Performance of the HWRF Rapid Intensification Analog Ensemble (HWRF RI-AnEn) during the 2017 and 2018 HFIP Real-Time Demonstrations

WILLIAM E. LEWIS

Space Science and Engineering Center, University of Wisconsin–Madison, Madison, Wisconsin

CHRISTOPHER ROZOFF AND STEFANO ALESSANDRINI

National Center for Atmospheric Research, Boulder, Colorado

LUCA DELLE MONACHE

Center for Western Weather and Water Extremes, Scripps Institute of Oceanography, University of California, San Diego, La Jolla, California

(Manuscript received 27 February 2019, in final form 25 February 2020)

ABSTRACT

The performance of the Hurricane Weather Research and Forecasting (HWRF) Model Rapid Intensification Analog Ensemble (RI-AnEn) is evaluated for real-time forecasts made during the National Oceanic and Atmospheric Administration (NOAA)'s 2018 Hurricane Forecast Improvement Program (HFIP) demonstration. Using a variety of assessment tools (Brier skill score, reliability diagrams, ROC curves, ROC skill scores), RI-AnEn is shown to perform competitively compared to both the deterministic HWRF and current operational probabilistic RI forecast aids. The assessment is extended to include forecasts from the 2017 HFIP demonstration and shows that RI-AnEn is the only model with significant RI forecast skill at all lead times in the Atlantic and eastern Pacific basins. Though RI-AnEn is overconfident in its RI forecasts, it is generally well calibrated for all lead times. Furthermore, significance testing indicates that for the 2017–18 Atlantic and eastern Pacific sample, RI-AnEn is more skillful than HWRF at all lead times and better than most of the other probabilistic guidance at 48 and 72 h. ROC curves reveal that RI-AnEn offers a good combination of sensitivity and specificity, performing comparably to SHIPS-RII at all lead times in both basins. With respect to specific high-impact cases from the 2018 Atlantic season, performance of RI-AnEn ranges from excellent (Hurricane Michael) to poor (Hurricane Florence). The multiyear assessment and results for two high-impact case studies from 2018 indicate that, while promising, RI-AnEn requires further work to refine its performance as well as to accurately situate its effectiveness relative to other RI forecasts aids.

1. Introduction

Operational forecasts of tropical cyclone (TC) track and intensity have improved significantly over the past decade as guidance available to forecasters has increased both in sophistication and reliability (DeMaria et al. 2014; Simon et al. 2018). Nevertheless, significant challenges remain. Foremost among these is rapid intensification (RI), or, more broadly, rapid intensity change (RIC). Progress forecasting RIC has been grudging indeed (e.g., Gall et al. 2013) and, given the potential logistical difficulties produced by a TC intensifying (or weakening) rapidly in the period immediately preceding landfall, the National Hurricane Center (NHC) has made improving the prediction of RIC one of its highest priorities. Likewise, the National Oceanic and Atmospheric Administration (NOAA) Hurricane Forecast Improvement Program (HFIP) has identified RIC as one of its primary research foci.

While progress in RIC prediction has been slow, there has been significant investment in TC-related numerical weather prediction (NWP) in recent decades. High-performance computing systems increasingly permit operational NWP models in both deterministic and ensemble settings to better resolve near-convective-scale motions within TCs (Biswas et al. 2018). Positive strides

Corresponding author: William E. Lewis, welewis@facstaff.wisc.edu

DOI: 10.1175/WAF-D-19-0037.1

have also been made in parameterization of the TC planetary boundary layer and microphysics (Wang et al. 2018; Mehra et al. 2018). Finally, data assimilation continues to advance and is resulting in better model initialization. All of these areas of improvement in NWP have yielded positive results in intensity prediction (DeMaria et al. 2014; Liu et al. 2018).

While ongoing improvements in NWP models will likely continue to provide incremental improvements in track and intensity forecasts, the potential for NWP models alone to improve RIC forecasts is by no means certain. First, intensity predictability studies with dynamical models show that multiscale internal TC dynamics cause NWP intensity forecast error saturation to occur within a few days, and even more quickly in cases of higher rates of intensification (Judt and Chen 2016). In the absence of external forcing, the intrinsic predictability of the TC azimuthal wind is limited to about three days (Hakim 2013), and environmental factors such as vertical wind shear, particularly in cases of moderate shear, can reduce predictability considerably (e.g., Zhang and Tao 2013; Finocchio and Majumdar 2017). These predictability studies imply there is an intrinsic limit of intensity and RI predictability in NWP models. Second, given that interaction of the mesoscale convective and synoptic regimes has been demonstrated to impact atmospheric predictability (e.g., Boer 1994; Zhang et al. 2007; Judt 2018), regime-dependent systematic errors will likely continue to exist as long as there are imperfections in NWP models. For the latter challenge especially, statistical postprocessing methods for NWP form an attractive set of complementary forecast aids.

Postprocessing techniques developed using NWP model forecast output have demonstrated success in the realm of TC intensity prediction. One successful approach is a consensus derived from the intensity predictions of multiple models (e.g., Sampson et al. 2008; Goerss and Sampson 2014; Krishnamurti et al. 1999; Williford et al. 2003; Simon et al. 2018; Ghosh and Krishnamurti 2018), with optimized multimodel ensembles producing some of the best intensity forecasts. Statistical and probabilistic techniques for intensity and RI prediction have also been derived using global model analyses (e.g., DeMaria et al. 2005; DeMaria 2009; Kaplan et al. 2015) and are important stalwarts in operational forecast centers since they have been competitive with NWP predictions. In real-time forecast settings, however, these empirical models typically use global NWP model forecast fields for their time-varying input predictors instead of predictors representing only the initial analysis fields; in this way, they are hybrid NWP–postprocessing techniques. More recently, logistic regression (Onderlinde and DeMaria 2018) and feed forward neural network (Cloud et al. 2019) approaches have been applied to high-resolution model output with promising results in probabilistic RI prediction.

Another NWP postprocessing framework has emerged in the atmospheric sciences is based on the identification of analogs in historical NWP forecasts that, through consideration of the corresponding verifying observations, help quantify analog-dependent errors. In particular, the analog ensemble (AnEn) of Delle Monache et al. (2013) is a technique that provides a prediction of some variable of interest (e.g., 2-m temperature, 10-m wind speed) from a single deterministic NWP forecast by identifying a set number of closest matching analogs from the same NWP model's historical forecasts. The verifying observations that accompany these historical forecast analogs are used to create an ensemble prediction for the chosen variable of interest. The utility of the AnEn as a postprocessing tool for the Hurricane Weather Research and Forecasting (HWRF) Model has been demonstrated recently for TC intensity prediction in Alessandrini et al. (2018), who used a variety of environmental and inner-core predictors that were derived from HWRF reforecast data to construct an AnEn for intensity prediction. This implementation of the AnEn was competitive with, or superior to, the forecast skill of the baseline HWRF and also provided useful uncertainty information. In fact, the ensemble displayed excellent dispersion properties in that the model spread matched well with the root-mean-square error of the intensity prediction at all lead times. In this study, we build upon the promising results of Alessandrini et al. (2018) for TC intensity prediction by deriving an AnEn for intensity change with specific application to RI. Furthermore, we show the real-time performance of this intensity change-based AnEn for TCs in the 2018 Atlantic and eastern Pacific hurricane seasons as well as for extended samples with forecasts made during the 2017 Atlantic and eastern Pacific seasons.

The remainder of the manuscript is organized as follows: a description of the methodology and analytical tools employed is presented in section 2; results from two cases chosen from the 2018 Atlantic season are presented in section 3; results for all HFIP real-time demo forecast cases from the 2017 and 2018 Atlantic and eastern Pacific seasons are presented in section 4; and, a discussion and conclusions are given in section 5.

2. Methodology

a. *The analog ensemble*

In contrast to other postprocessing methods currently used to generate objective forecast guidance, the AnEn technique uses a single operational forecast to generate

an ensemble of *observed values* corresponding to a forecast predictand (e.g., the change in maximum sustained wind speed $\Delta V_{\max}$ for a given forecast lead time). This task is accomplished by exploiting an extant set of historical NWP model forecasts (e.g., the 2018 HWRF preimplementation test set), which can be searched for forecasts that are close matches for the operational forecast. Briefly, a set of m matches (analogs) is selected from the historical forecast set by minimizing the following Euclidean norm with respect to a set of N_v predictors:

$$\|F_t, A_t\| = \sum_{i=1}^{N_v} \frac{w_i}{\sigma_{f_i}} \sqrt{\sum_{j=-\tilde{t}}^{\tilde{t}} (F_{i,t+j} - A_{i,t'+j})^2}, \qquad (1)$$

where F_t is a current HWRF forecast valid at the lead time t, A_t is the corresponding prospective analog valid at the same forecast lead time, w_i are the predictor weights, σ_{f_i} is the standard deviation of the time series of past forecasts of a given predictor at the same forecast lead time, and $\tilde{t}$ is an integer equal to the half-width of the lead time window over which the norm is computed ($\tilde{t} = 12,\ 24,$ and 36 h for forecast lead times of 24, 48, and 72 h, respectively). The m historical forecasts thus represent the m closest matches to the operational forecast with respect to the predictors chosen. Once the analog forecasts are determined, the AnEn itself is simply the set of m observations corresponding to each analog forecast. Since the AnEn consists of observed (not modeled) values, it represents a naturally downscaled and well-calibrated ensemble at essentially zero cost.

