The Prediction of Extratropical Storm Tracks by the ECMWF and NCEP Ensemble Prediction Systems

LIZZIE S. R. FROUDE, LENNART BENGTSSON, AND KEVIN I. HODGES

Environmental Systems Science Centre, University of Reading, Reading, United Kingdom

(Manuscript received 30 May 2006, in final form 22 September 2006)

ABSTRACT

The prediction of extratropical cyclones by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) ensemble prediction systems (EPSs) has been investigated using an objective feature tracking methodology to identify and track the cyclones along the forecast trajectories. Overall the results show that the ECMWF EPS has a slightly higher level of skill than the NCEP EPS in the Northern Hemisphere (NH). However in the Southern Hemisphere (SH), NCEP has higher predictive skill than ECMWF for the intensity of the cyclones. The results from both EPSs indicate a higher level of predictive skill for the position of extratropical cyclones than their intensity and show that there is a larger spread in intensity than position. Further analysis shows that the predicted propagation speed of cyclones is generally too slow for the ECMWF EPS and shows a slight bias for the intensity of the cyclones to be overpredicted. This is also true for the NCEP EPS in the SH. For the NCEP EPS in the NH the intensity of the cyclones is underpredicted. There is small bias in both the EPS for the cyclones to be displaced toward the poles. For each ensemble forecast of each cyclone, the predictive skill of the ensemble member that best predicts the cyclone's position and intensity was computed. The results are very encouraging showing that the predictive skill of the best ensemble member is significantly higher than that of the control forecast in terms of both the position and intensity of the cyclones. The prediction of cyclones before they are identified as 850-hPa vorticity centers in the analysis cycle was also considered. It is shown that an indication of extratropical cyclones can be given by at least 1 ensemble member 7 days before they are identified in the analysis. Further analysis of the ECMWF EPS shows that the ensemble mean has a higher level of skill than the control forecast, particularly for the intensity of the cyclones, from day 3 of the forecast. There is a higher level of skill in the NH than the SH and the spread in the SH is correspondingly larger. The difference between the ensemble mean error and spread is very small for the position of the cyclones, but the spread of the ensemble is smaller than the ensemble mean error for the intensity of the cyclones in both hemispheres. Results also show that the ECMWF control forecast has ½ to 1 day more skill than the perturbed members, for both the position and intensity of the cyclones, throughout the forecast.

1. Introduction

The atmosphere is a chaotic system and therefore has a finite limit of predictability (Lorenz 1963). Even with a perfect forecast model, small errors in the initial conditions will grow rapidly, resulting in a total loss of predictability at higher lead times. From a practical point of view, today's models are not perfect and deficiencies in the parameterization schemes, used to represent unresolved atmospheric processes, will cause further forecast error. Ensemble prediction provides a method of extending the intrinsic limit of predictability of a single deterministic forecast. By integrating an ensemble of forecasts, each started from slightly different initial conditions, an estimation of the probability density function of forecast states can be obtained (Leith 1974). One of these forecasts is integrated from the analysis and is referred to as the control forecast. The initial conditions for the other ensemble members are obtained by applying perturbations to the analysis, with the aim of sampling the probability density function of the errors in the initial state.

A variety of methods have been developed, and implemented operationally, for generating initial con-

Corresponding author address: Lizzie S. R. Froude, Environmental Systems Science Centre, University of Reading, Harry Pitt Building, Whiteknights, P.O. Box 238, Reading RG6 6AL, United Kingdom.
E-mail: lsrf@mail.nerc-essc.ac.uk

DOI: 10.1175/MWR3422.1

dition perturbations. For example, the European Centre for Medium-Range Weather Forecasts (ECMWF) uses a singular vector method, which selectively samples the perturbations with fastest linear growth over a finite time interval (Buizza and Palmer 1995; Molteni et al. 1996). At the National Centers for Environmental Prediction (NCEP), the bred-vector method is used, which aims to randomly sample fast growing analysis errors using the full nonlinear model (Toth and Kalnay 1993, 1997). The singular vector and bred-vector approaches are similar in the sense that they both select fast growing perturbations. At the Meteorological Service of Canada (MSC), a different approach is used, which involves the assimilation of randomly perturbed observations (Houtekamer et al. 1996). The question of how to best perturb the initial conditions is a subject at the forefront of current research (Hamill et al. 2000).

Current ensemble prediction systems (EPSs) of operational meteorological centers differ not only in how the initial condition perturbations are generated, but by many other factors as well, such as the resolution of the forecast model and the number of ensemble members integrated. As well as perturbing the initial conditions, some centers, such as ECMWF, also perturb the forecast model (Buizza et al. 1999). It is important that the forecasts generated from the different EPS are verified and compared, to explore the impact these factors have on forecast skill and to determine how the forecast systems could be improved. There are a number of conventional methods used to verify ensemble forecasts, such as the root-mean-square (rms) error, the Brier score (Brier 1950), pattern anomaly correlation (PAC; see, e.g., Wei and Toth 2003), and relative operating characteristic (ROC) curves (e.g., Mason and Graham 2002). In a recent study, Buizza et al. (2005) compared the ECMWF, NCEP, and MSC EPS using all of the above forecast verification measures.

Although there have been some studies that focus on individual weather elements, such as precipitation (Ebert 2001), the verification of EPS is often based on the 500-hPa geopotential height. This tends to focus on the large-scale aspects of the general circulation. In this paper an alternative method for assessing forecast skill, which focuses on individual synoptic-scale weather systems, has been applied to the ECMWF and NCEP EPS. The method involves the identification and tracking of extratropical cyclones (Hodges 1995, 1999) along forecast trajectories. Statistics can then be generated to determine the rate at which the position and intensity of the forecast storm tracks diverges from the analyzed tracks with increasing lead time. Diagnostics for other storm characteristics such as their growth and speed can also be produced. This storm-tracking approach to forecast verification was first used in a recent study by Froude et al. (2007, hereafter FBH) to explore the prediction of storms and the impact that different types of observation have on their prediction, using the 40-yr ECMWF Re-Analysis (ERA-40; Simmons and Gibson 2000). The study provided detailed information concerning the prediction of extratropical cyclones, which could not be obtained from other more conventional forecast verification measures. For example, the study showed a higher level of predictive skill for the position of the cyclones than for their intensity, which we believe indicates that the vertical structure of the storms may not be properly represented by forecast models. The study also showed that the propagation speed of the forecast cyclones was generally too slow.

Extratropical cyclones are fundamental to the day-to-day weather in the midlatitudes, and we therefore believe that this analysis method provides a good measure of the ability of NWP to predict the weather. The method is also perhaps particularly well suited for the analysis of an EPS constructed from dynamical perturbations, since analysis errors located in regions of high baroclinic instability, where cyclogenesis is favorable, tend to amplify and grow rapidly throughout the forecast. Indeed, studies have shown that both singular vector and bred vector perturbations tend to be located in highly baroclinic regions (Hartmann et al. 1995; Hoskins et al. 2000; Toth and Kalnay 1997). It is therefore suggested that the storm-tracking analysis methodology may provide a more useful measure of the growth and development of these dynamical perturbations than the error growth of the 500-hPa geopotential height field.

There have been several case studies of the prediction of extratropical cyclones by EPS. Recently Jung et al. (2005) investigated the prediction of three severe European storms by the ECMWF EPS. Other studies include that of Buizza and Hollingsworth (2002) and Buizza and Chessa (2002), which explored the prediction of some particularly destructive storms, also by the ECMWF EPS, that hit Europe and the United States, respectively. These studies have shown that ensemble prediction can provide additional information to that obtained from a deterministic forecast, which can be valuable to the forecaster. One particular advantage of ensemble forecasting highlighted was the measure of risk that can be provided via probabilities. Although these case studies clearly illustrate the value of ensemble forecasting for the prediction of intense cyclones, a statistical analysis of a large number of storms is required to provide a truly complete and objective assessment of the prediction of cyclones by EPS. This is the motivation for this study.

The main aims of this paper are to determine the ability of the ECMWF and NCEP EPS to predict extratropical cyclones and to explore the benefits an EPS can offer over a single deterministic forecast in the prediction of these cyclones. We currently have a larger amount of ECMWF than NCEP EPS data. Since the storm-tracking methodology used in this paper requires a larger amount of data than other forecast verification methods (see FBH), it has not been possible to produce all of the diagnostics for the NCEP EPS. This paper is concerned with the ability of an EPS to predict extratropical cyclones and does not address the reliability of such predictions. Since this paper is the first statistical analysis of the prediction of extratropical cyclones by EPS, a more complete comparison of the two systems and other EPS systems will constitute future work.

The paper continues with a description of the data we have used in section 2. The analysis methodology is discussed in section 3, the results are presented in section 4, and a discussion of the results and conclusions are given in section 5.

2. Data description

The ECMWF EPS data used in this study are from the time period of 6 January–5 July 2005. During this time period the ensemble forecasts were integrated at a spectral resolution of T255 with 40 vertical levels using the operational forecast model. The ECMWF EPS consists of 50 perturbed members integrated out to 10 days at 0000 and 1200 UTC. There is also a control forecast, integrated from the unperturbed analysis, but at the same resolution as the perturbed members. This analysis is obtained from the operational early delivery four-dimensional variational data assimilation (4DVAR) system first introduced in June 2004 (Haseler 2004). A singular vector–based method is used to perturb the initial conditions in the extratropics (Buizza and Palmer 1995; Molteni et al. 1996). The singular vectors are computed at a horizontal resolution of T42 with 40 vertical levels using a model with simplified physics that does not currently include moist processes. A total energy norm (Buizza and Palmer 1995) is used as the measure of growth in the singular vector computation, with an optimization time of 48 h. Both initial and evolved singular vectors are used, corresponding to those perturbations that grow fastest in the next 48 h of the forecast and the 48 h prior to the forecast start time, respectively (Barkmeijer et al. 1999). In addition to the initial condition perturbations, random perturbations are applied to the parameterized physical processes (stochastic physics; Buizza et al. 1999) to represent the model uncertainty.

The NCEP EPS data used in this study are from the time period of 6 January–5 April 2005. During this time period the ensemble forecasts were integrated out to 16 days at a resolution of T126L28 for the first 7.5 days, from which point the resolution was reduced to T62L28 for the remaining 8.5 days of the forecast. NCEP has a smaller ensemble than ECMWF with just 10 perturbed members, but it is run more frequently, every 6 h at 0000, 0600, 1200, and 1800 UTC. There is also a control forecast integrated from the unperturbed analysis at the same resolution as the perturbed members, but only once a day at 0000 UTC. NCEP uses a three-dimensional variational data assimilation (3DVAR) system to generate their analyses, known as the spectral statistical interpolation (SSI) analysis system (Parrish and Derber 1992). A methodology known as the bred-vector method (Toth and Kalnay 1993, 1997) is used to perturb the initial conditions. The method selects growing errors generated in the data assimilation process by running breeding cycles. A breeding cycle is initiated by adding/subtracting a random perturbation to/from the analysis at time t_0. The full nonlinear model is then integrated forward in time from both these initial states for 24 h to time t_1. The difference between the two forecasts is computed and rescaled using a regional rescaling algorithm (Toth and Kalnay 1997). This difference is then added and subtracted from the analysis at time t_1 to form two bred-vector perturbations and the process is then repeated forward in time. NCEP uses five breeding cycles to generate their initial condition perturbations, and unlike ECMWF they do not perturb the forecast model. The NCEP EPS data, used for this study, were accessed directly by the storm-tracking software using the Open-source Project for a Network Data Access Protocol (OPeNDAP; see online at http://www.opendap.org/) from the NCEP nomad5 server (available online at http://nomad5.ncep.noaa.gov/ncep_data/index.html).

Initially the results of this paper were computed with just the first 3 months of the ECMWF data we have available (i.e., the same time period as the NCEP data). The results were the same as those obtained with the 6 months used in this paper, but the statistics were less stable. Unfortunately we only have 3 months of NCEP data available to us at the moment, but we believe that the conclusions of the study are unaffected by this. We have not been able to produce all the statistics for the NCEP EPS. A considerably larger data sample than 6 months (1 yr or more) would be required to generate these missing statistics for two reasons. First the lower frequency of the control forecast, compared with the ensemble members, means it has not been possible to generate statistics for the control forecast. The second

reason is due to the considerably smaller number of ensemble members in the NCEP EPS and is discussed at the end of the next section.

3. Analysis methodology

The analysis methodology used in this study was first implemented in FBH. A brief description of the method will now be given, but the reader is referred to the previous referenced study for a more in-depth description and discussion.

