Objective verification of deterministic forecasts

Adrian Simmons
European Centre for Medium-Range Weather Forecasts

1. Introduction

There are many ways in which deterministic weather forecasts can be verified objectively. Verification of direct weather-element forecasts is discussed in a separate contribution to these Proceedings. Here we are concerned with verification of dynamical upper-air fields. We concentrate on two widely used measures of skill: the root-mean-square (rms) error and the anomaly correlation, the correlation between forecast and analysed (or observed) deviations from climatology. Although there is much that can be said about these two measures, it must be noted that there are variants and other gross measures such as tendency correlation, rms error normalized by the rms error of a persistence forecast and S1 skill score which might also have been discussed. Note should also be made of alternative, synoptically-based verification methods. One approach involves, for example, either manual or numerical identification of cyclone centres, and accumulation of statistics such as track and intensity error. Examples of studies for the ECMWF system are those by Akyildiz(1985), who examined errors in extratropical forecasts, and Serrano(1997), who compared tropical cyclone tracks in analyses with manually-derived "best track" data. Another approach is to derive measures based on the optimal displacements and amplitude adjustments needed to bring forecast fields into close agreement with analyses or a set of observations (Hoffman and Grassotti, 1996).

The presentation at the Seminar covered the following topics:

- •Root-mean-square error and anomaly correlation - Definitions, properties, verification against analyses and observations, and the smoothing of forecasts;
- •Differences between consecutive forecasts and implications for predictability;
- •Trends in objective scores from operational forecasting;
- •Use of objective verification in forecasting-system development.

Much of the material presented was drawn from a lecture delivered to the 1995 ECMWF Seminar on Predictability. Reference should be made to the comprehensive write-up of that lecture (Simmons, 1996) for basic definitions, properties and examples of applications of the skill measures, for discussion of the dependence of skill measures on scale, season and smoothing (or elimination of unpredictable scales), and for discussion of the differences between consecutive forecasts and error-growth modelling. In the present contribution, we supplement the paper of Simmons(1996) with some additional topics and results discussed at the 1999 Seminar. Some of the examples have been updated to include results not available at the time of the Seminar itself. Almost all results have been obtained using the operational ECMWF verification software developed originally by Nieminen(1983).

2. Verification against analyses

One question that is raised from time to time concerns the extent to which skill scores based on comparisons of forecasts with analyses depend on the choice of verifying analysis. A data assimilation system produces an analysis by using observations to correct a short-range "background" forecast started from the preceding analysis in the sequence. If the same model is used in the data assimilation as is used to produce the forecast to be verified, then the verifying analysis may inherit some of the errors of the forecast, particularly where observations are scarce or if the assimilation gives low weight to the observations. Verification statistics would then give erroneously low indications of forecast error.

In practice, this is not a major concern for the verification of medium-range forecasts, although when a major change to a forecasting system is tested, improvements in the short range forecasts in the tropics, southern hemisphere or high in the stratosphere may be masked if the verification is carried out against operational analyses rather than the test analyses from the new system. Similar considerations apply when comparing the verification scores of forecasts from different centres.

To illustrate this we present some results from the forecasting systems of ECMWF, the Met Office (UKMO) and the German weather service (DWD). We have chosen to verify the 28 cases in the period 19 October to 9 December 1999 for which complete sets of forecasts and verifying analyses are available in the ECMWF archives. The starting date was chosen so that all forecasts were made after the important upgrades that the three centres made to their operational forecasting systems in October 1999.

Fig. 1 presents mean anomaly correlations of 500hPa height for the extratropical northern hemisphere, comparing ECMWF and UKMO and ECMWF and DWD forecasts. Each centre's forecast is shown verified against its own analysis and against the analysis of the other centre in the pair. The differences in mean forecasting performance of the centres in the medium range is clearly insensitive to the verifying analysis for this measure and domain. Slight differences can be detected; the UKMO forecasts score marginally better when verified against the UKMO rather than the ECMWF analyses, while the ECMWF forecasts score similarly against the ECMWF and UKMO analyses, and very sightly better against the DWD analyses beyond day 5.

Much larger differences can be seen in Fig. 2, which shows rms errors of 500hPa height forecasts for the more poorly observed southern hemisphere. In the short range there is a strong sensitivity to verifying analysis, indicating that there is a significant correlation between differences in the short-range forecasts of two centres and differences in the verifying analyses of the same two centres. Note that each centre's "day-0" forecast is simply the analysis of that centre, so the rms error at day 0 for each centre is zero when verified against its own analysis and equal to the rms analysis difference when verified against the other centre's analysis. The ECMWF and UKMO analyses are evidently more similar to each other than are the ECMWF and DWD analyses.

The rms errors for the southern hemisphere show less sensitivity to verifying analysis in the medium range, but relative differences are much larger than for the anomaly correlations of the northern hemisphere forecasts shown in Fig. 1. It can again be seen that later in the range the ECMWF forecasts verify better against the DWD analyses than against the ECMWF analyses.

The latter result may be explained by differences in the variance of the verifying analyses. Rms anomalies of the forecasts and verifying analyses are presented in Fig. 3. Analysed and forecast levels of variance are very similar for the ECMWF and UKMO systems, and significantly lower for the DWD system. Lower variance in forecasts or verifying analyses implies lower asymptotic values for rms error at large forecast range. This provides a likely explanation not only for the lower rms errors of day-6 and day-7 ECMWF forecasts when verified against DWD rather than ECMWF analyses, but also for the much lower rate at which rms error grows in the DWD forecasts beyond day 4.

3. Verification against observations

Verification against a set of observations made over a particular domain is an alternative to verification against analyses over that domain. This typically involves verification against radiosonde measurements in the case of upper-air fields. The machinery of data assimilation is, however, constructed to enable comparison of short-range background forecasts with a variety of observations such as satellite-measured radiances, aircraft-measured winds and temperatures, and cloud-track winds. It could thus be adapted to provide a more comprehensive verification of longer-range forecasts against observations.

Verification against observations provides a more independent measure of forecast skill than verification against analyses. Interpretation of results does require some care however, as account has to be taken of the often far-from-uniform spatial distribution of the observations, of the differences in availability of observations at the different synoptic hours, and of the biases that can exist in different components of the observing system. Decisions have also to be made concerning the application of quality controls to the verifying observations. Again, problems are most evident for regions where observations are relatively scarce.

Fig.4 shows verification scores for the extratropical southern hemisphere, computed as part of the assessment of a new scheme for the assimilation of raw (level-1c) radiances from the TOVS and new ATOVS instruments (McNally et al., 1999). Forecast scores from the new scheme (introduced operationally in May 1999) are compared with those from the former operational scheme which assimilated pre-processed radiances from the TOVS instruments.The upper curves show rms 500hPa height forecast errors computed against analyses, and the lower curves the corresponding results for verification against radiosondes. In this case both verifications indicate that the new system gives a small overall improvement in forecast accuracy. However, the generally lower absolute values of error seen in the verification against radiosondes is indicative of a distribution of radiosonde measurements that under-samples where variance is relatively large over the southern oceans. All forecasts in this set of cases were run from analyses for 12UTC, and the diurnal oscillation in the scores against radiosondes is presumably a consequence of the different availability of radiosonde observations at 00 and 12UTC. This has to be borne in mind in particular when comparing forecast performance at a particular range from two operational centres, one of which runs from 00UTC and one of which runs from 12UTC, or indeed in comparing the skill of the forecasts from 00 and 12UTC produced by one centre.

Fig. 5 shows rms errors of day-3 850hPa vector wind forecasts for the tropics. Results are shown for ECMWF and UKMO forecasts run from 12UTC analyses, and also for UKMO forecasts from 00UTC. Unlike earlier results which were computed applying ECMWF verification software to fields received over the GTS, these results are based on verification scores computed by each of the centres for their own forecasts, following standards set by WMO/CBS. The upper plot shows verification of each centre's forecasts against its own analyses, and according to this measure, the 12 UTC UKMO forecasts are superior to the 12UTC forecasts of ECMWF. The lower plot shows verification against radiosonde measurements. This measure indicates conversely that the 12UTC ECMWF forecasts are superior to those of UKMO. The 00UTC forecasts from UKMO are poorer than the 12UTC UKMO forecasts according to both measures. Against analyses, the ECMWF forecasts appeared to improve substantially relative to those of UKMO in 1997, associated principally with introduction of a new formulation of the background-error constraint in the variational data assimilation (Derber and Bouttier, 1999). Errors measured against radiosondes also decline over this period for the ECMWF forecasts, but a similar decline is seen also for the UKMO forecasts, particularly those from 12UTC. Alternative verification measures and other diagnostics must be examined to gain a reliable and comprehensive picture of relative performance in such cases. Note, however, that Fig. 5 does indicate a general trend towards lower errors for both centres, and some degree of agreement between the two centres' results and between the two sets of verifications as regards interannual variations in skill.

Fig. 6 shows measures of forecast bias near the tropical tropopause. Time series of mean 100hPa temperature errors are shown for several forecast ranges, for verification against analyses (upper panel) and against radiosonde measurements (lower panel). In many respects there is a large degree of agreement between the two measures. Both indicate, for example, steady growth of warm bias in the forecasts in 1994 which was much reduced following model changes made in April 1995 (Miller et al., 1995). Both also indicate development of cold bias in mid 1999, a problem associated with the change from 31- to 50-level resolution in March (Untch and Simmons, 1999). This was alleviated by a parametrization change introduced with a 60-level model version in October (Jakob et al., 2000).

Differences in the two sets of time series can also be seen in Fig. 6. In particular, since 1997 forecasts have developed a larger cold bias measured against radiosondes than measured against analyses. Correspondingly, the ECMWF analyses have developed a cold bias when compared with radiosonde measurements. Moreover, sample re-analyses recently produced in preparation for ERA-40, a new analysis for the years since mid-1957, are colder by about 1K in the tropical mean than the ERA-15 analyses (Gibson et al. 1997) produced with close to the mid-1995 version of the operational system. Results of trials of the individual changes made to the forecasting system over the past three years have to be examined to establish the factors responsible for this change in character. It should be noted, however, that no bias correction of the verifying radiosonde data is made in the software used here, so a component of the apparent cold bias in the analyses and forecasts could arise from warm bias in the set of verifying radiosonde measurements, due to absence of a correction of radiation effects in some observations.

