JOINT MEDIUM RANGE ENSEMBLES FROM THE UKMO AND ECMWF MODELS

R E Evans, M S J Harrison and R J Graham
United Kingdom Meterological Office
Bracknell, U.K.

1 Introduction

The European Centre for Medium Range Weather Forecasting (ECMWF) has produced operational medium-range ensembles since December 1992 (Molteni *et al.* 1996). Dynamically active perturbations are derived from linear combinations of singular vectors calculated from the adjoint of the linear tangent model. The current operational ensemble prediction system (EPS) at ECMWF uses a T42 adjoint model to generate 25 pairs of perturbations; these are added to the operational analysis and integrated out to Day 10 using the T_L159 model.

Deficiencies in parametrisations, systematic or regime dependent model errors can severely affect the skill of the ensemble (Molteni *et al.* 1996; Toth and Kalnay 1995) but the EPS at ECMWF does not consider sensitivity to the model formulation. Both model and analysis dependencies are introduced in the Canadian System developed at Recherche en Prévision Numérique (RPN), where several versions of the operational model with modified physics are used to produce both analyses and ensemble members (Houtekamer *et al.* (1996), and Lefaivre in this volume).

Another possible method of including model uncertainties (in addition to perturbations in initial conditions) in the EPS is to incorporate members run using additional NWP models. Harrison *et al.* (1995) demonstrate the benefits of multi-model and multi-analysis (MMMA) ensembles using low resolution UKMO and ECMWF models initialised from their own analyses but using the same perturbations. Over the 38 cases studied, the joint ensembles provided improved information over both single-system ensembles, and this improvement was found to be beyond that due to a simple increase in ensemble size.

This study builds on the work of Harrison *et al.* (1995); the benefit of MMMA ensembles is investigated with higher resolution versions of the ECMWF and UKMO models run from their own analyses. Initially the investigation concentrates on deterministic forecasts generated by ensembles,

i.e. the ensemble mean. This leads on to an examination of the spread around the ensemble mean and its value as an indicator of ensemble mean skill. But, of course, one of the main aims of ensemble forecasting is to generate an accurate forecast of the probability density functions of various components of the atmosphere. So, in Section Five, a number of verification methods are used to examine the potential benefits of joint ensembles for probabilistic forecasting.

The proven benefits of MMMA ensembles may stem from intrinsic differences between the individual systems with previous studies providing support for the hypothesis that combining forecasts produced by different methods, or models represent a means of enhancing forecasting performance (Brown and Murphy 1996; Vislocky and Fritsch 1995). So, in Section Six the extent of independent information provided by the ECMWF and UKMO systems is examined.

Cost considerations mean that it is important to assess if both analyses and models are required for the largest benefits, i.e. could similar gains in skill be achieved though use of either a single model with two analyses or a single analysis with two models? Attempts to answer this lead on to investigations of the relative importance of model and analysis dependencies. Rabier *et al.* (1996) use adjoint techniques to demonstrate the importance of analysis dependencies in producing forecast error, estimating that analysis imperfections account for around 10% of the 2-day error in the ECMWF model. However it is also acknowledged that model error can severely affect forecast skill (Molteni *et al.* 1996). In the last part of this paper the relative importance of model and analysis dependencies is discussed using results from hybrid ensembles - UKMO model run from ECMWF analyses but again using ECMWF perturbations.

2 Data and analysis methods

The UKMO component of the experiment is the HADAM2b version of the Unified Model (UM) (Hall *et al.* 1995) at 1.25° by 0.83° horizontal resolution with 19 vertical levels. The ECMWF integrations were generated using the T_L106 model, which corresponds to an approximate grid resolution of 1.70° by 0.85°, with 31 vertical levels. Each model has been run from the operational analysis of the 'home' centre and in all cases the same 16 pairs of ECMWF perturbations calculated at T42 have been added to the analyses. In addition a number of forecasts have been produced using the UM initialised from the ECMWF analysis - the hybrid ensemble. Brief studies have indicated that additional perturbations created by interpolating ECMWF analyses over the UM orography dissipated within 12 hours.

Ensembles were run for ten cases (Table 1), with initial dates chosen in consultation with ECMWF. Due to computer resource limitations there are only 4 complete 33-member ensembles of the hybrid configuration (UM model run from ECMWF analysis). And for Case 1 only 17 members of the UKMO and ECMWF configurations are available (8 perturbation pairs plus the control run). The ensemble systems are referred to by 2 letters - the first indicating the model used and the second indicating the model used to create the initial analysis (e.g. UU = UKMO model from UKMO analysis). Four types of combined ensemble are available for assessment:

- Combination of each model off the analysis of the home centre using all members
  - referred to as EEUU66 - 66 member ensemble - assess total impact.

- Combination of each model off the analysis of the home centre using just the control and first 8 pairs of perturbations from each system
  - referred to as EEUU34 - 34 member ensemble - assess potential benefit of MMMA ensembles achievable without a significant increase in overall ensemble size and computational costs.

- Both models run from the ECMWF analysis, the control and first 8 pairs of perturbations.
  - referred to as EEUE - 34 member ensemble - assess impact of adding a second model to the ECMWF system.

- UM run from both analyses, the control and first 8 pairs of perturbations.
  - referred to as UUUE - 34 member ensemble - assess impact of adding a second analysis to the UKMO system.

In addition, model dependencies can be examined by comparisons between ECMWF and hybrid systems (both run from ECMWF analysis but using different models), and similarly analysis dependencies can be examined by comparing UKMO and hybrid ensembles (both systems use UKMO model but run from different base analyses).

Table 1. Ensemble members available for the 3 ensemble types for the 10 cases.

UU - UKMO model from UKMO analysis.

EE - ECMWF model from ECMWF analysis.

UE - UKMO model from ECMWF analysis

CASE	DATE	NUMBER OF MEMBERS AVAILABLE		
		UU	EE	UE
1	04 Nov 94	17	17	17
2	09 Nov 94	33	33	33
3	11 Nov 94	33	33	17
4	27 Nov 94	33	33	17
5	10 Nov 94	33	33	17
6	12 Dec 94	33	33	19
7	18 Dec 94	33	33	33
8	24 Dec 94	33	33	33
9	31 Dec 94	33	33	17
10	15 Jan 95	33	33	33

UKMO 0000Z analyses are used for the verification of 500 hPa height field, while 850 hPa temperature fields are verified against ECMWF 0000Z analyses. Ensembles were initialised at 1200 UTC, and forecasts verified at 0000 UTC at daily intervals; forecast ranges are represented in terms of hours (NN) using the convention T+NN. The analysis concentrates on Europe and the North Atlantic, although extra-tropical Northern Hemisphere fields are also examined. A number of measures are used to compare forecast performance for both deterministic and probability forecasts. Relative forecasting performance can also be assessed in terms of percentage improvement over some standard of reference using the skill score defined in Stanski *et al.* (1989),

$$SS = 100 * \left(\frac{SC - ST}{PS - ST}\right) \qquad (1)$$

where SC is the numerical value of performance measure for the forecast of interest, ST is the corresponding value achieved by the standard forecast and PS is the score for a perfect forecast. This

measure is used throughout the paper with the ECMWF forecast as the standard forecast in order to indicate the percentage skill improvement/reduction over ECMWF ensembles.

For the majority of analyses the results are based on the nine cases where the full 33 member EE and UU ensembles are available (Table 1). But for a number of methods results from all available members from all 10 cases are utilised; the number of cases used is described in each section.

3 Improvements to deterministic forecasts

In the following section the ensemble mean (EM) forecast produced by each configuration is verified using the basic measure of root mean squared error (RMSE) from verifying analysis. Deterministic forecasts from ensemble means represent a simplification of the information contained in the ensembles, but verification of the EM is a useful and simple way of comparing the properties of the systems. In addition, the skill of each model is assessed by comparing the results with those from an internally consistent ensemble (ICE), these are estimated by taking one member at random to be the verifying analysis, and averaging results over many Monte Carlo-type iterations. All results in this section refer to the European area (20°W-40°E, 75°-30°N) and are averages over the 9 cases for which the full 33 member EE and UU ensembles are available.

3.1 500 hPa height

At all lead times the joint ensemble mean of EEUU66 has lower RMSE than either individual system (Fig. 1a). The performance of the EEUU34 system is, for practical purposes, equivalent to that of EEUU64 for this diagnostic - a result that demonstrates that the improved skill of the joint ensemble mean can be achieved without a (significant) increase in the size of the ensemble. This improvement achieved by the EEUU34 ensembles is equivalent to a gain in lead time of around 18 hours at T+168. Recall that 500 hPa height is verified against UM analyses which may give the UU ensemble an advantage in the very early stages of the forecast - indeed UU does have lower RMS errors than EE until around T+24, but after this time EE is the most skilful individual ensemble. Calculation of skill score with EE as the standard reference quantifies the clear improvement in RMS errors achieved by the joint ensemble; until T+204 the RMS errors of EEUU34 ensemble mean forecasts generally represent a 5% improvement over the EE ensemble mean errors (Fig. 1b).

Comparison with ICE mean errors provides further evidence of the superior performance of the MMMA ensemble mean (Figs 1c-h). The EM RMS error of both individual systems is higher than

Figure 1a. Root Mean Squared Error of ensemble mean 500 hPa forecasts over Europe from T+12 to T+228 (averaged over 9 cases (cases 2-10)). Results for four configurations are shown: UU, EE, EEUU34 and EEUU66.

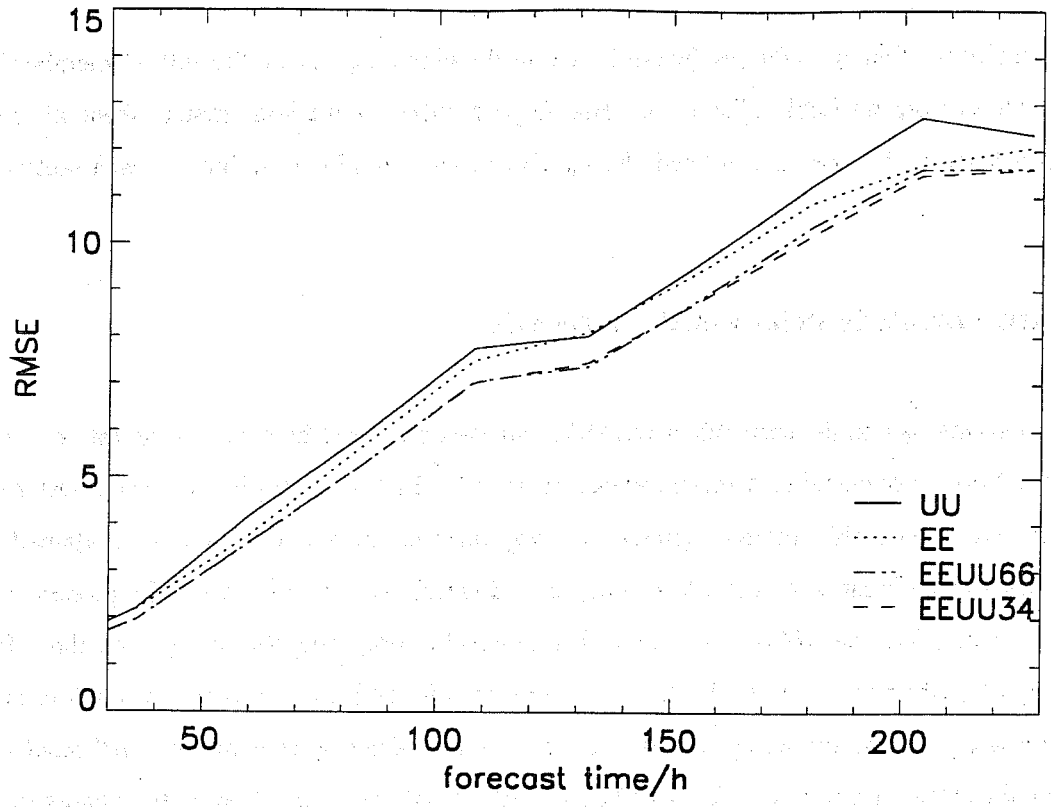


Figure 1b. As Figure 1a. but results are presented as percentage improvement over EE.

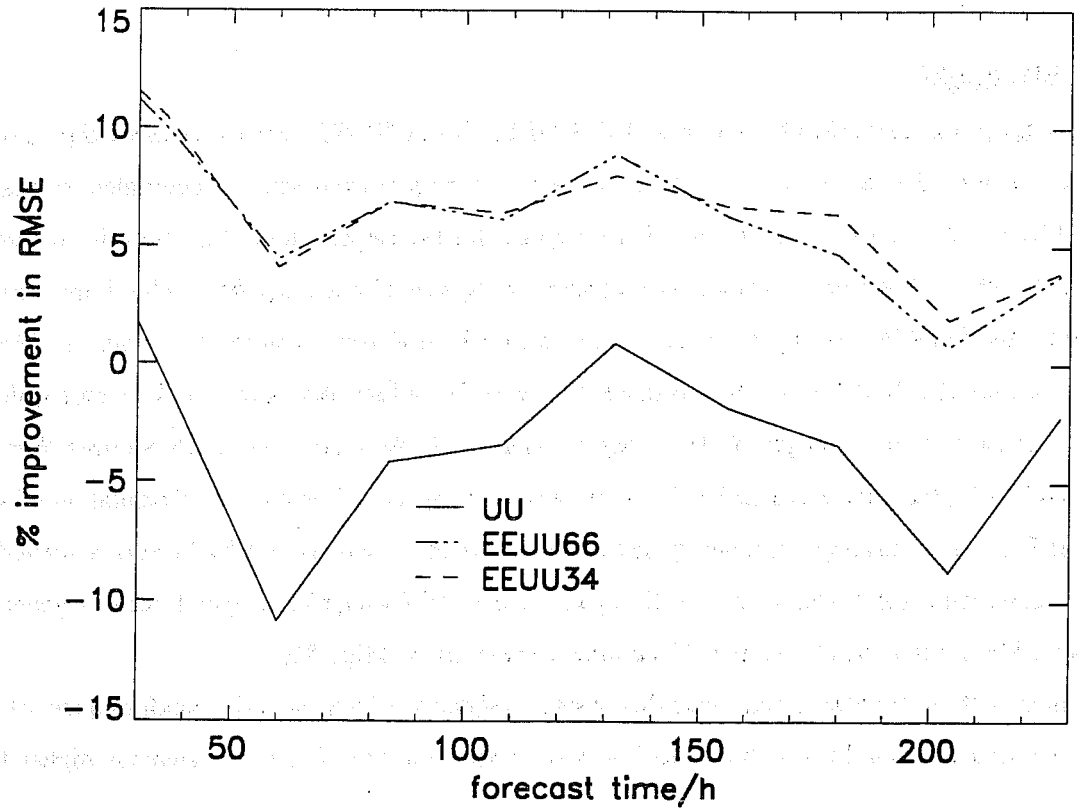


Fig. 1c. As in Figure 1a but for UU and internally consistent ensemble (ICE)

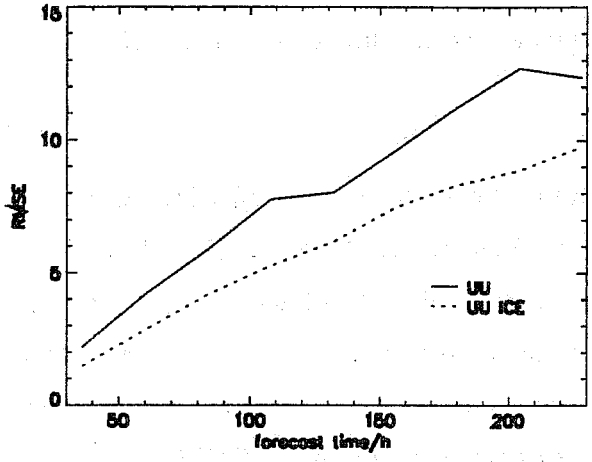


Fig. 1d. RMSE of ICE mean as percentage improvement over UU

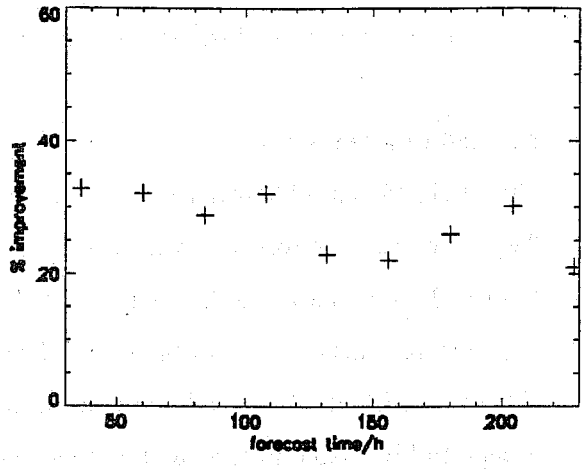


Fig. 1e. As in Figure 1a but for EE and ICE

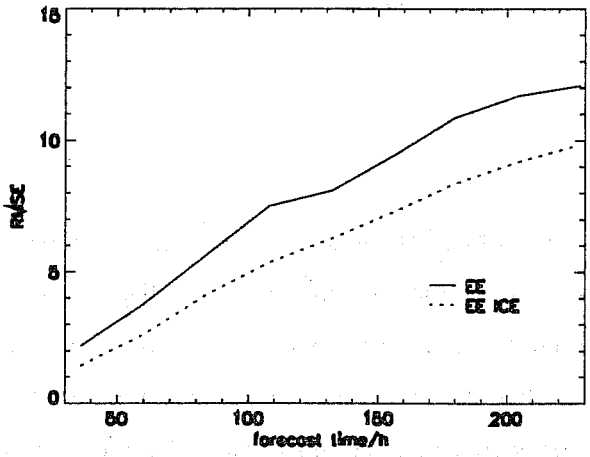


Fig. 1f. RMSE of ICE mean as percentage improvement over EE

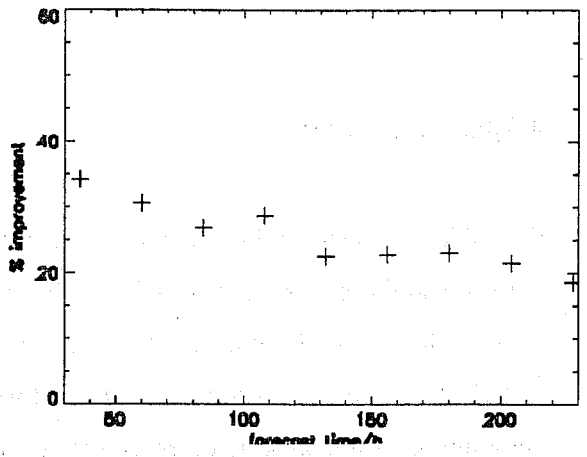


Fig. 1g. As in Figure 1a but for EEUU34 and ICE

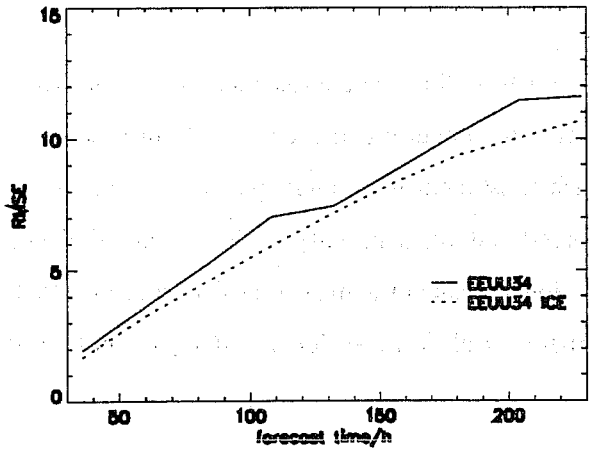
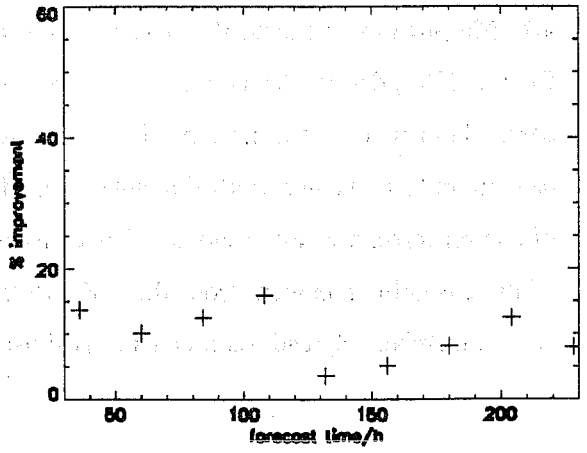


Fig. 1h. RMSE of ICE mean as percentage improvement over EEUU34



the corresponding ICE error, in fact, averaged over all forecast times RMS errors for the ICE systems are over 25% better than those of the individual systems. In contrast, the EEUU34 ensemble mean errors are considerably closer to those of the ICE, with an average difference of under 10%.

3.2 850 hPa temperature

The EEUU34 ensemble mean error is a clear improvement over the EE error after Day 2, (Fig. 2a); the joint ensemble errors are over 15% lower than the EE errors after T+48 (Fig. 2b). Inspection of Figure 2a indicates that the gain in lead time is of the order of 1-2 days after T+96. This improvement may be due to the clear advantage of the UU scores over the EE scores after Day 2 although the EEUU34 ensembles produce a further improvement over UU scores at all forecast times except T+136. Note that, as 850 hPa temperature fields are verified against ECMWF analyses, this may disadvantage the UU ensembles in the initial stage of the forecast. As with 500 hPa height, the errors of the joint ensemble are considerably closer to those of the corresponding ICE than either of the individual systems (Figs 2c-h).

4 Ensemble dispersion

Ensembles are designed in principle to estimate the sensitivity of predictions to errors in the forecast process, and this can be used to indicate the confidence of a prediction - the larger the dispersion the greater should be the uncertainty in the forecast process. It is also important that the spread is sufficient to cover all uncertainties in the forecast - so that the probability forecasts sample the full population. Three aspects of ensemble spread are considered here: (1) magnitude of the ensemble spread; (2) relationship between spread and skill; (3) coverage of observations. All results are averages over the 9 cases for which the full 33 member EE and UU ensembles are available.