The extratropical cyclones were identified and tracked along the 6-hourly forecast trajectories of each of the perturbed ensemble members and the control forecasts in both hemispheres using the tracking scheme of Hodges (1995, 1999). Before the cyclones were identified, the resolution of the data was reduced to T42 and the planetary scales with total wavenumber less than or equal to 5 were removed (Hoskins and Hodges 2002, 2005). Removing the planetary scales is necessary so that the cyclones can be identified as extrema without being masked by the larger scales. Initially the identification and tracking was performed with both the 850-hPa relative vorticity (ξ_{850}) and mean sea level pressure fields. However, since the results from the two fields were rather similar, this paper concentrates on just the ξ_{850} field. There are several reasons for choosing vorticity rather than pressure (Hoskins and Hodges 2002). Vorticity features with a magnitude exceeding 1.0×10^{-5} s^{-1} were identified as maxima in the NH and minima in the SH and considered cyclones. Once the cyclones had been identified the tracking was performed, which involves the minimization of a cost function (Hodges 1999) to obtain smooth trajectories (storm tracks). The tracking was performed separately in the NH and the SH. Only those storm tracks that lasted at least 2 days, traveled farther than 1000 km, and had a majority of their life cycle in 20°–90°N or 20°–90°S were retained for the statistical analysis. The identification and tracking was also performed with the ECMWF and NCEP operational analyses for the selected time periods to use for the verification.

The forecasted storm tracks, from the two different EPS, were validated against their own analyses using a matching methodology. While it is common practice for operational centers to validate their forecasts against their own analyses, this may result in higher levels of skill than those that would be obtained if the forecasts were validated against an independent data source. This issue was addressed in the study of Bengtsson et al. (2005), which showed that errors in the earlier part of the forecast were most affected by this bias. We have also validated the NCEP EPS against the ECMWF analyses. There was very little difference in the results, but the errors were again slightly larger for the earlier part of the forecast. While we are confident that any biases introduced into the results from verifying the two EPS against their own analyses will have no impact on the conclusions of this paper, the predictive skill of both systems may be a little on the optimistic side.

A forecast storm track was considered to be the same system as an analysis storm track (i.e., matched) if the two tracks met certain predefined spatial and temporal constraints. A forecast track was said to match an analysis track if

1) At least T% of their points overlapped in time, that is, $100 \times [2n_m/(n_A + n_F)] \geq T$, where n_A and n_F denote the total number of points in the analysis and forecast tracks, respectively, and n_m denotes the number of points in the forecast track that overlapped in time with the analysis track.
2) The geodesic separation distance d between the first k points of the forecast track (which coincide in time with the analysis track) and the corresponding points in the analysis track was less than $S°$, that is, $d \leq S°$.

The geodesic separation distance between a point A, on a forecast track, and a point B, on an analysis track, is defined as the great-circle distance between the two points. It is calculated (assuming the earth is a perfect sphere) as $\cos^{-1}(\mathbf{P}_A \cdot \mathbf{P}_B)$, where $\mathbf{P}_A$ and $\mathbf{P}_B$ are unit vectors directed from the center of the earth to points A and B, respectively. We used this geodesic measure to avoid any biases caused by working with projections (i.e., a separation distance of $S°$ corresponds to the same distance in kilometers at any latitude). Since the geodesic separation distance is measured between analysis and forecast points that occur at the same time, it will include components of both along-track error (error occurring purely because the forecasted storm is propagating at a different speed than the analyzed storm) and cross-track error (error occurring because the forecasted storm takes a different path than the analyzed storm). These two types of error can also be measured separately (see sections 4c and 4e).

The forecast tracks that matched analysis tracks were used to generate diagnostics concerning the position and intensity of the storms. In FBH the sensitivity of the diagnostics to the choice of parameters k, T, and S was explored in detail. They found that although the number of forecast storm tracks that matched analysis tracks varied with different choices of the parameters, the diagnostics produced from the matched tracks were unaffected. In this study the diagnostics were initially

produced for three different levels of matching criteria, which are listed below:

1) $k = 4$, $T = 60\%$, and $S = 2°$
2) $k = 4$, $T = 60\%$, and $S = 4°$
3) $k = 4$, $T = 30\%$, and $S = 4°$

As with FBH the number of forecast tracks that matched varied considerably for the different matching criteria, but there was very little difference between the diagnostics generated from the matched tracks. For this reason when we present the results concerning the number of tracks that match we show all 3 matching criteria; however, the diagnostics produced from the matched tracks are shown for just criterion 2. We note that the parameter k is set to 4 (1 day) in all 3 matching criteria. In FBH the same criteria but with $k = 1$ were also explored and were found to have no impact on the diagnostics produced from the matched tracks, but it did cause a large number of tracks to be incorrectly matched. We have therefore kept $k = 4$ for this study.

As an additional constraint, only those storms whose genesis occurs within the first 3 days of the forecast or that already existed at time 0 were considered. Results from the study of Bengtsson et al. (2005) indicated that the skill in predicting storm tracks after 3 days is relatively low. If a storm was generated in a forecast at a lead time greater than 3 days, and matches a storm in the analysis, then it was probably more due to chance than an accurate prediction. Although this may not be the case for the more recent forecast and analysis systems used in this study, we keep this constraint so that the methodology is consistent with FBH and the statistics of two studies can therefore be compared.

The number of ensemble members that match will vary for different storms and forecast start times (see section 4b). In addition the storm tracks of the different ensemble members will be different lengths and so the number of ensemble member tracks available decreases with increasing lead time. The statistics of section 4 therefore only include those data points where at least 5 perturbed member tracks are available, since calculating these values from less than 5 members would not be very informative. Restricting the diagnostics further by increasing this value of 5 does have some effect on the diagnostics, but it also limits the amount of data available, particularly at the higher lead times. We believe that as long as all the diagnostics are produced with the same restriction, the conclusions will be unaffected by the choice of value. Since the NCEP EPS has a smaller number of ensemble members than the ECMWF EPS, there will be less data points (particularly at higher lead times) where at least 5 matching perturbed members are available. This is the second reason (see section 2) why a much larger data sample would be required to produce some of the statistics of this paper for the NCEP EPS.

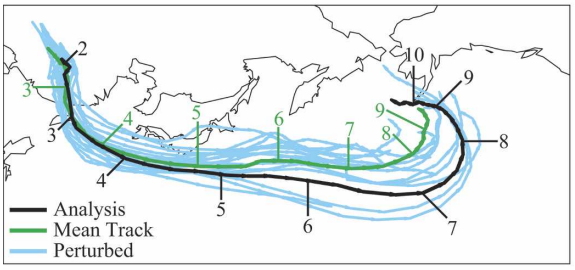
4. Results

a. An example of a Pacific storm

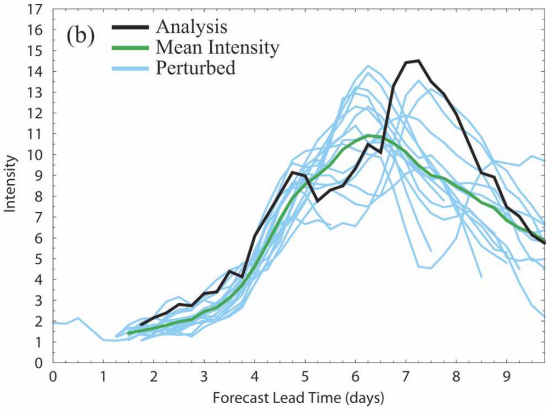
Before presenting the statistical analysis, an example of an intense Pacific storm predicted by both the ECMWF and NCEP EPS (Fig. 1) is discussed to help put the statistics into context. The storm was first identified in the ECMWF analysis at 0600 UTC 1 March 2005 in southeast China. It then traveled over Taiwan and across the Pacific south of Japan reaching its peak 5 days later. The track of the storm then curved in to the northwest while decaying over the next 3½ days.

Figure 1a shows the track of the storm in the ECMWF analysis and the tracks from the ECMWF ensemble forecast started at 1200 UTC 27 February 2005 (1.75 days before the storm was identified in the ξ_{850} analysis) that match the analysis track for matching criterion 2. Figure 1b shows the corresponding intensities (in units of 10^{-5} s^{-1} of ξ_{850} relative to background field removal) as a function of forecast lead time. The mean track and mean intensity of all matching ensemble member tracks is also shown. Sixteen of the perturbed members contain a track that matches the analysis track in this earlier forecast. The control forecast does not have a matching track for this particular forecast. This is due to the generation of a double ξ_{850} center early in the forecast, causing the track to be separated into two parts and not satisfy the matching criterion. Relaxing the matching criterion does result in a larger number of matched ensemble member tracks for this particular storm. However, as explained previously, the statistics produced from the matched tracks were virtually identical for the different matching criteria and since we have chosen criterion 2 for the statistics, we use it also in this example for consistency. The genesis of the storm occurs a little earlier in time and farther inland than the analyzed storm for most of the perturbed members. A majority of the tracks predicted by the perturbed members lie to the left of the analyzed track, which causes the mean track to also lie to the left. It can also be seen from the numbers indicating the forecast lead time on the mean track and analyzed track that the ensemble member storms are generally moving at a slower speed than the analyzed storm. The initial growth of the cyclone, until around day 5 of the forecast, is predicted well by all the matching ensemble members. At this point the analyzed cyclone weakens slightly before further increasing in intensity. Although a couple of ensemble members do predict the double-

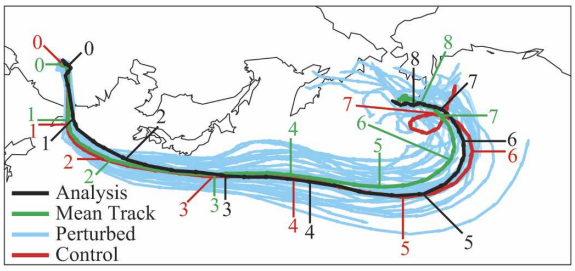
(a) ECMWF - forecast started 1200 UTC 27 Feb 2005



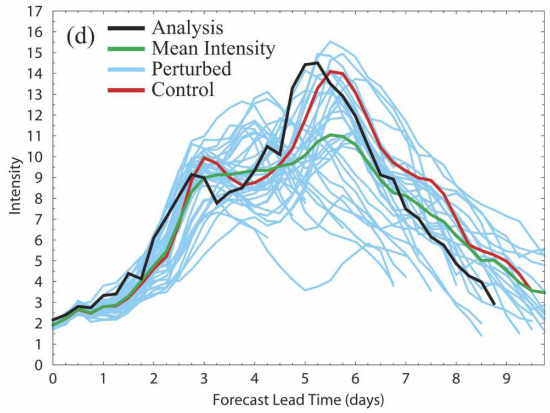
(b)



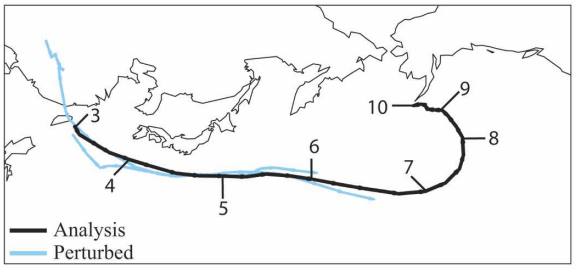
(c) ECMWF - forecast started 1200 UTC 1 Mar 2005



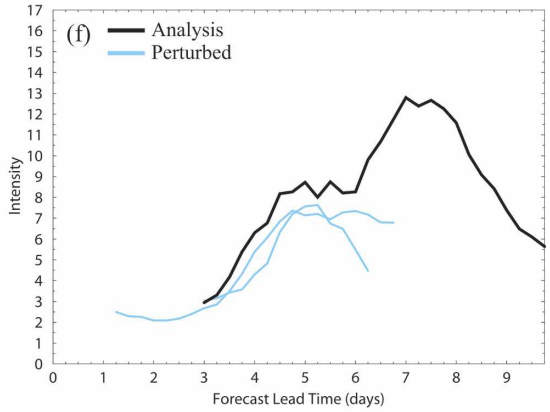
(d)



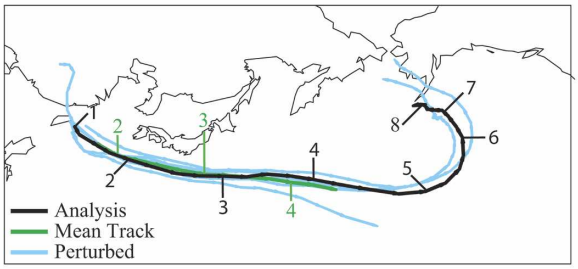
(e) NCEP - forecast started 1200 UTC 27 Feb 2005



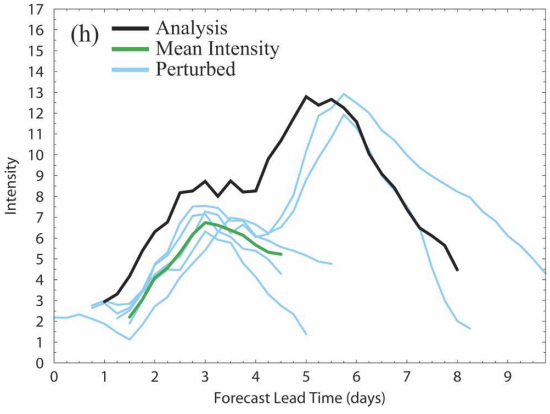
(f)



(g) NCEP - forecast started 1200 UTC 1 Mar 2005



(h)



peaked form of the analyzed cyclone quite well, most of the ensemble members continue to increase in intensity at day 5, without first decaying slightly. They then reach their peak too early at around day 6 of the forecast. It can be seen from the mean intensity that the amplitude of the cyclone is generally underpredicted.

