

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

Chi Yang<sup>1</sup>, Jing Xu<sup>2</sup>, and Jianming Yin<sup>3</sup>

<sup>1</sup>College of Global Change and Earth System Science, Beijing Normal University

<sup>2</sup>Chinese Academy of Meteorological Sciences

<sup>3</sup>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.

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

3   **Chi Yang<sup>1</sup>, Jing Xu<sup>2,4</sup>, and Jianming Yin<sup>3</sup>**

4   <sup>1</sup>College of Global Change and Earth System Science, Beijing Normal University, Beijing  
5   100875, China.

6   <sup>2</sup>State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences,  
7   Beijing 100081, China.

8   <sup>3</sup>China Re Catastrophe Risk Management Company LTD., Beijing 100033, China.

9   <sup>4</sup>Fujian Key Laboratory of Severe Weather, Fuzhou 350001, China.

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

11   **Key Points:**

- 12       • Multivariate functional principal component analysis approach to the full-track  
13       simulation of tropical cyclones for risk assessment
- 14       • Tropical cyclone annual occurrence model using Conway–Maxwell–Poisson distributions
- 15       • Fully automated program for the generation of high-performance synthetic tropical  
16       cyclones  
17

## 18 **Abstract**

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

## 35 **Plain Language Summary**

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

## 50 **1 Introduction**

51 Tropical cyclones (TCs) are one of the biggest threats to life and property around the  
52 world. Over the past 50 years, there have been nearly 2,000 disasters linked to tropical cyclones,  
53 causing nearly 780,000 deaths and US\$ 1,500 billion in economic losses (World Meteorological  
54 Organization, 2020). However, the infrequent nature of catastrophic TCs invalidates the standard  
55 actuarial loss estimation approaches. Computer models that are able to simulate tens, even  
56 hundreds, of thousands of synthetic TC tracks were developed in the past to compensate the  
57 scarcity of historical TC loss data, and to achieve more stable TC loss estimates for regions  
58 where little to no historical data exist. For insurance and reinsurance companies, it is necessary  
59 to evaluate the TC hazard risks as precisely as possible to quantify, manage and mitigate  
60 financial losses. TC risk assessment requires estimation of catastrophic TCs having a very low  
61 occurrence probability, or equivalently a very long return period (e.g., 10000 years). Since

62 reliable TC data are available only for recently decades, and landfalling TCs are relatively few in  
63 nature, stochastic modeling and simulation turned out to be an effective approach to more stable  
64 TC risk estimates for regions where little or no historical TC records are available. The common  
65 practice consists of two stages (Vickery et al. 2009). The first stage is to fit a basin-wide TC full-  
66 track model to TC track data, to generate hundreds of thousands of synthetic TCs that can make  
67 up for the sparseness of TC observations while still comply with the statistical characteristics of  
68 the observed TCs. The second stage is to couple these synthetic TCs with TC wind field models,  
69 either to simulate landfall TC wind fields for wind hazard estimation, or to further drive storm  
70 surge models for coastal flood estimation. Therefore, the performance of the synthetic TCs is  
71 crucial to the respective risk estimates. The full-track TC data consist of at least the TC center  
72 coordinates, maximum wind speed (MWS) as a measure of intensity and/or minimum central  
73 pressure (MCP) observed along TC tracks. A full-track model should be able to represent these  
74 elements and be used for simulation.

75 Vickery et al. (2000) published the first full-track model for the North Atlantic (NA)  
76 basin within the regression framework. The track heading, speed and intensity were determined  
77 for each  $5^{\circ} \times 5^{\circ}$  grid over the entire basin individually. This approach was later adapted by James  
78 and Mason (2005) for the Coral Sea, Yin et al. (2009), Li and Hong (2016) and Chen and Duan  
79 (2018) for the western North Pacific (WNP) basin, respectively. Casson and Coles (2000)  
80 generated TC tracks for the NA basin simply by sampling the historical tracks and then  
81 translating by a normally distributed random displacement with the standard deviation less than  
82 100 nm (1 nm = 1.852 km) and used a simple empirical model to simulate the central pressure  
83 depth with land effects. Emanuel et al. (2006) presented two different track models for the NA  
84 basin: a stochastic Markov chain model and a deterministic beta and advection model. The  
85 former propagates tracks by sampling a transition matrix that relates prior track speed and  
86 direction to the new speed and direction; the latter determines the TC motion by the weighted  
87 average of TC-ambient flow at 850 and 250 hPa plus a beta-drift correction. The TC intensity  
88 along tracks was obtained by coupling each synthetic track to a numeric model developed by  
89 Emanuel et al. (2004). Following this work, several Markovian-type TC track models were  
90 developed, e.g., Hall and Jewson (2007), Rumpf et al. (2007, 2009), Yonekura and Hall (2011),  
91 Kriesche et al. (2014) and Nakamura et al. (2015), for the NA or WNP basin or both. Emanuel et  
92 al. (2008) further developed a statistical-deterministic model for downscaling TC climatologies  
93 from global analyses, using a random seeding method to initiate the storm, and a beta and  
94 advection model to propagate the storm. Following this approach, Lee et al. (2018) and Jing and  
95 Lin (2020) developed similar TC hazard models, either of which is comprised of three  
96 component models for TC genesis, track and intensity, respectively, dependent upon local  
97 environmental conditions.

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

108 multivariate functional linear regression with spatially varying coefficients, highlighting the  
109 representation of complex spatial-temporal dependency of TC tracks.

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

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

## 140 **2 Data**

141 The only raw material we use to construct the simulation model is the historical best-  
142 track (reanalyzed) data of TCs. The data sets for the NA and WNP basin were derived from the  
143 Atlantic hurricane database (HURDAT) and Joint Typhoon Warning Center (JTWC),  
144 respectively, and were redistributed through the International Best Track Archive for Climate  
145 Stewardship (IBTrACS, Knapp et al., 2010). The period since 1980 is generally considered the  
146 modern era when geostationary satellite coverage has been nearly global and polar orbiting  
147 satellite data has been more widely available than the prior years. Therefore, we take data from  
148 1980 till the recent year available, which is 2019 for the NA basin and 2018 for the WNP basin,  
149 respectively. The TC information in the best-track data includes storm type, date and time, center  
150 coordinates longitude/latitude (LON/LAT), MWS, MCP, and average translation speed and  
151 direction inferred from center coordinates, recorded every 6 hours. For the two USA agencies,

152 MWS is defined as the maximum 1-minute sustained wind speed at 10 m above the surface. For  
 153 the WNP basin, MCP is available only since 2001. Only those TCs with their lifetime maximum  
 154 intensity (LMI) reaching the tropical storm (TS) level (34 kt or  $17.5 \text{ m s}^{-1}$ ) or above are chosen  
 155 as sample observations for modeling. As a result, sample data for the NA and WNP basin are  
 156 comprised of 513 and 1035 TCs, respectively.

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

### 167 3 Methods

#### 168 3.1 Multivariate FDA

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

180 Multivariate functional data (MFD) take multiple functions at the same time into account.  
 181 Each observation unit consists of a fixed number of functions  $p$ , and is assumed to be a  
 182 realization of a random process  $X = (X^{(1)}, \dots, X^{(p)})$ , where  $X^{(k)} = \{X^{(k)}(t), t \in \mathcal{T}\}$ ,  $k =$   
 183  $1, \dots, p$ . As only observed discretely, MFD are of the form  $(t_{ij}^{(k)}, x_{ij}^{(k)})$ ,  $i = 1, \dots, N$ ,  $j =$   
 184  $1, \dots, S_i$ ,  $k = 1, \dots, p$ . Each element function can be represented separately by its observation  
 185 points and the observed values. The full MFD sample is a collection of all the  $p$  element  
 186 functions.

