

IS TROPICAL CYCLONE INTENSITY GUIDANCE IMPROVING?

BY MARK DeMARIA, CHARLES R. SAMPSON, JOHN A. KNAFF, AND KATE D. MUSGRAVE

Contrary to what is often reported in the literature, tropical cyclone intensity forecast models have improved over the past two decades at a rate that is statistically significant.

Tropical cyclone (TC) intensity is defined as the 1-min maximum surface (10 m) wind by the National Hurricane Center (NHC), the Central Pacific Hurricane Center (CPHC), and the Joint Typhoon Warning Center (JTWC). TC intensity is measured in terms of knots (kt; $1 \text{ kt} = 0.514 \text{ m s}^{-1}$), so we will use knots, the native unit, for the remainder of the paper. These operational TC forecast centers monitor the location and intensities of all active TCs in their respective areas of responsibility.¹ The centers also issue intensity and track forecasts every 6 h and contain projections valid 12, 24, 36, 48, 72, 96, and 120 h after the forecast's nominal initial time (0000, 0600, 1200, or 1800 UTC). All three centers create postseason best tracks² and verify their forecasts versus the track and intensity information contained in those best tracks. The best-track intensity values are provided to the nearest 5-kt interval (Landsea and Franklin 2013; Landsea et al. 2013; Chu et al. 2002).

It is well accepted that operational track forecasts have improved dramatically in the last decade or so, but verifications of operational intensity forecasts over these same time periods have shown, by these centers' own assessments, little or no improvement

(Cangialosi and Franklin 2012b; Falvey 2012). For example, Fig. 1 shows the time series of the yearly mean absolute intensity errors through 2012 in the North Atlantic from NHC and for the western North Pacific from JTWC. Atlantic intensity forecasts along with trend lines show that there has indeed been relatively little improvement in the official NHC forecasts in the North Atlantic at the shorter ranges (24–48 h). A similar picture emerges from the JTWC official intensity forecast verifications in Fig. 1, with little short-term improvement.

Intensity verifications of short-term intensity forecasts such as these have led to the conclusion that little or no progress is being made in improving the methods or models that produce TC intensity forecasts, which we refer to as intensity guidance. This perception is reflected by statements such as, “The state of affairs in [storm] track is steady improvement, while the state of affairs in [storm] intensity has been stagnant,” said Frank Marks (Freedman 2012). Many similar statements appear in the refereed literature. For example, Harnos and Nesbitt (2011, p. 1) state, “While trends in TC track forecasts have seen an improvement in accuracy of nearly 50% at time periods of 24–72 h from 1990

¹ NHC's area of responsibility includes the North Atlantic and the eastern North Pacific east of 140°W, and the JTWC area of responsibility includes the western North Pacific Ocean west of the date line; the northern Indian Ocean; and the Southern Hemisphere, excluding the South Atlantic. CPHC's area includes the eastern North Pacific from 140°W to the date line.

² Best tracks are the best estimate of tropical cyclone position and intensity based on current operational methods and all the data that are available at the end of the season. Best tracks are typically created with 6-hourly temporal resolution.

to 2008, little gain in accuracy of intensity forecasts over this same time frame have been achieved.” Zhang and Tao (2013, p. 975) state, “For the past few decades, despite large improvement in the track forecast of tropical cyclones (TCs), there is almost no improvement in the intensity forecast for all lead times.” Similar statements can be found in many other recent papers on tropical intensity change prediction.

The purpose of this study is to perform a quantitative analysis to determine if there have been any improvements in TC intensity guidance accuracy over the past two decades, and to relate those results to the trends in the operational intensity and track forecast errors. The dataset is presented in the following section, the verification procedures are described in the “verification methods” section, results of the trend analysis are presented in the “results” section, and the model and operational forecast trends are compared in the “comparison of model and official forecast error trends” section. Conclusions are summarized in the final section.

DATA USED IN THIS STUDY. The starting point for the trend analysis of intensity guidance errors is 1989, which is the year that the Automated Tropical Cyclone Forecast (ATCF; Sampson and Schrader 2000) system first became available at the operational forecast centers. Intensity guidance data for the 24-yr period ending in 2012 were included. The ATCF contains the forecasts from all available intensity guidance and those from the official forecasts. The verifying TC intensities come from the final best-track datasets created by NHC and JTWC. The ATCF data are the basis for the verification statistics like those shown in Fig. 1.

The forecasts in the Atlantic (AL), eastern North Pacific (EP), western North Pacific (WP), and Southern Hemisphere (SH) will be considered. The

SH includes the South Pacific and the southern Indian Oceans. Although CPHC produces forecasts for the region between 140°W and the date line and JTWC provides forecasts for the northern Indian Ocean, the numbers of forecasts made in those regions are too small to evaluate intensity guidance error trends. During the first half of the 24-yr verification period, the official forecasts for the AL, EP, and WP basins extended to 72 h. These were increased to 120 h on an experimental basis in 2001 and became fully operational in 2003. Thus, the models will be verified out to 72 h for 1989–2000 and to 120 h for 2001–12 for those three basins. The SH forecasts only extended to 48 h from 1989 to 1994. From 1995 to 2009, 72 h was added for a small subset of storms on an “as needed” basis. Beginning in 2010, the SH forecasts were extended to 120 h, similar to the other basins. For this reason, the SH trend analysis is only performed through 48 h.