Figures 1c,d show the tracks and intensities predicted by the ECMWF ensemble forecast started at 1200 UTC 1 March, which is the first forecast to be made when the ξ_{850} center is present in the initial conditions. Here the control forecast predicts the track of the storm very well. The control also predicts the amplitude of the storm very well; it is just slightly out of phase with the analysis. Thirty-seven of the perturbed members predict the storm. The ensemble member tracks are more evenly distributed about the analyzed track than in the earlier forecast, resulting in a high-quality mean track prediction that is better than the control forecast at the high lead times. Again the forecasted cyclones are moving at a slower speed than the analyzed cyclone. Some of the perturbed members, like the control, predict the double-peaked shape of the analysis, while others reach their peak a day or more early (as in the earlier forecast) and then completely decay without first regaining intensity. The mean intensity is actually very similar to the earlier forecast, showing a significant underprediction of the storm's amplitude by the ensemble members, between day 4 and 6, around the storm's peak.

Figures 1e,f show the track and intensity of the storm identified with the NCEP analysis. The storm is not identified until 1200 UTC 2 March, 30 h after the ECMWF analysis, and it does not reach as high a peak in intensity. This is probably due to the lower resolution of the NCEP analysis system to that of ECMWF. The 1200 UTC 27 February 2005 NCEP EPS has two matching perturbed ensemble members. Both of the ensemble members only predict the earlier part of the storm. A low-resolution control forecast is not available for the NCEP EPS at 1200 UTC.

Figures 1g,h show the tracks and intensity of the storm predicted by six ensemble members of the NCEP EPS started at 1200 UTC 1 March. The mean track and intensity is shown when there are at least five ensemble members available to average. Two of the ensemble members predict the track and intensity of the storm very well, but the other four members do not regain intensity after the initial growth and have decayed completely by day 5 or 6 of the forecast. The ensemble members underpredict the amplitude of the cyclone. As with the ECMWF EPS, the forecasted cyclones are propagating at a slower speed than the analyzed cyclone on average.

b. Number of forecast storm tracks that match

In this section the impact that the 3 different matching criteria have on the number of forecast storm tracks that match analysis storm tracks is investigated. Table 1 shows the percentage of tracks from all the perturbed ensemble members that match with analysis tracks, for the ECMWF and NCEP EPS, for each of the 3 matching criteria. It also shows the percentage of tracks from just the control forecast that match with analysis tracks. The second column shows the total number of forecast tracks, which exist at time 0 or whose genesis occurs within the first 3 days of the forecast. Since this was a restriction introduced in the matching methodology (see section 3a), the other columns of the table show the percentage of these tracks that match with analysis tracks.

The percentages increase steadily as the matching criteria are relaxed and are comparable in both hemispheres. A higher percentage of the perturbed ensemble member tracks and the control forecast tracks match for the ECMWF EPS than for the NCEP EPS. The difference between the percentage of ECMWF and NCEP perturbed member tracks that match is comparable to the difference between the percentage of ECMWF and NCEP control forecast tracks that match. This suggests that the superior skill of the ECMWF EPS is due to the higher-resolution model and 4DVAR data system rather than the perturbation methodology. *Another noticeable observation is that the percentage of control forecast tracks that match is consistently higher (by up to 10%) than that of the perturbed member tracks for both the ECMWF and NCEP EPS*. Since the control forecast has been generated from an optimal analysis, one might expect it to be better than the perturbed members for the earlier part of the forecast. The spatial matching focuses on the earlier parts of the forecast

←

FIG. 1. Example of the prediction of a Pacific storm by the ECMWF and NCEP EPS. The tracks and intensities, as a function of forecast lead time, of the analyzed storm and storm predicted by the ensemble members of the ECMWF EPS started at 1200 UTC (a),(b) 27 Feb 2005 and (c),(d) 1 Mar 2005 and by the NCEP EPS started at 1200 UTC (e),(f) 27 Feb 2005 and (g),(h) 1 Mar 2005 are shown. The mean track and mean intensity of the ensemble members is also shown when at least 5 ensemble members are available. Units of intensity are 10^{-5} s^{-1} (relative to background field removal). The numbers along the tracks correspond to the forecast lead time (in days).

TABLE 1. Percentage of ECMWF and NCEP perturbed ensemble member tracks and control ensemble member tracks that match with analysis tracks for each of the three matching criteria in the NH and SH. The second column in the table shows the total number of forecast tracks that begin within the first 3 days of the forecast and the other columns show the percentage of these tracks that match analysis tracks.

	No. of tracks in forecasts	% match $S = 2°$, $T = 60\%$	% match $S = 4°$, $T = 60\%$	% match $S = 4°$, $T = 30\%$
NH:				
ECMWF perturbed	414 762	26.3	38.9	56.3
ECMWF control	8239	35.5	47.5	64.8
NCEP perturbed	83 928	20.9	36.5	52.1
NCEP control	2089	29.1	42.6	58.5
SH:				
ECMWF perturbed	432 389	26.9	39.7	57.7
ECMWF control	8683	36.1	48.8	66.8
NCEP perturbed	85 201	21.1	34.3	51.8
NCEP control	2120	28.6	42.3	61.7

tracks and this may therefore be the reason why a higher percentage of control forecast tracks match.

We note that all the forecast storm tracks (except those whose genesis occurred at a lead time greater than 3 days) were compared with all of the analysis tracks. This means that the percentages of Table 1 include forecast storm tracks that have been identified in forecasts, integrated from initial states that occur before and after the vorticity center was first identified in the 850-hPa analysis (see the discussion of FBH's Table 2 for further clarification).

It is apparent that ensemble forecasts, integrated from analyses in which the ξ_{850} centers already exist, are likely to have a larger number of matching ensemble member tracks than ensemble forecasts started from earlier analyses. This is investigated in Fig. 2, which shows the average percentage of perturbed ensemble members that match analysis tracks as a function of the number of days M the forecast was started after the ξ_{850} center was first identified in the analysis. Positive values of M correspond to forecasts that are integrated from analyses after the ξ_{850} center was first identified. Negative values correspond to forecasts that are integrated from analyses before the ξ_{850} center was first identified. A value of 0 corresponds to forecasts that are integrated from analyses in which the ξ_{850} center was first identified. Since the ECMWF EPS has 50 perturbed members and NCEP has just 10, the percentages correspond to different numbers of ensemble members for the 2 systems. For example, a value of 50% corresponds to 25 ECMWF ensemble members and 5 NCEP ensemble members. The results are shown for both the ECMWF and NCEP EPS for each of the 3 matching criteria in both hemispheres.

The shape of the curves is very similar for the ECMWF and NCEP EPS. Relaxing matching criterion 1 to criterion 2 (i.e., the spatial condition has been relaxed from 2° to 4°) has significantly more impact on forecasts for which $M \leq 1$. This is not surprising since the forecasts with $M > 1$ have been generated from analyses that contain well-developed ξ_{850} centers and the positions of the cyclones are likely to be well predicted for the first part of the forecast. Relaxing the spatial criteria will therefore have a limited effect. Forecasts with $M \leq 1$ will have been integrated, either from analyses that contain weak and undeveloped ξ_{850} centers, or from analyses that do not contain ξ_{850} centers at all. This means the positions of the cyclones will probably not be as well predicted in general and relaxing the spatial criterion will therefore have more of an impact. Relaxing matching criterion 2 to 3 (i.e., the temporal criterion has been relaxed from 60% to 30%) causes the percentages to further increase, but has the most impact when $M \geq 2$. This is because forecasts with high values of M can only predict the later part of the storm tracks. The number of points that coincide with the analysis tracks will therefore be limited, meaning that a temporal criterion with a high value of T (i.e., 60%) may not be satisfied. A similar argument would apply to forecasts for which $M \leq -2$, since these forecasts can only predict the earlier part of the storm tracks. However, we only considered storm tracks whose genesis occurs within the first 3 days of the forecast and consequently very few ensemble member tracks match for $M \leq -3$. The few that do correspond to forecast tracks that begin earlier in time than the corresponding analyzed tracks.

The percentage of ensemble members that match is slightly higher in the SH than the NH for both the ECMWF and NCEP EPS. This could be because SH storms are subjected to less changeable surface boundary conditions, in the form of coastlines and varying

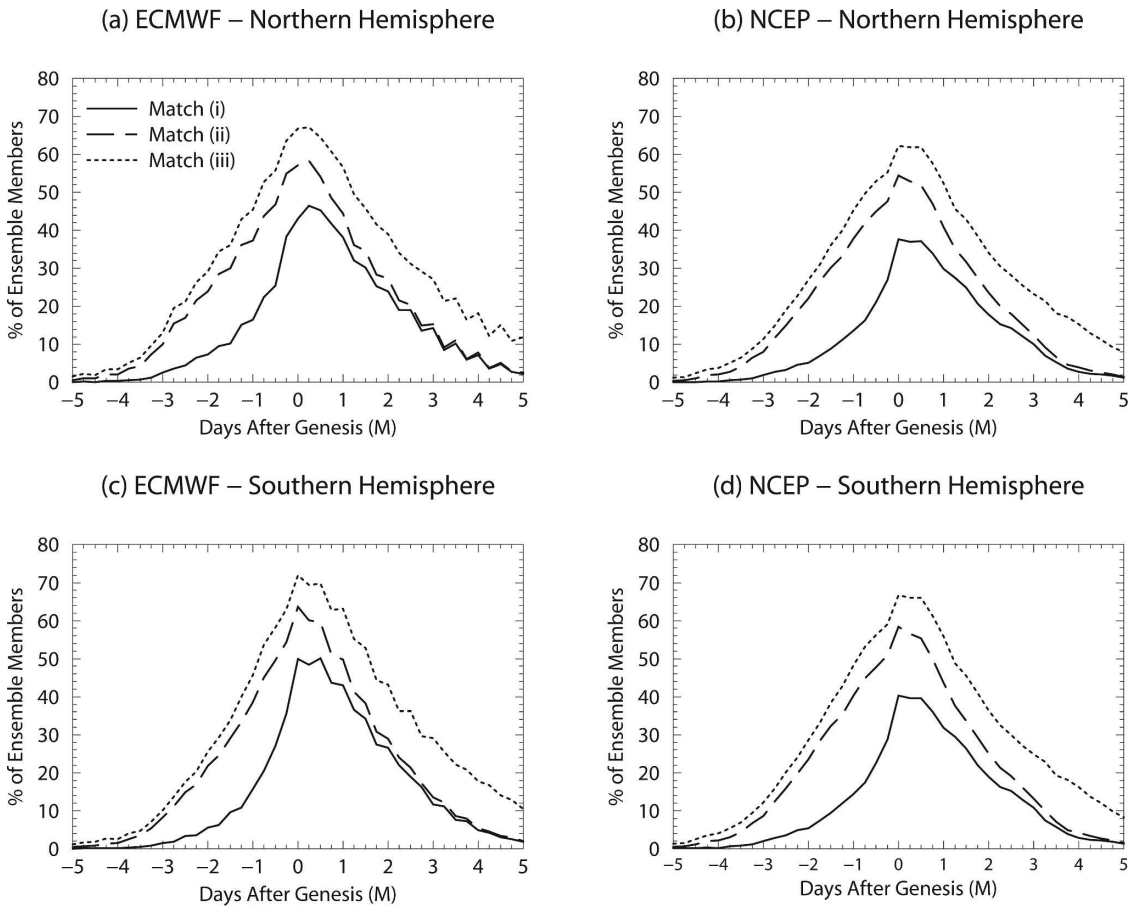


FIG. 2. Average percentage of perturbed ensemble members that have a storm track that matches the analysis track for the ECMWF EPS in the (a) NH and (c) SH, and for the NCEP EPS in the (b) NH and (d) SH for each of the three matching criteria. The *x* axis corresponds to the number of days *M* after the cyclones have been identified as 850-hPa vorticity centers in the analysis. Negative numbers correspond to forecasts made before the cyclones have been identified in the analysis.

orography, than NH storms (this idea was also discussed in the Bengtsson et al. 2005 study). NH storm tracks may therefore be more likely to get broken into smaller sections and consequently not satisfy the temporal matching criterion. In agreement with the results of Table 1, the percentage of ensemble members that match is higher for the ECMWF EPS than for the NCEP EPS.

c. *Forecast skill of the perturbed members and the control*

As explained previously in section 3, the matched tracks were used to generate further diagnostics. The statistics of this subsection and the following subsections have been generated from tracks that matched using criterion 2. All the forecast tracks for different values of *M* (see above) have been included in the diagnostics. Ideally it would be better to separate the diagnostics for different values of *M*, but this would require a larger data sample than we currently have available (see FBH for more discussion of this).

