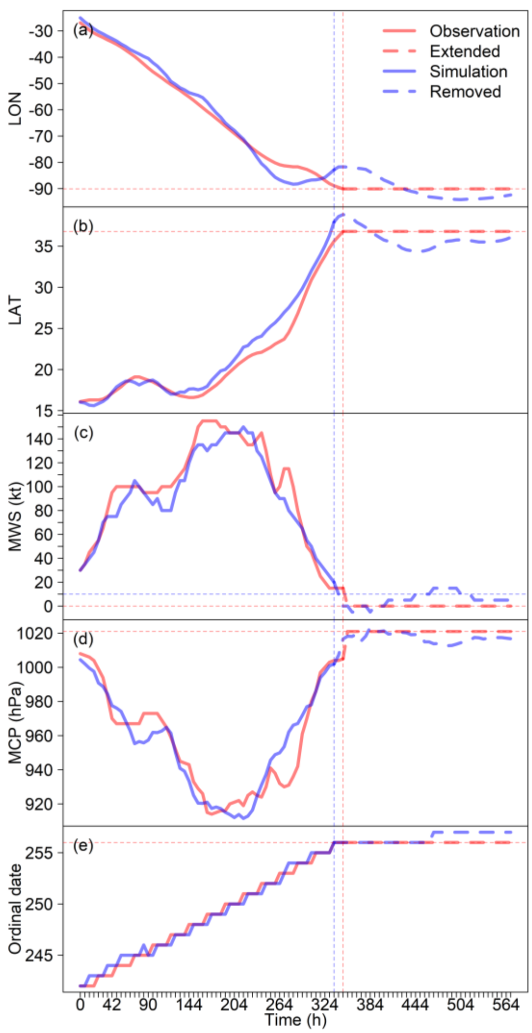



**Figure 1.** Example of pre- and post-processing procedures. The observation is hurricane Irma (2017242N16333) recorded from 00Z 30 August to 12Z 13 September 2017, a period of 348 hours (red solid curves). The vertical red dashed line indicates the time point of lysis. By extending all the variable records with constant values to 570 hours (red dashed curves), the observation is transformed into a multivariate functional datum for modeling. Blue curves are a simulation by using Irma as the "seed". The vertical blue dashed line indicates the last time point at which the simulated MWS is greater than the threshold of 10 kt (indicated by the horizontal blue dashed line in 1c). Blue dashed curves are removed in post-processing. The remaining blue solid curves constitute a synthetic TC.

4.2 Model summary

Table 1 summarizes the primary information about the model fitting and simulation. In this study, the TC MFD for the NA basin consists of nine element functions while that for the WNP basin consists of eight, due to the shortage of the MCP records in the WNP basin. With MFPCA, the TC MFD can be represented as Eq. (6) which serves as the fitted model in this study. If we divide the summation in the right-hand side of Eq. (6) into two parts, the first $M$ leading eigenvectors as the signal part and all the others as the noise part, it turns out to be the simulation model expressed as Eq. (10). Only about 54 % and 60 % of the total $M_+$ $(= p \times (S+2))$ eigenvalues are nonzero for the NA and WNP basin, respectively. As we pointed out in section 3.2, $(S+2)$ number of degrees of freedom for each element function are obviously redundant due to the fact that most of the actual TC lifetimes are less than $\mathcal{T}$, hence the rank of the joint covariance matrix $\hat{Z}$, or correspondingly the number of nonzero eigenvalues, is much smaller than $M_+$. However, by means of the ladle estimator, only the first 22 leading eigenvectors that explains about 93% of total variance are recognized to constitute the signal part of the simulation model, coincidently for both the two basins (Fig. 2). The rest of eigenvectors with nonzero eigenvalues then constitute the noise part.

**Table 1** Summary of data, model fitting and simulation

| | NA | | WNP | |
|---|---|---|---|---|
| | Observation | Simulation | Observation | Simulation |
| Period (years) | 40 (1980–2019) | 1000 | 39 (1980–2018) | 1000 |
| Total number of TCs | 513 | 12931 | 1035 | 26578 |
| Occurrence mean $\mu$ | 12.8 | 12.9 | 26.5 | 26.6 |
| Occurrence variance $V(\mu,\nu)$ | 22.3 | 24.3 | 21.6 | 22.9 |
| Occurrence dispersion $\nu$ | 0.56 | 0.51 | 1.23 | 1.16 |
| No. of element functions $p$ | 9 (LON, LAT, MWS, MCP, SIN, COS, VX, VY and RLT) | | 8 (LON, LAT, MWS, SIN, COS, VX, VY and RLT) | |
| No. of track points $S$ during the lifetime $\mathcal{T}$ | 96 | | 104 | |
| No. of total eigenvectors $M_+$ | 882 | | 848 | |
| No. of nonzero eigenvalues | 472 | | 512 | |
| Optimal truncation lag $M$ (Percentage of total variance) | 22 (92.9%) | | 22 (92.8%) | |

The fitted CMP distributions for the two basins reveal that the annual occurrence of TC in the NA basin is overdispersed ($\nu < 1$), whereas that in the WNP basin is underdispersed ($\nu > 1$). Note that the dispersion is roughly but not exactly the simple ratio of mean to variance. Such difference in the dispersion property of TC occurrence between the two basins may imply that the TC-environmental conditions modulating the TC occurrence is more stable in the WNP basin than in the NA basin.

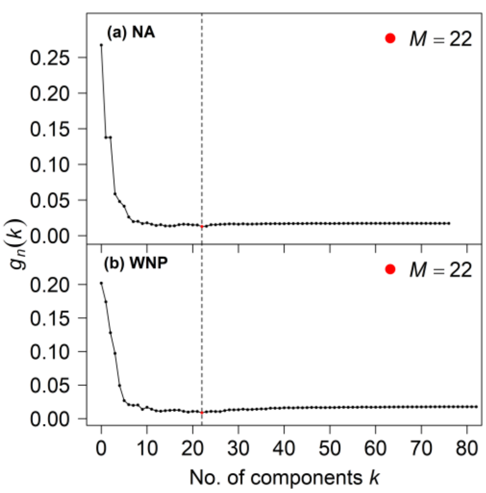



**Figure 2.** Ladle estimates of the optimal truncation $M$ in the simulation model for the two basins, respectively. Both the results are 22, coincidently.

4.3 Model validation

4.3.1 Spatial pattern and annual occurrence

In order to validate the model, we simulate 1000-year worth of TCs for each of the two basins and then compare with their respective best-track data. best-track records with MWS equal to or less than 10 kt are also removed to make a fair comparison with synthetic

TCs. To present a general picture of the performance of the described approach, Figure 3 shows the spatial patterns of the fitted and simulated vs. the best-track (observed) TC tracks, for both the NA (3a and 3b) and WNP (3c and 3d) basins, respectively. By using the total $M_+$ eigenvectors, the fitted model can faithfully reconstruct the best-track data. It can be seen that the observed and fitted TC tracks are overlapped so well that they can hardly be distinguished one from the other. Simulated TC tracks are much denser than the observed TC tracks, but still resemble them in spatial pattern, curvature and re-curvature, genesis and lysis features.

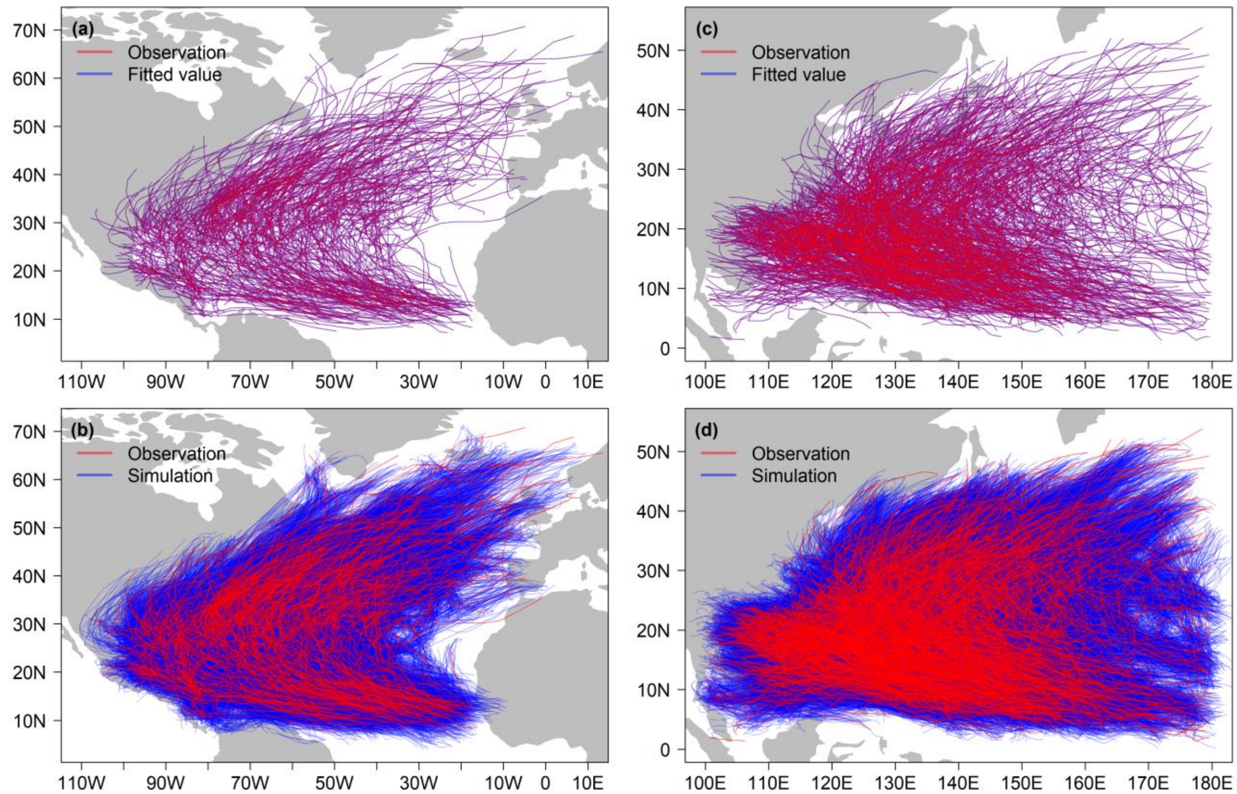



**Figure 3.** Comparisons of fitted values and simulations to observations (best-tracks) for the NA (left panes) and WNP (right panes) basin, respectively. Note that observations and fitted values are actually overlapped.

The synthetic TCs also well capture the historical features of the annual TC occurrence. Comparison of the CMP distributions fitted to observations and simulations (Table 1) shows that for each basin, the occurrence mean of simulations is quite close to that of observations; the occurrence variance of simulations is a little higher than that of observations, which is probably due to the removal of track points in post-processing that may result in track splitting or track removal.

