Predicting Subseasonal Tropical Cyclone Activity Using NOAA and ECMWF Reforecasts

MATTHEW B. SWITANEK,[a,b] THOMAS M. HAMILL,[b] LINDSEY N. LONG,[c,d] AND MICHAEL SCHEUERER[e]

[a] *CIRES, University of Colorado Boulder, Boulder, Colorado*
[b] *NOAA/Physical Sciences Laboratory, Boulder, Colorado*
[c] *Climate Prediction Center, NOAA/NWS/NCEP, College Park, Maryland*
[d] *ERT, Inc., Laurel, Maryland*
[e] *Norwegian Computing Center, Oslo, Norway*

ABSTRACT: Tropical cyclones are extreme events with enormous and devastating consequences to life, property, and our economies. As a result, large-scale efforts have been devoted to improving tropical cyclone forecasts with lead times ranging from a few days to months. More recently, subseasonal forecasts (e.g., 2–6-week lead time) of tropical cyclones have received greater attention. Here, we study whether bias-corrected, subseasonal tropical cyclone reforecasts of the GEFS and ECMWF models are skillful in the Atlantic basin. We focus on the peak hurricane season, July–November, using the reforecast years 2000–19. Model reforecasts of accumulated cyclone energy (ACE) are produced, and validated, for lead times of 1–2 and 3–4 weeks. Week-1–2 forecasts are substantially more skillful than a 31-day moving-window climatology, while week-3–4 forecasts still exhibit positive skill throughout much of the hurricane season. Furthermore, the skill of the combination of the two models is found to be an improvement with respect to either individual model. In addition to the GEFS and ECMWF model reforecasts, we develop a statistical modeling framework that solely relies on daily sea surface temperatures. The reforecasts of ACE from this statistical model are capable of producing better skill than the GEFS or ECMWF model, individually, and it can be leveraged to further enhance the model combination reforecast skill for the 3–4-week lead time.

KEYWORDS: Hurricanes/typhoons; Tropical cyclones; Forecast verification/skill; Numerical weather prediction/forecasting; Statistical forecasting

1. Introduction

Tropical cyclones (TCs) are one of the most powerful and destructive natural disasters on Earth. At the same time, the economic impact of these events has steadily risen in the United States since 1980 (NOAA 2021). Given the massive influence these events have on life, property, and our economies, it is no surprise that a great amount of effort has gone into improving TC forecasts (Cangialosi et al. 2020; Klotzbach et al. 2019; Lee et al. 2020).

Just over a decade ago, Knutson et al. (2010) stated that it was unclear whether historical trends in TC frequency and intensity can be attributed to rising levels of atmospheric greenhouse gases. In their study, the focus was on the global changes to the maximum intensity of TCs. While the attribution may remain unclear, sea surface temperatures (SSTs) in the main development region of the North Atlantic Ocean show a statistically significant positive trend (see Fig. 1; Huang et al. 2017; Knapp et al. 2010, 2018; NOAA 2021). There is, likewise, an accompanying statistically significant increasing trend in the accumulated cyclone energy (ACE) in the Atlantic basin (where ACE is a function of both event intensity and duration). Furthermore, there is a statistically robust correlation between the yearly values of the Atlantic ACE (Bell et al. 2000) and the costs incurred by these TCs.

Given the reasonable expectation that the conditions favoring more extreme TC events will become increasingly likely in the twenty-first century (Alexander et al. 2018), better subseasonal forecasts can lead to better preparedness (Molina et al. 2021) and have the potential to save lives and reduce the costs of these large-scale natural hazards.

Over the last 70 years, incredible progress has been made with respect to short-term forecasting of TCs (Aberson 1998; Cangialosi et al. 2020; Komaromi and Majumdar 2014; Mehra et al. 2018). The National Hurricane Center (NHC) began in the 1950s by making 24-h operational forecasts of TCs. Next, over the period of the 1960s–90s, operational forecasts were extended out through 72 h. At the end of the twentieth century, Aberson (1998) proposed that some skill should be expected up to a 5-day or 120-h lead time. The NHC followed up in the 2000s by providing operational forecasts out through 120 h. Cangialosi et al. (2020) recently summarized the improvements of the official NHC's forecasts over the period 1954–2019. The authors showed that forecasts over the 2010–19 period were better in both the track and intensity errors at 120 h than were the 72-h forecasts over the 1990–99 period. Complementing the gains made by numerical weather prediction, a number of statistical models have used predictor information such as SSTs and vertical wind shear of the horizontal wind to produce skillful short-term forecasts of hurricane track and intensity (DeMaria and Kaplan 1994; Emanuel et al. 2006).

In addition to the progress that has been made toward improving short-term forecasting, seasonal forecasts of TCs
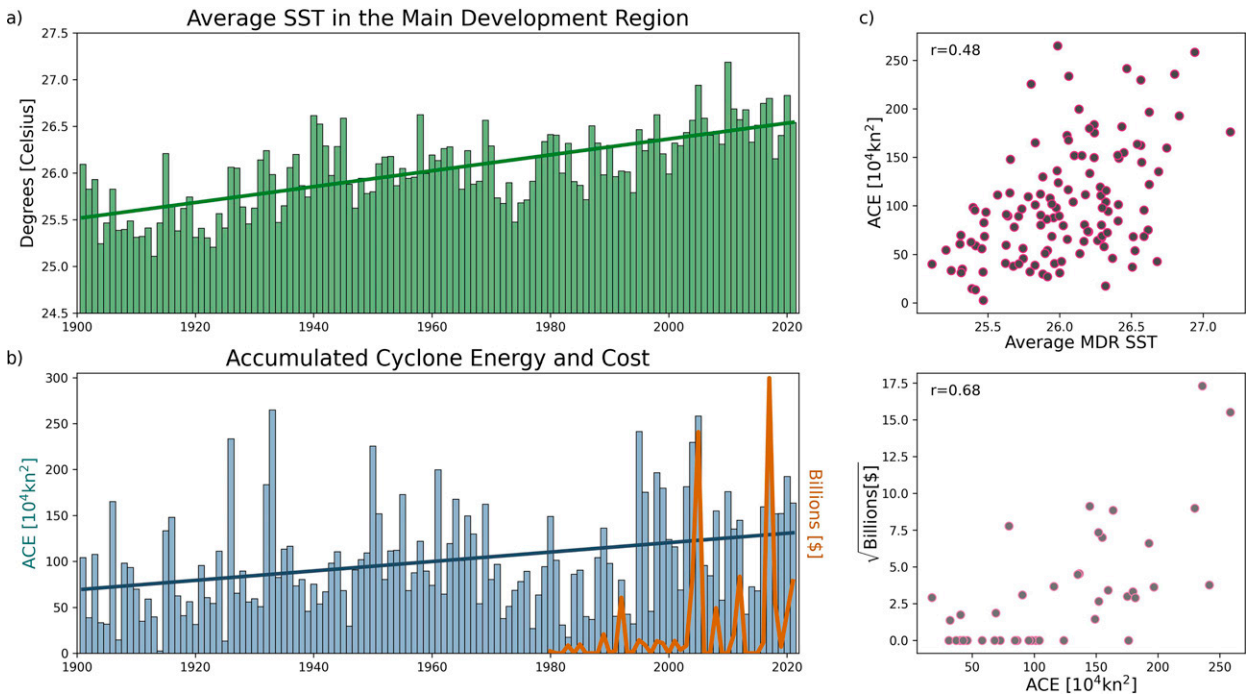
FIG. 1. (a) The time series (1901–2021) of yearly sea surface temperatures averaged across the main development region of the North Atlantic. The trend line is also plotted, which is statistically significant ($p < 0.01$). (b) The time series of accumulated cyclone energy (ACE) in the Atlantic (blue bars) and the inflation-adjusted costs of these events in the United States (orange line). The time series of the costs begins in the year 1980. As with (a), the trend line of ACE is shown, and it is statistically significant ($p < 0.01$). (c) The plots of paired years from (a) and (b), and (d) the pairings of ACE and the square root of cost values from (b) for the period 1980–2021.

have also long been an area of active research. Some early work, such as Gray et al. (1993), used a statistical model to produce seasonal TC forecasts. Since then, there have been a variety of additional statistical, dynamical, and statistical-dynamical hybrid methods that have been developed (Camargo et al. 2007; Camargo and Barnston 2009; Chen and Lin 2013; Klotzbach 2011; Wang et al. 2009). Implementing high-resolution dynamical models has also been shown to produce skillful forecasts of seasonal TC activity (Vecchi et al. 2014). Though statistical and dynamical models by themselves may be skillful, hybrid statistical–dynamical models can often produce better seasonal forecasts than either method alone (Goerss 2000; Murakami et al. 2016). Many of the different worldwide entities that currently engage in real-time seasonal TC forecasts were outlined in Klotzbach et al. (2019), which found that while forecasts of North Atlantic TC activity or intensity are often skillful when made in the month of April, the skill improves when forecasts are made at a later date (e.g., June or August).

