Meteorol Atmos Phys 97, 19–28 (2007)
DOI 10.1007/s00703-006-0241-4
Printed in The Netherlands

Meteorology
and Atmospheric
Physics

[1] NOAA/NESDIS, Fort Collins, CO, USA
[2] CIRA/Colorado State University, Fort Collins, CO, USA
[3] Naval Research Laboratory, Monterey, CA, USA

Evaluation of long-term trends in tropical cyclone intensity forecasts

M. DeMaria[1], **J. A. Knaff**[2], and **C. Sampson**[3]

With 5 Figures

Received August 15, 2006; accepted September 17, 2006
Published online: March 14, 2007 © Springer-Verlag 2007

Summary

The National Hurricane Center and Joint Typhoon Warning Center operational tropical cyclone intensity forecasts for the three major northern hemisphere tropical cyclone basins (Atlantic, eastern North Pacific, and western North Pacific) for the past two decades are examined for long-term trends. Results show that there has been some marginal improvement in the mean absolute error at 24 and 48 h for the Atlantic and at 72 h for the east and west Pacific. A new metric that measures the percent variance of the observed intensity changes that is reduced by the forecast (variance reduction, VR) is defined to help account for inter-annual variability in forecast difficulty. Results show that there have been significant improvements in the VR of the official forecasts in the Atlantic, and some marginal improvement in the other two basins. The VR of the intensity guidance models was also examined. The improvement in the VR is due to the implementation of advanced statistical intensity prediction models and the operational version of the GFDL hurricane model in the mid-1990s. The skill of the operational intensity forecasts for the 5-year period ending in 2005 was determined by comparing the errors to those from simple statistical models with input from climatology and persistence. The intensity forecasts had significant skill out to 96 h in the Atlantic and out to 72 h in the east and west Pacific. The intensity forecasts are also compared to the operational track forecasts. The skill was comparable at 12 h, but the track forecasts were 2 to 5 times more skillful by 72 h. The track and intensity forecast error trends for the two-decade period were also compared. Results showed that the percentage track forecast improvement was almost an order of magnitude larger than that for intensity, indicating that intensity forecasting still has much room for improvement.

1. Introduction

The improvement in tropical cyclone (TC) track forecasting is one of the great success stories in the field of meteorology. For example, the average 72-h National Hurricane Center (NHC) official Atlantic track forecast error of ~380 nmi for the period 1970–1979 was reduced to ~160 nmi by 2000–2005 (www.nhc.noaa.gov). This error reduction is primarily due to improved TC track prediction models (McAdie and Lawrence, 2000; DeMaria and Gross, 2003). In the 1970s, track forecasts were primarily based on statistical forecast techniques. Through improvements in computer technology, numerical modeling techniques, in situ and satellite observations and data assimilation, accurate TC track forecasts are currently available from a number of global and regional numerical weather prediction models. The statistical track models are now primarily used as a baseline for evaluation of forecast skill. Because of these improvements and after a two-year evaluation period in 2001–2002, the U.S. TC forecast centers (NHC, the Central Pacific Hurricane Center (CPHC), and the Joint Typhoon Warning Center (JTWC)) extended their forecasts from 3 to 5 days beginning in 2003. The 5-day Atlantic track forecasts in the 2000s are more accurate than the 3-day forecasts in the 1980s.

Despite the major improvements in the TC modeling systems, the intensity forecasts have not shown dramatic improvement. In fact, it is often stated in research studies that intensity forecasts have little or no skill (e.g., Park and Zou, 2004), but without quantitative evidence. In this paper, the long-term trends in the intensity forecasts from the three most active northern hemisphere TC basins (the Atlantic, the eastern North Pacific, and the western North Pacific) are examined in detail to determine if there has been any intensity forecast improvement over the last two decades. The Atlantic and eastern North Pacific (east of 140° W) intensity forecasts that will be evaluated are from NHC, and the western North Pacific forecasts are from JTWC. The CPHC, which is part of the National Weather Service (NWS) Forecast Office in Honolulu, has responsibility for TCs from 140° W to the dateline. However, the CPHC forecast sample sizes are generally too small to reliably evaluate forecast trends.

Intensity forecasts are typically evaluated in terms of a mean absolute error (MAE), which is the difference between the forecasted 1-minute maximum sustained surface wind and that from the post-analysis "best track", where both are measured in knots rounded to the nearest 5. This metric is analogous to the mean absolute distance error, which is used to evaluate the track forecasts. When the forecast improvement is large, as is the case for track forecasting, these metrics work well for evaluating long-term trends. However, there is variability in the forecast difficulty from year to year, which can sometimes make it more difficult to evaluate trends. Various methods have been developed to normalize for forecast difficulty (McAdie and Lawrence, 2000; Neumann, 1981), which typically rely on forecasts based on simple input from climatology and persistence. In this paper, the traditional MAE will be used, and a new metric will be introduced that calculates how much of the variance in the observed intensity changes is reduced by the forecasts (variance reduction). The variance reduction also helps to account for year to year variability in forecast difficulty.