### 4.3.2 Marginal distributions

Next we examine the performance of synthetic TCs in more detail, by comparing the empirical distributions of LON, LAT, MWS, MCP, ordinal dates and lifetime from simulations to those in the observations. Figures 4 and 5 show the comparisons in terms of empirical probability density function (EPDF) and empirical cumulative distribution function (ECDF) for the NA and WNP (for which MCP is not available) basin, respectively. Histograms represent

EPDFs, and curves represents ECDFs corresponding to EPDFs. It can be seen that for each basin, EPDFs for observations and simulations are almost overlapped while the two ECDFs are quite close to each other. The agreement between the observations and simulations show the capability of the simulation model in capturing the marginal distribution features of the TC variables. Discrepancies such as in the lower/upper tail of MWS/MCP distributions are mostly related to the noises just exceeding the 10-kt threshold and thus can be ignored. As for the seasonality and lifetimes of the simulated TCs, comparisons to the observations show satisfying results as well. Particularly for the seasonality of TC activity, simulations almost reconstructed the distribution of ordinal dates from the observations (Fig. 4d and Fig. 5d), which is helpful for assessing the TC risk on a monthly or even a shorter-term basis.

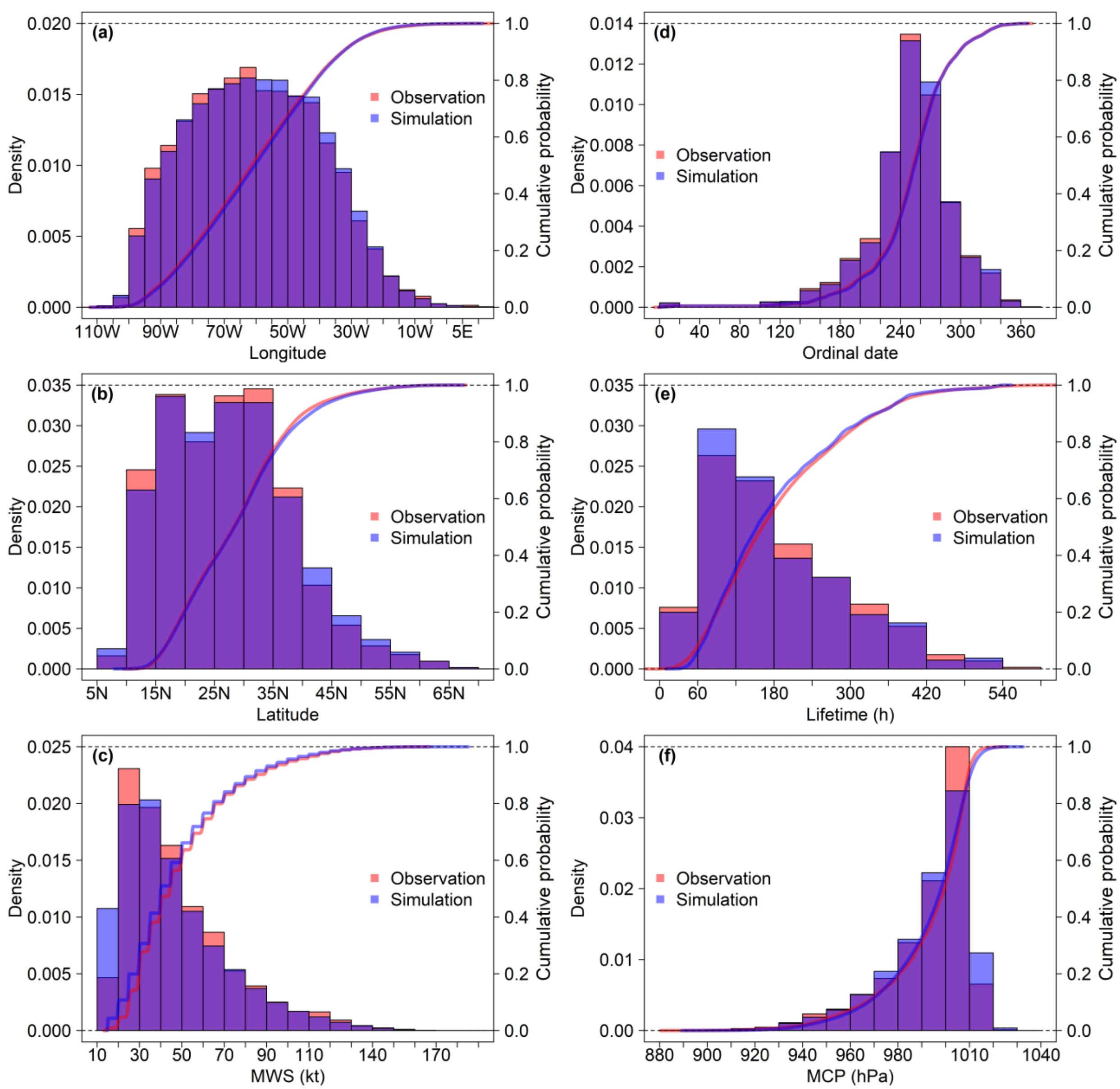



**Figure 4.** Comparisons between observations and simulations in terms of EPDF (histogram) and ECDF (curve) of TC variables for the NA basin. EPDF and ECDF values are indicated by the left and right ordinate, respectively.

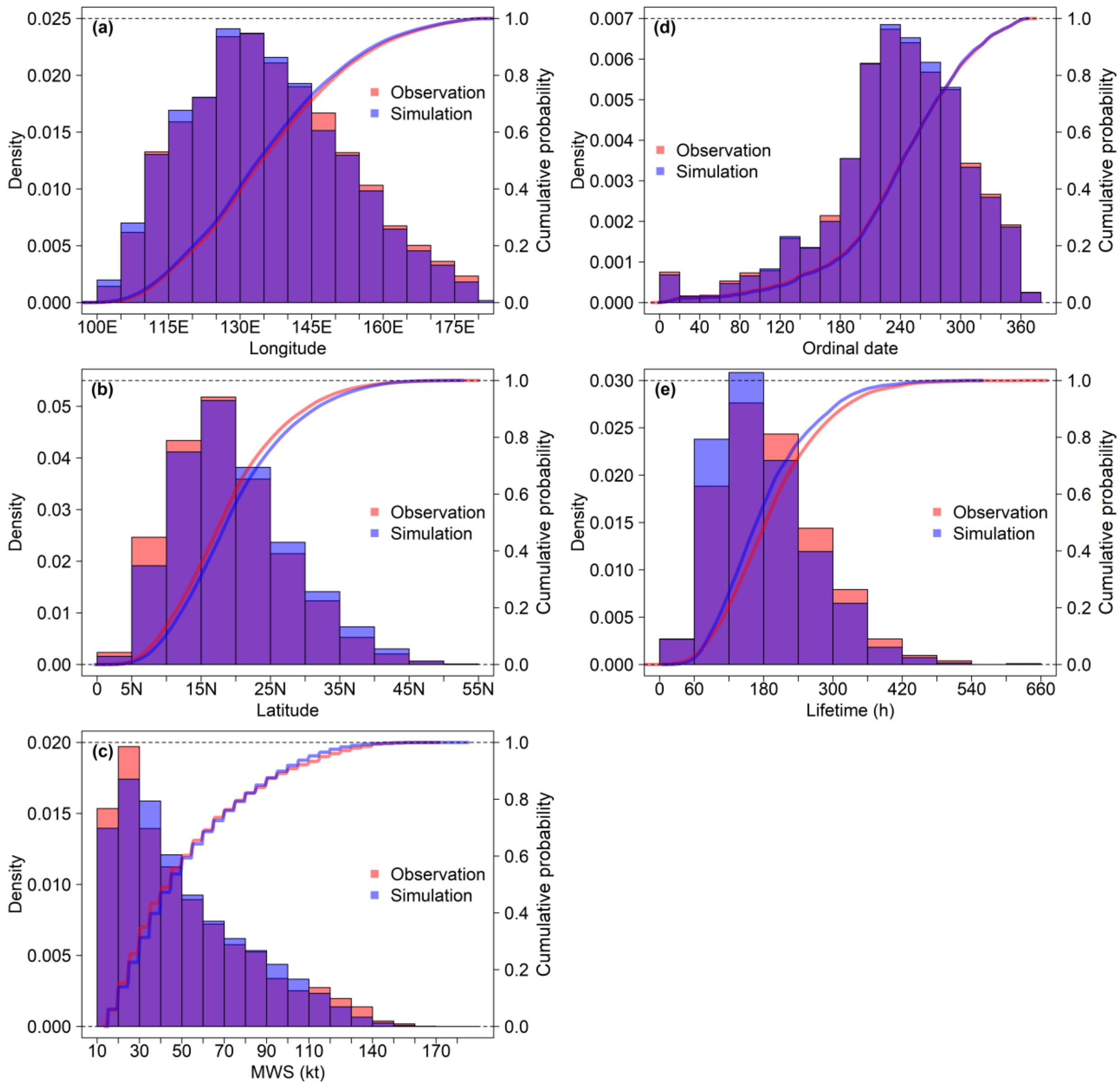



**Figure 5.** Same as Fig. 4 but for the WNP basin. Note that MCP is not available for the WNP basin.

4.3.3 Spatial distributions

We then check further the joint distributions of the TC variables with a focus on the spatial distribution features of TC density and intensity. Joint distributions of TC variables are estimated using the R package "ks" (Duong, 2021) by means of multivariate kernel smoothing (Chacon and Duong, 2018). First, we compare the annual mean spatial densities of track points from observations and simulations, derived from the joint distribution of LON and LAT (Figs. 6a–6d). It can be seen that for each basin, the spatial density of simulations matches well as a whole with that of the observations, even though the time span of simulations is much longer than that of the observations. For the assessment of economic and insurance losses, the TC landfall locations are of particular interest. Figures 6e and 6f compare the empirical distributions of the observed and simulated landfall locations as functions of longitude along the thick coastline for the two basins, respectively. These two coastlines are more liable to be attacked by TCs among others in their respective basins. Histograms represent EPDFs of landfall locations, and curves represents ECDFs corresponding to EPDFs. Once again, a high consistency exists between observations and simulations in terms of TC landfall locations. Particularly for the WNP basin (Fig. 6f) where the TC landfalls are more frequent than in the NA basin, there are more data available for modeling, resulting in reduced model uncertainty and smaller simulation bias.

