# The 2012 Triply Nested, High-Resolution Operational Version of the Hurricane Weather Research and Forecasting Model (HWRF): Track and Intensity Forecast Verifications

STANLEY B. GOLDENBERG,* SUNDARARAMAN G. GOPALAKRISHNAN,* VIJAY TALLAPRAGADA,+ THIAGO QUIRINO,* FRANK MARKS JR.,* SAMUEL TRAHAN,+ XUEJIN ZHANG,# AND ROBERT ATLAS@

*NOAA/Atlantic Oceanographic and Meteorological Laboratory/Hurricane Research Division, Miami, Florida
+ NOAA/Environmental Modeling Center/National Centers for Environmental Prediction, Washington, D.C.
# University of Miami/Cooperative Institute for Marine and Atmospheric Studies, Miami, Florida
@ NOAA/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida

(Manuscript received 27 August 2014, in final form 23 December 2014)

ABSTRACT

The Hurricane Weather Research and Forecasting Model (HWRF) was operationally implemented with a 27-km outer domain and a 9-km moving nest in 2007 (H007) as a tropical cyclone forecast model for the North Atlantic and eastern Pacific hurricane basins. During the 2012 hurricane season, a modified version of HWRF (H212), which increased horizontal resolution by adding a third (3 km) nest within the 9-km nest, replaced H007. H212 thus became the first operational model running at convection-permitting resolution. In addition, there were modifications to the initialization, model physics, tracking algorithm, etc. This paper compares H212 hindcast forecasts for the 2010–11 Atlantic hurricane seasons with forecasts from H007 and H3GP, a triply nested research version of HWRF. H212 reduced track forecast errors for almost all forecast times versus H007 and H3GP. H3GP was superior for intensity forecasts, although H212 showed some improvement over H007. Stratifying the cases by initial vertical wind shear revealed that the main weakness for H212 intensity forecasts was for cases with initially high shear. In these cases, H212 over- and under-intensified storms that were initially stronger and weaker, respectively. These results suggest the primary deficiency negatively impacting H212 intensity forecasts, especially in cases of rapid intensification, was that physics calls were too infrequent for the 3-km inner mesh. Correcting this deficiency along with additional modifications in the 2013 operational version yielded improved track and intensity forecasts. These intensity forecasts were comparable to statistical–dynamical models, showing that dynamical models can contribute to a decrease in operational forecast errors.

## 1. Introduction

The National Hurricane Center (NHC) of the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS) uses input from various operational hurricane prediction models to produce its official tropical cyclone (TC) forecasts. For track forecasts, these range from relatively simple statistical models such as Climatology and Persistence (CLIPER5) to sophisticated global and regional forecast models such as NOAA's Global Forecast System (GFS), the NOAA/Geophysical Fluid Dynamics Laboratory's (GFDL) regional hurricane forecasting model (Bender et al. 2007), and others (Cangialosi and Franklin 2013, hereafter CF2013). For intensity forecasts, the models range from the basic statistical climatology–persistence Statistical Hurricane Intensity Forecast model (DSHIFOR5; Knaff et al. 2003) to sophisticated statistical–dynamical models such as the Statistical Hurricane Intensity Prediction Scheme (DSHIPS; DeMaria et al. 2005)[1] and the Logistic Growth Equation Model (LGEM; DeMaria 2009) that predict storm intensity using statistical relationships with

*Corresponding author address:* Stanley B. Goldenberg, AOML/HRD, 4301 Rickenbacker Cswy., Miami, FL 33149.
E-mail: stanley.goldenberg@noaa.gov

DOI: 10.1175/WAF-D-14-00098.1

[1] The "5" in CLIPER5 and DSHIFOR5 refers to the fact that these two models now produce 5-day (120 h) forecasts. The original versions provided forecasts out to only 3 days (72 h). The "D" in DSHIFOR5 (i.e., Decay-SHIFOR5) and DSHIPS (i.e., Decay-SHIPS) refers to adjustments for the decay of storms when they move inland, according to DeMaria et al. (2006).

climatological, persistence, and numerical model predictors, to the various complex dynamical prediction models such as those listed above for track forecasts (a complete listing of these models is found in CF2013). To account for the strengths and weaknesses of the extensive suite of models and to maximize the information obtained from them, NHC also uses various consensus-type forecasts, created by simple and complex combinations of the output from the various track and intensity models (CF2013).

Through the use of newer models and improved versions of older models, NHC has experienced substantial reductions in track forecast errors over the last several decades. In addition, enhanced use of satellite data and the use of data collected by NOAA and U.S. Air Force hurricane hunter aircraft, especially from their synoptic surveillance missions, have produced further significant reductions in track forecast errors (Aberson 2010). Operational track forecasts from dynamical models have shown substantial improvements (generally at least 50% at 72 h compared to CLIPER5) and these improvements have been reflected in the substantial improvements to NHC's official operational track forecasts (CF2013). These improvements to track forecast errors have produced a reduction in warning and evacuation areas, resulting in substantial saving of lives and property (Rappaport et al. 2009).

In spite of all of the new and improved models, however, NHC has not experienced a comparable reduction in its official intensity forecast errors, and the overall improvements in skill (versus a no-skill baseline like DSHIFOR5) have been nominal (CF2013). This is because the processes controlling intensity changes in TCs are much more complex than those controlling the TC tracks. Intensity forecasts are a multiscale problem, requiring not only three-dimensional knowledge of the large-scale storm environment, but also the dynamical and thermodynamical processes that affect the storm's inner-core region at relatively small spatial scales. In addition to the issues of sufficient temporal and spatial resolution (both horizontal and vertical) and adequate modeling of the convection-scale processes in and around the storm, there is also the issue of obtaining more accurate initial conditions of the three-dimensional structure of the storm (Hendricks et al. 2011) and its surrounding environment and assimilating this information into the models.

Ongoing efforts to meet the challenge of significantly improving intensity forecasts have been driven by the importance of accurately predicting TC intensity, especially in the case of rapid intensification (RI) or weakening just prior to landfall. These rapid changes in TC intensity are problematic for preparations and evacuation decisions that must be made by emergency managers and other government officials (Kaplan and DeMaria 2003). For instance, all of the hurricanes known to have made landfall in the United States at the category 5 intensity—the Labor Day (Florida Keys) hurricane (1935), Hurricane Camille (1969), and Hurricane Andrew (1992)—rapidly intensified from tropical storms to category 5 hurricanes in less than 3 days (about 66, 36, and 42 h, respectively). And, despite all of the intensity guidance available to NHC, the rapid weakening of Hurricane Lili (2002) from a strong category 4 hurricane to a category 1 hurricane just prior to its Louisiana landfall still came as a surprise. Obviously, there is a need to produce more reliable intensity forecasts in real time (operationally), especially for cases involving rapid changes in intensity.

To address the deficiencies in intensity prediction skill (with an emphasis on RI events), NOAA established the Hurricane Forecast Improvement Project (HFIP) in 2007. (The HFIP 5-yr strategic plan is available online at http://www.hfip.org/documents/hfip_strategic_plan_yrs1-5_2010.pdf). Since then, HFIP has supported an unprecedented multiagency effort to accelerate improvements to TC predictions (Gall et al. 2013). This effort has focused on providing resources to evaluate and improve several hurricane track and intensity forecast models.

One of the dynamical models being improved as part of HFIP is the Hurricane Weather Research and Forecasting Model (HWRF). HWRF was developed at the NOAA/NWS/Environmental Modeling Center (EMC) as a highly advanced system with the goal of using it as a platform to test and implement the latest improvements in hurricane prediction modeling (Gopalakrishnan et al. 2010). HWRF, with its 27-km outer domain and 9-km moving nest (27:9), first became operational at the National Centers for Environmental Prediction (NCEP) in 2007 (Gopalakrishnan et al. 2010) and has been used as part of NHC's suite of models since that time. The current study outlines the series of modifications to the 27:9 HWRF to create a new operational version (Table 1). These modifications were developed as a collaborative effort primarily between the NWS/EMC and the Hurricane Research Division (HRD) of the Atlantic Oceanographic and Meteorological Laboratory (AOML). The foundational improvement is that the horizontal resolution for HWRF was increased to a 27-km outer domain with 9-km intermediate and 3-km innermost moving nests, making it the first operational model running at a convection-permitting resolution (Xue et al. 2013; Tallapragada et al. 2014; Zhang et al. 2015, manuscript submitted to *Wea. Forecasting*). These modifications and a description of the various versions of HWRF will be outlined in section 2. Section 3 compares the verifications of track and intensity forecasts from the HWRF versions for retrospective runs from the 2010 and 2011 North

TABLE 1. List of differences between the original operational HWRF (H007) and the experimental and operational triply nested high-resolution HWRF versions, H3GP and H212, respectively. Microphysics scheme is described in Ferrier (1994). POM-TC stands for the Princeton Ocean Model for Tropical Cyclones.

| | H007: original (2007) operational HWRF | H3GP: experimental high-resolution HWRF | H212: operational (2012) high-resolution HWRF |
|---|---|---|---|
| Spatial resolution | 27 km, 77.76° × 77.76° | 27 km, 77.76° × 77.76° | 27 km, 77.76° × 77.76° |
| Horizontal grid | 9 km, 7.2° × 6.0° | 9 km, 10.56° × 10.2° | 9 km, 10.56° × 10.2° |
| | | 3 km, 7.6° × 6.04° | 3 km, 6.16° × 5.44° |
| Vertical levels | 42 hybrid levels, 10 <850 hPa | 42 hybrid levels, 10 <850 hPa | 42 hybrid levels, 10 <850 hPa |
| Time steps | 27 km, 54 s; 9 km, 18 s | 27 km, 45 s; 9 km, 15 s; 3 km, 5 s | 27 km, 45 s; 9 km, 15 s; 3 km, 5 s |
| Vortex initialization | 27–9 km, yes | 27–9 km, yes | 27–9 km, yes |
| | | 3 km, no (downscaling) | 3 km, yes |
| Cycling | Yes (vortex only) | Yes (9-km domain vortex only) | Yes, down to 3 km |
| Ocean coupling (POM-TC) | 27–9 km, yes | 27–9 km, yes | 27–9 km, yes |
| | | 3 km, no (downscaling) | 3 km, no (downscaling) |
| Physics schemes | | | |
| Microphysics | Ferrier | Ferrier | Ferrier |
| Short- and longwave radiation | GFDL | GFDL | GFDL |
| Surface scheme | GFDL (2011 implementation) | GFDL (HR implementation) | GFDL (HR implementation) |
| PBL scheme | GFS | GFS (HR implementation) | GFS (HR implementation) |
| | | $\alpha = 0.25$ | $\alpha = 0.50$ |
| Cumulus parameterization (CP) | SAS | SAS, 27–9 km | SAS, 27–9 km |
| | | No CP, 3 km | No CP, 3 km |
| Land surface | GFDL slab | GFDL slab | GFDL slab |
| Gravity wave drag | Yes, 27 km; no, 9 km | Yes, 27 km; no, 3–93 km | Yes, 27 km; no, 9–3 km |
| Physics call frequency | 27 km, 108 s | 27 km, 3 min | 27 km, 3 min |
| (Note: temperature tendencies are updated every time step) | 9 km, 18 s | 9 km, 3 min | 9 km, 3 min |
| | | 3 km, 30 s | 3 km, 3 min |

Atlantic hurricane seasons. Section 4 discusses causes for some of the deficiencies in the 2012 operational version as well as areas for additional improvements.

## 2. HWRF versions

### a. Pre-2012 operational (27:9 km) version (H007)

The first operational version of HWRF was built upon the foundation of the Weather Research and Forecasting (WRF) Model, a general purpose, multi-institutional mesoscale modeling system (Gopalakrishnan et al. 2010). The key modification within the WRF–Nonhydrostatic Mesoscale Model (WRF-NMM; Janjic 2003) system to effectively address hurricane forecasting was the creation of a moving nest capability (Gopalakrishnan et al. 2006). This HWRF version used an outer coarse mesh with a resolution of 27 km and a movable, inner fine mesh of 9 km (27:9). The 27:9 operational version of HWRF, referred to in this study as H007 [or HOPS in Tallapragada et al. (2014)], was operationally implemented by NWS beginning with the 2007 hurricane season and experienced occasional minor modifications through 2011 (see Table 1 for details about this and the other HWRF versions verified in this study). H007's 9-km inner nest had the potential to improve hurricane intensity forecasts by permitting nonhydrostatic scales of motion within the hurricane inner core and was initially designed to replace the hydrostatic GFDL model (Bender et al. 2007). Additional description of HWRF is found in Gopalakrishnan et al. (2010) and at the Development Testbed Center website (http://www.dtcenter.org/HurrWRF/users/docs/scientific_documents/HWRF_final_2-2_cm.pdf).