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

## 194 3.2 MFPCA

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

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

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

$$\Gamma \psi_m = v_m \psi_m \quad (2)$$

208 where  $v_m$  are eigenvalues and  $v_m \rightarrow 0$  for  $m \rightarrow \infty$ , and  $\rho_m$  are zero mean random variables with  
 209  $\text{cov}(\rho_m, \rho_n) = v_m \delta_{mn}$ . Moreover,

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

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

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

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

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

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

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

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

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

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

223 respectively.

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

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

225 where  $\hat{\psi}_m = (\hat{\psi}_m^{(1)}, \dots, \hat{\psi}_m^{(p)})$  having the same multivariate structure of  $X$ . The R package  
226 “MFPCA” (Happ-Kurz, 2020) provides an easy way to implement the above algorithm.

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

### 243 3.3 Order determination

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

$$X = Z + \epsilon \quad (7)$$

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

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

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

$$g_n(k) = f_n(k) + \phi_n(k) \quad (9)$$

257 where  $f_n(k)$  and  $\phi_n(k)$  represent the bootstrap eigenvector variability and sample eigenvalues,  
 258 respectively,  $n$  is the number of bootstrap samples (half the number of data by default),  $k =$   
 259  $0, \dots, p - 1$ . Refer to Eqs. (4) and (5) in Luo and Li (2016) for the mathematical forms of the two  
 260 terms. The eigenvalue term  $\phi_n(k)$  is large when  $k < d$ ; the eigenvector term  $f_n(k)$  is large when  
 261  $k > d$ ; but both are small when  $k = d$ . Therefore,  $g_n(k)$  is expected to reach its minimum  
 262 approximately at  $d$ . The function curve of  $g_n(k)$  resembles a ladle, hence the name. The R code  
 263 provided in the Supplementary material of Luo and Li (2016) can be adapted to estimate the  
 264 optimal truncation lag  $M$  in the MFPCA context.

### 265 3.4 Full-track simulation

#### 266 3.4.1 Simulation model

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

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

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

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

#### 284 3.4.2 Occurrence model

285 For count data like the annual number of TCs, Poisson distribution is usually the  
 286 preferred model in which the expected value stands for the annual rate of occurrence. Poisson

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

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

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

301 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  
 302 to overdispersion, equidispersion (Poisson distribution) and underdispersion, respectively. Huang  
 303 (2017) suggested a reparameterization of CMP distributions with mean  $\mu$  and dispersion  $\nu$ , which  
 304 is more suitable for fitting Generalized Linear Models. As a result, the variance of the CMP  
 305 distribution is a function of  $\mu$  and  $\nu$ , or  $V(\mu, \nu)$ . In this work, we fit the CMP distribution in the  
 306  $\mu$ - $\nu$  form to the annual TC number sequence as the occurrence model with the help of the R  
 307 package “mpcmp” (Fung et.al, 2020), the R implementation of Huang (2017).

## 308 4 Results

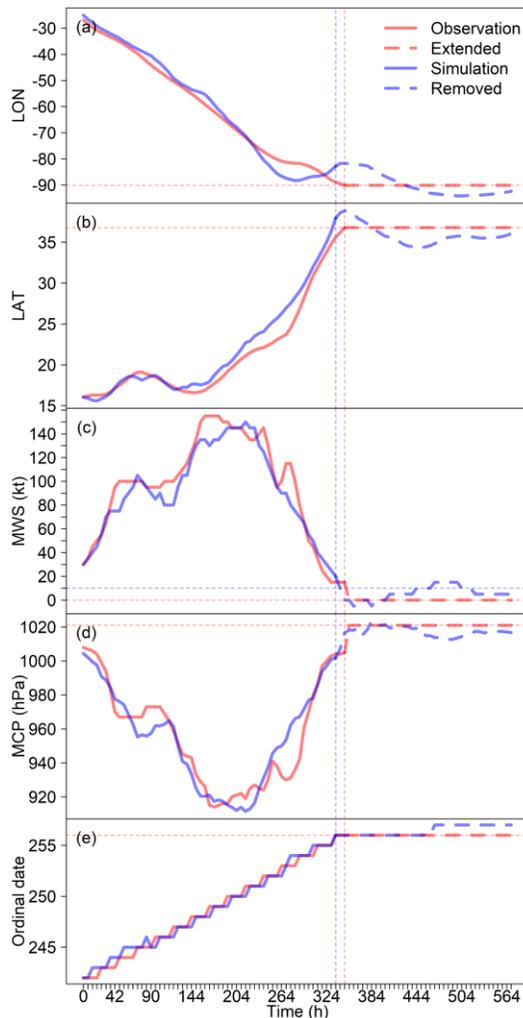
### 309 4.1 Pre-processing of best-track data and post-processing of simulations

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

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

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

336 Figure 1 illustrates the above pre- and post-processing procedures, using the hurricane  
 337 Irma (2017242N16333) as an example. The time span for observation is 00Z 30 August to 12Z  
 338 13 September 2017, a period of 348 hours. In order to prepare the TC MFD for modeling, all the  
 339 variable records are extended with constant values to 570 hours, the interval  $\mathcal{T}$  on which the  
 340 MFD are defined for the NA basin. The simulation is randomly generated by using Irma as a  
 341 “seed”. Simulated track points with MWS equal to or less than 10 kt are removed, resulting in  
 342 two track segments. The one with LMI less than 34 kt is also discarded. The remaining one  
 343 finally becomes a synthetic TC. Note that the simulated dates are not a strictly regular step  
 344 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 a “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

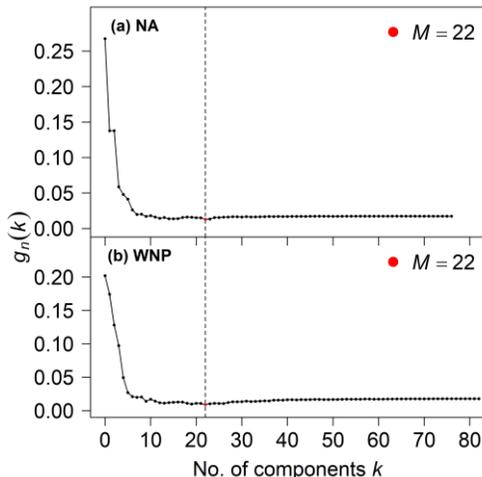
Table 1 summarizes the primary information about the model fitting and simulation. In this study, the TC MFD for the NA basin consists of nine element functions while that for the WNP basin consists of eight, due to the shortage of the MCP records in the WNP basin. With MFPCA, the TC MFD can be represented as Eq. (6) which serves as

370 the fitted model in this study. If we divide the summation in the right-hand side of Eq. (6) into  
 371 two parts, the first  $M$  leading eigenvectors as the signal part and all the others as the noise part, it  
 372 turns out to be the simulation model expressed as Eq. (10). Only about 54 % and 60 % of the  
 373 total  $M_+ (= p \times (S + 2))$  eigenvalues are nonzero for the NA and WNP basin, respectively. As  
 374 we pointed out in section 3.2,  $(S + 2)$  number of degrees of freedom for each element function  
 375 are obviously redundant due to the fact that most of the actual TC lifetimes are less than  $\mathcal{T}$ ,  
 376 hence the rank of the joint covariance matrix  $\hat{Z}$ , or correspondingly the number of nonzero  
 377 eigenvalues, is much smaller than  $M_+$ . However, by means of the ladle estimator, only the first  
 378 22 leading eigenvectors that explains about 93% of total variance are recognized to constitute the  
 379 signal part of the simulation model, coincidentally for both the two basins (Fig. 2). The rest of  
 380 eigenvectors with nonzero eigenvalues then constitute the noise part.

