Seasonal Forecasts of Tropical Cyclones using GFDL SPEAR and HiFLOR-S

Hiroyuki Murakami[1], Thomas L. Delworth[1], Nathaniel C. Johnson[1], Feiyu Lu[1,2], Colleen E. McHugh[1,3], and Liwei Jia[1]

[1]National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

[2]University Corporation for Atmospheric Research, Boulder, CO, USA

[3]Science Applications International Corporation, Reston, VA, USA

Revised on October 29, 2024

Submitted to *Journal of Climate* on July 5, 2024

Corresponding author address: Hiroyuki Murakami, NOAA/GFDL, 201 Forrestal Rd., Princeton, NJ 08540-6649
E-mail: hir.murakami@gmail.com
Tel: 609-452-5824

Abstract

The seasonal prediction skill of tropical cyclone (TC) activity is evaluated using the Seamless System for Prediction and Earth System Research (SPEAR), a modeling system developed at the Geophysical Fluid Dynamics Laboratory (GFDL) for experimental real-time seasonal forecasts. Compared with previous GFDL seasonal prediction models, SPEAR demonstrates improved skill in predicting TC activity for the western North Pacific, while exhibiting comparable or slightly degraded skill for the eastern North Pacific and North Atlantic. These changes in prediction skill do not always align with changes in prediction skill in large-scale variables, particularly over the North Atlantic. This study highlights that changes in the model's response of TCs to large-scale variables, as well as the changes in the amplitude of interannual variations in TC genesis frequency, are crucial for the changes in TC prediction skill. Using the predicted sea surface temperatures from SPEAR as lower boundary conditions, the High-Resolution Forecast-Oriented Low Ocean Resolution model (HiFLOR-S) was employed to predict intense TCs, demonstrating skillful predictions of major hurricanes that are comparable to the previous HiFLOR coupled model predictions.

Significance Statement

This study reveals the prediction skill in seasonal forecasting of tropical cyclones using a new experimental real-time seasonal prediction system developed at the Geophysical Fluid Dynamics Laboratory. The new system demonstrates skillful prediction of tropical cyclones in the western North Pacific, eastern North Pacific, and North Atlantic a few months before the hurricane season, with notable differences in the skill compared to the previous prediction system. The findings suggest that higher prediction skill in large-scale variables, such as vertical wind shear and sea surface temperatures, do not necessarily lead to higher prediction skill for tropical cyclones. This underscores that even when a model accurately predicts large-scale variables, its predictions of tropical cyclones could still be inaccurate. Our findings emphasize the need to refine the model's response of tropical cyclones to specific large-scale environments, rather than focusing only on improving large-scale environment predictions, to enhance the accuracy of dynamical seasonal predictions for tropical cyclones.

1. Introduction

Tropical cyclones (TCs), defined as storms with a maximum wind speed of ≥17.5 m s^{-1}, are the costliest natural disasters worldwide, making the prediction of TC activity on a seasonal time scale of vital socio-economic interest. Since Gray (1984a, b), numerous studies have attempted to develop seasonal TC predictions. Comprehensive reviews of seasonal TC predictions over the past 40 years are available in Camargo et al. (2007), Klotzbach et al. (2019), and Chu and Murakami (2022). Specifically, dynamical seasonal TC predictions began in 2001 at the European Centre for Medium-Range Weather Forecasts (ECMWF) (Vitart and Stockdale 2001). Since then, many dynamical models have demonstrated skillful predictions of TC activity a few months in advance from the storm season, specifically over the North Atlantic (NA) (e.g., LaRow et al., 2008; Zhao et al., 2010; Chen et al., 2011 and 2013; Camp et al., 2015; Befort et al., 2022).

However, most seasonal predictions have focused on forecasting TC activity based on basin-wide statistics, such as the basin-total frequency of named storms (with a maximum wind speed ≥17.5 m s^{-1}), hurricanes (with a maximum wind speed ≥34.0 m s^{-1}), major hurricanes (with a maximum speed ≥49.4 m s^{-1}), and Accumulated Cyclone Energy (ACE; Bell et al., 2000) (Klotzbach et al., 2019; Takaya et al., 2023). These basin-wide variables have also been the targets for predicting seasonal hurricane outlooks produced by the National Oceanic and Atmospheric Administration (NOAA) (Klotzbach et al., 2019). However, the World Meteorological Organization (WMO) has suggested exploring beyond the predictions of basin-wide statistics, such as sub-basin scale information like landfalling TCs, which are more relevant to society and stakeholders (Klotzbach et al., 2019; Takaya et al., 2023).

The NOAA Geophysical Fluid Dynamics Laboratory (GFDL) is one of the U.S. research institutions contributing to the North American Multi-Model Ensemble Project (NMME; Kirtman et al., 2014). Among the NMME models, GFDL models incorporate the highest horizontal resolution (i.e., 50-km mesh), enabling direct prediction of TCs. These real-time and retrospective TC predictions from GFDL have been shared with the experts at the NOAA Climate Prediction Center (CPC) and the National Hurricane Center (NHC), supporting their seasonal hurricane outlook, issued each May and updated in August. Previously, GFDL had used the Forecast-oriented Low Ocean Resolution of GFDL Coupled Model version 2.5 (FLOR; Vecchi et al., 2014) and the high-resolution version of FLOR

(HiFLOR) (Murakami et al., 2015, 2016a) for real-time TC predictions. Both FLOR and HiFLOR showed reasonable skill not only for basin-wide named storms, major hurricanes, and ACE, but also for regional TC frequency of occurrence (Vecchi et al., 2014; Murakami et al., 2016a,b; Zhang et al., 2017; G. Zhang et al., 2019), TC rainfall (W. Zhang et al., 2019), and extratropical transition of TCs (Liu et al., 2018).

In January 2021, GFDL upgraded its real-time experimental seasonal to decadal prediction system to Seamless System for Prediction and Earth System Research (SPEAR; Delworth et al., 2020; Lu et al., 2020), replacing FLOR. The predictions from the new SPEAR system demonstrated good skill in predicting climate variability, such as ENSO (Lu et al., 2020), and hydroclimate extremes, including heatwaves (Jia et al., 2022), atmospheric rivers (Tseng et al., 2021), Arctic and Antarctic sea ice (Bushuk et al., 2021, 2022), and wintertime temperature swings (Yang et al., 2022). While SPEAR was not specifically optimized for improving TC predictions relative to FLOR, the prediction skill of seasonal TC activity by SPEAR has not been investigated or reported previously.

In this study, we assess the prediction skill of TCs using SPEAR and compare these evaluations with those from previous GFDL prediction models, FLOR and HiFLOR. The predictions target seasonal mean TC activities, including basin-total TC genesis frequency for different storm intensity categories, Accumulated Cyclone Energy (ACE), and Power dissipation Index (PDI), as well as regional TC occurrence and landfalling frequencies in the western North Pacific (WNP), eastern North Pacific (ENP), and North Atlantic (NA) basins (see Fig. 3 in Murakami et al., 2015 for regional boundaries). Additionally, we demonstrate prediction skill through HiFLOR downscaling from SPEAR's predicted sea surface temperatures (SSTs). Furthermore, we examine the causes of differences in prediction skill for TC variables between the new and previous prediction models, particularly in relation to changes in the skill of large-scale variables. A unique case from the 2023 predictions is also presented, in which the two models in the new prediction system provided differing predictions for the hurricane season, with possible reasons for these discrepancies explored. Section 2 describes the methods, including models, seasonal predictions, TC detection method, observed datasets, and forecast skill metrics. Section 3 presents the results, with a summary provided in Section 4.

2. Methods

a. Dynamical Models

The dynamical models used in this study include FLOR (Vecchi et al., 2014), HiFLOR (Murakami et al., 2015, 2016a), and SPEAR (Delworth et al., 2020), all developed at GFDL. FLOR comprises 50-km mesh atmosphere and land components coupled with 100-km mesh sea ice and ocean components. The atmosphere and land components are adapted from the Coupled Model version 2.5 (CM2.5; Delworth et al., 2012), while the sea ice and ocean components are derived from the Coupled Model version 2.1 (CM2.1; Delworth et al., 2006). HiFLOR is nearly identical to FLOR, except for the horizontal resolution of the atmosphere and land components, which employs a 25-km mesh, along with some minor adjustments in parameters in the dynamical core and physical parameterizations (Murakami et al., 2015; Vecchi et al., 2019).

The GFDL SPEAR incorporates a coupled atmospheric-oceanic model consisting of the new AM4-LM4 atmosphere and land-surface model (Zhao et al., 2018), coupled with the MOM6 ocean model and SIS2 sea-ice model (Adcroft et al., 2019). Similar to FLOR, SPEAR employs a 50-km mesh for the atmosphere and land components and a 100-km mesh for sea ice and oceanic components.

b. Retrospective Seasonal Predictions

For each year and month from 1992 to 2020, 12-month retrospective seasonal predictions were generated by initializing each model to observationally constrained conditions for the ocean and sea ice components (Vecchi et al., 2014; Murakami et al., 2015, 2016; Lu et al., 2020). A summary of the seasonal predictions is provided in Table 1.

Table 1 Prediction configurations. For each previous prediction system (i.e., FLOR and HiFLOR) and new prediction system (i.e., SPEAR and HiFLOR-S), the following are listed: horizontal resolution of atmosphere and land components; horizontal resolution of ocean and sea-ice components; number of ensemble members for the predictions; methods to generate ocean initial conditions; methods to generate atmosphere and land initial conditions; period for retrospective predictions; initial months; methods for ocean bias adjustments during forecasts; and reference for the model and predictions.

	Previous Prediction System		New Prediction System	
	FLOR	HiFLOR	SPEAR	HiFLOR-S
Atmosphere and land resolution	50 km	25 km	50 km	25 km
Ocean and sea-ice resolution	100 km	100 km	100 km	100 km
Ensemble member	12	12	15	15
Ocean IC	ECDA (Zhang and Rosati, 2010)	ECDA (Zhang and Rosati, 2010)	SPEAR_ECDA (Lu et al., 2020)	N/A
Sea ice IC	ECDA (Zhang and Rosati, 2010)	ECDA (Zhang and Rosati, 2010)	SPEAR nudged (Lu et al. 2020)	N/A
Atmosphere and land IC	SST-forced AMIP simulaitons	SST-forced AMIP simulaitons	SPEAR nudged (Lu et al. 2020)	SST-forced AMIP simulaitons
Initial years	1992-2020	1992-2020	1992-2020	1992-2020
Initial months	Each month of January-December	January, April, June, July	Each month of January-December	April, May, July
Ocean adjustment during forecasts	Flux adjustiment (Vecchi et al. 2014)	N/A	OTA (Lu et al. 2020)	Nudged to the SPEAR predicted SST
Reference	Vecchi et al.(2014)	Murakami et al. (2015,2016)	Lu et al. (2020)	N/A

For the FLOR and HiFLOR predictions, the 12-member initial conditions for the ocean and sea ice were generated using the GFDL's ensemble coupled data assimilation system (ECDA; Zhang and Rosati 2010; Chang et al., 2013). The atmosphere and land components were initialized from a suite of SST-forced atmosphere-land-only simulations (Vecchi et al., 2014). HiFLOR provides forecasts initialized on the first day of the month only from July, June, April, and January, whereas FLOR offers forecasts starting every month. To mitigate climatological biases in SSTs and the associated model drift with increasing lead time, seasonal predictions by FLOR were conducted using "flux adjustment", which adjusts the model's air-sea fluxes of momentum, enthalpy, and freshwater to align the long-term climatology of SST and surface wind stress with the observations (Vecchi et al., 2014). HiFLOR predictions do not apply flux adjustment.

For the SPEAR predictions, the 15-member initial ocean conditions were generated with SPEAR_ECDA (Lu et al., 2020). The atmosphere and land components, as well as the sea ice component for SPEAR, were initialized from restoring simulations, where the SSTs were nudged to the values of Optimum Interpolation Sea Surface Temperature (OISST, Reynolds et al., 2002). The SPEAR predictions incorporate ocean tendency adjustment (OTA; Lu et al., 2020) to reduce three-dimensional oceanic biases, improving SST climatology and variability.

