Improving Tropical Cyclone Intensity Forecasts with PRIME

KIERAN T. BHATIA AND DAVID S. NOLAN

Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida

ANDREA B. SCHUMACHER

Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado

MARK DEMARIA

NOAA/National Hurricane Center, Miami, Florida

(Manuscript received 2 February 2017, in final form 14 April 2017)

ABSTRACT

The Prediction of Intensity Model Error (PRIME) forecasting scheme uses various large-scale meteorological parameters as well as proxies for initial condition uncertainty and atmospheric flow stability to provide operational forecasts of tropical cyclone intensity forecast error. PRIME forecasts of bias and absolute error are developed for the Logistic Growth Equation Model (LGEM), Decay Statistical Hurricane Intensity Prediction Scheme (DSHP), Hurricane Weather Research and Forecasting Interpolated Model (HWFI), and Geophysical Fluid Dynamics Laboratory Interpolated Hurricane Model (GHMI). These forecasts are evaluated in the Atlantic and east Pacific basins for the 2011–15 hurricane seasons. PRIME is also trained with retrospective forecasts (R-PRIME) from the 2015 version of each model. PRIME error forecasts are significantly better than forecasts that use error climatology for a majority of forecast hours, which raises the question of whether PRIME could provide more than error guidance. PRIME bias forecasts for each model are used to modify intensity forecasts, and the corrected forecasts are compared with the original intensity forecasts. For almost all basins, forecast intervals, and versions of PRIME, the bias-corrected forecasts achieve significantly lower errors than the original intensity forecasts. PRIME absolute error and bias forecasts are also used to create unique ensembles of the four models. These PRIME-modified ensembles are found to frequently outperform the intensity consensus (ICON), the equally weighted ensemble of DSHP, LGEM, GHMI, and HWFI.

1. Introduction

An accurate forecast of a major tropical cyclone (TC) landfall represents one of the most remarkable feats of the earth sciences. Forecasting agencies can now produce skillful 120-h forecasts of the intensity, timing, and location of a TC making landfall. These long-range TC forecasts appear particularly impressive within the context of other natural disasters, such as earthquakes, tornadoes, tsunamis, and volcanic eruptions, which can only be diagnosed hours or sometimes seconds beforehand. The ability to provide track and wind speed predictions at longer forecast times has prompted emergency managers, public officials, businesses, and citizens to make numerous costly and life-changing decisions several days in advance of landfall. As a result, the largest evacuations (between 1.5 and 3 million people) in U.S. history[1] were all triggered by forecasts of potential TC landfalls: Hurricanes Rita, Floyd, Georges, and Gustav (Urbina and Wolshon 2003; Cutter and Smith 2009; Litman 2006). Based on the influence of these forecasts on their end users, it is imperative that they are reliable and accurate.

At first glance, the recent efforts of the scientific community (Gall et al. 2013) to lower the average errors of Atlantic and east Pacific intensity (Bhatia and Nolan 2015, hereafter BN15; DeMaria et al. 2014) and track (Cangialosi and Franklin 2016; see review in

Corresponding author: Kieran T. Bhatia, kbhatia@princeton.edu

[1] Official evacuation numbers for Hurricane Matthew in October 2016 were not available at the time this manuscript was submitted.

DOI: 10.1175/WAF-D-17-0009.1

introduction of DeMaria et al. 2014) model guidance appear successful. However, TC intensity models have only improved at about ⅓ to ½ of the rate observed for the track models between 24 and 72 h and are lagging at longer forecast hours as well (DeMaria et al. 2014). The lower rate of improvement for intensity forecast guidance is visible in the National Hurricane Center (NHC) official (OFCL) forecast verification statistics. From 2007 to 2015, OFCL track forecasts have approximately averaged, depending on the basin forecast hour, 50%–250% more skill than OFCL intensity forecasts, when comparing each set of forecasts to their respective benchmark models (Cangialosi and Franklin 2016; Cangialosi and Franklin 2013; Franklin 2010). Although some studies conclude that the plateauing intensity forecast performance could be the result of a 2–3-day intrinsic predictability limit for TC intensity (Hakim 2013; Brown and Hakim 2013), recent work by Emanuel and Zhang (2016) suggests that intensity forecasts could be more skillful out to 7 days. Additionally, Judt et al. (2016) showed that the large-scale components of Hurricane Earl's wind field were skillfully forecasted up to 7 days in advance.

The purpose of this study is to reduce the gap between current intensity forecast performance and the theoretical limits proposed by Judt et al. (2016) and Emanuel and Zhang (2016). Specifically, we try to lower the absolute error (AE) of the intensity forecasts produced by the Decay Statistical Hurricane Intensity Prediction Scheme (DSHP), Logistic Growth Equation Model (LGEM), Geophysical Fluid Dynamics Laboratory Interpolated Hurricane Model (GHMI), and Hurricane Weather Research and Forecasting Interpolated Model (HWFI) and outperform the equally weighted ensemble of the four models, intensity consensus (ICON), in the Atlantic and east Pacific basins. To achieve this goal, the meteorological research community typically recommends finer spatial and temporal resolution for dynamical models, more advanced data assimilation techniques, and the acquisition of additional observations by deploying more instrumentation (Zhang et al. 2011). Forecast postprocessing such as bias corrections and ensembles are inexpensive solutions that can supplement these intensity forecast improvement techniques. Our forecast modification strategies differ considerably from other statistical studies because they are situation-dependent, defined by the nature of the TC and its surroundings.

Bhatia and Nolan (2013, hereafter BN13) provided the foundation for this research by demonstrating that the high variance in intensity forecast performance between different storms, models, and days is dependent on TC attributes and synoptic conditions. BN13 concluded that the average forecast error of a model often served as a poor guide for how an individual forecast would perform and suggested that certain variables could potentially anticipate high and low error forecasts. Based on the results of BN13, BN15 developed the Prediction of Intensity Model Error (PRIME) model to forecast the bias and AE of DSHP, LGEM, GHMI, and HWFI intensity forecasts in the Atlantic basin. In addition to the dynamical parameters considered in BN13, proxies for atmospheric flow stability and initial condition error served as predictors in PRIME's stepwise multiple linear regression formula. Independent verification of PRIME predictions of AE and bias revealed PRIME had significantly[2] lower errors than climatological forecasts[3] for all forecast hours and models from 2007 to 2014. A second version of PRIME[4] called Retrospective PRIME (R-PRIME) was developed using the retrospective forecasts of models that were operational during 2014. The R-PRIME model also performed very well, and the AE of its error forecasts were lower than PRIME.

Our work updates BN15 by adding the 2015 hurricane season to the PRIME data sample and developing PRIME for the east Pacific basin. The following section shows PRIME and R-PRIME's performance in both basins from 2011 to 2015. For all basins, versions of PRIME, and models, PRIME outperforms climatological forecasts of bias and AE. The encouraging performance of PRIME indicates its output could not only be used as a forecasting supplement that quantifies how much confidence should be placed in an intensity forecast but also as an intensity forecast improvement tool. Section 3 explores different ways PRIME forecasts can be manipulated to increase the accuracy of intensity forecasts. First, PRIME bias forecasts are tested as corrections to intensity forecasts. The errors of the bias-corrected models are then compared with the errors of the original models to determine whether TC intensity forecasts are significantly improved. The second part of section 3 is devoted to ensembles that are modified based on PRIME error forecasts. Seven ensembles composed of DSHP, LGEM, GHMI, and HWFI are

[2] Unless specified, significance in this study means a paired t test [e.g., Eq. 5.11 from Wilks (2011)], adjusted for serial correlation, outputs a 95% likelihood that the null hypothesis is invalid.

[3] Note that here the term climatological forecasts refers to the prediction of intensity forecast errors based on a multiyear average of past forecast performance.

[4] Used throughout the manuscript as an umbrella term for all forecasts created by the PRIME forecasting scheme. Unless contextually specified, "PRIME" refers to both the PRIME and R-PRIME models.

TABLE 1. Dynamical and proxy predictors for PRIME. Abbreviations are listed for each predictor. Boldface predictor abbreviations indicate predictors whose 0-h and forecast average (with a letter A added as a prefix to the abbreviations in the manuscript) values are both used. If the predictor value varies depending on the model, a Y is listed in the third column. An N is used if the same GFS output from the SHIPS files is used to produce the predictor for every model.

Dynamical predictors	Abbrev	Changes with model? (Y/N)	Proxies	Abbrev	Changes with model? (Y/N)
Percentage of area of GOES cold pixels	GCLD	N	Std dev of the intensity forecast ensemble	SPRD	N
Std dev of GOES brightness temperature	GBRT	N	Deviation of intensity forecast from ensemble mean	DFEM	Y
850–200-hPa shear magnitude	**SHR**	N	Absolute DFEM	ADEM	Y
Storm speed	**SSPD**	N	Deviation of track forecast from ensemble mean	DTRK	Y
Sin(850–200-hPa shear direction)	**SHRDIR**	N	Forecasted intensity change	FIC	Y
Ocean heat content	**OHC**	N	Absolute FCIC	AFIC	Y
Potential intensity	**POT**	N	Previous 12-h intensity change	P12C	Y
850-hPa vorticity	**VOR**	N	Previous 12-h error	P12E	Y
200-hPa divergence	**DIV**	N	Distance to land	**LDIS**	N
850–700-hPa relative humidity	**RH**	N	Forecasted distance to land	FLND	Y
Initial intensity	0INT	Y			
Forecast intensity	FINT	Y			
Latitude	**LAT**	N			
Longitude	**LON**	N			

assembled and their skill score (SS) relative to ICON is presented. Finally, conclusions and future work are provided in section 4.

2. Updated PRIME and R-PRIME

BN15 showed R-PRIME and PRIME were able to skillfully predict the AE and bias of DSHP, LGEM, GHMI, and HWFI intensity forecasts over a 7- and 8-yr independent verification sample. PRIME was subsequently tested as an operational forecasting tool to communicate the expected error of TC intensity forecasts. R-PRIME was utilized to produce error forecasts for the 2015 Atlantic basin hurricane season, which were available at the Cooperative Institute for Research via the Atmosphere real-time products web page.[5] The methodology and data sources used in BN15 were followed to create operational PRIME forecasts. Table 1, copied from the corresponding table in BN15, lists all the predictors considered for the creation of operational PRIME. The second-order polynomial transformation of the initial distance to land (LDIS) predictor was added to the bias predictor pool because it was the only nonlinear predictor, for either predictand, that improved the forecasts of all models, forecast intervals, and versions of PRIME. Unfortunately, the error statistics for the 2014 version of the models were used to develop operational R-PRIME because the 2015 retrospective data were not available.[6] The outdated retrospective data coupled with the small sample size for the 2015 Atlantic basin hurricane season led to a lack of statistically significant results. As a result, the verification figures for the operational error forecasts in 2015 are omitted here.

This study focuses on the performance of PRIME during 2011–15. Retrospective runs, created by applying the 2015 version of each model to storms during 2011–14, are used to train R-PRIME. The length of the training period is shortened compared with BN15, because fewer seasons of retrospective forecasts were produced for GHMI and HWFI. In addition to the Atlantic basin, PRIME is also developed for the east Pacific basin during this time frame. The same statistical framework outlined in BN15 is used to produce PRIME forecasts of AE and bias for DSHP, LGEM, GHMI, and HWFI in both basins. Optimal predictors are selected

[5] Information online at http://rammb.cira.colostate.edu/products/tc_realtime/. For each storm, choose "Model Products" and scroll down to the bottom of the web page.

[6] Obtaining the appropriate retrospective forecasts will not be an issue in subsequent years.

TABLE 2. The average bias and AE (kt, where 1 kt = 0.51 m s^{-1}) for HWFI, GHMI, LGEM, and DSHP are calculated using the operational forecasts of Atlantic basin storms between 2011 and 2015. The first column represents the number of verified real-time forecasts. For each model and forecast interval, the first entry in a table cell is the bias and the second entry is the absolute error.

		BIAS, AE (kt)			
No. of cases	No. of hours	HWFI	GHMI	LGEM	DSHP
1219	12	−0.6, 6.2	−0.6, 6.7	−0.8, 6.2	−0.3, 6.1
1093	24	−0.7, 8.8	−1.8, 9.7	−0.7, 9.1	0.8, 9.0
964	36	0.4, 10.9	−1.7, 12.4	−0.3, 11.2	2.0, 11.0
841	48	1.9, 12.6	0.4, 14.1	0.0, 13.2	2.9, 12.9
728	60	3.4, 14.2	2.6, 15.6	0.1, 14.5	3.5, 14.3
630	72	4.4, 15.7	4.9, 17.6	0.4, 15.2	3.8, 15.1
538	84	5.0, 16.7	7.2, 19.1	1.1, 15.5	3.8, 15.5
460	96	5.0, 17.7	8.6, 20.7	1.7, 16.2	3.2, 16.0
397	108	5.2, 19.4	10.3, 22.1	1.9, 16.7	2.3, 16.8
348	120	5.7, 21.0	12.2, 23.5	1.1, 17.2	0.6, 16.7

from Table 1 for each model, version of PRIME, basin, and predictand. Unlike BN15, where nonlinear functions are used to modify certain independent variables, no fitted predictors are considered here because operational convenience and reproducibility is prioritized. PRIME forecasts are only computed for a particular forecast time if the TC is at least a tropical or subtropical depression initially and at verification, all four models have an intensity forecast available in the National Oceanographic and Atmospheric Administration's (NOAA) Automated Tropical Cyclone Forecast (ATCF; Sampson and Schrader 2000) "a-deck" files, and a verifying TC intensity is available in the best-track dataset (Landsea and Franklin 2013). These verification rules result in a dataset that is completely homogeneous among the models. Tables 2–5 examine the forecasts from each model that meet these criteria.

