atmosphere

MDPI

Article

Validation of Ensemble-Based Probabilistic Tropical Cyclone Intensity Change

Ryan D. Torn [1,*,†,‡] and Mark DeMaria [2,‡]

[1] Department of Atmospheric and Environmental Sciences, University at Albany, State University of New York, Albany, NY 12222, USA
[2] Cooperative Institute for Research in the Atmosphere, Colorado State University/CIRA, Fort Collins, CO 80521, USA; mark.demaria@noaa.gov
* Correspondence: rtorn@albany.edu
† Current address: ES 351, 1400 Washington Avenue, Albany, NY 12222, USA.
‡ These authors contributed equally to this work.

Abstract: Although there has been substantial improvement to numerical weather prediction models, accurate predictions of tropical cyclone rapid intensification (RI) remain elusive. The processes that govern RI, such as convection, may be inherently less predictable; therefore a probabilistic approach should be adopted. Although there have been numerous studies that have evaluated probabilistic intensity (i.e., maximum wind speed) forecasts from high resolution models, or statistical RI predictions, there has not been a comprehensive analysis of high-resolution ensemble predictions of various intensity change thresholds. Here, ensemble-based probabilities of various intensity changes are computed from experimental Hurricane Weather Research and Forecasting (HWRF) and Hurricanes in a Multi-scale Ocean-coupled Non-hydrostatic (HMON) models that were run for select cases during the 2017–2019 seasons and verified against best track data. Both the HWRF and HMON ensemble systems simulate intensity changes consistent with RI (30 knots 24 h^{-1}; 15.4 m s^{-1} 24 h^{-1}) less frequent than observed, do not provide reliable probabilistic predictions, and are less skillful probabilistic forecasts relative to the Statistical Hurricane Intensity Prediction System Rapid Intensification Index (SHIPS-RII) and Deterministic to Probabilistic Statistical (DTOPS) statistical-dynamical systems. This issue is partly alleviated by applying a quantile-based bias correction scheme that preferentially adjusts the model-based intensity change at the upper-end of intensity changes. While such an approach works well for high-resolution models, this bias correction strategy does not substantially improve ECMWF ensemble-based probabilistic predictions. By contrast, both the HWRF and HMON systems provide generally reliable predictions of intensity changes for cases where RI does not take place. Combining the members from the HWRF and HMON ensemble systems into a large multi-model ensemble does not improve upon HMON probablistic forecasts.

Keywords: tropical cyclones; intensity change; ensemble forecasting

Citation: Torn, R.D.; DeMaria, M. Validation of Ensemble-Based Probabilistic Tropical Cyclone Intensity Change. *Atmosphere* **2021**, *12*, 373. https://doi.org/10.3390/atmos12030373

Academic Editor: Sundararaman Gopalakrishnan

Received: 12 February 2021
Accepted: 9 March 2021
Published: 12 March 2021

1. Introduction

Despite numerous advances in numerical weather prediction and physical understanding, the prediction of tropical cyclone (TC) intensity change remains a challenging problem, e.g., [1]. This issue is particularly acute for rapid intensification (RI), which is often defined as the 95th percentile of the climatological intensity change [typically defined as 30 kt 24 h^{-1} or 15.4 m s^{-1} 24 h^{-1}] [2]. Furthermore, TCs that undergo RI just prior to landfall, such as Hurricanes Harvey [2017] [3], Maria [2017] [4] and Michael [2018] [5] pose an especially important societal challenge because of the threat to lives and property along coastlines. As a consequence, it is critical to provide timely and accurate estimates of rapid intensity change, which in turn will inform other hazard predictions, such as wind, freshwater flooding due to rain, landslides, and storm surge.

There are numerous potential reasons why RI might be especially difficult to predict, which is likely at least partially related to the interaction of multiple scales of motion that may be responsible for RI, e.g., [6]. On the synoptic scale, it is generally recognized that RI is more likely to take place in favorable environments, which includes low vertical wind shear, high mid-tropospheric moisture content, and large difference between the TC intensity and the maximum potential intensity, e.g., [2,7,8]. Progressing to the mesoscale and convective scale, the structure and extent of convection in the TC, particularly with respect to the shear vector appears to play a critical role. In particular, TCs with faster intensification rates are characterized by greater areal extent of convection within the upshear-left quadrant, e.g., [9–12], which is likely related to more efficient projection of latent heat onto the symmetric component of the TC circulation, e.g., [13,14]. As a consequence, it seems important to accurately represent the interaction of all of these scales of motion. Furthermore, since RI is at least partially tied to convection, a probabilistic approach is likely more appropriate since convection is generally considered less predictable, although deterministic forecasts are made by operational centers.

The distinct preference for RI to take place in specific synoptic environments has resulted in a number of statistical models that try to estimate the probability of RI. The first of these models was the Statistical Hurricane Intensity Prediction System (SHIPS) Rapid Intensification Index (RII), which made use of a combination of large-scale predictors, such as the vertical wind shear, ocean heat content, 850–700 hPa water vapor, and the distance that the TC is away from its maximum potential intensity (MPI), as well as statistics of the infrared brightness temperature from geostationary satellite imagery [8]. A more recent version of this model includes predictors that measure the water vapor in the upshear quadrant and the EOF of the infrared brightness [6]. This approach has also been used to produce probabilistic forecasts in other ocean basins with predictors that vary by basin, e.g., [15]. Such forecasts were shown to provide skillful predictions of RI relative to climatology out to 48 h. Other statistical-based approaches have also emerged, which use input from numerical weather prediction models, including logistical regression-based methods that convert from deterministic to probabilistic guidance [DTOPS] [16], analog-based methods that constructs probabilistic forecasts based on predictors that look for analogs based on past forecasts, e.g., [17], and neural network approaches, e.g., [18].

Given that RI likely involves the interaction of multiple scales of motion, the most accurate probabilistic RI forecasts are likely to come from dynamical model ensemble prediction systems, which can simulate the evolution of both the large-scale environment and convection. Such convection-allowing (<4 km horizontal grid spacing) ensemble prediction systems are very computationally expensive; therefore, these systems would need to provide skillful forecasts to justify their cost. Despite the potential of these systems, many of the prototype ensemble prediction systems which have been run over a variety of cases, are characterized by a lack of variability in TC intensity relative to the ensemble-mean errors, e.g., [19–21]. There are numerous reasons for this result, including systematic biases, particularly within the parameterization schemes, and insufficient treatment of all sources of forecast uncertainty, including in the atmosphere, ocean, and physical parameterizations. In particular, adding more sources of uncertainty leads to more skillful ensemble intensity forecasts, e.g., [20,22].

Many of these previous studies of dynamical ensemble prediction systems have focused on verification of the instantaneous TC intensity; however, few have investigated whether these models have skill at predicting intensity changes. Furthermore, many models show difficulty at replicating intensity changes of the same magnitude as RI, e.g., [23–26]. As a consequence, it is difficult to know whether these dynamical models have any skill at predicting RI or any other intensity change category. The goal of this study is to validate a large number of probabilistic TC intensity change forecasts, which are derived from the quasi-operational Hurricane Weather Research and Forecasting (HWRF) and Hurricanes in a Multi-scale Ocean-coupled Non-hydrostatic (HMON) ensemble prediction systems, which were run as part of the Hurricane Forecast Improvement Project (HFIP)

demonstration system [27] for select TCs during the 2017–2019 seasons. The goal of HFIP is to reduce TC track and intensity errors by 50% over 10 years. In addition, these forecast products were produced as part of the HFIP Ensemble Products Tiger Team, whose goal was to assess new ensemble-based forecast products that could be used by TC forecasters. These probabilistic forecasts are compared to operational RI predictions to understand their relative skill, and to probabilistic forecasts that represent a combination of the ensemble members from both the HWRF and HMON prediction system.

The remainder of the paper proceeds as follows. Section 2 describes the model output and methods used in this particular study. The role of bias correction on the dynamical ensemble output and its validation against the statistical model guidance is provided in Section 3. Finally, a summary and conclusions are provided in Section 4.

2. Data and Methods

This study makes use of the quasi-operational ensemble forecast products that were produced as part of the HFIP demonstration system, as well as comparable-time operational RI forecasts from the 2017–2019 seasons. The main focus of this study is on ensemble forecasts from the HWRF and HMON models. Table 1 provides the list of storms and number of cycles per storm that are used in this study. For computational considerations, these ensemble systems were only run for a subset of potential cycles (489 initialization times), and primarily consists of Atlantic Basin storms.

Table 1. TC cases used in this study. The values inside the parenthesis note the TC identification number, while the second is the number of initialization times for each storm (489 total initialization times).

2017	2018	2019
Emily (AL06; 3)	Florence (AL06; 58)	Barry (AL02; 16)
Franklin (AL07; 14)	Gordon (AL07; 5)	Dorian (AL05; 56)
Gert (AL08; 14)	Isaac (AL09; 29)	Gabrielle (AL08; 13)
Harvey (AL09; 16)	Kirk (AL12; 16)	Humberto (AL09; 23)
Irma (AL11; 45)	Leslie (AL13; 8)	Jerry (AL10; 29)
Jose (AL12; 16)	Michael (AL14; 18)	Lorenzo (AL13; 32)
Maria (AL15; 37)	Hector (EP10; 8)	Nestor (AL16; 3)
Nate (AL16; 10)	Lane (EP14; 10)	Erick (EP06; 10)

When the SHIPS-RI guidance was first developed, RI was defined as a 30 kt or greater increase in 24 h, which roughly corresponds to the 95th percentile of the Atlantic basin 24 h intensity change distribution [2]. That definition was based on forecaster feedback, and the method only considered the first 24 h of the forecast. That RI threshold has since become a common RI definition used in a number of studies not directly tied to the original intent of a forecaster tool. The probabilistic intensity change forecasts from the two ensemble systems will be validated for each 24 h period starting from 0 to 24 and ending at 96–120 h for the RI category. Three other intensification categories will also be evaluated, as shown in Table 2. The steady category includes intensity changes that are below the clearly-detectable intensity change threshold based on current observational capabilities, e.g., [28,29]. A 5th category was originally included corresponding to the 5% of the climatological CDF (rapid weakening), but that was combined with the 4th category in Table 2 (intensity changes < -10 kt) due to the small sample size. As forecasters gained experience with the SHIPS-RII guidance, the 0–24 h time frame was considered too restrictive, so additional thresholds were added in later versions of the method [6]. Those defined RI over longer time intervals, such as 55 kt increase in 48 h and a 65 kt increase in 72 h. Those definitions will also be included in the ensemble validation so they can be compared with the statistical-dynamical RI model forecasts.

