

The Predictability of Extratropical Storm Tracks and the Sensitivity of their Prediction to the Observing System

Lizzie S. R. Froude^{*}, Lennart Bengtsson and Kevin I. Hodges

Environmental Systems Science Centre (ESSC), University of Reading, Harry Pitt
Building, Whiteknights, PO Box 238, Reading, RG6 6AL, UK

* Corresponding Author:

Email: lsrf@mail.nerc-essc.ac.uk

Abstract

A new method for assessing forecast skill and predictability, which involves the identification and tracking of extratropical cyclones, has been developed and implemented to obtain detailed information about the prediction of cyclones that cannot be obtained from more conventional analysis methodologies. The cyclones were identified and tracked along the forecast trajectories and statistics were generated to determine the rate at which the position and intensity of the forecasted storms diverges from the analysed tracks as a function of lead time. The results show a higher level of skill in predicting the position of extratropical cyclones than the intensity. They also show that there is potential to improve the skill in predicting the position by 1-1½ days and the intensity by 2-3 days, via improvements to the forecast model. Further analysis shows that forecasted storms move at a slower speed than analysed storms on average and that there is a larger error in the predicted amplitudes of intense storms than the weaker storms. The results also show that some storms can be predicted up to 3 days before they are identified as an 850 hPa vorticity center in the analyses. In general the results show a higher level of skill in the northern hemisphere (NH) than the southern hemisphere (SH); however, the rapid growth of NH winter storms is not very well predicted. The impact that observations of different types have on the prediction of the extratropical cyclones has also been explored, using forecasts integrated from analyses that were constructed from reduced observing systems. A terrestrial, satellite and surface based system were investigated and the results showed that the predictive skill of the terrestrial system was superior to the satellite system in the NH. Further analysis showed that the satellite system was not very good at predicting the growth of the storms. In the SH the terrestrial system has significantly less skill than the satellite system highlighting the dominance of

satellite observations in this hemisphere. The surface system has very poor predictive skill in both hemispheres.

1 Introduction

Extratropical cyclones are important constituents of the general circulation of the atmosphere and are important for the day to day weather of the extratropics via their presence or absence. They can be both beneficial, in that they bring most of the rainfall required to sustain human activities such as agriculture, and destructive through excessive rainfall leading to floods and damaging winds. It is therefore important that these storms are predicted as accurately as possible, by numerical weather prediction (NWP), to provide the best estimates of their locations and intensities. This paper introduces a new approach to forecast verification, which focuses directly on these extratropical cyclones.

Current operational NWP systems combine observations from disparate sources with a previously obtained model atmospheric state via data assimilation to provide the initial conditions from which a new model forecast can be made. Reducing the errors in the initial conditions and improving the models will hopefully lead to better forecasts useful for longer forecast lead times. For example Simmons and Hollingsworth (2002) showed, using the European Centre for Medium Range Weather Forecasts (ECMWF) operational analyses and forecasts, that large scale forecast skill has improved by 3-4 days in the last 25 years. In the northern hemisphere (NH) this improvement can mainly be attributed to improvements in the models and data-assimilation systems. In the southern hemisphere (SH) forecasts now have almost the same level of skill as those in the NH, but this is mainly due to the introduction of satellite observations covering those areas where terrestrial observations are sparse.

Conventional methods for estimating forecast skill involve the use of root mean square (RMS) error or anomaly correlation to verify forecasts of varying lead times against the corresponding analyses. Often the 500mb geopotential height is chosen as a representative field (for example see Simmons and Hollingsworth 2002), but other smaller scale fields such as vorticity or precipitation can also be used. In this paper a new and alternative method for assessing forecast skill has been designed and implemented, which focuses on the extratropical cyclones. It involves the identification and tracking of extratropical cyclones along forecast trajectories to produce a set of forecast storm tracks (we note that here the term storm track refers to an individual storm trajectory, rather than the average track many storms pass along for which the term is also often used). The tracking is also performed on the analyses, from which the forecasts are initialized, to produce a set of analysis tracks to use for verification. Statistics can then be generated to determine the rate at which the position and intensity of the forecasted storms diverge from the analysed tracks with lead time. Diagnostics for other storm attributes, such as their growth rates and speeds, can also be produced.

This storm tracking methodology gives very different information to that obtained from the traditional 500 hPa geopotential height RMS and anomaly correlation approaches. Whilst these Eulerian based methods provide information about the prediction of general weather patterns, the new analysis methodology gives direct information about the prediction of extratropical cyclones and therefore may provide a better measure of the models ability to predict the weather. As with any analysis methodology the method is not perfect and does have some limitations and biases that should be taken into consideration. These issues are discussed as they arise throughout the text and summarised in the discussion section at the end of the paper.

This paper has two main objectives. The first is to use the storm tracking methodology to explore the predictability of extratropical storms. The second is to confirm and extend the results of a recent study by Bengtsson et al. (2005), which investigated the impact that different types of observation have on forecast skill. The Bengtsson et al. (2005) study followed on from a previous study by Bengtsson et al. (2004b), which explored the sensitivity of the analyses to the observing system. In this study the ECMWF 40 year Re-Analysis System (ERA40) (Simmons and Gibson 2000) was used to construct analyses from different observing systems. Using both Eulerian and feature tracking diagnostics they showed the equal importance to the analyses of the terrestrial and satellite observations in the NH (the terrestrial system becomes more important at the smaller synoptic scales) and the dominance of the satellite data in the SH. The surface observations were shown to provide limited information on their own in the NH and virtually none in the SH.

Bengtsson et al. (2005) used the analyses from the Bengtsson et al. (2004b) study to generate forecasts to determine the impact observations of different types have on forecast skill. They showed that this was in accord with the relative importance of the different observing systems to the analyses. They also considered the predictability as determined by the Lorenz (1982) approach and showed that there was scope to improve the forecasts by a day or more by improving the initial state or forecast model. Their study used both the RMS approach and a storm tracking approach; however, the storm tracking methodology, which differed to the new method used in this paper, had some problems and limitations meaning that the conclusions that could be drawn were somewhat tentative (see section 2 for more details). The method used in this paper does not suffer from these limitations and provides far more detailed information about the prediction of extratropical cyclones.

There have been other studies concerning the prediction of extratropical cyclones. For example Xiao et al. (2002) investigated the impact that satellite derived winds have on the prediction of a mid-Pacific cyclone. They found that the satellite wind observations increased the cyclonic zonal wind shear and cross-front temperature gradient associated with the cyclone and consequently improved the predicted position and intensity of the cyclone. In a recent study Zhu and Thorpe (2005) investigated forecast error growth, due to errors in the initial conditions and model deficiencies, by following the development of an extratropical cyclone in a simulation obtained by applying upper level potential vorticity (PV) perturbations to an idealized two-dimensional baroclinic jet initial state. There have been many other studies of a similar vein to these, studying the prediction of individual cyclones or cyclone simulations, but this paper is the first to provide a statistical measure of the forecast skill and predictability of extratropical cyclones from an operational system.

The method of tracking cyclones along forecast trajectories is currently used, as a verification tool for tropical cyclone forecasting, by some of the operational meteorological centers (for example ECMWF (Van der Grijn 2002), the National Center for Environmental Prediction (Marchok 2002) and the UK Meteorological Office (Heming 1994)). These centers mainly use this method to study tropical cyclones on a case-by-case basis, producing diagnostics such as strike probability maps (see Van der Grijn 2002), rather than a statistical analysis of a large number of storms. Performing a statistical study of the prediction of tropical cyclones would require a considerably larger time period of data than for extratropical cyclones, because of the comparatively small number of tropical cyclones that occur in a selected time period. Such a study may however constitute future work.

The paper continues with a discussion of the data and how they were generated in section 2, a description of the analysis methodology is given in section 3 and the results are presented in section 4. A final discussion and conclusions are given in section 5.

2 Data Description

The analysis data for the different observing systems used in this study and in the previous Bengtsson et al. (2004b, 2005) studies were generated using the ERA40 reanalysis system (Simmons and Gibson 2000). This is a three-dimensional variational (3D Var) data assimilation system, which uses version 23R4 of the ECMWF operational Integrated Forecast Model (IFS) (White 2000). The model is spectral semi-Lagrangian with a resolution of triangular truncation 159 with 60 levels in the vertical (T159L60). The system includes terrestrial observations of temperature, pressure, wind and humidity together with a significant quantity of satellite observations. Satellite radiance data from the TIROS Operational Vertical Sounder (TOVS) are assimilated directly, rather than assimilating retrievals of temperature and humidity. Other derived satellite quantities such as cloud motion winds are also assimilated.

Different types of observing system were constructed by systematically removing observations from the ERA40 observation database and then re-running the data assimilation. The control system consisted of all the observations used in ERA40 apart from humidity, since these observations were found to have very limited impact on the quality of the analyses (Bengtsson et al. 2004a; Bengtsson and Hodges 2005b). Three different observing systems were considered: a terrestrial system, a satellite system and a surface system. The terrestrial system was obtained by removing all the

satellite observations, the satellite system was obtained by removing all the terrestrial observations apart from surface pressure and the surface system consisted of just surface observations.

The predictability study of Bengtsson et al. (2005) used these analyses as initial states to produce forecasts. They used a later and further improved version (26R3) of the forecast model than that used for the data assimilation (Bengtsson et al. 2004b), but it was integrated at the same horizontal and vertical resolution of T159L60. Figure 1(a) illustrates the experiment setup for the 1st December 1990 – 28th February 1991 season used in this previous study, where each cross in the diagram represents one time frame of data. The forecast model was run from each 6 hourly analysis out to 7 days (the diagonal dashed lines in the figure) and selected fields were archived daily. The data generated were combined to form 7 three month forecast datasets corresponding to the solid horizontal lines in the diagram. Extratropical cyclones were then identified and tracked along each of these 7 forecast datasets. The problem with this method was that each time step, within a forecast dataset, was generated from different initial conditions. This led to some difficulty in tracking individual cyclones through the forecast datasets of higher lead times and resulted in a probable underestimate of the predictive skill and predictability of storm tracks. This previous approach was also limited in terms of the diagnostics that could be produced.

For this study the forecasts have been re-run from each 6 hour analysis out to 14 days using the same model and resolution, but archiving selected fields every 6 hours of the forecast. This allowed the extratropical cyclones to be tracked along the forecast trajectories (see section 3). The analyses and forecasts were produced, for all the observing systems, for the selected seasonal periods of 1st December 1990 – 28th February 1991 and 1st December 2000 – 28th February 2001. They were also

generated, for just the control system, for the 1st June 1991 – 31st August 1991 and 1st June 2000 – 31st August 2000 time periods.

3 Analysis Methodology

3.1 Storm Tracking Methodology

The extratropical cyclones were identified and tracked using the method of Hodges (1995, 1999). This method has been used extensively in other studies of extratropical cyclones (e.g. Bengtsson et al. 2004b; Hoskins and Hodges 2002, 2005). 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 (for further details see Hoskins and Hodges (2002, 2005)). Initially the identification and tracking was performed with both the 850 hPa relative vorticity (ξ_{850}) and Mean Sea Level Pressure (MSLP) fields, but since the results for each field were very similar this paper focuses on the ξ_{850} field. Vorticity features with a magnitude exceeding $1.0 \times 10^{-5} \text{ s}^{-1}$ were identified as maxima in the NH and minima in the SH and considered as 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, travelled further than 1000 km and had a majority of their lifecycle in $20^\circ\text{N} - 90^\circ\text{N}$ or $20^\circ\text{S} - 90^\circ\text{S}$ were retained for the statistical analysis.

Figure 1(b) shows the experiment set-up used for this study for the 1st December 1990 – 28th February 1991 season, where the forecast datasets (described above) are represented by the diagonal lines. The cyclones were identified and tracked through

each of these datasets to obtain a total of 360 ensembles of storm tracks. This procedure was performed for each observing system for the DJF (December January February) seasons and additionally for the control system for the JJA (June July August) seasons. The control analysis storm tracks used in Bengtsson et al. (2004b) were used to verify the forecast storm tracks. These were obtained by identifying and tracking cyclones through the analysis time steps for each 3 month time period (illustrated by the horizontal line in Figure 1(b)). To help set the scene and illustrate the storm tracking we present an example of a storm track in the following section.

3.2 Example Storm Track

Figure 2 shows an example of an intense fast moving storm identified in the control analysis. The analysis shows the storm initially developing (as an 850 hPa vorticity center) just off the northeast coast of North America on the 6th January 1991 at 12UTC. Inspection of the upper troposphere fields suggests that the storm was initiated by an upper level disturbance. The storm travelled across the Atlantic, gradually intensifying until it reached its peak of $7.5 \times 10^{-5} \text{ s}^{-1}$ over the UK on the 9th January at 06UTC. It then moved over Scandinavia and into northwest Russia, weakening over the next 2 days.

