Evaluation of RI/RW for HWRF and other HFIP models

Mrinal K Biswas, John H. Gotway, and Tressa L. Fowler

Developmental Testbed Center, NCAR, Boulder

POC: Mrinal K Biswas biswas@ucar.edu

Date: September 30, 2015

Introduction

The National Hurricane Center (NHC) forecast verification shows large improvement in the official tropical cyclone (TC) track forecasts over the last decade. However, intensity forecasts remain a challenge. While some improvements in TC intensity forecasting has been demonstrated (DeMaria et al. 2014), predicting large changes in TC intensity remains very problematic. This is particularly true for episodes of rapid intensification (RI), which is NHC's highest operational forecasting priority (Rappaport et al. 2009). The difficulty of predicting RI is due in large part to the multi-scale nature of the problem with environmental, oceanic, and inner-core processes all likely playing important roles in determining if and when a TC will undergo RI. Rapid weakening (RW) is also a difficult and important forecasting challenge. The deficiencies of models to forecast these large changes leads to significant intensity errors and contributes to the problem of intensity forecasting. This study establishes benchmarks for model performance for large changes in intensity, addressing both RI and RW. Additional analyses are included to document the characteristics of the forecasting errors. The primary focus is on the HWRF model, though other forecasts from operational models are compared and contrasted.

Data

The primary focus of this evaluation is the HWRF 2014 pre-implementation retrospective runs (H214) conducted by Environmental Modeling Center (EMC) (Tallapragada et. al. 2014). The forecasts were submitted to the National Center for Atmospheric Research's Tropical Cyclone Modeling Team (TCMT) along with the 2014 operational runs (hereby referred to as HWRF-14). For purpose of comparison, the NHC official forecasts (OFCL) and forecasts from the Coupled Ocean Atmosphere Mesoscale Prediction System for Tropical Cyclones (COAMPS-TC) model are evaluated. The 2014 COAMPS-TC retrospective evaluations (covering storms from 2011-2013) (CTCX) and real-time forecasts of 2014 storms are both included (hereby referred to as CTCX-14). Model forecast intervals are available more often and have more lead times available than the OFCL forecasts (available 12-h). The models are also run for "invest" storms, however, official forecasts are not issued for these storms. Thus, the individual issue and lead times do not always match between forecasts, and the OFCL forecasts are unavailable for several lead times. Table 1 describes the datasets included in HWRF-14, CTCX-14 and

OFCL, which were used to verify RI/RW events in the Northern Atlantic (AL) and eastern North Pacific (EP) basins.

Table 1: Datasets included for HWRF-14, CTCX-14, and OFCL

Years	Model Versions		
	HWRF-14	CTCX-14	OFCL
2011	H214	CTCX	Operational
2012	H214	CTCX	Operational
2013	H214	CTCX	Operational
2014	Operational	Operational	Operational

The HWRF model is upgraded every year to include new innovations. Therefore, the 2015 pre-implementation retrospective runs (H215) were compared with the HWRF-14 dataset to see if the upgrades improve RI prediction. Both datasets cover AL and EP storms from 2011-2014.

For the Western Pacific (WP) basin, only two years (2013 and 2014) of HWRF forecasts cases were available from realtime runs for evaluation.

The categories of tropical cyclones that are included in the sample are tropical depressions, tropical storms, and hurricanes.

Methodology

Kaplan and DeMaria (2003) defined RI as the 95th percentile of all 24-h intensity change episodes over water of tropical systems in the AL basin. This translates to an increase of 30 kt (15.4 m/s) in a 24-h period for the storms in their sample, covering years 1989-2000. The same criterion is applied to define RW as a 30 kt decrease in intensity in a 24-h period.

The observation errors are $\pm$5 kt, therefore, a flexible criteria for identifying RI are used to verify the RI forecasts including both a 20 kt and 30 kt change of VMAX (i.e. intensity defined as maximum 10-m sustained wind) in a 24-h period. Although these values correspond to approximately the 90th and 97th percentiles in the AL as reported by DeMaria and Kaplan (2010) based on storms occurring in 1989-2006 and may not be applicable in the EP and WP basins due to different TC climatology in this region, these criteria will be considered sufficient hereinafter to characterize the rapid development of TCs in this basin due to the lack of the climatology for 24-h intensity change in the EP and WP basins. Following Tallapragada and Kieu (2014), the verification of the RI (RW) forecasts for each RI (RW) threshold is quantified in terms of a binary event (yes or no) based upon which contingency table statistics can be computed (Wilks, 2011).

Additionally, to diagnose possible sources of model error, conditional error evaluations were conducted for each combination of forecast and observed event. These evaluations answer the following questions:

- When an RI event is both forecasted and observed, is the forecast increase biased?
- When a model forecast misses an RI (RW) event, does it do so by large or small amounts?
- Are correctly forecast null events biased?
- How large are the overforecasts when the model issues a false alarm?

Further, the RI criteria are relaxed in time and intensity to determine if small errors in timing or amount are leading to large errors in the event statistics. These types of analyses aim to answer:

- Does the model correctly forecast more events within a 30-h window? In other words, do some cases have timing errors of less than 6-h?
- Does the model forecast an increased intensity, but not large enough to be considered an event? In other words, does the model identify potential events but suffer from a low bias in the intensity change?

These diagnostic evaluations can help focus future efforts to improve model forecasts. The evaluations were performed using the Model Evaluation Tools – Tropical Cyclone (MET-TC). The new enhancements and capabilities will be available in the MET v5.1 release.

Results

RI Verification in AL and EP basins (2014 model versions)

HWRF RI/RW forecasts are the primary focus in this study with OFCL and COAMPS-TC forecasts included for comparison. However, many fewer cases are available for these comparison forecasts, so the full set of cases is used for all forecasts, without regard to matching exactly the time and event of each case for each model.

Table 2 shows the counts of HWRF forecast and observed RI events and non-events for both AL and EP basins for all lead times combined. Overall, RI events are relatively rare, making up less than 5% of cases. Forecasts of RI from the HWRF are even more rare, and comprise less than 1% of forecasts. Thus, the HWRF RI forecasts have a considerable low frequency bias. The probability of detection (POD) and false alarm rate (FAR) are 7.3% and 66.4%, respectively for HWRF. While the hits are low and the false alarms are high, this is quite common among forecasts of rare events.

Tables 3 and 4 show the contingency table counts for RI events and non-events as well as event summary statistics for CTCX-14 and OFCL, respectively. The CTCX-14 (OFCL) RI POD is 4.7% (7.4%) and the FAR is 86.9% (18.4%). The lower FAR of OFCL forecasts can be attributed to a conservative approach towards forecasting RI events.

