Experiments with a Simple Tropical Cyclone Intensity Consensus

CHARLES R. SAMPSON

Naval Research Laboratory, Monterey, California

JAMES L. FRANKLIN

National Hurricane Center, Miami, Florida

JOHN A. KNAFF AND MARK DEMARIA

NOAA/NESDIS/Center for Satellite Applications and Research, Fort Collins, Colorado

(Manuscript received 21 March 2007, in final form 25 July 2007)

ABSTRACT

Consensus forecasts (forecasts created by combining output from individual forecasts) have become an integral part of operational tropical cyclone track forecasting. Consensus aids, which generally have lower average errors than individual models, benefit from the skill and independence of the consensus members, both of which are present in track forecasting, but are limited in intensity forecasting. This study conducts experiments with intensity forecast aids on 4 yr of data (2003–06). First, the skill of the models is assessed; then simple consensus computations are constructed for the Atlantic, eastern North Pacific, and western North Pacific basins. A simple (i.e., equally weighted) consensus of three top-performing intensity forecast models is found to generally outperform the individual members in both the Atlantic and eastern North Pacific, and a simple consensus of two top-performing intensity forecast models is found to generally outperform the individual members in the western North Pacific.

An experiment using an ensemble of dynamical model track forecasts and a selection of model fields as input in a statistical–dynamical intensity forecast model to produce intensity consensus members is conducted for the western North Pacific only. Consensus member skill at 72 h is low (−0.4% to 14.2%), and there is little independence among the members. This experiment demonstrates that a consensus of these highly dependent members yields an aid that performs as well as the most skillful member. Finally, adding a less skillful, but more independent, dynamical model-based forecast aid to the consensus yields an 11-member consensus with mixed yet promising performance compared with the 10-model consensus.

Based on these findings, the simple three-member consensus model could be used as a standard of comparison for other deterministic ensemble methods for the Atlantic and eastern North Pacific. Both the two- and three-member consensus forecasts may also provide useful guidance for operational forecasters. Likewise, in the western North Pacific, the 10- and 11-member consensus could be used as operational forecast aids and standards of comparison for other ensemble intensity forecast methods.

1. Introduction

The value of consensus forecasting has been recognized in meteorology for decades. Sanders (1973) demonstrated that a simple average of the forecasts from a group of forecasters was routinely superior to even the best of the individual forecasters. Leslie and Fraedrich (1990), Mundell and Rupp (1995), and Goerss (2000) applied consensus techniques to tropical cyclone track forecasting and found that a consensus is more accurate, on average, than forecasts from individual models. Goerss's (2000) study focused on dynamical track models because they were the best performers, on average.

Many of these dynamical models also produce forecasts of tropical cyclone intensity (maximum 1-min mean wind at 10-m elevation) along with the track,

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

DOI: 10.1175/2007WAF2007028.1

even though many of the operational models are handicapped by resolution, initialization, and parameterizations of the smaller-scale processes (Knaff et al. 2007) and, thus, cannot simulate the inner core of a tropical cyclone. Consequently, the only operational dynamical models showing skill in intensity forecasting are high-resolution models designed specifically for tropical cyclone forecasting (DeMaria et al. 2007). Dynamical models that routinely produce intensity forecasts in the Atlantic and eastern North Pacific basins are the Naval Operational Global Atmospheric Prediction System (NOGAPS; Hogan and Rosmond 1991; Goerss and Jeffries 1994), the Geophysical Fluid Dynamics Laboratory Hurricane Prediction System (GFDL; Kurihara et al. 1993, 1995, 1998), a version of GFDL run with NOGAPS initial and boundary conditions (GFDN; Rennick 1999), the Met Office global model (UKM; Cullen 1993; Heming et al. 1995), the National Weather Service (NWS) global spectral model [the Global Forecast System (GFS); Lord 1993], and the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5; Grell et al. 1995) run operationally by the Air Force Weather Agency (AFWA). Table 1 provides a summary of these and other aids used in this study.

The Statistical Hurricane Intensity Prediction System (SHIPS; DeMaria et al. 2005; DeMaria et al. 2006; Kaplan and DeMaria 1995, 2001) also produces intensity forecasts for the Atlantic and eastern North Pacific. This statistical–dynamical model does not attempt to resolve the tropical cyclone inner core; rather it forecasts changes in intensity through regression of large-scale environmental parameters. One of the main weaknesses of this model is that even though it generally has lower mean absolute error than NWP models, it does not forecast rapid intensification (Knaff et al. 2007). For an intensity skill baseline we will use the simple climatology and persistence statistical model, the 5-day Statistical Hurricane Intensity Forecast (SHF5; Knaff et al. 2003). This model is also a poor predictor of rapid intensification and decay since it is designed to minimize mean forecast errors, yet its seasonal average performance is still competitive (DeMaria et al. 2007).

