Australian Meteorological and Oceanographic Journal 58 (2009) 19-27

Southern hemisphere tropical cyclone intensity forecast methods used at the Joint Typhoon Warning Center,
Part III: forecasts based on a multi-model consensus approach

Charles R. Sampson
Naval Research Laboratory, Monterey, California, USA
and
John A. Knaff
NOAA/NESDIS, Center for Satellite Research and Applications, Fort Collins, Colorado, USA

(Manuscript received June 2008;
Revised October 2008)

Consensus forecasts aids, those derived from forecasts from several models, and ensemble forecasts aids, those derived from several forecast created by a single model, have become commonplace. Consensus forecast aids are now an integral part of operational tropical cyclone forecasting at the Joint Typhoon Warning Center (JTWC). These consensus aids generally have lower average errors than individual forecast aids and benefit from the skill and independence of their members. This paper discusses the performance of one ensemble forecast aid and one consensus forecast aid run in real-time and made available to JTWC during the 2007 and 2008 southern hemisphere tropical cyclone seasons. The ensemble forecast aid is shown to be as skillful as the top performing ensemble member at each forecast time. The consensus forecast aid is shown to be the most skillful aid available to JTWC during the 2007 and 2008 seasons. Further experiments indicate that adding more forecast aids to the intensity consensus may marginally improve both skill and forecast availability of the consensus forecast aid.

Introduction

In this paper, combinations of tropical cyclone forecasts are discussed. These combinations have been referred to as consensus methods and/or ensemble methods and those terms are often used interchangeably in the literature. To offer some clarity to readers of this paper, the authors provide definitions for these terms. The term 'consensus' refers specifically to the combination of forecasts made by different models or techniques, while the term 'ensemble' refers to the combination of forecasts made with the same technique or model. These definitions will be used throughout this manuscript.

The meteorological community recognized the benefits of consensus forecasting as far back as the 1970s (Sanders 1973; Thompson 1977). A subjective form of consensus forecasting has been applied to tropical cyclone track forecasting for decades, and more recently objective consensus methods have become popular (Burton et al. 2007). The more successful consensus forecasting efforts focused on combining dynamical track models (Goerss 2000; Williford et al. 2003) because track models have outperformed intensity forecasts (DeMaria et al. 2007).

Corresponding author address:
Charles R. Sampson, NRL,
7 Grace Hopper Ave. Stop 2, Monterey, CA 93943-5502, USA
Email: sampson@nrlmry.navy.mil

Some of these dynamical track models also produce forecasts of tropical cyclone intensity (maximum 1-min mean wind at 10 m elevation). However, most of the dynamic models cannot simulate the inner core dynamics of a tropical cyclone because of limited horizontal resolution, inadequate initialization and inappropriate parameterizations of the smaller scale processes (Knaff et al. 2007). Consequently, the only skillful intensity forecast models are high-resolution numerical models and statistical-dynamical models designed specifically for tropical cyclone forecasting (DeMaria et al. 2007).

Dynamical models that routinely produced intensity forecasts for the southern hemisphere and were available to the Joint Typhoon Warning Center (JTWC) during the 2006 through 2008 seasons are: the Naval Operational Global Atmospheric Prediction System (NOGAPS; Hogan and Pauley 2007), the Geophysical Fluid Dynamics Laboratory Hurricane Prediction System run with NOGAPS initial and boundary conditions (GFDN; Rennick 1999), the United Kingdom Meteorological Office global model (UKM; Heming 2008), the National Weather Service (NWS) global spectral model (GFS; Moorthi et al. 2001), the fifth-generation Pennsylvania State University/National Center for Atmospheric Research Mesoscale Model (MM5; Grell et al. 1995) run operationally by the US Air Force Weather Agency (AFWA), and finally the TC-Limited Area Prediction System (TC-LAPS; Davidson and Weber 2000) and the Tropical eXtended Area Prediction System (TXLAPS; Australian Bureau of Meteorology 2005) run by the Australian Bureau of Meteorology. Table 1 provides a summary of these and other aids examined in this study. A more exhaustive overview and reference list for many of these models is discussed in Heming and Goerss (2009).

Table 1. Objective tropical cyclone intensity guidance techniques available at the Joint Typhoon Warning Center between 2006 and 2008, its interpolated aid, a brief description, and the year of first availability. + STIPS ensemble members were run at NRL in real-time, but only a descriptive text file was delivered (via email) to JTWC.

