16A.1 PREDICTION OF TROPICAL CYCLONE INTENSITY FORECAST ERROR

Kieran T. Bhatia* and David S. Nolan
University of Miami, Miami, Florida

1. INTRODUCTION

Although the skill of operational tropical cyclone (TC) track forecasts have increased considerably over the last twenty years in the Eastern Pacific and Atlantic basins, intensity forecasts have only shown mild improvement in the eastern Pacific Basin and, depending on the forecast time, little improvement to worsening in the Atlantic basin (Cangialosi and Franklin 2013). TC intensity forecast performance can also differ considerably between models, years, and storms. The lack of quality and consistency decreases the value of TC intensity forecasts and demands attention from the scientific community. One approach to increasing the value of TC intensity forecasts with the resources currently available is to create real-time error predictions that help forecasters and end users know whether a particular model forecast will be more or less skillful than average. This a priori expectation of forecast performance would combat the adverse effects of the substantial day-to-day and storm-to-storm fluctuations in forecast quality.

As a first step towards quantifying the expected intensity forecast error of a TC, Bhatia and Nolan (2013, hereafter BN13) studied the relationship between synoptic parameters, TC attributes, and forecast error. This study represented one of the first attempts to understand how TC intensity forecast performance is connected to these storm-specific characteristics. Model performance was binned according to these defining features, and *t* tests were used to measure the robustness of the results. The statistical significance established between bins conveyed that there is evidence that forecast error is often related to the nature of the particular storm and surrounding atmospheric environment. Based on these conclusions, it seems likely that variables capturing a tropical cyclone's environment can be used to skillfully predict intensity forecast error.

*Corresponding author address: Kieran T. Bhatia, Univ. of Miami, RSMAS, email: kieran.bhatia@gmail.com

Since then, the authors have found that parameters representing initial condition error and atmospheric stability ("proxies") are also linked to forecast error. These empirically-derived proxies along with the synoptic variables are used to predict tropical cyclone intensity forecast error.

2. DATA AND METHODOLOGY

In BN13, the situation-dependent performance of the Logistic Growth Equation Model (LGEM), Decay Version of the Statistical Hurricane Intensity Prediction Scheme (DSHP), Geophysical Fluid Dynamics Laboratory hurricane model (GFDL), and the National Hurricane Center's official (OFCL) forecast were evaluated at 24, 48, and 72 hours by computing the mean absolute error (MAE), bias, and percent skill (PS). BN13 focused on the five Atlantic basin hurricane seasons between 2006 and 2010. For this study, the results have been expanded to include 96- and 120-hour forecasts, the Hurricane Weather Research and Forecasting (HWRF), GHMI, and HWFI models, and the 2011 and 2012 hurricane seasons. GHMI and HWFI represent the "early" versions of the HWRF and GFDL models, which are created by interpolating the version of HWRF and GFDL from the previous forecast. The intensity forecasts for all models are located in the National Oceanographic and Atmospheric Administration's (NOAA) Automated Tropical Cyclone Forecast (ATCF) "a-deck" files. The forecasts are verified with National Hurricane Center (NHC) best-track data.

The focus of this study moves away from analysis of model performance and towards the prediction of forecast error. As a result, only the early models, DSHP, LGEM, HWFI, and GHMI are considered because these forecasts are produced at synoptic times. The dynamical predictors used for these predictions include initial and forecast intensity and initial and forecast average (e.g., the forecast average shear of a 24-hour forecast is the average of initial shear and each 6-hourly forecasted shear until 24 hours): potential intensity, storm speed, 850-200 hPa wind shear magnitude, 850-200 hPa wind shear direction, longitude, latitude, 850

hPa vorticity, 200 hPa divergence, 700-500 hPa relative humidity, and distance to land. The predictor values are computed using output from the National Centers for Environmental Prediction's (NCEP) Global Forecast System (GFS) and are available in the stext (SHIPS) files*.