Tables 2 and 3, respectively, list the AE, bias, and sample size for the real-time intensity forecasts and retrospective forecasts of DSHP, LGEM, GHMI, and HWFI in the Atlantic basin. Tables 4 and 5 contain the corresponding information for the east Pacific basin. As in BN15, real-time cases are added to the retrospective cases to increase the sample sizes in Tables 3 and 5. GHMI retrospective data for the 2013 hurricane season are missing in the Atlantic basin, and GHMI retrospective data were not generated in the east Pacific basin during 2011 and 2013. The real-time GHMI forecasts from the missing years are added to the sample because the other three models produced retrospective forecasts during these years. PRIME forecasts are only issued when all four models have an intensity forecast available, so the additional GHMI forecasts enhanced the sample size of all models. This expanded dataset led to better results in both basins so all the presented R-PRIME analyses use an augmented sample.

There are still slightly fewer cases in the retrospective sample compared with the real-time sample, because retrospective HWFI forecasts are only generated when tail-wind Doppler radar and retrospective Global Forecast System (GFS) data are available. However, the overall trends in the real-time model performance are almost identical when considering only the cases that are available for both the retrospective and real-time forecasts. In the Atlantic basin, recent upgrades to HWFI

TABLE 3. As in Table 2, but for entries that correspond to the statistics of verified retrospective forecasts for each model at each forecast interval. These values are calculated using Atlantic basin storms between 2011 and 2015.

		BIAS, AE (kt)			
No. of cases	No. of hours	HWFI	GHMI	LGEM	DSHP
931	12	−1.1, 5.8	−0.1, 6.4	−0.8, 6.2	−0.3, 6.1
838	24	−1.3, 8.1	−0.1, 8.8	−0.7, 8.8	0.9, 8.9
759	36	−1.2, 9.9	0.6, 11.2	−0.6, 10.8	1.8, 11.0
677	48	−1.5, 10.8	1.2, 12.9	−0.5, 12.3	2.5, 12.8
595	60	−1.4, 11.3	1.8, 14.2	−0.9, 13.2	3.1, 14.1
524	72	−1.2, 11.7	2.6, 15.6	−1.0, 13.5	3.5, 14.6
462	84	−0.4, 12.0	3.8, 16.6	−0.4, 13.7	4.0, 15.0
401	96	−0.2, 12.7	5.0, 18.2	−0.1, 14.7	3.8, 15.9
348	108	0.0, 14.0	6.9, 19.7	−0.1, 15.5	3.1, 16.6
306	120	1.1, 15.3	8.4, 20.9	−0.8, 16.3	2.0, 16.7

TABLE 4. As in Table 2, but for the average bias and AE calculated using east Pacific basin storms between 2011 and 2015.

BIAS, AE (kt)					
No. of cases	No. of hours	HWFI	GHMI	LGEM	DSHP
1634	12	−2.4, 7.3	−3.1, 8.0	−1.6, 7.2	−1.2, 7.0
1451	24	−4.2, 11.4	−6.8, 12.8	−3.4, 11.5	−1.6, 11.2
1274	36	−5.6, 14.3	−9.3, 16.5	−5.2, 14.5	−2.0, 13.8
1103	48	−6.5, 16.5	−8.8, 18.1	−6.4, 16.8	−2.4, 15.4
954	60	−6.5, 17.6	−7.2, 18.6	−6.9, 17.8	−2.7, 16.2
817	72	−5.9, 17.9	−5.0, 18.9	−6.6, 17.7	−2.5, 16.5
690	84	−5.9, 18.3	−3.1, 20.0	−6.1, 17.6	−2.4, 16.5
575	96	−6.2, 18.4	−2.5, 20.4	−6.0, 17.4	−1.9, 17.1
471	108	−5.0, 18.7	−1.8, 20.4	−5.7, 17.0	−1.7, 17.1
379	120	−3.8, 18.6	−1.6, 19.7	−5.9, 15.9	−0.8, 16.4

(Atlas et al. 2015) have resulted in forecasts with significantly lower AE and biases. LGEM, GHMI,[7] and DSHP Atlantic basin retrospective forecasts also displayed lower AE than the real-time forecasts but not to the same degree as HWFI. For all models, the largest improvements are observed for longer forecast hours. The models exhibit similar behavior in the east Pacific basin, but the retrospective forecasts show less improvement over the real-time forecasts. Additionally, LGEM, GHMI, and HWFI 12–60-h forecasts have large negative biases for both the real-time and retrospective forecasts whereas in the Atlantic basin, the biases are mainly positive and smaller.

As in BN15, our primary concern is providing insight on how PRIME will perform in an operational setting, so all error forecasts are independently verified using cross validation (Wilks 2011). For our cross validation, all but one of the years are used as the training data, and then the excluded year is used for validation; this procedure is repeated for all years. Figures 1–4 show the average AE of R-PRIME, PRIME, and climatological forecasts of AE and bias in the Atlantic and east Pacific basin from 2011 to 2015. The sample size is limited to only include the cases available for both PRIME and R-PRIME. For almost all forecast intervals, models, predictands, and basins, R-PRIME outperformed PRIME,[8] and both versions of PRIME had smaller errors than their respective climatological forecasts.

Figure 1 shows the 2011–15 average AE of R-PRIME, PRIME, and climatological AE predictions in the Atlantic basin. The number of cases for each forecast interval is approximately equal to those listed in Table 3. For all models and versions of PRIME, the AE of 0–96-h forecasts[9] is significantly less than the AE of the corresponding climatological forecasts. Additionally, GHMI PRIME and R-PRIME AE forecasts are significantly better than climatological forecasts at all forecast hours. Figure 2 is similar to Fig. 1 except it shows the average AE of PRIME and R-PRIME bias predictions compared with climatological bias forecasts. Besides 96–120-h R-PRIME forecasts of HWFI bias, PRIME and R-PRIME bias forecasts are significantly better than the corresponding climatological forecasts. As expected from Tables 2 and 3, switching from real-time to retrospective models improved the climatological and PRIME error forecasts for the dynamical models more than the statistical models. The large modifications to the model configurations of real-time HWFI and GHMI over the time series caused inconsistent forecast performance, which makes it more difficult for climatological forecasts to anticipate error based on past statistics. At the same time, PRIME suffers because the derived predictor–error relationships for these models are not representative of the whole forecast sample. DSHP and LGEM have more consistent formulations throughout the sample, which results in PRIME and R-PRIME behaving similarly for these models.

Compared with the results in BN15, the skill of R-PRIME and PRIME forecasts diminished. There are three main reasons for this behavior. First, we selected a more current time frame for the analysis of PRIME, even though all of the evaluated models have produced real-time intensity forecasts since 2007. The models that produce the real-time forecasts from 2011 to 2014 more closely resemble the 2015 version of the models, so PRIME performance becomes more similar to R-PRIME. The real-time forecasts are formulated more consistently

[7] GHMI retrospective results are negatively affected by the addition of real-time cases.

[8] Except for 24-h forecasts of HWFI AE in the east Pacific basin.

[9] Except for 96-h R-PRIME forecasts of HWFI.

TABLE 5. As in Table 4, but these values are calculated using retrospective forecasts for east Pacific basin storms between 2011 and 2015.

		BIAS, AE (kt)			
No. of cases	No. of hours	HWFI	GHMI	LGEM	DSHP
1359	12	−3.4, 7.6	−3.0, 8.1	−1.6, 7.1	−1.2, 7.1
1213	24	−6.1, 11.7	−6.4, 13.0	−3.4, 11.2	−1.6, 11.4
1068	36	−8.0, 14.4	−8.4, 16.2	−4.8, 14.0	−1.7, 14.1
931	48	−8.6, 16.0	−7.8, 17.6	−5.8, 16.0	−1.7, 15.8
800	60	−8.7, 16.7	−6.2, 18.4	−6.1, 17.2	−1.2, 16.8
687	72	−7.5, 16.0	−3.4, 18.4	−5.5, 16.9	0.0, 16.8
579	84	−6.4, 15.2	−0.3, 19.1	−5.1, 16.3	1.2, 16.3
485	96	−5.7, 14.6	1.2, 19.3	−5.1, 15.7	1.8, 16.6
404	108	−4.9, 15.0	1.2, 19.4	−5.5, 15.2	1.7, 16.2
331	120	−4.3, 15.5	0.6, 19.0	−6.5, 14.4	1.7, 15.5

in a shorter and more current analysis sample, which leads to better climatological error forecasts. At the same time, the shorter training period can cause both versions of PRIME to regress, because robust statistical relationships are harder to develop with fewer cases.

Second, the gap between GHMI performance and the rest of the models widened. Tables 2 and 3 reveal that GHMI is the worst-performing model at almost[10] every forecast hour for both real-time and retrospective forecasts, and at some forecast hours, the next worst model is over 25% better than GHMI. Deviation of track forecast from the ensemble mean (DTRK), deviation of intensity forecast from the ensemble mean (DFEM), absolute DFEM (ADEM), and standard deviation of the ensemble of intensity forecasts (SPRD) are predictors that rely on all four evaluated models, and their relationships with bias and AE are predicated on the fact that larger deviations from the ensemble mean are damaging to forecast accuracy. However, in recent years, it has often been advantageous to deviate from GHMI, so these predictors are less effective. The third explanation for the slumping performance of PRIME is the lack of fitted predictors. Depending on the forecast hour, adding the fitted predictors to the predictor pool can improve PRIME skill over 10%. These fitted predictors are omitted because one of the goals of this manuscript is to demonstrate the effectiveness of PRIME using a simple methodology.

Figure 3 is similar to Fig. 1 but applies to the east Pacific basin. For all models and versions of PRIME, 12–48-h AE forecasts have significantly lower AE than climatological AE forecasts. In general, PRIME AE forecasts for dynamical models are more skillful than those for statistical models. For GHMI and HWFI, both versions of PRIME are significantly better than their respective climatological forecasts for all forecast intervals except 120-h R-PRIME forecasts of HWFI AE (significant at the 90% level). LGEM only has one other forecast interval outside of 12–48 h, 60-h R-PRIME forecasts, where a version of PRIME has significantly lower average AE than climatological forecasts, and DSHP has no other significant results.

Figure 3 reveals that the AE of PRIME AE forecasts between 12 and 60 h are negatively correlated with forecast length, which is opposite of the relationship beyond 60 h. The accuracy of the models' intensity forecasts also follows this trend, which lowers the variance and mean AE of the forecast sample for longer forecast intervals. This reversal in model performance with forecast length is not observed in the Atlantic basin, and it opposes conventional wisdom that uncertainty increases with forecast horizon (Lorenz 1963). With less variability in forecast performance and better forecasts, forecast error becomes increasingly dominated by noise. In other words, random errors, originating from imprecise best-track intensities (Landsea and Franklin 2013) and the chaotic nature of the atmosphere, become a larger percentage of total error, and PRIME is less successful at explaining the variance of the predictand.

Potential explanations for this unique model behavior involve the strong sea surface temperature (SST) gradient that exists off the west coast of North America [comparison to other basins' development regions first shown in Fig. 18 of Gray (1968)] and the extreme shear surrounding the island of Hawaii [highlighted in Fig. 25 of Gray (1968)]. Unlike the Atlantic basin, hospitable shear values and SSTs for TCs are present over a small portion of the east Pacific basin during the months of peak TC formation. There are three potential pathways through which these environmental conditions in the east Pacific basin could lead to better long-range intensity forecasts.

1) A large portion of the verified long-range forecasts involve TCs that are weakening because of cool ocean waters. In these situations, it is hypothesized that some of the poorly understood mechanisms controlling TC intensity do not need to be resolved

[10] Except for 24-h retrospective forecasts.

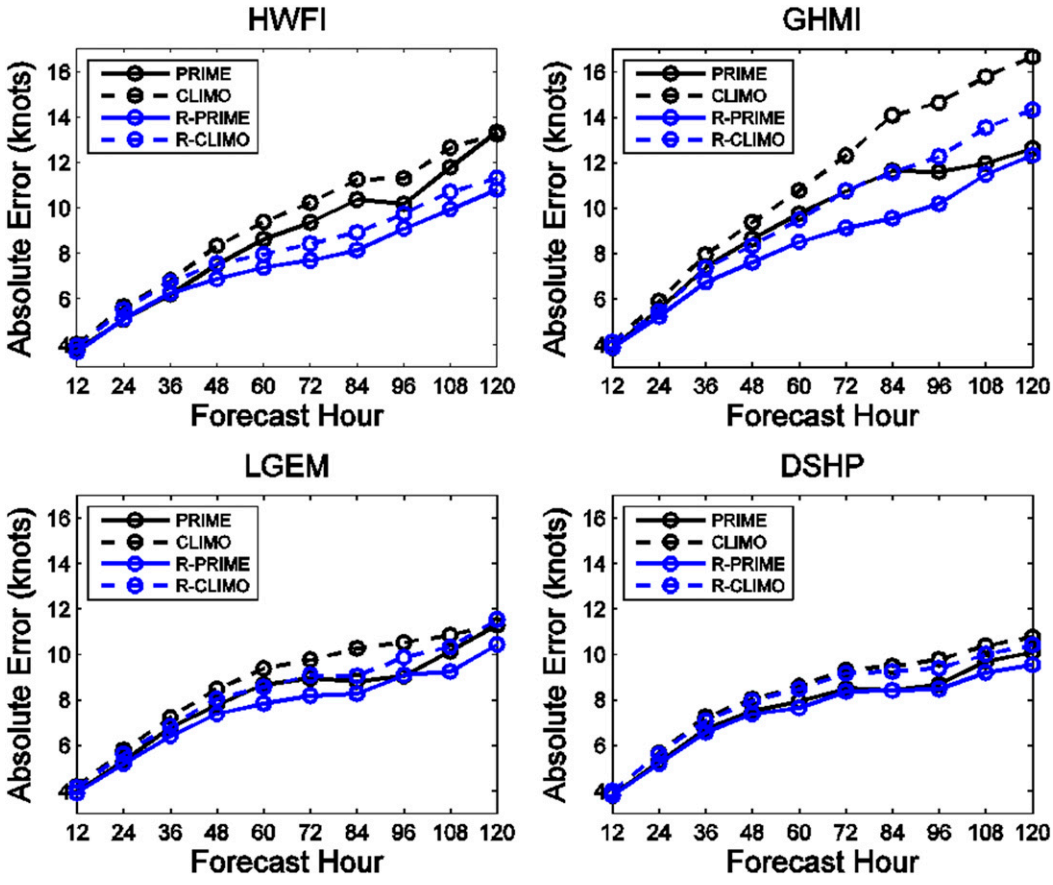

FIG. 1. The average AE of PRIME AE forecasts, R-PRIME AE forecasts, and CLIMO and R-CLIMO AE forecasts for HWFI, GHMI, LGEM, and DSHP. In each model subplot, the blue and black lines indicate data originating from R-PRIME and PRIME, respectively. The solid lines illustrate PRIME results, while the dashed lines represent climatological results. Error statistics are calculated using cases from 2011 to 2015 in the Atlantic basin that are available for both PRIME and R-PRIME.

for forecast accuracy. For example, capturing small-scale dynamical processes or shear interactions (Tao and Zhang 2014; Tao and Zhang 2015; Judt et al. 2016) with the TC moisture field might not be necessary when SST is such an overwhelming negative deterrent for intensification.

