Tellus (2005), 57A, 515–527
Printed in Singapore. All rights reserved

TELLUS

Global observations and forecast skill

By LENNART BENGTSSON, KEVIN I. HODGES* and LIZZIE S. R. FROUDE, *Environmental Systems Science Centre (ESSC), University of Reading, Harry Pitt Building, Whiteknights, Reading, RG6 6AL, UK*

(Manuscript received 10 November 2004; in final form 4 February 2005)

ABSTRACT

The impact of selected observing systems on forecast skill is explored using the European Centre for Medium-Range Weather Forecasts (ECMWF) 40-yr reanalysis (ERA-40) system. Analyses have been produced for a surface-based observing system typical of the period prior to 1945/1950, a terrestrial-based observing system typical of the period 1950–1979 and a satellite-based observing system consisting of surface pressure and satellite observations. Global prediction experiments have been undertaken using these analyses as initial states, and which are available every 6 h, for the boreal winters of 1990/1991 and 2000/2001 and the summer of 2000, using a more recent version of the ECMWF model. The results show that for 500-hPa geopotential height, as a representative field, the terrestrial system in the Northern Hemisphere extratropics is only slightly inferior to the control system, which makes use of all observations for the analysis, and is also more accurate than the satellite system. There are indications that the skill of the terrestrial system worsens slightly and the satellite system improves somewhat between 1990/1991 and 2000/2001. The forecast skill in the Southern Hemisphere is dominated by the satellite information and this dominance is larger in the latter period. The overall skill is only slightly worse than that of the Northern Hemisphere. In the tropics (20°S–20°N), using the wind at 850 and 250 hPa as representative fields, the information content in the terrestrial and satellite systems is almost equal and complementary. The surface-based system has very limited skill restricted to the lower troposphere of the Northern Hemisphere. Predictability calculations show a potential for a further increase in predictive skill of 1–2 d in the extratropics of both hemispheres, but a potential for a major improvement of many days in the tropics. As well as the Eulerian perspective of predictability, the storm tracks have been calculated from all experiments and validated for the extratropics to provide a Lagrangian perspective.

1. Introduction

Systematic validation of numerical prediction has demonstrated that large-scale forecast skill has improved by 3–4 d over the last 25 yr (Bengtsson, 1999; Simmons and Hollingsworth, 2002). Repeating operational forecasts for past periods using reanalyses and recent forecasting models has shown that a considerable part of the improvements are due to better models and data assimilation systems. Using present models and data assimilation systems to produce predictions for the Northern Hemisphere (NH) extratropics for the pre-satellite period 1950–1970, for example, has resulted in the useful predictive skill being extended by about two additional days (Kistler et al., 2001). This demonstrates that the observing system consisting of radiosondes and surface observations had more potential than could be realized at the time. However, for the Southern Hemisphere (SH) the situation is different. Here the pre-satellite observing system was so poor that not even more accurate models and assimilation systems would be expected to improve the quality of the forecasts except marginally. For the SH, the huge improvements in predictive skill, now displaying virtually the same forecast quality as for the NH (Simmons and Hollingsworth, 2002), are mainly the result of satellite observations with global coverage.

In a recent study, Bengtsson et al. (2004a) described a series of data assimilation experiments based on the European Centre for Medium-Range Weather Forecasts (ECMWF) 40-yr reanalysis (ERA-40) system using different atmospheric observations, supposed to mimic typical meteorological observing systems for the last 100 yr. This was accomplished by systematically removing observations from the data base used in the ERA-40 reanalyses (Simmons and Gibson, 2000). The reduced systems included a surface system where only surface observations were used, a terrestrial system which used the radiosondes and aircraft observations in addition to the surface observations, a satellite system which used only satellite data apart from pressure data used to constrain the surface, and finally a control system where all available observations typical of the last 10 yr were used apart from humidity data (Bengtsson et al., 2004b; Bengtsson and Hodges, 2005). All of the subsystems were compared to the control. The assimilation study (Bengtsson et al., 2004a) demonstrated that the surface system had severe limitations in reconstituting the

*Corresponding author.
e-mail: kih@mail.nerc-essc.ac.uk

flow pattern in the upper troposphere and the stratosphere with the surface observations having only a minor impact in the NH and virtually zero impact in the SH. The terrestrial system deviated only little from the control for most of the NH, except for the central part of the Pacific Ocean, but had significant deficiencies over the ocean regions of the SH. The satellite-based system, on the other hand, was almost as accurate as the terrestrial system for the NH but was shown to dominate in the SH. However, a more detailed comparison of the terrestrial and the satellite systems in the NH showed that small-scale features, as represented by the vorticity field were better analysed by the terrestrial than by the satellite system. Another result was that it was not possible to analyse the quasi-biennial oscillation (QBO) from satellite-based data only, when the model itself cannot simulate a QBO. In this case, observations from the equatorial radiosondes are essential to provide the wind information not available from the satellites.

While an evaluation of the accuracy of the analyses provides useful information of the information content of the different observing systems, its value is nevertheless limited as we do not yet know how the initial errors will grow with time. To answer this question, forecast experiments are needed. The assessment and validation of such prediction experiments are the objectives of this study.

Here we report on a series of prediction experiments undertaken by using the assimilated data sets described in Bengtsson et al. (2004a) as the initial states. We have carried out a set of prediction experiments for each of the reduced observing systems consisting of 360 global forecasts of length 7 d and starting from every 6 h analysis.

There are three main objectives of this study. First, we wish to explore the predictive skill and how it depends on the different observing systems. Verification is undertaken with the control as well as different assimilated data sets in order to determine how much the skill assessment depends on the data sets used in the validation. This includes the use of National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis data to validate the predictions. This is done because of the practice in operational numerical weather prediction (NWP) to validate forecasts against analyses obtained using the same assimilation model. We have reasons to believe that this practice is biased towards higher skill scores, in particular in data sparse regions. (see, for example, fig. 7 of Kistler et al., 2001). Verification against station data is also possible but is often subject to the errors inherent in the observations, both random and systematic. It may also be of limited use in regions of poor observational coverage, e.g. the SH.

Secondly, we wish to revisit the issue of predictability along the same line as was done by Lorenz (1982) and recently by Simmons and Hollingsworth (2002). We extend this approach further by comparing how the predictions from different observing systems deviate from each other with time.

Thirdly, we wish to extend the concept of predictive skill and predictability to storm tracks (Hodges et al., 2003, 2004). While the traditional validation of geopotential fields provides a representative measure of the skill to predict the large-scale flow and the general weather type, validation of storm tracks displays the forecast skill of individual weather systems. This is a much tougher validation test, but is likely to better portray the model's capability to predict the weather, as this generally evolves along the storm track. The storm tracks will be evaluated in a Lagrangian sense, meaning that we will specifically identify and validate the trajectories of centres of surface pressure or centres of relative vorticity and how they evolve in time. Figure 1 shows examples from each hemisphere of analysed storm tracks and the storm amplitudes by the different observing systems based on the mean sea level pressure (MSLP) field.

The paper continues as follows. In Section 2 we explain the experimental set-up and the validation method, in Section 3 the results of the predictive skill and predictability, and in Section 4 the prediction of storm tracks. A summary of the results, a general discussion and some concluding remarks follow in Section 5.

2. Experimental set-up and diagnostics

Forecasts have been made for each of the reduced observing systems and the control using the six-hourly analyses produced for the study of Bengtsson et al. (2004a) as the initial states. The forecasts have been integrated forward in time to 7 d and selected fields archived daily. The model used for this study is the ECMWF operational Integrated Forecast Model (IFS; White, 2000), version 26R3. This is a spectral semi-Lagrangian model and was integrated at the same horizontal and vertical resolution of T159L60 as was used for the assimilation and ERA-40 to avoid interpolation problems and to restrict excessive use of computer time. This version is a later and further improved model than that used for the data assimilation (Bengtsson et al., 2004a). The periods covered by these experiments are 1 December 1990–28 February 1991, 1 June 2000–31 August 2000 and 1 December 2000–28 February 2001. As the results from the three periods are rather similar, we focus here on the first period, the winter of 1990/1991.