4.1 Magnitude of ensemble spread over Europe

Buizza (1997) defines the spread of an ensemble of forecasts as the average distance of the perturbed ensemble members from the control - where the control is the unperturbed member. Joint ensembles have more than one unperturbed member - one from each constituent individual system - so the choice of control forecast is not obvious. Hence ensemble spread is defined initially as the average distance of the ensemble members from the EM, while in Section 4.3 spread is measured from the ECMWF control member. Spread has been calculated using anomaly correlation coefficient (ACC) for both 500

Figure 2a. Root Mean Squared Error of ensemble mean 850 hPa temperature forecasts over Europe from T+12 to T+204 (averaged over 9 cases (cases 2-10)). Results for four configurations are shown: UU, EE, EEUU34 and EEUU66.

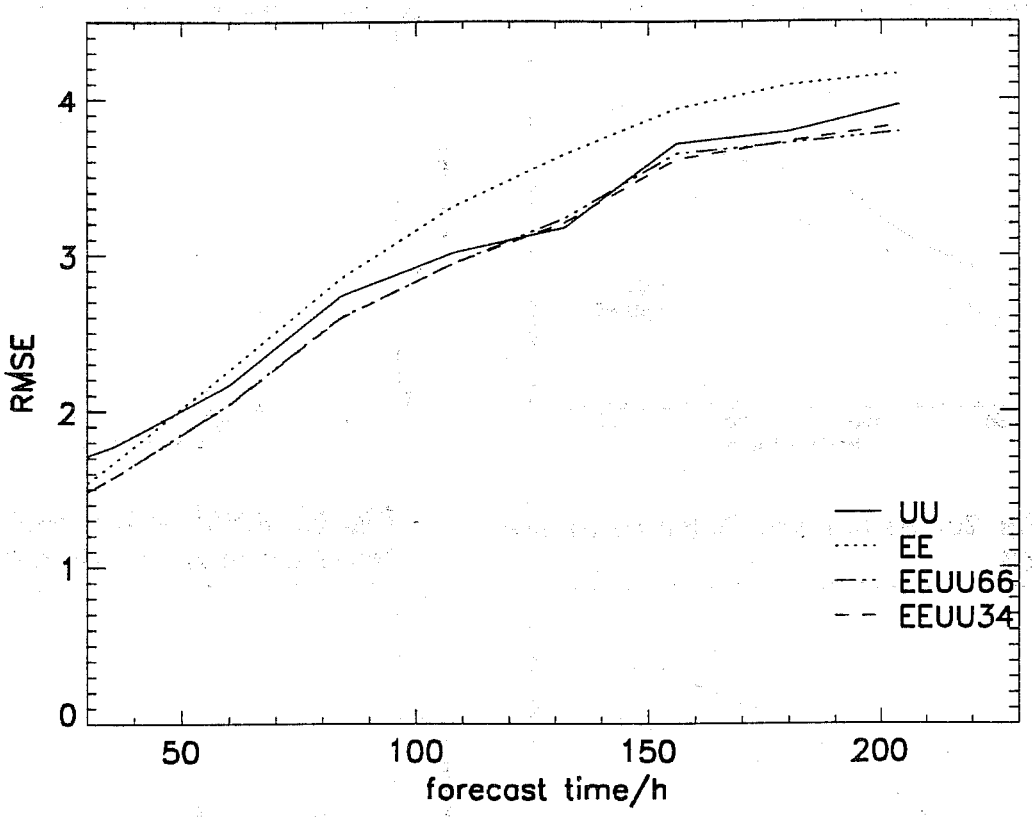


Figure 2b. As Figure 2a. but results are presented as percentage improvement over EE.

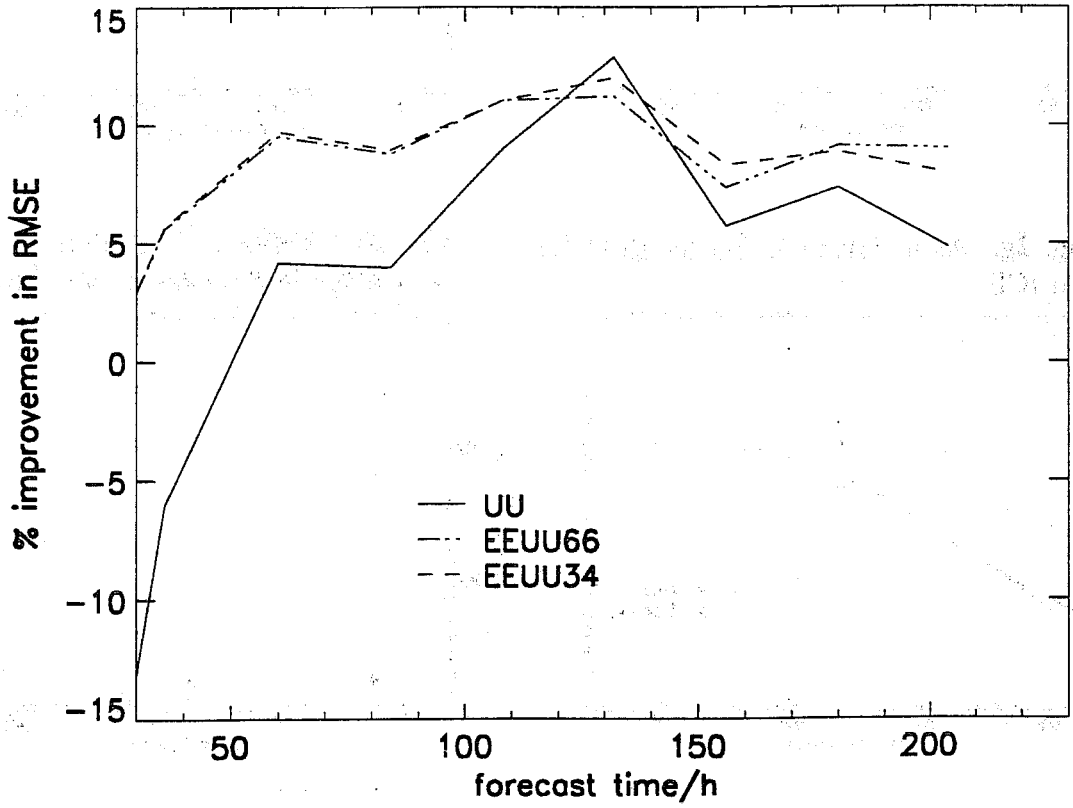


Fig. 2c. As in Figure 2a but for UU and ICE

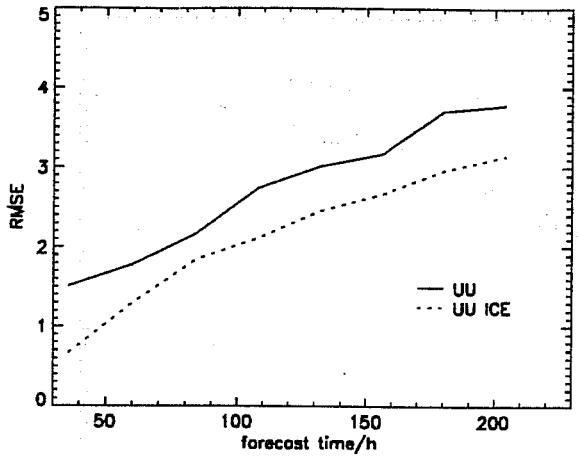


Fig. 2d. RMSE of ICE mean as percentage improvement over UU

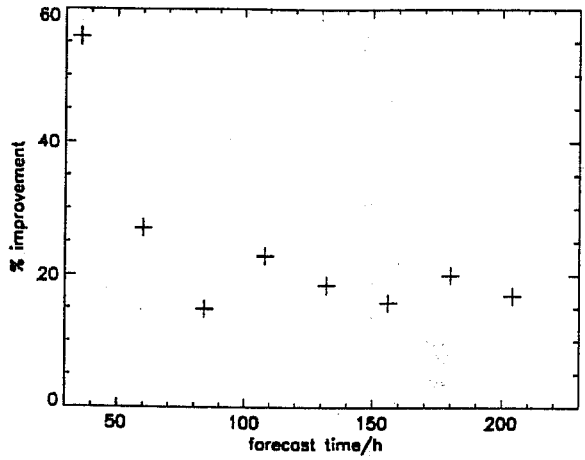


Fig. 2e. As in Figure 2a but for EE and ICE

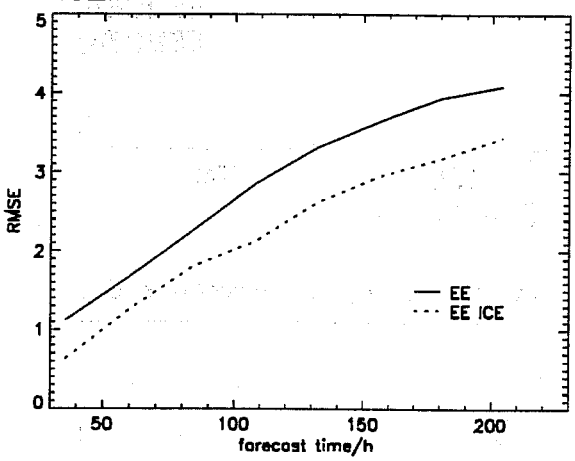


Fig. 2f. RMSE of ICE mean as percentage improvement over EE

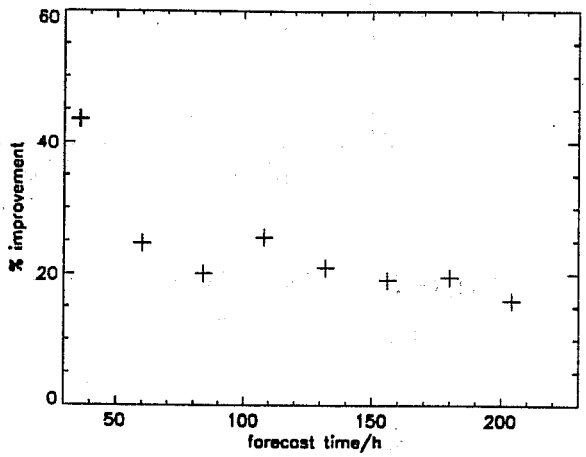


Fig. 2g. As in Figure 2a but for EEUU34 and ICE

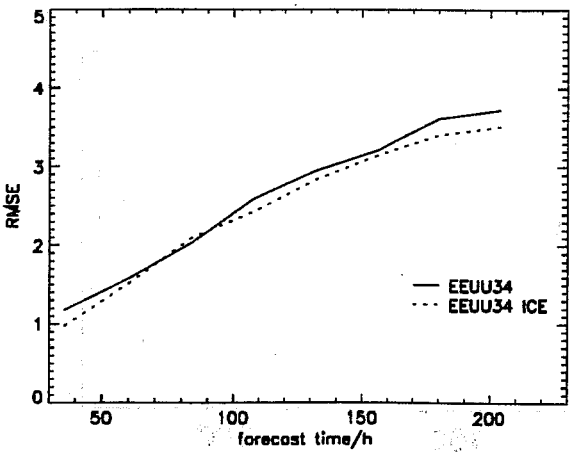
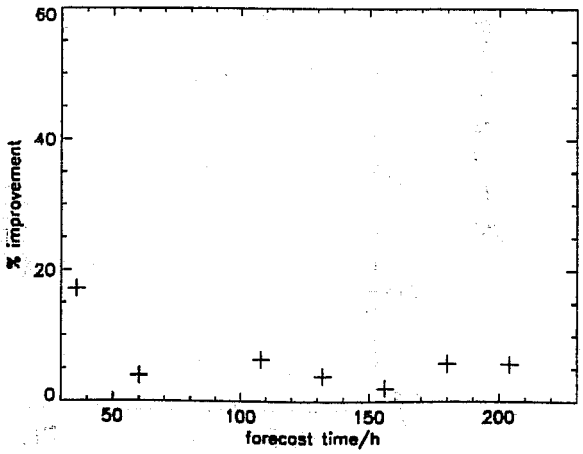


Fig. 2h. RMSE of ICE mean as percentage improvement over EEUU34



hPa height and 850 hPa temperature (see Buizza (1997) for more details of calculations). Spread has also been examined using RMS distance with results broadly similar to those below.

The 500 hPa height ACC spread of the EEUU66 ensemble is larger than that of both single-system ensembles throughout the forecast, and this increase in spread is only marginally dependent on the increased number of members (cf. spread of EEUU34 in Figure 3a). The increased spread of the MMMA ensemble, EEUU34, relative to that of the ECMWF system alone grows with forecast time, reaching over 5% by T+180 and over 10% by T+204 (Fig. 3b). In fact EE has the smallest spread of either the individual ensemble systems, but the spread in the UKMO ensembles is never more than 2% greater than that of the ECMWF system.

Similarly for 850 hPa temperature the MMMA ensembles have greater spread than either of the individual models; in fact the EEUU34 ensembles provide even larger increases in spread relative to the ECMWF system (Figs 4a and b). By T+204 the spread within the EEUU34 ensembles is over 20% larger than that in the EE system alone. For both 500 hPa height and 850 hPa temperature this increase in spread is achieved without loss of ensemble mean skill (as shown in previous section). One of the benefits of this increased spread is examined in Section 4.4.

4.2 Spread - Skill relationships

Previous studies of the EPS and UKMO ensembles (Buizza 1997; and Harrison *et al.* 1995) suggest that there is some correspondence between small ensemble spread and high skill of the deterministic forecast derived from the ensemble (control or mean), but that high spread ensembles can produce both skilful and unskilful deterministic forecasts. To some extent this is to be expected, as explained by Molteni *et al.* (1996) - when spread is large the deterministic forecast could be skilful by chance. In order to account for sampling problems of this kind results are again compared with those from a ICE. As used previously, this ICE is defined by taking one member, at random, from each ensemble to be the verifying analysis and averaging results over several iterations (Molteni *et al.* 1996). Skill is defined here as the ACC of the ensemble mean, and spread, as before, is the average ACC between the ensemble members and the ensemble mean. Again results are averaged over 9 forecasts and are calculated over Europe.

Figure 3a. Spread (average anomaly correlation coefficient of ensemble members with ensemble mean) for forecasts of 500 hPa height over Europe for T+12 to T+228, for UU, EE, EEUU66 and EEUU34 ensembles, averaged over 9 cases.

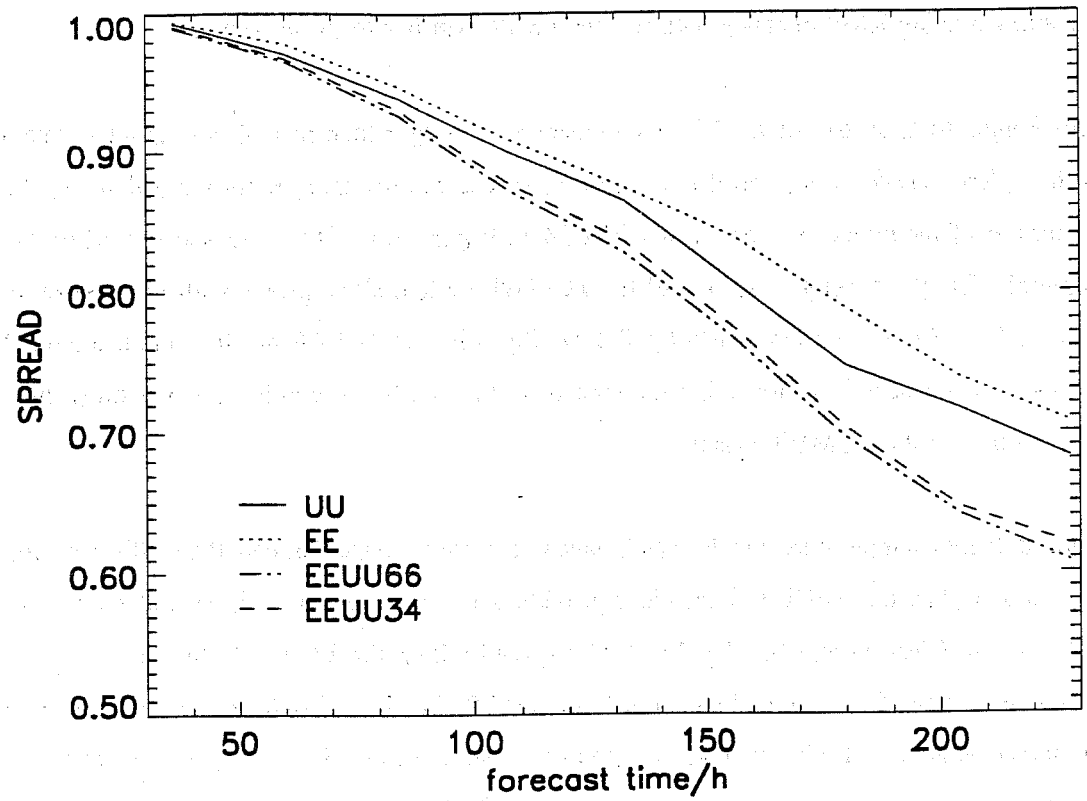


Figure 3b. As Figure 3a but results are presented as percentage improvement over EE.

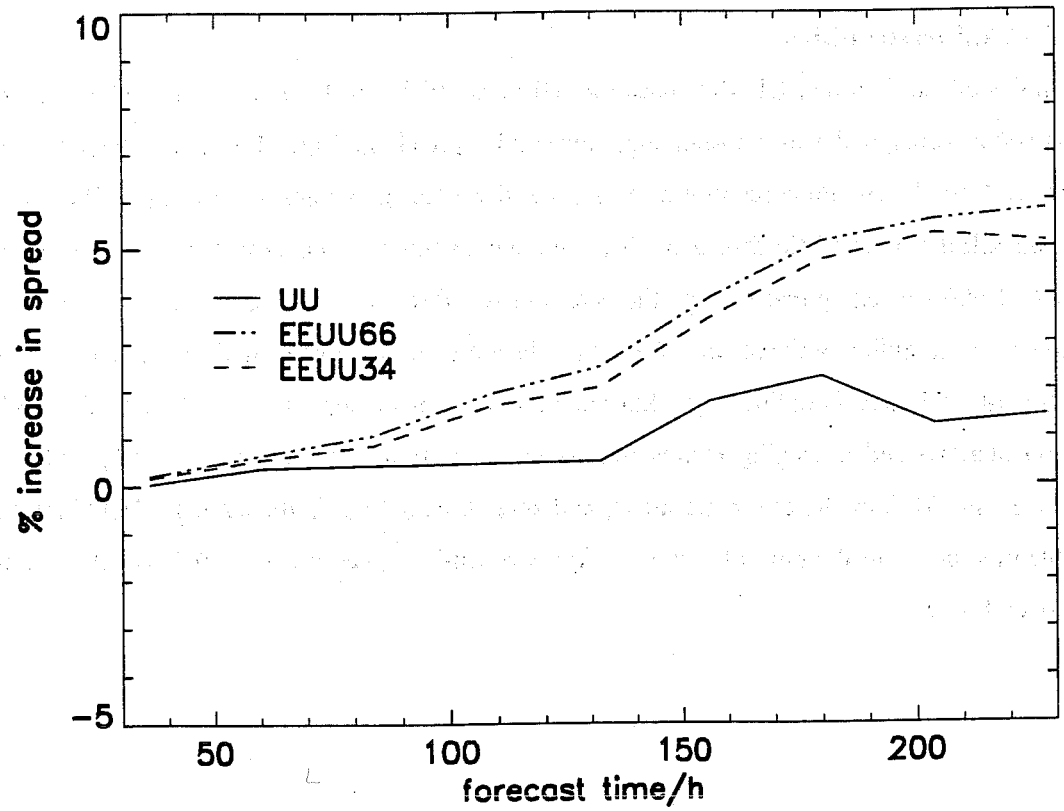


Figure 4a. Spread (average anomaly correlation coefficient of ensemble members with ensemble mean) for forecasts of 850 hPa temperature over Europe for T+12 to T+204, for UU, EE, EEUU66 and EEUU34 ensembles, averaged over 9 cases.