To complement SPEAR for intense TC predictions, we conducted HiFLOR predictions forced with the predicted SSTs by SPEAR (HiFLOR-S). These HiFLOR-S predictions were not initialized with data assimilation experiments, although simulated SSTs were nudged to SPEAR-predicted SSTs at a 5-day time scale. The initial conditions of ocean and sea ice components for HiFLOR-S were derived from an arbitrary year in a HiFLOR long-term control climate simulation. For example, ensemble member 1 is initiated from the restart file of year 101, while ensemble member 2 is initiated from that of year 111. However, our preliminary assessment revealed that the choice of years has little impact on the results of TC predictions, as prescribing SSTs from the SPEAR-predicted values is more critical for TC predictions than the differences in ocean initial conditions. Meanwhile, the atmosphere and land initial conditions were derived from the SST nudged experiments in which the SSTs were nudged to the values of OISST.

We primarily compare the predictions of TC activity in the WNP, ENP, and NA in the boreal summer and early fall season (i.e., July–November). Forecasts initialized in July (January) are defined as lead-month 0 or L0 (6 or L6) forecasts. Since the retrospective predictions by FLOR and HiFLOR are only available for the period 1992–2020, we compare these predictions with the predictions by SPEAR and HiFLOR-S over the same period. Given the limited computational resources, retrospective predictions are only available for L0, L2, and L3 for HiFLOR-S, and L0, L1, L2, L5, and L6 for HiFLOR, although retrospective predictions are available for every initial month between L0 and L6 for SPEAR and FLOR. Additional prediction differences for the summer of 2023 will be shown for SPEAR and HiFLOR-S in Section 3. c.

Vecchi et al. (2014) revealed that the prediction skill in the basin-wide frequency of hurricanes in the NA by FLOR showed comparable or higher prediction skill compared with other state-of-the-art prediction systems (e.g., Vitart et al., 2007; Klotzbach and Gray 2009; Zhao et al., 2010; LaRow et al., 2010; Wang et al., 2009; Chen and Lin 2013; see Fig.9 in Vecchi et al., 2014). Therefore, the prediction skill of FLOR can serve as a reference for typical skill obtained by dynamical TC seasonal predictions. As also noted by Befort et al. (2022), prediction skill for TC activity is relatively higher in the NA than in other ocean basins like the WNP and ENP for most of the dynamical model predictions.

c. TC Detection Method

The detection of model-generated TCs followed the method outlined by Harris et al. (2016) and Murakami et al. (2015). Briefly, the tracking scheme employs the flood fill algorithm to identify closed contours of a specified negative sea level pressure (SLP) anomaly with a warm core (temperature anomaly higher than 1K for FLOR and SPEAR and 2K for HiFLOR and HiFLOR-S). Additionally, the detection scheme requires that a TC must persist for at least 36 hours while maintaining its warm core, along with meeting a specified surface wind speed criterion (15.75 m s^{-1} for FLOR and SPEAR and 17.5 m s^{-1} for HiFLOR and HiFLOR-S). These thresholds were determined by the previous studies of FLOR and HiFLOR (Murakami et al. 2015). Because the horizontal resolution of FLOR and SPEAR is 50-km mesh and unable to represent intense TCs, the warm core and wind speed threshold were relaxed from those for HiFLOR and HiFLOR-S as in the previous studies (Murakami et al. 2015).

d. Observational Datasets and Large-scale Variables

The observed TC "best-track" data for the period 1992–2023 were obtained from the International Best Track Archive for Climate Stewardship (IBTrACS v04r00) (Knapp et al., 2010). We use a compilation from the NHC and Joint Typhoon Warning Center (JTWC), identified by the flag 'usa' in the IBTrACS dataset. We considered TCs with tropical storm intensities or stronger, defined as TCs possessing 1-minute sustained surface winds of 17.5 m s^{-1} or greater.

We utilized the OISST (Reynolds et al., 2002) and the Japanese 55-year Reanalysis (JRA-55) (Kobayashi et al., 2015) for the period 1992–2023 as observed SST and atmospheric large-scale variables, respectively. To elucidate the factors contributing to the differences in the prediction skill in TCs among the GFDL models, we compared the prediction skill in key large-scale variables. These large-scale variables include vertical wind shear between 850hPa and 200 hPa (V_s), relative humidity at 600 hPa (RH_{600}), absolute vorticity at 850 hPa (ζ_{a850}), Maximum Potential Intensity (*MPI*; Bister and Emanuel, 1998), vertical motion at 500 hPa (ω_{500}), shear vorticity of zonal winds at 500 hPa (U_{y500}), and SST anomaly (*SST*), which are commonly used for tropical cyclone genesis potential indices (e.g., Emanuel and Nolan, 2004; Murakami and Wang, 2010; Wang and Murakami, 2020; Murakami and Wang, 2022). Here, anomalies are defined as the deviations from the mean climatology of 1992–2020, with climatology calculated separately for each lead month prediction. These large-scale variables were evaluated exclusively over the main development region of TCs for each WNP (10–25°N, 110–150°E), NA (10–25°N, 80–20°W), and ENP (5–25°N, 130–100°W) ocean basin.

e. Metrics for Evaluation of Forecast Skill

As in Murakami et al., (2016a), storms are classified into three categories based on their lifetime maximum intensity: Tropical Storms (TCS; ≥17.5 m s^{-1}); Hurricanes (HUR; ≥32.9 m s^{-1}); and Category 3–5 (or major) hurricanes (C345; ≥49.4 m s^{-1}). We note that while a hurricane is referred to as a "typhoon" in the WNP, we use the term "hurricanes" for WNP typhoons in this study. Additionally, we considered ACE, defined as the sum of the square of the maximum surface wind velocity throughout the lifetime of a TC, normalized by a factor of 10^5 (10^5 m^2 s^{-2}; Bell et al., 2000). Along with ACE, we evaluated PDI, which is similarly defined, but as the sum of the cube of the maximum surface wind velocity throughout the lifetime of a TC, normalized by a factor of 10^7 (10^7 m^3 s^{-3}; Emanuel, 2005, 2007). We examined the prediction skill in interannual variation of the basin-wide frequencies for TCS, HUR, C345, ACE, PDI, and the landfalling TCs over the Continental U.S. (US), Caribbean Islands (CAR), and Hawaiian Islands (HI).

As outlined in Murakami et al. (2016a), we employed two scores to assess prediction skill for the TC activity relative to observed values: Spearman's rank correlation coefficient (RCOR) and mean square skill score (MSSS) (Kim et al., 2012; Li et al., 2013). Following Vecchi et al. (2014), we chose Spearman's rank correlation instead of Pearson's correlation as our correlation metric because we do not expect the ensemble-mean forecasts of TC counts and the observed annual TC counts (integer values) to follow a Gaussian distribution. Additionally, Pearson's correlation is sensitive to outliers, which are common in TC data, as extreme values can disproportionately influence the coefficient and distort the perceived relationship between predictions and observations. In contrast, RCOR measures the forecast system's ability to correctly identify the relative ranking of years from least to most active in the observed record.

MSSS is defined by the following equation:

$$MSSS \equiv 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(f_i^{obs} - f_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(f_i^{obs} - f^{obs})^2}, \qquad (1)$$

where n is the total number of years, f_i^{obs} and f_i are the values from observations and predicted values for the i^{th} year, respectively, and f^{obs} is the observational mean. The MSSS compares the model's skill against climatological forecasts, with values greater than zero indicating better predictive skill than a climatological forecast (Kim et al., 2012; Li et al., 2013).

Throughout the analysis, unless presenting raw predicted results, both TC and large-scale variables are normalized by subtracting the climatological mean and dividing by the standard deviation, with these mean and standard deviation values specific to each model's lead month. After normalization, RCOR and MSSS are computed. We assess the statistical significance of RCOR using a two-tailed test, with the test statistic asymptotic t-distributed with n–2 degrees of freedom, where n is the sample size, adjusted for observed autocorrelation (Siegel and Castellan, 1988).

We also used the bootstrap method proposed by Murakami et al. (2013) to evaluate the statistical significance of the mean difference between model experiments. Two tested populations were resampled in pairs 2,000 times, and the mean difference for each pair was calculated, creating a new distribution with 2,000 samples. A 95% confidence interval was derived from this distribution and compared with the original mean difference.

3. Results

a. Retrospective Forecast of Basin-Wide TC Activity

We first compare the retrospective forecast skill in basin-wide seasonal TC activity over the NA between FLOR and SPEAR and between HiFLOR and HiFLOR-S. Figure 1 shows the time series of observed and predicted TCS, HUR, C345, and ACE from the July initial predictions (i.e., L=0). Generally, the new prediction system (i.e., SPEAR and HiFLOR-S) exhibited similar though usually slightly lower skill than the previous prediction system (i.e., FLOR and HiFLOR), although both systems show statistically significant correlations, covering the observations within their 90% range estimated from the ensemble members. There are some clear differences in active seasons between SPEAR and FLOR. For example, SPEAR predicted a higher number of HUR for 1995 than FLOR (Fig. 1b). However, this feature is inconsistent; for example, FLOR predicts a higher number of HUR for 2005 than SPEAR.

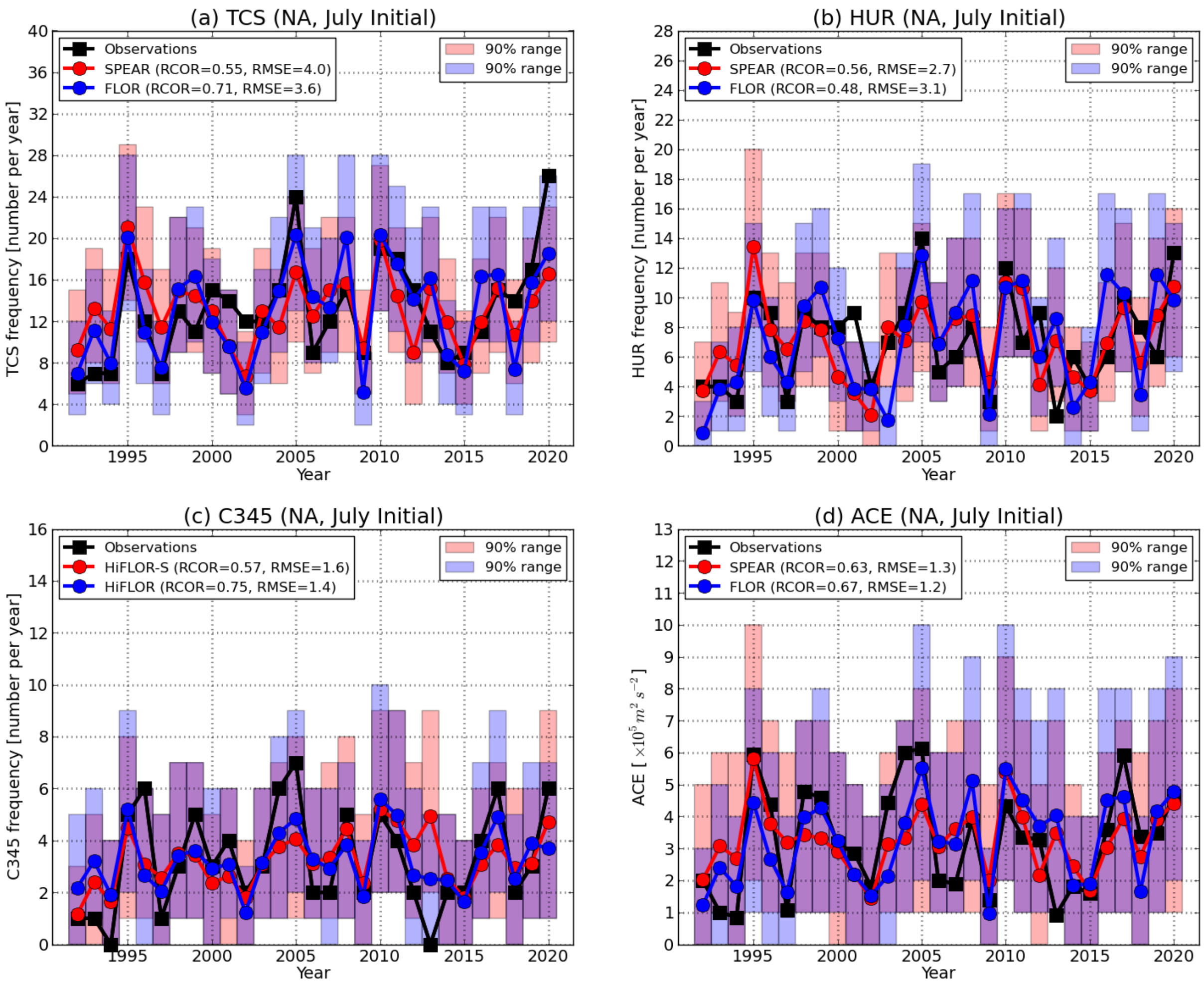

FIG. 1 Retrospective predictions of (a) basin-wide frequency of TCS, (b) HUR, (c) C345, and (d) ACE in the NA during the peak TC season of July–November for the period 1992–2020 initialized in July. The black lines represent the observed values, the red lines represent the mean values of the new prediction system (SPEAR or HiFLOR-S), and the blue lines represent the mean values of the previous prediction system (FLOR or HiFLOR). Shading indicates the 90% confidence intervals computed by convolving inter-ensemble spread based on the Poisson distribution. The values of “RCOR” and “RMSE” in each panel indicate the rank correlation coefficient and root-mean-square error between the predictions and observations, respectively. Units: number per year for TCS, HUR, and C345 and 10^5 m^2 s^{-2} per year for ACE.