All forecasts are verified against the corresponding National Hurricane Center (NHC) best track values. In order to limit the impact of land, only forecast lead times where the

best track TC position is at least 20 km from land are considered. Except where noted, this study employs the "late" version (i.e., non-interpolated) version of the HWRF and HMON ensemble forecasts, which in turn could potentially identify any spin-up issues in the model and remove the complication of applying an interpolation scheme, except when necessary. TC forecast centers, such as NHC, use "early" versions of the model, which are employed because numerical models are not available until after the forecasters are required to make a forecast. For example, these ensemble forecasts were not available until 9 h after the forecast initialization time (i.e., a forecast initialized at 0000 UTC is available at 0900 UTC). To account for this, "early" versions are produced, which adjust the older 12 h intensity forecast to the analyzed value, while the same 12 h intensity difference is added to all subsequent forecast times. A summary of each of the models is provided below.

Table 2. Intensity change categories used in this study.

Name	24-h Maximum Wind Speed Change	Climatological Frequency
Rapid Intensification	δI >= 30 kts	5%
Intensification	10 kts <= δI < 30 kts	29%
Steady	−10 kts < δI < 10 kts	39%
Weakening	δI <= −10 kts	27%

2.1. HWRF Ensemble

The HWRF ensemble is a system that includes uncertainty in the large-scale environment, TC vortex, and model through a variety of methods that are mostly documented in [19]. This 21-member ensemble forecasting system (20 perturbed initial condition and 1 control member) uses much of the same configuration as the deterministic HWRF system for that season, but it was run at a slightly larger grid spacing for computational considerations. A brief description of the ensemble system is provided; the interested reader is directed to the HWRF scientific document for specific information on the model, including the physics parameters, vortex initialization, etc. [30]. While the ensemble system employs the same physics and ocean model as the deterministic model, it was run at 3 km horizontal grid spacing with 61 vertical levels (compared to 2 km horizontal grid spacing and 75 vertical in the deterministic model), with the same domain sizes (75° × 75° outer domain, 8° × 8° highest resolution inner domain). Large scale initial condition and lateral boundary condition variability is obtained by pairing each member of the HWRF ensemble with a corresponding 0.5° Global Forecast System (GFS) ensemble forecast member. Vortex-scale initial condition variability is achieved through perturbing the operational estimates of TC position and intensity (i.e., TCVitals) that are used in the vortex relocation process. The perturbation TC position and maximum wind speed values are obtained by sampling from a Gaussian distribution with the mean as the operational value and the standard deviation taken from [28]. Furthermore, each ensemble member's ocean initial state is perturbed in the manner described in [20], which uses scaled deviations from climatology. Finally, model error is represented by adding white noise stochastic perturbations to the momentum drag coefficient, boundary layer height, and convective trigger.

2.2. HMON Ensemble

The HMON ensemble system is designed using a similar setup as the HWRF ensemble system, with the exception being in the model error strategy. A description of this ensemble system is provided below; further documentation of HMON model is available from [31]. While the HMON ensemble has the same horizontal resolution and grid sizes as the deterministic HMON model, all three domains of the ensemble configuration have 8% fewer grid points compared to the deterministic version. Similar to the HWRF ensemble, initial and boundary conditions for the 11 members (10 perturbed initial condition and one control member) are obtained from the GFS Ensemble Prediction System (GEFS), with vortex-scale uncertainty using perturbed TC position and intensity. By contrast, each

member of the HMON ensemble uses a different combination of physics parameterization packages, which in turn means that this system is a multi-model ensemble. In particular, the HMON members use three different cumulus schemes, and two different turbulent mixing, land surface, microphysics, and surface/enthalpy exchange coefficient formulations, while the control member (member 0) has the same physics configuration as the deterministic HMON.

2.3. ECMWF Ensemble

In addition to evaluating HWRF and HMON ensemble output, similar methods are applied to output from the European Centre for Medium Range Weather Forecasting (ECMWF) ensemble produced during the 2019 season. At this time, the ECMWF ensemble [32] has 51 ensemble members (50 perturbed initial condition and one control member) initialized at 0000 and 1200 UTC. Each member of the ensemble has 18 km horizontal resolution and 91 vertical levels. Initial conditions perturbations are generated based on singular vectors, which represent the fastest growing error structures over a 48-h period on a hemispheric scale and around the TC itself, e.g., [22]. Subgrid-scale model errors are represented using the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme.

3. Results

3.1. Bias Correction

Previous studies have suggested that numerical models can have difficulty simulating rapid intensification, e.g., [23–26]. As a consequence, this could make it difficult to obtain skillful probabilistic RI guidance. In order to determine each model's ability to replicate observed intensification rates, the cumulative distribution function (CDF) of 24 h intensity changes was computed from retrospective forecasts of each modeling system during each 24 h period (i.e., 0–24 h, 6–30 h, 12–36 h), as well as the corresponding best track intensity changes. This is done to provide an independent validation of the model's intensity change rates, which will be used for a bias correction method that will be described later. Figure 1 shows the 0–24 h HWRF and HMON intensity changes computed from retrospective forecasts generated prior to the 2017 season, which includes Atlantic and Eastern Pacific storms from 2014 to 2016 (over water times only). For HWRF, the 0–24 h intensity change CDF generally matches the best track values up to the 70th percentile, after which the HWRF CDF is steeper than the best track CDF. As a consequence, the HWRF 95th percentile value (the RI threshold) is 28 knots 24 h^{-1}, compared to 33 knots 24 h^{-1} from best track. This result suggests that the HWRF model does a fairly good job replicating the observed intensity changes up to roughly 14 knots 24 h^{-1}, but beyond that, the model appears to systematically under-estimate intensity changes. By contrast, HMON model (Figure 1b) appears to overlap the best track CDF for most intensity changes for the same period of time, suggesting that this model does a better job at replicating the variability of observed intensity changes.

At later lead times, both models tend to systematically produce too narrow of a distribution of intensity change values. For the HWRF model (Figure 1c), the model and best track CDFs are similar up to the 65th percentile, but the model CDF curve is steeper both above and below this value, such that the HWRF does not replicate the frequency of weakening storms (HWRF 10th percentile −20 kts 24 h^{-1} vs. −25 kts 24 h^{-1} for best track) and strengthening storms (the HWRF 95th percentile is 23 kts 24 h^{-1} vs. 32 kts 24 h^{-1} for best track). Similar results are obtained for the HMON model during the same lead times (Figure 1d); therefore, it appears that both models have a limited ability to replicate the observed intensity change variability after running for some period of time.

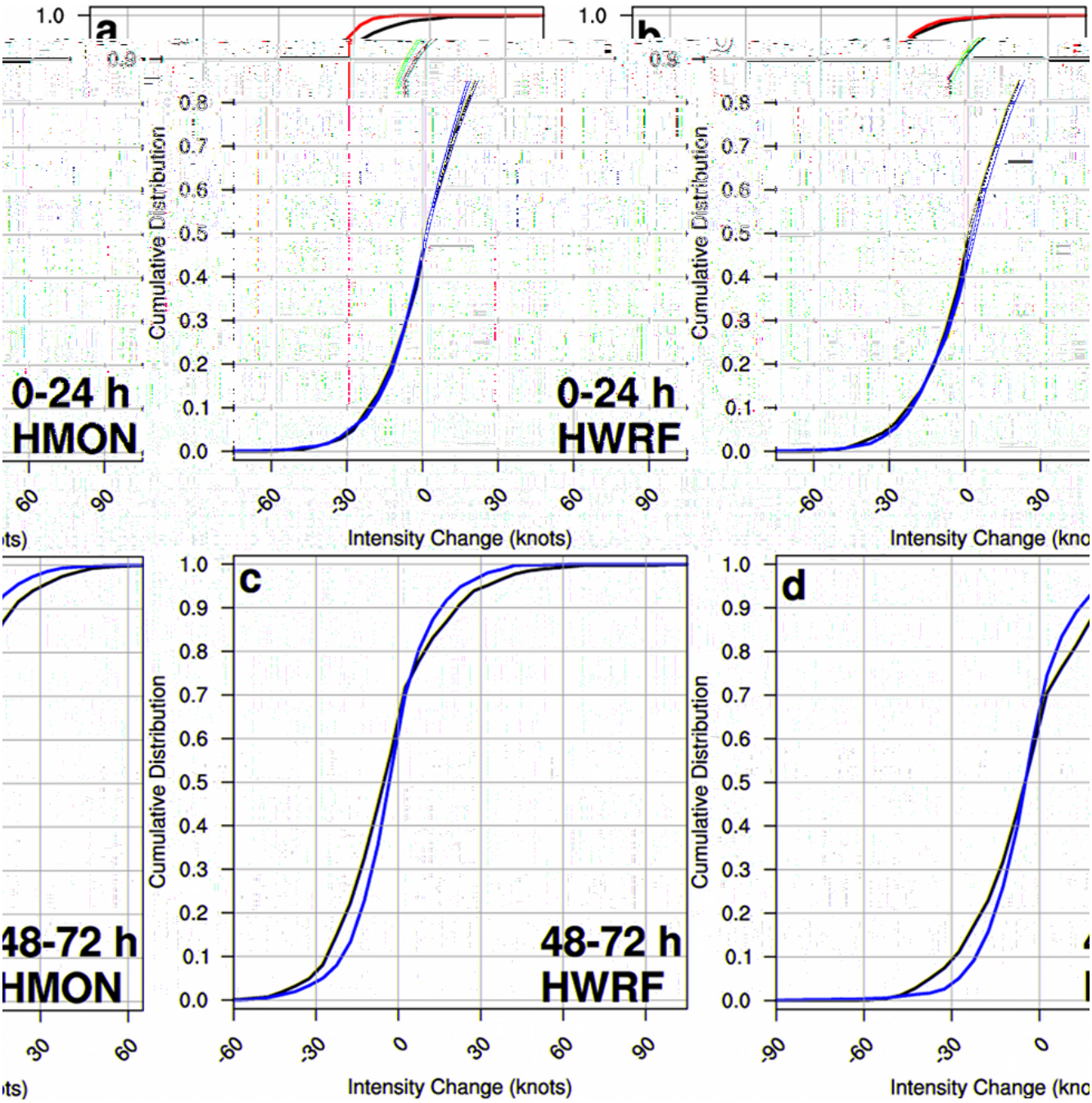

Figure 1. 0-24 h (**a**) HWRF and (**b**) HMON maximum wind speed change CDFs based on retrospective forecasts from the 2014–2016 cases (red line). The black line denotes the corresponding best track maximum wind speed change CDF. (**c**,**d**), as in (**a**,**b**), but for the 48–72 h forecasts.