To supplement the synoptic predictors, proxies for atmospheric stability and initial condition error were also computed. The proxies for atmospheric stability include the forecasted intensity change during the forecast period, deviation of each model's intensity forecast from the mean of all the models (deviation from ensemble mean: DFEM), the standard deviation of the intensity forecasts (spread), and the spread of the track forecasts. The only proxies for initial condition error considered were the previous 12-hour intensity change and the error of the previous 12-hour forecast that verifies at the time when the error prediction is made.

In a manner similar to the development of SHIPS (DeMaria and Kaplan 1994), standard multiple linear regression models were created from the twenty-nine dynamical parameters and proxies to predict the absolute error (AE) of the different tropical cyclone intensity models. The 2007-2011 Atlantic basin hurricane seasons served as the training period for the regression and the 2012 season was used as the independent verification period. Predictions of the AE were created for 24- through 120-hour intensity forecasts (with 24 hour increments). The model development sample excluded low pressure systems ("LO" in ATCF files), extratropical storms ("EX" in ATCF files), and storms that pass over land during the forecast period. As mentioned by DeMaria and Kaplan (1994), statistical properties of storms that are over land are different from the properties of storms over the ocean, so training the regression formula using both cases is not physically justified.

The standard multiple regression techniques that were used to develop the SHIPS model were followed here. Dependent and independent variables are normalized so the regression coefficients for different variables, forecast intervals, and forecast hours can be compared. A backward stepping stepwise regression procedure was used to select the predictors. The regression equation starts with all of the predictors and then the least significant predictor is removed. This process is repeated until the weighting coefficients associated with the predictors are all different from 0 at the 95% confidence level (F-statistic used). The preliminary results revealed that some predictors were never found to be significant or if found significant, were associated with physically unjustifiable coefficients. As a result, distance to land, longitude, previous 12-hour intensity change, and previous 12-hour intensity error were excluded.

Additionally, adjustments to predictors were made before making our error predictions. Initial and forecast latitude squared was added as a predictor because AE exhibited a nonlinear relationship when transitioning from lower to higher latitudes throughout the models. Storms at lower latitudes had much higher AEs in all the models. Additionally, sine of shear direction replaced shear direction as a predictor to avoid the illusion that shear pointing from 359 degrees is significantly different from shear pointing from 1 degree. With these changes, there were 24 independent variables that were inputted into the multiple linear regression formula.

3. RESULTS

For each of the four models and five forecast times, stepwise linear regression was carried out using the 24 synoptic variables and proxies to predict AE during the 2012 Atlantic basin hurricane season. The percent variance of intensity forecast AE that could be explained for the different models ranged from 0-4% for 24-hour forecasts, 2-9% for 48-hour forecasts, 4-13% for 72-hour forecasts, 4-18% for 96-hour forecasts, and 2-31% for 120-hour forecasts. The exact R^2 values for each of the models and forecast times are included in Table 1. Table 1 shows that the percent variance explained is larger for longer forecast periods. This trend is likely due to two reasons. First, the intensity values in the best-track database are rounded to 5-knot increments; this value represents about 33%-50% of the average forecast error for short-range forecasts. Shorter forecasts of AE are likely affected by this noise (DeMaria and Kaplan 1994). Secondly, the variance in the true AE is much larger for longer forecast intervals, which can lead to higher R^2 values.

* Files are available at ftp://rammftp.cira.colostate.edu/demaria/SHIPS/stext_oper/.

	Forecast Period (hr)				
	24	48	72	96	120
DSHP	0.01	0.09	0.13	0.06	0.14
LGEM	0.02	0.06	0.12	0.18	0.31
HWFI	0.02	0.06	0.09	0.07	0.17
GHMI	0.00	0.02	0.04	0.04	0.02

Table 1. R^2 between predicted AE and true AE for DSHP, LGEM, HWFI, and GHMI at each forecast period for intensity forecasts in the Atlantic basin (2012 hurricane season).

