NOTES AND CORRESPONDENCE

Consensus of Dynamical Tropical Cyclone Track Forecasts—Errors versus Spread

RUSSELL L. ELSBERRY AND LESTER E. CARR III

Department of Meteorology, Naval Postgraduate School, Monterey, California

27 December 1999 and 2 May 2000

ABSTRACT

The relationships between consensus spread of five dynamical model tracks and the consensus mean error is explored for a western North Pacific tropical cyclone database of 381 cases. Whereas a small spread of the five tracks is often indicative of a small consensus track error, some cases with large errors also are found even though the consensus spread is small. Some of the success of the dynamical model consensus approach arises because a substantial number (21%) of the cases with a large consensus spread have reduced errors after the consensus averaging. In nearly all the cases in this sample, the best of the five models has a 72-h track error of less than 300 n mi, but no tools are available to allow the forecaster to always select this best model. It is demonstrated that the forecaster can also add value by forming a selective consensus after first eliminating one or more likely erroneous track(s) and averaging the remaining tracks. Conceptual models and symptoms in the predicted fields to assist the forecaster in this error detection have been separately described by the authors, and their successful application would result in more accurate selective consensus forecasts than nonselective consensus forecasts.

1. Background

Goerss (2000) has proposed a simplified, economical consensus of tropical cyclone track forecasts from up to three global models at 0000 and 1200 UTC and two regional models at 0600 and 1800 UTC. On average, the Goerss consensus track errors are smaller than those of all the individual models, which indicates that the averaging process provides a smoothing of the random error components as in an ensemble prediction system. This consensus approach yields either the most accurate or secondmost accurate track forecast in more than 70% of the cases (J. Goerss 1999, personal communication). While the consensus track forecast can never be the worst track, the forecaster needs guidance as to when the consensus may be accurate or may not be that useful.

Even though the Goerss ‘‘ensemble’’ contains only a small number of ‘‘members,’’ each model is a sophisticated baroclinic model with small systematic errors. The combination of perturbed initial conditions as represented by the different center analyses and the multiple, sophisticated models in the Goerss consensus approach gives it some of the qualities of an ensemble prediction system. However, the distinctions between a consensus of track forecasts and an ensemble prediction system are noteworthy. Whereas the consensus is of a small number of tropical cyclone track positions, an ensemble prediction system generally addresses an area-averaged (often a hemisphere) measure of model skill. That is, one midlatitude cyclone could be poorly predicted and not contribute greatly to a hemisphere-mean skill measure if the remainder of the circulation was well forecast.

Another essential objective of an ensemble prediction system is to predict the probability density function of the solution. If the prediction model were perfect, this probability density function would be uniquely related to the uncertainty in the initial conditions. Thus one approach is to define perturbed initial conditions that represent the uncertainty in the analysis and then integrate these ensemble members with a single model. Since the model is not perfect, some of the uncertainty in the solution is model related. Thus, a second ensemble approach is to combine the solutions of multiple models. The spread in an ensemble prediction system typically is a second-order moment over the ensemble member solutions, as well as over the hemispheric domain. In the consensus track approach, the spread is defined as some geometrical measure (Goerss used the average) of the dispersion of the cyclone positions about the consensus mean position at each time. This spread definition is clearly not analogous to the spread of the ensemble prediction system.

Corresponding author address: R. L. Elsberry, Department of Meteorology, Naval Postgraduate School, Code MR/Es, 589 Dyer Rd., Room 254, Monterey, CA 93943-5114.
E-mail: elsberry@met.nps.navy.mil

One of the desired qualities of an ensemble prediction system (and the consensus approach) is that the spread of the ensemble member solutions about the ensemble mean will contain the true solution, which may require large numbers of members. Thus, the ensemble prediction system developer is generally trying to increase the spread to ensure the true solution is within the ensemble solution swarm.

Since the errors in the initial analysis grow most rapidly in regions of instability, a greater spread among the ensemble members may indicate a greater uncertainty. Unfortunately, this spread is not necessarily a good indicator of the likely error of the ensemble mean (Buizza 1997; Whitaker and Loughe 1998).

It would help the tropical cyclone forecaster if the spread among the consensus dynamical model tracks would provide an indicator of track accuracy. In the authors' experience, many tropical cyclone forecasters believe that situations with small spread in track guidance are the ones in which they are more successful. Conversely, large spread in their model track guidance is believed to indicate the potential for large track forecast errors. These large spread situations are when the forecaster needs the most help in interpreting the guidance. Thus, it is desirable to determine if this perceived spread–error relationship holds for the dynamical model consensus.

Given the small (either two or three) consensus sizes, it is not surprising that Goerss (2000) found the spread of the tropical cyclone tracks was not well correlated with the consensus mean track errors. Whereas a small spread seemed to be associated with smaller consensus track errors, it was not true that large errors necessarily were associated with large spreads.