More complete descriptions of the AnEn method and its adaptation to the TC intensity prediction problem can be found in Delle Monache et al. (2013) and Alessandrini et al. (2018), respectively.

b. *The HWRF Rapid Intensification Analog Ensemble (RI-AnEn)*

The AnEn method described above is adapted to the prediction of intensity change $\Delta V_{\max}$ at three forecast lead times (t = 24, 48, and 72 h) by developing optimized predictand–predictor relationships for each lead time t. The probability of RI at forecast lead time t is then calculated from the analog ensemble according to

$$p(\mathrm{RI}_t) = \frac{N(\Delta V_{\max} \geq \Delta V_{\mathrm{thresh}})}{m}, \qquad (2)$$

where m is the size of the ensemble and $N(\Delta V_{\max} \geq \Delta V_{\mathrm{thresh}})$ represents the number of analog ensemble members for which forecast $\Delta V_{\max}$ meets or exceeds the RI threshold $\Delta V_{\mathrm{thresh}}$, where the three thresholds considered in this study define the 95th percentile of intensity change for a given time period (e.g., Kaplan et al. 2015) and are 30 kt (1 kt $\approx$ 0.51 m s^{-1}) per 24-h period, 55 kt per 48-h period, and 65 kt per 72-h period.

For the 2018 HFIP demonstration, the ensemble size is m = 20, and the model is trained with 1146 and 1372 Atlantic and eastern Pacific reforecasts (from the years 2015–17), respectively, from the HWRF preimplementation test (i.e., H218). An objective forward feature selection method described in Alessandrini et al. (2018) is employed here to find a small set of predictors from a large set of kinematic and thermodynamic predictors derived from the HWRF forecast fields. The optimal set of predictors and their weights in the analog search metric are chosen such that the mean absolute error over an independent portion of the training period is minimized. The set of optimal predictors that emerged from the training process and which are used in the 2018 HFIP demonstration is given in Table 1. Additionally, to address limitations imposed by the rather small sample size resulting from a single season of forecasts, we also include results from the 2017 HFIP real-time demonstration. This version of the RI-AnEn was trained on the 2017 HWRF preimplementation test set (H217) and included 858 and 1630 Atlantic and eastern Pacific reforecasts (from the years 2014–16), respectively. The set of optimal predictors used in the 2017 RI-AnEn is given in Table 2. The leading predictor at all lead times for both the 2017 and 2018 iterations of RI-AnEn is the HWRF $\Delta V_{\max}$ (i.e., the predicted intensity change for a particular forecast lead time as computed from the HWRF deterministic model output). Other predictors are derived from full 3D HWRF Model output and include shear, relative humidity, inertial stability, and vertical motion. A full investigation of why some predictors are selected for a given lead time (and why others are not) is beyond the scope of this paper. Each of the predictors does possess a recognized phenomenological relationship to TC intensity change, however, and the authors hypothesize that the variability in predictor selection from cycle to cycle is due to a combination of statistical (i.e., sampling-related) and process-related reasons.

c. *Test and verification datasets*

For verification purposes, the HURDAT database, available from the NHC (https://www.nhc.noaa.gov/data/#hurdat), is used to determine the timing of RI events. HURDAT contains the final best track datasets for both 2017 and 2018. For comparison purposes, the performance of two operational statistical models, SHIPS-RII and the SHIPS-based RI consensus (SHIPSCON) (Kaplan et al. 2015) is evaluated from SHIPS text files also available from NHC (https://ftp.nhc.noaa.gov/atcf/stext/). SHIPS-RII is a linear-discriminate analysis-based

TABLE 1. Optimal predictors chosen for the 2018 HWRF RI-AnEn.

Forecast lead time	Atlantic basin	Eastern Pacific basin
24 h	ΔV_{max} (HWRF), 500–250-hPa relative humidity, inner-core sensible heat flux (r = 0–50 km), 850–200-hPa vertical shear magnitude	ΔV_{max} (HWRF), inner-core sensible heat flux (r = 0–50 km), 500–250-hPa relative humidity, 850–200-hPa vertical shear magnitude
48 h	ΔV_{max} (HWRF), maximum potential intensity, inertial stability (r = 0–100 km), 500–250-hPa relative humidity	ΔV_{max} (HWRF), maximum potential intensity, 850–200-hPa vertical shear magnitude
72 h	ΔV_{max} (HWRF), inertial stability/vertical motion coupling symmetry (r = 100–250 km)	ΔV_{max} (HWRF), MPI, inner-core average vertical motion (r = 0–50 km), inertial stability (r = 0–100 km), HWRF V_{max} (t = 0)

model that uses a mixture of global model and satellite-derived predictors, while SHIPSCON incorporates other SHIPS models (including those using logistic regression and Bayesian techniques) into a consensus forecast. In contrast to the SHIPS models, which employ as many as 10 predictors, the RI-AnEn is somewhat more parsimonious (several predictors per lead time) and also incorporates predictors derived from the TC inner core using HWRF Model output.

In addition to the RI-AnEn and SHIPS models, the deterministic (i.e., operational) HWRF Model is also considered. The RI-AnEn is built upon a foundation of HWRF historical forecasts, and it is necessary to compare RI-AnEn performance with HWRF performance to adjudge the value-added benefit of RI-AnEn.

d. Assessment tools

Forecasts obtained from RI-AnEn, SHIPS-RII, SHIPSCON, and HWRF are evaluated using several verification tools. The Brier skill score (BSS; Wilks 2006) measures the degree of improvement achieved by a set of probabilistic forecasts relative to a climatological baseline:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{forecast}}}{\mathrm{BS}_{\mathrm{climatology}}}, \tag{3}$$

where $\mathrm{BS}_{\mathrm{forecast}}$ and $\mathrm{BS}_{\mathrm{climatology}}$ are the Brier scores (Brier 1950) obtained for the set of forecasts and for climatology, respectively. A BSS value greater than 0 indicates that the forecasts are skillful (with 1 being a perfect BSS value), whereas a negative value indicates that the simple climatological forecast is superior. The climatological RI probabilities for the Atlantic and eastern Pacific basins are computed from the HURDAT dataset over the period 1987–2017 in accord with the criteria set forth in Kaplan and DeMaria (2003). These are given in Table 3 and are in good agreement with the climatological RI probabilities computed from the SHIPS developmental dataset and that appear in the SHIPS text files.

In addition to forecast skill, we verify the calibration of the probabilistic models (RI-AnEn, SHIPS-RII, SHIPSCON) by constructing reliability diagrams (Hartmann et al. 2002). Forecasts are gathered into four bins ($p \le 0.01, 0.01 < p \le 0.3, 0.3 < p \le 0.6, p > 0.6$) and the number of verifying RI events corresponding to those forecasts is then divided by the total number of forecasts. Ideally, the observed frequency of RI would match the predicted probability when averaged over the entire sample. Ordered pairs lying above the diagonal of a reliability diagram indicate the observed frequency exceeds the model prediction (i.e., underconfidence of the underlying forecast model), whereas values lying below the diagonal indicate overconfidence of the underlying model. Given the sensitivity of reliability to sample size, this assessment will only be performed with the sample composed of forecasts obtained for the 2017 and 2018 HFIP real-time demonstrations

TABLE 2. Optimal predictors chosen for the 2017 HWRF RI-AnEn.

Forecast lead time	Atlantic basin	Eastern Pacific basin
24 h	ΔV_{max} (HWRF), symmetry of low-level inflow (r = 0–100 km)	ΔV_{max} (HWRF), minimum sea level pressure
48 h	ΔV_{max} (HWRF), convective available potential energy (r = 200–600 km), latent heat flux (r = 0–50 km)	ΔV_{max} (HWRF), total condensate (r = 0–100 km)
72 h	ΔV_{max} (HWRF), storm translation speed, latent heat flux (r = 0–50 km)	ΔV_{max} (HWRF), inertial stability (r = 0–100 km)

TABLE 3. Baseline climatological RI probabilities (%) for the Atlantic and eastern Pacific basins computed for the years 1987–2017. The observed frequency of RI for the samples analyzed in the manuscript is shown in parentheses (2018) and brackets (2017–18).

Lead time/ΔV_{max}	Atlantic	Eastern Pacific	Atlantic + eastern Pacific
24 h/30 kt	6.6 (6.1) [11.1]	8.3 (11.7) [12.1]	7.4 [11.6]
48 h/55 kt	4.6 (6.3) [7.7]	6.4 (6.8) [7.6]	5.6 [7.6]
72 h/65 kt	5.1 (4.5) [3.8]	5.2 (7.0) [7.2]	5.1 [5.4]

for both the Atlantic and eastern Pacific basins (see section 2e below).

Receiver operating characteristic (ROC) (Wilks 2006) curves are used to further evaluate the performance of the probabilistic models with respect to resolution (i.e., their ability to discriminate between events and nonevents). The ROC provides a compact representation of model performance at various probability thresholds and can prove useful in configuring such models for use in operations when a trade-off between model sensitivity and specificity (i.e., hit rate and false alarm rate) must be considered. To provide a standardized representation of the ROC results, we also compute ROC skill scores (ROCSS). The ROCSS (Mason and Graham 1999) is related directly to the area under the ROC curve (AUC):

$$\mathrm{ROCSS} = 2\,(\mathrm{AUC} - 0.5). \tag{4}$$

The ROCSS is somewhat easier to interpret than the ROC curve since a value equal to 1 implies perfect forecasts and a ROCSS lower than 0 indicates performance worse than that obtainable by random draws from a uniform distribution [i.e., $U(0,1)$]. As with reliability, the ROC and ROCSS are computed only for the larger 2017–18 Atlantic and eastern Pacific sample.

