A Quantitative Method to Evaluate Tropical Cyclone Tracks in Climate Models

YIXUAN SHEN

College of Meteorology and Oceanography, National University of Defense Technology, Nanjing, China

YUAN SUN

Joint International Research Laboratory of Climate and Environmental Change (ILCEC), Nanjing University of Information Science and Technology, and College of Meteorology and Oceanography, National University of Defense Technology, Nanjing, China

SUZANA J. CAMARGO

Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York

ZHONG ZHONG

College of Meteorology and Oceanography, National University of Defense Technology, Nanjing, China

(Manuscript received 4 April 2018, in final form 9 August 2018)

ABSTRACT

The ability to simulate tropical cyclones (TCs) realistically is an important factor in the performance evaluation of climate models. In previous studies, indirect evaluation methods have been proposed that are based on the comparison of TC-related background circulation between model results and observations. Direct model evaluation methods, in most cases, are limited to the model skill in simulating the TC frequency, intensity, and track density. Here we propose a new method to quantitatively and directly evaluate the ability of climate models in simulating TC tracks. The method consists of two indicators that account for the model performance in simulating TC track density and the geographic properties of TC tracks, respectively. This method is applied to evaluate the skill of climate models in simulating TC tracks over the western North Pacific Ocean. The explicit models include seven from phase 5 of the Coupled Model Intercomparison Project and eight from the U.S. CLIVAR Hurricane Working Group (HWG), as well as four downscaled HWG models. Our results indicate the order of these 15 explicit models according to their ability to simulate TC tracks. In addition, we show that, for one of the models, the TC track simulation is greatly improved by using downscaling.

1. Introduction

Tropical cyclones (TCs) are probably the most devastating of natural disasters, posing great threats to life and property along their paths (Tonkin et al. 1997; Henderson-Sellers et al. 1998). With the increase of horizontal resolution in recent years, climate models have become a powerful tool to study tropical TC activity, and the evaluation of the model skill in simulating TCs is increasingly important. In many previous case studies, the model skill for TC simulations was evaluated based on a direct comparison between observed and simulated TC tracks (Landman et al. 2005; Rogers 2010; Sun et al. 2017). However, this method is not appropriate for evaluating the TC track simulations in climate models, since climate models target the simulation of TC climatological characteristics rather than simulations of real TC cases. The purpose of evaluating TC tracks in climate model simulations is to evaluate the climatological characteristics of the TC tracks such as TC genesis locations, track lengths, and types.

One indirect method to evaluate the model skill of simulated TC activity is based on the comparison of observed and simulated circulation backgrounds associated with TC frequency (Zhou and Xu 2017; Zhou 2012) or the TC genesis potential index calculated from

Corresponding author: Yuan Sun, sunyuan1214@126.com

DOI: 10.1175/JTECH-D-18-0056.1

the large-scale environmental fields from models and reanalysis products (Song et al. 2015). However, these methods do not consider the simulated TCs per se. Several previous studies also used statistical analysis to compare the TCs' genesis location, occurrence frequency, intensity, and lifetime models and observations (Camargo et al. 2005; Bengtsson et al. 2007; Zhao et al. 2009). The most common diagnostics of model skill include a comparison of the number of simulated TCs, as well as the genesis position spatial distribution of TC with that of observations, their differences, and/or correlation coefficients (Camargo et al. 2005; Murakami et al. 2012, 2014; Camargo 2013; Wang and Wu 2015). TC track density is another important element in evaluating the skill of climate models in simulating TCs (Camargo et al. 2005; Kim et al. 2012; Strazzo et al. 2013), since it contains information on TC frequency, position, and duration. Some studies considered the difference in the TC track density between simulations and observations (Strazzo et al. 2013; Shaevitz et al. 2014; Kossin et al. 2016). Strazzo et al. (2013) proposed a method to compare a suite of models with each other and with observations using a hexagon spatial lattice framework first introduced into TC studies by Elsner et al. (2012). The hexagon lattice has the same function as the latitude–longitude grids. However, although this method is universal and provides a uniform framework, it evaluates only the attributes of the TC track that are included in TC track density without considering the geographical properties of TC tracks.

The geographical property is an important feature of the TC tracks. In previous studies, TC tracks were classified either over a study area that was geometrically divided into several segmentations on the basis of latitude/longitude or over a study area that was determined from the spatial distribution of TC occurrence (Wu and Wang 2004; Wu et al. 2005). Objective analysis has also been applied for TC track classification; for example, Elsner (2003) implemented the *k*-means clustering method (MacQueen 1967) to classify TCs over the North Atlantic. Following that work, Nakamura et al. (2009) optimized the parameters used in the *k*-means clustering for storm classification by developing a method to distill the track shape and length using mass moments. Alternatively, Camargo et al. (2007a,b) classified TCs over the western North Pacific Ocean (WNP) using a regression mixture cluster model proposed by Gaffney et al. (2007). This latter cluster method was used together with track moments in Nakamura et al. (2017) to analyze WNP model tracks. In Daloz et al. (2015) and Nakamura et al. (2017), the model skill was evaluated based on the full TC track, using a regression mixture cluster analysis to classify observed and simulated TC tracks. The model's performance was evaluated based on the comparison of the cluster memberships between observations and simulations. However, the results in the studies were still somewhat qualitative.

Thereby, it is necessary to develop a new method that can objectively evaluate the ability of climate models to stimulate not only the TC track density but also the geographical properties of TC tracks in a comprehensive and quantitative way. The aim of the present study is to propose a new method that can directly and quantitatively evaluate the performance of climate models in simulating TC tracks. This new method will then be implemented to evaluate the TC tracks from models from phase 5 of the Coupled Model Intercomparison Project (CMIP5) (Taylor et al. 2012) and the U.S. CLIVAR Hurricane Working Group (HWG) (Shaevitz et al. 2014; Walsh et al. 2015; Daloz et al. 2015; Nakamura et al. 2017).

The paper is organized as follows. Section 2 introduces the data used in the present study. The method proposed to evaluate model skill in simulating TC tracks is provided in section 3. A comparison of the simulations and observations and a quantitative evaluation of the model skill are presented in section 4. Conclusions and a discussion are given in section 5.

