P1.27

THE HFIP HIGH-RESOLUTION HURRICANE FORECAST TEST: BEYOND THE TRADITIONAL VERIFICATION METRICS

L. Nance[1*], L. Bernardet[2&], S. Bao[2&], B. Brown[1], T. Fowler[1], C. Harrop[2&], E. Szoke[2#], E. Tollerud[2], J. Wolff[1], H. Yuan[2&]
[1]National Center for Atmospheric Research, Boulder, CO
[2]NOAA Earth System Research Laboratory, Boulder, CO
[&]Also affiliated with CIRES, University of Colorado, Boulder, CO
[#]Also affiliated with CIRA, Colorado State University, Fort Collins, CO

1. INTRODUCTION

In the last 10 years, hurricane track forecast errors have been reduced by about 50% through improved model guidance, enhanced observations, and increased forecaster expertise, whereas little progress has been made toward reducing forecasted intensity errors. Recent research suggests that prediction models with grid spacing less than 1 km in the inner core of the hurricane may provide a substantial improvement in intensity forecasts (Powers and Davis 2002, Hendricks et al. 2004, Yau et al. 2004, Braun et al. 2006, Chen et al. 2007, Davis et al. 2008, Rotunno et al. 2009). The 2008-09 staging of the Hurricane Forecast Improvement Project (HFIP) High Resolution Hurricane (HRH) Test focused on quantifying the impact of increased horizontal resolution in numerical models on hurricane forecasting, with a special focus on intensity forecasting. The HRH test plan assembled by the Developmental Testbed Center (DTC) was developed jointly by a broad range of community members, including specialists in hurricanes, numerical modeling, and forecast verification. The focus of this test was intra-model differences resulting from changes in horizontal resolution, rather than inter-model comparisons. More detailed information about this project can be found at http://www.dtcenter.org/plots/hrh_test.

The HRH test focused on 69 retrospectives cases from the 2005 and 2007 hurricane seasons. These cases include a diverse set of storms and time periods from Wilma, Philippe, Rita, Karen, Katrina, Humberto, Felix, Ingrid, Emily and Ophelia featuring a number of Rapid Intensification (RI) and Rapid Weakening (RW) events. The DeMaria-Kaplan RI criteria defines an RI event as a 30 kt increase in maximum sustained surface wind (MSSW) in a 24-h period, whereas an RW event is defined as a 25 kt decrease in MSSW in a 24-h period. Both RI and RW events are restricted to time periods for which the storm is over water.

Six independent modeling groups participated in this test 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). The range of horizontal resolutions each modeling group provided is summarized in Table 1.

The DTC was tasked with providing objective verification statistics for the retrospective forecasts submitted by each modeling group. Bernardet et al. (2010) presents an overview of the HRH test and a summary of the intra-model differences in track and intensity errors. This paper discusses results obtained by applying new verification tools developed by the DTC that assess the dependency of RI/RW forecast skill and forecast consistency on horizontal resolution. Note that the high-resolution configuration of the UWM model will not be discussed in this paper because the sample size was too small for the RI/RW metrics to justify consideration.

Table 1. Modeling groups that participated in the HRH Test 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

*Corresponding author address: Louisa B. Nance, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307; email address: nance@ucar.edu

2. METHODOLOGY

The input to the HRH evaluation system consisted of gridded data files at 30-min intervals provided by the modeling groups and Best Track (BT) and TC Vitals storm message files provided by the National Hurricane Center (NHC). A modified version of the GFDL Vortex Tracker (Marchok 2002, Gopalakrishnan et al. 2010) was used to locate the storm and determine the basic properties of the storm in each forecast. Due to concerns about the potential for significant temporal variability in the instantaneous forecasted maximum surface winds (MSW) for tropical storms in high resolution models, the maximum wind or intensity used for this test was based on the average of the maximum wind output from the Vortex Tracker over a two-hour window centered at the verification time.