381 **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%)	

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



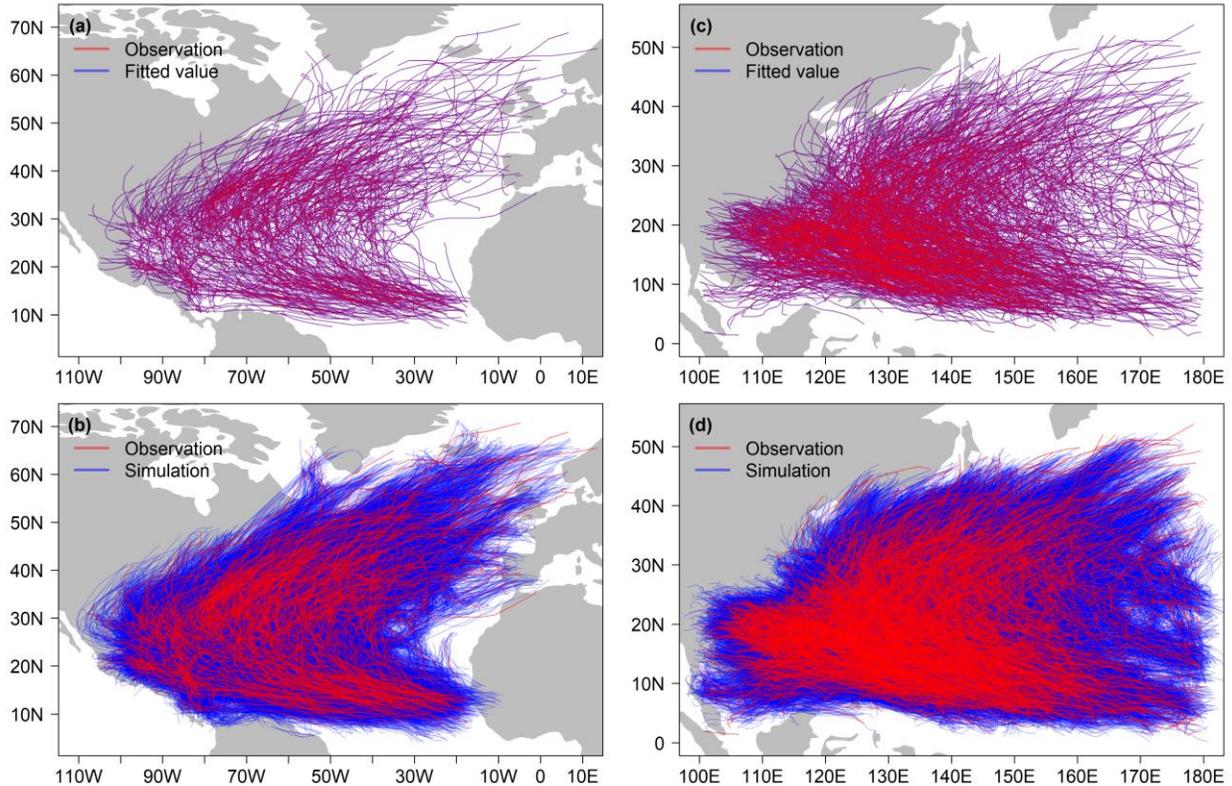
**Figure 2.** Ladle estimates of the optimal truncation  $M$  in the simulation model for the two basins, respectively. Both the results are 22, coincidentally.

#### 4.3 Model validation

##### 4.3.1 Spatial pattern and annual occurrence

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

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



406

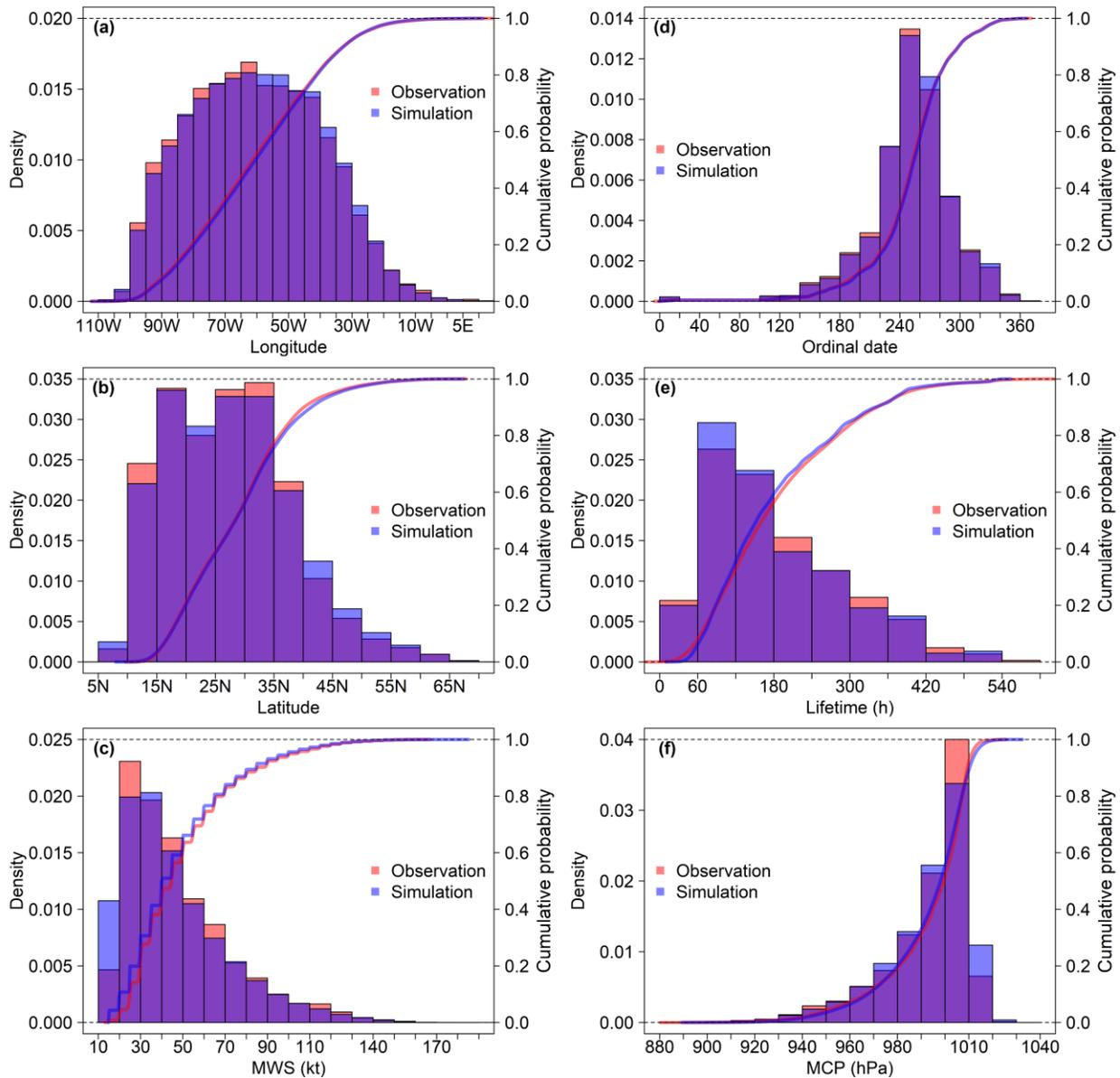
407 **Figure 3.** Comparisons of fitted values and simulations to observations for the NA and WNP  
 408 basin, respectively. Note that observations and fitted values are actually overlapped.

