3A.9 PREDICTING THE EXPECTED SKILL OF TROPICAL CYCLONE INTENSITY FORECASTS USING ENVIRONMENTAL PARAMETERS

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

1. INTRODUCTION

It is commonly accepted that an a priori expectation of a forecast's skill is a necessary part of every forecast (Kalnay and Dalcher 1987; Molteni and Palmer 1991; Tennekes et al 1987; Palmer and Tibaldi 1988). The degree of skill in forecasting the atmosphere varies based on initial condition errors, imperfect model formulations, and the inherent uncertainty of different atmospheric states (Kalnay and Dalcher 1987). An estimation of the effects of such uncertainties on forecast accuracy is imperative to quantify the amount of confidence that should be allotted for an individual forecast. Even with the obvious benefits of skill predictions, to the author's knowledge, there have been no studies applying skill predictions to hurricane intensity forecasts. This is particularly surprising because operational tropical cyclone track forecasts have improved dramatically in the past 20 years, while intensity forecasts have lagged behind and in some cases regressed. According to the National Hurricane Center's (NHC) official verification results, the average forecast error from 1990 to 2010 for the 24 hour to 72 hour official (OFCL) intensity forecasts have shown below 1 knot improvement and in the case of the 24 hour forecast, the forecast error has increased (Cangialosi and Franklin 2011).

Therefore, there could be great value of forecasts of forecast skill for tropical cyclone intensity models. Firstly, by knowing when models are consistently underperforming or succeeding, forecasters can know what situations produce forecasts that deserve higher or lower confidence. For example, if a tropical cyclone is approaching land and models are in a high confidence regime, then emergency managers can bolster their evacuations and storm preparations accordingly as a result of the larger reliability of a land-falling prediction. Secondly, a handful of statistical, dynamical, and "hybrid" (a mixture of the two) models have recently been developed but no individual model consistently outperforms the others (DeMaria and Gross 2003). If forecast solutions diverge, knowledge of which model is reliable in a given situation can help NHC forecasters decide which model to favor and consequently produce better verifying forecasts. Finally, if skill forecasts reveal that certain environmental conditions or the level of variability in the current flow pattern (consistency between adjacent forecasts, skill of earlier short-range forecasts, etc.) consistently lead to less or more reliable forecasts then further investigation into these regimes are obligatory. Modelers can focus their efforts into improving a model in these less reliable situations and explore the dynamical mechanisms that cause low confidence regimes.

As a result of the potential value of skill predictions, the main goal of this preliminary investigation into tropical cyclone intensity forecast improvement is to test a variety of environmental parameters' ability to predict forecast skill. These results should address conventional wisdom about which environmental conditions lead to better forecasts of hurricane intensity and highlight the different strengths of each model. Then, in future work, the author will statistically select which predictors of skill perform the best and produce a probabilistic forecast of confidence to accompany each intensity forecast.

In this study, the four hurricane intensity models that were operational for the duration of the 2006-2010 hurricane seasons, as well as the official forecast (OFCL), are evaluated based on different performance metrics. The four models include the Logistic Growth Equation Model (LGEM), the Statistical Hurricane Intensity Prediction Scheme (SHIPS), SHIFOR5 model (updated version of SHIFOR), and the GFDL hurricane model. The better-performing inland decay version (DSHP) of the SHIPS model was used instead of SHIPS. Each model's performance is assessed by computing the mean absolute error (MAE), bias, and skill relative to the SHIFOR5 model for 24, 48, and 72 hour forecasts in the Atlantic basin. These performance metrics are binned based on certain environmental conditions ("predictors") and computed for each of the different models.

2. METHODOLOGY

The 24, 48, and 72 hour intensity forecasts and 0 hour verification for the GFDL, OFCL, and

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

SHF5 models were obtained from NOAA's ATCF database. The predictor values as well as the DSHP and LGEM forecasts and verification data came from ftp://rammftp.cira.colostate.edu/demaria/ships/stext_oper/. The predictors at this site are outputs from the GFS model. The predictors tested include initial intensity, storm speed, initial shear (850-200 hPa), potential intensity, shear direction, latitude, and the average of each of these predictors during the forecast period (i.e. for a 24 hour forecast, the average of each 6-hourly forecasted shear up until 24 hours).

To evaluate the skill of intensity forecasts for each model, histograms were made for individual predictors by selectively binning a predictor and plotting the performance metrics for the forecasts based on those bins. The forecast error values necessary for computing MAE, bias, and skill were obtained by comparing each model's intensity forecast to the corresponding 0 hour operational analysis (not by verifying against best track). Two-variable plots were created by graphing the performance metrics against two predictors. Each square on these two variable plots represents a range of values for each predictor, and the square is shaded a different color to indicate the magnitude of MAE, skill, or bias. Binning was accomplished through two different methods: either dividing the data into 3 approximately equal sized bins or selecting arbitrary bin sizes to gain insight on whether certain ranges yielded anomalous results. The "equal sized" bins were determined by taking all the predictor values for the different models and splitting them up into thirds based on the values of the predictor.

