3C.2

THE HFIP HIGH-RESOLUTION HURRICANE FORECAST TEST: OVERVIEW AND RESULTS OF TRACK AND INTENSITY FORECAST VERIFICATION

L. Bernardet[1&], L. Nance[2], S. Bao[1&], B. Brown[2], L. Carson[2], T. Fowler[2], J. Halley Gotway[2], C. Harrop[1&], J. Wolff[2]
[1]NOAA Earth System Research Laboratory, Boulder, CO
[2]National Center for Atmospheric Research, Boulder, CO
[&]Also affiliated with CIRES, University of Colorado, Boulder, CO

1. INTRODUCTION

The NOAA Hurricane Forecast Improvement Project (HFIP) High-Resolution Hurricane (HRH) Test was conducted by the Developmental Testbed Center (DTC, Bernardet et al. 2008) and collaborators from March 2008 through September 2009 in order to assess the impacts of using higher horizontal resolution in hurricane numerical forecasting, with a special focus on intensity forecasting. The plan for this test was developed jointly by a broad range of community members, including specialists in hurricanes, numerical modeling, and forecast verification. The test focused on 69 retrospectives cases from the 2005 and 2007 hurricane seasons. Six independent modeling groups participated in this effort employing three configurations of the Weather Research and Forecasting (WRF) model, the operational Geophysical Fluid Dynamics Laboratory (GFDL) model, the Naval Research Laboratory (NRL) tropical cyclone model, and a model from the University of Wisconsin-Madison (UWM).

This paper provides an overview of the HRH Test and summarizes the *intra*-model differences in track and intensity errors between the lower and higher resolution configurations of the participating forecast models. It should be noted that *inter*-model comparisons were not an objective of this test. A comprehensive website for the HRH Test is at http://www.dtcenter.org/plots/hrh_test.

2. PARTICIPATING MODELS

Six modeling groups participated in the HRH test. The range of horizontal resolutions each modeling group provided is summarized in Table 1. Each participating model was run for at least two resolutions, with a couple of models run for three resolutions. For models at two resolutions, two separate runs were done for each case: the first at low resolution, and the second employing higher-resolution. Similarly, groups providing three resolutions ran the model using three different configurations for each case, using progressively higher resolution. These separate runs were necessary because the analysis of the low-resolution grids could be done on configurations contaning a high-resolution nest due to feedback from the high resolution nest.

For a detailed description of each model configuration and for the definition of the acronyms listed in Table 1, see http://www.dtcenter.org/plots/hrh_test/model/index.php.

Table 1. Modeling groups that participated in HRH and the models used to generate retrospective forecasts.

Institution (Contact)	Model	Grid Spacings (km)		
		Low	Mid	High
NOAA AOML (S. Gopalakrishnan)	HWRF-X	9	3	-
NCAR MMM (Chris Davis)	AHW	12	-	1.3
NRL (Melinda Peng)	COAMPS-TC	9	3	-
PSU (Fuqing Zhang)	WRF-ARW	13.5	4.5	1.5
URI (Isaac Ginis)	GFDL	9	6	-
UWM (Greg Tripoli)	UW-NMS	12	3	1

3. CASES

A diverse set of storms and time periods from each of these storms was selected to feature a number of Rapid Intensification (RI) and Rapid Weakening (RW) events, defined as an increase (decrease) of the maximum sustained surface wind of more than 30 (25) kt in 24 h, for a storm that is over water. A total of 69 cases were selected from the following storms of the 2005 and 2007 seasons: Wilma, Philippe, Rita, Karen, Katrina, Humberto, Felix, Ingrid, Emily and Ophelia. Not every case was run using all models, and some model forecasts ended before the Best Track did. Additionally, there was a small portion of the submitted forecasts that the DTC could not evaluate because of difficulties in extracting the storm track from the forecast fields. The actual number of forecasts used for the evaluation is listed on the included figures.