2) In the east Pacific basin, rapid weakening is often observed immediately following rapid intensification (RI), which could be another reason for the unique AE trend for long forecast intervals. The combination of land, hostile SSTs, and high shear values provides multiple barriers for TCs to maintain major hurricane status in the east Pacific basin. As a result, forecasts that incorrectly predict constant intensity throughout the forecast period can be inaccurate at shorter forecast intervals but fairly accurate for longer forecast intervals. Two recent examples from the 2015 season are Hurricane Andres and Hurricane Hilda. Figure 4 shows the intensity forecasts of LGEM and HWFI[11] as well as the 0-h intensity and track estimates for each storm. Both TCs exhibited a period of RI followed immediately by a period of rapid weakening because of either declining SSTs (Andres) or rising shear (Hilda). This storm behavior resulted in lower errors for the long-range forecasts and higher errors for the short-range forecasts.
3) Verification rules used here and followed by NHC prevent some of the long-range forecasts of TCs from being included in the forecast sample. When a TC travels through high shear, over land, or above very cold ocean waters, it typically decays and loses its tropical attributes within 1 or 2 days. At this

[11] These models were selected because they are the best-performing statistical and dynamical model, respectively. DSHP and GHMI display similar behavior but are omitted to avoid additional clutter in the figure.

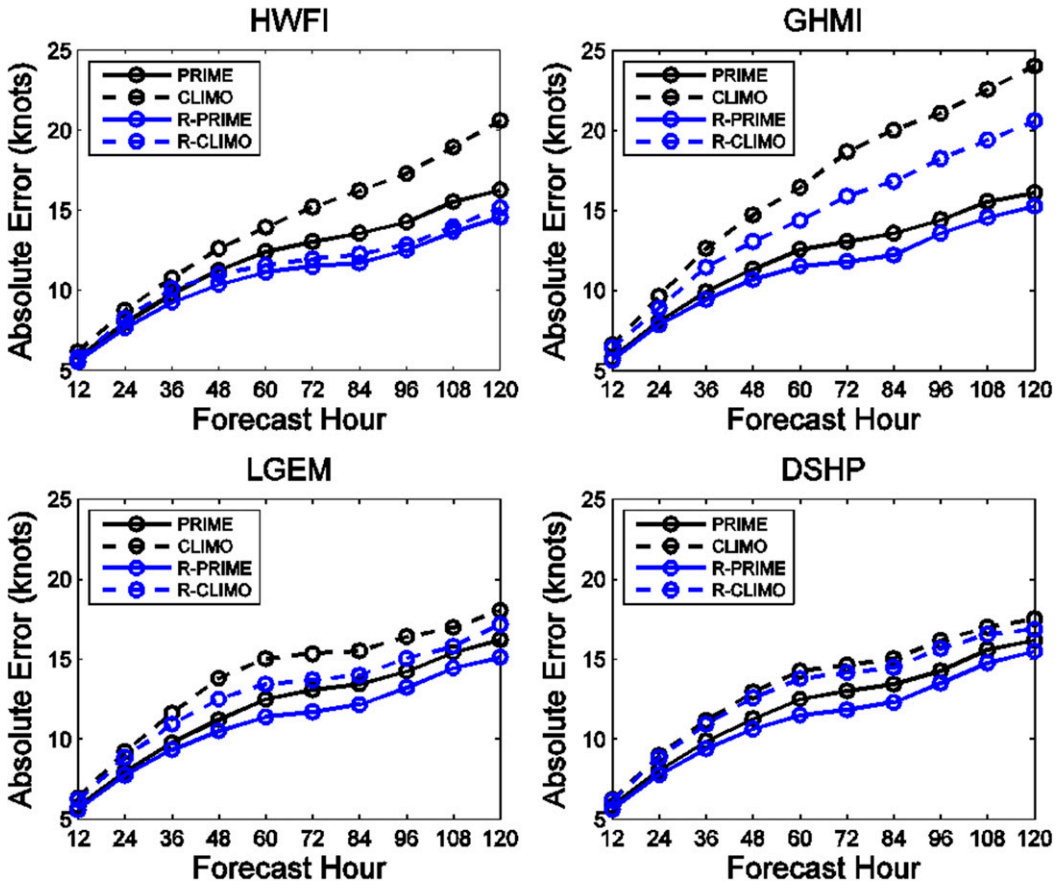

FIG. 2. As in Fig. 1, but the plots show the average AE of PRIME bias forecasts, R-PRIME forecasts, and CLIMO and R-CLIMO bias forecasts.

point, a TC will be listed as a "LO" (low pressure system) or not listed in the best-track file, which excludes these storms from verification. Hurricane Patricia in 2015 and Hurricane Jova in 2011 are just two examples of TCs whose long-range forecasts with large errors never verified even though their short-range forecasts are included in the analysis.

PRIME bias forecasts in the east Pacific basin are on average more skillful than PRIME AE forecasts for every model and version of PRIME. Figure 5 shows that the AE of the bias forecasts for both versions of PRIME is significantly less than the associated climatological forecasts at all forecast hours. As in Fig. 3, PRIME behavior for 72–120-h forecasts appears quite different to shorter forecast intervals. The relatively constant size and variance of the predictand for longer forecast intervals results in the performance of PRIME and the climatological forecasts barely changing for 72–120-h forecasts. This interesting trend is largely attributable to many of the same mechanisms discussed in the previous paragraphs.

A potential way to increase the value of PRIME while still incorporating a simple methodology is to combine the information contained in its bias and AE forecasts into one consistent message for end users. Figure 6 shows the SS of R-PRIME forecasts in the Atlantic basin for each model based on whether the R-PRIME AE and bias forecasts agree. SS essentially normalizes PRIME errors with climatology errors and is defined as

$$\mathrm{SS} = 100 \times \left[1 - \left(\frac{E_{\mathrm{PRIME}}}{E_{\mathrm{CLIMO}}}\right)\right], \qquad (1)$$

where E is the average error from PRIME or climatology at a given forecast interval. A positive SS represents an improvement upon climatology, with the highest SS being 100%. In Fig. 6, verified forecasts are partitioned into two main groups: FCST AGREEMENT and FCST DISAGREEMENT cases. Forecasts agree if both the R-PRIME forecast AE and the absolute value of the R-PRIME forecast bias are above or below the climatological

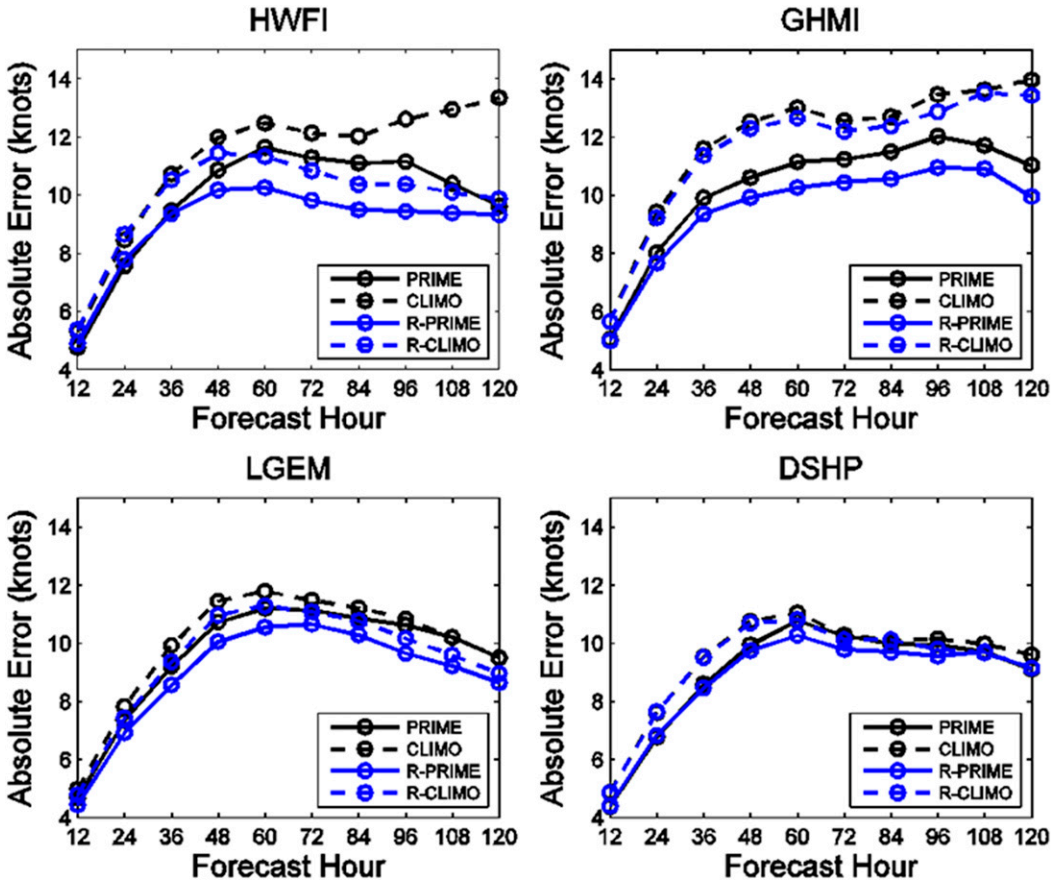

FIG. 3. As in Fig. 1, but for the east Pacific basin.

mean AE for a particular forecast hour–model pair. Besides 120-h HWFI forecasts, R-PRIME AE forecasts perform better when they agree with bias forecasts. By only consulting R-PRIME in these situations, end users could expect AE forecasts with significantly higher skill while sacrificing access to error forecasts for less than 20%–40% (depending on the model and forecast hour) of the total sample size. Results for R-PRIME bias forecasts as well as PRIME AE and bias forecasts are omitted here but show similar results.

Figure 7 shows that R-PRIME AE forecasts in the east Pacific basin are also more skillful when they agree with the simultaneously created R-PRIME bias forecasts. The data in Fig. 7 are formulated and presented in a manner similar to Fig. 6. The FCST AGREEMENT cases typically have their highest SSs for forecast hours and models where R-PRIME AE and bias forecasts are more skillful. In fact, there are multiple forecast hours where the SS of the FCST AGREEMENT cases are over 30% higher than the SS of the FCST DISAGREEMENT cases. Further investigation is needed for determining how dependent these results are on the threshold that defines forecast agreement and disagreement.

PRIME's skill, coupled with its real-time availability, demonstrates it can enhance the value of operational intensity forecasts in the Atlantic and east Pacific basins. For the remainder of this study, PRIME forecasts are tested for their ability to lower TC intensity forecast error.

3. PRIME applications

a. Optimal predictors

In this section, R-PRIME forecasts are applied to the data presented in Tables 3 and 5. Retrospective data are only considered here in order to keep forecast models consistent in the verification and training sample and better represent how error forecasts would operationally perform (BN15). The significant predictors for each model, basin, and predictand are almost identical to those used to create the R-PRIME forecasts in the

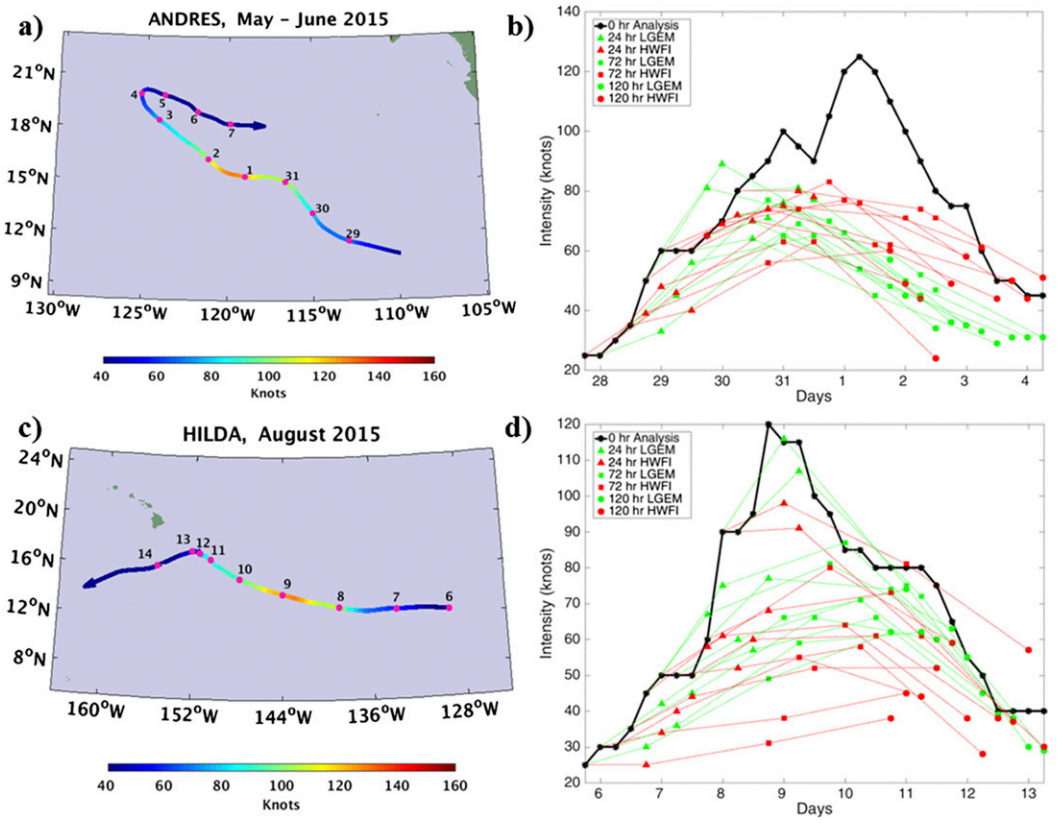

FIG. 4. The track and intensity of (a) Hurricane Andres and (c) Hurricane Hilda during the 2015 east Pacific hurricane season. The black numbers on each map represent the position of the TC at the start of each day for the month(s) listed in the title. The intensity scale is at the bottom of each map. Additionally, 0-h operational intensity estimates (solid black line) as well as the 24-, 72-, and 120-h retrospective HWFI and LGEM intensity forecasts for (b) Andres and (d) Hilda are also plotted. Data points along the black line are only plotted if all three forecasts are generated at the particular time for the two models or if a forecast is verified at the time. The 24-h forecasts are indicated by triangles, 72-h forecasts are indicated by squares, and 120-h forecasts are indicated by circles. HWFI forecasts are colored red, and LGEM forecasts are colored green.