Figure 2 compares the RCOR skill of TC activities for each initial month. While Fig. 1 indicates that the new prediction system worsens the prediction skill in the NA from the July initial conditions, this is not always the case for different initialization months. Overall, both SPEAR and FLOR demonstrate statistically significant skill in predicting TCS and HUR in the NA from lead-month 0 to 2 predictions (Figs. 2a,d). SPEAR also shows skillful predictions of

TCS and HUR at lead-month 4, although the skill at lead-month 3 is not statistically significant. Additionally, Fig. 2 displays prediction skill for the WNP and ENP, revealing that SPEAR generally outperforms (underperforms) FLOR for TCS and HUR predictions in the WNP (ENP). For the comparison of C345 predictions between HiFLOR and HiFLOR-S, both show comparable prediction skill across the three ocean basins (Figs. 2g–i). Generally, ACE predictions exhibit skill even from February's initial predictions (Figs 2j,k,l), indicating greater skill in ACE predictions compared with TC frequency predictions.

Previous studies have reported that ensemble means of multi-models often outperform individual models in TC seasonal predictions (e.g., Vitart 2006; Vitart et al. 2007). In this study, we also assessed the prediction skill of the ensemble means of SPEAR and FLOR (shown by the purple lines in Fig. 2). Our findings indicate that the prediction skill of the multi-model ensemble mean is not simply an average of the skill of the two individual models. In some instances, the multi-model ensemble mean outperforms both models, particularly for ACE predictions. This result highlights the potential for further improvements in prediction skill by utilizing a multi-model ensemble approach.

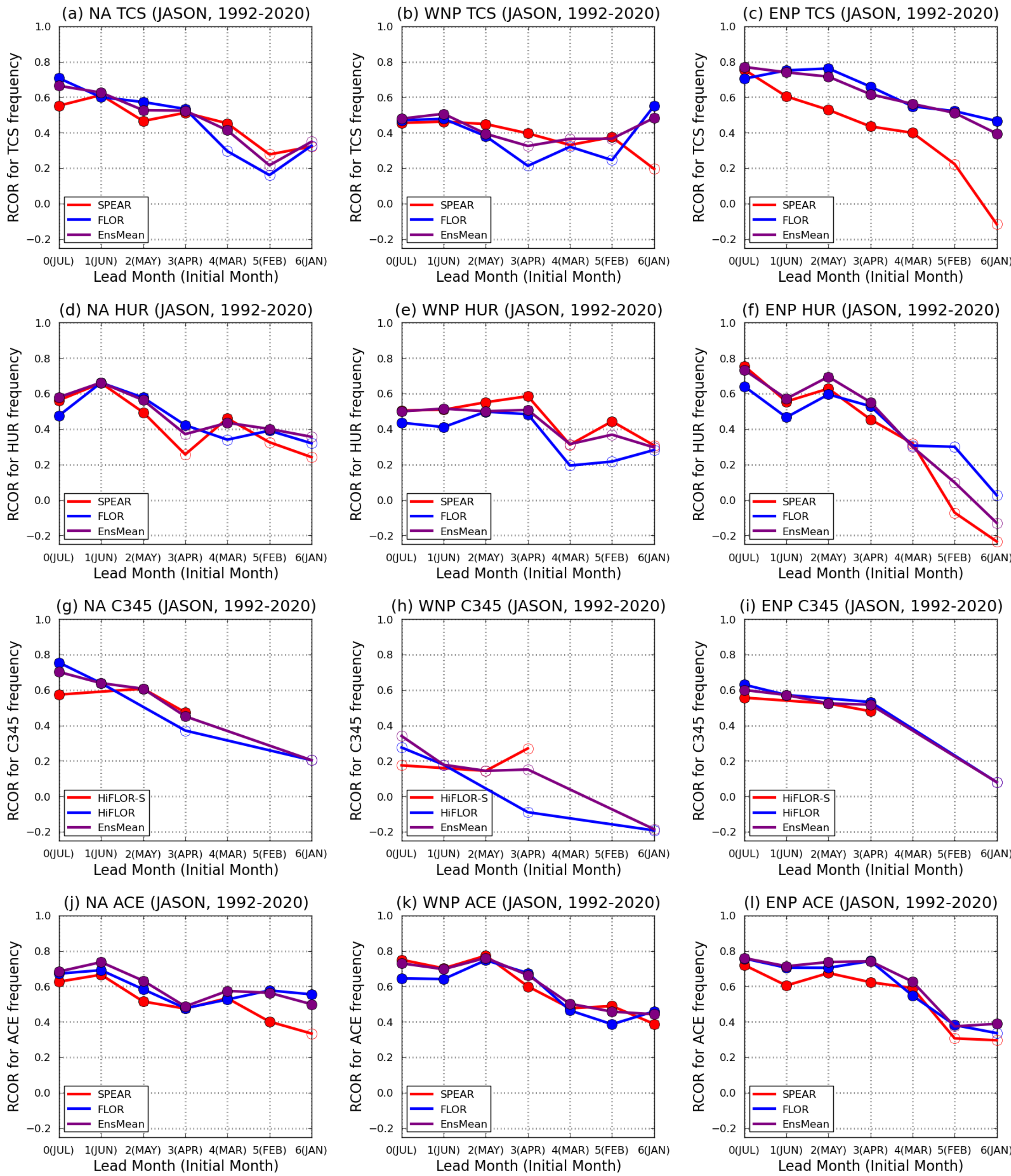

FIG. 2 RCORs between observed and predicted TC activity for each initial month from January (L6) to July (L0). (a)–(c) TCS, (d)–(f) HUR, (g)–(i) C345, and (j)–(l) ACE over (left) the NA, (middle) WNP, and (right) ENP. The red lines depict predictions by the new prediction system (SPEAR or HiFLOR-S), whereas the blue lines depict predictions by the previous prediction system (FLOR or HiFLOR). The purple lines are multi-model ensemble means of the new and previous prediction systems. Filled marks indicate statistically

significant RCORs at a 95% confidence level, whereas open marks denote non-significant RCORs.

To provide a more comprehensive quantification of how the TC metrics of the new prediction system compare with those of the previous prediction system, we display scatter plots of RCOR and MSSS in Fig. 3 for interannual variation of seasonal mean value between observations and predictions. Here, we compare basin-wide frequencies of TCS, HUR, landfalling frequencies of CAR and HI, and basin-total values of ACE and PDI. A maker above the diagonal line indicates that SPEAR outperforms FLOR for the TC metric at the specified lead month.

As expected, the shortest lead-month forecasts (e.g., L0 and L1) generally yield higher RCOR and MSSS than the longer lead months (e.g., L5 and L6) for most of the TC variables. It is also worth noting that models generally predict ACE better than TCS (Fig. 3), a finding consistent with previous studies (e.g., Murakami et al., 2016a). Overall, SPEAR outperforms FLOR for the TC predictions over the WNP (75–79%), whereas SPEAR underperforms FLOR over the NA (33–38%) and ENP (23–26%) where the parentheses indicate the fraction of the number of variables that SPEAR outperforms FLOR relative to the total number of the variables.

Similar trends are obtained for the comparisons between HiFLOR-S and HiFLOR (Supplementary Fig. 1). Generally, HiFLOR-S outperforms HiFLOR for the NA (60–62%), WNP (64–71%), but underperforms HiFLOR or comparable for the ENP (49%), where the parentheses indicate the fraction of the number of variables that HiFLOR-S outperforms HiFLOR.

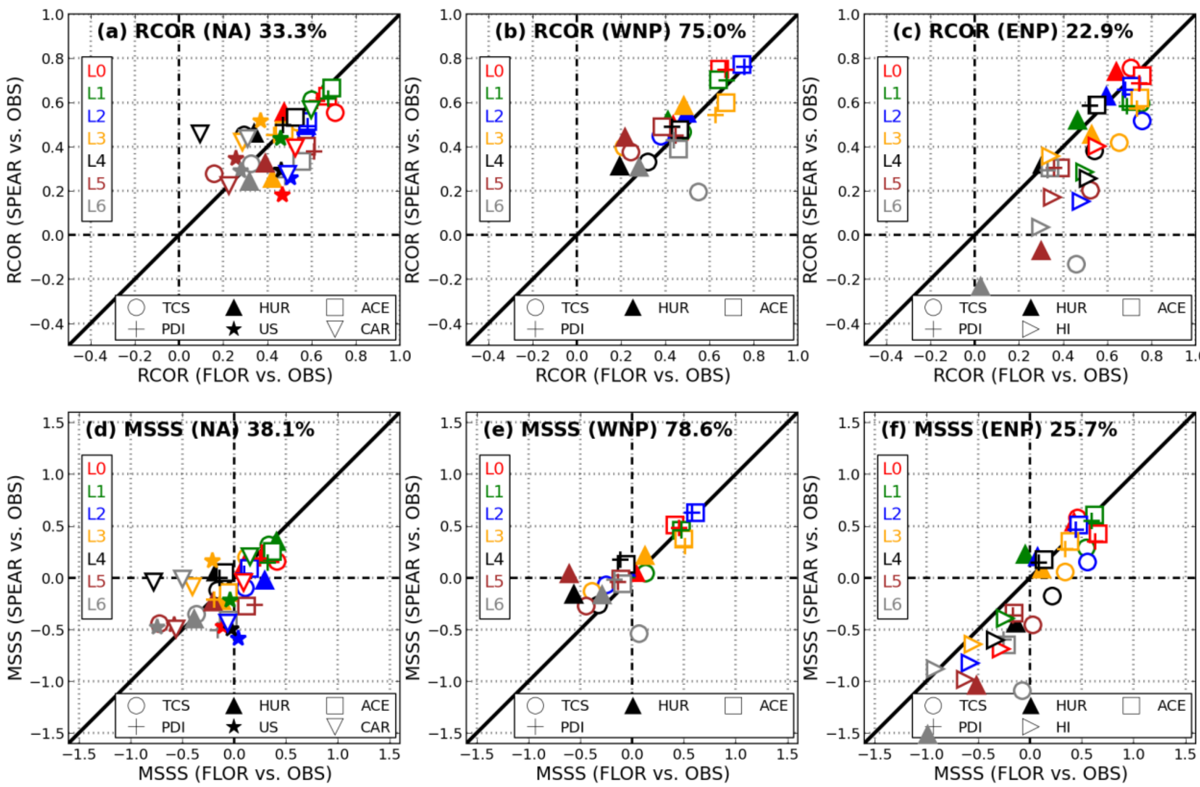

FIG. 3 Scatterplots of RCOR between SPEAR prediction and observations (*y*-axis) and FLOR prediction and observations (*x*-axis) for the (a) NA, (b) WNP, and (c) ENP. (d–f) As in (a–c), but for MSSS. A marker positioned above the diagonal line indicates that SPEAR exhibits higher skill than FLOR. The variables evaluated include basin-wide frequency of TCS, HUR, basin-wide values of ACE, PDI, and the landfalling TC frequency for the Continental United States (US), Caribbean Islands (CAR), and Hawaiian Islands (HI). Different colors represent different lead months (L0–L6). Percentages on the plots denote the fraction of variables in which SPEAR outperforms FLOR relative to the total number of variables evaluated.

b. Retrospective Predictions of Landfalling and Regional TC Activity

Beyond the prediction skill of basin-wide TC variables, we evaluate prediction skill in regional TC activity in terms of landfall TCs (i.e., US, CAR, and HI) and the frequency of TC occurrence.