2. Data

The data used in this study include the TC best-track data from observations and TC tracks from simulations of seven CMIP5 and eight HWG global climate models. The TC observations are extracted from International Best Track Archive for Climate Stewardship (IBTrACS), version v03r09, for the period of 1980–2005, which provides TC position and intensity information at 6-h intervals (Knapp et al. 2010). Similarly, TC model data used in this study include global TC genesis positions and tracks obtained from the outputs of CMIP5 and HWG models. Similar to observations, the model tracks provide information on latitude/longitude of the TC center, the maximum wind speed, and minimum sea level pressure. Differences in horizontal resolution among the climate models are taken into account by varying the thresholds for the tracking algorithm. Details can be found in Camargo (2013) for the CMIP5 models and Shaevitz et al. (2014) and Nakamura et al. (2017) for the HWG models. The sensitivity of the HWG models' TC statistics to different tracking algorithms is discussed in Horn et al. (2014). The models included in the present study are listed in Table 1. The horizontal resolutions of the CMIP5 models are

TABLE 1. List of the CMIP5 and HWG models analyzed this study. The columns show the model name, horizontal resolution, model type (i.e., CMIP5 or HWG), number of simulated TCs in the WNP, number of simulation years, annual mean number of TCs, and TC track density simulation index. Here, LR, MR, and HR indicate low, medium, and high resolution, respectively, and DX gives the name of a downscaled model corresponding to the original model X, e.g., HiRAM and DHiRAM.

Model	Resolution (°)	Type	No.	Years	Annual No.	DSI
CSIRO Mk3.6.0	1.9	CMIP5	459	26	17.65	0.42
CanESM2	2.9	CMIP5	67	26	2.58	0.18
FGOALS-g2	3.0	CMIP5	35	26	1.35	0.18
GFDL CM3	2.5	CMIP5	188	26	7.23	0.29
IPSL-CM5A-LR	3.7	CMIP5	18	26	0.69	0.17
MIROC5	1.4	CMIP5	55	26	2.12	0.18
MPI-ESM-LR	1.9	CMIP5	109	26	4.19	0.25
CAM5.1 HR	0.25	HWG	153	16	9.56	0.41
CMCC/ECHAM5	0.75	HWG	303	9	33.67	0.42
FSU	1	HWG	143	5	28.60	0.41
GFS	1	HWG	118	20	5.90	0.27
GISS	1	HWG	29	20	1.45	0.24
HadGCM3 MR	0.83	HWG	134	20	11.90	0.37
HiRAM	0.5	HWG	677	20	33.85	0.49
MRI	1.25	HWG	441	25	17.64	0.36
DCAM5	—	Downscaled HWG	2987	19	32.79	0.52
DCMCC	—	Downscaled HWG	2858	19	20.84	0.29
DGISS	—	Downscaled HWG	2799	19	28.50	0.27
DHiRAM	—	Downscaled HWG	2575	19	30.72	0.30

typically much lower (i.e., coarser) than those of the HWG models. More information about the HWG simulations can be found in various publications from the HWG, for example, Horn et al. (2014), Daloz et al. (2015), Walsh et al. (2015), and Nakamura et al. (2017). In addition to the CMIP5 models and HWG models, TC tracks produced by a statistical–dynamical downscaling method for four HWG models (i.e., DCAM5, DCMCC, DGISS, and DHiRAM) are also analyzed. The downscaling technique can be divided into three steps. First, origin points of the tracks are generated by a random seeding procedure. The survival of these seeds depends on the environmental conditions at the seeding location. Second, the storm movement is determined by the steering winds, with a correction for the beta drift (Holland 1983; Marks 1992). Third, storm intensity is determined by a coupled TC intensity model (Emanuel et al. 2004) that is run along the storm track. Details can be found in Emanuel (2006) and Emanuel et al. (2006). There are alternative ways to do each step of the synthetic tracks generation. For instance, Lee et al. (2018) genesis seeding is weighed by the tropical cyclone genesis index (Tippett et al. 2011).

The observations and simulations of CMIP5 models are analyzed for the period of 1980–2005. Only those observed TCs that reached the intensity of tropical storm ($\geq$35 kt; 17.85 m s^{-1}) are included in our analysis. During this period, the best-track data are considered to be the most complete with the highest quality in terms of both storm position and storm intensity as a result of the monitoring of geostationary satellites (Knapp and Kruk 2010; Kossin et al. 2014, 2016). The study period ends at 2005 to match the CMIP5 historical simulations, which cover the period of 1851–2005. For the HWG multimodel datasets, the simulations are forced with monthly varying climatological sea surface temperatures (SST); that is, the SST is constant from year to year. The number of years available is different for different models (Table 1). Although the SST climatological period (1985–2001) of the HWG models is not fully consistent with those in observations, it does not affect the reliability of the main conclusions regarding the model performance evaluation in the present study. As the HWG models are forced by climatological SST, the time-averaged results are also the climatological mean state of the WNP TC tracks. For the observations and simulations of CMIP5's multiple models, the 26-yr averaged results can be considered as the climatological mean state of the WNP TC tracks because the length of the dataset is longer than the periods of the dominant modes of natural variability (e.g., El Niño–Southern Oscillation and Pacific decadal oscillation).

3. Evaluation method

The evaluation metric is divided into two subindices—that is, the TC track density simulation index (DSI) and the index of geographical properties of the TC track

(GPT)—that jointly determine the final skill score of a given climate model.

a. TC track DSI

Only TCs that were active over the WNP region (0°–60°N, 100°E–180°) are considered in this study. This area is divided into 29 × 39 grid boxes with 2° × 2° horizontal resolution. The TC data at 6-hourly intervals are considered to be independent samples, and each occurrence of the TC center at a given grid is considered as one TC exposure. The annual average TC exposure at each individual grid box is then calculated.

Suppose that the model-simulated annual mean TC exposure at each grid box is m and that the observed exposure is n; both m and n have units of number per area per year. For the ith grid box, the skill score of the model can be expressed as

$$\mathrm{DSI}(i) = \begin{cases} n/m, & \text{if} \quad m > n, \quad m > 0 \\ m/n, & \text{if} \quad m \le n, \quad n > 0 \end{cases}. \tag{1}$$

Here a valid grid box is defined as the grid box in which there exists either observed or simulated TC exposure. Many previous studies have applied this method to evaluate short-term climate forecasts, for example, in the applications of the threat score (TS) (Palmer and Allen 1949) and the critical success index (CSI) (Donaldson et al. 1975). In a forecast skill score, the "hit" rate (or "correct forecast" rate) will be unrealistically high if "correct nulls" (or correct rejections), which represent the cases that the event occurs in neither observations nor forecasts, are considered as correct forecasts. For this reason, correct nulls are not counted as correct forecasts in the calculation of skill scores. Similarly, in our analysis, the grid boxes in which there exists neither observed nor simulated TC exposure are invalid and thus are excluded from our calculation of skill scores. In other words, if there exists at least one TC exposure at a given grid box in either observations or simulations from any of the 15 models, then this grid box is considered valid. Note that when comparing the simulation of one specific model with observations, there might exist cases in which the TC exposure from both observations and that specific model simulation is zero in a given grid box. However, if the TC exposure in any of the other model simulations (or in at least one model simulation) is not zero in that grid box, then that grid box is still considered as valid. For example, when we calculate the DSI of the CMCC, there might exist a few grid boxes where the TC exposure from both the CMCC and observations is zero ($m = 0$ and $n_{\mathrm{CMCC}} = 0$) but it is not zero from simulations of the CSIRO or some other model ($n_{\mathrm{CSIRO}} = 1$). This situation is also considered to be a correct forecast of the CMCC, and DSI = 1 in that grid point. The value of DSI(i) over the entire region is calculated following the approach described above, and the track DSI is the averaged value of DSI(i) over all valid grid boxes.