Table 2: Contingency table (a) and event summary (b) for HWRF-14 with observed and forecast events of an intensity increase of 30 kt or more in 24-h for all lead times combined.

(a)

		Observation		
		RI	***No RI***	***Total***
Model Forecast	***RI***	128 (0.3%)	253 (0.6%)	381 (0.9%)
	No RI	1623 (4.1%)	37654 (94.9%)	39277 (99%)
	Total	1751 (4.4%)	37907 (95.6%)	39658 (100%)

(b)

POD	7.3%
PODN	99.3%
FAR	66.4%
RI Event Rate	4.4%

Table 3: Same as Table 1, except for CTCX-14.

(a)

		Observation		
		RI	***No RI***	***Total***
Model Forecast	***RI***	26 (0.15%)	173 (0.9%)	199 (1.1%)
	No RI	524 (3.0%)	16608 (95.8%)	17132 (98.8%)
	Total	550 (3.1%)	16781 (96.8%)	17331 (100%)

(b)

POD	4.7%
PODN	98.9%
FAR	86.9%
RI Event Rate	3.1%

Table 4: Same as Table 1, except for OFCL.
(a)

		Observation		
		RI	***No RI***	***Total***
Model Forecast	***RI***	31 (0.3%)	7 (0.06%)	38 (0.03%)
	No RI	387 (3.8%)	9683 (95.7%)	10070 (99.6%)
	Total	418 (4.1%)	9690 (95.8%)	10108 (100%)

(b)

POD	7.8%
PODN	99.9%
FAR	18.4%
RI Event Rate	4.1%

When the cases are limited to the 24-h lead-time, the HWRF POD is 10.8% while the FAR is 68.6%, indicating that RI forecasting is more skillful earlier in the forecast period comparatively. For the OFCL forecasts, the POD is 13% and the FAR is 17%.

Relaxation of the intensity threshold and the timing results in somewhat better forecast performance. The POD increases to 38.6% with a smaller increase in FAR (78.0%) for all lead times combined when the model RI criterion is related to 20 kt in 30-h. This suggests that the models often forecast an intensity increase of between 20 and 30 kt within 6-h of the occurrence of an intensity increase of greater than 30 kt. Therefore, the timing and amount are somewhat in error, though the model gives an indication of the RI event. For many users, this indication may be a sufficiently useful forecast even though it underestimates the amount and has small timing errors. Further, model improvements may most easily be made on these types of cases with relatively small, specific errors. Improvements to cases where the forecast gives no indication of the RI event may prove more difficult.

In order to better understand model performance, additional evaluation beyond examination of POD and FAR are required. While categorical statistics can provide a nice summary of events, the use of thresholds can mask some errors while exaggerating others. For example, if an intensity forecast over a 24-h period is for an increase in wind

speeds of 80 kt, but the observed increase is only 30 kt, this RI forecast will be a "hit" although the error is quite large. However, for the same event, a forecast change of 25 kt would be a "miss" event though it is a reasonable forecast. In order to determine whether a model is able to properly capture the RI magnitude, the distributions of the difference of change in a-deck values (model forecasted maximum 10-m intensity) in the last 24-h and change in b-deck values (best track) in the last 24-h (hereby referred to as ABDEL) are examined. These values are computed separately for the 4 quadrants of the contingency table.

Figure 1 shows boxplots of the HWRF distribution of ABDEL for hits, misses, false alarms and correct rejections. The green boxes depicting the hits (FY_OY) are mostly below the zero line indicating that even when HWRF detects a RI, it underestimates the magnitude of the intensity change. As expected, the number of hit cases decreases rapidly with lead time. HWRF typically underestimates RI magnitude by around 8 to 10 kt, as measured by the median error. The red boxplots show the errors for the false alarms (FY_ON), whose median values indicate that HWRF tends to over-predict by 15-20 kt. Of course, by definition, a false alarm must be an overestimate; so all values will be above 0. Also, there are a large number of "misses" (FN_OY), with the errors shown in orange. By definition, a missed RI must be an underestimate so all values are below 0. HWRF typically underestimates intensities by around 25 kt, with some very small errors but also with some of more than 70 kt indicated in outliers.

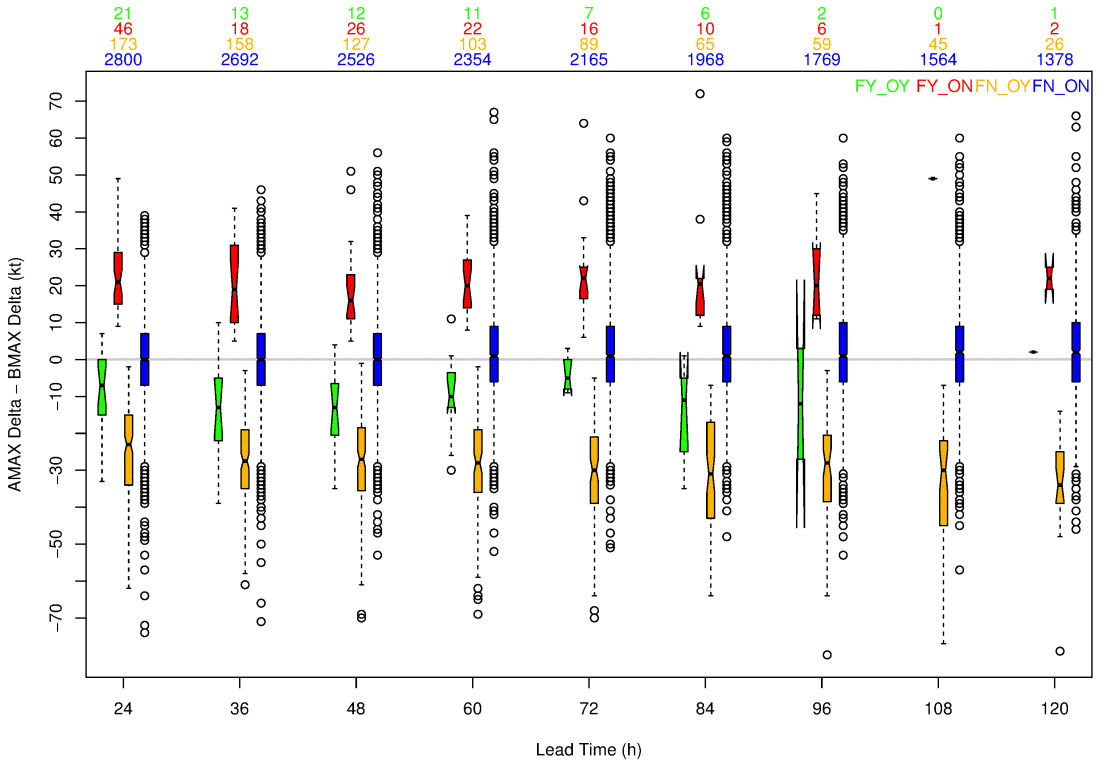

Figure 1. Boxplots showing distribution of difference of change in a-deck intensity values in the last 24-h and change in b-deck intensity values in the last 24-h for hits (green), false alarms (red), misses (orange) and correct nulls (blue) for HWRF by

lead time. The storms included are from AL and EP basins. The numbers of cases are indicated at the top of the figure by lead time.