A. RI/RW

The RI/RW parameters considered by the DTC included: frequency of occurrence, timing of onset, and event-based contingency table scores for matched pairs with time relaxation. Missed forecast events stemming from the forecasted track being shorter than the observed track (i.e., lead times for which the tracker did not produce a fix) are not included in the sample.

To explore the properties of the observed and forecasted RI and RW events, total counts of RI and RW events were compiled for the entire sample partitioned by model configuration and resolution. The sample for each model is defined by the times for which tracker output is available for both resolutions. Hence, the numbers for BT may vary between models. A comparison of the medium- or high- and low-resolution configurations in the context of the observed occurrence frequency provides insight into whether the forecasted events occur more frequently, less frequently or at the same frequency as that observed. Two methods for defining an event were considered: 1) episodes – define any sequence of one or more periods of rapid change to be one event, and 2) events - define every period of rapid change to be an event. Figure 1 provides a graphical illustration of these two definitions. This hypothetical situation has one RI episode composed of three RI events (green oval) and a one RW episode composed of a single RW event (blue circle). The episode approach does not penalize forecasts that capture the occurrence of the event but not the duration, whereas the individual approach considers both the number and duration of the forecasted events.

For evaluating the skill of predicting the onset of RI and RW events, the onset of an event is defined as the hour in which a single isolated event, or the first in a sequence of events, occurs. The timing performance was evaluated by preparing cumulative frequency plots for a 48-h window centered on the observed onset. For these cumulative frequency plots, the observed events appear as a step function at time zero corresponding to the total number of observed events. The count of onset occurrences predicted by the full set of runs for each model is incremented in the appropriate time period relative to the observed onset. Forecasted events for which the onset is earlier than observed appear as counts for negative lead times and forecasted events for which the onset is later than observed appear as counts for positive lead times. For uniformly perfect model timing, the forecasted events would also appear as a step function at time zero. This manner of presentation provides information on both timing errors and the number of missed events, but ignores false alarms.

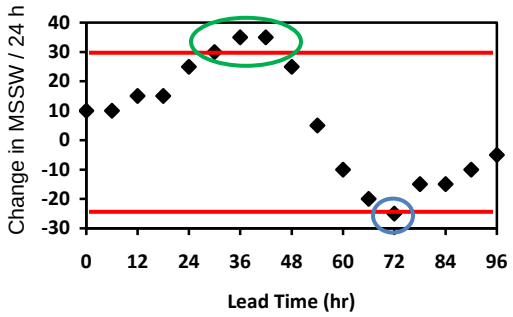

Figure 1: Hypothetical forecast of change in MSSW over 24 hours as a function of lead time to illustrate the difference between events and episodes. Red lines demarcate criteria for RI and RW. The green oval marks three consecutive RI events, which would be defined as a single RI episode. The blue circle marks an RW event that would also be defined as an RW episode.

Proportion Correct (PC), Probability of Detection (POD), Critical Success Index (CSI), and False Alarm Rate (FAR) were computed for each model configuration using event-based contingency tables for exact matches between forecast and observed RI and RW events (Wilks, 1995), as well as matched pairs obtained by considering successively longer time relaxation

windows. The search for matched pairs with several time relaxation windows was performed such that the total number of forecasted events is maintained (i.e., each event is only used once when searching for a match for a given time window). The shifted forecast sequence obtained through this methodology will only improve or leave the forecast skill unchanged.

The number of RW events ended up being too small to justify considering the onset of RW events and the RW event-based scores. Hence, the discussion of these metrics will focus on RI events.

B. Consistency

For this study, consistency was defined as the variability of the forecasted storm center among runs of a given model and resolution initialized at various times and valid at the same time. In other words, consistency refers to the differences in the forecast at multiple lead times for the same valid time. Higher consistency, or smaller variability from one initialization time to the next, is a desirable property for a set of forecasts. Consistency results for the low- and higher-resolutions of each model were inter-compared to determine if higher resolution led to higher consistency (lower variability). Location was the only variable considered.