Model	JTWC Aid ID	Interpolated	Description	First available
NOGAPS	NGPS	NGPI	US Navy global model	2004
UKM	UKM	UKMI	Met Office global model	2003
GFS	AVN	AVNI	NWS global model	2002
GFDN	GFDN	GFNI	Geophysical Fluid Dynamic Lab initialized by the navy Operational Global Analysis and Prediction system model	1998
TC-LAPS	TCLP	TCLI	Australian TC-Limited Area Prediction System	2002
TX-LAPS	TXLP	TXLI	Australian Tropical eXtended Area Prediction System	2005
US Air Force regional model	AFW1	AFWI	Air Force mesoscale model	2002
5-day Statistical Intensity Forecast	ST5D	None	Statistical model	2004
PEST	PEST	None	Consensus and probability aid	2005+
AFS1	AFS1	None	STIPS ensemble member	2006+
AVS1	AVS1	None	STIPS ensemble member	2006+
GFS1	GFS1	None	STIPS ensemble member	2006+
NGS1	NGS1	None	STIPS ensemble member	2006+
TCS1	TCS1	None	STIPS ensemble member	2006+
UKS1	UKS1	None	STIPS ensemble member	2006+
WBS1	WBS1	None	STIPS ensemble member	2006+
STIPS ensemble average	ST10	None	Average of all STIPS ensemble member forecasts	2006
ST11	ST11	None	Multi-model consensus that combines the ensemble members of ST10 and GFNI	2007
ST12	ST12	None	ST10 members, GFNI and CHII	Not Available
ST13	ST13	None	ST10 members, GFNI, CHII and TCLI	Not Available
ST14	ST14	None	ST10 members, GFNI, CHII, TCLI and UKMI	Not Available
CHIPS	CHIP	CHII	Coupled dynamical hurricane model	2003

Two simple models are designed specifically to produce intensity forecasts. The Southern Hemisphere Statistical Typhoon Intensity Prediction System (SH STIPS; Knaff and Sampson 2009a), as described in Part II, is a statistical-dynamical model that forecasts changes in intensity through regression of large-scale environmental parameters (e.g. vertical wind shear, sea surface temperature, relative humidity, temperature, and low-level vorticity) and an empirical inland decay model (DeMaria et al. 2006). Even though SH STIPS has lower mean absolute error than the NWP models, it does not forecast rapid intensification (Knaff et al. 2007) because statistical models are designed to minimize variance over the entire development data set. An ensemble version of the western North Pacific, STIPS (Knaff et al. 2005), has been implemented for the western North Pacific with limited success in that its forecasts perform as well as its most skillful member (Sampson et al. 2008). A similar methodology has been applied to create an ensemble using SH STIPS that is described in detail in Appendix A. Throughout the remainder of this paper the STIPS ensemble refers to the southern hemisphere version unless stated otherwise.

The Coupled Hurricane Intensity Prediction System (CHIPS; Emanuel et al. 2004) is the second model designed specifically to produce intensity forecasts, and unlike STIPS, it attempts to model the inner core region. Inputs to this model now include a thermodynamic state, vertical wind shear, sea surface temperature, climatological mixed layer depth and sub-mixed layer thermal stratification (Emanuel et al. 2008).

A simple climatology and persistence statistical model called the 5-day Statistical Hurricane Intensity Forecast (ST5D; Knaff and Sampson 2009b) is used as the intensity skill baseline. Although this model is a poor predictor of rapid intensification and decay, its seasonal average performance is still competitive. Other statistical intensity aids (e.g. climatology, climatology and persistence, analogs, extrapolation and hybrids) exist, but they do not perform as well as ST5D (Knaff and Sampson 2009b) and are not discussed further.

The first purpose of this study is to assess the skill of the existing guidance available to JTWC for forecasting southern hemisphere tropical cyclones. The second purpose is to determine whether superior skill can be obtained using a simple equally weighted consensus of the best performing members. Intensity forecasts for the northern hemisphere have been shown to have relatively little skill compared to skill baselines (Sampson et al. 2008; DeMaria et al. 2007), so it is likely that the skill of intensity forecast aids in the southern hemisphere is also low and that the benefits of equally weighted consensus aids will be small. Still, it is important to perform a study that sets an intensity consensus baseline to assess further improvements.

Data

The data used for this study are from the operational archive at the JTWC as stored on the Automated Tropical Cyclone Forecasting System (Sampson and Schrader 2000), for which a description is given in JTWC (2008). For this study a large homogenous data set is desirable so that any resulting statistics are stable. The GFDN intensity forecast aid was available as early as 1998, and by 2003 there were five intensity forecast aids available to JTWC (Table 1). However, the skill baseline (ST5D) only became available in 2004 and a few of the better performing aids only became available in 2005, 2006 and 2007. Hence, the period chosen for evaluation in this study was from 1 July 2005 to 30 June 2008.

Methods

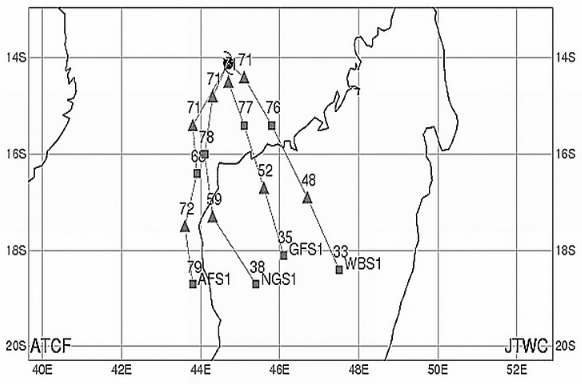
Intensity forecast aids are characterized as being either early or late depending on whether or not they are available to the JTWC forecaster during the forecast cycle. For example, consider the 1200 UTC forecast cycle, which begins with the 1200 UTC synoptic time and ends with the release of an official forecast at 1500 UTC. The 1200 UTC run of the GFDN model is not complete nor is its forecast intensity aid (also named GFDN) available to the forecaster until about 1600 UTC. This is about an hour after the official JTWC forecast is released, and thus the 1200 UTC GFDN would be considered a late forecast aid because it could not be used to prepare the 1200 UTC forecast.