For the western North Pacific, there are approximately 20 intensity forecast models. The NOGAPS, GFDN, UKM, and MM5 are all available in this region, as they are in the Atlantic and eastern North Pacific. Two models run operationally at the Japan Meteorological Agency (Kuma 1996), the global spectral model and the typhoon model, generate intensity forecast guidance. The Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS; Hodur 1997), the TC-Limited Area Prediction System (TC-LAPS; Davidson and Weber 2000) run by the Australian Bureau of Meteorology, and the Coupled Hurricane Intensity Prediction System (CHIPS; Emanuel et al. 2004) are also available. As is the case in the other two basins, a statistical–dynamical model similar to SHIPS (STIPS; Knaff et al. 2005) is available. Finally, the 5-day Statistical Typhoon Intensity Forecast (ST5D; Knaff et al. 2003), a climatology and persistence model, will be used for an intensity skill baseline. Other statistical intensity aids (e.g., climatology, climatology and persistence, analogs, extrapolation, and hybrids) exist, but they are considered predecessors to ST5D and not discussed further.

Unlike tropical cyclone track forecasts, intensity forecasts have relatively little skill when compared with baselines such as SHF5 or ST5D (Knaff et al. 2005; DeMaria et al. 2005, 2007), so it is not clear that the generalizations derived from consensus track forecasting found in Goerss (2000) apply to the intensity forecast problem. Some attempts at forming an intensity

TABLE 1. A list of tropical cyclone intensity forecast aids used in this study. The first column gives the name of the aid, the second column gives the name of the interpolated version of that forecast aid, and the final column gives a description of the numerical or statistical model that is the basis for those forecast aids.

Forecast aid ID	Interpolated version	Basis model
AFW1	AFWI	Air Force mesoscale model (MM5)
AVNO	AVNI	NWS global model (GFS, formerly the Aviation Model)
CHIP	CHII	Coupled hurricane model (CHIPS)
COWP	COWI	Navy mesoscale model (COAMPS)
DSHP		Decay SHIPS
GFDL	GFDI	GFDL hurricane model
GFDN	GFNI	Navy version of GFDL hurricane model
INT2		Two-member consensus (GFDI+DSHP)
INT3		Three-member consensus (GFDI+DSHP+GFNI)
JGSM	JGSI	Japanese global model
JTYM	JTYI	Japanese typhoon model
NGPS	NGPI	Navy global model (NOGAPS)
SHF5		Statistical Hurricane Intensity Forecast (SHIFOR5)
STFD		Upgraded version of decay STIPS
STID		Decay STIPS
ST10		10-member STIPS ensemble
ST11		10-member STIPS ensemble+GFNI
ST5D		5-day Statistical Typhoon Intensity Forecast
TCLP	TCLI	Australian typhoon model (TC-LAPS)
UKM	UKMI	U.K. global model
WBAR	WBAI	Weber barotropic model

consensus have shown limited increases in the skill of the consensus over its members. Weber (2005) found that an average of all intensity forecasts was generally among the top performers. The Joint Typhoon Warning Center (JTWC) recently reported that a method whereby the forecaster selects the members to form a consensus results in forecasts that were competitive with individual members (JTWC 2006). And finally, attempts at weighted consensus based on past performance demonstrated gains with respect to the individual members (Emanuel 2005; Biswas et al. 2006). Even though these methods used different members, different algorithms, and different weights, all reported improvements in skill by combining forecasts. None of these studies, however, examined an evenly weighted average of a predetermined set of members.

This study will address this issue. It will first explore the skill of the existing guidance, and then determine whether superior skill can be obtained using a simple equally weighted consensus of the most skillful members. Finally, baselines will be proposed for evaluating the more complex consensus techniques.

2. Data

The data used for this study are taken from the operational archives at the National Hurricane Center (NHC) and the JTWC as stored on the Automated Tropical Cyclone Forecasting System (ATCF; Sampson and Schrader 2000). The authors attempted to obtain a large dataset for this study so that statistics would be stable. The period chosen was 2003–06 for the following reasons: 1) the 120-h official forecasts from both agencies and skill baselines (SHF5 and ST5D) were available, 2) the version of STIPS used in the western North Pacific (STID; Knaff et al. 2005) was run during this period, 3) the forecasts from the GFDL and GFDN models were available, and 4) the NHC and JTWC official best tracks were complete and available. A subset of this dataset is used for evaluation of STIPS consensus forecasts ST10, which is described in detail in appendix A, and ST11, which is defined in Table 1.

3. Methods

Numerous objective forecast aids are available to help the NHC and JTWC in the preparation of official track and intensity forecasts. Forecast aids are characterized as either *early* or *late*, depending on whether or not they are available to the 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 GFS model is not complete nor is its forecast aid AVNO available to the forecaster until about 1600 UTC, about an hour after the forecast is released; thus, the 1200 UTC GFS would be considered a late forecast aid because it could not be used to prepare the 1200 UTC official forecast. This report focuses on the verification of early forecast aids unless otherwise stated.

Unlike statistical model forecast aids (e.g., SHF5 and ST5D) and some statistical–dynamical model forecast aids (e.g., DSHP), dynamical model forecast aids are generally late. Fortunately, a simple technique exists 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. In the example above, forecast aid for 6–126 h from the previous (0600 UTC) run of the GFS 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. The adjustment process creates an "early" version of the GFS forecast aid for the 1200 UTC forecast cycle that is based on the most current available guidance. The adjusted versions of the late forecast aids are known, for historical reasons, as interpolated forecast 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 consensus forecasts described in this paper are equally weighted averages of the consensus members. An attempt is made to compute a consensus forecast at each forecast period (12, 24, 36, 48, 72, 96, and 120 h). A consensus is computed if *two or more* consensus members exist for a given forecast period. If fewer than two members exist, the consensus is aborted for this and subsequent time periods. This procedure differs from that employed by some operational consensus techniques [e.g., the NHC consensus aids the GFDL–UKM–NOGAPS model ensemble average (GUNS) and the GFDI–UKMI–NGPI–AVNI model ensemble average (GUNA)] that require all ensemble members to be present.