AFFILIATIONS: DeMARIA AND KNAFF—NOAA/NESDIS, Fort Collins, Colorado; SAMPSON—Naval Research Laboratory, Monterey, California; MUSGRAVE—Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado

CORRESPONDING AUTHOR: Mark DeMaria, NOAA/NESDIS and CIRA, CSU, 1375 Campus Delivery, Fort Collins, CO 80525
E-mail: mark.demaria@noaa.gov

The abstract for this article can be found in this issue, following the table of contents.

DOI:10.1175/BAMS-D-12-00240.1

In final form 18 July 2013

©2014 American Meteorological Society

FIG. 1. Time series of operational mean absolute intensity errors (kt) from (top) NHC for AL TC intensity forecasts at 24, 48, 72, 96, and 120 h with linear trends indicated by the dashed lines, and (bottom) JTWC for the WP.

The official forecasts from NHC and JTWC must be issued by 3 h after synoptic time. The verification is restricted to those models that were available by the time official forecast is released (“early” models). All of the statistically based models are early models, but the dynamical models are almost never available by 3 h after synoptic time (“late” models). To better utilize the late models, postprocessing methods to interpolate the dynamical model forecasts from the previous cycle to the current forecast time have been developed. The version of the interpolator used in this study is similar to that described in Sampson et al. (2006). Only the interpolated (early) versions of the dynamical models were included in the verification. Also, a model was included in a given year only if the intensity forecasts were available for at least half of the cases that had an official forecast. All of the verification results for each year include a homogeneous sample of intensity guidance and official forecast cases.

Models included in the AL and EP verification are listed in Table 1. For the first few years of the verification period, the only available intensity model was the Statistical Hurricane Intensity Forecast model (SHIFOR; Jarvinen and Neumann 1979), which is listed as SHFR in Table 1. This model uses climatology and persistence variables to forecast intensity changes out to 72 h. A 5-day version of SHIFOR (SHF5) was implemented in 2001, as described in Knaff et al. (2003).

The Statistical Hurricane Intensity Prediction Scheme (SHIPS) first became available for the AL in 1991 on a limited basis, but the sample size did not contain at least half of the official forecasts until 1992. The EP version became available in 1996. The SHIPS model uses a statistical–dynamical approach, in which predictors of intensity change are obtained from climatology, persistence, global model forecast fields, and satellite data. The SHIPS model has undergone considerable changes since its introduction in 1991 (DeMaria et al. 2005). One of the most significant changes was the explicit treatment of the effects of land. That version is referred to as Decay-SHIPS (DSHP) and is tracked as a separate model in the verification.

The Logistic Growth Equation Model (LGEM) was implemented in 2006. It is a statistical–dynamical model that includes the same input as SHIPS, but it utilizes a more sophisticated prediction equation (DeMaria 2009).

Real-time forecasts from the Geophysical Fluid Dynamics Laboratory (GFDL) hurricane model first became available in 1992 on a limited basis,

TABLE 1. The intensity models included in the Atlantic and east Pacific verifications.

Model	Type	Years (AL)	Years (EP)
SHFR	Statistical	1989–2000	1989–2000
SHF5	Statistical	2001–12	2001–12
SHIP	Statistical–dynamical	1992–2012	1996–2012
DSHP	Statistical–dynamical	2000–12	2000–12
LGEM	Statistical–dynamical	2006–12	2006–12
GFDI	Dynamical	1996–2012	1997–2012
GHMI	Dynamical	2006–12	2006–12
HWFI	Dynamical	2007–12	2007–12
ICON	Consensus	2006–12	2006–12

and became fully operational in 1995. The interpolated (early) version (GFDI) also became available in 1995, but only for track. The interpolated version for intensity was added in 1996 for the AL and in 1997 for the EP. A modified version of the interpolator (GHMI) was added in 2006. For the GFDI, the difference between the current intensity and the 6-h GFDL forecast from the 6-h-old GFDL run is added as a correction out to 120 h. In GHMI the correction is phased out by 48 h. The GFDL hurricane model has undergone a number of changes since it was first transitioned to operations (Bender et al. 2007), including the addition of a coupled ocean model and upgraded physical parameterizations. The coupled ocean–atmosphere Hurricane Weather Research and Forecasting model (HWRF) and its interpolated version (HWFI) became operational in 2007 (Gopalakrishnan et al. 2010).

Once a number of intensity guidance models became available, it became possible to develop consensus forecasts. The first of these was ICON (intensity consensus) in 2006 (Sampson et al. 2008). The initial version was a simple average of DSHP and GFDI, but HWFI and LGEM were added beginning in 2008.

Two early models were not included in the analysis even though they were available for the AL and EP basins: the U.S. Navy’s version of the GFDL model (GFNI; Rennick 1999) and a variation on ICON that does not require all of the members to be available to provide a forecast [intensity variable consensus (IVCN)]. The GFNI model annual sample sizes were often smaller than those of the other models and generally had errors larger than GFDI and GHMI. IVCN became available in 2008 but, because all the verification results are for homogeneous samples, the results for IVCN and ICON were nearly identical.

TABLE 2. The intensity models includes in the WP and SH verifications. COTI = COAMPS-TC with Global Forecast System (GFS) initial conditions.