Diagnostics are produced from the forecast output based on both the Eulerian approach and the Lagrangian storm tracking of Hodges (1995, 1999). For the Eulerian approach, verification diagnostics are produced as root-mean-square (rms) errors by comparing the forecasts from the four experiments with the control analysis at the daily forecast intervals (forecasts for a particular time, e.g. day 1, are available as time series with six-hourly sampling, the same as the analyses). The rms is normalized by the standard deviation (StD) of the control analysis. As well as the predictive skill (verification) the predictability is also computed as described by Lorenz (1982). This compares forecasts separated in time, for example, the 2-d forecasts are verified against the 1-d forecasts, the 3-d against the 2-d and so on. This is done for the control only and rms diagnostics are produced as before.

The Lagrangian diagnostics follow a similar approach to the Eulerian approach using the six-hourly data for each forecast

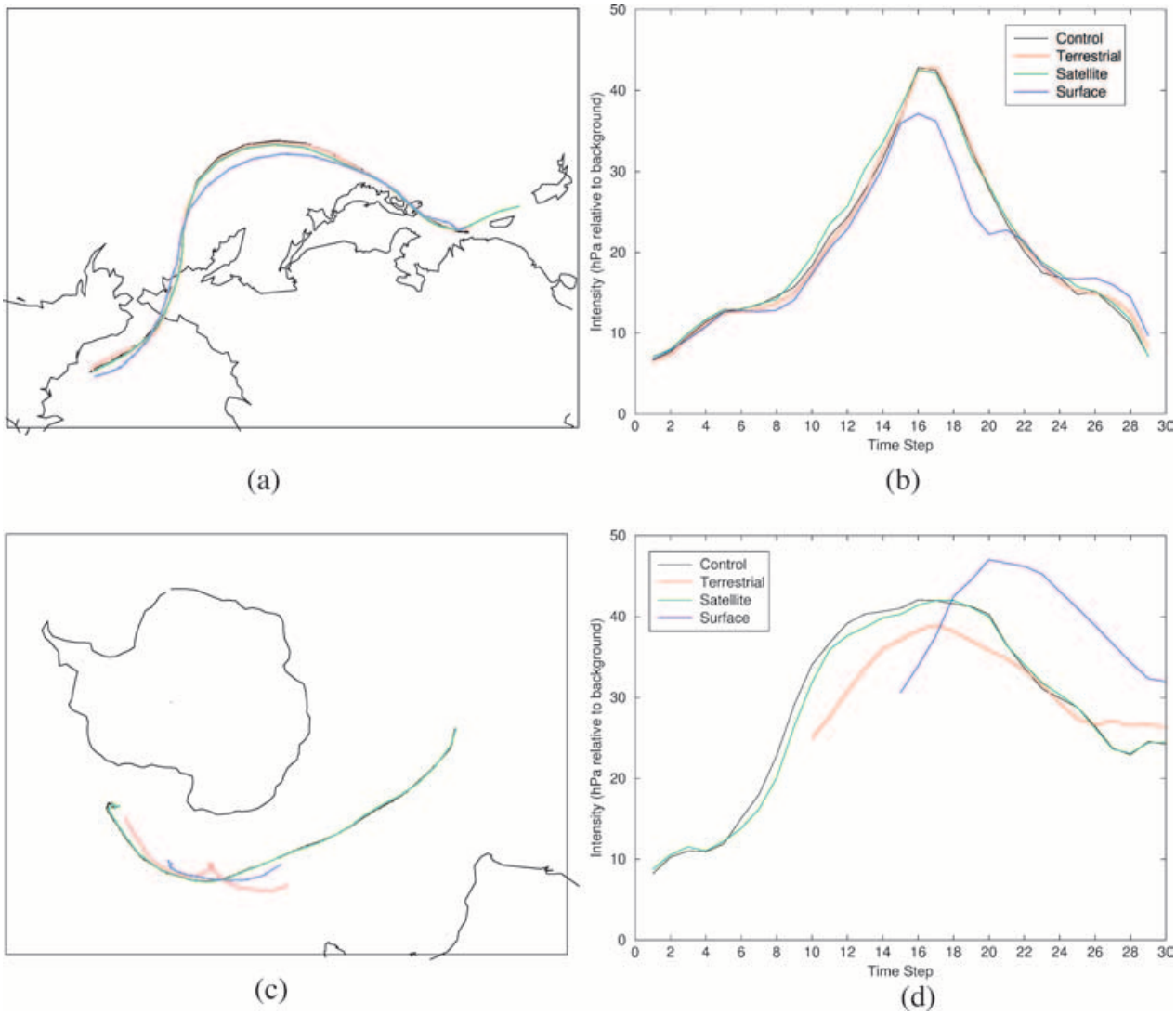

Fig. 1. Example storm tracks of extratropical cyclones over the North Pacific and SH as analysed by the control (black), terrestrial (red), satellite (green) and surface system (blue), respectively. For the North Pacific storm, (a) shows the trajectories for the different analyses and (b) the intensity in hPa versus the analysis time-step (6 h). For the SH storm, (c) shows the trajectories for the different analyses and (b) the intensity in hPa versus the analysis time-step. Intensities are relative to the removed background and converted to positive values.

instance, i.e. 1 d, 2 d, etc., to identify and track extratropical cyclones using the MSLP and relative vorticity at 850-hPa fields. The method used follows that of Hoskins and Hodges (2002) except no spatial statistics are computed; instead, the track ensembles are compared directly using the approach described in Bengtsson et al. (2004a) and Hodges et al. (2003, 2004) to produce distributions for the percentage of tracks that match between those in the control analysis and the forecasts. The comparison is made in the same way as for the Eulerian approach, with the forecast track ensembles verified against the control analysis track ensemble and the predictability estimated by verifying the 2-d track ensembles against the 1-d and so on. Ideally, if the forecast data were archived every 6 h, the tracking could also be performed as the forecast progresses; this will be explored in the future. A schematic diagram of the forecasts and how they are compared is shown in Fig. 2.

3. Predictability and predictive skill

We first assess the quality of the predictions from the four different data sets by considering the normalized rms error of the 500-hPa geopotential height field as a function of time; this is shown in Fig. 3. For the NH extratropics (Fig. 3a), the error of the control forecast increases linearly, reaching a value of around 0.85 at day 7. The forecast errors for the terrestrial system are slightly larger with the error curve virtually parallel to that of the control experiment. The reduction in predictive skill is around 6 h at any time of the whole forecast period. The forecast error

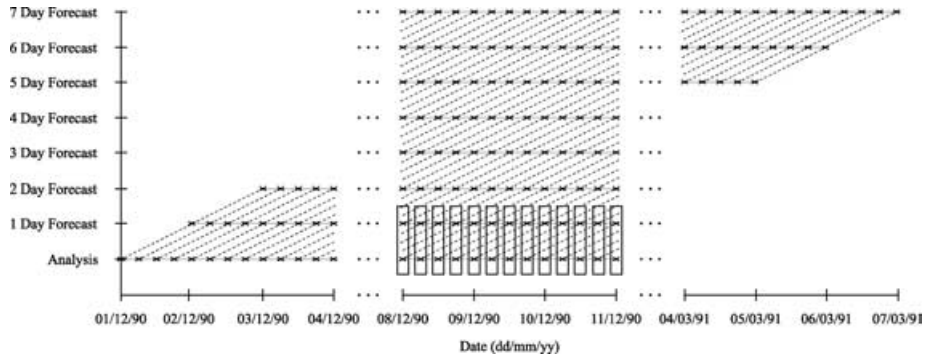

Fig. 2. Illustration of how the storm track ensembles were computed and compared. Each cross represents one time frame of data. The analysis data consists of six-hourly time frames for each experiment. The forecast model was run from each of these time frames to compute 1-, 2-, . . ., 7-d forecasts. This is illustrated by the diagonal dashed lines. The 1-d forecast data set consists of six-hourly time frames for the period 2 December 1990–1 March 1991, for example, the 2-d forecast data set consisted of six-hourly time frames for the period 3 December 1990–2 March 1991, and so on. These data sets are represented by the horizontal grey lines in the diagram. The storm tracking program was applied to these data sets (including the analysis). A storm track consists of a set of points in successive time frames. Two ensembles of tracks, *A* and *B*, were compared by computing the percentage of tracks in ensemble *A* that were also in ensemble *B*. A track in ensemble *A* was assumed to be the same as a track in ensemble *B* if the two tracks overlapped a certain amount in time and space. The vertical boxes show the pairs of points that are compared when a storm track in the analysis ensemble was compared with a storm track in the 1-d forecast ensemble.

curve of the satellite system is also parallel to that of the control experiment but with a larger reduction in skill corresponding to about 24 h through the forecast period. The forecast errors from the surface-based system are significantly larger, having an initial error of the same size as the control forecast at day 6, and is consequently of very limited value. Finally, we have estimated the predictability in the same way as Lorenz (1982) by comparing successive forecasts separated by 1 d; Fig. 3a shows this for the control experiment. The result suggests that there is still potential for a further improvement in forecast skill by about 2 d (the error at day 7 is the same as the error at day 5). The estimation of predictability is independent of the data set (not shown), meaning that the predictability estimate from the terrestrial and satellite experiments is the same as for the control.