Finally, to assess the significance of the skill scores discussed above, standard bootstrap confidence intervals (Davison and Hinkley 1997) are constructed for pairwise BSS differences between the RI-AnEn and the other models using 1000 bootstrap replicates. A significance level of $\alpha = 0.05$ is used, meaning that 95% of the replicates are contained within the displayed intervals. This permits testing of the hypothesis that the skill score computed from one of the model forecasts at a given lead time is significantly better than the same metric computed from one of the other models. For example, if the skill score of the RI-AnEn is greater than that of a competing model and the confidence intervals are entirely composed of positive values, then the RI-AnEn skill score is judged to be significantly better (at the 95% level). We also evaluate the overall skill of each of the model's RI forecasts using the Diebold–Mariano (DM) (Diebold and Mariano 1995) test, which is another method that allows for head-to-head comparisons of alternative forecasts. Originally developed to compare time series forecasts arising in economics applications, the DM test has recently been applied successfully to forecast output from NWP models (Sperati et al. 2015). The DM's null hypothesis [that the loss difference between two sets of model forecasts is statistically indistinguishable, i.e., distributed as $N(0,1)$] is tested for all model–model pairs.

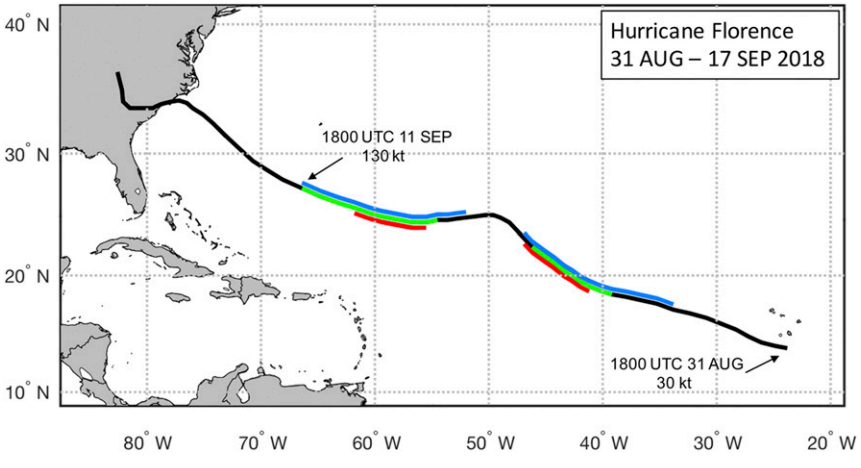

FIG. 1. Track of Hurricane Florence (2018) from its initial designation as a tropical depression on 31 Aug until its extratropical transition over the inland southeastern United States on 17 Sep. Periods during which Florence intensified at rates satisfying the 24-, 48-, and 72-h thresholds for rapid intensification are indicated in red, green, and blue, respectively. Maximum intensity (130 kt) was achieved at 1800 UTC 11 Sep.

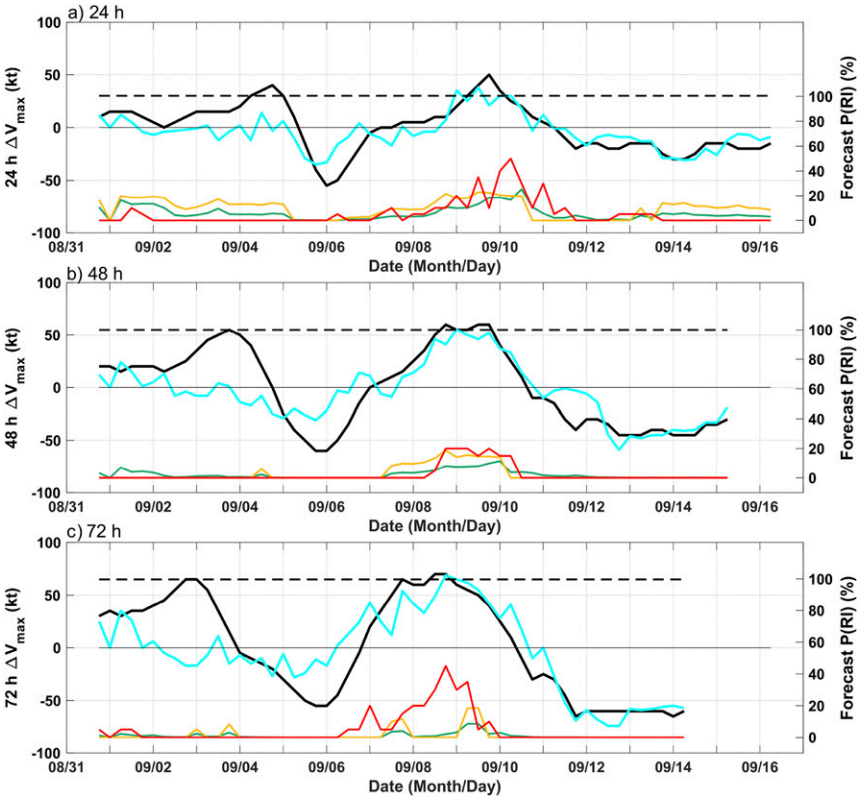

FIG. 2. Time series of intensity change ΔV_{max} for Hurricane Florence (2018) for forecast cycles valid at (a) 24, (b) 48, and (c) 72 h. Observed values are depicted by the heavy black line. The corresponding ΔV_{max} from the deterministic HWRF forecast is given by the cyan line. RI forecasts from the probabilistic models are superimposed (red, orange, and green indicating the RI-AnEn, SHIPS-RII, and SHIPSCON models, respectively) with corresponding scale on the right axis. Forecast cycles for which RI occurred are indicated by ΔV_{max} values that lie above the black dotted line. Labels on the abscissa indicate the initialization date for the forecasts.

e. Forecast period and case selection

RI-AnEn forecasts were generated in real time on the NOAA Jet supercomputer over the period 22 July–31 October 2018 for each available operational HWRF forecast. As currently configured, RI-AnEn is intended to demonstrate performance benefits relative to the deterministic HWRF and other RI guidance available at the given synoptic time. Full operationalization (to include development of an interpolated version of the AnEn, which would render it an "early" guidance product) has not been undertaken, but could readily be accomplished.

To ensure that the verification adheres to the definition of rapidly intensifying TCs set forth in Kaplan and DeMaria (2003), only those TCs that satisfy the following two criteria are retained for evaluation: 1) tropical depression strength or greater, and 2) over water for the duration of the forecast period of interest.

3. Case studies from the 2018 Atlantic season

a. Hurricane Florence (AL06)

Florence was a tropical cyclone typical of the "Cape Verde" type, developing from an African easterly wave in the far eastern tropical Atlantic and attaining hurricane intensity east of the Lesser Antilles (Fig. 1). Two distinct RI periods were observed with Florence. The first RI period, which began on 4 September, occurred despite large-scale environmental conditions (15–20 kt of southwesterly vertical wind shear, midlevel relative humidity $< 50\%$, sea surface temperature $< 27°C$) that were not conducive to significant strengthening (Stewart and Berg 2019). Forecasts for Florence (Fig. 2) reveal that the first RI is not well anticipated by HWRF or the HWRF RI-AnEn. RI-AnEn probabilities are near zero for nearly all 24-, 48-, and 72-h forecasts made during this period, and HWRF forecasts advertise only a mixture of slight intensification or weakening (±10–15 kt) at

each lead time. The large HWRF forecast errors are apparently due to the adverse impact of marginal to unfavorable environmental conditions, namely strong vertical wind shear and low midlevel relative humidity diagnosed in the model output fields. Given that the RI-AnEn uses these environmental predictors for its 24-h forecasts (i.e., relies upon the observed relationships between 24-h intensity change and HWRF vertical wind shear/midlevel relative humidity), it is understandable that RI-AnEn is unable to improve upon the HWRF forecasts for this RI event. We can only speculate the reasoning for this forecast difficulty, but it is possible the HWRF misdiagnosed the actual shear felt by Florence (e.g., Ryglicki et al. 2019), or similarly, incorrectly handled the water vapor content near the vortex core. There could be other internal dynamics that are not well captured by the HWRF as well. If, on the other hand, the HWRF is faithfully representing the TC environment, then there is also the possibility that a larger HWRF reforecast dataset may be necessary to provide a more adequate sampling of the TC RI climatology of which this particular event was an outlying sample. In contrast to the HWRF and RI-AnEn performance, the SHIPS-based models do indicate enhanced probabilities (~20%, or about three times the climatological mean) for the 24-h RI event, but these are part of a 4-day period of near-constant values and therefore do not resolve the timing of the event well. The 48- and 72-h forecasts of this event are exceedingly poor for all models, underscoring the peculiarity of the event and highlighting the need to better resolve the internal and external processes that control RI.

The second RI period (for forecasts initiated on 9, 8, and 7 September 2018 for lead times of 24, 48, and 72 h) occurred in a more favorable synoptic environment (vertical shear of 5–10 kt) and is consequently better predicted. All the models indicate elevated probabilities for 24 and 48 h, though these generally lag the onset of the event (i.e., they take a few cycles to ''catch up'' and therefore the maximum probabilities occur several cycles after the RI event ends). The exception is the SHIPS-RII, which shows its peak 48-h probability at 1800 UTC 8 September (a verifying 48-h RI event). At 72 h, the RI-AnEn alone among the models shows elevated probabilities during the broad period of intensification, but once more the peak forecast probability (45% for the forecast initialized at 1800 UTC 8 September) verifies at a lead time (1800 UTC 11 September) that is at the tail end of the RI period.

b. *Hurricane Michael (AL14)*

Unlike Florence, Michael developed much closer to the continental United States and began intensifying rapidly almost immediately. Figure 3 shows the track of Michael from its origins in the western Caribbean Sea to its ultimate landfall as a category 5 hurricane near Mexico Beach, Florida, only three days later. Perhaps what is most remarkable is the nearly continuous RI that occurred at all lead times for the duration of Michael's overwater traverse. With the exception of an 18-h period (0600 UTC 8 September–0000 UTC 9 September) during which the criterion for 24-h RI was not satisfied, the hurricane intensified rapidly with respect to RI thresholds at 24, 48, and 72 h. Like Florence during its first RI period, the large-scale environment in which Michael was embedded appeared less than ideal, with moderate-to-strong southerly and south-southwesterly vertical wind shear. There is some indication that favorable interaction with an upper-level trough over the Gulf of Mexico may have compensated for what otherwise appeared to be marginal conditions. In any case, the evolution of Michael was relatively well captured by most of the guidance, though there was a persistent low bias in intensity forecasts (Beven et al. 2019).