3. RESULTS

Figure 1 and 2 provide examples of how a histogram and a 2-variable plot demonstrate particular intensity models are more adept in certain environmental conditions as well as what conditions lead to low predictability for all models. Figure 1 contains four arbitrarily binned histograms for each model with skill plotted against 24 hour average forecast shear. The histograms show skill values that vary significantly across the bins. However, it is clear that the shear range 15 to 22.5 knots displays the lowest skill in every model. Yet, a paired t-test reveals that these differences among each model's bins are not significant at the 95% confidence level so they are not statistically robust. Figure 1 is particularly insightful because the differences among bins appear significant but a t-test confirms that this is a false conclusion.

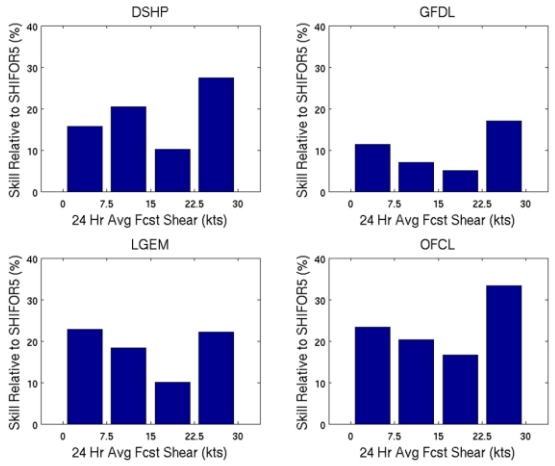

Figure 1. Skill of the 24 hour intensity forecast is plotted against the 24 hour average forecast shear for DSHP (top left), GFDL (top right), LGEM (bottom left), and OFCL (bottom right) models.

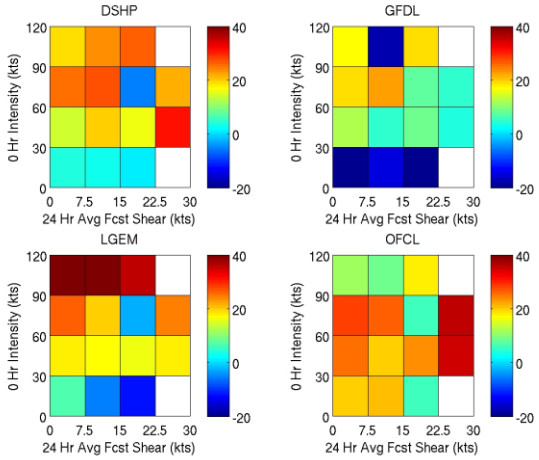

Figure 2. Skill of the 24 hour intensity forecast is contoured against 24 hour average forecast shear and 0 hour intensity for DSHP (top left), GFDL (top right), LGEM (bottom left), and OFCL (bottom right) models.

Adding a second predictor provides a more detailed description of the synoptic situations that are associated with the most skillful forecasts. Bin ranges that contain less than fifteen forecasts are excluded. A lot of information can be gleaned from Figure 2 but an especially unique pattern is visible for high intensity, low shear bins. These specific cases are forecasted exceptionally well by LGEM but poorly by OFCL and GFDL. The GFDL model is very unreliable (~ -20% skill) for all forecasts made with 24 hour forecast shear between 7.5 to 15 knots and initial intensity between 90 and 120

knots while LGEM is at a skill level in the 40 percent range. These results suggest that OFCL forecasts should weight LGEM more for these synoptic situations. The bootstrap method was used to compare the same bin between different models. This statistical technique, due to the fact there was less than 40 cases in this bin, showed that the difference in the mean skill of the mentioned bin for LGEM and GFDL is not significant at the 95 percent confidence level. More cases are needed before this result can become statistically significant.

In Figure 3, the intensity at the time of the forecast is plotted against the bias of the 48 hour intensity forecast. The large bias of the forecasts in the 90 to 120 knot intensity bin are very large in all models. In fact, for each model, the mean bias in this bin is significantly different at the 95 percent confidence level from the bias in all the other bins (one exception is the 60-90 knot intensity bin for GFDL). This statistical technique revealed the bias in the OFCL model's 90-120 knot intensity bin is significantly larger (at the 95% confidence interval) than the LGEM model. Figure 3 reflects that when the differences in the performance metric between bins becomes large, the sample size is sufficient to maintain statistical significance among the bins.