Model	Type	Years (WP)	Years (SH)
CLIM	Statistical	1989–2012	1989–1992
CLIP	Statistical	1994–12	—
ST5D	Statistical–dynamical	2001–12	2005–12
ST10	Statistical–dynamical	2005–12	2006–12
ST11	Statistical–dynamical	2007–12	2007–12
GFNI	Dynamical	2001–12	2003–12
COTI	Dynamical	2011–12	—
S511	Consensus	2008–12	2009–12

Models included in the verification for JTWC’s areas of responsibility are listed in Table 2. For the first half of the period, the only WP early models were the simple statistical Climatology from the Typhoon Analog model (CLIM; Sampson et al. 1990) and the Climatology and Persistence Model (CLIP; Chu 1994). Details of these models are provided in DeMaria et al. (2007). The GFNI and the 5-day Statistical Typhoon Intensity Forecast system (ST5D) became available for the WP in 2001. GFNI is the U.S. Navy’s version of GFDI described above for the Atlantic. ST5D is similar to the SHIFOR model (Knaff et al. 2003). The Statistical Typhoon Intensity Prediction Scheme (STIPS) model (Knaff et al. 2005), which is very similar to DSHP—albeit much simpler—was implemented in 2005. STIPS was not included in the verification because it is heavily utilized in the ST10, ST11, and S511 models described below, which generally have superior verification statistics. The interpolated version of the Coupled Ocean–Atmosphere Mesoscale Prediction System for Tropical Cyclones (COAMPS-TC) coupled ocean–atmosphere model (Doyle et al. 2011) became available in the WP and SH in 2011.

The consensus model approach for the WP and SH has taken a somewhat different approach from that for the AL and EP. Rather than simply averaging intensity forecasts from different models, the STIPS statistical–dynamical model has been run with a variety of inputs to create an ensemble of intensity forecasts, which are then averaged to create the consensus forecast. The primary inputs to STIPS are a track forecast and forecast fields from a numerical weather prediction model. For ST10, 10 combinations of tracks and global model input fields are used (Sampson et al. 2008). ST11 is similar to ST10 but with the GFNI added as the eleventh forecast in the consensus. S511 is similar to ST11 but with a version

of STIPS that includes an ocean heat content predictor (Goni et al. 2009).

The ATCF databases for the first half of the 24-yr period are much sparser in the SH than in the other basins. In fact, all of the intensity guidance was missing from 1993 to 1997, and there were no early models from 1998 to 2002. The CLIM model forecasts were available from 1989 to 1992 and so are considered representative of that time. Beginning in 2003, the SH models are similar to those for the WP, although the forecasts did not routinely extend beyond 48 h until 2010.

VERIFICATION METHODS. The intensity guidance discussed in section 2 is verified versus the intensity estimates in the final best tracks. The verifications are homogeneous, so they contain the same cases for each intensity guidance method in each year. Forecasts are verified only when the best-track intensity is greater than 20 kt and only when the system is classified as a tropical or subtropical cyclone. The cyclone must also be tropical or subtropical at the start of the forecast to be included. These are the same verification rules used by NHC and JTWC to determine the official forecast intensity errors shown in Fig. 1. Intensity guidance error (intensity error) is defined here as the absolute value of the difference between the forecast and best-track intensity at the forecast verifying time.

The intensity errors as a function of year are fitted using standard linear regression to determine the time trends. If the slope of the best-fit line is negative, then the errors are decreasing with time. The statistical significance of a decreasing trend is determined by considering the null hypothesis that $m \geq 0$, where m is the slope of the regression line. A decreasing error trend is considered statistically significant if the null hypothesis can be rejected at the 95% level using the Student’s t test.

The independent variable for the linear fit is the number of years from the start of the time series (1989 for the 24-, 48-, and 72-h forecasts and 2001 for the 96- and 120-h forecasts). The constant (b) of the best-fit line is its value at the start of the time series. The forecast improvement will be presented in units of percent per year, which is calculated from $-100(m/b)$.

As described in section 2, only simple statistical models were available in the first few years of the 24-yr verification period, statistical–dynamical and early versions of the dynamical models became available near the middle of the verification period, and consensus techniques became available in the latter part of the period. To determine the intensity

guidance error trend, the best model over a multiyear period where the set of intensity guidance was fairly constant was chosen. Although it is not possible to replicate the subjective method forecasters use to determine which models to use, picking one technique that was best on average over a multiyear period was considered better than just picking the best model in each year because it takes at least a season or two for forecasters to gain confidence in new techniques. The linear regression was then fit to the best model time series. Further details are provided in section 4.

RESULTS. For brevity the initial focus will be on the 48-h intensity forecasts in the four TC basins. Figure 2 shows the annual verifications for the AL and EP intensity guidance summarized in Table 1 and the official forecasts. For the AL the average yearly sample size was 195 with a range of cases from 55 in 1991 to 399 in 1995. For the EP the average yearly sample size was 193 with a range of cases from 84 in 1995 to 500 in 1992. SHFR was the only model during 1989–91, so it was considered the best. From 1992 to 1995, SHIP and SHFR were available, but the early version of SHIP did not improve upon SHFR, so SHFR was still considered the best model. From 1996 to 1999, SHFR, SHIP, and GFDI were available, with GFDI performing the best during that period. From 2000 to 2005, DSHP was the best-performing model, and ICON was best on average from 2006 to 2012. Fitting the trend line to the errors from those four techniques (i.e., SHFR, 1989–95; GFDI, 1996–99; DSHP, 2000–05; ICON, 2006–12) shows a downward trend, indicating that the intensity guidance has indeed improved over this 24-yr period. The official forecasts show less of a downward trend (red dashed lines). A similar analysis was performed for the EP (Fig. 2, bottom), and the best model error trend line also shows a downward trend. The best models in each year of the EP were very similar to those for the AL.