Supplementary Figs. 2 and 3 show results similar to Figs. 2 and 3, focusing exclusively on landfalling predictions (i.e., US, CAR, and HI). Regarding RCOR, SPEAR exhibits lower prediction skill for HI compared to FLOR across most lead-month predictions. For US and CAR, results are mixed: SPEAR outperforms FLOR in a few lead-month predictions (e.g., L3 or L4). In terms of MSSS, no clear differences are observed between SPEAR and FLOR.

Figure 4 displays the prediction skill as measured by RCOR between L0 predictions by the models and observations for each grid cell. Both SPEAR and FLOR demonstrate statistically significant skill in the central Pacific for TCS, particularly around Hawaii, indicating their ability to predict the frequency of landfalling TCs over the Hawaiian Islands. SPEAR also exhibits improved prediction skill for TCS and HUR near Japan relative to FLOR (Figs. 4a,b,d,e). In contrast, SPEAR shows degraded prediction skill for landfalling storms over the NA relative to FLOR. HiFLOR-S shows comparable skill to HiFLOR in terms of C345 in the Pacific Ocean, but HiFLOR-S demonstrates degraded prediction skill over the NA (Figs. 4c, f).

We counted the number of grids where the model shows statistically significant positive RCOR with observations (i.e., red and yellow shadings in Fig. 4). This number was then divided by the total number of valid grid cells where the observed frequency of occurrence is nonzero for at least 25% of years (i.e., seven years; all grids within the gray shading in Fig. 4). This fractional number is compared between the models on a global scale for each TC category and lead month (Fig. 5). Figure 5 indicates that SPEAR generally demonstrates a smaller area of skillful predictions for TCS and HUR relative to FLOR, although differences between HiFLOR-S and HiFLOR for C345 are marginal. Overall, we did not find clear improvements in prediction skill for TC activity at the regional scale with the new prediction system compared to the previous prediction system.

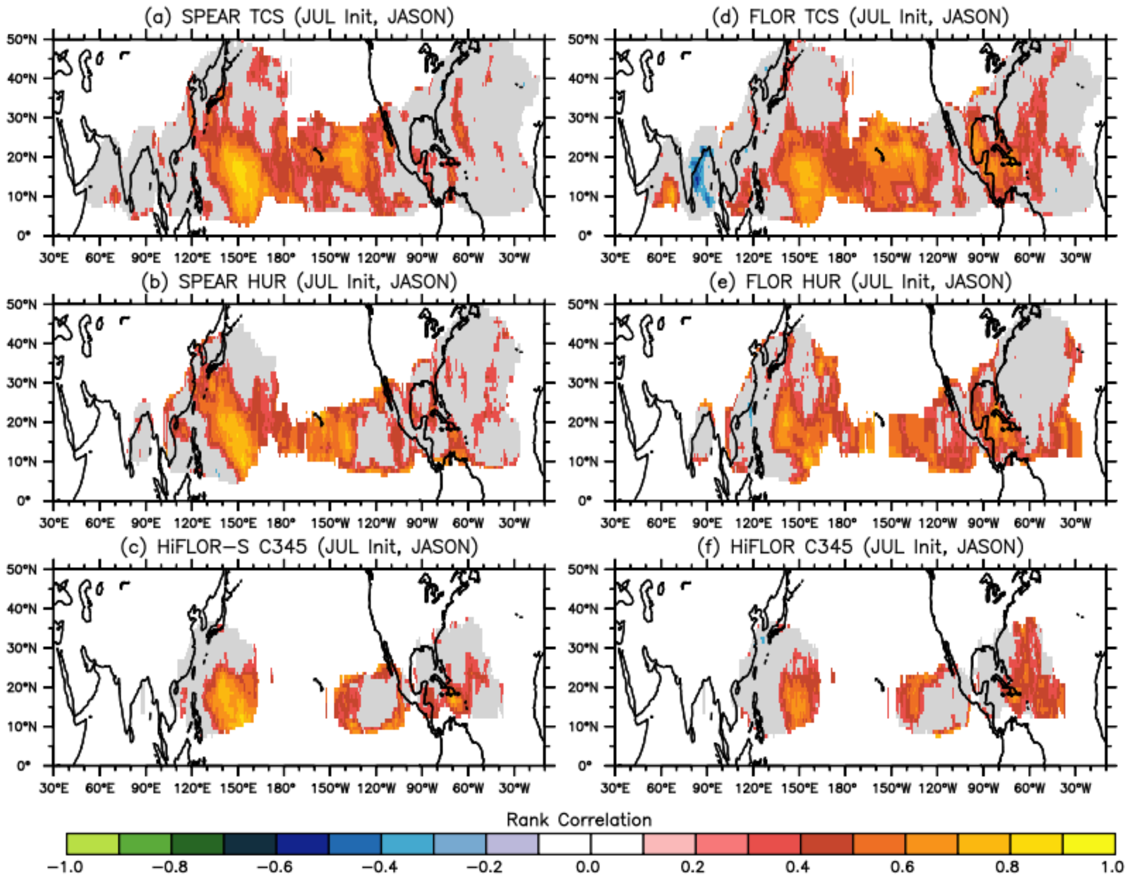

FIG. 4 Skill of frequency of occurrence of TCs during July–November 1992–2020 for the retrospective forecasts initialized in July. Shading indicates the retrospective RCOR of predicted versus observed TC frequency of occurrence (1°×1° grid box), masked at a two-sided p=0.1 level. Results are shown for (a) TCS for SPEAR, (b) HUR for SPEAR, and (c) C345 for HiFLOR-S. (d–f) As in (a–c), bur for FLOR and HiFLOR. Gray shading in all panels indicates that observed TC density is nonzero for at least 25% of years (i.e., seven years).

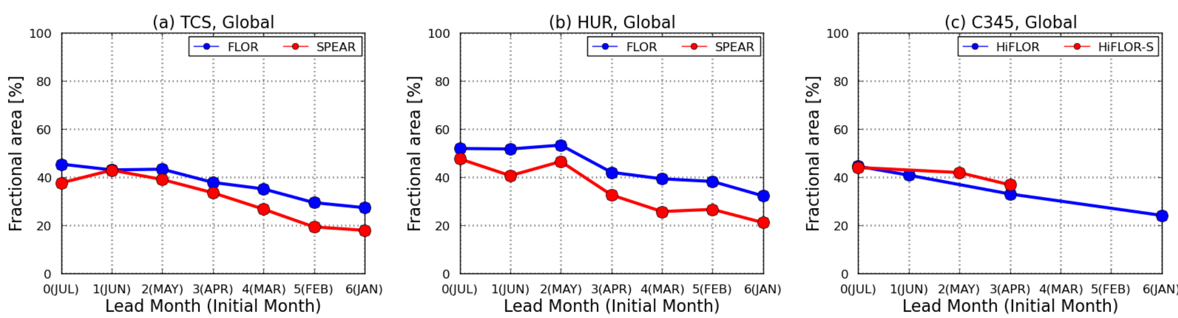

FIG. 5 Fractional number of grids with statistically significant positive RCOR between predictions and observations relative to the total number of valid grids on a global scale. Valid grids are defined as grids where the observed TC density is nonzero for at least 25% of the

years (i.e., seven years; gray areas in Fig. 4). Shown for (a) TCS for SPEAR and FLOR, (b) HUR for SPEAR and FLOR, and (c) C345 for HiFLOR-S and HiFLOR.

c. Retrospective Predictions of Large-scale Variables

Previous studies have suggested that improving the simulation of large-scale variables could result in improved simulations of TC activity (Vecchi et al., 2014; Murakami et al., 2015; Krishnamurthy et al., 2016). It is expected that improving prediction skill in large-scale variables should be linked to improving prediction skill in TC variables. However, this is not always the case. For example, Murakami et al. (2016a) revealed that the changes in prediction skill in large-scale variables are not always relevant to the changes in prediction skill in TC activity in the NA. To examine whether the differences in prediction skill in TC variables between the new and previous prediction systems, as shown in Section 3a,b are linked to the changes in prediction skill in large-scale variables, we compare the prediction skill in the TC-relevant large-scale variables.

Figure 6 compares the RCOR and MSSS between the observed and predicted large-scale variables in the key main development region for each basin by FLOR (x-axis) and between observed and predicted by SPEAR (y-axis).

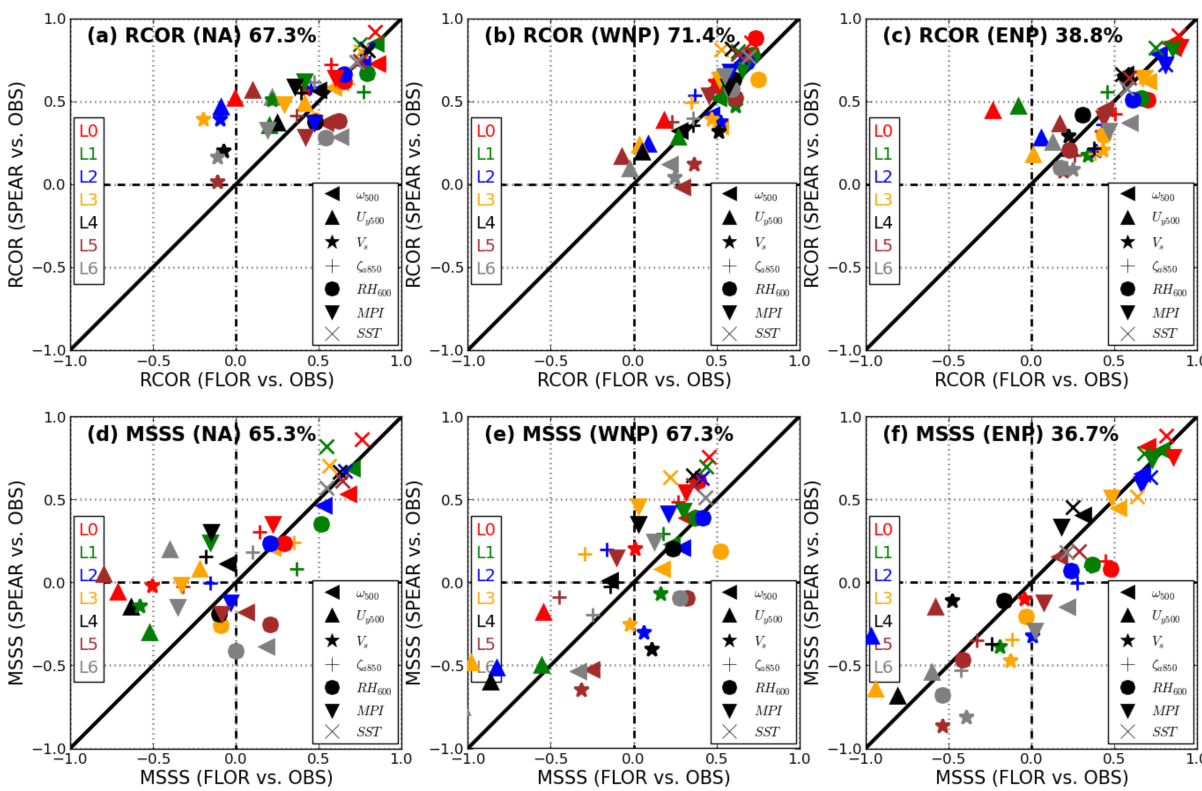

FIG. 6 As in Fig. 3, but for large-scale variables over the main development regions. Variables evaluated (symbols in the bottom right) are ω_{500}, U_{y500}, V_s, ζ_{a850}, RH_{600}, *MPI*, and *SST*.