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

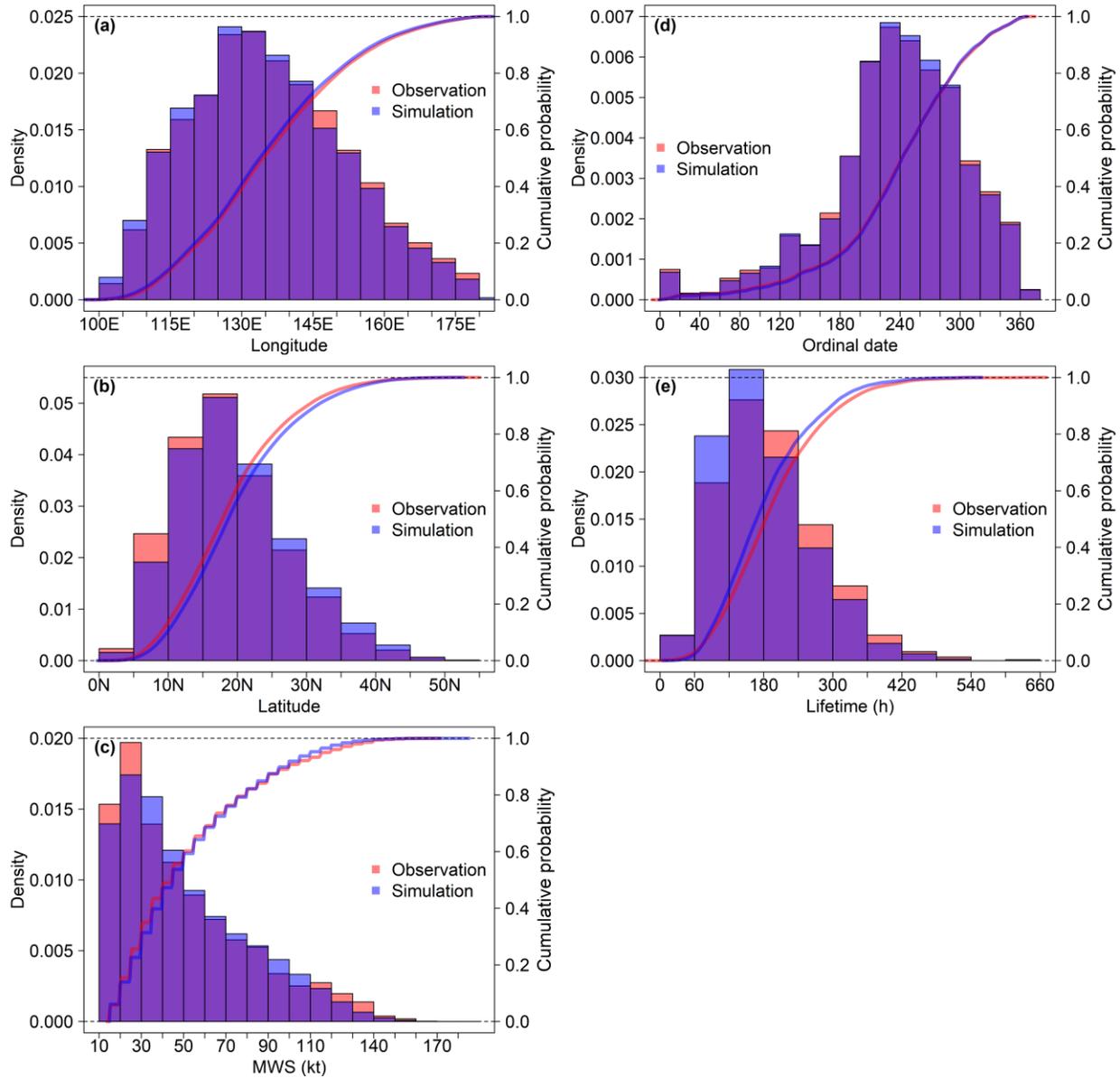
#### 415 4.3.2 Marginal distributions

416 Now we check the performance of synthetic TCs in more detail, by comparing the  
 417 empirical distributions of LON, LAT, MWS, MCP, ordinal dates and lifetime from simulations  
 418 to those from observations. Figures 4 and 5 show the comparisons in terms of empirical  
 419 probability density function (EPDF) and empirical cumulative distribution function (ECDF) for  
 420 the NA and WNP (for which MCP is not available) basin, respectively. Histograms represent  
 421 EPDFs, and curves represents ECDFs corresponding to EPDFs. It can be seen that for each

422 basin, histograms for observations and simulations are almost overlapped, and the two curves are  
 423 quite close to each other. These high consistencies show the capability of the simulation model to  
 424 capture the marginal distribution features of TC variables. Discrepancies such as in the  
 425 lower/upper tail of MWS/MCP distribution are mostly related to noises just exceed the 10-kt  
 426 threshold, thus can be ignored. As for the seasonality and lifetimes of the simulated TCs,  
 427 comparisons to observations also show satisfying results. Particularly for the seasonality of TC  
 428 activity, simulations almost reconstructed the distribution of ordinal dates from observations  
 429 (Fig. 4d and Fig. 5d), which is helpful for assessing the TC risk on a monthly or even shorter-  
 430 term basis.



431  
 432 **Figure 4.** Comparisons between observations and simulations in terms of EPDF (histogram) and  
 433 ECDF (curve) of TC variables for the NA basin. EPDF and ECDF values are indicated by the  
 434 left and right ordinate, respectively.