The purpose of this note is to explore further this error–spread relationship for dynamical track predictions of western North Pacific tropical cyclones. The first objective is to document the frequency of cases with small/large spread in relation to the small/large consensus errors and, thus, clarify why a simple numerical consensus (or nonselective NCON) may be associated with either small or large spreads among the dynamical model tracks. A second objective is to provide a motivation for searching for symptoms of large track errors, and after eliminating such tracks then form a ''selective consensus'' (SCON) track that is more accurate than the NCON track. The procedures by which the forecaster might recognize the large-error tracks are covered separately by Carr and Elsberry (1999) and will not be described here. Rather, the purpose here is to propose that the forecaster could in certain situations improve upon the simple Goerss consensus if the meteorological reasoning was available to form a selective consensus.

2. Data source

The database for this consensus study of dynamical track predictions is the same as in the Carr and Elsberry (1999) study of large track errors. The five models considered here are the U.S. Navy Operational Global Atmospheric Prediction System (NOGAPS), Geophysical Fluid Dynamics Laboratory Hurricane Prediction System—Navy version (GFDN); U.K. Met. Office (UKMO) global model, Japan Global Spectral Model (JGSM), and Japan Typhoon Model (JTYM). These are the same models as in the Goerss (2000), who formed a separate consensus for the global (NOGAPS, UKMO, and JGSM) models available at 0000 and 1200 UTC and for the regional models (GFDN and JTYM) available at 0600 and 1800 UTC.

The first step is to form a five-member consensus by relabeling the 0600 UTC (1800 UTC) regional model track positions that correspond to the 1200 UTC (0000 UTC) global model positions. To obtain a 72-h position from the regional models necessary for the consensus calculation, a 78-h position is extrapolated using the 60- and 72-h positions. Similarly, the time-coincident positions from the 0000 UTC (1200 UTC) global model forecast are relabeled to correspond to those of the 0600 UTC (1800 UTC) regional model forecast. Except for the JGSM, extrapolation is not necessary with the global models since those tracks generally extend to 120 h. Therefore, five consensus members are available every 6 h when all of the dynamical model tracks were received at the Joint Typhoon Warning Center (JTWC). Only those 381 cases for which all five model tracks were available will be considered here. Of course, these cases are not independent since some are separated by only 12 h. Carr and Elsberry (1999) found that large track error scenarios may suddenly begin or end in just 12 h, or may persist for a series of forecasts over several days.

The consensus track errors are calculated relative to the poststorm (best track) positions. Since at most five positions are available, the spread of the consensus is defined as the maximum displacement from the consensus (centroid) position. The maximum distance is selected rather than the root-mean square of the five model forecast positions because the focus here is on isolating the likely erroneous position, and this is frequently (but not always) the forecast with the maximum displacement. Such a definition also maximizes the likelihood of finding a spread–error relationship. Although 24- and 48-h consensus positions/errors and spreads are also calculated, the focus here is on the 72-h positions.

3. Consensus mean error versus spread

The consensus forecast track errors as a function of the consensus spread for the 381 cases in which all five models are available are presented in Fig. 1a. Only in a relatively small number of cases is the spread less than 100 n mi (185 km) when five members are present. Although the majority of the spreads are less than 300 n mi (555 km), values as large as 900 n mi (1575 km) are found. A few of the largest errors do occur along

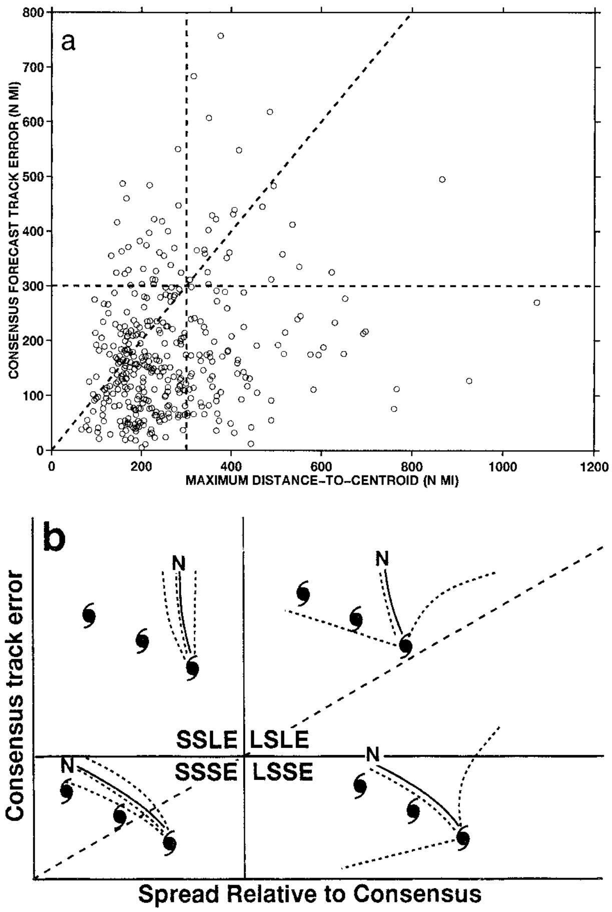

FIG. 1. (a) Five-member dynamical model consensus 72-h forecast track errors for a 1996–97 database of 381 western North Pacific tropical cyclone forecasts as a function of the maximum distance to the consensus mean position. For convenience, the thresholds for the large and small spreads and errors are both set at 300 n mi (555 km). (b) Schematics of a three-track (dashed lines) consensus (solid line labeled N) relative to the initial and verifying positions (filled typhoon symbols) to illustrate the small spread (SS) and large spread (LS) versus small error (SE) and large error (LE) relationships in (a). See text for description and Fig. 2 for actual tracks.