4. Trends in forecasting-system performance

Although some features of time series such as shown in Fig. 6 can be readily understood as consequences of specific changes to the forecasting system, variations in conventional tropospheric verification scores can be more difficult to understand, particularly later in the forecast range and for smaller regions. Fig. 7 shows time series of the range at which monthly-mean anomaly correlations of 500hPa height reach the 95% and 60% levels, for Europe and for the extratropical northern hemisphere. The improvement over the past two decades has been quite regular in the short to early-medium range, as represented by the results for the 95% level, rather more so for the northern hemisphere than for Europe. There is, however, considerably more variability at the 60% level, particularly for Europe, where isolated months with unusually good scores occur throughout the period. Indeed, the performance over Europe at the 60% level achieved in June 1980, within the first year of operations, has yet to be matched in any subsequent June. The performance in the June and more especially the August of 1999 were especially poor compared with preceding years, yet the following autumn and early winter scores provide an unmatched sequence of good scores in which the 60% level was reached beyond day 7 in each of the monthly means.

Overall, the improvement over the past twenty years has been by a little under 1.5 days at the 95% level and by a little over 1.5 days at the 60% level. At the 60% level there was no increase in skill between the mid 1980s and the early 1990s for Europe, and only a marginal increase over this period for the northern hemisphere, despite clear improvement at the 95% level. This was interpreted by Simmons et al.(1995) as a consequence of the development of an improved forecast model. The model became more realistic, and with accompanying data assimilation improvements led to more accurate short-range forecasts. It also led to a faster growth of error however, cancelling the benefit to medium-range forecast skill that would otherwise have resulted from improved analyses and short-range forecasts.

Comparison with the performance of other operational centres can help in assessing whether variations in operational performance result from changes made to the forecasting system or from external factors affecting predictability such as variations in circulation pattern or changes to the observing system. Since there is a tendency for most if not all major operational systems to perform relatively well or poorly in certain months or years, changes in the separation between the scores of two centres may indicate impact of changes to one or other of the forecasting systems. Interpretation is not always straightforward however, both because a forecasting system that performs relatively well in one particular synoptic regime may perform less well in another, and because all systems are subject to some degree of change from time to time, in data usage if not in the forecasting system itself.

By way of illustration, Fig. 8 presents annual running means of the anomaly correlations of the day-5 500hPa height forecasts of ECMWF and UKMO from 1991 onwards. Results are shown for the extratropical northern and southern hemispheres and for Europe and North America. Some of the short–period oscillations in these curves are common to both sets of forecasts, indicating months in which both centres produced forecasts with unusually good or poor anomaly correlations. Coherent longer–period fluctuations can also be seen, most clearly in the fluctuations in scores for North America since 1994.

Changes in relative performance can also be seen. The UKMO forecasts for Europe improved much more rapidly than those of ECMWF in the period from 1991 to 1994, whereas the ECMWF forecasts advanced relative to those of UKMO in 1995. The latter appears to be associated with the model changes made in April 1995, which are also the most likely reason for the improvement in summertime performance at the 60% level over Europe after 1994 shown in Fig. 7. The very marked divergence in performance over the southern hemisphere between 1995 and 1998 is almost certainly an indication of benefits from ECMWF's developments in modelling, from its various refinements to the use of satellite data and from its implementation of variational data assimilation (e.g. Rabier et al., 2000), including the direct variational assimilation of TOVS/ATOVS radiances. The UKMO introduced a variational data assimilation system in spring 1999, and direct assimilation of radiance data some six months later. These changes led to marked improvements in its southern hemisphere forecasts, as indicated by the sharp rise in 1999 of the UKMO curve in the lower-left panel of Fig. 8.

5. Error doubling times

Another approach to understanding variations in forecasting-system performance is to estimate the variations in intrinsic predictability that arise from low frequency intraseasonal and interannual variations in the atmospheric general circulation. This might be investigated, for example, by constructing a simple index of the baroclinic instability of the flow in terms of mean latitudinal temperature gradient and static stability, or by constructing an index based on the amplification factors of the leading singular vectors used to construct the initial perturbations for ensemble forecasting. An alternative was identified by Lorenz (1982), who estimated intrinsic forecast error growth rates by fitting a simple error-growth model to the differences between successive forecasts verifying at the same time. Lorenz' study was based on forecasts from a 100-day period beginning 1 December 1980. His conclusions were re-examined by Simmons et al. (1995; see also Simmons, 1996) in the light of subsequent operational performance. Diagnosis of recent forecasts gave estimates of intrinsic error doubling time of around 1.4 to 1.6 days, compared with the values of around 1.8 to 2 days found using the early operational forecasts. It was argued that this was likely to be due to the development of a more active (and more realistic) operational forecast model rather than to a shift in the predictability of the atmosphere itself.

Table 1 presents error doubling times computed for the extratropical northern hemisphere for each season of the past ten years. Generally, doubling times are relatively short in autumn and summer, most variable in spring and longest in winter. This is broadly consistent with the seasonal variations seen in the scores based on monthly anomaly correlations presented in Fig. 7. Some of the interannual differences in doubling time can be related to known consequences of model changes. In particular, the shorter doubling times beginning autumn 1991 occur following operational implementation of the Centre's T213L31 model, which was more active than its T106L19 predecessor. The early T213L31 version in fact tended to overdevelop systems due to problems both with its new semi-Lagrangian advection scheme (revised in August 1992; Ritchie et al., 1995) and with its parametrization of cloud/radiation interaction (revised in February 1993; Morcrette, 1993). This is consistent with the relatively low doubling times recorded for the first three seasons of 1992. One indication of the effect on scores of such variations in intrinsic error growth rate is provided in Fig. 10, which shows anomaly correlations derived from Lorenz' error-growth model for doubling times of 1.4, 1.5, 1.6 and 1.7 days, assuming a common initial rms error of 2.5% of the asymptotic limit.

Another noteworthy feature of Table 1 is that error doubling times for spring and summer are significantly lower in 1999 than in any other recent year, and the winter 1999 values are also relatively low. This cannot be linked in any obvious way to recent model changes, and provides some reassurance that the relatively poor ECMWF forecast performance in early and mid 1999 might have been due, in part at least, to an inherently less predictable synoptic regime. Some confirmation of this is provided by the day-5 northern hemisphere scores presented in the upper-left panel of Fig. 8, which show that the running annual-mean anomaly correlations of both ECMWF and UKMO forecasts were a minimum over the year to August 1999. This is no cause for complacency, however, as the ECMWF medium-range forecasts were especially poor over Europe from May to August, scoring worse than those of UKMO at day 6 for each month, both for rms error and for anomaly correlation of 500hPa height.

Table 1: Intrinsic error doubling times (days), following Lorenz(1982). Winter 1991 denotes the period from 1 December 1990 to 28 February 1991

	Winter Dec-Feb	Spring Mar-May	Summer Jun-Aug	Autumn Sep-Nov
1990	1.76	1.57	1.73	1.51
1991	1.75	1.65	1.71	1.45
1992	1.52	1.49	1.43	1.45
1993	1.61	1.52	1.54	1.46
1994	1.57	1.50	1.54	1.42
1995	1.51	1.50	1.50	1.53
1996	1.64	1.59	1.54	1.46
1997	1.63	1.60	1.53	1.48
1998	1.60	1.48	1.51	1.46
1999	1.51	1.42	1.45	1.47

6. Testing changes to a forecasting system

Numerical forecasts produced operationally are subject to considerable synoptic scrutiny by forecasters. This provides an independent view which is commonly in agreement with objective verification in identifying spells of much below or above average performance. Such extensive scrutiny cannot, in practice, be applied to extensive series of research experiments carried out as part of the advanced testing of potential changes to operational systems. Objective verification thus plays a particularly important role in the assessment of such changes.

In the first instance, unexpectedly poor skill scores from the initial tests of a trial system may provide the first signal that there has been a technical problem in the preparation of the trial. Thereafter, routine verification of an extended set of forecasts may reveal unexpected biases requiring further investigation of the change, or may confirm that a change has had the expected effect on biases. Its primary use, however, is to provide measures of the impact on the accuracy of tropospheric forecasts. In this case the risk of misleading results due to running the trial over too limited a period has to be kept in mind, particularly if one is seeking to measure improvement for a relatively small domain during a period of low predictability.

A pronounced example of the latter is provided by tests carried out at ECMWF over the summer of 1999. A particularly lengthy run of data assimilation and forecasts was carried out as part of the pre-operational testing of cycle 21r4 of the Centre's Integrated Forecasting System (IFS). This version was made operational in mid October, and involved major changes in vertical resolution, parametrization and data assimilation (Jakob et al., 2000). After shorter tests of the individual changes, the new cycle as a whole was run for the period from early May to mid October, and results were compared with those from cycle 21r2, which was the operational version of the IFS from mid July until mid October. Cycle 21r2 had been run earlier for more than two months as part of its own pre-operational testing.

Fig. 10 shows 500hPa height anomaly correlations computed for the European domain for the two cycles, with results averaged separately for the complete calendar months of June, July, August and September. They show a substantial variation in the impact of the change from cycle 21r2 to 21r4. Cycle 21r4 improves substantially over 21r2 in August, when operational performance for Europe was unusually poor. However, the converse was the case in September, when the operational performance was unusually good. The new cycle performed better in June, and depending on the forecast range had either neutral or slightly negative impact in July. It is rare for ECMWF (and other operational centres) to test forecasting-system changes over so long a period, and thus clearly difficult to identify the extent to which specific changes to the forecasting system have contributed to the long-term improvement of forecast accuracy over Europe.

Hemispheric scores provide a more reliable indication of whether a change increases overall forecast accuracy, although results for a particular month may still not be indicative of those for a larger sample. Fig. 11 shows the impact of the change from 21r2 to 21r4 on anomaly correlations for the northern hemisphere. Impact is positive in July as well as June and August. It is largest in August, and essentially neutral in September.

Variability of impact is, of course, not always as marked as shown in Fig. 10. An extreme counter example is provided by the impact on stratospheric forecasts of the change from 31- to 50-level vertical resolution made operationally in March 1999 (Untch and Simmons, 1999) and from parallel experimentation testing the new "1c" use of TOVS/ATOVS data for which impact on 500hPa height scores was presented in Fig. 4. The uppermost full model level in the 31-level resolution was located at 10hPa, and forecast scores at this level were, as expected, substantially improved by changing to the finer layer spacing and much higher top level of the 50-level resolution. Forecast scores in the stratosphere were improved further by the changed use of radiance data.

Fig. 12 shows rms errors of 10hPa temperature and vector-wind forecasts, verified against radiosonde data for the extratropical northern hemisphere, averaged for a large sample of cases from the winter of 1998/99. Substantial improvement of the 50-level system over the 31-level system and further improvement from the new radiance assimilation are evident. Fig. 13 shows scatter plots for the day-5 temperature verification, comparing separately the impacts of the resolution change and the radiance-assimilation change. Results of the statistical-significance tests incorporated in the verification software are included. The figure shows the highly systematic nature of the improvements. It also identifies a rogue case in which the 50-level system with the old radiance assimilation did unusually badly, with an rms error of more than 8K, about twice that of either the 31-level forecast or the 50-level forecast from the new radiance assimilation. This case is clearly one which merits further diagnosis.