The above process is repeated for the retrospective forecasts generated prior to the 2017, 2018, and 2019 seasons to understand whether the yearly model innovations have systematically improved each model's ability to simulate RI at the correct frequency. Figure 2a shows the 95th percentile as a function of the forecast lead time and season for the HWRF model and corresponding best track. It is worth pointing out that the set of retrospective cases changes from year to year (generally 2–3 years of past forecasts), so the set of cases is not homogeneous. For all three seasons, the HWRF model's 95th percentile intensity change is less than the best track at all lead times. Generally speaking, the best track 95th percentile shifts toward smaller values with increasing lead time. This likely occurs because RI occurs most frequently at the beginning of a TC's lifetime, e.g., [33]; therefore, RI is more

likely to occur at early forecast lead times, which are more likely to sample the early part of a TC lifecycle, compared to later lead times, which are more likely to sample the later part of a TC lifecycle. By contrast, the HWRF 95th percentile values show a substantial amount of year-to-year variability, with the 2017 values decreasing from 28 to 23 kts between 0 and 24 h to 48–72 h, while the 2019 values increase from 26 kts to 30 kts (the best track value is roughly 35 kts at all times in 2019). As a consequence, it appears that the 2019 HWRF retrospective forecasts are better able to simulate larger intensity changes compared to earlier years, which is likely a result of the difference in cases between the 2018 and 2019 retrospective forecasts since there were no major changes to the HWRF model between 2018 and 2019.

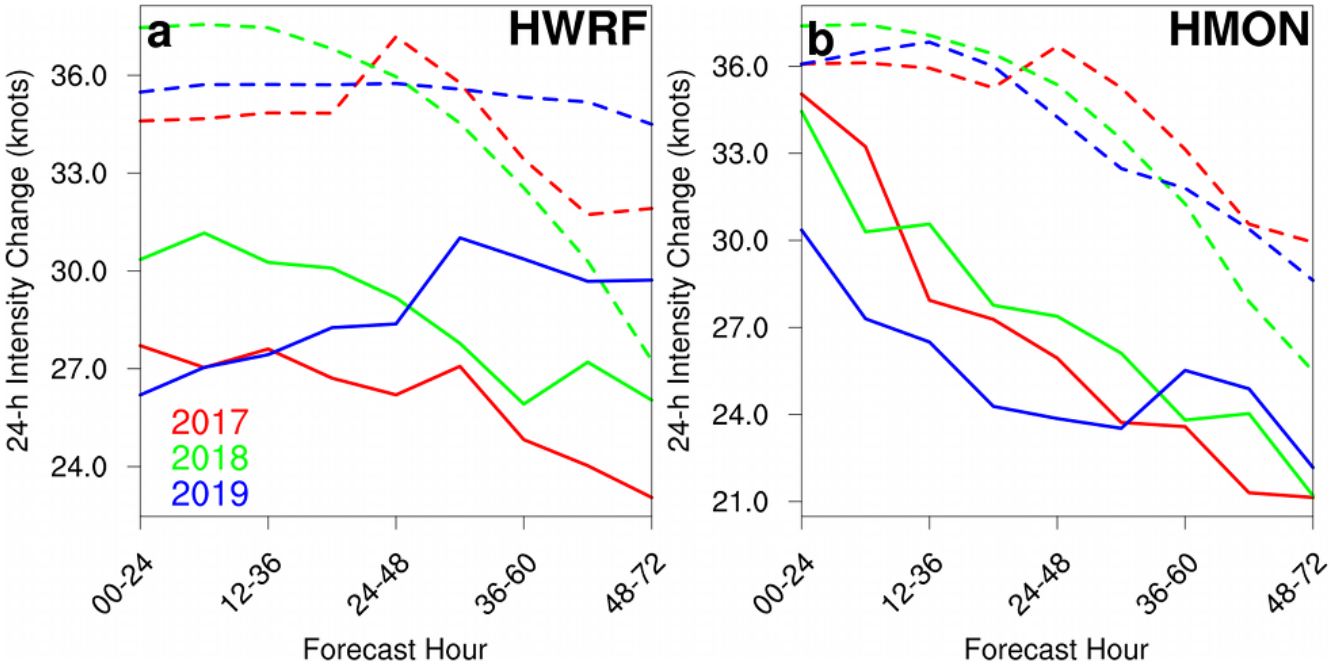

Figure 2. (**a**) HWRF 95th percentile 24-h TC maximum wind speed changes (solid line) based on retrospective forecasts prior to the 2017 (red), 2018 (green) and 2019 (blue) seasons as a function of forecast lead time. The dashed lines indicate the corresponding best track TC maximum wind speed changes for those years. (**b**) as in (**a**), but for the HMON model.

Whereas HWRF forecasts were characterized by improvements in the 95th percentile over this three year period, the HMON forecasts exhibit more year-to-year consistency (Figure 2b). For all three years, the best track 95th percentile decreases by roughly 6 kts 24 h^{-1} between 0 and 24 h to 48-72 h. By contrast, the HMON model 95th percentile from 2017 and 2018 decreases at a faster rate, from roughly 35 kts 24 h^{-1} to 21 kts 24 h^{-1} (a decrease of 14 kts 24 h^{-1}). This would suggest that HMON has greater difficulty simulating large intensity changes as the model is run for longer periods of time. By contrast, the 95th percentile in the 2019 HMON retrospective decreased at the same rate as the best track at all lead times, but where the HMON value is systematically 6 kts 24 h^{-1} lower than the best track.

The intensity change CDFs presented above suggest that HWRF and HMON forecasts replicate some intensity changes at the correct frequency as best track for some categories, but replicates others at a much lower frequency. As a consequence, both modeling systems could benefit from a conditional bias correction scheme that adjusts the model's intensity forecast more substantially for the intensity change values that the model has difficulty replicating (i.e., RI), while leaving others as is. One method for accomplishing this type of bias correction is a quantile-based approach, e.g., [34,35], whereby the model's intensity forecast is mapped to its CDF percentile based on the retrospective forecasts from that year (i.e., the 2017 forecasts use the CDF based on the retrospective forecasts produced before the 2017 season) and the model forecast intensity change is replaced by the corresponding

best track intensity change based on the forecast percentile. Returning to the results from 2017, this would mean a 0–24 h HWRF forecast of 10 kts 24 h^{-1} (a 69% intensity change) would be corrected to 11.5 kts 24 h^{-1}, while a HWRF forecast of 27 kts 24 h^{-1} (a 95% intensity change) would be adjusted to 33 kts 24 h^{-1}. HWRF and HMON forecasts are bias corrected against CDFs that are computed for each 24 h period starting from 0 to 24 h to 72–96 h, while 24 h periods beyond that time use the 72–96 h CDF due to the limited number of samples at longer lead times, which in turn leads to unstable statistics. This bias correction is applied independently to each ensemble member prior to calculation of forecast probabilities.

Application of quantile-based bias correction has a substantial impact on the skill of HWRF and HMON RI forecasts. This is assessed by comparing probabilistic RI forecasts for 2017–2019 HWRF and HMON ensemble systems with and without bias correction during each 24 h period from 0–24 h to 48–72 h each 6 h. The skill of each set of forecasts is assessed using the Brier Skill Score [36], which measures the skill relative to a climatological baseline, and is defined as:

$$BSS = 1.0 - \frac{BS_{forecast}}{BS_{climatology}}, \quad (1)$$

where $BS_{forecast}$ is the Brier scores [37] of the HWRF and HMON forecasts and $BS_{climatology}$ is the Brier score for a forecast based on the climatological frequency of RI (i.e., a 5% chance). Both HWRF and HMON forecasts have higher BSS with bias correction, with HWRF increasing its BSS by 67%, compared to 15% for HMON, though it should be noted that the BSS are still relatively small (Table 3). By contrast, the raw and bias-corrected BSS for other intensity change categories are relatively unchanged by the application of quantile mapping, which is not surprising given that this method adjusts large intensity change (>20 kts 24 h^{-1}) relative to smaller values (not shown). Given the superior performance of quantile-mapping bias correction, it is subsequently applied for all of the remaining results.

Table 3. Brier Skill Score of the Probability of RI based on 2017–2019 HWRF and HMON ensemble data during each 24 h period from 0 to 24 h to 48–72 h. The bias corrected forecasts use the quantile-mapping method, while the raw forecasts do not use any bias correction. For this calculation, the climatological frequency is assumed to be 5%.

Model	Raw Forecast	Bias Corrected
HWRF	0.006	0.010
HMON	0.121	0.139

3.2. HWRF and HMON Ensembles

Before proceeding to validate the ensemble-based probabilities over the three years of cases, it is worthwhile to show an example ensemble-based intensity change probabilities that includes a skillful prediction of the timing of RI. Figure 3 shows a stacked bar chart of the probability of different intensity change categories, e.g., [21] based on the HWRF and HMON ensemble for Hurricane Harvey initialized 1200 UTC 22 August 2017 (i.e., before Harvey was reclassified as a tropical depression for a second time). In this case, both the HWRF and HMON ensembles have the maximum probability of RI in the middle of the period where RI actually took place (there were eight 24 h periods that qualify for RI starting at 24–48 h and ending in the 60–84 h period), which immediately precedes landfall of the storm, after which time both ensembles correctly predict rapid weakening due to land interaction. It is worth pointing out that quantile mapping has a substantial impact on producing the high probabilities in this case. In the case of the HWRF ensemble, the raw forecast probabilities have a maximum probability of 32% during the 36–60 h period (not shown), compared to a maximum of 71% with the quantile mapping.