Figure 3b shows the normalized rms errors for the SH extratropics. The errors for the control forecasts increase with time almost at the same rate as for the NH. The forecasts for the satellite system have the next lowest errors, having a reduced skill of about 12 h compared to the control. The forecasts for the terrestrial system are poor with an initial error equivalent to the control at day 3.5. Forecasts from the surface-based system are already initially larger than 1.0 and have little forecast skill.

Figure 4 shows the verification of the 500-hPa geopotential height for four subregions: North America (15°N–70°N, 65°W–160°W), continental United States (25°N–55°N, 65°W–125°W), Europe (35°N–75°N, 12.5°W–42.5°E) and Australia–New Zealand (45°S–12.5°S, 120°E–175°E). These regions have reliable observational networks and the analysed state is virtually unbiased vis-à-vis the forecasting model. The results for the three areas of the NH are consistent with the hemispheric results. For the European region, the terrestrial system is almost identical to the control, suggesting that it may be the contribution from aircraft data as well as downstream propagation of information from North America which contribute to the impact of the terrestrial system. Alternatively, it could be an effect of reduced satellite information generally providing less information at higher latitudes. The surface-based system provides modest information for the first 2 d or so, and more for levels below 500 hPa (not shown). The forecast error from the surface-based system grows fast and provides no useful information beyond day 3. For the Australian–New Zealand region, the control experiment provides useful information until day 7. The terrestrial and the satellite systems have initially the same error but thereafter the error from the terrestrial system grows faster and at day 4 the skill of the satellite system is the same as the terrestrial system at day 3. It is suggested that the reason for this is the increasing influence from regions upstream where terrestrial data are few but satellite data abundant. The results for North America and continental United States are very similar to those for Europe.

Conventionally, in NWP, operational forecasts are verified against analyses which are produced using the same model in the data assimilation. This means that in areas with insufficient observations model information determines the initial state to a considerable degree, creating an artificial bias in reducing the real forecast error. This has nothing to do with forecast skill but is just an artefact of the assimilation procedure. In the extreme case of no observations, the forecast would be perfect for all times. The effect is minor in data dense regions but considerable when observations are sparse. Figure 5 shows the verification of the 500-hPa height forecasts from the terrestrial system for the two hemispheres, verified against both the control and the terrestrial analyses. Verifying the NH forecasts against the terrestrial system only affects the first day or so, while the effect at the SH is significant. The SH has a large growth rate, tending towards the error of the verification against the control at day 7. This just reflects the poor terrestrial data coverage in the SH and

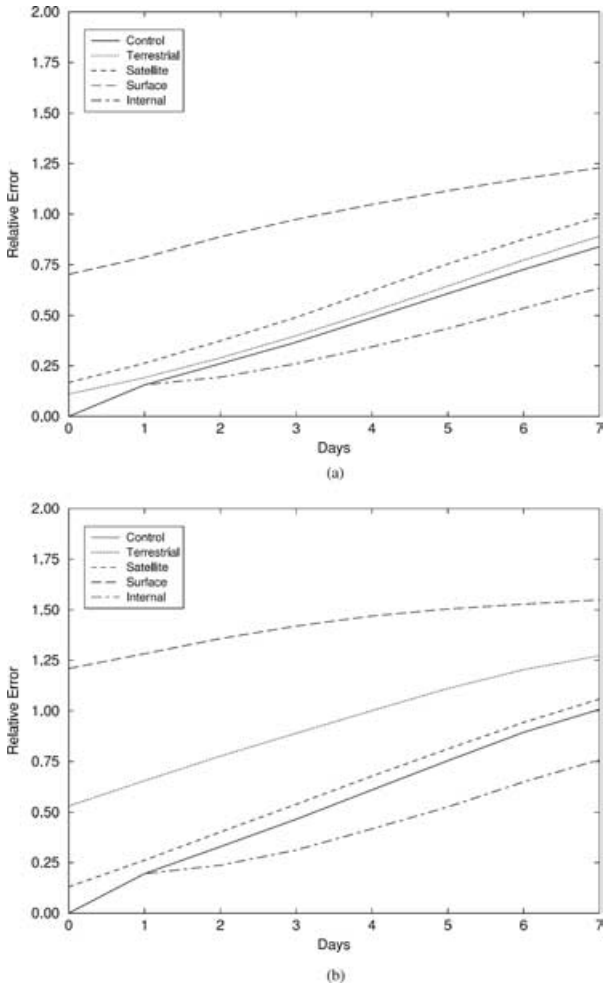

Fig. 3. Normalized forecast skill for the 500-hPa geopotential height for DJF 1990/1991 for the four experiments in (a) the NH extratropics (20°N–90°N) and (b) the SH extratropics (20°S–90°S): control (full line), terrestrial (dotted), satellite (short-dashed) and surface (long-dashed). The predictability is indicated by the dash-dotted line. Predictability has been estimated in the same way as in Lorenz (1982).

the fact that the assimilation system is not in an optimal configuration for these sparse observations. The initial error when verified against the control is larger than the forecast error at day 2 when verified against the terrestrial system. It is this circumstance that is the cause of the strange result in Kistler et al. (2001) where the forecasts in the SH are actually more accurate in the 1950s than 20 yr later. At the same time, we may conclude that forecast skill in general is likely to have improved more than previously recorded, because verification of earlier forecasts against present more accurate initial states are found to be less accurate than when verified against past analyses. In a similar way, we have in Fig. 6 also verified the control forecasts using the NCEP reanalyses. For the NH, the difference falls below 10% around day 2 and becomes insignificant at day 6 (less than 1%). However, for the SH, day 2 validated with the control is a more correct forecast than day 1 validated with NCEP. For longer forecasts, the verification bias diminishes gradually though, and at day 6 the difference is less than 5%.

For the tropics (20°S–20°N), we have verified the wind at 850 and 250 hPa, respectively, as neither the temperature nor the geopotential fields are suitable to describe tropical predictability and predictive skill. This is shown in Fig. 7. This shows the initial error at 850 hPa for both satellite and terrestrial systems to be ∼2 and 2.5 m s^{-1} at day 1, respectively. For a summer period (JJA 2000, not shown) this increases to 2.5 and 3.1 m s^{-1}, respectively. It is interesting to note that the satellite and terrestrial systems have almost identical skill at both levels. They are also apparently complementary, as the forecasts of the control system are clearly better than the forecasts from the reduced observing systems. It is interesting to note the proportionally large impact of the relatively few tropical radiosondes. Generally, the error growth is slow in the control, terrestrial and satellite systems, increasing linearly and with the same rate after 3–4 d into the forecast. The forecasts of the surface-based system have hardly any skill except for a day or so at 850 hPa only.

An estimate of the predictability undertaken in the same way as for the extratropical 500-hPa height (see above) indicates an even slower error growth rate, suggesting the potential for major improvements in the tropical forecasts. The increase in error between days 3 and 7 is only about 25%.

