Tropical Cyclone Track Forecasts Using an Ensemble of Dynamical Models

JAMES S. GOERSS

Naval Research Laboratory, Monterey, California

(Manuscript received 26 August 1998, in final form 14 May 1999)

ABSTRACT

The relative independence of the tropical cyclone track forecasts produced by regional and global numerical weather prediction models suggests that a simple ensemble average or consensus forecast derived from a combination of these models may be more accurate, on average, than the forecasts of the individual models. Forecast errors of a simple ensemble average of three models for the 1995–96 Atlantic hurricane seasons, and either three global models or two regional models for the western North Pacific during 1997, were compared with errors of the individual models. For the Atlantic, the mean errors for the joint ensemble were 120 km at 24 h, 194 km at 48 h, and 266 km at 72 h, which represent improvements of 16%, 20%, and 23% with respect to the best of the individual models. The joint ensemble also resulted in reduction in the standard deviation of the forecast error. The 95th percentile of forecast error for the ensemble was reduced 19%, 14%, and 23% with respect to the best of the individual models. The spread of the ensemble forecast was found to possess some potential for use by forecasters as a measure of confidence in the ensemble forecast. Similar results were found for the western North Pacific.

1. Introduction

The quality of tropical cyclone (TC) track forecasts from regional and global numerical weather prediction models has dramatically improved in the 1990s. Indeed, the significant enhancements in tropical cyclone track guidance provided by dynamical models foretold by Elsberry (1995) have been realized. As a result, TC forecasts from these models have become increasingly more important in recent years as guidance to forecasters at both the National Hurricane Center (NHC, now known as the Tropical Prediction Center), Miami, and the Joint Typhoon Warning Center (JTWC), Pearl Harbor. Forecasters at NHC routinely use TC track forecasts from the Geophysical Fluid Dynamics Laboratory Hurricane Prediction System (GFDL; Kurihara et al. 1993, 1995, 1998), the Navy Operational Global Atmospheric Prediction System (NOGAPS; Hogan and Rosmond 1991; Goerss and Jeffries 1994), and the United Kingdom Meteorological Office global model (UKMO; Cullen 1993; Heming et al. 1995). As documented by Kurihara et al. (1998), Goerss et al. (1998), and Heming and Radford (1998), the TC track forecasts of these models for the 1995 Atlantic hurricane season were excellent with mean forecast position errors on the order of 150–170 km at 24 h, 230–270 km at 48 h, and 350–400 km at 72 h. To provide some perspective and to illustrate in part the impact these models have had upon operational TC forecasting, the mean errors of the official NHC Atlantic TC forecasts for 1970–79 (Neumann and Pelissier 1981a), 1980–89, and 1990–96 (Lawrence et al. 1997) are displayed in Table 1. While the forecast errors for the 1980s were only slightly less than those for the 1970s, the forecast errors for 1990–96 were markedly reduced. When making comparisons of forecast skill, it should be noted that the forecasts from these dynamical models were not available to the operational forecasters until after the time they must make their forecasts, and that, in preparing their forecasts, they utilized 6-h old interpolated versions of the forecasts from these models.

Like their counterparts at NHC, forecasters at JTWC routinely use TC track forecasts from these models along with those from two models run operationally at the Japan Meteorological Agency (Kuma 1996): the global spectral model (GSM) and typhoon model (TYM). While the track forecasting skill of these individual models varies from year to year and basin to basin, their overall skill is comparable.

Sanders (1973) demonstrated that a consensus forecast may be beaten from time to time, but it is hard to beat the consensus in the long run. Thompson (1977) presented a theoretical discussion of the improvement in forecast accuracy that one can expect by combining independent forecasts. Leslie and Fraedrich (1990) and Mundell and Rupp (1995) applied this approach to TC track prediction and illustrated the forecast improvements that resulted from using linear combinations of

Corresponding author address: James S. Goerss, NRL, 7 Grace Hopper Ave., Stop 2, Monterey, CA 93943-5502.
E-mail: goerss@nrlmry.navy.mil

TABLE 1. Mean forecast errors (km) for the official NHC Atlantic TC forecasts for the periods 1970–79, 1980–89, and 1990–96.