More recently, several researchers have pursued subseasonal TC forecasts. These forecasts have a lead time greater than the 5-day short-term operational forecasts, and less than seasonal forecasts (currently, the extent of subseasonal dynamical forecasts is constrained by the forecast lead times of the models performing numerical weather prediction, or about six weeks). To achieve skill at this lead time, it is necessary to perform some amount of spatial and temporal

averaging. As a result, researchers are often aggregating the forecasts and observations over 1- or 2-week periods and across entire ocean basins, such as the Atlantic. Belanger et al. (2010) and Elsberry et al. (2010) were some of the first to show that TC forecasts could be skillful at the subseasonal time scale. However, those studies were only performed and validated over a couple of years. Given the reasonably slow-moving day-to-day persistence of phenomena such as MJO and ENSO, there has been substantial focus on the links between those teleconnective indices and TC activity at the subseasonal time scale (Pegion et al. 2008; Klotzbach 2010; Jiang et al. 2018; Camargo et al. 2019; Hansen et al. 2020; Robertson et al. 2020). Many numerical weather prediction models are now generating longer reforecast records out through lead times of at least four weeks. Using some of these longer records, a number of studies have demonstrated some level of subseasonal skill in TC forecasts in the North Atlantic (Yamaguchi et al. 2015; Wang et al. 2018; Lee et al. 2018; Vitart and Robertson 2018; Gao et al. 2019). Yamaguchi et al. (2015) looked at the skills of a few dynamical models for all global ocean basins out through two weeks. They used a 3-day temporal-aggregation window, and demonstrated that some basins exhibited modest skill out through two weeks, and that the ensemble mean of the models was superior to any model alone. Provided the limited success of subseasonal forecasts, which can vary by model, basin, and lead time, Wang et al. (2018) experimented with partitioning the skill as a function

of different tropical cyclogenesis pathways. They concluded that strong and weak tropical transition pathways had lower predictability than pathways such as low-level baroclinic or trough induced. Lee et al. (2020) found that subseasonal TC forecasts were quite capable of reproducing the seasonal cycle with the best performance being seen by the ECMWF model. However, these same models lacked skill beyond approximately two weeks when looking at the skill with respect to the month itself. Essentially, the authors found that these models were skillful in their ability to pick up on the seasonal variability of TC frequency/intensity, but that the models were not able to successfully predict whether events in September, for example, will be above or below normal with respect to a climatological September.

In this paper, we evaluate the subseasonal skill of the bias-corrected and cross-validated reforecasts of the GEFS and ECMWF models. We do this for reforecasts of ACE at lead times of 1–2 weeks (days 1–14) and 3–4 weeks (days 15–28), and using two regions: 1) a larger region encompassing the North Atlantic basin, and 2) a smaller subregion in the west Atlantic. A probabilistic forecast skill metric is used to determine if bias-corrected TC reforecasts of the GEFS and the ECMWF, whether individually or in combination, are an improvement over climatology. We also investigate how the skill varies throughout the peak hurricane season July–November. Skill values are computed with respect to a 31-day climatological window that reflects the variability of TCs throughout the peak season. Ultimately, our aim is to better understand where and when we have subseasonal skill, and subsequently we plan to transition these methods to produce real-time operational forecasts. Last, we compare the forecast skills of the GEFS and ECMWF models, along with their combination, to a statistical model that we developed that uses near-real-time daily SSTs.

2. Data

Observational TC data were obtained from IBTrACS (Knapp et al. 2010, 2018). The IBTrACS dataset includes the maximum sustained wind speeds and the minimum pressure values, along with the associated latitude and longitude coordinates, for each observed tropical cyclone over the period 1842–present. By the middle of the twentieth century, the dataset exhibits greater coverage and accuracy. Wind speeds and the associated coordinates are provided every six hours for speeds of three knots and greater. Using the IBTrACS dataset, running 2-week ACE quantities are calculated as

$$\text{ACE}_t = 10^{-4} \sum_{t=1}^{t=14} v_t^2, \tag{1}$$

where v are maximum sustained wind speeds at 6-h intervals, when wind speeds are 35 kt (1 kt $\approx$ 0.51 m s^{-1}) or greater, and the summation is performed over 2-week periods with t being day 1 of the 14-day period, where $t \in$ (1 January 2000, 2 January 2000, …, 31 December 2019). These quantities of ACE are calculated using the wind speeds that fall within the spatial boundaries of the Atlantic and west Atlantic regions (see Fig. 2). In addition to the entire North Atlantic region,

we will also compare the forecast skill in the west Atlantic region. This is because a TC that is found in the west Atlantic subdomain has a greater likelihood of making landfall in the United States than a TC found anywhere in the larger Atlantic domain. Using the IBTrACS dataset, we found that if a TC event had at least one time step with a wind speed of 35 kt or greater in the Atlantic domain, then there was approximately a 6% chance that that same event made landfall in the contiguous United States (CONUS) with wind speeds also equal to or exceeding 35 kt. That means for every 16 events that took place in the Atlantic region, there is approximately one that will have made landfall in CONUS. In contrast, we found that approximately one in nine west Atlantic events will have made landfall in CONUS. The likelihood of an event making landfall is about 80% greater when originating from the west Atlantic than an event originating anywhere in the larger Atlantic domain.

Over the period 2000–19, 96% and 95% of the total annual ACE falls within the months of July–November, for the Atlantic and west Atlantic regions, respectively (see Table 1). As a result, we focus our analysis on the subseasonal forecast performance during these months that comprise the height of North Atlantic hurricane season.

TC forecasts were computed from the raw GEFS and ECMWF output data from their ensemble set of reforecasts based on the algorithm presented in Camargo and Zebiak (2002). The TC events flagged by the algorithm were archived over their reforecast period 2000–19. Version 12 of the GEFS model was used to produce a fixed set of subseasonal reforecasts incremented weekly starting on 5 January 2000 and ending on 25 December 2019. Since the authors plan to transition these subseasonal forecasts into a real-time framework, it is important that our methodology can be implemented with real-time data. So, even though this study focuses on the subseasonal skill using the reforecasts, we additionally share details of the real-time data. Real-time GEFS forecasts are produced daily (Guan et al. 2022). In contrast, the ECMWF model with lead times greater than 10 days is run twice per week. For each forecast date, the ECMWF model is retrospectively run for the prior 20 years. For example, model forecasts made on 2 January 2020 would also be accompanied by reforecasts for 2 January 2000–19. The reforecasts of the ECMWF were produced using IFS Cycle 46r1 prior to 1 July, and IFS Cycle 47r1 beginning on 1 July. The GEFS and ECMWF models produced reforecasts with 11 ensemble members (while there are 31 and 51 ensemble members with the real-time forecasts, respectively). Similarly to the IBTrACS dataset, maximum sustained wind speeds and the minimum pressure values were calculated for each TC, for both GEFS and ECMWF, along with the associated latitude and longitude coordinates. The algorithm ingested data at 0.5° latitude by 0.5° longitude resolution from both models, and produced TC reforecasts for the GEFS at a 6-h temporal resolution and a lead time of up to 30 days, while it similarly produced reforecasts for the ECMWF at a 12-h temporal resolution and a lead time of up to 32 days. ACE is computed over 6-h intervals, while the temporal resolution of the ECMWF model is every 12 h. As a result, Eq. (1) is still used

a)
GEFS Reforecast for Sep. 21, 2000 - Oct. 4, 2000 (made Sep. 6, 2000)

b)
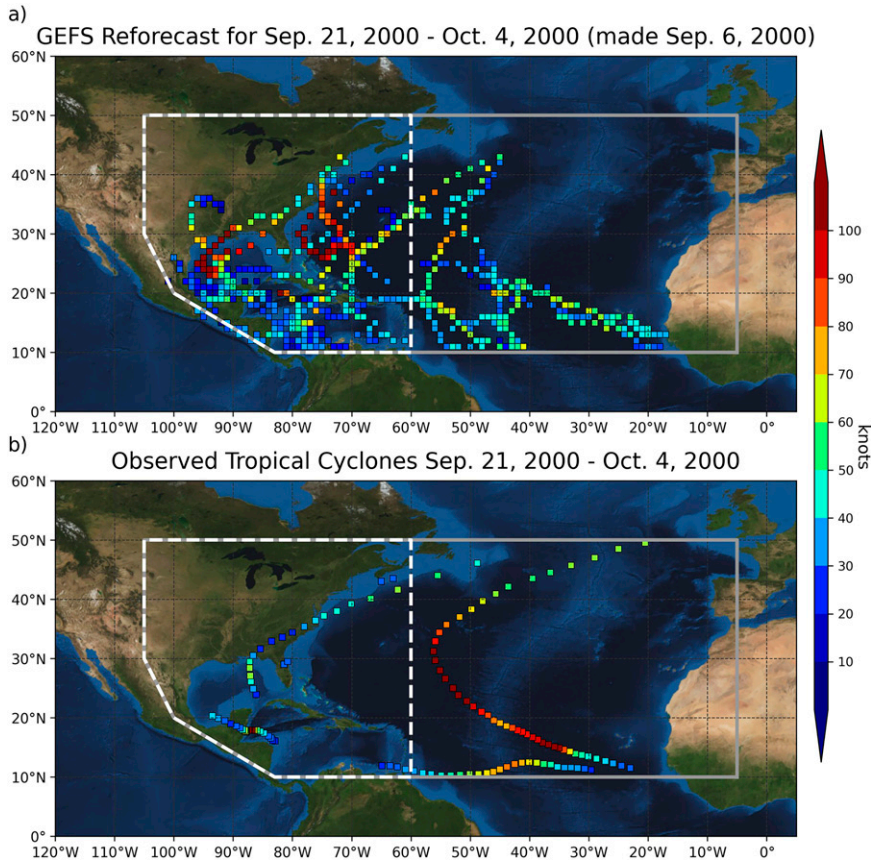## Observed Tropical Cyclones Sep. 21, 2000 - Oct. 4, 2000

FIG. 2. Model reforecasted and observed wind speeds are plotted in the Atlantic (region bounded by the solid gray line) and west Atlantic (region bounded by the dashed white line) basins. These values are the wind speeds at all 6-h time steps over the 2-week period 21 Sep–4 Oct 2000.

to calculate ACE for ECMWF using the 12-h data, but the resulting quantity is then also multiplied by a factor of 2.