7. Concluding remarks

In this paper we have discussed some aspects of objective verification of deterministic forecasts and presented some examples of its use. We have concentrated on recent cases, supplementing and updating the account given by Simmons (1996). Conventional objective skill scores provide a necessary and useful diagnostic of the basic health of an operational forecasting system and play an important role in the assessment of changes to the system. Objective scores do, however, need careful (inevitably subjective!) interpretation, as they favour smooth forecasts and underactive models, and as a large sample of forecasts may be needed to obtain a reliable signal. Objective verification can be used to indicate problems, raise questions, identify specific cases for investigation, and check solutions. Generally though, one needs to utilize other diagnostics to find the causes of problems and to justify solutions to them. We have illustrated a healthy overall trend in forecasting-system performance, and indicated how comparison of the results of different operational centres can help in the understanding of variations in forecast skill.

Acknowledgements

The operational verification software originally developed by Rauno Nieminen was subsequently maintained and extended by Brian Norris, and most recently by Claude Gibert.

References

Akyildiz, V., 1985: Systematic errors in the behaviour of cyclones in the ECMWF operational models. *Tellus*, **37A**, 297-308.

Derber, J. and F. Bouttier, 1999: A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, **51A**, 195-221.

Gibson, J.K., P. Kållberg, S. Uppala, A. Nomura, A. Hernandez, E. Serrano, 1997: ERA Description. ECMWF Re-Analysis Final Report Series, **1**, 72pp.

Hoffman, R.N. and C. Grassotti, 1996: A technique for assimilating SSM/I observations of marine atmospheric storms: Tests with ECMWF analyses. *J. Appl. Meteor.*, **35**, 1177-1188.

Jakob, C., E. Andersson, A. Beljaars, R. Buizza, M. Fisher, E. Gerard, A. Ghelli, P. Janssen, G. Kelly, A. McNally, M. Miller, A. Simmons, J. Teixeira and P. Viterbo, 2000: The new operational IFS cycle CY21R4. *ECMWF Newsletter*, **85**, *in press*.

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

McNally, A.P., E. Andersson, G. Kelly and R.W. Saunders, 1999: The use of raw TOVS/ATOVS radiances in the ECMWF 4D-Var assimilation system. *ECMWF Newsletter*, **83**, 2-7.

Miller, M., M. Hortal and C. Jakob, 1995: A major operational forecast model change. *ECMWF Newsletter*, **83**, 2-8.

Morcrette, J.-J., 1993: Revision of the clear-sky and cloud radiative properties in the ECMWF model. *ECMWF Newsletter*, **61**, 3-14.

Nieminen, R. 1983: Operational verification of ECMWF forecast fields and results for 1980-1981. *ECMWF Tech. Rep.*, **36**, 48pp.

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons, 2000: The ECMWF operational implementation of four dimensional variational assimilation. Part I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, **126**, *in press*.

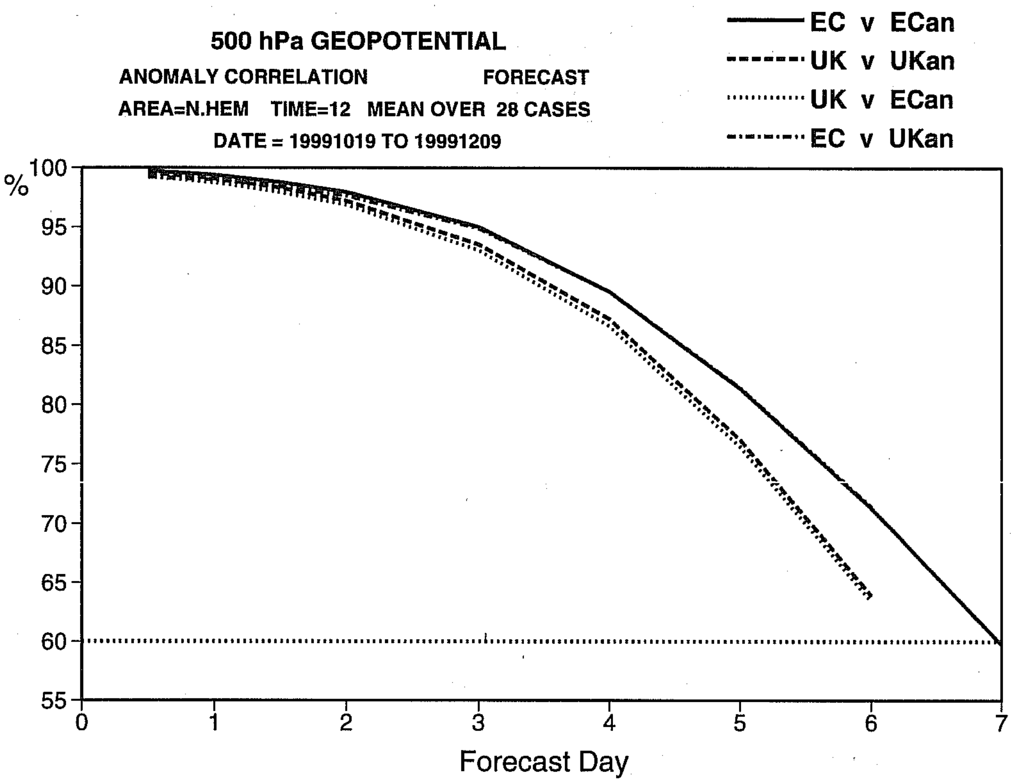
Ritchie,H., C. Temperton, A. Simmons, M. Hortal, T. Davies, D. Dent, and M. Hamrud, 1995: Implementation of the semi-Lagrangian method in a high resolution version of the ECMWF forecast model. *Mon. Wea. Rev.*, **123**, 489-514.

Serrano, E., 1997: Tropical cyclones. ECMWF Re-Analysis Final Report Series, **5**, 30pp.

Simmons, A.J., R. Mureau and T. Petroliagis, 1995: Error growth and predictability estimates for the ECMWF forecasting system. *Quart. J. Roy. Meteor. Soc.*, **121**, 1739–1771.

Simmons, A.J., 1996: The skill of 500hPa height forecasts. *Proceedings of 1995 ECMWF Seminar on Predictability, Vol. 1*, 19-68.

Untch, A., A.J. Simmons and colleagues, 1999: Increased stratospheric resolution in the ECMWF forecasting system. *ECMWF Newsletter*, **82**, 2-8.



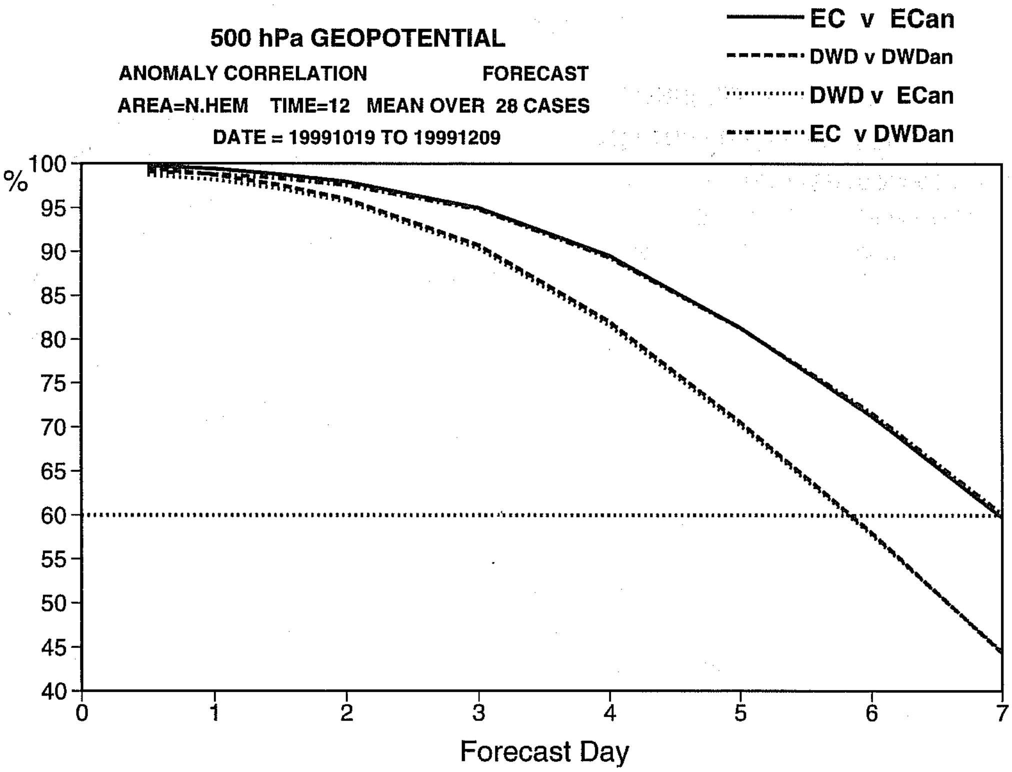


Fig. 1. Anomaly correlations of 500hPa height for the extratropical northern hemisphere as functions of forecast range, averaged over 28 forecasts within the period 19 October to 9 December 1999. The upper panel compares forecasts from ECMWF and UKMO, each verified against its own and the other centre's analyses. The lower panel presents a similar comparison for the forecasts of ECMWF and DWD.