Similar boxplots for the OFCL and CTCX forecasts are shown in Figures 2 and 3, respectively. After the 48-h lead time, there are no OFCL forecasts of RI events. Further, there are very few false alarms for the OFCL forecasts, so no analysis of this type of error was performed.

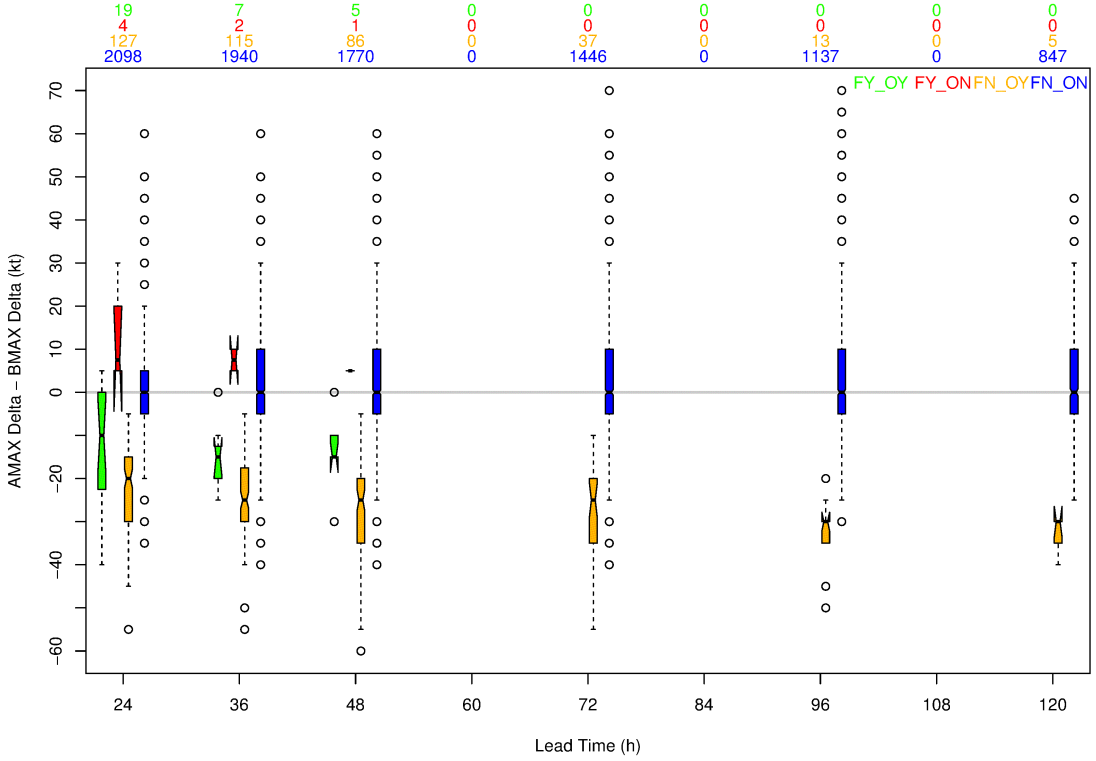

Figure 2. Same as Figure 1 except for OFCL.

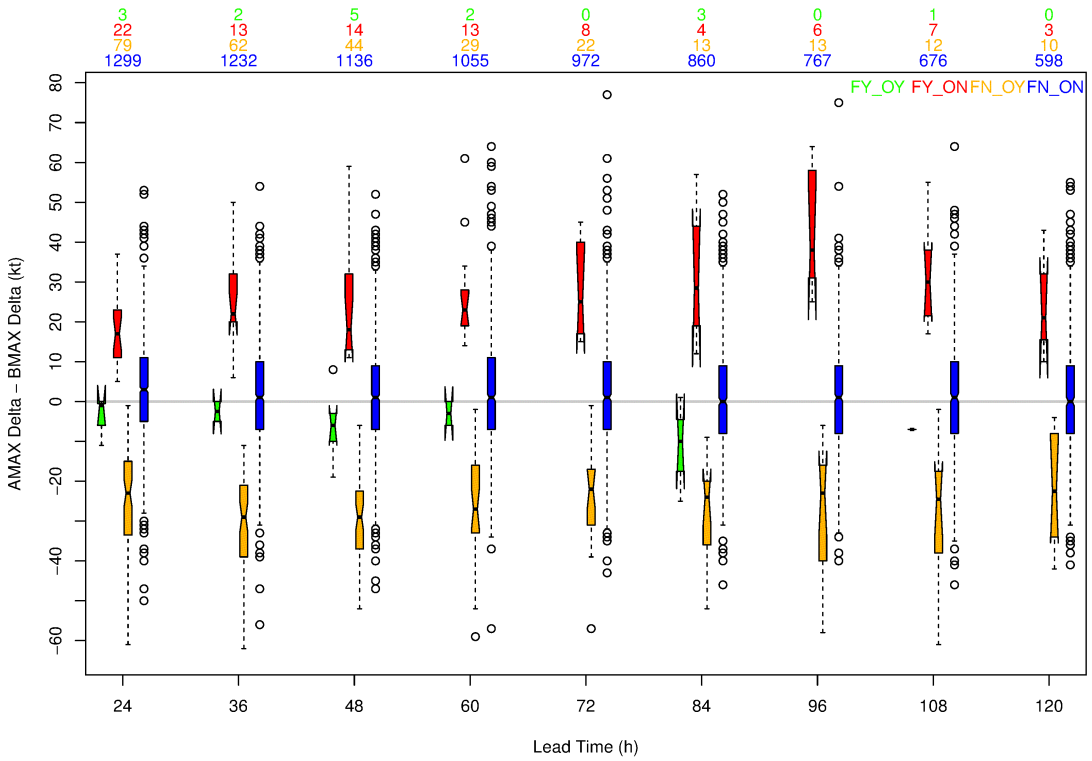

Figure 3. Same as Figure 1 except for CTCX.