### b. Experimental 27:9:3-km version (H3GP)

As a first step toward improving H007, an experimental version of HWRF (HWRFX) was developed at AOML/HRD and designed to run within an idealized research framework (Gopalakrishnan et al. 2011) as well as in forecast mode (X. Zhang et al. 2011). HWRFX was used to develop and test various changes to the model grid resolution, initial conditions, and model physics to better understand the influence of horizontal grid resolution on the dynamics of hurricane vortex intensification in three dimensions (Gopalakrishnan et al. 2011). The fine-grid version of HWRFX used a parent domain and moving nest with resolutions of 9 and 3 km, respectively. The 3-km inner mesh was designed to study the intensity change problem at convection-permitting scales. Details of verifications using HWRFX for a diverse sample of hurricane forecast situations are given in X. Zhang et al. (2011) and Gopalakrishnan et al. (2012).

The impacts of these modifications on the TC forecast (track, intensity, and structure) in HWRFX were used to provide guidance for improvements to the operational

HWRF. However, because of operational computational constraints, it was necessary to develop a triply nested version of HWRF to capitalize on the high resolution of the HWRFX 3-km inner mesh for operational forecasts (Zhang et al. 2015, manuscript submitted to *Wea. Forecasting*). This version (henceforth H3GP) used a parent domain with a horizontal coarse mesh of 27 km and a pair of two-way telescopic moving nests of 9 and 3 km (27:9:3 henceforth; see Table 1).

For H3GP, the cumulus-cloud scheme was switched off in the third (3 km) nest since earlier studies demonstrated that the scheme should not be used for such a high-resolution (HR) nest (Gopalakrishnan et al. 2013). Most of the model physics options used in H3GP were configured as closely as possible to H007. Nevertheless, significant physics improvements related to the high-resolution version of HWRF were required before its eventual transition to operations. In fact, the improved predictions from HWRF are partly attributed to a reconstruction of the surface and boundary layers on the basis of actual hurricane observations. HWRF incorporates the GFDL surface layer parameterization scheme, based on the Monin–Obukhov similarity theory, to provide estimates of the surface layer exchange coefficients, $C_d$ and $C_k$, for the computation of the surface-layer fluxes. The $C_d$ coefficient has produced values consistent with observations for both higher and lower wind speeds but, as a result of uncertainties in the observations, $C_k$ has been left unchanged (Bender et al. 2007). The 27:9 operational HWRF, H007, used the GFDL implementation of the surface layer scheme from 2007 to 2009. Note that starting in 2010, H007 was upgraded to include a function for $C_k$ that produced a value of about $1.3 \times 10^{-3}$, which was consistent with available observations at that time (Haus et al. 2010).

In H007, the GFS boundary layer formulation was used (Hong and Pan 1996). However, the scheme was found to be overly diffusive, especially for high-resolution applications, and this impacted both size and intensity forecasts. Inner-core flight-level data collected by NOAA WP-3D research aircraft at an altitude of about 500 m in category 5 hurricanes Allen (1980) and Hugo (1989) were used as the basis to redesign this scheme for high-resolution hurricane applications (J. Zhang et al. 2011). Eddy diffusivities for mass and moisture were, therefore, reduced to one-fourth of the original value for H3GP. Such a change, which best matched observations, provided a significant improvement in hurricane size prediction (Table 1; Gopalakrishnan et al. 2013).

### c. *The 2012 high-resolution 27:9:3-km operational version (H212)*

Starting with the experimental version (H3GP) developed at HRD as a foundation, EMC, in collaboration with several other NOAA and NOAA/HFIP-funded organizations, incorporated a number of modifications to produce the triply nested (27:9:3) operational version of HWRF (Bernardet et al. 2015). These modifications included improvements to the model computational efficiency, vortex- and large-scale initialization, ocean coupling, and model physics [such as for convection, the surface, and the planetary boundary layer; see Table 1; Tallapragada et al. (2014)]. The new version, referred to here as H212, replaced the previous operational version (H007) starting with the 2012 North Atlantic and east Pacific hurricane seasons.

As with H3GP, for H212 the cumulus parameterization scheme is turned off (i.e., explicit convection is used) in the 3-km domain (Table 1). The GFS shallow convection scheme was used with a slight variation that excludes precipitation from stratocumulus clouds less than 50 mb thick and when the cloud top is below the top of the planetary boundary layer (PBL). In addition, several of the microphysical parameters (e.g., maximum concentrations of large ice crystals, number concentrations of cloud droplets, and snowfall speed) were adjusted to better match the observed properties of stratiform precipitation of midlatitude mesoscale convective systems (Ferrier 1994) with further modifications to optimize the representation of precipitation in tropical cyclones. Under HFIP, work is ongoing to better determine the most physically reasonable values for these parameters in hurricanes. Based on analyses of actual hurricane observations, for the PBL the critical Richardson number was changed from 0.50 to 0.25 and $\alpha$ was set to 0.50.

The various upgrades to H212 were extensively tested and evaluated. Basic statistics for track and intensity forecast verifications for the final version incorporated all of the modifications discussed above for H212 and are shown in Tallapragada et al. (2014). In addition to these standard metrics, Tallapragada et al. (2014) showed how H212 produced major improvements over H007 in the prediction of storm size by verifying the 34-, 50-, and 64-knot (kt; 1 kt = 0.51 m s$^{-1}$) wind radii in different quadrants. It was suggested that these significant improvements to the storm structure were the result of H212's higher resolution and the new vortex initialization method. There were also similar size improvements for HWRFX through the use of a 3-km inner mesh (X. Zhang et al. 2011; Gopalakrishnan et al. 2012) measured by comparing the cumulative distribution function of the radius of maximum wind (at 10 m above the ground) in the model forecasts to HRD's Real-time Hurricane Wind Analysis System (H*WIND; e.g., Powell et al. 1998). In addition, Tallapragada et al. (2014) showed how the improvements incorporated into the H212 version reduced the problem of the erroneous initial spindown and spinup of the model vortex for initially stronger and weaker TCs, respectively [as discussed in Gopalakrishnan et al. (2012)]. The current study compares the verifications of the

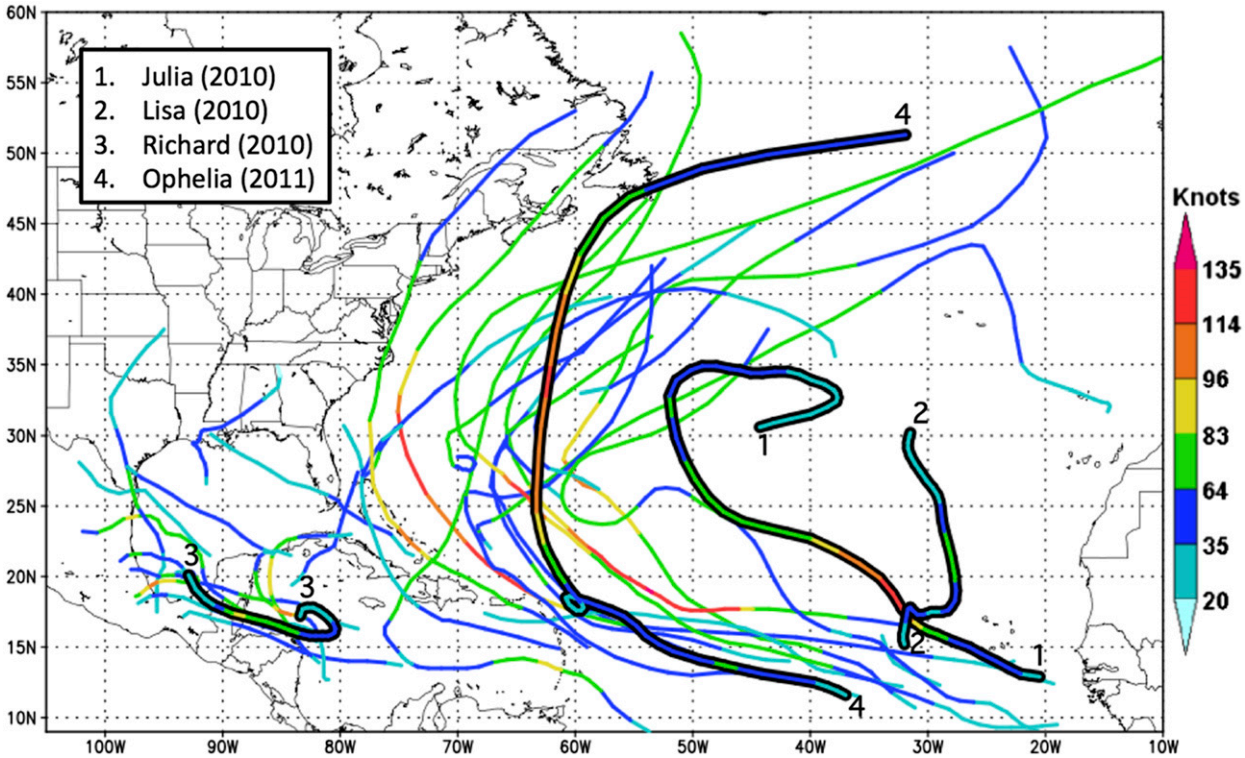


FIG. 1. Actual tracks and intensities (from NHC best-track data) of the 38 tropical cyclones from the 2010–11 North Atlantic hurricane seasons used in this study. Best-track max sustained surface wind speeds (1-min average at 10 m) are color coded (speeds on legend are in knots). Tracks of the four "problem" storms (see text and Fig. 10)—Julia, Lisa, and Richard (2010) and Ophelia (2011)—are shown with thicker track lines and labeled.

H212 hindcast forecasts for the 2010–11 North Atlantic hurricane seasons with forecasts from H007 and H3GP, the triply nested (27:9:3) research version of HWRF.

## 3. Track and intensity verifications

Details regarding the input data for the large-scale analyses in HWRF are described in Tallapragada et al. (2014). It should be noted that the GFS large-scale analyses used in the HWRF runs in this study were not homogeneous between the various versions included in this study.[2] Given the time scale of the evolution of the operational HWRF, it was not feasible to use the same GFS large-scale analysis. However, tests performed using different versions of the GFS large-scale analyses with the same version of HWRF showed minimal impact on the HWRF verification results for track and intensity. Therefore, it is expected that the main differences between the results for the various HWRF versions shown in this section are due to model differences rather than GFS large-scale analysis differences, especially for intensity forecasts.

For TC input data, all of the HWRF versions discussed here used the operational storm parameters provided by NHC in real time. The complete tracks of the storms verified in this study are shown in Fig. 1. A total sample of ~670 cases at 12 h (reduced to ~200 cases at 120 h) from the 2010–11 Atlantic hurricane seasons were used to test the differences in performance between the various versions of HWRF. These cases cover most of the Atlantic basin and include a wide range of initial intensities, intensifying and weakening TCs (including a number of RI events), and track types.

The model forecast results were compared with NHC's postprocessed, best-track storm and intensity data for the verifications. This study followed the same guidelines used by NHC in its official verifications, in that a forecast was verified only if a system was a tropical (or subtropical) cyclone (depression intensity or greater) at the initial time and the verification time (CF2013).

The primary goal of the modifications to H007 (and almost all other operational models) was to improve

[2] H007 and H3GP used the operational version of the GFS analysis for the 2010 and 2011 hurricane seasons. For the 2010 season, H212 also used the operational GFS input. However, for its 2011 retrospective runs, H212 used the hybrid GFS reruns (pre13r) for the period from 20 August through 11 October and pre13h, a slightly different version of the hybrid GFS (Wang et al. 2013), for the rest of the season.