CHIPS and all the dynamical model forecast aids available to JTWC are late models. CHIPS forecasts are judged to be late because they require the current forecast track as input and results are not complete in time to be considered for the intensity forecast, which at JTWC usually occurs immediately after the track forecast is constructed. To alleviate the problem, a simple method is used to take the latest available forecast aid from a run of a late model and adjust it to the current synoptic time and initial conditions. For example, the GFDN forecast aid for hours 6-126 from the previous (0600 UTC) run would be adjusted, or shifted, so that the 6 h forecast (valid at 1200 UTC) would exactly match the observed 1200 UTC position and intensity of the tropical cyclone. If 6 h interpolated forecast aids are not available, 12 h interpolated forecast aids are computed. The 12 h interpolations occur approximately 15 per cent of the time or less for models that are available every six hours.

The adjustment process creates an early version of the GFDN forecast aid for the 1200 UTC forecast cycle that is based on the most current available guidance. The adjustment algorithm is called “the interpolator” and the adjusted aids are called “interpolated” aids. The version of the interpolator used in this study is similar to that described in Sampson et al. (2006). The name of the interpolated forecast aid is usually the acronym of the late forecast aid with an “I” substituted for the last letter (Table 1). One exception used in this study is GFDN, for which GFNI is the acronym for the interpolated forecast aid.

The STIPS ensemble members (AFS1, GFS1, AVS1, NGS1, TCS1, UKS1 and WBS1) and thus the STIPS ensemble (ST10), described in Appendix A, are considered early models. They make use of interpolated tracks available before the official track forecast is produced, are run during the official track forecast construction, and return intensity forecasts to the JTWC forecaster in time for use in the official intensity forecast construction. A benefit of running an ensemble of SH STIPS in the southern hemisphere is that the forecasts are available every six hours and extend beyond 48 h. In contrast, running SH STIPS on the official JTWC track, as was done to evaluate the skill of SH STIPS in Knaff and Sampson (2009a), would produce an aid that is usually available every 12 hours and extends to 48 h. Since the JTWC track is not required to run SH STIPS in this manner, forecasts are also produced for weak tropical systems. These forecasts can serve as tropical cyclone formation guidance. Though this application has not been rigorously verified, the output has been available to JTWC and the preliminary results look promising. Another benefit running STIPS as an ensemble is that it produces a range of solutions. The ensemble size is small (at most seven members in the southern hemisphere), but the members could provide guidance for different forecast scenarios. An example of an ensemble of SH STIPS forecasts is shown for Tropical Cyclone Fame, the 13th tropical cyclone of the 2008 season at 1200 UTC 26 January (Fig. 1). In this example, SH STIPS is run using identical input fields (i.e. forecast fields from NOGAPS), but tracks from four different NWP models. Even with identical input fields, the ensemble member output intensities at 48 h are markedly different. While one of the solutions (AFS1) remains over water for the entire 48 h forecast, the others (NGS1, GFS1 and WBS1) make landfall between 24 and 36 h. The STIPS ensemble (ST10) for this example is an average of the four ensemble member intensities. At 48 h, the ST10 intensity is (79+38+35+33)/4 or 46 kn. One final advantage to running an ensemble is that at least a few of the members (NGS1, AVS1, and UKS1) are run with dynamic forecast fields that are consistent with the ensemble member tracks. This ensures that the shear computations in SH STIPS are computed with an annulus around the tropical cyclone instead of an annulus that potentially intersects the vortex.

Fig. 1 STIPS ensemble members for the 26 January 1200 UTC forecast for Tropical Cyclone Fame, the 13th tropical cyclone of the 2008 JTWC season. Forecast periods are 12, 24, 36 and 48 h and each position is labeled with corresponding STIPS ensemble member intensity forecast (kn). The solid typhoon symbol represents the initial position, and the four characters at the end of each forecast indicate the specific ensemble member.



ST11, a consensus aid which is an average of all available STIPS ensemble members and GFNI, is also an early model. ST11 was implemented and available to JTWC for the 2007 and 2008 southern hemisphere seasons based on results from evaluations in the western North Pacific. The 2006 season forecasts were recomputed for this evaluation to increase the dataset. The other consensus aids in Table 1 that are evaluated in this study (ST12, ST13 and ST14) were not available operationally. These consensus aids however were computed using equally weighted operationally available and independent forecasts (i.e. consensus members) and thus their evaluation statistics are representative of an actual operational capability. The authors intend to evaluate these consensus forecast aids based on real-time independent data in the near future.

All the ensemble and consensus forecasts discussed below are equally weighted averages of forecasts from the available members at each forecast period (12, 24, 36, 48, 72, 96 and 120 h). An ensemble or consensus forecast is only computed if two or more members exist for a given forecast period. If fewer than two members exist, the ensemble or consensus computation is aborted for this and subsequent time periods.