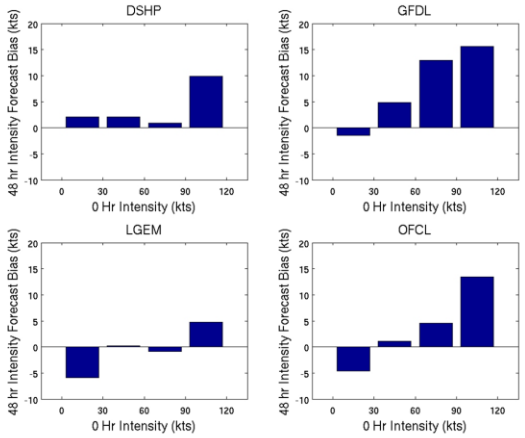

Figure 3. 48 hour forecast bias is plotted against the 0 hour intensity for DSHP (top left), GFDL (top right), LGEM (bottom left), and OFCL (bottom right) models.

Finally, by using fewer bins in the two figure plots, significance between bins can be reached. In Figure 4, MAE is plotted against 24 hour average forecast shear and 0 hour intensity. The binning ranges are derived from the previously mentioned "equal binning technique" for each predictor. By focusing on the low shear, high intensity (0-9 knots shear, greater than 61 knots initial intensity) for the GFDL, LGEM, OFCL, and DSHP models, it is clear that LGEM outperforms the other models. The MAE of the mentioned LGEM bin is significantly less than the corresponding OFCL bin at the 95 % confidence interval. With this level of confidence in the fact that the difference in the means of the bins is significant, one can realistically say at the time of a forecast that if your tropical cyclone is above 60 knots in intensity and is moving into region with shear between 0-9 knots that your OFCL forecast should rely heavily on the LGEM forecast. Similar conclusions can be made for other models in different circumstances.

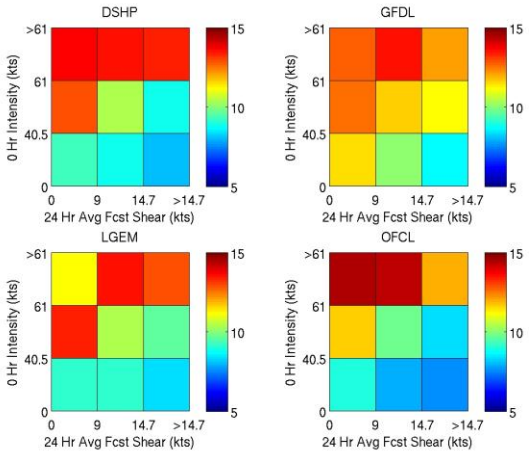

Figure 4. MAE of the 24 hour intensity forecast is contoured against 24 hour average forecast shear and 0 hour intensity for DSHP (top left), GFDL (top right), LGEM (bottom left), and OFCL (bottom right) models.

From these four figures alone, forecasters are able to discern what situations produce poor model forecasts and can gain intuition which models to trust and how much confidence to put in a forecast.

4. CONCLUSIONS AND FUTURE WORK

The results from this study offer a new approach for improving tropical cyclone intensity forecasts. For all the cases analyzed during this investigation, LGEM was the most skillful model for the 48 hour and 72 hour forecasts and almost identical in performance to the OFCL for the 24 hour forecasts. The GFDL model was significantly worse than the other models for all forecast times. Histograms and two-variable plots were created for a variety of predictors and three forecast times and countless conclusions are possible based on the results. Four figures were presented to provide

an example of the inferences possible from these plots. Expanding these results could lead to more knowledge about the reasons for intensity forecasts' poor performance in recent years and lead to better forecasts in the future.

5. REFERENCES

Cangialosi, J.P. and J.L. Franklin, 2011: 2010 National Hurricane Center Forecast Verification Report, National Hurricane Center, NOAA/NWS/NCEP/Tropical Prediction Center.

DeMaria, M., and J. M. Gross, 2003: Evolution of prediction models. Hurricane! Coping with Disaster: Progress and Challenges since Galveston, 1900. R. Simpson, Ed., Amer. Geophys. Union, 103–126.

DeMaria, M., Mainelli, M., Shay, L. K., Knaff, J. A. and Kaplan, J. 2005: Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). *Wea. Forecast.* **20**, 531–543.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.*, **115**, 349–356.

Molteni, F. and T. N. Palmer, 1991: A real-time scheme for the prediction of forecast skill. *Mon. Wea. Rev*., **119**, 1088–1097.

Palmer, T. N. and S. Tibaldi, 1988: On the prediction of forecast skill. *Mon. Weather Rev.*, **116**, 2453-2480.

Tennekes, H., Baede, A. P. M. and Opsteegh, J. D. (1987) Forecasting forecast skill. In: Proceedings ECMWF Workshop on Predictability, Reading, April 1986.