Results presented are from recomputed interpolated aids and consensus forecasts using methods described above as well as operational input. The purpose of this is 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 on the order of 1%.

Forecasts are verified only when the best-track intensity is greater than 20 kt (10.3 m s^{-1}) and only when the system is a tropical or subtropical cyclone. Interpolated forecast aids are used as described above. If 6-h interpolated forecast aids are not available, then 12-h interpolated forecast aids are used. The 12-h interpolations occur approximately 15% of the time or less for models that are available every 6 h. The dataset is further restricted in that there must be a verifying official forecast. Performance is discussed through the use of skill charts. The measure of skill in these charts is defined as

$$\text{skill} = 100 \times (\text{baseline error} - \text{model error})/\text{baseline error}. \quad (1)$$

Thus, skill is positive when the forecast aid error is less than that of the baseline forecast aid. A one-tailed Student's t test at the 95% level with serial correlation of 30 h removed (Neumann et al. 1977) is also employed as a method to test significance in forecast error differences between individual intensity forecast aids.

4. Results

The first step in forming a consensus was identifying those forecast aids that may be of use in a consensus. A comparative verification of the various intensity forecast aids is shown in Fig. 1. Intensity forecast skill for each forecast aid was generally less than the skill associated with its track forecasts at 72 h (Franklin 2007; JTWC 2006) where skill relative to the baselines was generally positive and approaches 50%. It is also apparent that the number of skillful forecast aids was limited to one or two in each basin. The DSHP was skillful in the Atlantic and eastern North Pacific basins. The GFDI was skillful at 72 h only in the Atlantic. In the western North Pacific there was only one skillful forecast aid (STID), but another (GFNI) was nearly so.

Requiring a prospective consensus member to be skillful would yield only a two-member consensus in the Atlantic and no consensus at all in the other basins. However, skill is an arbitrary benchmark and its sign is not an indicator of potential benefit to the consensus. Rather, we used Fig. 1 to simply identify the top-performing forecast aids in each basin as a first cut at potential consensus membership. In the Atlantic and eastern North Pacific, GFDI, DSHP, and GFNI were much more skillful than the remaining forecast aids. In the western North Pacific, STID and GFNI were clearly superior to the remaining forecast aids.

In addition to mean error (or skill), member independence is an important factor governing value to a consensus (Goerss 2000; Sampson et al. 2006). An equation (derived in appendix B) for the consensus mean error (μ_c) is

$$\mu_c = \mu/n^{1/2}, \quad (2)$$

where μ is the mean of the members (assumed to all be equal to each other) and n is the number of independent members. This shows that increasing the number

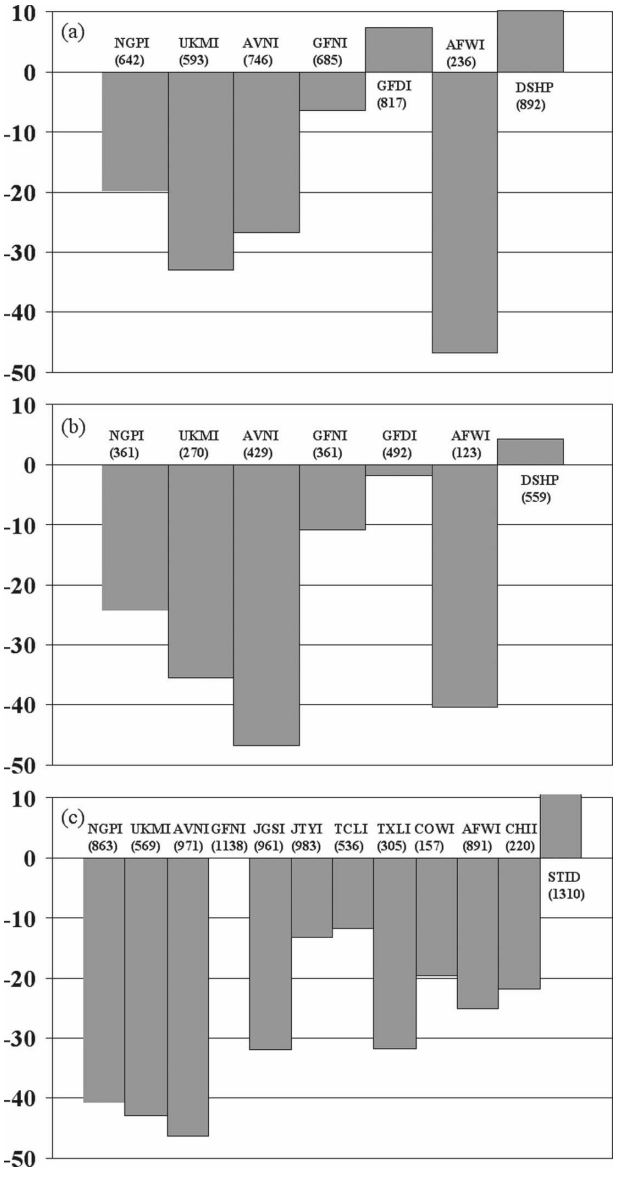

FIG. 1. The 72-h intensity forecast skill (%) for a subset of forecast aids available to operational forecast centers relative to (a) SHF5 in the Atlantic, (b) SHF5 in the eastern North Pacific, and (c) ST5D in the western North Pacific (2003–06 seasons). Acronyms are defined in Table 1 and numbers of cases are shown in parentheses.