4. METHODOLOGY

The input to the HRH evaluation system consisted of gridded data files in GRIB format provided by the modeling groups, and Best Track and TC Vitals storm message files provided by the National Hurricane Center (NHC). Each gridded data file delivered to the DTC contained the required input fields for the Vortex Tracker (zonal and meridional wind components and geopotential height at 850, 700 and 500 hPa, zonal and meridional wind components at 10-m, absolute vorticity at 850 and 700 hPa, and mean sea level pressure), plus temperature and dewpoint temperature at 2-m, 850, 700 and 500 hPa and 1-h accumulated precipitation used for plotting and display. These files contained forecasts every 30 minutes out to 126 h.

The HRH evaluation system (Fig. 1) was composed of the following elements: 1) Vortex Tracker, 2) intensity averager, 3) plotting, 4) NHC Verification System, 5) aggregation and statistical significance assessment, 6) RI/RW verification, 7) consistency verification, and 8) archival. Some of these steps are described in more detail below.

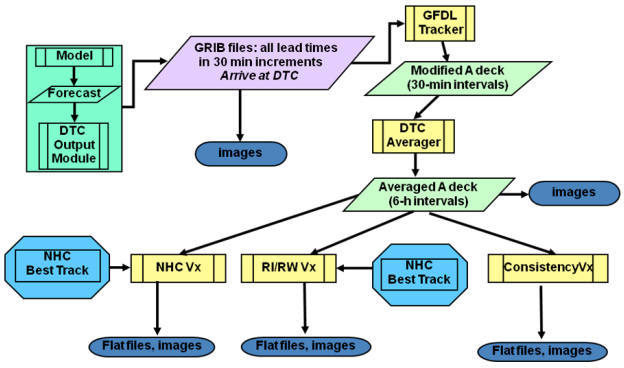

Figure 1. Schematic of the DTC evaluation system for HRH.

A. Vortex Tracker and averager

A revised version of the Geophysical Fluid Dynamics Laboratory (GFDL) Vortex Tracker (Marchok 2002, Gopalakrishnan et al. 2010) was used to locate the storm in each forecast. GFDL implemented several modifications to make the standard version of the Vortex Tracker suitable for the HRH Test, including the ability to read subhourly forecasts, and to process high-resolution moving nests. In spite of all the enhancements, the tracker was not able to follow all storms in the forecasts processed for the HRH Test. Forecasts for which the storm was weak or disorganized, even if matching the observed storm, could not be tracked. If the tracker could not find the storm at a given forecast lead time, it was not able to locate the storm at longer lead times, regardless of whether the storm became organized later. Hence, a disorganized storm that develops into an organized storm may not have been included in the evaluation.

Because the forecasted maximum surface winds (MSW) for tropical storms can vary significantly over a small time period, the verification of maximum winds was based on the average of the maximum winds over a two-hour window centered at the verification time. The 2-hour average was computed using tracker output data at 30-minute intervals (i.e., average over five data points). Data at minus 30 and 60 minutes were not available for forecast lead times at the beginning of the forecast, so a one-sided 1-hour average was computed for the initial time. Only maximum winds were averaged; that is, storm location and extent of wind radii were not averaged.

B. NHC Verification System

The NHC Verification System was used to verify the forecasts against the NHC Best Track. Only the tropical portion of the tracks was verified. Each case was processed separately; that is, the input for each run of the NHC Verification System included a single forecast for one model resolution. All forecasts were run through the NHC verification system twice: once for the complete forecast track and a second time for the portion of the track that was over water only. For the latter, only situations in which both the observed and forecast storm centers were over the ocean were considered. Variables verified include: storm location, averaged MSW, and extent of wind radii. Metrics generated included: absolute, cross- and along-track error, MSW error, and wind radii error.

C. Aggregation and significance assessment

Results from the individual runs of the NHC Verification System were aggregated using a script in the R statistical language. Given the distribution of errors and absolute errors at a given lead time, several parameters of the distribution were computed: mean, median, quartiles, and outliers. Confidence intervals (CI) on the median were computed using a parametric method. Only lead times and errors for which the distribution contained at least 11 samples were considered in the statistical significance discussions because the error

distribution parameters could not be properly computed for sample sizes less than that. Skill scores for all models at all resolutions were also computed against the Decay-SHIFOR5 (OCD5), but will not be discussed in this paper.