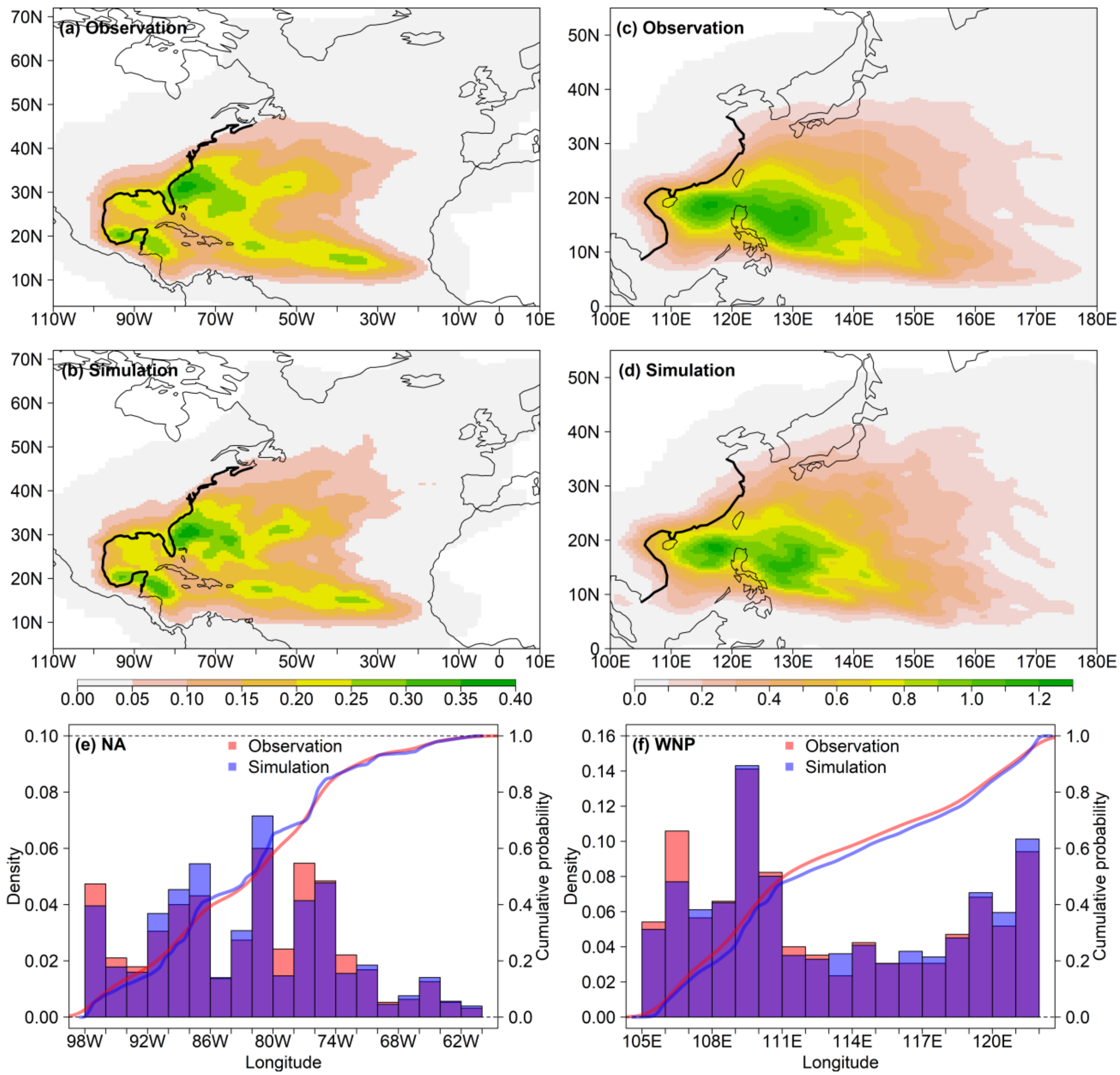



**Figure 6.** Annual mean spatial densities of track points (unit: degree$^{-2}$ yr$^{-1}$) from observations and simulations, and comparison between observations and simulations in terms of EPDF (histogram) and ECDF (curve) of landfall locations along the thick coastline with respect to longitude, for the NA (left panes) and WNP (right panes) basin, respectively.

Next, we examine the spatial distribution of the simulated MWS in terms of return periods of Saffir-Simpson hurricane intensity categories for the two basins. The wind speed ranges for categories 1–5 (Cat. 1–5) are 64 –82, 83–95, 96–112, 113–136 and > 137 kt, respectively. Figure 7 compares the return periods of simulated MWS using the lower limits of Cat. 1, 3 and 5 as thresholds to those from the observations, respectively, for the NA basin. Figure 8 is the same as Fig. 7 but for the WNP basin. With these statistics, the spatial distribution of MWS as a function of LON and LAT can be outlined. These results show that, although the time span of simulations is much longer than that of observations, simulations do not substantially deviate from observations in terms of statistical properties, which is essential for synthetic TCs to be used for risk assessment.

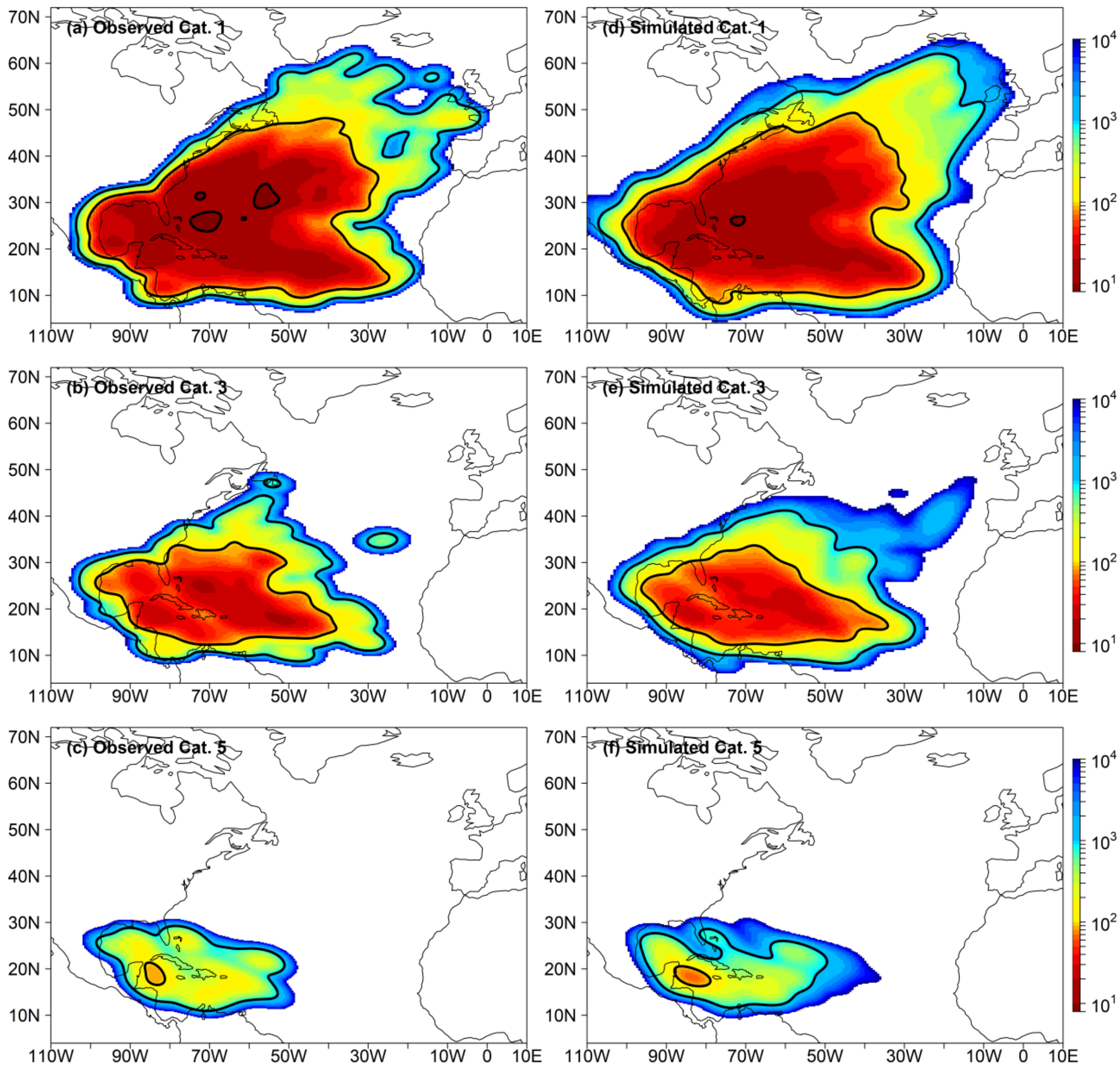



**Figure 7.** Return periods of observed and simulated MWS using the lower limits of Cat. 1, 3 and 5 as thresholds, respectively, for the NA basin. Black contours indicate 10-, 100- and 1000-year return periods.

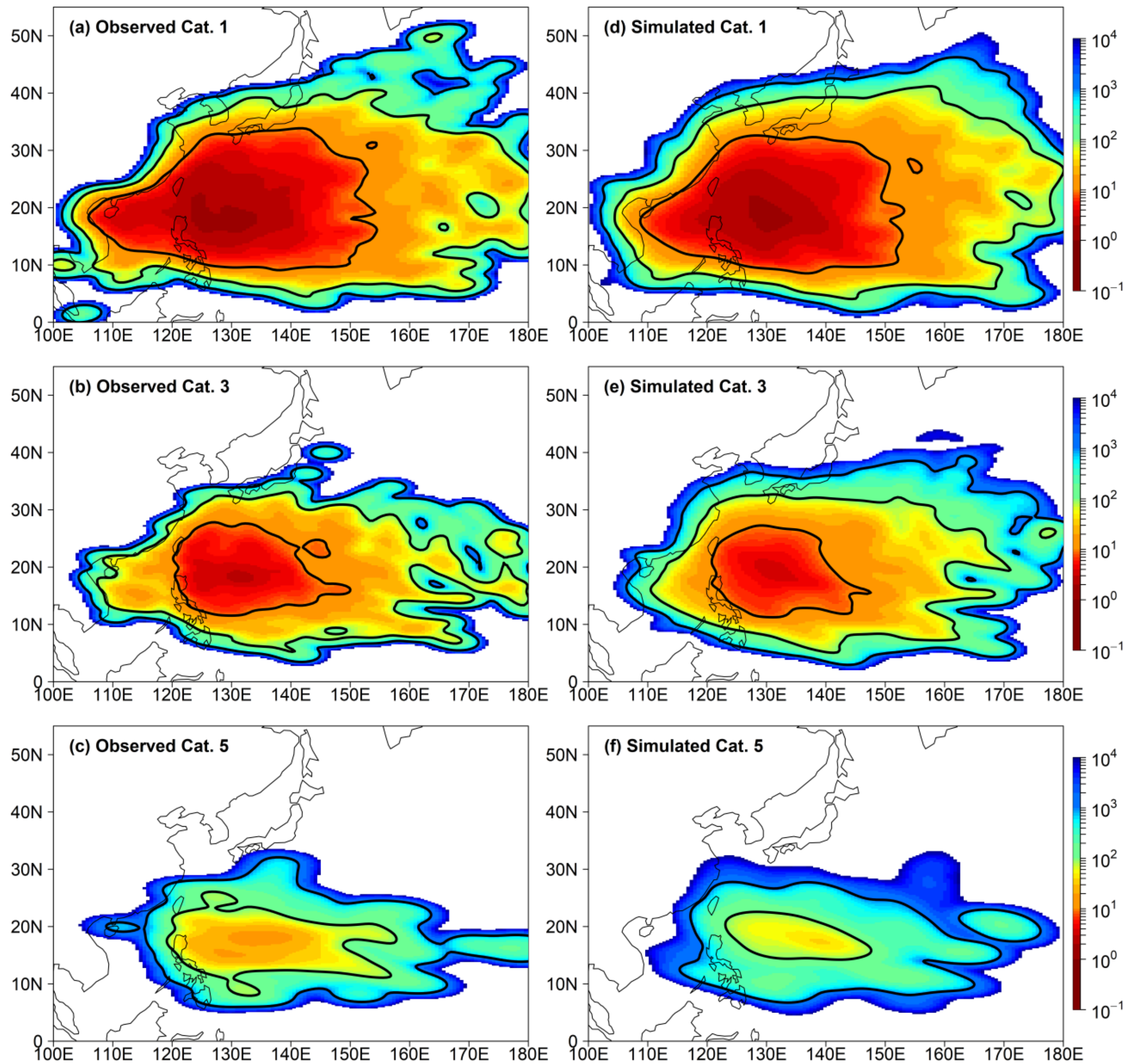



**Figure 8.** Same as Fig. 7 but for the WNP basin.

#### 4.3.4 Intensity extremes

Ideally, synthetic TCs for risk assessment should be able to present cases stronger than all observations, while they are still consistent with observations in terms of distributional features. Previous comparisons have shown that the latter requirement is well satisfied. We finally focus on the TC intensity extremes to complete the validation. Figures 9a–9d compares maxima of MWS over each individual 1°×1° grid squares from simulations to those from observations, for the NA and WNP basin, respectively. Obviously, simulated maxima are generally greater than observed ones as we desired. However, the maximum potential intensity (MPI) of TC is restricted by the TC-environmental conditions. Knaff et al. (2005) set 185 kt as the upper bound for MPI in the WNP basin according to an empirical relationship between MPI and SST. Similar relationship was also found in the NA basin (DeMaria and Kaplan, 1994). Coincidently, for both the two basins the simulated maximum MWS is 185 kt. This seeming coincidence actually indicates that the simulation model does grasp the empirical MPI by mining the best-track data.