The trend lines of the best models and official forecasts described above

were determined for the other forecast periods from 24 through 120 h, and the statistical significance tests on the trends were performed. Although not always true, the same models considered the best at 48 h were used as the best at the other time periods. With this simplification, Table 3 indicates that the best intensity guidance shows a statistically significant improvement at 48–120 h for the AL and at 24–72 h in the EP. The intensity guidance improvement rates for the times that are statistically significant range from 0.8% to 1.5% per year at 24–72 h, with larger improvement rates at 96 and 120 h.

As a basis for comparison, Table 3 also shows the improvement rates of the NHC official track forecasts for the same periods as the intensity guidance.

Fig. 2. Time series of annual 48-h intensity errors (i.e., MAE) associated with intensity forecast guidance and simple intensity consensus methods in (top) the AL, and (bottom) the EP. Linear trend lines for the best model in each year and the official forecasts are also shown (see text for explanation of how the best model was chosen).

TABLE 3. Average improvement rates (% yr⁻¹) of the best intensity guidance model (left number), official intensity forecast from NHC or JTWC (middle number), and official track forecast (right number) for the period 1989–2012 for the 24-, 48-, and 72-h forecasts and for the period 2001–12 for the 96- and 120-h forecasts. Improvements that are statistically significant at the 95% level are in bold. The official track error trends are not provided in the SH because of an incomplete time series.

Basin	Time (h)				
	24	48	72	96	120
AL	0.1/0.0/ 2.7	1.3 /0.6/ 3.0	1.1 / 0.9 / 3.1	4.3 / 2.2 / 4.1	4.9 / 2.5 / 3.8
EP	0.8 /0.5/ 2.2	1.2 / 1.1 / 2.3	1.5 / 1.5 / 2.4	0.8 / 2.3 / 3.6	1.3 / 2.3 / 4.2
WP	1.9 /–0.6/ 2.5	1.9 /0.0/ 2.7	2.1 /0.7/ 2.8	2.5 / 2.7 / 1.9	2.1 / 1.5 / 2.1
SH	1.8 /–0.6/*	1.5 /0.1/*	*	*	*

Although these trends are not direct measurements of the track model improvements, they are good estimates because the NHC official track errors have been very close to those of the finest track models for the past few decades (e.g., Cangialosi and Franklin 2012a; DeMaria and Gross 2003). All of the track improvement trends were statistically significant. The AL and EP intensity guidance improvements are 2–3 times smaller than the concurrent track improvements out to 72 h. The rates of intensity guidance improvement at 96 and 120 h are comparable to those of track for the AL but 3–4 times smaller for the EP.

Table 3 shows that the NHC official intensity errors for the AL also show a statistically significant downward trend at 72 h, but the improvements were smaller than those of the best model. For the EP, the official forecast improvements were significant at more time periods and were more closely tied to the intensity guidance trends, especially at 48 and 72 h. Thus, in addition to intensity guidance improvements, the NHC official intensity forecasts at the longer time periods have improved as well. It is possible that the emphasis on the short time periods and the smaller improvement rates compared with track errors have led to the improper conclusion that intensity guidance has not improved in several decades.

Figure 3 shows the yearly 48-h intensity errors for the WP and SH guidance and official forecasts. For the WP the average yearly sample size was 335 with a range of cases from 130 in 2010 to 700 in 1994. For the SH the average yearly sample size was 109 with a range of cases from 26 in 1990 to 160 in 2012. Similar to the AL and EP, the intensity guidance improved as the new forecast techniques became available, with the consensus methods being best in the past few years. For the WP, the best model trend line included CLIM from 1989 to 1993, CLIP from 1994 to 2000, ST5D from 2001 to 2004, ST10 from 2005 to 2007, and S511 from 2008 to 2012. The downward slope of the

best model trend line is a little steeper than for the AL, which is most likely due to the later implementation of the more sophisticated intensity guidance.

For the SH, the ATCF data were incomplete in 1993–2002, so those years were not included in the trend analysis. The best model included CLIM from 1989 to 1992, GFNI from 2003 to 2004, ST5D in 2005, ST10 in 2006, ST11 in 2007–08, and S511 in 2009–12. The downward slope of the best model for the SH is similar to that for the AL and EP and a little less steep for the WP. The downward slope of the best model was twice as large when only the period from 2003 to 2012 with more complete data was included in the linear trend.

The official forecast error trend lines for the WP and SH shown in Fig. 3 are very flat, indicating little change in accuracy. This result is confirmed Table 3, where the WP and SH official forecasts showed no significant improvement at 24–72 h. The WP official forecasts did show significant improvement at 96 h. In contrast, the best WP and SH intensity guidance showed significant improvements at 24–72 h. Possible reasons for the discrepancy between the best intensity guidance and official error trends will be discussed in more detail in the next section. Note that the rates of the WP official track error improvement in Table 3 have been comparable to those for the AL and EP out to 72 h, but they have been a little smaller at 96 and 120 h. Because of the incomplete time series, the official track error trends were not calculated for the SH at the longer ranges.