SST data are obtained from the NOAA High Resolution Optimum Interpolation SST (OISST) dataset (Reynolds et al. 2007; Huang et al. 2021) provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, from their website at https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html. These

data are produced at a daily temporal resolution beginning in September 1981 and they have a spatial resolution of 1/4° latitude × 1/4° longitude.

Validation skill metric

We use the continuous ranked probability skill score (CRPSS) metric to evaluate the probabilistic performance of the model forecasts of ACE. CRPSS is a function of the forecasted and climatological continuous ranked probability score (CRPS). It is defined as

$$\text{CRPSS} = 1 - \frac{\sum \text{CRPS}_{\text{fcst}}}{\sum \text{CRPS}_{\text{clim}}}, \qquad (2)$$

where the $\text{CRPS}_{\text{fcst}}$ and $\text{CRPS}_{\text{clim}}$ are the continuous ranked probability scores of the model forecasts and climatology, respectively. $\text{CRPS}_{\text{fcst}}$ is defined as

$$\text{CRPS}_{\text{fcst}}(t) = \int_{-\infty}^{\infty} [\text{ACE}_{\text{fcst}}(x,t) - \text{ACE}_{\text{obs}}(x,t)]^2 \, dx, \qquad (3)$$

where ACE_{fcst} is the cumulative distribution function (CDF) of forecasted ACE values at time t, and ACE_{obs} is the CDF

TABLE 1. Monthly averaged ACE over the period 2000–19.

Month	Atlantic ACE	West Atlantic ACE
Jan	0.5	0.2
Feb	0.0	0.0
Mar	0.0	0.0
Apr	0.1	0.1
May	0.7	0.7
Jun	2.3	2.0
Jul	4.8	3.7
Aug	22.6	13.4
Sep	52.4	28.0
Oct	19.9	13.7
Nov	5.2	3.2
Dec	1.0	0.3

(here just a single step) of observed ACE at time t. Similarly, CRPS$_{clim}$ is defined as

$$\text{CRPS}_{\text{clim}}(t) = \int_{-\infty}^{\infty} [\text{ACE}_{\text{clim}}(x,t) - \text{ACE}_{\text{obs}}(x,t)]^2 \, dx, \quad (4)$$

where ACE$_{obs}$ is the same as above, and ACE$_{clim}$ is the CDF of the climatological distribution of ACE values at time t. A CRPSS value of one reflects perfect forecasts, while a value of zero indicates the forecasts performed as skillfully as climatology, and negative values can be interpreted as the model forecasts performing worse than climatology. Model forecasts that perform better than climatology will have a CRPSS of greater than zero. We evaluate the skill of the reforecasts using the reforecast dates of the GEFS model during the months of July–November over the 2000–19 period of record.

3. Methods

a. Bias correction of the GEFS and ECMWF forecasts

Ultimately, our goal is to establish how skillful the TC forecasts of the GEFS and ECMWF models are over the height of the hurricane season in the North Atlantic, and subsequently transition these forecasts to a real-time operational product. To this end, we first want to establish how skillful these model forecasts are after correcting biases in the mean and variance. Then, we additionally compare the skill of a combination of the bias-corrected ensemble forecasts of these two models to that of each individual model alone.

Figure 2 shows all of the modeled and observed TC wind speeds for the same 2-week period. In Fig. 2a, reforecasted wind speeds are plotted for the GEFS model initialized on 6 September 2000. These wind speeds are for the 3–4-week (days 15–28) lead time that corresponds to 21 September–4 October 2000. All of the values that fall over this 2-week period within the larger Atlantic domain (demarcated by the gray line) are used to calculate the biweekly accumulation of ACE using Eq. (1). Likewise, all of the values that fall within the west Atlantic domain (the dashed white line) are used to calculate a separate biweekly accumulation of ACE. This process, discussed here and illustrated in Fig. 2, is repeated for all reforecast time steps. As a result, we then have four different time series: 1) GEFS reforecasts of 2-week accumulated ACE for the Atlantic, 2) GEFS reforecasts of 2-week accumulated ACE for the west Atlantic, 3) ECMWF reforecasts of 2-week accumulated ACE for the Atlantic, and 4) ECMWF reforecasts of 2-week accumulated ACE for the west Atlantic. This is done independently for each of the 11 ensemble members of both models.

As we pointed out in section 2, the ECMWF forecasts are made twice a week and always on the same date (e.g., 2 January for the period 2000–19). This is in contrast to the GEFS reforecasts that are made once a week over the period of 2000–19 and starting on 5 January 2000. For any lead time, such as 1–2 weeks (days 1–14) or 3–4 weeks (days 15–28), the GEFS and ECMWF forecast time series, respectively, have lengths of 1043 and 2100 values. Provided that we have the reforecasts for 11 ensemble members, we have matrices of the size (1043, 11) and (2100, 11) for our two models at our desired lead time, and for both the Atlantic and west Atlantic regions. For each GEFS reforecast date, we find the most recent preceding ECMWF reforecast date. Again, this is because we are validating the forecast performance with respect to the GEFS time series. The ECMWF reforecast date can fall on the same date as the GEFS or up to three days prior. Note that this is also the scenario one would encounter with real-time forecasts; the GEFS real-time forecasts are made daily, while the ECMWF real-time forecasts along with new reforecasts are made twice per week.

Continuing with the example illustrated in Fig. 2, consider that we want to make a probabilistic forecast of week-3–4 ACE for the Atlantic region, using the GEFS model, with a reforecast date of 6 September 2000. Our validation period for this particular forecast time is 21 September–4 October 2000. First, we establish the "centroid" of the validation period, where we define the centroid to fall on the eighth day of the 14-day period. That would be 28 September 2000. We then have ensemble forecasts for this time step in the current year (i.e., 2000 in this case), and we denote these by the vector $\mathbf{x}_{fcst}$. To remove systematic biases, we first perform a square root transformation:

$$\mathbf{x}_{\text{Tfcst}_i} = \sqrt{\mathbf{x}_{\text{fcst}_i}} \quad \text{for } i \in (1,\ldots,11), \quad (5)$$

in order to reduce the positive skewness of each ensemble member i. The transformed forecast vector $\mathbf{x}_{Tfcst}$ is then shifted and scaled such that the first two moments of the climatology of bias-corrected, transformed ensemble forecasts $\mathbf{x}_{Tfcst_{BC}}$ match those of the climatology of square root transformed observations $\mathbf{y}_T$:

$$\mathbf{x}_{\text{Tfcst}_{\text{BC}}} = (\mathbf{x}_{\text{Tfcst}} - \overline{\mathbf{x}}_{\text{TRfcst}}) \times \frac{\sigma_{\mathbf{y}_T}}{\sigma_{\mathbf{x}_{\text{TRfcst}}}} + \overline{\mathbf{y}}_T, \quad (6)$$

where $\overline{\mathbf{x}}_{TRfcst}$, $\overline{\mathbf{y}}_T$ and $\sigma_{\mathbf{x}_{TRfcst}}$, $\sigma_{\mathbf{y}_T}$ are, respectively, the means and standard deviations of the climatology of transformed reforecasts and observations. For further details concerning some common bias correction methods and their implementation, please refer to papers such as Teutschbein and Seibert (2012) and Switanek et al. (2017). In a nutshell, this method maps the mean and variance of the modeled distribution to match the mean and variance of the observed distribution by a shifting and scaling transformation. If the model has a positive bias in the mean and a negative bias in the variance, the model PDF (i.e., probability density function) is to the right and more narrow than that of the observed PDF. The bias correction shifts the PDF to the left and additionally stretches out the distribution to more accurately follow the observed mean and variance.

