NOAA Joint Hurricane Testbed (JHT)
Final Report

Date:	February 29, 2015
Reporting Period:	September 1, 2013 – December 31, 2015
Project Title:	*Guidance on Intensity Guidance*
Principal Investigators:	David S. Nolan, RSMAS, University of Miami, and
	Andrea Schumacher, CIRA, Colorado State University
Award Period:	September 1, 2013 – December 31, 2015

1. Long-term Objectives and Specific Plans to Achieve Them:

The goal of this project is to develop a system for real-time prediction of the expected errors of individual hurricane intensity forecast models and to use this information to improve operational forecasts. In the first year of the project, we built on the results of Bhatia and Nolan (2013) to construct a model that predicts the error of each intensity forecast model at each forecast interval. Error prediction models were developed for each of the "early" intensity forecast models that are available to forecasters: DSHP, LGEM, GHMI, and HWFI. During the second year, the models were fully developed and began running operationally in real-time from June 1st. The system makes predictions of both absolute error (AE) and bias for each of the four official intensity forecast models, as well as bias corrected forecasts for each model, and a weighted consensus model that weights each model according to its predicted AE. All of these outputs were made available in graphical and text form in real time at the CIRA model products web page.

2. PRIME and R-PRIME

a. How it works

The Prediction of Intensity Model Error (PRIME) is very similar to the highly successful SHIPS model (DeMaria and Kaplan 1994, 2005). It uses multivariate linear regression to make predictions of the absolute error (AE) and bias of each of the 4 early intensity models at each forecast time. The predictors are chosen from a long list of potential predictors, which include synoptic information such as environmental wind shear and MPI, and also information from the other models, such as the difference between an individual model intensity forecast and the

mean of the four-model ensemble. For almost all predictors, both the 0-hour value and the mean value over the forecast period (e.g., shear at 0 h, and shear averaged over 72 h) are considered as possible predictors. The predictors related to the dynamical features of the storm and the surrounding synoptic environment are available in the stext (SHIPS) files. The intensity forecasts for the models are located in the ATCF "a deck" files, while the intensity verification is in the NHC best-track digital database (Landsea and Franklin 2013).

The model is trained on several years of forecasts. Using the standard approach, the least significant predictor is eliminated and the regression is repeated until only predictors with impacts that are statistically significant over one or more of the forecast hours remain. The distribution of the AE predictand for a given time interval is not normal and causes errors that are heteroscedastic. As a result, a power transformation is necessary to transform the positively skewed AE distribution to an approximately Gaussian distribution, creating more homoscedastic data for the linear regressions. In the final step, the predictands are transformed back to their physical values. We also experimented with nonlinear transformations of some of the predictors. For example, small positive values of the distance to land predictor (DIST) have the largest correlations with high errors. Therefore the variable is transformed to a predictor that has its largest value when the distance to land is around 50 km. The particulars of these transformations and the development of the model are described in great detail in the recently published article by Bhatia and Nolan (2015).

PRIME as such only uses intensity forecast data from the real-time models in past hurricane seasons. This can not account for yearly changes to each model, which can be substantial, especially for the dynamical models. Fortunately, the outcome of retrospective forecasts using the 2015 versions of the models on several years of forecasts were made available to us, from which we were able to develop a retrospective version, R-PRIME. R-PRIME was generally more accurate than PRIME, although not always more skillful, because the retrospective intensity forecasts are more accurate so their average errors ("climatological errors") are smaller. Another trade-off is that there are not as many years of forecasts of R-PRIME as there are for PRIME which simply uses all the past forecasts as far back as 2007.

Unfortunately, only the 2014 retrospective runs were available to us before the start of the 2015 hurricane season. Despite having used models one year out of date, R-PRIME performed better in cross-validation ("leave one year out") testing, so it was used as the basis for the 2015 operational system.

b. Corrected-consensus models

Another goal of the project was to produce unequally-weighted ensemble forecasts based on the expected error of each model, i.e., the models with larger predicted absolute error are given less weight, and vice-versa. We also experimented with a bias-corrected ensemble, where the predicted bias is removed from each forecast before they are then averaged with equal weights. When using the version of PRIME based on the past forecasts, we found that the bias-corrected ensemble produces the most accurate forecasts. However, when using R-PRIME we found the most accurate forecasts were made using an unequally weighted ensemble where each member is weighted by the inverse of its predicted error squared. This is the version that was applied operationally in 2015.