435

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

438

### 4.3.3 Spatial distributions

439

440

441

442

443

444

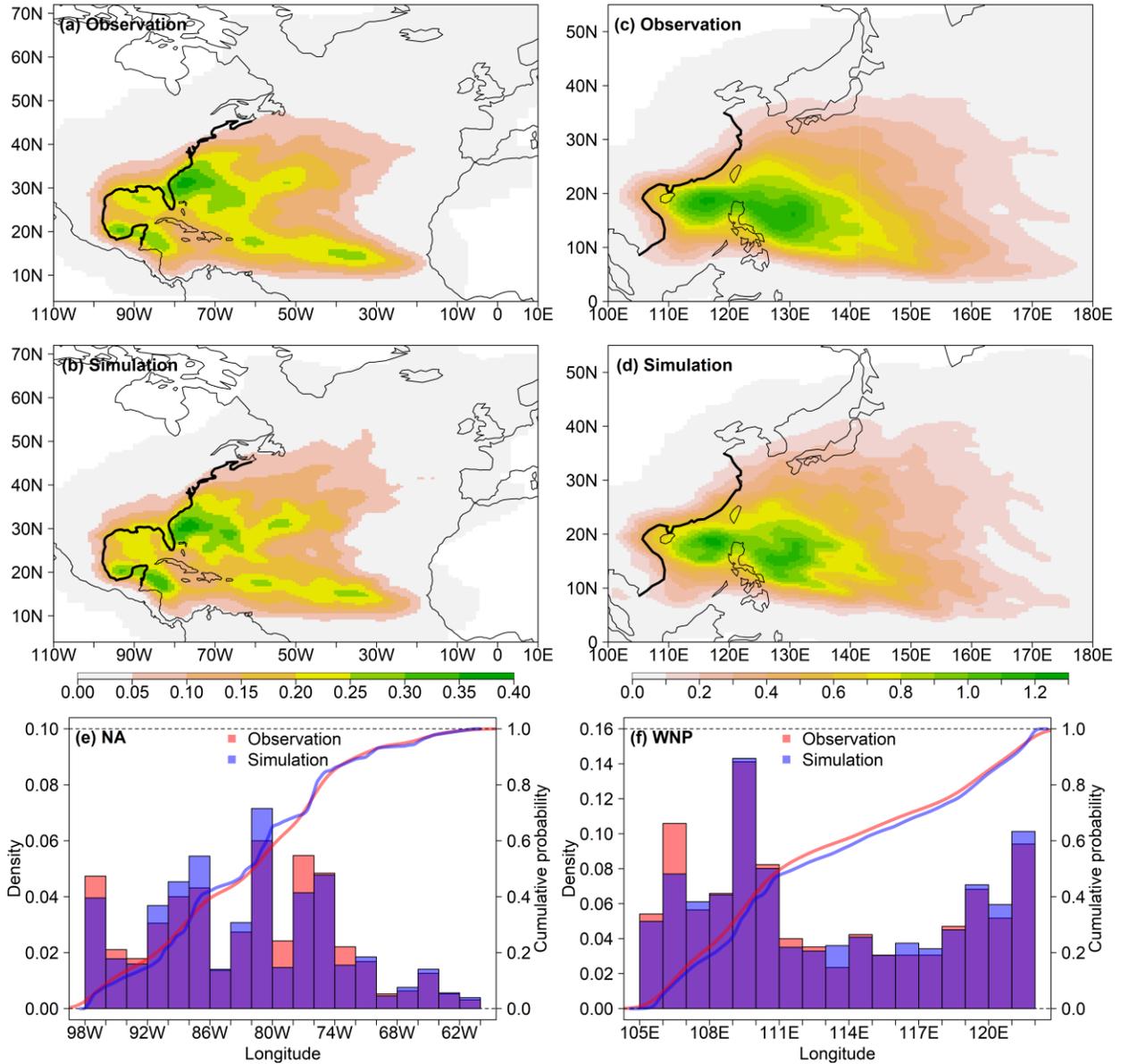
445

446

447

We check further the joint distributions of TC variables, focusing 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 observations, even though the time span of simulations is much longer than that of 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

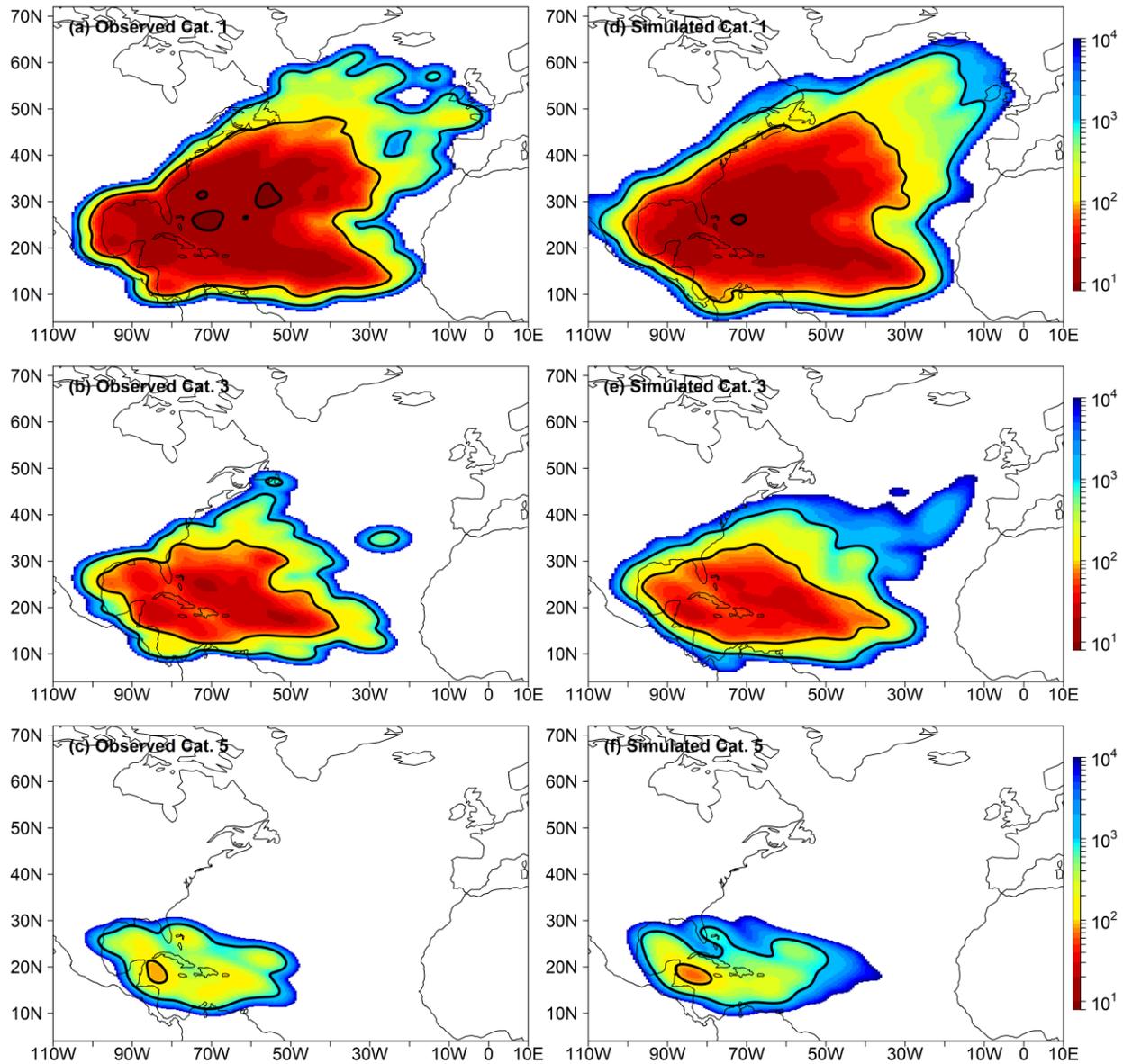
448 simulated landfall locations as functions of longitude along the thick coastline for the two basins,  
 449 respectively. These two coastlines are more liable to be attacked by TCs among others in their  
 450 respective basins. Histograms represent EPDFs of landfall locations, and curves represents  
 451 ECDFs corresponding to EPDFs. Once again, a high consistency exists between observations  
 452 and simulations in terms of TC landfall locations. Particularly for the WNP basin (Fig. 6f) where  
 453 the TC landfalls are more frequent than in the NA basin, there are more data available for  
 454 modeling, resulting in reduced model uncertainty and smaller simulation bias.