As indicated by the subscript "Rfcst", these moments are estimated from reforecasts (and corresponding observations), which in our example are composed from the years 2001–19. In the general cross-validation setting, they would be composed from all reforecast years except the one for which the bias-corrected forecast is constructed. To increase the sample size used to calculate these statistics we find all reforecasts

that fall within 15 days on either side of the calendar month and day. In our example, those would be all reforecasts with valid dates (centroids) between 13 September and 13 October for the years 2001–19. We make the assumption that the performance of any one ensemble member (1, 2, …, 11) is not systematically superior to the performance of any other ensemble member. Therefore, we treat all of the forecasts from each ensemble member as equally likely. The final, bias-corrected ensemble is obtained by back-transforming the elements of the moment-adjusted, transformed forecast from Eq. (6) via

$$\mathbf{x}_{\mathrm{fcst}_{\mathrm{BC}i}} = (\mathbf{x}_{\mathrm{Tfcst}_{\mathrm{BC}i}})^2 \text{ for } i \in (1, …, 11). \tag{7}$$

We calculate $\mathrm{CRPS}_{\mathrm{fcst}}(t)$ [from Eq. (3)] using the elements of $\mathbf{x}_{\mathrm{fcst}_{\mathrm{BC}}}$ at time step t along with the observed Atlantic ACE value at the same time step. Similarly, we calculate $\mathrm{CRPS}_{\mathrm{clim}}(t)$ using a climatological distribution built from values of $\mathbf{y}$ in our leave-one-year-out cross-validated framework. We repeat the same procedure for all time steps and all years. Then, using Eq. (2), we can obtain Atlantic and west Atlantic CRPSS values for both the GEFS and ECMWF. This is done for both reforecast lead times of weeks 1–2 and weeks 3–4.

We also measure the CRPSS of the combined forecast of the two models. This is calculated as pointwise average of the respective empirical CDFs resulting from the bias-corrected ensemble forecasts of ACE from each of the two models. As a result, we obtain the cross-validated probabilistic skill of the forecasts of GEFS, ECMWF, and the combination of the two.

b. A statistical benchmark

Beyond evaluating the bias-corrected dynamical model reforecasts of the GEFS, ECMWF, and the combination of those two models, we want to evaluate whether their forecast skill is an improvement with respect to a simpler statistical benchmark model. Daily SSTs (Reynolds et al. 2007; Huang et al. 2021) available at initialization time are used to produce statistical forecasts of ACE for weeks 3–4. Such a basic statistical model cannot be expected to be competitive at short lead times where predictability is still relatively high, and we therefore focus on comparing these statistical benchmarks to week-3–4 forecasts of GEFS and ECMWF.

After performing our analysis using data at both 0.25° and 5° resolutions, it was found that using the higher resolution did not yield better forecast skill. Therefore, processing the SST data began with spatially upscaling to 5° latitude × 5° longitude. Next, we constrain the spatial domain to include all longitudes and latitudes between the equator and 50°N. The SST statistical forecast model is a two-step process:

1) Construct a single predictor by spatially aggregating the SST data across all grid points that correlate strongly with the predictand (i.e., ACE in the Atlantic/west Atlantic region).

2) Use an analog method to construct an ensemble forecast for the predictand based on this aggregated SST predictor.

In the case of the statistical forecasts, we can leverage a larger observed sample size than the dynamical reforecasts (with daily SST and ACE data over the period 1982–2020). This study's focus is on the July–November peak Atlantic hurricane season, and as a result, 39 years (i.e., 1982–2020) of SST data and TC data are used to build the statistical forecast model. Consider an example case where SSTs are used to make a forecast on 26 July of ACE over the period 9 August to 22 August or a 3–4-week (15–28-day) forecast. To be conservative about the data availability and the typical lag time therein (~2–3 days), we use data from three days prior. In our example case, we would thus use SST data on 23 July to make a forecast for 9 August to 22 August, but still consider this a 3–4-week forecast for validation purposes. Our temporal centroid—as above, the eighth day in the 14-day period—of the observed ACE for this forecast is 16 August. Similarly to the method described in section 3a, we increase the training sample size by considering additional predictor-predictand pairs and collecting data from a 31-day window. Due to the time lag between the predictor (23 July, in our example) and predictand date (16 August, in our example), the time window is defined around those two dates, respectively, thus keeping the time lag fixed.

Denote by $\mathbf{P}_{t,g,w}$ the SSTs for year $t \in (1982, …, 2020)$, grid cell $g \in (1, …, 559)$, and time-window index $w \in (-15, …, 15)$. Standardized SST data $\mathbf{P}_{t,g,w}^Z$ are obtained via

$$\mathbf{P}_{t,g,w}^Z = (\mathbf{P}_{t,g,w} - \overline{\mathbf{P}}_g)/\boldsymbol{\sigma}_{\mathbf{P}_g}, \tag{8}$$

where $\overline{\mathbf{P}}_g$ and $\boldsymbol{\sigma}_{\mathbf{P}_g}$ are, respectively, the mean and standard deviation of SST at grid cell g. These statistics are calculated from training data over the time-window index $w \in (-15, …, 15)$ and all years except the one for which the forecast is constructed. For example, to calculate $\mathbf{P}_{t,g,w}^Z$ for the year 1982, training data from the years 1983–2020 is used to calculate the standardization statistics.

To illustrate the construction of a single SST predictor, we focus on the example period 9–22 August (centered on 16 August) in the year 1982. At each SST grid cell, compute the correlation coefficient between the associated standardized SSTs and the ACE predictand, using the training data composed from the years 1983–2020 and across the 31-day time window. The field of correlation coefficients are plotted in Fig. 3a, where the title of Fig. 3a reflects the 31-day window centered about 16 August. A single SST predictor is then constructed by averaging the standardized SST values corresponding to the 10% of grid cells with the highest correlation with the predictand (see Fig. 3 for an illustration). It may be counterintuitive to aggregate SSTs in that way, i.e., across potentially different parts of the globe, but we note that this is equivalent to regressing the individual SST values against the predictand under the constraint of a shared regression coefficient, and then averaging the corresponding predictions.

This aggregated SST predictor is now used in a second step as the single predictor in an analog method (Hamill and Whitaker 2006; Delle Monache et al. 2013). Training data are composed as above, i.e., using leave-one-year-out cross-validation and a 31-day time window around the forecast date. For each
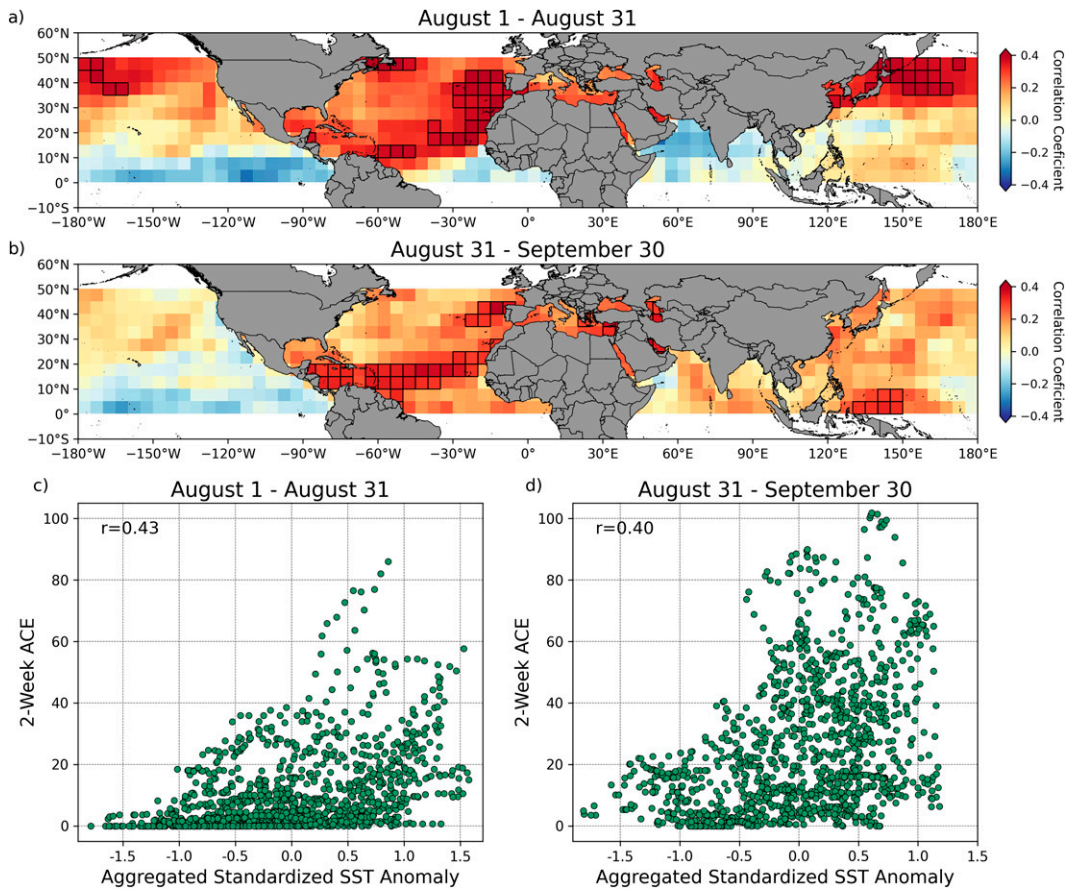
FIG. 3. (a),(b) The top 10% of SST grid cells (outlined by the black boxes), for two different 31-day time windows, with the greatest correlation over the years 1983–2020. The SST grid cells (outlined in black) are then used to produce one year's worth of aggregated SST predictors. (c),(d) The aggregated SST predictors, for all years, are plotted against observed ACE for the two time windows, respectively.

individual forecast date, Euclidean distances are calculated between the aggregated SST predictor associated with that date and the aggregated SST predictors in the training dataset. The third of the training dates with the shortest distance are selected, and a forecast ensemble is formed with the ACE values corresponding to these training dates (with the appropriate time-lag between SST predictors and the ACE centroid). An empirical CDF can be constructed from this ensemble and is used to compute values of CRPSS, using the same forecast validation dates as the GEFS model. To illustrate this step, one can imagine a case where the current year's forecast for the validation date, 16 August, is one standard deviation above normal (i.e., the point has a value of 1.0 along the x axis in Fig. 3c). Excluding the green points from the current year, the forecast ensemble distribution is constructed by using the ACE values along the y axis that correspond to the nearest third of points along the x axis, given the current aggregated SST predictor value of 1.0.