the diagonal with increasing spread. However, the correlation coefficient between spread and error is only 0.24 for the 381 cases, which implies that only about 5% of the variance in consensus mean error can be explained by the spread. A total of 81 cases have a spread greater than 300 n mi and yet have a consensus mean 72-h error of less than 300 n mi. Thus, a large spread does not necessarily indicate a large track error. Perhaps the most disconcerting situation for the forecaster is the approximately 30 cases of small consensus spread that have an error of more than 300 n mi. That is, a small spread in the track forecast guidance does not always mean a small track error.

The schematic in Fig. 1b is intended to assist in the interpretation of the errors in Fig. 1a. For ease of interpretation, only three tracks will be shown and the consensus track is indicated by the solid line. These three tracks might be considered as members of a ''tight'' cluster (small spread) or a ''broad'' cluster (large spread). These three tracks might also represent separate sets of the five dynamical model tracks (grouped into one to three tracks), except then the consensus position would be weighted appropriately.

Four regions are defined in the spread–error diagram (Fig. 1b): (i) small spread–small error (SSSE), (ii) small spread–large error (SSLE), (iii) large spread–large error (LSLE), and (iv) large spread–small error (LSSE). The three idealized forecast tracks in the SSSE region are quite similar (small spread) and the consensus track (labeled N) closely matches the actual positions (small error). An example of a SSSE situation is given in Fig. 2a. The 72-h consensus forecast of Typhoon (TY) Keith at 1200 UTC 31 October 1997 is one of the smallest (37 n mi, 68 km) errors in Fig. 1a. This tightly clustered group of five 72-h positions has a spread (maximum distance to the centroid) of only 84 n mi (155 km). The forecaster would have little reason to question this tightly clustered model guidance, and would be well advised to select the consensus track.

Three consistent tracks may also depart significantly from the actual track, which is the SSLE case in Fig. 1b. Two models that are not independent (e.g., a regional model driven by initial and boundary conditions from a global model) may share such a large error scenario. However, a cluster of erroneous tracks may also occur if all of the models are unable to predict an anomalous track scenario. Tightly clustered (spread of 167 n mi, 309 km) guidance from all five models initiated for TY Opal at 1200 UTC 17 June 1997 is indicated in Fig. 2b. All of these track errors exceed 300 n mi (555 km) as Opal has accelerated to the northeast, whereas most of the model guidance is for a continued slow, northward track.

The two outer tracks in the LSLE region in Fig. 1b suggest a bifurcation between a straight mover and a recurver track. The middle track, which is close to the consensus, is an intermediate track that may be highly unlikely to occur, and thus may have significant errors. Even though the intermediate track may be the one with statistically the smallest long-term average if such bifurcation situations are truly chaotic or indeterminate, it can be argued that the best warning strategy is to adopt one scenario, and present the other scenario as a lower probability alternative. As the sample size is increased, a probability distribution of the more (less) likely scenario might be provided to the forecaster. A large spread exists in the five model 72-h positions for TY Joan at 0000 UTC 18 October 1997 (Fig. 2c) because the JGSM has forecast a slow recurvature and the other four models have forecast an early and rapid recurvature. Even though the JGSM position is an outlier relative to the other four models, it is an excellent forecast with an error of only 77 n mi (142 km). However, the consensus forecast error is 483 n mi (894 km).