A pairwise technique was used to address the question of whether the differences between a medium- or high-resolution model configuration and their low-resolution counterpart are statistically significant (SS). For this technique, the absolute error of a given quantity (for example, track error) for a medium- or high-resolution forecast was subtracted from the same metric for the low-resolution forecast. This subtraction was done separately for each lead time of each case, yielding a distribution of forecast error differences. The parameters of this difference distribution were then computed using the same methodology applied to the error distributions for a single model resolution. The pairwise method enabled the identification of subtle differences between two error distributions that might have gone undetected when the mean absolute error or root mean square error of each distribution was computed and the overlap of the CIs for the mean was used to ascertain differences (Lanzante 2005).

A SS difference between the forecast verification metrics of the multiple resolutions for a given model was noted when it was possible to ascertain with 95% confidence that the median of the pairwise differences was not equal to zero. The median was chosen over the mean for this analysis because it is a robust measure, appropriate for this test, in which some distributions were normal while others differed from normality and presented outliers.

Boxplots provide a concise format for displaying the various attributes of the error distributions computed for the HRH Test. The mean of the distribution is depicted as a star and the median as a bold horizontal bar. The 95% CIs are shown by the height of the *notch* of the boxplot, and the outliers are represented by open circles. In the boxplots of error differences (Figs. 2b, 3b, and 4b), positive (negative) values indicate that the higher (lower) resolution configuration performed better.

5. RESULTS

Only a sample of the results will be discussed here. For detailed results, see Developmental Testbed Center 2009. Results from the high-resolution configurations run by PSU and UWM will not be presented since too few cases were submitted for evaluation. The results shown here include all forecasts for which the observed storm was in its tropical phase, regardless of whether the observed or forecast storm was over land or water.

A. HWRF-X model run by AOML

The median and spread of the track errors increase with lead time for both AOML configurations (Fig. 2a). The track error for the low-resolution configuration (AOM1) undergoes a larger increase than that for the medium-resolution configuration (AOM2), leading to a SS difference for which the medium resolution is up to 10 nm more accurate for lead times 30 to 48 h (Fig. 2b).

The median of the absolute intensity error distributions for the two AOML configurations does not exhibit any strong trends with lead time (Fig. 3a). SS differences for absolute intensity errors occur for 0 to 6 h and 24 to 30 h with intensity improvement for the medium-resolution configuration on the order of 5 kt (Fig. 3b).

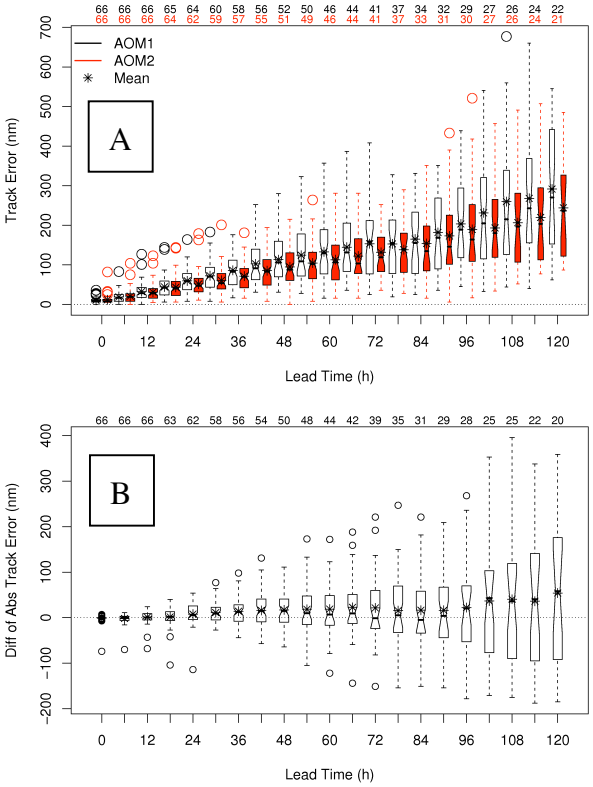

Figure 2. Track error distributions with respect to lead time for the a) low- (AOM1) and medium-resolution (AOM2) configurations of the AOML model and b) low - medium resolution (AOM1-AOM2) difference. The sample size is indicated on the upper part of the plot.