For the NA, more than half of the variables are located above the diagonal lines, indicating improved skill in the large-scale variables in SPEAR over FLOR (Figs. 6a and d), although SPEAR showed lower skill in TC metrics than FLOR (Figs. 3a and d). These results are consistent with those of Murakami et al. (2016a), who reported that the improvements in predicting TC activity over the NA are not directly related to the improvements in predicting large-scale variables. In contrast, the WNP and ENP are relatively consistent between large-scale variables and TC activity compared to the NA (Fig. 3 and Fig. 6). For the comparisons between HiFLOR-S and HiFLOR, differences in prediction skill for large-scale variables correspond well with differences in TC variables for RCOR (Supplementary Figs. 1 and 4).

Here, we aim to identify the reasons for the discrepancies in prediction skill between TC-related variables and large-scale variables when comparing SPEAR and FLOR in the NA. Differences in TC prediction skill between these models may stem from differences in the simulations of TC climatology and/or differences in how TC climatology responds to large-scale conditions. To start, we compared the spatial distributions of the climatological mean TC genesis frequency between observations and models, SPEAR and FLOR, in the NA (shadings in Fig. 7 and Table 2).

This comparison reveals that differences in the predicted climatological mean TC genesis frequency between the models do not fully explain why FLOR exhibits better NA TC prediction skill than SPEAR. For example, observations show frequent TC genesis in both the eastern tropical Atlantic (Domain A) and the western tropical Atlantic (Domain B), with slightly higher TC genesis frequency in Domain B than in Domain A (Fig 7a and Table 2). However, both SPEAR and FLOR display notable biases in the mean locations of TC genesis (Figs. 7b, c), underestimating TC genesis frequency in Domain A and showing increased frequency in the central tropical Atlantic compared to observations.

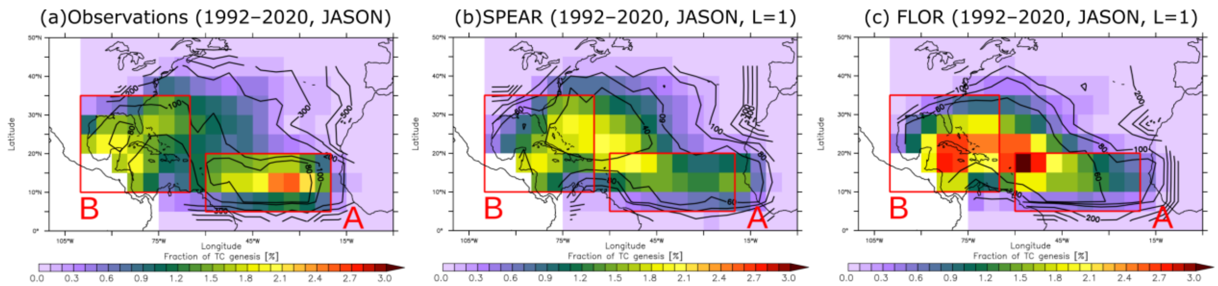

FIG. 7 Climatological mean TC genesis frequency and the standard deviation of interannual variability during July–November for the period 1992–2020. (a) Observations, (b) lead-month 1 predictions by SPEAR, and (c) lead-month 1 predictions by FLOR. Shadings represent the fraction of the climatological mean TC genesis frequency at each grid cell relative to the ocean basin total [Units: %]. Contours indicate the standard deviation of interannual variability, normalized by the climatological mean TC genesis frequency at each grid cell [Units: %]. Red rectangles highlight the main development regions, A and B.

Table 2 Climatological mean TC genesis frequency and the amplitude of interannual variation of TC genesis frequency for Domains A and B. Displayed are the fraction of climatological mean TC genesis frequency (total TC genesis frequency within a domain divided by the basin-total TC genesis frequency [%]) and the fraction of standard deviation relative to the climatological mean TC genesis frequency (standard deviation of interannual variation of total TC genesis frequency within a domain divided by the climatological mean TC genesis frequency for the same domain [%]).

	Fraction of climatological mean TC genesis frequency over a domain relative to the basin-total TC genesis frequency [%]		Fraction of standard deviation of interannual variation of TC genesis frequency relative to the climatological mean TC genesis frequency [%]	
	Domain A	Domain B	Domain A	Domain B
Observations	34.2%	38.6%	81.3%	104.7%
SPEAR	26.4%	35.4%	58.5%	50.6%
FLOR	28.4%	42.8%	59.3%	60.4%

On the other hand, substantial differences exist in the amplitude of interannual variation in TC genesis frequency between the models, which may further contribute to differences in TC prediction skill. For instance, observations show marked interannual variation in both Domains A and B, with the standard deviation exceeding 80% of the climatological mean TC genesis frequency (contours in Fig. 7 and Table 2). Although both

FLOR and SPEAR underestimate the amplitude of interannual variations in both domains, FLOR's amplitude is closer to observed values than SPEAR's, particularly in Domain B.

Furthermore, FLOR simulates a more accurate sensitivity of TC genesis frequency to large-scale variables in both Domains A and B than SPEAR (Table 3). For example, observations indicate that TC genesis frequency in Domain A is more highly correlated with thermodynamical variables (e.g., RH_{600} and SST) than with dynamical variables (e.g., V_s and ζ_{a850}). Conversely, in Domain B, it is more highly correlated with dynamical variables than thermodynamical ones. Although the RCORS produced by both models differ notably from observations, FLOR captures these observed tendencies better than SPEAR.

Table 3 RCORs of interannual variations between the TC genesis frequency and large-scale variables for each domain (1992–2020). The numbers in bold and underscore highlight the two highest correlations among the variables for each observation and model.

	V_s	ζ_{a850}	RH_{600}	MPI	SST
Domain A					
Observations	–0.24	+0.39	**<u>+0.56</u>**	+0.35	**<u>+0.42</u>**
SPEAR	**<u>–0.60</u>**	+0.52	+0.57	**<u>+0.82</u>**	+0.35
FLOR	–0.25	+0.33	**<u>+0.81</u>**	**<u>+0.83</u>**	+0.79
Domain B					
Observations	**<u>–0.43</u>**	**<u>+0.54</u>**	–0.11	–0.22	+0.00
SPEAR	–0.54	**<u>+0.89</u>**	**<u>+0.78</u>**	–0.31	+0.09
FLOR	**<u>–0.64</u>**	**<u>+0.90</u>**	–0.43	–0.32	+0.12

Previous studies suggest that ENSO, Madden Julian Oscillation (MJO), and tropical upper-tropospheric troughs (TUTT) associated with extratropical Rossby wave breaking influence wind shear and low-level vorticity in Domain A, while the Atlantic Meridional Mode (AMM) affects SST and relative humidity in Domain B (e.g., Maloney and Hartmann 2000; Kossin and Vimont 2007; Wang et al. 2020). Differences in teleconnection patterns or the influence of interannual climate modes on atmospheric conditions between the models may contribute to the variations in TC seasonal prediction skill in the NA.

With its higher horizontal resolution, HiFLOR-S is expected to outperform SPEAR in predicting TC variables, especially in intense storms such as C345. However, since the HiFLOR-S predictions were forced with SSTs predicted by SPEAR, differences in TC predictions between SPEAR and HiFLOR-S likely result from differences in the response of model-simulated TCs or large-scale variables to the same SSTs. Supplementary Figs. 5 and 6 display the same plots as Figs. 3 and 6, respectively, but for the comparisons between HiFLOR-S and SPEAR. Generally, the prediction skill differences between SPEAR and HiFLOR-S for TC variables do not align with those for large-scale variables except in the WNP. For example, the prediction skill of large-scale variables is lower (higher) in HiFLOR-S than in SPEAR in the NA (ENP). However, these skill differences in large-scale variables do not correspond to those of TC variables (Supplementary Fig. 5); HiFLOR-S generally outperforms (underperforms) SPEAR for TC variables in the NA (ENP). This finding reinforces the notion that higher prediction skill in large-scale variables do not necessarily lead to higher prediction skill in TC variables.

We compared the spatial pattern of the climatological mean TC genesis frequency and interannual variations between SPEAR and HiFLOR-S for L3 predictions, where HiFLOR-S outperforms better than SPEAR in TC predictions for the NA. Supplementary Fig. 7 indicates that HiFLOR-S has a less accurate spatial pattern of climatological TC genesis frequency than SPEAR. Specifically, TC genesis frequency in HiFLOR-S is heavily concentrated around the central tropical Atlantic, with a higher genesis frequency in Domain A than in Domain B (Supplementary Table 1). This again suggests that differences in climatological TC genesis frequency alone do not fully explain the variations in TC prediction skill. Meanwhile, the amplitude of interannual variation in TC genesis frequency in Domain B is larger and more aligned with observations in HiFLOR-S compared to SPEAR (Supplementary Table 1). Additionally, the RCORs of interannual variations between TC genesis frequency and large-scale variables are more accurate in HiFLOR-S than SPEAR for both Domains A and B (Supplementary Table 2).

Overall, these results emphasize that differences in TC predictions between models likely stem from biases in the models' sensitivity of TCs to large-scale variables, as well as biases in the amplitude of interannual variation in TC genesis frequency across the main development regions. This underscores that even when a model accurately predicts large-scale variables, its TC predictions could still be inaccurate.

d. Difference in 2023 summer predictions between SPEAR and HiFLOR-S

When we conducted real-time seasonal predictions for the summer of 2023, a notable discrepancy between SPEAR and HiFLOR-S in the TC predictions became apparent. The 2023 summer season was characterized by a strong El Niño development and warmer-than-average tropical North Atlantic (Fig. 8a). It is empirically known that, during El Niño developing summers, TCs are less active than normal over the NA due to strong vertical shears (e.g., Goldenberg et al., 1996; Smith et al., 2010). In contrast, previous studies have revealed that warmer tropical Atlantic conditions could lead to active TC seasons in the NA (e.g., Vecchi et al., 2011; Villarini et al., 2010; Murakami et al., 2018). Therefore, these contradicting SST conditions could result in either an active or inactive TC season in the NA.

As revealed in Figs. 8b,c, SPEAR accurately predicted the observed SST anomaly, even from the April 2023 initial predictions. Figure 8d highlights marked differences in the TCS predictions between SPEAR and HiFLOR-S. Until the May initial predictions, SPEAR predicted, in the ensemble mean, approximately 12 TCSs, whereas HiFLOR-S predicted around 17 TCSs. The observed TCS count was 17 in 2023, indicating that the HiFLOR-S predictions were more accurate than the SPEAR predictions. SPEAR adjusted its predictions to reflect a more active TC season from the June and July initial predictions compared to the previous month's predictions (Fig. 8d).