c. Implementation and operational appearance

The PIs (Nolan and Schumacher) and the graduate student (Kieran Bhatia) worked together to make PRIME and the corrected consensus models work in parallel with other real-time systems such as SHIPS. While PRIME was developed entirely using Matlab software at the University of Miami, calculation of the real-time error forecasts is straightforward and Fortran code that works on systems at NOAA has been developed to reproduce the results from UM. PRIME error and corrected consensus forecasts were available in real time from the CIRA web page. Six plots were produced for each forecast: 1) The corrected consensus using the inverse AE-squared weighting; 2) A histogram showing the predicted AE of each model for each forecast time; and (3)-(6) were bias-corrected forecasts for each of the four models. Examples of these figures (taken from the 30 Sept. 12Z forecasts of Joaquin) are shown below. A text file summarizing the forecasts, along with the values of the leading predictors, is also produced for each forecast. The example shown in Figure 2 shows the portion of the text file with HWFI results and the corrected consensus prediction. The data for the other models also appears in this file.

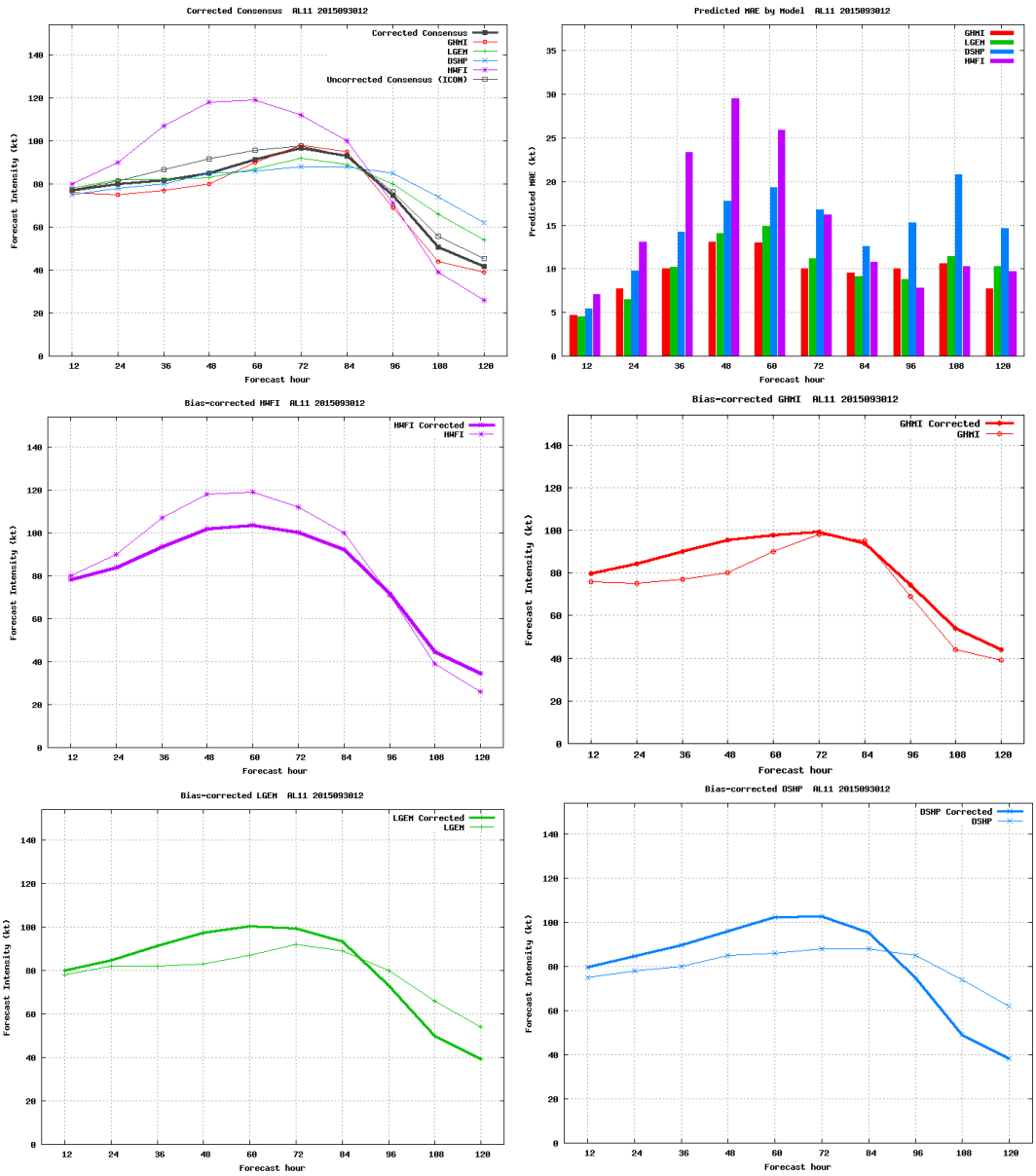

Fig. 1: Real-time output from PRIME taken from the CIRA web page for the Hurricane Joaquin 12Z forecasts on September 30.

```
                  *   PRedicted Intensity Model Error (PRIME)   *
                  *         AL112015  09/30/15  12 UTC          *
                  *                    HWFI                    *

TIME (HR)           12     24     36     48     60     72     84     96    108    120

V (KT) HWFI         80     90    107    118    119    112    100     71     39     26
V (KT) HWFI_BC      78     84     93    102    103    100     92     71     44     34

AERR (KT) PRED       7     13     23     29     25     16     10      7     10      9
AERR (KT) CLIM       5      7      8      9     10     10      9     10     10     11

BIAS PREDICTORS:
INIT DTL (NMI)   532.0  532.0  532.0  532.0  532.0  532.0  532.0  532.0  532.0  532.0
