Ensemble Forecasting of Tropical Cyclone Motion Using a Barotropic Model. Part I: Perturbations of the Environment

KEVIN K. W. CHEUNG AND JOHNNY C. L. CHAN

Department of Physics and Materials Science, City University of Hong Kong, Kowloon, Hong Kong, China

(Manuscript received 9 October 1997, in final form 7 May 1998)

ABSTRACT

The technique of ensemble forecasting is applied to the problem of tropical cyclone motion prediction. Three methods of generating perturbations for the environmental flow, Monte Carlo forecast (MCF), lagged-average forecast (LAF), and the breeding of growing modes (BGM), are tested with a barotropic model using 66 cases from the Tropical Cyclone Motion (TCM-90) Experiment. For the MCF, the ensemble mean forecast is almost identical to that without any perturbation. The other two methodologies are verified both under the perfect model assumption and using the best tracks. On average, in about half of the cases improvement in forecast can be demonstrated in the former verification. A high degree of correlation (with linear correlation coefficient >0.9) is also found between the spread of the ensemble and the root-mean-square forecast error. In the best-track verification, improvement in forecasts can also be obtained in 36% (42%) of all the cases using the LAF (BGM) technique. The spread-skill correlation is still significant (correlation coefficients vary from ~ 0.4 to 0.7 for different forecast times). An examination of the synoptic flow associated with cases in which the forecast is improved suggests some favorable conditions for the application of ensemble forecasting. These include a tropical cyclone (TC) making a transition from one synoptic region to another, an apparent break in the subtropical ridge (STR), a rapid strengthening/weakening of the STR, recurvature of a TC, and multiple-TC situations.

1. Introduction

The two major current directions of development in numerical weather prediction (NWP) are to improve the configuration of the model itself and to formulate it as a basic problem in nonlinear science. Benefiting from the advance of computing technology, the resolution of a current NWP (both global and mesoscale) model can be increased substantially. Research in improving numerical methods and physical parameterizations in NWP models is also continuously progressing. However, studies of the predictability problem indicate that these efforts will eventually reach a limit in improving the forecast due to the sensitivity of a prediction to its initial conditions. In other words, there is the danger that the forecast from a sophisticated high-resolution model can be accurate but not correct (Brooks and Doswell 1993). The strategy used to tackle this aspect of NWP since the last decade is ensemble forecasting. In this methodology, an ensemble of model runs, which start from slightly different initial conditions, are used to monitor the growth of errors inherent in the analyses. By taking the ensemble mean, contributions from some of these errors can be cancelled and thus, in general, a higher skill can be obtained.

The ensemble members are usually generated by adding perturbations to the analyses. Two attributes of these perturbations are obviously important to the quality of an ensemble: their representativeness of the analysis errors, and the sampling by the ensemble of the most essential growing patterns with respect to the weather system under study. Quite a few perturbation methodologies have been designed for different purposes. Some specifically treat global medium-/extended-range forecasts such as the singular vector perturbation currently used in the European Centre for Medium-Range Weather Forecasts (ECMWF) model (Molteni et al. 1996). Recently, some of the perturbation techniques have also been applied to short-range ensemble forecasting (Brooks et al. 1996), and the weather systems of concern are extended to those rare events like explosive cyclogenesis and heavy precipitation. The role of ensemble forecasting in this forecast range is to provide alternative scenarios of the evolution of these phenomena that are often missed in an ordinary NWP forecast.

Most of the weather systems that have been predicted with the ensemble forecasting technique are in the midlatitudes. In the Tropics, the tropical cyclone (TC) is the most significant system. A better prediction of its movement is obviously important. In this study, three different

Corresponding author address: Dr. Johnny C. Chan, Department of Physics and Materials Science, City University of Hong Kong, 83 Tat Chee Ave., Kowloon, Hong Kong, China.
E-mail: apjcchan@cityu.edu.hk

techniques to generate the perturbations are applied to a barotropic model for forecasting the motion of a TC. They are the Monte Carlo forecast (MCF), the lagged-average forecast (LAF), and the breeding of growing modes (BGM) method. The MCF (Leith 1974; Palmer et al. 1990) is a purely statistical method in which random noise is added to the analysis as perturbation. While it has the advantage of ease of generating a large number of members, it has little use in modeling highly unstable systems such as midlatitude baroclinic waves (Toth and Kalnay 1993). The random noise is difficult to organize into these unstable structures, which are often missed by the ensemble. This is the reason why MCF is not adopted as the operational scheme to generate perturbations by major centers such as the U.S. National Centers for Environmental Prediction (NCEP) and the ECMWF. In the short range, however, the applicability of MCF was reestablished recently. Ramamurthy and Shu (1996) reported that superior performance over an ordinary forecast of an ensemble with scale-dependent initial errors can be obtained for short-range forecasts on the NCEP Meso Eta Model. Du et al. (1997) also found that the quantitative precipitation forecast associated with cyclogenesis can be improved using the MCF in a mesoscale model. In light of these results, this option of generating perturbations is retained in our experiments.

The LAF was introduced as an alternative to the MCF (Hoffman and Kalnay 1983). The methodology uses forecasts that lag the initial time for different periods to be the ensemble members. Therefore, the perturbations are the short-range forecast errors resulted from the growth of the initial errors in the analyses according to the dynamics of the model. Dalcher et al. (1988) applied this technique to the ECMWF forecasts and obtained an improvement in skill over the operational ones after 5 days, although the correlation of the spread of the ensemble with the forecast skill is low. Application of this methodology in the extended range to enable the ensemble members to capture some specific features in the extratropics, for example, a blocking event, has also not been very successful (Branković et al. 1990). The reason is partially due to the systematic errors inherent in the model, but it also indicates the inadequacy of the LAF methodology in providing some skillful members within the ensemble. In view of this weakness of the LAF, another perturbation methodology, the BGM, was developed in the Environmental Modeling Center (EMC) of NCEP and adopted as their operational ensemble forecasting system since 1992. More details of this methodology will be given in section 3 where its implementation into our experiments is described.

The subject of study in this paper, the problem of TC motion prediction, has long been recognized as a multiscale problem. The size of the vortex of a TC can vary substantially. And during the lifetime of a cyclone, it will interact with different weather systems possessing different characteristic length scales, which together compose the synoptic environment in which the TC is embedded. Hence, for a numerical prediction of TC motion to be successful, ideally the model should have good skill in the whole spectrum of scales. An actual model, however, will certainly suffer from a number of deficiencies that may include, in this case, a resolution insufficient to resolve the structure of the TC vortex and inadequacy of the convective parameterization scheme in simulating a realistic warm-core structure. In addition, any problem that biases the forecast of the environmental flow and/or the vortex structure can ruin the final forecast. Hence, besides its original idea of dealing with the growth of analysis errors, ensemble forecasting can play a role in simulating the uncertainties due to model deficiencies. Such a role has been realized by other researchers who applied ensemble forecasting in the short-range domain and to mesoscale phenomena, because in these problems model deficiencies can often dominate the forecast error. Therefore, in order to generate efficient and useful perturbations, the model physics and/or the configuration of the model should also be varied. This can be accomplished by systematically changing the existing configuration of a model. Another way, as suggested more recently (Harrison et al. 1995; Richardson et al. 1996), is to form a hybrid ensemble that consists of analyses assimilated from different models.

Some work has already been carried out in applying ensemble forecasting ideas to TC-related problems. Aberson et al. (1995) used the NCEP ensemble members as initial conditions for the VICBAR barotropic model to forecast Atlantic hurricanes. Morison et al. (1996) applied the Monte Carlo method to western South Pacific TCs, with specific attention paid to landfall probability. Vitart et al. (1997) also used the Monte Carlo technique in a general circulation model to simulate the interannual variability of cyclogenesis.

To isolate the two main factors that determine the motion of TCs, namely, the environmental steering and the vortex structure, the present study is divided into two parts. Part I concentrates on the applications of the three perturbation methodologies to the environmental flow, and the intercomparison of their utility. The implementation of these perturbation schemes accordingly follows closely those of previous studies, though some modifications are necessary. In Part II (to be presented in a separate paper), the effect of perturbing the vortex structure will be the focus.

The flow of this paper (Part I) is as follows. The configuration and settings of the model are described in section 2. The performance of the model in predicting the cases chosen is also evaluated. In section 3, the procedures of implementing the three perturbation schemes with respect to our model are given. A major component in the procedure is the filtering algorithm used to separate the vortex and the environmental flow before bogussing is performed. This filtering algorithm will be briefly described. In section 4, verifications of the ensemble predictions are presented. First the statistical quantities used to evaluate an ensemble are defined. Then the skills of the MCF, LAF, and BGM ensembles are examined. Verifications are done in two different contexts. One is under the so-called perfect model assumption (PMA) in which one of the ensemble members is considered to be free of any error. The ensemble mean is then verified by this ''perfect'' forecast. The second part of verification is to use the actual best tracks of the TCs. Then the characteristics of some typical cases in which the ensemble forecast can outperform the control is highlighted. This can serve as a reference to forecasters when guidance to the relative importance of the ensemble products is needed. Finally, a summary and discussion of the results in this part of the study are given in section 5.