	1970–79	1980–89	1990–96
24 h	202	200	175
48 h	451	420	325
72 h	697	650	500

forecasts from various TC track prediction models. The operational availability of multiple dynamical models initiated with slightly different analyses provides the ability to routinely determine a simple ensemble average or consensus forecast. The aforementioned studies suggest that such a forecast may be more accurate, on average, than the forecasts of any of the individual models.

Currently, the focus of the ensemble forecast systems at operational numerical weather prediction centers is on extratropical weather forecasting (Molteni et al. 1996; Toth and Kalnay 1993). Following an approach suggested by Leith (1974) for numerical weather prediction, perturbations to the initial state of the numerical prediction model are generated, and multiple integrations of the model are made. Different methods have been used to generate these perturbations at the operational centers with strong emphasis on improving forecasts of midlatitude weather systems. Zhang and Krishnamurti (1997) described a perturbation method specifically designed for forecasting TC tracks and illustrated the technique for ensemble forecasts of three hurricanes. Although their results were quite promising, many interesting questions about predictability of TC tracks may only be answered by more exhaustive experiments. Aberson et al. (1998) conducted a more extensive study of ensemble forecasting of TC tracks using the GFDL model, which provided some insight into these questions. Whereas an approach perturbing only the initial conditions assumes a perfect model, Houtekamer and Lefaivre (1997) also use multiple versions of their model. This approach increases the spread of the ensemble forecasts. In fact, in their study using the European Centre for Medium-Range Weather Forecasts and the UKMO global models, Harrison et al. (1999) point to the superiority of a joint ensemble over all configurations of a single-model ensemble. It should be noted that the last two studies were for the extratropics and for the medium range, and that their results may not apply to TC track forecasting.

In this paper the TC forecast performance of a simple ensemble utilizing multiple numerical weather prediction models routinely used by the operational forecasters at NHC and JTWC is evaluated for the 1995–96 Atlantic hurricane seasons and for the western North Pacific in 1997. These models are integrated at the various operational centers from what may be considered perturbed initial conditions, since each center utilizes the available observational data in slightly different ways via their quality control and data assimilation procedures. In addition, each center utilizes different physical process representations and horizontal/vertical resolution in their numerical model, so that a combination of their forecasts may be considered to be an ensemble of multiple models in the spirit of Houtekamer and Lefaivre (1997). Even though only a small number of dynamical model TC forecasts are used here relative to the midlatitude ensemble prediction systems, the desired properties of a smaller ensemble mean error and an ensemble spread that is indicative of likely forecast error may be expected.

2. Theoretical background

Before looking at the actual performance of this simple ensemble, we first examine forecast position error from a theoretical viewpoint. The forecast position error for model i, E_i, is defined to be

$$E_i = (C_i^2 + A_i^2)^{1/2},$$

where C_i and A_i are the acrosstrack and alongtrack errors, respectively (Neumann and Pelissier 1981b). If, for simplicity we assume that, for every i, C_i and A_i are independent and normally distributed with zero mean and standard deviation σ, then E_i possesses a Rayleigh distribution with mean $\sigma(\pi/2)^{1/2}$ (Lindgren 1976). Since each ensemble forecast position is merely the mean of the individual model forecast positions, one can easily show that the ensemble acrosstrack and alongtrack errors, denoted by C_e and A_e, are simply the means of the acrosstrack and alongtrack errors of the individual models. Therefore, for an ensemble with n members, C_e and A_e are normally distributed with zero mean and standard deviation $\sigma/n^{1/2}$ (Hoel 1962), and the ensemble forecast error, E_e, possesses a Rayleigh distribution with mean $\sigma(\pi/2n)^{1/2}$. In practice, the C_i and A_i are not independent, and n denotes the effective degrees of freedom, a number less than the number of members in the ensemble.

