PERFORMANCE OF GLOBAL MODELS IN SUPPORT OF MAJOR USER SERVICES

A J Gadd
Meteorological Office
Bracknell, UK

1. INTRODUCTION

The first 10 years of medium-range numerical forecasting at ECMWF correspond more or less to the first 10 years of the operational use of global NWP models in support of national weather services. By 1982 global models were operational at ECMWF, Washington and Bracknell. These three centres were joined by Tokyo in 1987 and Paris in 1988, and now in 1989 a global model is running routinely at Melbourne and several other centres have plans for the implementation of global models.

The inception of global NWP models seems likely to be judged as a highly significant step in the development of meteorological science. It has brought a widening group of forecasters and modellers into day-to-day contact with weather systems in all parts of the world. The "other hemisphere" has become of continuing concern rather than of passing interest, whilst the tropical belt has been analysed and forecast routinely as never before.

The horizontal resolutions adopted for operational global models have increased from time to time during the decade as the necessary computing resources have been made available at the various centres, and further increases in resolution are in prospect. Although stretched coordinate systems have attracted interest, many of the models are likely to retain near-uniform resolutions for the entire globe, so that the results are suitable for regional application in any part of the world. (Sometimes this feature is, in effect, part of a user's requirement.) The availability of such widely applicable results from these global models implies that any regional model requires a very high resolution, and therefore a very powerful computer, in order to be competitive.

Right from the outset, global models have been applied to two different forecasting tasks. Thus, the design and operation of the ECMWF model have been tailored to the requirements of medium-range forecasting, with a single production run each day from a late data cut-off. At Bracknell, on the other hand, the main emphasis has been on short-range forecasting, with the global model running every 12 hours (probably every 6 hours eventually) from a relatively early data cut-off. At Washington both a medium-range run and a pair of short-range runs (known as the aviation runs) have been included in the daily schedule.

For medium-range forecasting a model with global coverage is required on meteorological grounds, since information initially located in any part of the globe could in principle affect the 10 day forecast for a chosen region of interest. For short-range forecasting the global coverage is a requirement of certain users, notably aviation users but in other sectors too when a customer wants short-range advice from a single source for widely spaced locations or areas.

This paper discusses aspects of the measurement of performance for global models. Some of the difficulties in obtaining satisfactory measures of performance are clarified. Standardized verification data from Bracknell and ECMWF are used for illustrative purposes.

2. THE GLOBAL FORECASTING SYSTEM

In considering the performance of global models, it is important to remember that we are really considering the performance of complete global forecasting systems with structures such as that illustrated in Fig. 1.

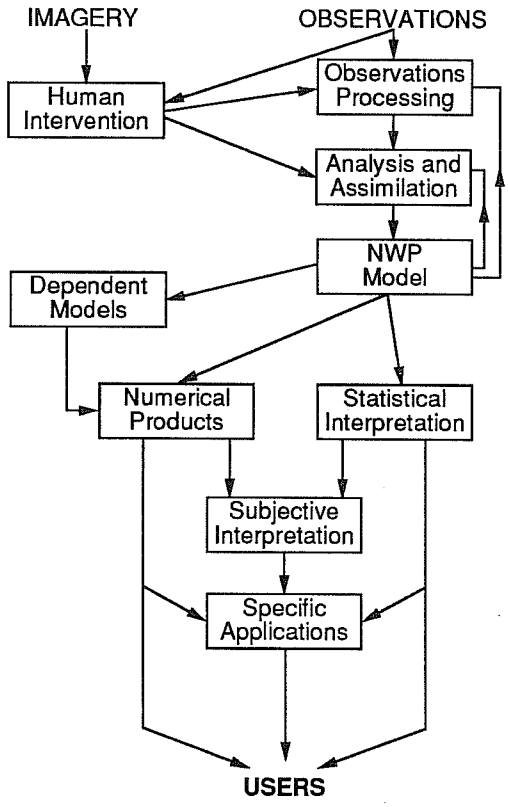


Figure 1 Sturcture of a complete global forecasting system

It is obvious enough that changes in the analysis and assimilation technique as well as changes in the NWP model itself can lead to changes in overall performance. Both these kinds of changes occur relatively infrequently. Changes in the observations processing, perhaps in response to routine data monitoring, are more frequent; often they have only a minor impact on performance but their cumulative effect can be significant. In some instances the human contributions to observations processing and data assimilation have an accepted ability to improve performance - the positioning of tropical cyclones is the notable example (Morris and Hall,1988).

From the users' perspective, changes in output processing (including statistical interpretation) or in dependent models can have important impacts on effective performance. Once again there is a human contribution where services rely on the interpretation of numerical forecasts; frost forecasts for coffee-growing regions is one example involving a global model.

3. SOME DIFFICULTIES ABOUT MEASURING PERFORMANCE

Suppose we wish to measure performance by comparing a set of forecast data f with suitable verifying data v, which itself constitutes an approximation to the truth t. A commonly used measure of performance is the root mean squared difference

$$\overline{\mathrm{SQRT}[(f-v)^{**}2]} \ ,$$

where the mean is taken over a sufficiently large set of locations distributed over space and/or time.