The interpolated, ensemble and consensus forecasts presented here have been recomputed using methods described above to ensure that all results are computed using the same version of the interpolator. Average differences in performance between recomputed interpolations and those produced in operations are generally less than 1 per cent.

Forecasts are verified only when the best track intensity (1-minute sustained) is greater than 20 kn (10.3 m/s) and only when the system is a tropical or subtropical cyclone. Performance is discussed through the use of skill charts. The measure of skill in these charts is defined as:

$$\text{Skill} = 100^{*}(\text{baseline error}-\text{model error})/\text{baseline error} \quad \ldots(1)$$

Thus, skill is positive when the forecast aid error is less than that of the baseline forecast aid, which in this study is the climatology and persistence technique ST5D. A one-tailed

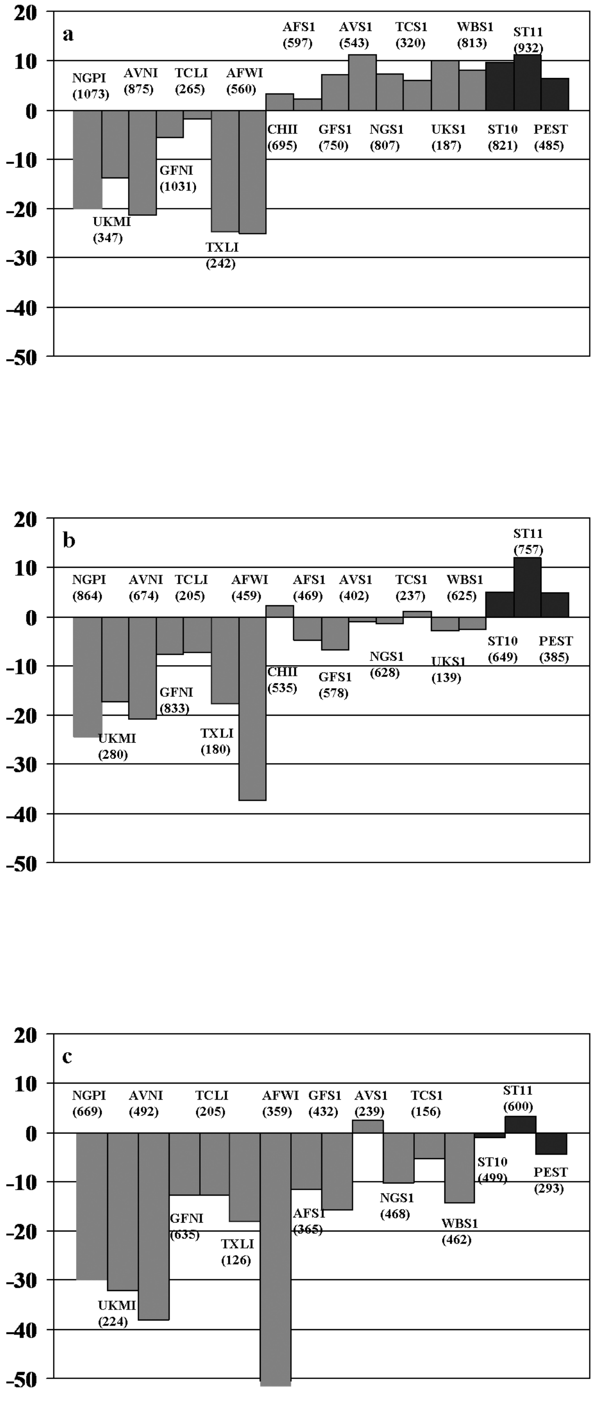
Student's t-test at the 95 per cent level is used to test the statistical significance of forecast error differences between intensity forecast aids. The effective sample size used for the Student's t test is estimated to be the number of 30 h samples contained in the dataset, which is described as the time between effectively independent samples (Leith 1973).

Results

A verification of the various intensity forecast aids available during the 2006 through 2008 seasons is shown in Fig. 2. In general, the intensity forecast skill is less than the skill associated with track forecasts. For example, GFNI track forecast skill at 48 h is approximately 25 per cent while its intensity forecast skill at 48 h is negative. However, a number of skillful intensity aids are available, and many of them are members of the STIPS ensemble described above and in more detail in Appendix A. Three of the best performers at all forecast times are consensus and ensemble methods. PEST (Weber 2005) is one of these consensus methods and it performs well, as expected. Most improvement or degradation seen in PEST relative to the other consensus aids can be attributed to the choice of input models, and results and diagnostics discussed below for the STIPS ensemble (ST10) and the STIPS-based consensus aids (e.g. ST11) would also apply to PEST.