For the GHMI and HWFI models, DFEM is consistently the leading predictor. This predictor has a weighting coefficient that is different from zero at the 99% significance level for both models at every forecast time. These results were expected based on DFEM's relationship with AE for the 2007-2011 hurricane seasons. During these seasons, the correlation between DFEM and AE for the GHMI and HWFI model ranged from 0.23-0.50 and 0.27-0.46 respectively (depending on the forecast period). However, this predictor might be weighted too heavily when being applied to independent data because HWFI and GHMI show low R^2 values for the 2012 AE predictions. This trend can be explained by HWFI and GHMI completing effective model updates that improved the models more than the statistical models (Cangialosi and Franklin 2013). As a result, these dynamical models often achieve lower AE in 2012 compared to 2007-2011 so that deviations from the ensemble mean likely indicate they are identifying a more accurate intensity change than the statistical models. If this is the case, deviating from the ensemble mean would lead to lower errors and the high positive correlations with AE would not be justified.

The statistical models, LGEM and DSHP, generally yield higher R^2 values. This result could be a byproduct of the more consistent model formulations throughout the training and verification period or the fact that GFS output is also used to train these models. For LGEM, forecast average divergence and DFEM were the leading predictors. Both of these predictors had weighting coefficients that are different from zero at the 99% significance level for four out of the five forecast times. For DSHP, forecast intensity and intensity forecast spread were the leading predictors. Both of these predictors also had weighting coefficients that are different from zero at the 99% significance level for four out of the five forecast times.

To better understand the varying levels of success between the different models and forecast times, two model-forecast hour combinations are selected for further analysis: 120-hour LGEM and 24-hour GHMI.

3.1 120-HOUR LGEM ERROR PREDICTIONS

The 120-hour LGEM AE predictions achieved the highest R^2 among all model-forecast hour pairs. Figure 1 shows a scatter plot displaying the relationship between LGEM AE and predicted error for 120-hour forecasts. The corresponding regression equation is included below:

Absolute Error = -0.21 X (0 hour Intensity) + 0.16 X (0 hour Shear) + 0.17 X (0 hour Shear Direction) + 1.58 X (0 hour Latitude) – 1.09 X (0 hour Latitude^2) + 0.28 X (0 hour Divergence) – 0.30 X (0 hour Storm Speed) - 0.54 X (0 hour Relative Humidity) - 0.80 X (Forecast Average Latitude) - 0.25 X (Forecast Average Divergence) + 0.23 X (Forecast Average Storm Speed) + 0.40 X (Forecast Average Relative Humidity) (1)

The training dataset consists of 318 verified forecasts and the testing dataset consists of 91