previous section.[12] Tables 6 and 7 specify the three most ''important'' predictors and the sign of their weighting coefficients for each model, basin, and predictand. Here, predictor importance is based on the average magnitude of the weighting coefficients over all forecast intervals and training samples. These averages help highlight the strongest predictor–error relationships because the optimal predictors can change for each training sample (i.e., predicting 2015 from the 2011–14 training sample can yield different optimal predictors than when predicting 2011 from 2012–15), and the weighting of predictors can change with forecast hour. The optimal number of predictors is specified in the table and varies among the models, predictands, and basin. Only three predictors are listed because they are significant at all forecast intervals, and the signs of their coefficients are typically consistent with physical reasoning.

The predictor information in Table 6 applies to R-PRIME forecasts in the Atlantic basin. ADEM, SPRD, forecast intensity (FINT), and forecasted intensity change (FIC) all have positive correlations with AE while forecasted distance to land (FLND) has a negative correlation; the potential mechanisms justifying the signs of the coefficients are outlined in BN15. For all the models, the most accurate R-PRIME bias forecasts are created when DFEM is selected as the one predictor. A majority of the optimal predictors for AE and bias represent the uncertainty in the atmospheric flow pattern, which implies

[12] R-PRIME performance in this section closely resembles the results shown in Figs. 1–3 and 5 but there are minor differences due to relaxing the requirement that both real-time and retrospective forecasts are available.

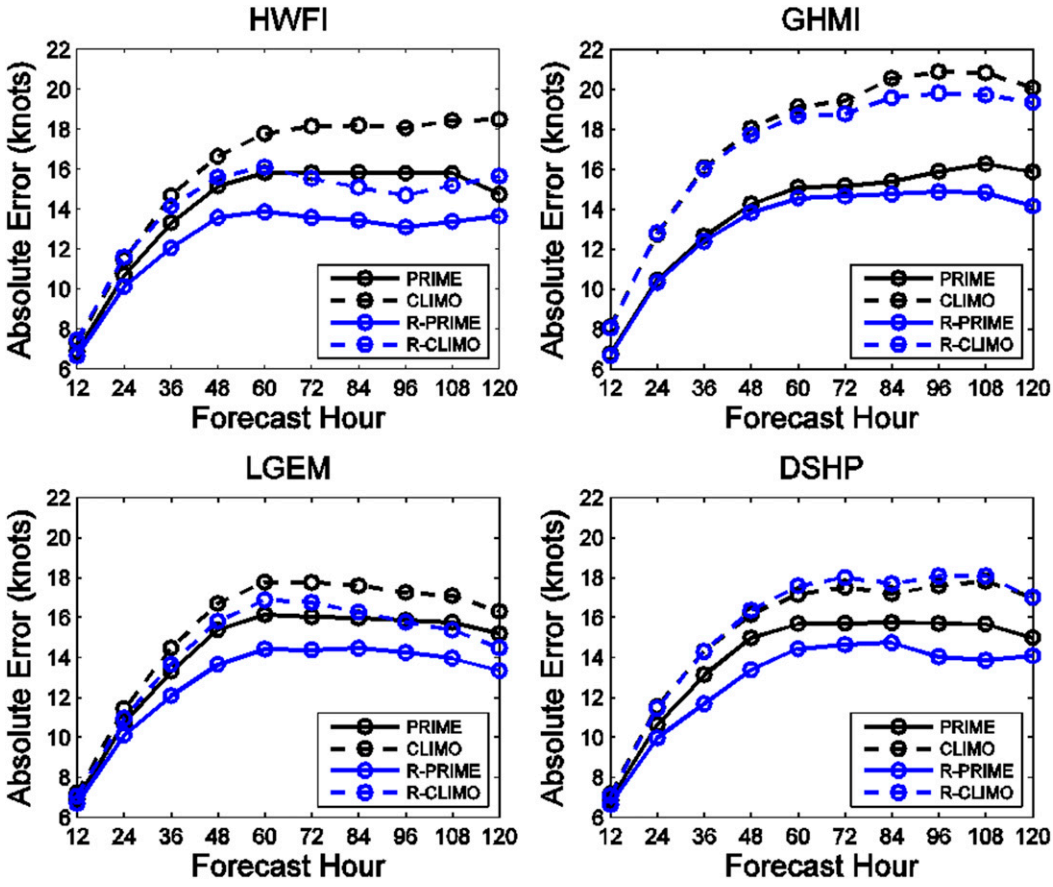

FIG. 5. As in Fig. 3, but for plots showing the average AE of PRIME, R-PRIME, and CLIMO and R-CLIMO bias forecasts.

that the statistical relationships between error and the synoptic predictors noted in BN13 are either not robust for the more recent forecast sample or due to the linear regression framework.

Table 7 highlights the important predictors for R-PRIME AE and bias forecasts in the east Pacific basin. Based on analysis in BN15, most of the AE predictors in Table 7 appear physically justified. Average potential intensity (APOT) and average divergence (ADIV) are introduced as two important predictors that have positive correlations with AE. APOT is also the only bias predictor deemed optimal that was not examined in BN15. The physical reasoning for the relationship of LGEM AE and ADIV is not clear. Initial analysis indicates that the formulation of LGEM is poorly capturing the connection between favorable upper-level conditions and TC intensity, but for a definitive explanation, further investigation is needed. Understanding the relationship of HWFI error and APOT is more straightforward. As discussed in section 2, there is a limited area in the east Pacific basin where SSTs, and thus APOTs, are sufficient for TC intensification. Therefore, TCs that undergo RI in the east Pacific basin avoid the ubiquitous colder SSTs at higher latitudes and are clustered around areas that have higher APOT. HWFI, like all operational models, struggles with predicting RI, which leads to high AEs and large negative biases for storms expected to traverse regions with higher APOT. Additionally, TCs with higher APOT normally have more potential for intensification and, thus, more room for error.

Figure 8 shows the correlations between APOT and HWFI AEs for 48- and 96-h forecasts in the east Pacific basin between 2011 and 2015. If a forecast records positive (negative) bias, the data point is red (black). The R values in the top-right corner of each plot capture the linear correlation of APOT and AE. Both plots show that HWFI forecasts have higher AE results for larger APOT values. Additionally, the grouping of the black dots for the larger AE and APOT values indicates that the HWFI forecasts with the largest AEs are negatively biased. Almost all of these forecasts involved HWFI not anticipating an RI event. Figure 9 contains the same two plots of

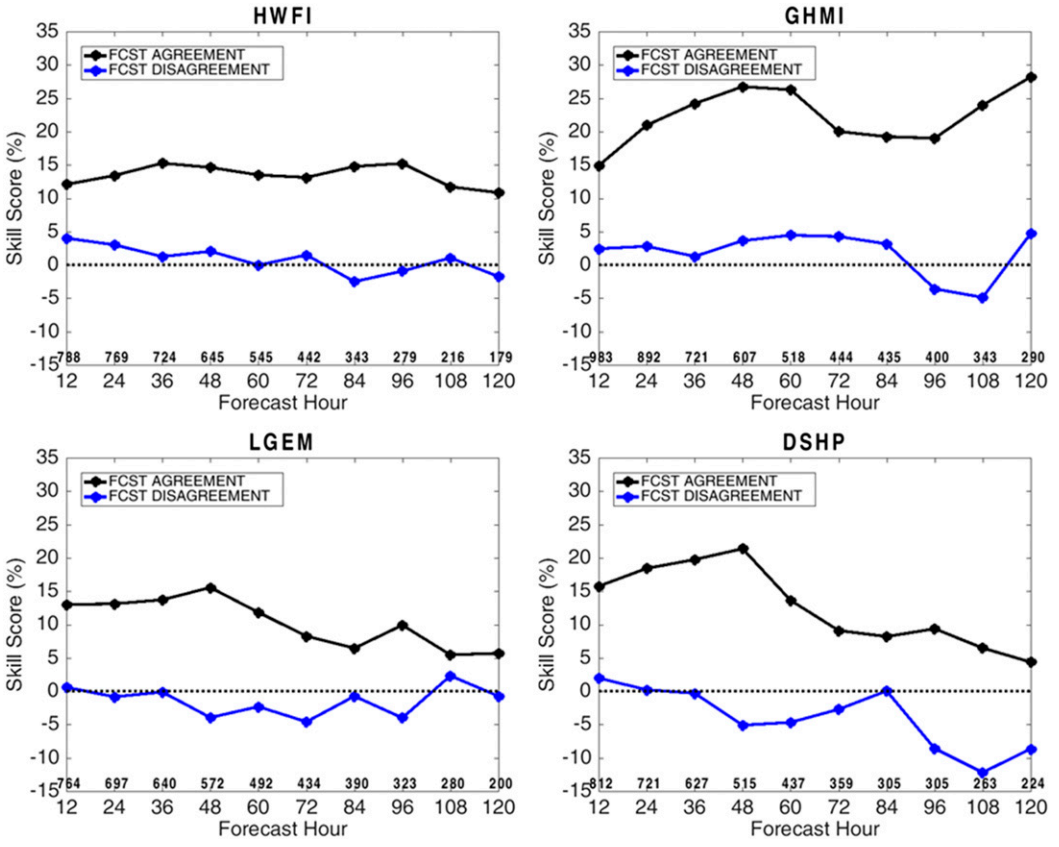

FIG. 6. The performance of R-PRIME AE forecasts in the Atlantic basin for HWFI, GHMI, LGEM, and DSHP when they agree and disagree with bias forecasts. Forecasts are considered FCST AGREEMENT cases if both the forecasted AE and absolute value of the forecasted bias are above or below the climatological mean AE of forecasts for the model–forecast hour combination. Black lines represent FCST AGREEMENT cases and blue lines represent FCST DISAGREEMENT cases. The boldface numbers at the bottom of each plot indicate the number of FCST AGREEMENT cases for each forecast hour.

AE versus APOT but for the Atlantic basin. Similar to the results in the east Pacific, APOT is positively correlated with AE but the correlation coefficients are much smaller. At both forecast times, the red and black dots are scattered with no trend for higher APOT values. The Atlantic basin has more homogeneous APOT values throughout the basin, so APOT is not as adept at diagnosing HWFI forecasts as likely to have high AE and negative bias.

APOT is just one example of a dynamical variable that is highly correlated with error for the R-PRIME forecasts of certain models in only one basin. In general, TC attributes and environmental parameters appear more often as significant predictors in the east Pacific compared with the Atlantic. Although the physical processes controlling intensity change in both basins should remain consistent, there is no theory that links forecast error variability across the global TC basins. This result is particularly important considering the recent idealized model studies that aim to correlate forecast uncertainty to a particular variable like shear, intensity, or moisture (Zhang et al. 2014; Zhang and Tao 2013; Tao and Zhang 2015; Emanuel and Zhang 2016). For the Atlantic basin operational intensity models, there is a lack of strong linear relationships between these types of predictors and error. BN15 and BN13 showed that more robust predictor–error relationships often involve nonlinear functions and capture how multiple predictors can covary with forecast uncertainty. Additionally, the predictor hierarchy of importance changes among basins and models, which implies that studies drawing sweeping conclusions from isolating one variable in one basin with one model could be oversimplifying a complex problem.

b. R-PRIME bias-corrected forecasts

Using the retrospective data and predictors described in the previous section, R-PRIME bias predictions are developed to correct DSHP, LGEM, GHMI, and HWFI intensity forecasts from 2011 to 2015. Figure 10 shows the average AE of retrospective intensity forecasts in the Atlantic basin with and without R-PRIME bias corrections.