One striking result seen from Fig. 2 is that ST11 performance is superior to ST10 performance even though the only difference is that the lacklustre-performing GFNI is added to the STIPS ensemble to form ST11. The authors attribute this effect to the relative independence between the GFNI aid and the STIPS ensemble members. As proposed in Goerss (2000), a successful consensus or ensemble forecast (i.e. one that outperforms other guidance) is constructed from members that are independent and have relatively small mean forecast errors. For intensity forecasts, the combined effects of these two factors is illustrated by the equation for the consensus mean error $\mu_c=\mu/(n)^{1/2}$, which is based on the central limit theorem and where μ is the mean error of the members (assumed to all be equal to each other) and n is the number of independent members/forecast aids (Sampson et al. 2008). This relationship implies that increasing the number of independent members reduces the mean error of the consensus. In operations, the intensity forecast errors are not entirely independent so n is replaced by the effective number of members or effective degrees of freedom n_e. Two members with errors that are completely independent (n_e = n = 2) can produce a consensus with a mean error reduction of approximately 30 per cent. On the other hand, two members with errors having little independence (e.g. n_e = 1.1) would only produce an improvement of approximately 5 per cent over the member mean. In this and other papers (Goerss 2000; Sampson et al. 2006; Sampson et al. 2008) n_e is computed from the mean error reduction achieved by com-

Fig. 2 Comparison of intensity forecast skill (per cent) for JTWC early forecast aids at (a) 24 h (b) 48 h and (c) 72 h. Data includes 2006-2008 southern hemisphere seasons. Skill is based on ST5D and the comparison is inhomogeneous. Acronyms are defined in Table 1 and Table A1. Number of cases is shown in parentheses. Consensus aids are shaded dark. ST11 consensus was not available to JTWC in 2006, and was recomputed for this evaluation.



bining the members rather than computing the mean error reduction achieved from n_e.

Our investigation of the independence of the STIPS ensemble ST10 is shown in Fig. 3. The results indicate that the STIPS ensemble performs about as well as the top performing members (AVS1 and TCS1), and in head-to-head comparisons outperforms four of the members at the 24, 48, and 72 h forecast periods as indicated by the filled circles. To explore whether the ensemble members lack independence, all possible two-member ensemble combinations (21 in all) were computed for the entire 2006-2008 seasons. The forecast results of each two-member ensemble were verified against the average of the mean forecast errors for its two input members. For two-member ensembles with 300 or more cases the forecast performance improvements in mean forecast errors ranged from 1 per cent to 3 per cent, indicating that independence is quite low (i.e. n_e ranging from 1.02 to 1.06). By comparison, forecast improvements for two-member consensus aids computed from other aids used in this study (300 or more cases) are higher (4 per cent to 8 per cent), which indicates more independence (i.e. n_e ranging from 1.06 to 1.17). These STIPS ensemble independence estimates are consistent with expectations because ensemble members essentially use the same model (SH STIPS), even though they have different track and NWP model input. For instance, the same exercise applied to a track forecasting consensus found the average effective degrees of freedom to be 1.54 for aids that included a generally less skillful barotropic model (WBAI) and 1.34 for aids that did not (Sampson et al. 2006). This result is important to the later discussions of the intensity consensus aids (ST11, ST12, ST13, and ST14) because it suggests that a model with relatively large errors can improve a consensus forecast when its forecasts are relatively independent when compared to the other consensus/ensemble members.

Forecast availability is also an important consideration because a forecast aid that performs well may not be as useful to a forecaster if it is only available for 50 per cent of the official forecasts. The availability of the STIPS ensemble at 48 h (651 cases) is approximately 5 per cent higher than the STIPS ensemble member with the highest availability (NGS1), and its availability is more than double the best performing member at 48 h (TCS1).

Figure 4 shows the impact of adding interpolated forecasts are added to the STIPS ensemble members one at a time, which resulted in the four consensus aids ST11 (ST10 members + GFNI), ST12 (ST11 members + CHII), ST13 (ST12 members + TCLI) and ST14 (ST13 members + UKMI). As mentioned above, ST11 has been available to JTWC since the 2007 season. Normally in a study like this, the best performing aid not already in ST10 (i.e. CHII) would be added to the ensemble first, but here we include GFNI first since that particular aid (ST11) was made available to JTWC in 2007. Immediately apparent from Fig. 4 is that improvements in

Fig. 3 (a) 24 h (b) 48 h and (c) 72 h intensity forecast skill (per cent) with respect to ST5D for 2006-2008 southern hemisphere seasons. Diamonds are individual STIPS ensemble members, circles are STIPS ensemble means for the same cases. Horizontal dashed line is ensemble mean skill when all cases are considered (821, 649, and 499 cases at 24, 48 and 72 h, respectively). Filled circles indicate that for that ensemble member the ensemble mean forecasts are improved in a statistical significance manner. Ensemble member acronyms are defined in Table A1. Number of cases is shown in parentheses.