The SH official forecast error trend line in Fig. 3 included the points from 1993 to 2002, when there was no intensity guidance available. For consistency with the intensity guidance, the official forecast error trends were recalculated by using only those years that also had intensity guidance errors. The trend lines changed very slightly, but the slopes remained statistically insignificant.

To further illustrate the recent improvements in the intensity guidance, Fig. 4 shows 4-yr error statistics (2009–12) for a variety of intensity guidance types. The errors for the SH only include a 3-yr sample, since those have only been routinely produced out to 5 days since 2010. The use of these longer periods lessens the impact of interannual variability of TC location and activity. In this figure, improvements are shown by the progression of the errors toward the bottom of the figure, with skill measured by improvements over the no-skill baseline guidance (SHF5 or ST5D). The AL (EP) sample includes 1171 (910) cases at 12 h, which decreases to 368 (173) cases by 120 h. The WP (SH) sample includes 1395 (567) cases at 12 h, which decreases to 368 (114) cases by 120 h.

The error statistics of LGEM in Fig. 4a show improvements over SHF5 through 120 h. The DSHP errors for the AL (not shown) are similar to those of LGEM through 72 h but larger at 96 and 120 h. This is a positive characteristic of the LGEM formulation, which is more sensitive to changes in the environment at longer lead times (DeMaria 2009) than DSHP. The most impressive performance is for ICON, which had the lowest errors at all times after 36 h. The addition of dynamical models in the consensus improves the ICON forecasts, even though their individual errors are generally higher than those of the statistical–dynamical models, (e.g., GHMI in Fig. 4a).

Figure 4b shows the 4-yr intensity error verification for the EP. The improvements over the no-skill intensity model (SHF5) are quite impressive. The overall performance of the intensity guidance is fairly similar to that for the AL, except that the ICON is not quite as accurate as LGEM at the longer forecast times and GHMI was better than SHF5 at all forecast times.

The HWFI errors for the AL were comparable to those of the GHMI out to 72 h and larger at 96 and 120 h (not shown). For the EP, the HWFI errors were larger than the GHMI at all forecast times. Despite these larger errors, the HWFI still contributed to an improvement in the consensus forecast.

Figure 4c shows the 4-yr intensity errors for the WP. The sample includes 1395 cases at 12 h, which decreases to 368 at 120 h. In this basin all intensity guidance except ST5D and the GFNI are based on intensity consensus methods. Here again there is evidence of a steady progression of skill relative to ST5D. ST10 shows skillful results at all forecast times. ST11 reduces the errors further by adding dynamical models to the consensus. Similar to the AL and EP, the dynamical guidance improves the consensus even though its mean absolute errors (MAEs) are fairly large, as can be seen for the GFNI in Fig. 4c. Finally, the improved formulation of STIPS [i.e., with oceanic heat content (OHC) as a predictor] used in S511 shows a small additional improvement. The picture that emerges is one of slow and steady improvements in the intensity guidance in this basin.

Figure 4d shows 3-yr intensity errors from the SH TC intensity verification. The sample includes 567

FIG. 3. As in Fig. 2, but for (top) the WP and (bottom) the SH.

cases at 12 h, which decreases to 114 at 120 h. Even in this TC basin, where development of TC intensity forecast techniques has lagged that in the other basins, there is clear evidence of a slow and steady improvement in the intensity guidance. ST10 shows dramatically smaller errors than the no-skill ST5D forecast out to 96 h. Adding dynamical forecasts to the ST11 consensus further reduces the intensity errors. However, the addition of the ocean heat content in S511 did not improve upon ST11, as was seen for the WP. Nonetheless, Fig. 4d illustrates steady improvements in the intensity guidance for the SH.

COMPARISON OF MODEL AND OFFICIAL FORECAST ERROR TRENDS. As indicated in Table 3 for the AL, the best intensity guidance errors decreased significantly at all forecast times from 48 through 120 h, but the NHC official forecast error reductions were only significant at 72 h. Furthermore, the best model improvement rates were larger than

those of the official forecasts at all forecast times. Thus, the trends in the AL official intensity errors are not representative of those of the best intensity guidance. The error trends for the WP and SH are similar to those for the AL, with the best model error reductions larger and more significant than those of the official errors at most forecast times. Only in the EP are the best model and official error trends similar at most forecast times, and both have significant improvements at several forecast times.

A natural question to ask is why are the official forecast improvements considerably less than those of the best guidance in most of the basins? Only the early models, those available to the forecasters before they make their forecasts, were included in this comparison. The answer to this question is evident in the time series of the best guidance and the official forecasts in Figs. 2 and 3. For example, the WP official intensity error trend (Fig. 3, top) has a slope very close to zero, while the best model intensity error

FIG. 4. Intensity forecast errors associated with forecast guidance and simple intensity consensus methods averaged for the period 2009–12 in the (a) AL, (b) EP and (c) WP and (d) averaged from 2010 to 2012 in the SH as a function of forecast time.

trend has a significant negative slope. The reason for the lack of improvement in the official forecasts is that, during the first half of the time series, the official forecasts were considerably better than all of the intensity guidance. Thus, the forecasters' subjective methods outperformed what could be achieved objectively during this early period. This is a credit to the forecasters, who were working with very limited tools. As skillful intensity guidance (better than CLIM and CLIP) became available after 2001, the official forecast errors became more comparable to the best guidance errors. In fact, the trend lines cross in about 2008, which suggests that the intensity guidance now has the potential to improve the official forecasts. This can also be seen in Fig. 4c, where the WP official intensity errors were a little larger than those of the best model (i.e., a consensus) during the past four years.

