Does Increased Horizontal Resolution Improve Hurricane Wind Forecasts?

CHRISTOPHER DAVIS, WEI WANG, AND JIMY DUDHIA

NCAR,* Boulder, Colorado

RYAN TORN

University at Albany, State University of New York, Albany, New York

(Manuscript received 19 March 2010, in final form 25 June 2010)

ABSTRACT

The representation of tropical cyclone track, intensity, and structure in a set of 69 parallel forecasts performed at each of two horizontal grid increments with the Advanced Research Hurricane (AHW) component of the Weather and Research and Forecasting Model (WRF) is evaluated. These forecasts covered 10 Atlantic tropical cyclones: 6 from the 2005 season and 4 from 2007. The forecasts were integrated from identical initial conditions produced by a cycling ensemble Kalman filter. The high-resolution forecasts used moving, storm-centered nests of 4- and 1.33-km grid spacing. The coarse-resolution forecasts consisted of a single 12-km domain (which was identical to the outer domain in the forecasts with nests). Forecasts were evaluated out to 120 h. Novel verification techniques were developed to evaluate forecasts of wind radii and the degree of storm asymmetry. Intensity (maximum wind) and rapid intensification, as well as wind radii, were all predicted more accurately with increased horizontal resolution. These results were deemed to be statistically significant based on the application of bootstrap confidence intervals. No statistically significant differences emerged regarding storm position errors between the two forecasts. Coarse-resolution forecasts tended to overpredict the extent of winds compared to high-resolution forecasts. The asymmetry of gale-force winds was better predicted in the coarser-resolution simulation, but asymmetry of hurricane-force winds was predicted better at high resolution. The skill of the wind radii forecasts decayed gradually over 120 h, suggesting a synoptic-scale control of the predictability of outer winds.

1. Introduction

There is little doubt that the prediction of hurricane intensity, as measured by the maximum sustained 10-m wind, remains a daunting challenge even after four decades of research. Improvement of intensity forecasts has been slow (DeMaria et al. 2007) compared to improvements in forecasts of storm position. Aberson (2008) showed that alternative verification methods do reveal improvements in some aspects of hurricane intensity forecasts. But there appear to be some fundamental limitations to predicting hurricane intensity as well (Zhang and Sippel 2009). The use of a highly localized attribute to define intensity has also contributed to predictability limitations.

To address the challenge of improving forecasts of hurricane intensity and rapid intensity change out to 5 days lead time, the 10-yr Hurricane Forecast Improvement Project (HFIP) has been organized by the National Oceanic and Atmospheric Administration (NOAA). An initial HFIP project is the determination of whether increases in horizontal resolution improve hurricane forecasts. In particular, the question was raised whether decreasing the grid spacing from roughly 10 to 1–4 km produced a measurable improvement in hurricane intensity and structure prediction, without degrading the track forecasts. During the past year, several groups have participated in what is referred to as the High-Resolution Hurricane (HRH) test. Groups were instructed to use model configurations of their choosing, with the constraint that the coarse and finescale configurations had to be identical apart from the addition of finer resolution. Typically, as in the present study, this enhanced

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Christopher A. Davis, NCAR, P.O. Box 3000, Boulder, CO 80307. E-mail: cdavis@ucar.edu

DOI: 10.1175/2010WAF2222423.1

© 2010 American Meteorological Society
resolution was achieved through moveable, nested grids. Moving nests are critical to obtaining significantly enhanced horizontal resolution within the hurricane inner core for a manageable increase in computational cost.

To understand part of the motivation for this test, it is useful to review the more general influence of varying model resolution on the prediction of phenomena that involve deep, moist convection. In the severe weather community there has been considerable debate for the past several years about whether decreasing the horizontal grid spacing in models improves forecast quality. Within this context, quality refers to the timing, location, and structure of deep convection. Some results (Done et al. 2004; Kain et al. 2006; Weisman et al. 2008) suggest that increasing horizontal resolution improves forecast quality while Schwartz et al. (2009) do not. The salient aspect of improved forecasts appears to be the removal of cumulus parameterization as the grid spacing decreases from around 10 to 4 km or less. Comparing forecasts with explicit convection at differing resolutions does not appear to yield significant differences as long as the resolutions compared are both in the range of 1–4 km (Schwartz et al. 2009).

There are numerous additional considerations that accompany an increase in horizontal resolution. The first is whether vertical resolution should always vary in proportion to the horizontal resolution. Examination of the sensitivity of moist convection to variations in vertical resolution is relatively scarce in the literature. However, based on the results of Aligo et al. (2009), it does not appear that forecasts of deep moist convection are systematically improved by increasing vertical resolution, provided that the vertical resolution is a few hundred meters or less. This contrasts with the well-known consistency of vertical and horizontal resolutions needed to avoid spurious oscillations in baroclinic waves and fronts (Lindzen and Fox-Rabinovitz 1989; Persson and Warner 1991).

Second is the fact that initialization data for models usually contain little information on the convective scale (Skamarock 2004). Hence, most resolution comparisons, including the present study, do not consider the effects of resolving additional scales of motion in the initial conditions. This constrains all differences to develop during the integration. A related point is the lateral boundary conditions that are typically required for simulations with grid increments of a few kilometers or less. Lateral boundaries have the effect of continually sweeping out finescale detail (Warner et al. 1997). Both of the above factors will influence the present study and must be recalled when interpreting our results.