of independent members will reduce the mean error of the consensus. The intensity forecast errors are not entirely independent, so n is replaced by the 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%. On the other hand, two members with errors having little independence ($n_e = 1.1$) would only produce an improvement of approximately 5% over the member mean. Model independence is generally not known a priori, and therefore a trial and error approach to find consensus members is generally required. Results of consensus trials are shown in Fig. 2.

In the Atlantic, the two-member consensus (INT2 = GFDI + DSHP) outperformed the best-member model by 0.7%, 3.0%, 6.2%, 12.5%, and 12.2% at 24, 48, 72, 96, and 120 h, respectively (Fig. 2a). The three-member consensus (INT3 = GFDI + DSHP + GFNI) outperformed its members by 1.4%, 5.0%, 7.5%, 3.3%, and 5.5% at 24, 48, 72, 96, and 120 h, respectively. When the previously defined t test was applied, the three-member consensus results were found to be significantly better than the individual members at 48 and 72 h in the Atlantic while the results at the other forecast times were not. Experiments run with a four-member consensus in the Atlantic by adding the next-best performer (NGPI) to the three-model consensus indicated that it degraded the consensus forecasts by 4.2%, 4.7%, 3.1%, and 0.2% at 24, 48, 72, and 96 h, respectively. Adding NGPI improved the 120-h forecast performance by 2.0%, but the result did not pass the significance test. Therefore, the four-member Atlantic consensus is not proposed as an objective aid.

In the eastern North Pacific (Fig. 2b), the two-member consensus INT2 outperformed its best member at 24, 48, 72, and 96 h by 2.5%, 2.8%, 6.8%, and 9.4%, respectively, and the best forecast aid outperformed the two-member consensus by 3.0% at 120 h. The three-member consensus INT3 outperformed its best member by 3.7%, 4.8%, 7.3%, 11.6%, and 2.0% at 24, 48, 72, 96, and 120 h, respectively. The three-member consensus results were significantly better than the best-member results only at the 96-h forecast period. Results at other times, although encouraging, did not pass significance tests. Experiments run with four-member consensus in the eastern North Pacific by adding the next-best performer (UKMI) to the three-member consensus indicated that it degraded the consensus forecasts by 6.0%, 8.5%, 12.0%, 14.1%, and 11.9% at 24, 48, 72, 96, and 120 h. Therefore, these results were not included in Fig. 2.

One advantage of the three-member consensus over the two-member consensus was that it was available