A comparison of the forecasts for DJF 2000/2001 (Fig. 8) suggest a slight deterioration of the terrestrial system in both hemispheres, but with virtually unchanged skill for the forecasts of the satellite-based systems. The error growth of the control system is slightly larger in the latter period, indicating lower predictability. However, it is probably not possible, in view of the small differences, to conclude whether this is representative or just an artefact of natural variations in the atmospheric circulation. However, the reduction in the number of radiosondes during the 1990s cannot be ruled out as a contributing factor (Bengtsson et al., 2004a). Apparently, if this is the case, the increase in aircraft observations may not be able to compensate for the reduction in the number of radiosondes. The regional verification shows differences between 1990/1991 and 2000/2001 but without any consistent trend (not shown). The latter period has higher skill for Europe and the earlier period has higher skill for North America. This includes all prediction and predictability scores (not shown).

Lorenz (1982) designed an elegant way to determine an upper bound on atmospheric predictability by making use of the daily archived analyses and forecasts at the ECMWF for a period of 100 d. This was done by considering the difference between the analysis for a given day and the 1-d forecast valid for the same day as a suitable perturbation. The growth of that perturbation was then examined by comparing the 1-d forecast with the 2-d forecast from the preceding day and so on until day 10. The error at day 1, 25 m, required 3.5 d to double, while larger errors amplified less rapidly. Recently, Simmons and Hollingsworth (2002) have repeated the Lorenz calculation for a series of more recent winter and summer periods. They found the error growth in the

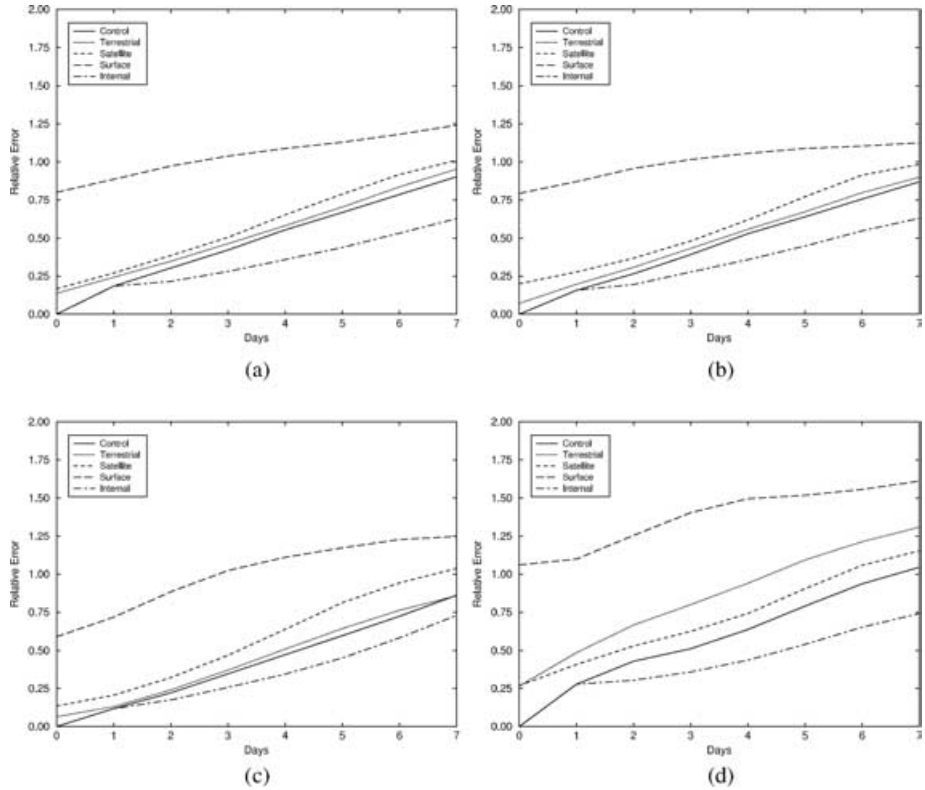

Fig. 4. Normalized forecast skill for the 500-hPa geopotential height for four subregions for DJF 1990/1991 for the four experiments: (a) North America (15°N–70°N, 65°W–165°W); (b) Continental US (25°N–55°N, 65°W–125°W); (c) Europe (35°N–75°N, 12.5°W–42.5°E); (d) Australia/New Zealand (45°S–12.5°S, 120°E–175°E). Control (full line), terrestrial (dotted), satellite (short-dashed), surface (long-dashed), and predictability (dash-dotted).

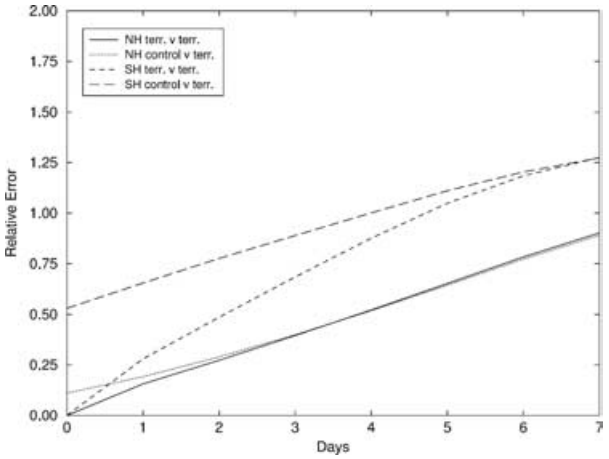

Fig. 5. Normalized forecast skill for the 500-hPa geopotential height for the extratropics of both hemispheres for DJF 1990/1991 of the terrestrial experiment. NH (20°N–90°N) verified versus the terrestrial system (full line), NH verified versus the control (dotted line), SH (20°S–90°S) verified versus the terrestrial system (short-dashed lines) and SH verified versus the control analyses (long-dashed lines).

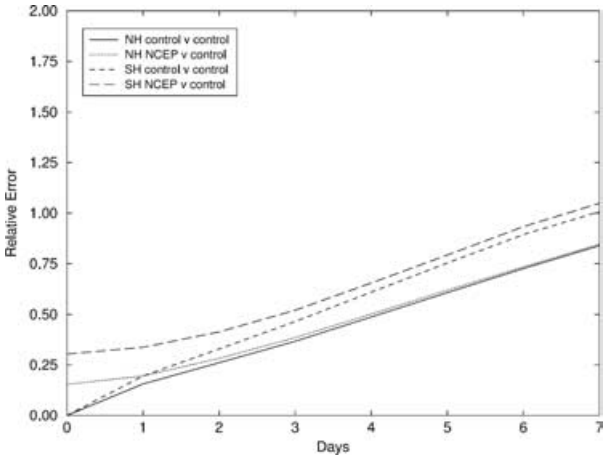

Fig. 6. Normalized forecast skill for the 500-hPa geopotential height for the extratropics of both hemispheres for DJF 1990/1991 of the control experiment. NH (20°N–90°N) verified versus the control (full line), NH verified versus the NCEP reanalyses (dotted line), SH (20°S–90°S) verified versus the control (short-dashed lines) and SH verified versus the NCEP reanalyses (long-dashed lines).

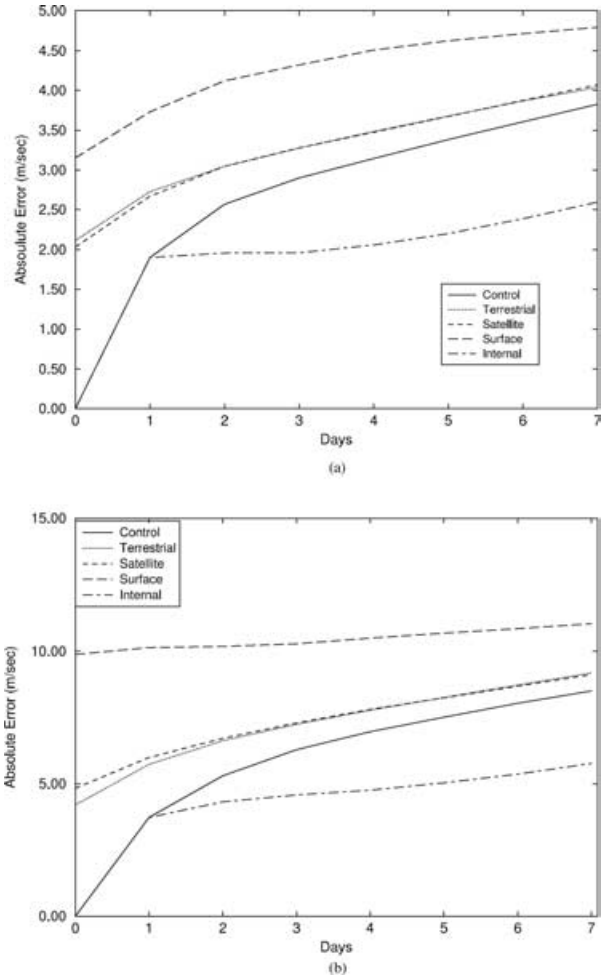

Fig. 7. Forecast skill for the tropics (20°S–20°N) for DJF 1990/1991 for the four experiments using winds for (a) 850 hPa and (b) 250 hPa. Control (full line), terrestrial (dotted), satellite (short-dashed) and surface (long-dashed). The predictability is indicated by the dash-dotted line. Predictability has been estimated in the same way as in Lorenz (1982).