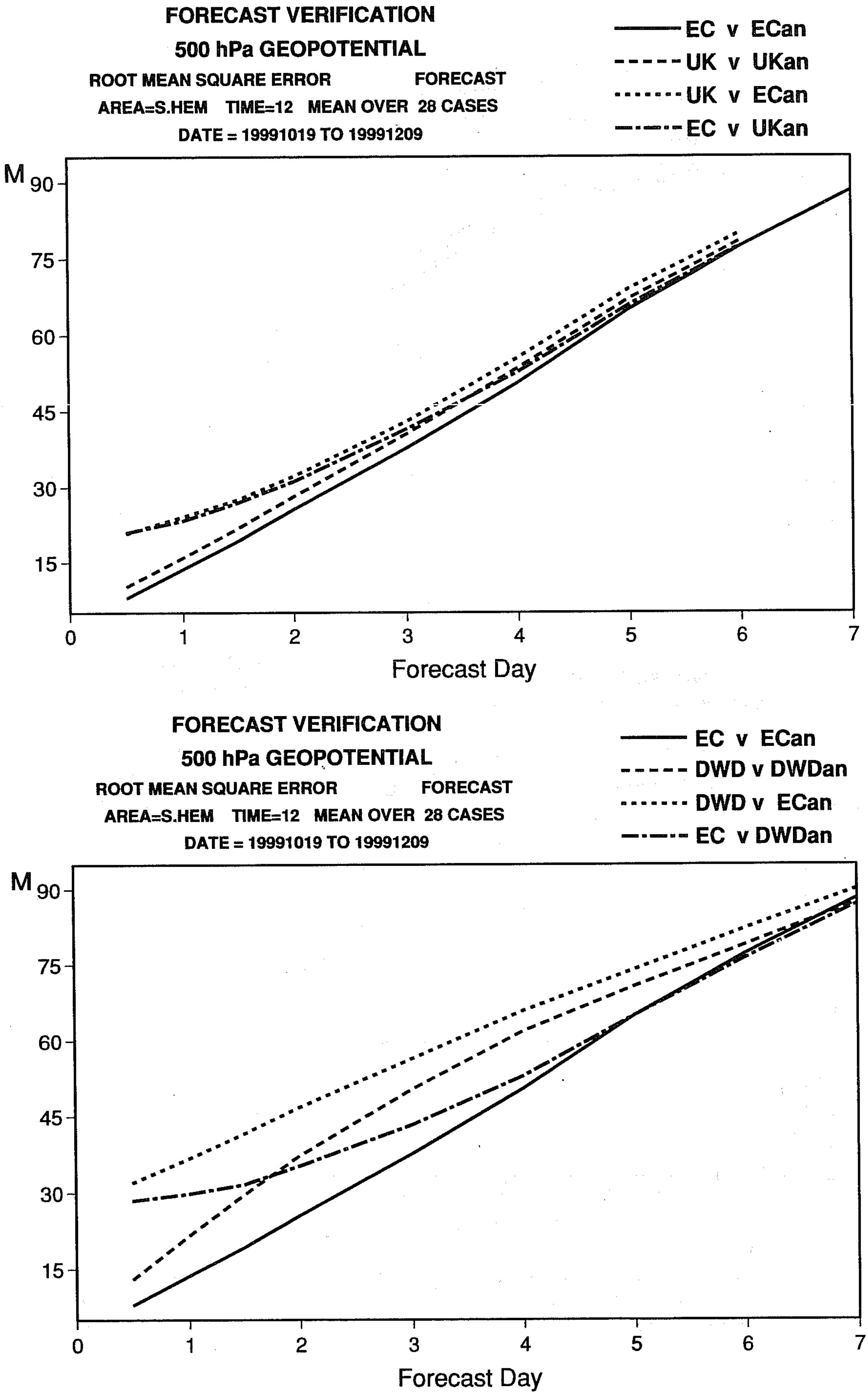
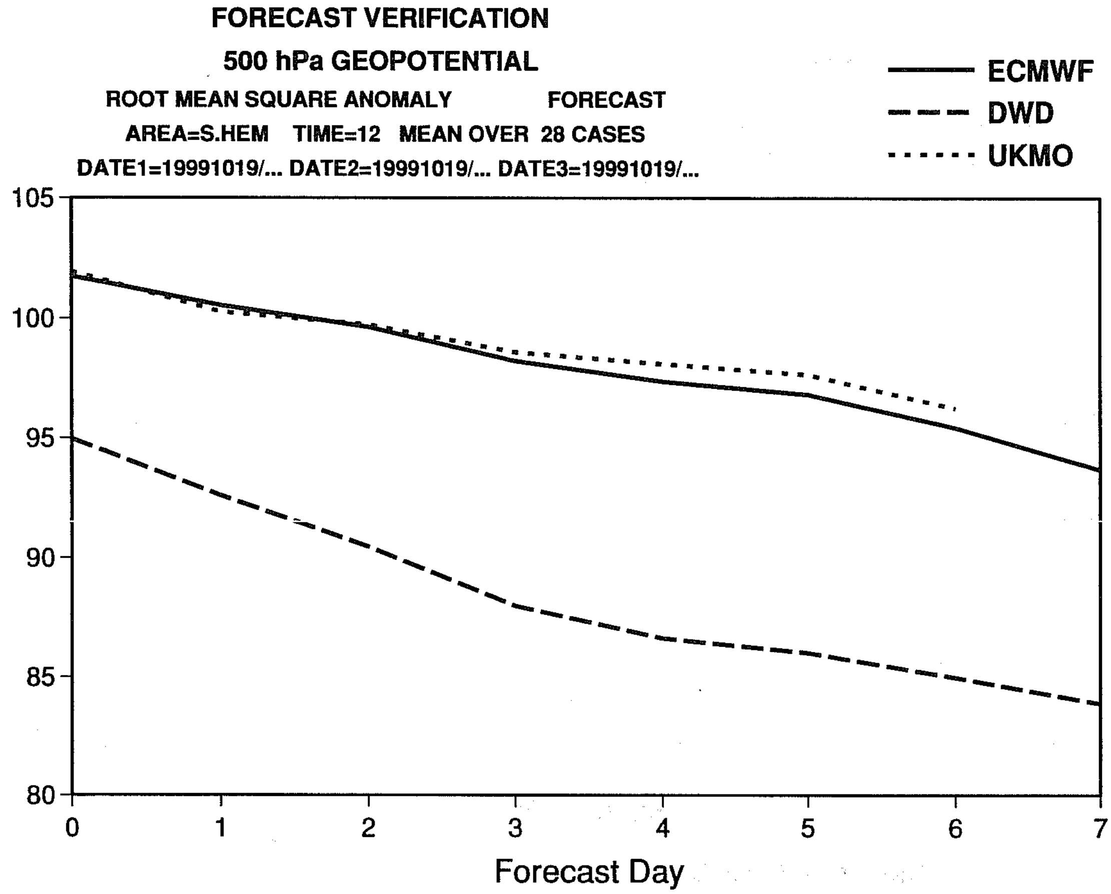


Fig. 2. As Fig.1, but for the rms error of forecasts for the extratropical southern hemisphere.



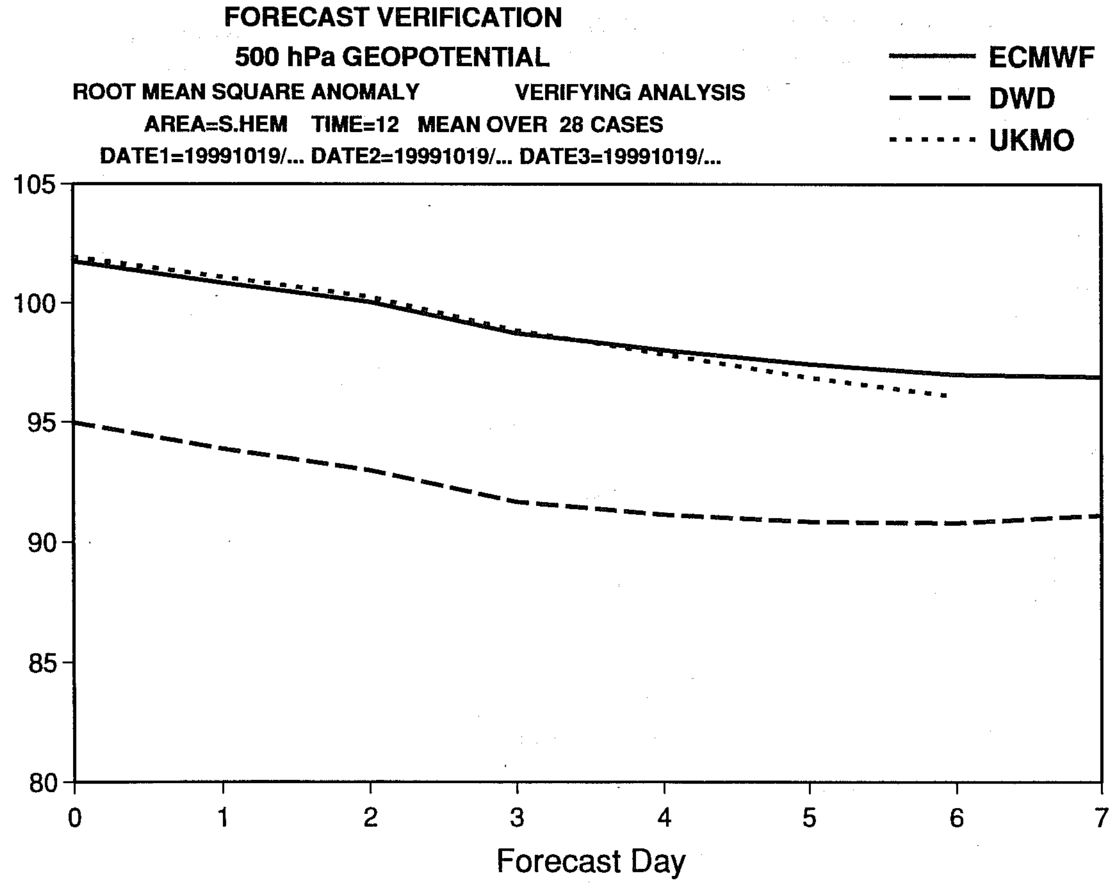
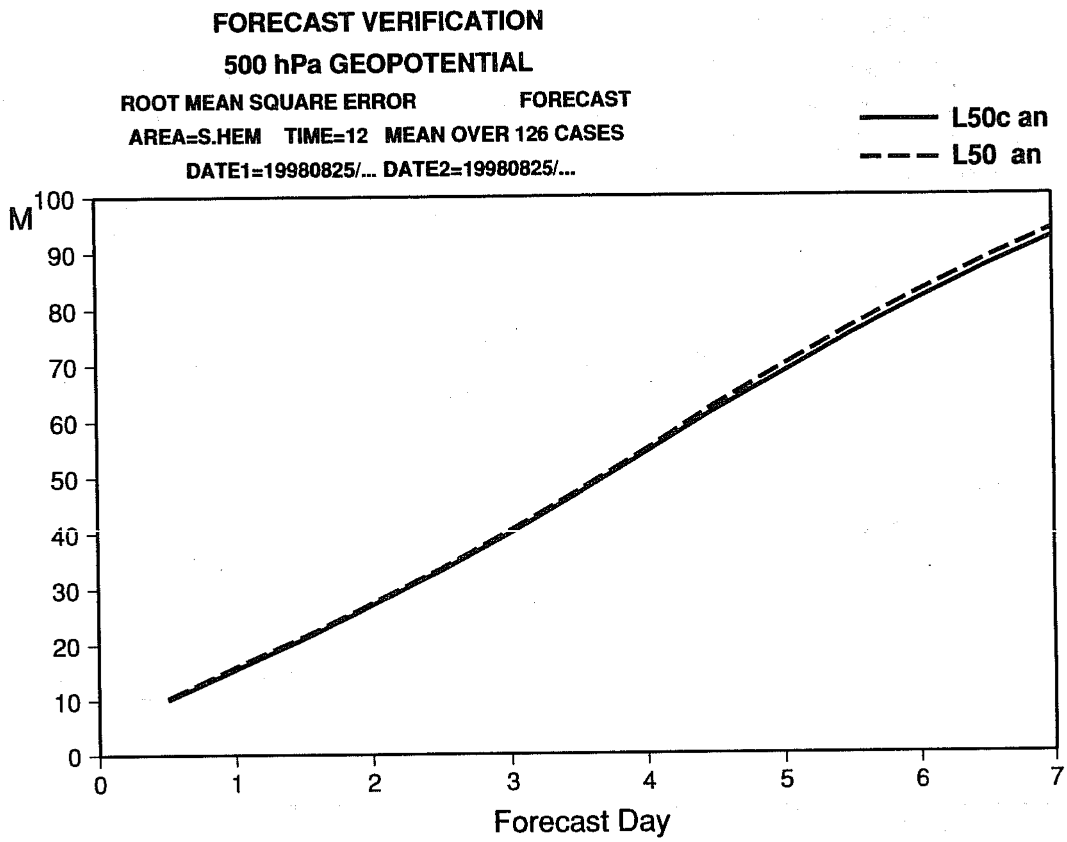


Fig. 3. Upper: The rms anomaly of the 500hPa height forecasts of ECMWF, DWD and UKMO for the extratropical southern hemisphere, averaged over the 28 cases used in Figs. 1 and 2.