The verification datasets used are described in Sect. 2, along with a brief summary of the intensity guidance models available to NHC and JTWC. In Sect. 3, the forecast metrics are described and the long-term trends are evaluated in Sect. 4. In Sect. 5, the intensity forecast trends and skill are compared with those of the track forecasts.

2. Datasets

As described in the Introduction, the intensity forecasts in the Atlantic (ATLC), eastern North Pacific (EPAC), and western North Pacific (WPAC) TC basins will be evaluated. To evaluate the forecast trends, it is desirable to have as long a time series as possible. The intensity forecasts from 1990–2005 are readily available in the automated tropical cyclone forecast (ATCF) system that was implemented at JTWC and NHC near the end of the 1980s (Sampson and Schrader, 2000). As part of the development of the operational statistical hurricane intensity prediction scheme (SHIPS), the ATLC intensity forecasts back to 1985 and EPAC forecasts back to 1988 were digitized from hard copy and converted to ATCF format. The WPAC intensity forecasts back to 1986 are also available in the ATCF format. NHC took over TC forecast responsibility for the EPAC from the Redwood City NWS forecast office in 1988, so the forecasts before 1988 will not be included. In summary, the evaluation period for this study includes 1985–2005 for the ATLC, 1988–2005 for the EPAC, and 1986–2005 for the WPAC.

The verification sample selection has varied over the years at operational forecast centers (e.g., DeMaria et al, 2005). For example, in past years, NHC restricted their verification to cases of at least tropical storm intensity. In this study, the verification sample utilizes the current NHC criteria. Tropical and subtropical cyclones of any intensity are included, but the extra-tropical, wave and disturbance stages are excluded. Cases from storms that were designated as a tropical cyclone but never got strong enough to be named are also included, except for the ATLC and EPAC prior to 1990, which were not available in the verification files.

Table 1. Operational intensity guidance models available in each forecast basin

Atlantic and East Pacific	
SHIFOR (1988–present)	Statistical hurricane intensity forecast, which uses simple climatology and persistence parameters
SHIPS (1991–present, ATLC) (1996–present, EPAC)	Statistical hurricane intensity prediction scheme, which uses climatology, persistence and real-time atmospheric and oceanic parameters
GFDL (1995–present)	Operational version of the geophysical fluid dynamics laboratory hurricane model
GFDN (2001–present)	GFDL model initialized from navy global model fields
SHIFOR5 (2001–present)	Updated version of SHIFOR with 5-day forecasts
West Pacific	
CLIM (1985–present)	Climatological analog model
STIFOR (1991–present)	Statistical typhoon forecast model, similar to SHIFOR
GFDN (1995–present)	GFDL model initialized from navy global model fields
AFW (2000–present)	MM5 mesoscale model adapted to typhoon forecasts
JTYM (2001–present)	Japanese Meteorological Agency limited area typhoon model
ST5D (2002–present)	Updated STIFOR model and extended to 5 days
STIPS (2003–present)	Statistical typhoon intensity prediction scheme, similar to SHIPS
ST10 (2005–present)	Ensemble version of STIPS

Although the emphasis of this study is on the NHC and JTWC official forecast intensity errors, it helpful to understand the intensity forecast models that were available during the period of this study, as shown in Table 1. For the ATLC and EPAC, there were no objective intensity models before 1988 that provided 72-h forecasts. The SHIPS model is described by DeMaria et al (2005), and has undergone many changes since 1991, the most significant of which are the inclusion of predictors from global model forecast fields (instead of just analyses) in 1997 and the inclusion of over-land decay effects beginning in 2000. The GFDL model was implemented operationally in 1995 (Kurihara et al, 1998), but some experimental real time forecasts were available to NHC beginning in 1992. The GFDL model has also undergone a number of changes, the most significant of which were the addition of a coupled ocean prediction in 2000 and major modifications to the physical parameterizations and initialization in 2003. The climatology and persistence model SHIFOR has been relatively constant, with an updated version (SHIFOR5) implemented in 2001 (Knaff et al, 2003).

For the WPAC, the first models were analogs, simple climatological models (e.g., Sampson et al, 1990) and the climatology and persistence model (Chu, 1994). The three-dimensional prediction system (GFDN) became available in 1995, followed by other limited area prediction models (MM5 in 2000 and the JTYM in 2001). These were followed by the development of a new climatology and persistence model (Knaff et al, 2003) and more sophisticated statistical model (STIPS) that is similar to SHIPS (Knaff et al, 2005). For the WPAC, an ensemble-based version of STIPS has recently been run in real time (Sampson et al, 2006) and has shown promise.

There are several other intensity forecast guidance methods that are not included in Table 1. These include the forecasts from operational global models, which tend to have errors larger than the models included in Table 1. Other techniques include the Florida state super-ensemble (Mackey et al, 2005), a version of SHIPS with input from microwave imagery (Jones et al, 2006) and a simple consensus forecasts (Sampson et al, 2006; Franklin, 2006). These models are showing promise, but have not yet been transitioned to operations. The Dvorak classification technique (Dvorak, 1975) also provides a short-term intensity forecast, which has been used by JTWC and NHC. However, these forecasts are not available in the ATCF, and do not provide predictions beyond 24 h.