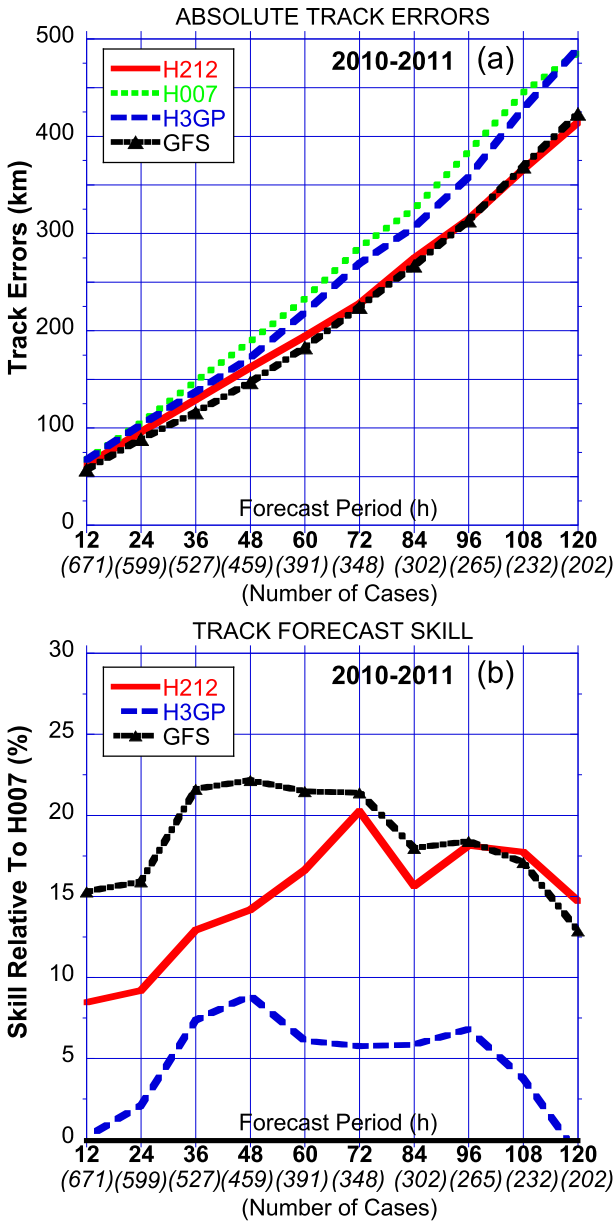


FIG. 2. Track forecast errors for H007, H3GP, and H212, plus GFS for the 38-storm homogeneous sample from the 2010–11 North Atlantic hurricane seasons. (a) Actual errors. (b) Skill of models relative to H007.

operational track and intensity forecasts. Therefore, the results presented in this section focus on the standard metrics of the average absolute track and intensity errors and intensity error bias.

### a. *Track errors*

Figure 2 provides an overview from the 2010–11 retrospective runs of the track error statistics for H007, H3GP, and H212, plus the operational GFS (i.e., the Aviation Model; AVNO). All of the track and intensity statistics are shown for the "late" versions[3] of these models. Figure 2 shows results for the total sample.

Figure 2a shows the actual average track errors. The errors for the complete sample of the models increased approximately linearly with time (Fig. 2a). The lowest errors for the HWRF versions were for H212 at all forecast times [see also Tallapragada et al. (2014)]. The improvement of H212 versus H007 was statistically significant[4] at all forecast times and for H212 versus H3GP at most times. (A listing of the statistical significance results of the comparisons between the average errors for the various models for Figs. 2–7, 10, and 12 is given in Table 2.) The average errors for H212 were comparable to errors from GFS, the model regarded as one of the best for track forecasts, from 72 to 120 h and only slightly higher at earlier forecast times.

Skill plots compare the errors of various models by normalizing (dividing by the baseline errors) the differences between results from a particular model and a selected baseline, where positive skill (given in percent) indicates improvements versus the baseline (CF2013). The standard benchmarks used by NHC and others for track and intensity skill plots are CLIPER5 and DSHIFOR5, respectively, where these models are viewed as no-skill baselines (e.g., Aberson 1998; CF2013). These NHC benchmarks help to measure how easy or difficult storm forecasts are for various storms or seasons. Gopalakrishnan et al. (2012) used DSHIPS as the baseline for intensity skill to compare the results against a higher standard, since DSHIPS is regarded by NHC as one of the most reliable intensity forecast models (CF2013). However, because the main issue addressed in this study and operational implementation has been to produce an improved operational version of

[3] The dynamical models are considered "late" models and use data from the current operational cycle but finish too late to be available to the hurricane specialists in time to provide guidance for their forecast package. "Early" models (e.g., CLIPER5 and DSHIPS) finish early enough for the specialists to use their output for the current operational cycle (CF2013). For track forecasts with the dynamical models, use of the early version versus the late version results shows a consistent reduction in skill of about 3%–5% (not shown here). However, intensity forecast verifications (not shown here) demonstrate no consistent reduction of skill at any forecast time for the early versions versus the late versions.

[4] This study used an a priori significance threshold of 0.10 (i.e., 90% confidence interval) where the statistical significance was determined by using a Student's $t$ test and the sample size was adjusted for 24-h serial correlation (Aberson and DeMaria 1994). Note that for skill plots, the statistical significance still refers to the differences between the average forecast errors.

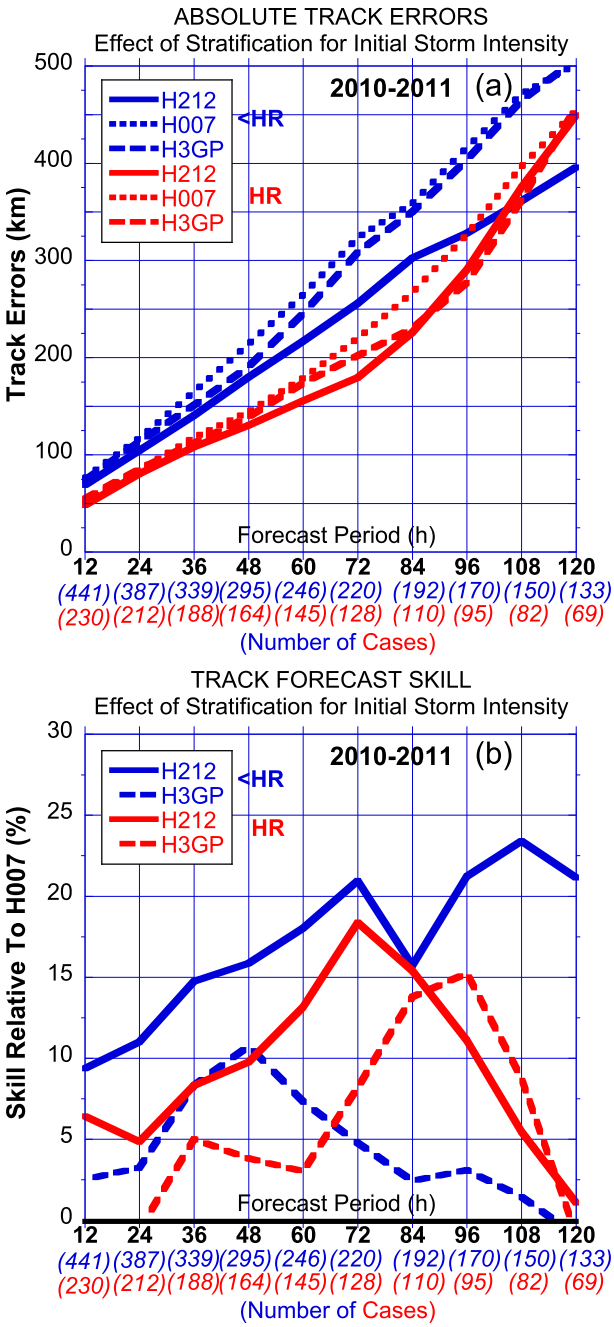


FIG. 3. Track forecast errors for H007, H3GP, and H212 for the 38-storm sample from the 2010–11 North Atlantic hurricane seasons stratified by initial storm intensity (<hurricane intensity, blue; hurricane intensity, red). (a) Actual errors. (b) Skill of the models relative to H007. Color-coded sample sizes are shown below the plot. The samples are only homogeneous within each stratification (i.e., same "color" lines).

HWRF, all skill plots presented here use H007 as the baseline to elucidate these improvements.

Figure 2b shows the track results normalized with respect to H007. The average track errors for H212 showed positive skill (improvement) of 8%–20%. An examination of the frequency of superior performance (FSP)[5] for the track forecasts from these two models (not shown) also showed that these improvements were very consistent, with H212 outperforming H007 between 60% and 65% of the time from 48 to 108 h. Although H3GP showed some skill relative to H007, the additional model modifications incorporated into H212 improved the track forecasts significantly (Table 2). GFS showed the highest overall skill at almost all forecast times, but H212 was extremely close from 72 to 120 h.

Stratification of the forecast samples based on various initial or future conditions has long been used to examine the performance of models in various situations (e.g., Aberson and DeMaria 1994; X. Zhang et al. 2011; Gopalakrishnan et al. 2012). Gopalakrishnan et al. (2012) demonstrated that the HWRF versions performed much better (especially for intensity forecasts) for tropical cyclones with stronger initial intensities. To further understand model behavior, Fig. 3 shows the results stratified for storms (as determined by NHC best-track data) with an initial intensity (maximum sustained surface wind speed) of $<33.4\,\mathrm{m\,s^{-1}}$ (i.e., less than hurricane intensity) and for storms with an initial intensity $\geq 33.4\,\mathrm{m\,s^{-1}}$ (i.e., hurricane intensity), indicated by blue and red lines, respectively. Results for GFS are only shown for the total sample (Fig. 2). Note that although the samples are homogeneous within each stratification (i.e., same "color" lines in Fig. 3), comparisons between the various stratifications must be performed with caution since those different samples are not homogeneous.

For all three versions of HWRF, with the exception of H212 from 108 to 120 h, the average forecast errors were the lowest for cases with initially stronger storms (ISSs; Fig. 3a). This is similar to the results of Gopalakrishnan et al. (2012), who showed various versions of HWRFX performing the best for the ISS cases. Although the samples for the stratifications were not homogeneous between them, these results do suggest that the track forecasts of the HWRF versions in this study were more reliable for initially stronger storms, possibly because the initial vortex was more developed. Weaker TCs often do not have a well-defined vortex center since they are still in the early stages of organization. Occasionally, NHC has had to relocate the operational estimate of the

[5] FSP was designed as a simple measure (in percent) of how often one model produces a better forecast than another (Velden and Goldenberg 1987). This statistic is best utilized in comparing only two models (paired) at a time. One point is given to a model for each superior forecast and 0.5 points are given to both models for a tie.

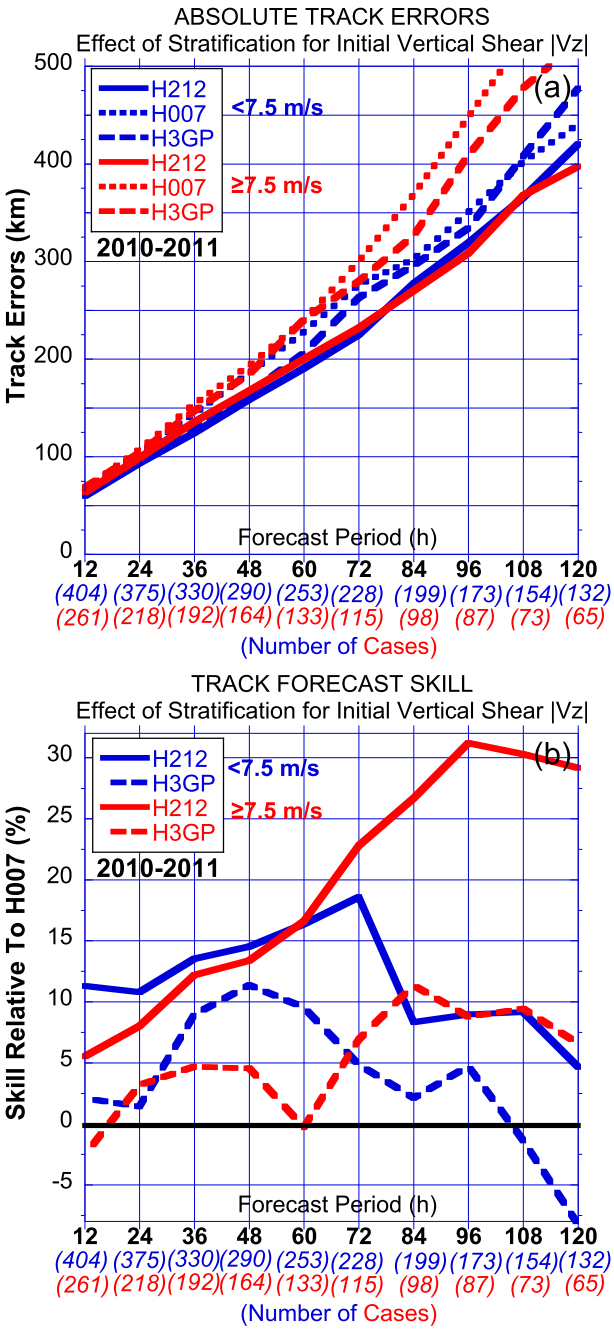


FIG. 4. As in Fig. 3, but for storms stratified by the magnitude of the 200–850-hPa vertical shear $|\mathbf{V}_z|$ over the storm (see text for explanation) at the initial time. Samples for $|\mathbf{V}_z|$ <7.5 and ≥7.5 m s$^{-1}$ are shown in blue and red, respectively.