For hurricane forecasts, there are no studies that measure the improvements of forecasts on either side of the transition to explicit convection (i.e., roughly 10 to 1–4 km) with a large number of cases (e.g., many tens or more). Numerous case studies show improvements in structural realism and intensity prediction with increasing horizontal resolution (e.g., Chen et al. 2007). Fierro et al. (2009) caution that varying the horizontal grid spacing between 1 and 5 km does not produce large changes in the quality of the simulation if precipitation is modeled explicitly throughout the range. This result echoes that of Schwartz et al. (2009) for continental moist convection. The result is somewhat surprising given that the eyewall is often poorly resolved on a 5-km grid, but generally well resolved on a 1-km grid.

However, it is difficult to make general statements about resolution dependence from case studies. There are cases (as we will show) for which increasing resolution produces a measurably worse forecast. The focus of the present article is to assess the systematic differences in hurricane structure, intensity, and track that arise from varying horizontal resolution. We do so by integrating the full suite of HRH test cases (section 2) and performing statistical verification of the results. Standard root-mean-square-error metrics are applied to track and intensity forecasts, but to verify a storm’s structure, new methods are devised to make use of the available estimates of the radial extent of various surface-wind thresholds in different quadrants of the storm (i.e., wind radii). These methods are outlined in section 2 as well. The statistical results, as well as a detailed evaluation of statistical significance, appear in section 3. A summary of comparisons between coarse- and fine-resolution forecasts appears in the final section.

2. Model and verification methods

The model that is the focus of this paper is the Advanced Hurricane-Research version (AHW) of the Weather Research and Forecasting Model (WRF). It is derived from the Advanced Research version (ARW; Davis et al. 2008) of the WRF. For the present test, the base version of ARW was 3.0.1.1, and the major upgrades were (i) the use of ensemble data assimilation to initialize the storm and (ii) an improved representation of the spatial variation of the upper-ocean mixed layer with a simple 1D ocean model. The basic configuration of the AHW is summarized in Table 1. The two sets of forecasts being compared differed only by whether two nests were present, with grid spacings 4 and 1.33 km, embedded within a coarse domain with a grid spacing of 12 km, or whether the coarse domain was integrated without nests. The former are referred to as the nested forecasts while the latter are referred to as the 12-km forecasts. The number (34) and spacing of vertical levels are the same in both sets of forecasts.
While the physical parameterizations for turbulence, microphysics, and radiation were unchanged from those used in Davis et al. (2008), the surface enthalpy flux formulation was slightly altered to produce the ratio of enthalpy (\(C_k\)) and drag (\(C_d\)) exchange coefficients shown in Fig. 1. Beyond a wind speed of about 30 m s\(^{-1}\), it is difficult to justify any particular treatment of enthalpy exchange. However, recent work that became known to the authors after the HRH testing was finished suggests that this coefficient tends to remain roughly constant out to at least 40 m s\(^{-1}\) (S. Chen, 2010, personal communication), which implies a nearly constant ratio \(C_k/C_d\).

Because the primary focus in the present paper is a resolution comparison that uses identical formulations for exchange coefficients at each resolution, the precise form of \(C_k\) at high wind speeds may not affect the comparison appreciably.

The 1D mixed-layer ocean parameterization is identical to that described in Davis et al. (2008) with two exceptions. First, the atmospheric frictional velocity used to drive the ocean model was limited empirically to 1.6 m s\(^{-1}\). Second, in the initial conditions we allow a horizontal variation of the mixed-layer depth derived from the horizontal variation of the ocean heat content. The ocean heat content, defined as the integral of the temperature in the upper 100 m of the ocean, \(Q = \int_{100m}^{Z} T \, dz\), contains significant horizontal variation, which greatly modifies the wind-driven mixing and cooling of the uppermost ocean layer. By decomposing the integrand into a variable mixed-layer depth, \(D\), and assuming a constant lapse rate \(\Gamma\) for the remainder of the 100-m layer, the expression for mixed-layer depth is

\[
D = Z - \left[\frac{2(T_s Z - Q)}{\Gamma}\right]^{1/2},
\]

where \(\Gamma = 0.2 \, \text{K m}^{-1}\), \(Z = 100\, \text{m}\), and \(T_s\) is the sea surface temperature. The heat content \(Q\) was obtained from the operational Hybrid Coordinate Ocean Model (HYCOM; Bleck 2002). Where the ocean depth is between 10 and 100 m, the limit of integration for \(Q\) is adjusted accordingly and \(Z\) is set equal to that depth. Where the depth of the ocean is less than 10 m, \(D\) is set to 10 m. A typical spatial distribution of mixed-layer depth appears in Fig. 2.

For the HRH test, we performed 69 pairs of simulations for a total of 10 Atlantic tropical cyclones. The model was initialized using an ensemble Kalman filter (EnKF) consisting of 96 members at 36-km grid spacing (Torn 2010). Assimilated observations included surface pressure, rawinsonde [including Gulfstream IV (G-IV) dropsondes], Aircraft Communicating, Addressing, and Reporting System (ACARS) winds and temperatures, cloud motion vectors, and tropical cyclone advisory data (location and minimum pressure). The update cycle for the ensemble assimilation was 6 h. The ensemble was initialized roughly 2 days prior to the disturbance being classified as a depression by adding balanced perturbations from the variational data assimilation system of the WRF (WRF-Var) to the Global Forecast System (GFS) 36-h forecast valid at the appropriate time. Using an old forecast with high-amplitude perturbations helped the ensemble develop a flow dependence more quickly than starting from forecasts of 6–12 h (Dirren et al. 2007).