b. Index of GPT

The k-means clustering is a method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. In previous studies (e.g., Elsner 2003; Nakamura et al. 2009, 2017; Yu et al. 2016), the k-means clustering has been used as an effective method to classify TC type. Here we consider the modified k-means method used by Nakamura et al. (2009) for TC track classification. Following Nakamura et al. (2009), the wind speed at each specific time and location of the track is used as a weighting factor in the computation of the centroid and variance ellipse of each individual TC track. The centroid and variance ellipse are characteristic variables that describe the TC track pattern, length, and location. The mass moment, which contains five elements—that is, the latitude and longitude of the TC centroid; and the variances of the TC centroid along the latitude, longitude, and diagonal directions—is used to describe the geographical properties of a full TC track such as its pattern and length in each of the clusters determined by the k-means method.

The five track moments elements are first normalized. The weights of the two elements associated with the centroid and that of the other three elements associated with the variances at different directions are set to 1/3 and 1/9, respectively, to relatively weaken the effects of TC track length, pattern, and direction represented by the variances, and the cosine distance metric is chosen as our distance metric. A detailed description of the method can be found in Nakamura et al. (2009, 2017).

The Nakamura k-means clustering is applied to the observed best-track data to classify all TCs that occurred during 1980–2005. The tracks are classified into k clusters; $F_{s,i}$ represents the fraction of simulated tracks in the ith cluster and $F_{o,i}$ is the actual fraction of tracks observed in the ith cluster. The index of geographical properties of the TC track (i.e., the accuracy of model in simulating geographical properties of the TC track) can be written as

$$\mathrm{GPT} = 1 - \left[\frac{1}{k}\sum_{i=1}^{k}(F_{s,i} - F_{o,i})^2\right]^{1/2}. \tag{2}$$

Note that GPT is merely the root-mean-square error of the cluster fractions.

c. CTI

Combining the TC track DSI and the index of GPT, we can define the comprehensive track index (CTI) as

$$\mathrm{CTI} = \mathrm{DSI} \times \mathrm{GPT}. \quad (3)$$

This definition of CTI as the product of DSI and GPT requires that the models have skill to simulate both the track density and the TC track characteristic, since both are important and should not be ignored.

4. Evaluation of the performance of 15 models for TC track simulations over the WNP

a. Comparison of TC track density between observations and simulations

To intuitively compare the difference of TC tracks between observations and simulations, Fig. 1 shows the TC track density distributions from observations and simulations of CMIP5 and HWG models. Compared with observations, the TC occurrence frequency over the WNP is underestimated by most of the CMIP5 models (Figs. 1b–h). This bias is mainly attributed to the low resolution of the CMIP5 models, although detection and tracking algorithms used to identify TCs may also contribute to the bias as well (Horn et al. 2014; Walsh et al. 2015). The TC track density distributions simulated by the HWG models (Figs. 1i–p) are better than those by the CMIP5 models, perhaps because of the relatively higher resolutions. However, there are still large differences between the HWG simulations and observations. For instance, while the simulations of CSIRO Mk3.6.0 (Fig. 1b) and HiRAM (Fig. 1p) are relatively consistent with observations, the observed maximum over the South China Sea is not simulated by HiRAM and CSIRO overestimates its value. Moreover, fewer TC occurrences are found in the HiRAM simulation over the northeastern part of the WNP compared with observations, which may be attributed to a shorter life cycle (or shorter tracks) of the modeled TCs. In addition, the values of TC track density in the CSIRO and HiRAM simulations are larger than observed in the low latitudes east of 160°E, which is due to the higher rate of TC genesis in this region. The performance of other models for the TC track simulation is even poorer.

For the HWG downscaled synthetic tracks (Figs. 1q–t), the number of generated storms is fixed globally. Therefore, the observations need to be compared with a normalized frequency of TCs. There are many ways to do that. For instance, Lee et al. (2018) normalized the TC number based on the tropical cyclone genesis index (Tippett et al. 2011). Here we normalize the TC number by the frequency of storm survival after being generated by the random seeding method (Emanuel 2006) to obtain a TC track density that is equivalent to observations. The calculation is as follows. First, the number of storms during the 19 years listed in Table 1 in the WNP is divided by the total global number of events for each simulation (i.e., 8000) to obtain the percentage of storms in the WNP for DCAM5, DCMCC, DGISS, and DHiRAM (37%, 36%, 35%, and 32%, respectively). Second, the frequency of surviving storms in the WNP is obtained by multiplying the mean number of the global events generated for each model and every year by the percentages calculated in the previous step. Finally, the original track density is normalized by multiplying by the quotient obtained by dividing the number of storms in Table 1 by the frequency values. Although the normalized TC track density in the synthetic TC tracks is generally larger than that in observations, they are more similar to the observations than the HWG explicit models' TC track densities. The synthetic TCs correctly reproduce the observed region of maximum track density, while the HWG explicit TC track density generally does not. However, in the northeastern portion of the WNP, the synthetic TC track density is overestimated.

Although we have qualitatively analyzed the model performance in simulating the TC track density, the above discussion is affected by subjective factors. To eliminate the influence of these subjective factors, we have also quantitatively analyzed the TC track simulations by all the models. The results are given in Table 1, which lists the track density indices (DSI) of all 15 models based on the mean number of TCs. Among all of the models, CSIRO, HiRAM, and FSU (17.65, 33.85, and 28.6 separately) agree relatively well with observations (25.38) in the mean annual number of TCs, and the simulated TC track density patterns are consistently close to observations (Figs. 1b,k,p). The number of TCs simulated by the IPSL (0.69) is the smallest in these models and also the most distinct when compared with observations. Meanwhile, the value of DSI is the highest (0.49) for HiRAM and the smallest (0.17) for IPSL-CM5A-LR among the explicit models (Table 1). The above results reflect the relationship between TC number and DSI. To further verify this point, the relationship between the TC number and DSI is illustrated in Fig. 2a, which shows that DSI is highly correlated with the annual number of TCs simulated by each model and the correlation coefficient between them is 0.76. In general, the models' DSI increases as the simulated annual number of TC becomes closer to observations. Models with high DSI values, such as CSIRO Mk3.6.0, CMCC, and FSU (0.4 and higher), have mean number of TCs closer to observations. Note that while HiRAM has