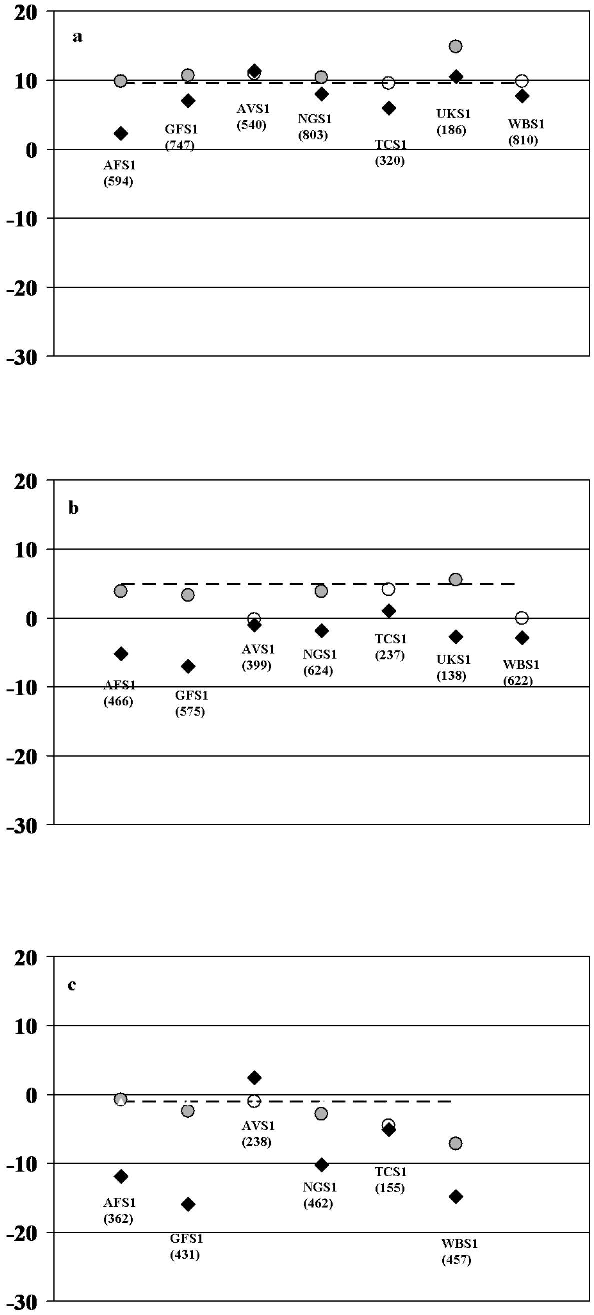
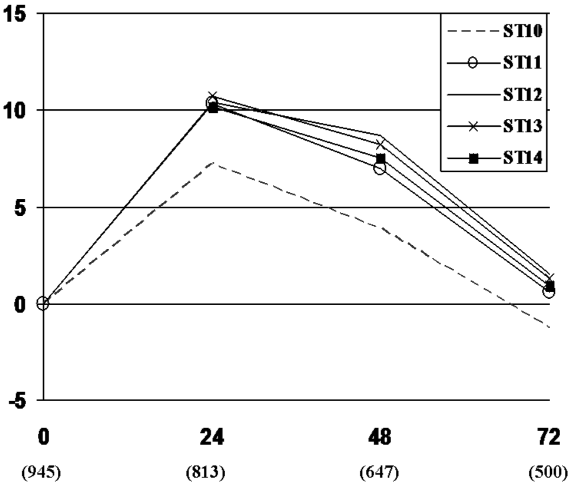


Fig. 4 Intensity forecast skill (per cent) relative to ST5D for one ensemble aid and four multi-model ensemble or consensus aids. Dataset is from the JTWC 2006-2008 southern hemisphere seasons and is homogeneous. Acronyms are defined in Table A1. Number of cases is shown in parentheses.



skill are small. The largest improvement in skill is gained by adding the first interpolated aid (GFNI) with the STIPS ensemble members to yield ST11. These improvements of approximately 2-3 per cent are significant at the 24 and 48 h forecast periods, but not at the 72 h forecast period. Addition of more aids to this consensus to produce ST12 through ST14 yields mixed results, none of which are statistically significant. The top performer at 24 h is ST13, while ST12 is the top performer at 48 and 72 h. The ST10, ST11, ST12, ST13 and ST14 forecasts were available for 74 per cent, 78 per cent, 89 per cent, 90 per cent, and 91 per cent of the 465 JTWC forecasts for 48 h, respectively. So from an availability standpoint, ST12, ST13, and ST14 provide the most reliable guidance.

Summary and conclusions

Intensity guidance available to JTWC during the southern hemisphere 2006-2008 seasons was evaluated against a skill baseline. An equally weighted consensus of STIPS ensemble members and GFNI (ST11) was found to be the top performer, though its skill relative to the statistical baseline (ST5D) was generally low compared to skill found in track consensus aids. Running a STIPS ensemble as described here has advantages over running SH STIPS on a single official forecast track including forecast availability, a range of forecast solutions, and timeliness; however there was no demonstrated improvement in skill. Upon further inspection we found that there was less independence among STIPS ensemble members than there was among aids from two different models (e.g. GFDN and CHIPS), hence there was little gain in skill by averaging the ensemble member forecasts.

Additional experiments were conducted involving the addition of the most skillful of the remaining intensity aids (CHII, TCLI and UKMI). Although the procedure was defined before the analysis, results from these experiments should be considered preliminary until they have been reproduced on real-time independent data. Adding additional models yielded mixed results, none of which were statistically significant. The STIPS ensemble with the addition of GFNI, CHII and TCLI was the top performer at 24 h, and the STIPS ensemble with the addition of GFNI and CHII was the top performer at 48 and 72 h. One advantage of adding CHII and TCLI in the consensus is that they increased the forecast availability by about 10 per cent while producing similar forecast skills. The consensus aids described in this evaluation could serve as benchmarks for other deterministic intensity forecasts from consensus and ensemble methods reviewed by Burton et al. (2007). They may also provide operational forecast guidance. It is suspected that improvements in the consensus members and additional members would further benefit this simple, equally weighted intensity consensus approach.