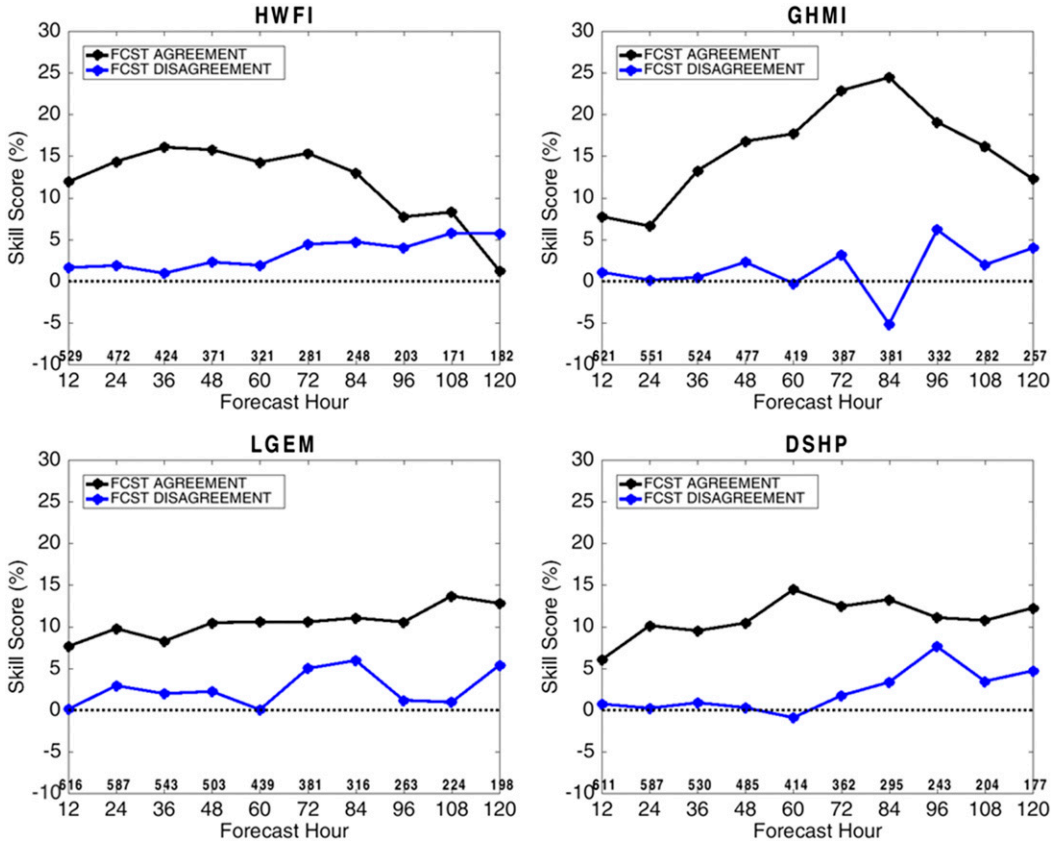

FIG. 7. The performance of R-PRIME AE forecasts in the east Pacific basin for HWFI, GHMI, LGEM, and DSHP when they agree and disagree with bias forecasts. The lines and numbers have the same interpretation as in Fig. 6.

The dashed lines represent the AE of the retrospective forecasts created with the 2015 version of each model. The solid lines capture the average AE of the forecasts that are bias corrected with R-PRIME. For all forecast intervals, the AE of bias-corrected GHMI is significantly lower than the AE of GHMI. Bias-corrected DSHP and LGEM forecasts have AEs that are significantly lower than their respective 2015 retrospective models for 12–84-h forecasts. Bias-corrected HWFI is only significantly better than HWFI at 24 and 36 h. The solid lines are grouped together because DFEM is the only predictor in the R-PRIME bias regression formula for all models and forecast hours. As a result, R-PRIME bias corrections will typically adjust forecasts closer to the mean of DSHP, LGEM, GHMI, and HWFI.

The equivalent figure for PRIME is omitted, but it is important to note that the AEs of the PRIME bias-corrected models are significantly lower than the AEs of the real-time models for all forecast hours.[13] The discrepancy between the number of significant forecast hours for R-PRIME and PRIME is likely attributable to the upgrades to the models. Tables 2 and 3 show the AE, and especially the bias, of all models decreases considerably when using the retrospective runs instead of the real-time runs. The smaller biases greatly reduce the variance in this predictand. As discussed in section 2, random errors become a larger percentage of total error when the magnitude of a predictand decreases, and it is more difficult for R-PRIME to explain the variance of a predictand. Hence, significant predictor–bias relationships disappear. HWFI showed the most improvement from using the retrospective runs, which could explain why R-PRIME bias corrections are the least skillful for this model.

Figure 11 is similar to Fig. 10 but conveys the performance of R-PRIME bias corrections in the east Pacific basin. The larger biases in the retrospective models for the east Pacific basin enable R-PRIME to produce more effective bias corrections. Again, GHMI has the worst-performing retrospective forecasts, and bias-corrected GHMI has significantly lower AE than GHMI at all forecast intervals. All other

[13] DFEM is also the one optimal predictor for PRIME bias forecasts.

TABLE 6. The three best predictors of AE and bias for each model in the Atlantic basin along with the sign of the weighting coefficients for each predictor. For each model, the predictors are ordered (top is most significant) based on the average magnitude of their weighting coefficients over all forecast intervals. The optimal number of predictors is listed for each model and predictand; they are derived from the different training datasets for R-PRIME during 2011–15.

Predictand	AE								
Model	HWFI		GHMI		LGEM		DSHP		
No. of optimal predictors	4		3		2		3		
Predictors/signs of weighting coefficients	FINT	+	ADEM	+	SPRD	+	SPRD	+	
	ADEM	+	SPRD	+	ADEM	+	ADEM	+	
	FIC	+	FIC	+	FLND	−	FINT	+	
Predictand	BIAS								
Model	HWFI		GHMI		LGEM		DSHP		
No. of optimal predictors	1		1		1		1		
Predictors/signs of weighting coefficients	DFEM	+	DFEM	+	DFEM	+	DFEM	+	
	FINT	+	FINT	+	FINT	+	FINT	+	
	FIC	+	FIC	+	ASHR	−	ASHR	−	

bias-corrected models are significantly better than their respective uncorrected models between 12 and 72 h. In general, PRIME bias corrections of real-time forecasts for the same time period in the east Pacific basin exhibit very similar trends in performance and statistical significance.

Figure 11 also highlights one of the primary themes of section 2: short- and long-range PRIME error forecasts in the east Pacific basin have different properties. Therefore, it could be advantageous for PRIME bias forecasts in the east Pacific basin to be developed separately for 12–60- and 72–120-h forecasts. Currently, when PRIME is formulated for a model and predictand, the selected predictors are designed to optimize performance over *all* forecast hours. In other words, the same number of predictors is used at every forecast hour for a model, and if a predictor is deemed unimportant at a particular forecast hour, the weighting coefficient ideally approaches zero. However, applying a large number of predictors that are uncorrelated with the predictand can introduce spurious predictor–predictand relationships. In the Atlantic basin, DFEM is the most important predictor for all forecasts but in the east Pacific basin, DFEM only dominates for long forecast intervals. For shorter forecasts, there are several synoptic predictors, such as ADIV, APOT, and average relative humidity (ARH), with significant correlations with bias. Adding these parameters to the predictor pool for short-range forecasts could be beneficial.

To test if bias forecasts in the east Pacific basin could be improved if R-PRIME was developed solely for short forecast intervals, we examine how SS varies with the number of predictors. In Fig. 12, the SSs of R-PRIME bias forecasts are plotted for 1, 2, 5, 8, 10, and 20 predictors. In general, R-PRIME bias forecasts improve

TABLE 7. As in Table 6, but for entries calculated using R-PRIME forecasts in the east Pacific basin between 2011 and 2015.

Predictand	AE								
Model	HWFI		GHMI		LGEM		DSHP		
No. of optimal predictors	1		3		5		5		
Predictors/signs of weighting coefficients	APOT	+	ADEM	+	ADEM	+	ADEM	+	
	ALAT	−	SPRD	+	ADIV	+	SPRD	+	
	ADEM	+	APOT	+	SPRD	+	FINT	+	
Predictand	BIAS								
Model	HWFI		GHMI		LGEM		DSHP		
No. of optimal predictors	2		1		3		3		
Predictors/signs of weighting coefficients	DFEM	+	DFEM	+	DFEM	+	DFEM	+	
	APOT	−	ARH	−	LAT	+	LAT	+	
	ALAT	+	ADIV	−	FINT	+	LDIS	−	

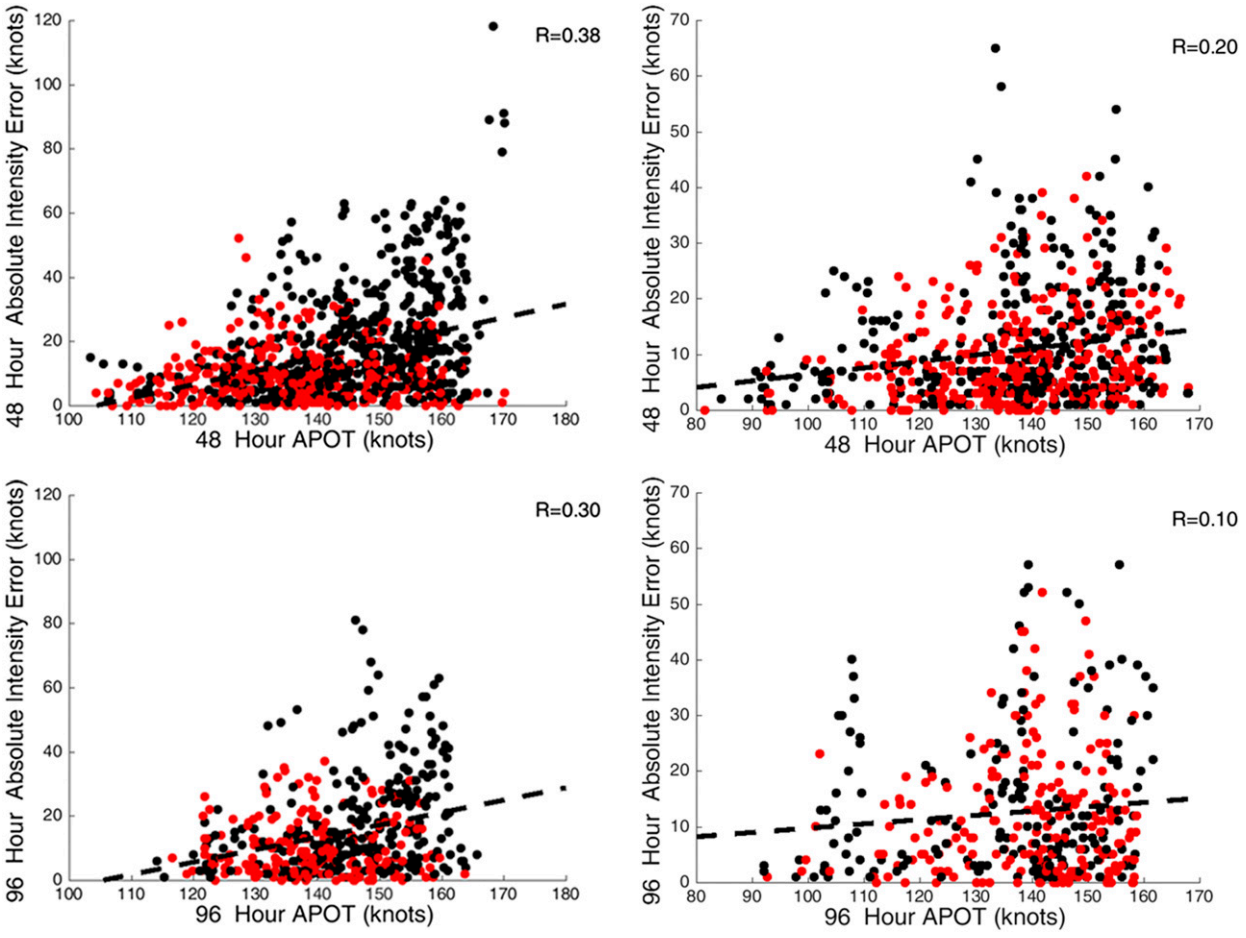

FIG. 8. (top) The 48-h retrospective HWFI AE vs 48-h APOT in the east Pacific basin. (bottom) The 96-h retrospective HWFI AE vs 96-h average APOT in the east Pacific basin. The black dots represent forecasts with a negative bias (underforecasting), and the red dots represent forecasts with a positive bias (overforecasting). The dashed lines represent the linear regression fits to the data, and the correlation coefficient is located in the top-right corner of each plot.

FIG. 9. As in Fig. 8, but for results that apply to the Atlantic basin.

when fewer predictors are used for long forecast intervals and more predictors are used for short forecast intervals. This behavior is visible to a lesser extent for R-PRIME AE forecasts in the east Pacific basin but not for R-PRIME AE or bias forecasts in the Atlantic basin (not shown). Still, it appears that optimizing R-PRIME for different forecast lengths could result in more skillful bias-corrected forecasts.