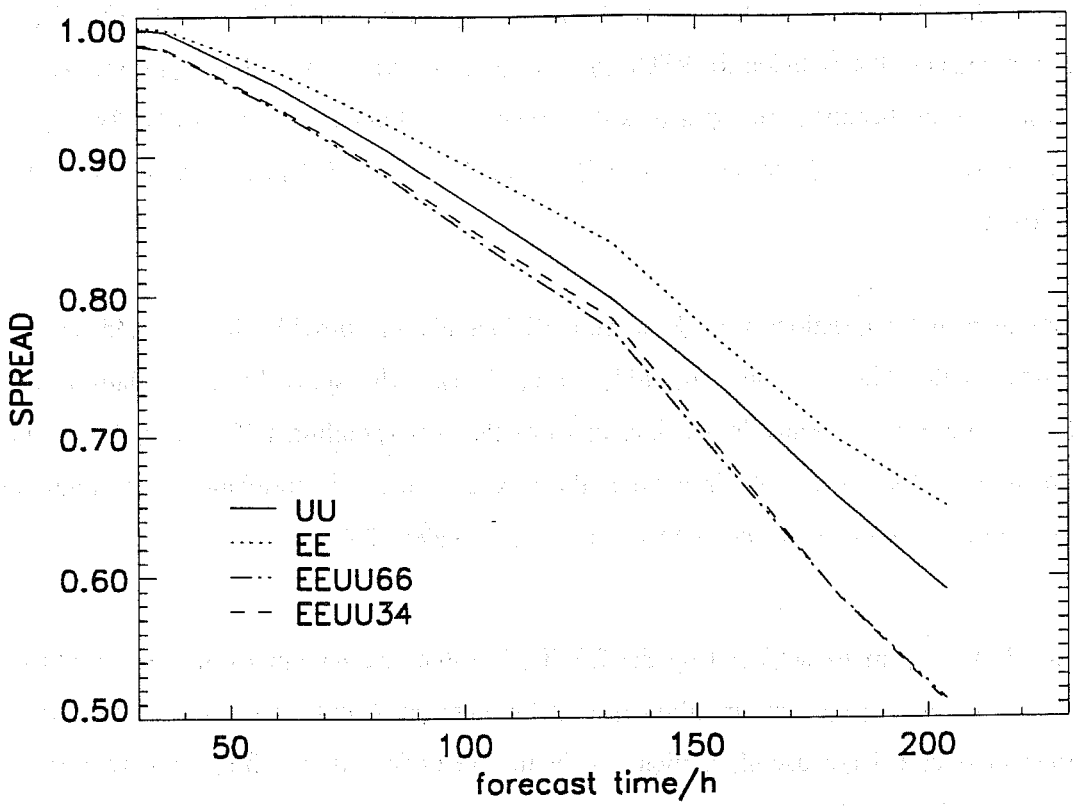
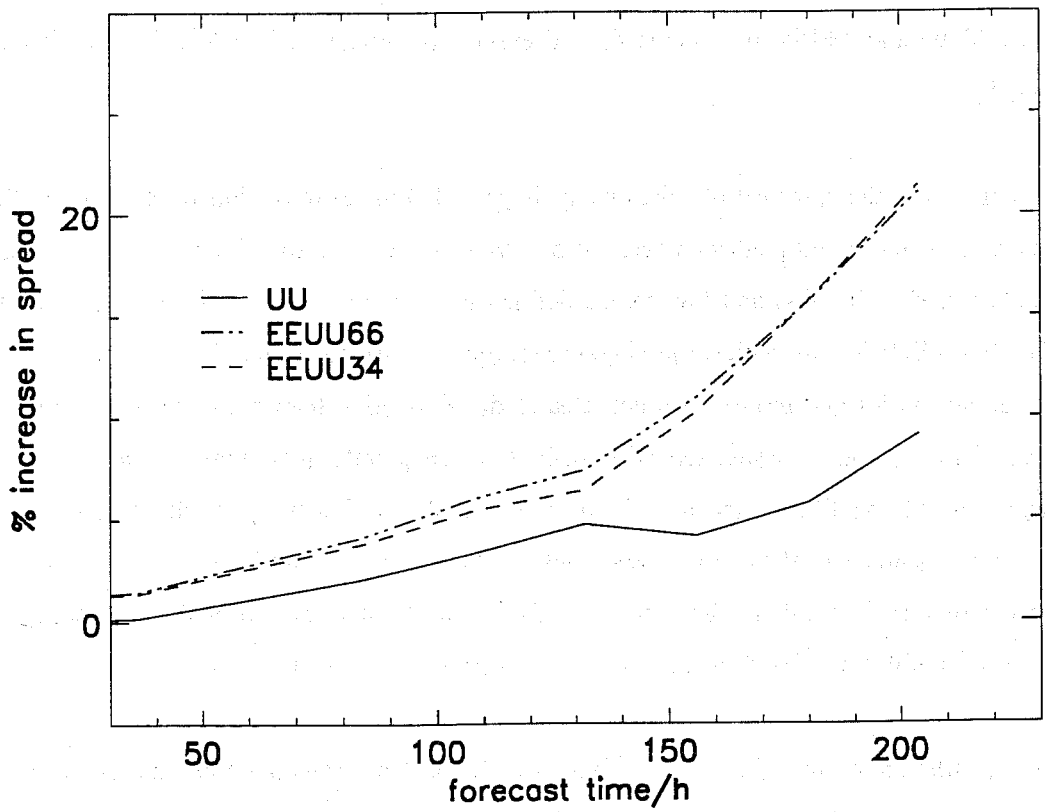


Figure 4b. As Figure 4a but results are presented as percentage improvement over EE.



4.2.1 500 hPa height

For 500 hPa height, the spread in the EEUU66 ensembles is a better predictor of forecast skill than the spread in either of the single-system ensembles at all ranges between T+108 and T+204 (Fig. 5a). In general the spread/skill correlation for EEUU66 is equal to, or greater than, that found for the single systems. Much of this benefit to the spread/skill correlation is obtained with the EEUU34 ensemble, with improvements of over 20% on the spread/skill correlation of the ECMWF system alone between T+108 to T+204.

Comparison of actual correlations with those from ICE simulations provides further evidence of the improvements achievable with joint ensembles (Figs 5b-e). The spread/skill correlations of the ECMWF system are considerably lower than those of the corresponding ICE - around 50% lower averaged between T+108 and T+204. In contrast the EEUU34 ensemble correlations are much closer to the ICE - on average the difference between the two is under 10%.

The spread/skill correlations achieved by the EEUU34 system are potentially useful. Wobus and Kalnay (1994) used agreement between different model systems as a skill indicator and suggested that correlations of over 0.4 are useful, particularly in the prediction of low frequency variability of forecast skill, while correlations of over 0.6 produced significant skill in forecasting day-to-day variability of forecast ACC scores. The EEUU34 ensemble produce spread/skill correlations of over 0.4 between T+96 and T+168; in contrast the EE ensemble correlation is below 0.3 for all forecast times (Fig. 5a).

Further insight into the spread/skill relationship is gained from examination of the distribution of spread and skill, with a skilful prediction taken as one with ensemble mean AC greater than the sample average and similarly with high and low spread defined in terms of average values across the sample (Table 2). For EEUU34 ensembles, the diagonal (stippled) entries are notably more populated than the off-diagonal entries and provide a much clearer discrimination than either of the single system ensembles. Low spread is associated with skill more frequently than with no skill in the UU ensembles at T+132 and T+156, and the EE ensembles at T+180, but the joint ensembles, EEUU34, provide a much clearer signal for low spread/high skill at all times. Similarly there are cases when high spread indicates low skill for both single systems, but the distribution of the high spread cases for the joint ensembles, EEUU34, again provides a much clearer discrimination.

The above results show that, for 500 hPa height at least, large spread within the joint ensemble provides a more reliable indicator of lack of skill and low predictability than large spread within an

Figure 5a. Correlation between skill (ACC of ensemble mean with analysis) and spread (average ACC of ensemble members with ensemble mean), for forecasts of 500 hPa height over Europe. Correlations are calculated over 9 cases for T+12-T+228 for UU, EE, EEUU66 and EEUU34 ensembles.

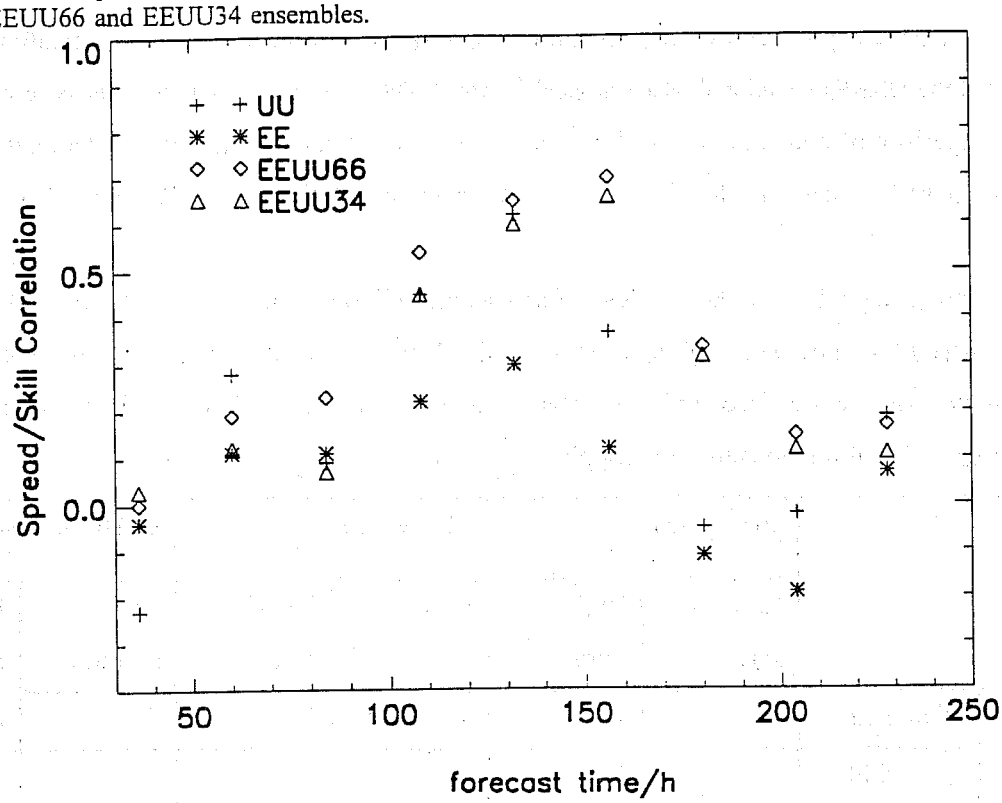


Fig. 5b. Spread/Skill correlations for EE and ICE for T+108 to T+204

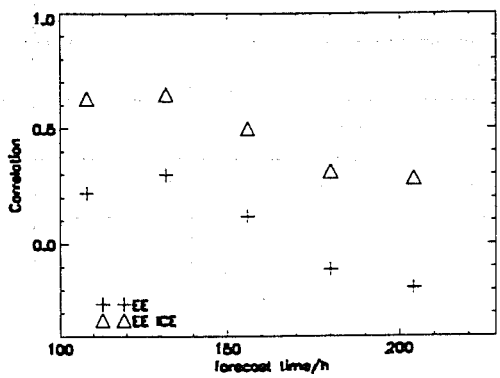


Fig. 5c. Spread/Skill correlations of ICE as percentage improvement over EE

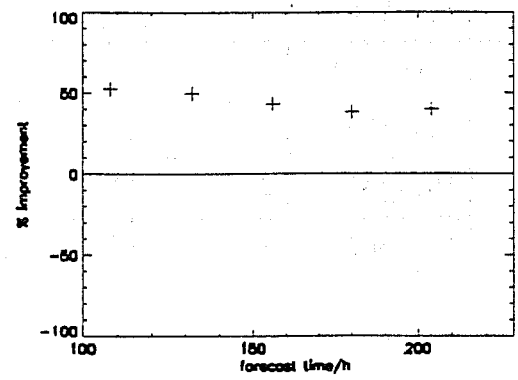


Fig. 5d. Spread/Skill correlations for EEUU34 and ICE for T+108 to T+204

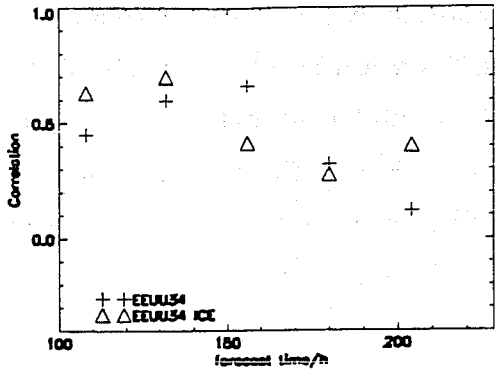
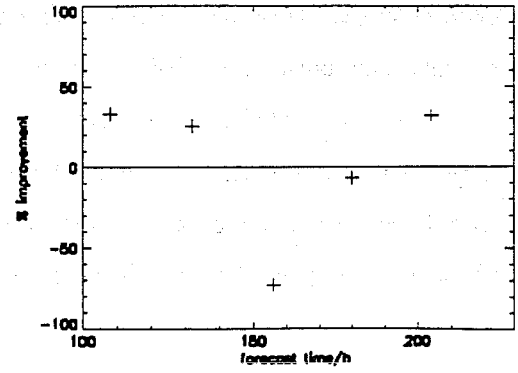


Fig. 5e. Spread/Skill correlations of ICE as percentage improvement over EEUU34



individual system. Similarly the correlation between low spread and high skill is strongest in the joint ensembles, particularly at later time-ranges. This suggests that comparisons between forecasts from different models can provide useful information on forecast confidence. In fact, medium-range forecasters at the UKMO's National Meterological Centre (NMC) already evaluate forecast confidence through comparison of operational model products from several centres, including: National Centers for Environmental Protection (NCEP), Deutscher Wetterdienst (DWD), ECMWF and UKMO.

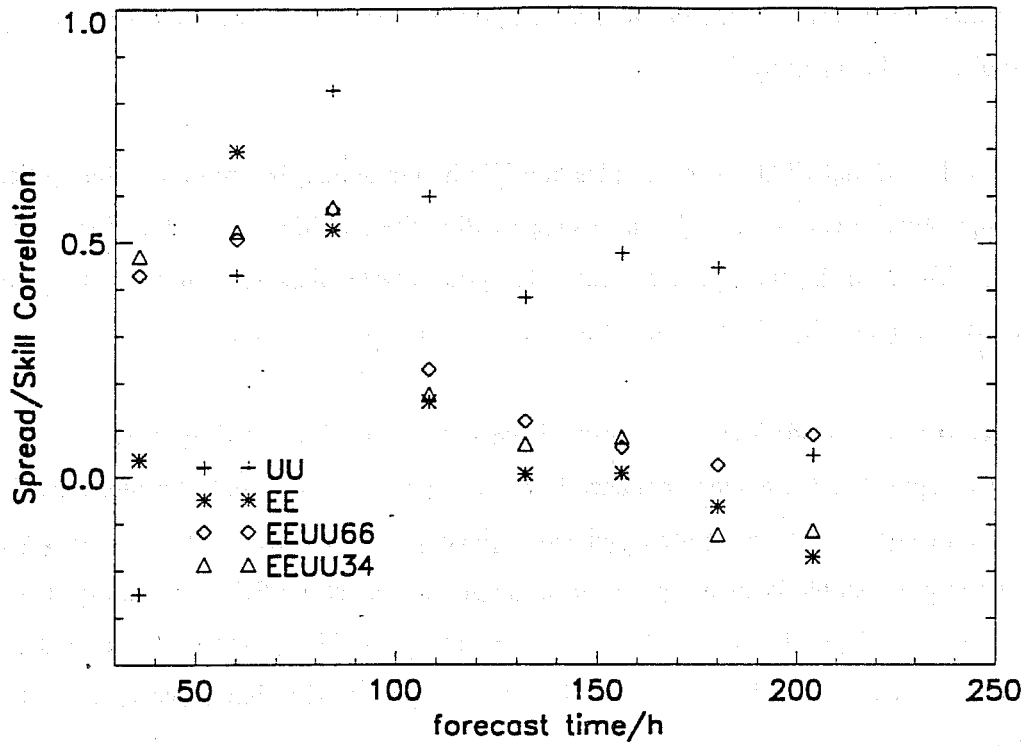
Table 2. Contingency tables at various days of spread/skill relationships for 500 hPa from the single and joint ensembles. Low and high spreads are defined in terms of averages across the 9 cases for each ensemble; skill is determined similarly. Diagonals in which observations should be concentrated for useful spread/skill relationships are stippled.

		UU System		EE System		EEUU34 system	
		Low spread	High Spread	Low spread	High spread	Low spread	High Spread
T+132	High Skill	4	1	3	2	5	0
	Low Skill	1	3	3	1	1	3
T+156	High Skill	4	2	3	2	5	0
	Low Skill	1	2	3	1	1	3
T+180	High Skill	2	2	4	2	5	1
	Low Skill	3	2	1	2	2	1
T+204	High Skill	2	2	2	2	4	1
	Low Skill	3	2	3	2	2	2

4.2.2 850 hPa Temperature

For 850 hPa temperature there is a considerable difference between the spread/skill correlations of the two individual systems (Fig. 6a). The correlations achieved by the UU ensembles are, on average, over 40% higher than those of the EE system between T+108 and T+204, compared with an average difference of 24% for 500hPa height. After T+108 the UKMO ensembles are the only system to achieve useful levels of correlation - that is above 0.4. This large difference in the performance of the two individual systems contributes to the more modest improvements in spread/skill gained with the joint ensemble for 850 hPa temperature compared with 500 hPa height.

Figure 6. Correlation between skill (ACC of ensemble mean with analysis) and spread (average ACC of ensemble members with ensemble mean), for forecasts of 850 hPa temperature over Europe. Correlations are calculated over 9 cases for T+12-T+228 for UU, EE, EEUU66 and EEUU34 ensembles.



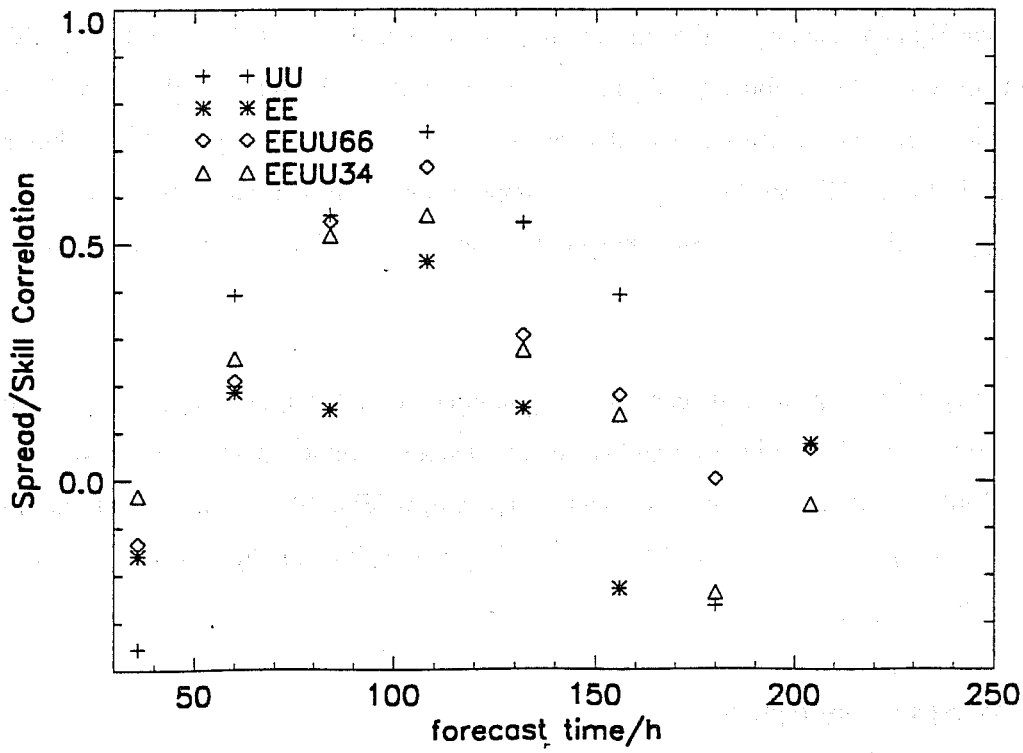


Figure 7. Correlation between 850 hPa temperature skill (ACC of ensemble mean with analysis) and spread of 500 hPa height (average ACC of ensemble members with ensemble mean). Correlations are calculated over 9 cases for T+12-T+204 for UU, EE, EEUU34 and EEUU66 ensembles over Europe.

The spread/skill correlation of the EEUU34 ensembles is generally equal to or slightly better than the EE system throughout the forecast, but after T+84 is substantially poorer than that achieved by the UU ensembles. Comparison with the ICE simulations confirms the superior performance of the UU ensembles for this measure (not shown).

For both the EE and EEUU34 ensembles 500 hPa height spread is a more reliable indicator of 850 hPa temperature ensemble mean skill than spread within the 850 hPa temperature forecasts themselves (Fig. 7). The MMMA ensembles provide substantial (greater than 10%) improvements over the EE ensemble but, generally, the UU ensemble achieves the highest correlations.

4.3 Spread and spread/skill as measured from the ECMWF control member.

Measuring spread as the average distance from the members to the control rather than the ensemble mean makes little difference to the conclusions given above. In fact the increase in spread achieved with the joint ensemble is even greater when measured against the ECMWF control than against the EM for both 500 hPa height and 850 hPa temperature (cf. Figs 8 and 9 with Figs 3 and 4). The increase in spread in the EEUU34 ensemble over EE grows with time reaching more than 10% by T+156.

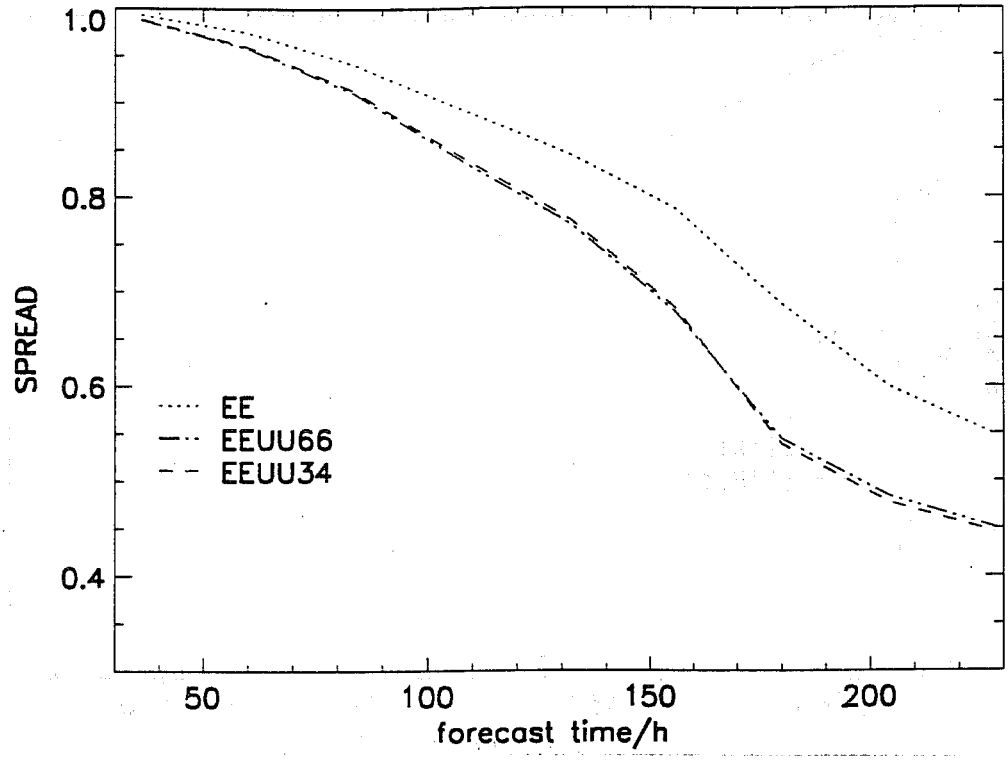
For spread/skill correlations, skill of the ensemble is measured as ACC between the verifying analysis and the ECMWF control forecast. As in the previous section, the spread/skill correlation of the joint ensemble, EEUU34, is greater than that of either individual system for 500 hPa height forecasts between T+108 and T+180 (Fig. 10a) with the improvement over EE during this time over 10%. The increased correlation of EEUU34 is potentially useful as it is above 0.4 between T+108 and T+228.

In contrast to the spread/skill correlations measured from the joint ensemble mean for 850 hPa temperature, (Fig. 7), the joint ensemble achieves useful increases in spread/skill correlation relative to the ECMWF system when measured against the control (Fig. 10b). After T+156 the improvement over the EE correlation is over 10%, and for the last 2 days of the forecasts the joint ensemble correlation is over 0.4.