Expected improvements should mostly come from NWP models with higher resolution, better specification of the tropical cyclone inner core dynamics, better initial conditions, better boundary conditions and other improvements. For example, a coupled version of the GFDN model is now available for evaluation, and there are ongoing improvements in models like TCLAPS that include superior initial and boundary conditions from the soon to be operational Unified Model. There are also improvements in the simplified models. For example, a version of STIPS run in the western North Pacific and that also makes use of ocean heat content as a predictor has shown small but significant increases in skill (Goni et al. 2008). The additional models and improvements to existing models should translate into an improved intensity consensus for use as guidance by the forecasters.

Acknowledgments

The authors would like to acknowledge Ann Schrader for her work with the ATCF, Jim Gross for development of the interpolator, Mike Fiorino for his vortex trackers, Mark DeMaria for his work with the statistical analysis code, and the JTWC for their efforts to maintain the ATCF best track and objective aid files. The authors also acknowledge Jeff Kepert, Noel Davidson and one anonymous reviewer for their careful reviews of the manuscript. This project is supported by grants from the Office of Naval Research and the National Oceanic and Atmospheric Administration. 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 US government position, policy, or decision.

APPENDIX A

The STIPS consensus

The STIPS consensus (ST10) is constructed using seven NWP model interpolated track forecasts available at approximately synoptic time + 1.5 hours. The interpolated track forecasts chosen were the seven members of the southern hemisphere operational track consensus (CONW) used at JTWC for the 2005-2006 seasons. These interpolated tracks then provide forecasts of storm locations from which SH STIPS produces intensity forecasts. The STIPS ensemble members are listed in column one and the forecast track aids used are those listed in column two of Table A1.

Ideally, a STIPS ensemble forecast would be formed using members with thermodynamic and dynamic fields from the model specified by the interpolated track. This would provide the most independence between the members, which could possibly lead to a larger reduction in the consensus mean. It would also provide model fields with a vortex structure consistent with the interpolated track, and thus should provide for more realistic vertical shear computations in SH STIPS. However, a complete suite of model forecast fields was not available for each ensemble member so a compromise solution was constructed. SH STIPS is run with the interpolated model track NGPI and field data from NOGAPS to produce NGS1. STIPS is run with the interpolated track UKMI, UK Met Office model forecast winds and NOGAPS data for all other input (temperature, relative humidity and geopotential height) to create the ensemble member UKS1. SH STIPS is run with the interpolated track AVNI, GFS forecast winds, and NOGAPS fields for all other input to create AVS1. The other four STIPS ensemble members (AFS1, GFS1, TCS1, and WBS1) are run using the interpolated tracks (GFNI, TCLI, and AFWI, respectively) and NOGAPS input fields. The field input for each ensemble member is also summarized in Table A1.

Table A1. STIPS ensemble members. The name of the individual ensemble member is given in the first column. The following columns describe the input data used in the STIPS model to create each of the ensemble members that are used in the ensemble average (ST10). Dynamic forecast fields refer to the specific forecast model that provides the forecasts of the winds. The NOGAPS is used to provide the thermodynamic, moisture, and SST fields for all the members.

ST10 member	Track input	Dynamic forecast fields
AFS1	AFWI	NOGAPS
AVS1	AVNI	GFS
GFS1	GFNI	NOGAPS
NGS1	NGPI	NOGAPS
TCS1	TCLI	NOGAPS
UKS1	UKMI	UKM
WBS1	WBAI	NOGAPS

A western North Pacific version of STIPS was originally installed at JTWC in 2004, and this aid used the JTWC forecast track and NOGAPS input fields. Changes in the JTWC operational configuration in 2006 required that the SH STIPS model, which was developed early in 2006, be run remotely, with the output delivered on a schedule to JTWC. This change made it unlikely that the SH STIPS based on the JTWC forecast track (STFD) would be available in time for use in the JTWC operational intensity forecast. The authors ran STFD at NRL in near real-time for comparison with ST10 during the 2006-2008 southern hemisphere seasons. ST10 and STFD mean absolute intensity forecast errors are within 3 per cent of each other at 12, 24, 36 and 48 h. The most notable difference is that STFD performance is better by 2.1 per cent at the 12 h forecast, which is to be expected since STFD uses the current JTWC track forecasts. On the other hand, the STIPS ensemble actually outperforms STFD by 1-2 per cent by 36 and 48 h, though the results do not pass statistical significance tests.

References

Australian Bureau of Meteorology. 2005. Operational Implementation of TXLAPS_PT375 Bulletin No. 59, May 2005 [Available on-line from http://www.bom.gov.au/nmoc/bulletins/59/].

Burton, A., Caroff, P., Franklin, J., Fukada, E., Lee, T. C., Sampson, B. and Smith, T. 2007. Sharing experiences in operational consensus forecasting. Sixth International Workshop on Tropical Cyclones, San Jose, Costa Rica [Available on-line from http://severe.worldweather.org/iwtc/].