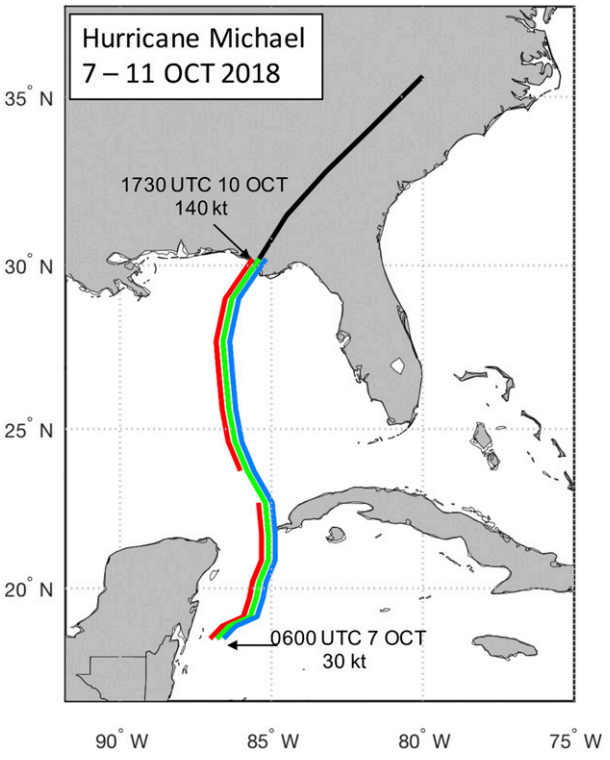

FIG. 3. Track of Hurricane Michael (2018) from its initial designation as a tropical depression on 7 Oct until its extratropical transition over the inland southeastern United States on 11 Oct. Periods during which Michael intensified at rates satisfying the 24-, 48-, and 72-h thresholds for rapid intensification are indicated in red, green, and blue, respectively. Maximum intensity (140 kt) was achieved at the time of landfall at 1730 UTC 10 Oct.

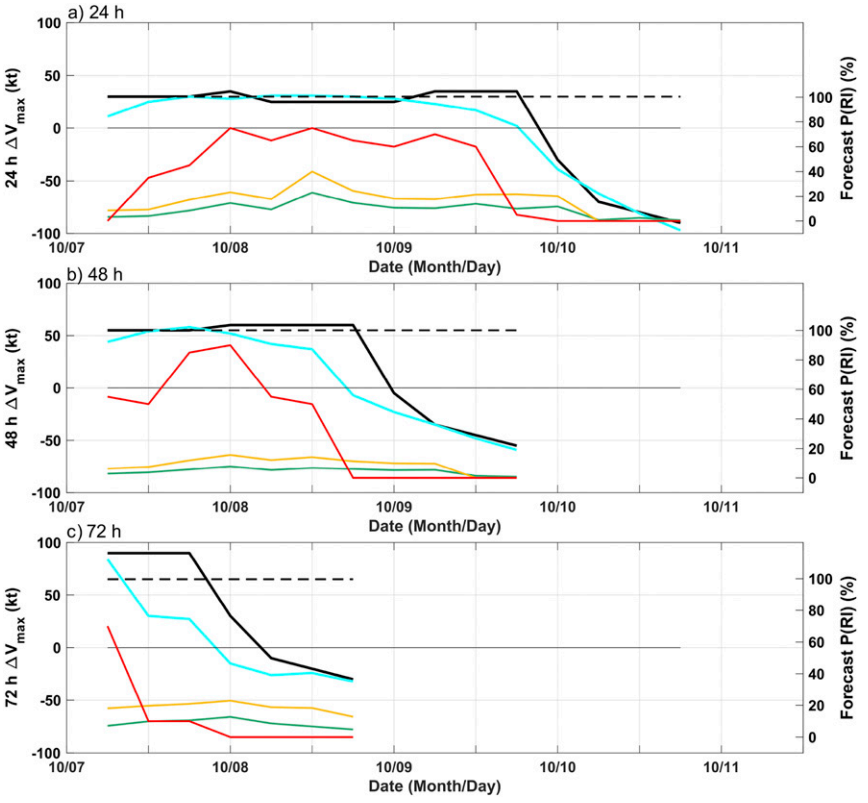

FIG. 4. As in Fig. 2, but for Hurricane Michael (2018).

Unlike the forecast failures that occurred during Florence's first RI, both the deterministic HWRF and the probabilistic models perform rather well for Michael (Fig. 4). For the 24-h lead time, the HWRF forecasts are excellent, matching the observed intensity change almost exactly for forecasts initialized from 1200 UTC 7 September to 0000 UTC 9 September. The RI-AnEn likewise shows probabilities in the 60%–70% range during this same period. SHIPS-RII and SHIPSCON also show elevated probabilities, but they are overall much lower (in the 20%–50% range). Each of the models is somewhat slow in recognizing the onset of RI (forecasts at 1200 UTC 7 September gave RI probabilities $< 40\%$) and is also hasty in forecasting its end, both of which are likely related to the evolving upper-level flow over the Gulf of Mexico. Though ultimately favorable for RI, the precise nature of Michael's interaction with the upper-level trough took a while for the models to resolve; likewise, small errors in the prediction of the steering flow led to the HWRF/Global Forecast System (GFS) forecasting landfall on the Florida Panhandle at a time earlier than what actually verified.

The 48-h forecasts are distinguished by a greater disparity between the RI-AnEn and SHIPS models. The former has greatly elevated probabilities (50%–80%) over most of the RI period, while the latter has only modestly elevated probabilities (<20%). HWRF again performs quite well, although as was the case with the 24-h lead time, poor timing of landfall results in a shorter period of RI than ultimately occurred. At 72 h, only the HWRF and RI-AnEn capture the rapid intensification signal, though, as is the case at 24 and 48 h, the signal is lost for forecasts verifying some 12 h prior to landfall.

c. *Ensemble $\Delta V_{\max}$ forecasts for Florence and Michael*

Since RI-AnEn produces ensemble forecasts of $\Delta V_{\max}$, it is useful to examine these forecast results for Florence and Michael as a supplement to the probabilistic RI results described above. To do so, we first compute binned spread–skill statistics for all 24-, 48-, and 72-h RI-AnEn $\Delta V_{\max}$ ensemble forecasts made as part of the 2017 and 2018 HFIP real-time demonstrations (Fig. 5). For this two-year forecast set, RI-AnEn tends to have poor spread–skill characteristics for ensemble spreads less than 15 kt (in these cases, RI-AnEn tends to be underdispersive). RI-AnEn remains slightly underdispersive for those cases that are intrinsically more dispersive (i.e.,

ensemble spreads > 15 kt), though the degree of this underdispersion is not significant at the 95% level as indicated by the overlap of the associated confidence intervals and the 1:1 line.

Ensemble ΔV_{max} forecasts for Florence (Fig. 6) show that the spread of the RI-AnEn forecasted change in intensity at the 24-, 48-, and 72-h lead times for the first RI event is both small and insufficient to account for the forecast error, consistent with the spread–skill results discussed above. Indeed, for all lead times in the first RI event, the spread of the RI-AnEn is constrained and unable to indicate any possibility of RI, although the RI-AnEn does show some heightened probabilities of greater intensification early on 1 September 2018 prior to the RI of Florence. In the second period of RI, while the median RI-AnEn prediction for the intensity change is too low, the spread indicates some analogs suggested RI would occur when it did. It is notable that the ensemble spread during the second RI period is considerably larger than during the first. This holds true for the 24-, 48-, and 72-h forecasts.

Turning now to Michael, the HWRF and RI-AnEn show premature weakening of Michael leading into 10 October 2018 at all forecast lead times (Fig. 7), due to the fact that the simulated HWRF landfall occurred sooner than was observed. Otherwise, consistent with the probabilistic forecast results discussed in section 3b, the RI-AnEn performs well, particularly for the 24- and 48-h lead times. The RI-AnEn spread straddles the best track data at most forecast times, and while the median RI-AnEn has some RI false alarms at 24 h (i.e., 24-h forecasts from 0600 UTC 8 October to 0000 UTC 9 October), the best track data show a storm intensifying at a rate almost great enough to qualify for RI. Overall, examination of forecast output from the ΔV_{max} ensemble spread perspective sheds light on nuances of the forecast not discernible from the probabilistic RI forecast vantage point. In this respect, RI-AnEn offers some of the same benefits provided by a full-physics dynamical ensemble.

4. Results from the 2017 and 2018 Atlantic and eastern Pacific seasons

Performance of the RI-AnEn relative to the deterministic HWRF and the SHIPS statistical models is first assessed using the BSS as described in section 2. Though not strictly defined for deterministic forecasts, the operational HWRF results are adapted for inclusion in the BSS comparisons by defining for each forecast a probability of 1 or 0 in the case that RI is or is not forecast, respectively.