4.4 Coverage of observations

The magnitude of spread in the MMMA ensembles is substantially larger than that of either individual ensemble. It is important to evaluate if this increase in spread is beneficial. One measure of benefit

Figure 8a. Spread from ECMWF control member (average anomaly correlation coefficient of ensemble members with EE control member) for forecasts of 500 hPa height over Europe for T+12 to T+228 (averaged over 9 cases). Results for three configurations are shown: EE, EEUU66 and EEUU34. (UU is not included as the ECMWF control member is not part of the UU ensemble.)



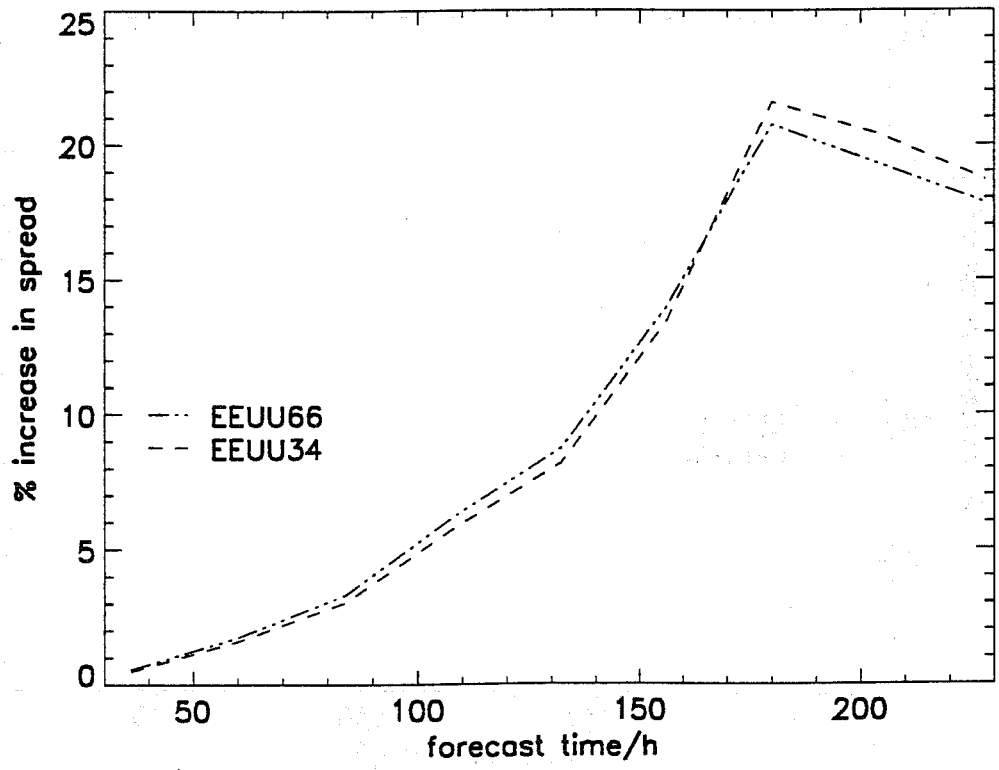
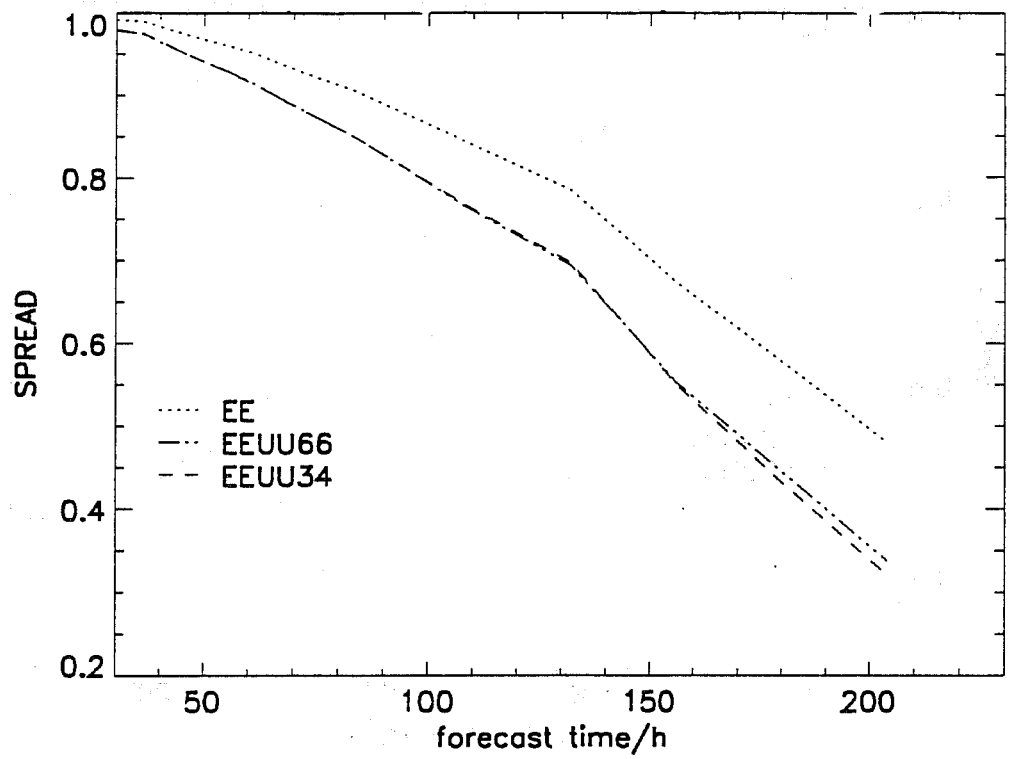


Figure 8b. As Figure 8a but results are presented as percentage improvement over EE.

Figure 9a. Spread from ECMWF control member (average anomaly correlation coefficient of ensemble members with EE control member) for forecasts of 850 hPa temperature over Europe for T+12 to T+204 (averaged over 9 cases). Results for three configurations are shown: EE, EEUU66 and EEUU34. (UU is not included as the ECMWF control member is not part of the UU ensemble.)



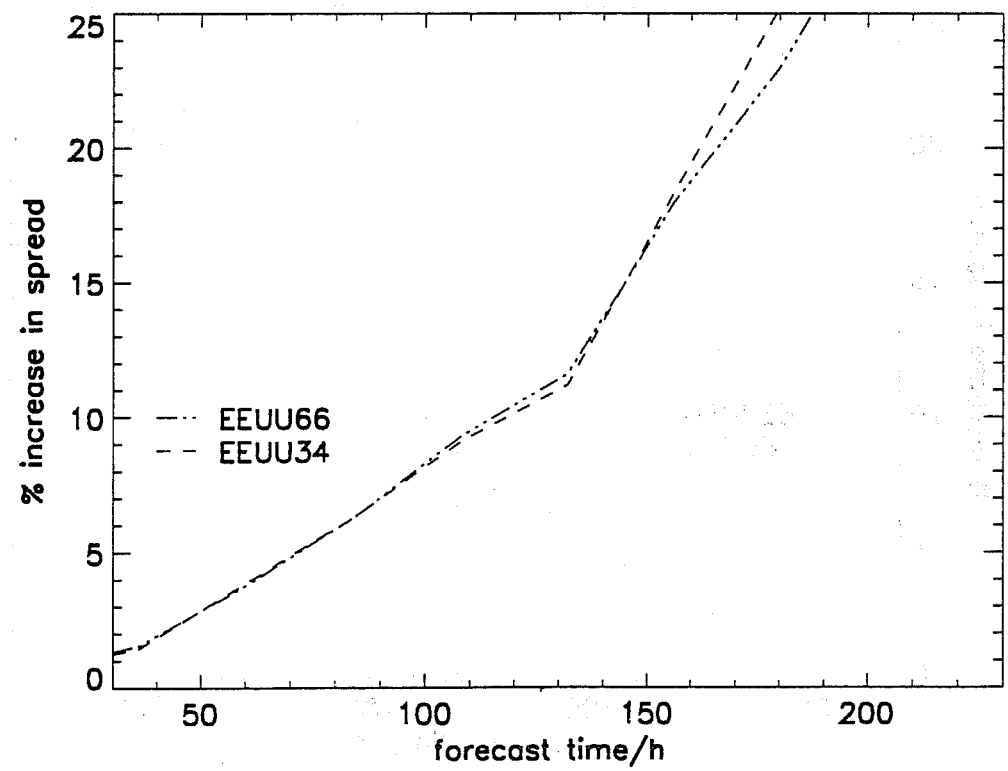
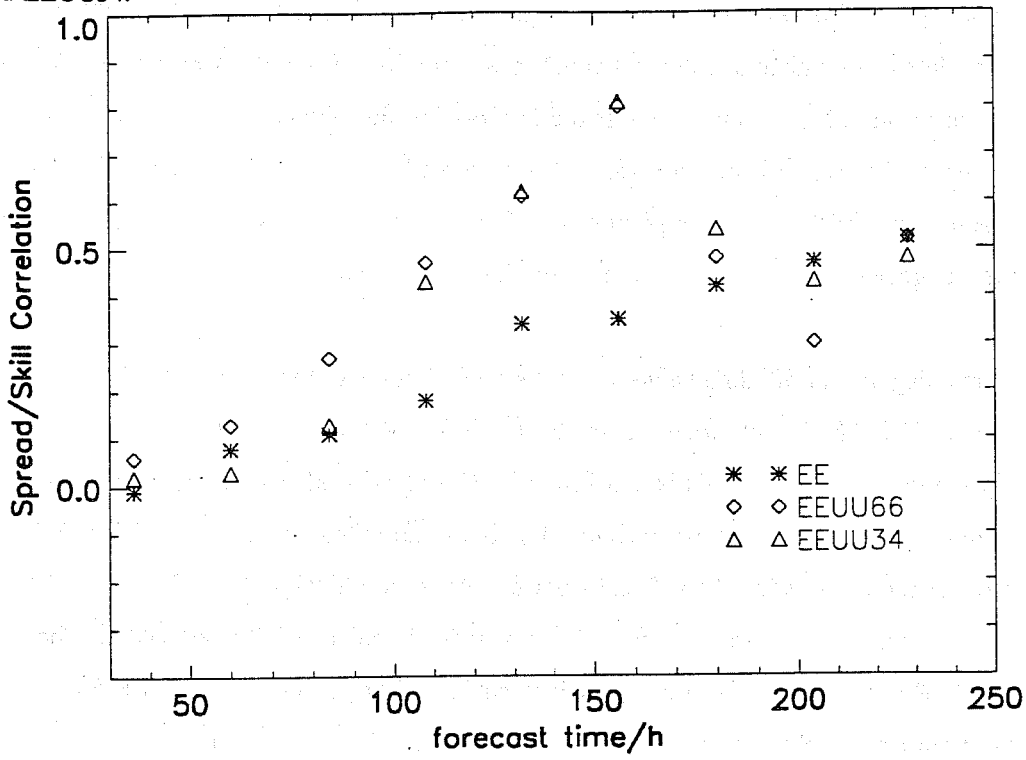


Figure 9b. As Figure 9a but results are presented as percentage improvement over EE.

Figure 10a. Correlation between skill (ACC of the ECMWF control member with analysis) and spread around the ECMWF control member (ACC of ensemble members with the EE control member). For forecasts of 500 hPa height over Europe for T+12 to T+228 (averaged over 9 cases). Results for three configurations are shown: EE, EEUU66 and EEUU34.



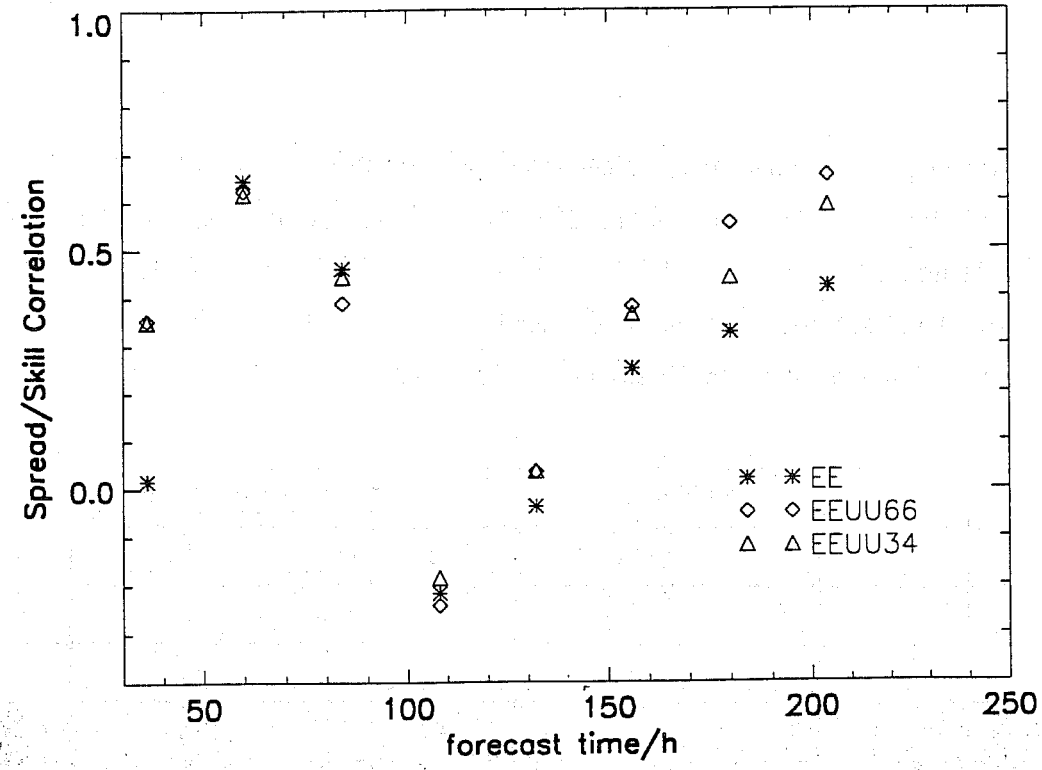


Figure 10b. As Figure 10a but for 850 hPa temperature for T+12 to T+204.

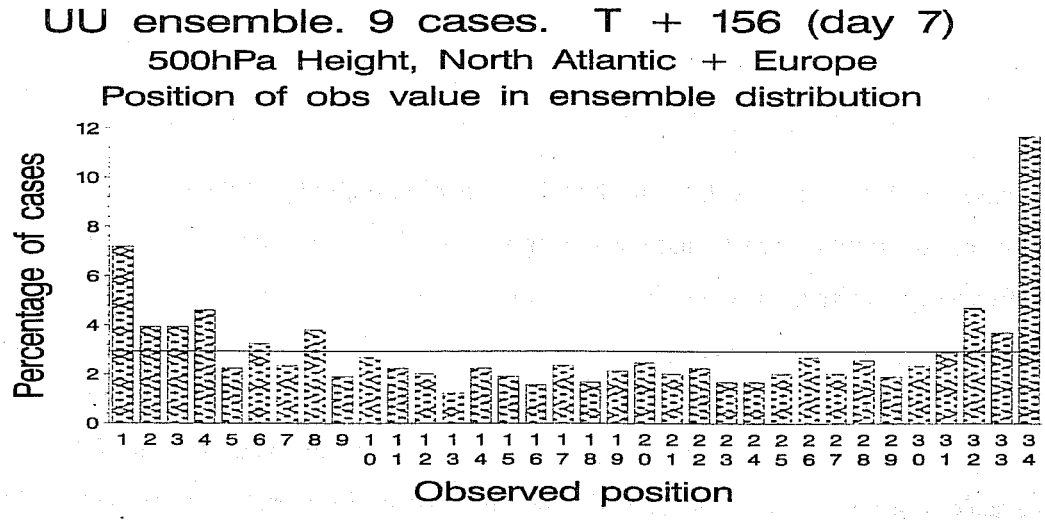
relates to the ideal that ensemble spread should be sufficient to cover all uncertainties in the forecast, with observed values falling uniformly into the intervals created by the ensemble. Talagrand diagrams provide a clear graphical representation of the distribution of observations relative to the ensemble members (Lazinger and Strauss 1995). At each grid point the forecast values from all members can be ordered to define a number of intervals equal to the number of members plus one. The observed value must lie in one of these intervals - if the observation lies outside the range of the ensemble it will lie in one of the two extreme intervals. For a correctly-formulated ensemble the spread of the members should be such that over a large number of cases the probability of the analysis being inside each of the categories (including the two extreme intervals) is equal.

The Talagrand diagram for 500 hPa height forecasts over Europe and the North Atlantic (65°W-35°E, 70°-25°N) at T+156 (Fig. 11) illustrates the relatively flatter, and therefore closer to ideal, distribution achieved by the joint ensemble. This is confirmed by Chi-squared statistical tests; the hypothesis that the distribution is uniform can be rejected for both individual systems at the 0.005 level of significance, but can be accepted at the 0.025 level for the joint EEUU34 ensembles. In particular the proportion of observations lying outside the ensemble is substantially reduced in the EEUU34 ensembles (Table 3). For both 500 hPa height and 850 hPa temperature, at T+156, the smallest proportion of outliers are provided by the MMMA ensembles, with reductions of over 40% compared with the EE ensembles.

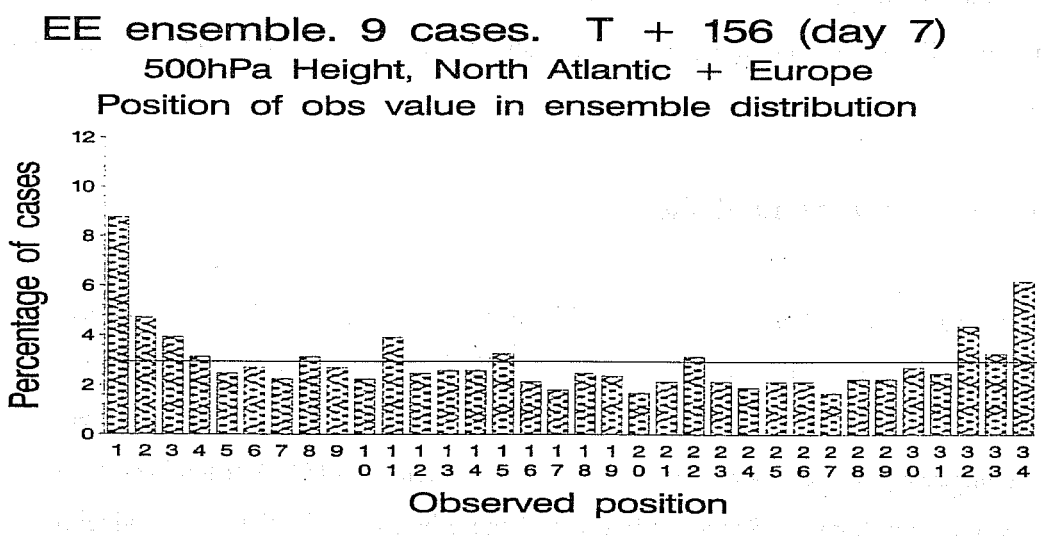
Table 3. Percentage of verifying analyses lying outside the ensemble distribution for various ensemble configurations, averaged over 9 cases, for both 500 hPa height and 850 hPa temperature forecasts over the North Atlantic/European area at T+156. Columns show percentage of outliers (N), percentage expected by chance (E), percentage above expectation (N-E), and the percentage above expectation expressed as an improvement with EE ensemble result as reference value, (% imp). Best values in bold.

		500 hPa height			850 hPa temperature		
	E	N	N-E	% imp	N	N-E	% imp
UU	5.9	18.9	13.0	-44.0	19.6	13.7	-11.3
EE	5.9	14.9	9.0	-	18.2	12.3	-
EEUU34	5.7	**8.9**	**3.2**	**+64.4**	12.8	**7.1**	**+42.3**
EEUU66	3.0	5.9	2.9	+67.8	**10.0**	7.0	+43.1

a)



b)



c)

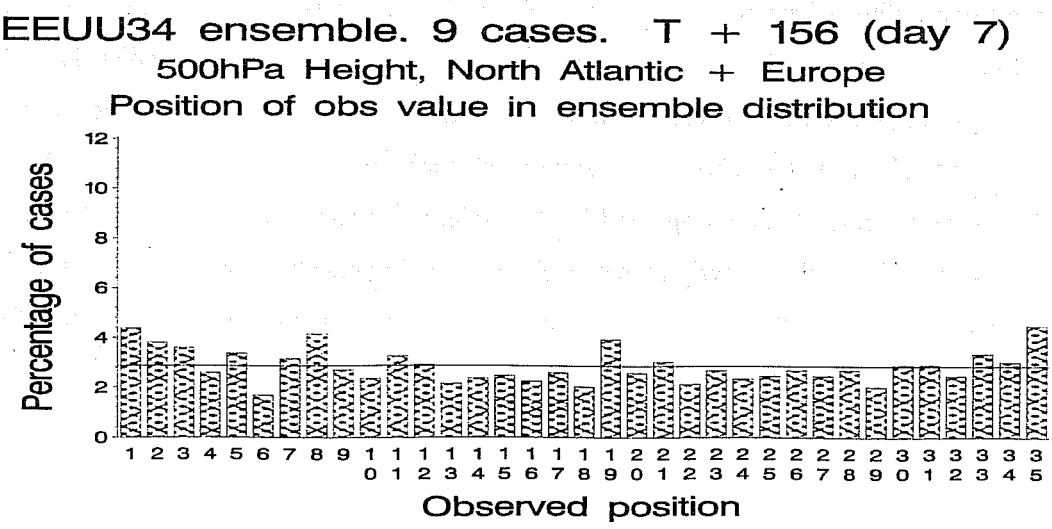


Figure 11. Talagrand diagram showing the frequency with which observations of 500 hPa height over the North Atlantic/European region lie within ensemble intervals at T+156. Horizontal lines indicate expected values. Results are for 9 cases.
a) UU ensemble, b) EE ensemble and c) for EEUU34 ensemble.

The joint ensemble brings improvement to the whole distribution, not just the outliers. Performance over the whole distribution can be assessed using the mean square frequency differences from the expected frequency (horizontal line on Figure 11). Scores have been calculated for T+108, T+156 and T+204, and are expressed as percentage improvement against the EE score (Fig. 12). For both 500 hPa height and 850 hPa temperature the EEUU34 ensembles fit the expected distribution more closely than either single systems; the improvement over EE exceeds 50% and is up to 75%. The UU ensemble achieves a flatter distribution for 850 hPa temperature forecasts compared with EE system, while for 500 hPa heights neither individual system consistently out performs the other. Talagrand diagrams for 850 hPa temperature forecasts (not shown) indicate that both models have a cold bias.