Figure 13 demonstrates the effects of creating separate R-PRIME regression formulas for long- and short-range bias corrections in the east Pacific basin. The dashed lines display the AEs of bias-corrected models created from a version of R-PRIME that optimizes 12–60-h forecasts while the solid lines represent the AEs of R-PRIME when it is designed to yield the lowest average AE over all forecast intervals. For all the models, 12–60-h R-PRIME forecasts have lower AEs when R-PRIME is developed only for short-range forecasts. The additional predictors for short-range forecasts reduce the weighting coefficient of DFEM and lead to more effective bias corrections that are not just nudging forecasts to the ensemble mean. Forecast end users would particularly benefit from upgraded short-range bias-corrected forecasts in the east Pacific basin, because bias is such a large percentage of the total error for these forecast intervals. Future research should prioritize the development of PRIME for forecast hour–model combinations with robust predictor–error relationships rather than optimizing all forecast hours.

c. *R-PRIME modified ensembles*

Multimodel ensembles are created by combining forecasts into a consensus forecast, typically by calculating the

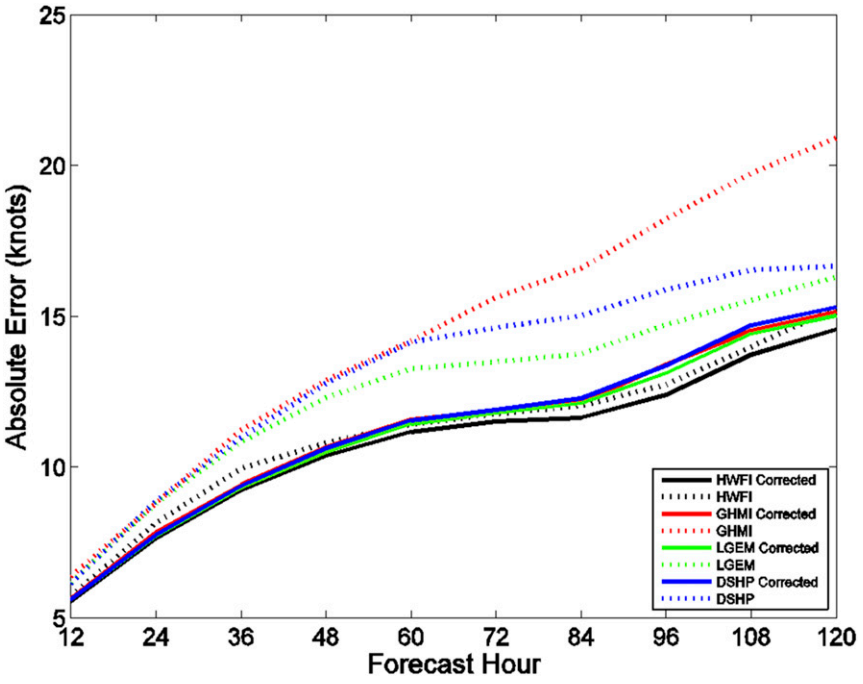

FIG. 10. The average AE of retrospective HWFI, GHMI, LGEM, and DSHP forecasts before and after R-PRIME bias corrections. The solid lines represent the bias-corrected models, and the dashed lines represent the original models. The plotted data are derived from Atlantic basin cases listed in Table 3.

ensemble mean from different models initialized at the same time. Using operational models as ensemble members often produces excellent results because each model generates forecasts that represent a realistic future state of the atmosphere, and the spread of the ensemble members captures the range of potential outcomes. The equally weighted multimodel consensus forecast, ICON, has provided TC intensity guidance in the Atlantic and east Pacific basins since 2007 and routinely registers the lowest errors out of all other intensity guidance (Cangialosi and Franklin 2016; Cangialosi and Franklin 2013; Franklin 2010). The Florida State Super Ensemble (FSSE) is the only unequally weighted ensemble that generates operational intensity forecasts (Krishnamurti et al. 2000). The FSSE technique uses the forecast errors of the individual models during the training period to create weighting coefficients, and the models are then linearly combined to predict future forecast error. As a result, models are weighted solely on how they have recently performed, and synoptic conditions experienced by the TC as well as storm-specific characteristics are not considered in the creation of FSSE forecasts.

With the success of consensus TC intensity forecasts, it is surprising that alternatives to FSSE and equal weighting have not surfaced. One of the potential explanations for the lack of research in this area is that many climate and economics studies have demonstrated that combining different forecast systems with equal weights is typically the best approach to multimodel forecasting (Doblas-Reyes et al. 2005; Wallis 2011). However, it is not clear whether the results in these disciplines are pertinent to the challenge of TC intensity forecasting. TC forecasts focus on a smaller portion of the globe than climate studies, and in limited regions, unequal weighting can provide significantly better results than equal weighting (DelSole et al. 2013). Also, many TCs are forecasted every year, and each TC is accompanied by numerous forecast verifications, which results in the sample sizes for TC studies typically being an order of magnitude larger than those observed for climate forecasting (BN13; Doblas-Reyes et al. 2005; Rodrigues et al. 2014). Additionally, economic research finds that unequal weights can be beneficial when ensemble members are developed from different information sources (Granger 1989). For TC intensity forecasting, each model uses different mathematical formulations to capture atmospheric processes, and therefore the best model can change based on the TC and the situation it is experiencing (BN13; BN15). The larger error variability between models increases the chance that unequal weighting will outperform equal weighting (DelSole et al. 2013).

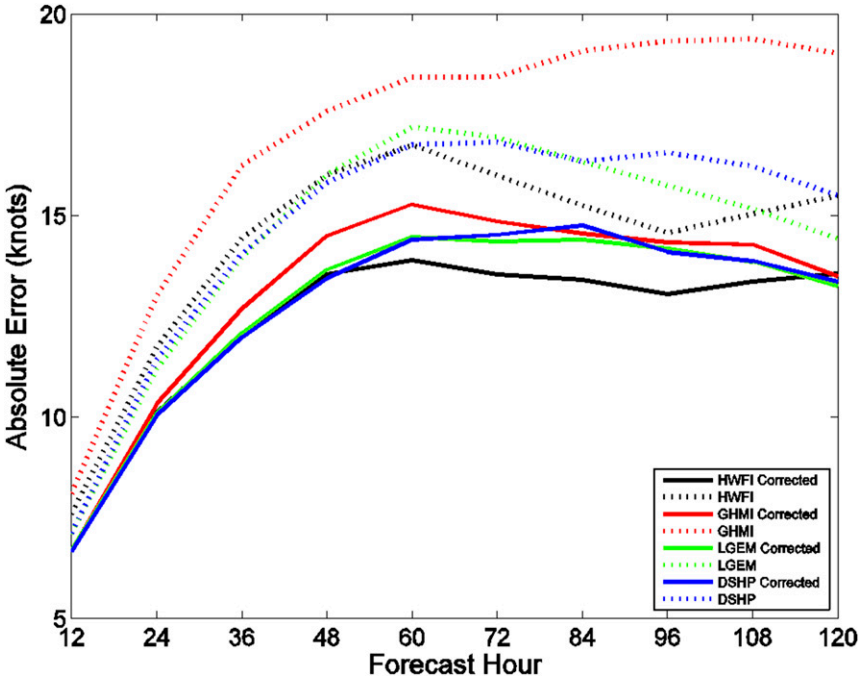

FIG. 11. As in Fig. 10, but for the east Pacific basin. The plotted data are derived from cases listed in Table 4.

As a result, we investigate whether R-PRIME output can be manipulated to create ensembles that are more skillful than ICON. Two main categories of ensemble modifications are applied to DSHP, LGEM, GHMI, and HWFI forecasts to produce seven unique ensembles. For the first group of ensembles, R-PRIME bias and AE forecasts are used to alter or remove individual models before taking the mean of the modified ensemble. Four ensembles are assembled based on this methodology: Correct Worst, Bias Corrected, Exclude Worst (AE), and Exclude Worst (bias). The second set of ensembles involves an unequal-weighting approach where R-PRIME AE and bias predictions are used to weight models inversely proportional (higher errors correspond to lower weights) to their expected error. This weighting technique is featured in three ensembles: Unequal (AE), Unequal (|Bias|), and Unequal (AE SQR). Figures 14 and 15 are included as diagrams that explain how each of the seven ensembles is created.

Figure 16 shows the SS of each of the seven ensembles relative to ICON for the Atlantic basin. Besides 120-h forecasts by the Unequal (|Bias|) ensemble, all of the ensembles' 60–120-h forecasts have lower AEs than ICON. Ensembles developed from R-PRIME AE forecasts generally outperform the ones developed from R-PRIME bias forecasts, which is likely a result of DFEM acting as the only significant predictor for R-PRIME bias forecasts. The Unequal (AE) and Unequal (AE SQR) ensembles have positive SSs at every forecast hour, and averaged over all forecasts, they achieve the highest SSs out of any ensemble. Neither ensemble has a forecast hour where its AE is significantly lower than ICON. However, if serial correlation is ignored when establishing statistical significance, the AE of the Unequal (AE) ensemble would be significantly lower than the AE of ICON for 24–120-h forecasts. Also, the 48- and 96–120-h forecasts produced by the Unequal (AE SQR) ensemble would be significantly better than the corresponding forecasts produced by ICON.

Figure 17 displays the performance of the R-PRIME modified ensembles in the east Pacific basin. Many of the observed trends in the Atlantic basin are reversed in the east Pacific basin. Although the Unequal (AE) ensemble is the only ensemble to record positive SSs at all forecast hours, the ensembles modified using R-PRIME bias, not AE, predictions generally produced the highest SSs. As expected from section 2, R-PRIME modified ensembles produce short-range forecasts that are considerably more skillful than their long-range forecasts. Excluding 72-h forecasts generated with the Exclude Worst (bias) ensemble, all of the ensembles' 12–72-h forecasts have lower AEs than ICON. The Bias Corrected, Unequal (AE), and Unequal (|Bias|) ensembles' 12–60-h forecasts have significantly lower AEs than ICON.

Following the methodology used to create the dashed lines in Fig. 13, R-PRIME 12–60-h bias forecasts can be

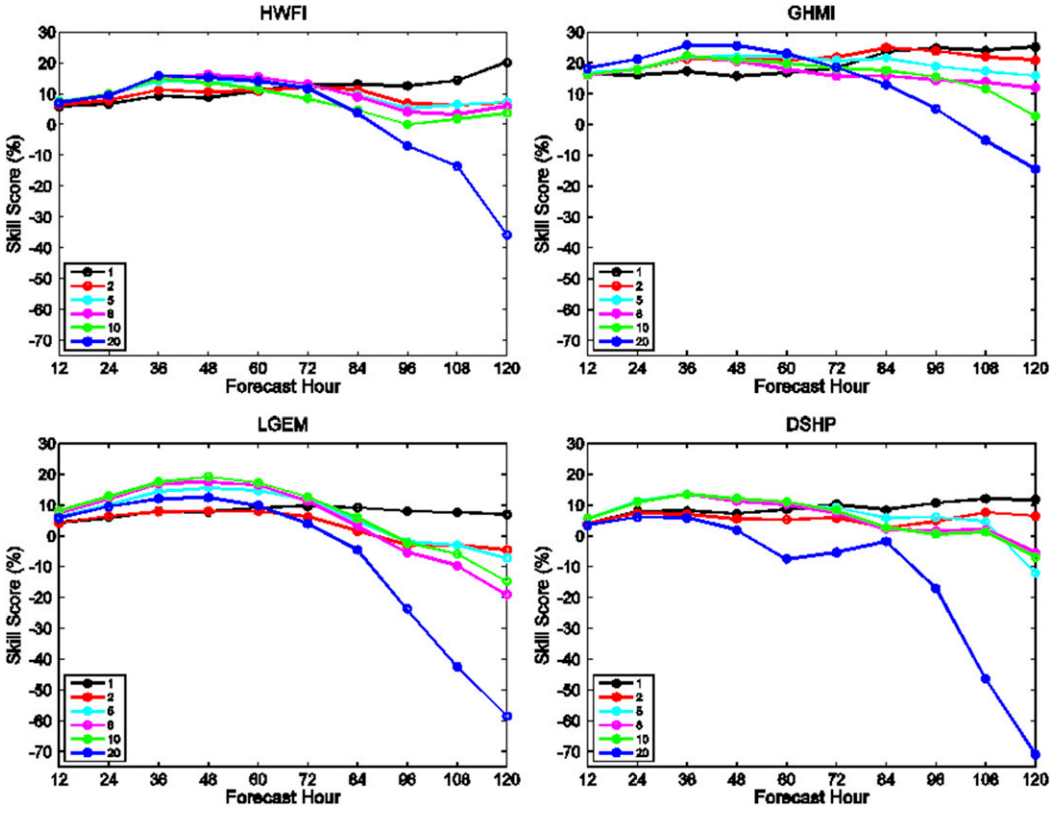

FIG. 12. The SS of R-PRIME bias forecasts relative to climatological bias forecasts in the east Pacific basin for each model. The different lines capture how the performance of R-PRIME changes with the number of predictors inputted into the stepwise multiple linear regression.

optimized to further improve the short-range ensemble forecasts in the east Pacific basin. Figure 18 shows the SSs of the Bias Corrected, Unequal (|Bias|), Correct Worst, and Exclude Worst (bias) ensembles for 12–60-h forecasts when R-PRIME is developed only for these forecasts. All of the ensembles are significantly better than ICON at the 99% confidence level for the plotted forecast intervals. The Bias Corrected ensemble provides the forecasts with the lowest AEs at all forecast intervals, and its SS values range from 6% to 15%.

In both basins, PRIME error forecasts for real-time models are also tested for their ability to produce skillful ensembles. The best-performing modified ensembles are almost identical for PRIME and R-PRIME, and similar relationships between forecast length and error are observed. The performance of the PRIME- and R-PRIME-modified ensembles varied based on how well the PRIME and R-PRIME error forecasts performed. As a result, the R-PRIME-modified ensembles have lower AEs than the PRIME-modified ensembles for a homogenous dataset. However, in the Atlantic basin, the PRIME ensembles generally have higher SSs than the R-PRIME ensembles, whereas in the east Pacific basin, the opposite behavior is observed. There are two mechanisms that explain this discrepancy.

First, HWFI is dramatically improved in the Atlantic basin when switching from real-time forecasts to retrospective forecasts, and no model in the east Pacific basin shows the same level of improvement. As a result, ICON performance greatly improves in the Atlantic basin when retrospective forecasts are used, which makes it more difficult for R-PRIME to produce modified ensembles with positive SSs. Second, replacing real-time forecasts with retrospective forecasts uniquely affects the variance of the ensemble members in each basin. In the Atlantic basin, the average ensemble variance for 36–120-h forecasts decreases between 6% and 36% when retrospective forecasts are used instead of real-time forecasts. When retrospective forecasts replace real-time forecasts in the east Pacific basin, the average ensemble variance for all forecast hours remains within from −6% to 9% of the original variance. Unequally weighted ensembles are typically the most skillful when there is higher variance between the forecasts of the ensemble members (DelSole et al. 2013).