2. Model description and performance

a. Configuration of the model

A semi-Lagrangian barotropic model (Holland et al. 1991) that was tested semioperationally in the 1990 season is used in this study. The barotropic model is used because of its computational efficiency and the relative ease in interpreting the dynamics (the nondivergent vorticity equation). The final analyses from the Tropical Cyclone Motion (TCM-90) Experiment (Rogers et al. 1993) are used as the initial input to the model. The model is integrated on a 50-km resolution Mercator grid covering approximately the domain 7°S–48°N, 100°–162°E. Wind fields from 850 to 300 hPa are vertically integrated (pressure weighted) to form a deep-layer mean.

A bogussing procedure is applied in this study. The original vortex in the analysis is first filtered using the vortex specification technique of Kurihara et al. (1993) and Kurihara et al. (1995) except that the determination of the filtering domain and the actual filtering are made from the deep-layer-mean wind. The axisymmetric vortex of Chan and Williams (1987) is then bogussed at the best-track position. This removes any forecast error due to an incorrect position. The profile of the tangential wind, v_t, is

$$v_t = v_{\max}\left(\frac{r}{r_{\max}}\right)\exp\left\{\frac{1}{b}\left[1 - \left(\frac{r}{r_{\max}}\right)^b\right]\right\}, \quad (2.1)$$

where $v_{\max}$ is the maximum wind speed, $r_{\max}$ the radius at which $v_{\max}$ occurs, and b a size factor ($0 < b < 1$) determining the boundary of the vortex. In the bogussing, these parameters are determined from known information of the cyclone. For instance, $v_{\max}$ is 0.8 times the warning intensity of the cyclone to account for the lower value of the vertically integrated mean than the observed surface wind (DeMaria et al. 1992), and $r_{\max}$ is set at 100 km. To find b using (2.1) requires one more point on the profile. This is taken to be the radius at which the azimuthally averaged tangential wind is 4 m s^{-1} (which is calculated in the Kurihara filtering algorithm). The properties of the bogus vortex remain unchanged in all the ensemble members.

The design of the model supports boundary data nesting, that is, some additional data can be blended into the boundary of the model domain during the integration of the model, with a preset width of the blending zone. This feature enables boundary condition control using a global model forecast in an operational setting. In our case, the boundary wind information from the analyses themselves is added to the domain. For the times between two consecutive analyses, the winds are obtained from a time interpolation.

b. Performance for tropical cyclones in 1990

Sixty-six forecast cases from nine TCs are chosen for study. Although the period covered by the dataset is only about two months, the nine TCs span a variety of track types including straight westward-moving, meandering, recurving, and northward-drifting cyclones. These different track characteristics reflect the different synoptic features affecting the motion of the TCs. Detailed descriptions of these features, together with their life histories, can be found in Elsberry et al. (1990) and the *Annual Tropical Cyclone Report* of the Joint Typhoon Warning Center (1990).

The cases were chosen such that at the time of starting a forecast, the TC had reached tropical storm intensity (i.e., a maximum sustained wind of 34 kt). Another factor to consider is that the track should not go too near (usually 10° latitude) the northern and/or western boundary of the model domain. Among the 66 cases, 41 allow forecasts up to 72 h, 15 stop at 60 h, and the remaining 10 at 48 h. A complete listing of the forecast cases is given in the appendix.

To provide an idea of the general performance of the model with respect to these chosen cases, the forecast tracks (termed the control forecasts) under the above configuration of the model are shown in Fig. 1. Some deficiencies of the model are already obvious. For example, a northern bias exists in many of the cases, especially for Tropical Cyclones Yancy, Becky, and Ed. Some imperfections of the configuration of the model can be pointed out, though they are not necessarily the reasons for these biases. The use of a purely axisymmetric vortex cannot account for the asymmetric structure (β gyres in barotropic dynamics) responsible for an additional northwestward propagation of the TC. Without this northwestward component, the TC will tend to move more northward, and thus to a certain extent this explains the northward bias observed. The scale-down factor of 0.8 of the intensity of the bogus vortex is probably not small enough because the observed intensity is the total wind, while an environmental flow resulting from the vortex filtering scheme already exists before the bogussing. This may make the symmetric vortex and the subsequently generated asymmetric β gyres too intense (and the winds in the outer core too strong, according to our profile). Further, no blending scheme to combine the environmental flow and the bogus vortex is applied before the integration starts and a period of adjustment is needed. This again may be a source of error of the subsequent motion.

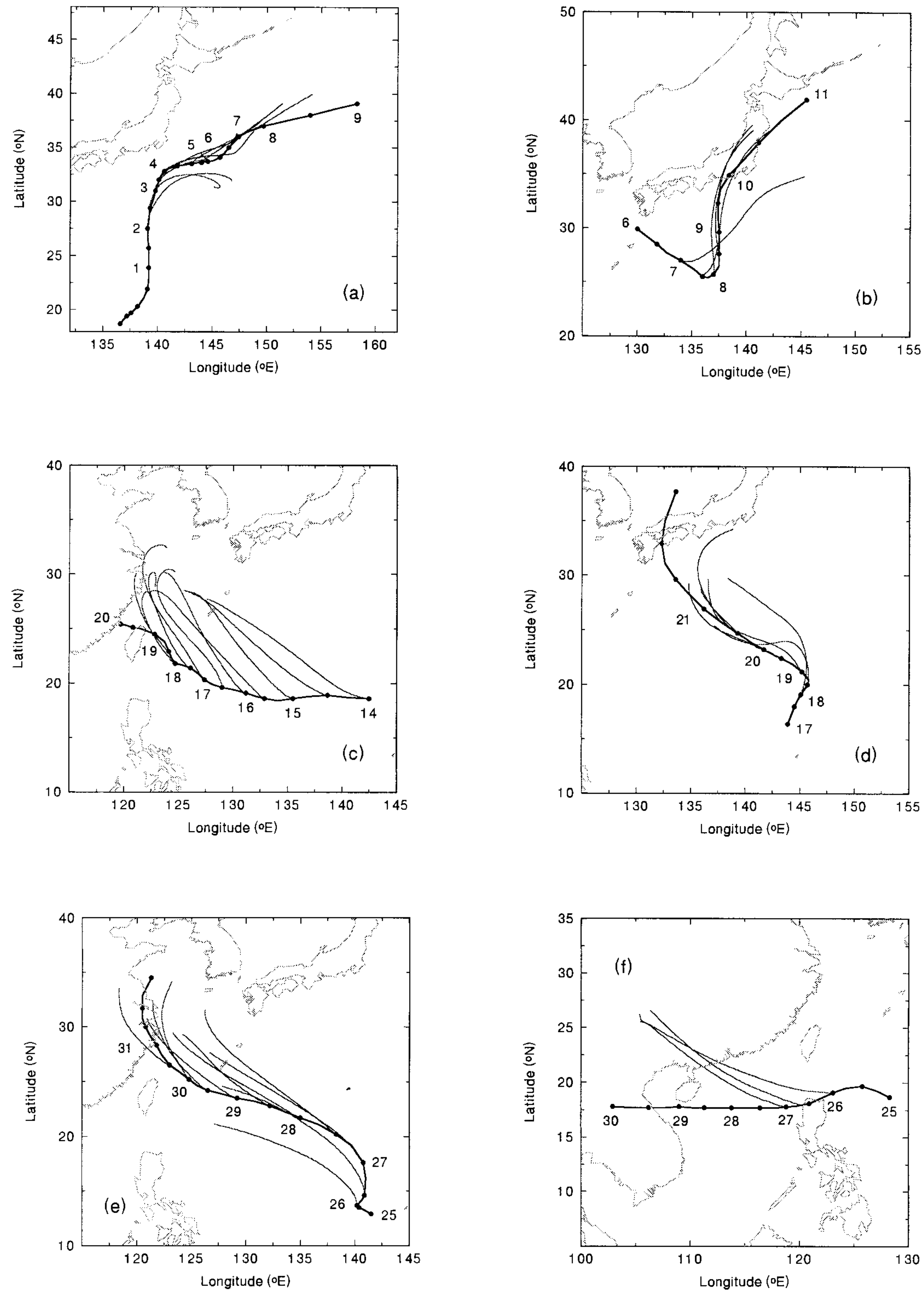

FIG. 1. Control forecasts for TC (a) Vernon, (b) Winona, (c) Yancy, (d) Zola, (e) Abe, (f) Becky, (g) Dot, (h) Ed, and (i) Flo. The numbers indicated are the dates in (a)–(e) August 1990 and (f)–(i) in September 1990.