RI performance by initial intensity

HWRF's RI performance for cases restricted to tropical storm strength or more (>=34 kt) at the initial time increased the POD to 8.8% and decreased the FAR to 50.8%. However, when the initial intensity is restricted to Category 1 or more, the POD drops significantly to 2.8% (detecting only 2 of 71 events) indicating HWRF has difficulty in predicting RI events when the storm is already hurricane strength at the initial time. CTCX has zero skill for the hurricane intensity threshold, while POD for OFCL was 1.3% (detecting 7 of 53 events). Note that the samples for the HWRF, CTCX and OFCL are different.

RI performance by storm location

Kaplan and DeMaria (2003) noted that the maximum frequency of RI occurs between 10° to 15°N. Analysis of RI performance based on the initial location of the storm indicates that the POD is higher (7.9%) when the initial location of the storm is between at 10° and 16°N. However, storms with initial position greater than 20°N has POD of 1.1% and FAR of 95% indicating HWRF has the capability to intensify the storm, but not at the right time and location. CTCX and OFCL RI performance decreases with increasing latitude. Relaxing the intensity threshold to a 25 kt increases the HWRF POD (FAR) to 12% (65%) when the storm's initial position is greater than 15°N, indicating that HWRF has difficulty in intensifying as much as the observed intensity change.

Rapid Intensification Examples

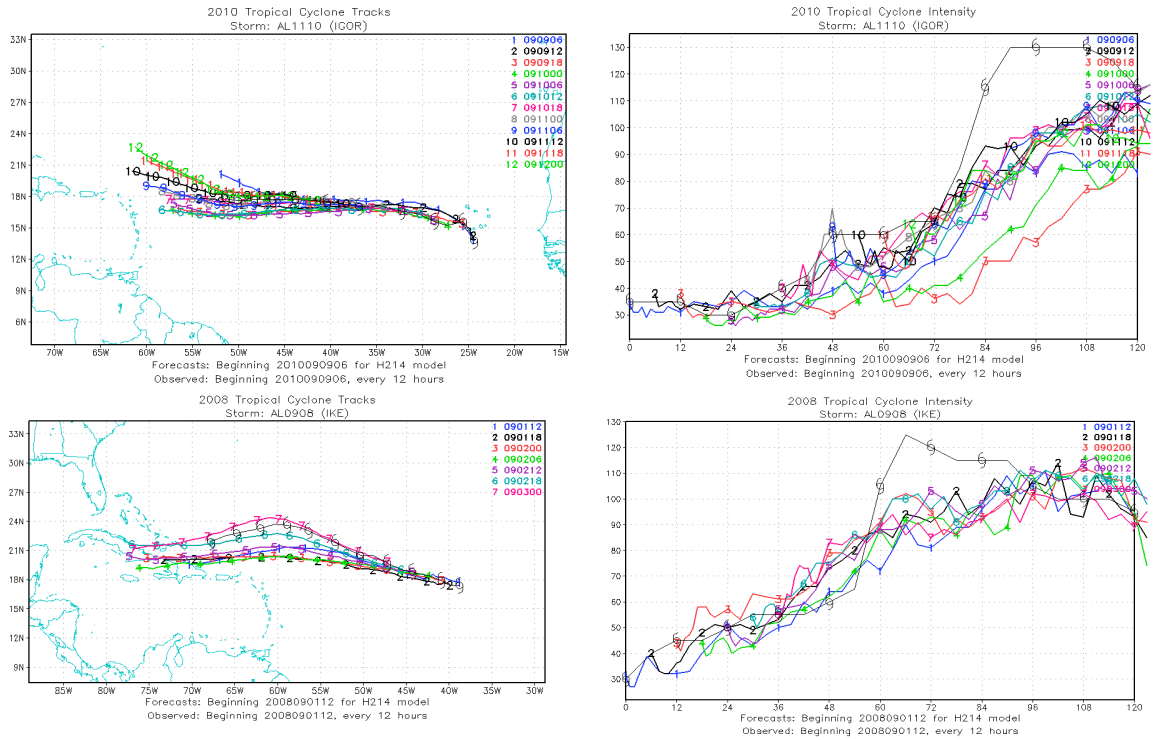

Figure 4. Example track and intensity plots for Igor (top) and Ike (bottom). Best track (black) and HWRF forecasts for various issue times (colors) are shown.

Example track and intensity plots for TCs Igor and Ike are shown in Figure 4. The best track values are in black and the colored lines represent HWRF forecasts from different initialization times. Because RI and RW events are defined over a 24-h period, it is often the case that several events happen in sequence as shown in this figure. For example, if storm intensities were 30, 35, 60 and 80 kt at the 12, 18, 36 and 42-h lead times; mathematically there is one RI event from 12 to 36-h (30 kt increase) and another from 18 to 42-h (45 kt increase). Cases analyzed in this example lack independence when several events occur in sequence from a single storm. A full listing of hits, false alarm and miss forecast cases are given in Appendix A, B and C respectively, which shows many cases exhibit this behavior.

RI Verification in WP basin

The HWRF WP forecast dataset has significantly fewer RI events (only two seasons evaluated) than the samples from the AL and EP basins, though the RI event rate is higher at 7.6%. HWRF achieves a POD (FAR) value of 28.7% (43.0%) for all lead times combined. The correctly forecast RI events (in green) show little bias (Figure 5). However, the false alarms (red) and missed events (orange) are off by around 20 kt in the early lead times. At the lead times greater than 60-h, too few events exist to draw any conclusions.