Figure 2(a) shows the track of the analysed storm and the track predicted by the control forecast beginning from 5th January 12UTC, 1 day before the storm was first identified in the 850 hPa analysis, and 2(b) shows the intensity of the analysed and predicted storm. The predicted storm is generated 6 hours (6th January 18UTC) after the analysed one at a lead time of 1.25 days. At a lead time of 4.25 days the track is cut short, because a double center is generated earlier in the forecast. One of these centers is tracked until the 4.25 day lead time, but after this point the cyclone becomes a single center again causing a discontinuity and truncated track. The track of the

storm is predicted very well, following the analysis track almost exactly, but at a slightly slower speed. The intensity of the storm is not as well predicted, it reaches its peak at the correct time, but it consistently underpredicts by up to $1.5 \times 10^{-5} \text{ s}^{-1}$.

Figure 2(c) and (d) show the track and intensity of the analysed storm and the storm predicted by the control forecast beginning from 6th January 12UTC, when the storm was first identified in the analyses. Again the track is predicted very well; the forecast follows the analysis closely until the day 3 lead time where it begins to deviate, moving south from the analysed track. The intensity is poorly predicted, worse in fact than the earlier forecast. The predicted storm intensifies at a much faster rate than the analysis, reaching a peak of $10.5 \times 10^{-5} \text{ s}^{-1}$, which is $3.0 \times 10^{-5} \text{ s}^{-1}$ higher than the peak of the analysed storm. It then decays much faster than the analysis, dropping below the analysis after the day 3 lead time.

Figure 2(e) and (f) show the tracks and intensities predicted by the terrestrial, satellite and surface forecasts beginning from 6th January 12UTC. The terrestrial forecast is similar to the control forecast predicting the track of the storm very well, but moving at a slightly slower speed than the analysed storm. Like the control forecast it also overpredicts the intensity, but to a lesser extent than that of the control forecast. The satellite system does not predict the storm as well; although it appears to predict the first part of the track well, it moves at a significantly slower speed than the analysed storm and is cut short at the day 3.25 lead time. Whereas the analysed storm lies over Norway at the day 3 lead time, the storm predicted by the satellite system is situated further upstream over Ireland. A feature that is almost certainly related is the underprediction of the storms amplitude; the predicted cyclone does not even begin to deepen until the day 2 lead time. In this particular example the satellite observations appear to actually degrade the forecast. Both the track and intensity of

the storm are predicted very poorly by the surface system. Although the track of the storm is predicted well for the first 1.5 days, from this point onwards it moves away from the track of the analysed storm, in completely the wrong direction, curving back up towards Greenland. The amplitude of the storm is extremely overpredicted. By inspection of figure 2(f) it can be seen that the initial amplitude of the cyclone differs significantly to the other observing systems, which probably contributes to the low quality forecast. The surface system does not include any upper air observations, which will significantly affect the forecast, especially since the storm appears to have been initiated from an upper level disturbance.

In this example the track of the storm is predicted very well by the control forecast, but the intensity is not, with an under prediction in the earlier forecast and an over prediction in the later one. The terrestrial system shows a higher level of predictive skill than the satellite system and the surface system shows a very low level of skill. The forecasted storms all move at a slower speed than the analysed one. One of the aims of this study is to determine whether this behavior is statistically typical. In the next section the method used to generate the statistics is described.

3.3 Validation of Storm Tracks

The forecast storm tracks were validated against the control analysis storm tracks using a matching methodology similar to that used by Bengtsson et al. (2004b, 2005). A forecast storm track was considered to be the same system as an analysis storm track (i.e. matched) if the two tracks satisfied certain predefined spatial and temporal criteria. The forecast storm tracks that matched control analysis storm tracks were then used to produce diagnostics to quantify the error in the predicted positions, amplitudes and other properties of the cyclones.

The temporal criteria used in this study are the same as that used previously in the Bengtsson et al. (2005) study, but the spatial matching criteria are different. The spatial matching of the previous study looked at the separation distance between two tracks over their whole lifetime. Since the storm tracks in this study were computed along the forecast trajectory they may begin very close to the corresponding tracks in the analysis but will probably diverge as the lead time increases. The spatial matching in this study therefore focuses on the first section of the forecast track rather than the whole track.

A forecast track was said to match an analysis track if

(i) At least $T\%$ of their points overlapped in time, i.e. $100 \times \left(\frac{2n_m}{n_A + n_F} \right) \geq T$ 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 that match.

(ii) 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° , i.e. $d \leq S^\circ$.

The geodesic separation distance between two points A and B on the earth (assumed to be a perfect sphere) is calculated by $\cos^{-1}(\mathbf{P}_A \cdot \mathbf{P}_B)$ 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 kilometres at any latitude. The distance in km can be obtained by simply multiplying by the radius of the earth.

Figure 3 shows a schematic of the spatial matching when $k=4$. The solid line represents an analysis track and the dashed lines correspond to forecast tracks. Tracks A , B and C would match the analysis track because their first 4 points (that overlap in time with the analysis track) are less than S° from the corresponding analysis points. Track D would not match, even though it is less than S° from the analysis by the 9th time step, because the separation distance is greater than S° initially.

Clearly the number of storm tracks in the forecasts that match with tracks in the analysis will depend on the values chosen for k , T and S . The diagnostics produced from the matched tracks may also be affected by the choice of these values. To determine how sensitive the diagnostics produced from the matched tracks are to the matching criteria we have explored 6 different criteria, which are listed below:

- (i) $k = 4, T = 60\%$ and $S = 2^\circ$
- (ii) $k = 4, T = 60\%$ and $S = 4^\circ$
- (iii) $k = 4, T = 30\%$ and $S = 4^\circ$
- (iv) $k = 1, T = 60\%$ and $S = 2^\circ$
- (v) $k = 1, T = 60\%$ and $S = 4^\circ$
- (vi) $k = 1, T = 30\%$ and $S = 4^\circ$

Since the first 3 criteria place a spatial restriction on 4 points of the forecast storm tracks, we thought this may cause some bias in the diagnostics concerning the position of the storms. We therefore also considered the same criteria, but with $k=1$ (criteria (iv)-(vi)). As an additional constraint, only those storms whose genesis occurs within the first 3 days of the 14-day forecast or that already existed at time 0 were considered. Results from Bengtsson et al. (2005) indicated that the skill in predicting storm tracks after 3 days is relatively low. If a storm is generated in a forecast at a

lead time greater than 3 days, and matches a storm in the analysis, then it is probably more due to chance than an accurate prediction.

Table 1 shows the percentage of control forecast storm tracks that satisfy the genesis constraint just discussed and match with control analysis tracks for each of the different matching criteria for the combined DJF seasons in the NH. The percentages increase steadily as the values of S and T are relaxed. More forecast tracks are matched when $k=1$ than when $k=4$. It is clear that the matching criteria will significantly affect the number of forecast tracks that are matched. However, it is the impact that the different criteria have on the diagnostics produced from the matched tracks that we are mainly interested in.

Figure 4(a) shows the mean geodesic separation distance between the matched control forecast tracks and corresponding analysis tracks, obtained with each of the different matching criteria, as a function of forecast lead time. Figure 4(b) shows the mean absolute intensity difference. It can be seen clearly from the figure that the choice of matching criteria makes hardly any difference to the diagnostics obtained with the matched tracks. In particular the figure shows that the spatial matching criterion has very little impact on the diagnostics concerning the position of the cyclones.

It is possible that a forecasted and analysed storm could have very similar tracks, but move at different speeds. For example in figure 2 the forecasted storms mainly follow the track of the analysed storm, but they move at a slower speed. If there is a large difference between the speed an analysis and forecast track move then they may not satisfy the spatial matching condition; we therefore also considered an alternative spatial matching methodology. Rather than comparing the position of an analysed storm with the position of a forecasted storm that is valid at the same time, the

position of a given point on a forecast track was compared with the position of the point on the analysed track that is perpendicular to it (using spherical geometry). Whilst this did have an impact on the number of matched tracks, it made no noticeable difference to the diagnostics obtained from the matched tracks.

Since the choice matching criterion has no significant impact on the error growth rates, the results presented in the next section of this paper are shown for just one matching criterion. We decided to set $k=4$, since we found, by studying many individual storms, that the 1 point matching often produced incorrect matches. Choosing the criterion for S and T is somewhat a balancing act. Using a strict criterion means that forecast track are more likely to be matched with the correct tracks, but there will be fewer tracks matched; a relaxed criterion will yield more matched tracks, but at the expense of some tracks being incorrectly matched. As a compromise we use matching criteria (ii).

At this point we note that the matching includes both temporal and spatial criteria, but there is no restriction on the difference in intensity between analysed and forecasted storms. We believe that the position and time criteria are sufficient to determine whether a forecasted storm corresponds to an analysed one and that an intensity criterion would introduce bias into the results. For example if we consider the storm of figure 2, it can be seen clearly from the storm tracks that both of the control forecast tracks correspond to the same analysed track (figure 2(a) and(c)). If an intensity criterion were introduced in the matching, the storm predicted by the later forecast (figure 2(c) and (d)) would match the analysed storm because they have very similar amplitudes at the beginning of the forecast. However the amplitude of the storm predicted by the earlier forecast (figures 2(a) and (b)) differs significantly from

the analysed storm at the beginning of the forecast and therefore it would probably not match.

In the next section the diagnostics obtained from the matched tracks are presented. Initially the results were generated separately for the 4 individual seasons discussed in section 2. A single season was found to provide insufficient data at the higher lead times to produce stable statistics, since only a limited number of storm tracks will last longer than 4 or 5 days. We therefore combined the two DJF seasons into one dataset and the two JJA seasons into another dataset. This provided more data points at the higher lead times and helped to stabilise the statistics; however, there is still insufficient data for some types of analysis that we would like to perform (see section 5). The results obtained from the control system are shown separately for the DJF seasons and the JJA seasons and the results for the other observing systems are shown for just the DJF seasons.

4 Results

4.1 The Prediction of Extratropical Cyclones

In this section we explore the prediction of extratropical cyclones by the control system. Table 2 shows the total percentage of forecast storm tracks that match with analysis storm tracks for the DJF and JJA periods in the northern and southern hemisphere. The percentage of matched tracks is larger for the winter seasons, i.e. there is a larger percentage for DJF than JJA in the NH and a larger percentage for JJA than DJF in the SH.

It should be noted 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 a forecast storm track that matches with an analysis track

could have been identified in a forecast which was integrated from an initial state that occurred before the vorticity/pressure center showed up in the 850 hPa analysis. In this case the first point in the forecast track, that coincides in time with the analysis track, will have a lead time greater than 0. Another possibility, and perhaps the most intuitive, is that the forecast storm track is identified in the forecast which was integrated from the initial state in which the center was first identified, and so the first point in the forecast track occurs at lead time 0. The final possibility is that the forecast storm track is identified in a forecast, which was integrated from an initial state that occurred a few days after the storm first showed up in the analysis and when the storm is more developed. In this case the forecast can only predict the later part of the storm track.

The solid lines in figure 5(a) and (c) show the mean separation distance between the matched forecast tracks and corresponding analysis tracks, as a function of forecast lead time, for the NH and SH respectively. Figure 5(b) and (d) show the mean absolute intensity difference between the matched forecast tracks and the analysis tracks. The dashed lines included in the figure will be discussed later. The separation distance curves take a very different shape to the intensity difference curves. Whilst the error in the position of the cyclones increases fairly linearly, with a slightly steeper gradient at the higher lead times, the error in the intensity increases faster initially and then levels off. This consistent difference between the error growth rates suggests a greater predictive skill in position than intensity.

The skill in predicting the position of the cyclones is very similar for the DJF and JJA periods in each hemisphere, with just slightly higher skill for the summer seasons (JJA in NH and DJF in SH), particularly at the higher lead times. More of a difference can be seen between the DJF and JJA seasons in the intensity diagnostics.

In the NH the skill is about 1 day higher for the JJA season than for the DJF season. There is less of a difference between the seasons in the SH; the skill is comparable until the day 2 lead time at which point the DJF season becomes about $\frac{1}{2}$ a day better. Beyond the day 5 lead time the statistics become noisy and less reliable due to insufficient data. Since the JJA seasons correspond to the NH summer and the DJF to the SH summer, the differences in the prediction of the intensity is presumably because there is a larger error in the predicted intensities of storms of higher amplitudes that occur in the winter seasons. The larger difference between the results of the DJF and JJA seasons in the NH than in the SH would then corresponds to the larger difference in the amplitudes of the storms between the two seasons in the NH.