From the previous discussion we note that the mean forecast error is directly related to the standard deviation of the acrosstrack and alongtrack errors. Thus, we see that the mean ensemble forecast error is dependent upon two things: 1) the mean forecast error of the individual models that make up the ensemble and 2) the degree of independence of the forecast errors of the individual models. Given the same number of effective degrees of freedom, the ensemble mean is smaller when the mean forecast errors of the individual models that make up the ensemble are smaller. However, given an ensemble of models with the same mean forecast error, it is the degree of independence of the forecast errors of the individual models that determines the effectiveness of the ensemble. If the forecast errors from the individual models are highly correlated, the effective degrees of freedom will be little more than one, and the mean ensemble error will only be slightly less than the mean model forecast error. On the other hand, if the forecast errors from the individual models are totally uncorrelated, the effective degrees of freedom will be equal to

TABLE 2. Homogeneous comparison of the GFDL model, NOGAPS, UKMO, the ensemble average (ENSM), and CLIPER TC position errors (km) for a sample of (*N*) forecasts of tropical storms and hurricanes during the 1995–96 Atlantic hurricane seasons.

	N	GFDL	NOGAPS	UKMO	ENSM	CLIPER
24 h	280	142	152	152	120	187
48 h	221	246	255	244	194	389
72 h	166	364	383	348	266	607

the number of individual models in the ensemble, *n.* In this case, the mean ensemble error will be reduced by a factor of $n^{1/2}$ with respect to the mean model forecast error.

3. Results

A homogeneous comparison of the TC track performance of GFDL, NOGAPS, UKMO, and a simple ensemble or consensus forecast is presented first for the 1995–96 Atlantic hurricane seasons (Table 2). For each model, TC track forecast errors were determined for all forecasts initiated and validated when the TC was of tropical storm strength or greater (winds greater than 34 kt). That is, forecasts when the TC was only a depression at the initial or final time were excluded. Ensemble forecast positions were determined by simply averaging the forecast positions from the GFDL, NOGAPS, and UKMO models whenever positions from all three models were available. A series of forecasts separated by only 12 h or even 24 h is not statistically independent because the initial conditions are a blend of observations and a prior forecast. Thus, the effective sample size was determined by requiring at least a 30-h separation between forecasts of the same storm in calculating statistical significance with the modified *t*-test as described by DeMaria et al. (1992).

While the forecast performance of the three individual models was quite similar for this sample of Atlantic TCs (Table 2), the ensemble forecast errors were improved by 16%, 20%, and 23% at 24, 48, and 72 h with respect to the best of the individual models. At every forecast time, these forecast improvements were significant at the 99% level. With respect to CLIPER (climatology and persistence), the ensemble forecasts were improved by 36%, 50%, and 56% at 24, 48, and 72 h.

To support the forecasters at JTWC, the GFDL model is integrated operationally at Fleet Numerical Meteorology and Oceanography Center for western North Pacific TCs utilizing NOGAPS fields for initial and boundary conditions (Rennick 1999), and is referred to as GFDN. While NOGAPS, UKMO, and GSM are initiated at 0000 and 1200 UTC, both GFDN and TYM are initiated at 0600 and 1800 UTC.

TABLE 4. As in Table 3 except for TYM, GFDN, the ensemble average (ENSM), and CLIPER.

	N	TYM	GFDN	ENSM	CLIPER
24 h	246	144	139	115	170
48 h	206	248	272	215	355
72 h	166	394	470	351	546

Ensemble forecast positions for western North Pacific TCs during 1997 were first determined by averaging the forecast positions for the GSM, NOGAPS, and UKMO models whenever positions from all three models were available (Table 3). Unlike the nearly uniform performance among the models for the Atlantic TCs, NOGAPS did not perform as well as GSM and UKMO. For example, the performance of UKMO was better than NOGAPS at the 90% significance level at 72 h, and the performance of GSM was better than NOGAPS at the 90% significance level at 24 h and at the 95% significance level at 48 h and 72 h. Although the GSM consistently had smaller errors than UKMO, the differences in the errors were not statistically significant for any forecast time. Notice the ensemble forecast errors were improved by l6%, 13%, and 12% at 24, 48, and 72 h, respectively, with respect to the GSM. These improvements were significant at the 99% level at 24 h, the 98% level at 48 h, and the 90% level at 72 h. With respect to CLIPER, the ensemble forecasts were improved by 28%, 39%, and 41% at 24, 48, and 72 h, respectively.