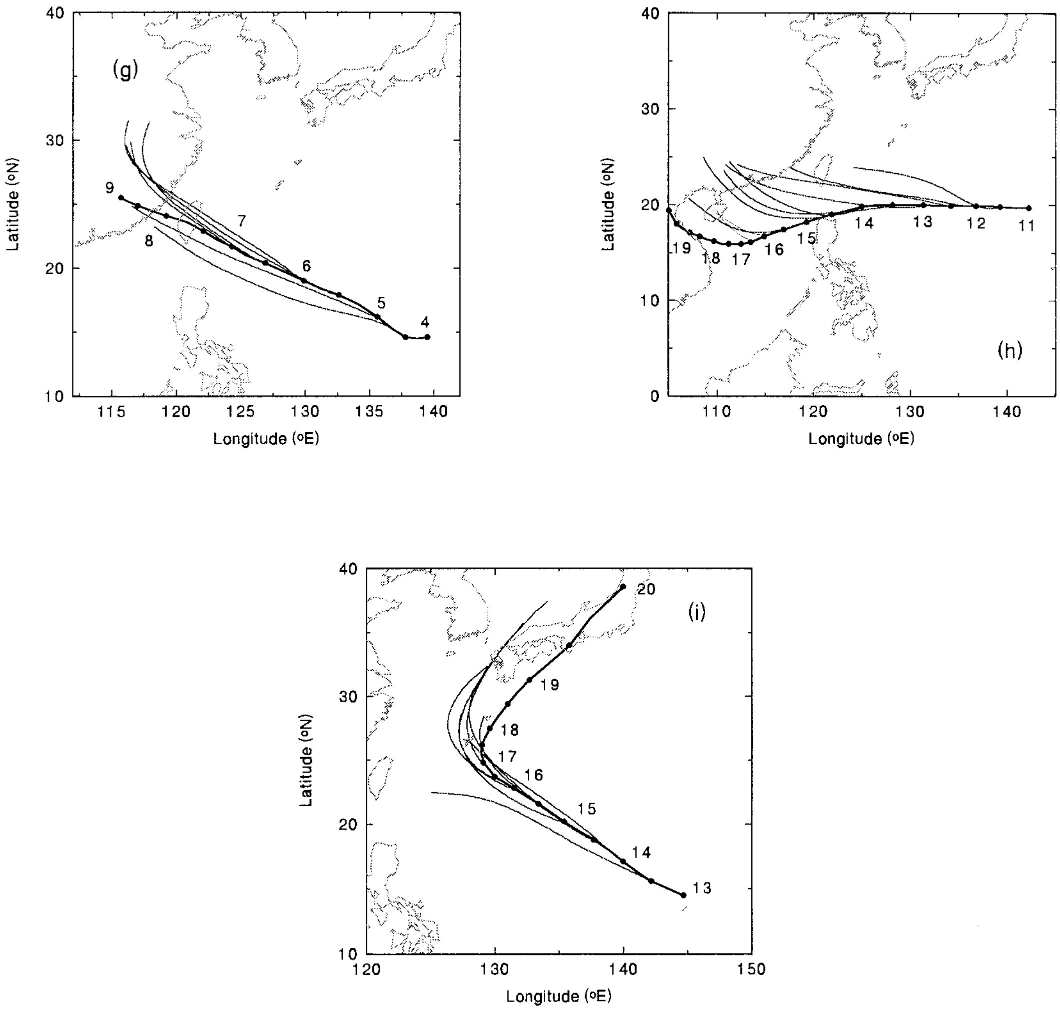

FIG. 1. (*Continued*)

To measure quantitatively the performance of the model, the control forecasts are compared with the climatology and persistence (CLIPER) forecasts. A relative skill score (RSS) similar to that in DeMaria et al. (1987) is calculated:

$$\mathrm{RSS}_{\mathrm{CONTROL}} = \frac{E_{\mathrm{CLIPER}} - E_{\mathrm{CONTROL}}}{E_{\mathrm{CLIPER}} + E_{\mathrm{CONTROL}}} \times 100\%, \qquad (2.2)$$

where E_{CLIPER} the forecast error of CLIPER and E_{CONTROL} the corresponding value of the control, both verified by the best track. A positive $\mathrm{RSS}_{\mathrm{CONTROL}}$ therefore indicates outperformance of the control over CLIPER. The values of $\mathrm{RSS}_{\mathrm{CONTROL}}$ for the 24-, 48-, and 72-h forecasts (Figs. 2a–c, respectively) suggest that the overall performance improves over CLIPER as the forecast time increases. The mean score averaged over all available cases is -12.6% at 24 h, -8.8% at 48 h, and $+2.1\%$ at 72 h. The variability of the performance over the cases is also large. In general, the forecasts are better for Vernon, Winona, Zola, and Abe, for which some positive scores can be obtained up to a lead time of 72 h. The skill for Becky, Dot, and Ed is moderate, with the score varying between $\pm 40\%$. The worst cases come from Yancy and Flo. In view of the possible deficiencies of the model discussed, this performance of the model is understandable. The main purpose of this study is not to optimize the model itself but to determine the extent to which the ensemble methodology can improve a forecast if no complicated measures to improve the model and design of the bogus vortex is adopted. Therefore, no further modification is made to the configuration of the model.

3. Perturbation methodologies

The perturbation methods used in this study are well documented in the literature. However, the application of these methods to the problem of TC track prediction needs some modifications, mainly in the environment-vortex separation and rebogussing procedures. The details of these procedures for each of the three schemes

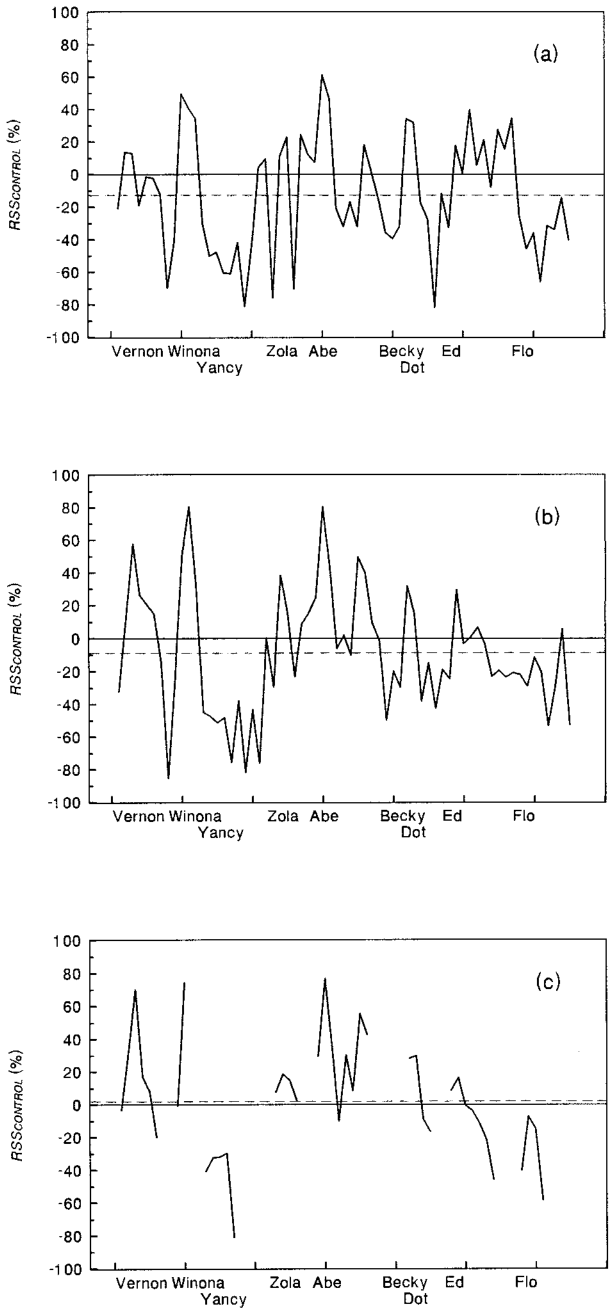

FIG. 2. Relative skill score of the control forecasts compared with CLIPER for a lead time of (a) 24 h, (b) 48 h, and (c) 72 h. See (2.2) for the definition of the score. The dashed line indicates the mean score of all the cases. The cases are arranged in chronological order and the first case for each TC is labeled by its name.

are therefore described below. The results are then presented in section 4.

a. Monte Carlo forecast (MCF)

In the MCF, random perturbations are added to the environmental zonal and meridional wind components. The root-mean-square (rms) magnitude of the perturbations is 3 m s^{-1}, which is about the square root of the error covariance of wind field at a midtropospheric level of a typical NWP model (Krishnamurti and Bounoua 1995). The rms difference over the model domain among analyses from various sources such as the ECMWF, the U.K. Meteorological Office (UKMO), and the U.S. Navy Operational Analysis and Prediction System (NOGAPS) analyses is also estimated to have an upper bound of about 3 m s^{-1}. Ninety-nine members with added perturbation are generated so that the total number of members of an MCF ensemble is 100 (including the control). While this number is chosen quite arbitrarily, the main consideration is the total integration time required. Tests have been performed by randomly selecting subensembles with fewer numbers of members from the original ensemble and examining their corresponding properties. It is found that the ensemble variance (i.e., the spread of the ensemble) remains about the same when the number of members is $\geq$50.