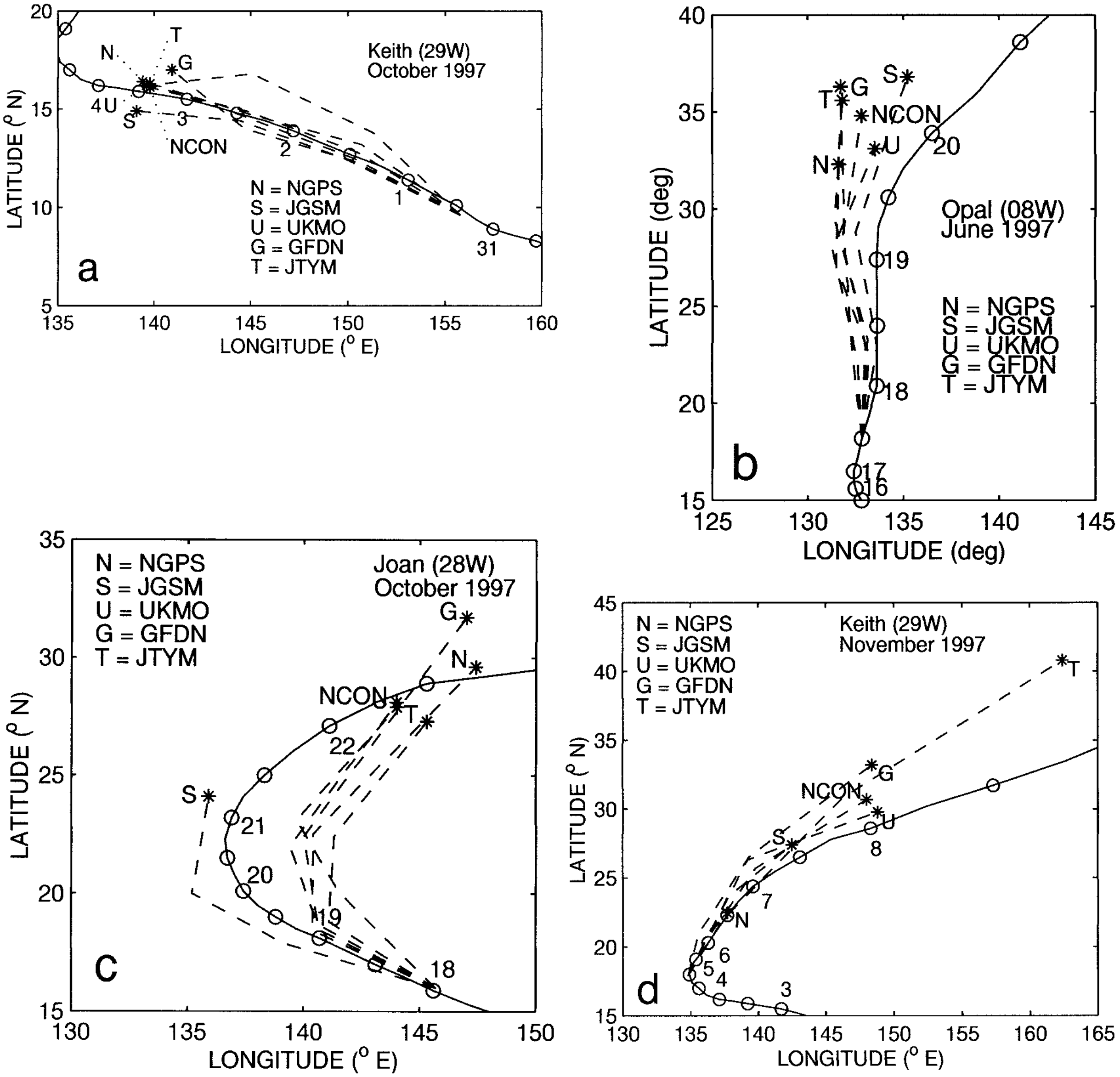

FIG. 2. Examples of five-member dynamical model (see inset for model descriptions) for the four spread–error categories defined in Fig. 1b. (a) SSSE case of TY Keith forecasts initiated at 1200 UTC 31 Oct 1997, (b) SSLE case of Opal forecasts initiated at 1200 UTC 17 Jun 1997, (c) LSLE case of Joan forecasts initiated at 0000 UTC 18 Oct 1997, and (d) LSSE case of Keith forecasts initiated at 0000 UTC 5 Nov 1997. The JTWC poststorm track (solid line) has circles at 0000 and 1200 UTC with the date adjacent to the 0000 UTC position.

In the LSSE region of Fig. 1b, the consensus track error is small even though the consensus spread is large. In this case, the cyclone path is between the two outer guidance tracks, which can occur when these two models have errors that happen to compensate. As indicated in the introduction, this is a difficult situation for the forecaster because the potential for a large error is great with the selection of an incorrect track. In the LSSE and SSSE cases, the five tracks may be considered to be simulating an ensemble prediction system and the consensus spread represents the growth in uncertainty from different initial conditions and different dynamical model characteristics. Other intelligence is required to determine if the present situation may be in the SSLE and LSLE categories, in which the consensus mean is not a good forecast. An example (Fig. 2d) is TY Keith at a later stage than in Fig. 2a after a sharp recurvature. A large spread (926 n mi, 1713 km) about the consensus mean exists because the JTYM has predicted excessive acceleration of the storm. However, the NOGAPS and JGSM 72-h positions lag the other models such that these compensating errors lead to a consensus forecast that has an error of only 127 n mi (234 km).