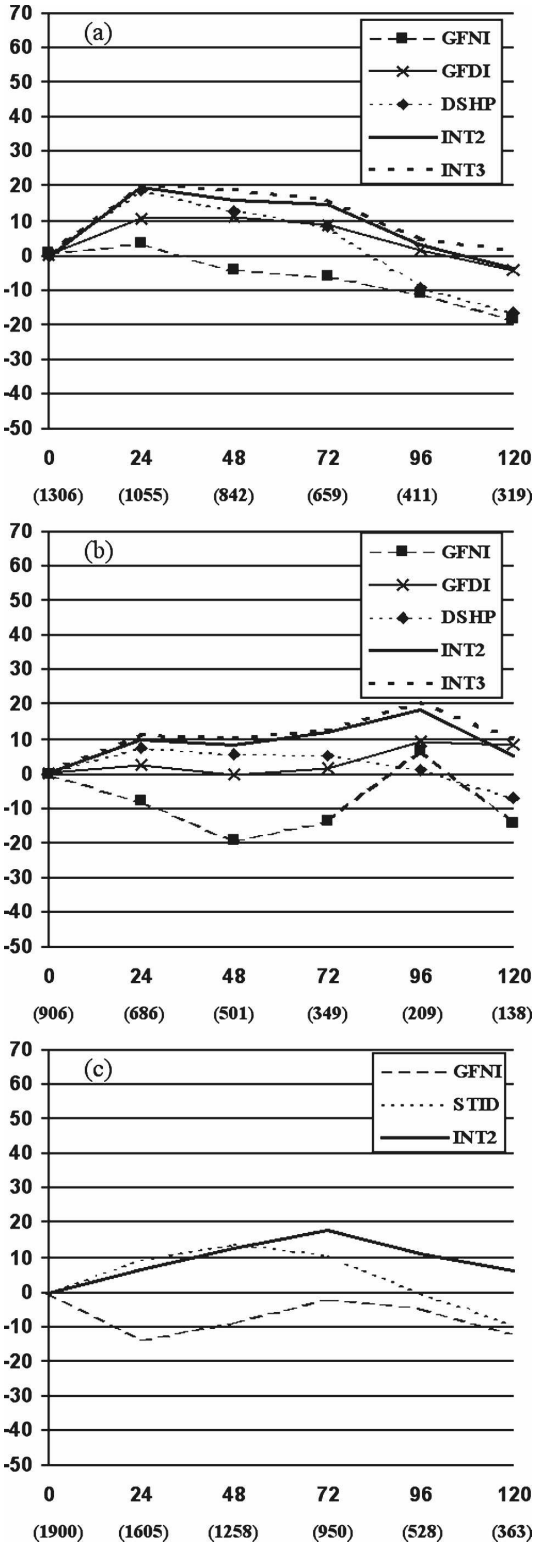

FIG. 2. Intensity forecast skill (%) of consensus and consensus members relative to (a) SHF5 in the Atlantic, (b) SHF5 in the eastern North Pacific, and (c) ST5D in the western North Pacific (2003–06 seasons). Acronyms are defined in Table 1 and numbers of cases are shown in parentheses.

more often than the two-member consensus since only two of the three members need to be available to compute the three-member consensus. For example, the three-member consensus was available for 91% of the Atlantic 2003–06 72-h official forecasts while the two-member consensus was available only 83% of the time. The disadvantage, however, is that a consensus made up of different members at different times is harder for forecasters to interpret.

In the western North Pacific (Fig. 2c), the two-member consensus outperformed its best member at 48, 72, 96, and 120 h by 1%, 7.3%, 11.3%, and 15.4%, respectively, and the best member outperformed the two-member consensus by 2.8% at 24 h. Results at 72, 96, and 120 h were significant (using the *t* test described above). Of note is that the STID forecast was actually significantly better than the consensus at 24 h. It is suspected that there was a problem with the GFDN in these early time periods that may have been solved in subsequent versions of the model. The performance of the GFDL model in the Atlantic and eastern North Pacific did not demonstrate a similar problem. On the contrary, the GFDL was skillful at 24 h in those basins. Experiments adding the next best performing forecast aids (JTYI and TCLI) to the western North Pacific consensus indicated that these forecast aids raised the average errors of the consensus by a minimum of 0.5%, 1.1%, and 1.7% at 24, 48, and 72 h, respectively. Neither the JTYI nor TCLI forecasts extended beyond 72 h. Hence, neither four-member consensus was included in Fig. 2, nor were JTYI or TCLI included in further tests.

Results from an experiment in which an ensemble of dynamical model track forecasts and a selection of model fields are used as input into an upgraded version of STIPS for the 2005–06 seasons are shown in Fig. 3. A more detailed description of the STIPS consensus (ST10) and its members is included in appendix A. Performance of individual ST10 members indicates that most are skillful with respect to ST5D (Fig. 3a), but that skill levels are low (−0.4% to 14.2%).

ST10 was among the top performers when compared with its members, outperforming the ST10 members by 0.5%–15.3% (Fig. 3b). As expected, this improvement was small because the ST10 members were highly dependent. Using Eq. (2) to estimate the effective degrees of freedom gave a value of approximately 1.05 for the entire set of members at the 72-h forecast. By comparison, Sampson et al. (2006) estimated the effective degrees of freedom of every possible two-member consensus the track members used in the STIPS consensus to range from 1.3 to 1.61 for the 72-h forecast. The 10