455

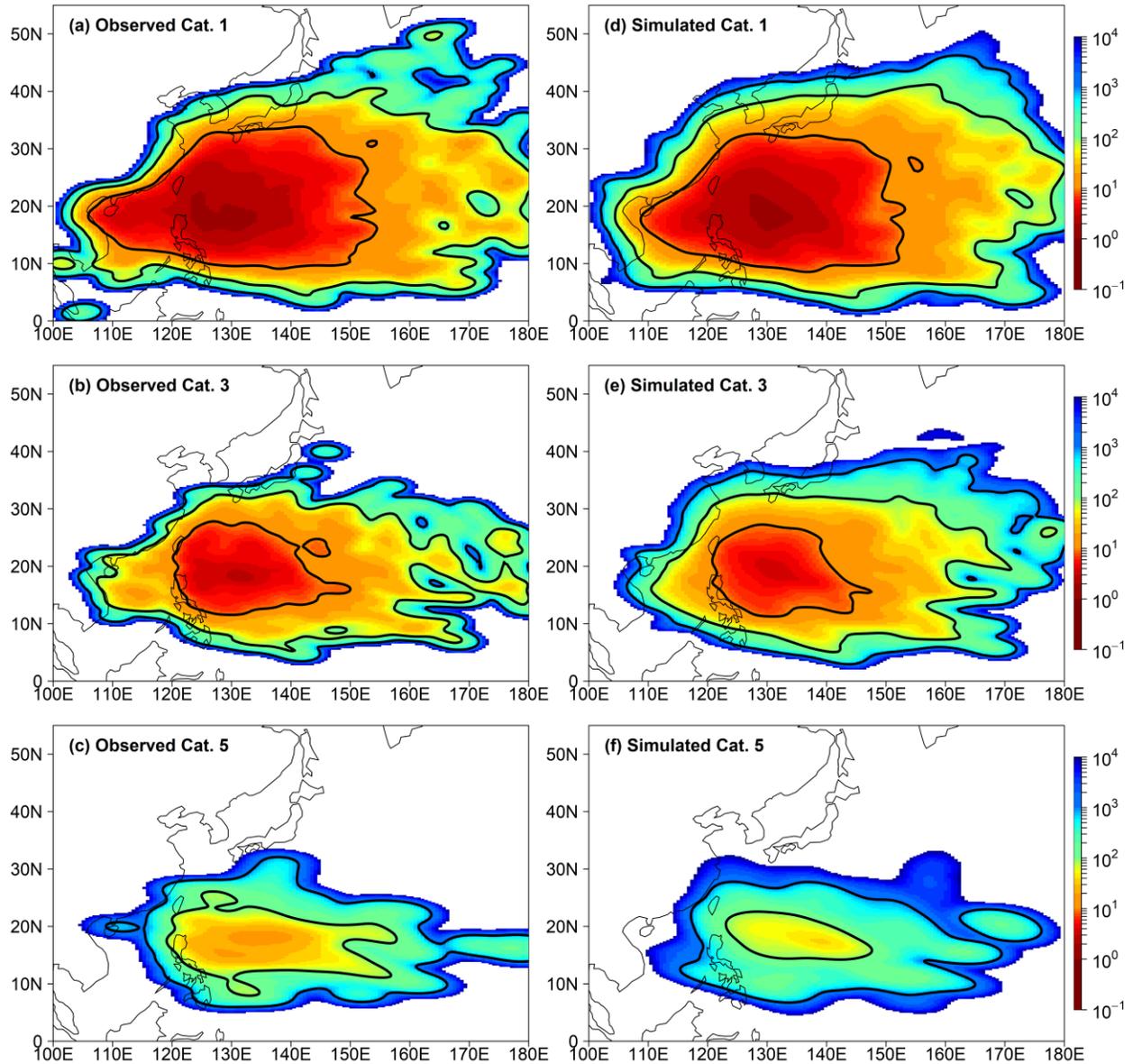
456 **Figure 6.** Annual mean spatial densities of track points (unit:  $\text{degree}^{-2} \text{yr}^{-1}$ ) from observations  
 457 and simulations, and comparison between observations and simulations in terms of EPDF  
 458 (histogram) and ECDF (curve) of landfall locations along the thick coastline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

460 Next, we examine the spatial distribution of the simulated MWS in terms of return  
 461 periods of categories from the Saffir-Simpson hurricane scale for both the two basins. The  
 462 ranges for categories 1–5 (Cat. 1–5) are 64–82, 83–95, 96–112, 113–136 and > 137 kt,  
 463 respectively. Figure 7 compares return periods of simulated MWS using the lower limits of Cat.  
 464 1, 3 and 5 as thresholds to those from observations, respectively, for the NA basin. Figure 8 is  
 465 the same as Fig. 7 but for the WNP basin. With these statistics, the spatial distribution of MWS  
 466 as a function of LON and LAT can be outlined. These results show that, although the time span  
 467 of simulations is much longer than that of observations, simulations do not substantially deviate  
 468 from observations in terms of statistical properties, which is essential for synthetic TCs to be  
 469 used for risk assessment.



470

471 **Figure 7.** Return periods of observed and simulated MWS using the lower limits of Cat. 1, 3 and  
 472 5 as thresholds, respectively, for the NA basin. Black contours indicate 10-, 100- and 1000-year  
 473 return periods.