The consistency assessment requires forecast cases with high-temporal frequency. While it is possible to apply this methodology to cases that are 24-h apart, the results are less relevant, since it is the short-term consistency (over 24-h) that has highest practical operational applicability. A single storm was chosen for the consistency assessment; Hurricane Felix was chosen because both NOAA operational hurricane models (GFDL and Hurricane WRF – HWRF) displayed dramatic run-to-run variability (low consistency). For Felix, four of the six models (AOML, MMM, NRL, and URI) participating in the HRH Test submitted a series of runs that were initialized at 6-hour intervals for a 30-hour period.

Figure 2 illustrates how consistency for a given valid time was assessed by creating a series of ten differences between storm position forecasts at various lead times. To further illustrate the approach, the storm positions used to compute these differences are indicated by a black rectangle in Fig. 3. Since the forecast lead times increase with valid time, an increase in the value of the differences towards later valid times is expected. Plots of differences as a function of valid time were created contrasting the results for the higher- and low-resolution configurations.

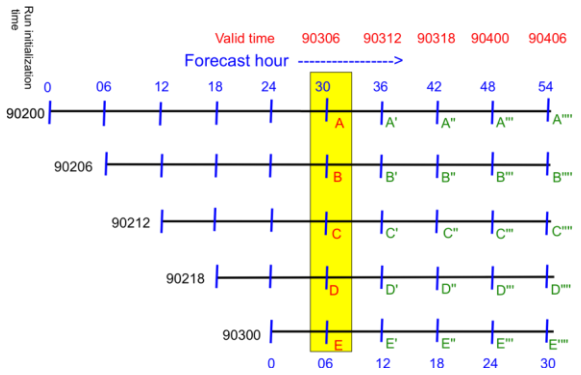

Figure 2. Black lines represent Hurricane Felix forecasts for a given model resolution. Model initialization times are in black to the left of the black lines, valid times are in red at the top, and lead times for the first/last run are depicted in blue above/below the corresponding black line, using the convention mddhh – month, day, and hh UTC. For each valid time, 10 differences in storm location are computed. For instance, for the forecasts valid at 90306 (highlighted in yellow), the following differences can be computed between the runs initialized at 90200 (A), 90206 (B), 90212 (C), 90218 (D), and 90300 (E): A-B, B-C, C-D, D-E, A-C, A-D, A-E, B-D, B-E, C-D, C-E, D-E.

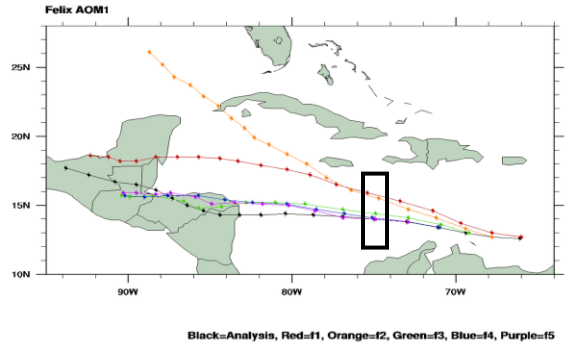

Figure 3. Plot of BT and AOML low-resolution track forecasts for Felix. Black rectangle indicates the storm positions that would be used to compute the differences illustrated in Fig. 2.

No comparison between the forecast and actual track of the storm was performed in the consistency evaluation because other metrics considered for this test addressed forecast accuracy. This approach leaves open the possibility that a model could have very consistent forecasts while producing larger errors.