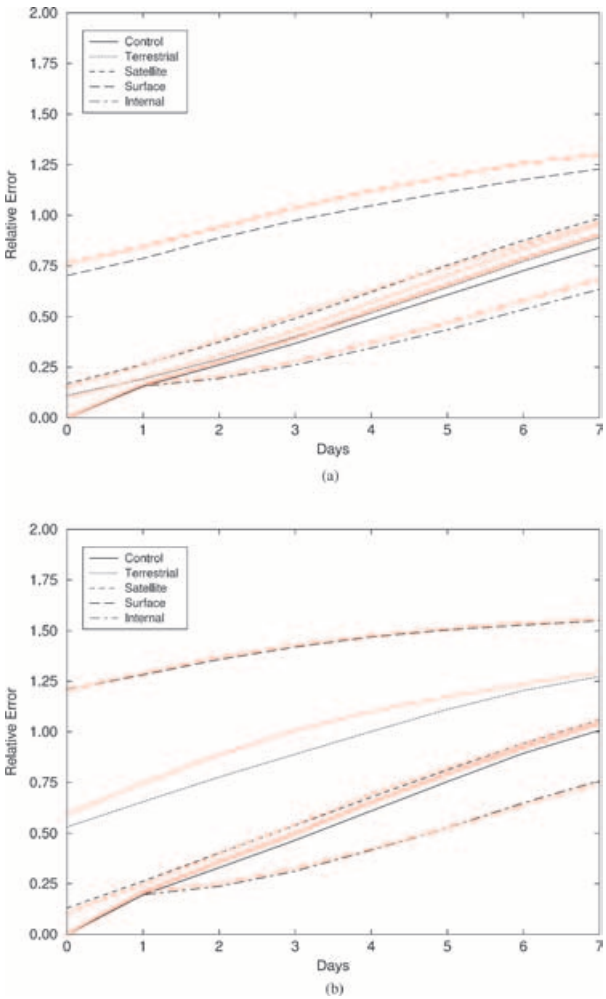

Fig. 8. The same as Fig. 3 with the superimposed results for DJF 2000/2001 inserted in red for (a) the NH and (b) the SH. The same notation are used for the curves in red.

early part of the forecasts was much faster, the error doubling in 1.3 d, but from a significantly smaller initial error at day 1 of about 11 m. By the time the error has reached 25 m, the doubling time is around 3 d in broad agreement with Lorenz early results. Here we have undertaken the calculation for extratropical domains poleward of 20°N and 20°S latitude, which is a slightly smaller area compared to the earlier studies referred to above. The three systems (control, terrestrial and satellite) have the same internal growth in both hemispheres corresponding to a doubling time of 1.4 at day 2 increasing to 2.8 at day 5 (not shown). The predictability is the same in both hemispheres and the same for the two periods 1990/1991 and 2000/2001 (see Fig. 8). The error doubling time is similar to what was found by Simmons and Hollingsworth (2002) but much shorter than in Lorenz (1982), demonstrating that when the initial error is further reduced the internal error growth rate increases. The predictability curves in the figures are calculated for the control integration.

4. Prediction and predictability of storm tracks

Storm tracks are suitable features to verify as they could be considered as substitutes for important weather elements. Extratropical weather patterns, clouds, precipitation, etc., are mostly coupled to transient vorticity centres where they evolve in a characteristic way during the lifetime of the cyclone. We have here undertaken a systematic calculation of storm tracks and verified them. This has been done by comparing each individual storm track identified in the control analysis with the storm tracks identified in sequences of the simultaneously valid forecasts (see Fig. 2). This is done by performing the cyclone tracking on an ensemble of forecast fields separated by 6 h for each forecast day, i.e. the crosses in Fig. 2. This is the closest analogy with the verification in the preceding sections for the storm tracks. Alternatively, it would have been possible to verify each

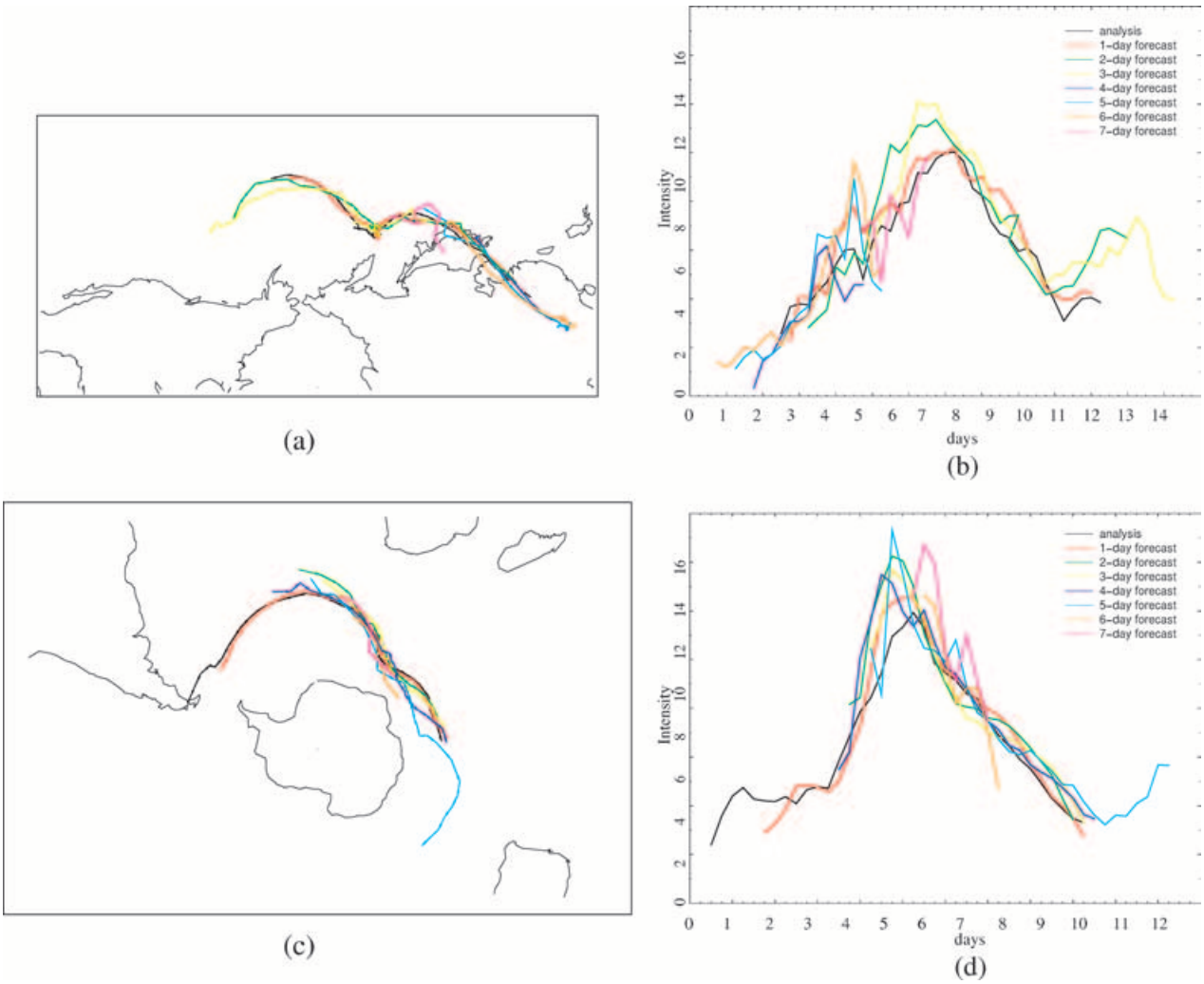

Fig. 9. Analysed and predicted storm tracks (ξ_{850}) from the control of an extratropical cyclone over eastern Asia and the North Pacific. Analysed (black), 1-d forecast (red), 2-d forecast (green), 3-d forecast (yellow), 4-d forecast (dark blue), 5-d forecast (light blue), 6-d forecast (orange) and 7-d forecast (pink). (a) The trajectories of the storm and (b) the intensity is in units of 10^{-5} s^{-1} relative to the background. Similarly, the analysed and predicted tracks for a SH storm are shown: (c) the trajectories and (b) the intensities converted to positive values.