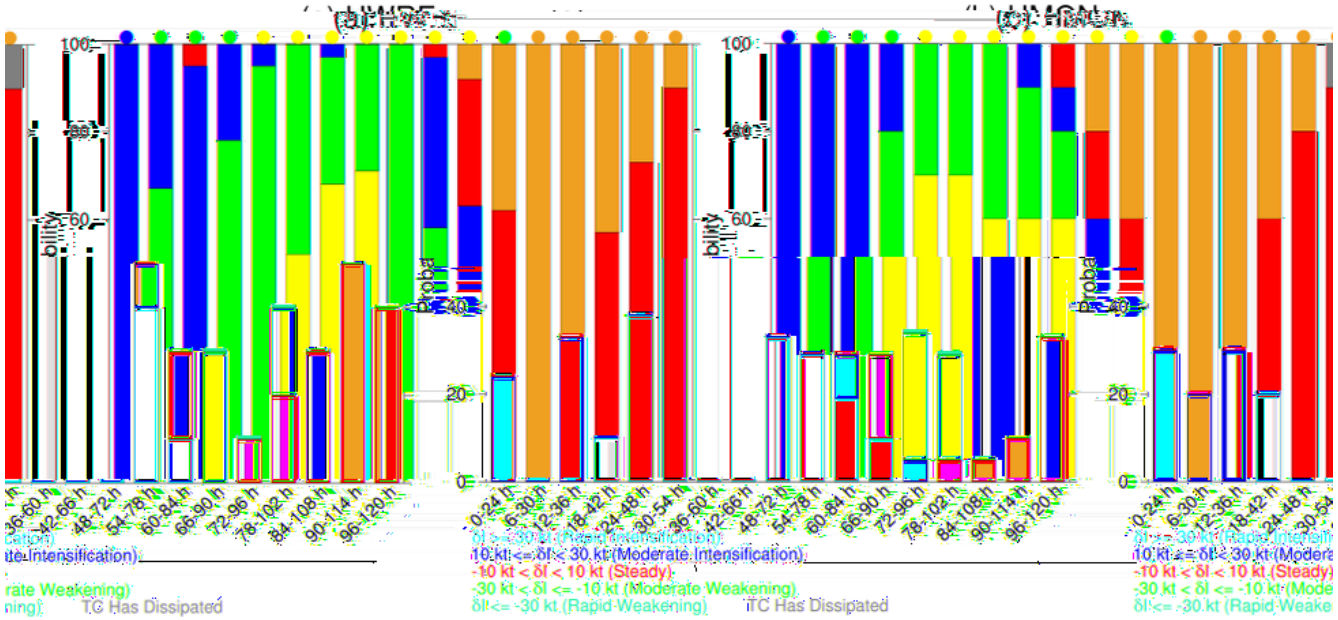

Figure 3. Stacked bar chart of the probability of the change in Hurricane Harvey's maximum wind speed as a function of lead time based on the (**a**) HWRF and (**b**) HMON ensemble initialized 1200 UTC 22 August 2017. The colored dots along the top of the figure denote the verifying change in Harvey's maximum wind speed based on best track data.

Turning to all forecasts for the 2017–2019, even with the application of quantile-mapping, both HWRF and HMON ensemble-based probabilistic forecasts do not exhibit RI forecast skill comparable to statistical models. Figure 4 shows reliability diagrams for probabilistic RI forecasts for the 0–24, 0–48, and 0–72 h periods, which equates to a 30, 55, and 65 knot intensity change, respectively, [2]. The HWRF and HMON ensemble forecasts are not available until 9 h after the initialization time; therefore, in order to provide a fair comparison between the statistical and dynamical models, the HWRF and HMON forecasts are "interpolated" by 12 h. For intensity changes, interpolation only involves time-shifting the forecasts because the intensity shift will uniformly adjust the intensity upward or downward by an equal amount at the start and end of the intensity change window. As a consequence, a hypothetical 1200 UTC SHIPS and DTOPS forecast is compared against the HWRF and HMON-based probabilities initialized at 0000 UTC, but valid at the appropriate time. Furthermore, given the relative infrequency of RI (42 periods over three years), the reliability diagrams use five bins, 0, 1–20%, 20–40%, 40–60%, 60–80% and >80%.

For the 0–24 h period, SHIPS-RII (RIOD) and DTOPS forecasts are characterized by reliability curves close to the 1:1 line. By contrast, both the HWRF and HMON ensemble-based probabilities are generally above the 1:1 line for forecast probabilities below 40%, meaning that RI happened more frequently than what was implied by the forecast probabilities. Furthermore, the HWRF and HMON probabilities are below the 1:1 line above the 40% forecast probability, meaning that RI happened less frequently than what was implied by the forecast probabilities, though it is difficult to assess probabilities above 60% due to the lack of cases. In addition, the Brier Skill Score for HMON forecasts is comparable to both SHIPS and DTOPS forecasts, with the HWRF ensemble substantially smaller (Table 4). Furthermore, 0–48 h RI forecasts (Figure 4b) show relatively similar behavior, though with the smaller number of cases (22 events during the three years), the curves are less consistent among the various bins. Both SHIPS-RII and DTOPS are close to the 1:1 line for low probabilities, while HWRF and HMON under-predict RI at the same level. Furthermore, HMON and DTOPS yield the highest BSS, with HWRF and SHIPS further behind. Finally, 0–72 h RI takes place less frequently than all four models suggest, as all of them are below the 1:1 line for nearly all categories (Figure 4c). Whereas SHIPS and HWRF ensemble-based probabilities exhibit BSS above zero, DTOPS and HMON ensemble guidance are slightly below.

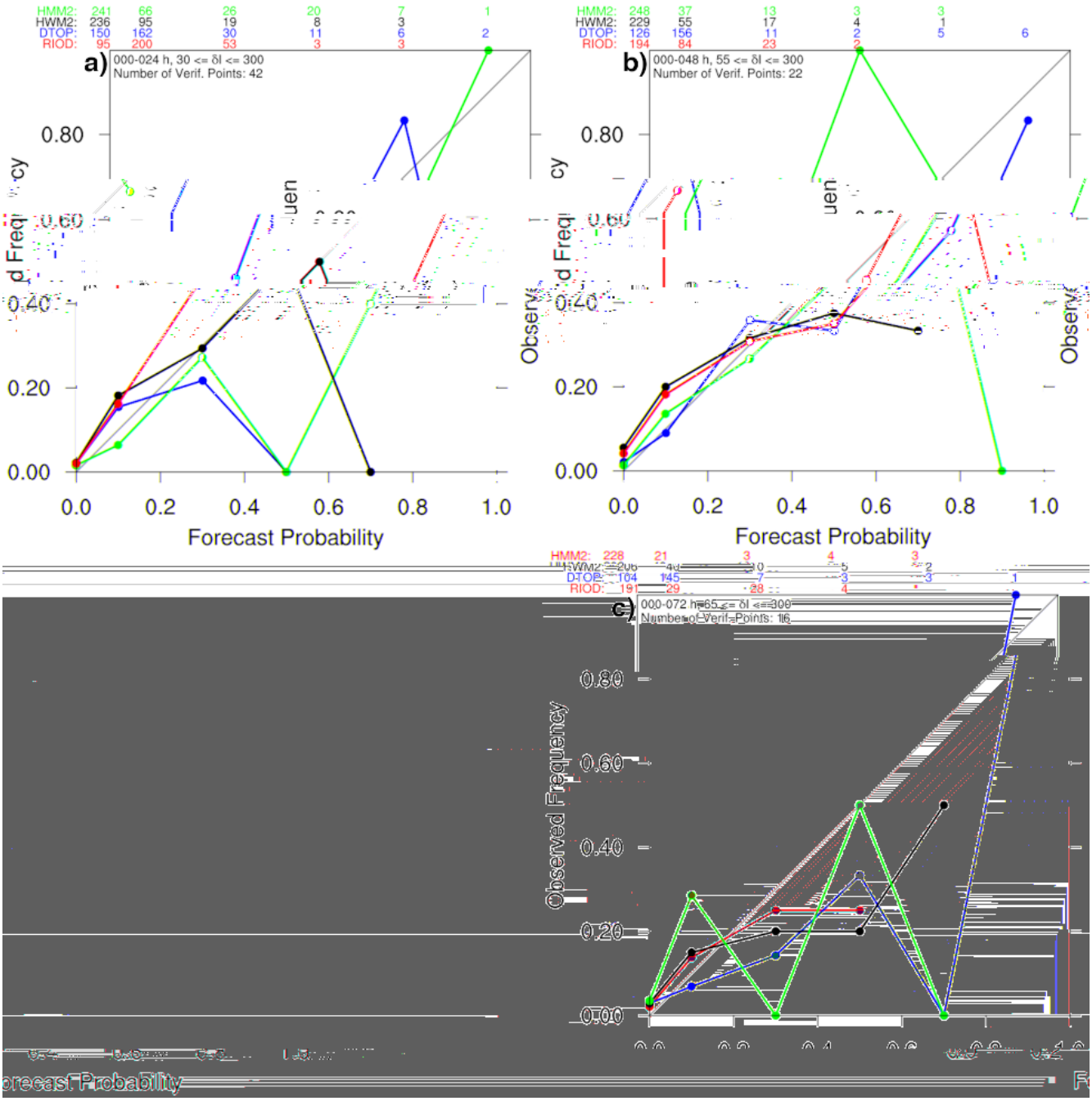

Figure 4. Reliability diagrams for (**a**) 0–24 h TC maximum wind speed change >= 30 knots, (**b**) 0–48 h TC maximum wind speed change >= 55 kts, and (**c**) 0–72 h TC maximum wind speed change >= 65 knots for interpolated HWRF ensemble (HWM2; black), interpolated HMON ensemble (HMM2; green), SHIPS Rapid Intensification Index (RIOD; red), and DTOPS (DTOP; blue) models during 2017–2019. The number of cases within each category is given along the top of each figure.