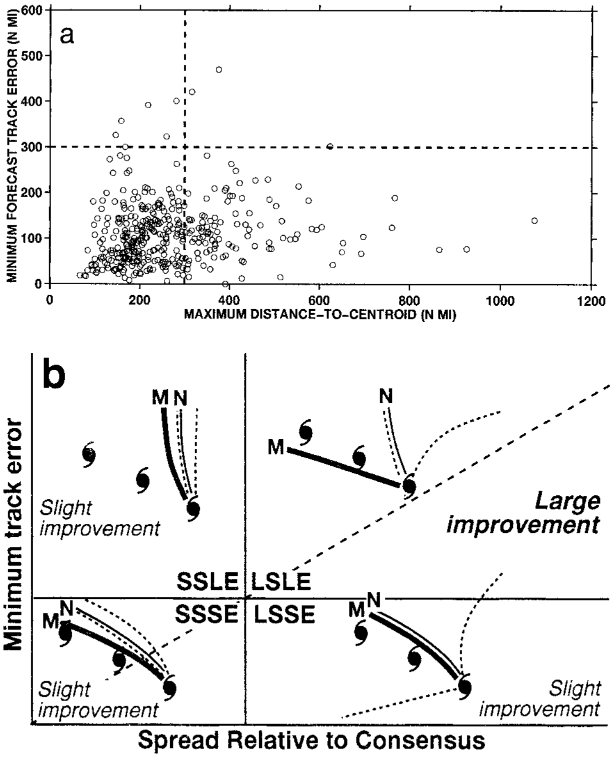

FIG. 3. As in Fig. 1 except for the minimum-error track forecast from the five-member consensus of dynamical models. (b) The solid track (labeled M) is the minimum-error track, and the potential improvement to be gained by its selection is indicated.

4. Searching for the minimum-error track

One motivation for the forecaster to spend the effort to improve on the consensus track forecast would be a demonstration that (at least) one of the model tracks is a better forecast than the consensus. If it had been possible for the forecaster to always be able to select that model track with the minimum 72-h error, the distribution of these (minimum) errors as a function of the consensus spread would be as shown in Fig. 3a. Almost all of these errors would be less than 200 n mi (370 km) and only seven would be greater than 300 n mi (555 km). This result is an indication of the improved quality of the dynamical track guidance in recent years. Even when the consensus spread is quite large, at least one of the model tracks is relatively close to the actual track.

While this is a motivation for the forecaster to ‘‘add value’’ relative to the consensus of the guidance, it is not expected that the forecaster can always select the model track with the minimum track error. Track schematics that illustrate possible error–spread distributions are shown in Fig. 3b. A common characteristic in these schematics is that the minimum-error track (labeled M) is closer to the actual track than is the consensus mean, which by definition has to lie near the midpoint of the three tracks in this schematic.

In the SSSE category of Fig. 3b, the spread is so small that the minimum-error model track selection is not going to provide much improvement. Nor is it likely that guidance can be provided to forecasters that would enable them to make consistently such a selection when the spread is small. In the SSSE illustration for the consensus mean errors in Fig. 2a, little advantage would be gained by selecting any one of the model 72-h positions in such a tightly clustered group, and the consensus position is as likely as any of the five models. In the SSLE category in Fig. 3b, the small spread among the three tracks about the consensus again means only a small improvement is possible from a minimum-error selection. The only way to decrease significantly the large forecast errors in the SSLE category is to reject the guidance provided by the dynamical models. Strong evidence from other guidance and good forecaster skill would be needed to make such a decision. In the SSLE illustration for the consensus mean errors in Fig. 2b, selection of the JGSM track would slightly decrease the 72-h error relative to the consensus and the other four models. If the forecaster has some evidence that the tropical cyclone was likely to turn to the northeast after passing 30°N, this might provide a basis to select the JGSM position over the consensus. While the JGSM often provides good guidance in such track scenarios, which might be considered to be a basis for selecting it as guidance, in other similar cases the JGSM may provide the worst guidance among the five models.

In the special bifurcation scenario in the LSSE three-track schematic in Fig. 3b, selection of the correct solution may again result in only a small improvement relative to the consensus track. In this schematic, the minimum-error track is close to the consensus track, which has only a small error. As described above, the large spread scenario is difficult for the forecaster since it is not obvious that the best solution is in the middle of widely spread tracks.

The LSLE category track schematic in Fig. 3b is also suggestive of a bifurcation scenario in which selection of the best of the three models would result in a large improvement over the consensus. If a large error is defined as 300 n mi (555 km) and five dynamical models are available, only 3 of the 381 minimum-error tracks in this sample fell in the LSLE category (Fig. 3a). If less than five dynamical models are available, more cases will fall in this LSLE category.

The minimum-error model errors in Fig. 3a indicate the consensus spread may be very large (>400 n mi; 740 km) and yet one of the model tracks will have quite a small (<225 n mi; 416 km) error at 72 h. An excellent example of how an outlier model (i.e., JGSM) 72-h position would have provided guidance with a very small error is given in Fig. 2c. The forecaster would need evidence that the other four models are incorrectly forecasting an early and rapid recurvature, while the outlier JGSM forecast of a delayed recurvature is indeed the correct guidance. Carr and Elsberry (1999) describe conceptual models and symptoms in the wind and sea level pressure fields that might assist the forecaster in such a situation. Whereas much is to be gained in this case from abandoning the consensus and choosing the outlier, a large error will result if the decision is wrong.