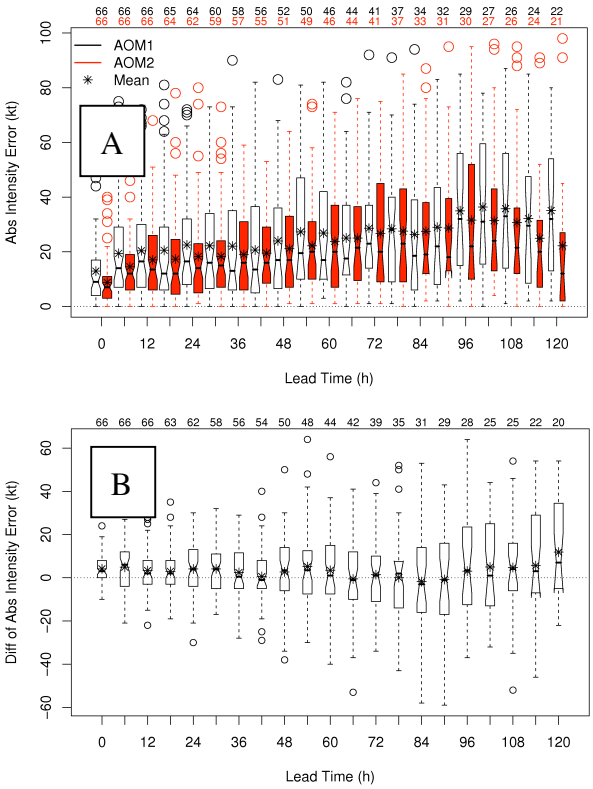

Figure 3. Absolute intensity error distributions with respect to lead time for the a) low- (AOM1) and medium-resolution (AOM2) configurations of the AOML model and b) low - medium (AOM1-AOM2) difference. The sample size is indicated on the upper part of the plot.

B. AHW model run by NCAR

The median of the track errors, as well as the spread in these errors increases with lead time for both MMM configurations (see Fig. 4a). The track error for the low-resolution configuration (MMM1) undergoes a larger increase than that for the high-resolution configuration (MMM3) at longer lead times, leading to SS differences for which the high resolution is more accurate for lead times 84 to 114 h (Fig. 4b). These SS differences correspond to a track improvement on the order of 25 nm.

The medians of the absolute intensity error distributions for the two MMM configurations do not exhibit any strong trends with lead time and the spread exhibits only a small increase (see Fig. 5). Only one SS difference occurs for absolute intensity errors (18-h lead time) with intensity improvement for the low-resolution configuration of less than 5 kts (not shown).

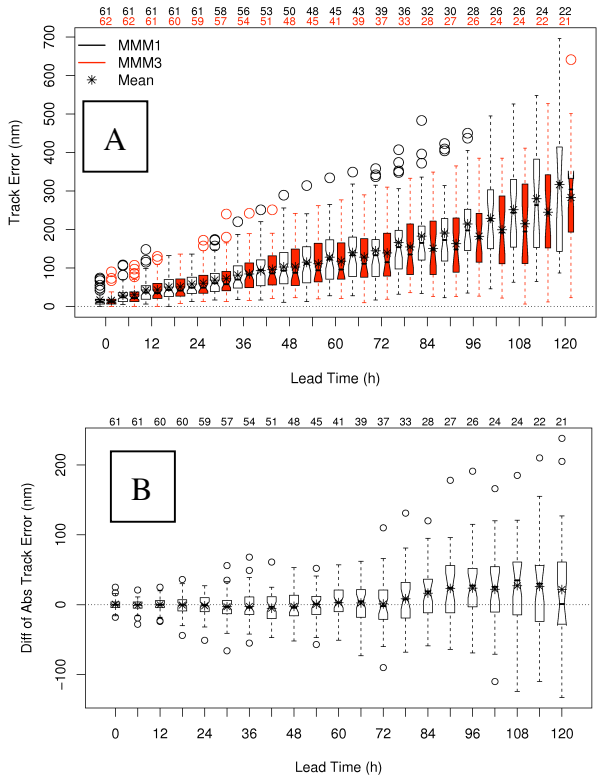

Figure 4. Same as Fig. 2, except for low- (MMM1) and high-resolution (MMM3) configurations of the MMM model.