particular forecast storm track (tracking performed along the forecast trajectory) against observations of the actual storm track (performing the tracking in fields represented by the dashed lines in Fig. 2), but this would have necessitated verifying the storm tracks for each successive 6 h, leading to excessive data handling requirements. However, we do not expect this to have any major effect on the overall assessment of the predictive skill of the storm tracks, although this approach may constitute a lower bound on storm track predictive skill. This second approach to verifying the forecast storm tracks constitutes ongoing work.

Figures 9a and b show examples of the analysed and predicted storm track and amplitude, respectively, for an intense storm which developed on the eastern slope of the Tibetan plateau moving eastward over China, Japan and into the northern part of the Pacific Ocean. Here we use the relative vorticity at 850 hPa (ξ_{850}) as an indicator of the storm track. The ξ_{850} analysis track reaches a maximum of 10×10^{-5} s^{-1} around time-step 25 (each time-step is 6 h) after which it gradually weakens and loses its identity some 20 time-steps (5 d) later. The 1-, 2-, 3- and 4-d forecasts reproduce the overall evolution rather well, while the later forecasts only reproduce parts of the storm track and its amplitude as the forecasts drift away from the analysis. This is to be expected, as each forecast for a particular forecast day is started from different initial conditions, resulting in the storm propagation appearing less smooth for the forecasts furthest away from the analysis. Whilst the tracking parameters could be relaxed to accommodate some of the loss of smoothness in the storm propagation, this would introduce subjectivity into the diagnostics; hence, the tracking parameters are kept the same as those used for the tracking of storms in the analyses. Both the forecasts of

Table 1. Verification of storm tracks for the control experiment using MSLP and ξ_{850}. For further information see text and Fig. 2

			Percentage of tracks in the analysis that match tracks in the forecast					
Forecast	Total number of tracks in ensemble		2° mean separation 60% time		2° min separation 60% time		2° min separation 30% time	
NH	MSLP	VOR850	MSLP	VOR850	MSLP	VOR850	MSLP	VOR850
Analysis	169	425						
1 d	167	428	72.2	55.8	82.2	66.6	84.6	72.0
2 d	158	348	37.9	23.8	66.3	48.0	71.6	56.2
3 d	153	325	12.4	6.4	50.9	35.3	58.0	46.6
4 d	107	223	1.2	0.9	24.9	15.5	34.3	22.6
5 d	76	143	0.0	0.5	13.0	7.5	20.7	12.9
6 d	43	95	0.0	0.2	5.3	4.9	8.3	7.5
7 d	14	53	0.0	0.0	0.0	1.4	1.8	2.8
SH	MSLP	VOR850	MSLP	VOR850	MSLP	VOR850	MSLP	VOR850
Analysis	158	364						
1 d	159	376	75.3	61.8	82.9	71.7	84.2	78.6
2 d	154	365	30.4	24.7	62.7	55.8	69.0	64.0
3 d	158	326	8.2	4.7	47.5	36.8	55.1	48.4
4 d	119	250	0.6	0.5	24.1	20.9	34.8	30.2
5 d	83	167	0.0	0.0	12.7	7.8	19.0	15.9
6 d	45	85	0.6	0.0	3.8	3.3	8.2	7.7
7 d	31	52	0.0	0.0	3.8	1.1	7.0	4.1

the track and the amplitude are well within a rather confined envelope for each forecast day. Figures 9c and d show the same diagnostics for an intense storm at a high latitude of the SH. Following the initial development at the most southern tip of South America, it undergoes a rapid development between time-steps 12 and 20, decreasing the ξ_{850} from about -5×10^{-5} s^{-1} to about -12×10^{-5} s^{-1}. It thereafter weakens over the next 4 d. Also, in this case the ensemble of 1, 2, 3 and 4 d describes the evolution rather well, suggesting an even more rapid deepening rate of the cyclone. As in the case from the NH, the track and the amplitude forecasts for all forecast days are within a well-confined envelope.

To provide diagnostics analogous to those of the previous Eulerian verification analysis, the same approach as that used in Bengtsson et al. (2004a) is applied to produce distributions for the percentage of tracks that match between those identified in the control analysis and those identified in the forecasts for the different observing systems. Because of the difficulty of precisely identifying the storm tracks in the ensemble of forecast days, we have used three levels of matching criteria: (i) the mean separation distance between tracks is less than 2° (mean geodetic distance for those points that correspond) and they must overlap in time for 60% of their points; (ii) the minimum separation distance between pairs of tracks is 2° with a 60% overlap of the points in time; (iii) the same as (ii) but with an overlap of 30% of the points in time (these are the same as in Bengtsson et al., 2004a). This is illustrated in Table 1 for the control experiment for both MSLP and ξ_{850} fields. Figures 10 and 11 show for MSLP and ξ_{850}, respectively, for each hemisphere and each of the three matching criteria, the percentage of matching tracks as a function of the forecast time for each of the four experiments, including the estimated predictability calculated analogous to Lorenz (1982). The number of matching pairs of storm tracks falls off rather fast due to the decrease in coherence of the storms with forecast time, but the overall result shows the same relative relation between the experiments as was found in the verification of the 500-hPa height. However, these results do not show the whole picture; the reduction in the number of storms with forecast time falls fastest, due to increased lack of propagation smoothness, for the weakest systems. Distributions based on the mean intensity of storms show that the tails of the distributions at the high-intensity end fall much slower than the parts of the distributions representing the weaker systems. The contrast in the results between those for MSLP and ξ_{850} highlights the importance of spatial scale when discussing predictive skill. MSLP tends to focus on systems at the large-scale end of the synoptic-scale range whilst ξ_{850} tends to focus on the smaller-scale end of the range.

We note that both the predictability as well as the predictive skill of the storm tracks is significantly less than that seen in the preceding sections and reaches only about 2 d for ξ_{850} in the control and 3 d in the MSLP, taking 50% as the useful limit of predictive skill and the middle set of matching criteria. Using the less stringent third set of criteria for coinciding storm tracks, the