3. Forecast metrics

The traditional method for evaluating intensity forecasts is to calculate the MAE between the predicted maximum sustained surface winds and that from the best track, which is the best estimate of the observed intensity based upon a post-storm analysis of all available information. The MAE was calculated for each of the three TC basins on a yearly basis for the time periods described in the previous section. For brevity, the MAE is evaluated for the 24, 48 and 72-h forecasts, even though the forecast centers also make predictions at 12 and 36 h. The MAE at 96 and 120 h since 2001 were also calculated and will be used to determine the current level of intensity forecast skill as described below.

Forecast skill is defined as the improvement over some baseline. For TC forecasts, the baseline is usually determined from forecasts based upon simple statistical models with parameters from climatology and persistence as input. This input includes the current position and intensity and their time tendencies, and the current date. In Table 1, the SHIFOR, SHIFOR5, STIFOR and ST5D models could be used as a baseline. In Sect. 4, the skill (S) of the intensity forecasts will be calculated using

$$\mathrm{S} = 100(\mathrm{E_b} - \mathrm{E_m})/\mathrm{E_b}, \qquad (1)$$

where $\mathrm{E_b}$ is the MAE from the baseline model and $\mathrm{E_m}$ is the MAE of the model being evaluated. The skill S in (1) is the percentage improvement in the error of the model relative to the error of the baseline, where positive S represents forecast skill.

As described in the Introduction, it is sometimes difficult to evaluate small trends in forecast errors because of the year-to-year variability in forecast difficulty. For track forecasts, climatology and persistence baseline models have been used to help account for the forecast difficulty. A problem with that approach in this study is that the baseline models were not available over the entire time periods being evaluated. In principle, it would be possible to re-run the baseline models using best track input. However, some of the forecasts at the earlier time periods were used to develop the baseline models. Thus, the early time periods would be dependent runs while the later time periods would be independent runs, which would further complicate the evaluation of trends. Because of these problems, a new method is proposed to help account for inter-annual forecast difficulty that is based on how much the model forecast reduces the variance of the observed intensity changes. The variance reduction metric (VR) is defined as

$$\mathrm{VR} = 100({\sigma_\mathrm{o}}^2 - {\sigma_\mathrm{e}}^2)/{\sigma_\mathrm{o}}^2, \qquad (2)$$

where

$$\sigma_\mathrm{o}^2 = \frac{1}{N}\sum_{n=1}^{N}(\Delta V_n - \overline{\Delta V_n})^2, \qquad (3)$$

$$\sigma_\mathrm{e}^2 = \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}_n^2, \qquad (4)$$

N is the number of forecasts in a given year, ΔV_n is the observed intensity change for an individual forecast, $\overline{\Delta V_n}$ is the annual mean intensity change for a given forecast interval, and E_n is the forecast error (the difference between the predicted and observed intensity change) for an individual forecast. If the forecasts were perfect, $\sigma_\mathrm{e}^2 = 0$ and $\mathrm{VR} = 100\%$. Thus, the forecasts eliminate all of the variance of the observed intensity changes. If ${\sigma_\mathrm{e}}^2 = {\sigma_\mathrm{o}}^2$ then $\mathrm{VR} = 0$ and the forecasts did not reduce the variance of the observed intensity changes. If the model forecasts are very poor it is possible for VR to be negative. In this case, the model increases the variance of the observed intensity changes. Note that (4) does not include the subtraction of $\overline{\mathrm{E}}$ inside the summation, as is usually included in the definition of variance. This factor is omitted because it would correct for the bias of the model forecasts, but the metric should penalize forecasts that have biases.

4. Intensity forecast analysis

4.1 Mean absolute error and variance reduction trends

Figure 1 shows the long-term trends in the MAE for the ATLC, EPAC and WPAC along with linear trend lines. In the ATLC all of the trend lines have a slight downward slope. In the EPAC and WPAC, there are small downward trends at 48 and 72 h, suggesting there has been some modest improvement. To determine the statistical significance of the trends, a one-sided *t*-test was performed on the slope of the regression line. Because the slopes are not very steep, a marginally significant level (80%) and highly significant level (95%) were utilized. Table 2 shows the slope values of the MAE trend lines and the re-