474

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

476

## 476 4.3.4 Intensity extremes

477

478

479

480

481

482

483

484

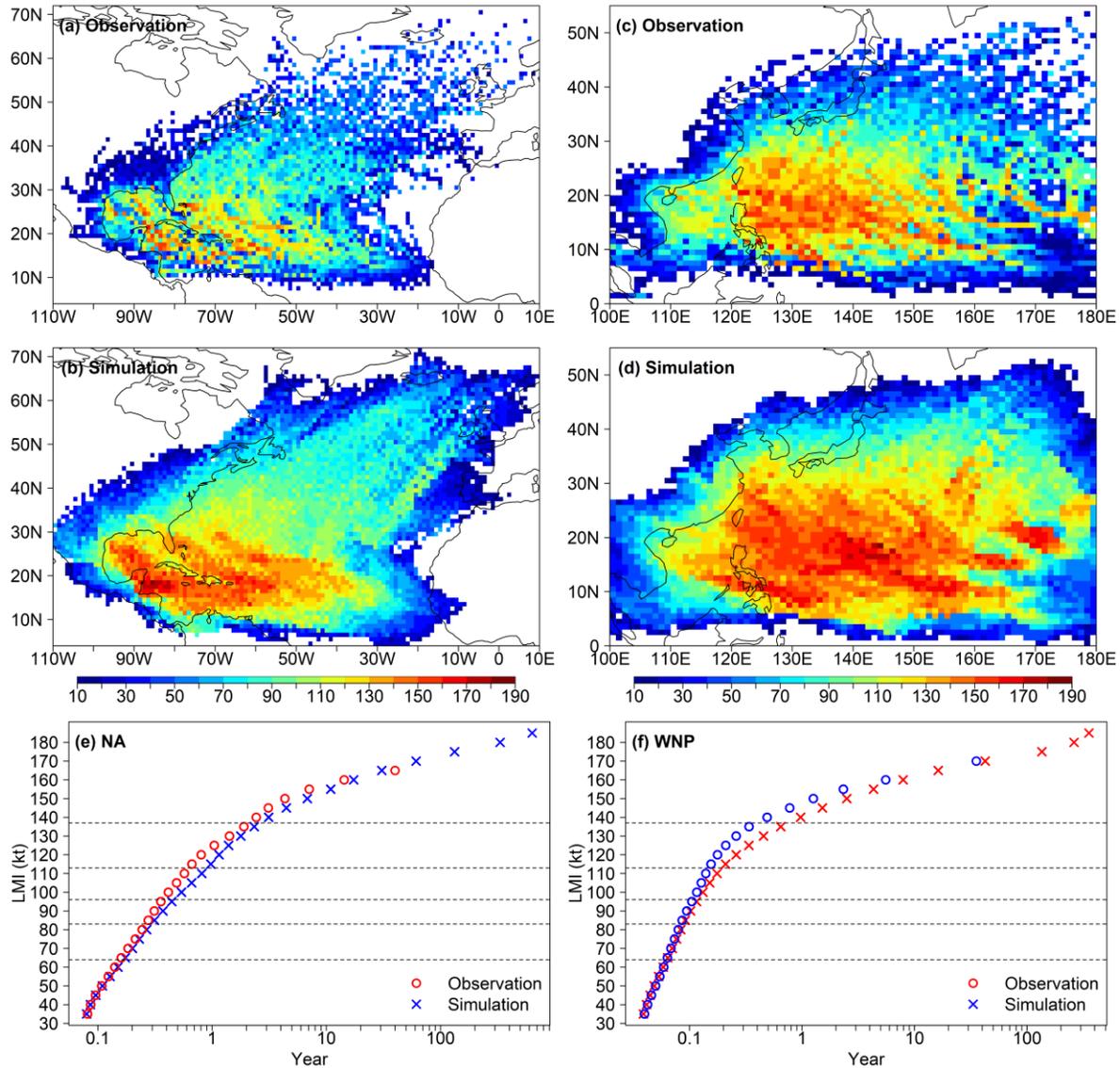
485

486

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 intensity extremes to complete the validation. Figures 9a–9d compares maxima of MWS over each individual  $1^\circ \times 1^\circ$  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). Coincidentally, for both the two

487 basins the simulated maximum MWS is 185 kt. This seeming coincidence actually indicates that  
 488 the simulation model does grasp the empirical MPI by mining the best-track data.

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



497  
 498 **Figure 9.** Maxima of MWS over each individual  $1^{\circ} \times 1^{\circ}$  grid squares from observations and  
 499 simulations, and comparison between observations and simulations in terms of return periods of  
 500 LMI, for the NA and WNP basin, respectively. Horizontal dashed lines indicate lower limits of  
 501 Cat. 1–5, respectively.

## 502 **5 Summary and Discussion**

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

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

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

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

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

543 Since the simulation model is purely empirical without external dynamic factors  
544 incorporated, it is not intended to be an all-purpose alternative to environmentally forced models  
545 such as those described in Emanuel et al. (2008), Lee et al. (2018) or Jing and Lin (2020),

546 particularly when these models are used for assessing TC risks projected by climate change  
 547 scenarios. To some extent, this approach is still capable of assessing TC risks modulated by  
 548 some climate variability, by sampling historical TCs subject to different phases such as El Nino  
 549 and La Nina separately during the simulation. A possible extension is the joint simulation of TCs  
 550 in different basins, such as the NA and East Pacific (EP) basins, by means of joint modeling of  
 551 annual TC occurrences in different basins. In doing so, TCs in different basins are simulated  
 552 synchronously with the inter-basin correlation of TC activity considered. This is helpful for  
 553 insurance/reinsurance companies to setup uniform standards for assessing risks for different  
 554 regions. These ideas will be implemented in our future work.

### 555 **Acknowledgments**

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

### 558 **Data availability statement**

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

### 562 **References**

- 563 Aberson, S. D. (1998). Five-day tropical cyclone track forecasts in the North Atlantic basin.  
 564 *Weather and Forecasting*, 13(4), 1005–1015. [https://doi.org/10.1175/1520-](https://doi.org/10.1175/1520-0434(1998)013<1005:FDTCTF>2.0.CO;2)  
 565 [0434\(1998\)013<1005:FDTCTF>2.0.CO;2](https://doi.org/10.1175/1520-0434(1998)013<1005:FDTCTF>2.0.CO;2)
- 566 Casson, E., & Coles, S. (2000). Simulation and extremal analysis of hurricane events. *Journal of*  
 567 *the Royal Statistical Society: Series C (Applied Statistics)*, 49, 227–245.  
 568 <https://doi.org/10.1111/1467-9876.00189>
- 569 Chacon, J. E., & Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications*. Boca  
 570 Raton: Chapman & Hall/CRC.
- 571 Chen, Y., & Duan, Z. (2018). A statistical dynamics track model of tropical cyclones for  
 572 assessing typhoon wind hazard in the coast of southeast China. *Journal of Wind Engineering and*  
 573 *Industrial Aerodynamics*, 172, 325–340. <https://doi.org/10.1016/j.jweia.2017.11.014>
- 574 DeMaria, M., & Kaplan, J. (1994). Sea surface temperature and the maximum intensity of  
 575 Atlantic tropical cyclones. *Journal of Climate*, 7, 1324–1334. [https://doi.org/10.1175/1520-](https://doi.org/10.1175/1520-0442(1994)007<1324:SSTATM>2.0.CO;2)  
 576 [0442\(1994\)007<1324:SSTATM>2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007<1324:SSTATM>2.0.CO;2)
- 577 Duong, T. (2021). *ks: Kernel Smoothing*. R package version 1.12.0. [https://CRAN.R-](https://CRAN.R-project.org/package=ks)  
 578 [project.org/package=ks](https://CRAN.R-project.org/package=ks)
- 579 Emanuel, K. (2004). Tropical cyclone energetics and structure. In E. Federovich, R. Rotunno, B.  
 580 Stevens (Eds.), *Atmospheric Turbulence and Mesoscale Meteorology: Scientific Research*  
 581 *Inspired by Doug Lilly* (pp. 165–192). Cambridge University Press.
- 582 Emanuel, K., Ravela, S., Vivant, E., & Risi, C. (2006). A statistical deterministic approach to  
 583 hurricane risk assessment. *Bulletin of the American Meteorological Society*, 87, s1–s5.  
 584 <https://doi.org/10.1175/BAMS-87-3-Emanuel>