Some potential problems come quickly to mind. For example v may contain information on scales that f excludes by definition. Such information may sometimes be important to particular users, but nevertheless the measurement of performance will inevitably be distorted by the inclusion of these scales. We hypothesize that there exists a model-dependent partitioning

$$x = [x] + x'$$

where x may be f,v or t, such that x' represents any modes or scales that are not physical solutions of the governing equations or that cannot be represented at the model's resolution. The most meaningful definitions of the errors of the forecast data and the verifying data are then

$$ef = [f] - [t] \qquad ev = [v] - [t]$$

and the primed quantities are to be regarded as noise as far as the measurement of performance is concerned.

Assuming that primed quantities are uncorrelated with square bracketed quantities the following relationship may be established.

$$\overline{(f-v)^{**}2} = \overline{ef^{**}2} + \overline{ev^{**}2} - 2*\overline{ef*ev} + \overline{f'^{**}2} + \overline{v'^{**}2} - 2*\overline{f'*v'}$$

(i) (ii) (iii) (iv) (v)

Thus we may regard $\overline{(f-v)^{**}2}$ as an estimate of $\overline{ef^{**}2}$ which is (i) increased by errors in the verifying data, (ii) decreased by correlation of errors in the forecast data with errors in the verifying data, and (iii)/(iv) increased by noise in the forecast data or noise in the verifying data, except (v) where these are correlated.

Note that terms (i) and (iv) depend only on the verifying data and so remain constant for all forecasts verifying at a given time.

In practice most models include effective procedures to control f′ by the elimination of computational modes. Initial noise f′(0) may be generated as a result of data assimilation, but can be expected to be dispersed and dissipated fairly rapidly as the forecast proceeds. In addition of course the reduction of f′(0) is an important motivation in the design of data assimilation techniques. These procedures to control f′ may however have side-effects in increasing the error ef of [f], and for this reason a decision may sometimes be made to accept a particular noise component on a temporary basis. Other than that, term (iii) in the above equation may be expected to become smaller as the forecast period lengthens.

Another term in the above equation that is expected to diminish during forecasts is (ii). Since this depends on the correlation of errors in the forecast data with errors in the verifying data, it is only likely to be non-zero when the verifying data are analyses produced using the model being verified. Even then, as a forecast proceeds, ef is expected to become dominated by errors that grow during the forecast and that are uncorrelated with ev.

Term (v) in the above equation may often be negligible except for the very early stages of forecasts, but needs to be retained when interpreting, for example, the fit of analyses to observations. Instances can also be envisaged where, for verification of forecasts against analyses, f′ and v′ remain correlated well into the forecast period.

The above equation is useful when assessing the merits of alternative sources of verifying data. The alternatives include the following in principle.

1. Observational data, normally subdivided by observing system.

2. Processed observational data, eg radiosonde data averaged over model layers.

3. Analyses obtained by data assimilation using the model being verified.

4. Analyses obtained from data assimilations carried out at other centres.

5. Objective analyses obtained without the use of a forecast model.

6. Subjective analyses.

None of these is ideal. Observational data are noisy, prone to error, and patchy in distribution. Processed observational data would be better except that they are frequently inappropriate as a basis for comparing models. Analyses obtained by data assimilation using the model being verified are suspect as regards correlations of forecast errors with analysis errors. Use of analyses from a single other centre implies a difficult choice among centres, so averaged analyses using results from several centres have seemed more attractive. Given a careful design that avoids inconsistencies arising from interpolations, this approach may have interesting potential. Objective analyses obtained without the use of a forecast model would be likely to have large errors in data sparse regions. Subjective analyses could be ideal in principle, but fields are rarely available in digital form.

4. THE CHOICE OF PERFORMANCE MEASURES

Performance measures for global NWP models need to be chosen according to the issues that are being addressed. Sometimes the measure needs to be closely related to the user requirement, as for example when assessing equivalent tailwinds for aviation flight planning. Sometimes the measure that is required refers to derived fields or to fields from dependent models, as with the verification against buoy data of wave heights calculated using NWP surface winds (Francis and Stratton,1989). Sometimes the performance assessment is specifically designed for particular weather systems, as with the verification of tropical cyclone tracks (Morris and Hall,1988) or of the explosive deepening of extratropical cyclones (Gadd et al,1989).

In addition, however, there is a need for basic information on model performance and this is provided in the form of conventional verification statistics that are calculated routinely over chosen geographical areas. A source of verifying data must be chosen, and different choices may be appropriate according to the motivation of the verification study.

Three of the principal motives for verification are as follows.