3. RESULTS

A. RI/RW frequency of occurrence

Table 2 shows the frequency of occurrence for RI and RW compiled using both the event and episode methodologies for all six models. Going to higher resolution improved the agreement between the frequencies of observed and forecasted RI events for all models except URI, for which resolution had negligible impact. The episode methodology reveals that the improvement in RI frequency of occurrence for AOML, MMM, and UWM is partially due to the higher-resolution configuration producing more RI episodes than observed. The forecasted episodes for NRL remained lower than observed for both resolutions, with the higher-resolution configuration capturing more episodes than the low resolution, whereas the occurrence of RI episodes for the high-resolution PSU configuration matched the observed.

Table 2.Total RI and RW counts for events and episodes found in BT and the forecasts.

	BT	High	Medium	Low
AOML				
RI Events	79		70	30
RI Episodes	27		32	15
RW Events	26		7	2
RW Episodes	18		6	2
MMM				
RI Events	77	48		24
RI Episodes	26	27		13
RW Events	24	9		8
RW Episodes	17	5		5
NRL				
RI Events	55		17	5
RI Episodes	20		13	4
RW Events	20		11	6
RW Episodes	14		6	5
PSU				
RI Events	16	10	6	2
RI Episodes	5	5	3	2
RW Events	10	0	0	0
RW Episodes	4	0	0	0
URI				
RI Events	94		33	30
RI Episodes	30		22	22
RW Events	27		2	1
RW Episodes	19		2	1
UWM				
RI Events	46		32	12
RI Episodes	13		18	11
RW Events	18		6	1
RW Episodes	12		5	1

Frequency of occurrence for the RW events showed varying degrees of improvement by going to higher resolution for AOML, NRL, and UWM, whereas the impact of resolution on RW events was negligible for MMM, PSU and URI, with all three PSU configurations failing to produce any RW events. The episode methodology reveals that AOML, URI, and UWM configurations basically produced isolated RW events for both resolutions, whereas both MMM configurations and the higher-resolution NRL configuration produced more consecutive RW events. All model configurations under-predicted the frequency of occurrence for both RW events and episodes.

B. RI onset timing

The cumulative frequency plots for onset timing of RI episodes revealed that the higher-resolution forecasts captured more of the observed RI episodes for AOML, MMM, NRL, PSU, and UWM, whereas the low-resolution configuration for URI captured more of the observed RI episodes (see Fig. 4). These plots also revealed a variety of timing behaviors. Timing errors tended to be very similar for both resolutions of AOML, NRL, URI; whereas forecasts produced by the higher-resolution MMM and UWM configurations showed a stronger tendency to lead the observed onset than the low-resolution configuration. All three PSU configurations have similar timing error distributions for which all onsets tend to be either on time or late.

C. RI event-based scores

The RI event-based scores for all model configurations generally improve as the relaxation window for matching is expanded (sample score distributions for MMM are shown in Fig. 5). POD and CSI generally favor the higher-resolution configuration of all models regardless of the relaxation window. PC is either insensitive to horizontal resolution (MMM, NRL, URI), favors the higher-resolution configuration (PSU) or transitions from favoring the low-resolution configuration for small relaxation windows to being indistinguishable or favoring the higher-resolution configuration for longer relaxation windows (AOML, UWM). Conversely, FAR favors the low-resolution configuration for all relaxation windows (NRL, UWM), transitions from either being indistinguishable or favoring the higher-resolution configuration for small relaxation windows to favoring the low-resolution configuration for longer relaxation windows (MMM, PSU, URI) or does not exhibit any consistent trend (AOML).