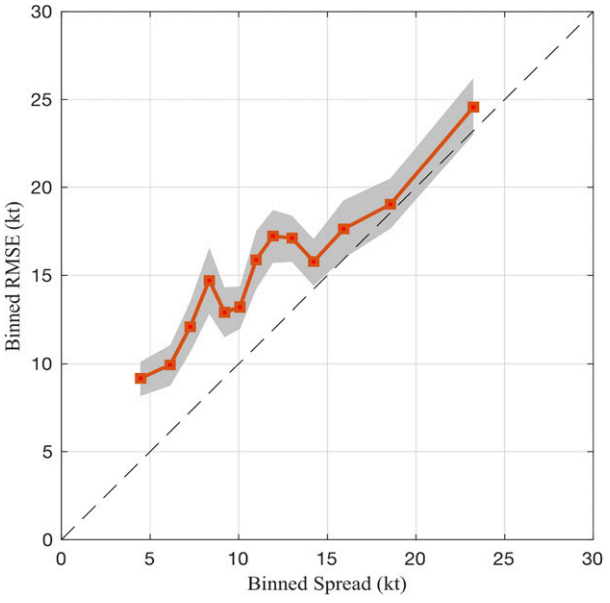

FIG. 5. Binned RI-AnEn spread–skill plot computed for the combined Atlantic and eastern Pacific basin forecast set for the 2017 and 2018 HFIP real-time demonstrations. Ensemble spread was computed for 24-, 48-, and 72-h ensemble forecasts of ΔV_{max} and sorted into 13 bins, each of which contains 172 samples. The ensemble-mean RMSE (i.e., the RMSE computed using the ensemble mean ΔV_{max} as the estimator) corresponding to each set of binned samples was then computed and is plotted on the ordinate. The 95% bootstrap confidence intervals (computed using 1000 replicates) for the RMSE are indicated by the gray shading.

For the Atlantic basin (Fig. 8a), at each lead time, the RI-AnEn is the most skillful of the models considered. The degree of separation is most pronounced at 48 h, where the RI-AnEn BSS is just under 0.5 (i.e., a 50% improvement over a forecast using the climatological probability of RI) and none of the other models provides more than an 8% improvement at any lead time. However, 48 h is the only lead time at which the RI-AnEn is significantly better than any of the other models at the 95th percentile level. This is confirmed by both bootstrap confidence intervals for the pairwise differences in BSS between RI-AnEn and the other models as well as by the DM test. Model performance was generally better in the eastern Pacific basin in 2018 (Fig. 8b). The SHIPS-RII, SHIPSCON, and RI-AnEn models are all skillful at 24 h. At 48 h, the SHIPS-RII, SHIPSCON, and RI-AnEn models continue to be skillful, while only the RI-AnEn model remains skillful at 72 h, and significantly more skillful than the other models analyzed. The deterministic HWRF model is not skillful at any lead time in the eastern Pacific. At no lead times are any of the models significantly better than the RI-AnEn.

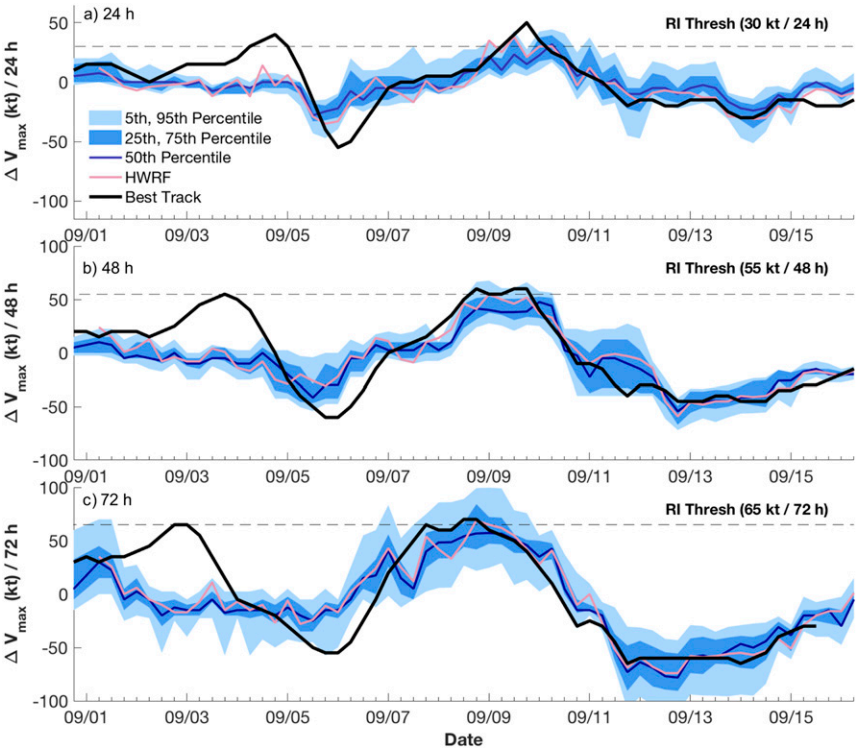

FIG. 6. The 24-, 48-, and 72-h ΔV_{max} predicted by HWRF and the RI-AnEn along with the accompanying verifying observations of intensity change for Hurricane Florence (2018). The RI threshold for each lead time is indicated by the black dashed line and the 5th–95th and 25th–75th percentiles of the RI-AnEn intensity change forecasts are indicated in light and dark blue shading, respectively.

As mentioned previously, the sample sizes in both the Atlantic and eastern Pacific are relatively small. To address this shortcoming, we augment the 2018 forecast set in each basin with corresponding forecast results from the 2017 HFIP real-time demonstration. The results for the Atlantic (Fig. 9a) are similar to those for the 2018 season (Fig. 8a) with the notable exception of the improved skill of the two SHIPS-based models at all lead times. The RI-AnEn is significantly more skillful than the HWRF at 24 h and significantly more skillful than all models at 48 h. The deterministic HWRF again fails to exhibit skill at any lead time. In the eastern Pacific (Fig. 9b), the results are fairly similar to those for 2018 alone (Fig. 8b), due primarily to the small sample size of the 2017 HFIP real-time forecasts in that basin. To achieve the best possible assessment of performance, the 2017–18 forecasts for the Atlantic and eastern Pacific basins are combined into a single set with 907, 735, and 594 forecasts valid at 24, 48,and 72 h, respectively. The BSSs computed for this sample (Fig. 10) show that the probabilistic models are tightly packed at 24 h, each exhibiting significant skill in excess of 20%. The RI-AnEn, however, is the only model we analyzed to exhibit significant skill at all lead times for the combined 2017–18 Atlantic–eastern Pacific sample. In fact, the RI-AnEn performs significantly better than all models at 72-h lead times and significantly better than the HWRF and SHIPSCON at the 48-h lead time in terms of both bootstrap confidence intervals and the Diebold–Mariano test. In both Figs. 9 and 10, none of the models are more skillful than the RI-AnEn at any lead time in a statistically significant sense.

An important component of the Brier score is reliability (the degree to which forecast probability agrees with observed frequency), and results for the 2017–18 Atlantic–eastern Pacific sample are summarized in Fig. 11. At 24 h, each of the models exhibits good reliability for forecast probabilities less than 0.6. Beyond that threshold, the RI-AnEn's tendency to be too aggressive (i.e., overconfident) becomes more pronounced, while the SHIPSCON model's underconfidence is likewise amplified. Of the three models considered here, SHIPS-RII is the most reliable at 24 h. At 48 h, SHIPS-RII and RI-AnEn both perform well, though for the majority of forecasts ($p \leq 0.6$) the SHIPS models tend to be underconfident while the RI-AnEn is once again

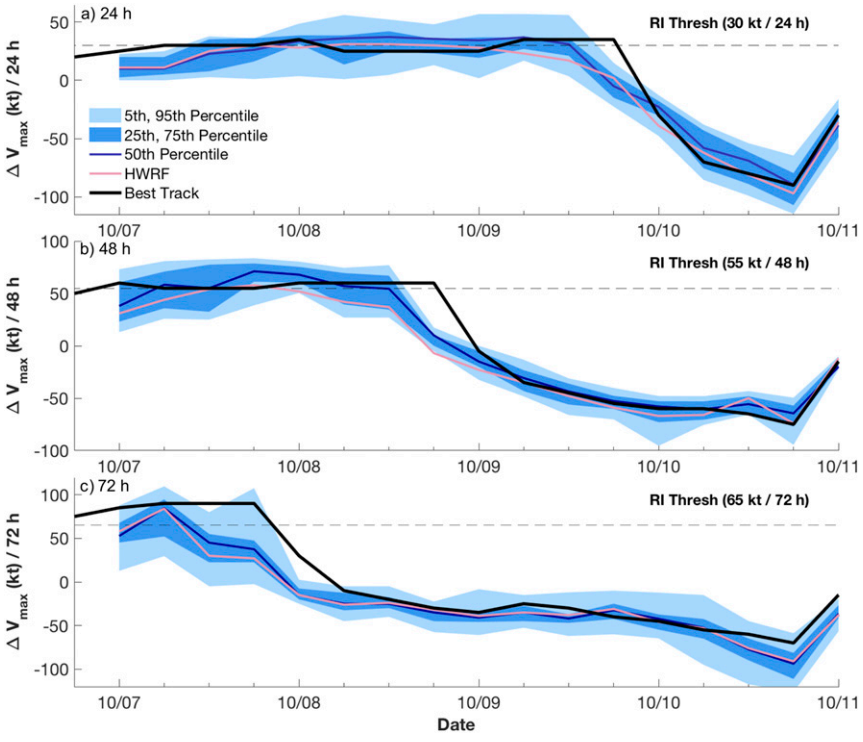

FIG. 7. As in Fig. 6, but for Hurricane Michael (2018).

overconfident. The impact of undersampling begins to become evident for the 72-h forecasts, as there simply are not enough SHIPS-RII or SHIPSCON forecasts at the higher probability thresholds to generate useful reliability information. The RI-AnEn, while once again tending to overstate the likelihood of RI, is relatively reliable for forecast probabilities greater than 0.3. Comparisons of model reliability are limited to a qualitative basis, as the sample sizes at the higher probability thresholds are insufficient to generate representative uncertainty statistics.