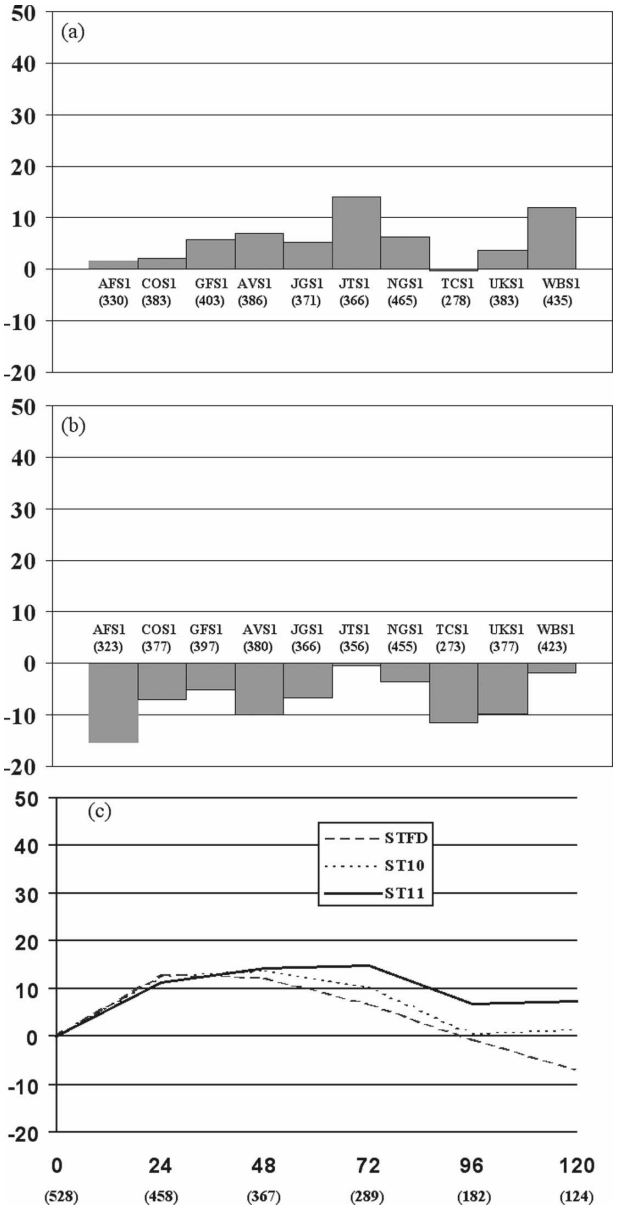

FIG. 3. (a) The 72-h intensity forecast skill (%) relative to ST5D for STIPS consensus members, (b) the 72-h intensity forecast skill relative to a consensus aid constructed from all 10 members (ST10), and (c) the homogeneous comparison of forecast skill relative to ST5D at 0–120 h for STIPS run on the official JTWC track and NOGAPS fields (STFD), and the consensus aids ST10 and ST11. The analysis is for the 2005–06 seasons in the western North Pacific (some dates missing due to operational issues). Acronyms for STIPS consensus members are defined in Table A1 and numbers of cases are shown in parentheses.

dynamical models that produce tracks for the STIPS consensus garnered independence from a number of possible differences (data assimilation, physics, parameterizations, resolution, and many others), while the STIPS consensus members had limited potential for in-

dependence since the independence was only to be obtained through differences in tracks and forecast winds. The effects of those differences on tracks and forecast winds were also limited since the STIPS model forecasts were largely dependent on the analysis fields. *Thus, this experiment demonstrated that forming a consensus from skillful members was not sufficient to reduce the consensus mean error. The members must also demonstrate independence from each other.* This is why multimodel consensus approaches generally outperformed single-model ensembles, both in tropical cyclone track forecasting and midlatitude meteorology (Fritsch et al. 2000).

A comparison of the performance of ST10, ST11 (formed by adding GFNI to the STIPS consensus), and STFD (the upgraded STIPS model run using the official track) is shown in Fig. 3c. Results indicated that ST10 outperformed STFD by 1.3%, 3.7%, 0.7%, and 8.2% at 48, 72, 96, and 120 h, respectively, and STFD outperformed ST10 by 0.4% at 24 h. None of the differences in errors passed the significance test. The ST11 outperformed STFD by 1.8%, 7.9%, 7.5%, and 14.1% at 48, 72, 96, and 120 h, and STFD outperformed ST11 by 2.8% at 24 h. The results at 72 and 120 h passed the *t*-test significance test. The impact of GFNI on the consensus was greatest at 72, 96, and 120 h where ST11 outperforms ST10 by 4.2%, 6.3%, and 5.9% (Fig. 3c) and passed the significance test at 72 and 96 h.

5. Summary and conclusions

Experiments with evenly weighted consenses were conducted on operational objective aid data (2003–06) for the Atlantic, eastern North Pacific, and western North Pacific basins. First, 72-h intensity forecast errors from several forecast aids were evaluated to find potential candidates for the formation of a consensus. Then, the most skillful candidates were added to the consensus in rank order until one is found that reduces skill rather than increasing it. For the Atlantic and eastern North Pacific, a three-model consensus of DSHP, GFDI, and GFNI was found to perform best while a two-member consensus of STID and GFNI was found to perform best for the western North Pacific. A further experiment was performed for the western North Pacific involving a consensus of 2–10 members constructed from assorted NWP model tracks and NWP model fields as input to a statistical–dynamical model (STIPS), as described in appendix A. Most individual members of the STIPS consensus were skillful at 72 h, yet no additional skill was attained by forming a consensus of the members. This result is consistent with the lack of independence of the consensus members. Adding an independent dynamical model forecast aid (GFNI) to the STIPS consensus produced significant improvement in skill at 96 h, but degraded the forecast at 24 h. It is suspected that subsequent improvements to the GFDN model addressed its performance at 24 h so that it becomes a positive contributor to a consensus at all forecast lengths.

The consensus forecast aids described above (INT3, the evenly weighted average of GFDI, DSHP, and GFNI in the Atlantic and eastern North Pacific; ST11, the evenly weighted 11-model consensus in the western North Pacific) are intensity forecast aids likely to have higher skill than the individual members alone. *These equally weighted consensus aids could serve as deterministic intensity forecast benchmarks for other consensus or ensemble methods and may also provide operational forecast guidance.* NHC postseason analysis (Franklin 2006, 2007) found that the performance of a simple, evenly weighted, intensity consensus (INT2) was competitive with the more complex method of Biswas et al. (2006). This is a remarkable result given that Biswas et al. (2006) use the interpolated 6-h-old official forecast (i.e., subjective expert information) in addition to the objective guidance discussed previously.

It is suspected that improvements in the consensus members and the addition of other independent, skillful forecast aids would further benefit this simple, evenly weighted intensity consensus. Some questions, however, remain regarding consensus forecasting. Is there a way to predict a model's impact on a consensus before running experiments? Can weighted consensus forecasting methods be constructed that outperform the evenly weighted average? If so, can the consensus member weights be designed so that they do not require updates? These questions will be investigated in future studies.