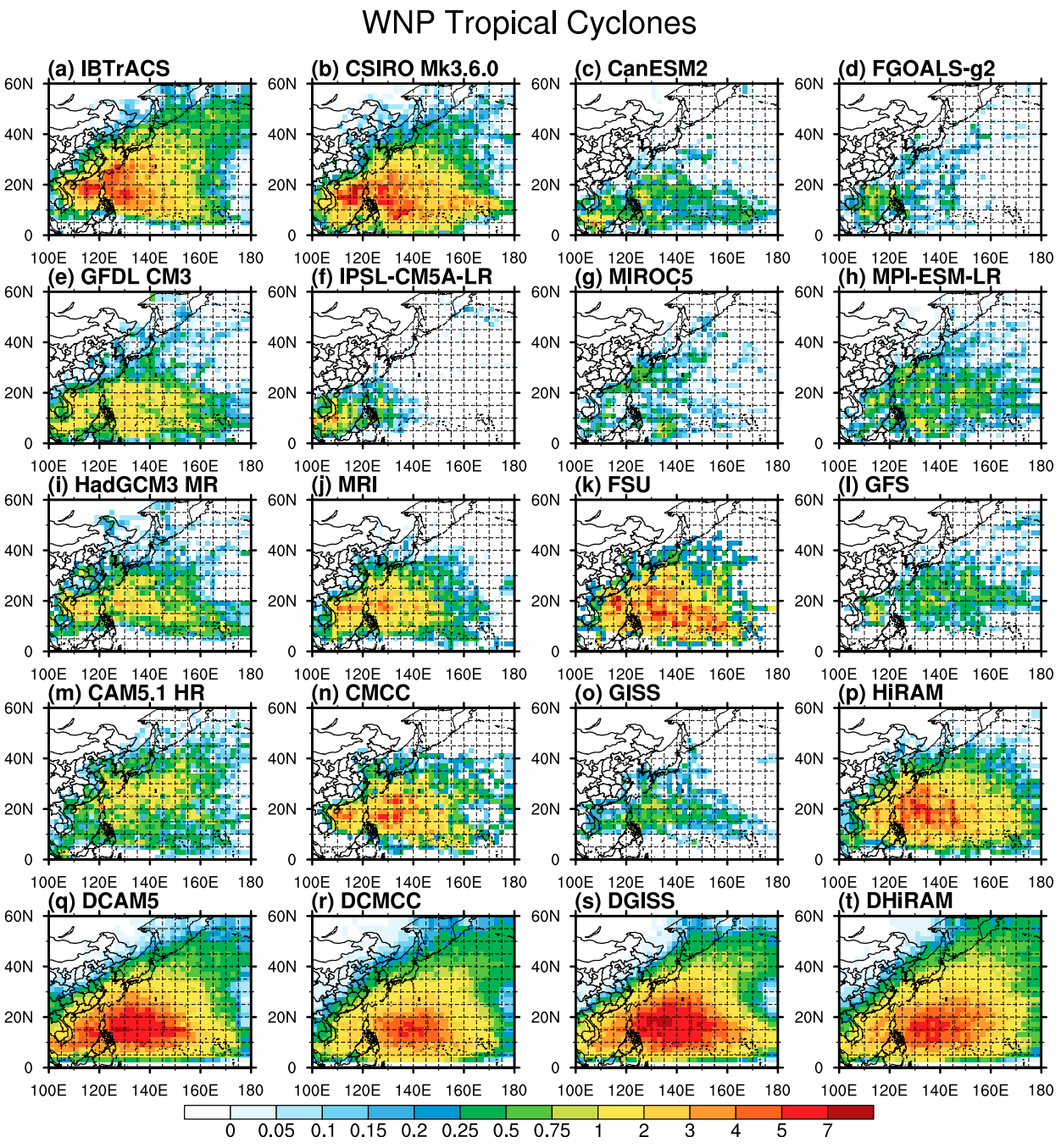

FIG. 1. Observed and simulated TC track density distributions measured by the number of events per 2° × 2° grid box per year: (a) observations (IBTrACS), (b)–(h) CMIP5 models, (i)–(p) HWG models, and (q)–(t) HWG downscaled models.

the highest DSI value, its annual number of TCs is not the closest to observations. Moreover, although CMCC and HiRAM have very similar annual mean number of TCs (33.67 and 33.85, respectively), their DSIs are different (0.42 and 0.49, respectively). Thereby, the value of DSI is also influenced by other factors (e.g., TC genesis position, TC lifetime).

As for the downscaled models (i.e., DCAM5, DCMCC, DGISS, and DHiRAM), we normalized TC track density and calculated the DSI based on the normalized TC track density. As shown in Table 1 and Fig. 2a, the DSI of DCAM5 (0.52) is the highest among these models and is significantly higher than that of its corresponding model CAM5 (0.41). The DSIs of the DGISS and GISS are very