In an ideal situation, we would like to integrate the entire ensemble forward using fine and coarse resolutions to understand the dependence of predictability on resolution. However, computational constraints forced the selection of a single ensemble member for a deterministic forecast. This member was chosen as the one closest to the observed intensity at initialization time. This choice, rather than using the member closest to the ensemble mean, was motivated by the fact that the ensemble on a 36-km grid had a negative intensity bias. Pairs of forecasts, one with a single 12-km grid, the other with storm-centered, moving nests of 4- and 1.33-km grid spacings, were integrated to 126 h or until the time

---

**Table 1.** The AHW configuration. The model top is at 20 hPa. Here, KF refers to the Kain–Fritsch cumulus scheme (Kain 2004), WSM5 is the WRF single-moment microphysics scheme with four categories of condensed water (rain, snow, cloud water, and cloud ice), and YSU is the Yonsei University planetary boundary layer (PBL) scheme Hong et al. (2006). The 1.33-km grid is centered within the 4-km grid and both move with the storm as detailed in Davis et al. (2008).

<table>
<thead>
<tr>
<th>Domain size</th>
<th>12 km</th>
<th>4 km</th>
<th>1.33 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of vertical levels</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Time step (s)</td>
<td>60</td>
<td>20</td>
<td>6.7</td>
</tr>
<tr>
<td>Cumulus scheme</td>
<td>KF</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Microphysics scheme</td>
<td>WSM5</td>
<td>WSM5</td>
<td>WSM5</td>
</tr>
<tr>
<td>PBL scheme</td>
<td>YSU</td>
<td>YSU</td>
<td>YSU</td>
</tr>
</tbody>
</table>

**FIG. 1.** Ratio of \(C_k/C_d\) as a function of wind speed.
the observed storm dissipated. Lateral boundary conditions were obtained from the GFS forecast initialized at the same time as AHW. Table 2 summarizes the storms and the number of forecasts for each.

The present article focuses on verification of standard parameters including storm position, maximum wind, and minimum sea level pressure. Position errors are considered as both root-mean-square errors and biases along and across the track of the storm. The storm motion is estimated from the best-track average movement over the 12 h ending at the verification time. Position and intensity errors are compared for the high-resolution forecasts (also termed nested forecasts), 12-km grid spacing (hereafter “12-km forecasts”) and the official forecasts (OFCL) produced by the National Hurricane Center. Provided there was a best-track representation of the storm, verification was performed. The only exception was for Hurricane Ophelia when it approached the northern boundary of the coarse domain during extratropical transition. Unless otherwise stated, all samples are homogeneous across different forecasts.

For root-mean-square errors (RMSEs), the statistical significance of the differences between forecast sources is assessed using a bootstrap resampling method where random samples (with replacement) are generated from the distribution of RMSEs. In this particular application, the entire distribution is resampled 10,000 times and bootstrapping is performed on the differences between the squared errors of paired samples. The level of statistical significance is then based on the rank at which the resampled distribution of the differences crosses zero. This approach differs from the more standard bootstrapping method, in which resampling of two distributions is performed separately. In cases with two distributions, each with a large variance but a small systematic offset between them, bootstrapping the difference yields notably larger significance levels than bootstrapping the distributions separately. Only significance levels of 90% or greater are considered meaningful herein.

Despite the emphasis on the maximum wind, aspects of the wind field such as radial extent and asymmetry can

---

1 Because no official forecast was issued for Felix corresponding to the numerical forecasts initialized at 1200 UTC 31 Aug 2007, there are only 68 cases in the three-way intercomparison instead of the full 69 cases in the HRH test.

2 There is some degree of serial correlation among forecasts. However, the highly irregular temporal spacing of the forecasts in our sample makes it problematic to implement traditional blocked bootstrap approaches (Wilks 1997) that account for such correlation. Therefore, blocked bootstrapping is not attempted.
strongly affect the destructive potential of hurricanes as manifested through storm surge (Irish et al. 2008), waves, duration, and area affected (Powell and Reinhold 2007; Maclay et al. 2008). This motivates evaluation of forecasts of spatial attributes of the surface wind field. The attributes on which we focus are the radial extent of 64-, 50-, and 34-kt (32.9, 25.7, 17.5 m s$^{-1}$) winds in each of the four directional quadrants (northeast, northwest, southwest, and southeast), known generically as wind radii. Estimates of wind radii derived from observations were obtained from the extended best-track data (Demuth et al. 2006). As noted in Knaff et al. (2007), the estimates of wind radii are rather uncertain, especially when data from reconnaissance aircraft are not available. However, the availability of Quick Scatterometer (QuikSCAT) and Advanced Microwave Sounding Unit (AMSU) data allow reasonable estimates of at least the radius of 34-kt winds even for storms where reconnaissance is not available. Most of the storms in the HRH dataset were sampled many times by reconnaissance aircraft. Particularly well sampled storms include Emily, Katrina, Rita, Ophelia, and Wilma from 2005, and Felix and Humberto from 2007. All 10 storms had some data taken from reconnaissance.

For the model, at each time and in each quadrant, the maximum radial distance of a given wind speed is computed. This computation is consistent with the operational definition of wind radii. Wind radii appear as zero, and this category is distinct from all other values of wind radius.