The more pronounced difference between the forecast performance of NOGAPS and that of the other two models for the western North Pacific TCs during 1997 raises the interesting question of whether an ensemble using only the UKMO and GSM forecasts would outperform an ensemble using all three model forecasts. Despite the better forecast performance of UKMO and GSM with respect to NOGAPS, the three-model ensemble showed consistent, but not statistically significant, forecast improvement when compared with the two-model ensemble (Table 3).

In a second ensemble for 0600 and 1800 UTC, forecast positions were averaged for GFDN and TYM, whenever positions from both models were available (Table 4). Whereas neither model performed significantly better than the other at 24 h or 48 h, the performance of TYM at 72 h was significantly better than

TABLE 3. Homogeneous comparison of GSM, NOGAPS, UKMO, the ensemble average (ENSM), the GSM–UKMO average (ENS2), and CLIPER TC position errors (km) for a sample of (*N*) forecasts of tropical storms and typhoons in the western North Pacific during 1997.

	N	GSM	NOGAPS	UKMO	ENSM	ENS2	CLIPER
24 h	265	141	157	146	118	122	163
48 h	219	246	289	259	215	216	350
72 h	175	366	438	385	322	331	546

TABLE 5. Homogeneous comparison of the means (standard deviations) of the cross-track errors (km) of the GFDL model, NOGAPS, UKMO, and the ensemble average (ENSM) for a sample of (*N*) forecasts of tropical storms and hurricanes during the 1995–96 Atlantic hurricane seasons. Negative values indicate a forecast to the left of the actual track.

	N	GFDL	NOGAPS	UKMO	ENSM
24 h	280	−24 (118)	−28 (107)	−7 (113)	−20 (91)
48 h	221	−13 (226)	−22 (170)	−22 (192)	−19 (142)
72 h	166	−2 (331)	7 (296)	−48 (281)	−14 (220)

TABLE 6. As in Table 5 except for the alongtrack errors, where a negative value indicates a forecast slower than the actual track.

	N	GFDL	NOGAPS	UKMO	ENSM
24 h	280	−18 (117)	−50 (129)	−30 (139)	−33 (105)
48 h	221	−11 (194)	−102 (226)	−43 (222)	−52 (168)
72 h	166	30 (281)	−126 (314)	−57 (289)	−51 (215)

GFDN at the 95% level. The ensemble forecast errors were improved by 17%, 13%, and 11% with respect to the better of the two models at 24, 48, and 72 h, respectively. These improvements were significant at the 99% level at 24 and 48 h and at the 90% level at 72 h. With respect to CLIPER, the 0600 and 1800 UTC ensemble forecast errors were improved by 32%, 39%, and 36% at 24, 48, and 72 h, respectively.

To better illustrate the characteristics of the ensemble forecasts, a more extensive investigation of the GFDL, NOGAPS, UKMO, and ensemble forecasts for the 1995–96 Atlantic hurricane seasons is presented. Similar results (not shown) were found with the sample of western North Pacific TCs during 1997.

First, each envelope of 72-h forecast tracks from the individual models (166 cases) was visually examined to determine whether it included the actual TC track. A track was considered to be inside the envelope if the lines connecting the best-track positions were within the lines connecting the forecast points, and there was no indication of a gross error in speed. Despite the fact that each envelope only contained three tracks, we found that the actual TC track was included 79% of the time (131 cases). The 72-h ensemble forecast error was 218 km when the actual TC track was included in the ensemble envelope and 444 km when it was not. The pronounced difference in forecast error is a good indicator that the subjective determination of TC track inclusion was successful. Using an objective technique, we found that 39% of the 72-h ensemble mean forecast positions (64 cases) were contained in a box defined by the minimum and maximum forecast latitudes and longitudes of the ensemble members. The 72-h ensemble forecast error was 166 km when the forecast position was contained in the box and 328 km when it was not. Finally, we found that 81% of the 72-h ensemble mean forecast positions (133 cases) were within 160 km of such a box, and that the respective forecast errors were 212 and 487 km. Clearly, the effectiveness of the spread of the simple joint ensemble plays an important role in the significant reduction in forecast error demonstrated by this approach.