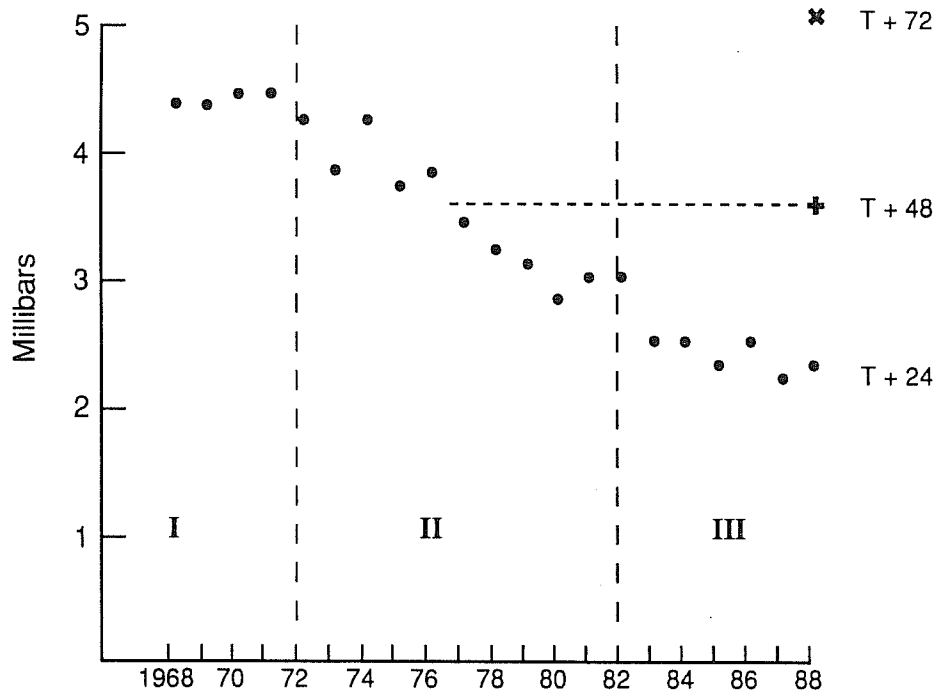
1. The documentation of trends in performance at a given centre over time, essentially to provide assurance that quality has been maintained or improved and that research and development have been beneficial.

2. The provision of evidence in support of proposed changes in an operational NWP system.

3. The monitoring of performance in relation to that at other centres, to assess the impact of changes and to detect new problems that may arise.

5. LONG TERM TRENDS IN PERFORMANCE

Figs. 2 and 3 show results from verifying Bracknell operational numerical forecasts since 1968 against gridpoint values extracted from subjectively prepared analyses of sea level pressure in the UK region. Although restricted to a single variable and a small geographical region, this source of verifying data is valuable for comparing recent performance with that of previous years. The two sequences illustrated are the annual values of the 24 hour rms pressure error and the 24 hour pressure change correlation.



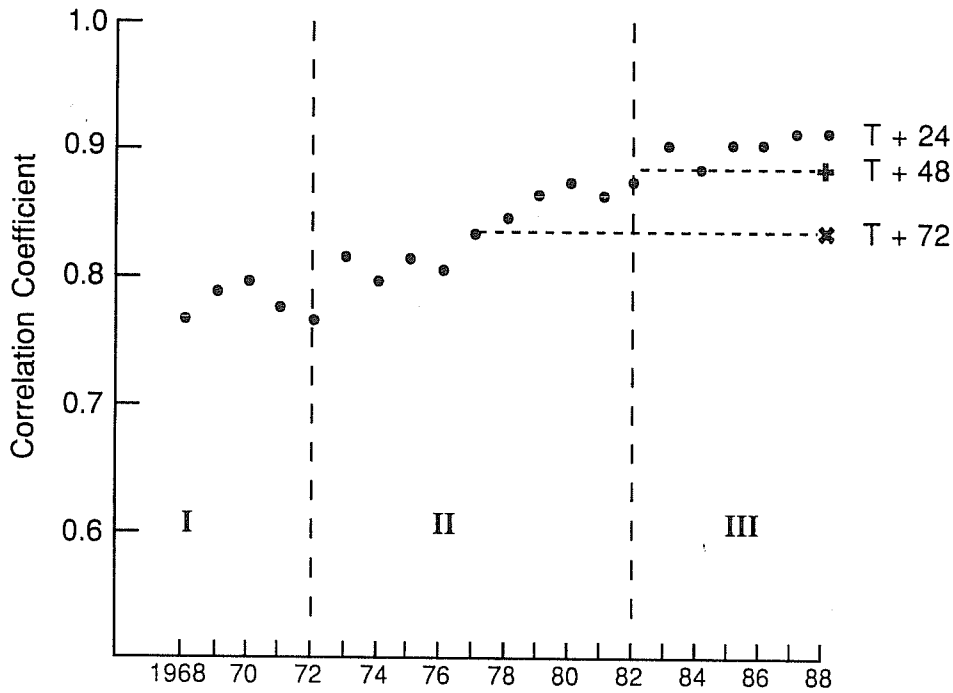
Root mean square errors of 24 hour forecasts of sea level pressure. (Annual values for 48 points near the UK.) Also of 48 and 72 hour forecasts for 1988 only.

I 3-level model
II 10-level model
III 15-level model

Figure 2

Both sequences show the advantages gained by the transition from a 3-level quasigeostrophic regional model (I) to a 10-level, primitive equation, extratropical northern hemisphere model (II) and to a 15-level global model (III). A trend of improving performance from the cumulative effect of smaller operational changes is also evident throughout. Other features in the curves arise mainly from interannual variability of the atmosphere in a small region.