Table 2. Storms and the numbers of forecasts for each storm.

<table>
<thead>
<tr>
<th>Storm</th>
<th>No. of forecasts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emily (2005)</td>
<td>10</td>
</tr>
<tr>
<td>Katrina (2005)</td>
<td>6</td>
</tr>
<tr>
<td>Philippe (2005)</td>
<td>6</td>
</tr>
<tr>
<td>Rita (2005)</td>
<td>7</td>
</tr>
<tr>
<td>Ophelia (2005)</td>
<td>11</td>
</tr>
<tr>
<td>Wilma (2005)</td>
<td>11</td>
</tr>
<tr>
<td>Felix (2007)</td>
<td>8</td>
</tr>
<tr>
<td>Humberto (2007)</td>
<td>2</td>
</tr>
<tr>
<td>Ingrid (2007)</td>
<td>4</td>
</tr>
<tr>
<td>Karen (2007)</td>
<td>4</td>
</tr>
</tbody>
</table>

The method of determining wind radii and the asymmetry is summarized visually in Fig. 3. Shown is the 10-m wind speed for a 24-h forecast of Katrina valid at 0000 UTC 28 August 2005. The range rings (bins) are marked every 30 n mi in the northeast and southwest quadrants, where 30 n mi is the interval chosen to divide the distribution of 34-kt wind radii. The northeast quadrant, the 34-kt radius extends to the seventh bin from the center, but only to the fourth bin in the southwest
quadrant. Based on the extended best-track data, the observed 34-kt winds extend only to the fifth bin in the northeast quadrant and to the fourth bin in the southwest quadrant. Thus, the joint histogram bin (7, 5) is incremented by one, as is the bin (4, 4), the latter constituting a hit. The same procedure is followed for the other quadrants.

The asymmetry parameter $A$ is computed from the difference in wind radii between opposite quadrants. In the example, the maximum radius of the 34-kt winds in the northeast quadrant would be subtracted from the maximum radius of 34-kt winds in the southwest quadrant. These radii are indicated by the arrows. The difference between the northeast and southwest quadrants is larger than the difference between the other two opposing quadrants, and hence this difference is used to define $A$. In the example in Fig. 3, that difference is about 100 n mi, whereas the average radial extent of 34-kt winds is about 150 n mi. This yields a value of $A$ near 0.7. This value indicates moderate asymmetry, near the median of the overall distribution of asymmetry values (not shown).

3. Results

a. Position

Comparing RMS position errors among the nested, 12-km, and official forecasts, the official forecasts were superior through 72-h lead time. Overall, there was little difference between nested and 12-km forecasts (Fig. 4). The bootstrapping method outlined in section 2 was applied to the paired differences in position error, and it was found that the differences between nested and 12-km forecasts did not reach 90% confidence at any time. The overall errors in all forecasts, including the official forecasts, are rather large compared to the climatological errors in the official forecasts after 72 h. Some large errors were contributed by Ophelia late in its life cycle after it underwent extratropical transition.

Position biases are quantified by decomposing errors into along-track and across-track errors (Fig. 5). In this relative coordinate system, symbols denoting forecasts that move the storm systematically too quickly and to the right of the track will appear in the upper-right quadrant. A slow and leftward bias will appear in the lower-left quadrant. At each time the nested simulations have a slightly smaller bias, based on a smaller distance to the origin. For the most part, the bias is to the left of the observed track, although by 120 h, the bias lies to the right and ahead of the observed position. Applying a one-sided $t$ test to the distributions of position errors at different lead times, it turns out that both 12-km and nested forecasts produce statistically significant left-of-track biases through 72 h. Here, significance is defined as 95% confidence in correctly rejecting the null hypothesis that mean errors are indistinguishable from zero. Confidence levels for the 12-km forecasts exceed 99.5%. Along-track biases and official forecast biases are not significant at any lead time. Differences in cross-track errors between the 12-km and nested forecasts are also not significant. This last point is consistent with the overall finding that track forecasts from 12-km and nested forecasts are statistically indistinguishable.
b. Intensity

Errors in maximum wind are examined using several metrics. A standard metric is the RMSE of the maximum 1-min sustained wind. From the numerical forecasts, this is computed as the maximum of the wind at 10-m altitude. Both instantaneous values and the average of nine maximum wind values spaced 15 min apart over a 2-h period centered on the valid time were computed and the results differed little. In what follows, the single instantaneous value is used.

Errors in the official forecast are significantly smaller than in both sets of numerical forecasts at 12 h (Fig. 6), and remain significantly smaller than the errors in the 12-km forecasts through 36 h. Here, “significance” has the precise meaning defined in section 2 based on bootstrapping of pairwise differences in errors. Intensity errors in the nested and official forecasts are indistinguishable from 24 to 48 h. At 72 h, the official forecasts have significantly larger errors than either numerical forecast. At this time, the 12-km forecasts also have larger errors than the nested forecasts at a significant level. Finally, at 120 h, the nested forecasts again have significantly smaller errors than the 12-km forecasts. At no time are the 12-km forecasts distinguishably better than the nested forecasts. Aggregating all lead times, the RMSE of the nested forecasts is about 8% less than the RMSE of the 12-km forecasts.