5 Probabilistic measures of skill

It was demonstrated in Section 4 that improved spread and coverage of observations was achieved by the MMMA ensembles. It should, however, be remembered that large spread and flat Talagrand distributions are not necessarily sufficient; since both could be achieved, without improvement in skill, by simply increasing spread towards climatology. The benefit of MMMA ensembles for deterministic forecasts was demonstrated in Section 4 but one of the primary motivations for producing ensembles is to forecast probabilities of meterological events. Probability forecasts can be easily generated from ensemble forecasts by simply taking the percentage of ensemble members which predict that an event will occur. In this Section the benefits of MMMA ensembles to probability forecasts for 500 hPa height and 850 hPa temperature above or below one standard deviation of normal at T+108, T+156 and T+214, across 9 cases, over Europe and North Atlantic is examined. A number of verification methods are used - for further descriptions of techniques and applications see, Murphy and Winkler (1992) and Stanski *et al.* (1989). For ease of comparison, the probabilistic verification scores for UU, EEUU34 and EEUU66 are presented as percentage improvement on the EE score, that is as skill scores with EE as standard (see Eqn 1).

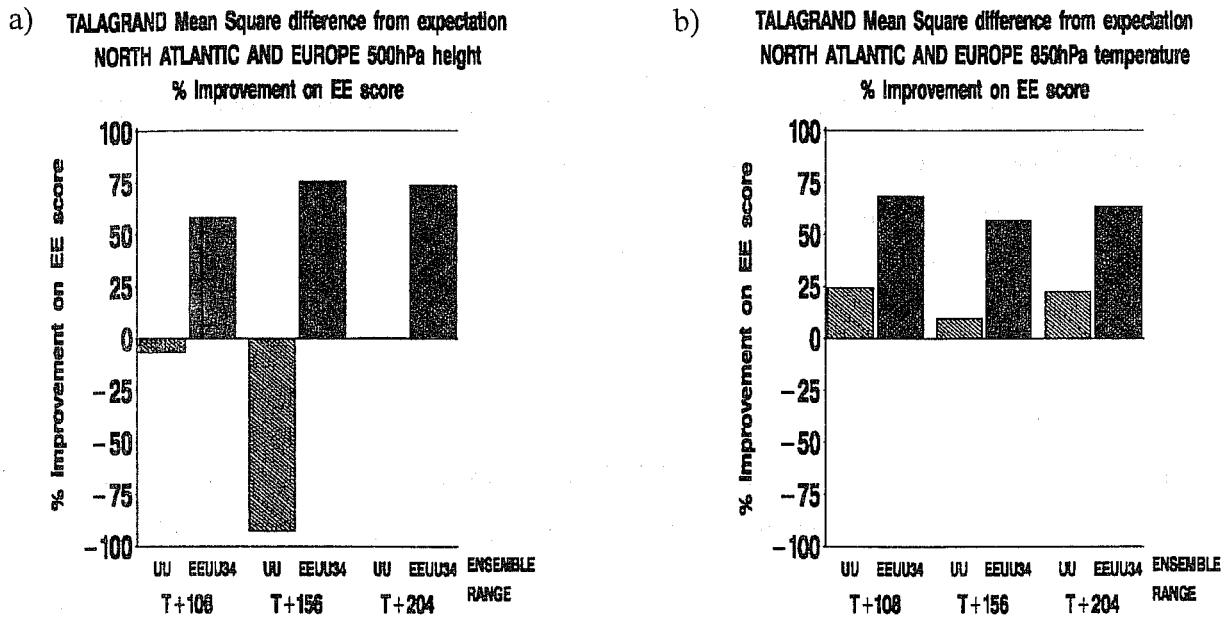


Figure 12. Mean square difference from expectation of ensemble distribution (Talagrand diagrams) for T+108, T+156 and T+204, over North Atlantic/European region (for 9 cases). Results are presented as percentage improvement over EE ensemble - UU (hatching), EEUU34 (cross hatching).
a) 500 hPa height, b) 850 hPa temperature.

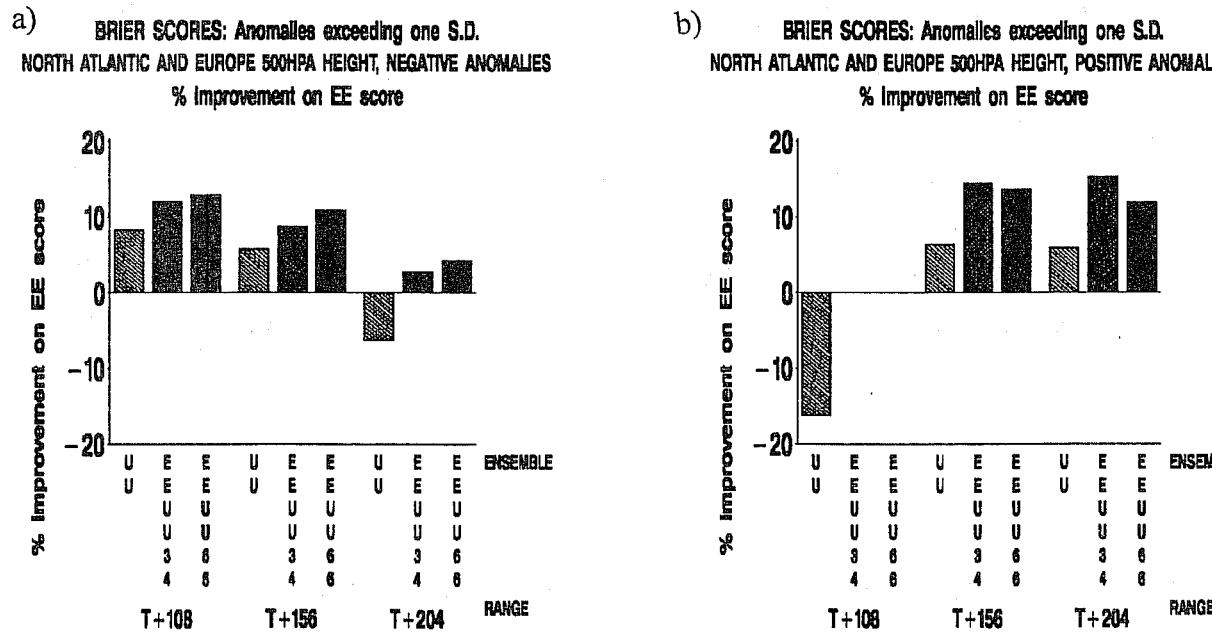


Figure 13. Brier Scores for T+108, T+156 and T+204 predictions of 500 hPa height anomalies over the North Atlantic/European region exceeding one standard deviation (for 9 cases) - a) negative anomalies b) positive anomalies. Results are presented as percentage improvement over EE ensemble - UU (hatching), EEUU34 (cross hatching), and EEUU66 (solid).

5.1 Brier Scores

A simple measure of probability forecast accuracy is the Brier Score (Brier 1950), which can be thought of as the mean square error of the probabilistic forecast. For a set of n forecasts of a binary event the Score is written as:

$$BS=\frac{1}{n}\sum_{k=0}^{n}(f_k-o_k)^2 \qquad (2)$$

where f_k is the forecast probability and o_k is 1 if the event occurred and 0 if not. The Brier Score is negatively orientated, i.e. zero represents a perfect score while one is the worst possible score.

5.1.2 500 hPa height

In all cases Scores for the joint ensembles EEUU34 and EEUU66 are equal to or better than the Score for the most skilful individual system (Fig. 13). The performance of the EEUU34 ensemble is similar to that of the EEUU66 ensemble, suggesting that the benefits are due to the combination of model systems, rather than from increasing the size of the ensemble. On average the EEUU34 ensembles achieve around a 10% improvement (maximum 20%) in Brier Score over the EE ensembles for these cases. Inspection of the actual Brier Scores for each system (not shown) suggests that this improvement equates to a gain in predictability of the order of 1 day.

Murphy (1973) decomposed the Brier Score into the sum of 3 components: reliability, resolution and uncertainty. Suppose that forecast probabilities are allowed to take on one of a range of values, f_i, where i=1, I - for example, the issued forecast probability f_i may be 0%, 10%, 20% etc. Then for each forecast category the average observation (conditional on the forecast category) is defined as:

$$\bar{o}_i=\frac{1}{n}\sum_{k=1}^{n}o_k \qquad (3)$$

Then, after some algebra, the two category form of the Brier score in Eqn 2 can be rewritten as:

$$BS=\frac{1}{n}\sum_{i=1}^{I}N_i(f_i-\bar{o}_i)^2-\frac{1}{n}\sum_{i=1}^{I}N_i(\bar{o}_i-\bar{o})^2+\bar{o}(1-\bar{o}) \qquad (4)$$

where $\bar{o}$ is the average frequency of occurrence over the whole sample.

The first, second and third terms on the right-hand side are referred to respectively as the reliability, resolution and uncertainty terms. Reliability indicates the correspondence between forecast probability and the actual observed frequency of occurrence of the event, while resolution is the ability of the forecast to resolve the set of sample events into subsets with different frequency distributions. The uncertainty is the variance of observations in the sample and so is independent of the forecasts. The reliability and resolution scores are included here to determine the source of improvement to the Brier Score provided by the MMMA ensembles.

The joint model EEUU66 is more reliable than either individual model in all cases and its improvement over the EE ensembles is at least 40% (Fig. 14). Generally, halving the number of members in the full MMMA ensemble of 66 members does reduce the reliability slightly, but EEUU34 still provides a substantial improvement over EE, with an average improvement (over positive and negative anomalies) of more than 55% over the 3 times and 9 cases studied. Of the individual systems UU is on average more reliable than the EE ensembles, with average improvement against EE of 26% for positive anomalies and 41% for negative anomalies, a non-symmetry perhaps associated with differences in model biases.

On average, the EEUU34 ensembles resolution is 6% better than that of EE (Fig. 15). For this measure, unlike Brier and reliability, the performance of the UU ensemble is worse than the EE ensemble such that a reduction in resolution results from MMMA in some cases. However in general these reduction are minimal and are outweighed by the improved resolutions achieved.

The above results show that improvements in Brier Scores for 500 hPa height anomalies achieved by the EEUU34 ensembles result from improvements in both reliability and resolution and not because of improved spread producing drift towards climatological distributions. Note that it is possible to improve reliability by simply improving the climatology of the forecast system. However, this is not the case for the EEUU34 forecasts; since a forecast of observed climatology has low resolution, while improved reliability in the EEUU34 ensemble is not at the expense of resolution (in fact resolution is improved by over 5% against the single EE system). The benefits to probability forecasts gained by combining model systems may be due to the sampling of different, skilful populations provided by the individual systems - the extent of independent information contained in the two systems is examined in Section 6.

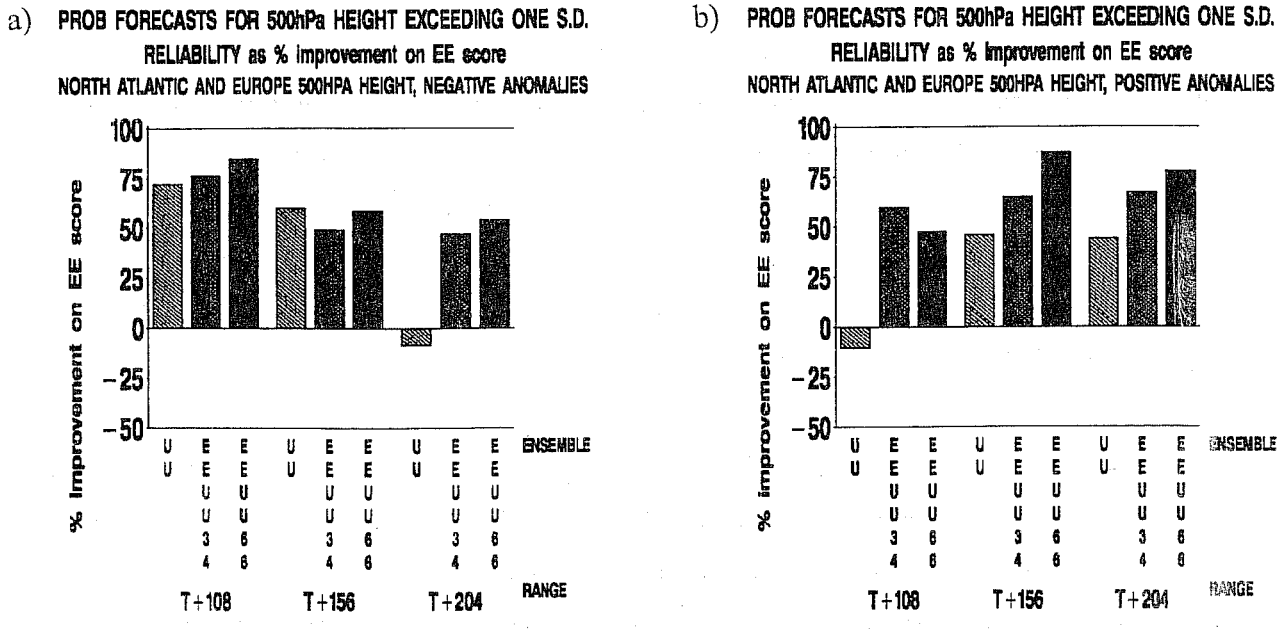


Figure 14. as Figure 13, but for reliability (note change in vertical scale).

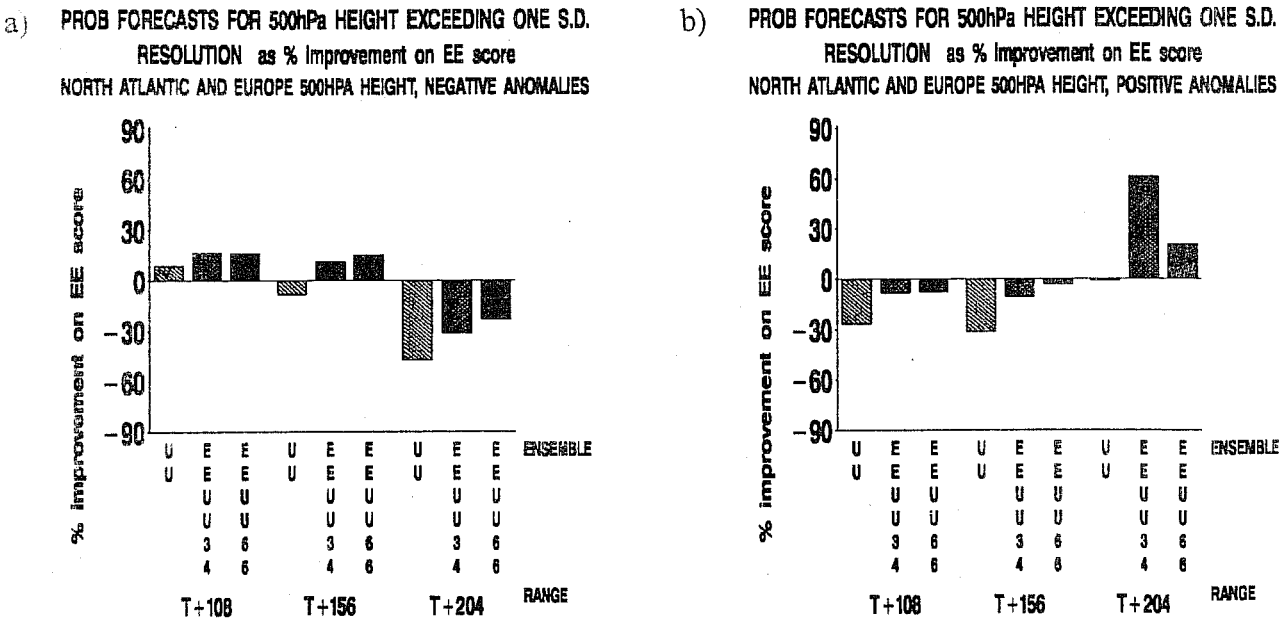


Figure 15. as Figure 13, but for resolution (note change in vertical scale).

5.1.2 850 hPa temperature

As with 500 hPa height the MMMA ensembles produce large improvements in Brier Scores over the ECMWF system for 850 hPa temperatures (Fig. 16), with these improvements equating to a gain in lead time of around 1 day at Day 7 (not shown). For positive anomalies the EEUU34 ensembles achieve a 5% average improvement, while for negative anomalies the improvement is over 14%. This non-symmetry is also present in UU ensemble skill scores; for positive anomalies the UKMO and ECMWF systems have similar Brier scores, but for negative anomalies the UU scores are, on average, over 10% better than those from EE.

The EEUU34 system also gives unsymmetrical improvements in reliability (Fig. 17). For negative anomalies the EEUU34 ensemble produces improvements over EE of at least 50% for all three times examined. For positive anomalies the improvement is 5% at T+108, rising to over 50% at T+204. Again this non-symmetry of may be associated with different biases in the two individual systems. As with 500 hPa height this improvement in reliability is not achieved at the cost of resolution; as for both positive and negative anomalies the EEUU34 ensemble produces an average improvement in resolution of more than 15% over the EE ensemble (Fig. 18).

5.2 Relative Operating Characteristics (ROC)

The Relative Operating Characteristic curve, taken from signal detection theory, indicates the performance of a system in predicting a particular event in terms of hit and false alarm rates (stratified by observations) (Stanski *et al.* 1989). Each point on the curve is located by plotting the false alarm rate against the hit rate for probabilities at or greater than a specific value, Figure 19 provides a sample curve. Ideally the percentage of hits will always exceed the percentage of false alarms and the curve will lie in the upper left hand corner of the diagram; in fact a perfect forecast will have no false alarms and a hit rate of 1 for all thresholds and so is represented by a curve that stretches from (0,0) to (0,1) to (1,1). The standardised area enclosed beneath the curve is a simple quantitative measure associated with the ROC, with a range of 0 to 1, where 1 is a perfect score. In contrast a system with no skill will achieve hits at the same rate as false alarms and so its curve will lie along the 45° line and enclose a standardised area of 0.5. In the following, area under the curve is used to create a skill score with the EE score as standard.

This measure is similar to resolution, as it assess how well the system can discriminate between occurrences and non-occurrences of an event. As ROC is based on stratification by observation it

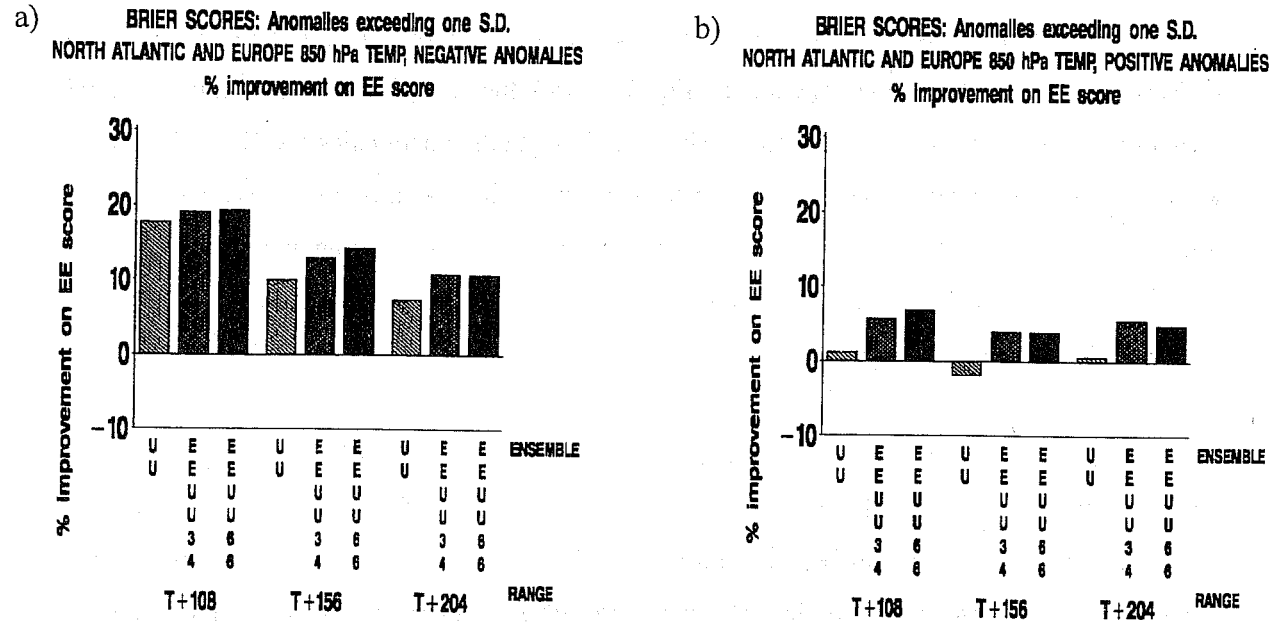


Figure 16. Brier Scores for T+108, T+156 and T+204 predictions of 850 hPa temperature anomalies over the North Atlantic/European region exceeding one standard deviation (for 9 cases) - a) negative anomalies b) positive anomalies. Results are presented as percentage improvement over EE ensemble - UU (hatching), EEUU34 (cross hatching), and EEUU66 (solid).

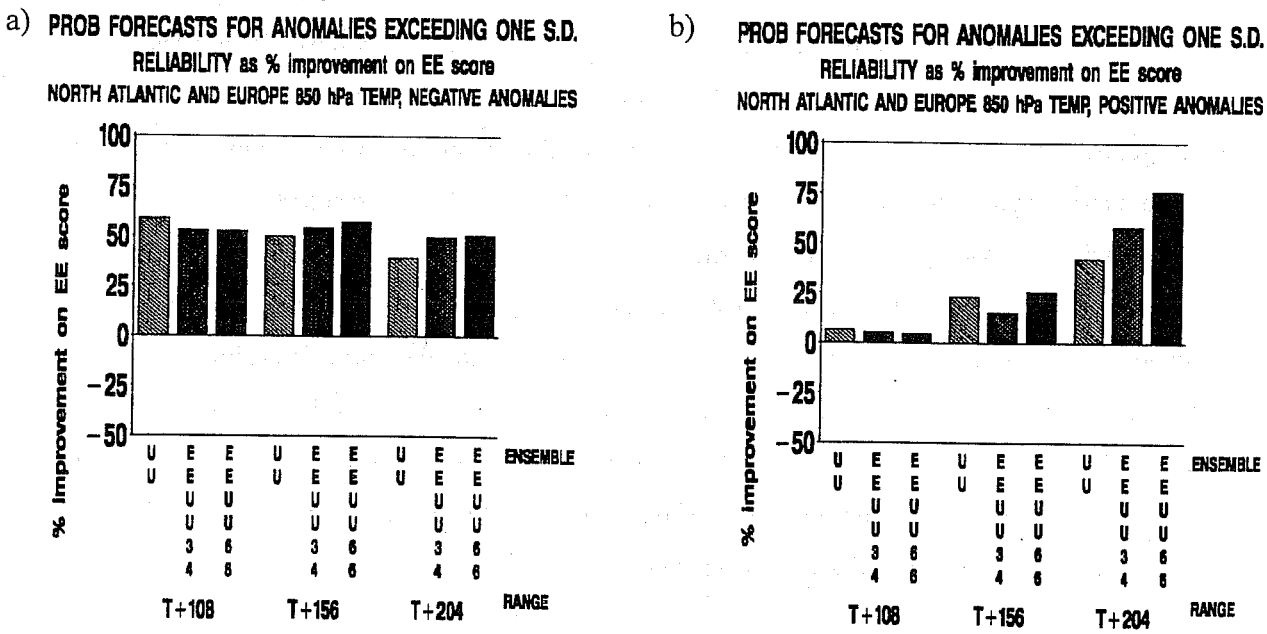


Figure 17. as Figure 16, but for reliability (note change in vertical scale).