By comparing the two hemispheres we see that there is a higher level of skill in the NH than the SH for both the position and intensity of the storms. There is approximately 1 day more skill for the prediction of their position in the NH than in the SH. Since there is a significant difference in the intensity diagnostics between the DJF and JJA seasons, it is less straight forward to compare the two hemispheres than for the position diagnostics. The skill for the NH DJF season is very similar to that of the SH; however, this is perhaps not the fairest comparison, since we are comparing the NH winter with the SH summer. If the NH DJF season is compared with the SH JJA season (i.e. the two winter seasons are compared) then there is about $\frac{1}{2}$ a day more skill in the NH than in the SH. Similarly if the NH JJA season is compared with the SH DJF season (i.e. the two summer seasons are compared) then there is about 1 day more skill in the NH than in the SH. This difference in skill between the hemispheres is probably due to the larger number of terrestrial observations in the NH, particularly the wind observations provided by radiosondes and aircraft. The

larger number of observations will allow accurate initial states to be produced and will result in higher quality forecasts.

The solid lines in figure 5 include extratropical cyclones of all amplitudes. To determine whether the differences in the intensity diagnostics discussed above were due to the amplitude of the storms we also considered the prediction of just the intense storms. Storms that reached an amplitude of $8.0 \times 10^{-5} \text{ s}^{-1}$ or more in their lifetime were selected from the analysis storm tracks and the predictive skill diagnostics were re-generated for just these intense storms. This chosen value of $8.0 \times 10^{-5} \text{ s}^{-1}$ lies around the 80th percentile of the maximum intensity distribution of the analysis tracks. Storms that reach or exceed this amplitude can therefore be considered more intense than normal. The results obtained for just these intense storms are shown as dashed lines in figure 5. By filtering the data according to the amplitude of the cyclones we have significantly reduced our sample size and therefore do not consider the intense storm diagnostics to be reliable beyond the day 5 lead time.

Considering the intense storms separately makes little difference to the cyclone position diagnostics, although the position of the intense storms does seem to be slightly better predicted. The skill in predicting the amplitude of the intense storms is reduced by about 1 day in the NH for both the DJF and JJA seasons. There is also a reduction in skill in the SH for the JJA season, but not by as much as in the NH and there is no reduction in skill for the DJF season. This is perhaps due to the different nature of the storms in the two hemispheres. The growth rate of cyclones can be significantly larger in the NH than the SH; the mean growth rates in the main baroclinic regions of the NH winter are $\sim 1 \text{ day}^{-1}$ (Hoskins and Hodges 2002), whereas in the SH the large ocean regions have mean growth rates of $\sim 0.5 \text{ day}^{-1}$ (Hoskins and

Hodges 2005). If the forecast model incorrectly predicts the fast growth rates associated with intense storms in the NH then this would lead to a larger error in the predicted amplitude of these storms.

To explore this further the rate of change of intensity of the cyclones was calculated at each time step by $\frac{ds}{dt} = \frac{s(t_2) - s(t_1)}{t_2 - t_1}$ where $s(t)$ denotes the amplitude of a cyclone at time t . Figure 6(a) and (c) shows the mean absolute difference in the rate of change in intensity of the forecast storm tracks and the corresponding analysis tracks for the all storms and for just the intense storms in the NH and SH respectively. We only considered storms that were growing, i.e. those for which the rate of change of intensity is positive. The results are only shown to the day 3 lead time. Since most storms reach their peak within the first 3 days of the forecast, there is insufficient data at higher lead times. There is a large error in the rate of change of intensity at the beginning of the forecast and the error increases slightly with forecast lead time. The errors are larger when just the intense storms are considered. The errors for the JJA season in the NH and both the DJF and JJA seasons in the SH are of a similar magnitude, but the error is significantly larger for the DJF period in the NH. This suggests that the forecast model is unable to accurately predict the rapid growth rates associated with NH winter storms.

When considering the error in the predicted intensity and the rate of change of intensity we have used the absolute difference as a measure of error. Although this provides a measure of the magnitude of the error it does not provide any information about whether the intensity of storms is in general being overpredicted or underpredicted. The mean signed intensity difference was also considered (not shown) and the results indicated that the intensity of the storms was generally overpredicted in the SH, but was more variable in the NH. We have not presented

these results because our sample size is insufficient to obtain reliable and consistent results, but it would be interesting to investigate this further with a larger dataset.

Figure 6(b) and (d) shows the mean absolute difference in propagation speed between the forecast and analysis tracks for the NH and SH respectively. The error growth is significantly faster in the SH than the NH. Although the initial error is slightly larger in the NH, the faster error growth in the SH means that, from the day 3 lead time, the error is larger in the SH. There is a slightly larger error for the winter seasons than for the summer seasons in both hemispheres; however there is slightly less error for the intense storms.

In the example of figure 2 the forecasted storms all move at a slower speed than the analysed storm. To determine whether this was the case in a statistical sense, the mean signed speed difference between the forecast tracks and corresponding analysed tracks was calculated and is shown in figure 7 for the DJF and JJA seasons in both hemispheres. The difference is consistently negative showing that in general the forecasted storms are moving slower than the analysed storms. This bias is very small in the SH, but is significantly larger in the NH.

Before continuing to the next section it is worth noting that the x-axes in figures 5, 6 and 7 correspond to the lead time of the forecast irrespective of when in the forecast the storm tracks start. We realise that this is not the ideal solution since tracks which begin at the beginning of the forecast will have had more time to diverge, from the corresponding analysis track, by the day 7 lead time than those tracks that do not begin until the day 3 lead time. Ideally it would be better to separate the data and consider storms which start at each lead time separately; however, the amount of data used in this study is not sufficient for this type of analysis. The purpose of figure 5 is to compare the prediction of the position of the cyclones with that of the intensity for

the DJF and JJA seasons in the two hemispheres and to determine whether the more intense storm are predicted any better than the weaker ones. Since all the results presented in the figure are subject to the same limitations we believe that combining storms which start at different lead times will have very little impact on the conclusions. Similar arguments apply to the other figures. The conventional forecast verification methods of RMS error and anomaly correlation suffer from a similar problem, since any forecast field of a given lead time will consist of a mixture of weather systems at different stages of development.

4.2 The Predictability of Extratropical Cyclones

Lorenz (1982) devised a method for quantifying upper and lower bounds of atmospheric predictability. The lower bound, or predictive skill, was determined by calculating the RMS difference between forecast data, of varying lead times, and analysis data valid at the same time. The upper bound, or predictability, was determined by calculating the RMS difference between consecutive pairs of forecasts, valid at the same time, but with lead times differing by some fixed time interval. For example if this interval were 1 day, then the analysis for a given day would be compared with the 1-day forecast valid for the same day, then this 1-day forecast would be compared with the 2-day forecast valid for the same day and so on. Lorenz argued that if the forecast model was realistic enough that two forecasts started from similar initial states (i.e. forecasts separated by 1 day) diverged at a similar rate to that at which two similar but distinct atmospheric states diverged, then the predictability measure described above could not be improved unless the 1-day forecast error was reduced. The method provides a convenient way to determine how errors of different sizes grow with increasing forecast lead time and gives a measure of potential forecast

skill (Bengtsson et al. 2005; Bengtsson and Hodges 2005a; Simmons and Hollingsworth 2002).

In this section the predictability measure described above is extended to storm tracks, by an internal comparison of the control forecast storm tracks, in order to obtain an estimate of how much potential there is to improve the prediction of the position and intensity of extratropical cyclones by changes to the model. The storm tracks identified in the forecast started from the $(i+j)$ th time step were matched with the storm tracks identified in the forecast started from the i th time step for all time steps i with $j=1,2,3,4$. Since the time steps are 6 hours apart j corresponds to 6,12,18 and 24 hour intervals between the forecasts. Considering different values of j enables the growth of errors of different sizes to be explored. The matched storm tracks were used to generate predictability diagnostics analogous to the predictive skill results of figures 5. Figure 8 shows the results obtained for both the NH and SH for the DJF season. The predictive skill curves from figure 5 are also included for comparison as the lower bound.

The results show an increase in both separation distance and intensity difference as the time interval j between the forecasts being compared is increased (i.e. as the size of the initial error is increased the size of the errors at higher lead times increase). In the NH the day 2 doubling time (the time taken for the error at day 2 to double) for the position of the cyclones is about 2 days regardless of the size of the initial error estimate. In the SH the $j=6$ hour curve is almost the same as that of the NH, whereas the $j=24$ hour curve increases at a faster rate for higher lead times. The doubling times are slightly shorter than in the NH.

The NH mean absolute intensity difference curves take a different shape to the predictive skill curves barely increasing until day 2. The curves lie parallel to each

other, meaning that the doubling times will vary for different values of j . When $j=24$ the day 2 doubling time is 2.5 days and when $j=6$ it is 4 days. In the SH the intensity difference increases at a faster and more steady rate with day 2 doubling times of 2 and 3.5 days for $j=24$ and $j=6$ respectively ($\frac{1}{2}$ a day less than in the NH).

An unexpected result is the slow growth rate of both the separation distance and intensity difference at the lower lead times. An analogous result was also found using the conventional Eulerian approach by Bengtsson and Hodges (2005a). The result is particularly noticeable for the NH intensity difference, which as previously mentioned does not really start to increase until the day 2 lead time. For further discussion please see the Bengtsson and Hodges (2005a) study.

Comparing the predictive skill curves with the $j=24$ hour predictability curve shows, for the DJF season, that the skill in predicting the position of the storms could potentially be improved by about 1.5 days in both hemispheres whereas the intensity could be improved by 2 days in the SH and by 2.5 - 3 days in the NH. If the $j=6$ hour curve, having the smallest initial error estimate, is used instead as the upper bound on predictability then these estimates increase by about $1\frac{1}{2}$ days.

The predictability estimates were also calculated for the JJA season (not shown). As for the DJF season the results showed that the skill in predicting the position of the storms could potentially be improved by about 1.5 days in both hemispheres. The potential improvement for the intensity was also the same 2 days in the SH, but in the NH it was only 1.5 – 2 days (1 day less than for the DJF season). This again perhaps highlights the models inability to predict the rapid deepening of intense NH winter storms.

4.3 Predicting Storms before their Genesis

In the introduction to this paper we discussed how various operational meteorological centers use cyclone tracking as a method for verifying tropical cyclone predictions. Such centers only consider cyclones that have already been identified in the analysis cycle and are therefore present in the initial conditions of the forecast. They do not consider cyclones which are generated later in the forecast. It is clearly important, that once a cyclone (tropical or extratropical) has formed, to predict both the track and its amplitude as accurately as possible. However, it is also important that an indication of such cyclones, particularly those that are likely to be intense, is given as far in advance as possible by forecasts made before the cyclone has been identified in the analysis cycle. In this section some results are presented addressing the issue of how far in advance of their genesis, extratropical cyclones can be predicted.

The genesis of a storm was taken to be the first point in its analysed 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 $1.0 \times 10^{-5} \text{ s}^{-1}$ (relative to the large scale background field removal) to be considered a cyclone. To determine whether the cyclone was predicted N days in advance of its genesis, the forecast storm tracks identified in the forecast made N days before the genesis of the analysis storm track, were examined to see if any of them matched the analysis track.

We realise that the identification of a 850 hPa vorticity center is not the only indication of a developing storm. There will almost certainly be other upper level precursors that can be identified in the analysis cycle before the 850 hPa vorticity center. However, our definition of storm genesis marks a specific stage of cyclone

development, when a cyclone can easily be identified in the 850 hPa level of the initial conditions.

Figure 9(a) and (b) show for both the NH and SH the percentage of control analysis storm tracks, which are predicted by the control system as a function of the number of days (N) before the storms genesis occurred in the analysis for the DJF and JJA seasons respectively. The value when $N=0$ is the percentage of analysed storms that are predicted by forecasts integrated from the analyses in which the storms genesis occurs (i.e. the initial state is the analysis which contains the first point of the analysed track). This should not be confused with the percentages in tables 1 and 2, which are the percentage of forecast tracks that match with analysis tracks and includes forecasts that were started before and after the storms were first identified in the analysis.

In the NH about 60% of the storms are predicted when the forecast model is integrated from the analysis in which the storm was first identified ($N=0$), about 40% of the storms are predicted 1 day before ($N=1$) and less than 10% are predicted 3 days before ($N=3$). In the SH the control predicts more of the storms than in the NH when N is less than 1, but it predicts less than in the NH when N is larger. The results are comparable for the DJF and JJA seasons. These percentages will clearly vary considerably when different matching criteria are used; however, the general results for the different hemispheres and seasons stay the same. Even with the very relaxed matching criterion (iii), the results still show that the majority of storms are not predicted more than 3 days before they are identified in the analysis.