Figure 3 shows some diagnostics for the cyclones predicted by the ECMWF EPS. The solid lines of Figs. 3a,c show, in the NH and SH, respectively, the minimum, mean, and maximum separation distance of the matched perturbed member storm tracks from the analysis tracks as a function of forecast lead time. As with the matching criteria these separation distances are calculated between points on the analysis and forecast tracks that occur at the same validation time and therefore include components of along-track and cross-track error. It should be noted that the mean curve is the mean separation distance of the perturbed members from the analysis and not the separation distance of the mean track from the analysis, which is investigated in the next subsection of this paper. The mean

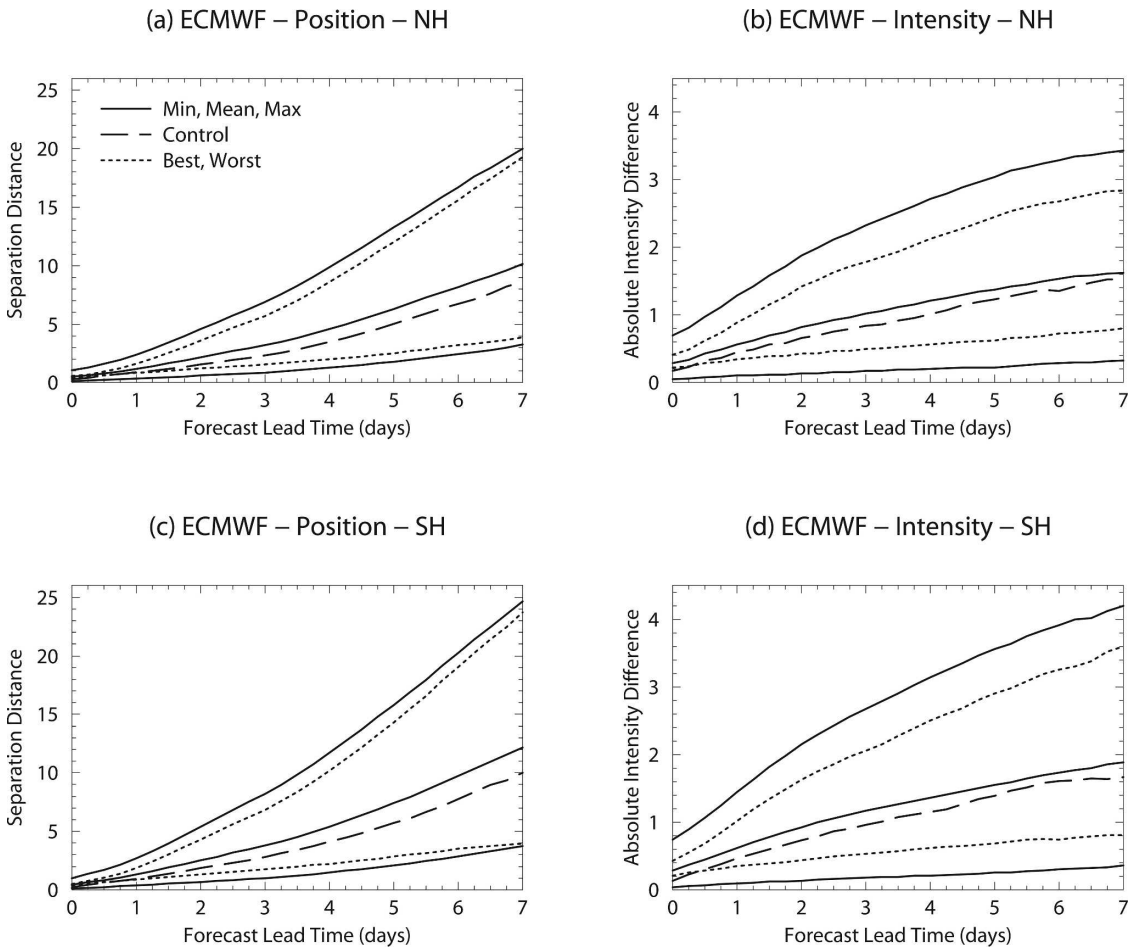


FIG. 3. Ensemble diagnostics for the ECMWF EPS. The solid lines show the minimum, mean, and maximum separation distance in the (a) NH and (c) SH, and absolute intensity difference in the (b) NH and (d) SH, between the matched perturbed member tracks and analysis tracks as a function of forecast lead time. The dotted lines show the separation distance/absolute intensity difference between the best and worst perturbed ensemble members (for details see text) and analysis tracks and the dashed lines show the separation distance/absolute intensity difference between the control forecast tracks and analysis tracks. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal), respectively.

curve in the figure therefore corresponds to the average skill of the perturbed member forecasts. The difference between the minimum and maximum curves provides a measure of the ensemble spread. In this subsection the term "spread" is used to refer to this difference. This is different than the usual meaning of ensemble spread, which is used to refer to the average distance of the ensemble members from the ensemble mean and is explored in the next subsection of this paper.

The dotted lines of Figs. 3a,c show the separation distance of the best and worst track of the ensemble from the analysis. To calculate the best and worst track, the average separation distance of each ensemble member track from the corresponding analysis track over its whole lifetime was computed. The ensemble members with the lowest and highest values were taken to be the best and worst tracks, respectively. The minimum (and maximum) error of the ensemble will be obtained from different ensemble member tracks at different lead times, whereas the best (and worst) ensemble member error is obtained from the same ensemble member at all lead times. The diagnostics described above have included the perturbed ensemble members only and not the control. In Figs. 3a,c, the dashed line shows the separation distance of the matched control forecast tracks from the analysis tracks. Figures 3b,d show the same diagnostics, but for absolute intensity difference rather than separation distance. Here the best and worse ensemble member tracks are determined by the average absolute intensity difference between forecast

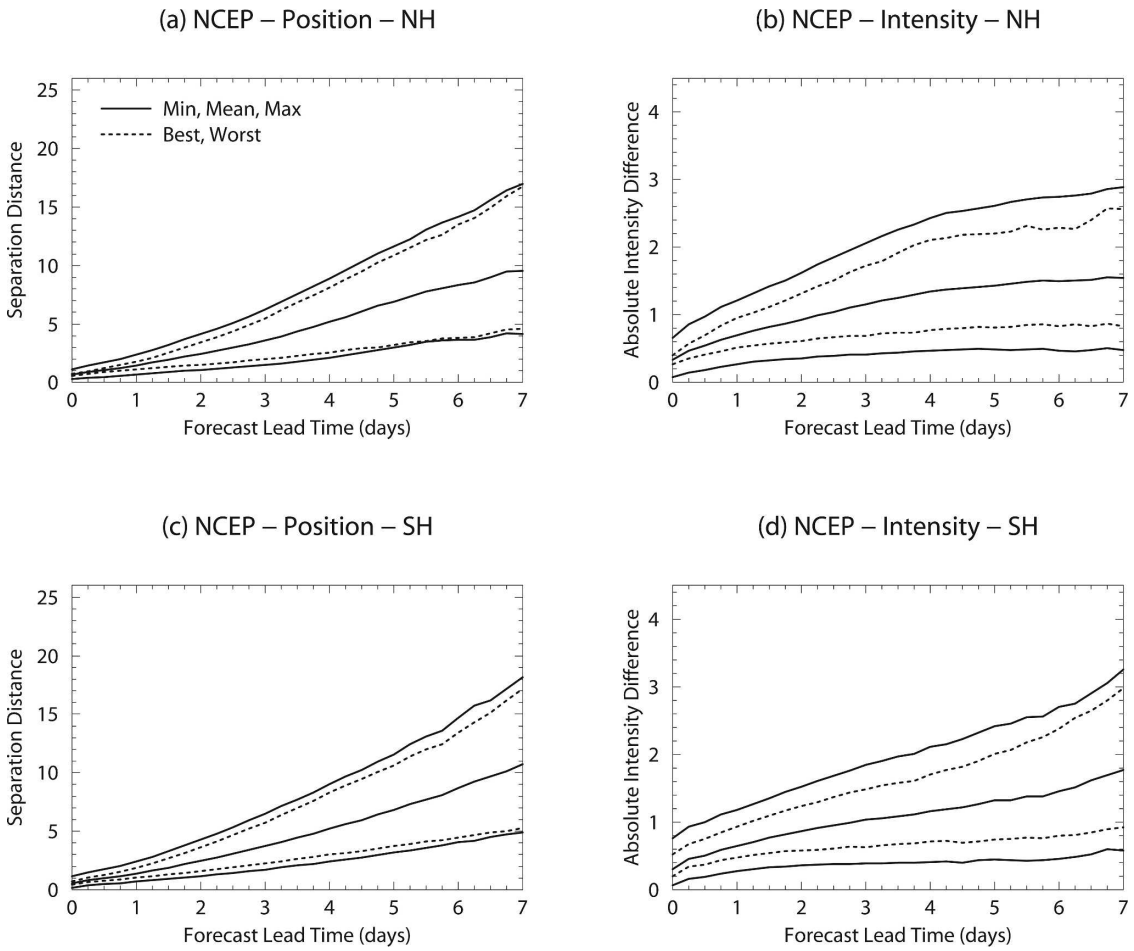


FIG. 4. The same as in Fig. 3 but for the NCEP EPS. The diagnostics for the control forecast are not included.

and analysis tracks over their whole lifetime. Hence the best ensemble member for the separation distance results is not necessarily the best member for the intensity results. In fact, calculations show that only about 12% of ensemble members that are best for position are also best for intensity. Figure 4 shows the diagnostics for the NCEP EPS. Since the NCEP control forecast is only run once a day, the NCEP EPS data we currently have available are not sufficient for producing the diagnostics for the control forecast.

By comparing the separation distance curves with the intensity curves, it can be seen that the error growth rates of the storm's intensity differ considerably with those of the storm's position for both the ECMWF and NCEP EPS. For intensity the errors grow rapidly in the earlier part of the forecast, but at higher lead times the errors grow at a slower rate and the curves become almost flat, showing signs of saturation. The error in position, on the other hand, grows quite slowly at the beginning of the forecast but becomes faster at the higher lead times. This difference between the error growth rates was also found in FBH and suggests a higher level of skill in the prediction of the position of the cyclones than in their intensity. The spatial matching may have some impact on the growth of the separation distances. However, in FBH the diagnostics were also produced using just one point for the spatial matching criterion ($k = 1$) and there was virtually no difference in the position or intensity error growth rates. There is also a larger spread (more uncertainty) in intensity than position. This is deduced by comparing the minimum and maximum curves with the mean curves. For example, for the ECMWF EPS in the NH (Figs. 3a,b), the maximum error in position at day 3 is about 7° and the mean error reaches this value about 2½ days further into the forecast. Similarly the mean error at day 3 is about 3° and the minimum error curve reaches this value about 3½ days further into the forecast. Now considering the intensity curves, the maximum error at day 3 is about 2.4×10^{-5} s^{-1}, but the

mean error does not reach this value in the 7-day forecast range plotted. A similar result is found when comparing the minimum and mean intensity errors.

The ECMWF control forecast has consistently higher skill than the perturbed members by ½–1 day for both position and intensity. This is to be expected in the earlier part of the forecast, since the initial part of the control forecast has been optimized by the 4DVAR system. If the perturbed ensemble members were obtained using only initial condition perturbations (and not model perturbations), then it might be expected that the error of the control forecast would converge to that of the perturbed members at higher lead times. However, the ECMWF EPS also includes model perturbations, which may have an impact on the skill of the perturbed members at higher lead times. Unfortunately we have insufficient data for the NCEP control forecast to determine whether it is also consistently better than the perturbed members.

The difference between the mean curves and the best ensemble member curves is significant, showing that the best ensemble member can provide a much better prediction of a cyclone than a single deterministic forecast. For the ECMWF EPS the day 5 skill of the best ensemble member is comparable to the day 3 skill of the control and to the day 2½ skill of the mean curve for the position of the cyclones. For the intensity of the storms these improvements are even greater, increasing by about 1 day. Similar improvements are also gained from the best ensemble member of the NCEP EPS. The high level of skill of the best ensemble member is encouraging in itself because it indicates that the errors in the initial state are being sampled effectively. However, from a practical point of view, the question of how soon into the forecast the best ensemble member can be determined is more important. If, for example, the ensemble member that is best for the first day of the forecast is still best (or better than the average ensemble member) some time further into the forecast, then this would provide helpful information to an operational weather forecaster. This was investigated by selecting the ensemble member that was best for the first day and first 2 days of the forecast and then computing the error growth of this ensemble member (not shown). Rather disappointingly, the error of these selected ensemble members diverges very quickly to that of the average ensemble member.

The perturbed members and the control forecast of the ECMWF EPS have slightly less skill in the SH than the NH. A more noticeable feature is that the spread of the ensemble is considerably larger in the SH than the NH. The NCEP perturbed members also have slightly less skill in the SH than the NH, but the spread is very similar in the two hemispheres.