The means and standard deviations of the acrosstrack and alongtrack errors for each of the models and the ensemble are displayed in Tables 5 and 6. In section 2, with some simplifying assumptions, we defined a relationship between mean forecast position error and the standard deviation of acrosstrack and alongtrack error. If we average the standard deviations of the 24-h acrosstrack and alongtrack errors for the different models and the ensemble displayed in Tables 5 and 6 and apply this relationship, we obtain mean forecast position errors of 147, 148, 158, and 123 km for GFDL, NOGAPS, UKMO, and the ensemble, respectively. These values compare favorably with the actual mean position errors displayed in Table 2 (142, 152, 152, and 120 km, respectively). For each forecast time, the mean acrosstrack (alongtrack) error for the ensemble is simply the average of the mean acrosstrack (alongtrack) errors of the three models. Therefore, these mean ensemble errors are not as good as the best but not as bad as the worst. However, the standard deviation of the acrosstrack (alongtrack) errors for the ensemble is smaller than for each of the individual models. These reductions in error distribution, which are illustrated graphically by the scatterplots in Fig. 1, are the primary reason for the forecast improvement obtained using the ensemble. As discussed in section 2, if the errors of the three models were completely independent, the effective number of degrees of freedom would be three and the ensemble standard deviations would be reduced by the square root of three. The reduction in the ensemble standard deviations with respect to those from the individual models displayed in Tables 5 and 6 indicate that the number of effective degrees of freedom for the ensemble is approximately 1.5 at 24 h, 1.65 at 48 h, and 1.9 at 72 h. The consistency among the model forecast tracks is also indicated by the error correlations between the models (Table 7). These acrosstrack and alongtrack error correlations decrease with forecast length just as the effective degrees of freedom for the ensemble increases. Notice the errors of the two global models are the most highly correlated, while the GFDL and UKMO errors are the least highly correlated.

One important research goal to emerge from the U.S. Weather Research Program (USWRP) Hurricane Landfall Workshop (Elsberry and Marks 1998) was the reduction of the area of overwarning associated with TC landfall forecasts. The extent of these coastal warning areas is directly related to the 95% level of uncertainty in TC track forecasts. One suggested approach to reducing this level of uncertainty is ensemble forecasting. The 95th percentile for the 24-, 48-, and 72-h forecast errors was empirically determined for the ensemble and the individual models (Table 8). As measured by this quantity, the reduction in uncertainty at 24 h for the