Thus, the large-spread–large-error situations offer opportunities for decreased track errors relative to a consensus forecast—provided a technique is available for the forecaster to isolate consistently the dynamical model guidance with the minimum forecast error. If instead of three individual tracks as illustrated in Fig. 1b, there are two clusters of guidance, this conclusion may be stated that selection of that cluster of forecasts in the large consensus spread scenario with the minimum track forecast error prior to the calculation of a consensus (average) track will decrease significantly the number of busted forecasts. Rather than taking the path of least regret based on a consensus of all track guidance in the large-spread scenario, a ‘‘selective consensus’’ would be a significant improvement.

5. Searching for the maximum-error track

A similar motivation for improving track prediction via a selective consensus exists if the forecaster can eliminate the track (or the cluster of tracks) that has the maximum track error prior to calculating the consensus (average). Here, the goal would be to select (and reject) the worst track position, or that cluster of predicted tracks, for which evidence exists that erroneous processes in the dynamical model are probably contributing to large errors. Carr and Elsberry (1999) have described a series of conceptual models based on the physical processes contributing to tropical cyclone motion that may be handled erroneously by the dynamical model. Recognition (subjectively or objectively) of such an erroneous track forecast scenario based on specific track changes (curvature, deceleration/acceleration, departures from past track, etc.), or circulation features in the predicted wind and sea level pressure fields, will be required to decide that the large-error conceptual model applies so that the erroneous track guidance can be rejected.

Unless all of the five consensus members are identical, at least one member must also have a larger track error than the consensus error. By eliminating an erroneous track(s) prior to calculating the (selective) consensus, the track error will usually be reduced. As shown in Fig. 4a, these maximum-error 72-h positions are typically greater than 300 n mi (555 km) and may be as large at 1300 n mi (2405 km)! While a general trend exists of increasing maximum error with increasing consensus spread, the scatter of errors for the same spread is large. Notice that most of these maximum errors are larger than the consensus spread (i.e., to the left of the diagonal). Only a few cases exist in which the maximum error is less than the spread.

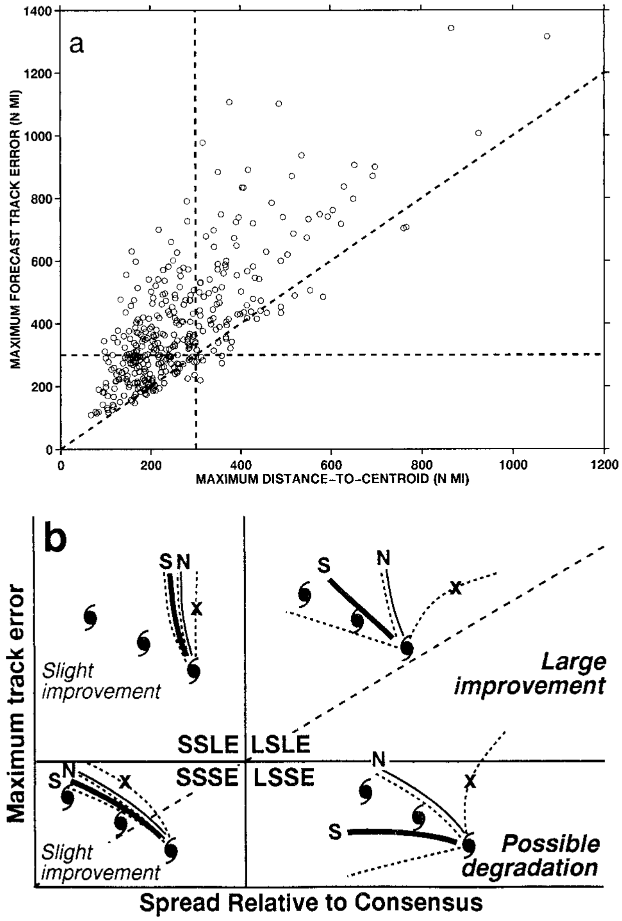

FIG. 4. As in Fig. 3 except for the maximum-error track forecast from the five-member consensus of dynamical models. (b) The solid track (labeled S) is the selective consensus, which is the average of the two remaining tracks after removal of the maximum-error track (dashed line with x to indicate deletion).

The schematics (Fig. 4b) to help visualize these maximum-error situations are similar to Fig. 3b, with the maximum-error track marked by the x to indicate deletion. In the SSSE category, not much is to be gained from forming a selective consensus by rejecting a maximum-error track that is not much different from the minimum-error or consensus track because the spread is small. As stated above, it will be difficult to select consistently from among the model tracks when the spread is small. The example in Fig. 2a applies here; even the maximum-error 72-h position is not that poor of a forecast. Specifically, eliminating the largest 72-h error by the GFDN model in Fig. 2a would not decrease much the consensus track error. Similarly, the small-spread large-error schematic in Fig. 4b indicates that, while it is desirable to eliminate the maximum-error track, it will not result in the (large) consensus track error being much improved. An example would be eliminating the NOGAPS 72-h error in Fig. 2b. As stated previously, the only way to decrease significantly the large forecast errors in the SSLE category is to reject the guidance provided by all of the five dynamical models, which is a decision that would require strong evidence and good forecaster skill.