We should additionally comment on the use of other predictors. We also investigated using Niño-3.4 and different MJO time series as predictors. However, we found that none of these yielded any positive-valued probabilistic skill for either the Atlantic or the west Atlantic regions.

4. Results

a. Bias-corrected forecasts of GEFS and ECMWF

In Figs. 4a and 4b, the skill of week-1–2 bias-corrected Atlantic ACE reforecasts is shown for the GEFS and ECMWF models. The CRPSS values are the average skill over a series of 31-day moving windows centered on the dates shown, where the sample size at each time step is between 87 and 90 (which is approximately equal to the window size divided by the frequency of the forecasts multiplied by the number of reforecast years, which is $31/7 \times 20$). As previously mentioned, we focus our analysis on the months July–November. Both the bias-corrected forecasts of the GEFS and ECMWF are seen to exhibit good skill throughout the season for week-1–2 forecasts. The bias-corrected skill of the combined forecasts of GEFS and ECMWF can be seen in Fig. 4c (i.e., average of CDFs). One can observe a noticeable improvement in skill with the combined forecast with respect to the GEFS and ECMWF alone, where the CRPSS in this case is greater than 0.20 for every time window over the entire season. The CRPSS values, computed over the entire season, are summarized in Table 2. Due to the strong seasonality of ACE, these seasonally computed skill values are automatically weighted more
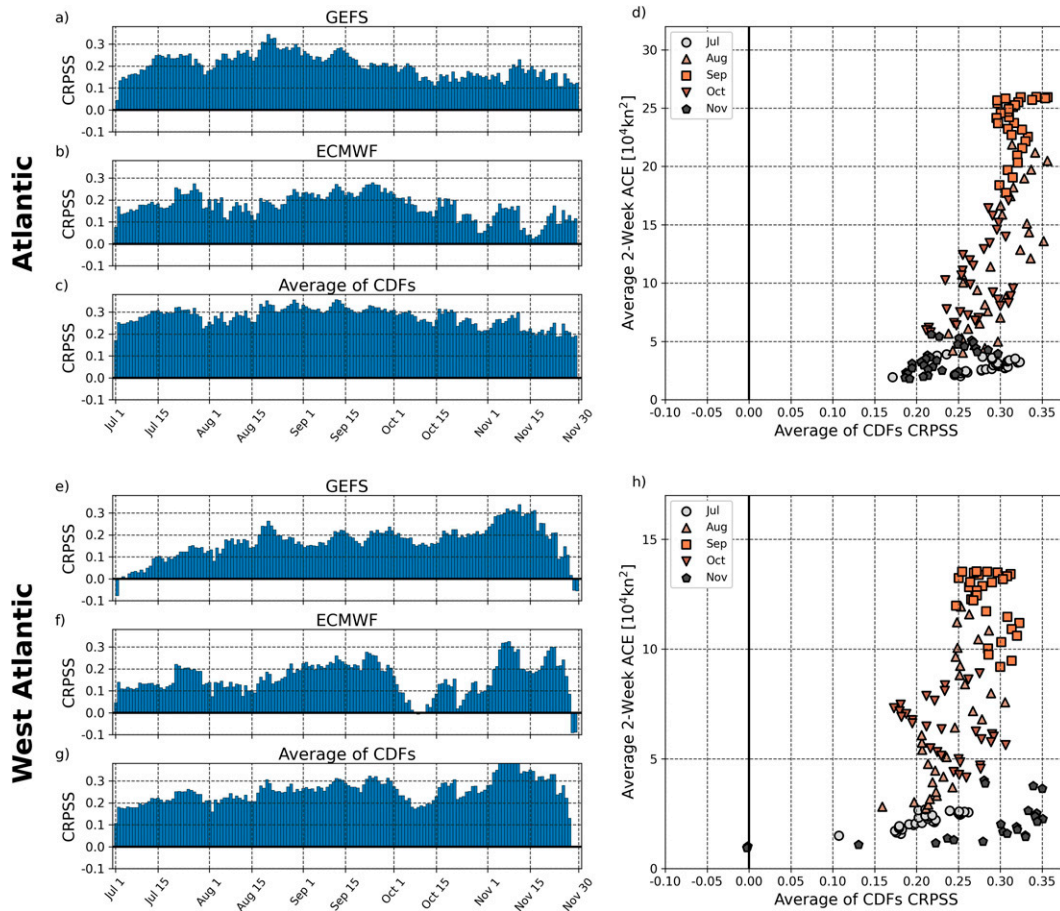
FIG. 4. The seasonality of the CRPSS skill metric is plotted for the Atlantic and west Atlantic basins for week-1–2 reforecasts. (a)–(c) The skill of the forecasts in the Atlantic when using the GEFS, ECMWF, and the combination of the GEFS and ECMWF (i.e., average of CDFs), respectively. (d) The combined CDF forecast skills from (c) are plotted on the x axis against the observed average 2-week accumulated ACE on the y axis. (e)–(h) As in (a)–(d), but for the west Atlantic.

heavily in the months of August–October, for example, than the months of July or November. This results from the fact that the errors between forecasted and climatological ACE are greater when the values of observed ACE are larger. Figure 4d reflects this seasonality of observed ACE, and it plots the pairings of the combined forecast skill on the x axis along with the average biweekly accumulations of observed ACE on the y axis (the 2-week ACE values are averaged over the 31-day window and

TABLE 2. The CRPSS values, computed over the entire season for the GEFS, ECMWF, and the combination of the two models (i.e., average of CDFs). The skill scores are split by region for the lead times of 1–2 weeks and 3–4 weeks.

| | Atlantic | | West Atlantic | |
	Weeks 1–2	Weeks 3–4	Weeks 1–2	Weeks 3–4
GEFS	0.219	0.021	0.180	−0.025
ECMWF	0.197	0.026	0.168	−0.037
Average of CDFs	0.297	0.075	0.262	0.026

over the 20-yr reforecast period of record, 2000–19). With Fig. 4d, one can now evaluate the skill as a function of the intensity of TCs (one can see that the points are concentrated on the right side of the subplot, indicating positive skill throughout the season). Here, we have a couple of notes concerning the skill. First, a Monte Carlo experiment was implemented using randomly selected standardized forecasts, and it was determined that any seasonally computed CRPSS value greater than 0.00 was statistically significant (p value < 0.01). And second, the week-1–2 forecast skills of the raw (or the non-bias-corrected) reforecasts are not shown. For weeks 1–2, the raw reforecasts performed about as well as the bias-corrected ones. However, for weeks 3–4, the raw reforecasts exhibited poorer performance in contrast to those of the bias-corrected reforecasts (see the discussion of Fig. 5). As a result, we compare the skills across the common framework of using the bias-corrected reforecasts.