Note that the two statistics lead to different quantitive estimates of the extent to which forecast accuracy has improved. Judged by the rms errors, 48 hour forecasts in 1988 were as accurate as 24 hour forecasts around 1976. The pressure change correlations give a more favourable result, with 1988's 48 hour forecasts matching the 24 hour forecasts as recently as 1982, whilst 1988's 72 hour forecasts match the 24 hour forecasts around 1977.



Correlations of forecast and analysed 24 hour changes in sea level pressure. (Annual values for 48 points near the UK.) Also of 48 and 72 hour changes for 1988 only.

I 3-level model
II 10-level model
III 15-level model

Figure 3

6. ASSESSMENT OF PROPOSED CHANGES

Assessment of a proposed change in an operational global NWP system requires the study of statistics with a wide geographical coverage, at several levels in the atmosphere, and preferably for a large number of independent cases. Some changes have the awkwardness of improving some aspects of model performance whilst making others worse, and the differing priorities of various centres then become relevant.

Limitations on the computing resources available for testing proposed changes have frequently presented difficulties, and various strategies have been adopted in different centres. Verifications against observations, despite their drawbacks, have certain attractions in this particular context. This is because verifications against analyses can be difficult to interpret since the verifying data as well as the forecast data are then affected by the proposed change.

A recent example of a change that was introduced largely on the basis of verifications against observations was the "analysis correction" method of data assimilation that has been used in the Bracknell global NWP system since 30 November 1988 (Lorenc et al, 1989). Particular weight was given to the improved verifications of 6 hour forecast fields against radiosonde observations, whereas the fit of the analyses to the observations was in general less close than with the previous data assimilation method.

7. ROUTINE MONITORING OF PERFORMANCE

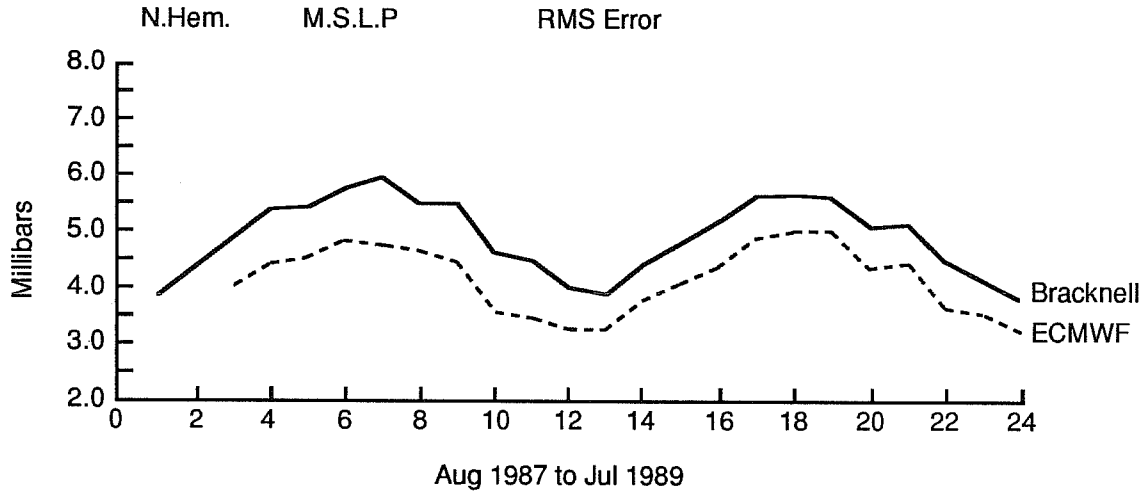
The ongoing comparison of verification scores from different centres is a valuable technique for monitoring the effects of operational changes and for detecting new problems arising from data or software. WMO's Commission on Basic Systems has agreed standard procedures for centres to use in verifying their own forecasts both against their own analyses and against synoptic observations (WMO, 1986). Beginning in 1987, there has been a gradual implementation of these standard verifications at the global modelling centres.

The CBS standard verification against synoptic observations is confined to four regions with good radiosonde coverage. These regions are located in North America, Europe, Asia and Australia/New Zealand. Even for these regions, the procedure is flawed at present because the standardization does not include the techniques to be used to exclude the effects of rogue observations. A variety of ad hoc techniques for this purpose are still in use at centres, and these distort any comparisons of results.

The CBS standard verification against analyses permits comparison of global models on a planetary scale. The regions used are the extratropical northern hemisphere (NH,90-20N), the

tropics (TR,20N-20S) and the extratropical southern hemisphere (SH,20-90S). The CBS standard is a successor to the WMO/CAS NWP Data Study that produced results (for the extratropical northern hemisphere only) from 1979 onwards. Over the years it has become customary to compare extratropical forecasts at the 72 hour forecast stage in the first instance.

Fig. 4 shows the CBS standard results for Bracknell and ECMWF 72 hour forecasts from 12 UTC data for extratropical sea level pressure over a two year period ending in July 1989. The familiar advantage to ECMWF in the northern hemisphere curve is evident. At a qualitative level this advantage seems to have been unchanged by the major operational changes at ECMWF in January 1988 (no vertical diffusion above the boundary layer) and May 1988 (analysis changes) or at Bracknell in November 1988 (new assimilation technique). An advantage to ECMWF is evident also in the southern hemisphere curve, though to a lesser extent. (The Bracknell results for the southerm hemisphere were affected by an error that degraded the sea surface temperature fields for most of 1988.)