The large number of LSLE cases with a consensus spread >300 n mi (555 km) and maximum-error values >300 n mi in Fig. 4a indicates the importance of detecting and eliminating such model guidance before calculating the selective consensus. As shown in the LSLE region of the schematic in Fig. 4b, these large-spread cases are likely to be bifurcation scenarios. Especially in the two-branch bifurcation scenario, it is important to reject the ‘‘wrong branch’’ with the maximum error. These situations are precisely the focus of the large-error conceptual models of Carr and Elsberry (1999). Their premise is that it will be easier for the forecaster to detect particularly erroneous tracks than it will be to try to detect the best model track in each situation. Granted it will be a challenge in some situations to eliminate the wrong branch, but the reward of eliminating the magnitude of errors that are found in the LSLE region of Fig. 4a make the effort worthwhile, because to make such large errors would be damaging to forecaster credibility.

An example where eliminating the maximum-error model (GFDN) in an LSLE situation would not have much impact is shown in Fig. 2c. This case is one in which the model (JGSM) 72-h position that represents the large spread from the consensus is not the model with the large error. Thus, rejecting the wrong branch in this case means eliminating four models that are fairly well clustered, and then accepting as the correct guidance a real outlier (JGSM).

At least for the sample in Fig. 4a, it is extremely rare for a five-member consensus to have a maximum error in the large-spread–small-error category where the spread and error magnitudes are both defined to be equal to 300 n mi (555 km). The schematic tracks in the LSSE category in Fig. 4b illustrate a possible degradation in the selective consensus if two models with large compensating errors are present. Eliminating just one of the two erroneous tracks must then degrade the accuracy. The case in Fig. 2d is a useful illustration in that eliminating just one of the fast (JTYM) or the slow (NOGAPS) 72-h positions would actually degrade the consensus of the remaining four.

In summary, the identification of the dynamical model track guidance with the maximum error is a worthy goal for the forecaster. While this goal will not reduce the track forecast error as much as being able to always pick the best of a five-member consensus, the forecaster can add value by detecting erroneous model tracks. Forming a selective consensus by first removing the erroneous track(s) prior to the calculation of the average of the remaining tracks will usually reduce the error.

6. A simple selective versus nonselective consensus comparison

As a simple illustration of the potential error reductions from a selective consensus, assume that the maximum-error model track position in Fig. 4a could always be detected and eliminated before calculating the consensus of the remaining four members. Subtracting this simple SCON track error from the NCON errors results in a large number of reduced errors (Fig. 5). Many of these reductions are only of the order of 0–50 n mi (0–92 km) when the NCON track errors are less than 200 n mi (370 km). Many of these cases would be the small consensus spread cases in Fig. 1a, and not much is to be gained in rejecting the maximum-error position. An example is the SSSE case in Fig. 2a. Although not as numerous, other cases of an SCON improvement over an NCON forecast have values from 50 to 200 n mi. Particularly when this improvement is relative to an already good NCON forecast of say less than 200 n mi, the percentage improvement from a 50–200 n mi error reduction would be good. By contrast, some cases in which the NCON forecast is not good (72 h errors >300 n mi), the SCON improvement by eliminating just one maximum-error forecast does not result in much of an improvement. An example is shown in Fig. 2c in which four models would have to be rejected to get an accurate consensus forecast. Such scenarios are infrequent with the skill of these five dynamical model forecasts.

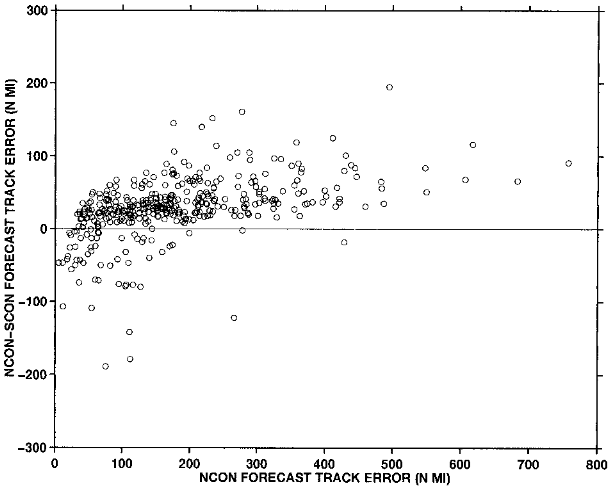

FIG. 5. Difference between NCON and an idealized four-member SCON 72-h track forecast error when the maximum-error track among the five-member consensus is eliminated as a function of the NCON error. The database is the same as in Figs. 1–4.