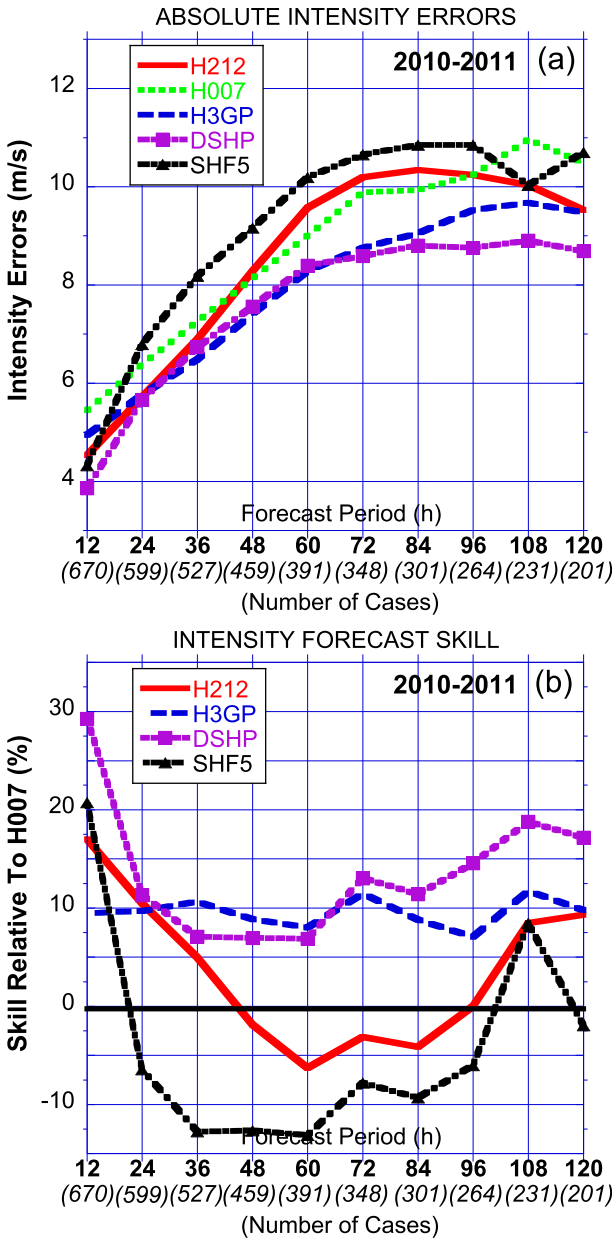


FIG. 5. Absolute intensity forecast errors for H007, H3GP, and H212, plus DSHP and SHF5 for the 38-storm homogeneous sample from the 2010–11 North Atlantic hurricane seasons. (a) Actual errors. (b) Skill of the models relative to H007.

center position of one of these weaker systems as much as about a degree compared to the previous operational position, whereas significant relocation would be a rare event for a TC of hurricane intensity. In addition, the models have a more difficult time following and maintaining the vortex of weaker TCs.

A key to understanding the results for the skill of the average forecast errors of the stratified (for initial intensity) samples (Fig. 3b) is to remember the baseline for each sample (stronger, weaker) consists of the H007 results for that stratified sample (Fig. 3a); that is, there is a different baseline for each group. For instance, in the case where a model does much better for initially weaker storms (IWSs), if the baseline, H007, also does much

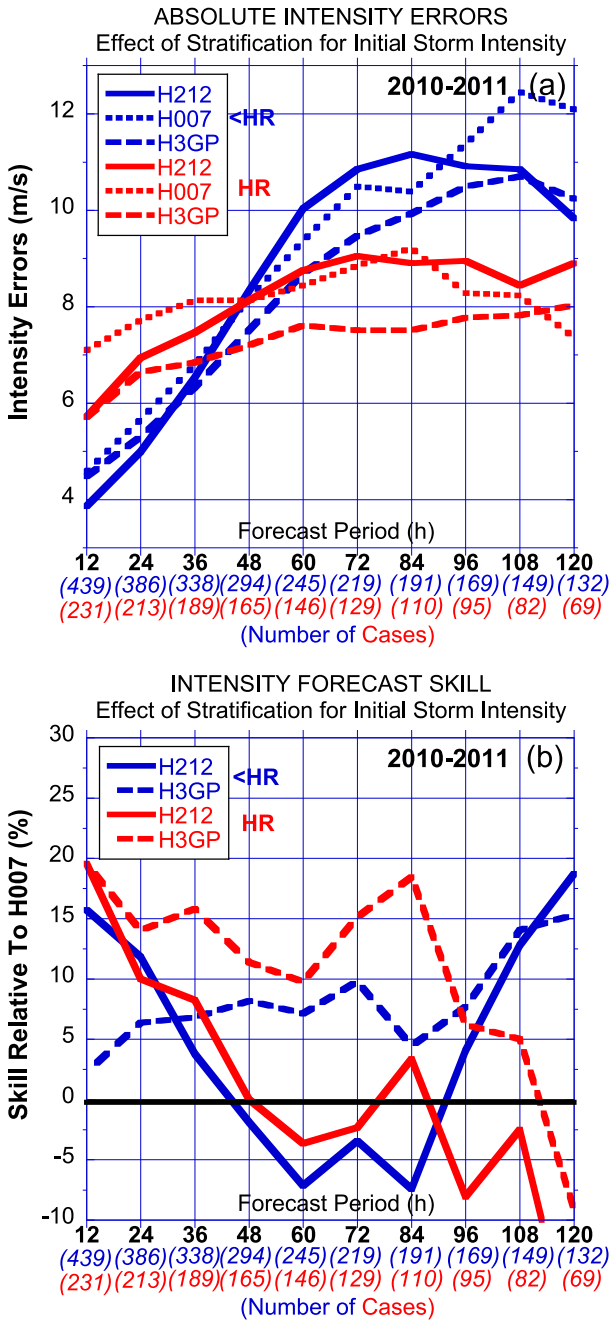


FIG. 6. As in Fig. 5, but for storms stratified by initial storm intensity (<hurricane intensity, blue; hurricane intensity, red). See Fig. 3 for an explanation of color-coded sample sizes.

better, the skill may or may not increase depending on the relative improvement of the two models.

As with the results for the complete sample (Fig. 2b), the skill of H212 versus H007 was large for all forecast times. Although the smallest errors for H212 were for the ISS cases (Fig. 3a), the greatest improvement (skill) versus H007 was for the IWS cases (Fig. 3b). For the IWS cases, the improvements for H212 versus H007 and H3GP were significant for most of the forecast times whereas only a few of these improvements were statistically significant with the ISS sample (Table 2). Note that since almost two-thirds of the overall sample size consisted of IWS cases, these results were closer to the results from the overall sample. H3GP also outperformed H007 for the full and stratified samples at virtually all forecast times, although the improvements were only significant at a few forecast times.

Although there are numerous other parameters that can be used to stratify the forecast samples (e.g., Aberson and DeMaria 1994), the vertical shear of the horizontal wind $\mathbf{V}_z$ has long been recognized as a critical factor affecting hurricanes, especially intensity (e.g., Kaplan and DeMaria 2003). Hence, $\mathbf{V}_z$ was selected as the other parameter used for stratification of the verifications in this study. Values that were used to determine the magnitude of the deep-layer (200 − 850 hPa) $\mathbf{V}_z$ (i.e., $|\mathbf{V}_z|$) were obtained from the SHIPS database (M. DeMaria 2012, personal communication). The parameter from the database used in this study is SHRD, defined as $|\mathbf{V}_z|$ from the GFS analysis fields for the initial forecast time,[6] averaged for 0–500 km relative to the 850-hPa center where the average is calculated after the storm vortex is removed (Kaplan et al. 2010).

Tropical cyclone response (mainly intensification) to $|\mathbf{V}_z|$ can be divided into basic high- and low-$|\mathbf{V}_z|$ ranges (e.g., Zhao et al. 2006; Kaplan et al. 2010; Paterson et al. 2005). The stratification based on $|\mathbf{V}_z|$ for this study was divided between initially "high-" and "low-" shear cases, which had values of $\geq 7.5$ and $<7.5\,\mathrm{m\,s^{-1}}$, respectively. Figure 4 shows the average track errors stratified for initial vertical shear. One of the main features of the actual average track errors (Fig. 4a) is that the H212 results showed very little difference for initially high-shear (IHS; red lines) or initially low-shear (ILS; blue lines) samples. H3GP was somewhat worse for the IHS sample while H007 showed a substantial degradation for the IHS sample past 72 h. Therefore, the skill of H212 versus H007 (Fig. 4b), although positive for all samples, was largest for the IHS cases. The improvements in H212 versus H007 were

[6] Obviously $|\mathbf{V}_z|$ changes with time in the actual and in the model forecast fields during the forecast period. However, to use $|\mathbf{V}_z|$ at the actual verification time is somewhat problematic for several reasons (e.g., the difference in the location of the actual storm center and the predicted storm center). Therefore, the stratifications in this study use only the $|\mathbf{V}_z|$ values at the initial time.

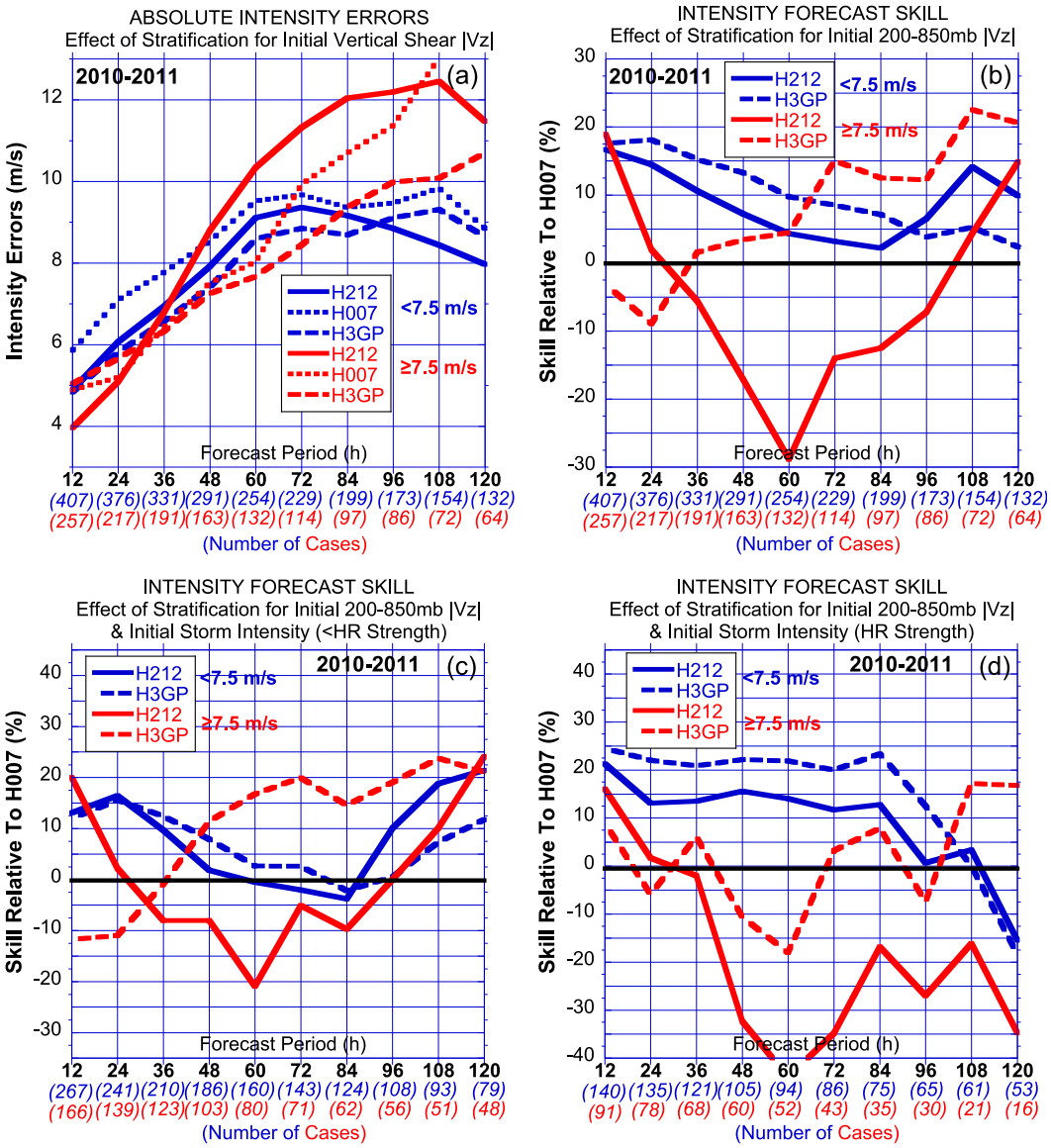


FIG. 7. Absolute intensity forecast errors for H007, H3GP, and H212 for the 38-storm sample from the 2010–11 North Atlantic hurricane seasons stratified by the magnitude of the 200–850-hPa vertical shear $|\mathbf{V}_z|$ over the storm (see text for explanation) at the initial time. Samples for $|\mathbf{V}_z| < 7.5$ and $\geq 7.5$ m s$^{-1}$ are shown in blue and red, respectively. (a) Actual errors [note that H007 for high shear (red) goes to 13.5 m s$^{-1}$ at 120 h]. (b) Skill of models relative to H007. (c),(d) As in (b), but for cases where the initial storm intensity is less than the hurricane intensity and hurricane intensity, respectively. See Fig. 3 for an explanation of color-coded sample sizes.

significant for most forecast times for the IHS and ILS cases (Table 2).