The relative infrequency of RI can make it difficult to assess whether these models can provide useful probabilistic guidance. One way to increase the number of potential cases is to consider multiple 24 h periods simultaneously, namely from 0 to 24 h to 48–72 h, which in turn yields 9 potential verification times for each individual forecast. Figure 5a shows the RI reliability diagram for all 24 h periods from 0 to 24 to 48–72 h into the forecast, which yields 311 RI "events". Similar to Figure 4a, both the HWRF and HMON ensembles are characterized by forecast probabilities that are lower than the observed frequency when the forecast probability is below 20%, while the ensemble-based forecast probabilities are higher than the observed frequency when the forecast probability exceeds 50%, with the

HMON forecast closer to the 1:1 line. In turn, the HMON BSS is 0.14, compared to 0.01 for the HWRF system. As a consequence, it appears that the HMON ensemble-based probabilities are more skillful forecasts of RI relative to the HWRF ensemble.

Table 4. Brier Skill Score (BSS) for the probability of TC maximum wind speed changes for varied time periods.

Model	0–24 h 30 kts	0–48 h, 55 kts	0–72 h, 65 kts
HWRF Ensemble	0.070	0.061	0.041
HMON Ensemble	0.170	0.216	−0.007
SHIPS RI Index	0.167	0.061	0.080
DTOPS	0.174	0.216	−0.011

Other forecast change categories are characterized by mixed skill relative to the RI category. Figure 5b shows the probability of a 24 h intensity change from 10 knots up to 30 kts. Similar to RI, this intensity change occurs more frequently than forecast up to the 50% forecast probability, while above that threshold, this intensity change occurs less frequently then the forecast implies. For a less than 10 kt change in maximum wind speed, both models exhibit a relatively flat reliability curve, with under-forecasting below 30% and over-forecasting above 30%, such that both models have BSS below zero. By contrast, both models exhibit skillful forecasts of weakening (24 h intensity changes $<= -10$ kts) as the reliability curves are close to the 1:1 line for all forecast probabilities (Figure 5d).

The previous results suggest that both the HWRF and HMON models have varying ability to provide useful probabilistic guidance on the correct intensity change category. One way to evaluate this is to compute rank probability skill score (RPSS; [38]) as a function of forecast period. RPSS is defined as

$$RPSS = 1.0 - \frac{RPS_{forecast}}{RPS_{climatology}}, \quad (2)$$

where $RPS_{forecast}$ is the rank probability score of the ensemble forecast computed from the probabilities of the four categories in Table 2, and $RPS_{climatology}$ is the rank probability score of a forecast based on the climatological frequency of each category, which is provided in Table 2. Similar to BSS, RPSS ranges from negative infinity to one, where a score of 1 is perfect and less than zero means no skill relative to a forecast based on the climatological frequency (Figure 6). Prior to 36–60 h, both the HWRF and HMON ensembles have positive RPSS scores, with a value of approximately 0.15 for 0–24 h, linearly decreasing to zero at 36–60 h. Beyond that lead time, both models have a skill score that is indistinguishable from zero; therefore, these models appear to provide their most skillful forecasts of the intensity change category right after initialization, but where skill is lost by the 36–60 h and beyond.

As would be expected from the reliability diagrams, the most difficult categories are the intensification and steady categories. Examining the cases where the 48–72 h HWRF ensemble has the greatest number of members in the steady category ($-10 < \delta I < 10$ kts; 169 cases), a steady intensity change is observed in 40% of those cases, while 14% of those cases actually experienced either rapid intensification ($\delta I >= 30$ kts) or even rapid weakening ($\delta I <= -30$ kts). Similarly, when the HWRF ensemble had the highest probability of intensification ($10 <= \delta I < 30$ kts) for the same time period, intensification was observed 40% of the time; however, steady was observed 33% of the time. Similar results are obtained for both the HMON ensemble and for other lead times (not shown); therefore, it suggests that these two ensemble systems have difficulty at producing reliable forecasts of the steady and intensification categories.

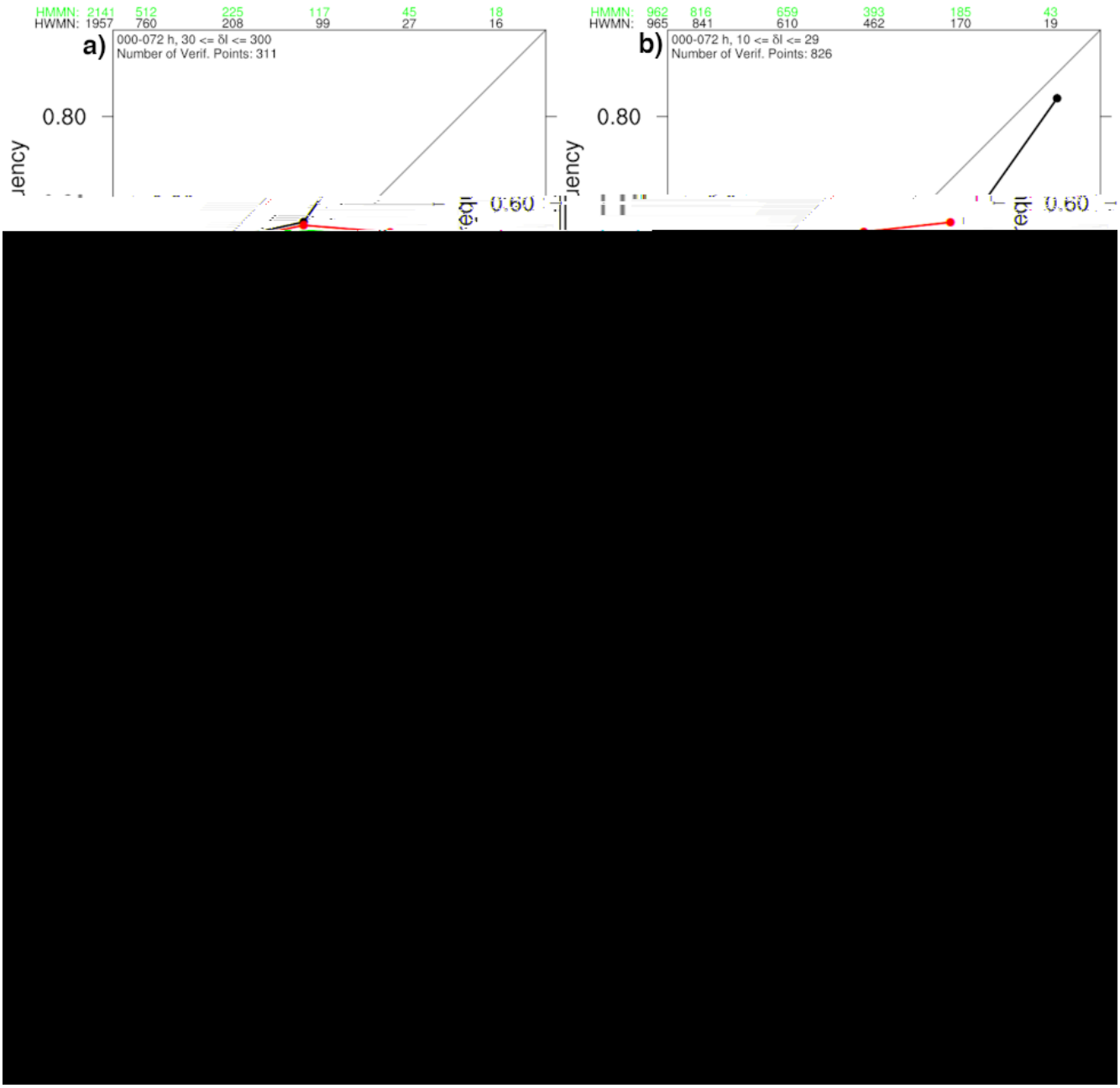

Figure 5. HWRF (black) and HMON ensemble (green) reliability diagrams for 24-h TC maximum wind speed change (**a**) >= 30 knots, (**b**) 10 <= δI < 30 kts, (**c**) −10 < δI < 10 kts, and (**d**) δI <= −10 kts during each 24 h period from 0 to 24 h to 48–72 h during 2017–2019. The number of cases within each category is given along the top of each figure.

Having two distinct ensemble prediction systems, each with some skill at predicting RI and other intensity change categories, provides the possibility of combining all of the ensemble members from both systems into a single multi-model ensemble probabilistic product. The hope would be that the combined ensemble will have more skill than any individual system. This hypothesis is tested by computing the probability of RI from the 21 HWRF members and 11 HMON members for the same set of cases. Given the larger number of HWRF members, the probabilities from the combined ensemble are going to be weighted toward the HWRF probabilities, which as seen above, has less skill at predicting RI relative to the HMON ensemble system. Figure 7 shows the reliability diagram of the individual ensemble-based products as well as the combined ensemble. For lower probabilities (<30%), the combined ensemble has a relatively close to the 1:1 line and similar to the individual ensembles. At higher forecast probabilities, the combined

ensemble falls off the 1:1 line, but not as quickly as the HWRF system, but is still further away from the 1:1 line relative to the HMON ensemble, and the BSS of the combined ensemble (0.113) is lower than the BSS of the HMON ensemble itself. As a consequence, these results suggest that adding the HWRF ensemble members to the HMON ensemble members does not lead to a more skillful ensemble.

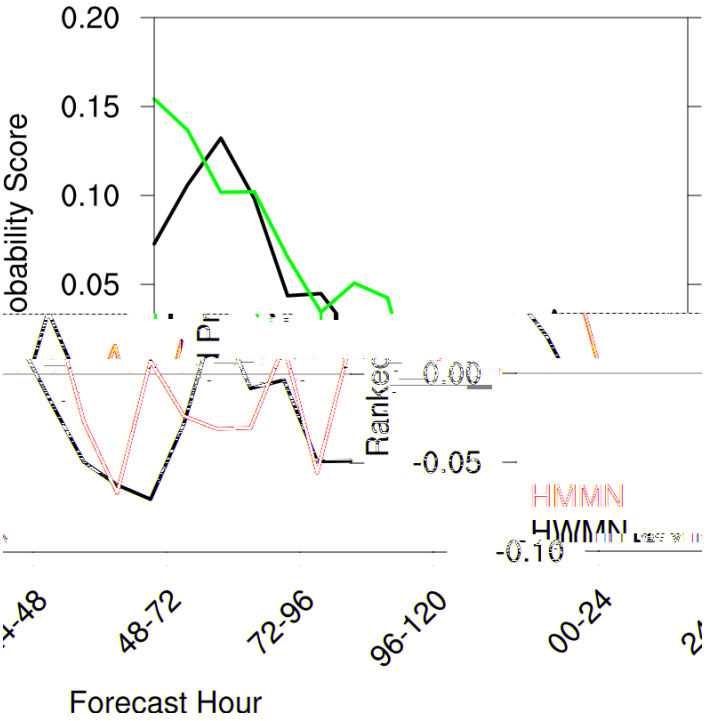

Figure 6. Ranked Probability Skill Score (RPSS) of HWRF and HMON ensemble 24 h maximum wind speed change as a function of forecast lead times from2017 to 2019.