By comparing Fig. 3 with Fig. 4 we see that the NCEP EPS has a smaller spread than the ECMWF EPS. This is because the NCEP EPS has far fewer ensemble members and this is therefore not a very fair or objective comparison. In Fig. 5, the minimum, mean, and maximum diagnostics are shown for the NCEP EPS and for a 10-member version of the ECMWF EPS, obtained by randomly selecting 10 of the 50 ECMWF ensemble members. In the NH, the ECMWF perturbed ensemble members have approximately ½ a day more skill in predicting the position of the cyclones and approximately 1 day more skill in predicting the intensity of the cyclones than the NCEP perturbed ensemble members. However, in the SH, the NCEP perturbed ensemble members are about ½ a day better, from day 4 of the forecast, at predicting the position of the cyclones and are about 1 day better, from day 3 of the forecast, at predicting the intensity. As with the full 50-member ECMWF EPS, the spread of the 10-member ensemble is larger in the SH than the NH. This is investigated in more detail in the next subsection.

As mentioned previously, the error in cyclone position discussed so far includes both along-track and cross-track error (referred to as total position error). The two components of the total position error were also computed separately. Figure 6 shows a schematic illustrating the along-track, cross-track, and total position error. As with the total position error, the along-track and cross-track errors were calculated using the geodesic separation distance defined in section 3. The cross-track intersection point on the analysis track (labeled X on the schematic) is also calculated using spherical geometry. Figure 7 shows the minimum, mean, and maximum diagnostics for along-track and cross-track error for the ECMWF EPS. The along-track errors are larger than the cross-track errors in both hemispheres. This shows that errors in the propagation speed of the forecasted cyclones are having a larger impact on the total position error than errors in the track the storm takes. These errors are investigated in more detail in section 4e. The along-track and cross-track errors were also computed for the NCEP EPS (not shown). As with the ECMWF EPS, the along-track errors were larger than the cross-track errors.

The diagnostics of this subsection include cyclones of all amplitudes, but they were also produced for just the intense cyclones (not shown) using the same methodology as FBH. As with this previous study, the results were comparable for the position of the storms, but

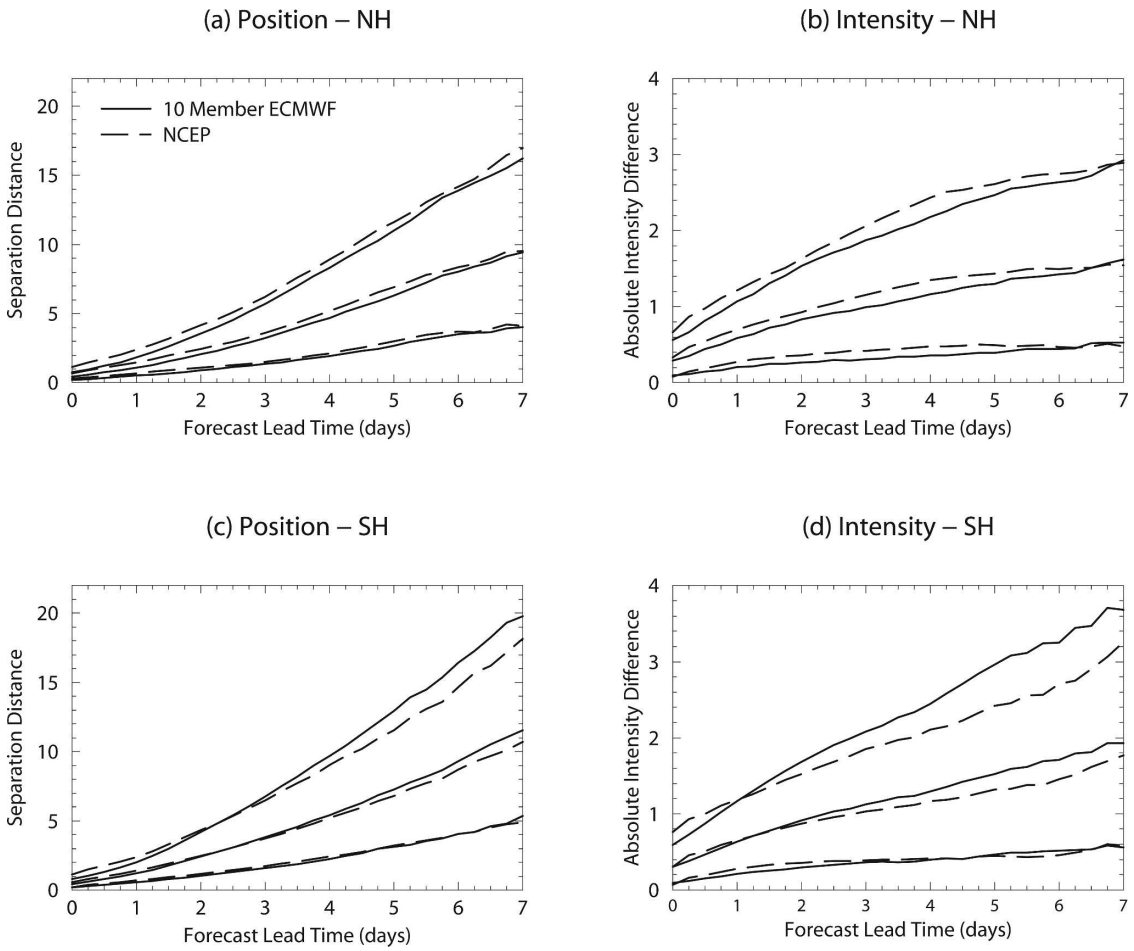


FIG. 5. Minimum, mean, and maximum separation distance in the (a) NH and (c) SH, and absolute intensity difference in the (b) NH and (d) SH, between the matched perturbed member tracks and analysis tracks for the NCEP EPS (dashed lines) and for 10 randomly selected perturbed members of the ECMWF EPS (solid lines) as a function of forecast lead time. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal), respectively.

there was a larger absolute error in the predicted intensities of the higher-amplitude storms.

d. The ensemble mean and ensemble spread

One of the aims of ensemble prediction is that the average of the ensemble forecasts will provide a forecast that is, although somewhat smoothed, superior to the control forecast (Leith 1974; Toth and Kalnay 1993, 1997). This is investigated in this subsection. The results are only presented for the ECMWF EPS, since we have insufficient data at this time to generate the diagnostics for the NCEP EPS. For each ensemble forecast, the mean track and mean intensity of the ensemble member tracks (including the control) that matched were computed. Figures 8a,c show the mean separation distance of the mean tracks and control tracks from the analysis tracks in the NH and SH, respectively. The skill of the mean track and control track is almost identical until day 4 of the forecast, from which point the error growth of the control track becomes slightly larger. By day 7 of the forecast, the mean track has about a ½ a day advantage over the control forecast in both hemispheres. However, this may be of little benefit, since forecasts of this high lead time will have low levels of skill in general.

Figures 8b,d show the results for absolute intensity difference. The difference between the error growth rates of the ensemble mean and control is more significant for the cyclone intensity than for position. From day 2 of the forecast the error growth of the control forecast is larger than that of the ensemble mean, and by day 7 of the forecast the ensemble mean has about 2 days more skill than the control forecast. The skill of the ensemble mean is higher in the NH than the SH for

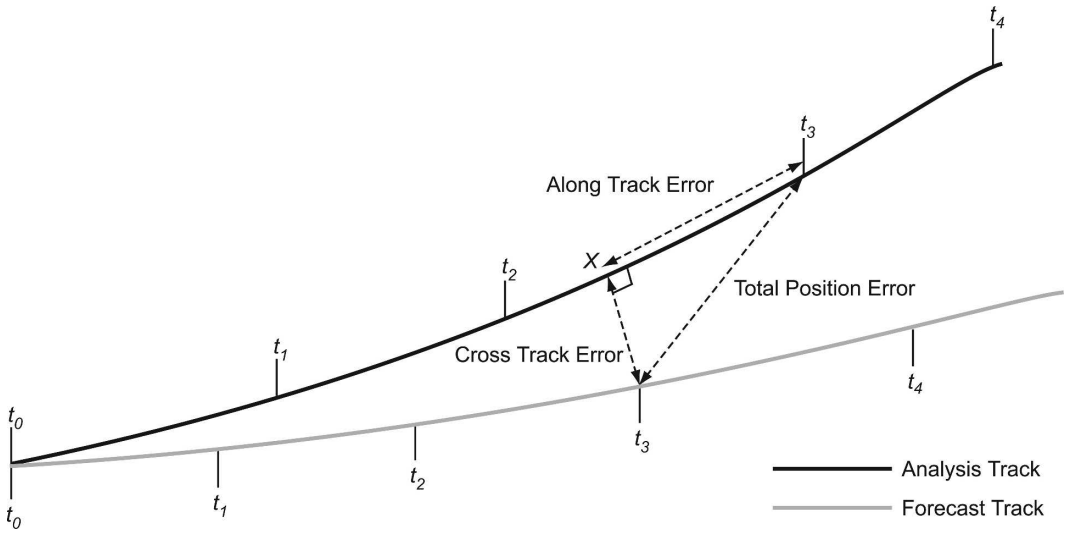


FIG. 6. Schematic illustrating the along-track and cross-track error.

both position and intensity of the cyclones. This corresponds to the larger spread in the SH than the NH shown in Figs. 3, 5.

Although the difference between the ensemble minimum and maximum error curves of Figs. 3–5 provides a measure of the ensemble spread in relation to the analysis, ensemble spread is generally measured in relation to the ensemble mean. For an EPS to be statistically reliable, the average distance of the ensemble mean from the analysis should be equal to the average distance of the ensemble members from the ensemble mean (i.e., the error of the ensemble mean should be equal to the ensemble spread). Ensemble spread was calculated as the mean separation distance/absolute intensity difference of the ensemble member tracks from the ensemble mean track. Figure 9 shows the ensemble spread and ensemble mean skill for the position and intensity of the cyclones in both hemispheres. The difference between the ensemble mean error and spread is very small for the position of the cyclones. In the NH the spread is almost identical to the skill until day 4, when the ensemble mean error begins to exceed the spread ever so slightly. In the SH the results are even more encouraging, since very little difference can be seen between the two curves. The lower level of skill in the SH corresponds well with the larger spread.

There is a much larger difference between the mean error and spread curves for the intensity of the cyclones. The error growth of the ensemble mean is larger than the spread in both hemispheres, but there is more of a difference in the NH. This shows that the intensities predicted by the ensemble members are not evenly distributed about the analyzed intensities. This underdispersion of the ECMWF EPS was also found by Buizza et al. (2005) for the 500-hPa geopotential height field, from about day 5 of the forecast. The Buizza et al. (2005) study also found this to be the case, and to a greater extent, for the NCEP and MSC EPS.

Since the ensemble mean and spread results of this subsection are computed from ensemble member tracks that match with analysis tracks, they are unknown before the forecast verification time has past. It is suggested that in an operational forecast situation, measures of ensemble mean and spread could instead be obtained by matching perturbed ensemble member tracks with control forecast tracks, rather than with analysis tracks. Since the diagnostics of this subsection were unaffected by the choice of matching criteria (see section 3), we believe this could provide a useful practical measure of ensemble mean and spread.

e. Intensity, propagation speed, and track bias

In this subsection we investigate whether there is any bias in the prediction of cyclone intensity, propagation speed, and track. Figures 10a,c show the mean signed intensity difference between the matched ECMWF and NCEP perturbed ensemble member tracks and corresponding analysis tracks for the NH and SH, respectively. The difference between the matched ECMWF control tracks and analysis tracks is also shown, but we have insufficient data to show this for the NCEP control. In the NH, the ECMWF system shows very little bias, but there is a small positive bias in the SH showing that the perturbed forecasts and the control forecast are in general slightly overpredicting the amplitude of the storms. However, the small magnitude of these biases