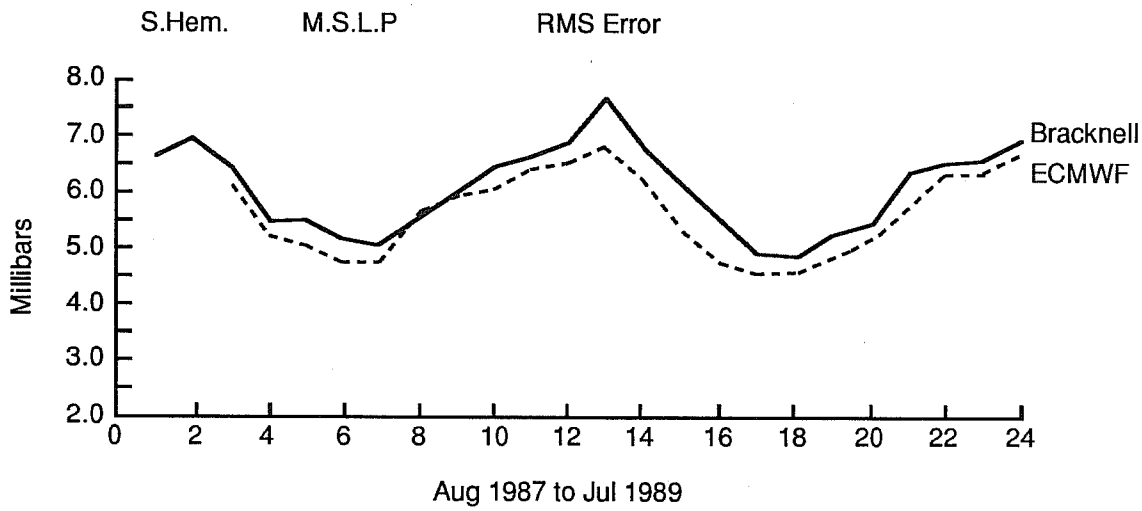


Figure 4 Standard verification vs analyses for 72 hour forecast of sea level pressure

Fig 5 shows similar comparisons but this time for the 250hPa wind. In this case results for the tropics are also part of the CBS standard. The change in January 1988 shows up very clearly in the verifications for the tropics, consistent with the effects noted in 1984 when vertical diffusion was introduced in the tropics of the Bracknell model (Watkins,1987). The change in November 1988 also now shows up clearly in the results for the northern hemisphere.

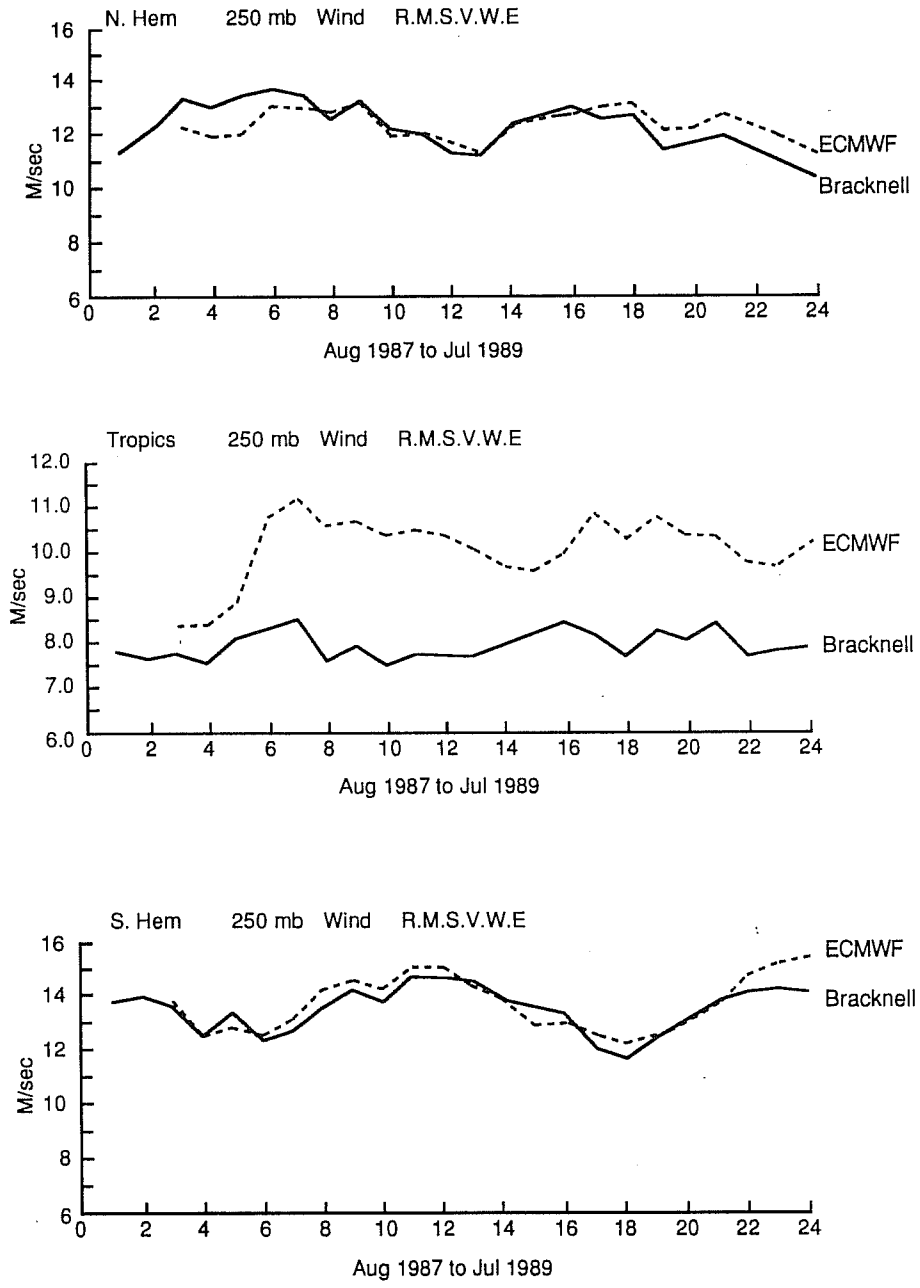
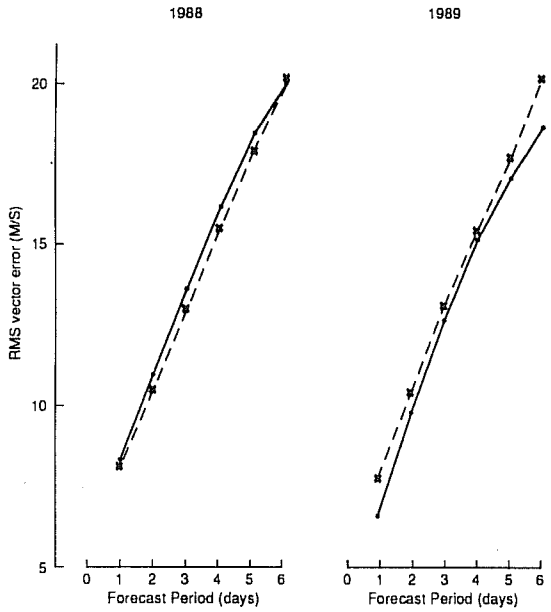