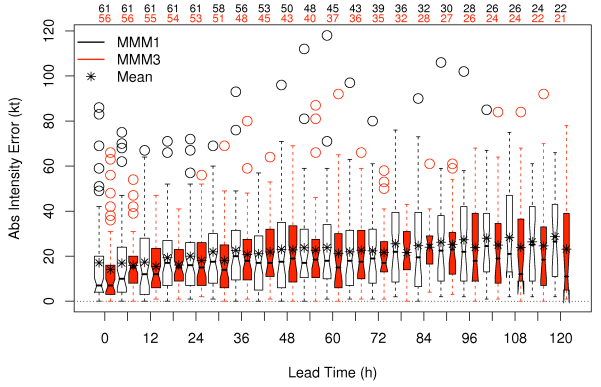

Figure 5. Same as Fig. 3a, except for low- (MMM1) and high-resolution (MMM3) configurations of the MMM model.

C. COAMPS-TC model run by NRL

The median and spread of the track errors increase with lead time for both NRL configurations (see Fig. 6). Median errors start near zero and grow to 300 nm for the 5-day forecast. The track errors for the two configurations increase at differing rates such that SS track error differences exist at lead times

24, 42, 54, and 96 h, for which the higher-resolution configuration is more accurate at 24 h and the high-resolution configuration degrades the forecast for the latter three lead times which exhibited SS differences (not shown). These SS results correspond to a maximum track error difference of 20 nm.

The absolute intensity errors for the two configurations do not grow in time. Rather, their median peaks at the three-day forecast and decreases thereafter (Fig. 7). The absolute intensity errors are mainly due to bias (not shown): both configurations systematically underpredict intensity out to 90 h, and the low-resolution configuration extends this underprediction to the 5-day forecast. SS differences for absolute intensity errors occur at 0-, 6-, 24, and 48-h lead times, with the medium resolution improving intensity by up to 5 kt by reducing the underprediction (not shown).

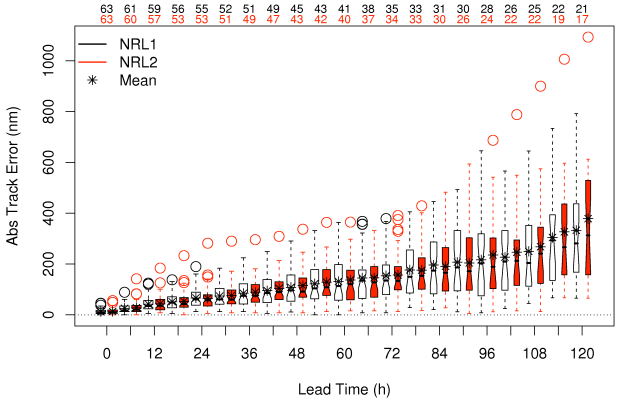

Figure 6. Same as Fig. 2a, except for the low-(NRL1) and medium-resolution (NRL2) configurations of the NRL model.

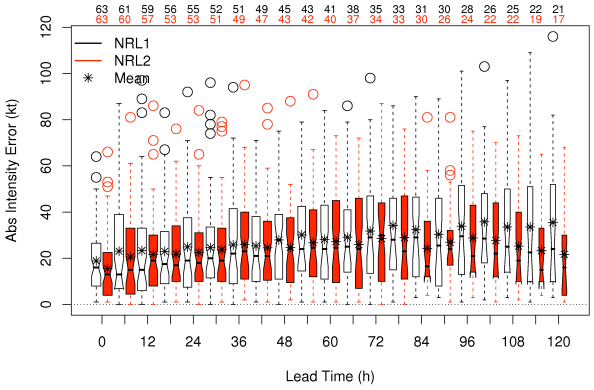

Figure 7. Same as Fig. 3a, except for the low-(NRL1) and medium-resolution (NRL2) configurations of the NRL model.