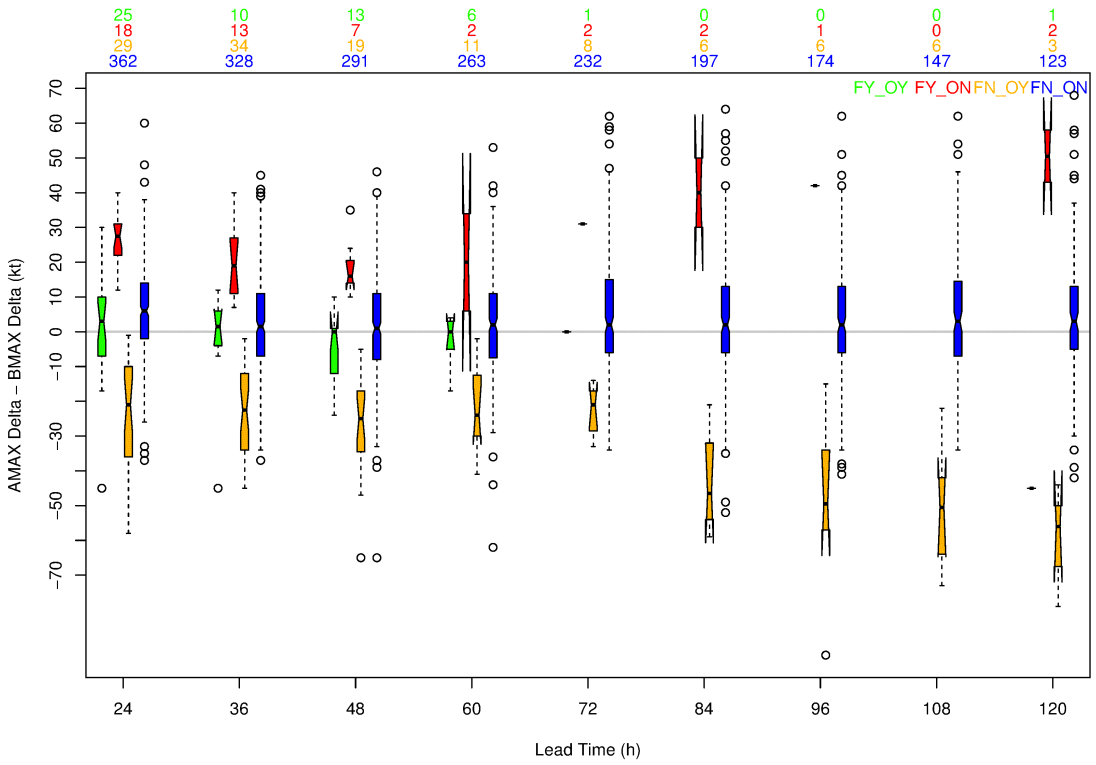

Figure 5. Same as Figure 1, except for Western Pacific basin.

Rapid Weakening

The incidence of RW is similarly rare to that of RI, occurring in about 5% of cases. Once again, HWRF has an under-forecasting frequency bias, forecasting RW events in only about 2% of cases. However, the statistics are somewhat more encouraging for RW than for RI. HWRF has a much higher POD (15.3%) with a slightly lower FAR (58.7%) and a very high PODN (98.8%) (Table 5). Relaxation of the intensity threshold and the timing (observed events of 30 kt decrease in 24-h and forecast events of 20 kt decrease in 30-h) results in somewhat better forecast performance, with POD increasing to 42.4% with a smaller increase in FAR (76.8%) (Table 6). This suggests that the models often forecast an intensity decrease between 20 and 30 kt within 6-h of the occurrence of an intensity decrease of greater than 30 kt. Therefore, the timing and amount are somewhat in error, though the model gives an indication of the event.

Table 5: (a) Contingency table and (b) event summary for HWRF with both observed and forecast events of 30 kt decrease in 24-h.

(a)

		Observation		
		RW	***No RW***	***Total***
Model Forecast	***RW***	324 (0.8%)	461 (1.2%)	785 (2.0%)
	No RW	1793 (4.5%)	37080 (93.5%)	38873 (98.0%)
	Total	2117 (5.3%)	37541 (94.7%)	39658 (100%)

(b)

POD	15.3%
PODN	98.8%
FAR	58.7%
RW Event Rate	5.3%

Table 6: Event summary for HWRF with observed events of 30 kt decrease in 24-h and forecast events of a 20 kt decrease in 30-h.

POD	42.4%
PODN	91.9%
FAR	76.8%

Figures 6 to 8 show the boxplots of ABDEL for hits, misses, false alarms and correctly forecast nulls (non-events) for HWRF-14, OFCL and CTCX-14. The correctly forecast RW events for HWRF lack significant bias. Thus, when HWRF detects RW events, it gets reasonable average intensity change. In contrast, the OFCL forecasts tend to underestimate the reduction in intensity by 5 to 10 kt, even when the RW event has been forecast. When the HWRF forecast misses an event, it does so by around 10 kt at the 24-h lead time with the amount increasing to around 25 kt at later lead times. False alarms are off by around 20 kt with little change at increasing lead times. There are a handful of false alarms that have errors near 70 kt. Note that the RW events are defined as a 30 kt decrease in intensity in a 24-h period.