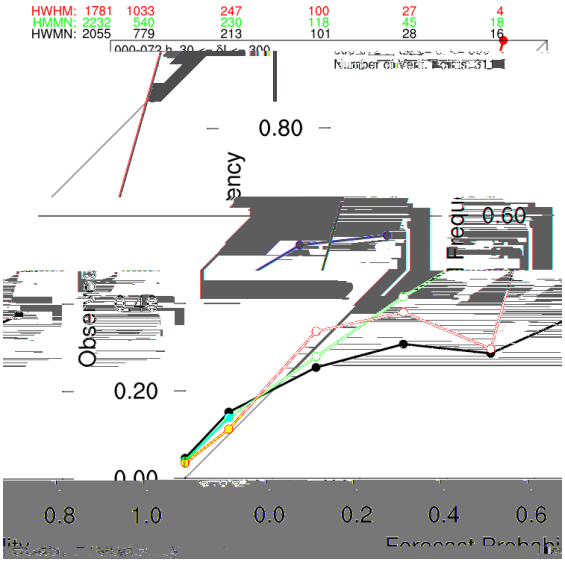

Figure 7. HWRF (black), HMON ensemble (green), and combined HWRF and HMON ensemble (red) reliability diagrams for the probability of 24-h TC maximum wind speed change >= 30 knots during each 24 h period from 0 to 24 h to 48–72 h during 2017–2019. The number of cases within each category is given along the top of each figure.

There are at least two potential reasons for the lack of additional skill in the multi-model combined ensemble. One possibility is that the HWRF and HMON forecasts are not sufficiently independent of each other, which is necessary for skillful consensus forecasts,

e.g., [39]. This possibility is assessed by computing the correlation between the error in the HWRF control member 24 h intensity changes and the error in HMON control member 24 h intensity changes for each period. For all time windows, the correlation coefficient exceeds 0.65 (not shown); therefore, it suggests that the two models may not provide a particularly independent set of forecasts.

Another possible explanation for the lack of higher skill in the combined ensemble could be the higher weight placed on the HWRF ensemble due to the greater number of members. This possibility is tested by computing the probability of RI based on a combination HWRF/HMON ensemble where equal weight is given to the HWRF and HMON probabilities. While the reliability curves for the equal weight probability are closer to the 1:1 line (not shown), the RI BSS for the multi-model ensemble where the HWRF and HMON ensemble probabilities receive equal weight is 0.025 higher compared to computing the probabilities based on combining all HWRF and HMON ensemble members into one multi-model ensemble (hence giving more weight to the HWRF ensemble because of its 21 members). As a consequence, it appears that some minor gains can be made with a multi-model ensemble with more weight given to the HMON ensemble.

3.3. Application to ECMWF Ensemble

While the HWRF and HMON ensemble show some promise to predict TC intensity changes, these are not operational ensemble prediction systems and hence are not run on all cases. Given this, and the potential benefit of the quantile-based TC intensity bias correction, it is worth trying to apply quantile-based bias correction to a lower-resolution operational ensemble prediction system, such as the ECMWF ensemble, and evaluate its ability to provide probabilistic TC intensity change guidance. During 2019, the ECMWF ensemble system had a native grid spacing of 18 km, thus although this not sufficient grid spacing to fully resolve TC structures, it is possible that this grid spacing could still resolve TC maximum wind speed tendencies.

This possibility is assessed by comparing the CDF of the ECMWF 24-h maximum wind speed changes for all Atlantic and Eastern Pacific TCs during 2017 and 2018 seasons against the corresponding best track data (Figure 8). During the 0–24 h period, the ECMWF model CDF has a sharper slope than the corresponding best track data. Specifically, the ECMWF model maximum wind speed changes appear to be within $\pm$30 kts 24 h^{-1}, which is a smaller range than the best track data. Furthermore, the slope of the ECMWF CDF appears to parallel the best track values up to a maximum wind speed change of 0 kts, but is higher above that, meaning that the ECMWF model does not capture intensification as frequently as the best track. Finally, the 95th percentile is roughly 20 kts in the ECMWF, compared to almost 40 kts for the best track data for these cases. The 48–72 h forecast CDF is qualitatively similar to the 0–24 h version, particularly for the largest intensity changes, the range of intensity changes where the slope is parallel to the best track, and the difference in the 95th percentile maximum wind speed change.

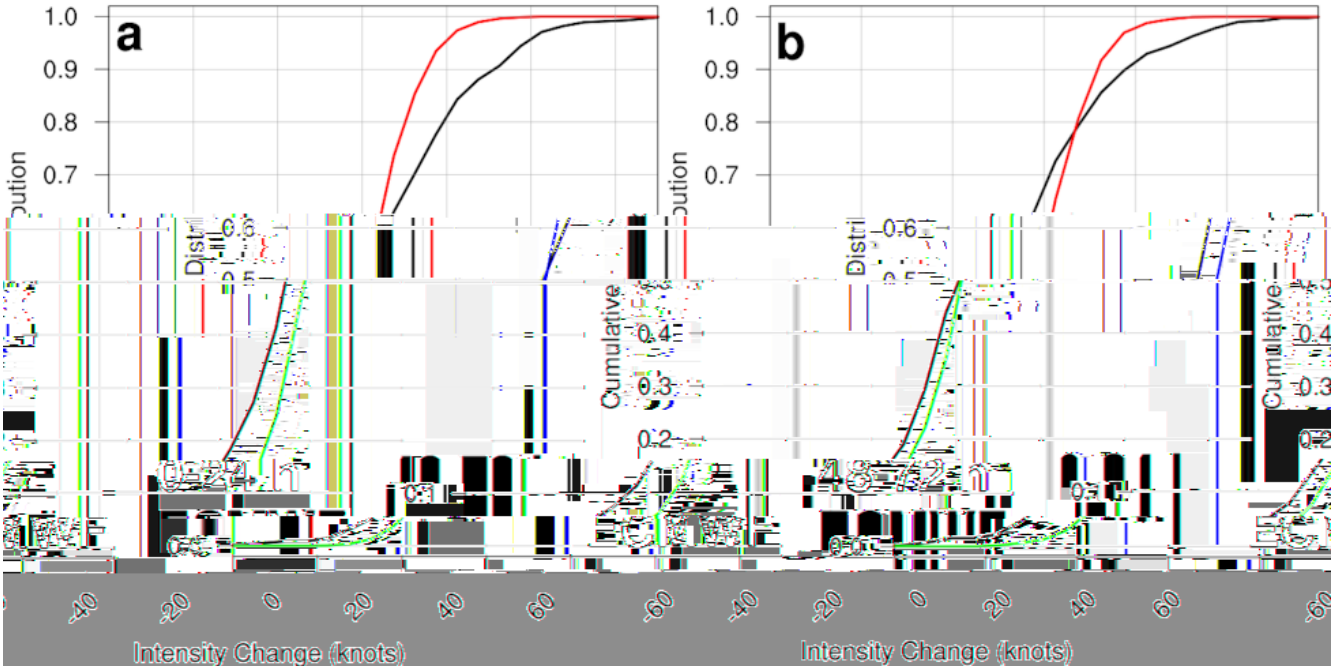

Figure 8. 0-24 h ECMWF maximum wind speed change CDFs based on retrospective forecasts from the 2017–2018 seasons (red line). The black line denotes the corresponding best track maximum wind speed change CDF. (**b**) as in (**a**), but for the 48–72 h forecasts.

Applying quantile-based bias correction to the ECMWF forecasts does not yield skillful probabilistic forecasts of RI. Figure 9 shows reliability diagrams of each intensity change category during each 24 h period from0 to 24 h to 48–72 h for all Atlantic and Eastern Pacific cases from 2019 (there are 5700 cases with a 0000 or 1200 UTC initialization time). Whereas the RI and steady intensity change categories exhibit no skill, as demonstrated by the relatively flat reliability curves (Figure 9a,c), and BSS that are less than zero, both the intensification and weakening categories are categorized by reliability diagrams near the 1:1 line (Figure 9b,d) and the BSS are 0.127 and 0.234, respectively. These results suggest that the ECMWF ensemble exhibits limited ability to predict the correct category beyond whether the TC is intensifying or weakening. One possible reason is the substantial bias correction, particularly for large intensity changes. For example, if the ECMWF predicts a 20 kt intensity change, the quantile method will correct that to 40 kts. In this case, the bias correction scheme cannot distinguish between the model correctly predicting a 20 kt intensity change (as occurs for some cases) or situations where there is a 40 kt intensity change. As a consequence, it appears that the quantile-based intensity change method might perform best when the intensity change CDF is relatively close to the verification CDF.