If picking LMI as independent extreme values for each basin, then return periods of LMI from observations and simulations using unique LMI values as thresholds can be compared as shown in Figs. 9e and 9f, for the NA and WNP basin, respectively. It can be seen that, for each basin within the time span of observations, return periods from simulations are quite consistent with those from observations, although for return periods shorter than 10 years, LMIs are a little underestimated by simulations. However, for each basin, simulated LMIs that are greater than the observed maximum LMI all have return periods beyond the time span of observations, manifesting the capability of this approach to suggest potential risks for assessment.

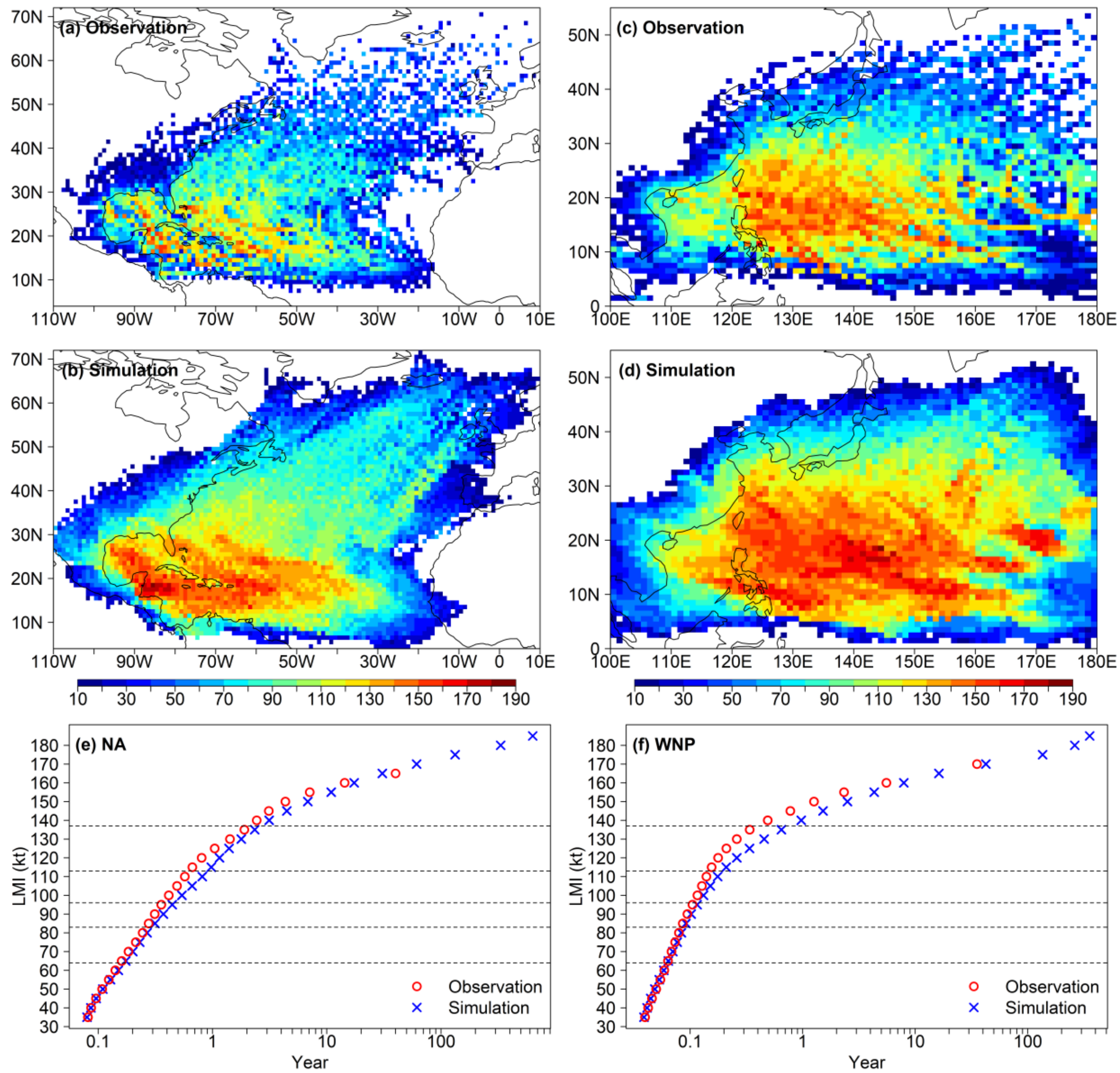



**Figure 9.** Maxima of MWS over each individual 1°×1° grid squares from observations and simulations, and comparison between observations and simulations in terms of return periods of LMI, for the NA (left panes) and WNP (right panes) basin, respectively. Horizontal dashed lines indicate lower limits of Cat. 1–5, respectively.

## 5 Summary and Discussion

In this study, we present a MFPCA approach to the full-track simulation of TC for risk assessment. The novelty of this approach is that elemental variables of TC along the track necessary for risk assessment, such as center coordinates LON/LAT, MWS and/or MCP and ordinal dates, can be simulated simultaneously at one go, yet using solely the best-track data with no data supplemented from any other sources. The simulation model is flexible and extensible, depending on the data availability for the basin of interest. With the help of ladle estimator, the optimal model is determined objectively so that the whole procedure can be programmed with little manual intervention needed.

We also introduce a novel TC occurrence model using CMP distributions, of which underdispersion, equidispersion and overdispersion are special cases. The annual occurrence of TC in the NA basin is overdispersed, whereas that in the WNP basin is underdispersed. This phenomenon might be an indicator of the variability of the TC-environmental conditions modulating the TC occurrence, deserving of further study. Within the framework of CMP distributions, annual TC occurrence in different basins with different dispersion features can be modelled uniformly and be compared with each other. Combining with the occurrence model, the full-track simulation of TC can be proceeded on an annual basis.

The performance of synthetic TCs is validated by comparison to the best-track data, in terms of annual occurrence, marginal distributions of TC variables, spatial distributions of TC density and intensity, and intensity extremes. High consistency between observations and simulations presents in distributional features for comparison, even though the two data sets have quite unbalanced time spans. As for intensity extremes, synthetic TCs with LMI greater than all observations also have return periods beyond the time span of observations, meanwhile they are still restrained from being unrealistic. These results show that the simulation model is able to generate synthetic TCs consistent with observations in terms of distributional features, but of large-enough size to include potentially extremer cases, which is essential for risk assessment.

There are some local biases in different aspects revealed through comparisons. The main source of such biases is apparently the truncation of total hundreds of eigenvectors to only a few leading ones of them to constitute the simulation model. Figure 3 actually demonstrates the effect of such a truncation. Nonetheless, just because when viewed as MFD, basin-wide best-track data can be encoded by only a few leading eigenvectors, the convenience of this approach is manifest.

Moreover, all the algorithms are implemented using the freely available R statistical software packages, with a little programming in the R language. The modeling and simulation process is fully objective and automated, which greatly improves the modeling efficiency and reduces turnaround time, especially when newly available TC data are incorporated periodically into the model. In a word, our proposed approach to the full-track simulation of TC not only generates high-performance synthetic TCs for risk assessment, but also makes this work simpler. These synthetic TCs can be used in conjunction with wind field and engineering vulnerability models to estimate economic and insurance losses for governments and insurance/reinsurance industry.

Since the simulation model is purely empirical without external dynamic factors incorporated, it is not intended to be an all-purpose alternative to environmentally forced models such as those described in Emanuel et al. (2008), Lee et al. (2018) or Jing and Lin (2020), particularly when these models are used for assessing TC risks projected by climate change scenarios. To some extent, this approach is still capable of assessing TC risks modulated by some climate variability, by sampling historical TCs subject to different phases such as El Nino and La Nina separately during the simulation. A possible extension is the joint simulation of TCs in different basins, such as the NA and East Pacific (EP) basins, by means of joint modeling of annual TC occurrences in different basins. In doing so, TCs in different basins are simulated synchronously with the inter-basin correlation of TC activity considered. This is helpful for insurance/reinsurance companies to setup uniform standards for assessing risks for different regions. These ideas will be implemented in our future work.

## Acknowledgments

This study has been supported in part by the National Natural Science Foundation of China under Grant 41875057, 41675044 and 41730960.

## Data availability statement

The tropical cyclone best-track data set IBTrACS can be accessed from the National Climatic Data Center (https://www.ncdc.noaa.gov/ibtracs/). The synthetic tropical cyclone data sets analyzed in section 4 are available through http://doi.org/10.5281/zenodo.4580315.

## References

Aberson, S. D. (1998). Five-day tropical cyclone track forecasts in the North Atlantic basin. *Weather and Forecasting*, *13*(4), 1005–1015. https://doi.org/10.1175/1520-0434(1998)013<1005:FDTCTF>2.0.CO;2

Casson, E., & Coles, S. (2000). Simulation and extremal analysis of hurricane events. *Journal of the Royal Statistical Society*: *Series C* (*Applied Statistics*), *49*, 227–245. https://doi.org/10.1111/1467–9876.00189

Chacon, J. E., & Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications*. Boca Raton: Chapman &Hall/CRC.

Chen, Y., & Duan, Z. (2018). A statistical dynamics track model of tropical cyclones for assessing typhoon wind hazard in the coast of southeast China. *Journal of Wind Engineering and Industrial Aerodynamics*, 172, 325–340. https://doi.org/10.1016/j.jweia.2017.11.014

DeMaria, M., & Kaplan, J. (1994). Sea surface temperature and the maximum intensity of Atlantic tropical cyclones. *Journal of Climate*, 7, 1324–1334. https://doi.org/10.1175/1520-0442(1994)007<1324:SSTATM>2.0.CO;2

Duong, T. (2021). *ks*: *Kernel Smoothing*. R package version 1.12.0. https://CRAN.R-project.org/package=ks

Emanuel, K. (2004). Tropical cyclone energetics and structure. In E. Federovich, R. Rotunno, B. Stevens (Eds.), *Atmospheric Turbulence and Mesoscale Meteorology*: *Scientific Research Inspired by Doug Lilly* (pp. 165–192). Cambridge University Press.