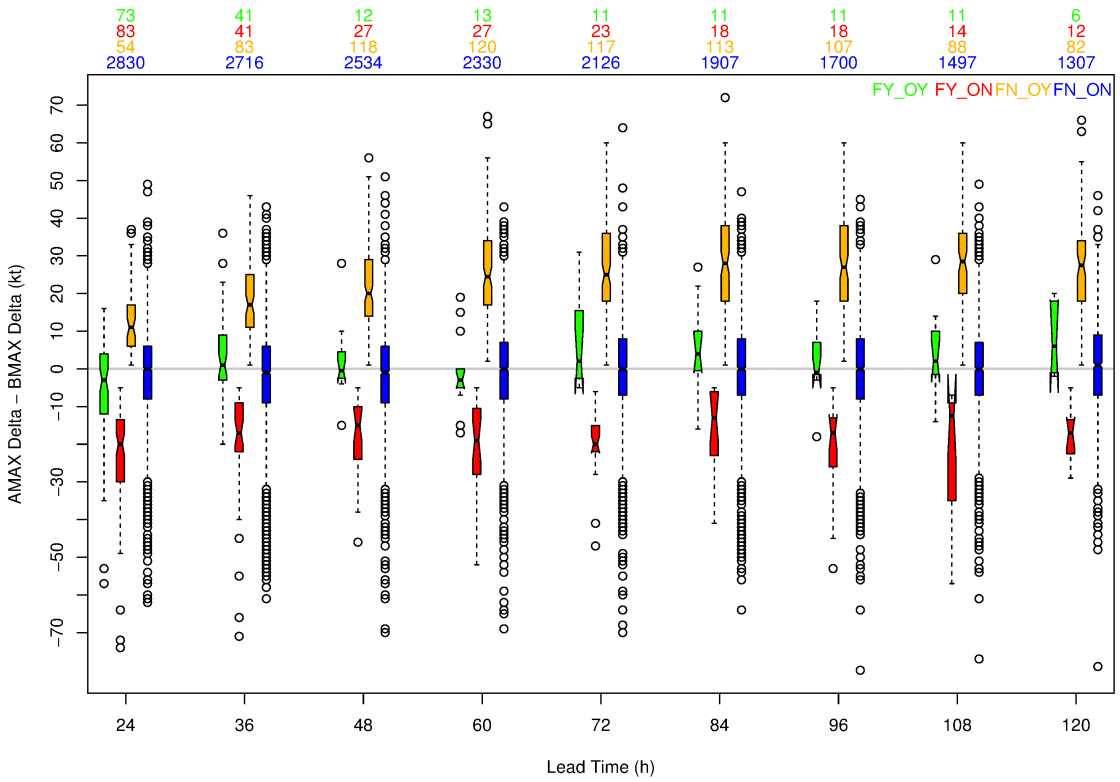

Figure 6. Boxplots showing distribution of difference of change in a-deck intensity values in the last 24-h and change in b-deck intensity values in the last 24-h for hits (green), false alarms (red), misses (orange) and correct nulls (blue) for HWRF by lead time. The storms included are from AL and EP basins. The numbers of cases are indicated at the top of the figure by lead time.

For the OFCL forecasts, even the correctly forecast RW events (green) were underestimated by 5 to 10 kt. The false alarms were only off by about 10 kt, indicating that weakening occurred, but the change was not large enough to count as an RW event. The missed events tended to be off by about 20 kt while correctly forecast non-events are unbiased. Compared to these OFCL forecasts, HWRF had less bias in the hits, larger errors in the false alarms, and similar errors for misses and correct nulls.

For the CTCX forecasts, the correctly forecast RW decreases (green) are very close the actual observed change. The false alarms were only off by about 15 kt, indicating that weakening occurred, but the change was not large enough to count as an RW event. The missed events tended to be off by about 20 to 30 kt, while correctly forecast non-events are unbiased. Compared to the HWRF, the hits and correct nulls are similar, with all being relatively unbiased. The false alarms are typically a bit smaller for CTCX than for HWRF. HWRF has smaller errors in the miss category at the 24-h lead time than CTCX, but the models have similar errors at the other lead times.

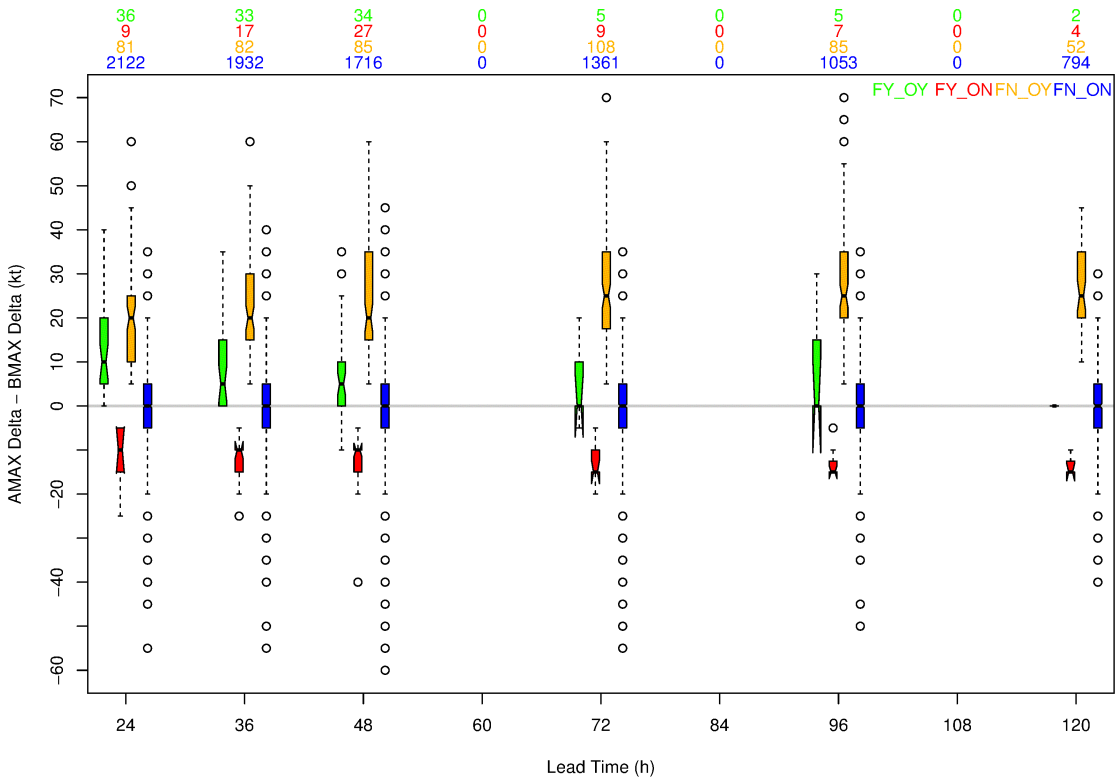

Figure 7. Same as Figure 6 except for OFCL.