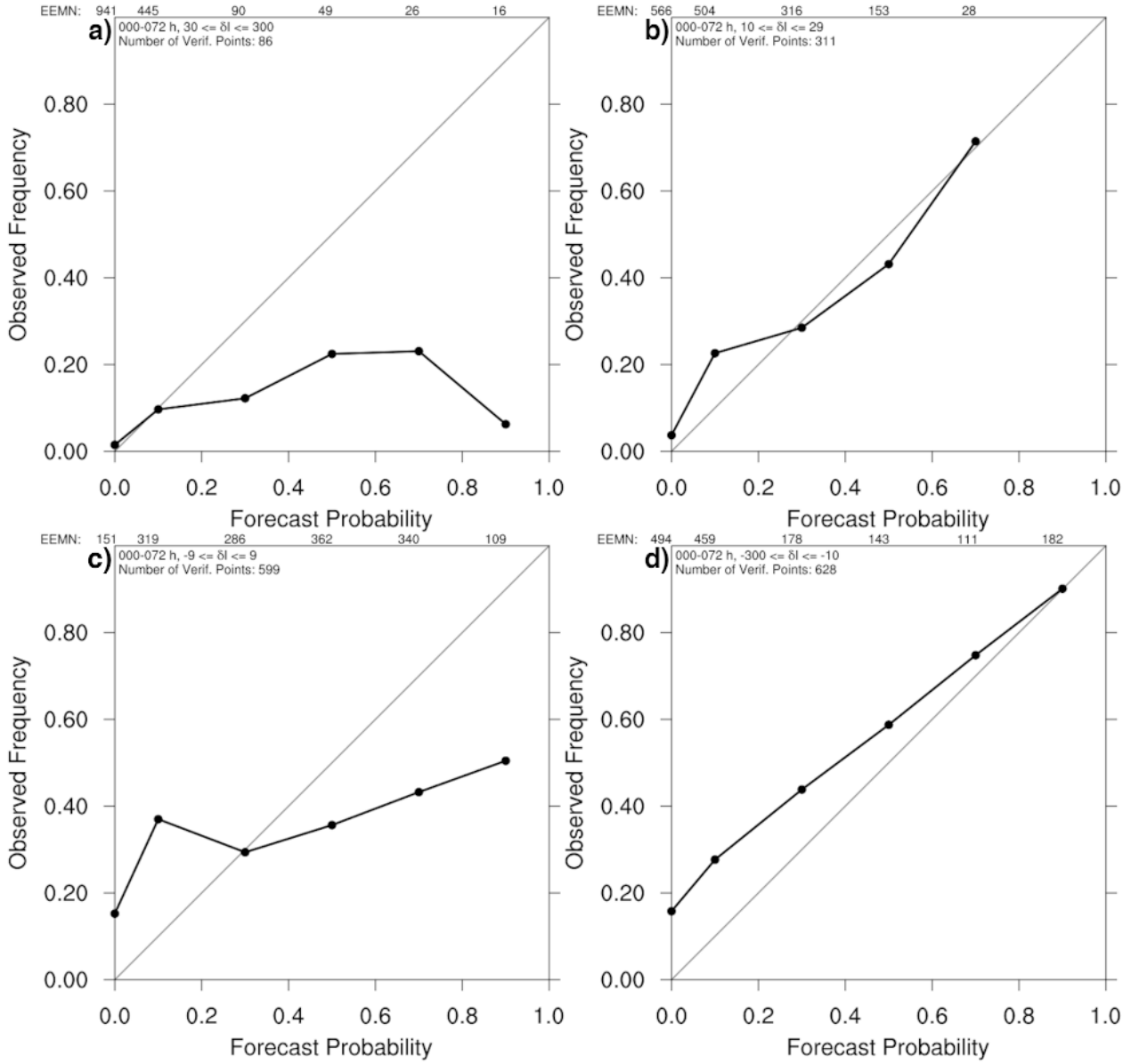

Figure 9. ECMWF ensemble reliability diagrams for a 24 h TC maximum wind speed change (**a**) >=30 knots, (**b**) 10 <= δI < 30 kts, (**c**) −10 < δI < 10 kts, and (**d**) δI <= −10 kts for each 24 h period from0 to 24 h to 48–72 h for all Atlantic and Eastern Pacific cases during 2019. The number of cases within each category is given along the top of each figure.

4. Summary and Conclusions

The goal of this study was to evaluate the hypothesis that specific dynamical ensemble prediction systems can provide skillful probabilistic predictions of TC intensity change relative to the current suite of operational guidance. This hypothesis is tested using 489 cases from the HWRF and HMON ensemble data from the HFIP demonstration system during the 2017–2019 seasons and verifying the results against best track data. These forecasts are validated in four different intensity change categories that are of potential interest to forecasters and comparable to the definitions used in operational statistical RI models (SHIPS-RII and DTOPS).

Intensity change CDFs for both the HWRF and HMON models indicate that these two systems generally replicate the observed frequency of weakening, steady, and moderate intensification events; however, RI events occur less frequently than observations. Rather than bias-correcting the model intensity change by a fixed amount for each forecast lead time, a quantile-based bias correction scheme was developed that adjusts the model

intensification rate to the best track value for a given percentile. For both HWRF and HMON, this results in the biggest change occurs for near-RI levels, while making small changes for other values. Applying this bias correction results in improved probabilistic RI predictions for both models.

Ensemble-based probabilistic guidance have limited skill relative to climatology for all intensity change categories. For RI, both the SHIPS-RII and DTOPS predictions are more reliable then either dynamical model, with HMON exhibiting greater skill relative to the HWRF ensemble. One potential reason for the superior performance of the HMON ensemble is the use of multiple physics packages, which other studies have shown generally results in greater ensemble standard deviation for a variety of mesoscale phenomenon, e.g., [40–42], but typically at the cost of maintaining multiple packages and having an ensemble where the members are not equally-likely. Of the remaining intensity change categories, both models performed best at predicting weakening, and worst for steady intensity changes, with the latter exhibiting no skill relative to climatology. Furthermore, both models have the greatest skill at predicting the intensity change category from 0 to 24 h and lose their ability to distinguish between the categories by 48–72 h, which is primarily the result of the relatively poor performance of the steady category. The higher skill in the first 24 h is somewhat surprising given that TC modeling systems often exhibit spin-up issues in the first 6–12 h, e.g., [43–45]. The lack of skill of the steady category is likely a consequence of the lack of ensemble spread in intensity (not shown), as the biggest issue is that verification falls outside of that category too frequently when the model predicts steady (including some cases where the TC experiences rapid intensification or weakening). Unfortunately, quantile-based bias correction will not address this flaw, thus it will be necessary to address the underlying model physics. The variety of physics packages employed by the HMON ensemble may explain its superior performance compared to the HWRF system, which parameterizes model error by applying white noise stochastic perturbations to various parts of the physics parameterizations. Combining the HWRF and HMON ensemble output into a single multi-model ensemble system provides more skillful forecasts relative to the HWRF ensemble; however, RI forecasts of the combined ensemble are worse than the HMON model itself, even when accounting for the extra weight given to the HWRF ensemble due to its larger ensemble size.

Although the ECMWF model has been shown to provide skillful probabilistic forecasts for a number of fields and phenomenon, including TCs, e.g., [46], applying quantile bias correction does not yield skillful intensity change forecasts. The main reason for this is that the ECMWF intensity change CDF is too narrow relative to observations. Furthermore, this approach makes it difficult to distinguish between cases where the model correctly predicts a moderate intensity change and situations where the moderate intensity change should be corrected to an RI. More than likely, the ECMWF ensemble will require additional resolution upgrades to achieve an intensity change CDF that is comparable to the HWRF and HMON ensemble and hence something closer to observations. Until then, one possibility is to use machine learning approach, e.g., [17,18], which would allow for a greater number of bias correction predictors.

Author Contributions: Conceptualization, R.D.T. and M.D.; Methodology, R.D.T. and M.D.; Software, R.D.T.; Validation, R.D.T.; Formal Analysis, R.D.T. and M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NOAA grant numbers NA16NWS4680025 and NA18NWS4680060.

Data Availability Statement: Forecast e-deck files are available from the authors upon request.

Acknowledgments: The authors thank Zhan Zhang and Weiguo Wang (NOAA/EMC) for providing HWRF and HMON ensemble output, respectively, and for Matt Onderlinde for providing the SHIPS-RI and DTOPS data and for discussions of the results. John Knaff, Kate Musgrave and three anonymous reviewers provided valuable input on earlier versions of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest. Furthermore, NOAA had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BSS	Brier Skill Score
HMON	Hurricanes in a Multi-scale Ocean-coupled Non-hydrostatic
HWRF	Hurricane Weather Research and Forecasting
RPSS	Ranked Probability Skill Score