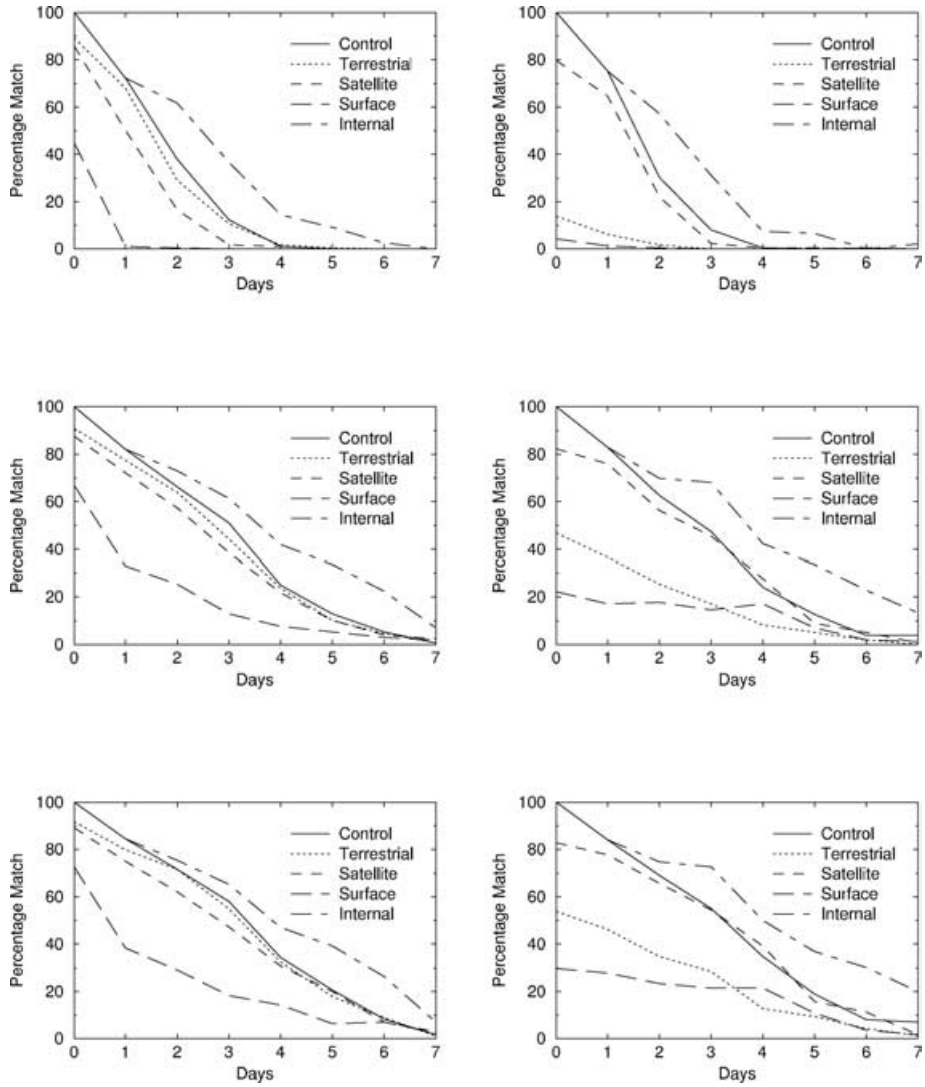

Fig. 10. Percentage of matching tracks between analysed storm tracks and forecast tracks as a function of time (for further explanation, see Fig. 2) for MSLP. Three levels of matching criteria have been used: (top) the mean separation between tracks is less than 2° (geodetic distance) and this is true for 60% of the points; (middle) the minimum separation is 2°, also true for 60% of the points; (bottom) the same as (top) but only valid for 30% of the points. The figures to the left show the result for the NH and figures to the right for the SH. Control (full line), terrestrial (dotted), satellite (short-dashed) and surface (long-dashed). The predictability is indicated by the dash-dotted line. Predictability has been estimated analogous to Lorenz (1982).

predictive skill is extended to nearly 3 d for ξ_{850} in the control and 3.5 d in the MSLP. It is interesting to note that the predictability of the storm tracks is somewhat higher for the SH than for the NH, both for the control and for the satellite-based forecasts and for both MSLP and ξ_{850}. We believe this is due to the higher consistency in the storm tracks at the SH being less disrupted by changing surface boundary conditions. The satellite information hardly adds any information to the NH, and the terrestrial system hardly adds anything for the SH. However, there is nevertheless scope for considerable improvements through better models by more than a day based on the predictability estimate.

5. Discussion

5.1. Extratropical predictive skill

The predictive skill of the control system is essentially identical to the ERA-40 (Bengtsson and Hodges, 2005). The predictive skill of 500-hPa height, as measured by the normalized error, is larger in the NH than the SH, but the error level in m at day 7 (103 and 83 m, respectively) is about the same as the variance in DJF 1990/1991. The skill for the regional areas (Europe, continental US and Australia) is broadly typical for respective hemispheric

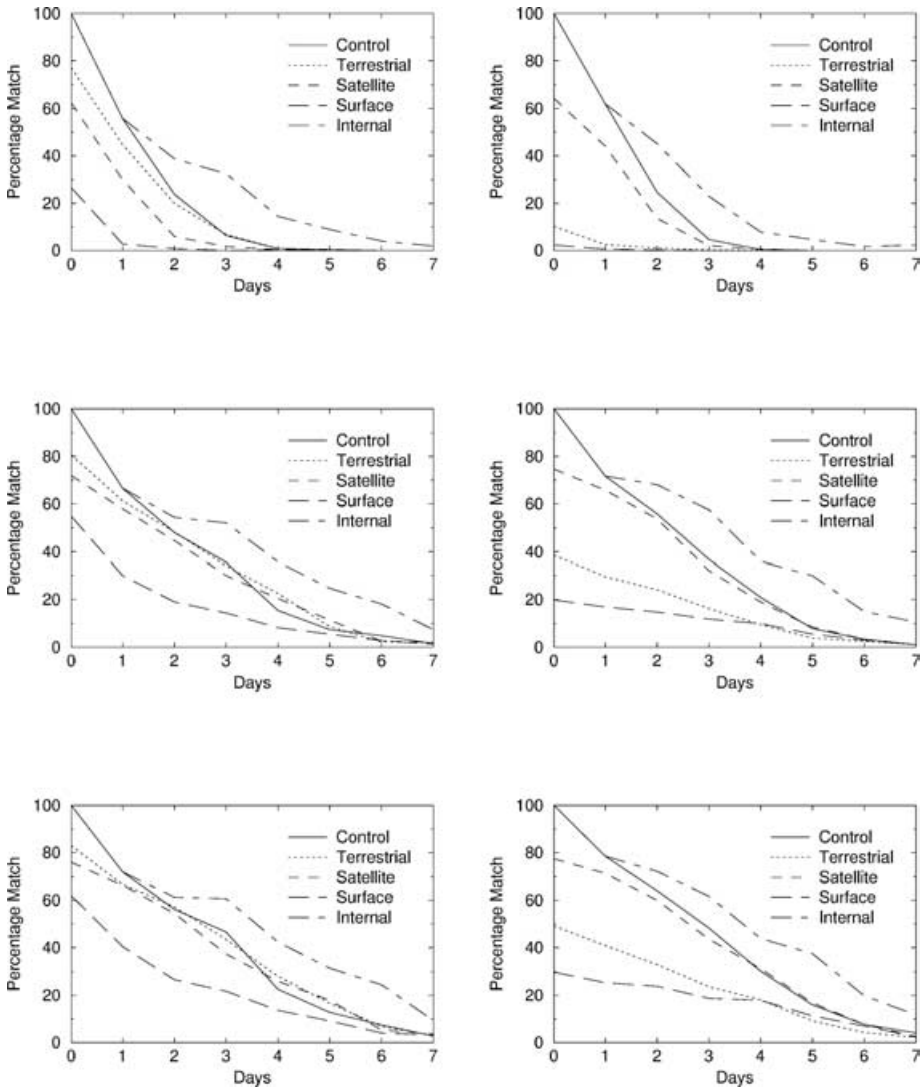

Fig. 11. Same as Fig. 10 except for ξ_{850}.

averages. There is a bias towards higher verification scores when verifying against analyses by the same data assimilation system, as evident from Fig. 6 where we have verified the control forecasts with the NCEP reanalyses. Ideally, the forecasts could also be compared with station data and radiosondes to provide some verification independent of the data assimilation. However, this has its own uncertainties in the form of both random and systematic bias errors in the observations. The forecast scores for the control prediction when verified against the control analyses are better for all time-scales than when verified against the NCEP analyses. The difference is small for the NH but significant for the SH, with an initial difference in the relative error of 0.30 between the two analyses, a value which is about equal to the error of the control forecast at day 2. This larger error in the SH was also apparent in the Lagrangian analysis of Hodges et al. (2003, 2004). The error growth is slightly larger in DJF 2000/2001 but presumably this is related to natural variations in predictability, although it cannot be ruled out that it could be a consequence of changes in the terrestrial and satellite systems between 1990 and 2000 (Bengtsson et al., 2004a).

The predictions by the terrestrial system show very little difference to the control in the NH. The control prediction is always better, but the extension of predictive skill is only marginal, corresponding to an increase of a few hours. This is also the case over Europe with an ocean region to the west where satellite information dominates. As the difference appears to be smallest at 250 hPa (not shown), it is suggested that aircraft data may have contributed positively. The terrestrial system has a slightly larger error growth in 2000/2001. The terrestrial system has only modest skill at the SH. The initial error is the same as the control forecast at around 3.5 d, a difference which has increased further in 2000/2001. This is most likely due to the deterioration of the terrestrial system, such as the reduced number and coverage of radiosondes during the 1990s.