In a few cases, the SCON forecast defined by eliminating the one model position with the maximum error results in a degradation (Fig. 5). Most of these degradations of less than 50 n mi (92 km) are for cases in which the NCON forecast errors were less than about 50 n mi. These cases reflect a somewhat random nature of a consensus average of five members. The variability among the members is such that the consensus averaging results in a cancellation of errors. Omitting the maximum-error position from the consensus then may actually be a degradation if another model error is no longer canceled. Some of these large SCON forecast degradations of 50 n mi (92 km) to more than 200 n mi (370 km) when the NCON forecast error was less than 100 n mi (185 km) in Fig. 5 are from eliminating only one of two models with compensating errors. An excellent example is in Fig. 2d. Eliminating only the highly erroneous (fast) JTYM forecast without eliminating the (slow) NOGAPS forecast would degrade the excellent NCON forecast.

This simple demonstration of the potential benefit of a selective consensus vis-à-vis a nonselective consensus illustrates that almost all cases are improved. The degradation cases tend to be small and are mostly for smaller NCON errors. Just eliminating one model track when another model has a counterbalancing error will also degrade the NCON forecast. A model traits knowledge base to assist the forecaster in detecting these likely erroneous track forecasts is provided by Carr and Elsberry (1999).

7. Summary

This research illustrates the benefit of a consensus tropical cyclone track prediction approach based on three global and two regional models that are relabeled and extrapolated as required to provide five tracks each 6 h. This relabeling to a common time is a slight extension of the Goerss (2000) consensus approach in which three global models are averaged at 0000 and 1200 UTC and two regional models are averaged at 0600 and 1800 UTC.

The primary objective has been to describe the relationship of the consensus track error to the spread of the consensus. A small spread of tracks (tightly clustered) is often an indication of a small error of the consensus track forecast. However, the approximately 8% cases of small consensus spread that have large (>300 n mi; 555 km) track errors are difficult situations for the forecaster as the storm track typically lies outside of the model guidance. Rejecting all of the dynamical guidance is usually not advisable, but this drastic measure is the only procedure to reduce the large consensus track errors in these small consensus spread situations.

One benefit of the five dynamical model consensus approach is the substantial number (approximately 21%) of cases with small consensus track errors even though the consensus spread is large. This averaging to reduce random errors is a characteristic of ensemble prediction systems with many more members. Here, a favorable result is achieved with only five dynamical models that have slightly different initial conditions and different model characteristics. While these dynamical model tracks are normally quite good, their guidance can also have large errors when the consensus spread is large. Another objective of an ensemble prediction system to relate the spread to the mean error is even less successful in this consensus approach. Since the consensus spread explains only 5% of the variance in the mean forecast error, the spread cannot be used directly to evaluate likely forecast error.

Two ways that a forecaster can add value over a consensus forecast are illustrated. A significant reduction in the occurrence of large track errors could be achieved if the forecaster had guidance to always pick the best of the five dynamical model tracks. In nearly all cases in this western North Pacific database, the best of the five models will have a 72-h track error of less than 300 n mi. Unfortunately, no accurate and consistent tool or guidance is available to assist the forecaster in selecting the best model in every situation.

Another value-added approach in consensus forecasting is to reject the likely erroneous track, which frequently has a 72-h track error exceeding 300 n mi (555 km). Eliminating such erroneous tracks prior to calculating the average of the remaining tracks is defined here as a ‘‘selective consensus.’’ A simple demonstration of such a selective consensus by removing just the maximum-error model and calculating the average of the four remaining model tracks nearly always reduces the track error over the five-member consensus. Carr and Elsberry (1999) provide conceptual models and symptoms in the predicted fields that the forecaster may use to detect large-error situations. If the forecaster can learn to use such guidance to recognize model tracks with large (and possibly compensating) errors, and then eliminate the erroneous tracks, the selective consensus approach will result in a further tropical cyclone track prediction improvement over the nonselective consensus.

Acknowledgments. This research has been sponsored by the Office of Naval Research Marine Meteorology. Comments by Dr. Jim Goerss and three anonymous reviewers improved the manuscript. Mark Boothe provided a much improved display for Figs. 1b, 3b, and 4b and made comments on the manuscript, which was skillfully prepared by Mrs. Penny Jones.

REFERENCES

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **125,** 99–119.

Carr, L. E., III, and R. L. Elsberry, 1999: Systematic and integrated approach to tropical cyclone track forecasting. Part III: Traits knowledge base for JTWC track forecast models in the western North Pacific. Tech. Rep. NPS-MR-99-002, Naval Postgraduate School, Monterey, CA, 227 pp. [Available from R. L. Elsberry, Dept. of Meteorology, Naval Postgraduate School, Code MR/Es, 589 Dyer Rd., Room 254, Monterey, CA 93943-5114.]

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

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.,* **126,** 3292–3302.