### b. *Intensity errors*

Figure 5 shows results for the complete 2010–11 retrospective runs sample of the average absolute intensity error statistics for H007, H3GP, and H212, plus DSHIPS (i.e., DSHP) and SHIFOR5 (i.e., SHF5).[7] Figure 5a shows the actual average absolute intensity errors. The differences are more easily seen in the skill plot (Fig. 5b),

[7] As in footnote 3, justification for showing early (SHIFOR5 and DSHIPS) and late (H007, H212, and H3GP) results together for intensity.

TABLE 2. Statistical significance of the differences between average errors (of two models) at 12–120-h forecast times for the results shown in Figs. 2–7, 10, and 12. Differences that are statistically significant at an a priori significance threshold of 0.10 (i.e., 90% confidence interval) are indicated with a cross. See text for additional details on the calculation of statistical significance.

| | 12 | 24 | 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|
| Fig. 2 | | | | | | | | | | |
| H212:H007 | X | X | X | X | X | X | X | X | X | X |
| H212:H3GP | X | X | | | X | X | X | X | X | X |
| H212:GFS | X | X | X | X | | | | | | |
| H3GP:H007 | | | X | X | | | | | | |
| Fig. 3 (<HR) | | | | | | | | | | |
| H212:H007 | X | X | X | X | X | X | X | X | X | X |
| H212:H3GP | X | X | X | | X | X | X | X | X | X |
| H3GP:H007 | | | X | X | X | | | | | |
| Fig. 3 (HR) | | | | | | | | | | |
| H212:H007 | | | | | X | X | X | | | |
| H212:H3GP | X | | | | | | | | | |
| H3GP:H007 | | | | | | | | X | | |
| Fig. 4 (low shear) | | | | | | | | | | |
| H212:H007 | X | X | X | X | X | X | | | | |
| H212:H3GP | X | X | | | | X | | | | X |
| H3GP:H007 | | | X | X | X | | | | | |
| Fig. 4 (high shear) | | | | | | | | | | |
| H212:H007 | | X | X | X | X | X | X | X | X | X |
| H212:H3GP | X | | X | X | X | X | X | X | X | X |
| H3GP:H007 | | | | | | | X | | X | |
| Fig. 5 | | | | | | | | | | |
| H212:H007 | X | X | | | | | | | | |
| H212:H3GP | | | | X | X | X | X | | | |
| H3GP:H007 | X | X | X | X | | X | | X | X | X |
| H212:DSHP | X | | | | X | X | X | X | | |
| H3GP:DSHP | X | | | | | | | | | |
| H007:DSHP | X | X | | | | X | X | X | | |
| H212:SHF5 | | X | X | X | | | | | | |
| H3GP:SHF5 | X | X | X | X | X | X | X | X | | |
| H007:SHF5 | X | | X | X | X | | | | | |
| DHSP:SHF5 | X | X | X | X | X | X | X | X | | X |
| Fig. 6 (<HR) | | | | | | | | | | |
| H212:H007 | X | X | | | | | | | X | X |
| H212:H3GP | X | | | X | X | X | | | | |
| H3GP:H007 | | | | | | | | | X | X |
| Fig. 6 (HR) | | | | | | | | | | |
| H212:H007 | X | | | | | | | | | |
| H212:H3GP | | | | | | X | | | | |
| H3GP:H007 | X | X | X | | | X | X | | | |
| Figs. 7a,b (low shear) | | | | | | | | | | |
| H212:H007 | X | X | X | | | | | | X | |
| H212:H3GP | | | | | | | | | | |
| H3GP:H007 | X | X | X | X | X | | | | | |
| Figs. 7a,b (high shear) | | | | | | | | | | |
| H212:H007 | X | | | X | X | X | X | | | |
| H212:H3GP | X | | | X | X | X | X | X | X | |
| H3GP:H007 | | | | | | X | X | X | X | X |
| Fig. 7c (<HR, low shear) | | | | | | | | | | |
| H212:H007 | X | X | | | | | | | X | X |
| H212:H3GP | | | | | | | | | | |
| H3GP:H007 | X | X | X | | | | | | | |
| Fig. 7c (<HR, high shear) | | | | | | | | | | |
| H212:H007 | X | | | | X | | | | | X |
| H212:H3GP | X | | | X | X | X | X | X | X | X |
| H3GP:H007 | | | | | X | X | X | X | X | X |

TABLE 2. (*Continued*)

| | 12 | 24 | 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|
| Fig. 7d (HR, low shear) | | | | | | | | | | |
| H212:H007 | X | | | | | | | | | |
| H212:H3GP | | | | | | | | | | |
| H3GP:H007 | X | X | X | X | X | X | X | | | |
| Fig. 7d (HR, high shear) | | | | | | | | | | |
| H212:H007 | | | | X | X | X | | X | | |
| H212:H3GP | | | | X | | X | X | | X | X |
| H3GP:H007 | | | | | | | | | | |
| Fig. 10 (without four TCs) | | | | | | | | | | |
| H212:H007 | X | X | X | | | | | | X | |
| H212:H3GP | X | | | | | | | | | |
| H3GP:H007 | X | X | X | X | X | X | X | | X | X |
| Fig. 12a (track with H213) | | | | | | | | | | |
| H213:H007 | X | X | X | X | X | X | X | X | X | X |
| H213:H212 | | | X | X | X | | | | X | |
| H213:H3GP | X | X | X | X | X | X | X | X | X | X |
| H213:GFS | | | | | | | | | X | |
| Fig. 12b (intensity with H213) | | | | | | | | | | |
| H213:H007 | X | | | X | X | X | X | X | X | X |
| H213:H212 | | | | X | X | X | X | X | X | X |
| H213:H3GP | X | | | | | | | X | X | X |
| H213:DSHP | X | | | X | X | | | | X | X |
| H213:SHF5 | | X | X | X | X | X | X | X | X | X |

especially for the shorter forecast times.[8] With the track forecasts, H212 was the superior HWRF version, but the results for intensity were mixed [see also Tallapragada et al. (2014)]. The lowest errors for the HWRF versions were for H3GP at all forecast times except for 12 and 120 h when H212 had the lowest errors. The differences between H3GP and H007 were significant at almost all forecast times (Table 2). The H3GP results were significantly better than those for H212 in the mid-forecast times. Although H007 was more skillful than H212 from 48 to 84 h, the differences were not significant. H212 was the best of the HWRF versions for the earlier forecast times, probably as the result of its improved initialization scheme producing less of a problem in the initial spinup of the TCs. H3GP further demonstrated its skill for intensity forecasts by being close to DSHP at almost all forecast times. DSHP errors were lower than the H007 and H212 errors at all forecast times, and half of those differences were significant. However, H212 and H007 both produced lower average errors for almost all forecast times than SHF5, the model used as the baseline by NHC for its intensity verifications.

Figure 5b shows the skill of the various models versus H007. H3GP, the best HWRF version in this study for intensity forecasts, showed skill of about 10%. An examination of FSP for the intensity forecasts of H3GP versus H007 (not shown) showed that these improvements were fairly consistent, with H3GP outperforming H007 between 53% and 62% at all times except for 60 h. FSP for H212 versus H007 was only above 50% at 12–36 and 120 h and fell to as low as 43% at 60 h. FSP for H212 versus H3GP was <50% for all times except for 12 h. H3GP was clearly the superior model for intensity forecasts of the HWRF suite in this study and comparable to a respected intensity forecast model like DSHP.

Figure 6 shows the results stratified by initial storm intensity, IWS and ISS (blue and red lines, respectively), for H007, H3GP, and H212. As with the track forecasts, before looking at skill results, it is important to observe the actual average errors, especially when examining the skill of the stratified samples (Fig. 6a). For all three HWRF versions, the average absolute intensity errors

[8] Note that statistical models (e.g., DSHP and SHF5) start off with the real-time initial intensity and position, whereas dynamical models (e.g., the HWRF versions) usually have track and intensity errors at the early forecast times because of the model initialization/spinup. "Early" (interpolated) versions of a dynamical model (e.g., HWRF) have better results (versus the "late" version of the same model) for the earliest forecast times because of use of the interpolation scheme that employs the 6–126-h forecasts from the previous cycle as the 0–120-h forecasts for the current cycle. The scheme adjusts model-derived track and intensity values at the 6-h forecast time from the previous cycle to match the 0-h operational initial values. Values for some or all of the other model forecast times are also adjusted depending on the interpolation scheme being used (CF2013).

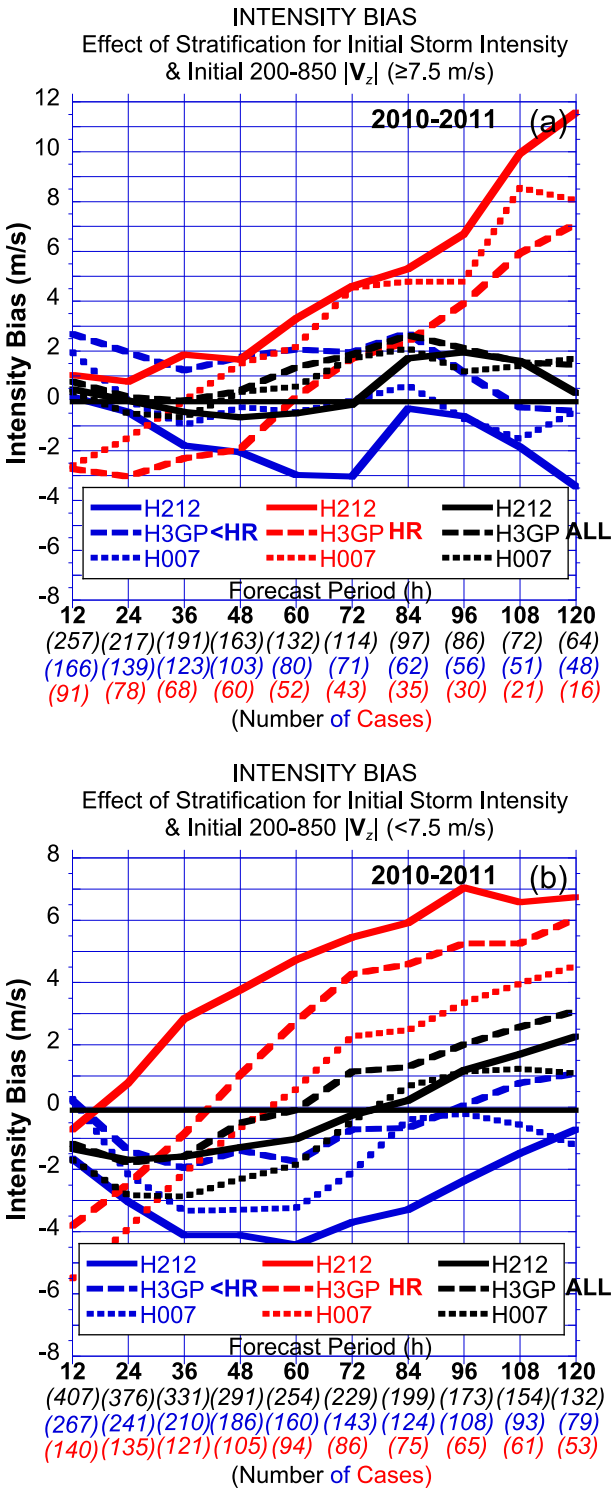


FIG. 8. Average bias of intensity forecast errors for H007, H3GP, and H212 for the 38-storm sample from the 2010–11 North Atlantic hurricane seasons stratified by initial storm intensity (all cases, black; < hurricane intensity, blue; and hurricane intensity, red) for cases where the magnitudes of the 200–850-hPa vertical shear $|\mathbf{V}_z|$ over the storm (see text for explanation) at the initial time were (a) ≥7.5 and (b) <7.5 m s$^{-1}$. Note the different scales for the bias between (a) and (b). See Fig. 3 for an explanation of color-coded sample sizes.

were lower for the IWS sample from 12 to 36 h and lower for the ISS sample from 48 to 120 h. The differences between the ISS and IWS results at the earlier and later forecast times were greatest for H007. The smallest differences were for H212 (except for 108–120 h). This is similar to the results of Gopalakrishnan et al. (2012), where the HWRF versions showed much lower skill (versus DSHP) for the IWS cases. Note that because the baseline used here (H007) had a large variation between average errors for IWS versus ISS cases, the skill shown in Fig. 6b has to be treated carefully. For the ISS cases, H3GP showed skill versus H007 (except at 120 h) but H212 was skillful for only the earlier forecast times and marginally at 84 h. For the ISS cases, H3GP performed better than H212 at all forecast times except at 12 h. Results for the IWS cases were similar. Of the HWRF versions for IWS cases, H3GP was still the most skillful overall and was better than H007 at all forecast times, but the differences were only significant for 108–120 h (Table 2). H3GP was also more skillful than H212 for most forecast times. H212 was skillful versus H007 at the earlier and later forecast times. H212 showed negative skill versus H007 at all other times. Once again, H3GP trumped the other HWRF versions in this study for both ISS and IWS.