Lower: The rms anomaly of the verifying analyses of the three centres.



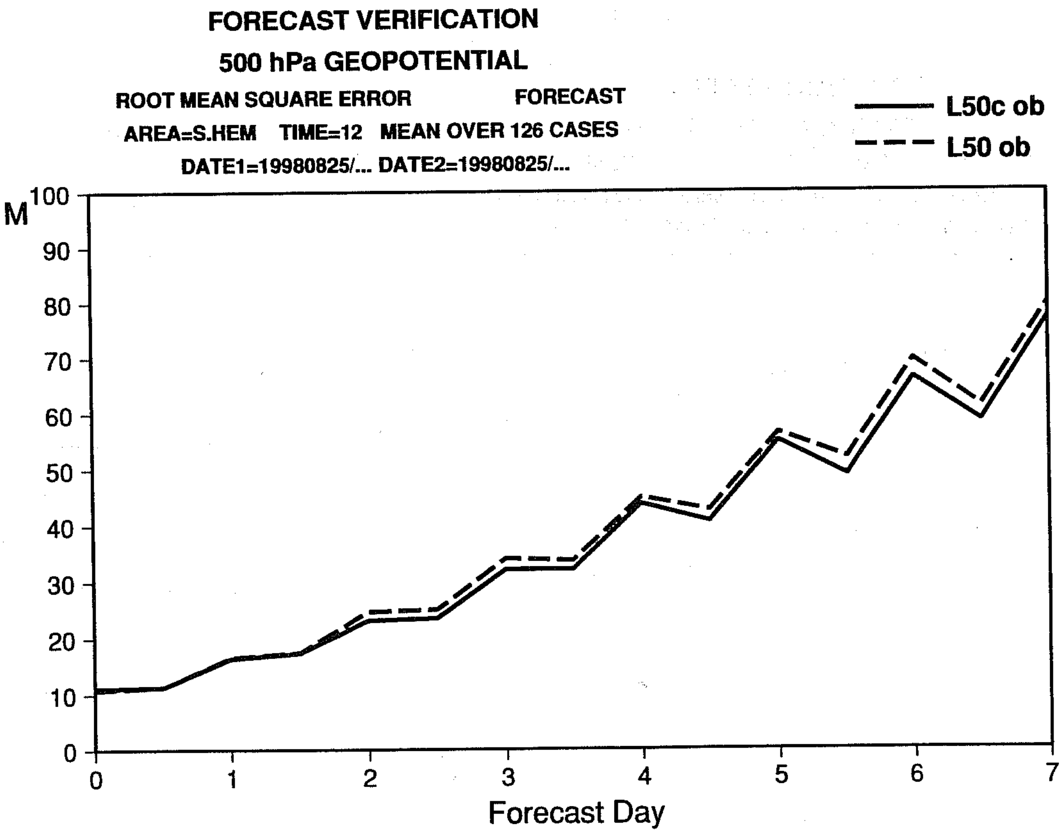


Fig. 4. Rms errors of 500hPa height forecasts for the extratropical southern hemisphere comparing the impact of a changed assimilation of satellite radiance data, averaged over 126 cases with starting dates between 25 August 1998 and 8 March 1999. The upper curves show the verification against analyses (using where possible verifying analyses produced by the version of the assimilation system from which the forecasts were made), and the lower curves show the mean of the verification against all radiosonde observations made over the region.

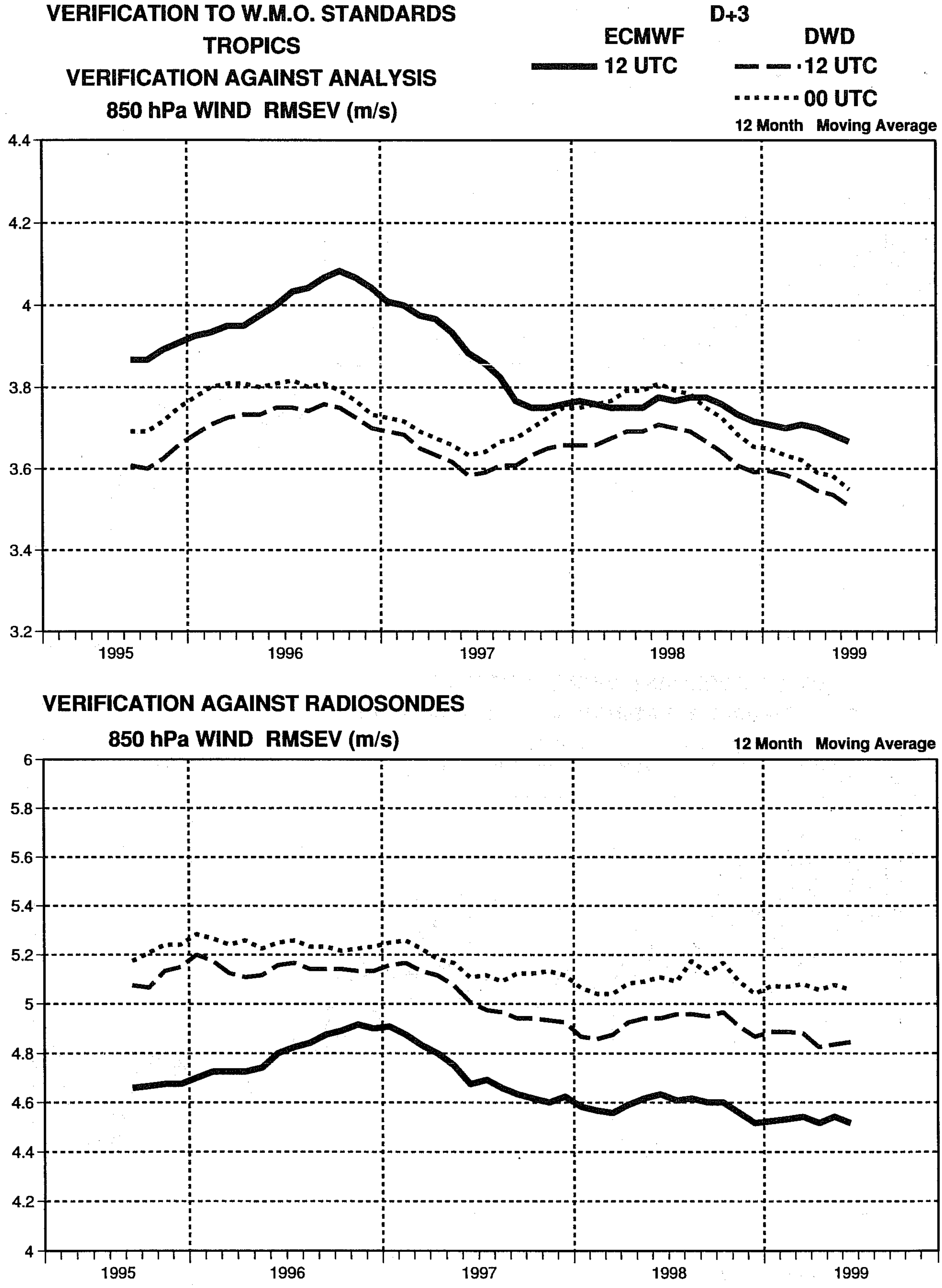
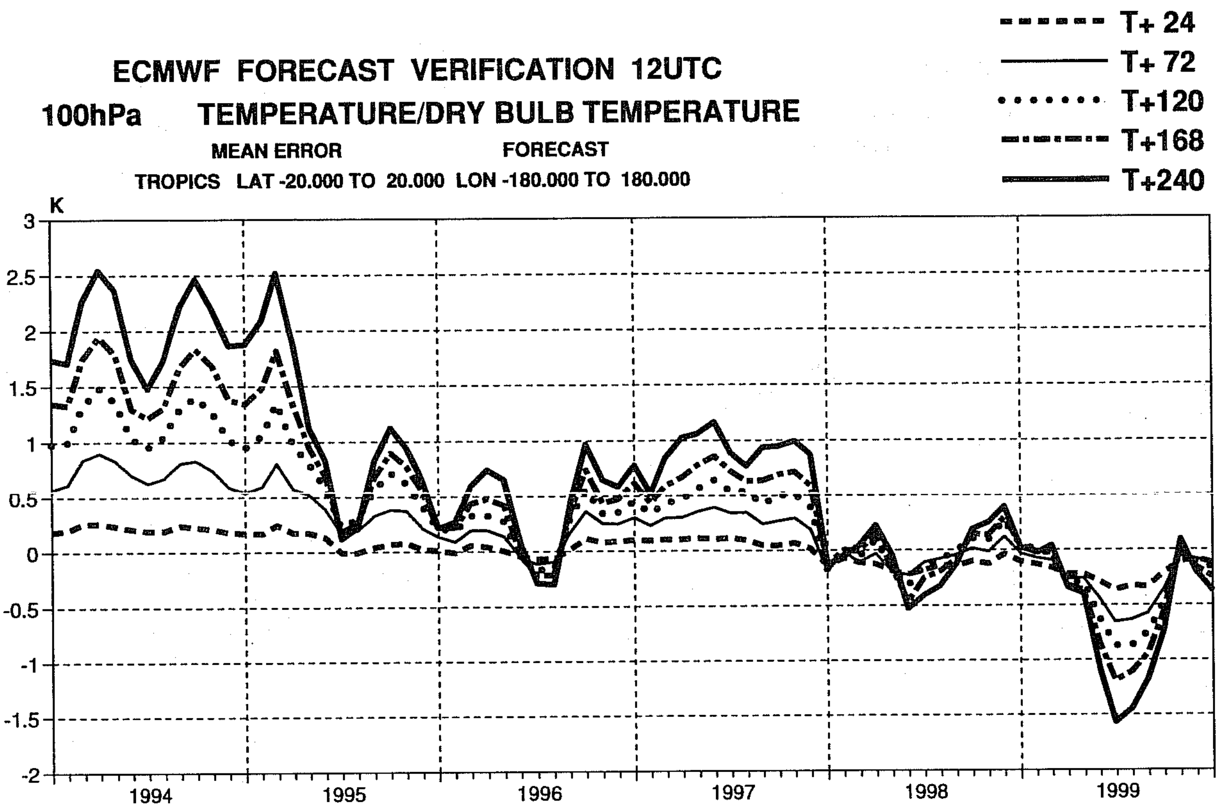


Fig. 5. Rms day-3 error of 12UTC ECMWF and 12UTC and 00UTC UKMO forecasts of 850hPa vector wind for the tropics, verified against analyses (upper) and radiosondes (lower). The annual running mean is plotted, using monthly values available from April 1995 to January 2000.