Only one case from each TC is used in the study of MCF, as shown in Table A1. The cases chosen usually have large errors in the control forecast, so that any possible improvement by the MCF can be easily identified. For those TCs with similar performance in the control forecasts, one case is chosen randomly. Experiments performed on these 9 cases, however, indicate that in general the ensemble members spread symmetrically around the control. This leads to the conclusion that the ensemble mean is always close to the control forecast. In other words, the ensemble forecast does not provide any improvement. The average distances (over the nine cases) between the ensemble mean positions and the corresponding control forecasts show that the ensemble mean seldom deviates from the control center by more than 100 km, even if the variation among the cases is considered (not shown). This result demonstrates that the MCF ensemble technique has little forecast utility in this context, and thus no further analysis of the results will be presented.

b. Lagged-average forecast (LAF)

In LAF, the ensemble members consist of those forecasts starting from earlier times [labeled T(−06), T(−12), T(−18), and T(−24) in Fig. 3a]. These are termed simple-lag (SL) forecasts (L1–L4, respectively). For those cases in which analyses at 0600 or 1800 UTC are not available, a temporal interpolation between the analysis at 0000 and 1200 UTC is made. As in the control forecast, vortex filtering and bogussing are performed on the analysis from which the lag forecast is made. This simulates an operational bogussing cycle, in that when the most updated information about the cyclone is obtained, it is used in the next bogussing step. This is, in fact, one advantage of the LAF perturbation scheme, as little change to the existing operational procedure is needed. To form the perturbed en-

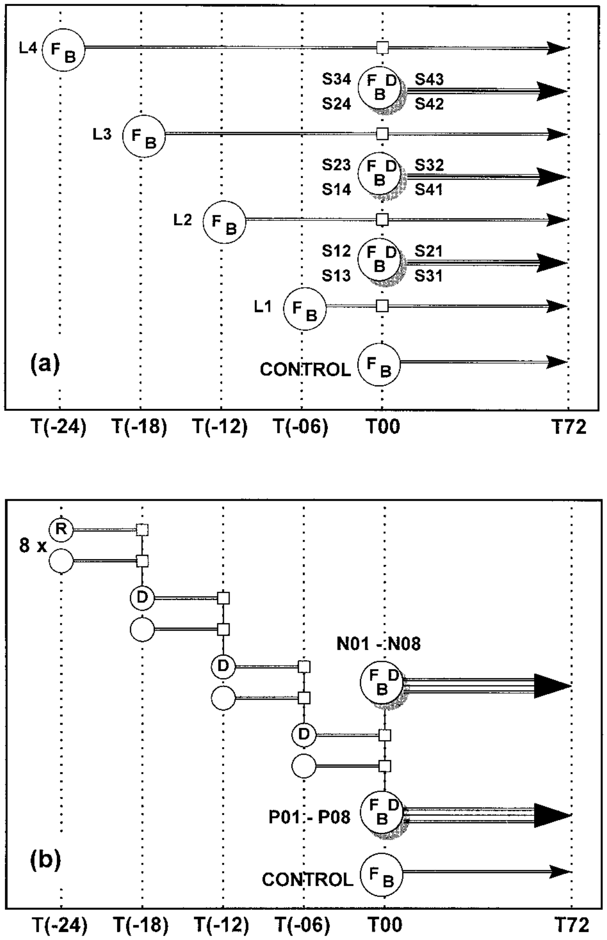

FIG. 3. Schematic diagram of the perturbation methodology for (a) LAF and (b) BGM. See text for the explanation of the symbols and labels.

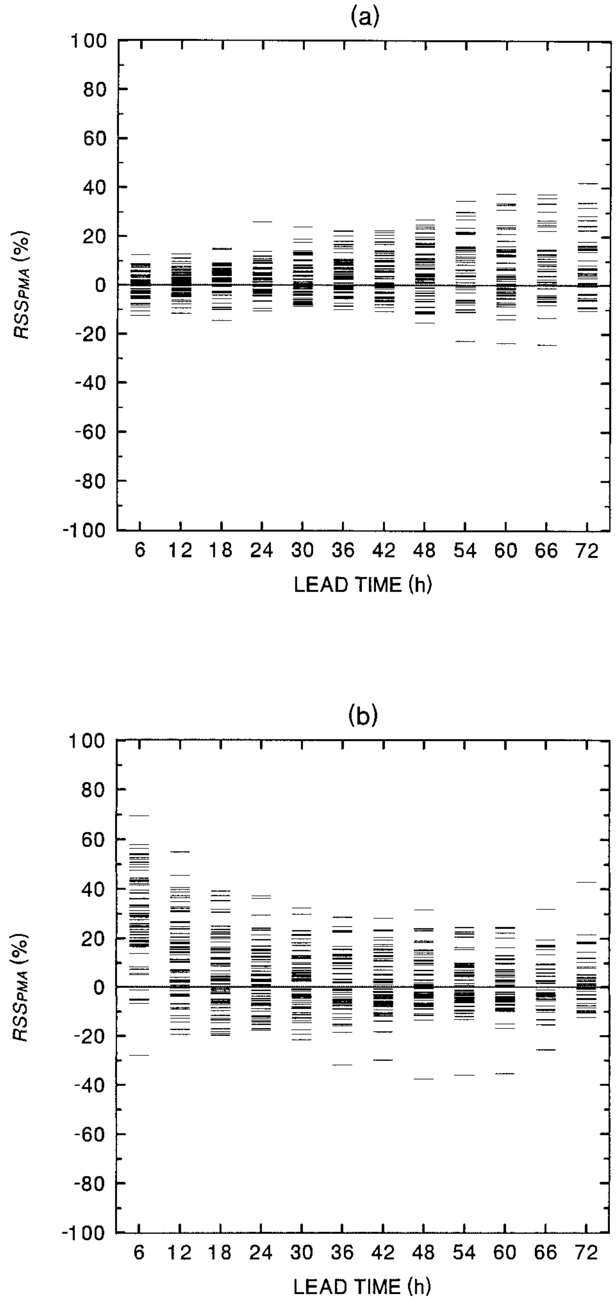

FIG. 4. Relative skill score under the PMA [defined in (4.6)] of all the cases and lead times for (a) LAF and (b) BGM ensembles. Each horizontal bar represents one case.

vironment, the forecast vortex verified at T00 (denoted by a small square) is first filtered. Its difference with the filtered analysis at T00 is then treated as the perturbation. The initial condition is therefore the sum of this perturbation, the filtered analysis, and an appropriate bogus vortex.

Additional perturbations are also generated by finding the so-called short-range forecast difference (SRFD), formed from the difference (denoted by a D) of the predicted environments (after the forecast vortices verified at T00 are filtered) of any two of the SL forecasts. This gives a total of six pairs of members with reverse sign in each (S12 and S21, S13 and S31, and so on). A bogus vortex the same as that in the SL members (denoted by B), together with the difference field, is then added at the best-track position to a filtered (denoted by F) analysis. Thus, every LAF ensemble consists of 17 members, including the control. The rms magnitude of all perturbations to a wind component is again controlled at 3 m s^{-1}, as in the MCF.

c. Breeding of growing modes (BGM)

The BGM perturbation scheme was developed at NCEP/EMC and has been its operational ensemble forecasting system since 1992. The background philosophy and original design can be found in Toth and Kalnay (1993), while Kalnay and Toth (1996) and Toth et al. (1997) provide more updated information and an evaluation of the products. The design of the BGM scheme

TABLE 1. Medium scores for the LAF and BGM under the perfect model assumption and best-track verification, respectively, at various lead times. The numbers in parentheses are percentages of the total number of cases with positive score.