The explanation described above for the WP best model–official error trend difference also applied to the AL (Fig. 2, top) and SH (Fig. 3, bottom). The subjective official forecasts are better than those of all the intensity guidance in the early part of the time series, resulting in much shallower official intensity error trends. The crossing of the trend lines in these two basins does not occur until near the end of the time series, when an intensity guidance-based consensus approach became available. However, the fact that the best models from the past few years in Figs. 4a and 4c (ICON in the AL and S511 in the WP) have lower errors than the official forecasts indicates that continued improvements in the guidance should start to lower the official intensity errors.

The differences in the best model and official error trends described above were not seen in the EP (Fig. 2b). For the EP, the two error trends are similar, which is attributed to the fact that EP TCs usually spend most of their life cycle at low latitudes, where the simpler statistical intensity guidance in the early part of the time series produced intensity errors comparable in accuracy to those of the official forecasts. As the EP intensity guidance improved, the official forecasts improved at a comparable rate.

CONCLUDING REMARKS. Evidence has been presented that is contrary to the common expressions that little or no progress has been made in forecasting TC intensity change. While improvements in operational forecasts of TC intensity from NHC and JTWC have been marginal for the 24-h forecast period, (Fig. 1), the best available intensity guidance has shown considerable advances over the last 24 years. Although improvement rates ($1\%–2\% \text{ yr}^{-1}$) have only

been about one-third to one-half those of the track model improvements at 24–72 h, these improvements are statistically significant at the 95% level at most forecast times. The best intensity model improvement rates at 96 and 120 h ($2\%–4\% \text{ yr}^{-1}$) were larger than those at the earlier times and even comparable to the track forecast improvement rates. The official forecasts were also significantly improved at some intermediate forecast periods (48–96 h).

These intensity guidance improvements have resulted from the transition from very simple climatology and persistence models (no-skill models) in the first part of the 24-yr period to statistical–dynamical and dynamical models in the middle of the period and the use of consensus forecast techniques near the end of the period. Some of the intensity guidance improvement might also be a result of the improved track forecasts, especially for landfall forecasts and at the longer ranges. DeMaria (2010) showed that the statistical–dynamical LGEM intensity errors decreased by up to 30% when the forecast tracks were replaced by the best-track positions while keeping all other model inputs the same as what were available in real time. Thus, the large reduction in track errors in the past few decades likely contributed to the intensity guidance improvements by reducing errors in the timing of landfall, changes in sea surface and subsurface ocean properties, and interactions with the adjacent synoptic environment.

It may seem paradoxical that the intensity guidance has generally improved at a much faster rate than the official intensity forecasts despite the fact that the official forecasts make use of this same intensity guidance. This seeming paradox is explained by the lack of skillful intensity guidance techniques in the early part of the 24-yr sample. For the AL, WP, and SH, the subjective official forecasts were more accurate than the guidance. Thus, the official intensity errors have shown much slower rates of improvement than the intensity guidance during these 24 years. Only in recent years have the improvements in the guidance had a large influence on the official errors. For most forecast intervals, the linear trend lines for the official forecasts and best available intensity guidance crossed in the past few years, which indicates a new era in which further intensity guidance improvements should also be realized in the subjective official forecasts. This improvement trend should now be similar to that for the EP, where the performance of intensity guidance and operational forecast errors had similar trends for most of the 24-yr period.

Although this study has demonstrated statistically significant improvements in intensity

guidance, the intensity forecast problem is far from solved. Improvements in the warning time frame (24–48 h) have been the slowest, and skill in the prediction of rapid intensity changes is still very poor (Kaplan et al. 2010). Fortunately, the Hurricane Forecast Improvement Program (HFIP) is a well-coordinated effort to continue track and intensity forecast improvements through research in development in data assimilation, model advances, and statistical postprocessing (Marks and Ferek 2013). HFIP is a 10-yr National Oceanic and Atmospheric Administration (NOAA) program designed to improve TC track and intensity forecasts by 50% by addressing the forecast guidance available to forecasters and accelerating the transition of research into operations (Gall et al. 2013). HFIP has provided support to improve TC initialization in dynamical models through improved data assimilation of satellite and aircraft reconnaissance observations (Zhang et al. 2011), improving dynamical models (COAMPS-TC, GFDL, and HWRF), and the development of new global and regional ensemble systems (Hamill et al. 2011). This extensive prediction improvement effort is also expected to improve statistical–dynamical forecasting techniques, and preliminary improvements of up to 5%–10% over existing statistical–dynamical models have been achieved (Musgrave et al. 2012).