Intensity forecast biases are not large in general (Fig. 6). The 12-km forecasts maintain a negative bias at all lead times. The nested and official forecasts have either a small positive or nearly 0 bias through 96 h. All forecasts exhibit a negative bias at 120 h. Applying a $t$ test to assess the statistical significance of the biases and denoting the significance based on 95% confidence of correctly rejecting the null hypothesis, we find that only at 24 h is the bias significant in the nested forecasts, while from 12 to 48 h, and again at 120 h, the negative bias of the 12-km forecast is significant. Biases in the official forecast intensity are not significant.

Further examination of the error distribution at 72 h is accomplished by a rank ordering of the absolute errors (Fig. 7). Intensity biases are generally small at this time (Fig. 6), and much of the difference in the pattern of behavior is determined by the presence or absence of a few large errors. The largest errors in the official forecast come from Felix, for which significant intensification was not anticipated in early forecasts, but in fact, Felix was a category 5 hurricane 72 h after the forecast initialization time from 1200 UTC 1 September 2007.

The RMSEs for individual storms (Fig. 8), aggregated over lead times from 12 to 120 h, indicate a substantial
variation in performance of each of the three forecasts relative to each other. The official forecasts for Felix were worse than those of the models, primarily due to a poor forecast of the initial, rapid intensification on 1 and 2 September. Numerical forecasts of Karen and Ingrid were substantially worse than the corresponding official forecasts. Both were highly sheared storms with large asymmetries. Early forecasts of these storms over-intensified them, perhaps because the vertical wind shear was not strong enough in the forecast. Karen was also the only storm for which the high-resolution forecast fared worse than the 12-km forecast. The 12-km forecast was poorer than the nested forecast for Felix, Wilma, and Ingrid. In Felix, the small scale of the inner core was not well resolved on a 12-km grid. Because of their longevity, Ophelia, Wilma, and Emily contributed the most to the overall statistics. The improvement of the official forecast over the numerical forecasts for these storms is reflected in the overall statistics (Fig. 6).

To quantify whether model errors depend on storm intensity and, therefore, to assess whether a given set of forecasts enjoys a systematic advantage, we constructed a joint histogram of the observed intensity versus the difference of the absolute value of the forecast intensity error (Fig. 9). Torn (2010) found that with EnKF on a 36-km grid there was a significant low intensity bias for storms of category 3 or greater. The present comparison of 12-km versus nested forecasts suggests a similar result. Most of the larger errors for the 12-km forecasts occur for maximum winds exceeding 100 kt. Errors for nested forecasts tend to be larger for weak storms. These errors are consistent with the intensity bias: the 12-km forecasts have a low intensity bias for strong storms and the nested forecasts have a high bias for weak storms (not shown).

Minimum sea level pressure (SLP) errors from the numerical forecasts were also compared. Errors in both models are relatively large at $t = 0$ due to the high SLP bias in the ensemble (Torn 2010). The nested forecasts were nonetheless significantly better than the 12-km forecasts at 12 and 24 h. Lower SLP errors were also found from 36 to 72 h, but differences were not statistically significant. The improvement of the nested forecasts at early times suggests that the increased horizontal resolution allows the model to adjust faster to the coarse initial conditions, especially for the more intense storms.

Rapid intensification was also examined as a dichotomous variable. Here, we define rapid intensification as

![Fig. 6. The RMS intensity errors vs forecast lead time for a homogeneous sample. Gray diamonds denote high-resolution (nested) forecasts, black squares denote 12-km forecasts, and OFCL are denoted with black triangles. Also shown are biases with unfilled shapes. Units are kt. Number of forecasts appears across the bottom of the graph. Gray arrows indicate lead times when the nested forecast is significantly better than the 12-km forecast (90% confidence).](image1)

![Fig. 7. Ranked absolute intensity errors for all 56 valid times of the 72-h forecasts.](image2)
an increase of maximum wind of 25 kt or more in 24 h. This is 5 kt less than the operational definition. The lower threshold is chosen to increase the number of events and improve the statistics. The standard equitable threat score is defined based on a 2 x 2 contingency table where hits are correct forecasts of rapid intensification (at the correct time) and correct negative forecasts are correct forecasts of a lack of rapid intensification in a 24-h period. Despite restricting the valid times to 24, 36, 48, 72, 96, and 120 h, there were still a total of 55 observed instances of rapid intensification, of which the nested forecasts predicted 21, the 12-km forecasts predicted 10, and the official forecasts predicted 3 (Table 3). The nested forecasts also had the greatest number of false alarms but still retained the largest equitable threat score compared with the 12-km nest and the official forecast. The nested forecasts also predicted approximately the correct number of rapid intensifications (51 forecast versus 54 observed), which was not true of the 12-km and official forecasts.

c. Wind radii and asymmetries

Before considering a deterministic evaluation of wind radii forecasts, the overall distributions of wind radii are examined (Fig. 10). The histograms shown in Fig. 10 cover all forecast lead times from 12 to 120 h. The 34- and 64-kt radii distributions are quantized differently because the 34-kt wind radius typically greatly exceeds the 64-kt wind radius. The width of the bins is determined mainly by a desire to separate approximately the range of the distribution into a number of partitions great enough to reveal some detail, but not so small as to lose signal. Furthermore, counts of a few hundred or so will prove important for maintaining robust statistics in the verification that follows.
It is also tempting to interpret the width of the partitions as related to uncertainties in estimating wind radii. Unfortunately, we do not know what this uncertainty is. Moyer et al. (2007) compared the extended best-track estimates of wind radii to objective calculations from the H*Wind Tropical Cyclone Observing System application of NOAA’s Hurricane Research Division (Powell et al. 1998) and found that the best-track estimates of 34-kt wind radii were smaller than those from H*Wind, and the difference was roughly 25%, although it was indeterminate at radii less than about 120 n mi. Possible errors in estimates of the extent of 50- and 64-kt winds are unknown.