D. GFDL model run by URI

The median track errors for both URI configurations are near zero at initialization time and grow with lead time, along with the error spread, to approximately 200 nm for the 5-day forecast (Fig. 8). The absolute track error showed no SS differences between the errors for the low- (URI1) and medium-resolution (URI2) configurations (not shown).

The intensity error distributions for the two URI configurations (not shown) indicate both resolutions tend to underpredict storm intensity for a few early lead times (0, 24, and 30 h), while overpredicting in the fourth day of forecasting. SS differences in absolute intensity errors between the resolutions (not shown) are noted at the initialization time (when the medium-resolution minimized the under forecasting) and at the 90- and 96-h lead times (when the medium-resolution exacerbated the overprediction by about 6 kt).

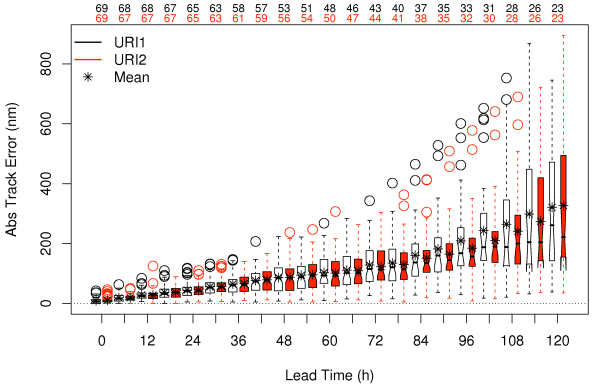

Figure 8. Same as Fig. 2a, except for the low-(URI1) and medium-resolution (URI2) configurations of the URI model.

E. NMS run by UWM

The median of the track errors for the low- and intermediate-resolution UWM configurations increases with lead time (Fig. 9). Only one SS difference between the low- (UWM1) and medium-resolution (UWM2) configurations occurs for track error (78 h), and it favors the low resolution (not shown).

The intensity error distributions for the low- and medium-resolution UWM configurations show that the low resolution tends to underpredict storm intensity for most lead times, whereas the intermediate resolution tends to overpredict the intensity for shorter lead times (see Fig. 10). These trends in error distributions are such that five SS differences favoring the

medium resolution occur in the 48 to 84 h time frame.

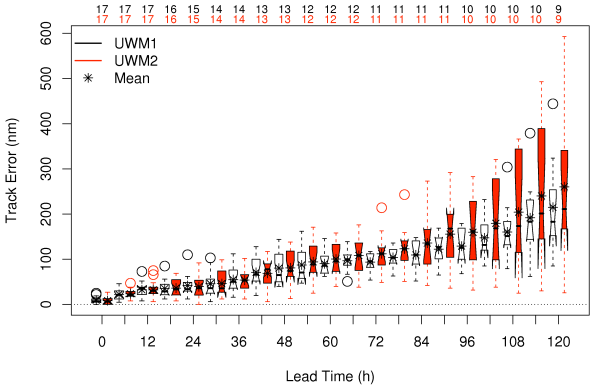

Figure 9. Same as Fig. 2a, except for the low- (UWM1) and medium-resolution (UWM2) configurations of the UWM model.

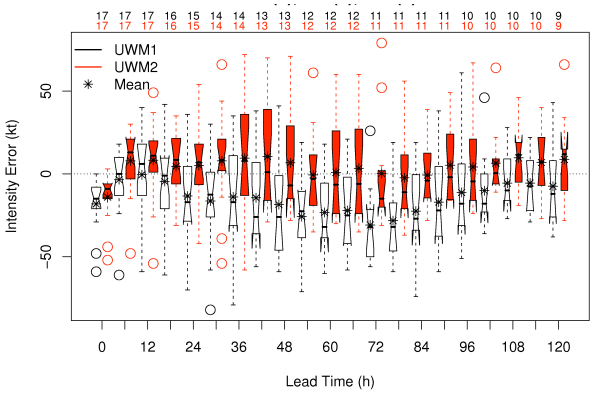

Figure 10. Same as Fig. 3a, except for the low- (UWM1) and medium-resolution (UWM2) configurations of the UWM model.