When the forecast storm tracks were matched with the control analysis storm tracks, only those storms whose genesis occurred within the first 3 days of the forecast were considered (see section 3.3). The results in figure 9 were also produced

without applying this restriction. The percentage of tracks predicted was very low for $N > 3$ indicating that storms are rarely predicted more than 3 days before they are first identified in the 850 hPa analysis. This result may be different for forecasts obtained from more modern data assimilation systems (4D-Var) from more recent time periods that have more observations and will be investigated in future work.

4.4 The Impact of Observations on the Prediction of Extratropical Cyclones

The results to this point have concerned the prediction of extratropical cyclones by the control system. In this section we explore the impact that observations of different types have on the prediction of the cyclones in the DJF seasons. Table 3 shows the total percentage of forecast storm tracks that match with control analysis storm tracks for each of the observing systems. In the NH the terrestrial system is almost the same as the control and is better than the satellite system. The satellite system has the highest percentage in the SH, but is significantly lower than the control system. The surface only system is very poor in both hemispheres. These results are all in agreement with those of Bengtsson et al. (2005).

Figure 10(a) and (c) show the mean separation distance between the matched forecast tracks and corresponding control analysis tracks as a function of forecast lead time, for the control, terrestrial and satellite systems in the NH and SH respectively. The surface system is not included, since only a limited number of the storm tracks identified matched with storm tracks in the control analysis (see table 3). In the NH the terrestrial system is almost identical to the control. The satellite system increases at the same rate as the control but has about $\frac{1}{2}$ a day less skill. In the SH the satellite system is almost identical to the control and the terrestrial system has significantly less skill due to the dominance of satellite observations in the SH. It should be noted that in the NH there is not much of a reduction in the percentage of tracks that match

from the control to the terrestrial system, but in the SH there is a larger reduction from the control to the satellite (see table 3). There is an increase in the difference between the terrestrial system and control as the lead time increases. At day 0 there is a difference of less than 2° , but at day 7 this difference has increased to 5° . This will be partly due to the spatial matching used at the beginning of the tracks and also due to reduced data at the higher lead times, which leads to less statistically stable results. The effect is less pronounced when matching criterion (iii) is used (not shown), which is probably the best criterion to use for the terrestrial system in the SH, because the number of tracks that match is very low compared to the other systems (see table 3). With this very relaxed criterion, the terrestrial system consistently has a separation distance of about 2° more than the control, which equates to a 1 day reduction in predictive skill.

Figure 10(b) and (d) show the mean absolute intensity difference between the matched forecast tracks and the control analysis tracks for the different observing systems in the NH and SH respectively. As with the separation distance diagnostics in the NH, the terrestrial system has a very similar level of skill to the control and is better than the satellite. Beyond the day 5 lead time the statistics become unstable, but up to this point the satellite system shows a reduction in predictive skill of about $\frac{1}{2}$ a day to the control, which is the same as for separation distance. In the SH the satellite mean intensity curve is very close to the control curve from the day 1 lead time. There is a noticeable difference between the control and the satellite system for the first day of the forecast showing that the limited terrestrial observations available in the SH are having an impact on the quality on the earlier part of the forecasts. The predictive skill of the terrestrial system is about 3 days less than the control for intensity. Hence the SH terrestrial system has a much lower level of predictive skill

for intensity than for position. Since there are very few terrestrial observations in the SH, the state of the atmosphere is probably not represented accurately by the terrestrial system, causing large errors in the predicted intensities.

Figure 11(a) and (c) show the mean absolute difference in the rate of change in intensity of the forecast storm tracks and the corresponding analysis tracks for the different observing systems in the NH and SH respectively. In the NH the terrestrial curve is again almost the same as the control curve. The satellite system has a considerably larger error for the first 2 days of the forecast. This may be because the satellite system does not provide a sufficient vertical resolution of observations to accurately represent the storms vertical structure. This error in the predicted growth of the storms could be one of the main reasons why the satellite system has less predictive skill than the terrestrial system in general in the NH.

In the SH the satellite system has a larger error than the control system for the first day of the forecast, which corresponds to the larger error in the predicted intensity in the first day of the forecast (see figure 10(d)). The terrestrial system again has a much larger error. As with the control system, the errors are smaller in the SH than in the NH for the other observing systems indicating that the model is unable to accurately predict the large growth rates of some NH winter storms. This was discussed in more detail previously in section 4.1 for the control system.

Figure 11(b) and (d) show the mean absolute difference in propagation speed between the forecast storm tracks and analysis tracks. A similar relationship exists between the different observing systems to that of the other diagnostics. As with the control system the satellite system has a larger error in the NH than the SH for the first 3 days of the forecast.

5 Discussion and Conclusions

In this paper a new method for measuring forecast skill and predictability involving the identification and tracking of extratropical cyclones, has been developed and implemented to obtain detailed information about the prediction of extratropical cyclones. The method provides a direct measure of how forecasted weather systems deviate from their analysed counterparts with increasing lead time, which can not be obtained from the more conventional analysis methodologies. Since extratropical cyclones play a large role in determining the weather in the mid-latitudes we believe the new method provides a good measure of the ability of NWP to predict the weather.

As with any data analysis methodology, the method does have some limitations and biases that should be taken into account. The main limitation is the large amount of data required to get reliable results. We believe that our sample size is large enough for most diagnostics; however the current statistics include a mixture of storms forecasted from different stages in their life cycle. It would be better to separate storms that are forecasted from the initial state in which the storm was first generated, from those that are predicted from earlier and later analyses, but several years worth of data would be required to produce reliable statistics. Another potential limitation is the matching criteria. By restricting the separation distance between the first 4 points of the forecast and analysis track we may have introduced some bias into the results, but any type of matching will inevitably introduce bias of some sort. Many sensitivity tests were performed to explore the affect the matching has on the diagnostics. For example the impact of performing the spatial matching with 1 point rather than 4 was explored (figure 4) and was found to have very little impact on the error growth rates of the predicted positions and intensities of the storms. An

additional point to note is that the diagnostics have been produced from only those tracks that match. So we have essentially taken an optimistic viewpoint, only looking at those storm tracks that are “well predicted”. The percentage of forecast storm tracks that match with the analysis (tables 2 and 3) should therefore always be taken into account with the other diagnostics.

The new methodology presented in this paper does have a number of limitations and biases, but traditional approaches also have deficiencies. For example if a storm is well predicted, but just misplaced slightly this will significantly affect the RMS error and anomaly correlation. Although these methods can be applied to a variety of different scale fields, they are often only applied to the 500 hPa geopotential height field. This focuses on large scale aspects of the weather rather than the weather systems themselves and may therefore give comparatively high measures of forecast skill. Considering smaller scale fields, such as vorticity, may give very different results.

We finish this paper with a summary and discussion of the main results. The skill in predicting the position of extratropical cyclones is significantly higher than that of the intensity. This can be seen clearly from figure 2 for one particular storm (it is also true for other individual storms we have examined) and the statistical results of section 4.1 show this to be the case for the majority of storms. The predictability calculations also indicate that there is more potential to improve the prediction of the intensity than the position of the storms. They show that, without improving 24 hour forecasts, there is potential to increase the skill of forecasts of higher lead times by 1 - 1½ days for position and 2 - 3 days for intensity via improvements to the forecast model. These values increase by about 1½ days if the predictability is calculated by comparing forecasts only separated by 6 hours (i.e. without improving 6 hour

forecasts). The sensitivity of these results to the matching was explored extensively (section 3.3) and we found that varying the matching criteria had very little impact on the position and intensity diagnostics, which gives us confidence in our results. We believe that the lower level of skill in predicting the intensity of the storms may be because the vertical structure of the storms is incorrectly represented. The vertical tilt is critical to the storms development and if incorrect will cause errors in the predicted amplitudes of the storms (see Holton 2004, chapter 8). The position of the storms, on the other hand, is mainly determined by the large scale flow pattern and will consequently be less affected by an incorrect tilt. The results also show that forecasted storms move at a slower speed than the corresponding analysed storms on average. This bias is relatively small, but is larger in the NH than the SH. The reason for this is currently unclear and will require further study. One possible explanation could be some type of numerical error caused by the truncation used in the model.

When considering just the high intensity storms we find that there is a significant reduction in the skill in predicting their amplitude. This explains why the amplitude of the winter storms is not as well predicted as that of the summer storms. Further analysis showed that there is a larger error in the predicted rate of change in intensity of the intense storms than the weaker ones in both hemispheres. The error is significantly larger for the NH DJF seasons suggesting that the model is unable to accurately predict the very fast growth of intense NH storms.

Storms that are predicted before they are identified in the initial conditions (at the 850 hPa level) were considered. Most storms are not predicted more than 1 day before their genesis, but a few storms are predicted as much as 3 days before. It would be interesting to know whether improvements to forecast models, such as increased resolution and better data assimilation methods, would extend this 3 day

limit. Ensemble prediction could also potentially extend this limit and is being investigated as future work.

The results relating to the impact that the different observing systems have on the predictability of the storm tracks confirm and extend those of the previous Bengtsson et al. (2005) study. In the NH the terrestrial system has almost the same level of skill as the control system and the satellite system has about $\frac{1}{2}$ a day less skill. Further analysis shows large errors, in the earlier part of the forecasts, for the predicted growth of the storms by the satellite system and could be part of the reason for the reduction in skill from the terrestrial to the satellite system in the NH. In the SH hemisphere the dominance of the satellite observations is apparent, but the terrestrial observations do have a noticeable impact on the quality of the forecasts. The surface system has little skill in either hemisphere.

The results of this paper indicate that in general the predictive skill, with respect to extratropical cyclones, is higher in the NH than in the SH. This is probably mainly due to the larger number of terrestrial observations in the NH, allowing more accurate initial states to be produced and resulting in higher quality forecasts. Improving forecast models ability to predict the growth of storms, particularly those that are more intense, could significantly improve NH forecast skill. Indeed the predictability calculations indicate that there is potential to improve the prediction of the intensity of NH winter storms by $\frac{1}{2}$ - 1 day more than that of SH winter storms. As far as SH forecasts are concerned, improvements to the observing network would probably still be more beneficial than improvements to the model.

In future work we will explore the vertical tilts of the storms to hopefully obtain further information about why the position of extratropical cyclones is predicted better than the intensity. The diagnostics of this study will hopefully be re-produced

for a 4D-Var analysis system, with a more recent observing system and with a longer time period. Longer time periods would allow us to perform some regional analysis so that ocean and land based systems could be considered separately. With a larger dataset we could also partition the storms according to their stage of development in the initial conditions and generate diagnostics for the different stages.

Acknowledgements

We would like to thank ECMWF for making the ERA40 analysis and forecast systems available to us and for providing support in running the experiments. We would also like to thank the anonymous reviewers for their thoughtful and interesting suggestions that helped to improve this paper.

References

Bengtsson, L. and K. I. Hodges, 2005a: A Note on Atmospheric Predictability. *Tellus*, **V58A**, 154-157.

Bengtsson, L. and K. I. Hodges, 2005b: On the Impact of Humidity Observations in Numerical Weather Prediction. *Tellus*, **57A**, 701-708.

Bengtsson, L., K. I. Hodges and L. S. R. Froude, 2005: Global Observations and Forecast Skill. *Tellus*, **57A**, 515-527.

Bengtsson, L., K. I. Hodges and S. Hagemann, 2004a: Sensitivity of Large Scale Atmospheric Analyses to Humidity Observations and its Impact on the Global Water Cycle and Tropical and Extratropical Weather Systems. *Tellus*, **56A**, 202-217.

Bengtsson, L., K. I. Hodges and S. Hagemann, 2004b: Sensitivity of ERA40 reanalysis to the observing system: determination of the global atmospheric circulation from reduced observations. *Tellus*, **56A**, 456-471.

Heming J., 1994: Keeping an eye on the hurricane – Verification of tropical cyclone forecast tracks at the Met. Office. *NWP Gazette*, **1**, 2-8.

Hodges, K. I., 1995: Feature Tracking on the Unit-Sphere. *Monthly Weather Review*, **127**, 3458-3465.

Hodges, K. I., 1999: Adaptive Constraints for Feature Tracking. *Monthly Weather Review*, **127**, 1362-1373.

Holton, J. R., 2004: *An Introduction to Dynamic Meteorology*. 4 ed. Elsevier Academic Press, 535pp.

Hoskins, B. J. and K. I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *Journal of the Atmospheric Sciences*, **59**, 1041-1061.