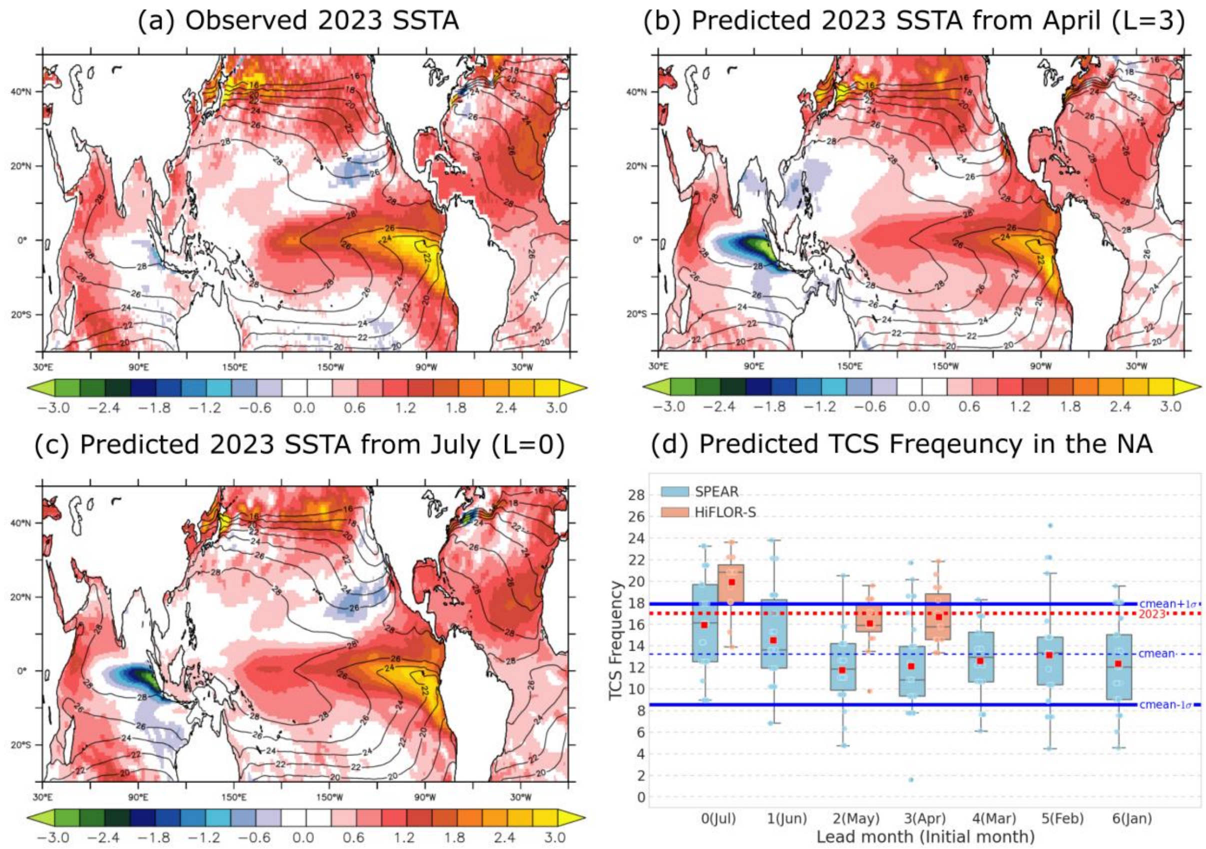

FIG. 8 Observed and predicted SST anomaly and TCS frequency over the NA during July–November 2023. (a) Observed 2023 SST anomaly, predicted 2023 SST anomaly from (b) April and (c) July initial conditions by SPEAR, and (d) Observed and predicted TCS frequency over the NA for each lead month prediction by SPEAR and HiFLOR-S. Shadings and contours in (a)–(c) represent SST anomalies and climatological mean SSTs, respectively. The dashed red line in (d) represents the 2023 observed TCS frequency, while the dashed blue line represents the observed climatological mean TCS frequency. Blue solid lines in (d) indicate the range of $\pm 1\sigma$ of the observed interannual variation. The red squares in (d) represent the ensemble mean values, whereas the dots represent values for each ensemble member. The boxes in (d) represent the lower and upper quartiles, with the horizontal lines showing the median value and the end lines showing the lowest datum still within the 1.5 interquartile range (IQR) of the lower quartile and the highest datum still within the 1.5 IQR of the upper quartile.

To assess the relative influence of the 2023 El Niño and warmer Atlantic SSTs on TCS frequency in the NA, we conducted idealized real-time attribution experiments using SPEAR and HiFLOR-S (Murakami et al., 2017, 2018; Qian et al., 2019; Nasuno et al., 2022). Similar to the HiFLOR-S predictions, we performed predictions using SPEAR and HiFLOR-S, which were forced with the predicted SSTs derived from the real-time 2023 April initial predictions by SPEAR but with some modifications. We conducted 15-member ensemble experiments from the 15-member SSTs predicted by SPEAR. Specifically, we replaced the SSTs over the tropical Pacific with climatological mean values to eliminate the 2023 El Niño conditions, denoted as the TPACCLIM experiment (Fig. 9b). Similarly, we removed the anomalously warm tropical Atlantic conditions, referred to as the MDRCLIM experiment (Fig. 9c). These experiments were compared with those using the original 2023 predicted SSTs, termed the SSTA2023 experiment (Fig. 9a), and the climatological man SSTs, termed the CLIM experiment.

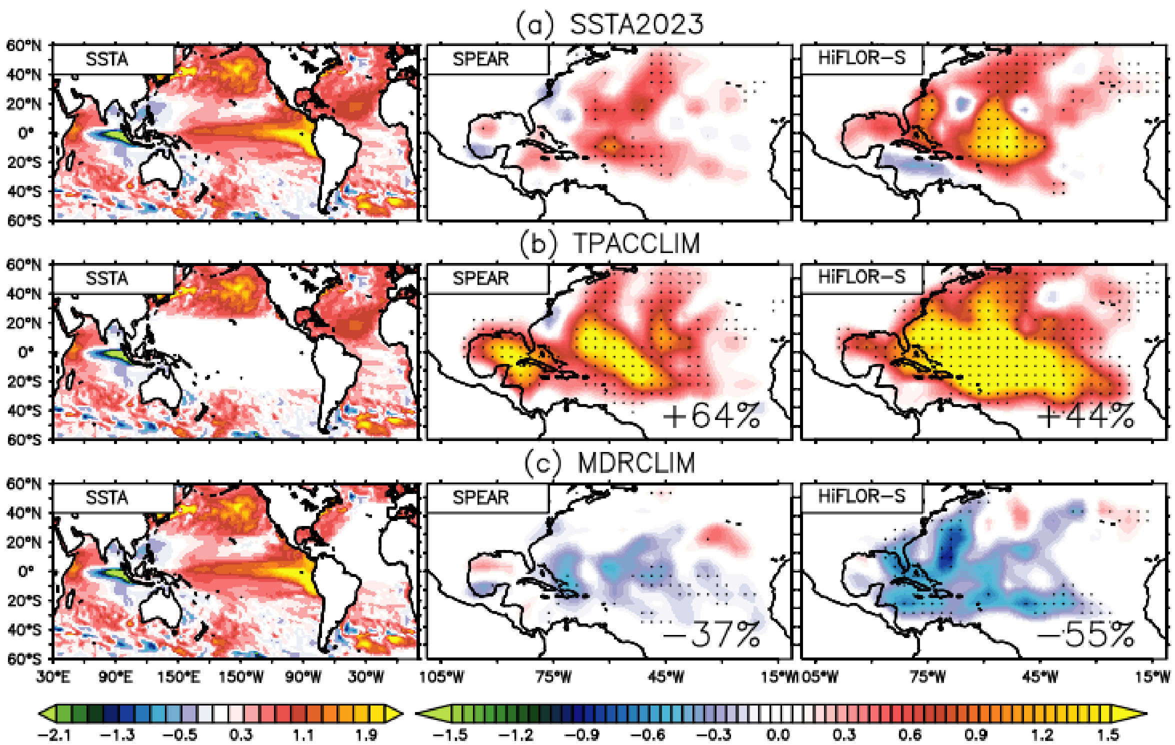

FIG. 9 Prescribed idealized SST anomaly (SSTA) and simulated anomaly of TC frequency of occurrence. Idealized seasonal predictions are conducted by prescribing the idealized SSTs in which SSTAs (left panels; units: K) are superimposed onto the climatological mean SST (CLIM). The resultant predicted TC frequency of occurrence anomalies relative to the CLIM experiment are shown by the shading in the middle- and right-hand panels (units: number per season every 5°×5° grid cell). The prescribed SSTAs are (a) all 2023 anomalies (SSTA2023); (b) as in SSTA2023, except the tropical Pacific SSTAs are set to zero (TPACCLIM); (c) as in SSTA2023, except the tropical Atlantic SSTAs are set to zero (MDRCLIM). Dots in the middle- and right-hand panels indicate the predicted change relative to the CLIM experiment is statistically significant at the 95% confidence level or above by a bootstrap method. The numbers in (b) and (c) denote fractional changes in TCS frequency relative to the SSTA2023 experiments.

Because El Niño conditions are expected to suppress TC activity in the NA, removing the 2023 El Niño through the TPACCLIM experiments is expected to result in more TCS frequency in the NA than in the SSTA2023 experiments. Likewise, removing the tropical Atlantic SST anomaly through the MDRCLIM experiments is expected to result in lower TCS frequency than in the SSTA2023 experiments. As expected, TCS frequency increases by about 64% in the SPEAR TPACCLIM experiments relative to the SSTA2023 experiments (Fig. 9b). In contrast, TCS frequency decreases by about 37% in the SPEAR MDRCLIM experiments (Fig. 9c). These magnitudes of the changes indicate that SPEAR is more sensitive to the El Niño condition than the tropical Atlantic SST for the TC activity in the NA. Meanwhile, TCS frequency increases by about 44% in the HiFLOR-S TPACCLIM experiments (Fig. 9b). However, the magnitude of the change is less than in the MDRCLIM experiments, in which TCS frequency was decreased by 55% (Fig. 9c). Therefore, in contrast to SPEAR, HiFLOR-S is more sensitive to the tropical Atlantic SST than the El Niño condition for TC activity in the NA.

Figure 10 illustrates the RCORs between Niño-3.4 SST and TC metrics in the NA compared with the RCORs between MDR SST and TC metrics for the observations and the retrospective seasonal predictions by SPEAR and HiFLOR-S during 1992–2020. Observations reveal that RCORs for most TC metrics other than US are around +0.4 with MDR SST and +0.5 with Niño-3.4 SST with a flipped sign (Fig. 10a). The April initial predictions by SPEAR (orange marks in Fig. 10b) reveal RCORs around +0.2 with MDR SST and +0.65 with Niño-3.4 SST with a flipped sign, indicating SPEAR is more sensitive to Niño-3.4 SST than MDR SST for NA TC variables compared to the observations. In contrast, those by HiFLOR-S (orange marks in Fig. 10c) show RCORs around +0.4 with MDR SST and +0.5 with Niño-3.4 with a flipped sign, closer to the observations than SPEAR. It is noted that shorter lead month predictions from SPEAR (e.g., red marks of L0) are relatively closer to the observations and HiFLOR-S than the longer lead month predictions (e.g., black marks of L4). These results are consistent with the 2023 summer predictions (blue plots in Fig. 8d), in which SPEAR changed to predict a more active season in the shorter lead month predictions than the longer lead month predictions. These results highlight that even given the same SST conditions, models would respond differently to the SST, resulting in different TC predictions.

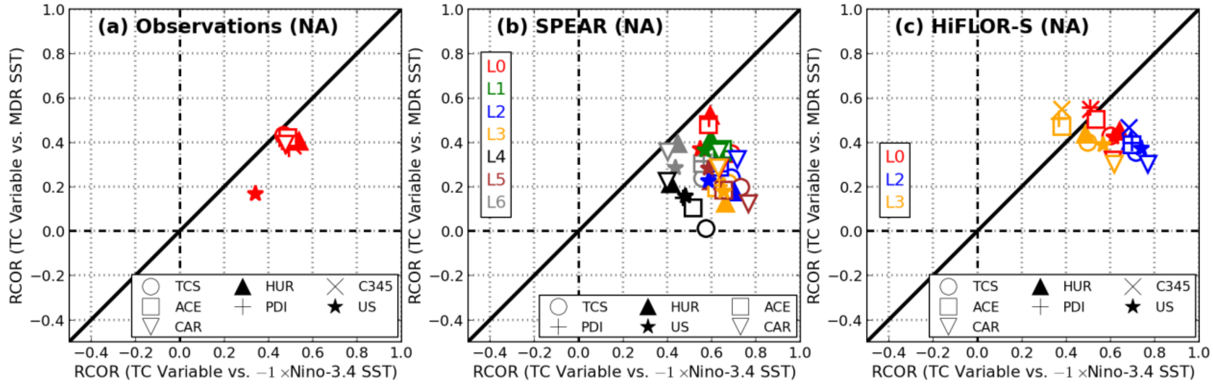

FIG. 10 Scatterplots of RCORs between TC variables and MDR SST (*y*-axis) and TC variables and Niño-3.4 SST with the reversed sign (*x*-axis) for the NA TC activity. (a) Observations from 1992–2020. Markers above the diagonal lines indicate a stronger relationship with the MDR SST compared with the Niño-3.4 SST. (b) Retrospective seasonal predictions by SPEAR and (c) HiFLOR-S during 1992–2020. Different colors indicate different lead month predictions (L0–L6). Evaluated TC variables are the same as those in Fig. 3.