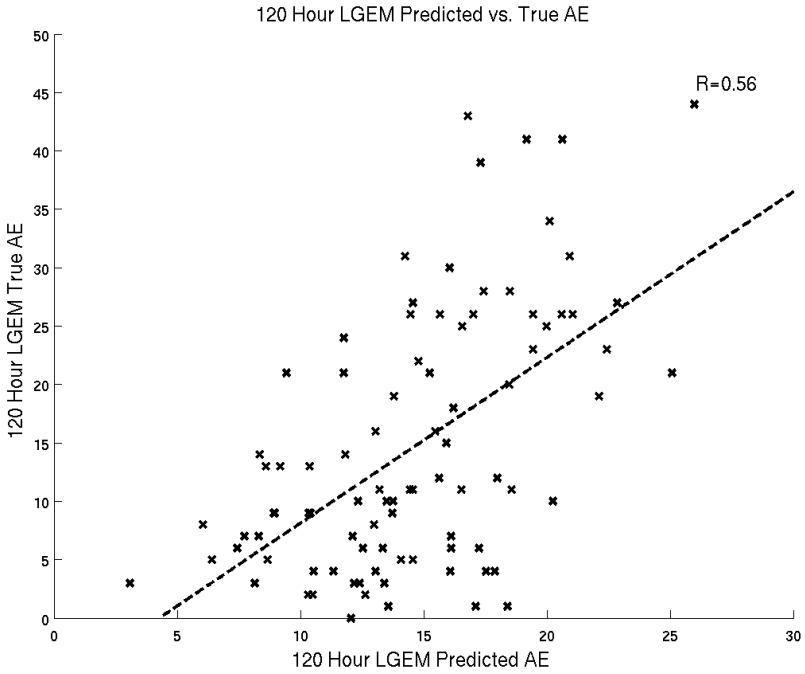

FIG. 1. Observed AE for 2012 120-hour LGEM forecasts as a function of predicted AE derived from multiple linear regression of the 2007-2011 data. The dashed line indicates the least squares regression line, and the R value of the line is included in the top right corner of the figure.

verified forecasts. Each 'x' in Figure 1 represents one verified forecast where the dependent variable is the true AE and the independent variable is the predicted AE using regression equation (1). The dashed line represents the least squares regression line of the data.

Even though the predicted AE appears to fit the true AE well, all the coefficient weights in equation (1) do not follow physical reasoning. The regression technique identified twelve predictors as statistically significant but three of these variables involve latitude, two involve divergence, two involve storm speed, and two involve relative humidity. More importantly, the predictor coefficients for these cases have opposite signs but in reality, do not have inverse relationships with forecast error. This inconsistency is compensated for in equation 1, because the larger weighting coefficient(s) of the pair (or three) is consistent with the correlations observed during the training dataset. For example, the coefficient for 0 hour storm speed is a larger negative value than the positive coefficient for forecast storm speed. During the training dataset, LGEM AE has a negative correlation with both 0 hour and forecast average storm speed. Ideally, the regression technique should select only the more significant predictor from the forecast average and initial value or include both with the correct sign.

This error is an artifact of the regression overfitting the results based on the multiple predictors involving the same dynamical variable. For the regression to perform the best on an independent dataset, the coefficients need to have physically justified magnitudes and signs. Regardless, the R^2 for the error predictions in Figure 1 is 0.31 which indicates this method has promise in an operational setting.

3.2 24-HOUR GHMI ERROR PREDICTIONS

The 24-hour GHMI AE predictions achieved the lowest R^2 among all model, forecast hour pairs. Figure 2 shows a similar scatter plot to Figure 1 except it displays the relationship between GHMI AE and predicted AE for 24-hour forecasts. The regression equation used to predict AE is included below:

Absolute Error = 0.15 X (Forecast Intensity) – 0.12 X (Forecast Latitude^2) + 0.22 X (Forecast Intensity Spread) + 0.1 X (DFEM) (2)