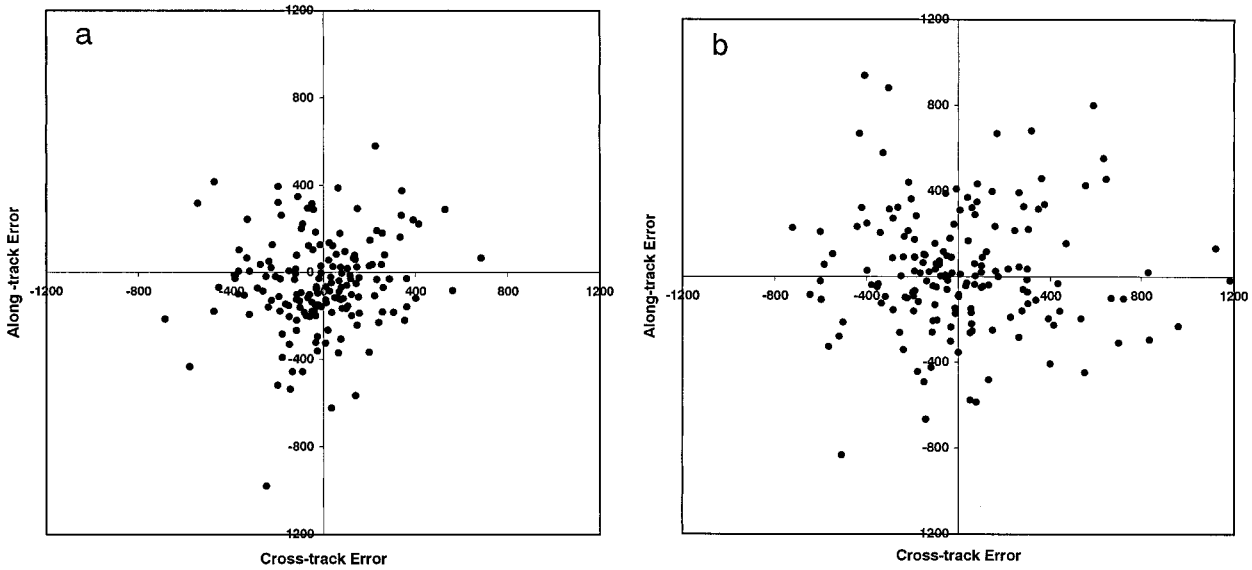

FIG. 1. Scatter of 72-h alongtrack error vs acrosstrack error (km) for (a) the ensemble average and (b) the GFDL model for the 1995–96 Atlantic hurricane seasons. Negative values of the along- and acrosstrack errors are behind and to the left of the actual storm position.

ensemble was approximately 20% with respect to the individual models. By way of comparison, the 95th percentile for the NHC 24-h forecast errors was 368 km for the period from 1988 to 1997 (M. DeMaria 1998, personal communication). Similar reductions are seen in Table 8 for the 48- and 72-h forecast errors. Thus, for all forecast times the uncertainty of the forecast error was reduced for the ensemble.

Another objective of the ensemble is prediction of the accuracy of the forecast. As discussed by Buizza and Palmer (1998), one would expect that a small ensemble spread should indicate a small ensemble mean forecast error, but that a large ensemble spread may not necessarily imply a large ensemble forecast error. However, the ensemble spread should be useful in approximating the upper bound of the ensemble forecast error. To test these expectations for the simple ensemble used here, the ensemble spread is defined to be the average distance of the ensemble member forecasts from the ensemble mean forecast. For each forecast time, the ensemble spread and forecast error were compared. The results displayed in Fig. 2 for the 72-h forecasts were typical of the earlier forecast times and similar to those shown by Aberson et al. (1998). As expected, there is no clear correlation between spread and error, but there does appear to be a relationship between the spread and the upper bound of error. If one ignores the seven outliers in the lower-right portion of Fig. 2, one can see that the upper bound of the ensemble error is roughly twice the ensemble spread. Incidentally, the seven outliers, which possess an unusually high ratio between error and spread, all come from cases where the actual TC track was not contained by the ensemble envelope. It is encouraging that only 7 of 166 cases possess this undesirable property of low spread and high error. To further quantify the relationship between ensemble spread and error, Table 9 has been constructed. When the ensemble spread was small (less than or equal to 300 km), the 72-h ensemble forecast error was less than 300 km 70% of the time and, more importantly, greater than 450 km only 8% of the time. On the other hand, when the ensemble spread was large (greater than 300 km), the 72-h ensemble forecast error was less than 300 km only 53% of the time and greater than 450 km 22% of the time. The mean of the 72-h ensemble forecast errors for the 116 cases with small spread was 241 km while the mean for the 50 cases with large spread was 327 km. The respective medians were 210 km and 295 km. Thus, in a broad sense, a forecaster could use the ensemble spread, a quantity that can be determined be-

TABLE 7. Correlations of acrosstrack (alongtrack) errors between the three forecast models for the sample of Atlantic tropical storms and hurricanes during the 1995–96 seasons.

	NOGAPS–GFDL	NOGAPS–UKMO	GFDL–UKMO
24 h	0.46 (0.52)	0.62 (0.65)	0.39 (0.45)
48 h	0.36 (0.44)	0.47 (0.57)	0.13 (0.27)
72 h	0.34 (0.28)	0.48 (0.42)	0.01 (0.18)

TABLE 8. The 95th percentile (km) for the forecast errors of the GFDL model, NOGAPS, UKMO, and the ensemble average (ENSM) for the sample of Atlantic tropical storms and hurricanes during the 1995–96 seasons.