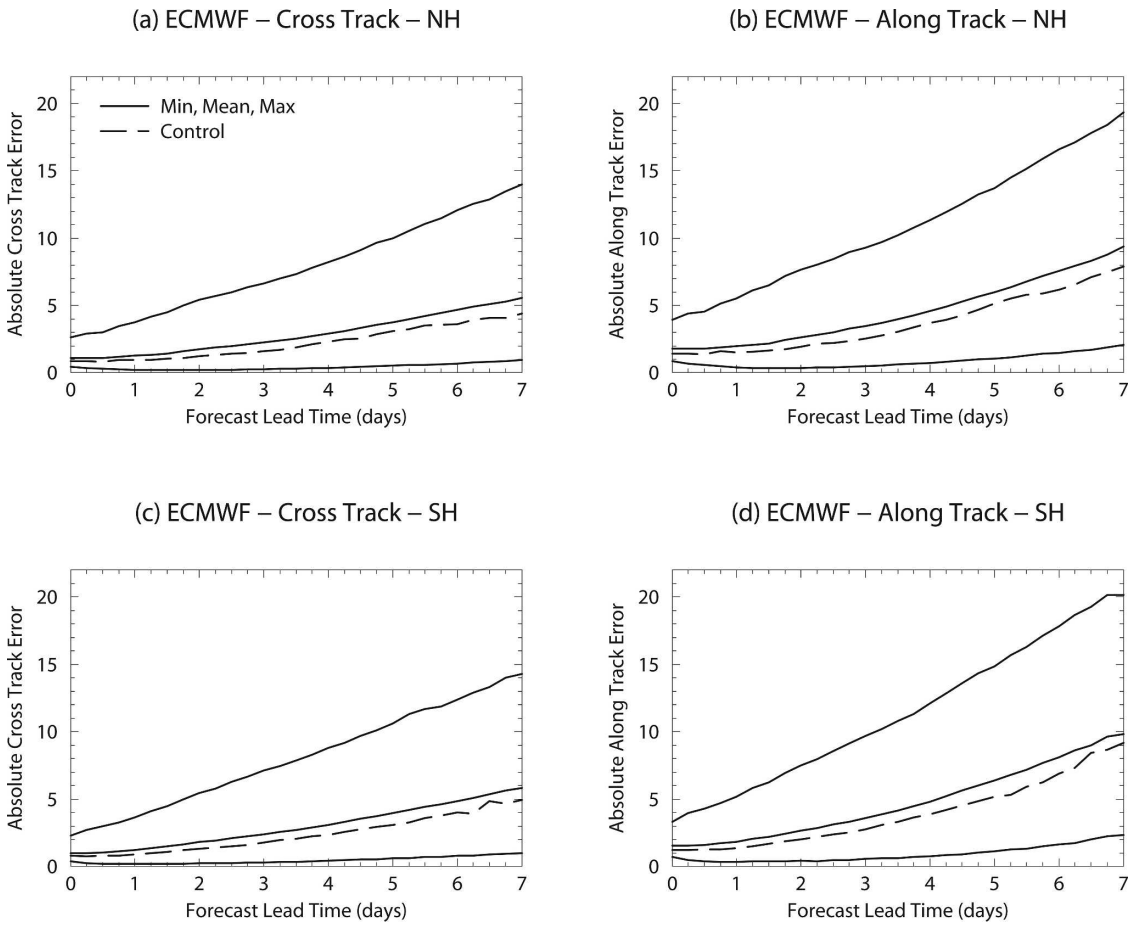


FIG. 7. Minimum, mean, and maximum cross-track error in the (a) NH and (c) SH, and along-track error in the (b) NH and (d) SH, of the perturbed members storm tracks of the ECMWF EPS. The errors are also shown for the control forecast. Units of cross-track and along-track error are geodesic degrees.

should be noted. The perturbed member bias grows faster than the control bias over the first 2 days of the forecast in both the NH and SH. It seems likely that this corresponds to the initial condition perturbations, which have been selected because they grow fastest (in terms of the total energy norm) over the first 2 days of the forecast.

The results are very different for the NCEP EPS. In the NH, the bias of the perturbed members becomes increasingly negative with forecast lead time, whereas in the SH there is a very small positive bias at the beginning of the forecast. This could perhaps be due to the lower resolution of the NCEP EPS, which may be unable to capture the rapid growth of some NH storms. Indeed, FBH showed that the growth of intense NH storms was badly predicted. In this previous study, the forecast model was integrated at a resolution of T159L60, which is still higher than the T126/T62L28 resolution of the NCEP EPS.

FBH also showed that the forecasted storms in general moved at a slower speed than the analyzed storms. To determine whether this was also the case for the EPS, we calculated the propagation speeds of the analysis and forecast storms at each point on their tracks, by comparing the position of consecutive points on the tracks. Since the points on the tracks are 6 h apart, the speed calculated at each point corresponds to the average propagation speed of the storm in the next 6 h. Figures 10c,d show the mean signed speed differences between the matched forecast tracks and analysis tracks in the NH and SH, respectively. Although the bias is small in magnitude, it is consistently negative for the ECMWF EPS in both hemispheres and is larger in the NH. This difference between the hemispheres was also found in FBH, but the magnitude of the bias was slightly larger in both hemispheres. The NCEP EPS has a similar bias in the SH, but in the NH the bias is negative initially and then becomes positive from day 3 of the forecast. This may simply be because the smaller data sample we have available for the NCEP EPS is

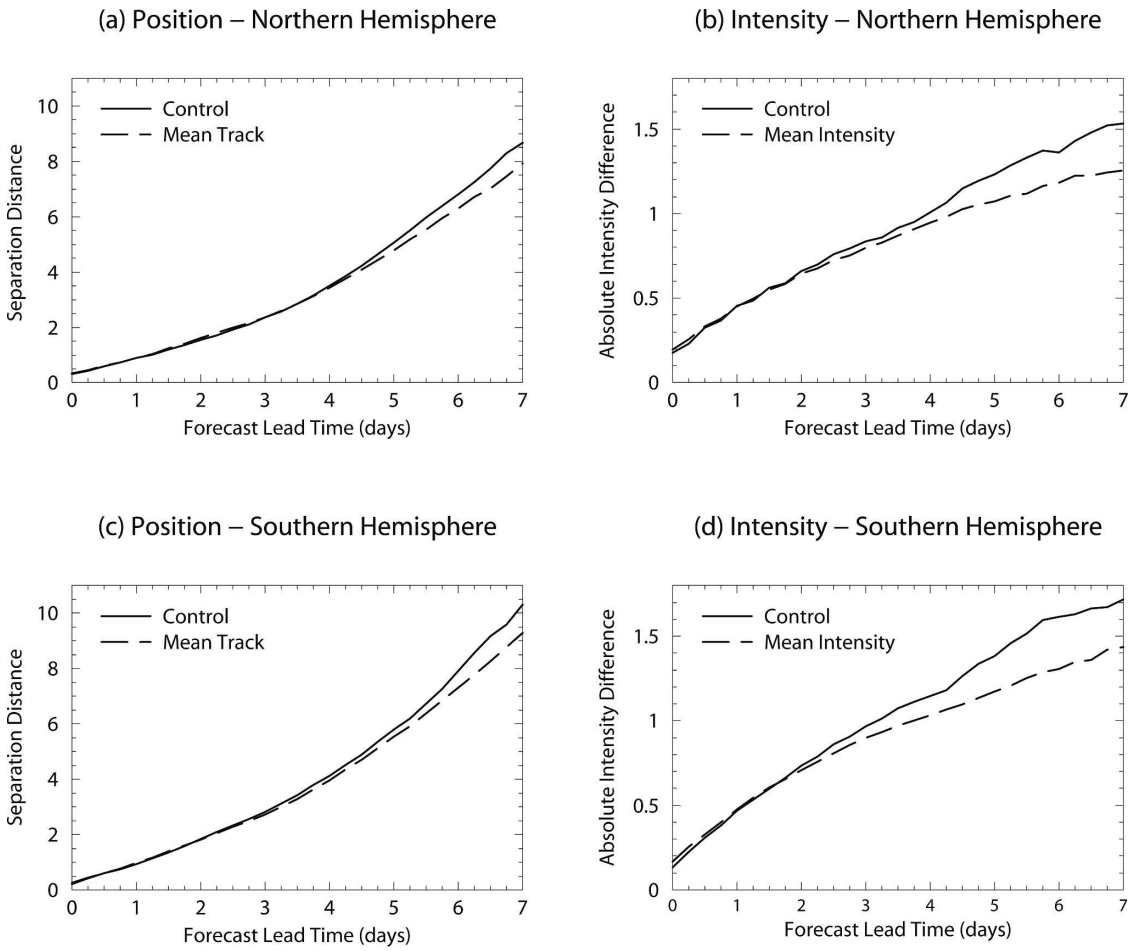


FIG. 8. Separation distance in the (a) NH and (c) SH, and absolute intensity difference in the (b) NH and (d) SH, between ECMWF ensemble mean tracks (computed from the matched perturbed ensemble members and control) and analysis tracks (solid lines) and ECMWF control forecast tracks and analysis tracks (dashed lines) as a function of forecast lead time. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal), respectively.

insufficient to obtain a clear signal of the bias. However, there does seem to be some correspondence between the speed bias and the intensity bias. It appears that an overprediction of cyclone intensity corresponds to a propagation speed that is too slow, whereas an underprediction of cyclone intensity corresponds to a propagation speed that is too fast. This seems feasible when the growth of a baroclinic disturbance is interpreted in terms of a pair of counterpropagating Rossby waves (Hoskins et al. 1985). This theory was first introduced by Bretherton (1966) with a two-layer model but has recently been extended to the primitive equations by Methven et al. (2005). The theory shows that the near-surface Rossby wave will propagate eastward more rapidly on its own than when it interacts with the upper-level Rossby wave. Once the lower wave couples with the upper wave, the phase locking will reduce the speed and the disturbance will intensify. Hence an overprediction of intensity would imply an underprediction of propagation speed and vice versa.

To investigate any bias in the track that the forecasted cyclones take, the cross-track error presented in section 4c was used. Cross-track errors were assigned positive (negative) values if the forecasted cyclone lay to the left (right) of the analyzed cyclone so that a bias could then be calculated (Figs. 10e,f). There is a small positive bias in the NH and small negative bias in the SH for both the ECMWF and NCEP EPS. This corresponds to a slight poleward bias in both hemispheres.

f. How far in advance can storms be predicted?

The statistics of Fig. 2 show that a proportion of cyclones are predicted by forecasts made before the ξ_{850} center has been identified in the analysis cycle ($M < 0$).

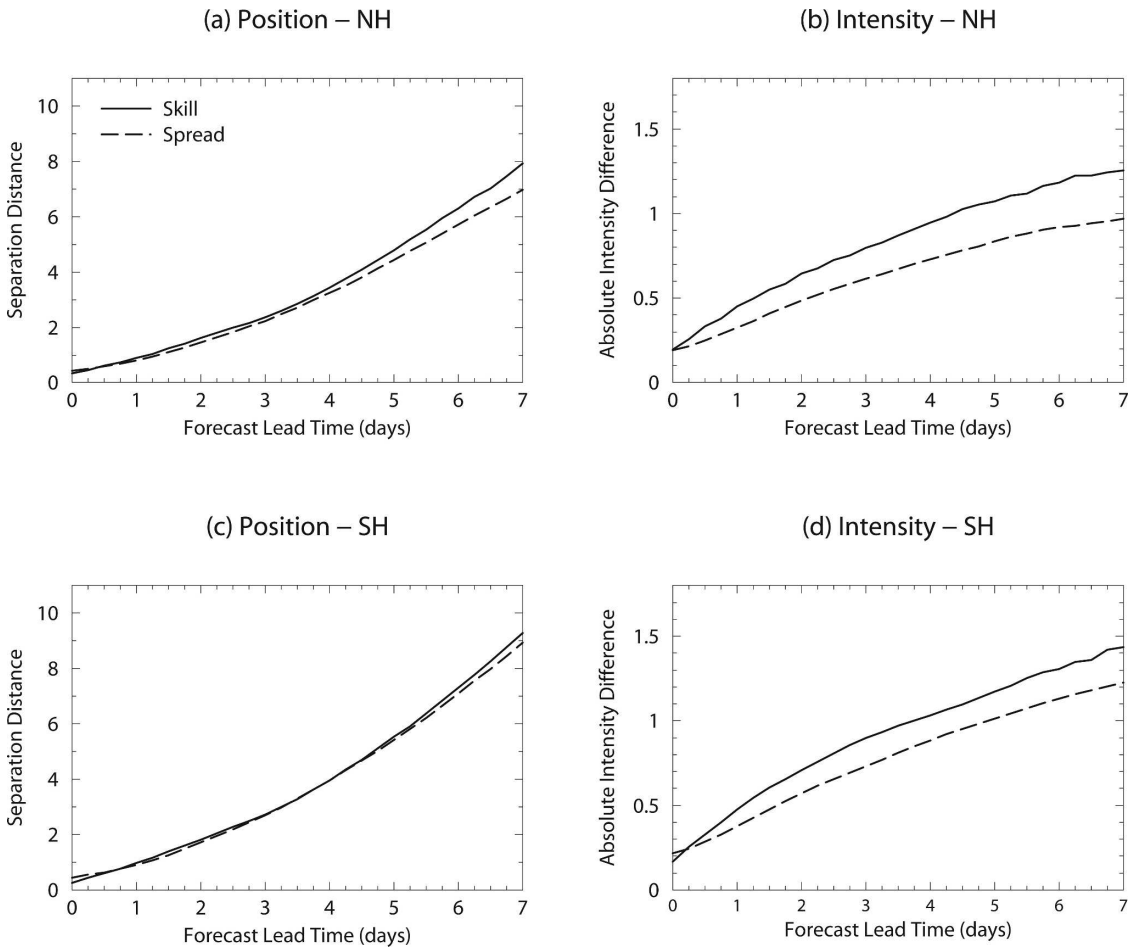


FIG. 9. Spread-skill diagnostics for ECMWF EPS. The solid lines show the skill and are the same as the dashed lines of Fig. 8. The dashed lines show spread calculated as the mean separation distance and mean absolute intensity difference between the matched perturbed ensemble member tracks and ensemble mean track. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal), respectively.