AVG LAT (N)       24.6   24.4   24.4   24.5   24.8   25.2   25.8   26.5   27.3   28.1
ENSMN V (KT)       2.8    8.8   20.5   26.5   23.5   14.5    7.0   -5.3  -16.8  -19.3

AERR PREDICTORS:
AVG LAT (N)       24.6   24.4   24.4   24.5   24.8   25.2   25.8   26.5   27.3   28.1
AVG MPI (KT)     163.7  163.4  163.2  163.3  163.4  162.3  160.0  157.8  155.1  151.7
STDEV BT (C)      -2.0   -2.0   -2.0   -2.0   -2.0   -2.0   -2.0   -2.0   -2.0   -2.0
VMX CHANGE (KT)   10.0   20.0   37.0   48.0   49.0   42.0   30.0    1.0  -31.0  -44.0
T=0 V (KT) HWFI   70.0   70.0   70.0   70.0   70.0   70.0   70.0   70.0   70.0   70.0
V (KT) HWFI       80.0   90.0  107.0  118.0  119.0  112.0  100.0   71.0   39.0   26.0
ENSMN V (KT)       2.8    8.8   20.5   26.5   23.5   14.5    7.0   -5.3  -16.8  -19.3


                  *   PRedicted Intensity Model Error (PRIME)   *
                  *           Corrected Consensus               *
ICON (KT)         77.2   81.2   86.5   91.5   95.5   97.5   93.0   76.2   55.8   45.2
CCON (KT)         77.0   80.0   81.5   84.9   91.3   96.8   93.1   74.6   50.8   41.7
```

Fig. 2: A portion of the text file that is also produced in real time.

3. Outcome and Validation

In the following sections, we assess a) how well the operational version of PRIME worked in real-time in 2015; b) how well future versions of PRIME might work; and c) results for PRIME developed for the East Pacific.

a. Real-time, operational PRIME in 2015: Error forecasts and weighted ensembles

As noted above, the operational version of PRIME used in 2015 is based on R-PRIME, but using retrospective forecasts from the 2014 models, not the 2015 models. The results of the AE forecasts are summarized in Fig. 2. The plots show mean absolute error of forecasts of AE for each of the 4 models, for each forecast time. The dashed lines show the mean absolute error of error forecasts that simply use the average error based on the 4 years of retrospective forecasts; this mean error is the “climatological error” and will hereafter be referred to as “climatology.” Assessments of skill are made by comparing the mean AE or bias of PRIME to climatology.