4. Summary

In this study, we evaluated the skill of retrospective seasonal predictions of TC activity using the new seasonal prediction system (SPEAR and HiFLOR-S) compared to the previous seasonal prediction system (FLOR and HiFLOR) developed at GFDL. Our analysis focused on predicting various aspects of TC activity, including basin-wide frequency of different categories of TC intensity, ACE, PDI, and landfalling TCs. Additionally, we examined relevant large-scale variables from July–November across the NA, WNP, and ENP ocean basins.

SPEAR consistently demonstrates skillful predictions of TC activity across the three ocean basins. Regarding basin-wide TC frequency, SPEAR exhibits statistically significant rank correlation skill up to lead month 4 (i.e., March initial conditions), with rank correlation coefficients ranging from +0.4 to +0.6 for the NA, +0.4 to +0.5 for the WNP, and +0.4 to +0.8 for the ENP. However, when compared to FLOR, SPEAR yields comparable or lower skill in TC activity for the NA and ENP but exhibits higher skill for the WNP. Similarly, like HiFLOR, HiFLOR-S demonstrates statistically significant rank correlation skill in predicting major hurricanes in the NA, even from April's initial predictions, with rank correlation coefficients ranging from +0.4 to +0.6. HiFLOR-S generally exhibits higher skill in TC activity for the NA and WNP but demonstrates comparable skill in the ENP compared to

HiFLOR. Our analysis also indicates that the multi-model ensemble mean can sometimes outperform individual model predictions, underscoring the potential for enhancing prediction skill by integrating multiple models.

We further examined the prediction skill of regional TC activity in terms of TC frequency of occurrence and landfalling storms. SPEAR generally underperforms FLOR in landfall predictions in the coastal areas of the U.S., Caribbean islands, and Hawaii. While SPEAR exhibited smaller areas of skillful predictions of regional TC activity compared to FLOR, SPEAR exhibited skillful predictions of regional TC activity near Japan. This suggests skillful landfalling TC predictions in the region.

We assessed prediction skill in TC-relevant large-scale variables to determine whether the different prediction skill in TC variables between the previous and new prediction systems could be attributed to differences in prediction skill in large-scale variables. However, this analysis revealed that the two do not always correspond, particularly for the NA, which aligns with findings from previous studies (e.g., Murakami et al., 2016a). Further analysis indicated that the amplitude of interannual variations in TC genesis frequency plays a crucial role in prediction skill. Moreover, the sensitivity of TCs to large-scale parameters varies by region. For instance, TC genesis frequency over the eastern tropical NA is more sensitive to thermodynamical variables than to dynamical variables, while the opposite is true for the western tropical NA. Accurately simulating these sensitivities is key to improving TC prediction.

Through idealized and retrospective seasonal predictions, SPEAR demonstrates greater sensitivity to El Niño conditions, while HiFLOR-S shows less sensitivity to El Niño compared to warmer SSTs in the MDR for predicting NA TC variables. This sensitivity discrepancy resulted in conflicting TC predictions for the 2023 summer season when both El Niño conditions and warmer MDR SSTs in the NA were predicted simultaneously. This underscores the importance of not only improving the prediction skill of SSTs themselves but also enhancing the model's response of TCs to such large-scale conditions like SSTs to achieve further improvement in TC prediction skill at a seasonal time scale.

Acknowledgments

The statements, findings, and conclusions are those of the authors and do not necessarily reflect the views of the National Oceanic and Atmospheric Administration, the U.S. Department of Commerce, or the U.S. Army Corps of Engineers. We thank Drs. Jaeyeon Lee and Jan-Huey Chen for their valuable comments and suggestions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Hiroyuki Murakami.

Data Availability

The observed TC data (IBTrACS) are publicly available at https://www.ncdc.noaa.gov/ibtracs/. The observed SST data (OISST) are available at https://www.ncei.noaa.gov/products/optimum-interpolation-sst. The JRA-55 reanalysis datasets are available at https://jra.kishou.go.jp/JRA-55/index_en.html. The model outputs by SPEAR are available through the NMME webpage: https://www.cpc.ncep.noaa.gov/products/NMME/data.html. The datasets analyzed during the current study are available at https://doi.org/XXX (The data will be uploaded upon the acceptance of this manuscript). These uploaded files are freely available.

References

Adcroft, A., and Coauthors, 2019: The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *J. Adv. Model. Earth Syst.,* **11**, 3167–3211, doi:10.1029/2019MS0017 268.

Befort, D. J., K. I. Hodges, and A. Weisheimer, 2022: Seasonal Prediction of Tropical Cyclones over the North Atlantic and Western North Pacific. *J. Climate,* **35,** 1385–1397, doi:10.1175/JCLI-D-21-0041.1.

Bell, G. D. *et al.*, 2000: The 1999 North Atlantic and Eastern North Pacific hurricane season [in "Climate Assessment for 1999"]. *Bull. Amer. Meteor. Soc.,* **81 (6),** S19–S22, doi:10.1175/15200477(2000)081<1328:CAF>2.3.CO;2.

Bister, M., and K. A. Emanuel, 1998: Dissipative heating and hurricane intensity. *Meteor. Atmos. Phys.,* **52,** 233–240, doi:10.1007/BF01030791.

Bushuk, M., M. Winton, A. F. Haumann, T. Delworth, F. Lu, Y. Zhang, L. Jia, L. Zhang, W. Cooke, M. Harrison, B. Hurlin, N. C. Johnson, S. Kapnick, C. McHugh, H. Murakami, A. Rosati, K. Tseng, A. T. Witternberg, X. Yang, and F. Zeng, 2021: Seasonal prediction and predictability of regional Antarctic sea ice. *J. Climate,* **34,** 6207–6233, doi:10.1175/JCLI-D-20-0965.1.

Bushuk, M., Y. Zhang, M. Winton, B. Hurlin, T. Delworth, F. Lu, L. Jia, L. Zhang, W. Cooke, M. Harrison, N. C. Johnson, S. Kapnick, C. McHugh, H. Murakami, A. Rosati, K. Tseng, A. T. Wittenberg, X. Yang, and F. Zeng, 2022: Mechanisms of regional Arctic sea ice predictability in two dynamical seasonal forecast systems. *J. Climate,* **35,** 4207–4231, doi:10.1175/JCLI-D-21-0544.1.

Camargo, S. J., A. G. Barnston, P. J. Klotzbach, and C. W. Landsea, 2007: Seasonal tropical cyclone forecasts. *WMO Bull.*, **56,** 297–309.

Camp, J., M. Roberts, C. MacLachlan, E. Wallace, L. Hermanson, A. Brookshaw, A. Arribas, and A.A. Scaife, 2015: Seasonal forecasting of tropical storms using the Met Office GloSea5 seasonal forecast system. *Quart. J. Roy. Meteor. Soc.,* **141,** 1975–2470, doi:10.1002/qj.2516.

Chang, Y. -S., S. Zhang, A. Rosati, T. L. Delworth, and W. F. Stern, 2013: An assessment of oceanic variability for 1960–2010 from the GFDL ensemble coupled data assimilation. *Clim. Dyn.*, doi:10.1007/s00382-012-1412-2.

Chen, J.-H. and S.-J. Lin, 2011: The remarkable predictability of inter-annual variability of Atlantic hurricanes during the past decade. *Geophys. Res. Lett.,* **38,** L11804, doi:10.1029/2011GL047629.

Chen, J.-H. and S.-J. Lin, 2013: Seasonal predictions of tropical cyclones using a 25-km-resolution general circulation model. *J. Climate,* **26,** 380–398, doi:10.1175/JCLI-D-12-00061.1

Chu, P. and H. Murakami, 2022: “Climate variability and tropical cyclone activity.” Cambridge University Press. 320 pp, doi:10.1017/9781108586467.

Delworth, T. L., and co-authors, 2006: GFDL’s CM2 global coupled climate models. Part I: Formulation and simulation characteristics. *J. Climate*, **19,** 643–674, doi:10.1175/JCLI3629.1.

Delworth, T. L., and co-authors, 2012: Simulated climate and climate change in the GFDL CM2.5 high-resolution coupled climate model. *J. Climate*, **25,** 2755–2781, doi:10.1175/JCLI-D-11-00316.1.

Delworth, T. L., and coauthors, 2020: SPEAR – the next generation GFDL modeling system for seasonal to multidecadal prediction and projection. *J. Adv. Model. Earth Syst.,* **12,** e2019MS001895, doi:10.1029/2019MS001895.

Emanuel, K. A., and Nolan, D. S., 2004: Tropical cyclone activity and global climate. Preprints, 26th Conf. on Hurricanes and Tropical Meteorology, Miami, FL, Amer. Meteor. Soc., 240–241.

Emanuel, K., 2005: Increasing destructiveness of tropical cyclones over the past 30 years. *Nature,* **436,** 686–688, doi:10.1038/nature03906.

Emanuel, K., 2007: Environmental factors affecting tropical cyclone power dissipation. *J. Climate,* **20,** 5497–5509, doi:10.1175/2007JCLI1571.1.

Gray W. M., 1984a: Atlantic seasonal hurricane frequency. Part I: El Nino and 30 mb quasi-biennial oscillation influences. *Mon. Wea. Rev.*, **112,** 1649–1668, doi:10.1175/1520-0493(1984)112<1649:ASHFPI>2.0.CO;2.

Gray W. M., 1984b: Atlantic seasonal hurricane frequency. Part II: Forecasting its variability. *Mon. Wea. Rev.*, **112,** 1669–1683, doi: 10.1175/1520-0493(1984)112<1669:ASHFPI>2.0.CO;2.

Goldenberg, S. B., L. J. Shapiro, 1996: Physical mechanisms for the association of El Niño and west African rainfall with Atlantic major hurricane activity. *J. Climate,* **9,** 1169–1187, doi:10.1175/1520-0442(1996)009<1169:PMFTAO>2.0.CO;2.

Harris, L.M, S.-J. Lin, and C.Y. Tu, 2016: High-resolution climate simulations using GFDL HiRAM with a stretched global grid. *J. Climate*, **29,** 4293–4314, doi:10.1175/JCLI-D-15-0389.1.

Jia, L., T. L. Delworth, S. Kapnick, X. Yang, N. Johnson, W. Cooke, F. Lu, M. Harrison, A. Rosati, F. Zeng, C. McHugh, A. T. Wittenberg, L. Zhang, H. Murakami, and K. Tseng, 2022: Skilful seasonal prediction of North American summertime heat extremes. *J. Climate,* **35,** 4331–4345, doi:10.1175/JCLI-D-21-0364.1.

Kim, H.-S., C.-H. Ho, J.-H. Kim, and P.-S. Chu, 2012: Track-pattern –based model for seasonal prediction of tropical cyclone activity in the western North Pacific, *J. Climate*, **25,** 4660–4678, doi:10.1175/JCLI-D-11-00236.1.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bull. Amer. Meteor. Soc.,* **95,** 585–601, doi:10.1175/BAMS-D-12-00050.1.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann (2010), The international best track archive for climate stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bull. Amer. Meteor. Soc.,* **91,** 363–376, doi:10.1175/2009BAMS2755.1.

Klotzbach, P.J., and W. M. Gray, 2009: Twenty-five years of Atlantic basin seasonal hurricane forecasts (1984–2008). *Geophys. Res. Lett.,* **36,** L09711, doi:10.1029/2009GL037580.