DeMaria, M., Knaff, J. A., and Kaplan, J. 2006. On the decay of tropical cyclone winds crossing narrow landmasses. *J. App. Meteor. and Clim., 45,* 491–9.

DeMaria, M., Knaff, J. A. and Sampson, C. 2007. Evaluation of long-term trends in tropical cyclone intensity forecasts. *Meteor. Atmos. Phys., 97,* 19-28.

Davidson, N. E. and Weber, H. C. 2000. The BMRC high-resolution tropical cyclone prediction system: TC-LAPS. *Mon. Wea. Rev., 128,* 1245-65.

Emanuel, K., Desautels, C., Holloway, C. and Korty, R. 2004. Environmental control of tropical cyclone intensity. *J. Atmos. Sci., 61,* 843-58.

Emanuel, K., Sundarajan, R. and Williams, J. 2008. Results from downscaling IPCC AR4 simulations. *Bull. Amer. Meteor. Soc., 89,* 347-67.

Goerss, J. S. 2000. Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev., 128,* 1187-93.

Goni, G., DeMaria, M., Knaff, J., Sampson, C., Ginis, I., Bringas, F., Mavume, A., Lauer, C., Lin, I., Ali, M., Sandery, P., Ramos-Buarque, S., Kang, K., Mehra, A., Chassignet, E. and Halliwell, G. 2008. Applications of satellite-derived ocean measurements to tropical cyclone intensity forecasting, Final GODAE Symposium 2008, Nice, France. [Available on-line from http://www.godae.org].

Grell, G. A., Dudhia, J. and Stauffer, D. R. 1995. *A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5).* NCAR Tech. Note NCAR/TN-398 + STR, 122 pp.

Heming, J. T. 2008. Tropical cyclone initialization in the Met Office global model: Is it still necessary and can it be improved? Amer. Meteor. Soc.

28th Conference on Hurricanes and Tropical Meteorology, Orlando, USA.

Heming, J. T. and Goerss, J. S. 2009. Track and structure forecasts of tropical cyclones, Chapter 11 of Global Perspectives on Tropical Cyclones - From Science to Mitigation. J. C. L. Chan editor, World Scientific Publishing Company, Singapore (In press).

Hogan, T. F. and Pauley, R.L. 2007. The impact of convective momentum transport on tropical cyclone track forecasts using the Emanuel cumulus parameterization. *Mon. Wea. Rev., 135,* 1195-207.

Joint Typhoon Warning Center (JTWC). 2008. The Annual Tropical Cyclone Reports [Available on-line from https://metocph.nmci.navy.mil/jtwc.php].

Knaff, J., Guard, C., Kossin, J., Marchok, T., Sampson, B., Smith, T., and Surgi, N. 2007. Operational guidance and skill in forecasting structure change. Sixth International Workshop on Tropical Cyclones, San Jose, Costa Rica [Available on-line from http://severe.worldweather.org/iwtc/].

Knaff, J. A., Sampson, C. R. and DeMaria, M. 2005. An operational statistical typhoon intensity prediction scheme for the Western North Pacific. *Wea. Forecasting, 20,* 688-99.

Knaff, J. A., and Sampson, C. R. 2009a. Southern hemisphere tropical cyclone intensity forecast methods used at the Joint Typhoon Warning Center, Part II: forecasts based on a statistical-dynamical approach. *Aust. Met. Oceanogr. J., 58,* 9-18

Knaff, J. A., and Sampson, C. R. 2009b. Southern hemisphere tropical cyclone intensity forecast methods used at the Joint Typhoon Warning Center, Part I: control forecasts based on climatology and persistence. *Aust. Met. Oceanogr. J., 58,* 1-7

Leith, C.E. 1973. The standard error of time-average estimates of climate means. *J. Appl. Meteor., 12,* 1066-9.

Moorthi, S., Pan, H. and Caplan, P. 2001. Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. *Technical Procedures Bulletin 484, National Weather Service*. [Available online at http://www.nws.noaa.gov/om/tpb/.]

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 Schrader, A. J. 2000. The automated tropical cyclone forecasting system (Version 3.2). *Bull. Amer. Meteor. Soc., 81,* 1231-40.

Sampson, C. R., Goerss, J. S. and Weber, H. C. 2006. Operational performance of a new barotropic model (WBAR) in the western North Pacific basin. *Wea. Forecasting, 21,* 656-62.

Sampson, C. R., Franklin, J. L., Knaff, J. A. and DeMaria, M. 2008. Experiments with a simple tropical cyclone intensity consensus. *Wea. Forecasting, 23,* 304-12.

Sanders, F. 1973. Skill in forecasting daily temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc., 54,* 1171-9.

Thompson, P. D. 1977. How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev., 105,* 228-9.

Weber, H. C. 2005. Probabilistic prediction of tropical cyclones, Part II: Intensity. *Mon. Wea. Rev., 133,* 1853-64.

Williford, C. E., Krishnamurti, T. N., Correa-Torres, R., Cocke, S., Christidis, Z. and Kumar, T. S. V. V. 2003. Real-time multimodel superensemble forecasts of Atlantic tropical systems of 1999. *Mon. Wea. Rev., 131,* 1878-94.