Hoskins, B. J. and K. I. Hodges, 2005: A New Perspective on Southern Hemisphere Storm-Tracks. *Journal of Climate*, **18**, 4108-4129.

Lorenz, E. N., 1982: Atmospheric Predictability Experiments with a Large Numerical Model. *Tellus*, **34**, 505-513.

Marchok, T. P., 2002: How the NCEP tropical cyclone tracker works. *Preprints*, 25th conference on hurricanes and tropical meteorology, San Diego, CA, 21-22.

Simmons, A. J. and J. K. Gibson, 2000: The ERA-40 Project Plan, ERA-40 Report Series No. 1, ECMWF, Shinfield Park, Reading, 63pp.

Simmons, A. J. and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **128**, 647-677.

Van der Grijn G., 2002: Tropical cyclone forecasting at ECMWF: new products and validation. ECMWF Technical Memoranda, ECMWF, Shinfield Park, Reading.

White, P., 2000: IFS Documentation Part III: Dynamics and Numerical Procedures (CY21R4), Meteorological Bulletin M1.6/4. ECMWF, Shinfield Park, Reading, UK.

Xiao, Q., X. Zou, M. Ponca, M. A. Shapiro and C. Veldon, 2002: Impact of GMS-5 and GOES-9 Satellite-Derived Winds on the Prediction of a NORPEX Extratropical Cyclone. *Monthly Weather Review*, **130**, 507-528.

Zhu, H. and A. Thorpe, 2005: Predictability of extratropical cyclones: the influence of initial condition and model uncertainties. *Journal of the Atmospheric Sciences*, in press.

Table Captions

Table 1. Percentage of NH control forecast tracks that satisfy the genesis constraint and match analysis tracks with the six different matching criteria for the two DJF seasons. The total number of forecast tracks that satisfy the genesis constraint is shown below the table.

Table 2. Percentage of control forecast tracks that satisfy the genesis constraint and match analysis tracks for the DJF and JJA seasons and for the NH and SH. The total number of forecast tracks that satisfy the genesis constraint is shown in brackets.

Table 3. Percentage of forecast tracks that satisfy the genesis constraint and match analysis tracks for each observing system in the NH and SH for the two DJF seasons. The total number of forecast tracks that satisfy the genesis constraint is shown in brackets.

Figure Captions

Figure 1. (a) Illustration of forecast experiment and tracking set-up used in previous Bengtsson et al. (2005) study for the 1st December 1990 – 28th February 1991 season. Each cross represents one time frame of data. The analysis data consisted of 3 months of 6 hourly time frames for each experiment. The forecast model was run from each of these time frames out to 7 days and this is represented by the diagonal dashed lines in the diagram. The forecast data was archived daily to form 7 three month forecast datasets corresponding to the solid horizontal lines in the diagram. The storm tracking program was applied to these datasets, including the analysis, to produce 8 ensembles of storm tracks for each observing system.

(b) Illustration of forecast experiment and tracking set-up used in this study for the 1st December 1990 – 28th February 1991 season. The analysis dataset is the same as in (a) and is represented by the horizontal line in the diagram. The forecast model was re-run out to 14 days from each of the analysis 6 hourly time frames, but this time the data was archived every 6 hours to allow the storm tracks to be computed along the forecast trajectories (the diagonal lines in the diagram). This resulted in 360 ensembles of storm tracks for each observing system.

Figure 2. Example of analysed and predicted storm track beginning on 6th January 1991 at 12UTC. Panels (a) and (b) show the track and intensity of the analysed storm and storm predicted by the control forecast beginning from 5th January 12UTC, (c) and (d) show the control forecast started from 6th January 12UTC and (e) and (f) show the terrestrial, satellite and surface forecast started from 6th January 12UTC. The

units of intensity are 10^{-5} s^{-1} relative to the background field removal and the numbers marked on the storm tracks correspond to the forecast lead time in days.

Figure 3. Schematic to illustrate spatial matching with $k=4$. The solid curve represents an analysis storm track and the dashed curves represent forecast storm tracks. Tracks A, B and C would match, but D would not.

Figure 4. Mean separation distance (a) and mean absolute intensity difference (b) between the matched forecast tracks, obtained with each of the six matching criteria, and the analysis tracks as a function of forecast lead time for the control system in the NH. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal) respectively.

Figure 5. Mean separation distance between the matched control forecast tracks and analysis tracks, for all cyclones and for just the intense cyclones, as a function of forecast lead time in the NH (a) and SH (c) for both the DJF and JJA seasons. Panels (b) and (d) are as (a) and (c) but show the mean absolute intensity difference. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal) respectively.

Figure 6. Mean absolute error in rate of change of intensity between the matched control forecast tracks and analysis tracks, for all cyclones and for just the intense cyclones, as a function of forecast lead time in the NH (a) and SH (c) for both the DJF and JJA seasons. Panels (b) and (d) are as (a) and (c) but show the mean absolute

speed difference. Units of rate of change of intensity and speed difference are $10^{-5} \text{ s}^{-1} \text{ day}^{-1}$ and kmh^{-1} respectively.

Figure 7. Mean speed difference between matched control forecast tracks and analysis tracks as a function of forecast lead time in the NH and SH for both the DJF and JJA seasons. Units of speed difference are kmh^{-1} .

Figure 8. Predictability curves obtained with the control system for the DJF season by comparing forecast storm tracks separated by 6, 12, 18 and 24 hours (for details see text). The mean separation distance between the matched forecast tracks is shown, as a function of forecast lead time, in the NH (a) and SH (c). Panels (b) and (d) are as (a) and (c) but show the mean absolute intensity difference. The predictive skill curves of figure 5 are also included for comparison. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal) respectively.

Figure 9. Plots showing the percentage of control analysis storm tracks predicted by the control system, as a function of the number of days (N) before the storms genesis occurred in the analysis, in both the NH and SH for the DJF (a) and JJA (b) seasons.

Figure 10. Mean separation distance between the matched control, terrestrial and satellite forecast tracks and the control analysis tracks as a function of forecast lead time in the NH (a) and SH (c) for the DJF seasons. Panels (b) and (d) are as (a) and (c) but show the mean absolute intensity difference. Units of separation distance and

intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal) respectively.

Figure 11. Mean absolute error in rate of change of intensity between the matched control, terrestrial and satellite forecast tracks and the control analysis tracks as a function of forecast lead time in the NH (a) and SH (c) for the DJF seasons. Panels (b) and (d) are as (a) and (c) but show the mean absolute speed difference. Units of rate of change of intensity and speed difference are $10^{-5} \text{ s}^{-1}\text{day}^{-1}$ and kmh^{-1} respectively.

Table 1. Percentage of NH control forecast tracks that satisfy the genesis constraint and match analysis tracks with the six different matching criteria for the two DJF seasons. The total number of forecast tracks that satisfy the genesis constraint is shown below the table.

Matching Criteria	% Match
(i) $j=4, S=2^\circ, T=60\%$	29.0
(ii) $j=4, S=4^\circ, T=60\%$	41.6
(iii) $j=4, S=4^\circ, T=30\%$	58.6
(iv) $j=1, S=2^\circ, T=60\%$	38.4
(v) $j=1, S=4^\circ, T=60\%$	46.5
(vi) $j=1, S=4^\circ, T=30\%$	65.5

Total Number of Forecast tracks = 16353

Table 2. Percentage of control forecast tracks that satisfy the genesis constraint and match analysis tracks for the DJF and JJA seasons and for the NH and SH. The total number of forecast tracks that satisfy the genesis constraint is shown in brackets.

	NH	SH
DJF	41.6 (16353)	36.2 (15390)
JJA	36.3 (14911)	39.3 (17691)

Table 3. Percentage of forecast tracks that satisfy the genesis constraint and match analysis tracks for each observing system in the NH and SH for the two DJF seasons. The total number of forecast tracks that satisfy the genesis constraint is shown in brackets.

Observing System	NH	SH
Control	41.6 (16353)	36.2 (15390)
Terrestrial	39.6 (16156)	8.5 (15358)
Satellite	31.7 (16550)	30.8 (15558)
Surface	11.6 (16844)	3.1 (16453)

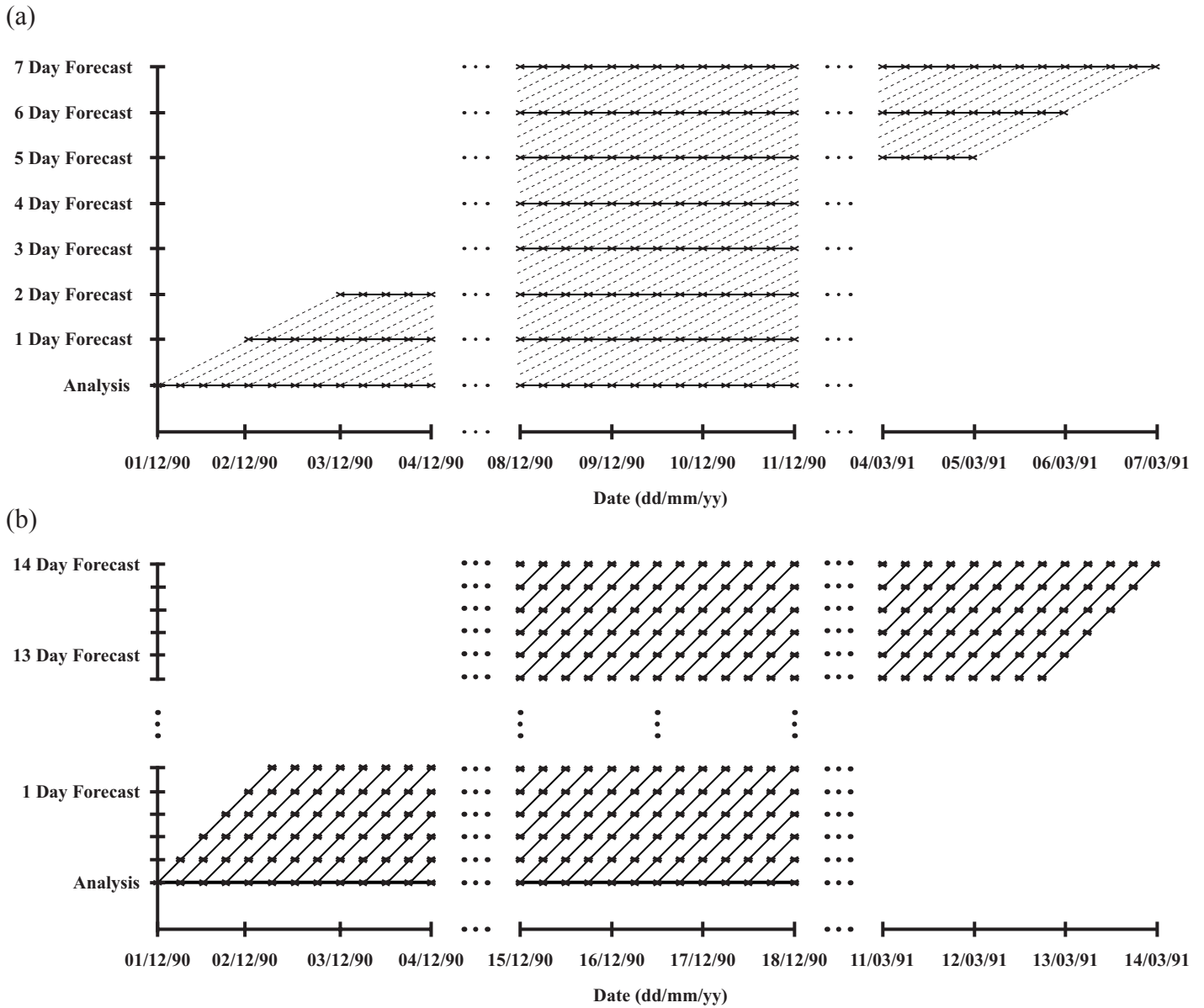
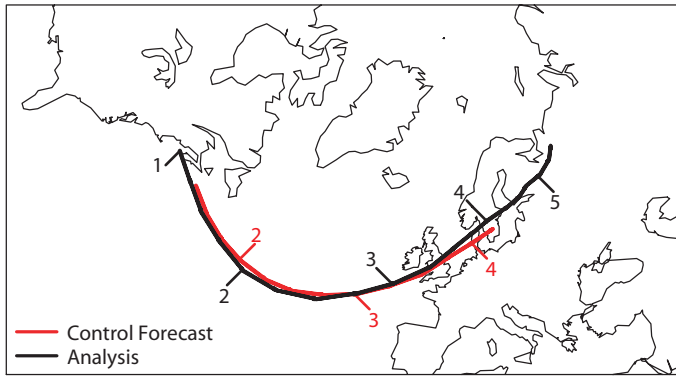
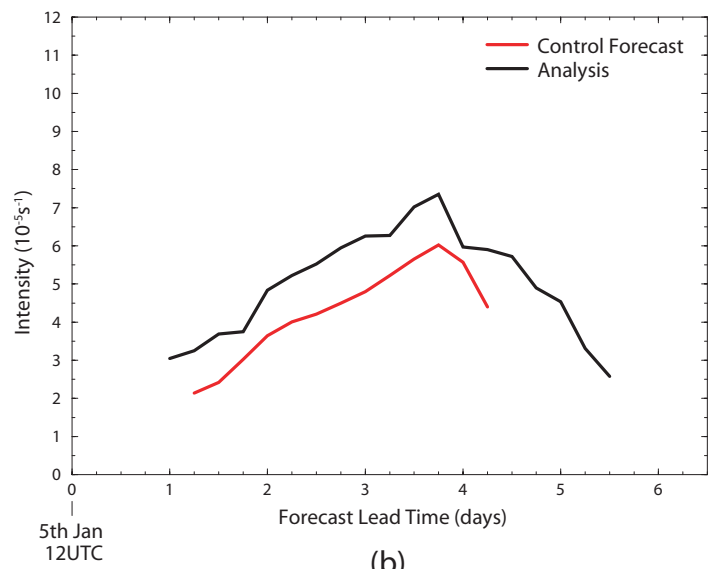