Acknowledgments. The authors acknowledge Ann Schrader and Chris Sisko for their work with the ATCF, Jim Gross for development of the interpolator, Mike Fiorino for his vortex trackers, and the staffs of NHC and JTWC for their diligent efforts with the ATCF. The authors also wish to acknowledge John Cook, Ted Tsui, Eric Blake, Dave Roberts, and two anonymous reviewers for their thoughtful comments. This project is supported through a grant from the Joint Hurricane Testbed and funding from the Office of Naval Research. The views, opinions, and findings contained in this article 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.

TABLE A1. STIPS consensus members. The name of the individual consensus member is given in the first column. The following columns describe the input data used in the STIPS model to create each of the consensus members. Dynamic forecasts fields refer to the specific forecast model that provides the forecasts of the winds, and other forecast fields refer to the model that provides the thermodynamic, moisture, and SST fields.

ST10 member	Track input	Dynamic forecast fields	Other forecast fields
AFS1	AFWI	NOGAPS	NOGAPS
AVS1	AVNI	NWS global	NOGAPS
COS1	COWI	COAMPS	NOGAPS
GFS1	GFNI	NOGAPS	NOGAPS
JGS1	JGSI	Japanese global	NOGAPS
JTS1	JTYI	Japanese global	NOGAPS
NGS1	NGPI	NOGAPS	NOGAPS
TCS1	TCLI	NOGAPS	NOGAPS
UKS1	UKMI	U.K. global	NOGAPS
WBS1	WBAI	NOGAPS	NOGAPS

APPENDIX A

The STIPS Consensus

The STIPS consensus (ST10) is constructed using 2–10 NWP model interpolated track forecasts available at an approximate synoptic time of +1.5 h. The interpolated track forecasts chosen were those of the operational track consensus used at JTWC for the 2005–06 seasons.

Ideally, consensus members should be run through STIPS with thermodynamic and dynamic input from the model corresponding to the interpolated track. This would provide the most independence in the members, which should lead to a larger reduction in the consensus mean. It would also provide model fields with a vortex structure collocated with the interpolated track and, thus, should provide for more realistic STIPS computations (e.g., shear computation) for that member. Because the authors could not obtain complete model field input for all of the member models, a compromise solution was constructed. For six of the interpolated model tracks (NGPI, GFNI, UKMI, AVNI, JGSI, and COWI) STIPS is run with dynamic fields (u and v components of the wind) from the model and NOGAPS data for the other STIPS field data input (temperature, relative humidity, and geopotential height). For the Japanese Typhoon Model interpolated track (JTYI), STIPS is run with Japanese global model dynamic model fields and NOGAPS data for other STIPS field data input. And finally, NOGAPS was used for all field data input to run the remaining three interpolated tracks (GFNI, TCLI, and AFWI). Table A1 provides an overview of the STIPS consensus members and their input.

The version of STIPS used for ST10 has upgrades regarding decay effects over land (DeMaria et al. 2006). A forecast aid run with this newer version of STIPS on the JTWC track (STFD) was also produced for comparison with ST10. The comparison is not entirely fair since the current operational configuration delays STFD sufficiently so that it is produced about an hour a later than the operational intensity forecast and is, therefore, a late forecast aid.

APPENDIX B

Relationship between Mean Errors, Consensus Errors, and Independence

Assume that for each consensus member the errors are normally distributed around a zero mean (no bias) with a standard deviation σ so that the mean deviation is defined in Spiegel (1961) as

$$\mu = \sigma(\pi/2)^{1/2}. \tag{B1}$$

Then, the errors of a consensus would also be normally distributed around a zero mean with a standard deviation:

$$\sigma = \sigma/n^{1/2}, \tag{B2}$$

where σ is the standard deviation of the members and n is the number of independent models (Hoel 1962). The consensus mean for this normal distribution is defined as

$$\mu_c = \sigma_c/(\pi/2)^{1/2}, \tag{B3}$$

and substituting (B2) into (B3) and solving for σ yields

$$\sigma = \mu_c(2n/\pi)^{1/2}. \tag{B4}$$

Finally, substitution of (B4) into (B1) and solving for the consensus mean yields

$$\mu_c = \mu/n^{1/2}. \tag{B5}$$

This result implies that increasing the number of independent models will reduce the mean error of the consensus.

REFERENCES

Biswas, M. K., B. P. Mackey, and T. N. Krishnamurti, 2006: Performance of the Florida State University Hurricane Superensemble during 2005. *Proc. 60th Interdepartmental Hurricane Conf.,* Mobile, AL, Office of the Federal Coordinator for Meteorological Services and Supporting Research. [Available online at http://www.ofcm.gov/ihc06/linking_file_ihc06.htm.]

Cullen, M. J. P., 1993: The Unified Forecast/Climate Model. *Meteor. Mag.,* **122,** 81–122.

Davidson, N. E., and H. C. Weber, 2000: The BMRC high-resolution tropical cyclone prediction system: TC-LAPS. *Mon. Wea. Rev.,* **128,** 1245–1265.