Lead time (h)	Perfect model assumption		Best track verification	
	LAF	BGM	LAF	BGM
06	0.7 (56)	28.0 (91)	0.3 (50)	−38.0 (20)
12	1.8 (65)	11.5 (76)	−0.9 (41)	−10.0 (36)
18	2.2 (65)	3.3 (61)	−1.9 (38)	−2.2 (44)
24	2.6 (65)	2.4 (56)	−1.8 (44)	0.6 (52)
30	1.5 (64)	−0.5 (45)	−2.1 (41)	0.9 (50)
36	3.9 (64)	−0.9 (42)	−2.6 (39)	0.4 (52)
42	5.4 (67)	−0.9 (44)	−3.4 (38)	−1.2 (48)
48	4.7 (65)	−0.7 (44)	−2.5 (39)	−2.5 (44)
54	4.7 (66)	−1.5 (46)	−1.8 (43)	−4.7 (46)
60	5.4 (64)	−0.8 (50)	−2.0 (43)	−4.4 (38)
66	6.8 (73)	1.7 (51)	−3.5 (37)	−3.6 (39)
72	6.7 (71)	1.0 (54)	−6.7 (29)	−8.3 (32)

is such that it can well represent the errors in the NCEP analyses and the possible fastest-growing modes in the analyses are well sampled (Iyengar et al. 1996). The latter condition is essential for the ensemble prediction system to provide an alternate scenario of midlatitude disturbances. The application of the BGM to a pure barotropic flow, however, has to be justified a posteriori.

The method of generating the bred perturbations is similar to that of the EMC except that a series of breeding cycles has to be set up, instead of modifying the regular assimilation cycles. The procedure is shown in Fig. 3b. The breeding cycles start 24 h before the initial forecast time, with a rescaling of the bred error every 6 h. In the beginning, two forecasts are made, one from the analysis and one with random perturbations (R) added. As in generating the SRFD members in LAF, the forecast vortices (after 6 h) are filtered, and the difference (denoted by D) of the filtered predictions is used as a perturbation to the next breeding cycle. Depending on the sign of this perturbation (the prediction from the perturbed field minus that from the analysis, or vice versa), a pair of bred perturbations (N01 and P01) can be generated. Four breeding cycles like this are performed until the initial forecast time (T00) is reached. At each cycle, the bred error is scaled to the same rms magnitude of 3 m s^{-1}, and an appropriate bogus vortex at that time is also added. For all the cases considered, the change in this magnitude usually has little variation in the second or third cycle, so that by the time of reaching the initial forecast time, the error growth has already been saturated in a root-mean-square sense. This justifies the choice of a 24-h breeding period.

In order to compare with the results from the LAF perturbation method, the same number of ensemble members should be used. Thus, in total eight pairs (denoted by 8× in the figure) of these bred perturbations (N01–N08 and P01–P08) are generated. The number of BGM members is, therefore, again 17 including the control forecast. After these bred perturbations are formed, they (again denoted by D) are then added to the analyses with the original vortices filtered (F) and the bogus vortices included (B).

4. Verification of the ensembles

a. Ensemble statistics

Various statistical measures can be computed for an ensemble to explore its characteristics. The definitions are similar to those in Branković et al. (1990). If we denote F_i as a parameter of the ith ensemble member, we can calculate the ensemble centroid or ensemble mean:

$$\overline{F} = \frac{1}{N}\sum_{i=1}^{N} F_i, \tag{4.1}$$

where N is the total number of members. This ensemble mean has an error E when verified by the corresponding parameter A in the analysis:

$$E = |\overline{F} - A|. \tag{4.2}$$

In our case of TC motion forecast, the quantity of primary interest is the vortex center. Hence, the quantities F_i and A are position vectors and the norm (denoted by $|\,|$) in (4.2) is simply taken as the distance between two center positions.

In addition, we can calculate the mean-square error e^2 and error variance Δ^2 of the ensemble, defined as

$$e^2 = \frac{1}{N}\sum_{i=1}^{N} |F_i - A|^2 \tag{4.3}$$

$$\Delta^2 = \frac{1}{N}\sum_{i=1}^{N} |F_i - \overline{F}|^2. \tag{4.4}$$

Another quantity to measure the spread of the ensemble is the mean-square distance δ^2 among the members:

$$\delta^2 = \frac{1}{N(N-1)}\sum_{i,j=1}^{N} |F_i - F_j|^2. \tag{4.5}$$

Note that both Δ and δ are independent of the verifi-

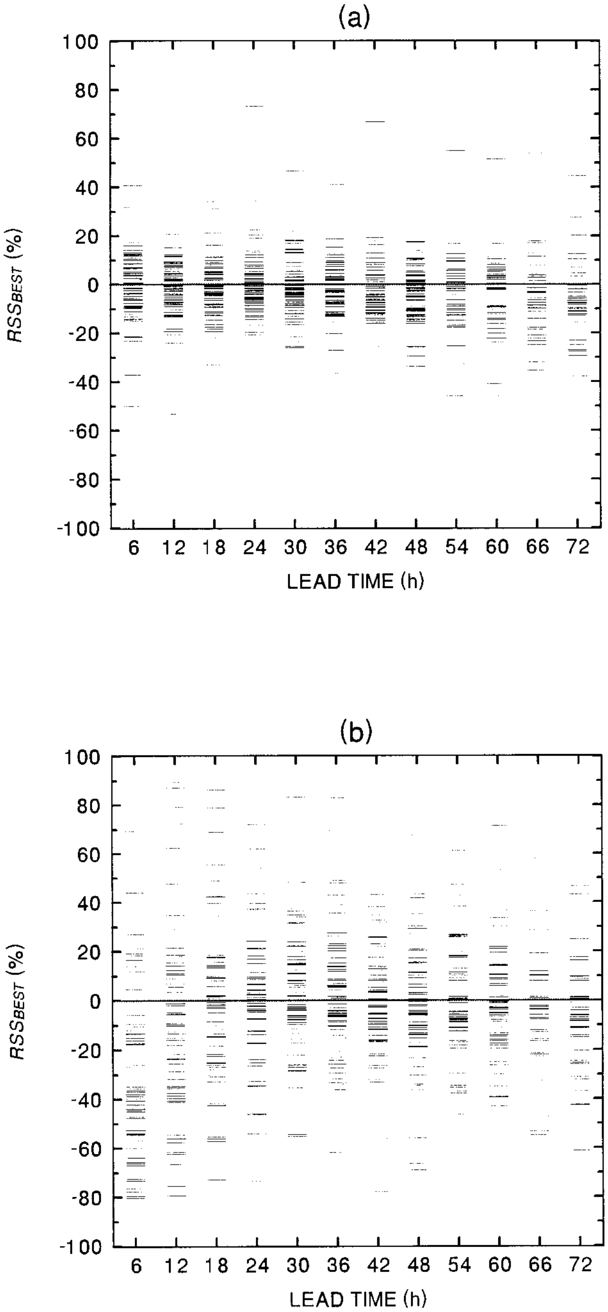

FIG. 5. As in Fig. 4 except that verification is by the best track for (a) LAF and (b) BGM ensembles.

cation. In a perfect model, δ can be a good estimation of the rms error e. This will be justified in the section discussing spread-skill correlation.

b. *Verification under the perfect model assumption*

The verification statistics of all the 66 cases studied using the LAF and BGM methodologies are presented in this and the next sections. The limitation of the present barotropic model has been shown in the verification of the control forecasts in section 2b. However, the relationship between analysis error and model deficiencies is complicated (Tribbia and Baumhefner 1988), and the extent to which ensemble forecasting can deal with model errors is generally unknown. In this section, the ensembles are verified under the PMA in which one of the ensemble members is chosen as the perfect forecast for verifying the control and the mean forecast of the remaining members. With this verification, the score obtained is then a pure measure of the performance of the ensemble methodology free of any model error. The disadvantage is that the ensemble can also include the verification, which is not the case in reality. Obviously, the verification results will depend on which member is chosen. In order to reduce this sensitivity, an average score over all possible realizations (i.e., each member other than the control is chosen as the verifying member once) is taken.

The relative skill score for this perfect model assumption ($\mathrm{RSS_{PMA}}$) is calculated as

$$\mathrm{RSS_{PMA}} = \frac{E_{\mathrm{CONTROL}} - E_{\mathrm{ENSEMBLE}}}{E_{\mathrm{CONTROL}} + E_{\mathrm{ENSEMBLE}}} \times 100\%, \qquad (4.6)$$

where E_{CONTROL} is the error of the control run, E_{ENSEMBLE} the average error of the ensemble mean, both verified by the ‘‘chosen reality.’’ The calculation of all the other quantities described in section 4a is the same except that the verification is changed. Again, a positive value of $\mathrm{RSS_{PMA}}$ indicates an outperformance of the ensemble mean over the control. The scores of all the cases in LAF are shown in Fig. 4a, with lead time varying from 6 to 72 h. The median score for each lead time, together with the percentage of the total number of cases with positive score, is given in Table 1 (first two columns). It can be seen that the median scores are positive for all times and within 10%. It is lower in the first 36 h but improves in later lead times. The percentage of the cases with a positive score is over 50% throughout the whole forecast period. A detailed examination of the temporal evolution of the score of each case shows that when a forecast starts with a positive score, the improvement generally continues to a longer forecast period. On the other hand, those without any superior performance in the beginning usually have little chance to be correct at a later stage.