The observed peak in 34-kt wind radius (Fig. 10a) occurs between 90 and 120 n mi from the storm center. Both models essentially reproduce this peak. There are many more instances where the predicted 34-kt wind radius lies outside 240 n mi than are observed, especially in the 12-km forecasts. These may include some cases where the 34-kt wind radius extended beyond what would reasonably be considered the tropical cyclone circulation due to anomalously strong synoptic-scale surface winds favoring a particular quadrant. It is also likely that these large wind radii would be more consistent with those derived from H*Wind given the results from Moyer et al. (2007).

For the extent of hurricane-force winds (Fig. 10b), there is a clear discrepancy at smaller radii where far more observed wind radii occur than are predicted between 10 and 30 n mi. Much of this bias arises from the relatively coarse initial conditions used for the model, wherein the inner-core structure is marginally resolved (not shown). The bias is greatest at short lead times and actually becomes less as lead time increases. There is also still a positive forecast bias at large wind radii, and it is larger than for the 34-kt wind radii.

It is also useful to document the absence of winds exceeding a given threshold in a quadrant (Fig. 10c). For the 50- and 64-kt wind thresholds, the forecasts too frequently predict the absence of wind speed reaching that threshold. The higher-resolution forecasts have larger errors in this aspect. The monotonic increase in missing wind radii for higher wind speeds reflects the fact that storms may be marginal tropical storms, in which none of the four quadrants would have hurricane-force winds, or may be asymmetric such that one or more of the quadrants have no winds exceeding 64 kt.

Recall that very limited information about the vortex structure was included in the data assimilation procedure. Only position and minimum sea level pressure estimates were included as direct observations. The wind strength in the outer radii would be affected by satellite-derived winds, dropsondes, and soundings, but no reconnaissance data within 200 km of the center were assimilated, nor was the maximum wind speed assimilated, so the pressure–wind relationship was also not a direct input to the initialization. The extent to which the forecast and observed distributions resemble each other at all is primarily because of the cycling of larger-scale information.

Deterministic verification of the wind radii is accomplished through the joint histogram approach described
in section 2. Displayed in Fig. 11 are counts of forecast–observation pairs accumulated in bin widths of 30 n mi for the 34-kt wind radii and 10 n mi for the 64-kt wind radii. For 50-kt wind radii, we use a box width of 15 n mi. In a perfect forecast, only the grid cells along the 1:1 line would contain values. An excess of filled boxes to the right of the line indicates that the forecast too often overestimated the radial extent of the winds of a given strength. The left-most column contains the accumulation of forecasts where the wind did not exceed a given threshold in a given quadrant. The bottom row is the analogy for the observations. The count in the box in the lower-left corner indicates forecasts in agreement with observations about the absence of winds exceeding the threshold in a particular quadrant.

Based on the clustering of many of the pairs near the 1:1 line, it is apparent that forecasts have some ability to predict wind radii. To more objectively assess skill, we define an equitable threat score based on concepts from dichotomous forecasts, but modified to account for multiple categories. In the joint histogram, the total count within each box along the diagonal (1:1 line) represents the total number of hits. Each remaining box along the row that contains a particular “hit box” contains misses. These are observations within a given radius bin that were not forecast to be in that bin. Similarly, each remaining
box in the column that intersects a particular hit box contains the false alarms. These are forecasts of a given wind radius that proved wrong. All boxes not part of the above row, column, or diagonal element define the correct negative forecasts, that is, forecasts and observed wind radii that both did not occur within the radius bin considered. This does not say that those forecasts represent hits; many do not. They are only correct negatives with respect to one wind radius bin.

We then sum the hits, misses, false alarms, and correct negatives over all wind radius bins. If we denote $\Sigma$ as the total number of forecast–observation pairs and $N$ as the number of bins, the equitable threat score (ETS) can be written as

$$ \text{ETS} = \frac{H - e}{2\Sigma - H - e} $$

where $e = \Sigma/N$ is the expected number of hits due to chance and $H$ is the sum of the counts in all boxes along the diagonal. Note that the denominator of (3) differs from the standard ETS definition. The number of misses equals the number of false alarms, $\Sigma - H$. The ETS can also be computed for any range of wind radius values. In particular, in addition to a total ETS, it is useful to consider an ETS for all categories except zero, where zero represents the absence of a wind exceeding a given threshold in a particular quadrant. This is denoted ETSP. Note that false alarms and misses may still fall into the zero bins, but ETSP does not count hits representing a correct forecast of the absence of a wind radius.

The ETSs for 34-kt wind radii, considering all lead times concurrently as in Fig. 11, are 0.11 for the nested forecasts and 0.07 for the 12-km forecasts, respectively. ETSP, which is always less than ETS, behaves similarly to ETS for 34-kt winds because nearly all quadrants in all storms contain gale-force winds. However, for 64-kt wind radii, ETSP is a notably more stringent metric than ETS because it does not give credit for the relatively easy forecast of wind below the hurricane threshold in weak storms. ETSP for 64-kt wind radii is small, 0.025 for nested forecasts and only 0.006 for 12-km forecasts.