	GFDL	NOGAPS	UKMO	ENSM
24 h	339	333	327	265
48 h	564	592	524	450
72 h	851	925	796	610

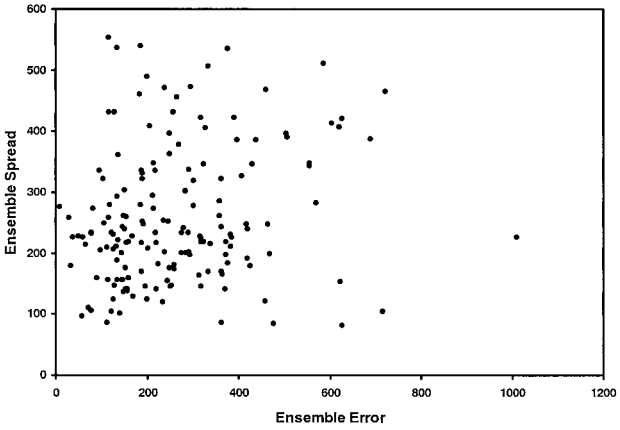

FIG. 2. Scatter of 72-h ensemble spread vs ensemble error (km) for the 1995–96 Atlantic hurricane seasons.

fore the forecast must be made, to obtain some measure of confidence to be attached to the ensemble forecast.

4. Summary and conclusions

In recent years, TC track forecasts from a number of global and regional numerical weather prediction models have become increasingly more important as guidance to forecasters at both NHC and JTWC. Forecast errors of a simple ensemble average of three models for the 1995–96 Atlantic hurricane seasons, and either three global models or two regional models for the western North Pacific during 1997, were compared with errors of the individual models. For the Atlantic, mean forecast track errors of the models (GFDL, NOGAPS, and UKMO) were on the order of 140–150 km at 24 h, 240–250 km at 48 h, and 350–380 km at 72 h. Mean errors for the ensemble were 120 km at 24 h, 194 km at 48 h, and 266 km at 72 h, which represent improvements of 16%, 20%, and 23% with respect to the best of the individual models. By comparison, mean errors of the official NHC Atlantic TC forecasts for the 1970s and 1980s were on the order of 200 km at 24 h, 420–450 km at 48 h, and 650–700 km at 72h. The 72-h ensemble forecasts were only a little worse than the 48-h model forecasts, which in turn were only a little worse than the 24-h official forecasts for the two previous decades.

For the western North Pacific, the forecast model performance was less uniform with mean position errors on the order of 140–160 km at 24 h, 250–290 km at 48 h, and 370–470 km at 72 h. Mean errors for the global model ensemble (GSM, NOGAPS, and UKMO) were 118 km at 24 h, 215 km at 48 h, and 322 km at 72 h, while those for the regional model ensemble (GFDN and TYM) were 115 km at 24 h, 215 km at 48 h, and 351 km at 72 h. The global model ensemble forecast errors were an improvement of 16%, 13%, and 12% at 24, 48, and 72 h, respectively, with respect to the GSM, which was the best of the individual models. The regional model ensemble errors were an improvement of 17%, 13%, and 11% with respect to the better of the two models at 24, 48, and 72 h. Again, to provide some perspective, mean errors of the JTWC forecasts during the 1970s and 1980s were on the order of 220 km at 24 h, 440 km at 48 h, and 660 km at 72 h. Thus, the 72-h model forecast errors were roughly comparable to the 48-h error of the JTWC forecasts for the two previous decades, while 72-h ensemble forecast errors were roughly comparable to the 36-h error.

TABLE 9. Distribution (percent) of 72-h ensemble forecast error (km) with respect to small (less than or equal to 300 km) and large (greater than 300 km) ensemble spread for the 1995–96 Atlantic hurricane seasons.

	<150	150–300	300–450	>450
Small	34	36	22	8
Large	14	39	25	22

In these comparisons the standard deviations of the forecast errors were reduced by the ensemble, which could have a direct impact on reducing the size of warning areas associated with TC landfall forecasts. The extent of these warning areas is directly related to the 95% level of uncertainty in the 24-h TC track forecasts, which for the Atlantic models ranged from 330 to 340 km. However, this quantity was 265 km for the ensemble, a reduction of approximately 20%. Such a reduction in the extent of TC coastal warning areas would result in significant savings in both monetary and human costs.

The spread of the ensemble forecast, defined to be the average distance of the ensemble member forecasts from the ensemble mean forecast, was found to possess some potential for use by forecasters as a measure of confidence in the ensemble forecast. For the 1995–96 Atlantic hurricane season, the ensemble spread was categorized as small (less than or equal to 300 km) or large (greater than 300 km). The mean 72-h ensemble forecast error was 241 km when the ensemble spread was small, compared with 327 km when the ensemble spread was large.