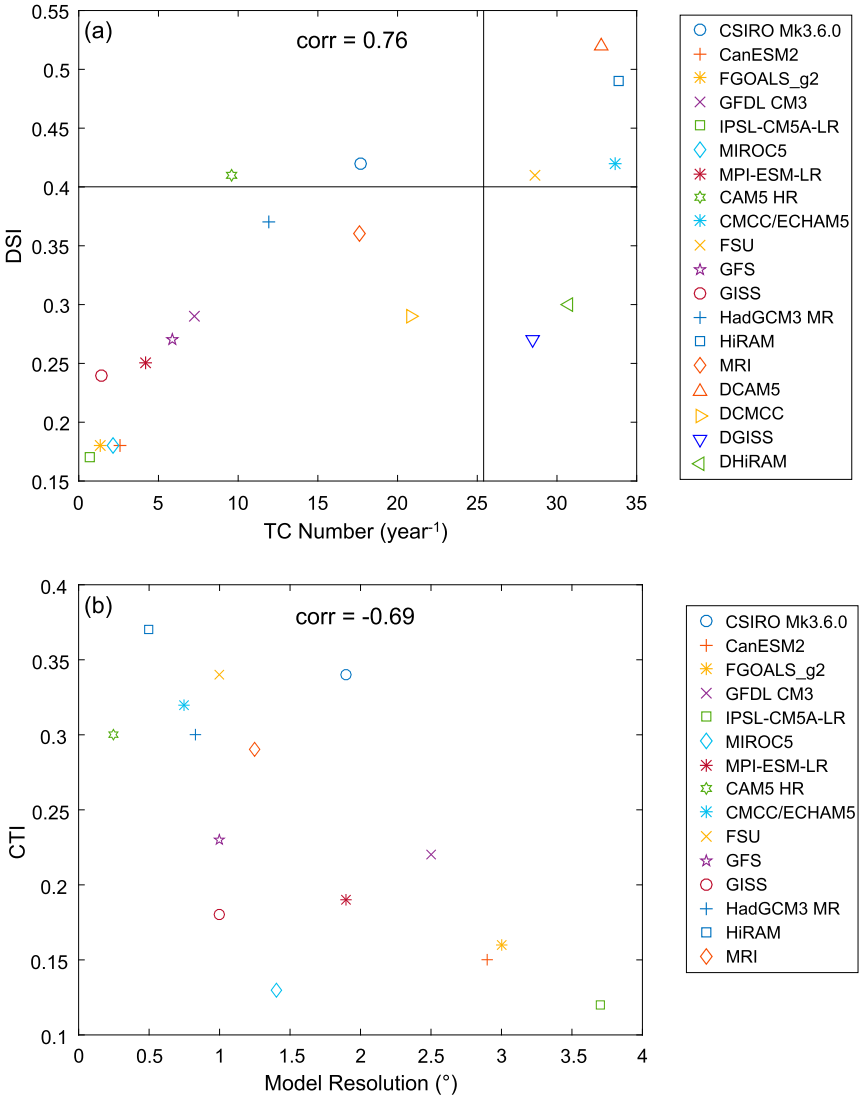

FIG. 2. (a) Scatterplot of the DSI vs TC number (per year) for 15 explicit models and 4 downscaled models. The horizontal line denotes DSI = 0.4, and the vertical line represents the observed TC number (25.38). (b) Scatterplot of CTI vs model resolution for 15 explicit models.

similar, while the DSIs of the DCMCC and DHiRAM are lower than that of CMCC and HiRAM. This indicates that the downscaling method does not always improve the model performance for the simulation of TC track density.

b. *Evaluation of the performance of 15 models on TC track type simulation based on IBTrACS dataset*

Results of the *k*-means clustering analysis of the IBTrACS best-track data are shown in Figs. 3a–c, which show the TC tracks, the average TC tracks, and the starting points of individual TCs in each individual clusters. We chose $k = 3$ for three reasons. First, the three patterns of TC track are very simple and straightforward; second, several previous studies have classified TC tracks into three track types, that is, westward, northwestward, and recurving (Elsner 2003; Wu and Wang 2004; Wu et al. 2005; Ying et al. 2011); third, for some models that simulate only a low number of TCs, three types of TC tracks may be more appropriate than a higher number of clusters, because the *k*-means clustering analysis is a hard clustering method, and there are potential uncertainties in the resulting clusters because of the random selection of initial centroids. However, multiple repetitive experiments indicate that the resulting clusters are robust for $k = 3$; that is, no matter how the initial centroids are

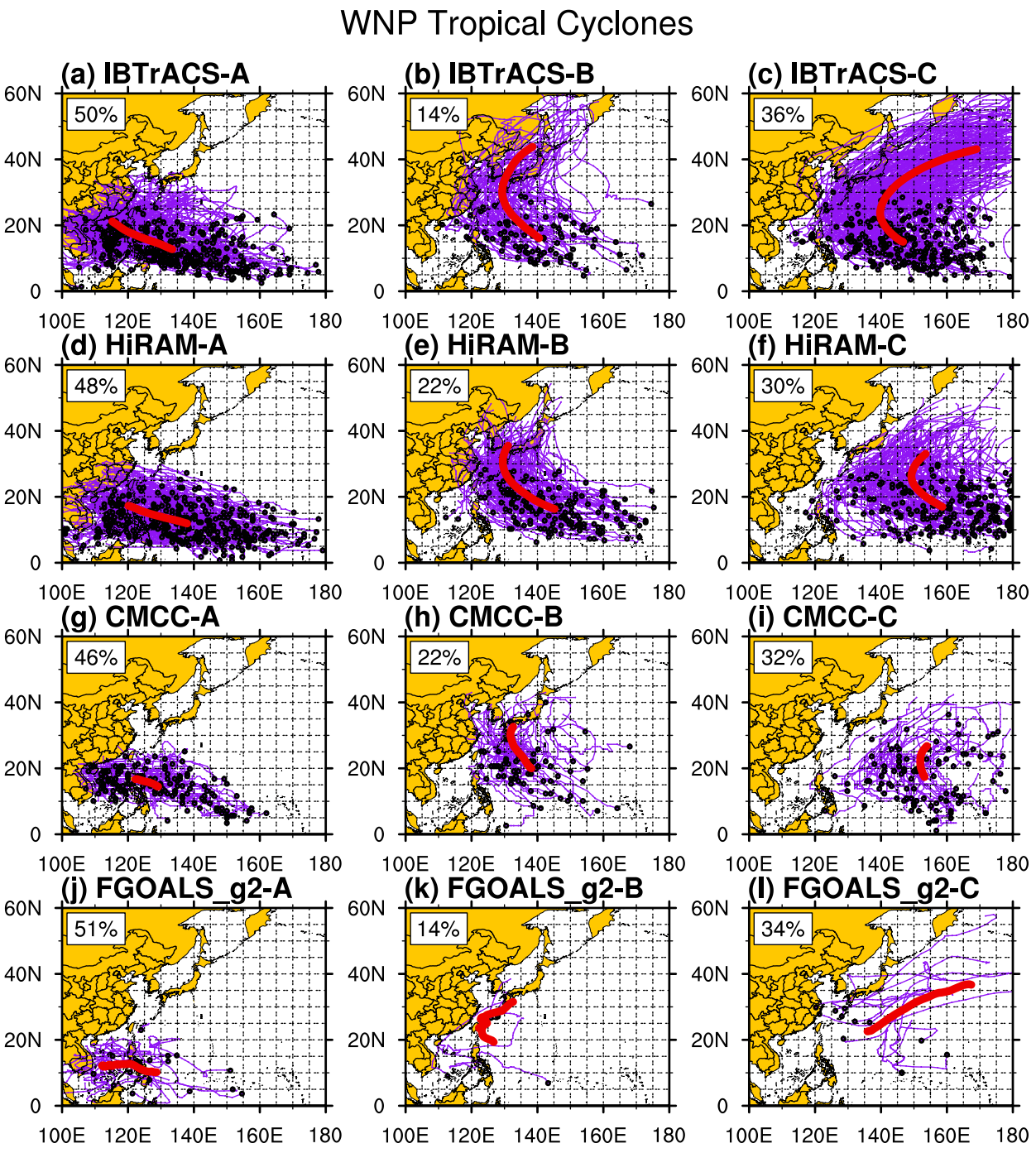

FIG. 3. TC tracks, initial positions, and mean tracks in three *k*-means clusters [track types (left) A, (center) B, and (right) C] for (a)–(c) observations (IBTrACS) and (d)–(l) three models (HiRAM, CMCC, and FGOALS-g2). The text box in the upper left-hand corner of each panel represents the percentages for each cluster.

chosen, a very high percentage of individual TCs are classified into the same specific cluster with the same centroid. This result indicates that the method proposed can be used to evaluate the performance of TC simulations in climate models.