Figure 1. (a) Illustration of forecast experiment and tracking set-up used in previous Bengtsson et al. (2005) study for the 1st December 1990 – 28th February 1991 season. Each cross represents one time frame of data. The analysis data consisted of 3 months of 6 hourly time frames for each experiment. The forecast model was run from each of these time frames out to 7 days and this is represented by the diagonal dashed lines in the diagram. The forecast data was archived daily to form 7 three month forecast datasets corresponding to the solid horizontal lines in the diagram. The storm tracking program was applied to these datasets, including the analysis, to produce 8 ensembles of storm tracks for each observing system.

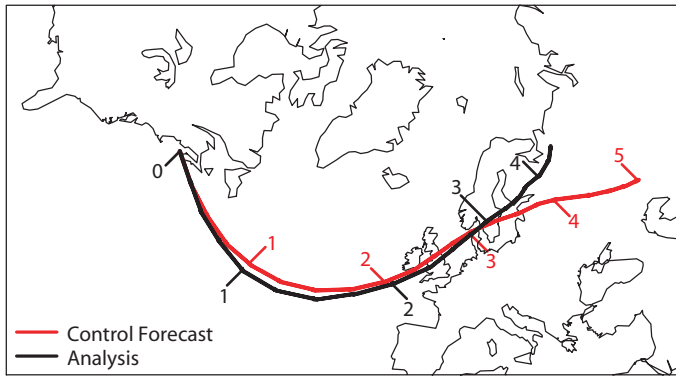
(b) Illustration of forecast experiment and tracking set-up used in this study for the 1st December 1990 – 28th February 1991 season. The analysis dataset is the same as in (a) and is represented by the horizontal line in the diagram. The forecast model was re-run out to 14 days from each of the analysis 6 hourly time frames, but this time the data was archived every 6 hours to allow the storm tracks to be computed along the forecast trajectories (the diagonal lines in the diagram). This resulted in 360 ensembles of storm tracks for each observing system.



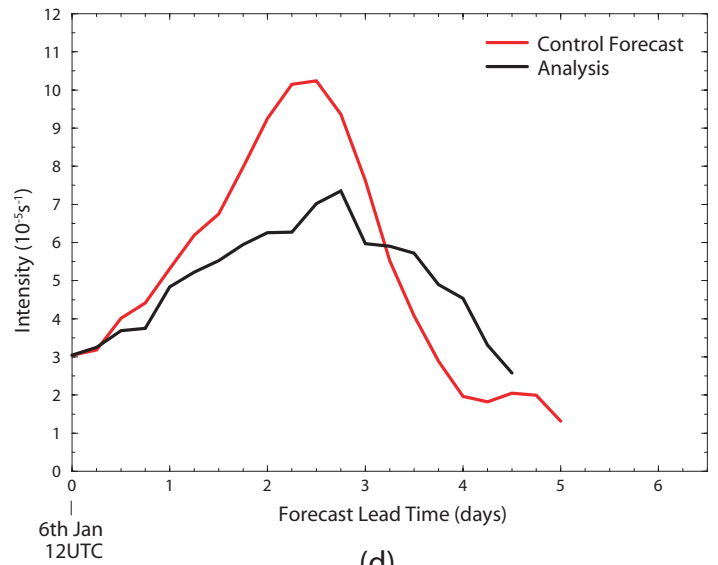
(a)



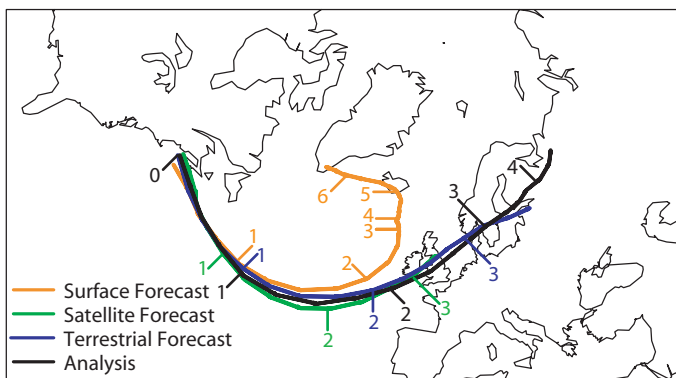
(b)



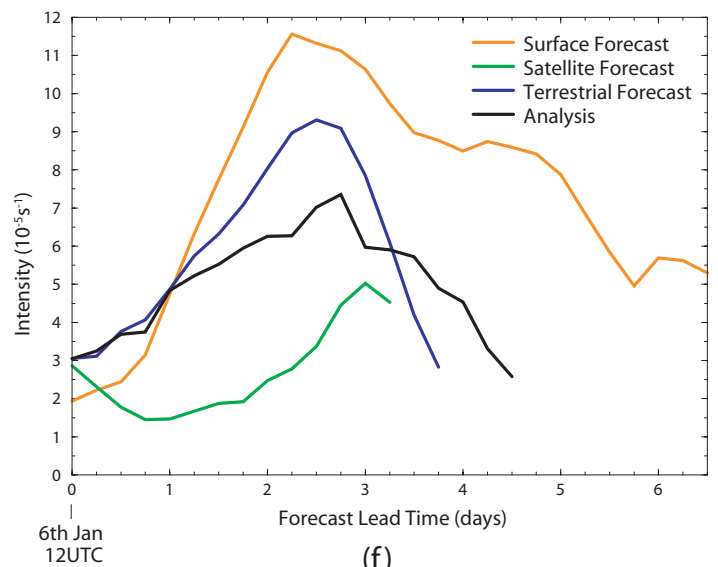
(c)



(d)



(e)



(f)

Figure 2. Example of analysed and predicted storm track beginning on 6th January 1991 at 12UTC. Panels (a) and (b) show the track and intensity of the analysed storm and storm predicted by the control forecast beginning from 5th January 12UTC, (c) and (d) show the control forecast started from 6th January 12UTC and (e) and (f) show the terrestrial, satellite and surface forecast started from 6th January 12UTC. The units of intensity are 10^{-5} s^{-1} relative to the background field removal and the numbers marked on the storm tracks correspond to the forecast lead time in days.

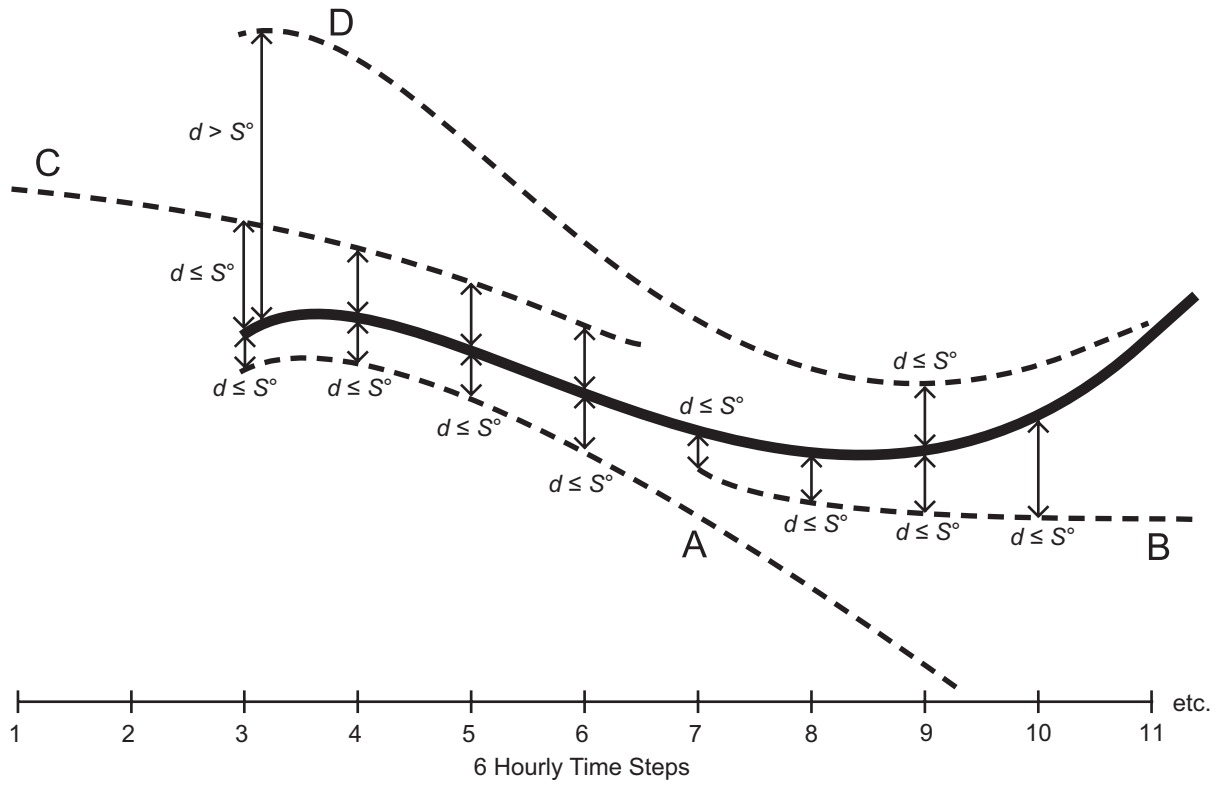


Figure 3. Schematic to illustrate spatial matching with $k=4$. The solid curve represents an analysis storm track and the dashed curves represent forecast storm tracks. Tracks A, B and C would match, but D would not.

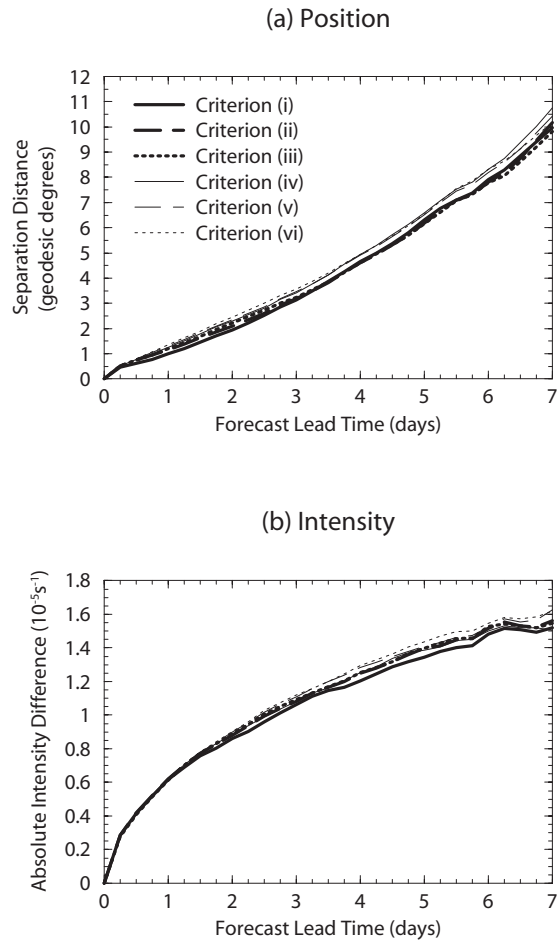


Figure 4. Mean separation distance (a) and mean absolute intensity difference (b) between the matched forecast tracks, obtained with each of the six matching criteria, and the analysis tracks as a function of forecast lead time for the control system in the NH. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal) respectively.