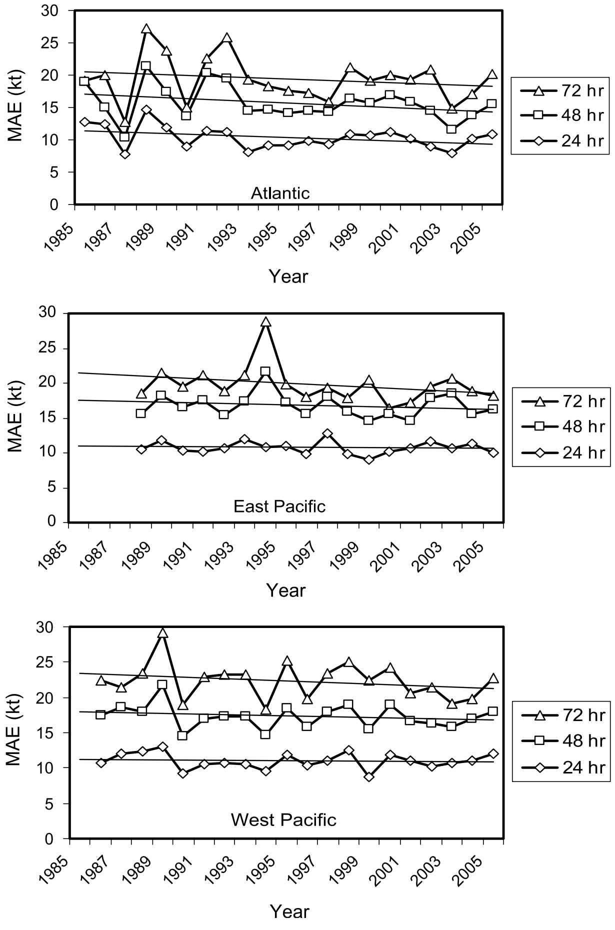


Fig. 1. The time evolution of the mean absolute error of operational intensity forecasts with linear trend lines

sults of the statistical significance tests. These results show that the downward trends are fairly small (~ 0.1 kt per year), but several are marginally significant. Although 0.1 kt is well below the noise level of the individual intensity estimates, the accumulated improvement over two decades is 2 kt, which should be detectable from a large sample of cases with individual accuracies on the order of 5 kt.

As summarized in Table 1, intensity guidance models beyond simple climatology and persistence techniques became available at NHC and JTWC in the early to mid-1990s. Figure 1 shows that the inter-annual variability in the MAE appears to decrease during this same time period, especially for the ATLC. The routine availability of this intensity guidance may have helped to eliminate the years with very large average errors, even though the effect on the downward trend of MAE is marginal.

Figure 2 shows the time evolution of the variance reduction due to the NHC and JTWC forecasts. For the first half of the ATLC sample the VR was negative in some cases, indicating that the NHC intensity forecasts increased the intensity change variance. However, there are no negative VR values after 1995, and the VR are generally larger in the second half of the ATLC sample. For the EPAC and WPAC, the VR in the first half of the time series do not show these negative values of VR, and the increasing trend is less obvious. Although the trend lines are not shown in Fig. 1 for the sake of clarity, the slopes of the trend lines and the statistical significance results are shown in Table 2. For the ATLC, the positive slopes of the trend lines are highly significant, consistent with the Fig. 2. The trends in the EPAC and WPAC are also positive, but not as large as for the ATLC. The EPAC and WPAC trends are marginally significant at some time periods. The increased significance of the slopes

Table 2. Slopes of the trend lines of intensity forecast mean absolute error and variance reduction and the results of statistical significance tests

Basin/time (h)	MAE slope (kt per year)	Significance		VR slope (% per year)	Significance	
		80%	95%		80%	95%
ATLC 24	−0.10	Yes	No	1.5	Yes	Yes
ATLC 48	−0.14	Yes	No	2.1	Yes	Yes
ATLC 72	−0.11	No	No	3.3	Yes	Yes
EPAC 24	−0.01	No	No	0.3	Yes	No
EPAC 48	−0.06	No	No	0.4	No	No
EPAC 72	−0.15	Yes	No	0.8	Yes	No
WPAC 24	−0.02	No	No	0.3	Yes	No
WPAC 48	−0.06	Yes	No	0.2	No	No
WPAC 72	−0.11	Yes	No	0.2	No	No

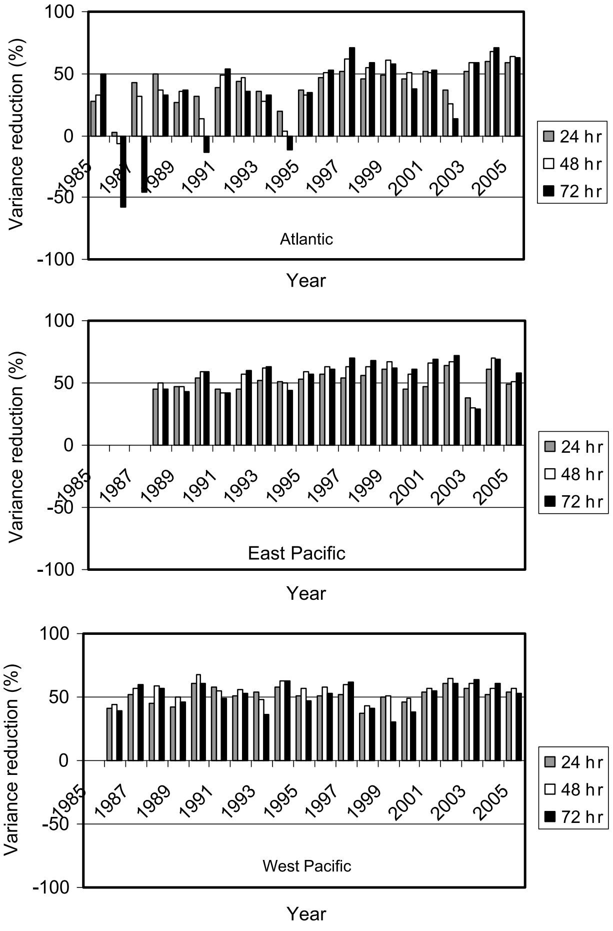


Fig. 2. The time evolution of the intensity variance reduction due to the NHC or JTWC forecasts

in VR for the ATLC compared with MAE shows the value of this metric in the detection of long-term trends of intensity forecasts.