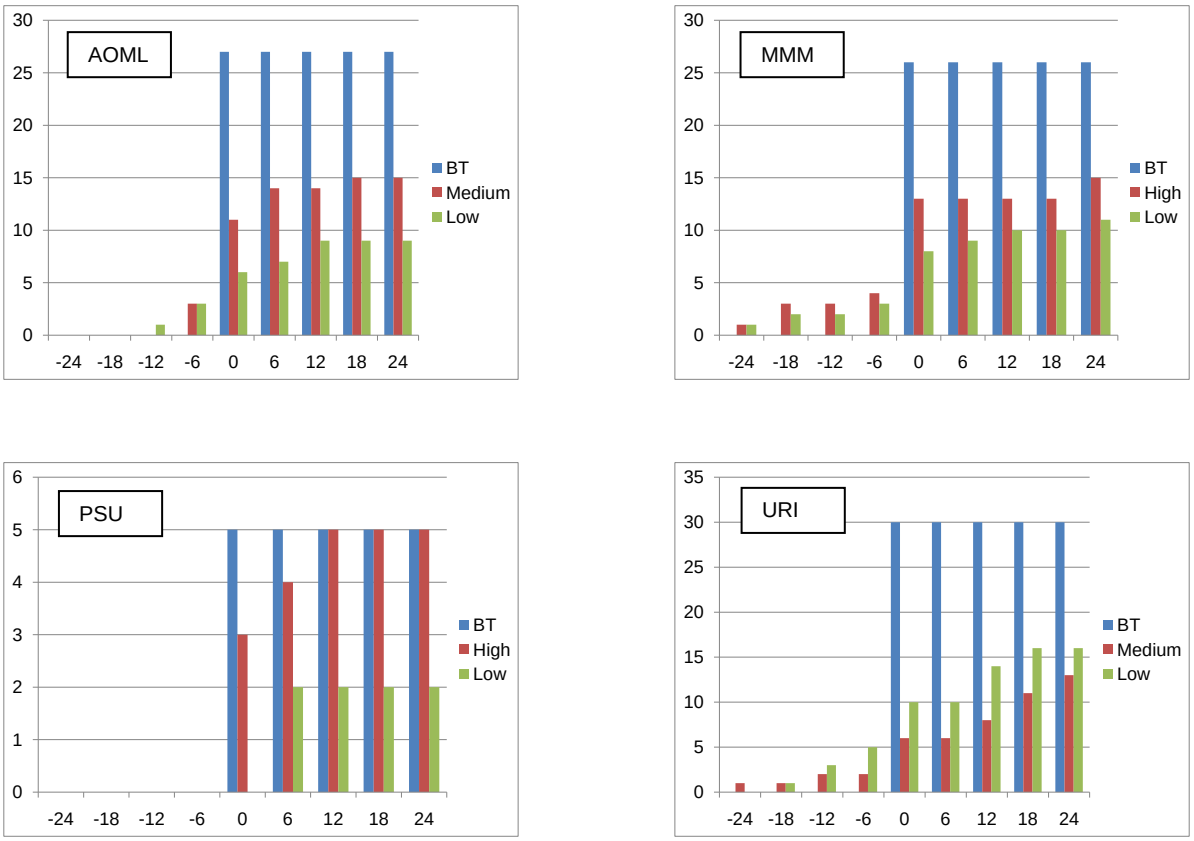

Figure 4. Cumulative counts of RI episodes for AOML, MMM, PSU, and URI composited relative to the observed onset times. Blue indicates observed, red higher-resolution, and green low-resolution. Perfect forecasts would be equivalent to the blue bars.

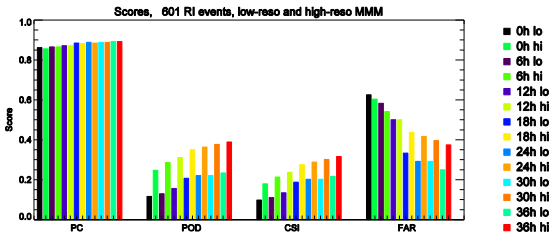

Figure 5. RI event-based scores for MMM. Bars for each score progress from exact match on the left to increasing time relaxation on the right. A pair of bars is shown for each time relaxation window corresponding to low- and high-resolution configurations.