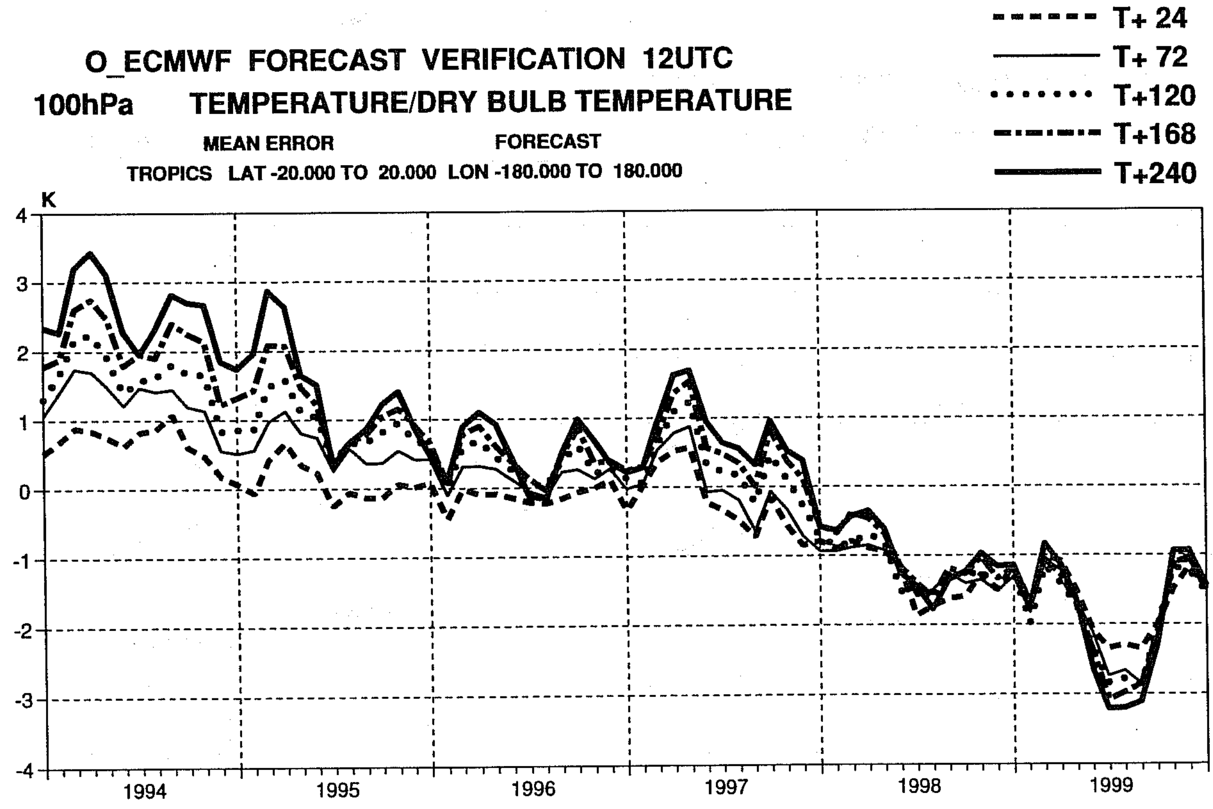


Fig. 6. Mean errors of operational 100hPa temperature forecasts for the tropics at forecasts ranges 24, 72, 120, 168 and 240 hours. Monthly averages for the period from January 1994 to January 2000 are shown. The upper panel is for verification against analyses, and the lower panel is for verification against radiosonde observations. The difference in scales for the two panels should be noted.

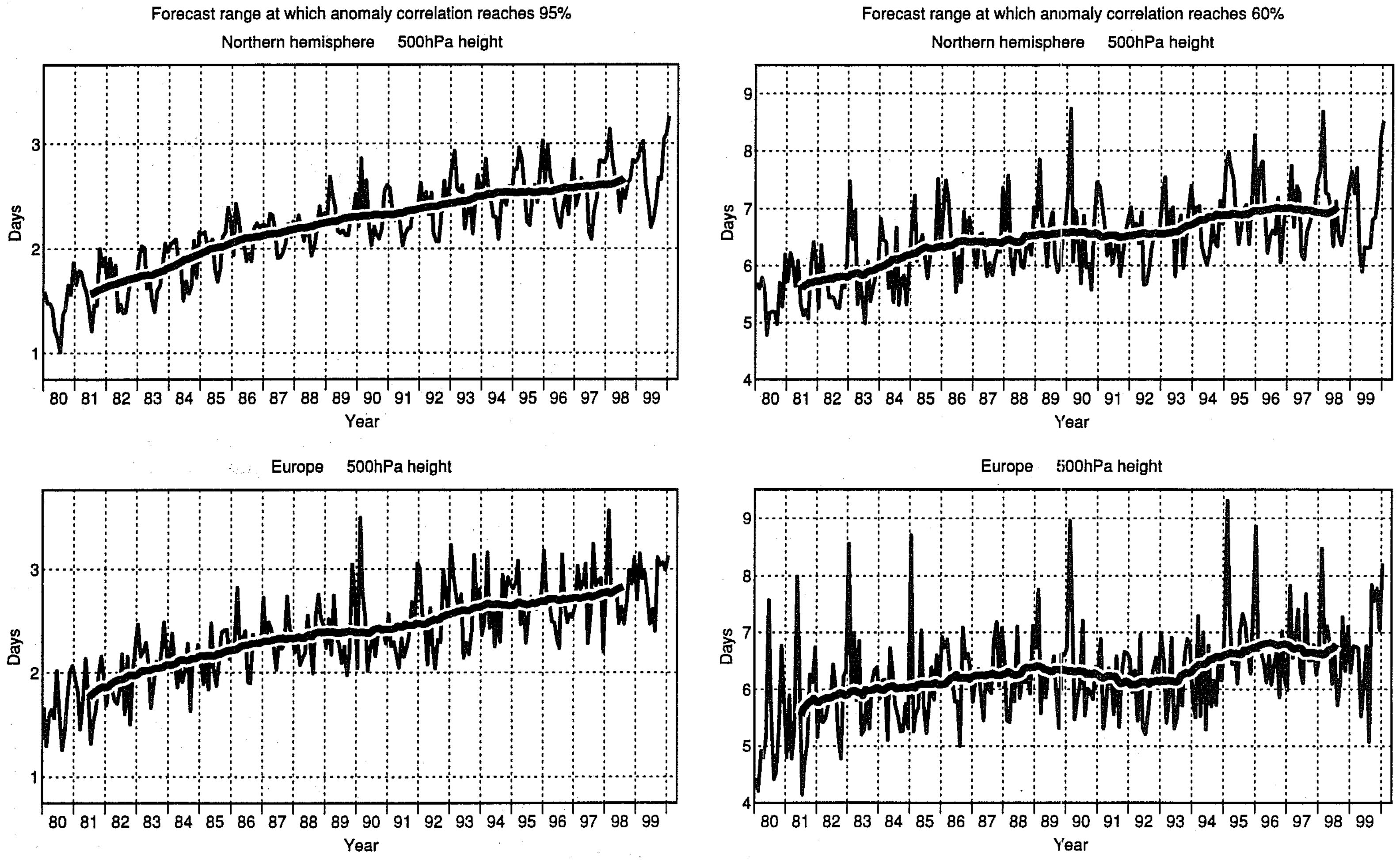


Fig.7. The forecast range at which the monthly mean 500hPa height anomaly correlation reaches 95% and 60%, for the extratropical northern hemisphere and for Europe, from January 1980 to January 2000. A three-year running mean is also shown.