DeMaria, M., 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 J. Kaplan, 2006: On the decay of tropical cyclone winds crossing narrow landmasses. *J. Appl. Meteor. Climatol.,* **45,** 491–499.

——, ——, and C. R. Sampson, 2007: Evaluation of long-term trends in tropical cyclone intensity forecasts. *Meteor. Atmos. Phys.,* **97,** 19–28.

Emanuel, K., 2005: Ensemble forecasting in hurricane intensity. *Proc. 59th Interdepartmental Hurricane Conf.,* Jacksonville, FL, Office of the Federal Coordinator for Meteorological Services and Supporting Research. [Available online at http://www.ofcm.gov/ihc05/linking_file_ihc05.htm.]

——, C. Desautels, C. Holloway, and R. Korty, 2004: Environmental control of tropical cyclone intensity. *J. Atmos. Sci.,* **61,** 843–858.

Franklin, J. L., cited 2006: 2005 National Hurricane Center forecast verification report. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2005.pdf.]

——, cited 2007: 2006 National Hurricane Center forecast verification report. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2006.pdf.]

Fritsch, J. M., J. Hilliker, and J. Ross, 2000: Model consensus. *Wea. Forecasting,* **15,** 571–582.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.,* **128,** 1187–1193.

——, and R. A. Jeffries, 1994: Assimilation of synthetic tropical cyclone observations into the Navy Operational Global Atmospheric Prediction System. *Wea. Forecasting,* **9,** 557–576.

Grell, G. A., J. Dudhia, and D. R. Stauffer, 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., J. C. L. Chan, and A. M. Radford, 1995: A new scheme for the initialization of tropical cyclones in the UK Meteorological Office Global Model. *Meteor. Appl.,* **2,** 171–184.

Hodur, R. M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.,* **125,** 1414–1430.

Hoel, P. G., 1962: *Introduction to Mathematical Statistics*. Wiley, 427 pp.

Hogan, T. F., and T. E. Rosmond, 1991: The description of the Navy Operational Global Atmospheric Prediction System's Spectral Forecast Model. *Mon. Wea. Rev.,* **119,** 1786–1815.

JTWC, cited 2006: The annual tropical cyclone reports. [Available online at https://metocph.nmci.navy.mil/jtwc.php.]

Kaplan, J., and M. DeMaria, 1995: A simple empirical model for predicting the decay of tropical cyclone winds after landfall. *J. Appl. Meteor.,* **34,** 2499–2512.

——, and ——, 2001: A note on the decay of tropical cyclone winds after landfall in the New England area. *J. Appl. Meteor.,* **40,** 280–286.

Knaff, J., M. DeMaria, C. R. Sampson, and J. M. Gross, 2003: Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting,* **18,** 80–92.

——, ——, and M. DeMaria, 2005: An operational statistical typhoon intensity prediction scheme for the western North Pacific. *Wea. Forecasting,* **20,** 688–699.

——, C. Guard, J. Kossin, T. Marchok, B. Sampson, T. Smith, and N. Surgi, 2007: Operational guidance and skill in forecasting structure change. *Proc. Sixth Int. Workshop on Tropical Cyclones,* San José, Costa Rica, WMO Tech. Doc. 1383. [Available online at http://severe.worldweather.org/iwtc/.]

Kuma, K., 1996: NWP activities at Japan Meteorological Agency. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J15–J16.

Kurihara, Y., M. A. Bender, and R. J. Ross, 1993: An initialization scheme of hurricane models by vortex specification. *Mon. Wea. Rev.,* **121,** 2030–2045.

——, ——, R. E. Tuleya, and R. J. Ross, 1995: Improvements in the GFDL hurricane prediction system. *Mon. Wea. Rev.,* **123,** 2791–2801.

——, R. E. Tuleya, and M. A. Bender, 1998: The GFDL hurricane prediction system and its performance in the 1995 hurricane season. *Mon. Wea. Rev.,* **126,** 1306–1322.

Leslie, L. M., and K. Fraedrich, 1990: Reduction of tropical cyclone position errors using an optimal combination of independent forecasts. *Wea. Forecasting,* **5,** 158–161.

Lord, S. J., 1993: Recent developments in tropical cyclone track forecasting with the NMC global analysis and forecast system. Preprints, *20th Conf. on Hurricanes and Tropical Meteorology,* San Antonio, TX, Amer. Meteor. Soc., 290–291.

Mundell, D. B., and J. A. Rupp, 1995: Hybrid forecast aids at the Joint Typhoon Warning Center: Application and results. Preprints, *21st Conf. on Hurricanes and Tropical Meteorology,* Miami, FL, Amer. Meteor. Soc., 216–218.

Neumann, C. J., M. B. Lawrence, and E. L. Caso, 1977: Monte Carlo significance testing as applied to statistical tropical cyclone models. *J. Appl. Meteor.,* **16,** 1165–1174.

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.

——, 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.

Sanders, F., 1973: Skill in forecasting daily temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.,* **54,** 1171–1179.

Spiegel, M. R., 1961: *Theory and Problems of Statistics*. McGraw-Hill, 359 pp.

Weber, H. C., 2005: Probabilistic prediction of tropical cyclones. Part II: Intensity. *Mon. Wea. Rev.,* **133,** 1853–1864.