D. Consistency

A comparison of the consistency measures for the two resolutions revealed two basic behaviors: no distinguishable change in the consistency behavior for three of the models (MMM, NRL, URI) and a higher degree of consistency for the higher-resolution forecasts from one of the models (AOML). The two basic types of difference distributions corresponding to these two behaviors are shown in Fig. 6. The differences for the high- and low-resolution MMM configurations do not show any distinct separation, whereas a portion of the differences for the low-resolution AOML configuration exhibit a distinct separation from the distribution for the higher-resolution AOML configuration. Closer scrutiny of the Felix forecasts submitted by AOML revealed the lower consistency (larger differences) for their low-resolution configuration stemmed from a single, rather errant, forecast (see orange forecast track in Fig. 3).

4. CONCLUSIONS

This paper presented the RI/RW error metrics of the HRH Test objective verification, as well as a consistency assessment. The tools developed

to explore the properties of the forecasted RI and RW events provided interesting insights into the impact of resolution on forecasts of rapid intensity change. A number of the RI/RW metrics presented in this paper suggest that running the models at higher-resolution will improve their ability to produce this type of event. On the other hand, the metrics also suggest the higher-resolution configurations may be producing more false alarms for these events. While these results are intriguing, the number of RI and RW events included in the sample ended up being inadequate to make any clear assessments with respect to impact of resolution on RI and RW forecasts. In addition, the RI/RW tools developed for this test only provide limited insight into the short-comings of the forecasts. Given the threshold nature of this metric, it would be useful to have tools that investigate the correlation between the temporal evolution of the observed and forecasted intensity changes in a context that would provide information on whether the forecast totally missed the intensity trend or simply falls slightly short of the threshold criteria, or produces multiple episodes during a single observed episode due to small changes in the rate of intensity change when near the threshold. More sophisticated matching and time relaxation methodologies for looking at timing errors might also provide useful information about this type of event.

The consistency methodology presented in this paper produced a useful indication of the differences in run-to-run variability stemming from changes in horizontal resolution or, for the most part, the lack of any major differences. Given the small sample size for this assessment, the consistency results in this paper should be seen more as a demonstration of a tool than actual robust results.

The results presented in this paper need to be analyzed in concert with the results from the other HRH Test tools (Bernardet et al, 2010; Developmental Testbed Center, 2009). When taking all these metrics into consideration, this test did not show that the use of higher resolution leads to an overall benefit in tropical cyclone forecasting. 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 high enough to resolve small-scale structures such as updrafts and meso-vortices that may need to be represented in order to improve intensity forecasting.