Figures 4e–h show week-1–2 reforecast skill for the west Atlantic region. In the case of this smaller domain, we see an accompanying decrease in skill (see Table 2 as well). Similarly to the larger Atlantic region, there is modestly better skill with
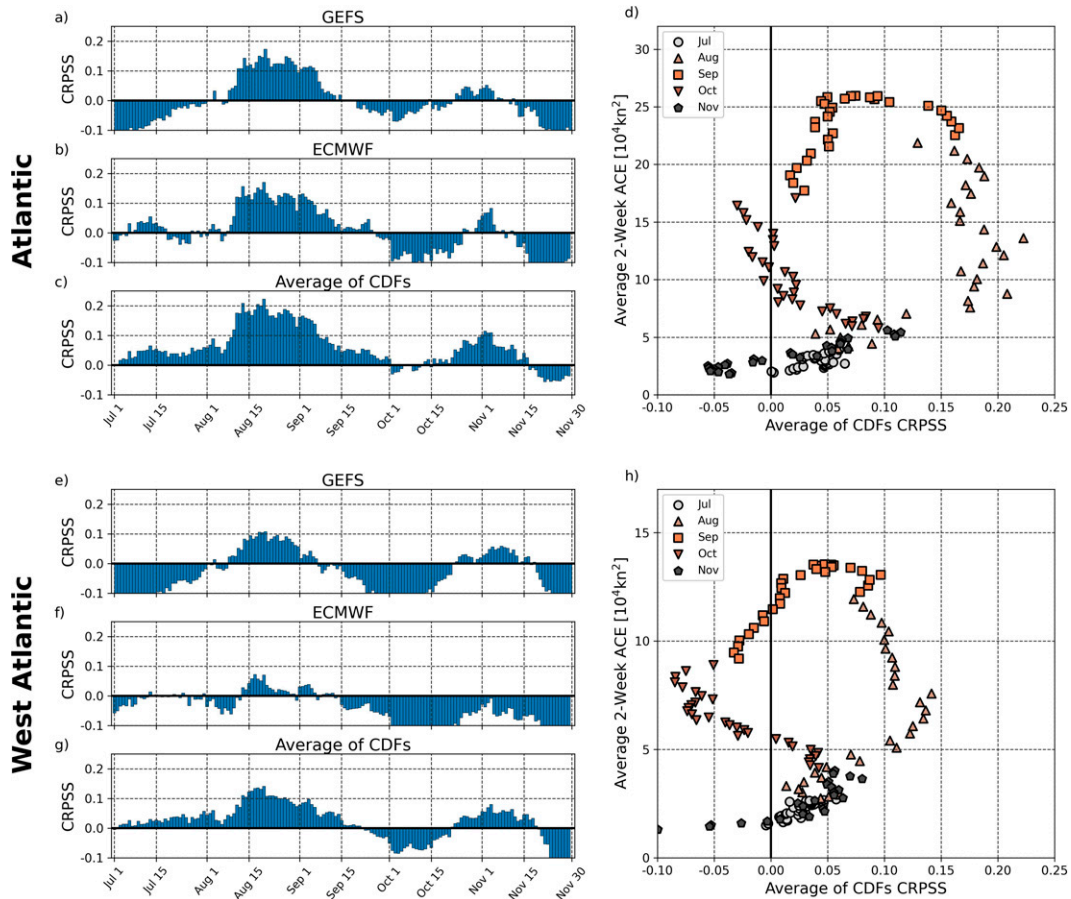
FIG. 5. As in Fig. 4, but for week-3–4 reforecasts.

the GEFS model than with the ECMWF. However, the GEFS model does not perform statistically significantly better, and the skill discrepancy most likely can be explained by the fact that we are anchoring our forecast times to the GEFS reforecast dates. This was necessary to simulate a forecast environment that can be applied daily and in real time. As we have already pointed out, we use the most recently available ECMWF reforecast for every GEFS reforecast date. Therefore, with our validation framework, the ECMWF 1–2-week forecasts have a lead time, on average, of 3–16 days (in contrast to 1–14 days for the GEFS).

In Fig. 5, we can see the results of the week-3–4 reforecasts. One can observe a marked decrease in skill between reforecasted ACE between weeks 1–2 and weeks 3–4 (also, see Table 2). However, the combined forecasts are still more skillful than climatology for much of the season for both the Atlantic and the west Atlantic. The periods where the combined forecasts perform worse than climatology can be seen at the end of September through early October and separately in the last half of November. Here, we make a couple of notes on the framework which we have used to compute the skill scores. First, we find that the bias-corrected model reforecasts perform better than the raw reforecasts. For weeks 3–4, the CRPSS values are −0.006 and −0.114, respectively, for the raw reforecasts of the GEFS and ECMWF in the Atlantic,

and −0.041 and −0.137, respectively, for the raw reforecasts of the GEFS and ECMWF in the west Atlantic. And second, the ECMWF model often has a disadvantage by analyzing the skill with respect to the GEFS reforecast dates. To be clear, the skill of the ECMWF model is better at shorter lead times. The reforecasts with lead times of 15–28 days are more skillful than those of 18–31 days, for example. Again, we chose to evaluate the model performance in a framework that can be applied in real time. And currently, real-time week-3–4 forecasts of the GEFS are performed daily, while ECMWF produces these at a frequency of twice per week. To produce forecasts for each day, then, some reforecast dates must rely on 1-, 2-, or 3-day-old ECMWF forecasts. We have simulated that real-time scenario with our reforecast evaluation. However, it can be of interest to compare the week-3–4 reforecast skills from Table 2 to those where ECMWF is at less of a disadvantage. Therefore, we compute CRPSS using only the GEFS reforecast dates where ECMWF had produced reforecasts on that day or one day prior. Using that smaller sample, the CRPSS values are 0.018 and 0.049, respectively, for the GEFS and ECMWF in the Atlantic, and −0.048 and −0.041, respectively, for the GEFS and ECMWF in the west Atlantic. Using only the 15–28- and 16–29-day reforecasts, the skill of the ECMWF model modestly outperforms the GEFS.
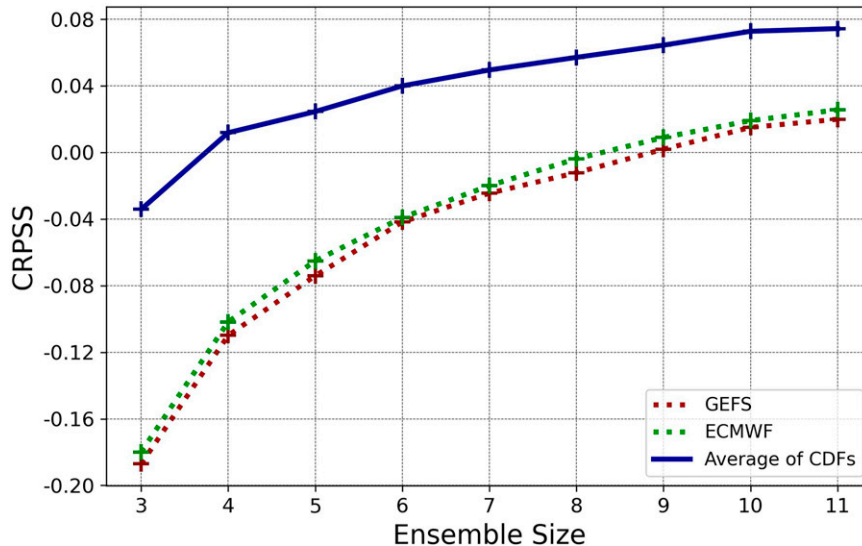
FIG. 6. The impact of ensemble size on the skill of the GEFS and ECMWF forecasts of ACE.

However, this difference in skill between the two models was not found to be statistically significant.

b. Impact of ensemble size on forecast skill

Figure 6 illustrates the impact that the ensemble size (Richardson 2001) has on TC forecast skill for weeks 3–4. This is shown for the Atlantic basin and using the GEFS and ECMWF forecasts, along with their combination. To find the relationship between ensemble size and skill, for each model we randomly selected a subset of the total number of ensembles. This was done 20 times for each ensemble size between 3 and 11. For ensemble size of 5, for example, we randomly selected 5 of the 11 ensemble members and calculated the corresponding CRPSS. Then, we repeated that procedure another 19 times. The average season-total CRPSS over the 20 randomly selected ensemble subsets are then plotted as a function of the ensemble size. One can clearly observe that the skill of the reforecasts increases as our ensemble size grows. The skill continues to increase at a relatively constant rate between 8 and 11 members, indicating that more ensemble members from both the GEFS and the ECMWF models would yield further gains in predictive skill at a lead time of 3–4 weeks. However, the gains of the model combination appear more asymptotic between 10 and 11 members. Additionally, the real-time forecasts, respectively, contain 31 and 51 members for the GEFS and ECMWF models. One would expect that with the real-time ensemble size, the skill would reach saturation as a function of ensemble size. However, in order to more accurately determine when additional ensembles do not provide additional skill, more ensembles in the reforecast time period are needed.

c. A statistical benchmark of subseasonal ACE based on sea surface temperature

Beyond evaluating the bias-corrected dynamical model forecasts of the GEFS, ECMWF and their combination, we want to compare their week-3–4 forecast skill to a simpler statistical benchmark model.

Figure 7 summarizes the monthly aggregated CRPSS values, for the GEFS, ECMWF, the statistical OISST model, the GEFS + ECMWF (the average of those two CDFs), and the GEFS + ECMWF + OISST (the average of those three CDFs). The OISST statistical model exhibits superior performance than either the GEFS or the ECMWF models, individually. The only cases where the statistical model does not perform better, is for the months of July and September, in the Atlantic, for the ECMWF model. In those cases, the ECMWF model is modestly more skillful than the OISST statistical model. In contrast, the statistical model has substantially more skill in the later months of October and November. The GEFS + ECMWF + OISST combination is more skillful than the GEFS + ECMWF combination in all cases, with some amount of positive skill seen in all months for both the Atlantic and west Atlantic regions.

Table 3 presents the skills of the different models computed over the entire season for weeks 3–4. The OISST statistical model was the best performing individual model for both the Atlantic and west Atlantic. These improvements of the OISST statistical model were found to be statistically significant (with $p < 0.01$ and $p < 0.01$, respectively, for the Atlantic and west Atlantic). Similarly, the best performing combination was the combination of the three models (with $p < 0.05$ and $p < 0.01$, respectively).