The BGM scores (Fig. 4b) generally have a larger variability among the cases than those for the LAF, especially in the first 48 h. The median score for the BGM, on the other hand, starts at a high value but has a decreasing trend. Negative values also appear at some of the forecast times (Table 1). The percentages of cases with positive scores are about the same as in the LAF for the first 24 h, but are slightly lower than those in the latter afterward. Similar to the situation in the LAF, superiority of an ensemble usually starts from the beginning of a forecast.

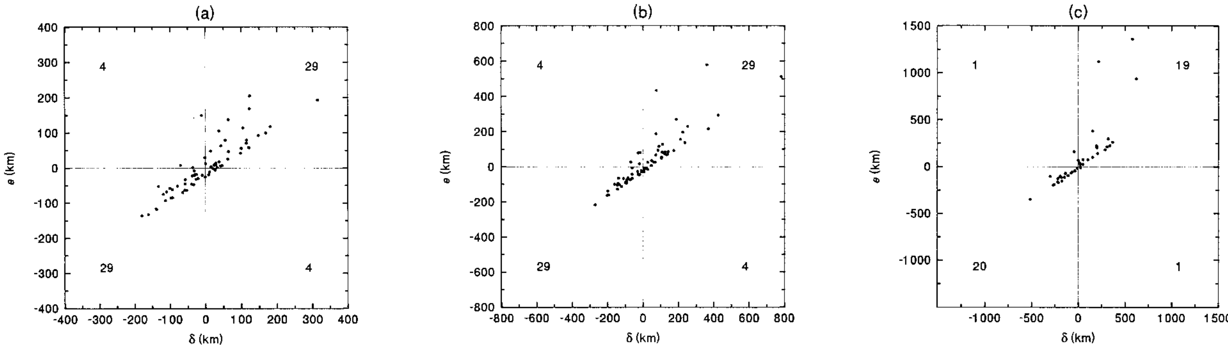

FIG. 6. Scatterplot for the e–δ pair from LAF under PMA at (a) 24 h, (b) 48 h, and (c) 72 h. The values plotted are deviations from the median. Contingency tables are constructed and the frequencies are shown in each quadrant.

To conclude, the results of verification under the PMA in this section indicate that if a skillful model is used, these two ensembles both have some potential to improve the forecast. It is not easy to explain the different performance of the two, and actually their difference in skill as measured by the RSS is not very significant. The decreasing median score in the BMG methodology may indicate problems of applying such a breeding scheme in a pure barotropic model. Nevertheless, its variability of skill for different situations is greater than that of the LAF. This is still true when the ensembles are verified by the actual reality.

c. Verification using best tracks

When the verification of the ensembles is made with respect to the best track, the score (termed $\mathrm{RSS}_{\mathrm{BEST}}$) to be calculated is the same as (4.6) except verified by the best track. The scores in both methodologies degrade substantially (Fig. 5a for LAF and Fig. 5b for BGM), with almost all the median scores below zero (Table 1, last two columns). The difference between the two schemes is not very significant (except at 6 h when the scores for the BGM are very low), although the BGM possesses a higher percentage of cases with a positive score at most lead times. While this degradation in score describes the effect of the model bias when the control forecasts are evaluated, a comparison with the results under the PMA needs careful interpretation. Tribbia and Baumhefner (1988) showed that if both the variability in forecast skill due to analysis error (characterized by a certain statistical distribution) and that due to model error (characterized by another distribution) are taken into account, the total forecast error is a function of both statistical distributions. The statistical quantities derived from an ensemble are then sampled estimates of the averages over both distributions. Thus, the decrease in scores using the best tracks as verification are not purely due to the model deficiencies, but a combined effect of such deficiencies and errors in the analyses. While this study concentrates on the skill of the ensemble mean, the characteristics of the model errors can actually be inferred by examining the error distribution within the ensemble as was done by Hamill and Colucci (1997). Based on such knowledge, the forecast may be corrected in a way similar to model output statistics. Such work will be reported later.

The verification results in this section show that the ensemble technique does not work under all circumstances, but its application should not be discouraged since its forecasts can provide additional information. Some characteristic synoptic situations associated with successful cases will be highlighted. Such an examination should be of value to forecasters when deter-

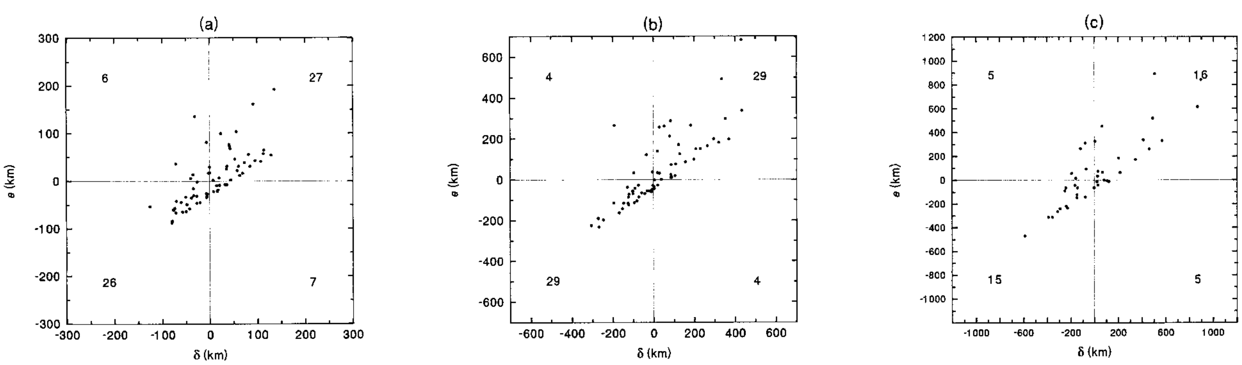

FIG. 7. As in Fig. 6 except for BGM.

TABLE 2. Values of the linear correlation coefficient ρ between δ and e for the LAF and BGM under the perfect model assumption and best-track verification, respectively, at 24, 48, and 72 h.

Lead time (h)	Perfect model assumption		Best track verification	
	LAF	BGM	LAF	BGM
24	0.88	0.73	0.60	0.42
48	0.88	0.82	0.53	0.51
72	0.82	0.86	0.37	0.68

mining the relative importance of the ensemble products under different synoptic conditions.

d. Spread-skill correlation

One utility of running an ensemble of forecasts with different initial conditions instead of a single one is to have some idea of the degree of predictability of the forecast. An underlying assumption (though without a solid theoretical foundation) of ensemble forecasting is that when the spread of the ensemble members is small, the situation is more predictable. In other words, there is a higher confidence that the ensemble mean forecast is skillful. On the other hand, if the spread among the ensemble members is large, the predictability is believed to be low and the forecast is likely to deviate considerably from the analysis. This idea forms the basis of forecasting the forecast skill using ensembles (Kalnay and Dalcher 1987). The applicability of this relationship between spread and skill is studied in this section.

As mentioned in section 4a, the spread among the ensemble members, δ, should be a good estimate of the rms error e in a perfect model. This is first examined using the statistics under the PMA. However, the averaging over all realizations in section 4b will make δ and e almost identical. Therefore, the values of e from a randomly chosen, single realization are used here. The scatterplots of δ and e (deviations from the corresponding median value) for all cases at three lead times, 24, 48, and 72 h, for the LAF (Figs. 6a, 6b, and 6c, respectively) demonstrate that correlation exists between these two quantities because most of the cases lie on the two diagonal boxes (the small-δ, small-e box and the large-δ, large-e one). Similar results are obtained for the BGM (Fig. 7). The values of the linear correlation coefficient, ρ, calculated for these times are given in Table 2. The correlation coefficients are all >0.8 except for the BGM at 24 h.

Under the best-track verification, more cases fall in the ''off-diagonal'' classes (Fig. 8 for LAF and Fig. 9 for BGM). The δ–e correlation coefficient at the three lead times decreases to a value from about 0.4 to 0.7 (Table 2). Contingency tables formed from the number of cases in each quadrant (shown in Figs. 6–9) suggest that differences between the best-track contingency tables (Figs. 8 and 9) and their PMA counterparts (Figs. 6 and 7) are significant based on the χ^2 tests, except for the BGM at 72 h. In that case, the two tables are identical, which indicates that the BGM has a slightly higher ability to sample large values from the underlying error distribution when the errors come from both the analyses and the model. However, the other test results suggest that in general the model errors are not well removed in terms of the spread-skill relation.