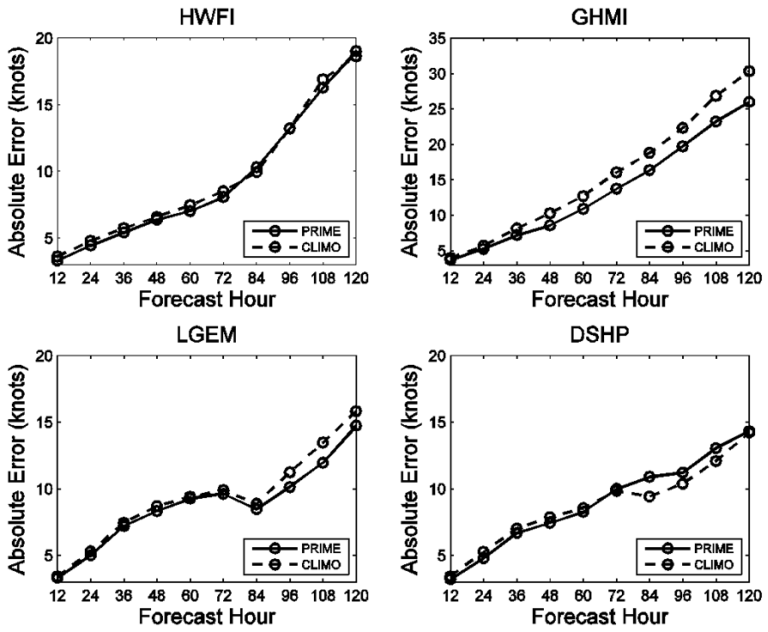

Fig. 3: Mean absolute error of forecasts of absolute error (AE) by PRIME (solid lines) for each model, along with the mean absolute error of AE forecasts using the climatology (mean errors) of each model. Note that the axes are different for GHMI.

Figure 3 shows mixed to positive results for PRIME forecasts of AE. Between 12 and 60 h, PRIME forecasts for all models show lower AE than climatology. Beyond 60 h, PRIME struggles for all models besides GHMI. The GHMI error forecasts are significantly better (at the 95% level, and hereafter all further significance comments will refer to this confidence level) at all times (note that the *y* axis extends to higher values). This is not surprising, as GHMI performed terribly in 2015. PRIME was correctly identifying its forecasts as likely to be erroneous, and therefore easily exceeded predictions based on the average error of GHMI.

Figure 4 shows the same results but for the absolute error of forecasts of model bias. Again, PRIME was very skillful at predicting the bias of GHMI, having correctly anticipated its very poor (usually overly intense) forecasts in 2015. Unfortunately, PRIME had negative skill for HWFI and LGEM, and mixed results for DSHP. In fact, the poor performance of GHMI is part of the reason that PRIME did not perform well predicting bias for the other models: one of the leading predictors for the bias of each model is its difference from the ensemble mean. If one model is consistently far from the other three, it will skew the bias forecasts of those models.

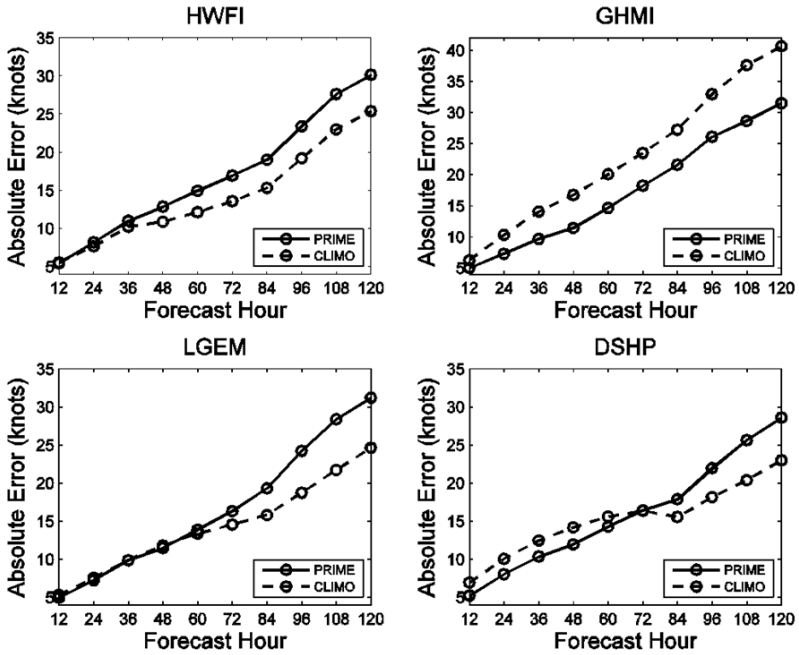

Fig. 4: Mean absolute error of forecasts of intensity model bias by PRIME (solid lines) for each model, along with mean (climatological) errors of each model. Note that the axes are different for GHMI.