Efforts sponsored by the U.S. Navy are underway to improve the WP and SH consensus forecasts by replacing the underlying statistical–dynamical model STIPS with the more advanced LGEM and SHIPS model formulations. Preliminary research results have indicated significant (~8%–10%) improvements, and transition to operations should improve the JTWC intensity forecasts. The successful Joint Hurricane Testbed (Rappaport et al. 2012) has also funded many intensity guidance improvements; highlights include incremental upgrades to existing dynamical models and statistical–dynamical techniques to forecast rapid intensification (Kaplan et al. 2010; Sampson et al. 2011). The statistical–dynamical techniques have shown promise in reducing bias and mean forecast errors for the operational intensity consensus in the AL and EP basins. Similar techniques have been applied to the WP basin.

New satellite observations also have great potential to improve track and intensity forecasts when assimilated in hurricane models. For example, the Advanced Technology Microwave Sounder (ATMS) and Visible Infrared Imaging Radiometer Suite (VIIRS) observations are now available from the recently launched *Suomi National*

Polar-Orbiting Partnership (S-NPP) satellite. In addition, the Advanced Baseline Imagery (ABI) and Geostationary Lightning Mapper (GLM) on the next-generation NOAA geostationary satellite system will become available within the next few years (Goodman et al. 2012). In addition to providing better spatial, spectral, and radiometric resolution, the GLM on Geostationary Operational Environmental Satellite-R (GOES-R) and *S-NPP* day/night band (Miller et al. 2012) provide new capabilities that can be exploited to improve TC prediction.

Given the current state of significantly improved intensity guidance and numerous ongoing efforts to improve the existing intensity guidance, the future is not nearly as bleak as has been purported in the popular press and refereed literature, which have stated that little or no progress has been made in TC intensity forecasting.

ACKNOWLEDGMENTS. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. government position, policy, or decision. We acknowledge the Joint Typhoon Warning Center and the National Hurricane Center for their tireless efforts to construct and support the archived data used in this study. We also acknowledge the support of the Office of Naval Research and the NOAA HFIP for funding this and other efforts to improve tropical cyclone intensity forecasting. The authors also thank Chris Landsea and two anonymous reviewers for providing valuable comments that improved this paper.

REFERENCES

- Bender, M. A., I. Ginis, R. 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.
- Cangialosi, J. P., and J. L. Franklin, 2012a: 2011 Atlantic and eastern North Pacific forecast verification. *Proc. 66th Interdepartmental Hurricane Conf.*, Charleston, SC, OFCM. [Available online at [www.ofcm.gov/ihc12/Presentations/01b-Session/03-IHC_2012_Verification_\(2012\)_v2.pdf](http://www.ofcm.gov/ihc12/Presentations/01b-Session/03-IHC_2012_Verification_(2012)_v2.pdf).]
- , and —, 2012b: 2011 National Hurricane Center forecast verification report. NOAA/NWS/NHC, 76 pp. [Available online at www.nhc.noaa.gov/verification/pdfs/Verification_2011.pdf.]
- Chu, J.-H., 1994: A regression model for the western North Pacific tropical cyclone intensity forecast. NRL Memo. Rep. NRL/MR/7541-94-7215, Naval Research Laboratory, 33 pp. [Available from Naval