Figure 7a shows the actual average absolute intensity errors for H007, H3GP, and H212 stratified by initial vertical shear, and Fig. 7b shows the skill of H3GP and H212 versus H007 for the stratified sample. It should be kept in mind that the ILS cases composed about two-thirds of the overall sample (i.e., nearly twice the number of cases in the IHS sample). In looking at the differences in performance of each model for the IHS (red lines) versus ILS (blue lines) cases in Fig. 7a, the average absolute intensity errors were lower for the IHS sample for all three models during the earlier forecast times (except for H3GP at 12 h) and then switched to errors being lower for the ILS cases through 120 h. The differences between the IHS and ILS results at the earlier times for each model were generally the greatest for H007 but, for the later forecast times when the average errors for the three versions were lower for the ILS cases, the differences were the greatest for H212 (except at 120 h when the IHS − ILS difference was slightly greater for H007). It is striking, however, that the performance of H3GP was very close for the IHS and ILS cases, with almost no difference out to 48 h, demonstrating a somewhat stable performance regardless of the vertical wind shear conditions. This is combined with the fact that H3GP showed the best overall performance for the intensity forecasts.

Looking at the comparison of vertical shear stratification between the three HWRF versions in actual errors and skill (versus H007) in Figs. 7a and 7b, respectively,

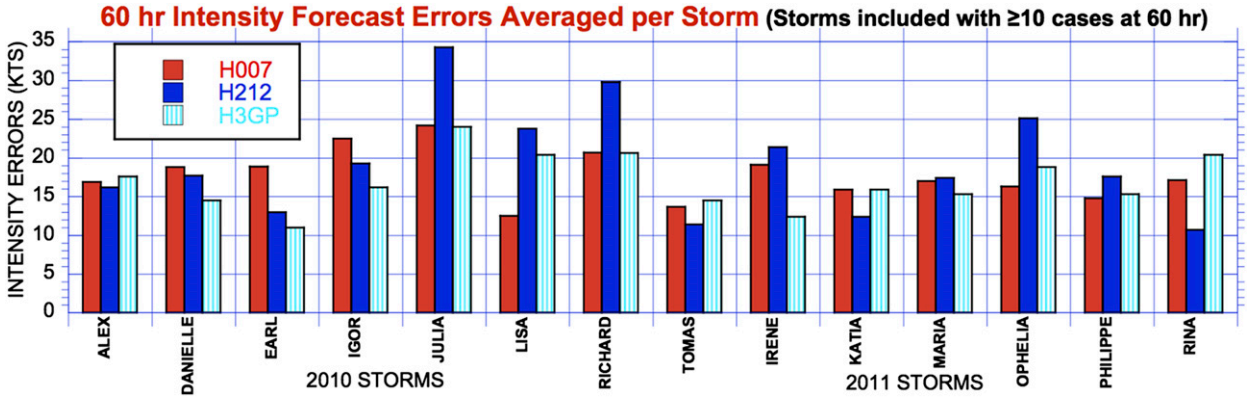


FIG. 9. Average intensity errors at 60 h for each storm for H007, H3GP, and H212 for the 2010–11 seasons. Only storms verifying for ≥10 cases at 60 h are shown.

H3GP showed the best overall performance for both the IHS and ILS cases, with positive skill except for 12 and 24 h for the IHS sample. For the ILS cases, H3GP showed positive skill for all forecast times (significant from 12 to 60 h), and higher skill than H212 for most forecast times. H212 also showed positive skill at all forecast times for the ILS sample. For the ILS sample, H212 was roughly equivalent to H3GP, although H3GP had a slightly better FSP result (not shown) versus H007.

The results were very different for the IHS cases, and this sample highlights the main weakness of H212 for intensity forecasts. The high average errors (and low skill) of H212 from 36 to 96 h for the IHS sample are apparent in Figs. 7a and 7b versus both H007 and H3GP. The negative skill for H212 was significant from 48 to 84 h. Note that H007 also had rather high errors for most of the forecast times for the IHS cases, and this affects the skill relative to H007. H3GP was superior to both H007 and H212 (except at 12 and 24 h), and most of the differences were significant (Table 2). Although the IHS cases composed only about one-third of the sample, the relatively poor performance of H212 in the mid-forecast times versus H007 and especially versus H3GP caused the results from the full sample (Fig. 5) to be poor in the mid-forecast times.

To better understand the nature of the relatively poor performance of H212 in the IHS cases, Figs. 7c and 7d show skill (versus H007) for a doubly stratified sample (by initial $|\mathbf{V}_z|$ and initial storm intensity); Figs. 7c and 7d show results stratified for $|\mathbf{V}_z|$ for the IWS and ISS cases, respectively. Once again, caution must be exercised in interpreting these results since the double stratification reduced the sample size even further. For IWS (Fig. 7c) with ILS, H212 and H3GP had comparable skill (i.e., none of the differences were significant) and were better than H007 at most forecast times. However, with IWS and IHS, H3GP was the best version (except from 12 to 36 h), while H212 showed negative skill from 36 to 84 h, and lower skill than H3GP at most times. For ISS (Fig. 7d) with ILS, once again H212 and H3GP had comparable skill, although H3GP is 5%–10% better at most forecast times. Both were skillful versus H007 at all forecast times, except for 120 h, and most of the differences were significant for H3GP versus H007. For ISS and IHS, the skill for H212 was extremely negative starting at 48 h. For this sample, although the H212 intensities were reasonable at the beginning, they became degraded (versus H007 and H3GP) with time, and the H212 results were significantly worse than those of H007 and H3GP for most of the forecast times starting at 48 h. H3GP was close to H007 for the ISS and IHS cases, seesawing from positive to

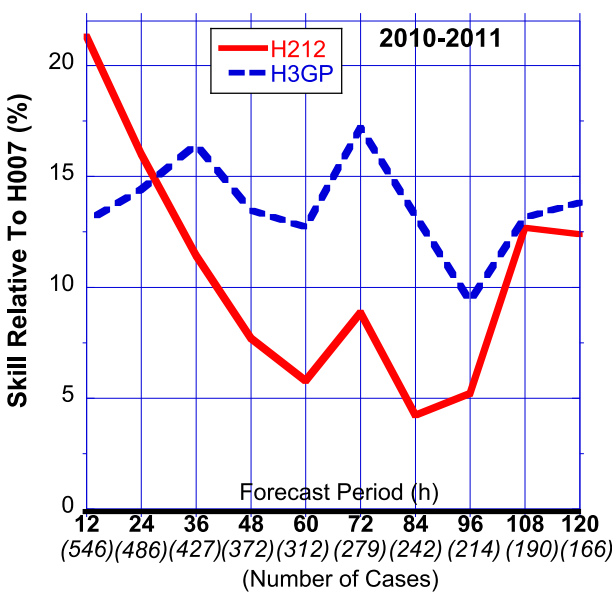


FIG. 10. Skill of the average absolute intensity forecast errors relative to H007 from H3GP and H212 for the 2010–11 North Atlantic hurricane seasons. The sample is the same as in Fig. 5, but without the four problem storms (see text and Fig. 1): Julia, Lisa, and Richard (2010) and Ophelia (2011).

negative skill from 12 to 120 h. It should be noted, however, that the ISS and IHS sample was the smallest of the four double stratifications in Figs. 7c and 7d.

Figure 7 shows that the greatest deficiency in the H212 intensity forecasts was for IHS, especially for ISS cases. Figure 8 looks at the bias of the intensity forecasts for the same stratifications as in Figs. 7c and 7d. For ILS (Fig. 8b), all HWRF versions showed an increasingly positive bias for ISS throughout most of the forecast time, indicating that for lower shear, the models were allowing for overintensification of ISS with time. For the IWS cases, all three versions had an increasingly negative bias during the first ~36 h, but then a positive trend until all were near zero bias at the later forecast times. After 60 h, the bias for H212 showed a positive trend until 120 h with a similar slope for ISS. Because of the usually opposite signs of the biases for the ISS and IWS cases, the total sample for ILS (i.e., combining ISS and IWS) was fairly neutral throughout.

Note that in Fig. 7 the worst performance of H212 was for the IHS cases. One would suspect that there would be a noticeable bias. However, for the overall sample in Fig. 8a, the bias was close to zero for all forecast times, but this was a case of two opposing trends canceling one another in the overall sample. It is only by examining the stratification for ISS and IWS that the picture becomes clear. For the H212 results, the bias started near zero for both IWS and ISS samples, but then trended negative for IWS and trended strongly positive for ISS, climbing to $\sim 12\,\mathrm{m\,s^{-1}}$ at 120 h. Thus, when initial shear was high, on average, H212 allowed the ISS to overintensify while over weakening the IWS; that is, the higher shear did not sufficiently dampen intensification for ISS, yet the shear had an overly suppressing impact for IWS. The H212 ISS cases had the largest bias. H007 and H3GP showed very little bias, except for the ISS cases when they also had a strong positive trend.

There are numerous ways of examining intensity and track verifications. This study has already presented the errors stratified by initial storm intensity and initial environmental vertical wind shear. Figure 9 shows a glimpse of the intensity results storm by storm for the 60-h forecast time. This time was chosen because it exhibited some of the worst results for H212 intensity errors (skill) versus H007 and H3GP (Figs. 5b, 6b, and 7b–d). In fact, the overall actual absolute intensity errors increased linearly through about 60 h and then leveled off, reaching the highest value at 84 h. Figure 9 shows the average absolute intensity errors for only the storms with at least 10 cases since storms with fewer cases have little impact on the overall results. Of the storms in the sample for this study, 14 had no cases at 60 h, and the majority of the storms not included in the figure had only between one and three cases. In comparing H212 to H007, H212 was superior for 7 out of the 14 storms, marginally worse for 3 storms, but significantly worse for 4 of the storms (i.e., Julia, Lisa, and Richard from 2010 and Ophelia from 2011; see Fig. 1 for the tracks of these four storms). As for H212 versus H3GP, H212 was only better for four storms, marginally worse for six storms, and significantly worse for four storms (three of them were the same as in comparison with H007). In fact the FSP for H212 versus H007 for these four storms was less than ~45% for all forecast times, dropping as low as 25% at 60 h; that is, the intensity errors for H212 were worse than for H007 in three out of four cases at 60 h. At 60–72 h, the average absolute intensity errors for the four-storm sample were about twice as large as the average errors for the rest of the storms. Two-thirds of the cases for the four storms were from Julia and Ophelia, both of which became category 4 major hurricanes and had long periods (>48 h) of RI. Lisa also went through a 24-h RI period, and Richard had a 24-h period that was close to RI. These cases highlight the need for further study of the challenge of more reliably predicting RI events with HWRF. Further comments on possible reasons for the relatively poor performance of H212 for these four storms are presented in section 4. At NHC, hurricane specialists make note of poor model performance for particular storms. Figure 10 shows the intensity error skill with the four "problem" storms removed. The sample sizes for the various forecast times were only reduced by <16% (at 60 h). With the revised sample, H212 had positive skill versus H007 for all forecast times, and several of these differences were significant. Although H3GP still had higher skill than H212 at all forecast times, except for 12 and 24 h, none of these differences were significant except at 12 h, when H212 was actually better.

## 4. Conclusions and further improvements to the operational HWRF

Results from retrospective runs for the 2010–11 North Atlantic hurricane seasons with H212, the first operationally implemented, triply nested (27:9:3 km) version of HWRF, were examined and compared with results from H007, the previous doubly nested (27:9 km) operational version of HWRF and H3GP, a triply nested research version of HWRF. The verifications from these three versions of HWRF were also compared to a few of the other operational models such as DSHP, SHF5, and GFS (also known as the Aviation Model; AVNO). The purpose of these verifications was to document the improved performance of H212, the version operationally implemented in 2012, while looking for deficiencies that might suggest additional areas for improvement. Since the primary focus of HFIP has been to reduce hurricane track and intensity forecast errors, the

standard metrics of track and intensity errors were examined. Since the main goal was to improve upon the previous operational version, skill was calculated versus H007. A closer look into model performance was also accomplished by stratifying the samples by initial storm intensity and environmental vertical shear of the horizontal wind.