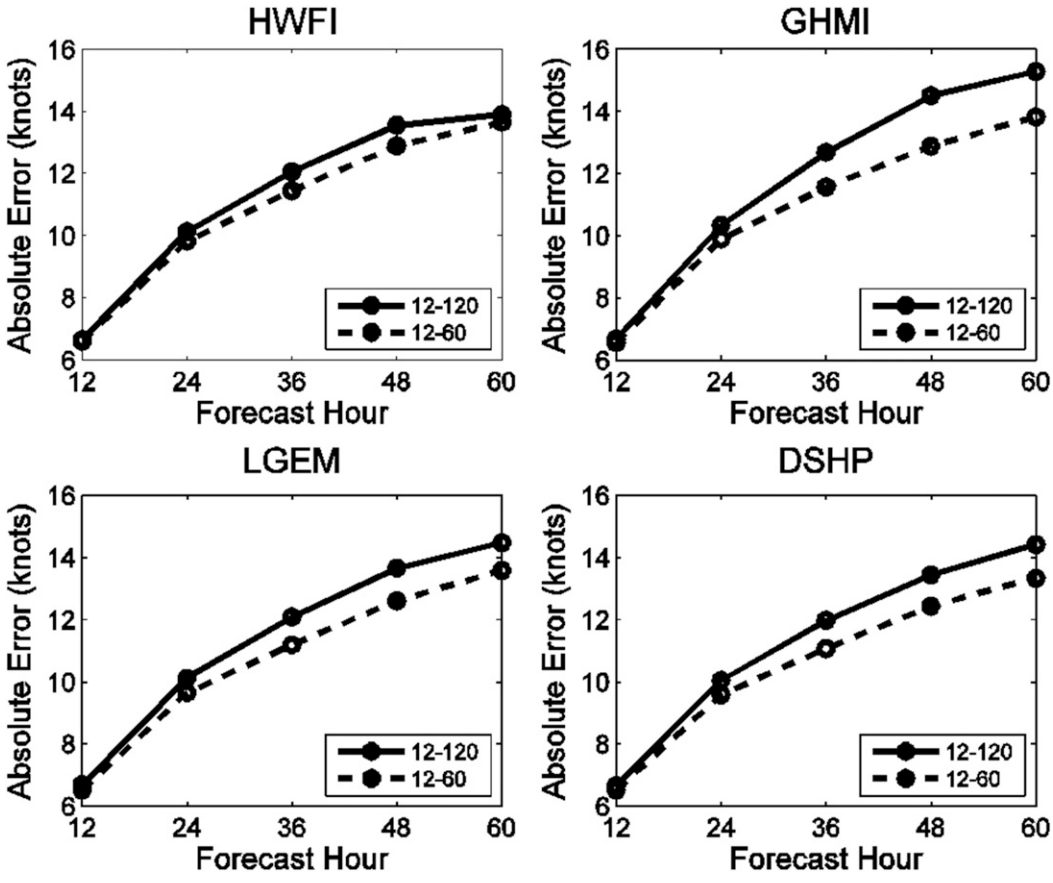

FIG. 13. The average AE of bias-corrected forecasts for two formulations of R-PRIME in the east Pacific basin. The solid black line for each model shows the average AE of R-PRIME bias-corrected forecasts when optimizing the independent verification results for all forecast hours while the dashed black line represents the average AE of bias-corrected forecasts developed solely for 12–60-h bias forecasts.

Therefore, when intensity forecasts produced by the different models show more agreement and are more accurate, there is less room for R-PRIME to detect error-prone situations.

Overall, the ensemble postprocessing techniques using PRIME yielded mixed results. Among all the tested ensembles in the Atlantic basin, the largest observed improvement for a forecast hour is only a 4% decrease in AE [Unequal (AE SQR) ensemble at 96 h] compared with ICON. However, for short-range forecasts in the east Pacific basin, the ensembles based on R-PRIME bias forecasts appear to be a significant upgrade over the current best-available intensity guidance, ICON. SSs in both basins can be increased by introducing nonlinear parameters to the predictor pool but these results are beyond the scope of this study. These predictors would increase the skill of PRIME forecasts and lead to more accurate modified ensembles. Additionally, Bhatia (2015) showed that PRIME- and R-PRIME-modified ensembles provided the biggest upgrade over ICON when their forecasts deviated more from ICON forecasts. Additional techniques for identifying better-performing PRIME ensemble forecasts will be explored in future research.

4. Conclusions and future work

An accurate portrayal of the limitations of a weather forecast is one of the simplest and most effective ways of enhancing forecast value. BN15 developed the PRIME model using statistical techniques and easily accessible data to communicate the expected error of TC intensity forecasts. PRIME was able to skillfully predict the bias and AE results of DSHP, LGEM, GHMI, and HWFI in the Atlantic basin, which paved the way for its operational testing in 2015. Expanding on the work of BN15, PRIME performance was examined in the east Pacific and Atlantic basins for the 2011–15 hurricane seasons. Additionally, R-PRIME was developed from retrospective forecasts using the 2015 version of each

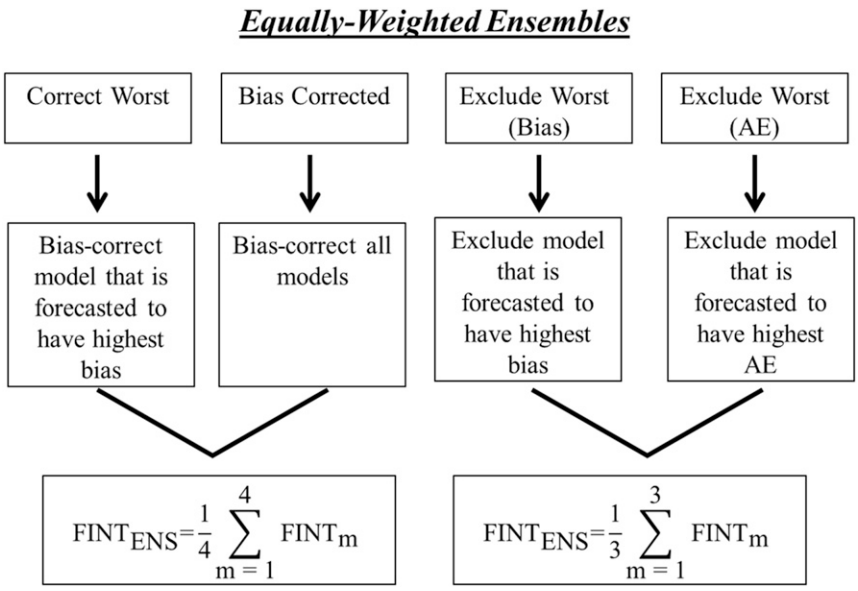

FIG. 14. The process for assembling the four equally weighted ensembles is depicted. Here, $FINT_m$ represents the forecasted intensity of each model m, and $FINT_{ENS}$ is the forecasted intensity of the R-PRIME modified ensemble. The Correct Worst and Bias Corrected ensembles are computed by averaging the four models' intensity forecasts after the applicable R-PRIME bias forecast(s) is/are used to correct the model(s). For the Exclude Worst ensembles, R-PRIME forecasts are used to isolate the model with the highest AE or bias and that model is excluded from the calculation of the ensemble mean.

model and also evaluated. In both basins, PRIME and R-PRIME forecasts outperformed climatological error forecasts for all forecast hours and were significantly better than climatological forecasts for a majority of forecast hours.

PRIME's ability to accurately predict error implied it could serve as a tool for improving intensity forecasts. As a result, R-PRIME bias forecasts were tested as corrections to model forecasts. The bias-corrected models achieved AE values that were lower than the original models' AE values for all forecast hours and basins. Finally, PRIME forecasts provided the foundation for removing, weighting, and correcting models in seven unique ensembles. Two groups of ensemble modifications were tested and compared with ICON. These ensembles outperformed ICON for a majority of forecast hours in both basins and registered statistically significant improvements for several forecast intervals.

Thus far, we have incorporated data from the four most skillful TC intensity models to assemble PRIME forecasts. Starting in 2017, the GHMI model will not be run operationally in the Atlantic or east Pacific basins, and therefore PRIME will need to be formulated using a different ensemble of models. Preliminary analyses indicate that PRIME forecasts for LGEM, DSHP, and HWFI actually improve when GHMI is excluded. This surprising result is likely due to GHMI performance considerably worsening in recent years and negatively affecting PRIME predictors (SPRD, DFEM, ADEM, etc.) for the other models. As mentioned by BN15, global model forecast performance has recently improved, and forecasts from these models are more consistently appearing in the a-decks. As a result, the European Centre for Medium-Range Weather Forecasts (ECMWF) model and GFS will replace GHMI in future iterations of PRIME.

In this study, much like in BN15, error forecasts had higher SSs when retrospective data were used to train PRIME instead of real-time data. Unfortunately, year to year, the length and size of the retrospective sample changes considerably for the dynamical models. For example, before the 2016 hurricane season, retrospective data for HWFI were only available for parts of 2013–15, and retrospective data for GHMI were only available for 2014 and 2015. These hurricane seasons were relatively inactive, so statistical relationships between predictors and errors were less robust. Therefore, the added benefit of having forecast models remain consistent in the verification and training sample might be outweighed by the smaller sample sizes. It is possible that training PRIME with

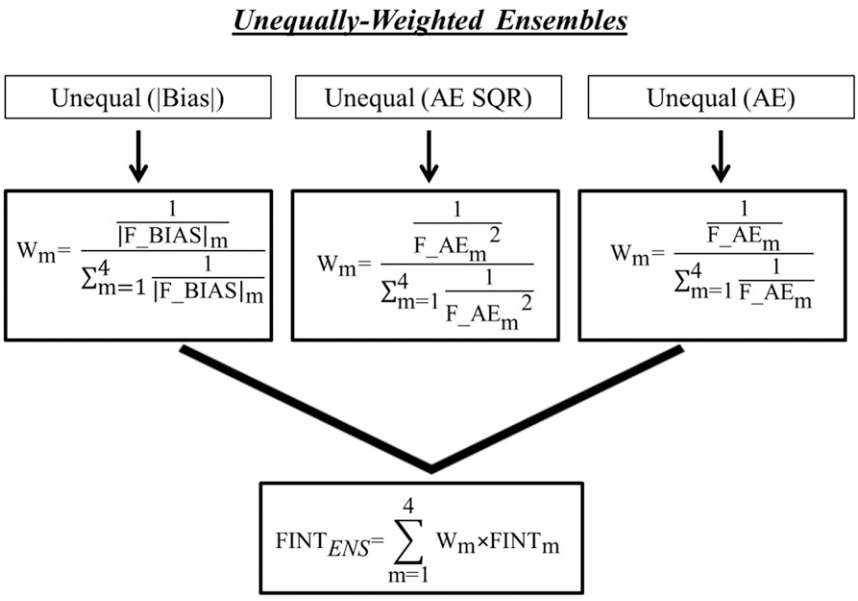

FIG. 15. An overview of how R-PRIME is used to create three unequally weighted ensembles. For the Unequal (AE) ensemble, PRIME AE forecasts F_AE_m quantify the weight W_m for each model m. A similar ensemble, Unequal (AE SQR) is computed using the same equation as Unequal (AE), but with the forecasted AE squared. This alteration adds additional weight to the models that R-PRIME expects to have the lowest AE. The Unequal (|Bias|) ensemble uses the absolute value of the R-PRIME bias forecasts $|F\_BIAS_m|$ to replace F_AE_m. Once the weights are calculated for each model, each model's FINT is multiplied by its computed W_m to obtain $FINT_{ENS}$.

real-time forecasts will lead to better operational results during the 2016 hurricane season. A comparison of how PRIME and R-PRIME bias and AE forecasts performed during 2016 will be shared in a future publication. Based on the potential value of R-PRIME forecasts, it would be beneficial at the start of each hurricane season to have retrospective forecasts for at least the previous five hurricane seasons.

For the entirety of this study, PRIME was developed using the same objective methodology discussed in BN15, with the exception that nonlinear predictors were omitted. The goal of this adjustment to PRIME was to demonstrate the effectiveness of a simpler multiple linear regression scheme. To counteract some of the accuracy lost by neglecting these extra predictors, we tested if combining the forecasts of the two predictands could isolate forecasts with higher skill a priori. In both basins, R-PRIME AE forecasts that agreed with R-PRIME bias forecasts typically resulted in SSs that were 5%–25% higher than SSs of AE forecasts that disagreed with R-PRIME bias forecasts. Additionally, for all forecast hours, the sample size for the forecast agreement cases was never less than 60% of the total forecast sample. This practice could easily be implemented in operations to obtain PRIME forecasts with higher SSs.

Isolating specific forecast intervals for PRIME development (BN15) is another approach for producing more skillful error forecasts while maintaining the same basic statistical framework. BN13 first observed how dynamical parameters exhibited different relationships with error based on forecast length. As a result, a version of R-PRIME was developed for just 12–60-h bias forecasts in the east Pacific basin. These specific forecast intervals were isolated in the east Pacific basin because PRIME bias forecasts and the bias of the models' intensity forecasts showed very different patterns of behavior for short- and long-range forecasts. When R-PRIME was optimized for these short forecast intervals, the AE of the R-PRIME bias forecasts was reduced, and the Bias-Corrected ensemble achieved an SS of nearly 15% relative to ICON for 48-h forecasts. Additional research into isolating other forecasts hours and predictands for PRIME development seems like a worthwhile endeavor.

The separate analysis and development of PRIME in the two basins provided some insights into what controls the performance of error forecasts. There were four key discrepancies in PRIME performance in the different basins:

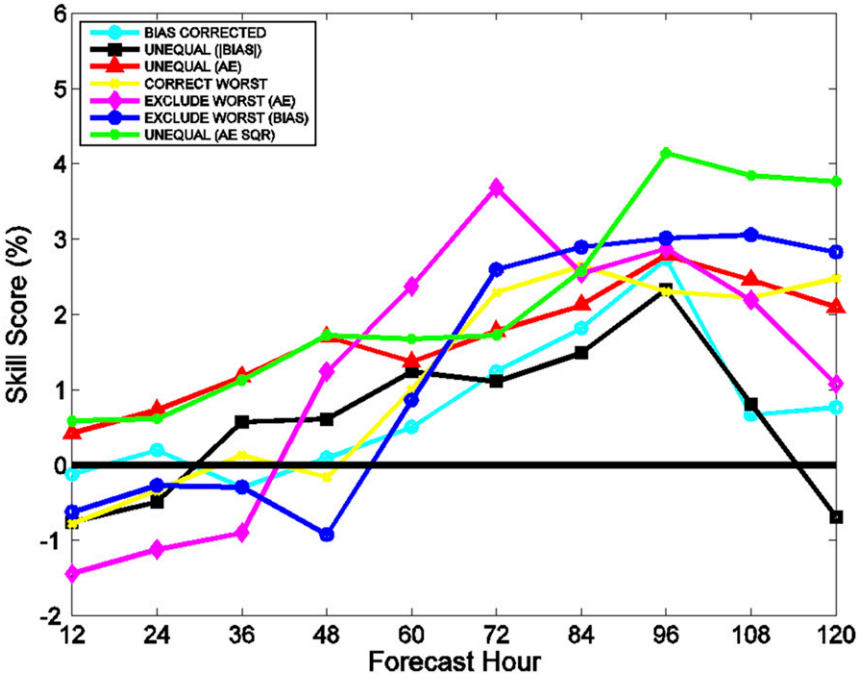

FIG. 16. The average skill score relative to ICON of seven ensembles modified by R-PRIME in the Atlantic basin. These ensembles are formulated using DSHP, LGEM, GHMI, and HWFI retrospective forecasts from 2011 to 2015 and are verified with best-track information.