This is also illustrated by Figs. 1a,b, which show that 16 of the 50 ECMWF perturbed members predict the cyclone 1¾ days before the storm is first identified in the analysis. In FBH, statistics were generated to determine how far in advance of their identification in the analysis cycle extratropical cyclones could be predicted. This idea has been extended to EPS and the results are presented in this subsection.

The first time a cyclone is identified in the analysis is defined to be the time of the first point in the analysis track. It is therefore defined by the parameters used in the cyclone identification and tracking methodology, which requires that the vorticity center must exceed a magnitude of 1.0×10^{-5} s^{-1} (relative to the large-scale background field removal) to be considered a cyclone. Although the identification of a ξ_{850} center is not the only indication of a developing storm, it marks a specific stage of cyclone development, which can easily be identified in the analysis. To determine whether a cyclone was predicted N days before it was first identified in the analysis, the forecast storm tracks identified in the forecast made N days before were examined to see if any of them matched the analysis track. The constraint that the forecast tracks must begin within the first 3 days of the forecast (see section 3) was removed, so that the possibility of predicting storms up to 7 days before they are first identified in the analysis could be considered.

Figure 11 shows, for the ECMWF and NCEP EPS in both hemispheres, the percentage of analysis storms that are predicted by at least 1 perturbed member and by at least 10%, 20%, 50%, and 80% of perturbed members as a function of N. The percentage of analysis storms predicted by the control is also shown for the ECMWF EPS. Since the NCEP EPS has 10 members, the curves corresponding to 1 member and 10% of

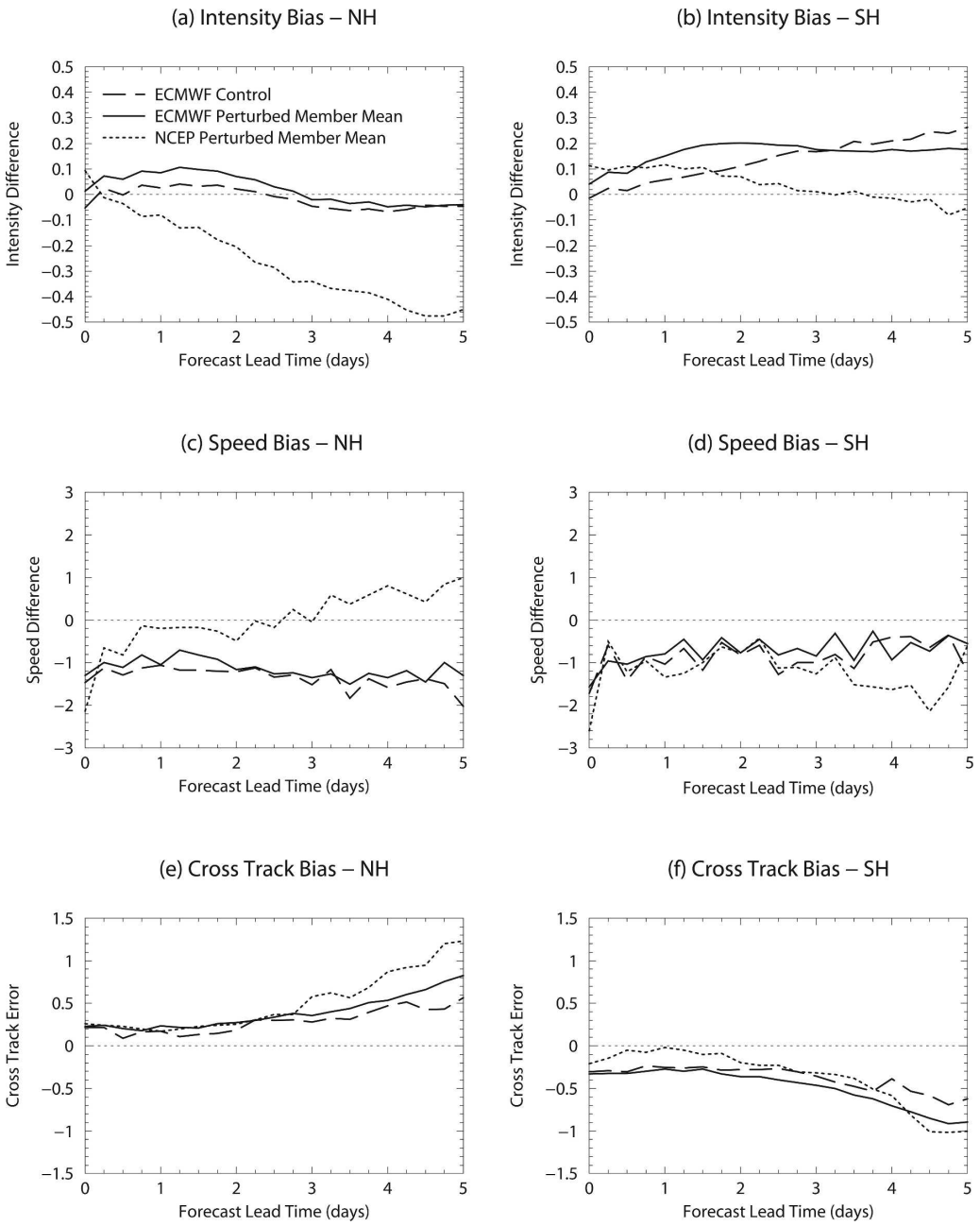


FIG. 10. Bias in intensity in (a) NH and (b) SH, propagation speed in (c) NH and (d) SH, and cross-track error in (e) NH and (f) SH of the perturbed member tracks for the ECMWF (solid lines) and NCEP (dotted lines) EPS. The biases are also shown for the ECMWF control forecast tracks (dashed lines). Units of intensity, speed, and cross-track error are 10^{-5} s^{-1} (relative to background field removal), kmh^{-1}, and geodesic degrees, respectively.

members are the same. The percentages will clearly depend on the choice of matching criteria; however, the general relationship between the different curves remains the same.

The results for the ECMWF EPS are discussed first. Comparing the percentage of cyclones predicted by the control forecast with the percentage predicted by at least 1 perturbed member illustrates how an EPS can extend the limit of predictability available from a single deterministic forecast. In the NH, approximately 20% of the cyclones are predicted by the control forecast 4 days before they have been identified in the analysis

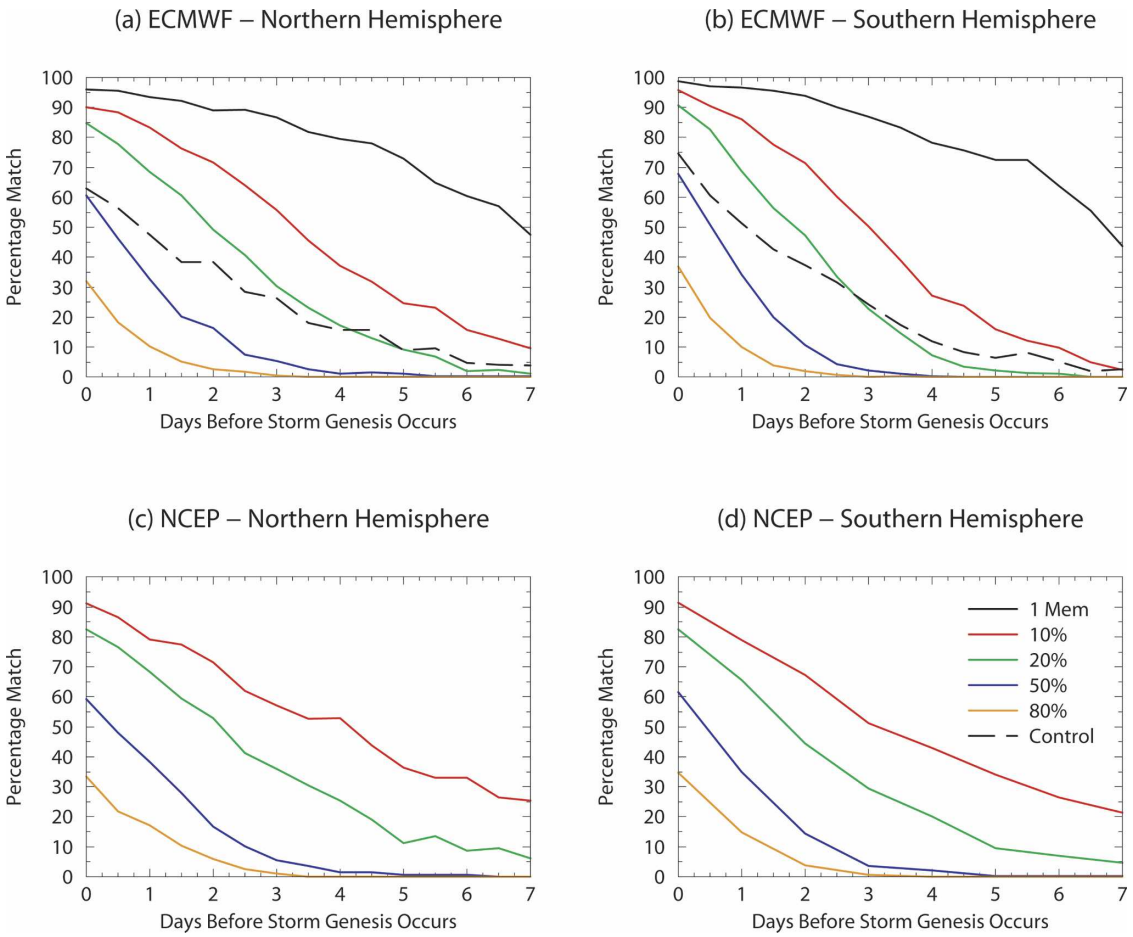


FIG. 11. The percentage of analysis tracks that are predicted by at least 1 perturbed member and by at least 10%, 20%, 50%, and 80% of perturbed members, as a function of the number of days N before the storm was first identified as an 850-hPa vorticity center in the analysis, for the ECMWF EPS in the (a) NH and (b) SH, and for the NCEP EPS in the (c) NH and (d) SH. The percentage of tracks predicted by the ECMWF control forecast is also shown by dashed lines. There are no solid black curves in the NCEP plots because 1 member is 10% of a 10-member ensemble and this is already shown by the red curve.

cycle, whereas about 80% are predicted by at least 1 of the perturbed ensemble members. It is very encouraging that a large number of cyclones can actually be predicted as much as 7 days before they are identified in the analysis cycle. However, the reliability of a prediction by one ensemble member needs to be considered, as it is possible that a significantly large number of false alarms also occur. The reliability aspect of cyclone prediction will be investigated in future work (see section 5 for further discussion).

When $N = 0$ the percentage of cyclones predicted by the control forecast is slightly higher than the percentage predicted by at least 50% of the perturbed members. However, as N increases the percentage predicted by 50% of perturbed members falls much faster than the percentage predicted by the control. This again highlights the superior quality of the control forecast to the perturbed ensemble members.

The percentages are higher in the SH than the NH when $N = 0$, but they fall faster in the SH as N increases and become lower than in the NH from $N = 3$ or 4. A similar difference between the hemispheres was also found in FBH. The reason why a larger number of cyclones are predicted in the NH than the SH when N is large could be because of the higher density of upper air observations available, particularly the radiosonde observations located in the upstream parts of the main storm-track regions. Extratropical cyclones are often initiated by perturbations in the upper-level flow. If such information is not available in the initial conditions of SH forecasts this could have a significant influence on the generation of a ξ_{850} center later in the

forecast. On the other hand, when cyclones are more developed in the initial conditions (i.e., N is small), the larger number that are predicted in the SH than in the NH (also seen in Fig. 2) could be due to the less varied surface boundary conditions of the SH discussed previously.

The most noticeable difference in the NCEP results is that the percentages fall less rapidly with increasing N. This is almost certainly because the NCEP forecasts are integrated to 16 days, whereas ECMWF forecasts are only integrated to 10. As N is increased, the temporal matching criterion (see section 3) is less likely to be satisfied by the ECMWF forecast tracks than the NCEP forecast tracks because the length of the tracks is limited. In future work we plan to repeat this analysis for the NCEP EPS with the forecasts truncated to 10 days. This might be considered a fairer comparison, but it could also be argued that the advantage the longer length of the NCEP forecasts has in giving early indications of future storms should be taken into consideration in the diagnostics. Unlike the results for the ECMWF EPS, the percentages in the NH and SH are comparable when $N = 0$, but the percentages do fall faster in the SH than in the NH.

5. Discussion and conclusions

This paper explores the prediction of extratropical cyclones by the ECMWF and NCEP EPS. The analysis methodology has enabled us to determine detailed and useful information about the prediction of the cyclones by the ECMWF EPS. We have currently not been able to produce a complete set of diagnostics for the NCEP EPS for a number of reasons, including the smaller number of ensemble members, the lower frequency of the lower-resolution control forecast, and the smaller size of our data sample. However, it has been possible to perform some preliminary comparisons of the NCEP EPS with the ECMWF EPS. This study is the first statistical analysis of the prediction of extratropical cyclones by EPS. A more complete comparison of the two systems, and other ensemble systems, will form the basis of future work. In particular we hope that the data provided via The Observing-System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) project will enable us to perform such a study (see http://www.wmo.ch/thorpex/pdf/tigge_summary.pdf for details of this project). The main results of the paper will now be discussed.