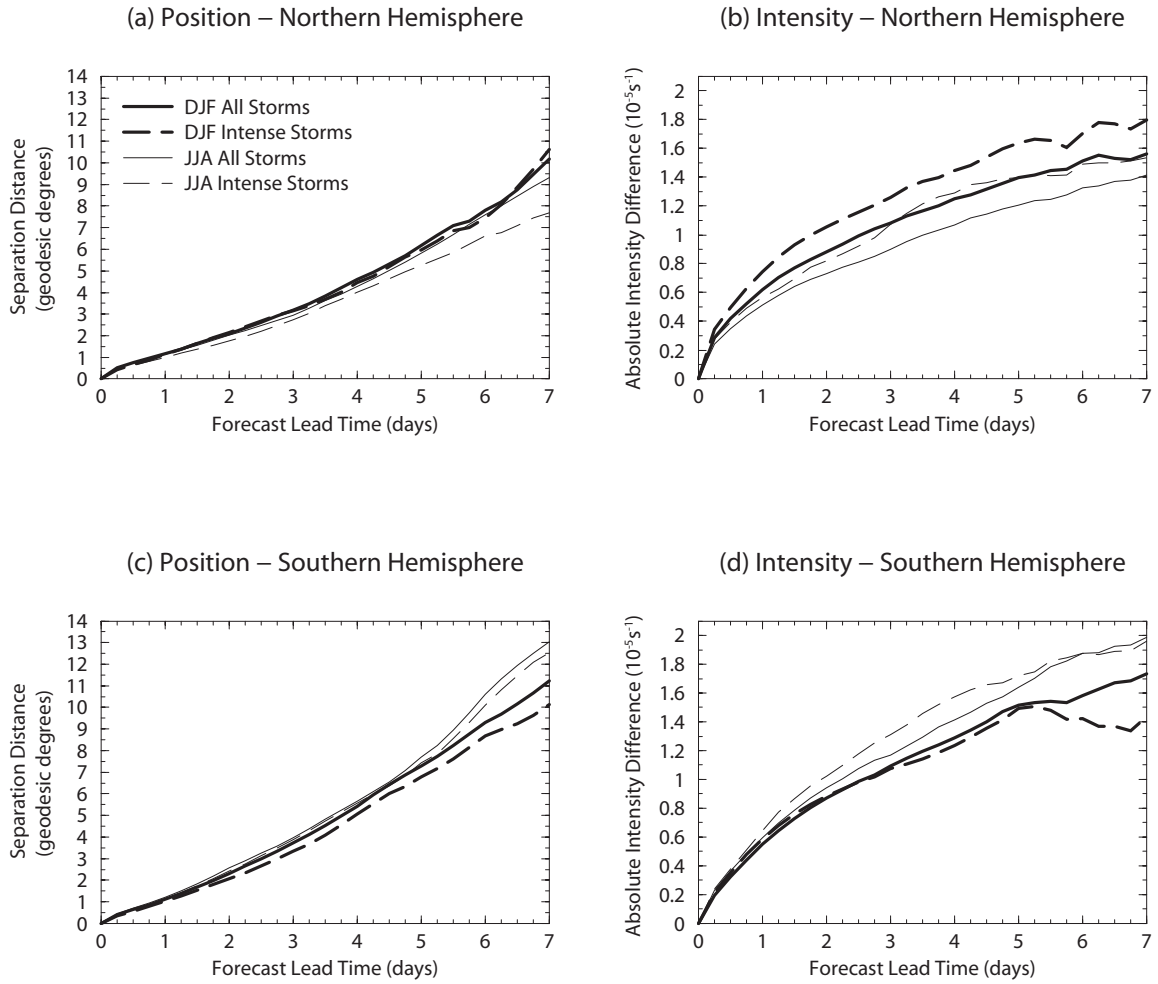


Figure 5. Mean separation distance between the matched control forecast tracks and analysis tracks, for all cyclones and for just the intense cyclones, as a function of forecast lead time in the NH (a) and SH (c) for both the DJF and JJA seasons. Panels (b) and (d) are as (a) and (c) but show the mean absolute intensity difference. Units of separation distance and intensity difference are geodesic degrees and $10^{-5} s^{-1}$ (relative to background field removal) respectively.

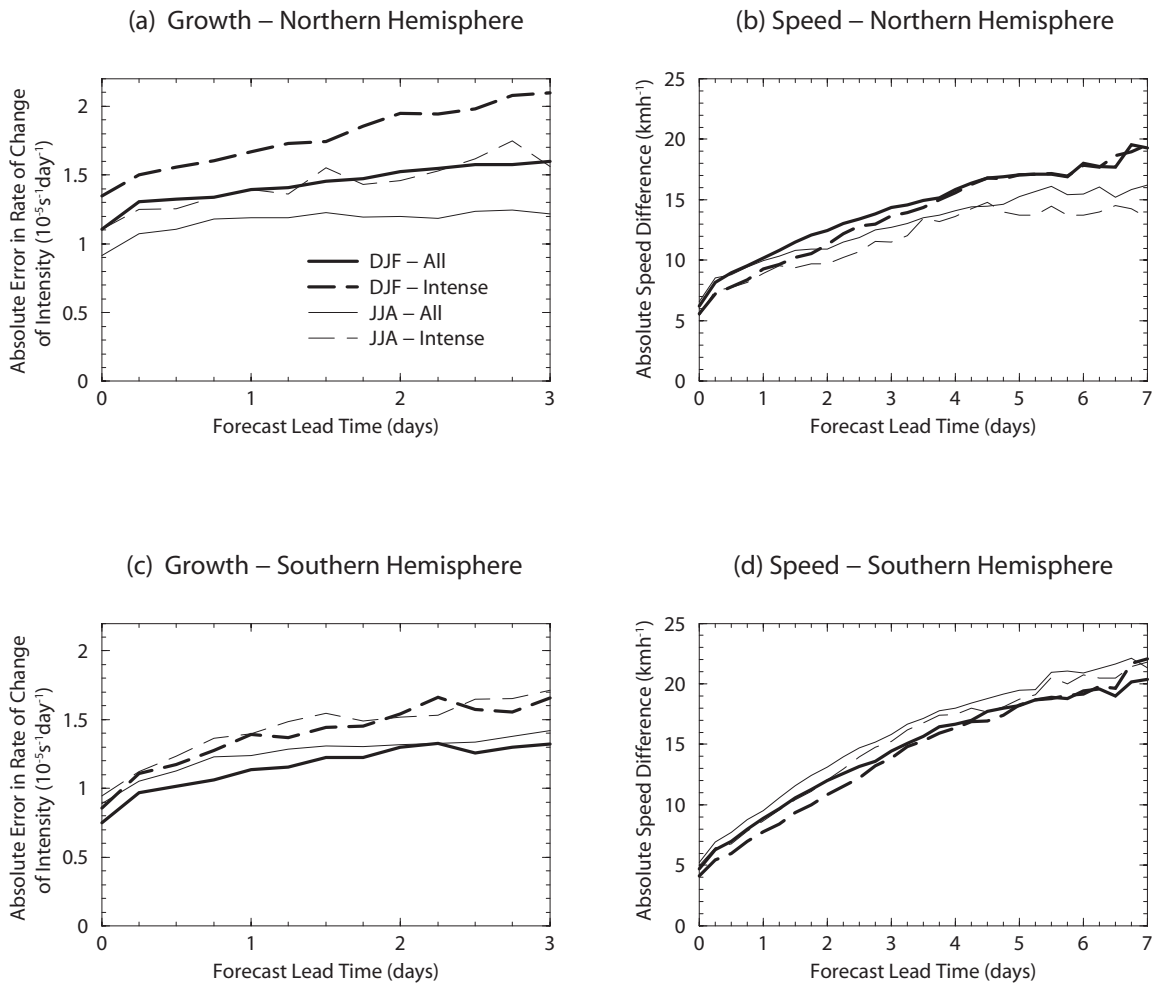


Figure 6. Mean absolute error in rate of change of intensity between the matched control forecast tracks and analysis tracks, for all cyclones and for just the intense cyclones, as a function of forecast lead time in the NH (a) and SH (c) for both the DJF and JJA seasons. Panels (b) and (d) are as (a) and (c) but show the mean absolute speed difference. Units of rate of change of intensity and speed difference are $10^{-5} \text{ s}^{-1} \text{ day}^{-1}$ and kmh^{-1} respectively.

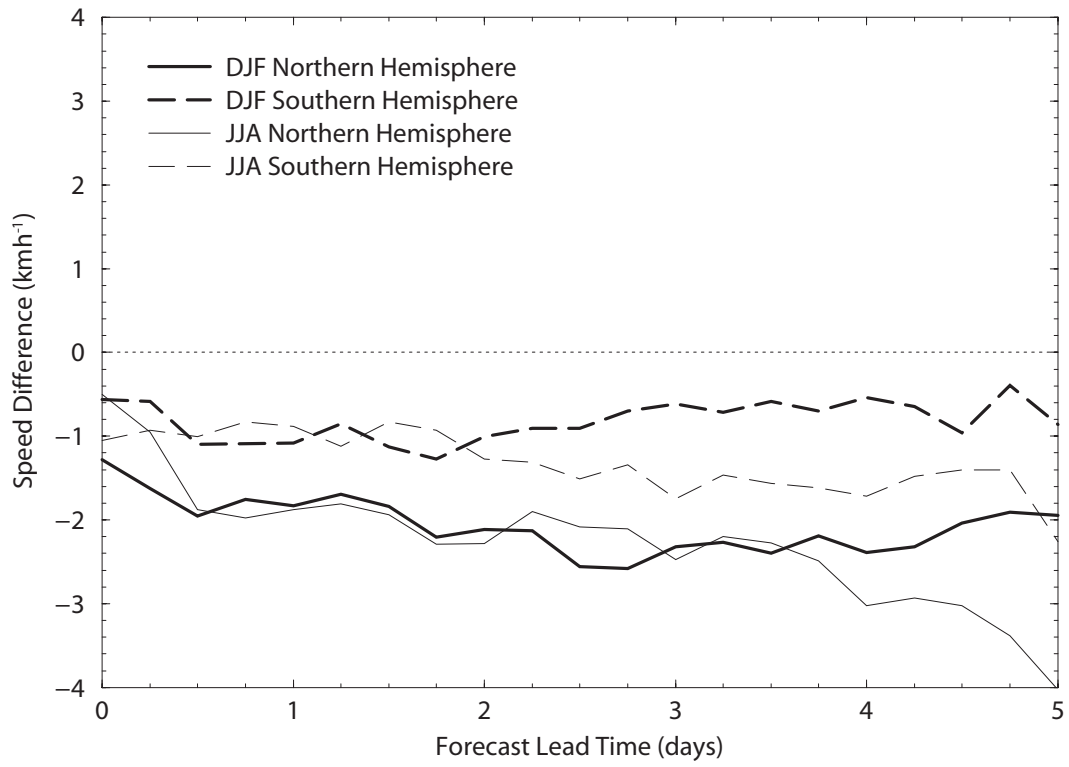


Figure 7. Mean speed difference between matched control forecast tracks and analysis tracks as a function of forecast lead time in the NH and SH for both the DJF and JJA seasons. Units of speed difference are kmh⁻¹.