It is speculated that the highly significant increase in the VR in the ATLC, and the more modest improvements in the EPAC and WPAC are due to the improved intensity guidance. To further investigate this possibility, the VR was calculated for each year for a homogeneous sample of cases that had the NHC or JTWC official forecasts and all of the models listed in Table 1 available. In order for a model to be included in the sample for a given year, the forecasts for at least 40% of the official forecasts were required. This fairly low threshold was used because some of the three-dimensional models were only run every other synoptic time in some years. For simplicity, only the 48-h forecast period is considered.

Figure 3 shows the time evolution of the 48-h VR for the official forecasts for each basin, and that from the corresponding best model. Table 3 shows which model provided the largest VR. For 1985–1987 in the ATLC, there were no intensity guidance models that produced a 48-h forecast. During this period the VR of the NHC subjective forecasts from 0 to ~40%. For the period 1988–1995 in the Atlantic, the simple SHIFOR model provided the maximum VR during most years, and the NHC forecasts were able to match or exceed that provided by SHIFOR. This situation changed in 1996–2005 when the SHIPS and GFDL models provided larger values of VR, and the NHC forecasts roughly matched the VR of these models. This result suggests that the significant slope in the trend line of VR in the ATLC (Table 2) was due to the improved intensity guidance models.

The trend in the VR in Fig. 3 for the EPAC is quite different than that in the ATLC. The simple SHIFOR model has much larger values of VR than in the ATLC. This is perhaps not too surprising because the East Pacific storms have fewer complications due to the interaction with land, extra-tropical transition, and re-curvature into the westerlies. Also, the sea surface temperature structure is less complicated in the east Pacific than in the Atlantic. Table 3 shows that in the latter part of the sample (1997–2005), the GFDL and SHIPS model provided larger values of VR than SHIFOR in most years, but the increase was much less dramatic than in the ATLC. Thus, the significant trend in the VR in the ATLC did not occur in the EPAC.

The trend in the VR in the WPAC is different than trends in both the ATLC and EPAC. In the early part of the sample (1986–1993), the simple CLIM model had negative or very small values of VR. Despite the lack of objective guidance, the subjective intensity forecasts from JTWC still had VR values of around 50% during this period. As shown in Table 3, the GFDN, STIFOR and STIPS models provided better guidance, and likely helped to increase the VR of the JTWC forecasts in the past few years. Thus, the fairly high VR values of the JTWC intensity forecasts in the early part of the time series (without much objective guidance) made the slopes of the trend lines only marginally significant at best. Because the JTWC and best model VR for the WPAC at the end of the time series are highly correlated, further improvements in the intensity guidance should lead to improved JTWC intensity forecasts.

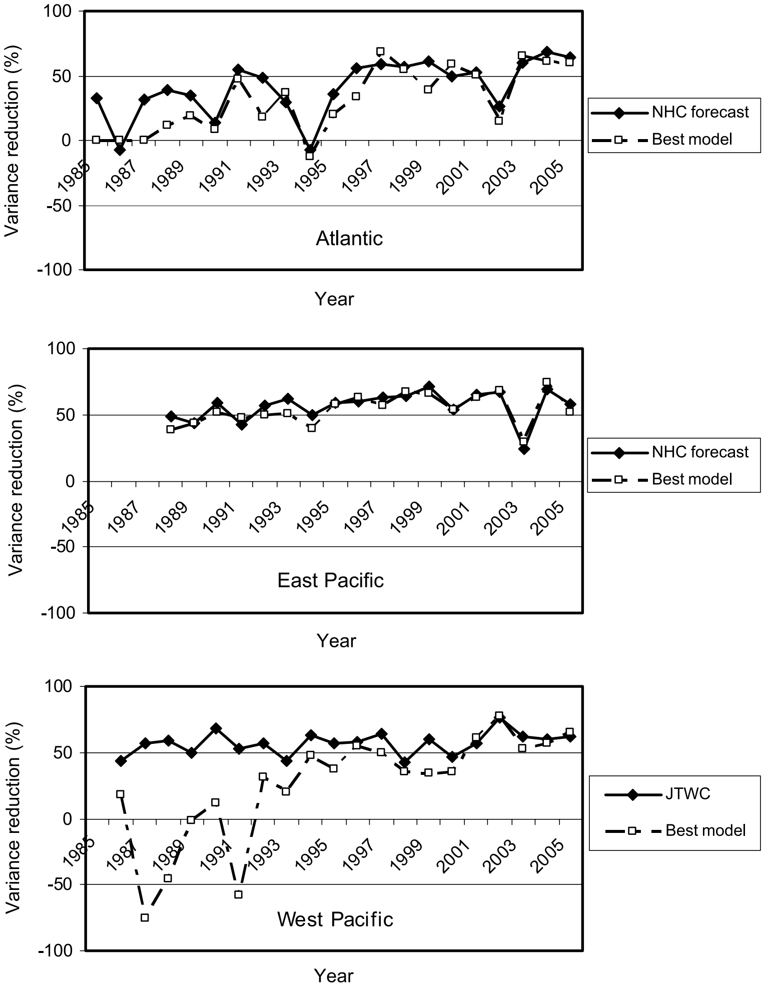


Fig. 3. The time evolution of the variance reduction of the NHC or JTWC official 48-h intensity forecast and that from the best guidance model