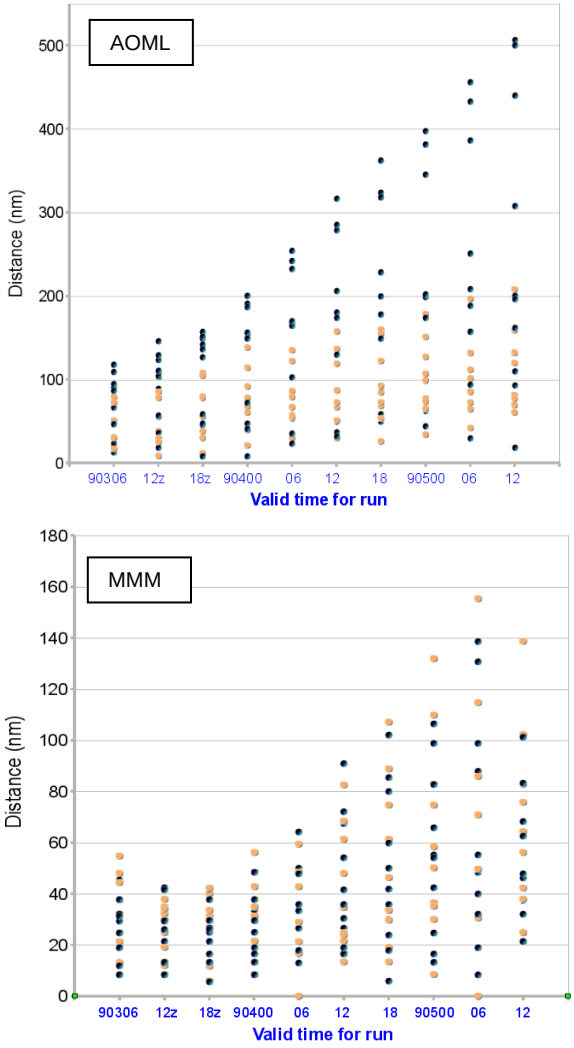

Figure 6. Distance (nm) between forecasts of Felix storm location initialized at the times shown in Fig. 2 and valid at the times listed on the x-axis. Black (yellow) represent the low- (higher-) resolution forecasts.

5. REFERENCES

Bernardet, L., L. Nance, S. Bao, B. Brown, L. Carson, T. Fowler, J. Halley Gotway, C. Harrop, and J. Wolff, 2010: The HFIP High-Resolution Hurricane Forecast Test: Overview and results of track and intensity forecast verification. *29th Conference on Hurricanes and Tropical Meteorology*,

Tucson, AZ. American Meteorology Society.
Braun, S. A., M. T. Montgomery, and X. Pu, 2006: High-resolution simulation of Hurricane Bonnie (1998). Part I: The organization of eyewall vertical motion. *J. Atmos. Sci.*, **63**, 19–42.
Chen, S. S., J. F. Price, W. Zhao, M. A. Donelan, and E. J. Walsh, 2007: The CBLAST-Hurricane Program and the next-generation fully coupled atmosphere-wave-ocean models for hurricane research and prediction. *Bull. Amer. Meteor. Soc*, **88**, 311-317.
Davis, C., W. Wang, S. Chen, Y. Chen, K. Corbosiero, M. DeMaria, J. Dudhia, G. Holland, J. Klemp, J. Michalakes, H. Reeves, R. Rotunno, and Q. Xiao, 2008: Prediction of landfalling hurricanes with the advanced hurricane WRF model. *Mon Wea. Rev*., **136**, 1990-2005.
Developmental Testbed Center, 2009; High Resolution Hurricane Test Final Report. http://www.dtcenter.org/plots/hrh_test/HRH_Report_30Sept.pdf.
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.
Hendricks, E. A., M. T. Montgomery, and C. A. Davis, 2004: On the role of vortical hot towers in hurricane formation. *J. Atmos. Sci.*, **61**, 1209–1232.
Marchok, T. P., 2002: How the NCEP tropical cyclone tracker works. Preprints, *25th Conf. Hurr. Trop. Meteor.*, 29 April – 3 May 2002, San Diego, CA, 21-22.
Powers, J. G., and C. A. Davis, 2002: A cloud-resolving, regional simulation of tropical cyclone formation. *Atmos. Sci. Lett*., doi.10.1006/asle.2002.0054
Rotunno, R., T. Chen, W. Wang, C. Davis, J. Dudhia and G. J. Holland, 2009: Large-eddy Simulation of an Idealized tropical Cyclone. *Bull. Amer. Meteor. Soc.*, **90**, 1783-1788.
Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.
Yau, M. K., Y. Liu, D.-L. Zhang and Y. Chen. 2004: A multiscale numerical study of Hurricane Andrew (1992). Part VI: Small-scale inner-core structures and wind streaks. *Mon. Wea. Rev.*, **132**, 1410–1433.

Acknowledgements: The authors would like to thank the participating modeling groups, as well as the following scientists for their support during this test: Timothy Marchok (GFDL), James Franklyn and Jack Beven (NHC), Mark DeMaria (NESDIS) and Mike Fiorino (ESRL). The DTC is funded by the National Oceanic and Atmospheric Administration, the Air Force Weather Agency, and National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation. This project was also supported by the NOAA Hurricane Forecast Improvement Project.