Emanuel, K., Ravela, S., Vivant, E., & Risi, C. (2006). A statistical deterministic approach to hurricane risk assessment. *Bulletin of the American Meteorological Society*, *87*, s1–s5. https://doi.org/10.1175/BAMS-87-3-Emanuel

Emanuel, K., Sundararajan R., & Williams, J. (2008). Hurricanes and global warming: Results from downscaling IPCC AR4 simulations. *Bulletin of the American Meteorological Society*, *89*, 347–368. https://doi.org/10.1175/BAMS-89-3-347

Fung, T., Alwan, A., Wishart, J., & Huang, A. (2020). *mpcmp*: *Mean-parametrized Conway–Maxwell Poisson Regression*. R package version 0.3.6. https://CRAN.R-project.org/package=mpcmp

Happ, C. & Greven, S. (2018). Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, *113*(522), 649–659. https://doi.org/10.1080/01621459.2016.1273115

Happ-Kurz, C. (2020). MFPCA: *Multivariate Functional Principal Component Analysis for Data Observed on Different Dimensional Domains*. R package version 1.3-6. https://CRAN.R-project.org/package=MFPCA

Hall, T. M., & Jewson, S. (2007). Statistical modelling of North Atlantic tropical cyclone tracks. *Tellus A*, *59*, 486–498. https://doi.org/10.1111/j.1600-0870.2007.00240.x

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, *17*(6), 359–380. https://doi.org/10.1177/1471082X17697749

James, M. K., & Mason, L. B. (2005). Synthetic tropical cyclone database. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *131*(4), 181–192. https://doi.org/10.1061/(ASCE)0733-950X(2005)131:4(181)

Jing, R., & Lin, N. (2020). An environment-dependent probabilistic tropical cyclone model. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001975. https://doi.org/10.1029/2019MS001975

Jolliffe, I. T. (2002). *Principal Component Analysis*, Second Edition. New York, NY: Springer.

Knaff, J. A., C. R. Sampson, & DeMaria, M. (2005). An operational statistical typhoon intensity prediction scheme for the western North Pacific. *Weather and Forecasting*, *440*(20), 688–699. https://doi.org/10.1175/WAF863.1

Knaff, J. A., & Zehr, R. M. (2007). Reexamination of Tropical Cyclone Wind–Pressure Relationships. *Weather and Forecasting*, *22*(1), 71-88. https://doi.org/10.1175/WAF965.1

Knapp, K. R., Knaff, J. A., Sampson, C. R., Riggio, G. M., & Schnapp, A. D. (2013). A Pressure-Based Analysis of the Historical Western North Pacific Tropical Cyclone Intensity Record. *Monthly Weather Review*, *141*(8), 2611-2631. https://doi.org/10.1175/MWR-D-12-00323.1

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The international best track archive for climate stewardship (IBTrACS). *Bulletin of the American Meteorological Society*, *91*, 363–376. https://doi.org/10.1175/2009BAMS2755.1

Kriesche, B., Weindl, H., Smolka, A., & Schmidt V. (2014). Stochastic simulation model for tropical cyclone tracks, with special emphasis on landfall behavior. *Natural Hazards*, *73*, 335–353. https://doi.org/10.1007/s11069-014-1075-x

Kueh, M. (2012). Multiformity of the tropical cyclone wind-pressure relationship in the western North Pacific: discrepancies among four best-track archives. *Environmental Research Letters*, *7*(2), 024015. https://doi.org/10.1088/1748-9326/7/2/024015

Lee, C.-Y., Tippett, M. K., Sobel, A. H., & Camargo, S. J. (2018). An environmentally forced tropical cyclone hazard model. *Journal of Advances in Modeling Earth Systems*, *10*, 223–241. https://doi.org/10.1002/2017MS001186

Li, S. H. & Hong, H. P. (2016). Typhoon wind hazard estimation for China using an empirical track model. *Natural Hazards*, *82*, 1009–1029. https://doi.org/10.1007/s11069-016-2231-2

Luo, W., & Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, *103*, 875–887. https://doi.org/10.1093/biomet/asw051

Nakamura, J., Lall, U., Kushnir, Y., & Rajagopalan, B. (2015). HITS: Hurricane Intensity and Track Simulator with North Atlantic Ocean Applications for Risk Assessment. *Journal of Applied Meteorology and Climatology*, *54*(7), 1620–1636. https://doi.org/10.1175/JAMC-D-14-0141.1

R Core Team (2021). *R*: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*, Second Edition. New York, NY: Springer,

Rekabdarkolaee, H. M., Krut, C., Fuentes, M., & Reich, B. J. (2019). A Bayesian multivariate functional model with spatially varying coefficient approach for modeling hurricane track data. *Spatial Statistics*, *29*, 351–365. https://doi.org/10.1016/j.spasta.2018.12.006

Rumpf, J., Weindl, H., Höppe, P., Rauch, E., & Schmidt, V. (2007). Stochastic modelling of tropical cyclone tracks. *Mathematical Methods of Operational Research*, *66*(3), 475–490. https://doi.org/10.1007/s00186-007-0168-7

Rumpf, J., Weindl, H., Höppe, P., Rauch, E., & Schmidt, V. (2009). Tropical cyclone hazard assessment using model-based track simulation. *Natural Hazards*, *48*(3), 383–398. https://doi.org/10.1007/s11069-008-9268-9

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S. & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society*: *Series C* (*Applied Statistics*), *54*: 127-142. https://doi.org/10.1111/j.1467-9876.2005.00474.x

Vickery, P. J., Masters, F. J., Powell, M. D., & Wadhera, D. (2009). Hurricane hazard modeling: The past, present, and future. *Journal of Wind Engineering and Industrial Aerodynamics*, *97*, 392–405. https://doi.org/10.1016/j.jweia.2009.05.005

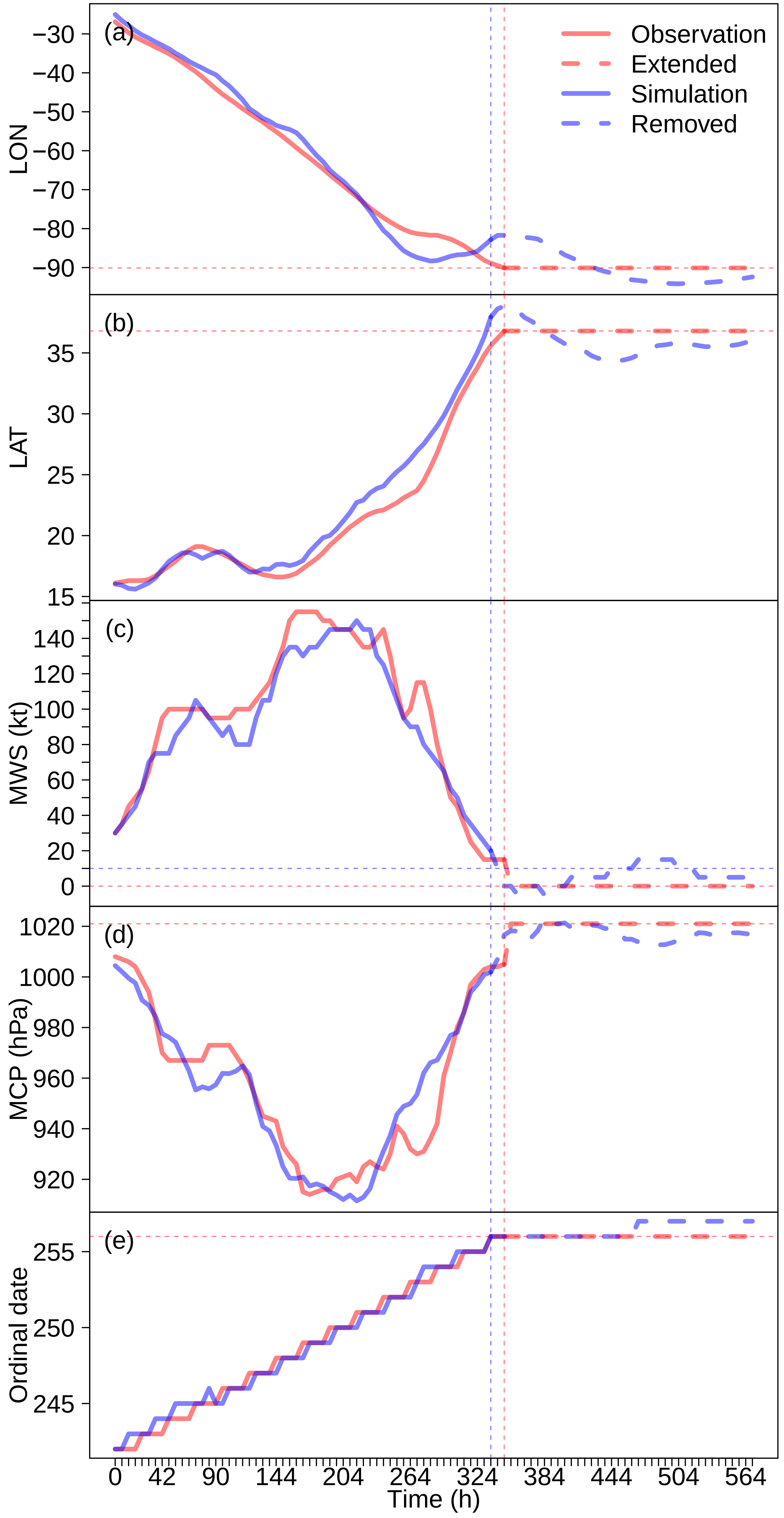
Vickery, P. J., Skerlj, P. F., & Twisdale, L. A. (2000). Simulation of hurricane risk in the U.S. using empirical track model. *Journal of Structural Engineering*, *126*(10), 1222–1237. https://doi.org/10.1061/(ASCE)0733-9445(2000)126:10(1222)

World Meteorological Organization (2020). *World's deadliest tropical cyclone was 50 years ago*. Published: 12 November 2020. https://public.wmo.int/en/media/news/world's-deadliest-tropical-cyclone-was-50-years-ago

Yin, J., Welch, M. B., Yashiro, H., & Shinohara, M. (2009). *Basinwide Typhoon Risk Modeling and Simulation for Western North Pacific Basin*. Paper presented at The Seventh Asia-Pacific Conference on Wind Engineering, Taipei, Taiwan.