The satellite system is less accurate than the terrestrial system in the NH with no noticeable improved contribution in 2000/2001 than in 1990/1991. However, the relative contribution of the satellite system is larger in both hemispheres as the difference in skill between the control and the satellite system is smaller during the later period. It is suggested that this may be a combined effect of a gradual increase in satellite information and reduction in the terrestrial information. Practically all the forecast skill in the later period in the SH comes from the satellite system.

The surface-based system has little predictive skill in the SH and only modest skill for a few days in the NH.

5.2. *Extratropical predictability*

Following Lorenz (1982), an upper bound on atmospheric predictability has been estimated by comparing the differences between consecutive forecast separated by 1 d. The three systems (control, terrestrial and satellite) have the same internal growth in both hemispheres corresponding to a doubling time of 1.4 at day 2 increasing to 2.8 at day 5 with a slight tendency to increase further towards the end of the interval (not shown). The predictability is the same in both hemispheres and the same for the two periods, 1990/1991 and 2000/2001. The error doubling time is similar to that found by Simmons and Hollingsworth (2002), but much shorter than in Lorenz (1982), demonstrating that when the model is improved the internal error growth tends to increase. In Fig. 3 we show the predictability calculated for the control integration. We have further undertaken an independent assessment of predictability by comparing the change in time of the differences between the control and the terrestrial and satellite runs, respectively. Between 6 and 24 h, the error growth is even faster (not shown), but reaches thereafter the same internal error growth rate as obtained with the Lorenz type of estimate.

5.3. *The tropics*

It is interesting to note that the terrestrial and satellite systems have practically identical skill for both the 850- and 250-hPa winds. This means that the rather limited number of soundings from radiosondes in the tropics provide the same information content as all the different satellite-based systems. Presumably this is due to the fact that the wind observations from the geostationary satellites are still inaccurate. To determine the wind field from temperature vertical retrievals does not work well because of the weak geostrophic relation in the tropics. As the control forecast is much better than any of the terrestrial or satellite forecasts, it is suggested that these two data sets complement each other. The error growth in all prediction experiments is very slow, increasing only by some 35% at 250 hPa between days 3 and 7.

Estimating the predictive skill similar to Lorenz (1982) indicates an even slower error growth, with an increase of the error of some 25% between days 3 and 7, suggesting that there is further scope for improving the tropical predictions by better models. The results from the tropical evaluation suggest that better wind information through the depth of the atmosphere combined with better models may imply significant increase in tropical predictive skill.

5.4. *Storm tracks*

Both the predictive skill and the predictability of storm tracks are less than that of 500-hPa height, but the relative relations between the different experiments do not change, showing that the satellite system dominates for the SH and the terrestrial system is better than the satellite system for the NH. It is interesting to note that the skill is slightly higher for the SH, which presumably is due to the larger coherence in the forecasts and fewer obstructions in the storm track, as coast lines and orographical obstacles play a smaller role than in the NH. However, the overall conclusions of the predictability of storm tracks is tentative due to the methods used in this study. We believe that the predictability of the storm tracks may be underestimated here due to the difficulty in performing the tracking as the forecast lead time increases. It cannot be excluded that the results may be different if the storm tracks were to be identified along forecast trajectories and validated against the consecutive 6-h analyses (see Fig. 2); this is currently being explored. The other aspect of this part of the study is the relative importance of the spatial scales of the features which are identified and tracked.

5.5. *Concluding remarks*

In this assessment of the information content in different observing systems, we have limited the evaluation to a few key parameters. In the extratropics, poleward of 20°N (20°S), we have used 500-hPa heights and in the tropics (20°N–20°S) we have used the vector wind at 850 and 250 hPa. We could have extended the evaluation to more parameters but we believe they are strongly related to those we have used and there are no indications that they would change the overall conclusions of this study. The results highlight the important role of satellite observing systems, in particular for the SH. However, a total dependency of satellite and surface observations in its present form would degrade forecast skill by about a day in the NH and even more in the tropics. At the same time, the terrestrial system on its own would only degrade the full system by 6–12 h in predictive skill in the NH but almost making NWP infeasible for the SH.

The surface system on its own is not feasible for NWP except for short-range prediction in the NH, and even there with very large errors. However, this may be slightly too pessimistic as the analyses might be improved by modifying the background error statistics, which would then have some impact on the forecasts. The fact that the surface system does have some skill in the short range implies that it may be possible to perform longer

reanalyses than currently available based on surface observations alone. Using an ensemble prediction system, Whitaker et al. (2004) have shown that the better treatment of the background error covariances in such a system does improve the assimilation of surface-only observations and by implication the forecasts.

In view of the comparatively high impact of the relatively small number of terrestrial-based observations, it is suggested that a rational strategy would be to emphasize the quality and all-weather capability of these observations. In the tropics for example, the terrestrial system provides the same predictive information as the satellite system in spite of the limited number of observations and large data sparse regions. It is suggested that satellite wind information still has large errors and that satellite temperature soundings are not effective to generate wind information. At the same time, predictability estimates indicate a slow internal error growth, suggesting great potential for improving tropical predictive skill. The results strongly suggest the need to improve the wind observing system in the tropics. An interesting possibility would be to use drones in combination with dropsondes for vertical wind profiles.

From these results it is possible to consider an observing system consisting of only satellite observations, at the expense of a small reduction in forecast skill in the NH. However, whilst this implies that no real-time terrestrial observations are required, data from radiosondes are still necessary to provide calibration of the satellite data. Hence, an observing system truly independent of terrestrial observations is not currently possible. In addition, the results suggest that observations without vertical coverage have limitations in providing high-quality predictions. It may be suggested that this is due to the inability to reproduce baroclinic structures, which often are associated with rapid development and also rapid error growth. This highlights the need for best possible profiles through the middle and lower troposphere, where baroclinic systems have a large tilt.

6. Acknowledgment

The authors would like to thank the ECMWF for making the ERA-40 system available to us and for providing support in running the experiments.

References

Bengtsson, L. 1999. From short-range barotropic modelling to extended-range global weather prediction. *Tellus* **51A**, 13–32.

Bengtsson, L. and Hodges, K. I. 2005. On the impact of humidity observations in numerical weather prediction. *Tellus* **57A**, in press.

Bengtsson, L., Hodges, K. I. and Hagemann, S. 2004a. Sensitivity of the ERA-40 reanalysis to the observing system: determination of the global atmospheric circulation from reduced observations. *Tellus* **56A**, 456–471.

Bengtsson, L., Hodges, K. I. and Hagemann, S. 2004b. 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.

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

Hodges, K. I. 1999. Adaptive constraints for feature tracking. *Mon. Wea. Rev.* **127**, 1362–1373.

Hodges, K. I., Hoskins, B. J., Boyle, J. and Thorncroft, C. 2003. A comparison of recent reanalysis data sets using objective feature tracking: storm tracks and tropical easterly waves. *Mon. Wea. Rev.* **131**, 2012–2037.

Hodges, K. I., Hoskins, B. J., Boyle, J. and Thorncroft, C. 2004. Corrigendum to "A comparison of recent reanalysis data sets using objective feature tracking: storm tracks and tropical easterly waves". *Mon. Wea. Rev.* **132**, 1325–1327.

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

Kistler, R., Collins, W., Saha, S., White, G., Wollen, J. and co-authors 2001. The NCEP/NCAR 50-yr reanalysis: monthly means CD-ROM and documentation. *Bull. Am. Meteorol. Soc.* **82**, 247–267.

Lorenz, E. N. 1982. Atmospheric predictability experiments with a large numerical model. *Tellus* **34**, 505–513.

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

Simmons, A. J. and Hollingsworth, A. 2002. Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.* **128**, 647–677.

Whitaker, J. S., Compo, G. P., Wei, X. and Hamel, T. M. 2004. Reanalysis without radiosondes using ensemble data assimilation. *Mon. Wea. Rev.* **132**, 1190–1200.

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