### a. Track forecasts

As demonstrated in this study, for the case of the 2012 operational version versus the previous operational version of HWRF (i.e., H212 versus H007), there were substantial, statistically significant reductions in track forecast errors at all forecast times, with the skill of H212 versus H007 increasing with time through 72 h. The improvement in track forecasts for H212 made the errors from 60 to 120 h comparable to those of the GFS model, which is generally regarded as one of the best models for track forecasts. The highest skill of H212 versus H007 was realized for the initially weaker (less than hurricane intensity) storm cases (which comprise about two-thirds of the sample) with >15% skill from 36 to 120 h, and skill as high as 23% at 108 h. Skill was still positive for initially stronger storms (hurricane intensity), and these cases had the lowest actual average errors, confirming previous studies that HWRF tends to perform better in general with initially stronger storms. Although H3GP had shown modest improvement in the track forecasts versus H007, further statistically significant improvements were demonstrated by H212 versus H3GP. Note that H007, H3GP, and H212 all had lower average track forecast errors for the initially stronger storm cases versus the initially weaker storm cases. Both H007 and H3GP had larger average errors for the initially high versus initially low vertical shear cases whereas the H212 average track forecast errors showed almost no sensitivity to initial vertical shear and had the greatest skill versus H007 and H3GP (past about 60 h) for the initially high-shear cases. Therefore, the higher skill for H212 with the initially high-shear cases is partly because H007 (the baseline for the skill plots) had the worst average track errors for those cases and partly because the modifications incorporated into H212 virtually removed the model's sensitivity to high shear with regard to track forecasts. Some possible reasons that the higher shear did not degrade the track forecasts with H212 are that the large-scale flow with H212 was not contaminated as much during the initialization because of the superior vortex.

Both versions of the triply nested (27:9:3) HWRF (H3GP and H212) saw improvements in their track forecasts, with the greatest improvement being for H212. Track improvements for the transition from H007 to H3GP are possibly primarily due to the increased resolution (from 27:9 to 27:9:3), as well as the physics package, which was modified to be consistent with the higher resolution. The better physics (which was added with H3GP and H212) combined with the better vortex initialization (in H212) allowed for more realistic multiscale interaction with the large-scale flow. Also, the intensity of the model vortex is often associated with the vertical extent of the storm, which impacts the depth of the large-scale flow that is steering the storm; that is, a more intense (weaker) storm will be steered by a deeper (shallower) layer of the atmosphere. Therefore, a more accurate intensity forecast, as well as a better forecast of the overall storm structure, could be one of the contributing factors to the improved track forecasts of H3GP and H212. In addition, track improvements from H3GP to H212 are the product of numerous additional modifications such as substantial improvements to the vortex initialization scheme, as well as possibly a small impact from the use of the newer version of the GFS large-scale analysis. A key factor in the track forecasts is the initial and forecast large-scale fields. The vortex initialization procedure can contaminate the large-scale field in the vicinity of the storm vortex that can then degrade the track forecasts. It is likely that changes to the initialization in H212 reduced this degradation of the large-scale field, thus improving the track forecasts.

### b. Intensity forecasts

Similar to what had been shown in earlier studies, all of the HWRF versions in this study performed better past 36 h at forecasting intensity of initially stronger storms. For the overall sample, H3GP exhibited a significant reduction in intensity forecast errors versus H007. However, although H212 showed some improvements in intensity forecasts versus H007, these were not significant at most forecast times, and the H212 intensity forecasts were actually degraded versus H007 in the midrange times. Although the H212 intensity forecasts started off better than both H007 and H3GP, possibly from the better vortex initialization discussed earlier, this improvement was quickly lost and the model showed the poorest skill from 48 to 84 h, while H3GP excelled at almost all times. H212 had a "head start" with better initial intensity, but the forecast integrations lost all of that gain. It was shown that much of these problems for H212 versus H007 were the results of several "problem" storms, and when these were removed (only 14% and 17% of the overall sample at 12 and 120 h, respectively), H212 showed positive skill at all forecast times. However, even with these storms removed, H3GP still outperformed H212. Therefore, much of the improvement to intensity forecasts realized in H3GP, the research 27:9:3 version, was lost in H212, the 27:9:3 operational version in use in 2012. Stratification of the sample by initial vertical shear elucidated the fact that the greatest problem for H212 was from initially high-shear cases.

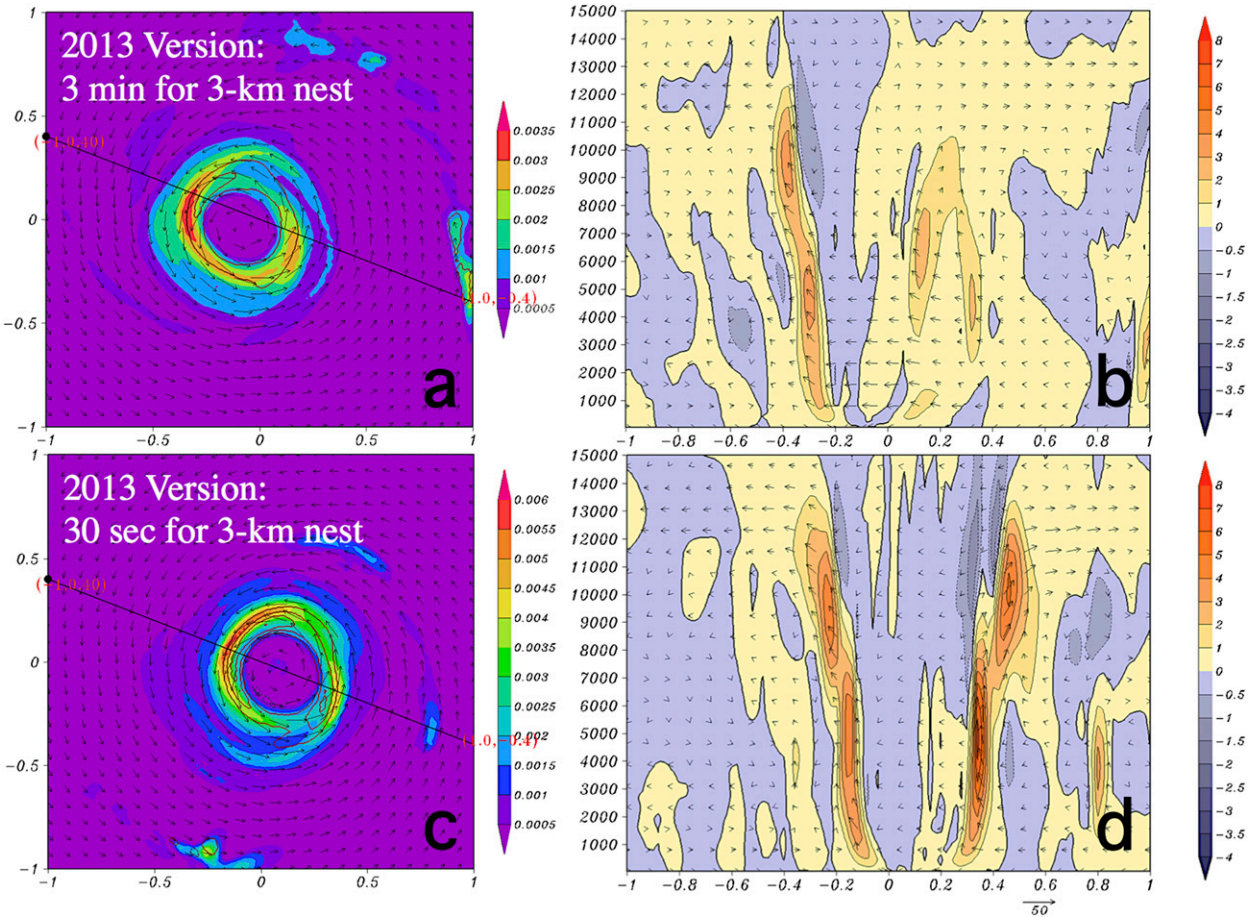


FIG. 11. Impact of frequency of physics calls on HWRF. All figures are for model-generated 92-h forecast fields using the idealized simulations with the 2013 operational version of HWRF, but (a),(b) with the physics call time in the innermost (3 km) nest reduced to only every 3 min (as in the 2012 version, H212) and (c),(d) at the actual frequency (every 30 s) for H3GP (see Table 1) and the operational 2013 HWRF. (a),(c) The vertical velocity field at 5-km height and (b),(d) the $X$–$Z$ cross section obtained across a vertical plane passing through the eyewall region indicated by the black line shown in (a) and (c). Note that the updraft–downdraft structure is much improved with the higher-frequency physics calls.

Intensity forecast results from H3GP showed very little sensitivity to initial vertical shear but, for H212, the actual errors (and skill versus H007) were much worse for the initially high-shear cases. Stratifying the sample by initial intensity and initial vertical shear showed that H212 had the worst skill (negative from 36 to 120 h) for the initially high-shear and initially stronger storms. An analysis of the intensity bias for this doubly stratified sample showed that, on average, in high-shear cases H212 allowed the initially stronger storms to overintensify (yet slightly over weaken the initially weaker storms).

To recover the intensity forecast improvements realized by H3GP, it is critical to uncover the reasons for the relatively poor intensity forecasts of H212 (compared to H3GP, its predecessor), especially in these high-shear cases. There were numerous modifications to H3GP (to create H212) that could have produced the degradation in intensity forecasts. However, it was discovered that as a result of computational restraints, EMC made the decision not to fully utilize the benefit of the new 3-km inner mesh and did not incorporate sufficient physics calls. Other recent HFIP-supported experiments performed at HRD using the idealized framework of HWRF indicate the importance of a sufficient frequency of these physics calls. The tests showed that although the physics tendencies of heating are stored between time steps, it is critical to incorporate these tendencies as frequently as possible. These experiments indicated that despite the higher horizontal resolution (H212 versus H007), the lack of temporal resolution of H212 versus H3GP in terms of physics calls (Table 1) had a negative impact on intensity forecasts. In fact, much of what was gained through the higher resolution was very likely lost as a result of the reduction in frequency of the physics calls. The fact that

H3GP better predicted the intensity of the four problem storms (Fig. 1), most of which underwent RI, suggests it is likely the relatively poor performance of H212 (versus H007 and H3GP) on these storms was the result of the reduction in the frequency of the physics calls. Figure 7b showed that H212 intensity forecasts were comparable to those from H3GP for the IWS cases but much worse for the IHS cases. The physics calls are possibly more critical with IHS since with IWS the hurricane intensity changes are driven mainly (more or less) by air–sea interaction, whereas with IHS the environment is much more critical. With IHS, the physics calls not only affect the vortex but could also affect the large-scale flow in the vicinity of the vortex in the innermost (3 km) grid.

Figure 11 shows the impact of the lower frequency of physics calls from idealized runs using an updated version of the model. Clearly, despite the 3-km horizontal resolution in the innermost mesh, the model is unable to produce the finescale updraft–downdraft structure that is key for transporting mass and moisture to and from the upper troposphere and would likely have an impact on intensity prediction, especially RI events. The vertical structure of the tropical cyclone is much improved when the physics call frequency is increased from 3 min to every 30 s.

This correction (increase) to the temporal resolution of the physics calls, as well as several other improvements (e.g., improved nest motion algorithm and further improvements to vortex initialization), was incorporated into a newer operational version of HWRF (H213; Tallapragada et al. 2013). The results from the H213 retrospective runs for the 60-h intensity forecasts of three of the problem storms (Julia, Lisa, and Ophelia) were much improved over the results from H212 and were close to the results for H3GP. One could expect that if the improvements to H3GP that produced the improved track forecasts in H212 were combined with the adequate physics calls of H3GP that provided the better intensity forecasts, the result would be a model with improved track and intensity forecasts. In fact, that was exactly what happened in the H213 version. The resulting reduction in track and intensity errors (for retrospective runs) using H213 was so dramatic that this new version of HWRF was operationally implemented for the 2013 hurricane season, replacing H212.