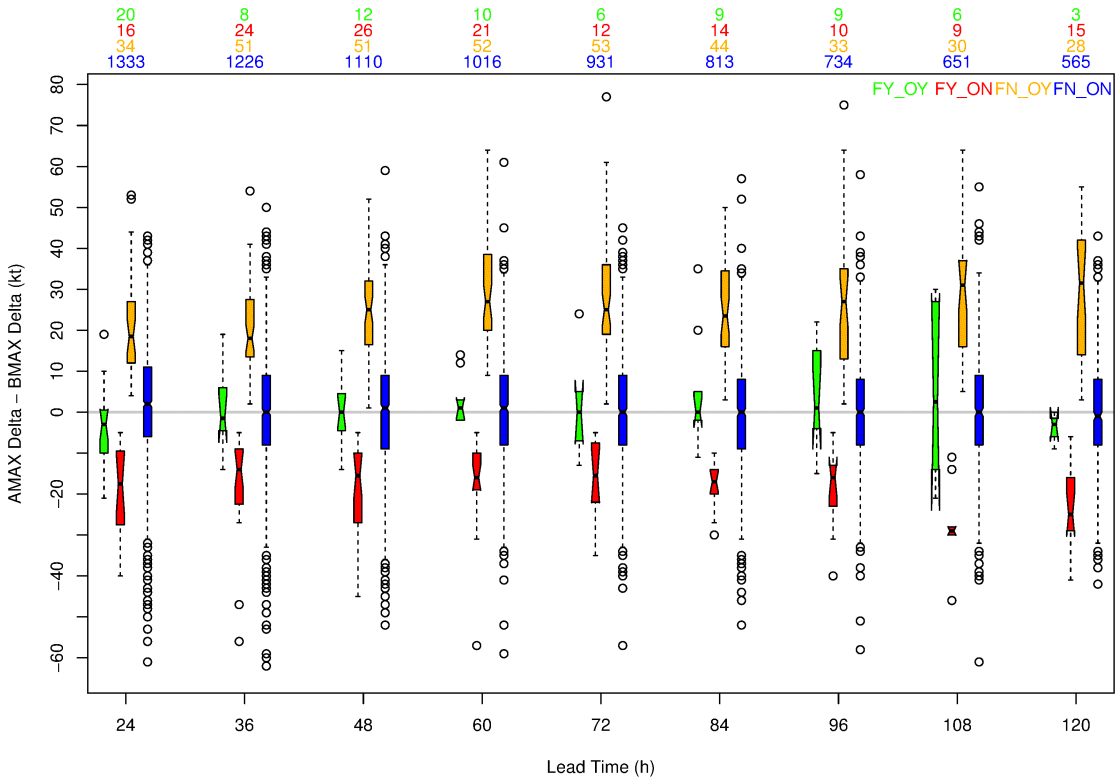

Figure 8. Same as Figure 6 except for CTCX.

The WP sample has fewer RW events than the samples of cases from the Atlantic and Eastern Pacific basins. The event rate is much higher however, with over 13% of cases having an RW event (Table 7). Forecasts for more common events typically perform much better than those for rare events, which is demonstrated with these results. HWRF correctly forecasts 37.7% of events across all lead times, with a false alarm rate of 38.4%. As shown in Figure 9, the correctly forecast events (in green) and non-events (in blue) show little bias. The missed events (orange) are off by about 25 kt at most lead times. False alarms are too few in number at any lead time to draw conclusions. At lead times above 60-h, too few hits exist to draw any conclusions.

Table 7: Contingency table (a) and statistics (b) for HWRF with observed and forecast events of 30 kt decrease in 24-h for all lead times combined in the WP.

(a)

		Observation		
		RI	No RI	Total
Model Forecast	RI	235 (5.1%)	147 (3.2%)	382 (8.2%)
	No RI	388 (8.5%)	3755 (82.9%)	4143 (91.5%)
	Total	623 (13.7%)	3902 (86.2%)	4525 (100%)

(b)

POD	37.7%
PODN	96.2%
FAR	38.4%
RI Event Rate	13.7%

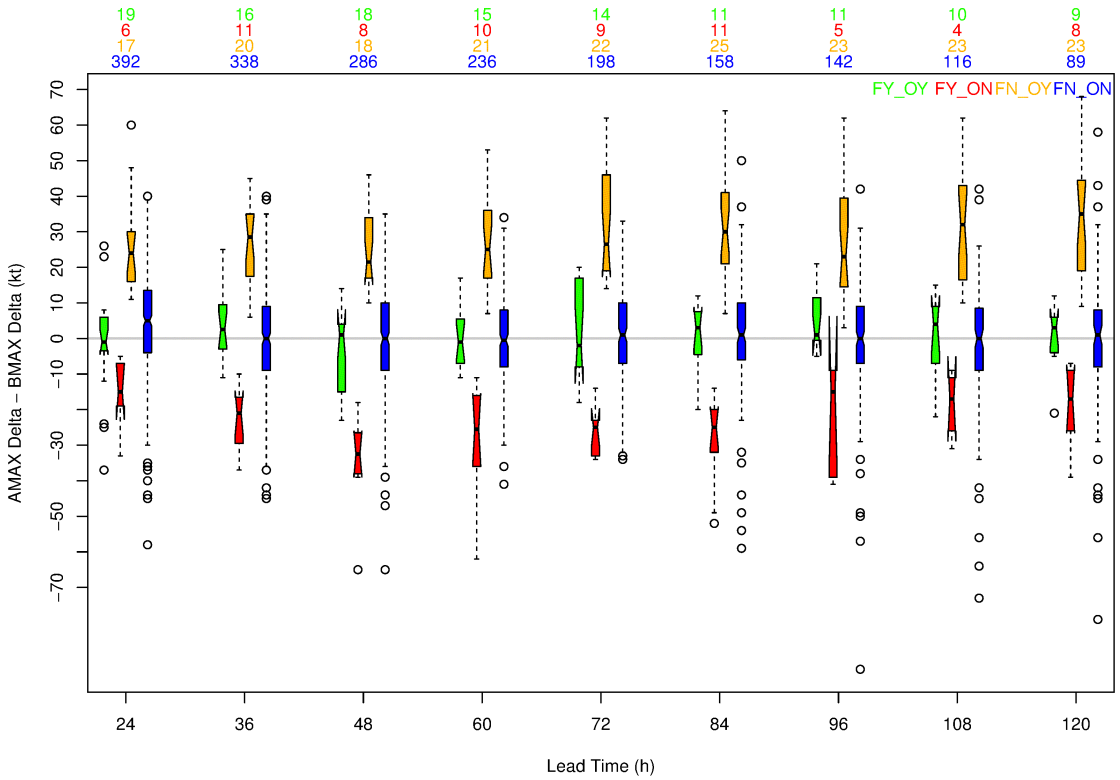

Figure 9. Same as Figure 6, except for WP basin.

Comparing H214 and H215 forecasts

The HWRF 2015 pre-implementation retrospective forecasts (H215) cover storms from 2011-2014. The POD and FAR for RI events are 6.9% and 69.3%, respectively. Relaxation of intensity to 25 kt in 24-h improves the performance with POD increasing to 14.1% and a smaller decrease in FAR (59.7%). For the H215 forecasts, the medians of the hits (green boxes) are within the observation error (±5 kt) (Figure 10). The False Alarms are similar to HWRF-14 forecasts showing that H215 also tends to over-forecast intensities by 15-20 kt.

When a homogeneous comparison of HWRF-14 and H215 forecasts is performed, HWRF-14 forecast hits (green) are closer to the zero line indicating that the ABDEL is within the observation error. The POD (FAR) for HWRF-14 forecasts are 4% (77%) compared to 7% (70%) for H215. Relaxing the intensity change threshold to 25 kt in 24-h increases the POD of HWRF-14 (H215) to 10% (14%).