Figure 5 Standard verification vs analyses for 72 hour forecasts of 250 hPa wind

Focussing further on the 250hPa wind field, and on the introduction of the analysis correction scheme at Bracknell in November 1988, Figs 6 - 11 display the evolution of rms vector errors to 6 days ahead in the three regions for the months of January and July in 1988 and 1989, ie before and after the change. A systematic shift in relative performance is indicated, with lower errors from Bracknell at all stages of the forecasts in all three regions in both months in 1989.

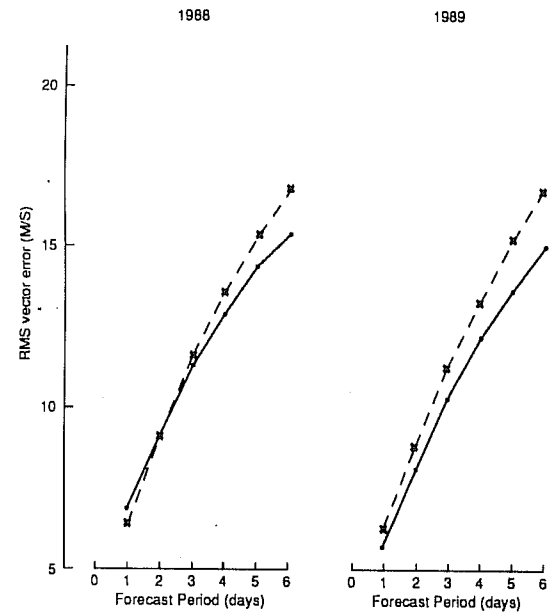
The shift in the Bracknell 250hPa rms wind errors is fairly consistent from month to month. The 6 month averages for 24 and 72 hour forecasts during the two periods December 1987 - May 1988 (period 1) and December 1988 - May 1989 (period 2) are tabulated below for the three areas.

	Area/Period						
	NH/1	NH/2	TR/1	TR/2	SH/1	SH/2	
T+24 :	8.0	6.4	6.0	5.2	7.8	5.9	m/s
T+72 :	12.9	12.0	8.0	8.1	13.4	12.8	m/s