Techniques of extracting additional information from an ensemble exist. For example, probabilistic regions based on the ensemble distribution can be formed; clusters of members can also be identified so that the most probable cluster can be used as substitution to the original deterministic forecast. However, these will not be explored in this paper.

e. Characteristics of the successful cases

The successful cases are defined as those in which the skill score (relative to the best track) is positive for more than half of the verification times. For the LAF, 24 such cases can be identified (Table 3), while for the BGM the number is 28 (Table 4). The distributions show that the cases include several TCs. For the LAF, the cases span quite uniformly among all TCs except none can be found for Becky, while for the BGM they are mainly from Vernon, Abe, and Flo. For each TC, the improvement in forecast occurs during some particular period. For example, the entire period during which Ver-

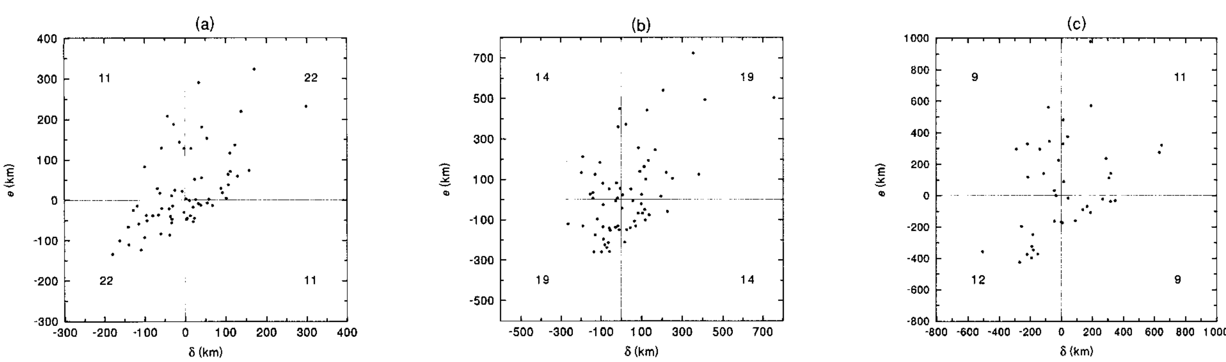

FIG. 8. As in Fig. 6 except verification is by the best track.

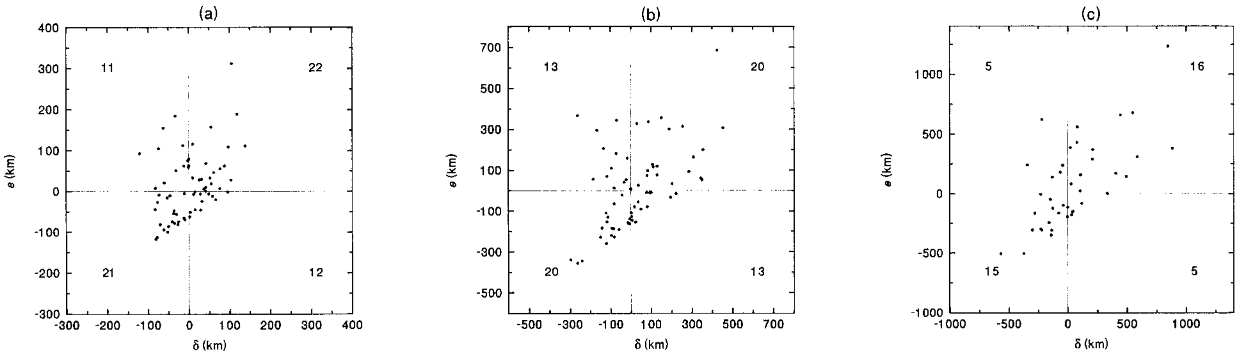

FIG. 9. As in Fig. 6 except for BGM and verification is by the best track.

non changed from moving northward to eastward appears in the list for the BGM (see Fig. 1 for the positions of the TC); the forecast of the recurvature of Flo in LAF and that of Abe in BGM are also quite successful. These statistics reveal the fact that some environmental situation or some transition from a particular situation to another can be favorable for the application of ensemble forecasting. However, some individual cases are also important, though the percentage of successful cases for that TC may not be particularly high. One example is the northwestward turning of Zola at 0000 UTC 18 August, which is well captured by both the LAF and BGM ensemble members. It is believed that these cases also possess some common features in the environmental flow when compared with the above.

When the synoptic flows associated with these successful cases are examined, it is found that they fall into one of the following characteristic situations.

1) A TC making a transition from a synoptic region (Carr and Elsberry 1994) to another.
2) Interaction with an apparent break in the subtropical ridge (STR), but no actual recurvature of the TC. Note that when no recurvature occurs, the TC remains in one synoptic region (that affected by the steering due to the STR).
3) A rapid strengthening or weakening of the STR. The latter situation should be responsible for an originally straight-moving TC to recurve (see the next point).
4) Potential recurvature of a TC. This is when the break in the STR is large enough for the TC to pass, either because of the passing of a midlatitude trough or because the northward steering of the TC is strong enough.
5) Multiple-TC case. One point worth mentioning is that if the bogussing procedure is applied to all existing TCs, the variability of the effect of one on the other may not be simulated in these environmental perturbation schemes. In this case, this multiple-TC effect has to be explicitly simulated using ensembles by means of perturbations to the vortex (see Part II of this study).

These characteristic synoptic situations all appear to have the common feature of a rapidly changing environment associated with the TC. Either the TC is moving to a different region or some major components in the synoptic pattern are changing. For the remaining cases (both the unsuccessful cases and those not considered above), relatively stable environmental conditions are found. For example, the early stages of Yancy, Abe, Dot, and Ed are all under the influence of a stable STR. Usually the model can already predict the motion well in this situation and, thus, the contribution of the ensemble methodology would be little. (In the successful cases of Dot for the BGM, the improvement is actually small.) For Winona, the cases chosen in this study are when it was drifting northward under a weak southerly steering. The synoptic pattern associated should also be a northward-oriented one (Carr and Elsberry 1994), but Winona did not make a transition from one synoptic region to another as Vernon did. Thus, although the forecasts of some of the Winona cases can be improved by the BGM ensembles, the improvements are not very

TABLE 3. Successful cases in LAF. The number in the last row is the percentage of successful cases for that particular TC. (See Table A1 for explanation of forecast case designator.)

Vernon	Winona	Yancy	Zola	Abe	Dot	Ed	Flo
V80212	W80712	Y81612	Z81800	A82700	D90412	E91212	F91312
V80312	W80800	Y81700	Z81812	A82712	D90500	E91300	F91400
V80400	W80812	Y81812			D90512		F91500
					D90600		F91512
							F91612
33%	75%	30%	33%	20%	67%	20%	63%

TABLE 4. As in Table 3 except for the BGM.

Vernon	Winona	Yancy	Zola	Abe	Dot	Ed	Flo
V80200	W80700	Y81500	Z81800	A82600	D90412	E91312	F91312
V80312		Y81700	Z81900	A82700	D90600	E91500	F91400
V80400		Y81800	Z82000	A82712			F91512
V80412				A82812			F91700
V80500				A82912			
V80512				A83000			
V80600							
78%	25%	30%	50%	60%	33%	20%	50%

significant. Note also that in the cases chosen for this study, there is no explicit transition from one synoptic pattern to another. The performance of ensemble forecasting in this circumstance therefore cannot be answered by these experiments.

5. Summary and discussion

a. Summary

The utility of ensemble forecasting of tropical cyclone motion with emphasis placed on the perturbations of the environmental flow is studied in this paper. A simple barotropic model is used as it is easier to interpret the dynamics. Three perturbation methodologies, the Monte Carlo forecast (MCF), the lagged-average forecast (LAF), and the breeding of growing modes (BGM) are tested and with their skills compared. Verifications are based on 66 cases chosen from the 1990 season in the western North Pacific. The main findings of this study are as follows.

1) The MCF ensemble mean positions are well within 100 km from the control centers in all the cases examined. Thus, the skill of the MCF is found to be comparable with that of the control.
2) The potential of the LAF and BGM methodologies is demonstrated under the perfect model assumption (PMA). The median score for the LAF as well as the number of ''good'' cases are slightly higher than those in the BGM, but the variability of performance for the latter is larger.
3) When verified by the best tracks, the average score of both methodologies decrease in all forecast periods. However, about 36% of the LAF and 42% of the BGM cases still outperform the corresponding control forecast.
4) A high degree of spread-skill correlation is demonstrated under the PMA for the quantity δ and the rms error e. The degree of correlation can be carried out to a forecast time of 72 h. In general, correlation is downgraded when verified by the best track. This should, however, be improved in a better model configuration.
5) The successful cases from the LAF and BGM techniques have one of these characteristic synoptic situations: 1) a TC making a transition from one synoptic region to another, 2) interaction with an apparent break in the STR, 3) a rapid strengthening/weakening in the STR, 4) potential recurvature of a TC, and 5) multiple-TC case. These patterns may therefore be labeled as being favorable for the ensemble forecasting approach.

b. Discussion

Some final remarks concerning the present study, and some perspective for the next part, are provided here. The ensemble size (17 members, including the control) chosen in the LAF ensemble is optimal for the following reason. Since TCs tend not to remain at the same intensity for too long and no model to date can make realistic predictions of intensity, extending the lagged period will increase the domination of the model deficiencies. Shortening the time lag between two successive SL members is of little use since the environment does not change significantly on such a short timescale. On the other hand, increasing the number of members in a BGM ensemble is straightforward, because that involves different projections of the starting random error on the fast growing modes.