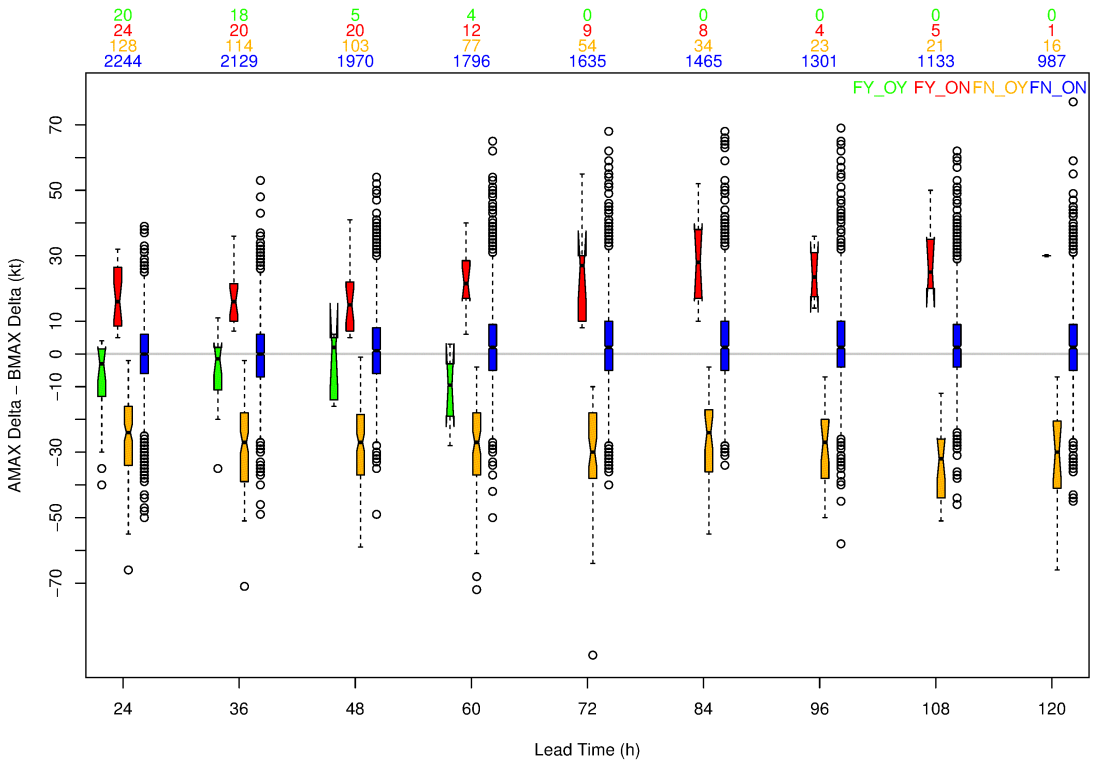

Figure 10. Same as Figure 1, except for H215.

Conclusions

Rapid intensification and weakening are important but rare events. In this study a baseline for future HFIP evaluation was created. However, forecasts are skillful only in a relative sense, and the forecasts evaluated here all demonstrate better than random skill at identifying both rapid intensification and weakening of tropical cyclones. Further, the definitions of these events are quite restrictive. Application of the same criteria to the forecasts makes lesser errors in timing and intensity equivalent to a forecast that gave no hint of such an event. Evaluating forecasts with relaxed criteria for RI and RW results in a substantial improvement in performance, most notably within +/-6 hours and only requiring a forecast change of 20 kt instead of 30. The cases with timing and intensity errors may prove easier to correct in the forecasts than the cases with no hint of a major change.

Due to the rarity of RI/RW events, homogeneous model comparisons were not conducted. When statistics are computed separately for each dynamical model, HWRF had higher POD and lower FAR compared to COAMPS-TC. The OFCL POD is comparable with HWRF. HWRF underestimates the intensity change even when it detects a RI (by definition) event. HWRF had difficulty in predicting RI events when the initial intensity is hurricane strength or greater. HWRF also performs better in RI prediction over the WP basin compared to AL and EP basins. The 2015 upgrades to the HWRF model increased the RI prediction skill compared to the 2014 version.

The RW POD is better than RI for HWRF. HWRF and COAMPS-TC predict the intensity change of RW relatively close to the observed change.

Yang et al. (2011) conclude that occurrences of intensification and weakening are most related to intensity change in past 12 hours. This evaluation considered subsequent 24-h periods without consideration of prior event occurrence. For future work, total periods of change may be considered to eliminate the potential dependence between events.
All the new enhancements to the RI/RW verification tools in MET-TC will be available in METv5.1.

Acknowledgement

The authors thank Ligia R. Bernardet and Kathryn Newman for providing valuable comments, which led to improvements in the manuscript. The authors also thank EMC and NRL for providing the dataset used in the study.

References

Kaplan, J. and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea Forecasting*, **18,** 1093-1108.

Wilks, D.S., 2011: Statistical Methods in the Atmospheric Sciences. *3rd Edition. Academic Press*, 704 pp.

Kaplan, J. M. DeMaria, and J. A. Knaff, 2010: A Revised Tropical Cyclone Rapid Intensification Index for the Atlantic and Eastern North Pacific Basins. *Wea. Forecasting*, **25**, 220–241.doi: http://dx.doi.org/10.1175/2009WAF2222280.1

Rappaport, E. N., et al. (2009), Advances and challenges at the National Hurricane Center, *Weather Forecast*., **24,** 395–419, doi:10.1175/ 2008WAF2222128.1

Tallapragada, V. and HWRF team, 2014: 2014 Upgrades to the HWRF modeling system-Further Enhancements to the High-resolution HWRF. *HFIP Telecon April 30, 2014*, (http://www.hfip.org/documents/minutes/09_Vijay_HFIP_30Apr2014.pdf).

Tallapragada, V. C. Kieu, 2014: Real-Time Forecasts of Typhoon Rapid Intensification in the North Western Pacific Basin with the NCEP Operational HWRF Model. *Tropical Cyclone Research and Review*, **3(2),** 63-77.

Yang, R., J. Tang, and D. Sun, 2011: Association rule data mining applications for Atlantic tropical cyclone intensity changes. *Wea. Forecasting*, **26,** 337–353. doi: http://dx.doi.org/10.1175/WAF-D-10-05029.1

Appendix A HWRF-14 Hits
http://www.dtcenter.org/eval/hwrf_rirw/docs/Appendix_A_Hits.pdf

Appendix B HWRF False Alarms
http://www.dtcenter.org/eval/hwrf_rirw/docs/Appendix_B_false_alarm.pdf

Appendix C HWRF-14 Misses
http://www.dtcenter.org/eval/hwrf_rirw/docs/Appendix_C_misses.pdf