5. Conclusions

This paper finds that the TC reforecasts of the GEFS and ECMWF models are skillful at lead times up to 3–4 weeks. We evaluated the performance of the reforecasts over the period 2000–19 for the season July–November. Additionally, the combined reforecasts of the two models are more skillful than either model alone. The TC reforecast skill of the combined
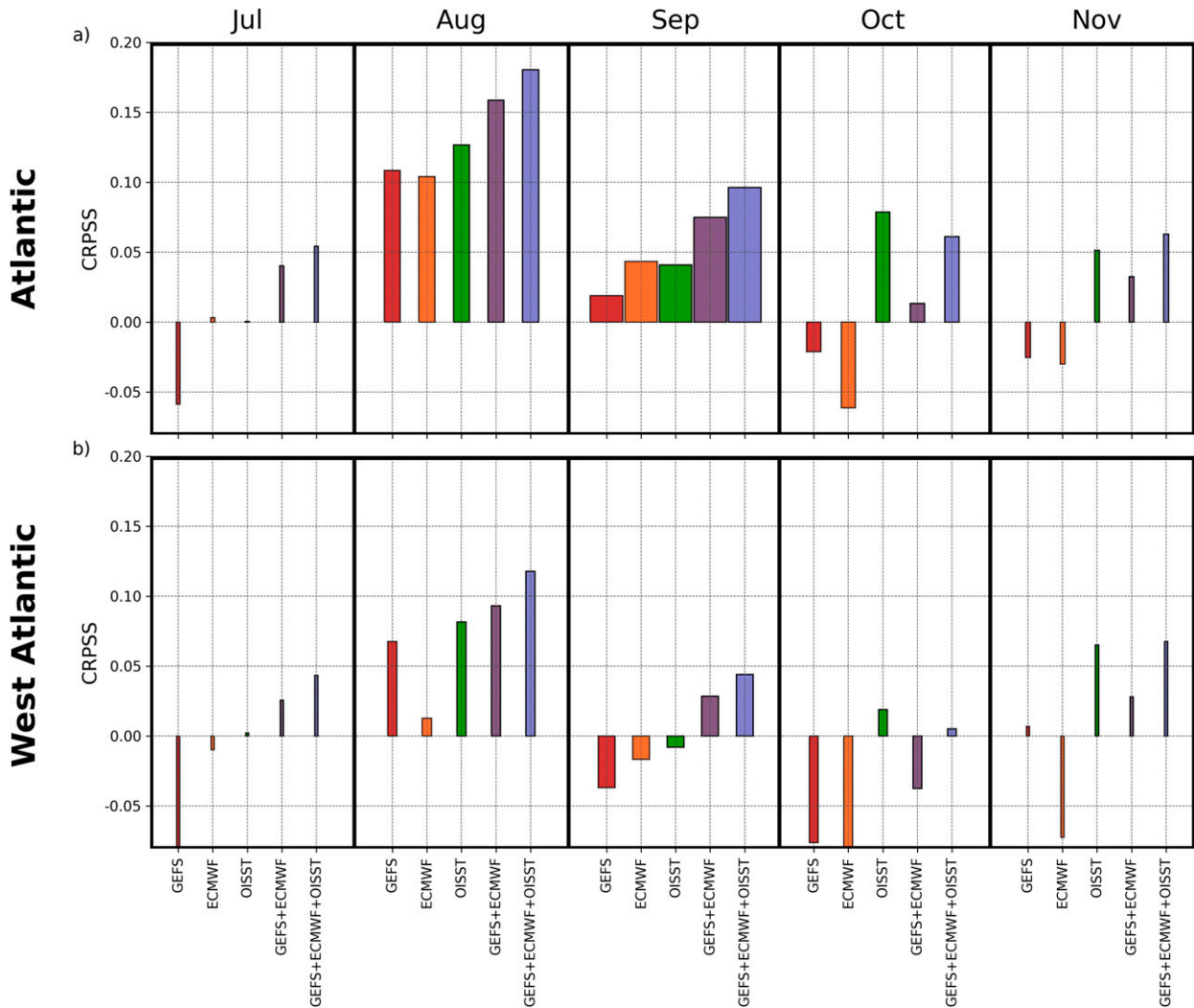
FIG. 7. The monthly aggregated skills of the different models are plotted. The colors correspond to different models, and the width of the bars reflect the average magnitude of ACE during that month (i.e., September is the most active month, with the greatest average ACE, and hence, it has the largest width). The performances for the different models are plotted for (a) the Atlantic region and (b) the west Atlantic.

model, for weeks 1–2, performs much better than climatology in both the Atlantic and west Atlantic regions. For week-3–4 reforecasts, there are times throughout the peak hurricane season where the combined model has historically exhibited

TABLE 3. The CRPSS values for weeks 3–4, computed over the entire season, using years 2000–19, for the individual GEFS, ECMWF, and OISST statistical models. Also, the skill scores are shown for the combinations of the two GEFS + ECMWF, and all three GEFS + ECMWF + OISST. The best performing single model and combination are highlighted in bold text.

	Atlantic	West Atlantic
GEFS	0.021	−0.025
ECMWF	0.026	−0.037
OISST	**0.067**	**0.023**
GEFS + ECMWF	0.075	0.026
GEFS + ECMWF + OISST	**0.102**	**0.051**

positive skill, though not at all times. At this longer 3–4-week lead time, there is modest positive forecast skill for both the Atlantic and west Atlantic regions in July, August, the first half of September, the last half of October, and the first half of November. In contrast, the reforecasts from the other times of the season (i.e., from middle of September through the middle of October, and the last half of November) were found to perform worse than climatology. The reforecasts in the larger Atlantic basin consistently exhibited greater skill than the west Atlantic region. This is true for both the 1–2- and 3–4-week reforecasts. This result is not surprising, since we should expect some amount of skill to be gained by increased spatial averaging.

Skill was additionally found to increase as a function of ensemble size for the week-3–4 reforecasts. One way to achieve further gains in TC predictive skill, would be to increase the ensemble sizes of the GEFS and ECMWF reforecasts. That way, one can more easily determine how many ensemble

members are required to reach saturation in our forecast skill. Another mechanism to provide better forecasts would be to have the forecast/reforecast framework on a daily time step. Currently, the GEFS real-time forecasts are daily. However, the GEFS is migrating its reforecasts and real-time forecasts with lead times greater than 15 days to match those of ECMWF. So, with the next version of GEFS, both models will produce real-time forecasts twice per week. What this means, is that some of the 15–28-day forecasts are in fact 16–29-, 17–30-, or 18–31-day forecasts. This will modestly diminish the skill level for weeks 3–4. Further gains in forecast skill can also be achieved by increasing the number of dynamical weather models that are used to generate the ensemble.

Last, we compared the reforecast skills of the GEFS and ECMWF, for week-3–4 reforecasts, to the skill achieved using a relatively simple statistical-analog model based on daily SSTs. The statistical OISST model was shown to outperform both the GEFS and ECMWF models, individually. This was true for both the Atlantic and the west Atlantic regions. Additionally, a combination of this model plus GEFS and ECMWF was found to statistically significantly improve the skill with respect to the combination which only uses the GEFS and ECMWF models. More sophisticated empirical-statistical algorithms (Rasouli et al. 2012; Leng and Hall 2020; Scheuerer et al. 2020) could potentially lead to further gains in forecast skill. However, our statistical model outlined here is not very difficult to implement, and we have shown it to be effective. We are in the process of transitioning these forecasts into an operational setup in order to make real-time subseasonal forecasts of week-3–4 ACE.

One potential explanation as to why such a simple statistical model can be shown to outperform the reforecasts of the dynamical models relates to how the problem is framed. The GEFS and ECMWF models perform numerical weather prediction whereby an entire state system is advanced through a set of equations governed by physical laws. These dynamical models are not solely optimized to provide the best TC forecasts. Rather, the models are attempting to simultaneously provide good forecasts for many different meteorological variables. As a result, these models output an entire space–time field of meteorological variables, and the values used to calculate TCs are just a few of the many variables that these models produce. On the other hand, we have conditioned our statistical model on SSTs with the explicit intention to provide the best TC forecasts.

Tropical cyclones are extremely destructive and costly. In the face of these natural disasters, improvements in forecasting these events can lead to better preparedness. In turn, we aim to minimize the devastation levied against affected communities.

Data availability statement. The data and code underlying the findings within the article can be accessed at https://github.com/mswitanek/tropicalcyclone-forecasts.

REFERENCES

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

Alexander, A. M., J. D. Scott, K. D. Friedland, K. E. Mills, J. A. Nye, A. J. Pershing, and A. C. Thomas, 2018: Projected sea surface temperatures over the 21st century: Changes in the mean, variability and extremes for large marine ecosystem regions of northern oceans. *Elementa*, **6**, 9, https://doi.org/10.1525/elementa.191.

Belanger, J. I., J. A. Curry, and P. J. Webster, 2010: Predictability of North Atlantic tropical cyclone activity on intraseasonal time scales. *Mon. Wea. Rev.*, **138**, 4362–4374, https://doi.org/10.1175/2010MWR3460.1.

Bell, G. D., and Coauthors, 2000: Climate assessment for 1999. *Bull. Amer. Meteor. Soc.*, **81** (6), S1–S50, https://doi.org/10.1175/1520-0477(2000)81[s1:CAF]2.0.CO;2.