The joint ensembles examined in this study are simple and inexpensive to produce because they may be constructed from forecasts that are routinely produced by the various operational centers. By comparison, the application of single-model ensemble techniques to the TC forecasting problem using models comparable to those examined in this study will be an expensive proposition. The present success of these simple ensembles points to the need for continued diversity of high quality TC forecast models. In the future, an intriguing possibility is the addition of other high quality forecast models to those used in this study, which is expected to further improve the performance of this joint ensemble. The ultimate goal is to gain enough knowledge about the forecast performance of the ensemble mean and the individual ensemble members in different synoptic situations so that a forecaster can determine, a priori, whether the best forecast is likely to be provided by the en-

semble mean or by one (or more) of the ensemble members.

Acknowledgments. Special thanks are given to Mark DeMaria of NHC for providing the data for the 1995–96 Atlantic hurricane seasons, to Buck Sampson of NRL Monterey for unlocking the secrets of the Automated Tropical Cyclone Forecasting System, and to Russ Elsberry of the Naval Postgraduate School for a thorough and constructive review of the manuscript. This work was supported by the Oceanographer of the Navy through the program office at the Space and Naval Warfare Systems Command (PMW-185), Program Element 0603207N.

REFERENCES

Aberson, S. D., M. A. Bender, and R. E. Tuleya, 1998: Ensemble forecasting of tropical cyclone tracks. Preprints, *12th Conf. on Numerical Weather Prediction*, Phoenix, AZ, Amer. Meteor. Soc., 290–292.

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.,* **126,** 2503–2518.

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

DeMaria, M., S. D. Aberson, K. V. Ooyama, and S. J. Lord, 1992: A nested spectral model for hurricane track forecasting. *Mon. Wea. Rev.,* **120,** 1628–1643.

Elsberry, R. L., 1995: Recent advancements in dynamical tropical cyclone track predictions. *Meteor. Atmos. Phys.,* **56,** 81–99.

——, and F. D. Marks, 1998: Hurricane Landfall Workshop report. National Center for Atmospheric Research Tech. Note NCAR/TN-442, 40 pp. [Available online from http://box.mmm.ucar.edu/uswrp/.]

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

——, C. S. Velden, and J. D. Hawkins, 1998: The impact of multispectral *GOES-8* wind information on Atlantic tropical cyclone track forecasts in 1995. Part II: NOGAPS forecasts. *Mon. Wea. Rev.,* **126,** 1219–1227.

Harrison, M. S. J., T. N. Palmer, D. S. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: Two transplant case studies. *Quart. J. Roy. Meteor. Soc.,* **125,** 2487–2515.

Heming, J. T., and A. M. Radford, 1998: The performance of the United Kingdom Meteorological Office global model in predicting the tracks of Atlantic tropical cyclones in 1995. *Mon. Wea. Rev.,* **126,** 1323–1331.

——, J. C. L. Chan, and A. M. Radford, 1995: A new scheme for the initialisation of tropical cyclones in the UK Meteorological Office global model. *Meteor. Appl.,* **2,** 171–184.

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.

Houtekamer, P. L., and L. Lefaivre, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev.,* **125,** 2416–2426.

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.

Lawrence, M. B., C. J. McAdie, and J. M. Gross, 1997: Operational tropical cyclone track forecast verification at the National Hurricane Center. Preprints, *22d Conf. on Hurricanes and Tropical Meteorology,* Fort Collins, CO, Amer. Meteor. Soc., 475.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

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.

Lindgren, B. W., 1976: *Statistical Theory.* Macmillan, 614 pp.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

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

Neumann, C. J., and J. M. Pelissier, 1981a: An analysis of Atlantic tropical cyclone forecast errors, 1970–1979. *Mon. Wea. Rev.,* **109,** 1248–1266.

——, and ——, 1981b: Models for the prediction of tropical cyclone motions over the North Atlantic: An operational evaluation. *Mon. Wea. Rev.,* **109,** 522–538.

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.

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

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.,* **105,** 228–229.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

Zhang, Z., and T. N. Krishnamurti, 1997: Ensemble forecasting of hurricane tracks. *Bull. Amer. Meteor. Soc.,* **78,** 2785–2795.