TCs in cluster A (Fig. 3a) are usually active in the southern part of the WNP, and they tend to move westward (with no TCs moving past 40°N), eventually making landfall in the Philippines, China, or Vietnam. TCs in cluster B typically move westward and make

landfall in Japan and the Korean Peninsula (Fig. 3b), affecting the East China Sea. TCs in cluster C typically have genesis east of the Philippines and move northwestward before turning northeastward at around 20°N (Fig. 3c), with some TCs affecting Japan and the Korean Peninsula.

Because of the large number of models, here we only show the tracks of three explicit models that produce large, medium, and small TC numbers, respectively (Figs. 3d–l). A model TC track is assigned to the nearest cluster using the distance metric. While the HiRAM TC frequency is close to the observed, there exists significant differences between the percentages of TCs assigned to specific clusters when compared with observations. From observations, there are more TCs in clusters A (50%) and C (36%) than in cluster B (14%); for HiRAM, however, the percentage of TCs in cluster B (22%) is close to that in cluster C (30%). Similar issues are found for TCs simulated by CMCC (46%, 22%, 32%, respectively), which has a bias in the simulated TC number. It is interesting to note that, while the total number of TCs simulated by FGOALS-g2 is also very small, the percentages of TCs in the three clusters (51%, 14%, 34% for clusters A, B, C, respectively) are more consistent with observations than those from the CMCC simulation. To quantitatively evaluate the model performance for simulating the geographic properties of TC tracks, we calculate the GPTs of all models in the present study (Table 2). The GPTs of the downscaled models (0.83, 0.78, 0.79, 0.82 for DCAM5, DCMCC, DGISS, DHiRAM, respectively) are higher than their corresponding explicit models (0.73, 0.76, 0.74, 0.77 for CAM5, CMCC, GISS, HiRAM, respectively), especially DCAM5, for which the normalized annual number of downscaled TCs is closer to the observations. Although it might appear that GPT may be related to the simulated TC number, this relationship does not hold in the explicit models. Among the explicit models, FGOALS-g2 has the best GPT performance, with a value of 0.88, although it only produces 1.35 TCs per year (Table 1). The annual TC number produced by FSU (28.6) is the closest to the observation (25.38) among all the models, whereas its GPT (0.83) ranks fourth. In addition, the GPT for CMCC (0.76) is smaller than that for FGOALS-g2, which is consistent with the qualitative analysis of the percentage of TCs per cluster above. Overall, the skill of HiRAM for partitioning the WNP TCs among different track types is not optimal although its TC frequency is close to observations. On the other hand, FGOALS-g2 shows the highest GPT score despite its relatively poor performance in simulating the TC number. Thereby, as expected, the GPT index is sensitive only to the percentage of total TCs in each TC track cluster and is not sensitive to the number of TCs, and the correlation coefficient between TC numbers and GPT is only 0.18.

TABLE 2. Skill scores of 15 models verified against observations (IBTrACS). The observational dataset or model name, percentage of TC track types in each cluster, index of geographical properties of TC track, and comprehensive index of TC track properties are given in individual columns. Models in boldface type are discussed in more detail in the text as representative. The first row indicates observations, and the last four rows indicate downscaled models.

Model	Cluster A (%)	Cluster B (%)	Cluster C (%)	GPT	CTI
IBTrACS	50	14	36		
HiRAM	48	22	30	0.77	0.37
CSIRO Mk3.6.0	45	18	37	0.82	0.34
FSU	45	18	36	0.83	0.34
CMCC	46	22	32	0.76	0.32
CAM5	41	24	35	0.73	0.30
HadGCM3MR	48	19	33	0.80	0.30
MRI	52	17	32	0.82	0.29
GFS	50	17	33	0.85	0.23
GFDL CM3	58	13	29	0.77	0.22
MPI-ESM-LR	53	22	32	0.77	0.19
GISS	45	24	31	0.74	0.18
FGOALS-g2	51	14	34	0.88	0.16
CanESM2	55	13	31	0.81	0.15
MIROC5	45	24	31	0.74	0.13
IPSL-CM5A-LR	61	6	33	0.73	0.12
DCAM5	51	17	32	0.83	0.43
DHiRAM	49	18	33	0.82	0.25
DCMCC	47	21	32	0.78	0.23
DGISS	47	20	33	0.79	0.21

c. *Comprehensive evaluation of 15 climate models simulating TC track*

The comprehensive index CTI combines the TC track density (i.e., DSI) and TC track properties (i.e., GPT). Table 2 indicates that, among the explicit models, the CTI of HiRAM is the highest (0.37), followed by that of CSIRO Mk3.6.0 (0.34) and FSU (0.34). These models all have relatively better ability for simulating the TC track density (i.e., relatively higher DSI) and TC track properties (i.e., relatively higher GPT). FGOALS-g2 has the highest GPT but a CTI of only 0.16 because of its poor performance in simulating TC track density (i.e., relatively lower DSI). A high CTI indicates a proper balance of high DSI and GPT. Among the downscaled models, the CTI of DCAM5 is the highest (0.43) and is higher than all explicit models. For GISS, the CTI has similar values for the explicit and downscaled TCs, while the CTIs of DCMCC and DHiRAM are lower than their corresponding models. Therefore, among the explicit models, HiRAM has the highest resolution and shows

the highest skill in simulating the WNP TC tracks, which agrees well with the results of Strazzo et al. (2013) that HiRAM matches well with observations over the WNP in terms of the area covered by TC tracks. IPSL-CM5A-LR, which is the model with the lowest resolution, also has the lowest skill. Consistent with the results of Nakamura et al. (2017), the performance of a given model in simulating the TC track improves as the model resolution increases. To further investigate this issue, we present Fig. 2b, which shows the relationship between model performance and model resolution. As model resolution increases, in general, the model's CTI also increases, and their correlation coefficient is −0.69. Apart from the model resolution, the number of TCs also plays an important role in simulating the TC track density and affects the model performance (Fig. 2a). Furthermore, for climate models that underestimate the number of TCs, the downscaling method can greatly improve the model performance (e.g., CAM5).

5. Conclusions and discussion

A new method to evaluate the skills of climate models in simulating TC tracks is proposed in this study. This method considers not only the TC track density but also the TC track geographical properties using objective skill scores. The WNP TC tracks by 15 models from CMIP5 and HWG are compared with the IBTrACS best-track data, using the comprehensive track index, which consists of the TC track density simulation index and TC track geographical properties index. Results indicate that, among the 15 explicit climate models, HiRAM has the best performance in simulating the WNP TC track density and IPSL-CM5A-LR has the poorest performance; FGOALS-g2 has the best skill score for TC track properties, followed by CMCC and CSIRO Mk3.6.0. When these two indexes are considered together, HiRAM has the best performance, followed by CSIRO Mk3.6.0 and FSU. Note that although the explicit model CAM5 gets a low comprehensive index of TC track properties (i.e., CTI), its corresponding downscaled model DCAM5 performs very well in simulating the TC track and the performance is even better than of HiRAM. This result indicates that, for some climate models (e.g., CAM5), the downscaling method can improve the performance of models in simulating TC tracks.