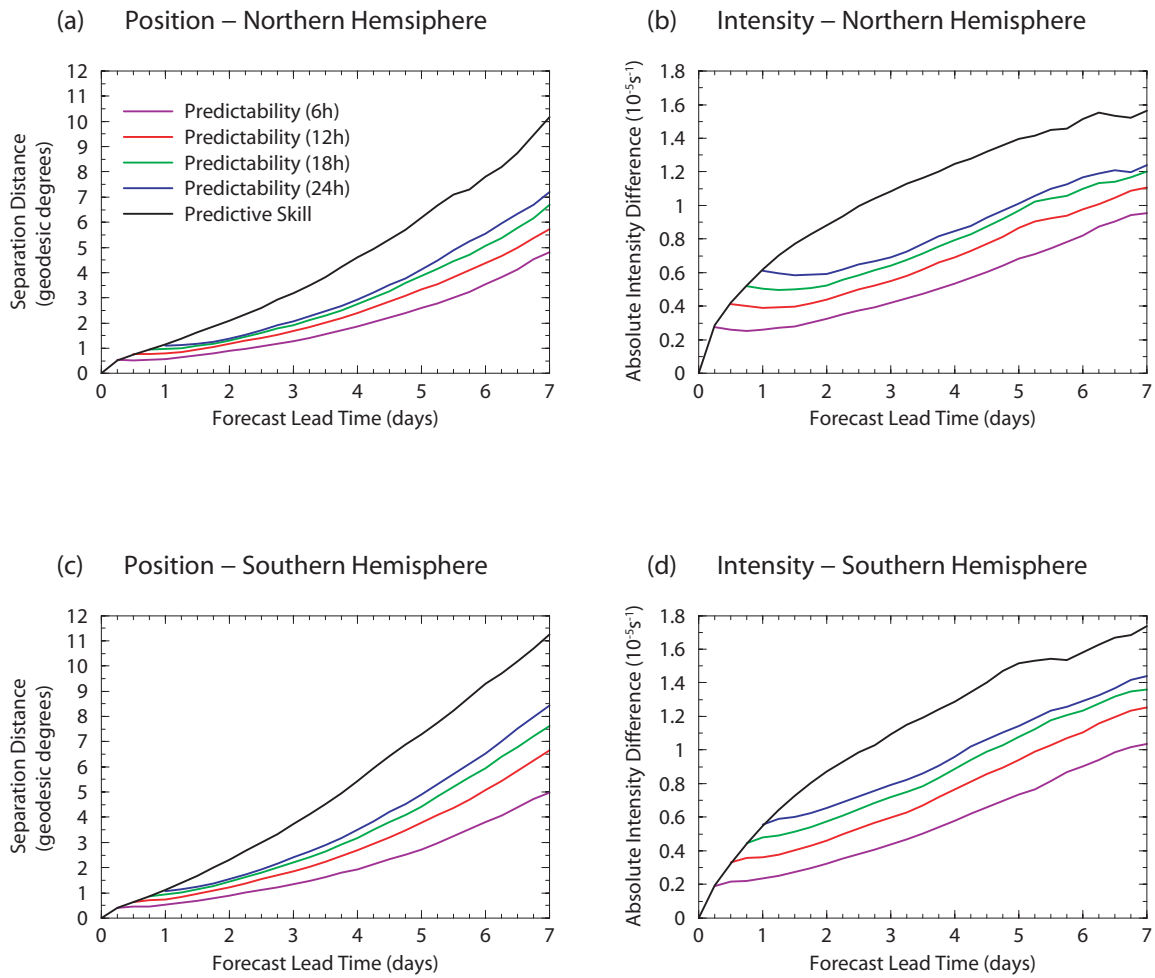


Figure 8. Predictability curves obtained with the control system for the DJF season by comparing forecast storm tracks separated by 6, 12, 18 and 24 hours (for details see text). The mean separation distance between the matched forecast tracks is shown, as a function of forecast lead time, in the NH (a) and SH (c). Panels (b) and (d) are as (a) and (c) but show the mean absolute intensity difference. The predictive skill curves of figure 5 are also included for comparison. Units of separation distance and intensity difference are geodesic degrees and 10^{-5}s^{-1} (relative to background field removal) respectively.

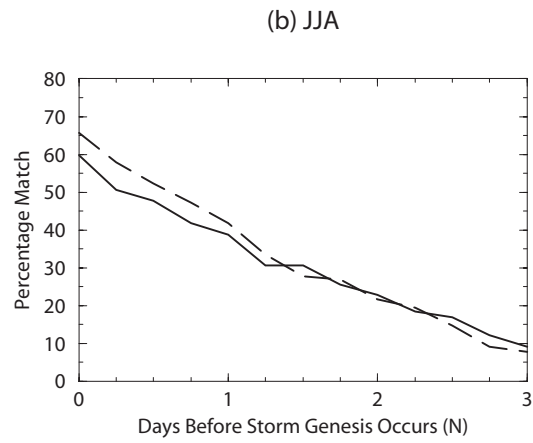
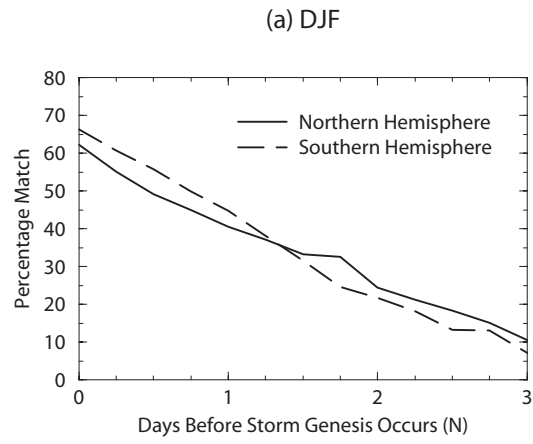


Figure 9. Plots showing the percentage of control analysis storm tracks predicted by the control system, as a function of the number of days (N) before the storms genesis occurred in the analysis, in both the NH and SH for the DJF (a) and JJA (b) seasons.

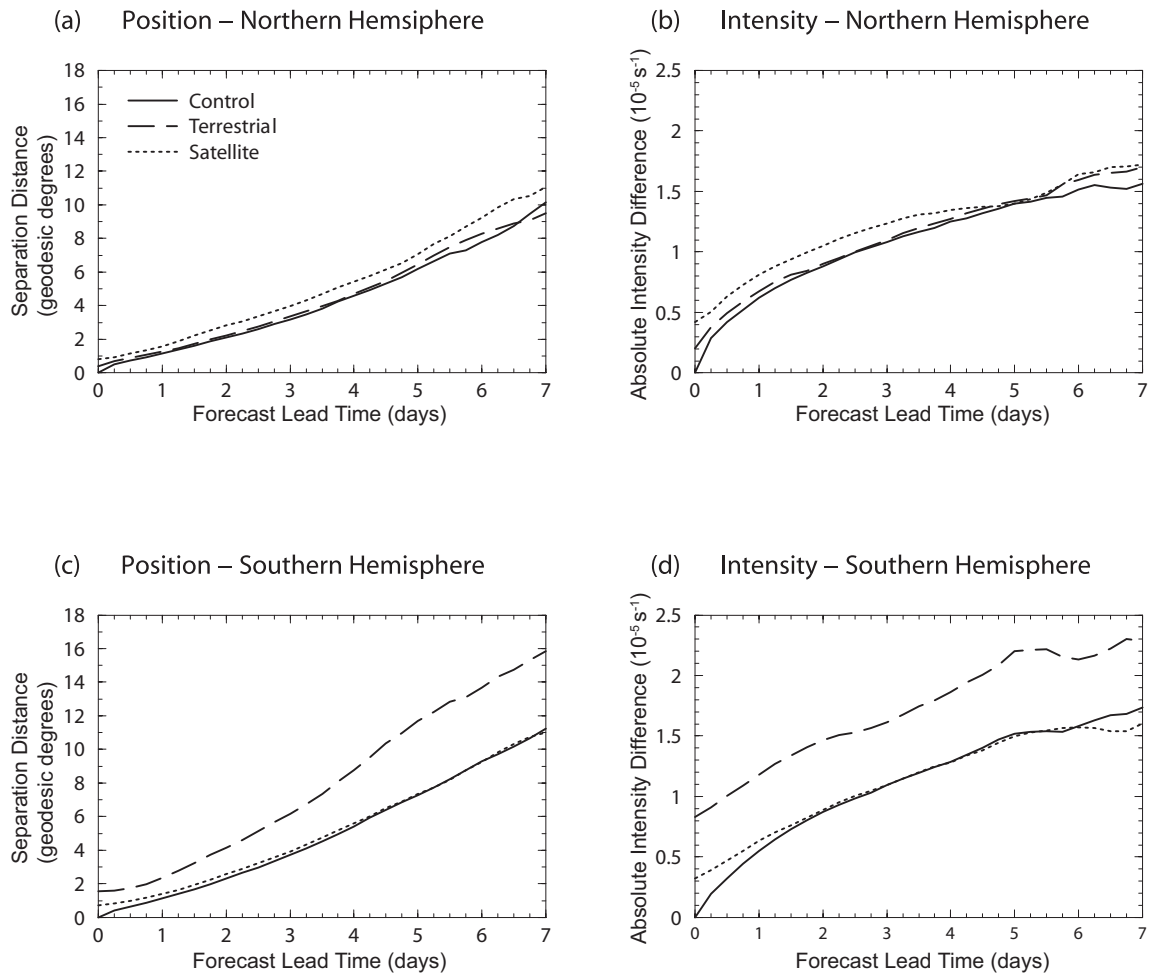


Figure 10. Mean separation distance between the matched control, terrestrial and satellite forecast tracks and the control analysis tracks as a function of forecast lead time in the NH (a) and SH (c) for the DJF seasons. Panels (b) and (d) are as (a) and (c) but show the mean absolute intensity difference. Units of separation distance and intensity difference are geodesic degrees and 10^{-5} s^{-1} (relative to background field removal) respectively.

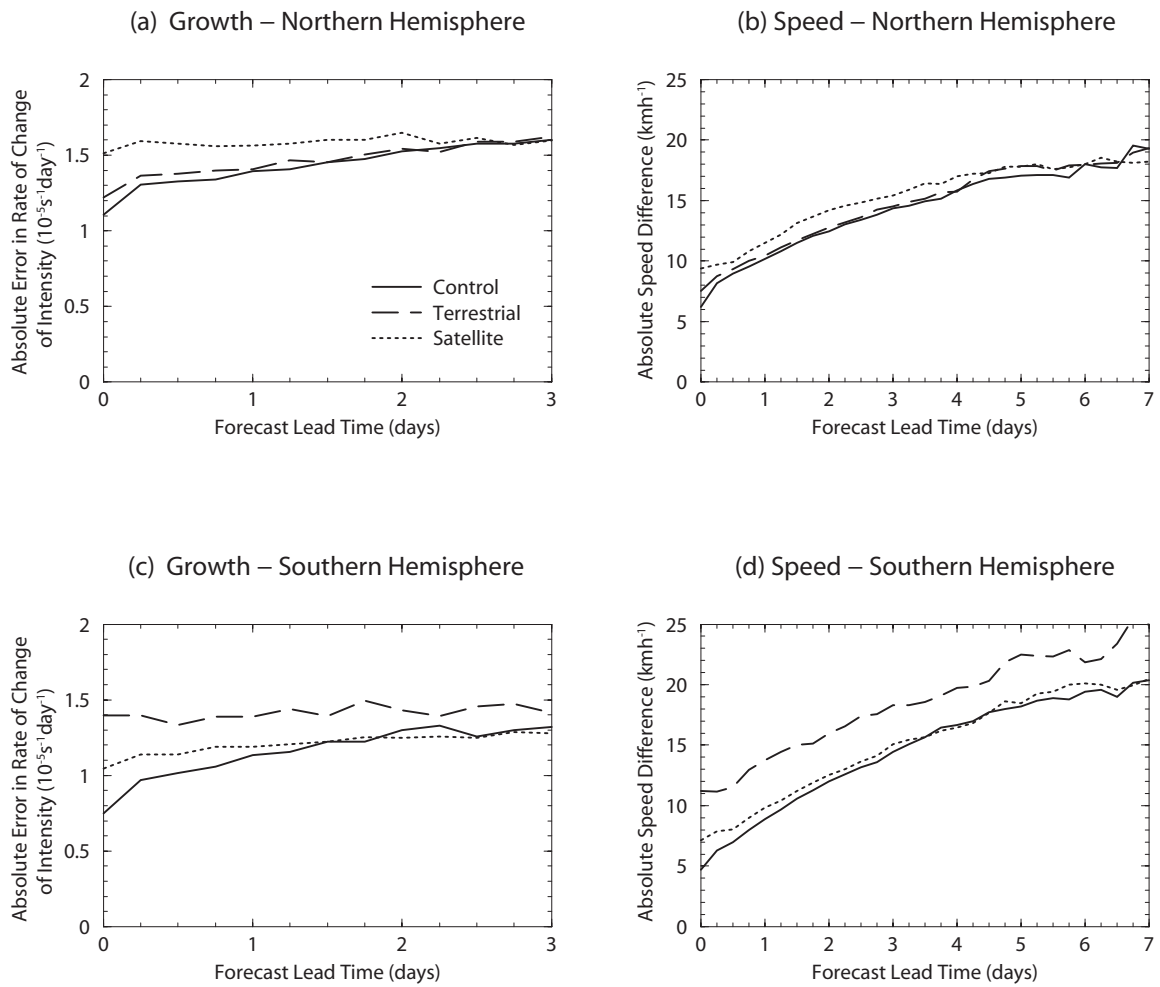


Figure 11. Mean absolute error in rate of change of intensity between the matched control, terrestrial and satellite forecast tracks and the control analysis tracks as a function of forecast lead time in the NH (a) and SH (c) for the DJF seasons. Panels (b) and (d) are as (a) and (c) but show the mean absolute speed difference. Units of rate of change of intensity and speed difference are $10^{-5} \text{ s}^{-1} \text{ day}^{-1}$ and kmh^{-1} respectively.