References

1. Cangialosi, J.P.; Blake, E.; DeMaria, M.; Penny, A.; Latto, A.; Rappaport, E.; Tallapragada, V. Recent Progress in Tropical Cyclone Intensity Forecasting at the National Hurricane Center. *Weather. Forecast.* **2020**, *35*, 1913–1922. [CrossRef]
2. Kaplan, J.; DeMaria, M. Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Weather Forecast.* **2003**, *18*, 1093–1108. [CrossRef]
3. Blake, E.S.; Zelinsky, D.A. Tropical Cyclone Report Hurricane Harvey (AL092017) 17 August September 2017. Technical report, NOAA/National Hurricane Center. 2018. Available online: https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf (accessed on 11 December 2020).
4. Pasch, R.J.; Penny, A.B.; Berg, R. Tropical Cyclone Report Hurricane Maria (AL152017) 16 September 2017. Technical Report, NOAA/National Hurricane Center. 2019. Available online: https://www.nhc.noaa.gov/data/tcr/AL152017_Maria.pdf (accessed on 11 December 2020).
5. Beven, J.L., II; Berg, R.; Hagen, A. Tropical Cyclone Report Hurricane Michael (AL142018) 7 October–11 October 2018. Technical report, NOAA/National Hurricane Center. 2019. Available online: https://www.nhc.noaa.gov/data/tcr/AL142018_Michael.pdf (accessed on 11 December 2020).
6. Kaplan, J.; Rozoff, C.M.; DeMaria, M.; Sampson, C.R.; Kossin, J.P.; Velden, C.S.; Cione, J.J.; Dunion, J.P.; Knaff, J.A.; Zhang, J.A.; et al. Evaluating Environmental Impacts on Tropical Cyclone Rapid Intensification Predictability Utilizing Statistical Models. *Weather Forecast.* **2015**, *30*, 1374–1396. [CrossRef]
7. Hendricks, E.A.; Peng, M.S.; Fu, B.; Li, T. Quantifying Environmental Control on Tropical Cyclone Intensity Change. *Mon. Weather Rev.* **2010**, *138*, 3243–3271. [CrossRef]
8. Kaplan, J.; Cione, J.; Leighton, P.; DeMaria, M.; Knaff, J.; Dunion, J.; Zhang, J.; Dostalek, J.; Solbrig, J.; Hawkins, J.; et al. Enhancements to the operational SHIPS Rapid Intensification Index. In Proceedings of the 29th Conference on Hurricanes and Tropical Meteorology, Tucson, AZ, USA, 10–14 May 2010.
9. Stevenson, S.N.; Corbosiero, K.L.; Molinari, J. The convective evolution and rapid intensification of Hurricane Earl (2010). *Mon. Weather Rev.* **2014**, *142*, 4364–4380. [CrossRef]
10. Zagrodnik, J.P.; Jiang, H. Rainfall, convection, and latent heating distributions in rapidly intensifying tropical cyclones. *J. Atmos. Sci.* **2014**, *71*, 2789–2809. [CrossRef]
11. Onderlinde, M.J.; Nolan, D.S. Tropical cyclone-relative helicity and the pathways to intensification in shear. *J. Atmos. Sci.* **2016**, *73*, 869–890. [CrossRef]
12. Rogers, R.F.; Reasor, P.D.; Zhang, J.A. Multiscale structure and evolution of Hurricane Earl (2010) during rapid intensification. *Mon. Weather Rev.* **2013**, *143*, 536–562. [CrossRef]
13. Nolan, D. What is the trigger for tropical cyclogenesis? *Aust. Meteorol. Mag.* **2007**, *56*, 241–266.
14. Jiang, H. The relationship between tropical cyclone intensity change and the strength of inner-core convection. *Mon. Weather Rev.* **2012**, *140*, 1164–1176. [CrossRef]
15. Knaff, J.A.; Sampson, C.R.; Musgrave, K.D. An Operational Rapid Intensification Prediction Aid for the Western North Pacific. *Weather Forecast.* **2018**, *33*, 799–811. [CrossRef]
16. Onderlinde, M.; DeMaria, M. Deterministic to Probabilistic Statistical rapid intensification index (DTOPS): A new method for forecasting RI probability. In Proceedings of the 33rd Conference on Hurricanes and Tropical Meteorology, Vedra Beach, FL, USA, 16–20 April 2018.
17. Lewis, W.E.; Rozoff, C.; Alessandrini, S.; Delle Monache, L. Performance of the HWRF Rapid Intensification Analog Ensemble (HWRF RI-AnEn) during the 2017 and 2018 HFIP Real-Time Demonstrations. *Weather Forecast.* **2020**, *35*, 841–856. [CrossRef]
18. Cloud, K.A.; Reich, B.J.; Rozoff, C.M.; Alessandrini, S.; Lewis, W.E.; Delle Monache, L. A feed forward neural network based on model output statistics for short-term hurricane intensity prediction. *Weather Forecast.* **2019**, *34*, 985–997. [CrossRef]
19. Zhang, Z.; Tallapragada, V.; Kieu, C.; Trahan, S.; Wang, W. HWRF based ensemble prediction system using perturbations from GEFS and stochastic convective trigger function. *Trop. Cycl. Res. Rev.* **2014**, *3*, 145–161.
20. Torn, R.D. Evaluation of atmosphere and ocean initial condition uncertainty and stochastic exchange coefficients on ensemble tropical cyclone intensity forecasts. *Mon. Weather Rev.* **2016**, *144*, 3487–3506. [CrossRef]

21. Komaromi, W.A.; Reinecke, P.A.; Doyle, J.D.; Moskaitis, J.R. The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System - Tropical Cyclone Ensemble (COAMPS-TC Ensemble). *Weather Forecast.* **2021**, *36*.
22. Lang, S.T.K.; Leutbecher, M.; Jones, S.C. Impact of perturbation methods in the ECMWF ensemble prediction system on tropical cyclone forecasts. *Q. J. R. Meteorol. Soc.* **2012**, *138*, 2030–2046. [CrossRef]
23. Goldenberg, S.B.; Gopalakrishnan, S.G.; Tallapragada, V.; Quirino, T.; Marks, F., Jr.; Trahan, S.; Zhang, X.; Atlas, R. The 2012 Triply Nested, High-Resolution Operational Version of the Hurricane Weather Research and Forecasting Model (HWRF): Track and Intensity Forecast Verifications. *Weather Forecast.* **2015**, *30*, 710–729. [CrossRef]
24. Tallapragada, V.; Kieu, C.; Trahan, S.; Liu, Q.; Wang, W.; Zhang, Z.; Tong, M.; Zhang, B.; Zhu, L.; Strahl, B. Forecasting Tropical Cyclones in the Western North Pacific Basin Using the NCEP Operational HWRF Model: Model Upgrades and Evaluation of Real-Time Performance in 2013. *Weather Forecast.* **2016**, *31*, 877–894. [CrossRef]
25. Zhang, J.A.; Rogers, R.F.; Tallapragada, V. Impact of Parameterized Boundary Layer Structure on Tropical Cyclone Rapid Intensification Forecasts in HWRF. *Mon. Weather Rev.* **2017**, *145*, 1413–1426. [CrossRef]
26. Hazelton, A.T.; Bender, M.; Morin, M.; Harris, L.; Lin, S. 2017 Atlantic Hurricane Forecasts from a High-Resolution Version of the GFDL fvGFS Model: Evaluation of Track, Intensity, and Structure. *Weather Forecast.* **2018**, *33*, 1317–1337. [CrossRef]
27. Gall, R.; Franklin, J.; Marks, F.; Rappaport, E.N.; Toepfer, F. The Hurricane Forecast Improvement Project. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 329–343. [CrossRef]
28. Torn, R.D.; Snyder, C. Uncertainty of tropical cyclone best track information. *Weather Forecast.* **2012**, *27*, 715–729. [CrossRef]
29. Landsea, C.W.; Franklin, J.L. Atlantic Hurricane Database Uncertainty and Presentation of a New Database Format. *Mon. Weather Rev.* **2013**, *141*, 3576–3592. [CrossRef]
30. Biswas, M.K.; Abarca, S.; Bernardet, L.; Ginis, I.; Grell, E.; Iacono, M.; Kalina, K.; Liu, B.; Liu, Q.; Marchok, T.; et al. *Hurricane Weather Research and Forecasting (HWRF) Model: 2018 Scientific Documentation*; Technical Report; NOAA: Washington, DC, USA, 2018.
31. Mehra, A.; Tallapragada, V.; Zhang, Z.; Liu, B.; Zhu, L.; Wang, W.; Kim, H.S. Advancing the State of the Art in Operational Tropical Cyclone Forecasting at NCEP. *Trop. Cyclone Res. Rev.* **2018**, *7*, 51–56.
32. Buizza, R.; Cardinali, C.; Kelly, G.; Thepaut, J.N. The value of observations. II: The value of observations located in singular-vector-based target areas. *Q. J. R. Meteorol. Soc.* **2007**, *133*, 1817–1832. [CrossRef]
33. Yaukey, P.H. Intensification and rapid intensification of North Atlantic tropical cyclones: Geography, time of year, age since genesis, and storm characteristics. *Int. J. Climatol.* **2014**, *34*, 1038–1049. [CrossRef]
34. Hamill, T.M.; Whitaker, J.S. Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Mon. Weather Rev.* **2006**, *134*, 3209–3229. [CrossRef]
35. Hopson, T.; Webster, P.J. A 1-10-day ensemble forecasting scheme for the major river basins of Bangladesh. *J. Hydrometeorol.* **2010**, *11*, 618–641. [CrossRef]
36. Wilks, D.S. *Statistical Methods in the Atmospheric Sciences*; Elsevier Academic Press: Amsterdam, The Netherlands, 2005; p. 648.
37. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [CrossRef]
38. Epstein, E.S. Stochastic-dynamic prediction. *Tellus* **1969**, *21*, 739–759. [CrossRef]
39. Sampson, C.R.; Franklin, J.L.; Knaff, J.A.; DeMaria, M. Experiments with a simple tropical cyclone intensity consensus. *Weather Forecast.* **2008**, *23*, 304–312. [CrossRef]
40. Eckel, F.A.; Mass, C.F. Aspects of effective mesoscale, short-range ensemble forecasting. *Weather Forecast.* **2005**, *20*, 328–350. [CrossRef]
41. Berner, J.; Ha, S.Y.; Hacker, J.P.; Fournier, A.; Snyder, C. Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Weather Rev.* **2011**, *139*, 1972–1995. [CrossRef]
42. Jankov, I.; Berner, J.; Beck, J.; Jiang, H.; Olson, J.B.; Grell, G.; Smirnova, T.G.; Benjamin, S.G.; Brown, J.M. A Performance Comparison between Multiphysics and Stochastic Approaches within a North American RAP Ensemble. *Mon. Weather Rev.* **2017**, *145*, 1161–1179. [CrossRef]
43. Pu, Z.; Zhang, S.; Tong, M.; Tallagragada, V. Influence of the self-consistent regional ensemble background error covariance on hurricane inner-core data assimilation with the GSI-based hybrid system for HWRF. *J. Atmos. Sci.* **2016**, *73*, 4911–4925. [CrossRef]
44. Tong, M.; Sippel, J.A.; Tallapragada, V.; Liu, E.; Kieu, C.; Kwon, I.; Wang, W.; Liu, Q.; Ling, Y.; Zhang, B. Impact of Assimilating Aircraft Reconnaissance Observations on Tropical Cyclone Initialization and Prediction Using Operational HWRF and GSI Ensemble-Variational Hybrid Data Assimilation. *Mon. Weather Rev.* **2018**, *146*, 4155–4177. [CrossRef]
45. Lu, X.; Wang, X. Improving Hurricane Analyses and Predictions with TCI, IFEX Field Campaign Observations, and CIMSS AMVs Using the Advanced Hybrid Data Assimilation System for HWRF. Part I: What is Missing to Capture the Rapid Intensification of Hurricane Patricia (2015) when HWRF is already Initialized with a More Realistic Analysis? *Mon. Weather Rev.* **2019**, *147*, 1351–1373.
46. Hamill, T.M.; Whitaker, J.S.; Fiorino, M.; Benjamin, S.J. Global ensemble predictions of 2009's tropical cyclones initialized with an ensemble Kalman filter. *Mon. Weather Rev.* **2011**, *139*, 668–688. [CrossRef]