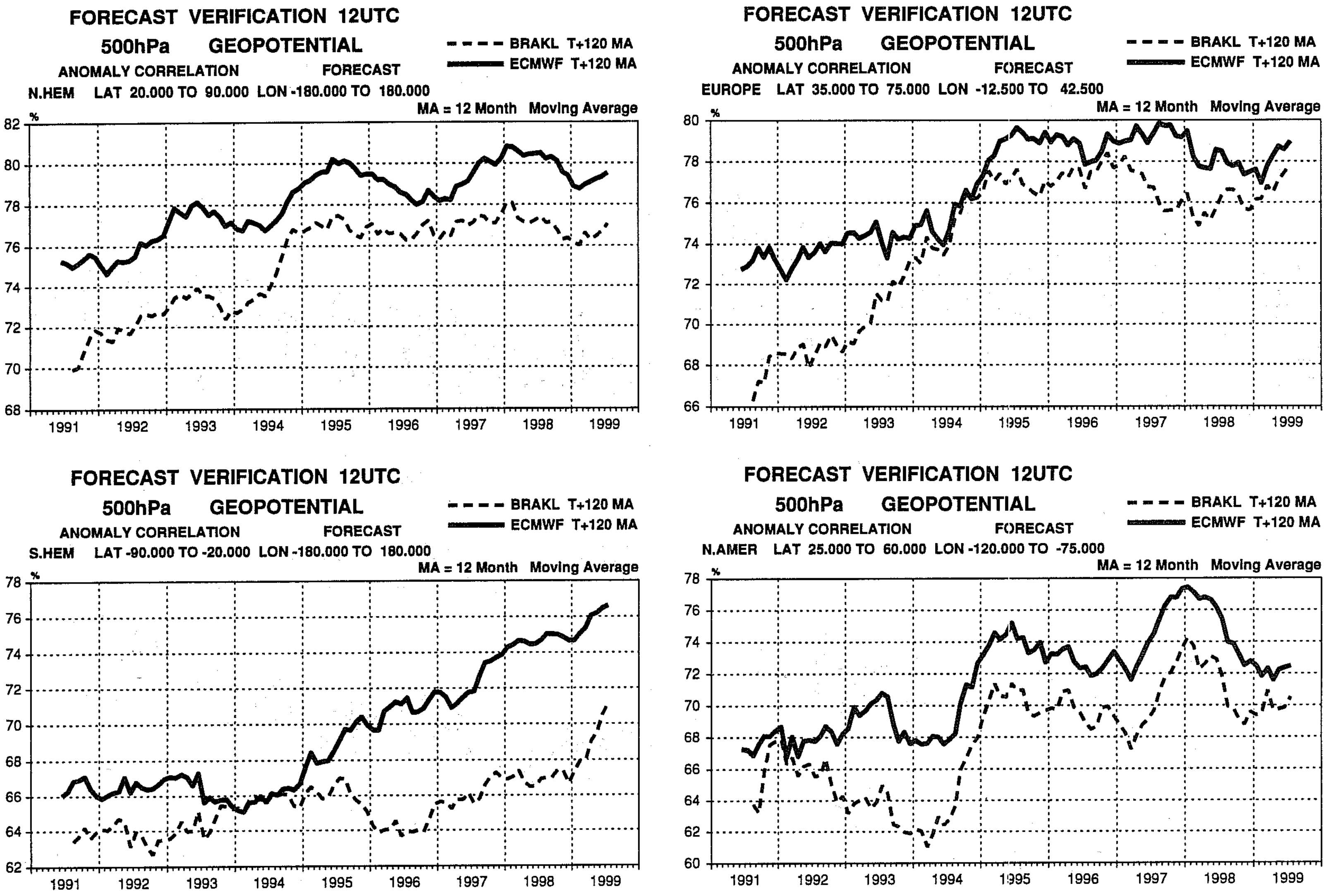


Fig. 8. Twelve-month running means of 500hPa height anomaly correlations for day-5 forecasts from ECMWF and UKMO, based on monthly-mean values from January '1991 to January 2000, for the extratropical northern and southern hemispheres and for Europe and North America.

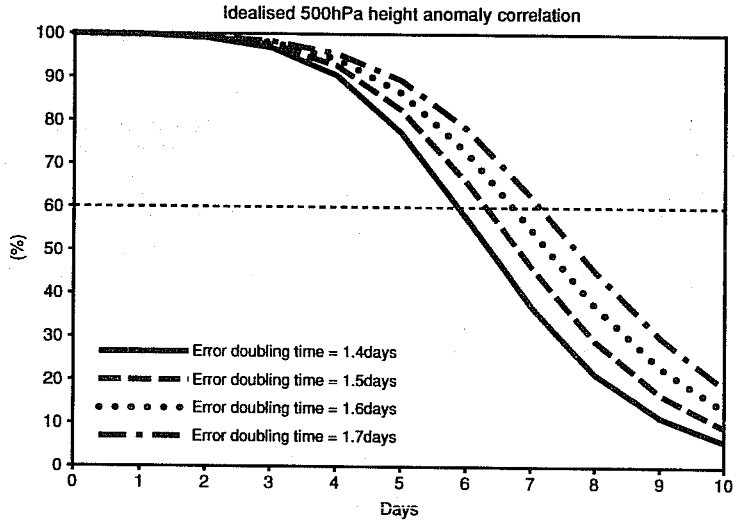


Fig. 9. Idealized 500hPa height anomaly correlations as a function of forecast range, indicating sensitivity to forecast error doubling time.

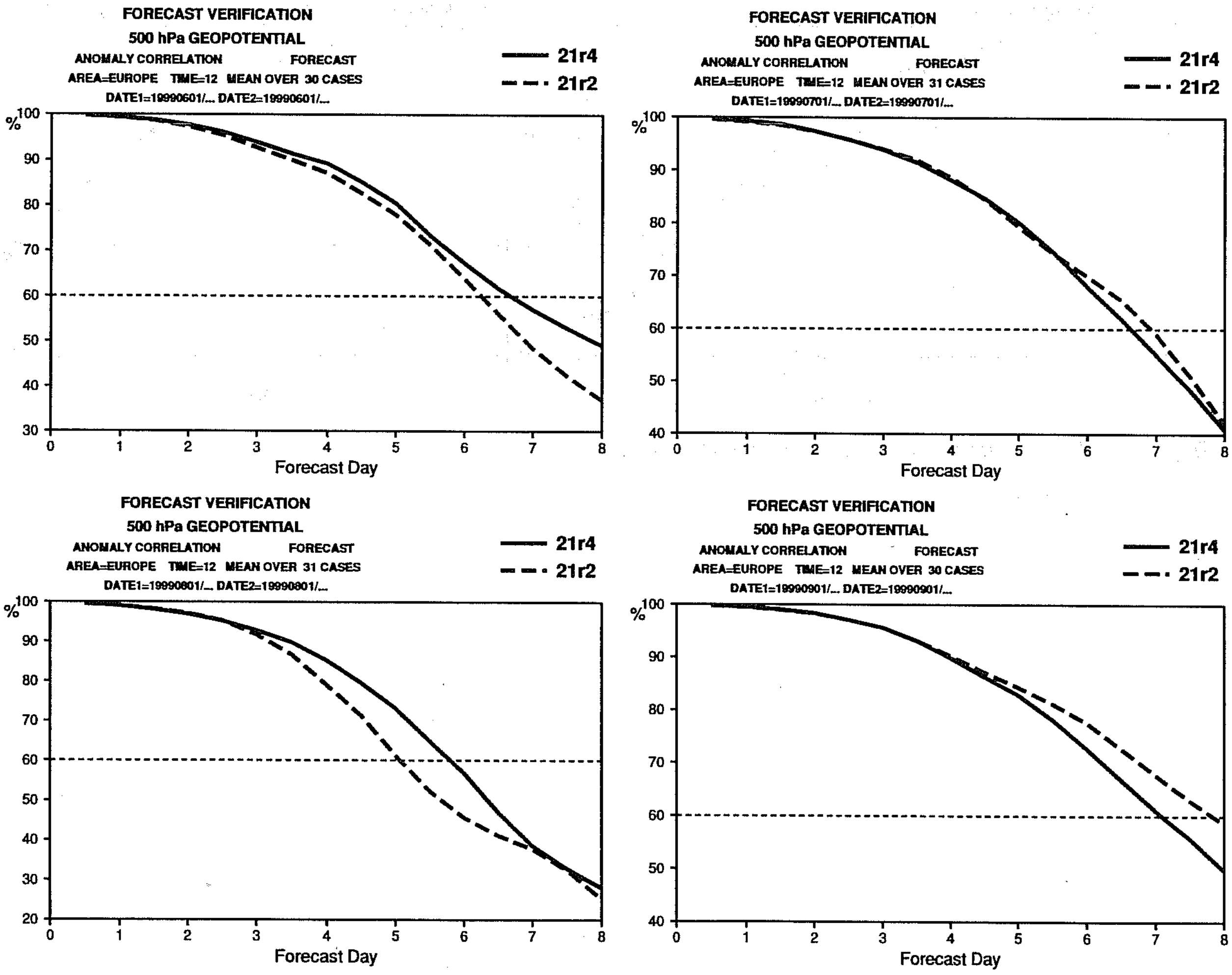


Fig. 10. 500hPa height anomaly correlation for Europe averaged for the months of June (upper left), July (upper right), August (lower left) and September (lower right), 1999. Results are shown for cycles 21r2 and 21r4 of the ECMWF forecasting system, using 50- and 60-level vertical resolutions respectively. Cycle 21r2 was operational from 13 July to 11 October 1999, after which it was replaced by 21r4.

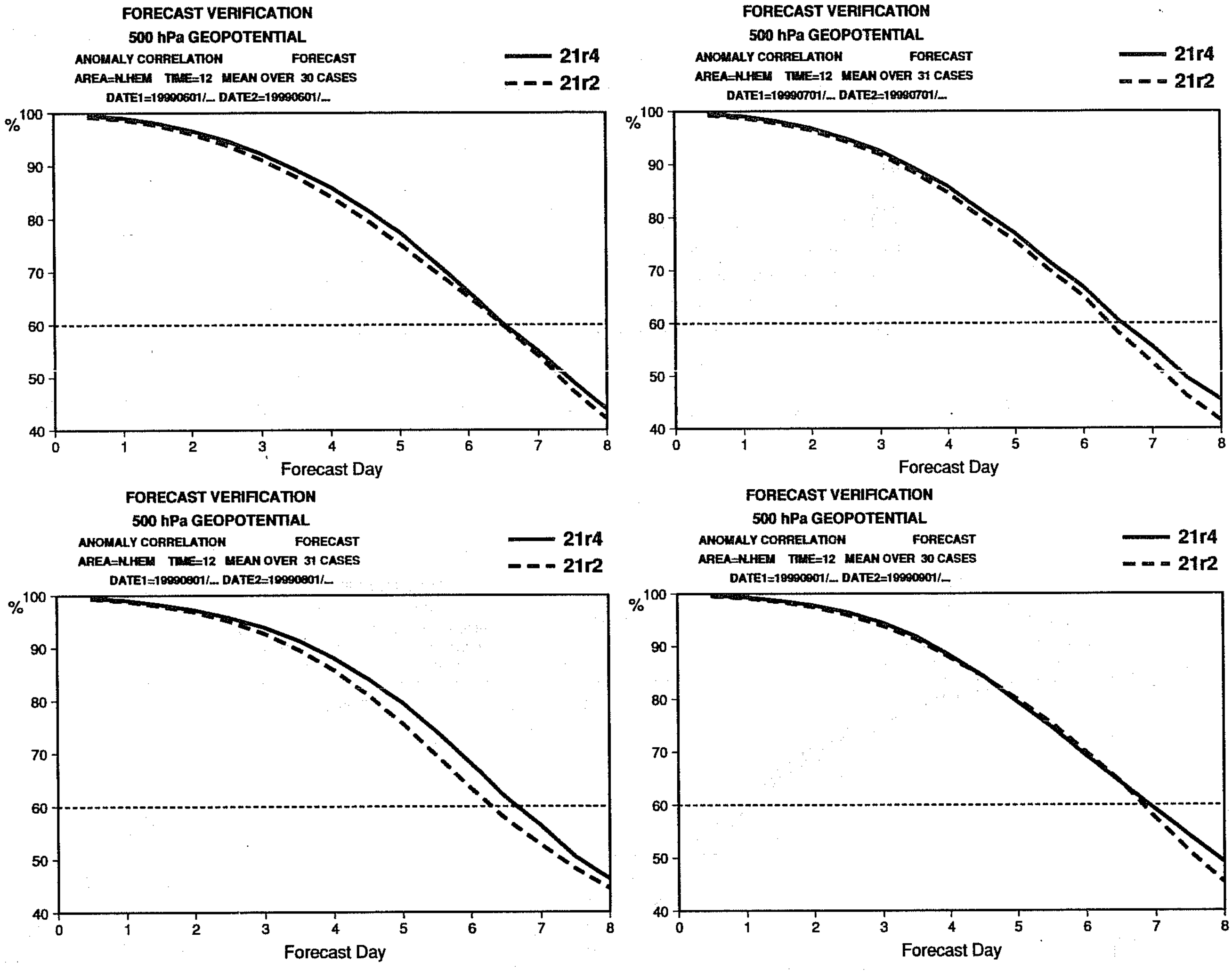
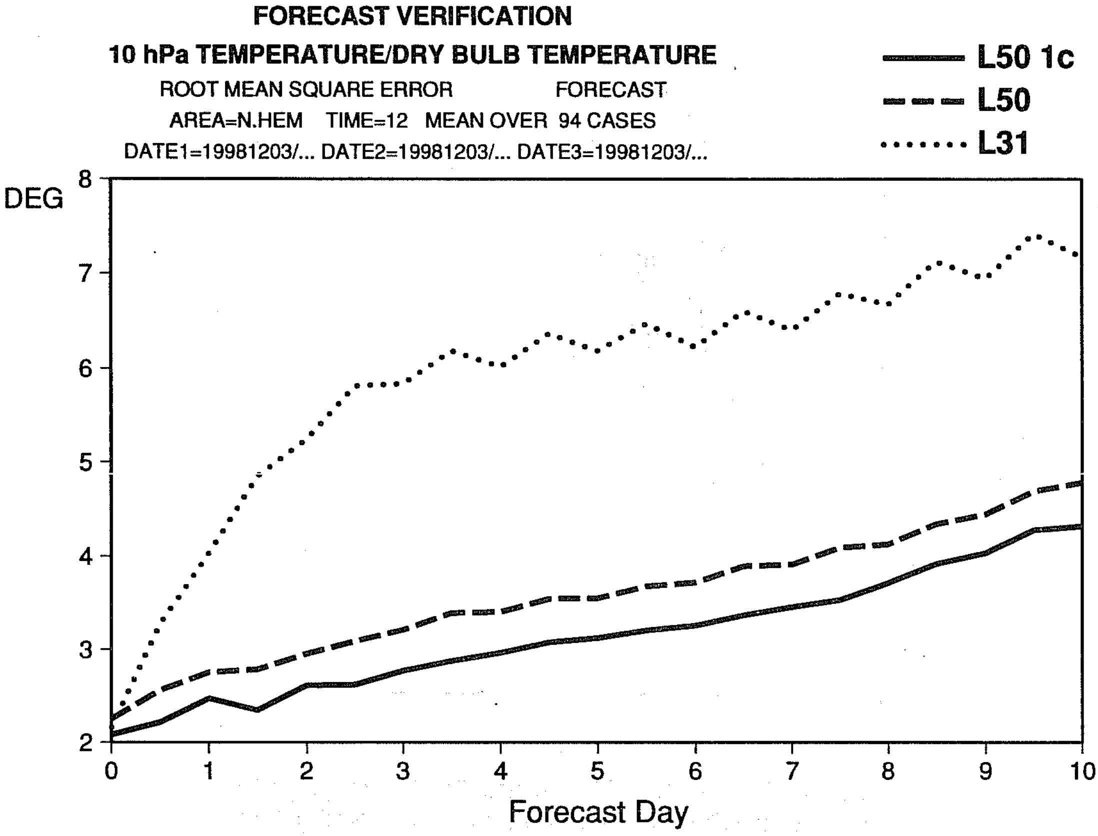