- 585 Emanuel, K., Sundararajan R., & Williams, J. (2008). Hurricanes and global warming: Results  
586 from downscaling IPCC AR4 simulations. *Bulletin of the American Meteorological Society*, 89,  
587 347–368. <https://doi.org/10.1175/BAMS-89-3-347>
- 588 Fung, T., Alwan, A., Wishart, J., & Huang, A. (2020). *mpcmp: Mean-parametrized Conway–*  
589 *Maxwell Poisson Regression*. R package version 0.3.6. [https://CRAN.R-](https://CRAN.R-project.org/package=mpcmp)  
590 [project.org/package=mpcmp](https://CRAN.R-project.org/package=mpcmp)
- 591 Happ, C. & Greven, S. (2018). Multivariate Functional Principal Component Analysis for Data  
592 Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*,  
593 113(522), 649–659. <https://doi.org/10.1080/01621459.2016.1273115>
- 594 Happ-Kurz, C. (2020). MFPCA: *Multivariate Functional Principal Component Analysis for*  
595 *Data Observed on Different Dimensional Domains*. R package version 1.3-6. [https://CRAN.R-](https://CRAN.R-project.org/package=MFPCA)  
596 [project.org/package=MFPCA](https://CRAN.R-project.org/package=MFPCA)
- 597 Hall, T. M., & Jewson, S. (2007). Statistical modelling of North Atlantic tropical cyclone tracks.  
598 *Tellus A*, 59, 486–498. <https://doi.org/10.1111/j.1600-0870.2007.00240.x>
- 599 Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for  
600 dispersed counts. *Statistical Modelling*, 17(6), 359–380.  
601 <https://doi.org/10.1177/1471082X17697749>
- 602 James, M. K., & Mason, L. B. (2005). Synthetic tropical cyclone database. *Journal of Waterway,*  
603 *Port, Coastal, and Ocean Engineering*, 131(4), 181–192. [https://doi.org/10.1061/\(ASCE\)0733-](https://doi.org/10.1061/(ASCE)0733-950X(2005)131:4(181))  
604 [950X\(2005\)131:4\(181\)](https://doi.org/10.1061/(ASCE)0733-950X(2005)131:4(181))
- 605 Jing, R., & Lin, N. (2020). An environment-dependent probabilistic tropical cyclone model.  
606 *Journal of Advances in Modeling Earth Systems*, 12, e2019MS001975.  
607 <https://doi.org/10.1029/2019MS001975>
- 608 Jolliffe, I. T. (2002). *Principal Component Analysis*, Second Edition. New York, NY: Springer.
- 609 Knaff, J. A., C. R. Sampson, & DeMaria, M. (2005). An operational statistical typhoon intensity  
610 prediction scheme for the western North Pacific. *Weather and Forecasting*, 44(20), 688–699.  
611 <https://doi.org/10.1175/WAF863.1>
- 612 Knaff, J. A., & Zehr, R. M. (2007). Reexamination of Tropical Cyclone Wind–Pressure  
613 Relationships. *Weather and Forecasting*, 22(1), 71–88. <https://doi.org/10.1175/WAF965.1>
- 614 Knapp, K. R., Knaff, J. A., Sampson, C. R., Riggio, G. M., & Schnapp, A. D. (2013). A  
615 Pressure-Based Analysis of the Historical Western North Pacific Tropical Cyclone Intensity  
616 Record. *Monthly Weather Review*, 141(8), 2611–2631. [https://doi.org/10.1175/MWR-D-12-](https://doi.org/10.1175/MWR-D-12-00323.1)  
617 [00323.1](https://doi.org/10.1175/MWR-D-12-00323.1)
- 618 Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The  
619 international best track archive for climate stewardship (IBTrACS). *Bulletin of the American*  
620 *Meteorological Society*, 91, 363–376. <https://doi.org/10.1175/2009BAMS2755.1>
- 621 Kriesche, B., Weindl, H., Smolka, A., & Schmidt V. (2014). Stochastic simulation model for  
622 tropical cyclone tracks, with special emphasis on landfall behavior. *Natural Hazards*, 73, 335–  
623 353. <https://doi.org/10.1007/s11069-014-1075-x>

- 624 Kueh, M. (2012). Multiformality of the tropical cyclone wind-pressure relationship in the western  
625 North Pacific: discrepancies among four best-track archives. *Environmental Research Letters*,  
626 7(2), 024015. <https://doi.org/10.1088/1748-9326/7/2/024015>
- 627 Lee, C.-Y., Tippet, M. K., Sobel, A. H., & Camargo, S. J. (2018). An environmentally forced  
628 tropical cyclone hazard model. *Journal of Advances in Modeling Earth Systems*, 10, 223–241.  
629 <https://doi.org/10.1002/2017MS001186>
- 630 Li, S. H. & Hong, H. P. (2016). Typhoon wind hazard estimation for China using an empirical  
631 track model. *Natural Hazards*, 82, 1009–1029. <https://doi.org/10.1007/s11069-016-2231-2>
- 632 Luo, W., & Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order  
633 determination. *Biometrika*, 103, 875–887. <https://doi.org/10.1093/biomet/asw051>
- 634 Nakamura, J., Lall, U., Kushnir, Y., & Rajagopalan, B. (2015). HITS: Hurricane Intensity and  
635 Track Simulator with North Atlantic Ocean Applications for Risk Assessment. *Journal of*  
636 *Applied Meteorology and Climatology*, 54(7), 1620–1636. <https://doi.org/10.1175/JAMC-D-14-0141.1>
- 638 R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for  
639 Statistical Computing, Vienna, Austria. <https://www.R-project.org>
- 640 Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*, Second Edition. New  
641 York, NY: Springer,
- 642 Rekabdarkolae, H. M., Krut, C., Fuentes, M., & Reich, B. J. (2019). A Bayesian multivariate  
643 functional model with spatially varying coefficient approach for modeling hurricane track data.  
644 *Spatial Statistics*, 29, 351–365. <https://doi.org/10.1016/j.spasta.2018.12.006>
- 645 Rumpf, J., Weindl, H., Höpfe, P., Rauch, E., & Schmidt, V. (2007). Stochastic modelling of  
646 tropical cyclone tracks. *Mathematical Methods of Operational Research*, 66(3), 475–490.  
647 <https://doi.org/10.1007/s00186-007-0168-7>
- 648 Rumpf, J., Weindl, H., Höpfe, P., Rauch, E., & Schmidt, V. (2009). Tropical cyclone hazard  
649 assessment using model-based track simulation. *Natural Hazards*, 48(3), 383–398.  
650 <https://doi.org/10.1007/s11069-008-9268-9>
- 651 Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S. & Boatwright, P. (2005). A useful distribution  
652 for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the*  
653 *Royal Statistical Society: Series C (Applied Statistics)*, 54: 127-142.  
654 <https://doi.org/10.1111/j.1467-9876.2005.00474.x>
- 655 Vickery, P. J., Masters, F. J., Powell, M. D., & Wadhwa, D. (2009). Hurricane hazard modeling:  
656 The past, present, and future. *Journal of Wind Engineering and Industrial Aerodynamics*, 97,  
657 392–405. <https://doi.org/10.1016/j.jweia.2009.05.005>
- 658 Vickery, P. J., Skerlj, P. F., & Twisdale, L. A. (2000). Simulation of hurricane risk in the U.S.  
659 using empirical track model. *Journal of Structural Engineering*, 126(10), 1222–1237.  
660 [https://doi.org/10.1061/\(ASCE\)0733-9445\(2000\)126:10\(1222\)](https://doi.org/10.1061/(ASCE)0733-9445(2000)126:10(1222))
- 661 World Meteorological Organization (2020). *World's deadliest tropical cyclone was 50 years*  
662 *ago*. Published: 12 November 2020. [https://public.wmo.int/en/media/news/world's-deadliest-](https://public.wmo.int/en/media/news/world's-deadliest-tropical-cyclone-was-50-years-ago)  
663 [tropical-cyclone-was-50-years-ago](https://public.wmo.int/en/media/news/world's-deadliest-tropical-cyclone-was-50-years-ago)

- 664 Yin, J., Welch, M. B., Yashiro, H., & Shinohara, M. (2009). *Basinwide Typhoon Risk Modeling*  
665 *and Simulation for Western North Pacific Basin*. Paper presented at The Seventh Asia-Pacific  
666 Conference on Wind Engineering, Taipei, Taiwan.
- 667 Yonekura, E., & Hall, T. M. (2011). A Statistical Model of Tropical Cyclone Tracks in the  
668 Western North Pacific with ENSO-Dependent Cyclogenesis, *Journal of Applied Meteorology*  
669 *and Climatology*, 50(8), 1725-1739. <https://doi.org/10.1175/2011JAMC2617.1>