Receiver operating characteristic (ROC) curves are plotted for the combined 2017–18 Atlantic–eastern Pacific sample in Fig. 12. Optimal performance of a probabilistic model from this perspective is indicated by a curve which cleaves closely to the left (false alarm rate = 0) and upper (hit rate = 1) axes. In other words, it is advantageous to maximize hit rate and minimize false alarm rate. Doing so results in a larger area under the ROC curve (AUC), the maximum of which can be unity. For the 24-h forecasts the SHIPS models clearly do a much better job than RI-AnEn of distinguishing between RI and non-RI events as evidenced by their larger AUC values (~0.88 for the former versus 0.83 for the latter). This pattern reverses at 48 h, where the RI-AnEn has a slightly larger AUC than either of the SHIPS models, and this remains true at 72 h, where RI-AnEn has an AUC more than 10% greater than SHIPS-RII. SHIPSCON, despite its degraded performance in 72-h forecast reliability relative to RI-AnEn, has an AUC nearly identical to RI-AnEn for the same lead time, which illustrates the importance of using a variety of metrics to evaluate forecasts.

To conclude the bulk assessment of model performance, the ROC curves considered in Fig. 12 are transformed to a standardized skill score using Eq. (4). The resulting ROCSS (Fig. 13) show that at 24 h, the SHIPS models are more skillful than RI-AnEn in distinguishing RI events from non-RI events. At 48 and 72 h, the RI-AnEn is more skillful than either of the SHIPS models.

5. Discussion and conclusions

The HWRF rapid intensification analog ensemble (RI-AnEn) is evaluated for real-time forecasts made during the 2018 HFIP real-time demonstration, and these results are further supplemented with results from the 2017 HFIP real-time demonstration to produce a sample which ensures more robust statistics.

Results obtained for two particular, high-impact cases (Hurricanes Florence and Michael from the 2018 Atlantic season) highlight the strengths and weaknesses of both RI-AnEn and probabilistic guidance in general. Florence,

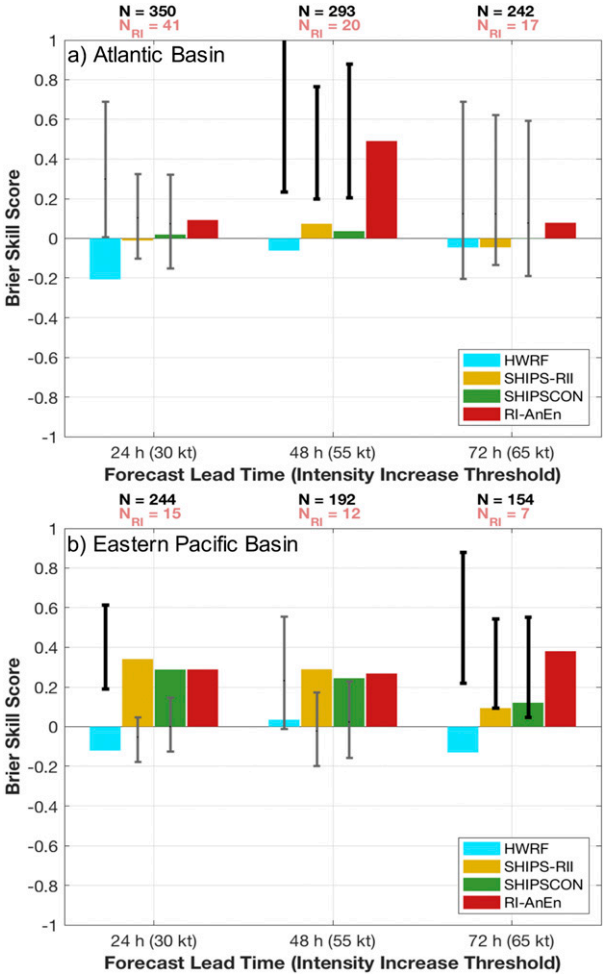

FIG. 8. BSSs for RI forecasts generated by the deterministic HWRF, SHIPS-RII, SHIPSCON and RI-AnEn models during the 2018 HFIP real-time demonstration. BSSs are shown for 24-, 48-, and 72-h lead times for the (a) Atlantic and (b) eastern Pacific basins and include 95% bootstrap confidence intervals (CIs) for pairwise RI-AnEn–HWRF, RI-AnEn–SHIPS-RII, and RI-AnEn–SHIPSCON BSS differences computed using 1000 replicates. In those instances where the Diebold–Mariano test indicates significance, the corresponding BSS CIs are shown in bold. The total number of forecasts N and number of RI events (Nri) are shown in black and magenta, respectively, across the top of each panel.

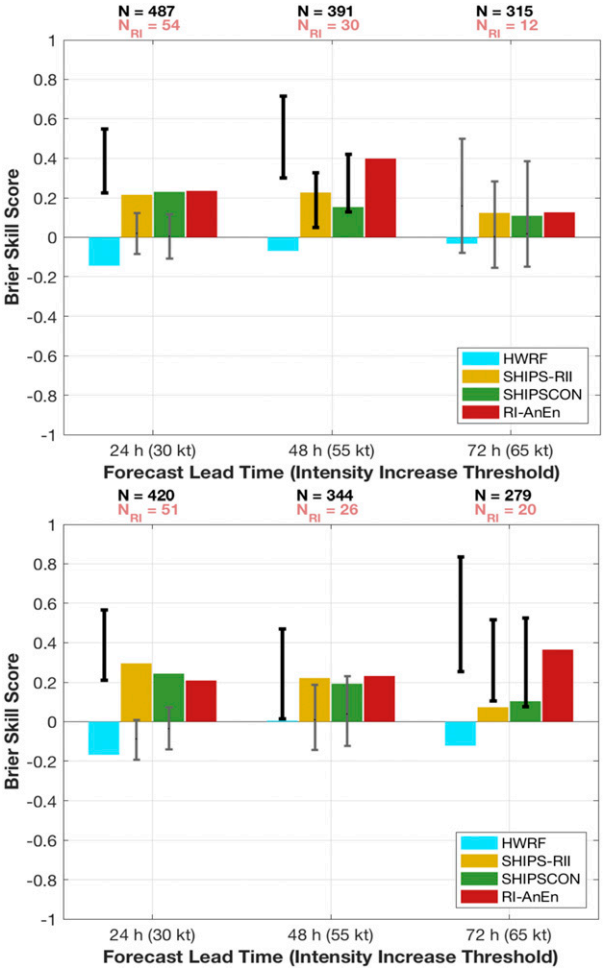

FIG. 9. As in Fig. 8, but for combined forecasts from the 2017 and 2018 HFIP real-time demonstrations.

which underwent RI under marginal (at best) environmental conditions, was poorly forecast, eluding every available forecast aid at the NHC's disposal. RI-AnEn was unable to provide useful guidance in this case, but performed comparably to other forecast aids for Florence's second RI event. Michael, on the other hand, was very well forecast by RI-AnEn. At 24- and 48-h lead times, the RI-AnEn forecast significantly elevated RI probabilities (>50%) for the majority of forecast cycles in which RI was ultimately confirmed. Furthermore, when the dispersion of the ΔV_{max} is analyzed (Figs. 6 and 7), it is shown that much of the same information contained in a full dynamical ensemble is also available from the RI-AnEn. Hence, while the RI probabilities themselves provide rapid access to the predicted likelihood of RI, examination of the median ΔV_{max} and ensemble spread provides an estimate of anticipated intensity change as well as its associated uncertainty. Given the documented spread–skill relationship of RI-AnEn over the entire, augmented sample, ensemble ΔV_{max} forecasts with spreads greater than 15 kt are anticipated to have better dispersion characteristics.

Using a suite of assessment tools (Brier skill scores, reliability diagrams, Diebold–Mariano test, ROC curves, ROC skill scores), the RI-AnEn is demonstrated to be very competitive with current operational deterministic and probabilistic RI forecast aids. From the Brier perspective, the RI-AnEn is the most skillful model at each lead time for Atlantic basin forecasts in 2018 and

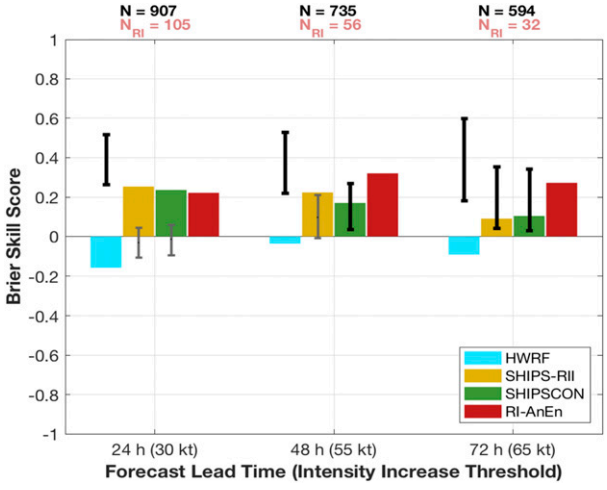

FIG. 10. As in Figs. 8 and 9, but for the combined Atlantic and eastern Pacific basin forecast set for the 2017 and 2018 HFIP real-time demonstrations.

demonstrates skill greater than 20% at each lead time for the eastern Pacific. Note that a model can be more skillful without the difference in skill necessarily being significant. That being said, there are significant differences in skill as indicated by both DM tests and bootstrap confidence intervals (both at the 95% significance level) for pairwise BSS differences. In the Atlantic, RI-AnEn is significantly better than all other guidance at 48 h for both the 2017 and 2018 seasons. In the eastern Pacific, RI-AnEn is significantly more skillful at 72 h for both seasons. DM tests and bootstrap confidence intervals for the combined 2017–18 Atlantic and eastern Pacific forecast sample show that the RI-AnEn significantly outperforms the underlying deterministic HWRF at all lead times and is significantly more skillful than the other probabilistic models at 72 h. Analysis of forecast reliability for this same sample reveals that, though the RI-AnEn tends to be overconfident in its RI forecasts, it is generally well calibrated (i.e., reliable), though sparsely sampled higher-probability events makes quantification of reliability differences among the models difficult even for the larger sample.