- Research Laboratory, 7 Grace Hopper Avenue, Monterey, CA 93943.]
- , C. R. Sampson, A. S. Levine, and E. Fukada, 2002: The Joint Typhoon Warning Center tropical cyclone best-tracks, 1945-2000. NRL Reference NRL/MR/7540-02-16. [Available online at www.usno.navy.mil/NOOC/nmfc-ph/RSS/jtwc/best_tracks/TC_bt_report.html.]
- DeMaria, M., 2009: A simplified dynamical system for tropical cyclone intensity prediction. *Mon. Wea. Rev.*, **137**, 68–82.
- , 2010: Tropical cyclone intensity predictability estimates using a statistical-dynamical model. Preprints, *29th Conf. on Hurricanes and Tropical Meteorology*, Tucson, AZ, Amer. Meteor. Soc., 9C.5. [Available online at https://ams.confex.com/ams/29Hurricanes/techprogram/paper_167916.htm.]
- , and J. M. Gross, 2003: Evolution of prediction models. *Hurricane! Coping with Disaster: Progress and Challenges since Galveston, 1900*, R. Simpson et al., Eds., Special Publications Series, Vol. 55, Amer. Geophys. Union, 103–126.
- , M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvement to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543.
- , J. A. Knaff, and C. R. Sampson, 2007: Evaluation of long-term trend in tropical cyclone intensity forecasts. *Meteor. Atmos. Phys.*, **97**, 19–28.
- Doyle, J. D., and Coauthors, 2011: Real-time tropical cyclone prediction using COAMPS-TC. *Adv. Geosci.*, **28**, 15–28.
- Falvey, R., 2012: Summary of the 2011 Western Pacific/Indian Ocean tropical cyclone season. *Proc. 66th Interdepartmental Hurricane Conf.*, Charleston, SC, OFCM. [Available online at www.ofcm.gov/ihc12/Presentations/01b-Session/05-JTWC_2012_IHC_Final.pdf.]
- Freedman, A., cited 2012: Storm intensity forecasts lag; Communities more at risk. [Available online at www.climatecentral.org/news/storm-intensity-forecasts-lag-putting-communities-more-at-risk/.]
- Gall, R., J. Franklin, F. Marks, E. N. Rappaport, and F. Toepfer, 2013: The Hurricane Forecast Improvement Project. *Bull. Amer. Meteor. Soc.*, **94**, 329–343.
- Goni, G., and Coauthors, 2009: Applications of satellite-derived ocean measurements to tropical cyclone intensity forecasting. *Oceanography*, **22**, 190–197.
- Goodman, S. J., and Coauthors, 2012: The GOES-R proving ground: Accelerating user readiness for the next-generation geostationary environmental satellite system. *Bull. Amer. Meteor. Soc.*, **93**, 1029–1040.
- Gopalakrishnan, S. G., 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. Development Testbed Center, 75 pp. [Available online at www.dtcenter.org/HurrWRF/users/docs/scientific_documents/HWR_final_2-2_cm.pdf.]
- Hamill, T. M., J. S. Whitaker, M. Fiorino, and S. G. Benjamin, 2011: Global ensemble predictions of 2009's tropical cyclones initialized with an ensemble Kalman filter. *Mon. Wea. Rev.*, **139**, 668–688.
- Harnos, D. S., and S. W. Nesbitt, 2011: Convective structure in rapidly intensifying tropical cyclones as depicted by passive microwave measurements. *Geophys. Res. Lett.*, **38**, L07805, doi:10.1029/2011GL047010.
- Jarvinen, B. R., and C. J. Neumann, 1979: Statistical forecasts of tropical cyclone intensity for the North Atlantic basin. NOAA Tech. Memo. NWS NHC-10, 22 pp. [Available from National Technical Information Service, 5285 Port Royal Rd., Springfield, VA 22161.]
- Kaplan, J., M. DeMaria, and J. A. Knaff, 2010: A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **25**, 220–241.
- Knaff, J. A., M. DeMaria, C. R. Sampson, and J. M. Gross, 2003: Statistical, five-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting*, **18**, 80–92.
- , C. R. Sampson, and M. DeMaria, 2005: An operational Statistical Typhoon Intensity Prediction Scheme for the western North Pacific. *Wea. Forecasting*, **20**, 688–699.
- Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592.
- , —, E. Blake, and R. Tanabe, 2013: The revised northeast and north central Pacific hurricane database (HURDAT2). NOAA, 4 pp. [Available online at www.nhc.noaa.gov/data/hurdat/hurdat2-format-nencpac.pdf.]
- Marks, F., Jr., and R. Ferek, 2013: Comparison of the 2010 and 2012 snapshots of tropical cyclone research and development. *67th Interdepartmental Hurricane Conf.*, OFCM. [Available online at www.ofcm.gov/ihc13/67IHC-Linking-File.htm.]
- Miller, S. D., C. L. Combs, S. Q. Kidder, and T. F. Lee, 2012: Assessing moonlight availability for nighttime environmental applications by low-light visible polar-orbiting satellite sensors. *J. Atmos. Oceanic Technol.*, **29**, 538–557.
- Musgrave, K. D., B. D. McNoldy, and M. DeMaria, 2012: Creation of a statistical ensemble for tropical

- cyclone intensity prediction. Preprints, *30th Conf. on Hurricanes and Tropical Meteorology*, Ponte Vedra Beach, FL, Amer. Meteor. Soc., 13A.1. [Available online at <https://ams.confex.com/ams/30Hurricane/webprogram/Paper204580.html>.]
- Rappaport, E. N., J.-G. Jiing, C. W. Landsea, S. T. Murillo, and J. L. Franklin, 2012: The Joint Hurricane Testbed: Its first decade of tropical cyclone research-to-operations activities reviewed. *Bull. Amer. Meteor. Soc.*, **93**, 371–380.
- Rennick, M. A., 1999: Performance of the navy’s tropical cyclone prediction model in the western North Pacific basin during 1996. *Wea. Forecasting*, **14**, 3–14.
- Sampson, C. R., and A. J. Schrader, 2000: The automated tropical cyclone forecasting system (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240.
- , R. J. Miller, R. A. Kreitner, and T. L. Tsui, 1990: Tropical cyclone track objective aids for the micro-computer: PCLM, XTRP, PCHP. Naval Oceanographic and Atmospheric Laboratory Tech. Note 61, 15 pp. [Available from Naval Research Laboratory, 7 Grace Hopper Avenue, Monterey, CA 93943.]
- , J. S. Goerss, and H. C. Weber, 2006: Operational performance of a new barotropic model (WBAR) in the western North Pacific basin. *Wea. Forecasting*, **21**, 656–662.
- , J. L. Franklin, J. A. Knaff, and M. DeMaria, 2008: Experiments with a simple tropical cyclone intensity consensus. *Wea. Forecasting*, **23**, 304–312.
- , J. Kaplan, J. A. Knaff, M. DeMaria, and C. A. Sisko, 2011: A deterministic rapid intensification aid. *Wea. Forecasting*, **26**, 579–585.
- Zhang, F., and D. Tao, 2013: Effects of vertical wind shear on the predictability of tropical cyclones. *J. Atmos. Sci.*, **70**, 975–983.
- , Y. Weng, J. Gamache, and F. Marks, 2011: Performance of convection-permitting hurricane initialization and prediction during 2008–2010 with ensemble data assimilation of inner-core airborne Doppler radar observations. *Geophys. Res. Lett.*, **38**, doi:10.1029/2011GL048469.