Fig. 11. As Fig. 10, but for the extratropical northern hemisphere.



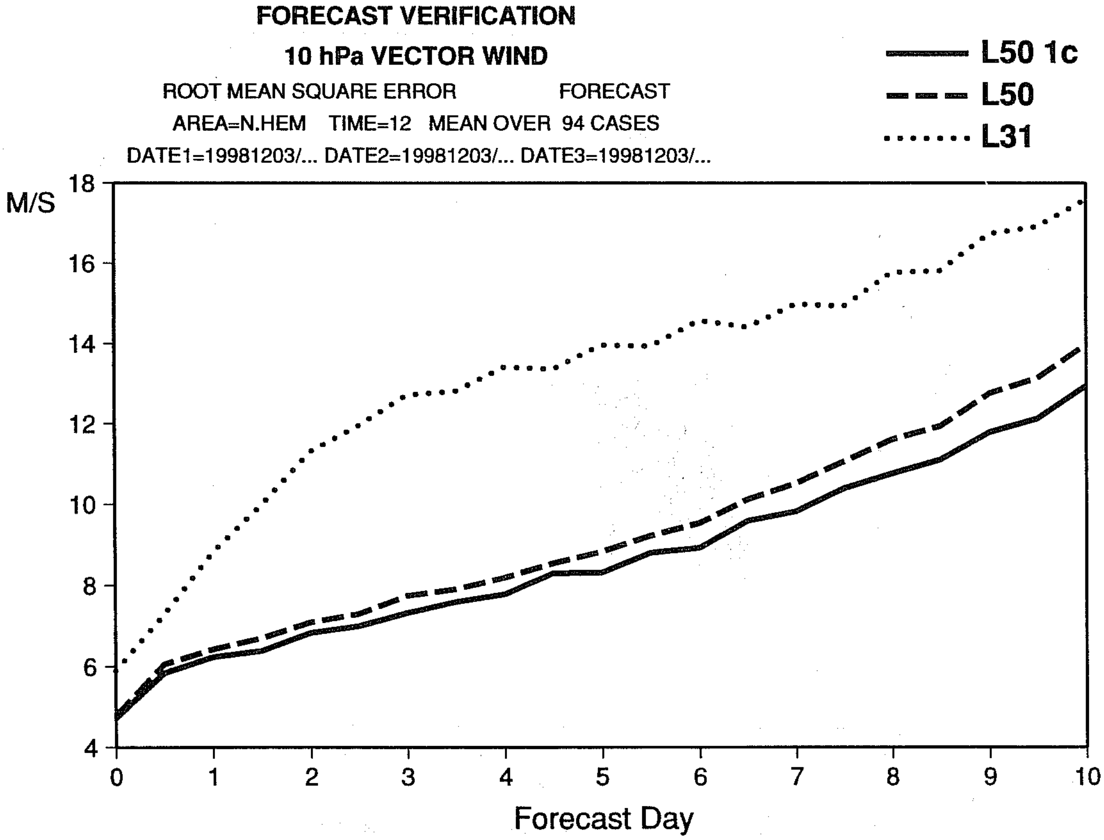
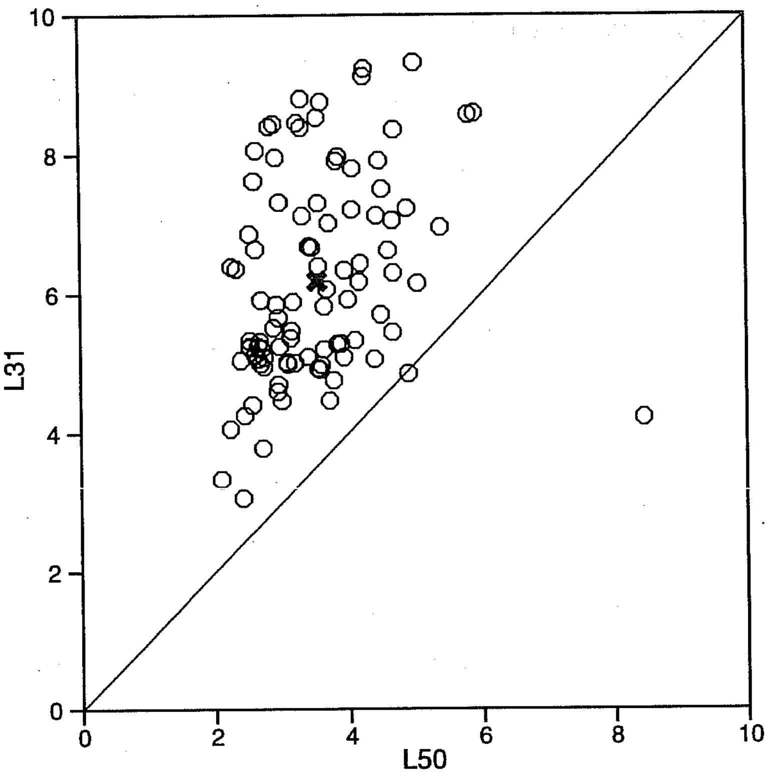
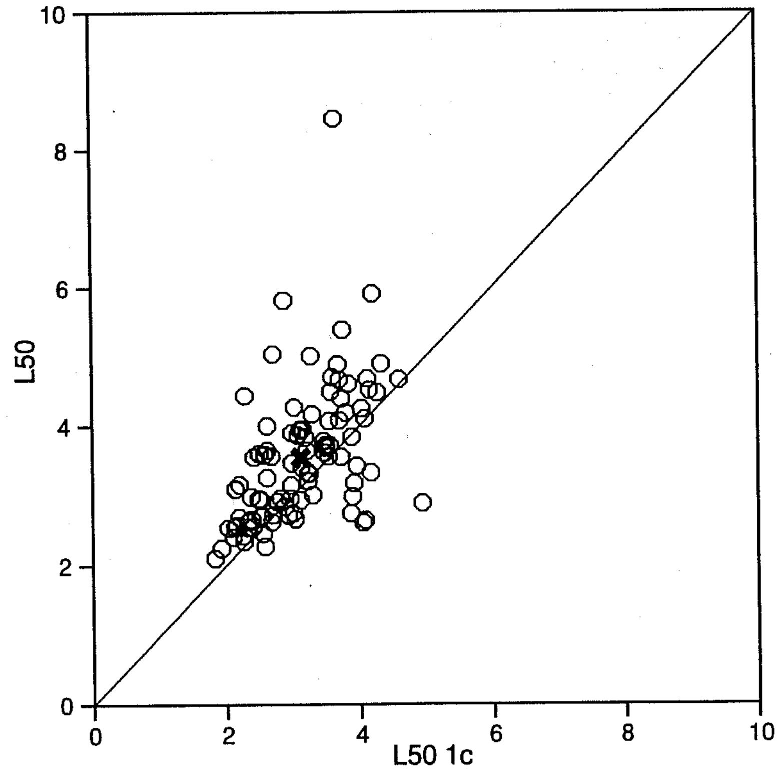


Fig. 12. Rms error of 10hPa temperature and vector wind forecasts verified against radiosondes, averaged over 94 cases within the period 3 December 1998 to 8 March 1999. Results are shown for 31- and 50-level vertical resolutions, the latter for two different versions of the assimilation of satellite radiance data.



L50 is BETTER than L31 at the 0.1% level (sign test)
L50 is BETTER than L31 at the 0.1% level (t test)



L50 1c is BETTER than L50 at the 0.1% level (sign test)
L50 1c is BETTER than L50 at the 0.1% level (t test)

Fig. 13. Scatter plots showing the distribution of rms errors of 10hPa temperature forecasts at day 5, for the experiments presented in Fig. 11.