ROC curves reveal that the RI-AnEn offers excellent resolution, namely, a combination of sensitivity (maximized for high hit rates) and specificity (maximized for small false-positive rates), performing competitively with respect to the SHIPS models in both basins. These results are supported by ROC skill scores, which indicate that at 48 and 72 h, the RI-AnEn is a more skillful model in terms of resolution. The ROCSS results are not statistically significant and await a larger, multiyear sample to more firmly establish the RI-AnEn's performance characteristics relative to other operational guidance. Nevertheless, the totality of evidence (BSS, reliability, DM, ROC, and ROCSS) suggests that RI-AnEn is a highly competitive model with very attractive advantages for longer-term (48 and 72 h) RI forecasts in both the Atlantic and eastern Pacific basins.

While the RI-AnEn performed well for both the 2018 and combined 2017–18 forecast samples, important challenges remain. In regard to operational applicability, the version of RI-AnEn evaluated here is derived using the late version HWRF output, meaning that a new version (using early HWRF output) would need to be derived for operational use. While this would likely lead to at least some changes in performance, it is worth noting that a neural network prediction tool Cloud et al. (2019) developed with the same HWRF predictor dataset showed that the early version of their model performed nearly as well as a late version when applied to RI prediction. It is also important to note that the relatively small size of the HWRF retrospective training set (only several years of forecasts) likely places an upper bound on current levels of performance. Insufficient sampling of the model climatology is highlighted by large forecast errors related to ''outlier'' events such as Florence's first RI. For this reason, extending the training set in various ways (including developing a lengthier reforecast dataset or developing a multimodel RI-AnEn) and incorporating more sophisticated machine learning tools will be of utmost importance for future performance, especially if cases such as 2018's Hurricane Florence are to be better forecast. Additional work also remains regarding the relationship between RI-AnEn performance and various aspects of the atmospheric and oceanic environment. Identification of biases with strong correlations to environmental conditions would represent actionable information for forecasters and future model developers alike.

It should be stressed that the RI-AnEn in particular, and the AnEn in general, is not intended as a replacement for any particular guidance product. Indeed, by its very nature, the AnEn is an adjunct of the dynamical model upon which it is built. This dependence serves to illustrate an important point, namely that the AnEn is a flexible tool that offers certain performance advantages over the parent deterministic model (as demonstrated here) at essentially zero runtime cost, meaning that the AnEn can be extended to other dynamical models (global as well as regional) with relative ease. As global models steadily increase in resolution and sophistication in the coming years, it is reasonable to expect that they will be relied upon for TC intensity

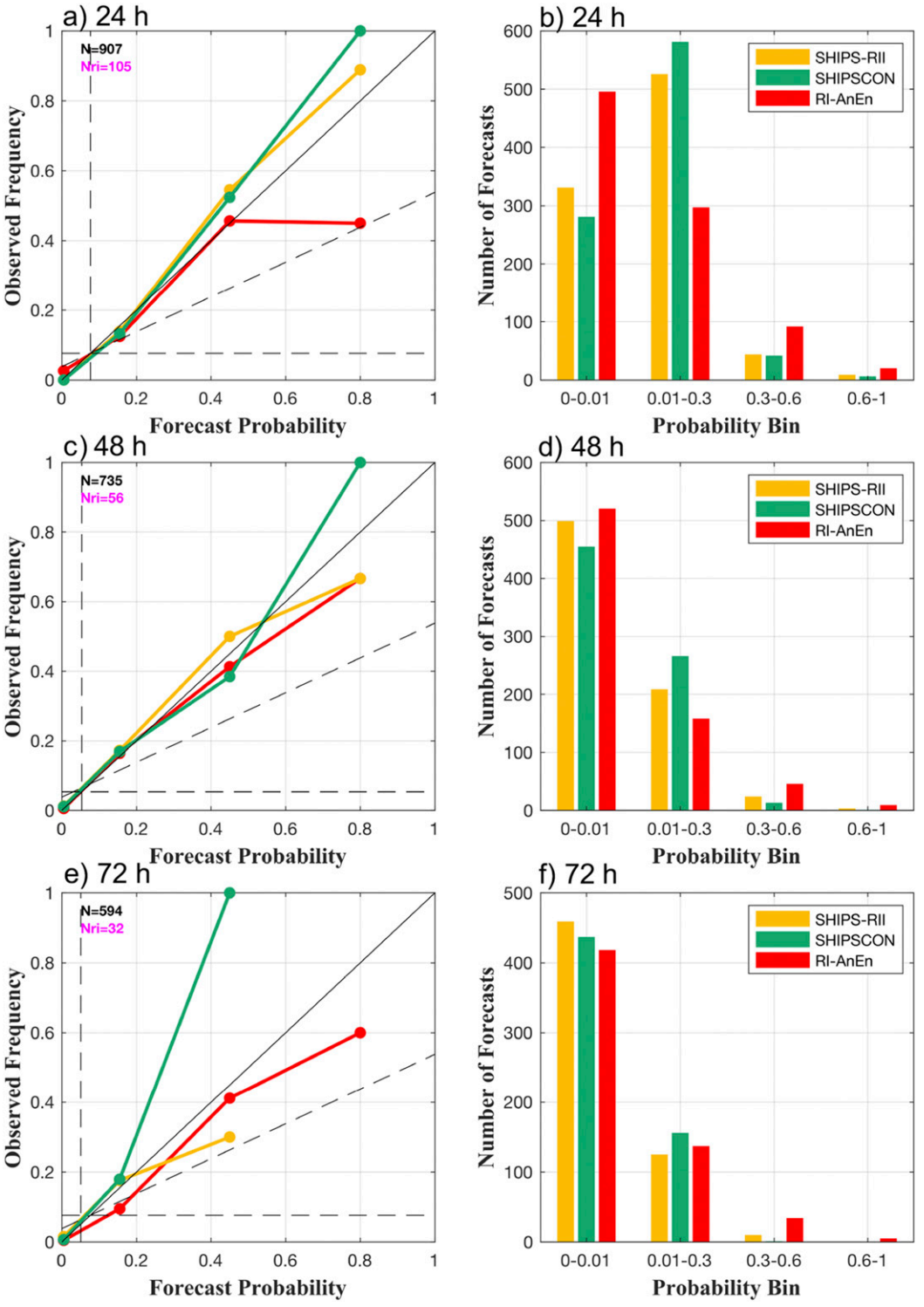

FIG. 11. Reliability diagrams for (a) 24-, (c) 48-, and (e) 72-h probabilistic RI forecasts made for the combined Atlantic and eastern Pacific forecast set for the 2017 and 2018 HFIP real-time demonstrations. The corresponding number of forecasts in each probability bin are shown for (b) 24, (d) 48, and (f) 72 h. The sample sizes (*N* and Nri) for each forecast lead time are shown in the upper right of (a), (c), and (e). The solid diagonal black line indicates perfect agreement between forecast probability and observed frequency. The horizontal dashed line depicts the baseline climatological RI probability for the given lead time, in this case given by the weighted average of the climatological RI probabilities in the Atlantic and Eastern Pacific as determined from HURDAT for the years 1987–2017. Plot points lying between the diagonal and vertical dashed lines indicate that forecasts in that probability bin are skillful.

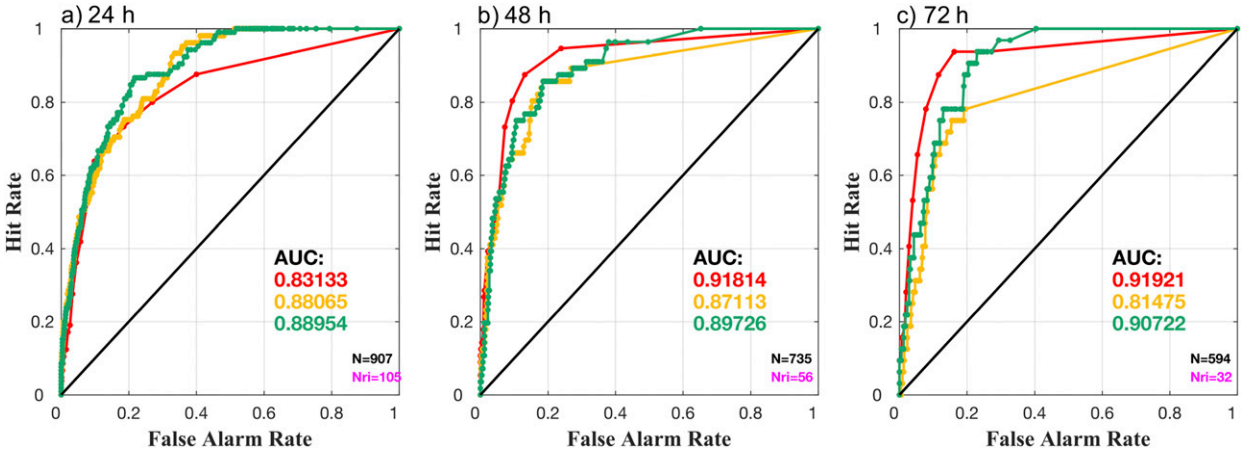

FIG. 12. ROC curves for the combined Atlantic and eastern Pacific forecast set for the 2017 and 2018 HFIP real-time demonstrations. ROC curves are shown for probabilistic RI forecasts made for (a) 24, (b) 48, and (c) 72 h for the RI-AnEn (red), SHIPS-RII (orange), and SHIPSCON (green) models. The area under the ROC curve (AUC) is shown for each model in the lower right of each panel with the corresponding sample sizes shown in the extreme lower right. Hit and false alarm rates are plotted at probability thresholds ranging from 0 to 1 using increments of 0.05 for RI-AnEn and 0.01 for SHIPS-RII and SHIPSCON.

forecasts with increasing frequency. As this transition occurs, AnEn models can be developed which provide well-calibrated ensembles of intensity change, thereby extending the value of the deterministic predictions and offering forecasters valuable insight into the associated uncertainty. Finally, it is also worth mentioning that AnEn models need not be restricted to predictions of TC intensity (as in Alessandrini et al. 2018) or intensity change as explored in this study. The AnEn can be extended to the prediction of TC track, structure, precipitation, and most any other field or parameter of importance in the TC prediction and warning process.