The corrected consensus model operational in 2015 was based on each model weighted by the inverse square of its forecasted AE. PRIME AE forecasts weighted each model with the equation:

$$W_m = \frac{\frac{1}{(\text{PRIME\_AE})_m^2}}{\frac{1}{\sum_{m=1}^{M}(\text{PRIME\_AE})_m^2}},$$

where *M* is equal to 4, the number of models. This creates "CCON" which can be compared to the standard, equally-weighted ICON, as shown in Figure 5. The performances of ICON and CCON are nearly identical out to 84 h, but for longer times CCON did perform better, with an improvement of 2 knots at 120 h. At 108-120 hr, these results are statistically significant.

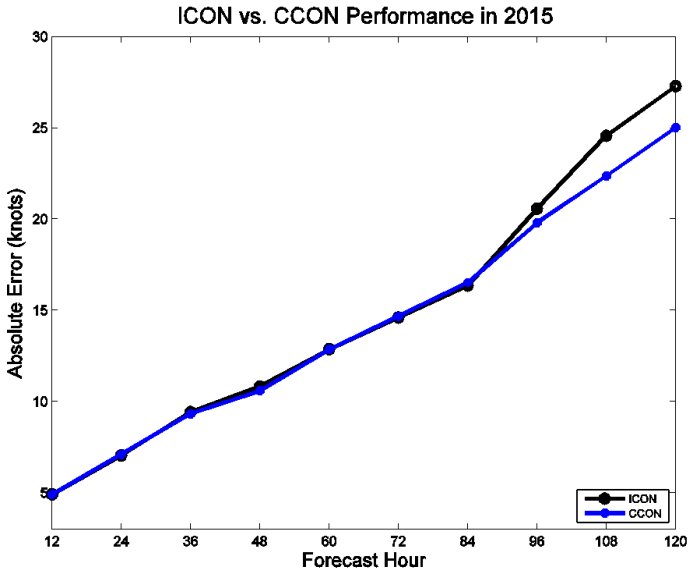

Fig. 5: Mean error of consensus forecast models of intensity: the standard ICON and the corrected consensus (CCON) using unequal weights.

b. Likely success of future implementations of PRIME

In a perfect world and with sufficient resources, PRIME would be updated before each hurricane season, much like the operational models. The choices of predictors and their coefficients would be recomputed based on the retrospective forecasts of the operational intensity models that are going to be used for that same hurricane season. In addition, these retrospective forecasts would be available for all 4 models for at least 4 full hurricane seasons.

Under those circumstances, how well would PRIME perform? Alternatively, how would PRIME perform if it were only based on real-time forecasts, without use of retrospectives?

To answer these questions, we repeated the development of PRIME using data from 2011 to 2015. Both PRIME and R-PRIME were developed: the former uses only data available in real time from each season, while the latter uses the retrospective forecasts of the 2015 models which became available to us later this past season. The following results use the standard "leave one year out" validation: the model is tested on each one of the years in 2011 to 2015 year using predictors and coefficients derived from the other 4 years, and the results are averaged over this process repeated over all 5 years. Figure 6 and Figure 7 show the results for AE and for bias.

Both figures show that both PRIME and R-PRIME have positive skill at all times for all models. Note that each version of the model is compared to its own climatology. While R-PRIME makes more accurate forecasts of AE and bias, it is generally not more skillful, because the retrospective (updated) models have less error variance in their forecasts. A paired t test, adjusted for serial correlation, determined that the differences between PRIME and climatology errors for all forecast intervals, predictands, models, and versions of PRIME were significant at the 95% level except 108-120 h AE forecasts of PRIME and R-PRIME for DSHP, LGEM, HWFI and 96-120 h bias forecasts of R-PRIME for HWFI bias. Additionally, both versions of PRIME were able to forecast the AE of the models' intensity forecasts significantly better than the models forecasted intensity.