Overall, the BGM scheme does not outperform the LAF much in the present study (except with a higher score in the PMA verification and a slightly larger number of successful cases). This is because the perturbations generated from both schemes give essentially the same type of error growth since the model is barotropic. However, when applied to a baroclinic model, the BGM scheme may be better suited for perturbing the environment because it can capture the fast-growing baroclinic modes better, as can be seen from the success of the application of the BGM scheme by NCEP with respect to some of the phenomena at higher latitudes.

The focus in this study is on the skill of the ensemble mean. However, previous research in short-range ensemble forecasting suggests that the mean may not be the most useful attribute of an ensemble. Rather, the possibility of an individual member within the ensemble to provide a correct forecast attracts more interest recently. (Refer back to the last comment in section 4e on the application of the spread-skill correlation.) A detailed examination of the tracks of our TC ensembles shows that if the past motion of the TC is taken into

TABLE A1. Forecast cases used in this study.

Vernon	Winona	Yancy	Zola	Abe	Becky	Dot	Ed	Flo
V80200[a]	W80700[a]	Y81400	Z81800[a]	A82600[a]	B82600[a]	D90412	E91112	F91312
V80212	W80712	Y81412	Z81812	A82612	B82612[b]	D90500	E91200	F91400
V80300	W80800[b]	Y81500[a]	Z81900	A82700	B82700[c]	D90512	E91212	F91412
V80312	W80812[c]	Y81512	Z81912	A82712		D90600[a]	E91300	F91500[a]
V80400		Y81600	Z82000[b]	A82800		D90612[b]	E91312	F91512[b]
V80412		Y81612[b]	Z82012[c]	A82812		D90700[c]	E91400[a]	F91600[b]
V80500[b]		Y81700[b]		A82900			E91412	F91612[c]
V80512[b]		Y81712[b]		A82912			E91500[b]	F91700[c]
V80600[b]		Y81800[c]		A83000[b]			E91512[b]	
		Y81812[c]		A83012[c]			E91600[c]	

[a] Used in the MCF.
[b] 60-h forecast.
[c] 48-h forecast.

consideration, some less skillful ensemble members can in fact be removed. Usually the past motion of a TC is taken into account in an operational model by adding in the area of the bogus vortex a persistence vector equal to the motion, so that the vortex will follow the steering of this vector immediately after the start of the model. This aspect of improving the ensemble will be included in the second part of this study.

Acknowledgments. The authors would like to thank Prof. L. Leslie of the University of New South Wales for providing the barotropic model and pointing out the error in the earlier manuscript. Dr. Pat Harr of the U.S. Naval Postgraduate School provided the final analyses dataset and Dr. Y. Kurihara of GFDL supplied the hurricane filtering code. Thanks also go to the following persons for providing the required references: Dr. J. Du of NCEP, Dr. J. L. Anderson and Dr. F. Vitart of GFDL, and Dr. M. S. J. Harrison of UKMO.

This research is sponsored by the U.S. Office of Naval Research Grant N00014-94-1-0824 and the Research Grants Council of the Universities Grants Committee of Hong Kong Grant 9040194.

APPENDIX

Listing of All Forecast Cases Considered

All the forecast cases are listed in Table A1 for reference. They are grouped with respect to individual TC and ordered chronologically. The total number is 66. Each forecast case is denoted as *Xmddtt,* where *X* is the first letter of the TC name, *m* the month, *dd* the day, and *tt* the UTC time. All cases are run up to 72 h except those specified. The selection criteria have been described in section 2b.

REFERENCES

Aberson, S. D., S. J. Lord, M. DeMaria, and M. S. Tracton, 1995: Short-range ensemble forecasting of hurricane tracks. Preprints, *21st Conf. on Hurricanes and Tropical Meteorology,* Miami, FL, Amer. Meteor. Soc., 494–496.

Branković, Č., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.,* **116,** 867–912.

Brooks, H. E., and C. A. Doswell III, 1993: New technology and numerical weather prediction—A wasted opportunity? *Weather,* **48,** 173–177.

——, D. J. Stensrud, and M. S. Tracton, 1996: Short-range ensemble forecasting pilot project: A status report. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J39–J40.

Carr, L. E., III, and R. L. Elsberry, 1994: Systematic and integrated approach to tropical cyclone track forecasting Part I. Approach overview and description of meteorological basis. NPS Rep. NPS-MR-94-002, 273 pp. [Available from Department of Meteorology, Naval Postgraduate School, Monterey, CA 93943.]

Chan, J. C. L., and R. T. Williams, 1987: Analytical and numerical studies of the beta-effect in tropical cyclone motion. Part I: Zero mean flow. *J. Atmos. Sci.,* **44,** 1257–1265.

Dalcher, A., E. Kalnay, and R. N. Hoffman, 1988: Medium-range lagged average forecasts. *Mon. Wea. Rev.,* **116,** 402–416.

DeMaria, M., 1987: Tropical cyclone track prediction with a barotropic spectral model. *Mon. Wea. Rev.,* **115,** 2346–2355.

——, 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.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.,* **125,** 2427–2459.

Elsberry, R. L., B. C. Diehl, J. C. L. Chan, P. A. Harr, G. J. Holland, M. Lander, T. Neta, and D. Thom, 1990: ONR Tropical Cyclone Motion Research Initiative: Field experiment summary. NPS Rep. NPS-MR-91-001, 107 pp. [Available from Department of Meteorology, Naval Postgraduate School, Monterey, CA 93943.]

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

Harrison, M. S. J., T. N. Palmer, D. S. Richardson, R. Buizza, and T. Petroliagis, 1995: Joint ensembles from the UKMO and ECMWF models. *Proc. ECMWF Seminar on Predictability,* Vol. 2, Reading, United Kingdom, ECMWF, 61–120.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35A,** 100–118.

Holland, G. J., L. M. Leslie, E. A. Ritchie, G. S. Dietachmayer, P. E. Powers, and M. Klink, 1991: An interactive analysis and forecast system for tropical cyclone motion. *Wea. Forecasting,* **6,** 415–424.

Iyengar, G., Z. Toth, E. Kalnay, and J. S. Woollen, 1996: Are the bred vectors representative of analysis errors? Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J64–J66.

Joint Typhoon Warning Center, 1990: Annual Tropical Cyclone Report. 278 pp. [Available from Joint Typhoon Warning Center, PSC 455, Box 12, FPO, AP 96540-0051.]

Kalnay, E., and A. Dalcher, 1987: Forecasting the forecast skill. *Mon. Wea. Rev.,* **115,** 349–356.

——, and Z. Toth, 1996: Ensemble prediction at NCEP. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J19–J20.

Krishnamurti, T. N., and L. Bounoua, 1995: *An Introduction to Numerical Weather Prediction Techniques.* CRC Press, 304 pp.

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.

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

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.

Morison, R. P., L. M. Leslie, and G. D. Hess, 1996: Development of tropical cyclone landfall strike probabilities using Monte Carlo simulations. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 414–415.

Palmer, T. N., R. Mureau, and F. Molteni, 1990: The Monte Carlo forecast. *Weather,* **45,** 198–207.

Ramamurthy, M. K., and J. Shu, 1996: Short-range ensemble forecast experiments with scale-dependent initial errors. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA Amer. Meteor. Soc., 60–61.

Richardson, D. S., M. S. J. Harrison, K. B. Robertson, and A. P. Woodcock, 1996: Joint medium-range ensembles using UKMO, ECMWF and NCEP ensemble systems. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J26–J28.

Rogers, E., S. L. Stephen, D. G. Deaven, and G. J. DiMego, 1993: Data assimilation and forecasting for the Tropical Cyclone Motion Experiment at the National Meteorological Center. Preprints, *20th Conf. on Hurricanes and Tropical Meteorology,* San Antonio, TX, Amer. Meteor. Soc., 329–330.

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

——, ——, S. M Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting,* **12,** 140–153.

Tribbia, J. J., and D. P. Baumhefner, 1988: The reliability of improvements in deterministic short-range forecasts in the presence of initial state and modeling deficiencies. *Mon. Wea. Rev.,* **116,** 2276–2288.

Vitart, F., J. L. Anderson, and W. F. Stern, 1997: Simulation of interannual variability of tropical storm frequency in an ensemble of GCM integrations. *J. Climate,* **10,** 745–760.