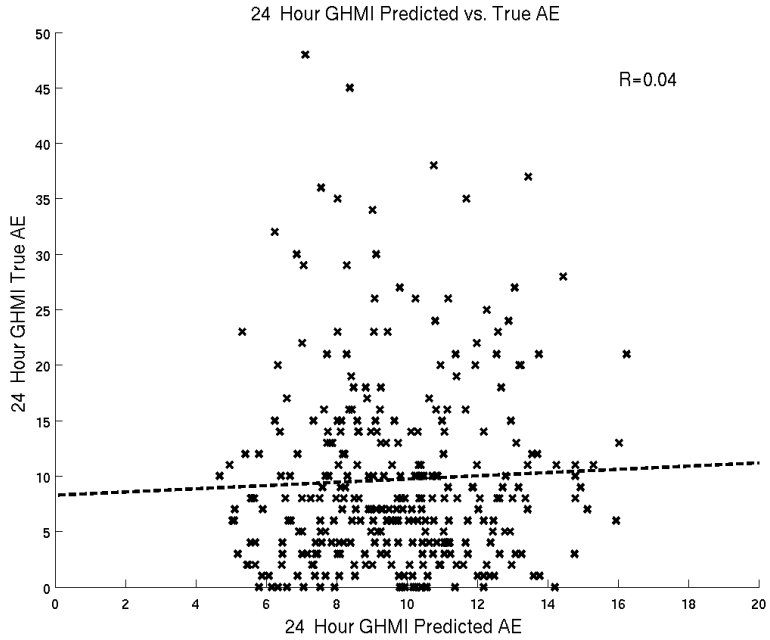

FIG. 2. Observed AE for 2012 24-hour GHMI forecasts as a function of predicted AE derived from multiple linear regression of the 2007-2011 data. The dashed line indicates the least squares regression line, and the R value of the line is included in the top right corner of the figure.

For this shorter forecast period, the datasets are much larger with a training dataset of 967 verified forecasts and a testing dataset of 307 verified forecasts. Clearly, for this model-forecast time pair, predicted AE does not explain the variance of true AE well. A possible explanation for this lower R^2 can be made by revisiting two already mentioned points. The noise in the data from the rounding of the best-track data adds some systematic error to short-range AE predictions. Secondly, the GFDL model received significant upgrades in 2012 as well some smaller upgrades throughout the training period (Bender 2012). Since GHMI is derived from GFDL, it has been affected by the same model upgrades as the parent model. The effects of these changes to the model are apparent when comparing the regression formula (not shown) obtained using the 2012 season as the training period instead of 2007-2011. The 2012 regression formula for 24-hour GHMI error predictions only has two of the same significant predictor variables, forecast intensity and DFEM, along with six new predictors. Clearly, GHMI AE is exhibiting different trends in the training and verification period.

Adding more skillful predictors that equally applicable to the training period and verification period might be difficult for the models when upgrades are made. Finding predictors less affected by model upgrades or adjusting the training period length (picking a training period where the dynamical model is similar to the verification period dynamical model) will be necessary to improve the R^2 values moving forward.

4. SUMMARY AND FUTURE WORK

A combination of dynamical, atmospheric instability, and initial condition uncertainty parameters were used as the independent variables in multiple linear regression formulas to predict TC intensity forecast AE. The percent variance of the GHMI, HWFI, LGEM, and DSHP intensity forecast AE that could be explained for the 2012 Atlantic season ranged from 0% to 31%. More methodical testing and careful analysis is necessary to select a list of the most physically justified predictors before finalizing a list of independent variables to be inputted into an operational multiple linear regression formula.

R^2 values can potentially be improved for all forecast models and hours because currently all dynamical variables besides latitude are

assumed to have a linear relationship with forecast error. A linear fit is not ideal for some of the synoptic predictors. For example, BN13 found that medium shear values have high AE while low and high shear values have lower AE; this trend is not described with a linear function.

It might be worthwhile to test additional dependent variables in the regression model as past studies have showed the benefits of using AAC and RMSE instead of AE (see review by Ehrendorfer 1997). In addition, probabilistic error forecasts will be added to the deterministic forecasts based on the selected predictors.

5. REFERENCES

Bender, Morris. “2012 Upgrades to the Operational GFDL/GFDN Hurricane Model.” 30th Conference on Hurricanes and Tropical Meteorology in Ponte Vedra Beach, Florida. Monday, 16 April 2012.

Bhatia, Kieran T., David S. Nolan, 2013: Relating the Skill of Tropical Cyclone Intensity Forecasts to the Synoptic Environment. Wea. Forecasting, 28, 961–980.

Cangialosi, J.P. and J.L. Franklin, 2013: 2012 National Hurricane Center Forecast Verification Report, National Hurricane Center.

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

DeMaria, M., and J. Kaplan, 1998: An updated statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic and eastern north Pacific basins. Wea. Forecasting, 14, 326-337.

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.

Ehrendorfer, M., 1997: Predicting the uncertainty of numerical weather forecasts: A review. Meteor. Z., 6, 147–183.