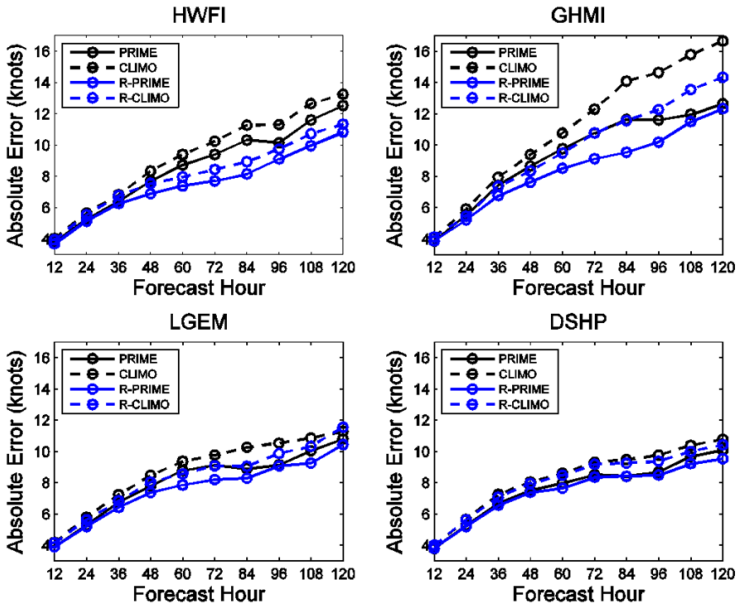

Fig. 6: Results for absolute error of forecasts of absolute error (AE) by PRIME (black curves) and R-PRIME (blue curves) developed using data from 2011-2015.

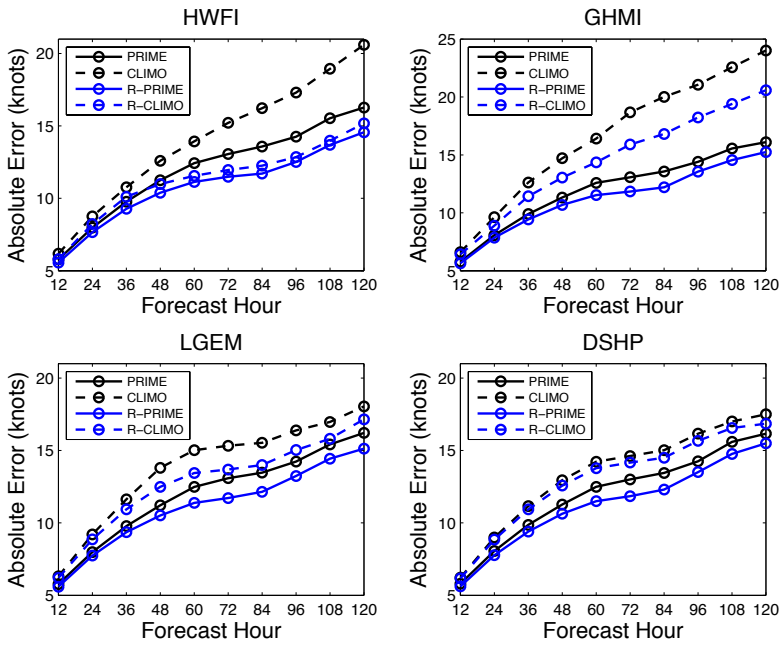

Fig. 7: As in Fig. 6, but for forecasts of model bias.

We also evaluate potential implementations of CCON for PRIME and R-PRIME, for each of the three different methods: bias correction before averaging with equal weights, unequal weighting by the inverse of AE, and unequal weighting by the squared inverse of AE. These are shown below in Fig. 8. All versions of CCON make small improvements over ICON for forecasts longer than 72 h, with average improvements reaching about 1 knot at 120 h. The PRIME modified ensembles are significantly better than ICON between 72-120 h. The R-PRIME modified Unequal SQR (MAE) ensemble is significantly better than ICON for 96-120 h (the other R-PRIME ensembles show no significant results).