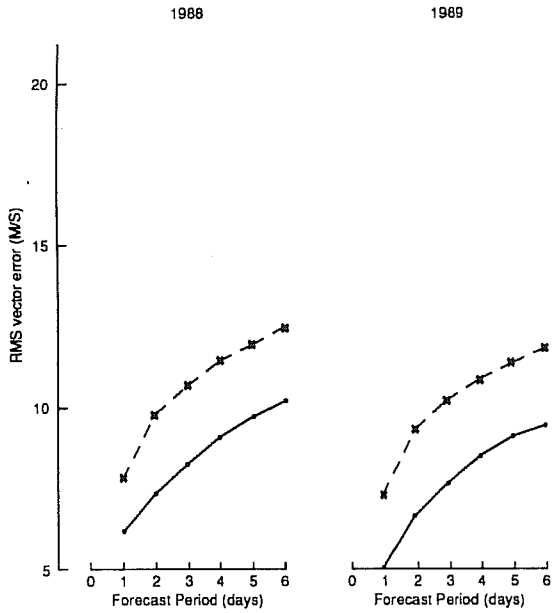
Standardized Verification vs Analyses
250 hPa Wind
Northern Hemisphere: January
Bracknell ECMWF

Figure 6



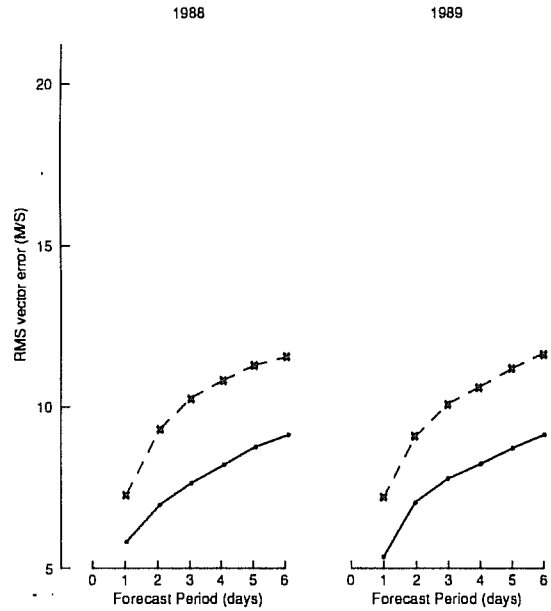
Standardized Verification vs Analyses
250 hPa Wind
Northern Hemisphere: July
Bracknell ECMWF

Figure 9



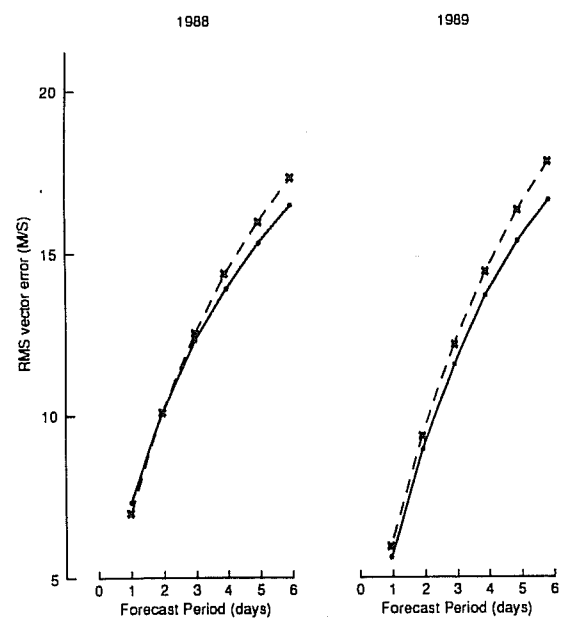
Standardized Verification vs Analyses
250 hPa Wind
Tropics: January
Bracknell ECMWF

Figure 7



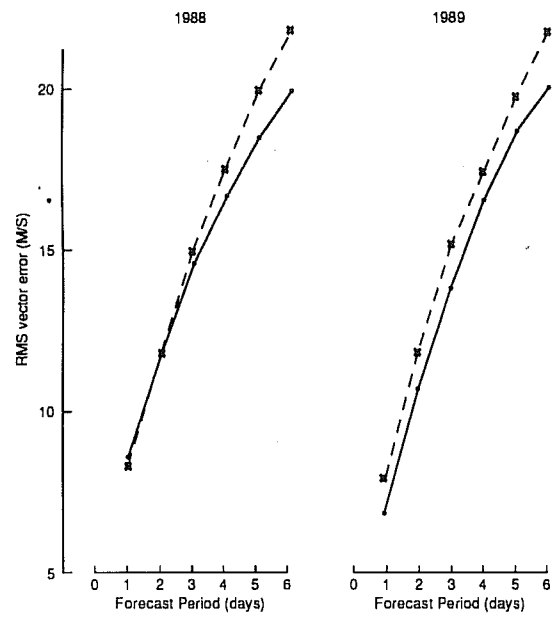
Standardized Verification vs Analyses
250 hPa Wind
Tropics: July
Bracknell ECMWF

Figure 10



Standardized Verification vs Analyses
250 hPa Wind
Southern Hemisphere: January
Bracknell ECMWF