Table 3. The intensity guidance model with the highest variance reduction for the 48-h forecast for each year

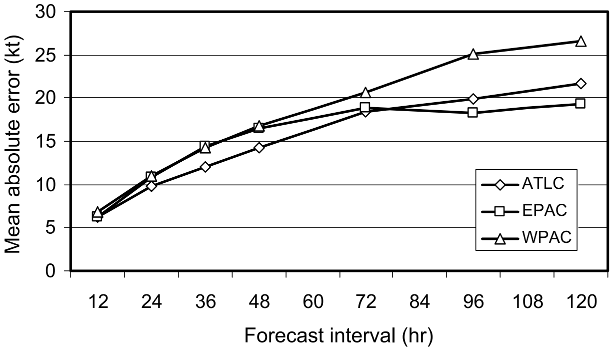
	Atlantic	East Pacific	West Pacific
1985	None	–	–
1986	None	–	CLIM
1987	None	–	CLIM
1988	SHIFOR	SHIFOR	CLIM
1989	SHIFOR	SHIFOR	CLIM
1990	SHIFOR	SHIFOR	CLIM
1991	SHIPS	SHIFOR	CLIM
1992	SHIFOR	SHIFOR	CLIM
1993	SHIFOR	SHIFOR	CLIM
1994	SHIFOR	SHIFOR	CLIM
1995	SHIFOR	SHIFOR	CLIM
1996	SHIPS	SHIFOR	GFDN
1997	GFDL	SHIPS	STIFOR
1998	SHIPS	SHIPS	STIFOR
1999	SHIPS	SHIPS	STIFOR
2000	SHIPS	SHIFOR	STIFOR
2001	SHIPS	SHIPS	STIFOR5
2002	SHIPS	GFDL	STIFOR5
2003	SHIPS	SHIFOR	STIPS
2004	SHIPS	GFDL	STIPS
2005	SHIPS	SHIPS	ST10

4.2 Forecast skill

The above results indicate that there has been some marginal improvements in the NHC and JTWC intensity forecasts over the past two decades. To determine if the recent forecasts have skill, their errors are compared with those from the 5-day versions of the SHIFOR and STIPS models using (1). The 5-year period from 2001–2005 was used because the 5-day forecasts were available during this period.

Figure 4 shows the MAE of the intensity forecasts at 12–120 h from the 5-year sample for each basin. The EPAC and WPAC errors are comparable through 48 h, but the WPAC errors are larger at later forecast times. This might be due to the fact that the EPAC storms do not stay as intense for as long as the WPAC systems due to the movement over cold water. The ATLC errors are smallest initially, but lie between the EPAC and WPAC errors at 96 and 120 h.

Figure 4 also shows the skill of the intensity forecasts. The statistical significance of the dif-



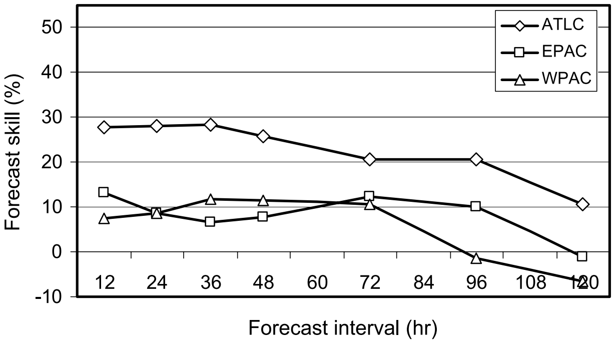


Fig. 4. The mean absolute error of the 2001–2005 NHC or JTWC intensity forecasts for the Atlantic, East Pacific and West Pacific (upper) and the corresponding forecast skill (lower)

ference between the means of the SHIFOR/ STIFOR errors and those from NHC/JTWC was determined using a standard *t*-test, where the sample size was adjusted for serial correlation using the method described by Franklin and DeMaria (1992). The 95th percentile was again used as the threshold for high significance. Figure 4 shows that the intensity forecasts in the ATLC were skillful out to 120 h, the EPAC were skillful to 96 h and the WPAC to 72 h. This skill was highly significant out to 72 h in the EPAC and WPAC, and out to 96 h in the ATLC. The skill of the EPAC and WPAC forecasts are comparable out to 72 h, but the ATLC skill is much higher. It is possible that the increased skill in the ATLC is due to the interaction with land. The baseline SHIFOR5 model does not include land effects, and there were many landfalls included in the ATLC sample, especially during 2003–2005.