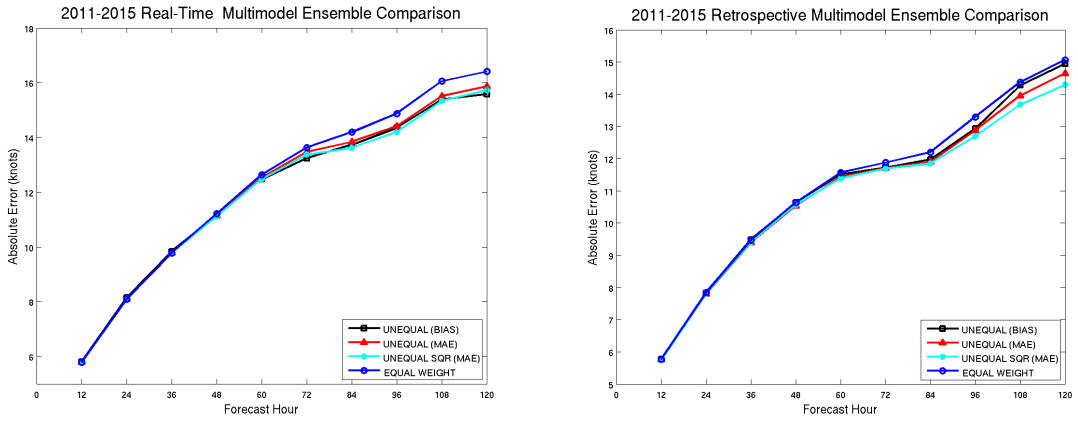

Fig. 8: Various versions of corrected consensus (CCON) models compared to ICON (blue curve) for PRIME (left) and R-PRIME (right). Note axes are different.

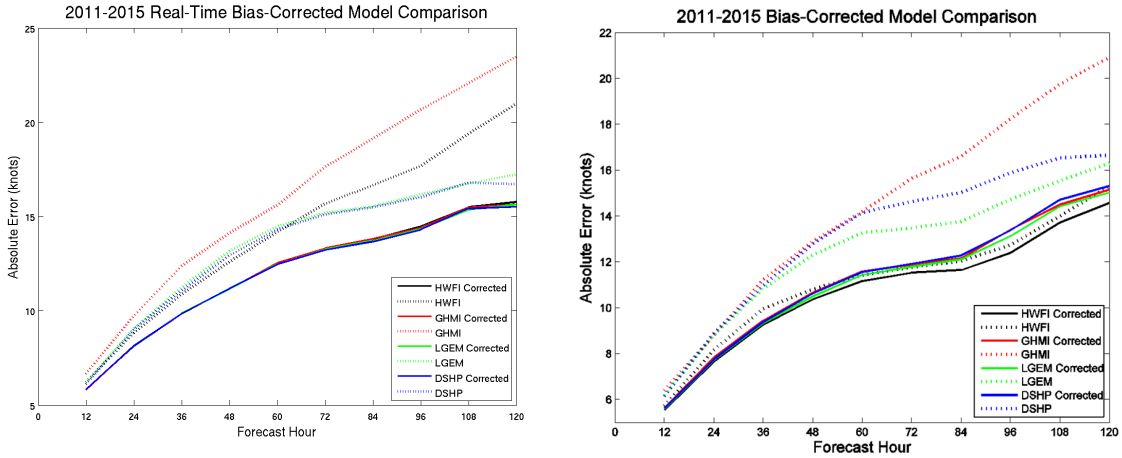

Fig. 9: Mean AE of forecast models, 2011-2015, after bias correction by PRIME (left) and R-PRIME (right). Note axes are different.

Finally, another way to see the potential impact of PRIME is to look at the mean errors of the forecast models after being bias-corrected by PRIME (or R-PRIME). These are shown in Fig. 9. Bias corrections lead to significant improvements in forecast error for every intensity model, with large improvements for the dynamical models at 120 h. The only exception appears to be for the retrospective HWFI forecasts. However, this is a result of the very good performance for HWFI on the right side of Figure 9. The upgrades to retrospective HWFI results in very low AE, which makes it very difficult for PRIME to detect significant error trends in the data.

Comparing the results of Figures 8 and 9, one might wonder why the CCON forecasts, especially those based on bias correction, do not lead to larger improvements over ICON. The reason is that the leading predictor of bias is the difference from the mean of the four intensity models. In other words, the models are all being adjusted towards the ensemble mean. Thus, CCON is often similar to ICON.

c. East Pacific PRIME

Although it was not available for real-time operations in 2015, in the last few months we have developed equivalent versions of PRIME and R-PRIME for the East Pacific. Fig 10 shows the mean errors of the AE and bias forecasts. The improvements over climatological error are almost uniformly positive, and are highly skillful for the dynamical models.