1) PRIME SSs relative to climatological forecasts typically increased with forecast length for both predictands in the Atlantic basin but decreased with forecast length in the east Pacific basin. This error behavior was attributed to the limited area conducive to intensification in the east Pacific basin compared with the Atlantic basin. As a result, storms in the east Pacific basin tend to rapidly weaken immediately

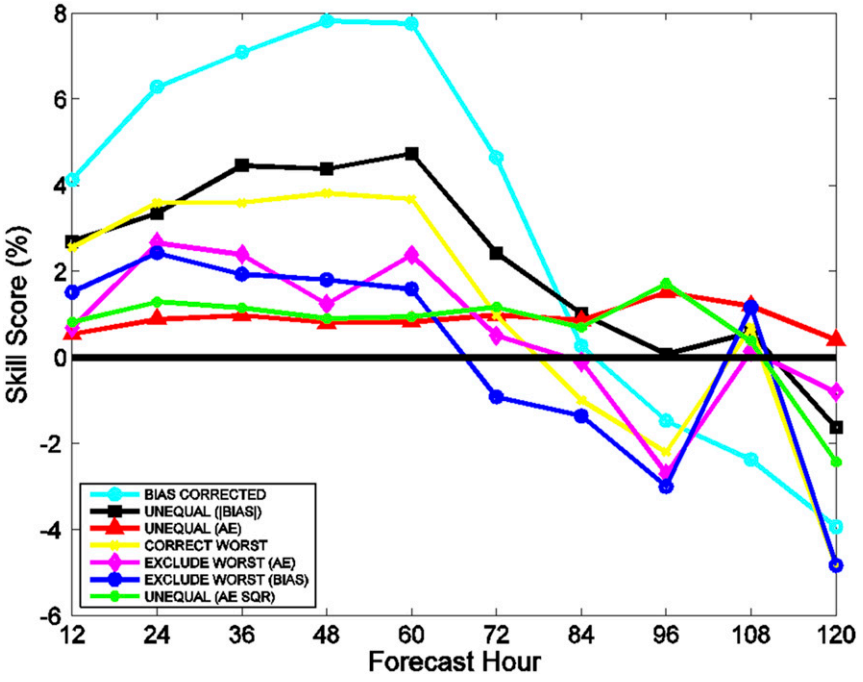

FIG. 17. As in Fig. 16, but for the east Pacific basin.

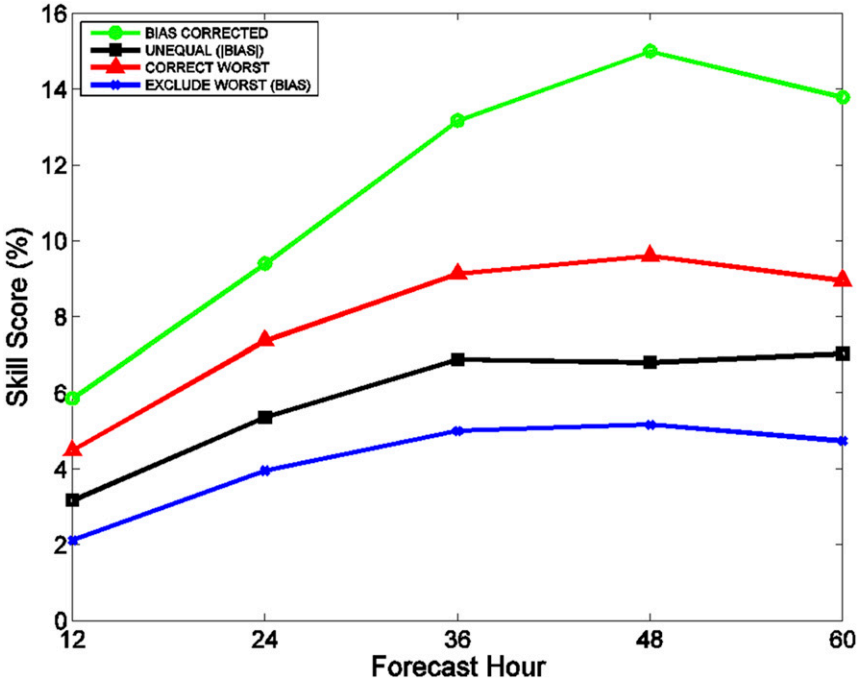

FIG. 18. The average skill score relative to ICON of the four best-performing ensembles modified by R-PRIME in the east Pacific basin. These ensembles are formulated from a version of R-PRIME that was designed to maximize performance between 12 and 60 h.

after rapid intensification and lose their TC status for verification. This unique TC behavior in the east Pacific basin reduces the error and error variance of long-range forecasts, which results in less skillful error forecasts.

2) The R-PRIME's AE forecasts in the Atlantic basin have comparable skill to the bias forecasts, but in the east Pacific basin, R-PRIME's bias forecasts have much higher skill than its AE forecasts. R-PRIME struggles to detect meaningful predictor–bias relationships in the Atlantic basin compared with the east Pacific basin because the models' intensity forecasts have considerably less bias and variance in the bias. Additionally, DFEM was determined to be the only significant predictor of bias in the Atlantic basin. In this scenario, intensity forecasts are forecasted to return to the ensemble mean, which often results in poor error forecasts.
3) The R-PRIME-modified ensembles showed similar SSs to PRIME-modified ensembles (not shown) in the east Pacific basin, while PRIME ensembles showed much higher SSs than R-PRIME ensembles in the Atlantic basin. In the Atlantic basin, the retrospective forecasts for the dynamical models provided significantly better forecasts than the real-time models. Similar verification trends were not observed in the east Pacific basin. This sharper decrease in error due to switching from real-time forecasts to retrospective forecasts greatly reduced the ensemble variance, making it harder for R-PRIME to correctly weight the models.
4) A different set of optimal predictors was used for R-PRIME forecasts in each basin. APOT was highlighted as a critical predictor for bias and AE in the east Pacific basin for multiple models but did not appear to be important in the Atlantic basin. This result suggests that the importance of parameters for error predictions can vary with the basin.

In conclusion, this study has confirmed that PRIME can predict intensity forecast error more accurately than climatological forecasts, and PRIME can serve as a tool to lower intensity forecast error. Multiple linear regression appears to be a sufficient benchmark model for forecasting TC intensity error but several more complex statistical schemes are available. Nonlinear methods and neural networking are two potential alternatives for producing error forecasts. Based on the fact that PRIME- and R-PRIME-modified ensembles varied according to how well PRIME and R-PRIME error forecasts performed, future work improving PRIME forecasts would likely lead to more accurate modified ensembles. After evaluating other error-forecasting techniques, guidance on TC intensity

forecast performance should also be produced in other TC-prone regions across the world. TC landfalls in the Atlantic basin represent less than 1/3 of the global landfall total (Weinkle et al. 2012), so reliable error forecasts would naturally be valuable in other basins. If successful, these forecasts could be produced globally and lead to more informed protocols for hurricane evacuations and storm preparations, which would ultimately save lives.

Acknowledgments. The authors thank Drs. Andrew Hazelton, Morris Bender, Edward Rappaport, Christopher Landsea, and Michael Brennan for their suggestions and comments during the internal review process. The authors also acknowledge that Fig. 4 was created with plotting tools available online (ftp://texmex.mit.edu/pub/emanuel/HURR/tracks/readme.pdf). The authors were supported by the NOAA Joint Hurricane Testbed through Grant NA13OAR4590188.

REFERENCES

Atlas, R., V. Tallapragada, and S. Gopalakrishnan, 2015: Advances in tropical cyclone intensity forecasts. *Mar. Technol. Soc. J.*, **49**, 149–160, doi:10.4031/MTSJ.49.6.2.

Bhatia, K. T., 2015: Tropical cyclone intensity forecast error predictions and their applications. Ph.D. thesis, University of Miami, 224 pp. [Available online at http://scholarlyrepository.miami.edu/oa_dissertations/1537.]

——, and D. S. Nolan, 2013: Relating the skill of tropical cyclone intensity forecasts to the synoptic environment. *Wea. Forecasting*, **28**, 961–980, doi:10.1175/WAF-D-12-00110.1.

——, and ——, 2015: Prediction of Intensity Model Error (PRIME) for Atlantic basin tropical cyclones. *Wea. Forecasting*, **30**, 1845–1865, doi:10.1175/WAF-D-15-0064.1.

Brown, B. R., and G. J. Hakim, 2013: Variability and predictability of a three-dimensional hurricane in statistical equilibrium. *J. Atmos. Sci.*, **70**, 1806–1820, doi:10.1175/JAS-D-12-0112.1.

Cangialosi, J. P., and J. L. Franklin, 2013: 2012 National Hurricane Center forecast verification report. NOAA/NWS/NCEP/Tropical Prediction Center, 79 pp. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2012.pdf.]

——, and ——, 2016: 2015 National Hurricane Center forecast verification report: 2015 hurricane season. NOAA/NWS/NCEP/Tropical Prediction Center, 69 pp. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2015.pdf.]

Cutter, S. L., and M. M. Smith, 2009: Fleeing from the hurricane's wrath: Evacuation and the two Americas. *Environment*, **51**, 26–36, doi:10.3200/ENVT.51.2.26-36.

DelSole, T., X. Yang, and M. K. Tippett, 2013: Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quart. J. Roy. Meteor. Soc.*, **139**, 176–183, doi:10.1002/qj.1961.

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

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting. Part II: Calibration and combination. *Tellus*, **57A**, 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

Emanuel, K., and F. Zhang, 2016: On the predictability and error sources of tropical cyclone intensity forecasts. *J. Atmos. Sci.*, **73**, 3739–3747, doi:10.1175/JAS-D-16-0100.1.

Franklin, J. L., 2010: 2009 National Hurricane Center forecast verification report. NOAA/NWS/NCEP/Tropical Prediction Center, 71 pp. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2009.pdf.]

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, doi:10.1175/BAMS-D-12-00071.1.

Granger, C. W. J., 1989: Invited review combining forecasts—Twenty years later. *J. Forecast.*, **8**, 167–173, doi:10.1002/for.3980080303.

Gray, W. M., 1968: A global view of the origin of tropical disturbances and storms. *Mon. Wea. Rev.*, **96**, 669–700, doi:10.1175/1520-0493(1968)096<0669:GVOTOO>2.0.CO;2.

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

Judt, F., S. S. Chen, and J. Berner, 2016: Predictability of tropical cyclone intensity: Scale-dependent forecast error growth in high-resolution stochastic kinetic-energy backscatter ensembles. *Quart. J. Roy. Meteor. Soc.*, **142**, 43–57, doi:10.1002/qj.2626.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216, doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, doi:10.1175/MWR-D-12-00254.1.

Litman, T., 2006: Lessons from Katrina and Rita: What major disasters can teach transportation planners. *J. Transp. Eng.*, **132**, 11–18, doi:10.1061/(ASCE)0733-947X(2006)132:1(11).

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Rodrigues, L. R. L., F. J. Doblas-Reyes, and C. A. S. Coelho, 2014: Multi-model calibration and combination of tropical seasonal sea surface temperature forecasts. *Climate Dyn.*, **42**, 597–616, doi:10.1007/s00382-013-1779-8.

Sampson, C. R., and A. J. Schrader, 2000: The Automated Tropical Cyclone Forecasting System (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240, doi:10.1175/1520-0477(2000)081<1231:TATCFS>2.3.CO;2.

Tao, D., and F. Zhang, 2014: Effect of environmental shear, sea-surface temperature, and ambient moisture on the formation and predictability of tropical cyclones: An ensemble-mean perspective. *J. Adv. Model. Earth Syst.*, **6**, 384–404, doi:10.1002/2014MS000314.

——, and ——, 2015: Effects of vertical wind shear on the predictability of tropical cyclones: Practical versus intrinsic limit. *J. Adv. Model. Earth Syst.*, **7**, 1534–1553, doi:10.1002/2015MS000474.

Urbina, E., and B. Wolshon, 2003: National review of hurricane evacuation plans and policies: A comparison and contrast of state practices. *Transp. Res.*, **37A**, 257–275, doi:10.1016/S0965-8564(02)00015-0.

Wallis, K. F., 2011: Combining forecasts—Forty years later. *Appl. Financ. Econ.*, **21**, 33–41, doi:10.1080/09603107.2011.523179.

Weinkle, J., R. Maue, and R. Pielke Jr., 2012: Historical global tropical cyclone landfalls. *J. Climate*, **25**, 4729–4735, doi:10.1175/JCLI-D-11-00719.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Elsevier, 676 pp.

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

——, Y. Weng, J. F. Gamache, and F. D. Marks, 2011: Performance of convection-permitting hurricane initialization and prediction during 2008–2010 with ensemble data assimilation of inner-core airborne Doppler radar observations. *Geophys. Res. Lett.*, **38**, L15810.

Zhang, Y. J., Z. Meng, Y. Weng, and F. Zhang, 2014: Predictability of tropical cyclone intensity evaluated through 5-yr forecasts with a convection-permitting regional-scale model in the Atlantic basin. *Wea. Forecasting*, **29**, 1003–1023, doi:10.1175/WAF-D-13-00085.1.