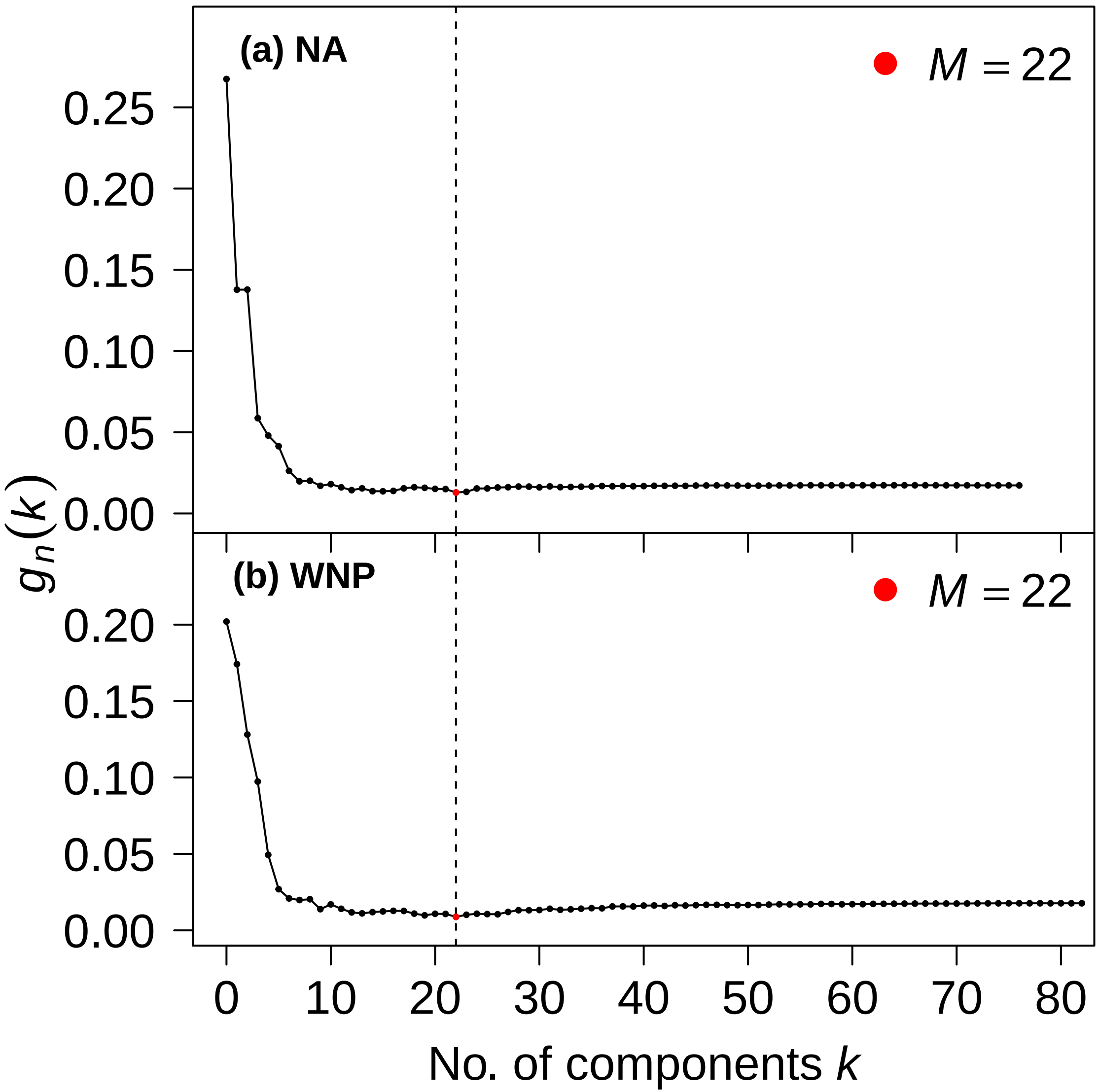
Yonekura, E., & Hall, T. M. (2011). A Statistical Model of Tropical Cyclone Tracks in the Western North Pacific with ENSO-Dependent Cyclogenesis, *Journal of Applied Meteorology and Climatology*, *50*(8), 1725-1739. https://doi.org/10.1175/2011JAMC2617.1

**Figure 1.**



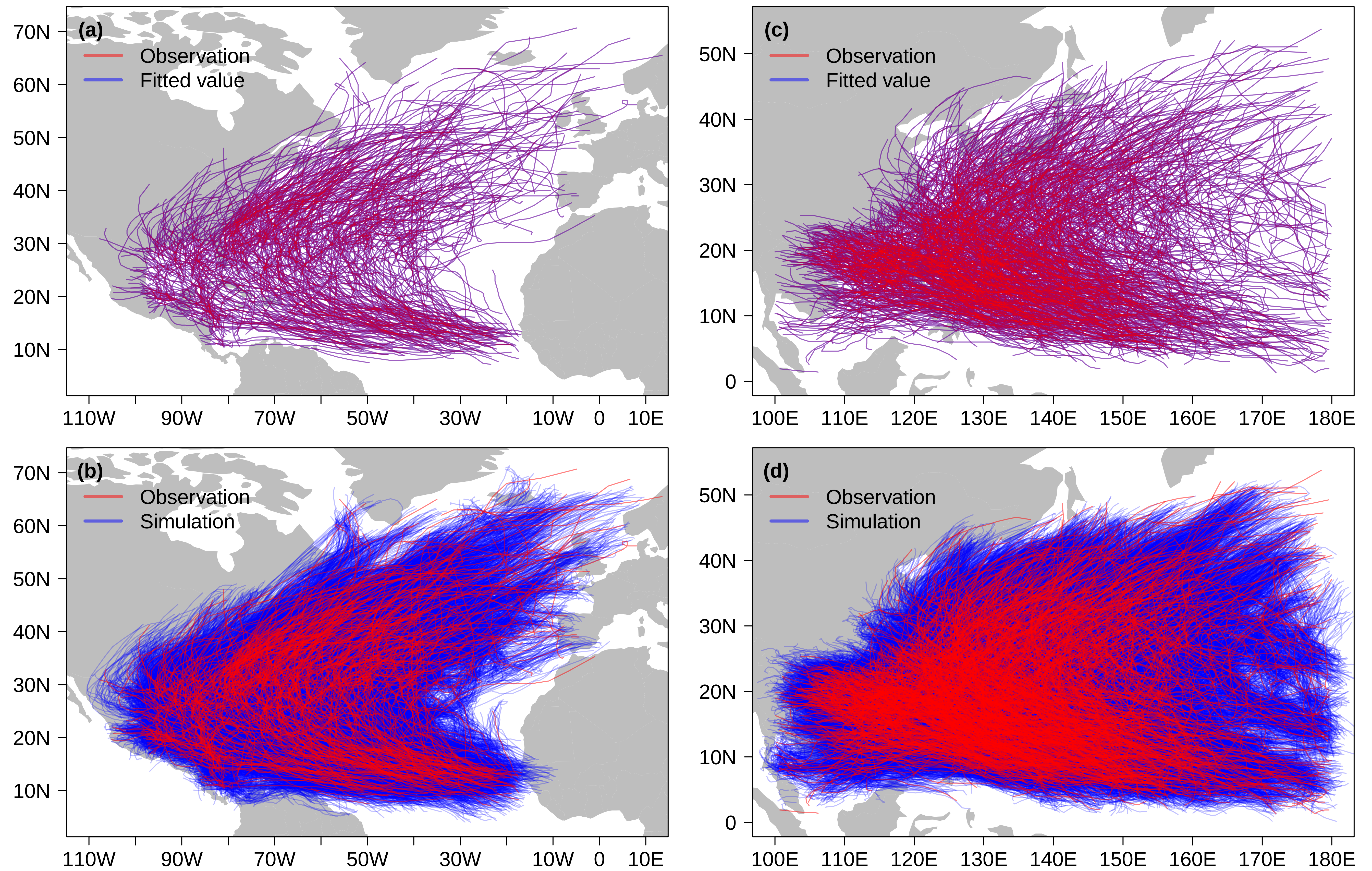
(a)
(b)
(c)
(d)
(e)
Observation
Extended
Simulation
Removed
LON
LAT
MWS (kt)
MCP (hPa)
Ordinal date
Time (h)

**Figure 2.**



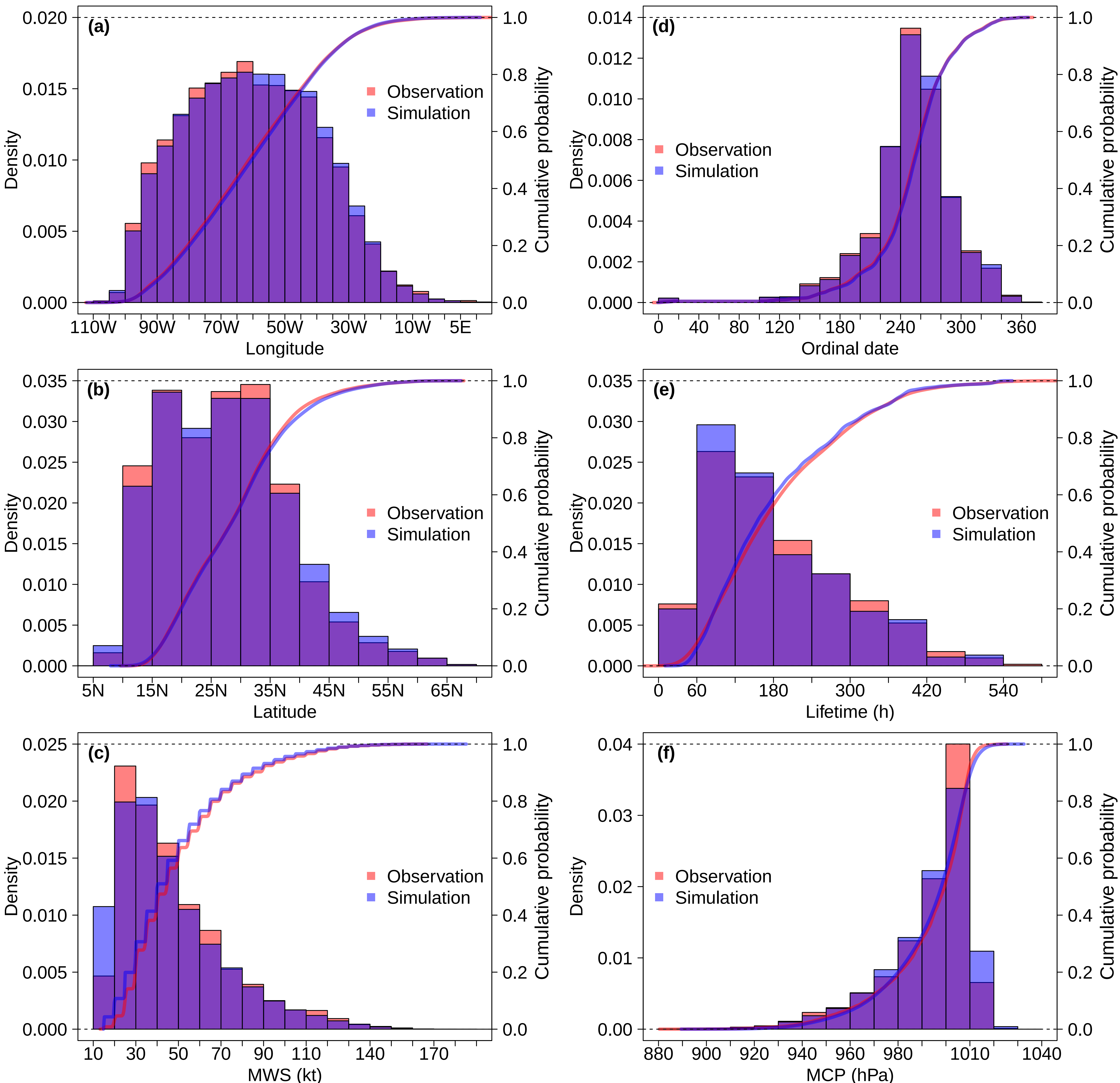
(a) NA
M = 22
0.25
0.20
0.15
0.10
0.05
0.00
(b) WNP
M = 22
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
60
70
80
No. of components k
g_n(k)