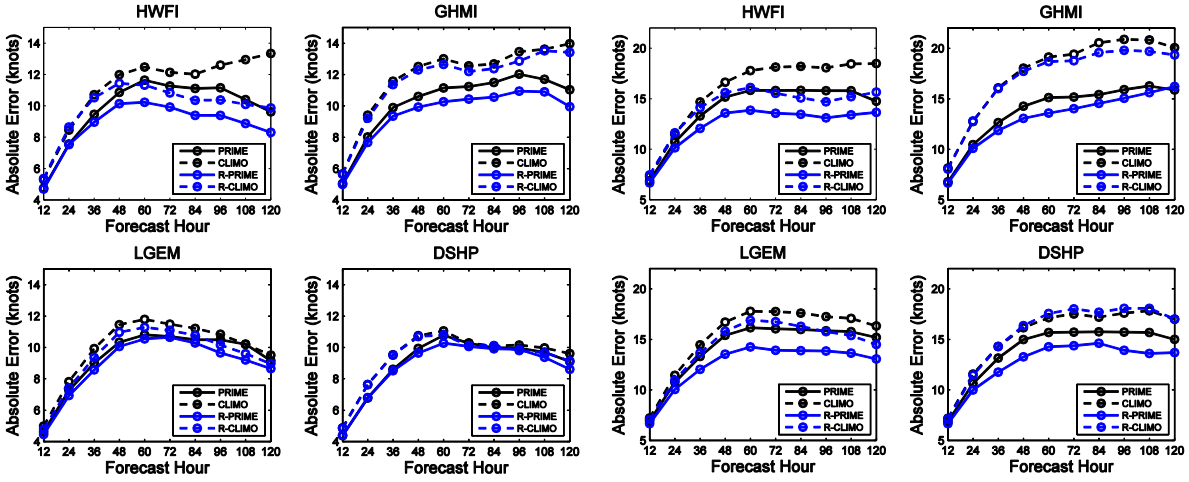

Fig. 10: Mean error of AE forecasts (left) and bias forecasts (right) for PRIME and R-PRIME developed for the East Pacific for 2011-2015.

4. Developer Recommendations

At the present time, NHC has no objective system to anticipate the errors and skill of the operational models on a forecast-by-forecast basis. In some forecast discussions, forecasts are described as “low-confidence” or “high-confidence” based on the situational awareness and experience of the forecasters. We do not discount the validity of these statements. Similarly, forecasters often use their own intuition to weight some models more heavily than others, especially when one of them appears to be an outlier. Operational implementation of PRIME

would provide objective guidance for statements of confidence and for model selection. Of course, an experience forecaster would always have the option to deviate from these forecasts (of error) as well.

Being a second-order modeling system (a forecast model of forecast errors of other models), PRIME is more complicated to update than first order models, like SHIPS and LGEM. A reliance on retrospective forecasts could be particularly problematic, because 1) the update cannot occur until after the retrospectives are completed, and 2) the potential benefits of retrospectives are greatly diminished because they are not generated for every storm and may not be generated for three or more hurricanes seasons. We have been informed that the retrospectives for the coming season will only be performed for the previous two seasons. An additional complication is the use of nonlinear adjustments to the predictors and predictands, which ideally would be updated every year for every model.

Nonetheless, given the relatively low activity of the 2015 Hurricane Season, and the potential benefits of PRIME, it is clearly worthwhile to evaluate PRIME operationally for an additional year. Therefore, we recommend that a simpler version of PRIME be developed that will be available in 2016 for the Atlantic. This version will 1) only use real-time forecasts, and 2) use only modifications to the predictors and predictands that are either very simple to update, or do not need to be updated for the foreseeable future. PRIME forecasts will be set up to appear with the operational model products on the CIRA web page just as they did in 2015. Without need for the retrospective forecasts and extensive tuning, we expect that these can be implemented in the next few weeks.

We are currently not aware to what extent PRIME was used by forecasters in 2015. Another recommendation is for the developers and the JHT contacts at NHC to work together in the coming spring and summer to make PRIME better-known among the hurricane specialists.

5. References

Bhatia, K. T., and D. S. Nolan, 2013: Relating the skill of tropical cyclone intensity forecasts to the synoptic environment. *Wea. Forecasting*, **28**, 961–980.

Bhatia, K. T., and D. S. Nolan, 2015: Prediction of intensity model error (PRIME) for Atlantic Tropical Cyclones. *Wea. Forecasting*, **30**, 1845-1865.

DeMaria, M., and J. Kaplan, 1994: A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209-220.

DeMaria, M., M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvement to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543.