The objective of the present study is to propose a straightforward method that can be used to quantitatively evaluate the skills of climate models in simulating TC tracks. Namely, by using the method, we can provide a score (i.e., CTI) for each model and then directly compare the skills of different models in simulating TC tracks. Note that there are still some shortcomings in this method. For example, the DSI of a given climate model is sensitive to the choice of models that we need to compare their abilities in simulating TC track. Namely, the DSI of a given climate model in a set of models may differ from the score of the same model in another set of models, because of the difference in valid grid boxes when calculating DSI with Eq. (1). Meanwhile, the classification of TC tracks (i.e., GPT) may change when the method is used in different areas and/or is based on different observed TC best-track datasets. As a result, the absolute values of DSI and GPT, and thus CTI, are sensitive to the models analyzed, the selected areas, and the observed TC best-track datasets. Thereby, it is meaningless to analyze the absolute value of CTI of a given model, but we can rank the selected models by comparing the CTIs of models. Although this method is not perfect and still needs improvement, it is a first attempt to shed light on this issue. For instance, some advanced cluster analysis may be more appropriate than the *k*-means clustering analysis in calculating GPT. In the future, we plan to explore ways to assess the reliability of the method proposed in this study.

Acknowledgments. The authors thank three anonymous reviewers for their constructive comments, which helped to improve the manuscript. The authors thank all of the members of the U.S. CLIVAR Hurricane Working Group for their contribution to this significant effort, in particular those who produced the model simulations used in this study. We also thank Naomi Henderson for managing the HWG dataset. This work is sponsored by China NSF Grants 41605072 and 41430426 and Jiangsu Projects BK20160768. Author SJC acknowledges support from the National Oceanic and Atmospheric Administration through Grants NA15OAR4310095 and NA16OAR4310079.

REFERENCES

Bengtsson, L., K. I. Hodges, and M. Esch, 2007: Tropical cyclones in a T159 resolution global climate model: Comparison with observations and re-analyses. *Tellus*, **59A**, 396–416, https://doi.org/10.1111/j.1600-0870.2007.00236.x.

Camargo, S. J., 2013: Global and regional aspects of tropical cyclone activity in the CMIP5 models. *J. Climate*, **26**, 9880–9902, https://doi.org/10.1175/JCLI-D-12-00549.1.

——, A. G. Barnston, and S. E. Zebiak, 2005: A statistical assessment of tropical cyclone activity in atmospheric general circulation models. *Tellus*, **57A**, 589–604, https://doi.org/10.3402/tellusa.v57i4.14705.

——, A. W. Robertson, S. J. Gaffney, P. Smyth, and M. Ghil, 2007a: Cluster analysis of typhoon tracks. Part I: General properties. *J. Climate*, **20**, 3635–3653, https://doi.org/10.1175/JCLI4188.1.

——, ——, ——, ——, and ——, 2007b: Cluster analysis of typhoon tracks. Part II: Large-scale circulation and ENSO. *J. Climate*, **20**, 3654–3676, https://doi.org/10.1175/JCLI4203.1.

Daloz, A. S., and Coauthors, 2015: Cluster analysis of downscaled and explicitly simulated North Atlantic tropical cyclone tracks. *J. Climate*, **28**, 1333–1361, https://doi.org/10.1175/JCLI-D-13-00646.1.

Donaldson, R. J., R. M. Dyer, and M. J. Krauss, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, *Ninth Conf. on Severe Local Storms*, Norman, OK, Amer. Meteor. Soc., 321–326.

Elsner, J. B., 2003: Tracking hurricanes. *Bull. Amer. Meteor. Soc.*, **84**, 353–356, https://doi.org/10.1175/BAMS-84-3-353.

——, R. E. Hodges, and T. H. Jagger, 2012: Spatial grids for hurricane climate research. *Climate Dyn.*, **39**, 21–36, https://doi.org/10.1007/s00382-011-1066-5.

Emanuel, K., 2006: Climate and tropical cyclone activity: A new model downscaling approach. *J. Climate*, **19**, 4797–4802, https://doi.org/10.1175/JCLI3908.1.

——, C. DesAutels, C. Holloway, and R. Korty, 2004: Environmental control of tropical cyclone intensity. *J. Atmos. Sci.*, **61**, 843–858, https://doi.org/10.1175/1520-0469(2004)061<0843:ECOTCI>2.0.CO;2.

——, S. Ravela, E. Vivant, and C. Risi, 2006: A statistical deterministic approach to hurricane risk assessment. *Bull. Amer. Meteor. Soc.*, **87**, 299–314, https://doi.org/10.1175/BAMS-87-3-299.

Gaffney, S. J., A. W. Robertson, P. Smyth, S. J. Camargo, and M. Ghil, 2007: Probabilistic clustering of extratropical cyclones using regression mixture models. *Climate Dyn.*, **29**, 423–440, https://doi.org/10.1007/s00382-007-0235-z.

Henderson-Sellers, A.,and Coauthors, 1998: Tropical cyclones and global climate change: A post-IPCC assessment. *Bull. Amer. Meteor. Soc.*, **79**, 19–38, https://doi.org/10.1175/1520-0477(1998)079<0019:TCAGCC>2.0.CO;2.

Holland, G. J., 1983: Tropical cyclone motion: Environmental interaction plus a beta effect. *J. Atmos. Sci.*, **40**, 328–342, https://doi.org/10.1175/1520-0469(1983)040<0328:TCMEIP>2.0.CO;2.

Horn, M., and Coauthors, 2014: Tracking scheme dependence of simulated tropical cyclone response to idealized climate simulations. *J. Climate*, **27**, 9197–9213, https://doi.org/10.1175/JCLI-D-14-00200.1.

Kim, J. H., C. H. Ho, H. S. Kim, and W. Choi, 2012: 2010 western North Pacific typhoon season: Seasonal overview and forecast using a track-pattern-based model. *Wea. Forecasting*, **27**, 730–743, https://doi.org/10.1175/WAF-D-11-00109.1.

Knapp, K. R., and M. C. Kruk, 2010: Quantifying interagency differences in tropical cyclone best-track wind speed estimates. *Mon. Wea. Rev.*, **138**, 1459–1473, https://doi.org/10.1175/2009MWR3123.1.

——, ——, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

Kossin, J. P., K. A. Emanuel, and G. A. Vecchi, 2014: The poleward migration of the location of tropical cyclone maximum intensity. *Nature*, **509**, 349–352, https://doi.org/10.1038/nature13278.