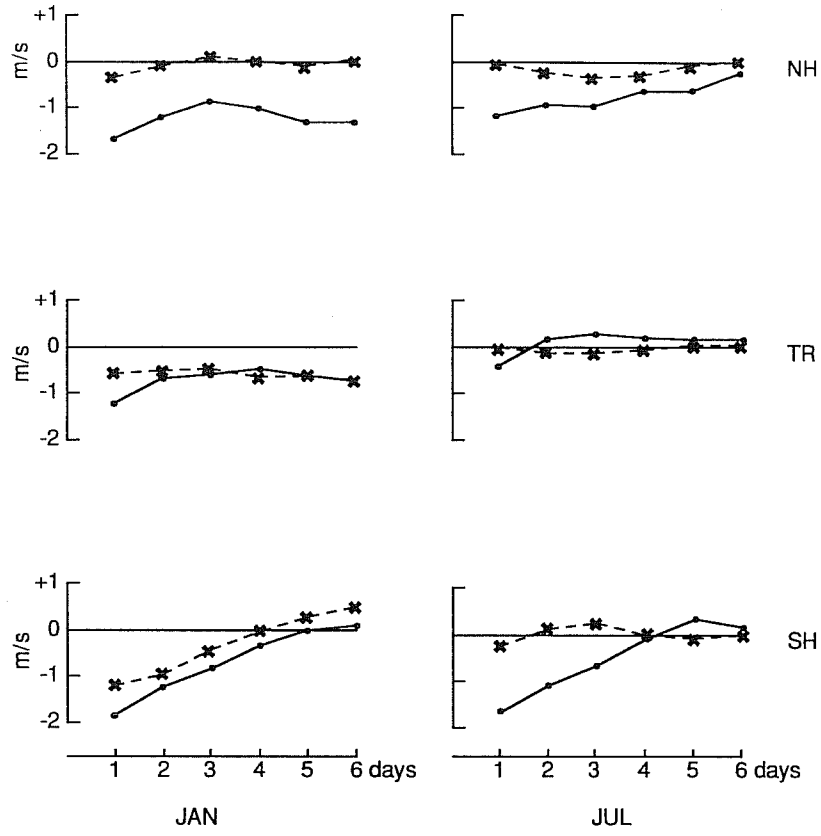
Figure 8



Standardized Verification vs Analyses
250 hPa Wind
Southern Hemisphere: July
Bracknell ECMWF

Figure 11

Fig. 12 compares the year-on-year differences for the Bracknell and ECMWF scores displayed in Figs. 6 - 11. Interpretation is not straightforward since the changes at ECMWF in May 1988 (analysis changes) and May 1989 (physics package) may have had some effect on this statistic and the possibility of a signal from interannual variability in the atmosphere should not be overlooked.



1989 minus 1988 values of RMS Errors for 250 hPa wind measured against analyses using the CBS standard verification.

Figure 12

The impression gained from Fig. 12 is of a different signature of the Bracknell change in each of the three regions. In the northern hemisphere an improved performance seems to be sustained throughout the 6 days, particularly in January. In the tropics the reduction in error is confined to the 24 hour forecasts. In the southern hemisphere the improved 1989 scores converge back to the 1988 level by about 4 days in July and by about 5 days in January, but in the latter case the ECMWF behaviour is similar. Further elucidation might have been possible using, say, the Tokyo verification data, but these are not available for January 1988. The same is true of the Paris results, which in any case only extend to 2 days ahead fom 12 UTC analyses. Washington does not yet distribute verifications against analyses.

Referring to the equation and discussion in section 3, these differing signatures may suggest differing contributions from

(i) a genuine improvement in skill;

(ii) a reduction in the forecast and/or analysis noise;

(iii) an increase in correlations of forecast error with analysis error.

A contribution from (ii) is almost certain and a contribution from (iii) is plausible from the known characteristics of the old and new data assimilation techniques (Lorenc et al 1989). The effects remain unquantified, though some estimates of analysis errors can be obtained from detailed study of the observations monitoring statistics that are now produced routinely at some NWP centres (Hall,1988).

8. CONCLUSIONS

Assessment of global model performance is intrinsically more complex than is often implied when verification scores are quoted or compared. The CBS standard verifications will produce a wealth of useful information that should assist centres to target onto key problems for each model. Although great progress has been made in the 10 years since ECMWF began operational medium-range numerical forecasting, much remains to be done if the best features in the performance of the several NWP systems are to be combined successfully.

Acknowledgement: I am grateful to Chris Hall, Wendy Adams and Simon Fuller for providing the verification data used in this paper.

References

Francis, P.E. and R.A. Stratton, 1989: Some experiments to investigate the assimilation of SEASAT altimeter wave height data into a global wave model. Submitted to QJ Roy Met Soc.

Gadd, A.J., C.D. Hall and R.E. Kruze, 1989: Operational numerical prediction of rapid cyclogenesis over the North Atlantic. To be published in Tellus.

Hall, C.D., 1988: Systematic errors in short-range forecasts of wind in the tropics. Workshop on systematic errors in models of the atmosphere, Toronto, 19-27 September 1988, WMO/TD-No.273.

Morris, R.M., and C.D. Hall, 1988: Forecasting the tracks of tropical cyclones with the UK global model. ESCAP/WMO Typhoon Committee Annual Review 1987.

Watkins, F.M., 1987: A comparison of the systematic errors of the UK global model in 1984 and 1985. WMO/IUGG NWP Symposium, Tokyo, 4-8 August 1986.

WMO, 1986: Commission for Basic Systems, Abridged Final Report of the Extraordinary Session, Hamburg, 21 October - 1 November 1985, WMO-No.654 (see pp 55-58).