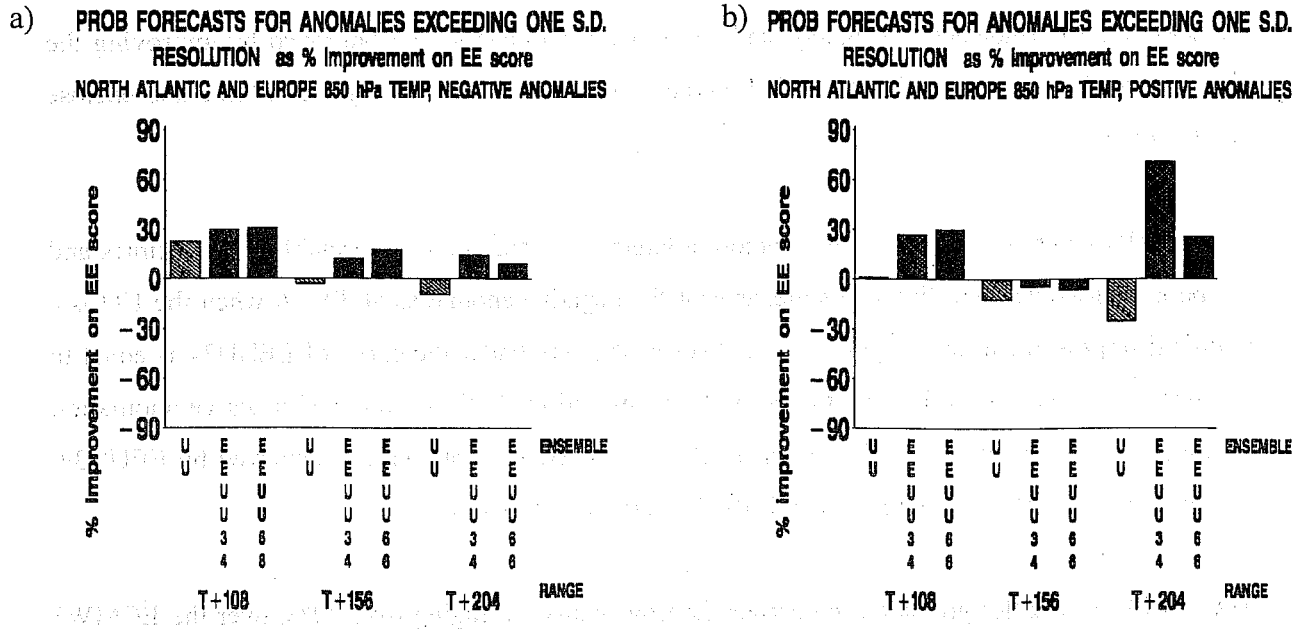


Figure 18. as Figure 16, but for resolution (note change in vertical scale).

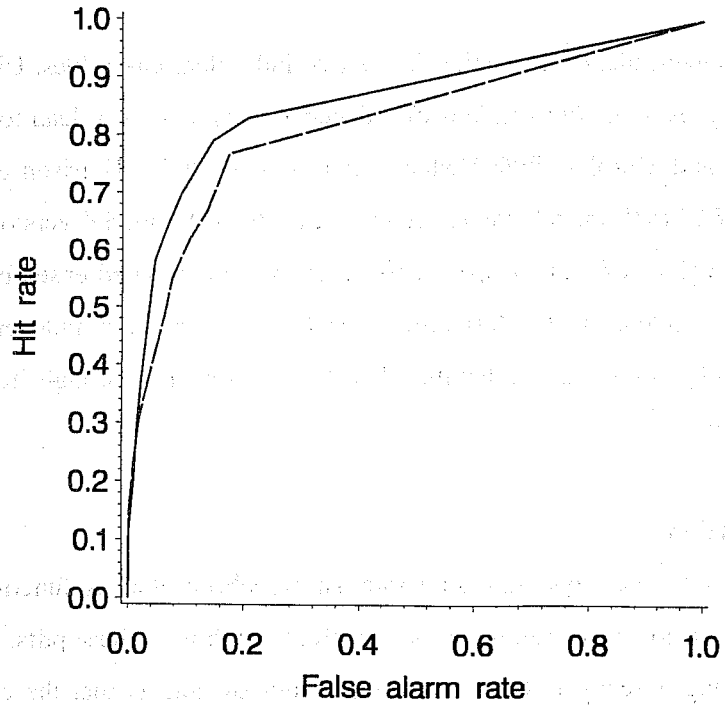


Figure 19. Sample Relative Operating Characteristic (ROC) curve for T+108 forecasts for 500 hPa height exceeding one standard deviation above normal over North Atlantic/European area. Results from 9 cases for UU ensembles (dashed) and EEUU34 ensembles (solid).

provides no information about reliability, and hence the curves can not be improved by improving the climatology of the system - a forecast of observed climatology will lie along the 45^0 line and enclose an area of 0.5.

For 500 hPa height the EE ensembles enclose a larger area than the UU ensembles for all times and for both positive and negative anomalies except for negative anomalies at T+156 when the UU is a marginal improvement on EE (Fig. 20). However, the area under the curve of EEUU34 is equal to or greater than that of the EE ensembles at all ranges and for both positive and negative anomalies, except for a marginal negative difference at T+204. The average improvement, achieved by EEUU34, is around 9% for positive anomalies and 6% for negative anomalies.

The MMMA ensembles produce even greater improvements, averaging over 12%, over the ECMWF system for 850 hPa temperature, and in contrast to Brier scores, the improvements in positive and negative anomalies are similar, as they are for resolution (Fig. 21).

6 Independence of information from different ensemble configurations.

In this Section, the extent of independent information in the two individual ensembles, UU and EE, is assessed. Previous studies suggest that combinations of independent forecasts can lead to enhanced forecast performance (Brown and Murphy 1996; Vislocky and Fritsch 1995). Harrison *et al.* 1995 showed that the UKMO and ECMWF models (at lower resolution than the model versions studied here) tend to explore different regions of a phase space defined from the combined ensemble system, and hence do indeed contain independent information. Here we assess the independence of information in the two systems by comparing probability density functions and through 'tubing' - a relatively new clustering technique.

6.1 Probability Density Functions

If the two individual ensemble systems always draw their forecast probability density functions (pdfs) from the same population then clearly there can be no benefit from combining these pdfs. Hence in order to benefit from combining ensembles it is necessary (but not sufficient) that the component systems sample different populations. Kolomogorov-Smirnov tests have been performed to test the hypothesis that forecast pdfs of the individual systems are taken from the same population. Frequently for both 500 hPa height and 850 hPa temperature, the two individual systems produce substantially

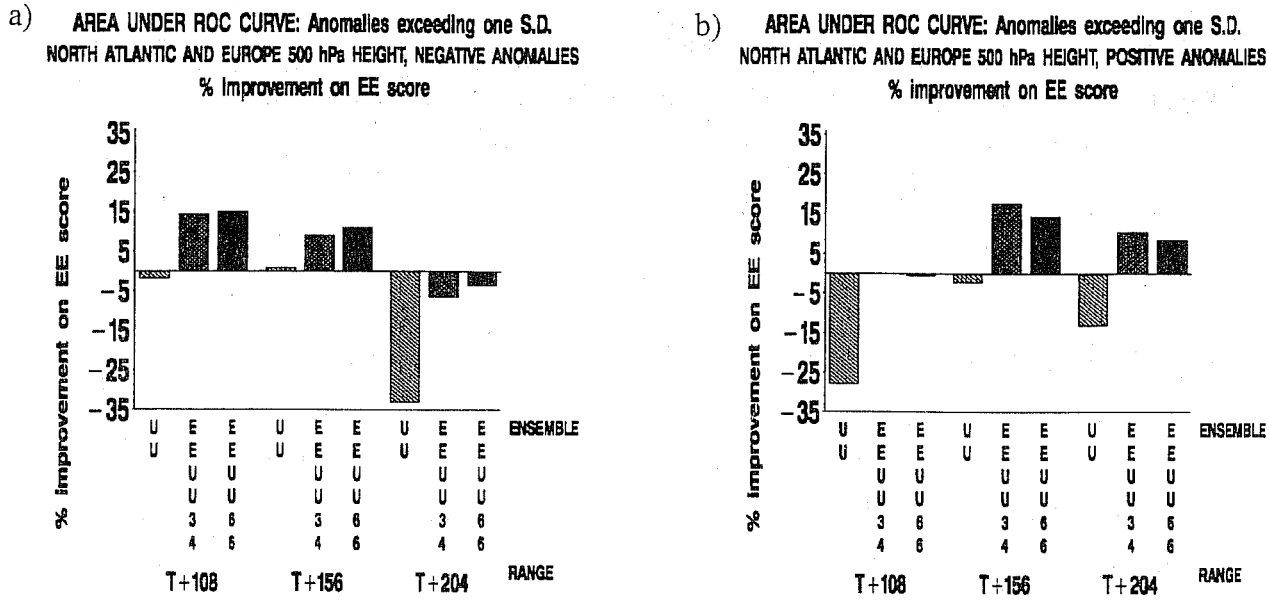


Figure 20. Area under ROC curves for T+108, T+156 and T+204 predictions of 500hPa height anomalies over the North Atlantic/European area exceeding one standard deviation. a) negative anomalies, b) positive anomalies. Results are presented as percentage improvement over EE ensemble (for 9 cases). UU hatching (left to right), EEUU34 cross hatching, EEUU66 solid.

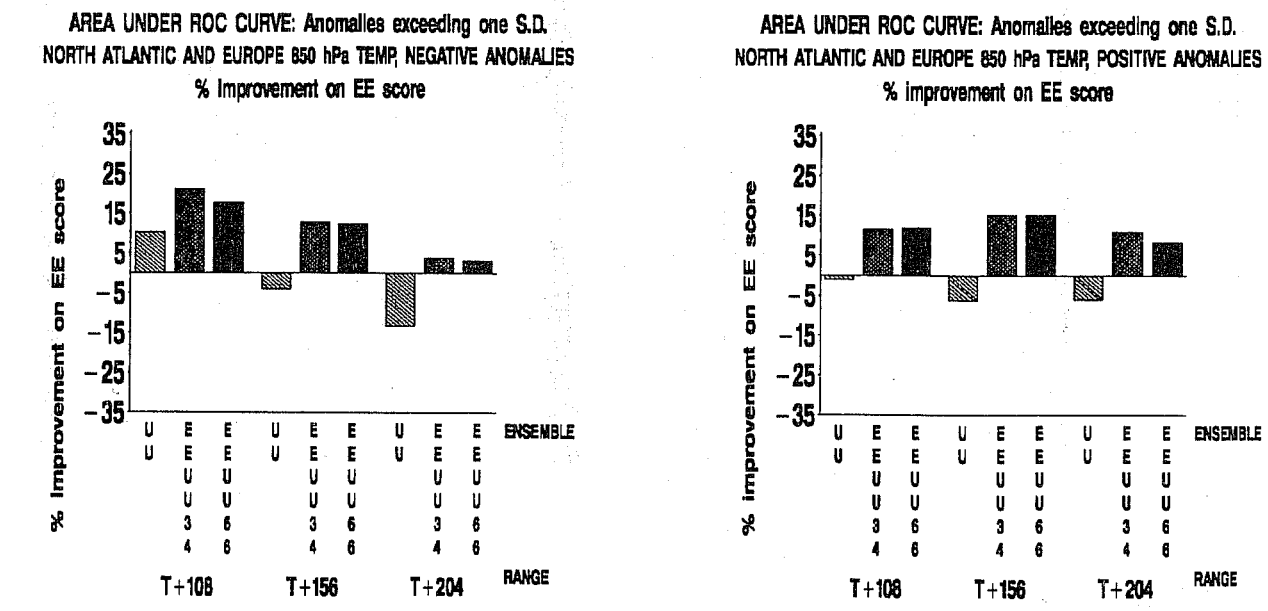


Figure 21. As Figure 20, but for 850 hPa temperature.

different pdfs. For 500 hPa height UU and EE forecast distributions are not drawn from the same population, at 95% significance level, over at least 40%, and up to 80%, of the extra-tropical Northern Hemisphere (>15°N) for all 9 cases and all forecast times (Table 4). For 850 hPa temperature the separation in the distributions is even stronger; significant differences exist between systems at over 50% of points. These values do not change substantially if the analysis is repeated after having removing model biases.

The geographical distribution of the areas of significance alters with time and between cases, but often includes Europe and North America. For example, for the forecast initialised on 24th December 1994 (Case 8), the null hypothesis - that UU and EE 850 hP temperature forecast distributions are drawn from the same population - can be rejected at the 95% level of significance over the European region, throughout the forecast (Fig. 22).

Table 4. Percentage of points in extra-tropical Northern Hemisphere where UU and EE pdfs are not drawn from the same population at 95% significance level.

CASE	500 hPa height			850 hPa temperature		
	T+108	T+156	T+204	T+108	T+156	T+204
2	56	70	58	73	69	67
3	66	68	64	77	73	65
4	51	41	40	64	60	52
5	61	51	40	66	54	50
6	68	56	44	73	73	70
7	82	66	60	79	68	63
8	75	66	50	77	72	59
9	63	65	65	78	70	58
10	79	71	67	79	61	56

6.2 Tubing

Atger (1996) developed tubing analysis as an alternative to traditional clustering techniques such as Ward's, these methods tend to average forecasts into small clusters and so lose information on extreme events. Tubing groups ensemble members into a central cluster around the ensemble mean and tubes originating from the central cluster. These tubes are designed to indicate possible ways in which the

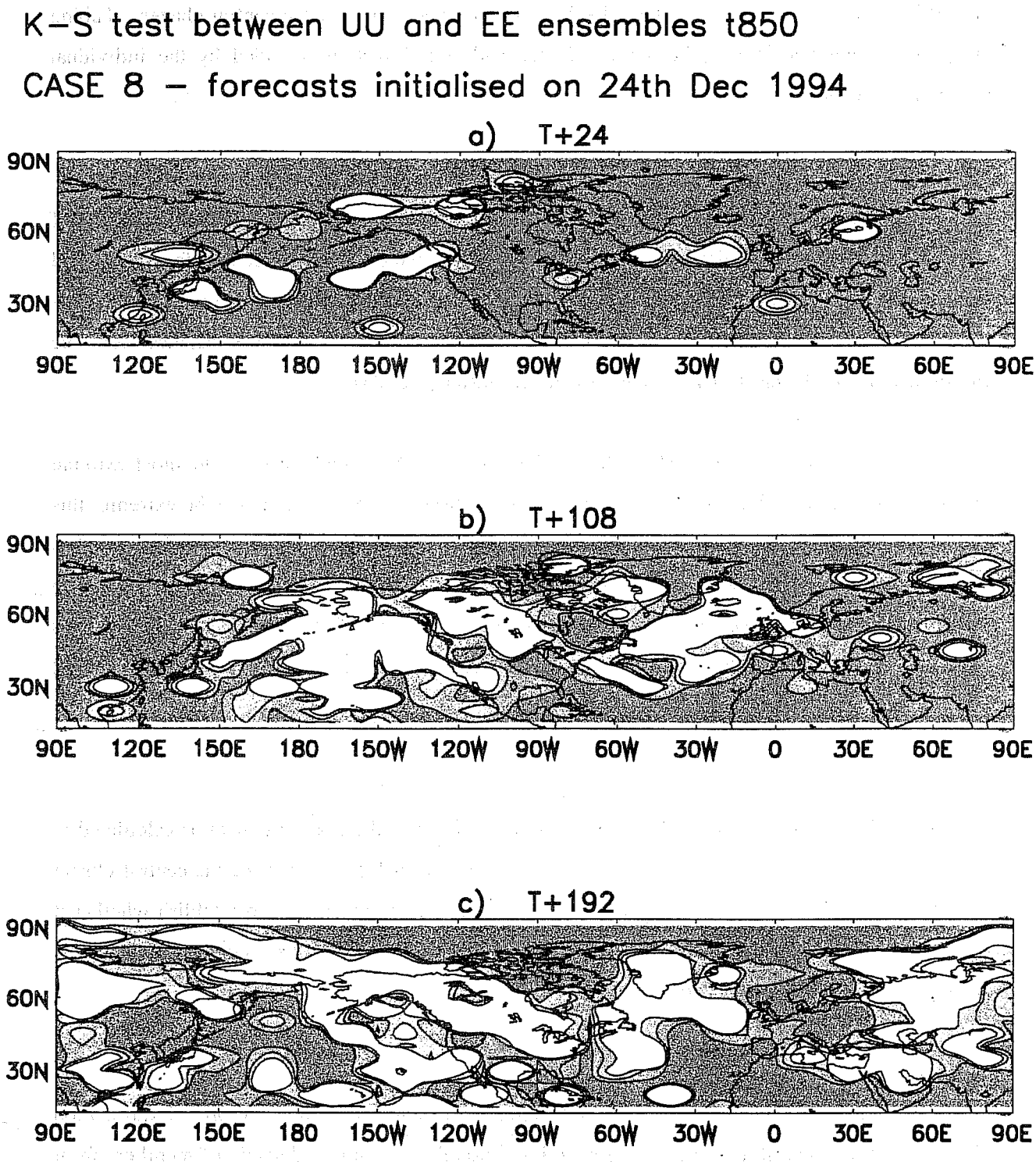


Figure 22. The geographical distribution of levels of significance of the Kolomogorov-Smirnov (KS) tests that UU and EE ensemble forecasts of 850 hPa temperature are from the same population. Significance shaded at 99%, 95% and 90% significance - darkest shading shows the regions where KS tests reject similarity with 99% significance. Results are for case 8 (initialised on 24 Dec 94); a) T+24, b) T+108, c) T+192

atmosphere may diverge from the most likely evolution represented by the central cluster. Tubing analysis provides insight into the degree of independent information provided by the individual ensemble systems.

6.2.1 The Tubing Method

The first stage of the process is to define the central cluster of the ensemble distribution as the set of members closest (in RMS distance) to the ensemble mean, which together account for a specified percentage of the variance of the entire ensemble (set to 50% for this study). The central cluster is represented by the mean of its members and its radius is defined as the RMS distance from the ensemble mean to the most extreme member of the cluster (Fig. 23).

Any remaining members are grouped into tubes originating from the central cluster. The most extreme member from the ensemble mean (measured with RMS distance) defines the first tube extreme; this first tube extends through phase space from the extreme to the central cluster with a radius equal to that of the central cluster. If the second most remote member from the ensemble mean does not lie inside the first tube (assessed by measuring the distance to the axis of the first tube), then it defines the extreme of the second tube. All remaining members are assessed in this way until they are all are accounted. Note that a member can belong to more than one tube but not to both the central cluster and a tube.

For verification the RMS distance from the verifying analysis to the ensemble mean is calculated to assess whether the analysis lies within the central cluster. If the analysis lies outside the central cluster then the distance between the analysis and each of the tube axes is measured to establish whether it lies inside any of the tubes. Note that the analysis can also lie in more than one tube or it can lie outside all the tubes and the central cluster.

6.2.2 Tubing results

Tubing has been applied to 500 hPa height and 850 hPa temperature fields over Europe. Examination of results suggests that if a tube from the combined ensemble is comprised solely of members from one system, then that system is supplying information on a possible synoptic outcome not available from the other. Over 40% of tubes generated from the EEUU66 ensemble forecasts of 500 hPa height at T+156 and T+228 consist of members from one individual system only (Table 5). UU ensembles frequently produce forecasts in directions not explored by the EE ensembles. For 850 hPa temperature

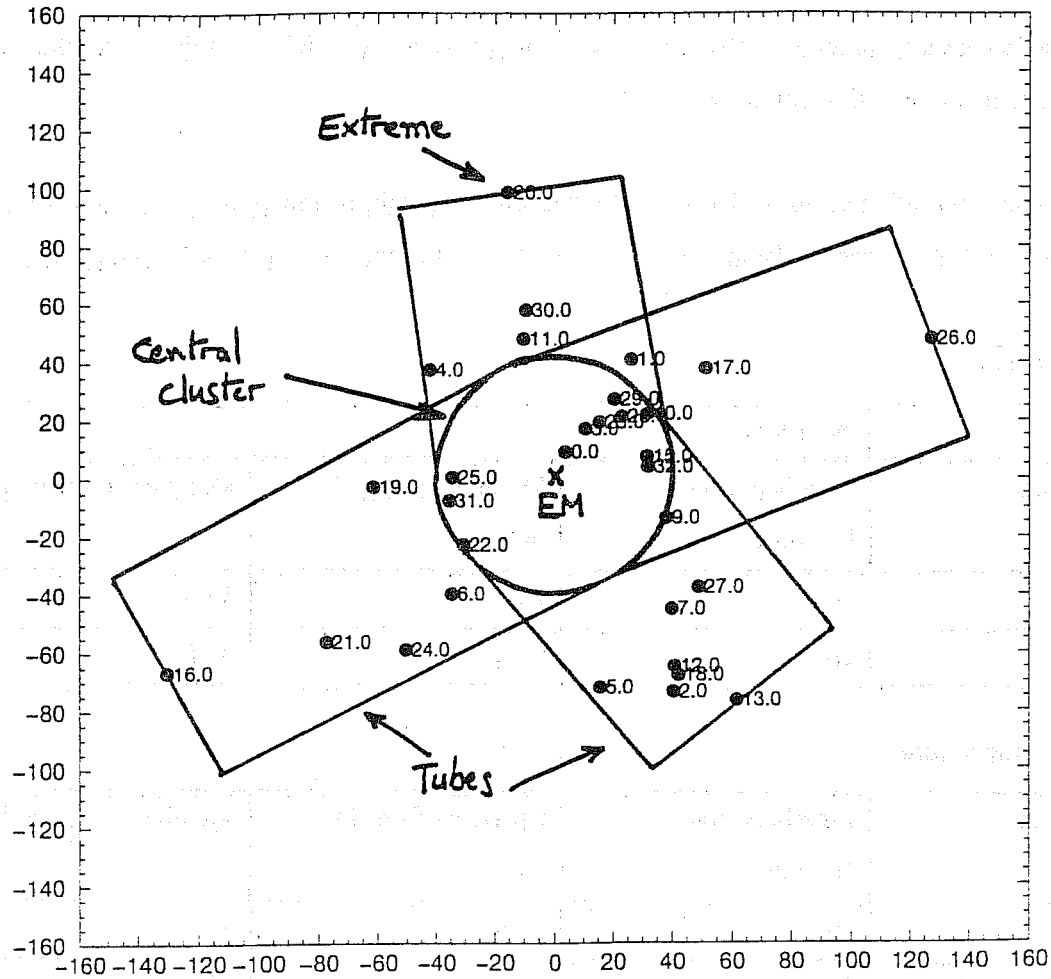


Figure 23. Schematic tubing applied to a random two-dimensional gaussian distribution. (Taken from Atger (1996)). Central cluster is a spherical cluster around ensemble mean (EM). Extremes are the most remote members from the EM and tubes originate from the central cluster, end on extremes, and have the same diameter as the central cluster.