**Figure 3.**



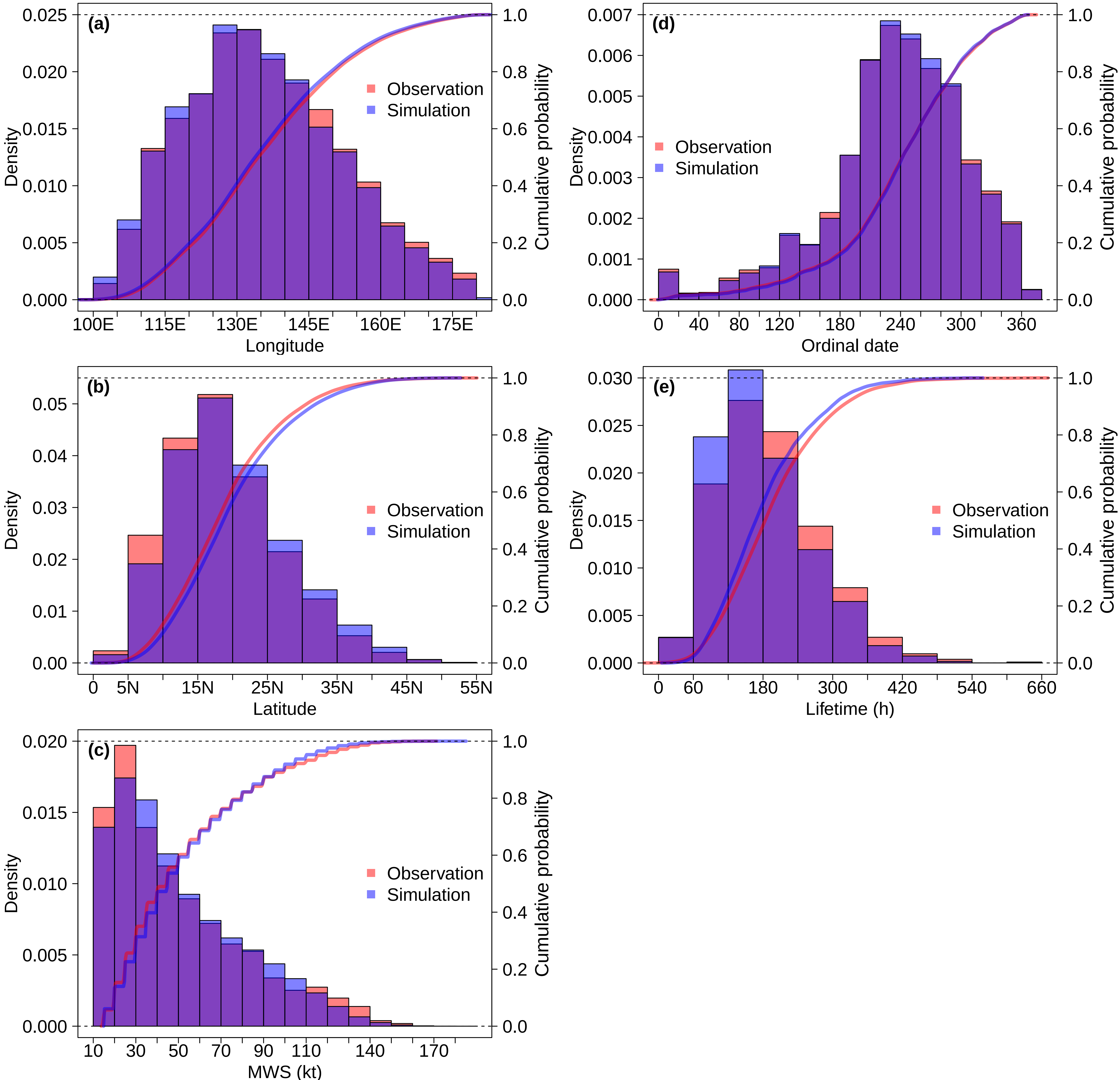
(a)
Observation
Fitted value
(b)
Observation
Simulation
(c)
Observation
Fitted value
(d)
Observation
Simulation
70N
60N
50N
40N
30N
20N
10N
0
110W
90W
70W
50W
30W
10W
0
10E
100E
110E
120E
130E
140E
150E
160E
170E
180E

**Figure 4.**



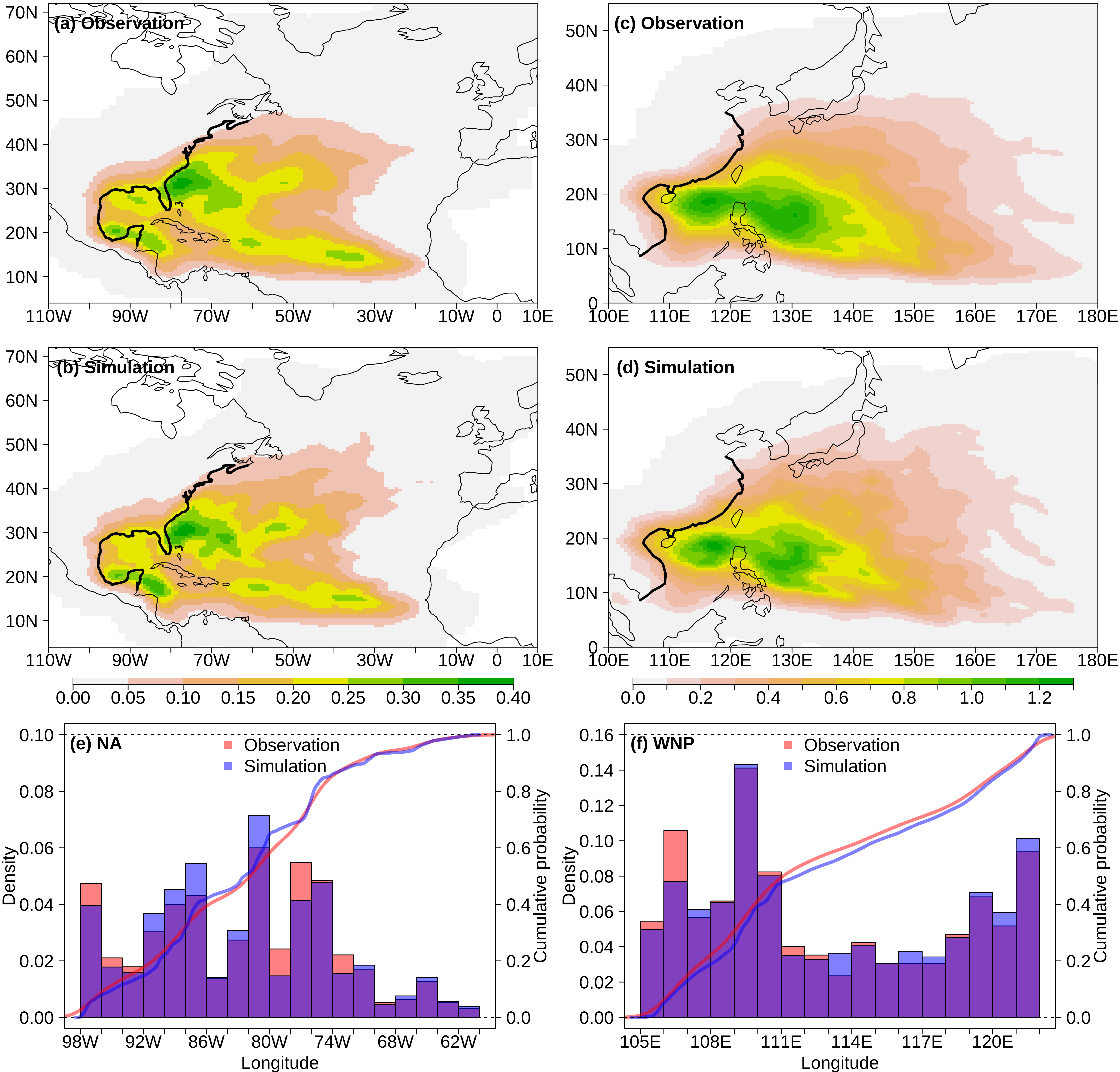
(a)
Observation
Simulation
Density
Cumulative probability
Longitude
110W 90W 70W 50W 30W 10W 5E
(b)
Latitude
5N 15N 25N 35N 45N 55N 65N
(c)
MWS (kt)
(d)
Ordinal date
(e)
Lifetime (h)
(f)
MCP (hPa)

**Figure 5.**



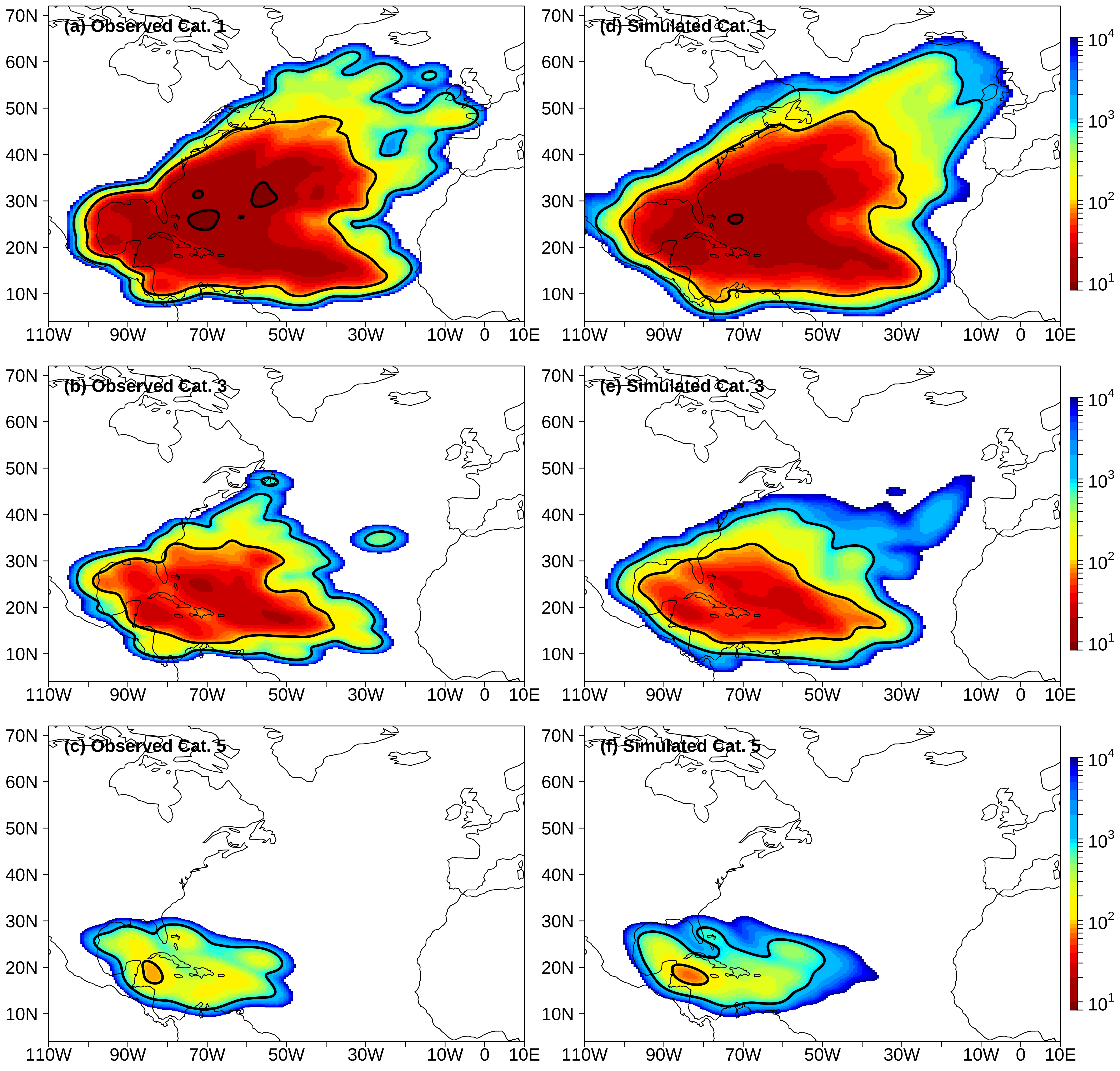
(a)
Observation
Simulation
Density
Cumulative probability
100E 115E 130E 145E 160E 175E
Longitude
(b)
Latitude
0 5N 15N 25N 35N 45N 55N
(c)
MWS (kt)
(d)
Ordinal date
(e)
Lifetime (h)