In general the ECMWF EPS has a higher level of skill in predicting the cyclones than the NCEP EPS. A higher percentage of ECMWF forecast storm tracks match with analysis storm tracks than NCEP forecast tracks in both hemispheres. The difference between the percentage of ECMWF and NCEP perturbed member tracks that match is comparable to the difference between the percentage of ECMWF and NCEP control forecast tracks that match, which suggests that the superior skill of the ECMWF EPS is due to the model and assimilation system rather than the perturbation methodology. Further analysis of the matched forecast tracks shows that in the NH the ECMWF perturbed ensemble members have slightly higher skill than the NCEP perturbed ensemble members for the prediction of both the position and intensity of the storms. However, in the SH the NCEP perturbed ensemble members have slightly more skill, particularly for the intensity. This is perhaps because the lower resolution of the NCEP system causes more difficulties in the prediction of the cyclones in the NH than in the SH. NH storms are influenced more by land–sea contrasts and changeable orography than SH storms and a higher resolution model may therefore be more important in the NH than in the SH. However, the large number of differences between the ECMWF and NCEP EPS makes it difficult to determine the exact causes of the differences in predictive skill in the two hemispheres. The ECMWF and NCEP EPS use completely different models, resolutions, data assimilation schemes, and perturbation methodologies. Indeed one of the main conclusions of the Buizza et al. (2005) study was the difficulty in comparing such different systems and that further studies, in which different perturbation methods are explored using a single analysis and forecast system, are necessary to determine the advantages and disadvantages of the different perturbation methodologies.

Both the ECMWF ensemble mean track and intensity have a higher level of skill than the control forecast from around day 3 of the forecast, but the difference is more significant for the intensity than position. The skill of the ensemble mean and control forecast is higher in the NH than the SH for both the position and intensity of the cyclones. This confirms the results of FBH, which showed that the high quantity of upper air observations available in the NH significantly improved the prediction of the storms. It is difficult to tell whether this is the case for the NCEP EPS, since we unfortunately have insufficient data to generate diagnostics for the ensemble mean or the control forecast. The diagnostics do, however, show that the perturbed ensemble members have slightly higher skill in the NH than in the SH, but the difference in skill is less than that of the ECMWF perturbed members.

The difference between the ensemble mean and spread of the ECMWF EPS is very small for the position of the cyclones. There is a larger spread in the SH than the NH, which corresponds well with the larger

error of the mean track. The difference between ensemble mean and spread is larger for the intensity of the cyclones. This shows that the analyzed intensities do not lie in the center of the envelope of ensemble member intensities on average and may possibly imply that the analyzed intensities of the cyclones are statistically different than the intensities predicted by the ensemble members. It reinforces the result that there is a higher level of skill in the prediction of cyclone position than intensity, which is suggested by the difference in the error growth rates.

In FBH it was suggested that the lower level of predictive skill for the intensity of cyclones was due to errors in the storm's vertical structure. The fact that the along-track error is a larger contributor to the total position error than the cross-track error (Fig. 7) also suggests that this is the case. Both the intensity and propagation speed of storms depends on the interaction of upper- and lower-level disturbances (see section 4e). Errors in the storm's vertical structure will therefore result in errors in both intensity and propagation speed. Since the along-track error is a result of propagation speed error, it will be affected by errors in the storm's vertical structure. The cross-track error, however, is caused by errors in the direction the storm propagates, which will be determined mainly by the 700-hPa wind field and less affected by errors in the vertical structure. Errors occurring in the vertical structure of forecast cyclones will be investigated in future work.

The low level of predictive skill for the intensity of cyclones is a trait common to both the ECMWF and NCEP models. It is possible that a higher resolution and/or frequency of upper air observations than currently available would be required to accurately predict the growth and development of extratropical cyclones. The use of targeted observations (e.g., Leutbecher et al. 2002) may improve the prediction of cyclone intensity considerably.

Another possible way of decreasing the difference between the ECMWF ensemble mean and spread for cyclone intensity would be the inclusion of moist physics in the computation of the extratropical singular vectors. This is currently being investigated at ECMWF. Coutinho et al. (2004) analyzed the impact that the moist physics had on the singular vectors. They found that the large-scale latent heat release led to larger growth and smaller horizontal scales. As a continuation of this study, Hoskins and Coutinho (2005) applied singular vector perturbations computed with both dry and moist physics to some extreme European cyclones. The singular vectors computed with moist physics were found to be much more relevant to the development of these intense cyclones than those computed without moist processes. Forecasts integrated from initial conditions perturbed with moist singular vectors were produced, and for each cyclone one of the perturbed forecasts was superior to the control forecast. If the growth of the cyclones is better represented by moist singular vectors then this could potentially decrease the difference between the ensemble mean and spread. The impact of moist singular vector perturbations on the prediction of extratropical cyclones by the ECMWF EPS will be investigated in future work.

The propagation speed of the cyclones predicted by both the perturbed members and the control forecast of the ECMWF is on average too slow, although the small magnitude of this bias should be noted. The results for the NCEP EPS in the SH also indicate that the cyclones are moving too slowly, but in the NH the results are less clear. A further result that is possibly related is that the NH cyclones predicted by the NCEP EPS are generally underpredicted, whereas the intensity of cyclones predicted by the NCEP EPS in the SH and those of the ECMWF EPS in both hemispheres is, if anything, overpredicted. It is possible that the resolution of the NCEP EPS is too low to accurately model the fast growth of NH storms. However, these results should be considered preliminary because of the smaller amount of NCEP data we currently have available. There is also a small bias in both the ECMWF and NCEP EPS for the forecasted cyclones to move too far toward the Poles.

The ECMWF control forecast has a consistently higher level of skill (½–1 day) than the perturbed members, throughout the first 7 days of the forecast, for both the position and intensity of the cyclones. It is to be expected that the control forecast is better than the perturbed members in the earlier part of the forecast, since it has been optimized to best fit available observations via 4DVAR. For an EPS constructed from initial condition perturbations alone, it would perhaps be expected that the error of the control forecast would converge to that of the perturbed members at higher lead times. However, the model perturbations included in the ECWMF EPS may cause the control forecast to have an advantage over the perturbed ensemble members at higher forecast lead times. It would be very interesting to compare the predictive skill of the perturbed ensemble members of other EPS with their control forecasts.

A very encouraging result is the very high level of skill provided by the best ensemble member for both the ECMWF and NCEP EPS. We remind the reader, however, that a rather low number of ensemble members that are best in terms of the position of the cyclones are also best in terms of their amplitude. It may therefore be better to use some type of combined mea-

sure of position and intensity to determine the best track. The difficulty with the best ensemble member is how to identify it at some useful time (i.e., before the final validation time of the forecast has past). Preliminary analysis suggests that this is very difficult; the error of the ensemble member, which is best for the first day or 2 of the forecast, rapidly approaches the error of the average ensemble member. However, the high skill of the best ensemble member should still be considered encouraging, since it suggests that the errors in the initial state are being effectively sampled.

Another encouraging result is the potential for EPS to provide early indications of cyclones. The results show that an indication by at least 1 ensemble member of a majority of cyclones can be given 7 days before they have been identified (as ξ_{850} centers) in the analysis cycle. An important question that needs to be considered is the reliability of such an indication by one ensemble member. It is possible that a large number of incorrect predictions (false alarms) also occur. To address this issue, probabilistic scoring methodologies, such as the Brier skill score (Brier 1950), will need to be invoked and will be investigated in future work. The interesting point about the diagnostics of Fig. 11 is that they show the potential of ensemble forecasting to extend the limit of predictability of a single deterministic forecast. It would be interesting to see how the percentage of cyclones predicted by at least one ensemble member would change as the number of ensemble members was increased. The reliability of the prediction of a storm by 1 ensemble member is clearly questionable. However, once the ensemble size approaches the limit at which 100% of cyclones are predicted by at least 1 member, it could be argued that the probability density function of the forecasts' states can be sufficiently estimated. It is therefore suggested that the diagnostics of the figure could potentially provide a useful guide in determining how many members an ensemble forecast should have.

The results presented in this paper could potentially be useful to both the developers of ensemble forecast systems and to the users of these systems. A developer could use the statistical approach of this paper to evaluate the impact of future upgrades to an ensemble forecast system. The user could use the storm-tracking method to help with the forecasting of individual storms. Modifications to the methodology presented in this paper would be required, such as matching ensemble member tracks with control forecast tracks to obtain measures of ensemble mean and spread for ensemble forecasts of individual storms (discussed in section 4d). The user should also be influenced by the results of the statistical analysis of the forecast developer.

Acknowledgments. Thanks go to ECMWF and NCEP for making the EPS forecast and analysis data available to us and to Brian Hoskins for discussions about storm dynamics. We would also like to acknowledge the reviewers for their helpful comments, which improved this paper.

REFERENCES

Barkmeijer, J., R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **125,** 2333–2351.

Bengtsson, L., K. I. Hodges, and L. S. R. Froude, 2005: Global observations and forecast skill. *Tellus,* **57A,** 515–527.

Bretherton, F. P., 1966: Baroclinic instability and the short wavelength cut-off in terms of potential vorticity. *Quart. J. Roy. Meteor. Soc.,* **92,** 335–345.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

Buizza, R., and T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.,* **52,** 1434–1456.

——, and P. Chessa, 2002: Prediction of the U.S. storm of 24–26 January 2000 with the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **130,** 1531–1551.

——, and A. Hollingsworth, 2002: Storm prediction over Europe using the ECMWF ensemble prediction system. *Meteor. Appl.,* **9,** 289–305.

——, M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **125,** 2887–2908.

——, P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.,* **133,** 1076–1097.

Coutinho, M. M., B. J. Hoskins, and R. Buizza, 2004: The influence of physical processes on extratropical singular vectors. *J. Atmos. Sci.,* **61,** 195–209.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.,* **129,** 2461–2480.

Froude, L. S. R., L. Bengtsson, and K. I. Hodges, 2007: The predictability of extratropical storm tracks and the sensitivity of their prediction to the observing system. *Mon. Wea. Rev.,* **135,** 315–333.

Hamill, T. M., C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.

Hartmann, D. L., R. Buizza, and T. N. Palmer, 1995: Singular vectors: The effect of spatial scale on linear growth of disturbances. *J. Atmos. Sci.,* **52,** 3885–3894.

Haseler, J., 2004: Early-delivery suite. ECMWF Tech. Memo. TM 454, 35 pp.

Hodges, K. I., 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.,* **123,** 3458–3465.

——, 1999: Adaptive constraints for feature tracking. *Mon. Wea. Rev.,* **127,** 1362–1373.

Hoskins, B. J., and K. I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *J. Atmos. Sci.,* **59,** 1041–1061.

——, and M. M. Coutinho, 2005: Moist singular vectors and the predictability of some high impact European cyclones. *Quart. J. Roy. Meteor. Soc.,* **131,** 581–601.

——, and K. I. Hodges, 2005: A new perspective on Southern Hemisphere storm tracks. *J. Climate,* **18,** 4108–4129.

——, M. E. McIntyre, and A. W. Robertson, 1985: On the use and significance of isentropic potential vorticity maps. *Quart. J. Roy. Meteor. Soc.,* **111,** 877–946.

——, R. Buizza, and J. Badger, 2000: The nature of singular vector growth and structure. *Quart. J. Roy. Meteor. Soc.,* **126,** 1565–1580.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Jung, T., E. Klinker, and S. Uppala, 2005: Reanalysis and reforecast of three major European storms of the twentieth century using the ECMWF forecasting system. Part II: Ensemble forecasts. *Meteor. Appl.,* **12,** 111–122.

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

Leutbecher, M., J. Barkmeijer, T. N. Palmer, and A. J. Thorpe, 2002: Potential improvement to forecasts of two severe storms using targeted observations. *Quart. J. Roy. Meteor. Soc.,* **128,** 1641–1670.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.,* **128,** 2145–2166.

Methven, J., E. Heifetz, B. J. Hoskins, and C. H. Bishop, 2005: The counter-propagating Rossby wave perspective on baroclinic instability. Part III: Primitive equation disturbances on the sphere. *Quart. J. Roy. Meteor. Soc.,* **131,** 1393–1424.

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.

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.,* **120,** 1747–1763.

Simmons, A. J., and J. K. Gibson, 2000: The ERA-40 project plan. ECMWF Reanalysis Rep. Series 1, 63 pp.

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

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Wei, M., and Z. Toth, 2003: A new measure of ensemble performance: Perturbation versus error correlation analysis (PECA). *Mon. Wea. Rev.,* **131,** 1549–1565.