CASE 6 (12th Dec 94) 500hPa height at T+156

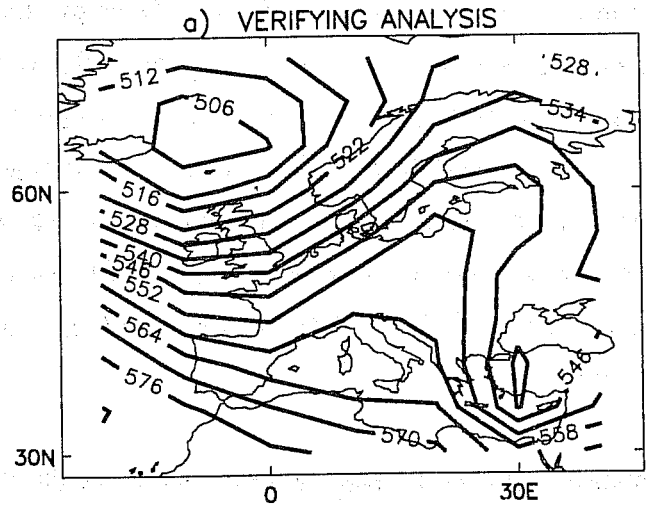


Figure 24a. Verifying analysis of 500 hPa height for T+156 for case 6 - forecast initialised on 12 Dec 94.

forecasts the separation between systems is even stronger, with over 60% of tubes consisting of members from just one individual system.

Table 5. Percentage of tubes formed from EEUU66 which contain members from one model only at T+156 and T+228 (9 cases). Tubing was performed over Europe and results are avenged over 9 cases.

a) 500 hPa height

	% of tubes		
Forecast time	Members from UU only	Members from EE only	Members from both
T+156	20.5	20.5	59.0
T+228	26.4	20.8	52.8

b) 850 hPa temperature

Forecast time	Members from UU only	Members from EE only	Members from both
T+156	29.0	44.9	26.1
T+228	22.9	41.4	35.7

6.2.3 Case of ensembles initialised on 12 December 94

The case of 12 December 94 provides one of the most substantial differences in forecast 500 hPa height synoptic patterns between the ensemble systems of any of the cases. All 4 tubes formed from EEUU66 at T+156 consist of members from one individual system. The verifying 500 hPa height field at T+156 has a low to the east of Iceland with a NE/SW ridge over Eastern Europe and a cut-off low over Turkey (Fig. 24a). The analysed field lies inside two of the tubes created by UU (Figs 24 b and c). One tube extreme captures the correct position of the low and has an RMSE of 11.7dam, and the other has an accurate representation of the ridge over Eastern Europe with RMSE of 8.9dam. In addition, the central cluster mean (RMSE of 8.6dam) has a good representation of the general trough-ridge pattern (Fig. 24d).

The analysis was not contained in any of the tubes or central cluster produced by the EE ensemble. Visual inspection of the tubes extremes and the central cluster mean confirms that none of the synoptic patterns resemble the observed field (Figs 24e-i). None of the scenarios forecast by EE capture the

CASE 6 500hPa height at T+156. Results from Tubing UU ensemble.

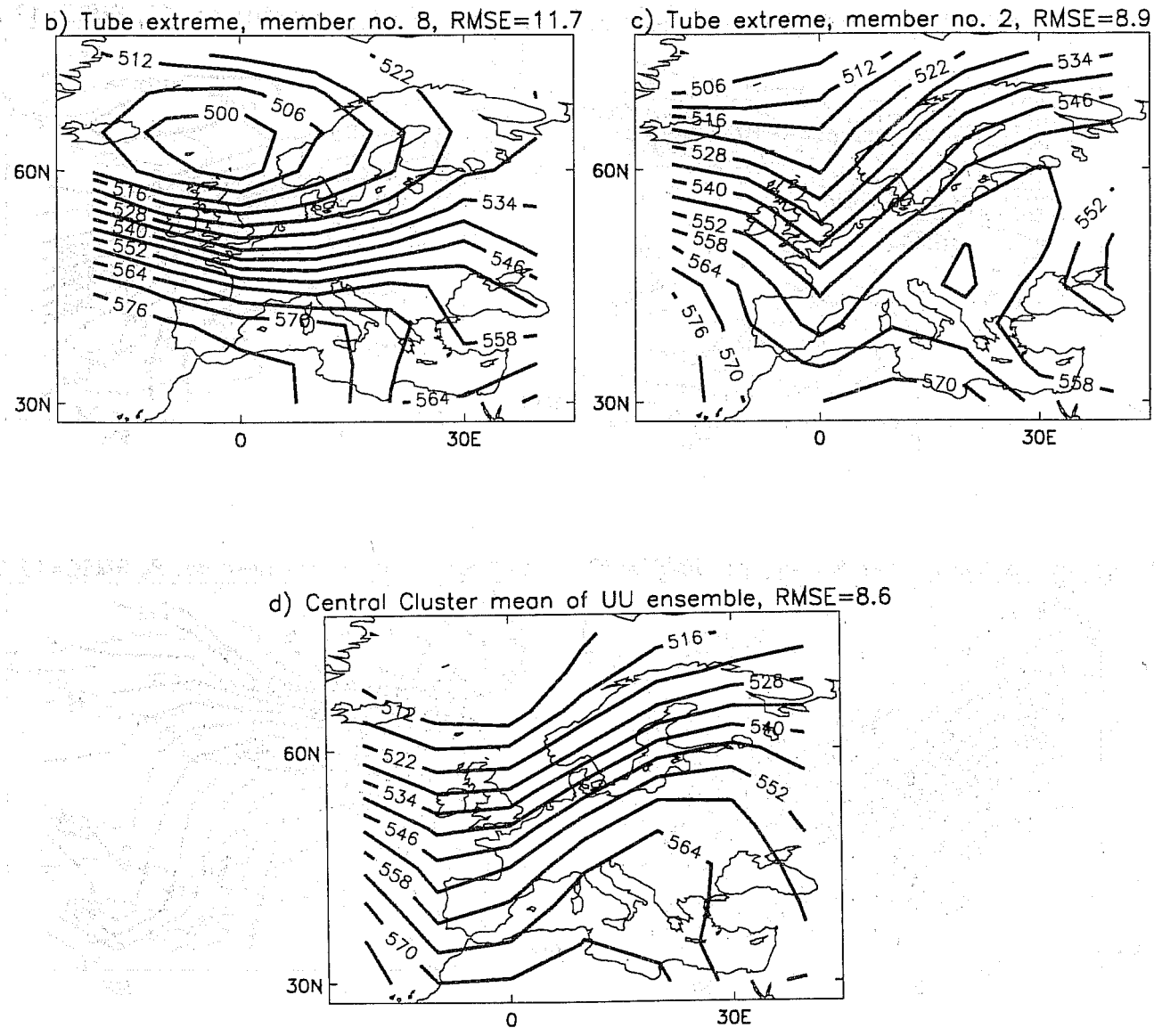


Figure 24 (continued). Results from tubing of UU ensemble from case 6 - forecast initialised on 12 Dec 94.

b) and c) Tube extremes of the two tubes, formed from UU ensemble, which contain the verifying analysis.

d) Central cluster mean from UU ensemble.

CASE 6 500hPa height at T+156. Results from Tubing EE ensemble.

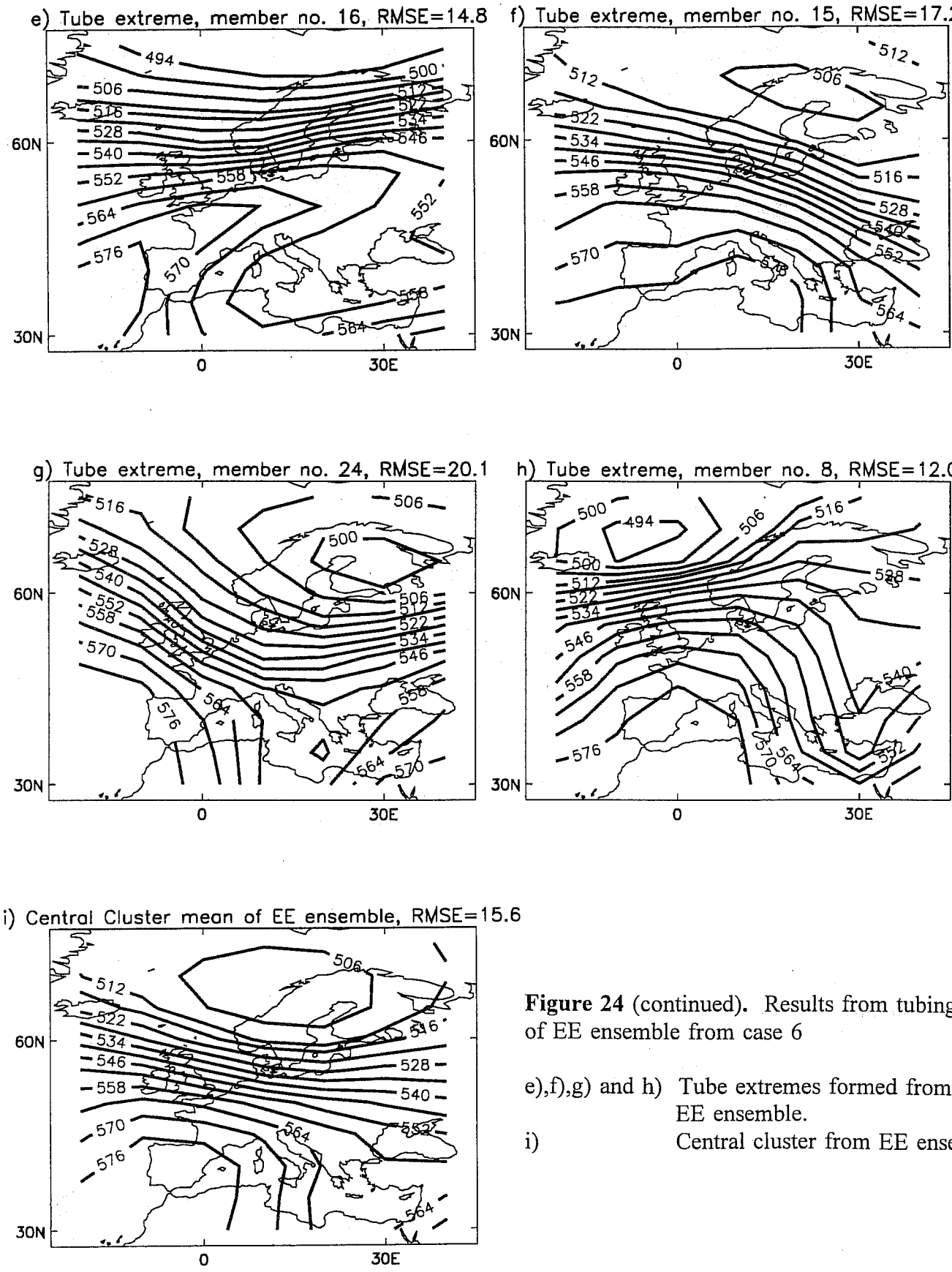


Figure 24 (continued). Results from tubing of EE ensemble from case 6

e),f),g) and h) Tube extremes formed from EE ensemble.

i) Central cluster from EE ensemble

full extent of the ridge over Central Europe. Member 8 (tube extreme, Fig. 24h) is the best forecast (measured using RMSE, member 8 has an RMSE of 12.0dam) with an accurate representation of the observed low off Iceland and slight ridging over Central Europe, but the flow over Northern Europe does not have the cyclonic curvature of the observed pattern and this forecast has not captured the magnitude of the observed ridge. Molteni *et al.* (1996) found that the frequency of this type of flow - strong ridge over Northern Europe and cut-off low over Southern Europe - is severely underestimated in the long term climatology of ECMWF's T63L19 model. This study uses a higher resolution, newer version of the ECMWF model, but in this case the ECMWF ensemble has failed to capture the observed pattern. In contrast the UKMO system contains successful representations of the solution and valuable synoptic information not present in the EE system. It is not clear whether these differences results from differences in analyses between systems, or differences in model formulation, or from a combination of both. The relative importance of model and analysis dependencies are examined in the following Section.

7 The relative benefits of model and analysis dependencies

7.1 Comparison of the amplitude of model and analysis dependencies

The difference between forecasts from corresponding members of the EE and UE ensembles gives an indication of the dependency on model formulation. Similarly the difference between forecasts from corresponding members of the UU and UE ensembles gives an indication of the dependency on the base analysis. A scatterplot of RMS for 500 hPa height over the extra-tropical Northern Hemisphere for all available member pairs of all 10 cases indicates that differences due to analysis dependencies (UU - EE) and model dependencies (EE - UE) are generally comparable (Fig. 25a). At early ranges (T+36) analysis dependencies have the larger amplitudes for the majority of members, but note that model dependencies are not insignificant, and there are members for which the amplitude of model dependencies are over twice the corresponding analysis amplitudes. Generally model effects become more important as forecast time increases. The average percentage of ensemble members for which model dependencies dominate increases through the forecast range, reaching about 50% by around Day 8, while by T+216 model dependencies tend to exceed analysis dependencies. Average RMS differences for each forecast range suggest that neither differences due to model or analysis dependencies saturate during the period, and that those due to model dependencies grow more quickly.

a)

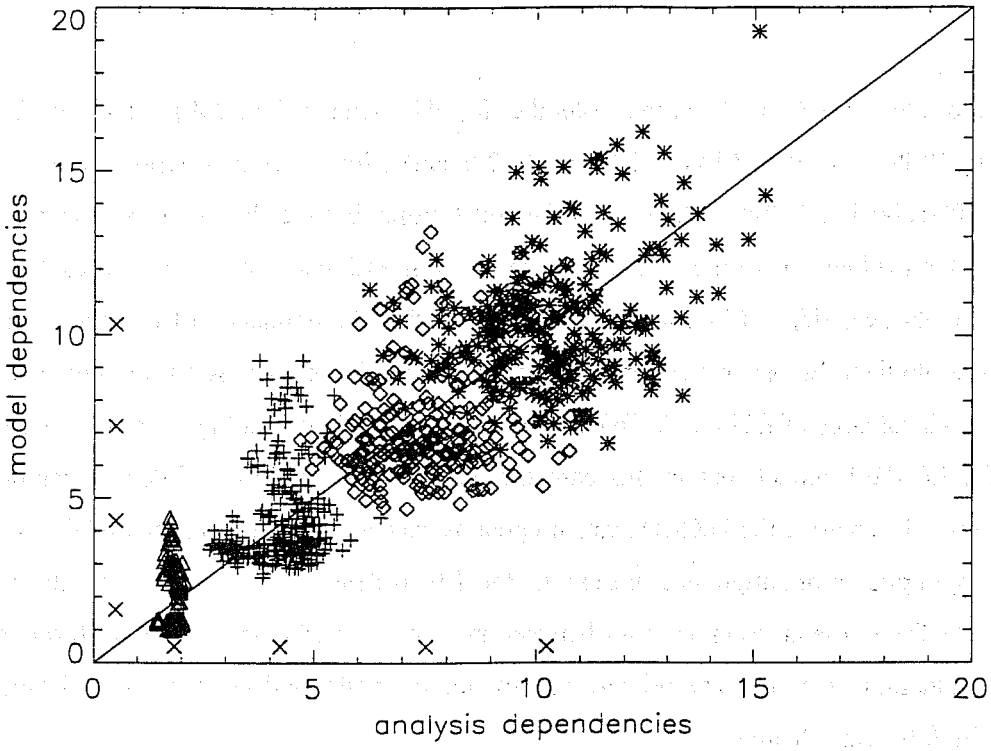


Figure 25a. Scatterplot of the impact of model dependencies against analysis dependencies on forecasts of 500 hPa height at T+36 (triangles), T+96 (pluses), T+156 (diamonds) and T+216 (asterisks) (average values for each forecast time are marked by large crosses). Model dependencies are represented by the RMS distance between paired members of EE and UE ensembles, and similarly analysis dependencies by the distance between UU and UE members. Results are for all available paired members from all 10 cases, over the Northern Hemisphere extratropics.

b)

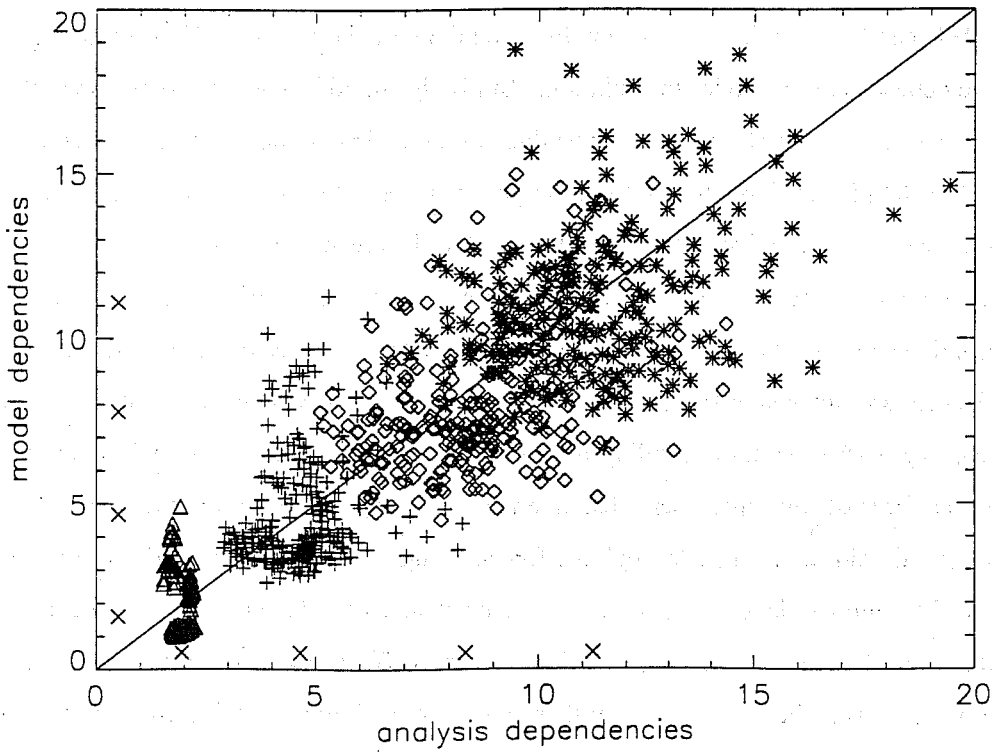


Figure 25b. As for fig. 25a but for standard deviation of model and analysis dependencies.

It is worth examining whether the increase in relative importance of model dependencies with forecast time is due to different biases in the two models - growth in systematic model error could result in increasing divergence between model forecasts. RMS difference EE - UE (model dependencies) may be decomposed into the contribution from mean differences between the models (as may arise from different model biases) and from the standard deviation (σ) from the mean - arising from local divergence between models:

$$Var(x)=E(x^2)-E(x)^2$$

or

$$\sigma^2=RMS^2-bias^2$$

A scatterplot constructed from standard deviation of model and analysis dependencies (Fig. 25b) is similar to that constructed from RMS differences (Fig. 25a), suggesting that the model dependency derives mainly from a tendency for local evolution differences between the models, rather than from any differences in model bias.

The information in the scatterplot of Fig. 25a may be summarised to show the frequency with which the analysis amplitude (defined, as above, by RMS distance) is larger than 110% and also 150% of the model amplitude. By T+96 the frequency for the 110% comparison has reached about 50%, indicating that on average the model dependency is only modestly smaller that the analysis dependency at this range (Fig. 26). Even at T+36 the amplitude of model sensitivity, when inflated by 150%, is more than the analysis amplitude for the majority of members, according to the 150% comparison. A case by case breakdown indicates that there are two cases where model dependencies dominate even at early stages of the forecast (Fig. 27) (forecasts initialised on 27 November 1994 and 18 December 1994). These two cases have been studied in more detail by Harrison *et al.* 1995 who also found substantial forecast sensitivity to model effects even at early forecast range. Thus although analysis dependencies appear to be the more prominent at the shorter ranges, model dependencies are distinctly non-negligible even at these times.