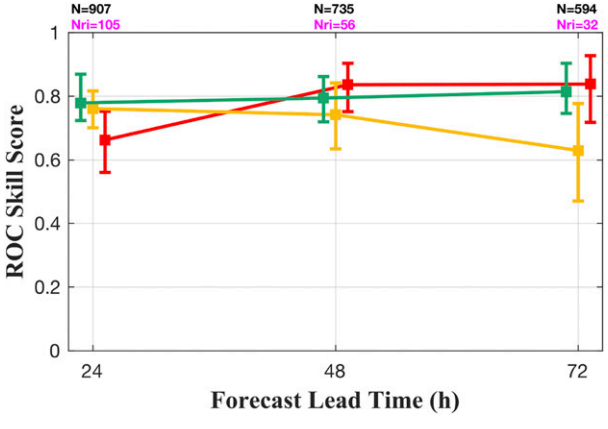

FIG. 13. ROCSS for the combined Atlantic and eastern Pacific forecast set for the 2017 and 2018 HFIP real-time demonstrations for the RI-AnEn (red), SHIPS-RII (orange), and SHIPSCON (green) models. The 95% bootstrap confidence intervals are shown for each lead time, and the corresponding sample sizes (*N* and Nri) are indicated along the top of the panel.

Acknowledgments. The authors wish to acknowledge the support of the National Oceanic and Atmospheric Administration (NOAA) and the Hurricane Forecast Improvement Program (HFIP) via NOAA HFIP Award NA16NWS4680027 and extensive use of the Jet supercomputer, as well as numerous fruitful discussions with fellow HFIP participants. The authors also extend their gratitude to John Kaplan as well as two anonymous reviewers for comments and insights that have substantially improved the quality of the manuscript.

REFERENCES

Alessandrini, S., L. D. Monache, C. M. Rozoff, and W. E. Lewis, 2018: Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Mon. Wea. Rev.*, **146**, 1723–1744, https://doi.org/10.1175/MWR-D-17-0314.1.

Beven, J. L., II, R. Berg, and A. Hagen, 2019: National Hurricane Center Tropical Cyclone Report: Hurricane Michael (7–11 October 2018). NOAA/NHC Tech. Rep. AL142018, NOAA/National Hurricane Center, 86 pp., https://www.nhc.noaa.gov/data/tcr/AL142018_Michael.pdf.

Biswas, M. K., and Coauthors, 2018: Hurricane Weather Research and Forecasting (HWRF) Model: 2018 Scientific Documentation. Accessed 15 January 2019, https://dtcenter.org/HurrWRF/users/docs/index.php.

Boer, G. J., 1994: Predictability regimes in atmospheric flow. *Mon. Wea. Rev.*, **122**, 2285–2295, https://doi.org/10.1175/1520-0493(1994)122<2285:PRIAF>2.0.CO;2.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Cloud, K. A., B. J. Reich, C. M. Rozoff, S. Alessandrini, W. E. Lewis, and L. Delle Monache, 2019: A feed forward neural network based on model output statistics for short-term hurricane intensity prediction. *Wea. Forecasting*, **34**, 985–997, https://doi.org/10.1175/WAF-D-18-0173.1.

Davison, A. C., and D. V. Hinkley, 1997: *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 592 pp.

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., 2009: A simplified dynamical system for tropical cyclone intensity prediction. *Mon. Wea. Rev.*, **137**, 68–82, https://doi.org/10.1175/2008MWR2513.1.

——, M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements in the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543, https://doi.org/10.1175/WAF862.1.

——, C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398, https://doi.org/10.1175/BAMS-D-12-00240.1.

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263.

Finocchio, P. M., and S. J. Majumdar, 2017: The predictability of idealized tropical cyclones in environments with time-varying vertical wind shear. *J. Adv. Model. Earth Syst.*, **9**, 2836–2862, https://doi.org/10.1002/2017MS001168.

Gall, R., J. Franklin, F. Marks, E. N. Rappaport, and F. Toepfer, 2013: The Hurricane Forecast Improvement Project. *Bull. Amer. Meteor. Soc.*, **94**, 329–343, https://doi.org/10.1175/BAMS-D-12-00071.1.

Ghosh, T., and T. N. Krishnamurti, 2018: Improvements in hurricane intensity forecasts from a multimodel superensemble utilizing a generalized neural network technique. *Wea. Forecasting*, **33**, 873–885, https://doi.org/10.1175/WAF-D-17-0006.1.

Goerss, J. S., and C. R. Sampson, 2014: Prediction of consensus tropical cyclone intensity forecast error. *Wea. Forecasting*, **29**, 750–762, https://doi.org/10.1175/WAF-D-13-00058.1.

Hakim, G. J., 2013: The variability and predictability of axisymmetric hurricanes in statistical equilibrium. *J. Atmos. Sci.*, **70**, 993–1005, https://doi.org/10.1175/JAS-D-12-0188.1.

Hartmann, H. C., T. C. Pagano, S. Sorooshiam, and R. Bales, 2002: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.*, **83**, 683–698, https://doi.org/10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2.

Judt, F., 2018: Insights into atmospheric predictability through global convection-permitting model simulations. *J. Atmos. Sci.*, **75**, 1477–1497, https://doi.org/10.1175/JAS-D-17-0343.1.

——, and S. S. Chen, 2016: Predictability and dynamics of tropical cyclone rapid intensification deduced from high-resolution stochastic ensembles. *Mon. Wea. Rev.*, **144**, 4395–4420, https://doi.org/10.1175/MWR-D-15-0413.1.

Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting*, **18**, 1093–1108, https://doi.org/10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2.

——, and Coauthors, 2015: Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Wea. Forecasting*, **30**, 1374–1396, https://doi.org/10.1175/WAF-D-15-0032.1.

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved skills for weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, https://doi.org/10.1126/science.285.5433.1548.

Liu, B., and Coauthors, 2018: Verification of 2018 HWRF and HMON performance. *HFIP Annual Review Meeting*, Miami, FL, NOAA, http://www.hfip.org/events/annual_meeting_nov_2018/.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2.

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.

Onderlinde, M., and M. DeMaria, 2018: Deterministic to Probabilistic Statistical rapid intensification index (DTOPS): A new method for forecasting RI probability. *33rd Conf. on Hurricanes and Tropical Meteorology*, Ponte Vedra, FL, Amer. Meteor. Soc., 16C.3, https://ams.confex.com/ams/33HURRICANE/webprogram/Paper339346.html.

Ryglicki, D., J. D. Doyle, D. Hodyss, J. H. Cossuth, Y. Jin, K. C. Viner, and J. M. Schmidt, 2019: The unexpected intensification of tropical cyclones in moderate vertical wind shear. Part III: Outflow–environment interaction. *Mon. Wea. Rev.*, **147**, 2919–2940, https://doi.org/10.1175/MWR-D-18-0370.1.

Sampson, C. R., J. L. Franklin, J. A. Knaff, and M. DeMaria, 2008: Experiments with a simple tropical cyclone intensity consensus. *Wea. Forecasting*, **23**, 304–312, https://doi.org/10.1175/2007WAF2007028.1.

Simon, A., A. B. Penny, M. DeMaria, J. L. Franklin, R. J. Pasch, E. N. Rappaport, and D. A. Zelinsky, 2018: A description of the real-time HFIP Corrected Consensus Approach (HCCA) for tropical cyclone track and intensity guidance. *Wea. Forecasting*, **33**, 37–57, https://doi.org/10.1175/WAF-D-17-0068.1.

Sperati, S., S. Alessandrini, P. Pinson, and G. Kariniotakis, 2015: The "Weather Intelligence for Renewable Energies" benchmarking exercise on short-term forecasting of wind and solar power generation. *Energies*, **8**, 9594–9619, https://doi.org/10.3390/en8099594.

Stewart, S. R., and R. Berg, 2019: National Hurricane Center Tropical Cyclone Report: Hurricane Florence (31 August–17 September 2018). NOAA/NHC Tech. Rep. AL062018, NOAA/National Hurricane Center, 98 pp., https://www.nhc.noaa.gov/data/tcr/AL062018_Florence.pdf.

Wang, W., J. A. Sippel, S. Abarca, L. Zhu, B. Liu, Z. Zhang, A. Mehra, and V. Tallapragada, 2018: Improving NCEP HWRF simulations of surface wind and inflow angle in the eyewall area. *Wea. Forecasting*, **33**, 887–898, https://doi.org/10.1175/WAF-D-17-0115.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Williford, C. E., T. N. Krishnamurti, R. C. Torres, S. Cocke, Z. Christidis, and T. S. Vijaya Kumar, 2003: Real-time multimodel superensemble forecasts of Atlantic tropical systems of 1999. *Mon. Wea. Rev.*, **131**, 1878–1894, https://doi.org/10.1175//2571.1.

Zhang, F., and D. Tao, 2013: Effects of vertical wind shear on the predictability of tropical cyclones. *J. Atmos. Sci.*, **70**, 975–983, https://doi.org/10.1175/JAS-D-12-0133.1.

——, N. Bei, R. Rotunno, C. Snyder, and C. C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594, https://doi.org/10.1175/JAS4028.1.