Figure 12 shows values of ETS and ETSP for the 34- and 64-kt wind radii as a function of forecast lead time. The ETSs for the 34-kt wind radii forecasts do not exceed 0.16 at any time. ETS and ETSP values are higher at nearly all lead times in the nested forecasts compared to the 12-km forecasts. Skill scores from both models tend to decrease toward zero with increasing forecast length. The rather slow decay of skill over 5 days, better defined for the 34-kt wind radii, suggests that errors in predicting outer wind radii are linked to synoptic-scale error growth.

To determine the statistical significance of differences in wind radii forecasts, we employed a bootstrapping method similar to that described in section 2. We performed bootstrapping on the full sample containing all lead times to assess the significance in the overall results. We also assessed significance of the differences at each lead time individually. For each of 10 000 samples, randomly generated with replacement, we computed the ETS value, paired the 12-km and nested ETS values, and computed the significance as described in section 2. For the 34-, 50-, and 64-kt wind radii, the ETSs for nested forecasts could be considered larger with greater than 99% confidence, both for the full sample containing all lead times, and for specific lead times between roughly 24 and 84 h. All differences between the two forecasts are 0 at $t = 0$, so it is perhaps not surprising that the greatest difference in skill occurs toward the middle of the forecast period. This represents the time by which the perturbations have grown, but also a time not so long that predictability is lost.

The criteria for a hit may appear rather stringent, and the resulting ETSs correspondingly low. Given the uncertainty in observed wind radii, it may be appropriate to modify the definition of a hit to allow a one-category error. For a given bin along the 1:1 diagonal, this amounts
to adding the counts in the surrounding four bins (up, down, left, and right of the diagonal bin) to the total of hits. Doing so markedly raises the ETSP scores for both models by roughly a factor of 4 (not shown). For the high-resolution forecasts of 34-kt winds, scores at early lead times exceed 0.4, and drop to roughly 0.2 by 120 h. As before the high-resolution forecasts have significantly better scores than the coarser-resolution forecasts at nearly every lead time.

There are two other relevant forecasts of wind radii whose skill may be compared with the numerical forecasts. One is the persistence forecast, defined as a perpetuation of the 6-h forecast of wind radii through 120 h. Verified using the same skill metric as above, persistence does nearly as well as the numerical forecasts for about 36 h (not shown). The skill of persistence drops abruptly after this time. This result is approximately the same for 34-, 50-, and 64-kt wind radii.

Another relevant forecast of wind radii is provided by official forecasts, although these are restricted to 72 h for 34-kt winds and 36 h for 64-kt winds. For a homogeneous sample of 34-kt wind radii forecasts, the ETSP reaches 0.41 at 12 h for the official forecasts compared to 0.09 for the nested forecasts and 0.06 for the 12-km forecasts. The ETSP values for official forecasts of 34-kt wind radii remain higher than the numerically predicted values at all times out to 72 h. The official forecasts have the advantage of starting with essentially 0 error. Because wind radii appear to evolve on time scales of a day or more, it is not surprising that the enhanced skill is retained. Furthermore, one could argue that the short-term official forecasts are not independent of the analysis.

The asymmetry parameter $A$ is evaluated using a similar joint histogram approach as for wind radii. Asymmetry is a nondimensional parameter with a range of 0–4. However, values exceeding 2 are very rare; most of the values are around unity or less. The increment chosen to partition the distribution of $A$ is 0.2 (this is the bin width). In Fig. 13 are shown the joint histograms for the 34- and 64-kt wind radii. There is a reasonable clustering of asymmetry values near the 1:1 line, suggesting that the forecasts do indeed provide some indication of the degree of asymmetry in a given storm at a given time. The dynamic range of the asymmetry is also similar in the two datasets, as are the median values. There is a tendency for the 12-km forecasts to produce excessive asymmetry for the 34-kt threshold. By contrast, the nested forecasts sometimes underestimate the asymmetry for hurricane-force winds.

Neither numerical forecast is obviously superior in predicting the degree of asymmetry for all wind thresholds (Table 4). A skill score analogous to that defined for wind radii was used to quantify the joint histogram results. Also included in Table 4 is a three-way comparison between the 12-km, nested, and official forecasts, although the sample size of this comparison is relatively small. The ETS score for the 34-kt joint histogram is slightly higher for the 12-km forecasts (0.10 versus 0.08) in the two-way comparison. A nearly opposite result is obtained in the three-way comparison. Because the three-way comparison is restricted in lead time, the result implies that higher resolution is beneficial for relatively short-range forecasts of asymmetry. In the three-way comparison, the scores for the official forecasts are again higher, but not by as much as for the wind radius verification. The number of hits in the three-way comparison drops to 15–20 for the 64-kt wind asymmetry, so results from the three-way comparison must be interpreted cautiously.