5. Comparison of track and intensity errors

The results in Fig. 4 show that the recent intensity forecast have skill out to about 72 h. The skill of the track forecasts were also calculated for this same 2001–2005 time period, where the mean absolute distance errors from NHC and JTWC were normalized with the corresponding errors from the 5-day version of the climatology and persistence track models (CLIPER; Aberson, 1998; Aberson and Sampson, 2003) for each basin. Figure 5 shows that all three basins have a high level of track skill, which was highly significant at every forecast interval. The skill of the ATLC and WPAC track errors are comparable, with the EPAC being a little lower. This difference is due to the fact that the CLIPER errors are smaller for the EPAC, again because of the lack of re-curving storms. Comparing Figs. 4 and 5 shows that for the ATLC, the track and intensity skill is similar at 12 h, but by 72 h, the track skill is a factor of 2 larger than the intensity skill. For the EPAC and WPAC, the track skill at 72 h is 3.3 and 4.7 times larger than the intensity skill, respectively.

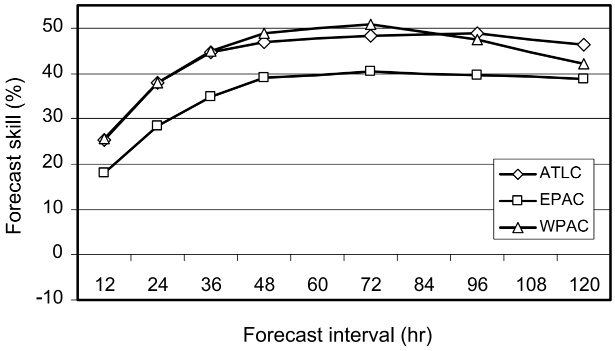


Fig. 5. The skill of the NHC (ATLC and EPAC) or JTWC (WPAC) track forecasts for 2001–2005

The linear trend lines of the 24, 48 and 72-h track forecast errors were also calculated for each basin, using the same years as for the in-

Table 4. Intensity and track forecast MAE trend line slopes in terms of percentage change per year

Basin/time (h)	Intensity slope	Track slope
ATLC 24	−1.0	−2.7
ATLC 48	−0.9	−3.4
ATLC 72	−0.6	−4.0
EPAC 24	−0.1	−2.1
EPAC 48	−0.4	−2.5
EPAC 72	−0.8	−2.8
WPAC 24	−0.2	−2.8
WPAC 48	−0.3	−3.6
WPAC 72	−0.5	−3.9

tensity trend analysis. To compare the track and intensity trends, the slopes of the trend lines were converted to a percentage per year, using the sample mean error at each forecast interval for each basin. The trend line slopes in terms of percentages are shown in Table 4. Using the same t-test as for intensity, the slopes of the track error trends were highly significant for every forecast interval in every basin. Table 4 shows that the intensity forecast improvement was at most 1% per year, which was for the Atlantic basin at 24 h. In contrast, the track forecast improvements ranged from 2 to 4% per year. In many cases, the track forecast improvements are almost an order of magnitude larger than those of the intensity forecasts. Thus, intensity forecasts have a long way to go, relative to the track forecasts.

6. Summary and discussion

The National Hurricane Center and Joint Typhoon Warning Center operational tropical cyclone intensity forecasts for the three major northern hemisphere tropical cyclone basins (Atlantic, eastern North Pacific, western North Pacific) for the past two decades were examined for long-term trends. Results show that there has been some marginal improvement in the mean absolute error at 24 and 48 h for the Atlantic and at 72 h for the east and west Pacific. The improvement in terms of the new metric that measures the variance of the observed intensity changes that is reduced by the forecast (variance reduction, VR) was more significant in the Atlantic. An examination of the VR for the intensity guidance models suggests that the modest improvements were due to the implementation of advanced statistical intensity prediction models (SHIPS and STIPS) and the operational version of the GFDL hurricane model in the mid-1990s. In the first part of the record (from the mid-1980s to mid-1990s), the operational intensity models consisted of fairly simple statistical techniques, which were largely ineffective. During this period, the subjective NHC and JTWC were generally much better than the guidance in terms of VR. In the latter half of the sample, however, the official intensity forecasts have VR values very similar to that of the intensity guidance. This result indicates that the current intensity guidance has utility and is driving the NHC and JTWC intensity forecasts so that improved models will lead to improved operational forecasts.

The skill of the operational intensity forecasts for the 5 year period ending in 2005 was evaluated by comparing the errors to those from simple statistical models with input from climatology and persistence. The intensity forecasts had significant skill out to 96 h in the Atlantic and out to 72 h in the east and west Pacific. These results show that some modest improvement has been made in operational intensity forecasting, and the predictions are now skillful.

To put these results in perspective, the intensity forecasts were compared to the track forecasts for the same data sample. The skill was comparable at 12 h, but the track forecasts were 2 to 5 times more skillful by 72 h, with the largest ratio in the west Pacific. The track and intensity forecast error trends for the two-decade period were also compared. Results showed that the percentage track forecast improvements were almost an order of magnitude larger than those for intensity, indicating that intensity forecasting still has a very long way to go.