Klotzbach, P. J., E. Blake, J. Camp, L. Caron, J. Chan, N. Kang, Y. Kuleshov, S. Lee, H. Murakami, M. Saunders, Y. Takaya, F. Vitart, and R. Zhan, 2019: Seasonal Tropical Cyclone Forecasting. *Trop. Cyclone Res. Rev.,* **8(3),** 134–149, doi:10.6057/2019TCRR03.03.

Kobayashi, S., and co-authors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, doi:10.2151/jmsj.2015-001.

Kossin, J.P. and D.J. Vimont, 2007: More general framework for understanding Atlantic hurricane variability and trends. *Bull. Amer. Meteor. Soc.,* **88,** 1767–1782, doi:10.1175/BAMS-88-11-1767.

Krishnamurthy, L., G. A. Vecchi, R. Msadek, H. Murakami, A. Wittenberg, and F. Zeng, 2016: Impact of strong ENSO on regional tropical cyclone activity in a high-resolution climate model. *J. Climate,* **29,** 2375–2394, doi:10.1175/JCLI-D-15-0468.1.

LaRow, T. E., Y. –K. Lim, D. W. Shin, E. P. Chassgnet, and S. Cocke, 2008: Atlantic basin seasonal hurricane simulations. *J. Climate*, **21,** 3191–3206, doi:10.1175/2007JCLI2036.1.

LaRow, T. E., L. Stefanova, D. –W. Shin, and S. Cocke, 2010: Seasonal Atlantic tropical cyclone hindcasting/forecasting using two sea surface temperature datasets. *Geophys. Res. Lett.,* **37,** L02804, doi:10.1029/2009GL041459.

Li, X., S. Yang, H. Wang, X. Jia, and A. Kumar, 2013: A dynamical-statistical forecast model for the annual frequency of western Pacific tropical cyclones based on the NCEP Climate Forecast System version 2. *J. Geophys. Res*., **118**, 12061–12074, doi:10.1002/2013JD020708.

Liu, M., G. A. Vecchi, J. A. Smith, H. Murakami, R. Gudgel, and X. Yang, 2018: Towards dynamical seasonal forecast of extratropical transition in the North Atlantic. *Geophys. Res. Lett.,* **45,** 12602–12609, doi:10.1029/2018GL079451.

Lu, F., and Coauthors, 2020: GFDL's SPEAR Seasonal Prediction System: Initialization and Ocean Tendency Adjustment (OTA) for Coupled Model Predictions. *J. Adv. Model. Earth Syst.,* **12,** e2020MS002, doi:10.1029/2020MS002149.

Maloney, E. D., and D. L. Hartmann, 2000: Modulation of hurricane activity in the Gulf of Mexico by the Madden-Julian oscillation. *Science*, **287 (5460)**, 2002–2004, doi:10.1126/science.287.5460.2002.

Murakami, H. and B. Wang, 2010: Future change of North Atlantic tropical cyclone tracks: Projection by a 20-km-mesh global atmospheric model. *J. Climate,* **23,** 2699–2721, doi:10.1175/2010JCLI3338.1.

Murakami, H., B. Wang, T. Li, and A. Kitoh, 2013: Projected increase in tropical cyclones near Hawaii. *Nat. Clim. Change,* **3,** 749–754, doi:10.1038/nclimate1890.

Murakami, H., G. A. Vecchi, S. Underwood, T. L. Delworth, A. T. Wittenberg, W. G. Anderson, J. -H. Chen, R. G. Gudgel, L. Harris, S. -J. Lin, and F. Zeng, 2015: Simulation and prediction of Category 4 and 5 hurricanes in the high-resolution GFDL HiFLOR coupled climate model. *J. Climate*, **28**, 9058–9079, doi:10.1175/JCLI-D-15-0216.1.

Murakami, H., G. A. Vecchi, G. Villarini, T. L. Delworth, R. Gudgel, S. Underwood, X. Yang, W. Zhang, and S. Lin, 2016a: Seasonal forecasts of major hurricanes and landfalling tropical cyclones using a high-resolution GFDL coupled climate model. *J. Climate,* **29,** 7977–7989, doi:10.1175/JCLI-D-16-0233.1.

Murakami, H., G. Villarini, G. A. Vecchi, W. Zhang, and R. Gudgel, 2016b: Statistical-dynamical seasonal forecast of North Atlantic and U.S. landfalling tropical cyclones using the high-resolution GFDL FLOR coupled model. *Mon. Wea. Rev.,* **144,** 2101–2123, doi:10.1175/MWR-D-15-0308.1.

Murakami, H., G. A. Vecchi, T. L. Delworth, A. T. Wittenberg, S. Underwood, R. Gudgel, X. Yang, L. Jia, F. Zeng, K. Paffendorf, and W. Zhang, 2017: Dominant role of subtropical Pacific warming in extreme eastern Pacific hurricane seasons: 2015 and the future. *J. Climate,* **30,** 243–264, doi:10.1175/JCLI-D-16-0424.1.

Murakami, H., E. Levin, T. L. Delworth, R. Gudgel, and P. -C. Hsu, 2018: Dominant effect of relative tropical Atlantic warming on major hurricane occurrence. *Science*, **362,** 794–799, doi:10.1126/science.aat6711.

Murakami, H. and B. Wang, 2022: Patterns and frequency of projected future tropical cyclone genesis are governed by dynamic effects. *Commun. Earth Environ.,* **3,** 77, doi:10.1038/s43247-022-00410-z.

Nasuno, T., M. Nakano, H. Murakami, K. Kikuchi, and Y. Yamada, 2022: Impacts of midlatitude western North Pacific sea surface temperature anomaly on the subseasonal to seasonal tropical cyclone activity: Case study of the 2018 boreal summer. *SOLA*, **18,** 88–95, doi:10.2151/sola.2022-015.

Qian, Y., H. Murakami, M. Nakano, P. -C. Hsu, T. L. Delworth, S. B. Kapnick, V. Ramaswamy, T. Mochizuki, Y. Morioka, T. Doi, T. Kataoka, T. Nasuno, and K. Yoshida, 2019: On the mechanisms of the active 2018 tropical cyclone season in the North Pacific. *Geophys. Res. Lett.,* **46,** 12293–12302, doi:10.1029/2019GL084566.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D.C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate,* **15,** 1609–1625, doi:10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

Siegel, S., and N. Castellan, 1988: "Nonparametric Statistics For The Behavioral Sciences", McGraw-Hill, New York, 235-244, 399 pp.

Smith, D.M. and co-authors, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nat. Geosci.,* **3,** 846–849, doi:10.1038/ngeo1004.

Takaya, Y., L. Caron, E. Blake, F. Bonnardotd, N. Bruneau, J. Camp, J. Chan, P. Gregory, J. J. Jones, N. Kang, P. Klotzbach, Y. Kuleshov, J. Lockwood, H. Murakami, A. Nishimura, D. R. Pattanaik, T. Philp, Y. Ruprich-Robert, R. Toumi, F. Vitart, S. Won, and R. Zhan, 2023: Recent advances in seasonal and multi-annual tropical cyclone forecasting. *Trop. Cyclone Res. Rev.,* **12,** 182–199, doi:10.1016/j.tcrr.2023.09.003.

Tseng, K., N. C. Johnson, S. B. Kapnick, T. L. Delworth, F. Lu, W. Cooke, A. T. Wittenberg, A. J. Rosati, L. Zhang, C. McHugh, X. Yang, M. Harrison, F. Zeng, G. Zhang, H. Murakami, M. Bushuk, and L. Jia, 2021: Are multiseasonal forecasts of atmospheric rivers possible? *Geophys. Res. Lett.,* **48,** e2021GL094000, doi:10.1029/2021GL094000.

Vecchi, G. A., and co-authors, 2011: Statistical–dynamical predictions of seasonal North Atlantic hurricane activity. *Mon. Wea. Rev.,* **139,** 1070–1082, doi:10.1175/2010MWR3499.1.

Vecchi, G. A., and co-authors, 2014: On the seasonal forecasting of regional tropical cyclone activity. *J. Climate*, **27,** 7994–8016, doi:10.1175/JCLI-D-14-00158.1.

Vecchi, G. A., and co-authors, 2019: Tropical cyclone sensitivities to CO_2 doubling: roles of atmospheric resolution, synoptic variability and background climate changes. *Clim. Dyn.,* **53(9-10)**, 5999–6033, doi:10.1007/s00382-019-04913-y.

Villarini, G., G. A. Vecchi, and J. A. Smith, 2010: Modeling of the dependence of tropical storm counts in the North Atlantic basin on climate indices. *Mon. Wea. Rev.,* **138,** 2681–2705

Vitart, F., and T. N. Stockdale, 2001: Seasonal forecasting of tropical storms using coupled GCM integrations, *Mon. Wea. Rev.*, **129,** 2521–2537, doi:10.1175/1520-0493(2001)129<2521:SFOTSU>2.0.CO;2.

Vitart, F., 2006: Seasonal forecasting of tropical storm frequency using a multi-model ensemble. *Quart. J. Roy. Meteor. Soc.* **132,** 647–666, doi:10.1256/qj.05.65.

Vitart, F., and co-authors, 2007: Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. *Geophys. Res. Lett.,* **34,** L16815, doi:10.1029/2007GL030740.

Wang, B. and H. Murakami, 2020: Dynamic genesis potential index for diagnosing present-day and future global tropical cyclone genesis. *Environ. Res. Lett.,* **15,** 114008, doi:10.1088/1748-9326/abbb01.

Wang, H., J.-K. E. Schemm, A. Kumar, W. Wang, L. Long, M. Chelliah, G.D. Bell, and P. Peng, 2009: A Statistical Forecast Model for Atlantic Seasonal Hurricane Activity Based on the NCEP Dynamical Seasonal Forecast. *J. Climate*, **22**, 4481–4500, doi:10.1175/2009JCLI2753.1.

Wang, Z., G. Zhang, T. J. Dunkerton, and F. F. Jin, 2020: Summertime stationary waves integrate tropical and extratropical impacts on tropical cyclone activity. *Proc. Natl. Acad. Sci. USA,* 117, 22 720–22 726, doi:10.1073/pnas.2010547117.

Yang X., T.L. Delworth., L. Jia, J.C. Johnson, F. Lu., and C. McHugh, 2022: On the seasonal prediction and predictability of winter surface temperature swing index over North America. *Front. Clim.* **4,** 972119, doi:10.3389/fclim.2022.972119.

Zhang, S. and A. Rosati, 2010: An inflated ensemble filter for ocean data assimilation with a biased coupled GCM. *Mon. Wea. Rev.*, **138,** 3905–3931, doi:10.1175/2010MWR3326.1.

Zhang, G., H. Murakami, R. Gudgel, and X. Yang, 2019: Dynamical seasonal prediction of tropical cyclone activity: Robust assessment of prediction skill and predictability. *Geophys. Res. Lett.,* 46, 5506–5515, doi:10.1029/2019GL082529.

Zhang, W., G. A. Vecchi, G. Villarini, H. Murakami, R. Gudgel, and X. Yang, 2017: Statistical-dynamical seasonal forecast of Western North Pacific and East Asia landfalling tropical cyclones using the GFDL FLOR Coupled Climate Model. *J. Climate,* 30, 2209–2232, doi:10.1175/JCLI-D-16-0487.1.

Zhang, W., G. Villarini, G. A. Vecchi, and H. Murakami, 2019: Rainfall from Tropical Cyclones: High-resolution Simulations and Seasonal Forecasts. *Clim. Dyn.,* **52(9-10)**, 5269–5289, doi:10.1007/s00382-018-4446-2.

Zhao, M., I. M. Held, and G. A. Vecchi, 2010: Retrospective forecasts of the hurricane season using a global atmospheric model assuming persistence of SST anomalies. *Mon. Wea. Rev.*, **138,** 3858–3868, doi:10.1175/2010MWR3366.1.

Zhao, M. and Coauthors, 2018: The GFDL global atmospheric and land model AM4.0/LM4.0 – Part I: Simulation characteristics with prescribed SSTs. *J. Adv. Model. Earth Syst.,* **10(3),** 735–769, doi:10.1002/2017MS001208.