The skill of the retrospective average track and intensity forecasts for the 2010–11 seasons for the 2013 version of HWRF versus the other HWRF versions presented in this study is shown in Fig. 12. For track forecasts (Fig. 12a), H213 outperforms all of the other models shown and these improvements are significant at all forecast times versus H007 and H3GP, and several of the times versus H212, although only at one time versus GFS (Table 2); that is, H213 is slightly better than even H212 for track forecasts

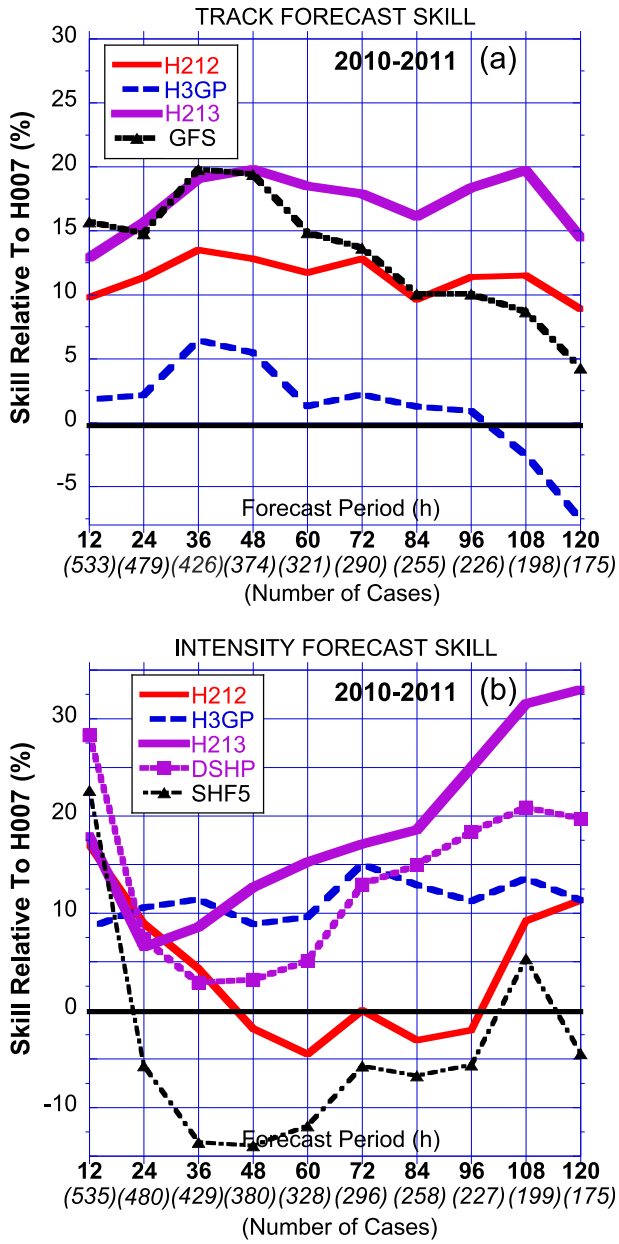


FIG. 12. Skill of (a) track and (b) intensity forecast errors relative to H007, as in Figs. 2b and 5b, respectively, but with the addition of H213 results. Note that some of the results look slightly different compared to results shown in Figs. 2b and 5b since the samples in Fig. 12 are not homogeneous with the samples in Figs. 2b and 5b as a result of the addition of the H213 retrospective runs.

and at least equivalent to GFS, which is considered to be one of the best models for track forecasts. That is a major achievement for HWRF considering that GFS is a global model and HWRF has a much more limited large-scale domain. Also, the results show that the correction of the problem from inadequate physics calls, plus other additional modifications incorporated into H213, resulted in a

model superior to all of the other models shown for intensity forecasts (Fig. 12b), and most of these improvements are statistically significant. In particular, the problem with the H212 intensity forecasts for the IHS cases was totally resolved in the H213 version (not shown), with H213 being comparable to H3GP for both the IHS and IWS samples and then superior to H3GP starting at 84 h. These ongoing improvements in HWRF intensity forecasts, with the results becoming comparable to the statistical–dynamical models, now show that dynamical forecast models can be viable intensity forecast tools and contribute to a decrease in the operational forecast errors. Details of the upgrades and verifications for the 2013 and subsequent operational versions of HWRF are available online (http://www.dtcenter.org/HurrWRF/users/docs/).

Ongoing research continues to examine and test additional areas for improvements to HWRF. These areas include a larger domain ("basin scale") combined with the ability to simultaneously cope with multiple storms, allowing for interactions between storms, multiscale interactions, and land interactions (Zhang et al. 2015, manuscript submitted to *Wea. Forecasting*). The improvements to track, intensity, and structure presented in this study demonstrate that substantial, significant improvements to forecasting with dynamical models are possible and that increased resolution holds the key to some of these improvements.

*Acknowledgments.* The authors acknowledge funding from NOAA's Hurricane Forecast Improvement Project that supported this work. The changes to H3GP to create the new H212 operational version and testing of H212 were primarily developed at the EMC. The authors wish to thank Drs. Rob Rogers, Sim Aberson (HRD), and three anonymous reviewers for their reviews of the manuscript; Gail Derr (AOML) for editorial assistance; Robert Black (HRD) for helpful comments; and James Franklin (NHC) for answering numerous questions along the way regarding verifications.

## REFERENCES

Aberson, S. D., 1998: Five-day tropical cyclone track forecasts in the North Atlantic basin. *Wea. Forecasting,* **13,** 1005–1015, doi:10.1175/1520-0434(1998)013<1005:FDTCTF>2.0.CO;2.

——, 2010: Ten years of hurricane synoptic surveillance (1997–2006). *Mon. Wea. Rev.,* **138,** 1536–1549, doi:10.1175/2009MWR3090.1.

——, and M. DeMaria, 1994: Verification of a nested barotropic hurricane track forecast model (VICBAR). *Mon. Wea. Rev.,* **122,** 2804–2815, doi:10.1175/1520-0493(1994)122<2804:VOANBH>2.0.CO;2.

Bender, M. A., I. Ginis, R. E. Tuleya, B. Thomas, and T. Marchok, 2007: The operational GFDL coupled hurricane–ocean prediction system and a summary of its performance. *Mon. Wea. Rev.,* **135,** 3965–3989, doi:10.1175/2007MWR2032.1.

Bernardet, L., and Coauthors, 2015: Community support and transition of research to operations for the Hurricane Weather Research and Forecast (HWRF) Model. *Bull. Amer. Meteor. Soc.,* doi:10.1175/BAMS-D-13-00093.1, in press.

Cangialosi, J. P., and J. L. Franklin, 2013: 2012 National Hurricane Center forecast verification report. NOAA/National Hurricane Center, 79 pp. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2012.pdf.]

DeMaria, M., 2009: A simplified dynamical system for tropical cyclone intensity prediction. *Mon. Wea. Rev.,* **137,** 68–82, doi:10.1175/2008MWR2513.1.

——, J. Kaplan, J. A. Knaff, M. Mainelli, and L. K. Shay, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting,* **20,** 531–543, doi:10.1175/WAF862.1.

——, J. A. Knaff, and J. Kaplan, 2006: On the decay of tropical cyclone winds crossing narrow landmasses. *J. Appl. Meteor.,* **45,** 491–499, doi:10.1175/JAM2351.1.

Ferrier, B. S., 1994: A double-moment multiple-phase four-class bulk ice scheme. Part I: Description. *J. Atmos. Sci.,* **51,** 249–280, doi:10.1175/1520-0469(1994)051<0249:ADMMPF>2.0.CO;2.

Gall, R., J. Franklin, F. D. Marks, E. N. Rappaport, and F. Toepfer, 2013: The Hurricane Forecast Improvement Project. *Bull. Amer. Meteor. Soc.,* **94,** 329–343, doi:10.1175/BAMS-D-12-00071.1.

Gopalakrishnan, S. G., N. Surgi, R. Tuleya, and Z. Janjic, 2006: NCEP's two-way-interactive-moving-nest NMM-WRF modeling system for hurricane forecasting. Preprints, *27th Conf. on Hurricanes and Tropical Meteorology,* Monterey, CA, Amer. Meteor. Soc., 7A.3. [Available online at https://ams.confex.com/ams/pdfpapers/107899.pdf.]

——, 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. Developmental Testbed Center Rep., L. Bernardet, Ed., 75 pp.

——, F. D. Marks Jr., X. Zhang, J.-W. Bao, K.-S. Yeh, and R. Atlas, 2011: The experimental HWRF system: A study on the influence of horizontal resolution on the structure and intensity changes in tropical cyclones using an idealized framework. *Mon. Wea. Rev.,* **139,** 1762–1784, doi:10.1175/2010MWR3535.1.

——, S. Goldenberg, T. Quirino, F. Marks Jr., X. Zhang, K.-S. Yeh, R. Atlas, and V. Tallapragada, 2012: Toward improving high-resolution numerical hurricane forecasting: Influence of model horizontal grid resolution, initialization, and physics. *Wea. Forecasting,* **27,** 647–666, doi:10.1175/WAF-D-11-00055.1.

——, F. Marks Jr., J. A. Zhang, X. Zhang, J.-W. Bao, and V. Tallapragada, 2013: A study of the impacts of vertical diffusion on the structure and intensity of tropical cyclones using the high-resolution HWRF system. *J. Atmos. Sci.,* **70,** 524–541, doi:10.1175/JAS-D-11-0340.1.

Haus, B. K., D. Jeong, M. A. Donelan, J. A. Zhang, and I. Savelyev, 2010: Relative rates of sea–air heat transfer and frictional drag in very high winds. *Geophys. Res. Lett.,* **37,** L07802, doi:10.1029/2009GL042206.

Hendricks, E. A., M. S. Peng, X. Ge, and T. Li, 2011: Performance of a dynamic initialization scheme in the Coupled Ocean–Atmosphere Mesoscale Prediction System for Tropical Cyclones (COAMPS-TC). *Wea. Forecasting,* **26,** 650–663, doi:10.1175/WAF-D-10-05051.1.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.,*

**124,** 2322–2339, doi:10.1175/1520-0493(1996)124<2322:NBLVDI>2.0.CO;2.

Janjic, Z. I., 2003: A non-hydrostatic model based on a new approach. *Meteor. Atmos. Phys.,* **82,** 271–285, doi:10.1007/s00703-001-0587-6.

Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting,* **18,** 1093–1108, doi:10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2.

——, ——, 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:10.1175/2009WAF2222280.1.

Knaff, J. A., M. DeMaria, B. Sampson, and J. M. Gross, 2003: Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting,* **18,** 80–92, doi:10.1175/1520-0434(2003)018<0080:SDTCIF>2.0.CO;2.

Paterson, L. A., B. N. Hanstrum, and N. E. Davidson, 2005: Influence of environmental vertical wind shear on the intensity of hurricane-strength tropical cyclones in the Australian region. *Mon. Wea. Rev.,* **133,** 3644–3660, doi:10.1175/MWR3041.1.

Powell, M. D., S. H. Houston, L. R. Amat, and N. Morisseau-Leroy, 1998: The HRD real-time hurricane wind analysis system. *J. Wind Eng. Ind. Aerodyn.,* **77–78,** 53–64, doi:10.1016/S0167-6105(98)00131-7.

Rappaport, E. N., and Coauthors, 2009: Advances and challenges at the National Hurricane Center. *Wea. Forecasting,* **24,** 395–419, doi:10.1175/2008WAF2222128.1.

Tallapragada, V., and Coauthors, 2013: Hurricane Weather and Research and Forecasting (HWRF) Model: 2013 scientific documentation. Development Testbed Center, 99 pp. [Available online at http://www.dtcenter.org/HurrWRF/users/docs/scientific_documents/HWRFv3.5a_ScientificDoc.pdf.]

——, C. Kieu, Y. Kwon, S. Trahan, Q. Liu, Z. Zhang, and I.-H. Kwon, 2014: Evaluation of storm structure from the operational HWRF Model during 2012 implementation. *Mon. Wea. Rev.,* **142,** 4308–4325, doi:10.1175/MWR-D-13-00010.1.

Velden, C. S., and S. B. Goldenberg, 1987: The inclusion of high density satellite wind information in a barotropic hurricane-track forecast model. Preprints, *16th Conf. on Hurricanes and Tropical Meteorology,* Houston, TX, Amer. Meteor. Soc., 90–93.

Wang, X., D. Parrish, D. Kleist, and J. S. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.,* **141,** 4098–4117, doi:10.1175/MWR-D-12-00141.1.

Xue, M., J. Schleif, F. Kong, K. K. Thomas, Y. Wang, and K. Zhu, 2013: Track and intensity forecasting of hurricanes: Impact of cloud-resolving resolution and ensemble Kalman filter data assimilation on 2010 Atlantic season forecasts. *Wea. Forecasting,* **28,** 1366–1384, doi:10.1175/WAF-D-12-00063.1.

Zhang, J. A., F. D. Marks Jr., M. T. Montgomery, and S. Lorsolo, 2011: An estimation of turbulent characteristics in the low-level region of intense Hurricanes Allen (1980) and Hugo (1989). *Mon. Wea. Rev.,* **139,** 1447–1462, doi:10.1175/2010MWR3435.1.

Zhang, X., T. S. Quirino, S. Gopalakrishnan, K.-S. Yeh, F. D. Marks Jr., and S. B. Goldenberg, 2011: HWRFX: Improving hurricane forecasts with high resolution modeling. *Comput. Sci. Eng.,* **13,** 13–21, doi:10.1109/MCSE.2010.121.

Zhao, B., Y. Duan, H. Yu, and B. Du, 2006: A statistical analysis on the effect of vertical wind shear on tropical cyclone development. *Acta Meteor. Sin.,* **20,** 383–388.