It is not surprising that the intensity forecast improvements have lagged behind the track improvements because a much wider range of processes must be accurately modeled to accurately predict intensity. The storm inner core structure, microphysical processes, air-sea energy exchanges, the ocean response, the interaction with land and the larger scale environment, and radiative effects can all impact intensity changes (e.g., Wang and Wu, 2004). To accurately model all of these processes will require an advanced coupled-ocean atmospheric prediction system with proper vertical and horizontal resolution and a data assimilation system that can utilize all available information, including in situ and remotely sensed observations in the inner core. The next generation national centers for environmental prediction hurricane model (the hurricane weather research and forecast (H-WRF) model), which will replace the GFDL model, has plans to include all of these factors. It remains to be seen if this new modeling system will provide significant intensity forecast improvement relative to statistical models, analogous to the transition that occurred for hurricane track forecasting in the 1990s.

Acknowledgments

The views, opinions, and findings in this report are those of the authors and should not be construed as an official NOAA and or U.S. Government position, policy, or decision.

References

Aberson SD (1998) Five-day tropical cyclone track forecasts in the North Atlantic basin. Wea Forecast 13: 1005–1015

Aberson SD, Sampson CR (2003) On the predictability of tropical cyclone tracks in the Northwest Pacific basin. Wea Forecast 131: 1491–1497

Chu J-H (1994) A regression model for the western North Pacific tropical cyclone intensity forecasts. NRL Memo. Rep. 7541-94-7215, Naval Research Laboratory, 33 pp [available from Naval Research Laboratory, 7 Grace Hopper Avenue, Monterey, CA 93943-5502, USA]

DeMaria M, Gross JM (2003) Hurricane! Coping with disaster. Chapter 4: Evolution of Tropical Cyclone Forecast Models (Simpson R, ed). American Geophysical Union, 360 p

DeMaria M, Mainelli M, Shay LK, Knaff JA, Kaplan J (2005) Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). Wea Forecast 20: 531–543

Dvorak VF (1975) Tropical cyclone intensity analysis and forecasting from satellite imagery. Mon Wea Rev 103: 420–430

Franklin JL (2006) 2005 National Hurricane Center forecast verification report [available online at http://www.nhc.noaa.gov/verification/pdfs/verification_2005.pdf]

Franklin JL, DeMaria M (1992) The impact of omega dropwindsonde observations on barotropic hurricane track forecasts. Mon Wea Rev 120: 381–391

Jones TA, Cecil DJ, DeMaria M (2006) Passive microwave-enhanced statistical hurricane intensity prediction scheme. Wea Forecast 21: 613–635

Knaff JA, DeMaria M, Sampson B, Gross JM (2003) Statistical, five-day tropical cyclone intensity forecasts derived from climatology and persistence. Wea Forecast 18: 80–92

Knaff JA, Sampson CR, DeMaria M (2005) An operational statistical typhoon intensity prediction scheme for the western North Pacific. Wea Forecast 20: 688–699

Kurihara Y, Tuleya RE, Bender MA (1998) The GFDL hurricane prediction system and its performance in the 1995 hurricane season. Mon Wea Rev 126: 1306–1322

Mackey BP, Biswas MK, Krishnamurti TN (2005) Performance of the Florida State University Superensemble during 2004, Presentation at the 59th Interdepartmental Hurricane Conference, Jacksonville, FL [available at http://www.ofcm.gov/ihc05/Presentations/02%20session2/s2-10mackey.ppt]

McAdie CJ, Lawrence MB (2000) Improvements in tropical cyclone track forecasting in the Atlantic basin, 1970–98. Bull Amer Meteor Soc 81: 989–998

Neumann CJ (1981) Trends in forecasting the tracks of Atlantic tropical cyclones. Bull Amer Meteor Soc 62: 1473–1485

Park K, Zou X (2004) Toward developing an objective 4DVAR BDA scheme for hurricane initialization based on TPC observed parameters. Mon Wea Rev 132: 2054–2069

Sampson CR, Schrader AJ (2000) The automated tropical cyclone forecasting system (Version 3.2). Bull Amer Meteor Soc 81: 1131–1240

Sampson CR, Knaff JA, DeMaria M (2006) A statistical intensity model consensus for the Joint Typhoon Warning Center. AMS 27th Conf. on Hurricanes and Tropical Meteorology, 24–28 April, Monterey, CA

Sampson CR, Miller RJ, Kreitner RA, Tsui TL (1990) Tropical cyclone track objective aids for the microcomputer: PCLM, XTRP, PCHP. Naval Oceanographic and Atmospheric Research Laboratory, Tech Note 61, 15 pp [available from Naval Research Laboratory, 7 Grace Hopper Avenue, Monterey, CA 93943-5502]

Wang Y, Wu C-C (2004) Current understanding of tropical cyclone structure and intensity changes – a review. Meteorol Atmos Phys 87: 257–278

Corresponding author's address: Mark DeMaria, NOAA/NESDIS/ORA, CIRA/CSU, West Laporte Avenue, Fort Collins, CO 80525, USA (E-mail: mark.demaria@noaa.gov)