——, ——, and S. J. Camargo, 2016: Past and projected changes in western North Pacific tropical cyclone exposure. *J. Climate*, **29**, 5725–5739, https://doi.org/10.1175/JCLI-D-16-0076.1.

Landman, W. A., A. Seth, and S. J. Camargo, 2005: The effect of regional climate model domain choice on the simulation of tropical cyclone–like vortices in the southwestern Indian Ocean. *J. Climate*, **18**, 1263–1274, https://doi.org/10.1175/JCLI3324.1.

Lee, C.-Y., M. K. Tippett, A. H. Sobel, and S. J. Camargo, 2018: An environmentally forced tropical cyclone hazard model. *J. Adv. Model. Earth Syst.*, https://doi.org/10.1002/2017MS001186, in press.

MacQueen, J., 1967: Some methods for classification and analysis of multivariate observations. *Statistics*, L. M. Le Cam, and J. Neyman, Eds.,Vol. 1, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 281–297, https://pdfs.semanticscholar.org/ac8a/b51a86f1a9ae74dd0e4576d1a019f5e654ed.pdf?_ga=2.178085605.1710985754.1535567185-1869641262.1535567185.

Marks, D. G., 1992: The beta and advection model for hurricane track forecasting. NOAA Tech. Memo. NWS NMC, 70, 89 pp., https://repository.library.noaa.gov/view/noaa/7184.

Murakami, H., and Coauthors, 2012: Future changes in tropical cyclone activity projected by the new high-resolution MRI-AGCM. *J. Climate*, **25**, 3237–3260, https://doi.org/10.1175/JCLI-D-11-00415.1.

——, P.-C. Hsu, O. Arakawa, and T. Li, 2014: Influence of model biases on projected future changes in tropical cyclone frequency of occurrence. *J. Climate*, **27**, 2159–2181, https://doi.org/10.1175/JCLI-D-13-00436.1.

Nakamura, J., U. Lall, Y. Kushnir, and S. J. Camargo, 2009: Classifying North Atlantic tropical cyclone tracks by mass moments. *J. Climate*, **22**, 5481–5494, https://doi.org/10.1175/2009JCLI2828.1.

——, and Coauthors, 2017: Western North Pacific tropical cyclone model tracks in present and future climates. *J. Geophys. Res. Atmos.*, **122**, 9721–9744, https://doi.org/10.1002/2017JD027007.

Palmer, W. C. and R. A. Allen, 1949: Note on the accuracy of forecasts concerning the rain problem. U.S. Weather Bureau Rep., 4 pp.

Rogers, R., 2010: Convective-scale structure and evolution during a high-resolution simulation of tropical cyclone rapid intensification. *J. Atmos. Sci.*, **67**, 44, https://doi.org/10.1175/2009JAS3122.1.

Shaevitz, D. A., and Coauthors, 2014: Characteristics of tropical cyclones in high-resolution models of the present climate. *J. Adv. Model. Earth Syst.*, **6**, 1154–1172, https://doi.org/10.1002/2014MS000372.

Song, Y., L. Wang, X. Lei, and X. Wang, 2015: Tropical cyclone genesis potential index over the western North Pacific simulated by CMIP5 models. *Adv. Atmos. Sci.*, **32**, 1539–1550, https://doi.org/10.1007/s00376-015-4162-3.

Strazzo, S., J. B. Elsner, T. Larow, D. J. Halperin, and M. Zhao, 2013: Observed versus GCM-generated local tropical cyclone frequency: Comparisons using a spatial lattice. *J. Climate*, **26**, 8257–8268, https://doi.org/10.1175/JCLI-D-12-00808.1.

Sun, Y., and Coauthors, 2017: Impact of ocean warming on tropical cyclone track over the western North Pacific: A numerical investigation based on two case studies. *J. Geophys. Res. Atmos.*, **122**, 8617–8630, https://doi.org/10.1002/2017JD026959.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Tippett, M., S. J. Camargo, and A. H. Sobel, 2011: A Poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis. *J. Climate*, **24**, 2335–2357, https://doi.org/10.1175/2010JCLI3811.1.

Tonkin, H., C. Landsea, G. J. Holland, and S. Li, 1997: Tropical cyclones and climate change: A preliminary assessment. *Assessing Climate Change: Results from the Model Evaluation Consortium for Climate Assessment*, W. Howe and A. Henderson-Sellers, Eds., Gordon and Breach, 327–360.

Walsh, K. J. E., and Coauthors, 2015: Hurricanes and climate: The U.S. CLIVAR Working Group on Hurricanes. *Bull. Amer. Meteor. Soc.*, **96**, 997–1017, https://doi.org/10.1175/BAMS-D-13-00242.1.

Wang, C., and L. Wu, 2015: Influence of future tropical cyclone track changes on their basin-wide intensity over the western North Pacific: Downscaled CMIP5 projections. *Adv. Atmos. Sci.*, **32**, 613–623, https://doi.org/10.1007/s00376-014-4105-4.

Wu, L., and B. Wang, 2004: Assessing impacts of global warming on tropical cyclone tracks. *J. Climate*, **17**, 1686–1698, https://doi.org/10.1175/1520-0442(2004)017<1686:AIOGWO>2.0.CO;2.

——, ——, and S. Geng, 2005: Growing typhoon influence on East Asia. *Geophys. Res. Lett.*, **32**, 109–127, https://doi.org/10.1029/2005GL022937.

Ying, M., E. J. Cha, and H. J. Kwon, 2011: Comparison of three western North Pacific tropical cyclone best track datasets in a seasonal context. *J. Meteor. Soc. Japan*, **89**, 211–224, https://doi.org/10.2151/jmsj.2011-303.

Yu, J.-H., Y.-Q. Zheng, Q.-S. Wu, and Z.-B. Gong, 2016: K-means clustering for classification of the northwestern Pacific tropical cyclone tracks. *J. Trop. Meteor.*, **22**, 127–135, https://doi.org/10.16555/j.1006-8775.2016.02.003.

Zhao, M., I. M. Held, S. J. Lin, and G. A. Vecchi, 2009: Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50-km resolution GCM. *J. Climate*, **22**, 6653–6678, https://doi.org/10.1175/2009JCLI3049.1.

Zhou, B.-T., 2012: Model evaluation and projection on the linkage between Hadley circulation and atmospheric background related to the tropical cyclone frequency over the western North Pacific. *Atmos. Oceanic Sci. Lett.*, **5**, 473–477, https://doi.org/10.1080/16742834.2012.11447036.

——, and Y. Xu, 2017: How the ''best'' CMIP5 models project relations of Asian–Pacific oscillation to circulation backgrounds favorable for tropical cyclone genesis over the western North Pacific. *J. Meteor. Res.*, **31**, 107–116, https://doi.org/10.1007/s13351-017-6088-4.