6. CONCLUSIONS

This paper presented the track and intensity error metrics of the HRH Test verification. These results indicate that running the models at higher resolution did not substantially improve the forecast. Some models showed limited improvements for track and/or intensity when higher resolution was used, but those were confined to a few lead times. The models from AOML, MMM and UWM had some positive results: higher resolution improved the AOML track forecast at two lead times, and the intensity forecast at four lead times, it improved the MMM track forecast at six lead times, and it improved the UW-NMS intensity forecast for 7 lead times. In the NRL model, higher resolution improved the intensity forecast but degraded track. URI showed little difference with resolution, perhaps because the model is hydrostatic.

These results need to be analyzed in concert with the results from the other HRH Test tools, which were used for verification of wind radii, RI/RW, and consistency (described in Developmental Testbed Center 2009). Wind radii forecasts in the higher-resolution configurations were degraded for the AOML, NRL, URI, and UW-NMS models (this verification was not computed for the MMM model), and the impact of high-resolution in the representation of RI/RW events and on forecast consistency was mixed (Nance et al. 2010).

In conclusion, the use of higher resolution in the participating models did not lead to an overall benefit in tropical cyclone forecasting as measured by the metrics used in this study. Improvement was noted for some metrics, lead times and models but the majority of results showed no SS difference in using higher resolution while a few, notably wind radii, presented consistent degradation when using higher resolution.

It is possible that the benefits of higher resolution were not fully realized in the participating models due to limitations, such as physics suites that are not appropriate for high-resolution, lack of a coupled ocean model, initialization techniques, or the model dynamics themselves (e.g., GFDL model is hydrostatic). Additionally, it is possible that the resolutions used in the test are not fine enough to resolve small scale structures such as updrafts and meso-vortices that may need to be represented in order to improve intensity forecasting.

We recommend diagnostic studies be conducted for a small sample of cases to determine if processes important to intensification are missing in the forecast. Once those are identified and addressed by the use of alternative physics suites and/or initialization techniques, new comprehensive tests can be conducted and it may then be possible that the benefits of high-resolution be realized.

7. REFERENCES

Bernardet, L., L. Nance, M. Demirtas, S. Koch, E. Szoke, T. Fowler, A. Loughe, J. L. Mahoney, H.-Y. Chuang, M. Pyle, and R. Gall, 2008: The Developmental Testbed Center and its Winter Forecasting Experiment. *Bull. Amer. Meteor. Soc.,* **89**, 611-627.

Developmental Testbed Center, 2009; High Resolution Hurricane Test Final Report.

Gopalakrishnan, S., Q. Liu, T. Marchok, D. Sheinin, N. Surgi, R. Tuleya, R. Yablonsky,

and X. Zhang, 2010: Hurricane Weather and Research and Forecasting (HWRF) Model scientific documentation. L. Bernardet, Ed., 75 pp.

Lanzante, J. R., 2005: A cautionary note on the use of error bars. *J. Climate*, **18**, 3699-3703.

Nance, L, L. R. Bernardet, S. Bao, B. G. Brown, T. L. Fowler, C. W. Harrop, E. J. Szoke, E. I. Tollerud, J. K. Wolff, and H. Yuan, 2010. The HFIP High Resolution Hurricane Forecast Test: Beyond the Traditional Verification Metrics. 29th Conference on Hurricanes and Tropical Meteorology, Tucson, AZ.

Acknowledgements: The authors would like to thank the participating modeling groups, as well as the support from the following scientists Timothy Marchok (GFDL), James Franklyn and Jack Beven (NHC), Mark DeMaria (NESDIS) and Mike Fiorino (ESRL). This work was supported by the NOAA Hurricane Forecast Improvement Project and was performed under the auspices of the DTC, which is funded by NOAA, the Air Force Weather Agency and NCAR.