Camargo, S. J., and S. E. Zebiak, 2002: Improving the detection and tracking of tropical cyclones in atmospheric general circulation models. *Wea. Forecasting*, **17**, 1152–1162, https://doi.org/10.1175/1520-0434(2002)017<1152:ITDATO>2.0.CO;2.

——, and A. G. Barnston, 2009: Experimental dynamical seasonal forecasts of tropical cyclone activity at IRI. *Wea. Forecasting*, **24**, 472–491, https://doi.org/10.1175/2008WAF2007099.1.

——, ——, P. J. Klotzbach, and C. W. Landsea, 2007: Seasonal tropical cyclone forecasts. *WMO Bull.*, **56**, 297–309.

——, and Coauthors, 2019: Tropical cyclone prediction on subseasonal time-scales. *Trop. Cyclone Res. Rev.*, **8**, 150–165, https://doi.org/10.1016/j.tcrr.2019.10.004.

Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Wea. Forecasting*, **35**, 1913–1922, https://doi.org/10.1175/WAF-D-20-0059.1.

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, https://doi.org/10.1175/JCLI-D-12-00061.1.

Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2.

Elsberry, R. L., M. S. Jordan, and F. Vitart, 2010: Predictability of tropical cyclone events on intraseasonal timescales with the ECMWF monthly forecast model. *Asia-Pac. J. Atmos. Sci.*, **46**, 135–153, https://doi.org/10.1007/s13143-010-0013-4.

Emanuel, K., 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.

Gao, K., J.-H. Chen, L. Harris, Y. Sun, and S.-J. Lin, 2019: Skillful prediction of monthly major hurricane activity in the North Atlantic with two-way nesting. *Geophys. Res. Lett.*, **46**, 9222–9230, https://doi.org/10.1029/2019GL083526.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193,

https://doi.org/10.1175/1520-0493(2000)128<1187:TCTFUA>2.0.CO;2.

Gray, W. M., C. W. Landsea, P. W. Mielke Jr., and K. J. Berry, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting*, **8**, 73–86, https://doi.org/10.1175/1520-0434(1993)008<0073:PABSTC>2.0.CO;2.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

Hansen, K. A., S. J. Majumdar, and B. P. Kirtman, 2020: Identifying subseasonal variability relevant to Atlantic tropical cyclone activity. *Wea. Forecasting*, **35**, 2001–2024, https://doi.org/10.1175/WAF-D-19-0260.1.

Huang, B., and Coauthors, 2017: NOAA extended reconstruction sea surface temperature (ERSST), version 5. NOAA/National Centers for Environmental Information, accessed 3 March 2022, https://doi.org/10.7289/V5T72FNM.

——, C. Liu, V. Banzon, E. Freeman, G. Graham, B. Hankins, T. Smith, and H.-M. Zhang, 2021: Improvements of the daily optimum interpolation sea surface temperature (DOISST) version 2.1. *J. Climate*, **34**, 2923–2939, https://doi.org/10.1175/JCLI-D-20-0166.1.

Jiang, X., B. Xiang, M. Zhao, T. Li, S.-J. Lin, Z. Wang, and J.-H. Chen, 2018: Intraseasonal tropical cyclogenesis prediction in a global coupled model system. *J. Climate*, **31**, 6209–6227, https://doi.org/10.1175/JCLI-D-17-0454.1.

Klotzbach, P. J., 2010: On the Madden–Julian oscillation–Atlantic hurricane relationship. *J. Climate*, **23**, 282–293, https://doi.org/10.1175/2009JCLI2978.1.

——, 2011: A simplified Atlantic basin seasonal hurricane prediction scheme from 1 August. *Geophys. Res. Lett.*, **38**, L16710, https://doi.org/10.1029/2011GL048603.

——, and Coauthors, 2019: Seasonal tropical cyclone forecasting. *Trop. Cyclone Res. Rev.*, **8**, 134–149, https://doi.org/10.1016/j.tcrr.2019.10.003.

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, https://doi.org/10.1175/2009BAMS2755.1.

——, H. J. Diamond, J. P. Kossin, M. C. Kruk, and C. J. Schreck, 2018: International Best Track Archive for Climate Stewardship (IBTrACS) project, version 4.0.0. NOAA/National Centers for Environmental Information, accessed 27 May 2021, https://doi.org/10.25921/82ty-9e16.

Knutson, T. R., and Coauthors, 2010: Tropical cyclones and climate change. *Nat. Geosci.*, **3**, 157–163, https://doi.org/10.1038/ngeo779.

Komaromi, W. A., and S. J. Majumdar, 2014: Ensemble-based error and predictability metrics associated with tropical cyclogenesis. Part I: Basinwide perspective. *Mon. Wea. Rev.*, **142**, 2879–2898, https://doi.org/10.1175/MWR-D-13-00370.1.

Lee, C.-Y., S. J. Camargo, F. Vitart, A. H. Sobel, and M. K. Tippett, 2018: Subseasonal tropical cyclone genesis prediction and MJO in the S2S dataset. *Wea. Forecasting*, **33**, 967–988, https://doi.org/10.1175/WAF-D-17-0165.1.

——, ——, ——, ——, J. Camp, S. Wang, M. K. Tippett, and Q. Yang, 2020: Subseasonal predictions of tropical cyclone

occurrence and ACE in the S2S dataset. *Wea. Forecasting*, **35**, 921–938, https://doi.org/10.1175/WAF-D-19-0217.1.

Leng, G., and J. W. Hall, 2020: Predicting spatial and temporal variability in crop yields: An inter-comparison of machine learning, regression and process-based models. *Environ. Res. Lett.*, **15**, 044027, https://doi.org/10.1088/1748-9326/ab7b24.

Mehra, A., V. Tallapragada, Z. Zhang, B. Liu, L. Zhu, W. Wang, and H.-S. Kim, 2018: Advancing the state of the art in operational tropical cyclone forecasting at NCEP. *Trop. Cyclone Res. Rev.*, **7**, 51–56, https://doi.org/10.6057/2018TCRR01.06.

Molina, R., D. Letson, B. McNoldy, P. Mozumder, and M. Varkony, 2021: Striving for improvement: The perceived value of improving hurricane forecast accuracy. *Bull. Amer. Meteor. Soc.*, **102**, E1408–E1423, https://doi.org/10.1175/BAMS-D-20-0179.1.

Murakami, H., G. Villarini, G. A. Vecchi, W. Zhang, and R. Gudgel, 2016: 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, https://doi.org/10.1175/MWR-D-15-0308.1.

NOAA, 2021: U.S. billion-dollar weather and climate disasters, 1980–present. NOAA/National Centers for Environmental Information (NCEI), accessed 15 September 2021, https://doi.org/10.25921/stkw-7w73.

Pegion, K., P. Pegion, T. DelSole, and M. Sirbu, 2008: Subseasonal variability of hurricane activity. Climate Test Bed Joint Seminar Series, NOAA/NCEP, Camp Springs, MD, 8 pp., https://www.nws.noaa.gov/ost/climate/STIP/FY09CTBSeminars/kpegion_121008.pdf.

Rasouli, K., W. W. Hsieh, and A. J. Cannon, 2012: Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol.*, **414–415**, 284–293, https://doi.org/10.1016/j.jhydrol.2011.10.039.

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007: Daily high-resolution-blended analyses for sea surface temperature. *J. Climate*, **20**, 5473–5496, https://doi.org/10.1175/2007JCLI1824.1.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489, https://doi.org/10.1002/qj.49712757715.

Robertson, A. W., F. Vitart, and S. J. Camargo, 2020: Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. *J. Geophys. Res. Atmos.*, **125**, e2018JD029375, https://doi.org/10.1029/2018JD029375.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Switanek, M. B., P. A. Troch, C. L. Castro, A. Leuprecht, H. I. Chang, R. Mukherjee, and E. M. C. Demaria, 2017: Scaled distribution mapping: A bias correction method that preserves raw climate model projected changes. *Hydrol. Earth Syst. Sci.*, **21**, 2649–2666, https://doi.org/10.5194/hess-21-2649-2017.

Teutschbein, C., and J. Seibert, 2012: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *J. Hydrol.*, **456–457**, 12–29, https://doi.org/10.1016/j.jhydrol.2012.05.052.

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

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *npj Climate Atmos. Sci.*, **1**, 3, https://doi.org/10.1038/s41612-018-0013-0.

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, https://doi.org/10.1175/2009JCLI2753.1.

Wang, Z., W. Li, M. S. Peng, X. Jiang, R. McTaggart-Cowan, and C. A. Davis, 2018: Predictive skill and predictability of North Atlantic tropical cyclogenesis in different synoptic flow regimes. *J. Atmos. Sci.*, **75**, 361–378, https://doi.org/10.1175/JAS-D-17-0094.1.

Yamaguchi, M., F. Vitart, S. T. K. Lang, L. Magnusson, R. L. Elsberry, G. Elliott, M. Kyouda, and T. Nakazawa, 2015: Global distribution of the skill of tropical cyclone activity forecasts on short-to-medium-range time scales. *Wea. Forecasting*, **30**, 1695–1709, https://doi.org/10.1175/WAF-D-14-00136.1.