### 4. Conclusions

The present paper has compared the accuracy of hurricane track, intensity, and structure in a set of 69 forecasts performed at each of two horizontal grid increments with the Advanced Hurricane Research component of the WRF (AHW) model. These forecasts covered 10 Atlantic tropical cyclones: 6 from the 2005 season and 4 from 2007. The forecasts were integrated from identical initial conditions produced by a cycling ensemble Kalman filter (Torn 2010). The high-resolution forecasts used moving, storm-centered nests of 4- and 1.33-km grid spacings. The coarse-resolution forecasts consisted of a single 12-km domain (which was identical to the outer domain in the forecasts with nests). Forecasts were integrated out to 126 h, or until dissipation of the observed storm, or until the forecast storm came within roughly 200 km of the coarse-domain boundary. Verification of the forecasts was performed using the best-track and extended best-track data. Verification samples were homogeneous.

Storm position errors showed no statistically meaningful differences between the two sets of forecasts. Root-mean-square position errors were greater in the numerical forecasts than in the official forecasts produced by NHC by a statistically significant margin as assessed using confidence intervals obtained from a bootstrapping technique.

Storm intensity, defined as the maximum instantaneous 10-m wind in the model, was slightly better forecast in the nested simulations than in the 12-km forecasts. The statistical significance of the differences was evident at 72- and 120-h lead times. Both sets of numerical forecasts were worse than the official intensity forecast through 24 h, but better at 72 h, again at a statistically significant level. The shortcoming early in the forecasts...
was due to a low bias of intensity that resulted from the coarse representation of storms on the 36-km grid used by the EnKF initialization method. The primary benefit of finer horizontal resolution was for storms of category 3 or greater intensity. The 12-km forecasts exhibited a negative intensity bias for strong storms whereas the nested forecasts revealed a small positive intensity bias for weak storms. Overall, the root-mean-square intensity error for nested forecasts was about 8% less than the error for the 12-km forecasts.

The radial extents of 34-, 50-, and 64-kt winds in each storm quadrant were also evaluated by using a joint histogram approach for depicting the distributions of observed and predicted wind radii. An equitable threat score was defined based on these joint histograms. The high-resolution forecasts were superior to the 12-km forecasts for all wind radii, and for all lead times between 24 and 84 h. Again, these results were determined to be statistically significant using a bootstrap technique. Skill in the wind radii forecasts decayed with time over 120 h, mainly in the last 60 h, in both sets of forecasts. This suggests a synoptic time scale associated with the predictability of wind radii, especially the outer wind radii. A similar result was found by Torn (2010). It is

![Fig. 13. Joint histograms of forecast and observed asymmetry A for (a) 34-kt wind radii asymmetry for nested forecasts and (b) 34-kt wind radii asymmetry for 12-km forecasts. (c),(d) As in (a),(b) but for 64-kt wind radii asymmetry. The asymmetry parameter is dimensionless.](image-url)
 TABLE 4. ETSP values for the asymmetry parameter. The model comparison columns include forecasts out to 120 h, whereas the model and OFCL comparison columns include only forecasts for when official radii forecasts were available. Because of the existence of missing radii (i.e., wind not exceeding the specified threshold in a given quadrant), the samples are not precisely homogeneous. The sample size for the two-way comparison is 3–4 times larger than the sample for the three-way comparison.

<table>
<thead>
<tr>
<th>Threshold (kt)</th>
<th>Model comparison</th>
<th>Model and OFCL comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nest 12 km</td>
<td>Nest 12 km OFCL</td>
</tr>
<tr>
<td>34</td>
<td>0.08</td>
<td>0.09 0.06 0.17</td>
</tr>
<tr>
<td>50</td>
<td>0.07</td>
<td>0.10 0.07 0.10</td>
</tr>
<tr>
<td>64</td>
<td>0.10</td>
<td>0.11 0.08 0.13</td>
</tr>
</tbody>
</table>

therefore somewhat surprising that the inclusion of high-resolution nested domains measurably improves the forecasting of outer wind radii.

An asymmetry parameter was defined based on the difference of wind radii in opposing quadrants of a storm. This parameter was also evaluated in the two sets of forecasts using a joint-histogram approach. The 12-km forecasts of asymmetry were better for the 34-kt winds, whereas the high-resolution forecasts produced better asymmetry forecasts for 64-kt winds. Statistical significance was not computed because the two samples were not completely homogeneous.

In summary, for all except the asymmetry of 34-kt wind radii, the high-resolution forecasts performed as well as, or significantly better than, the coarse-resolution forecasts. We believe this is the first demonstration of the superiority of high-resolution forecasts of hurricane intensity and structure in a relatively large sample of cases. All the improvement in the high-resolution forecasts developed after $t = 0$. We surmise that with high-resolution ensemble data assimilation, higher-resolution forecasts would be improved at short time ranges (36 h or less) by improving the representation of the initial inner-core structure. With cycling, the improvements in storm structure that we noted after 1–2 days of the forecasts would be present throughout the assimilation period. Such an improved background vortex structure may prove to be crucial for successfully assimilating data near the cyclone center. Furthermore, the relatively long time scale of the error growth for the outer wind radii suggests that improved initial vortex representation should also improve the outer wind forecasts over lead times of at least 2–3 days.

Acknowledgments. The authors acknowledge the helpful comments of Richard Rotunno of NCAR. Further, we are indebted to Sherrie Fredrick for performing many of the HFIP forecasts and much of the postprocessing of model output used in this study. We also thank Ginger Caldwell, Tom Engle, Marc Genty, and Sid Ghosh from the NCAR Computer and Information Services Laboratory for ensuring dedicated time on the NCAR bluefire IBM machine to do the retrospective simulations. This work was supported through the NOAA Hurricane Forecast Improvement Project.

REFERENCES