**Figure 6.**



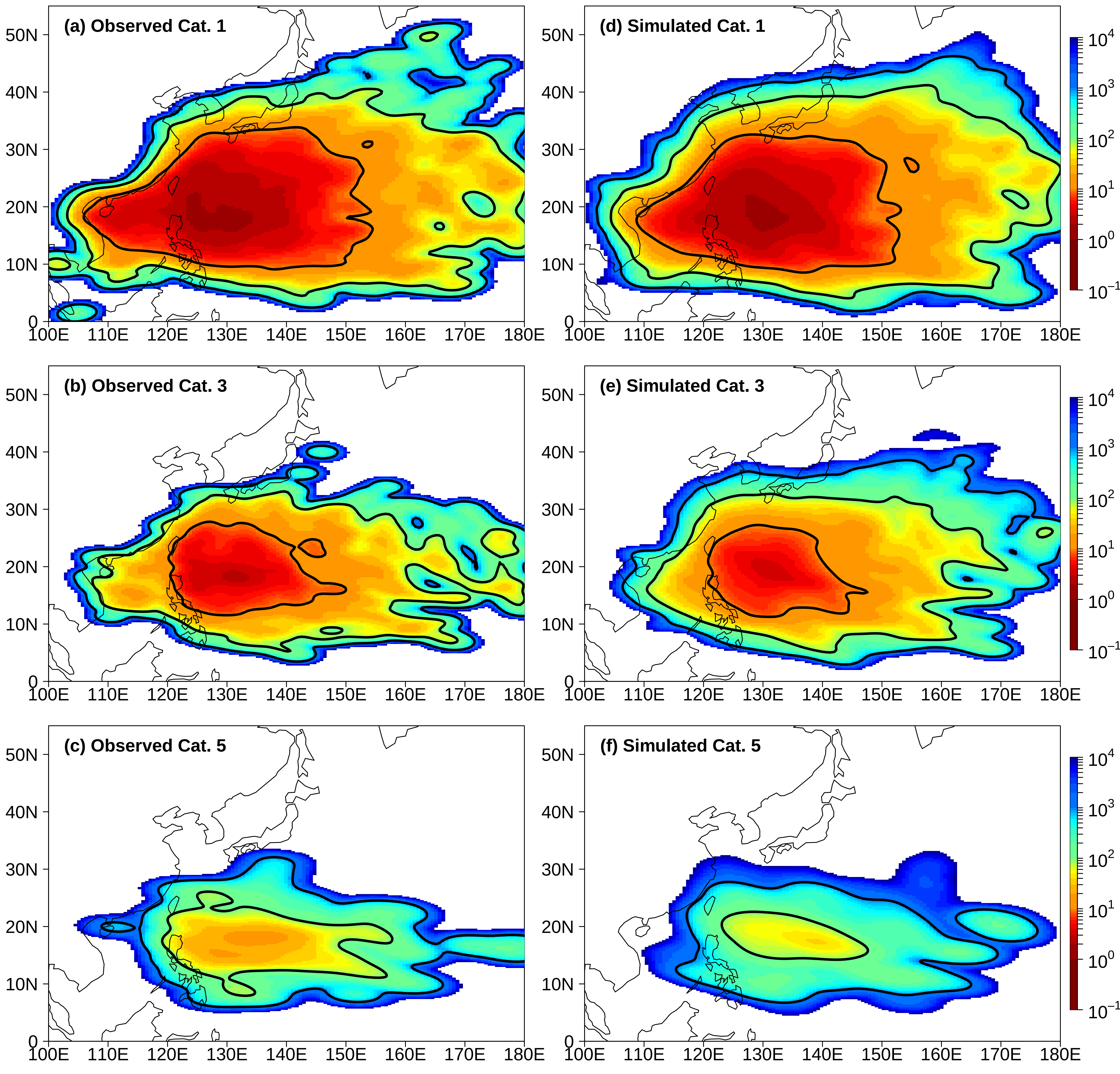
(a) Observation
(b) Simulation
(c) Observation
(d) Simulation
70N
60N
50N
40N
30N
20N
10N
0
110W
90W
70W
50W
30W
10W
10E
100E
110E
120E
130E
140E
150E
160E
170E
180E
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
0.0 0.2 0.4 0.6 0.8 1.0 1.2
(e) NA
(f) WNP
Observation
Simulation
Density
Cumulative probability
Longitude
98W
92W
86W
80W
74W
68W
62W
105E
108E
111E
114E
117E
120E

**Figure 7.**



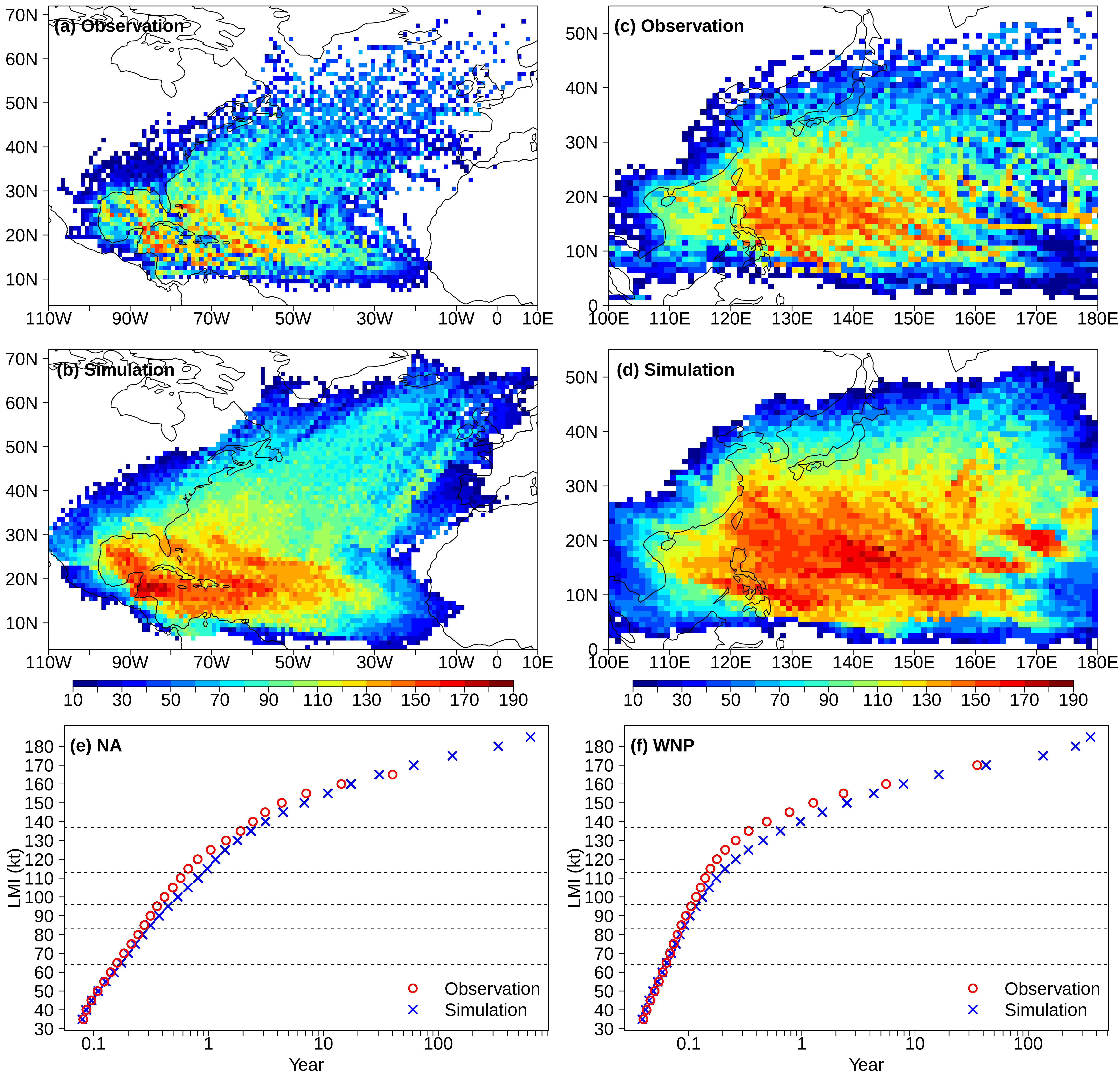
(a) Observed Cat. 1
(b) Observed Cat. 3
(c) Observed Cat. 5
(d) Simulated Cat. 1
(e) Simulated Cat. 3
(f) Simulated Cat. 5
70N
60N
50N
40N
30N
20N
10N
110W
90W
70W
50W
30W
10W
0
10E
$10^4$
$10^3$
$10^2$
$10^1$

**Figure 8.**



(a) Observed Cat. 1
(b) Observed Cat. 3
(c) Observed Cat. 5
(d) Simulated Cat. 1
(e) Simulated Cat. 3
(f) Simulated Cat. 5
50N
40N
30N
20N
10N
0
100E
110E
120E
130E
140E
150E
160E
170E
180E
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
$10^{-1}$

**Figure 9.**



(a) Observation
(b) Simulation
(c) Observation
(d) Simulation
70N
60N
50N
40N
30N
20N
10N
0
110W 90W 70W 50W 30W 10W 0 10E
100E 110E 120E 130E 140E 150E 160E 170E 180E
10 30 50 70 90 110 130 150 170 190
(e) NA
(f) WNP
LMI (kt)
Year
0.1 1 10 100
Observation
Simulation