For 850 hPa temperature forecasts, model dependencies are larger than analysis dependencies for most ensemble pairs throughout the forecast - even in the initial stages (Fig. 28a). The relative importance of model dependencies declines slowly with forecast time, but by T+216 the RMS amplitude of model

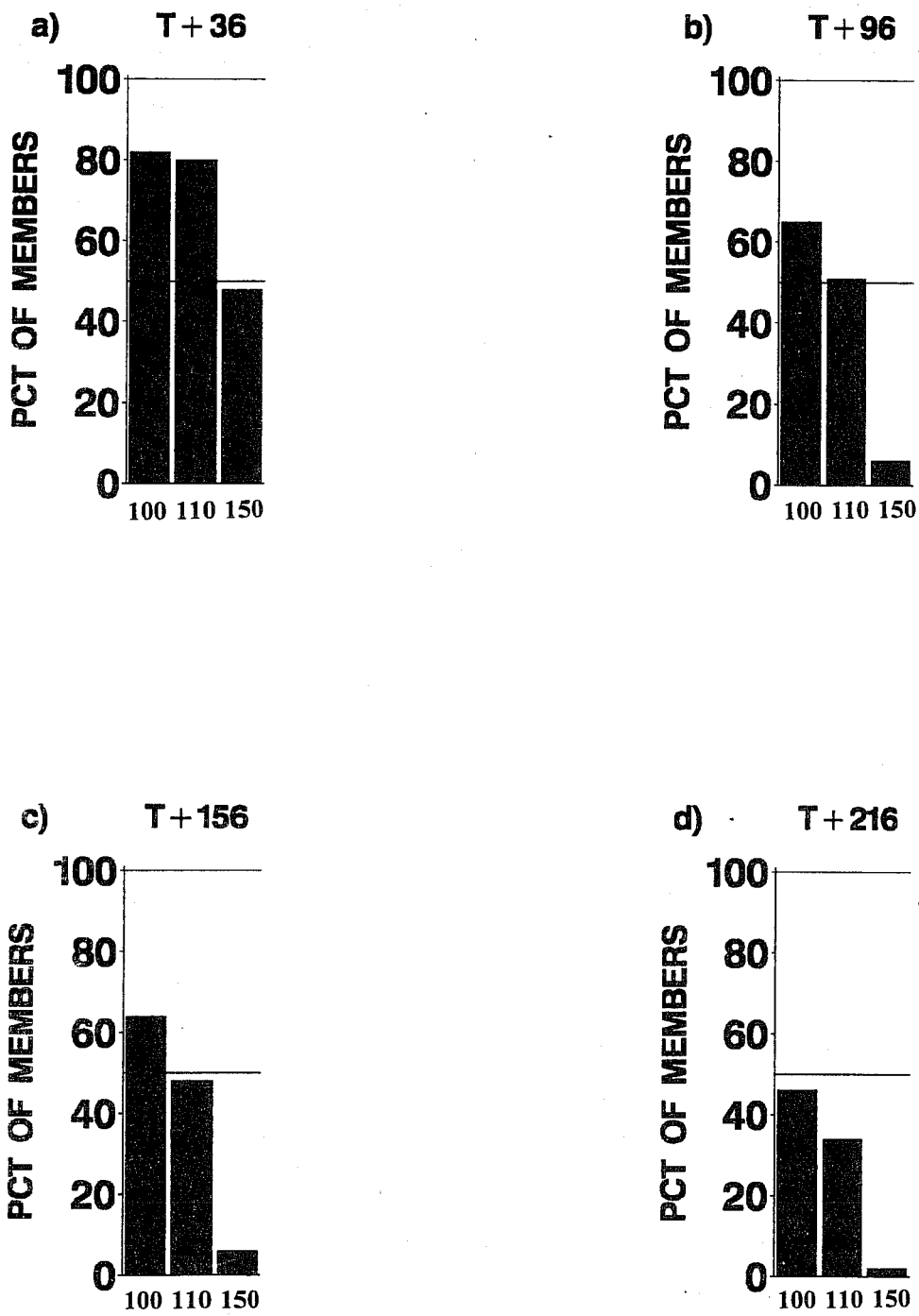


Figure 26. Average percentage of members with 500 hPa height RMS amplitude of analysis dependencies (UU-UE),
larger than RMS amplitude of model dependencies (EE-UE), (bar 1)
larger than 110% of model dependencies, (bar 2), and
larger than 150% of model dependencies, (bar 3).
Results are for all available paired members from all 10 cases, over the Northern Hemisphere extratropics. a) T+36, b) T+96, c) T+156, d) T+216.

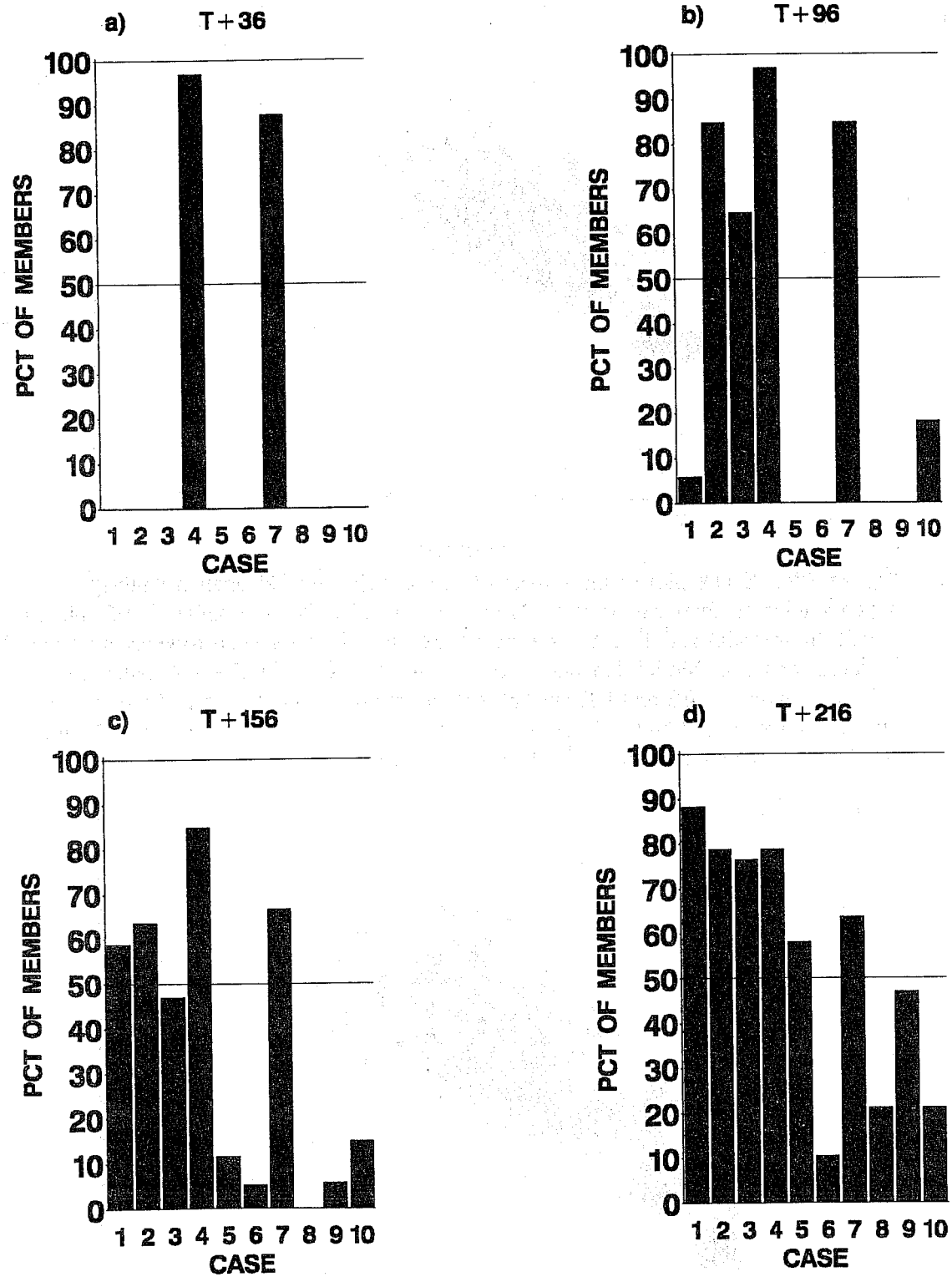


Figure 27. Percentage of members, for each case, with 500 hPa height RMS amplitude of model dependencies (EE-UE) larger than RMS amplitude of analysis dependencies (UU-UE). Calculated over Northern Hemisphere extratropics for all available paired members. a) T+36, b) T+96, c) T+156, d) T+216.

a)

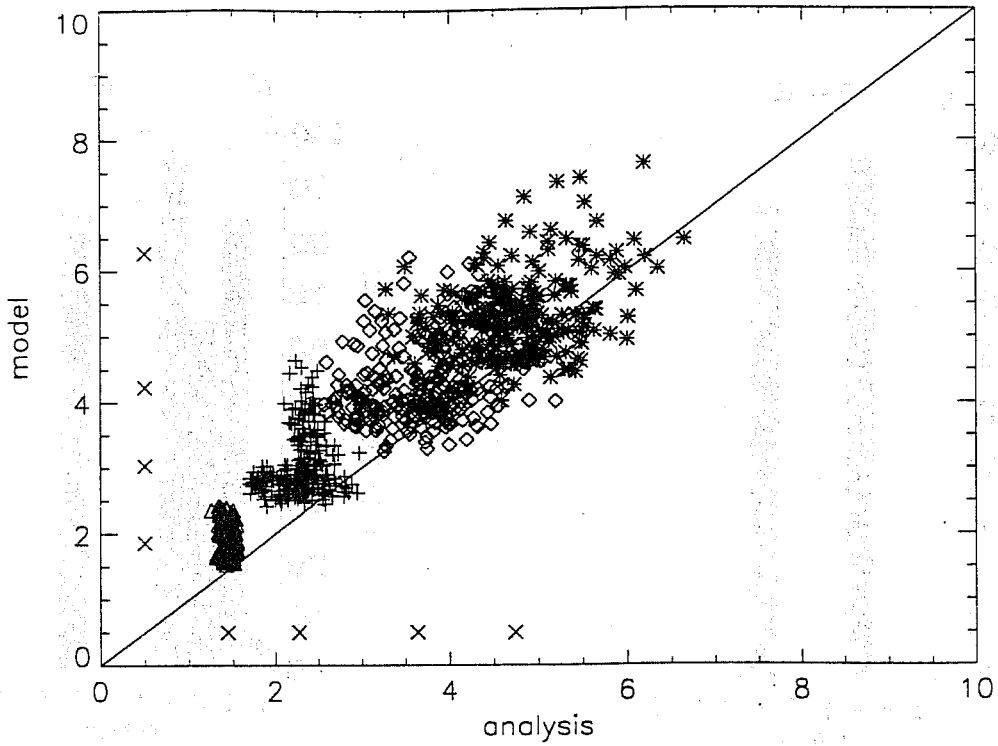


Figure 28a. Scatterplot of the impact of model dependencies against analysis dependencies on forecasts of 850 hPa temperature at T+36 (triangles), T+96 (pluses), T+156 (diamonds) and T+216 (asterisks) (average values for each forecast time are marked by large crosses). Model dependencies are represented by the RMS distance between paired members of EE and UE ensembles, and similarly analysis dependencies by the distance between UU and UE members. Results are for all available paired members from all 10 cases, over the Northern Hemisphere extratropics.

b)

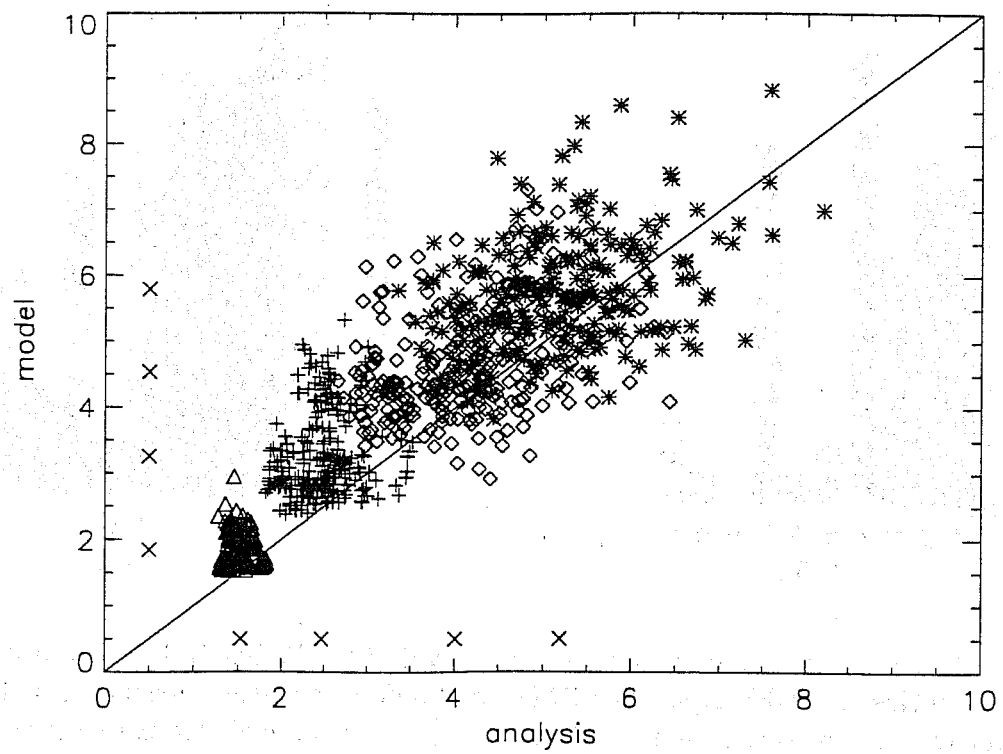


Figure 28b. As for Figure 28a but for standard deviation of model and analysis dependencies.

dependencies is still larger than that of analysis dependencies for the majority of member pairs. And removal of average biases only slightly reduces the relative importance of model dependencies (Fig. 28b).

Tubing analysis can also be used to investigate model and analysis dependencies. The percentage of UE members lying outside the tubes and central cluster formed by the EE ensembles provides a measure of the extent of synoptic divergence resulting from model differences, while the effect of analysis differences can be assessed using UE and UU ensembles. (Although note that comparisons between model and analysis dependencies are not direct, as model dependencies are measured against EE whereas analysis dependencies are against UU.) The analysis has been performed over Europe for the 9 cases where all 33 members of UU are available (Table 1).

For forecasts of 500 hPa height the largest percentages of outlying members are found when fitting members from UU to the tubes and central cluster generated from EE, and vice-versa (Fig. 29a). The independence of synoptic information is largest between ensembles run with different models and different analyses. Sensitivity to model effects alone is substantial, with over 30% of UE members outlying the tubes/central cluster formed by EE (bars labelled M). The additional information content, over the EE ensemble, gained from adding first an additional model and second an additional model and analysis may be appreciated by comparing the bars labelled M with bars labelled B.EE, and indicates that gains from model and analysis dependencies are comparable. To gain a fuller perspective the additional information content, over the UU ensemble, gained from adding first an additional analysis and second an additional model and analysis may be appreciated by comparing the bars labelled A with bars labelled B.UU; this comparison indicates that analysis dependencies contribute most to overall gains, but that gains from model dependencies are substantial, and increase with forecast time (the percentage of EE members lying outside tubes/central cluster from UU is 18 percentage points greater than for the UE ensemble by T+204).

Tubing analysis suggests model effects generate substantial synoptic differences in forecasts of 850 hPa temperature, over 40% of UE members lie outside the tubes and central cluster created by EE ensembles for forecasts between T+108 and T+204 (Fig. 29b, bars labelled M). As for 500 hPa height, differences in both model formulation and initial analysis are required for the largest synoptic differences between systems. Again this can be appreciated by comparing the bars labelled B against those labelled M and A; for example, using the UU tubes/central cluster as reference, the percentage

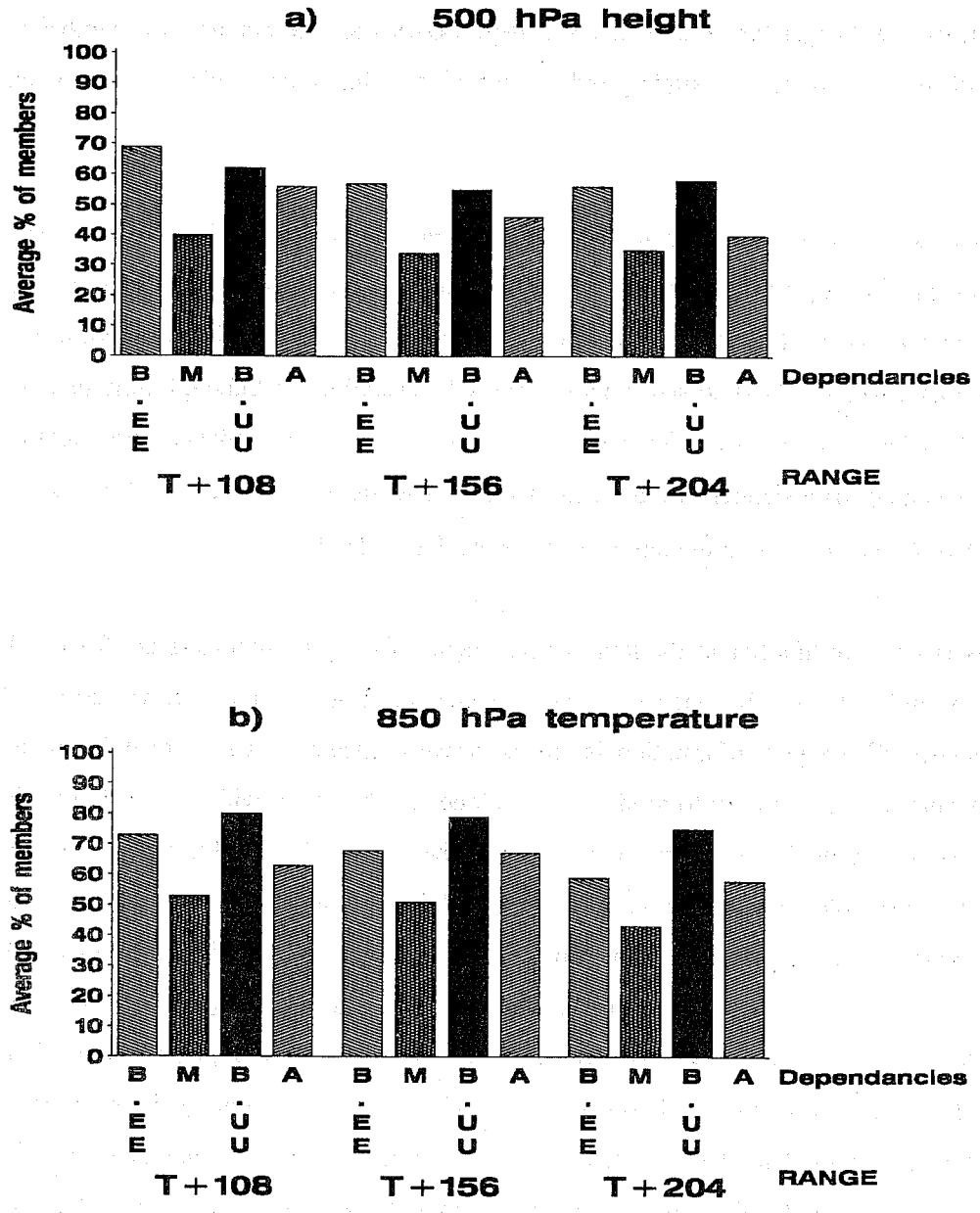


Figure 29. Tubing analysis for forecasts at T+108, T+156 and T+204, over Europe. Results averaged over 9 cases.

B.EE - **(Model and analysis dependencies)** Average % of UU members lying outside tubes and central cluster formed by EE.

M - **(Model dependencies)** Average % of UE members lying outside tubes and central cluster formed by EE.

B.UU - **(Model and analysis dependencies)** Average % of EE members lying outside tubes and central cluster formed by UU.

A - **(Analysis dependencies)** Average % of UE members lying outside tubes and central cluster formed by UU.

a) 500 hPa height
b) 850 hPa temperature

of EE members, on average, lying outside UU (model and analysis dependencies) is 17% greater than the number of UE members lying outside UU (analysis dependencies only).

7.2 Relative impact of model and analysis dependencies on forecast skill

The results described above suggest that the relative magnitude of model and analysis dependencies are broadly similar, and that the optimum joint ensemble will encompass both model and analysis dependencies. To test this hypothesis, probabilistic verification (Brier Score) has also been performed on the three 34 member joint ensemble sets; EEUU34 (MMMA), UUUE (multi-analysis) and EEUE (multi-model). This allows evaluation of improvements to the ECMWF system made by addition of a) UKMO model alone (EEUE) and b) UKMO model and analysis (EEUU34). And similarly, comparison between UU, UUUE and EEUU34 allows evaluation of the benefits of adding just the ECMWF analysis or both ECMWF model and analysis to the UKMO system. As before results are for probability forecasts of 500 hPa height and 850 hPa temperature above or below one standard deviation of normal, over the Europe and North Atlantic region, and averaged over 9 forecasts. The scores are presented as skill scores with EE as standard, and here results are averaged over three times (T+108, T+156 and T+204).

7.2.1 500hPa height

The MMMA ensemble (EEUU34) produces the largest consistent improvements in Brier Scores (Tables 6a and b). Introduction of the UK model alone to the ECMWF system (EEUE) improves average Brier Scores by around 4% relative to EE ensemble, introduction of both UKMO model and analysis (EEUU34) increases the benefit to nearly 9% (Table 6a). For this diagnostic the UKMO system (UU) is marginally more skilful than the EE system (skill score 0.68%). However, addition of the ECMWF analysis to the UKMO system (UUUE) increases the score to 4% (Table 6b) with a further increase to nearly 9% obtained with the EEUU34 system. These results again demonstrate that both an additional model and analysis are required for the maximum benefits.

Table 6. Percentage improvement over EE for Brier Scores of probability forecasts of 500 hPa height exceeding one standard deviation both above and below normal, scores are averaged over 3 forecast times (T+108, T+156 and T+216) and over 9 cases.

a)

Model system	Positive anomalies	Negative anomalies	Average
EEUE (model dependencies)	4.15	10.46	7.31
EEUU34 (both)	5.05	14.27	9.66

b)

Model system	Positive anomalies	Negative anomalies	Average
UU	-1.27	2.62	0.68
UUUE (analysis dependencies)	3.56	5.08	4.32
EEUU34 (both)	9.97	7.88	8.93

7.3.2 850 hPa temperature

Verification of probability forecasts of 850 hPa temperature exceeding one standard deviation from normal, indicate that the MMMA ensemble is consistently the most accurate of the three 34 member ensembles (Tables 7a and b). Introduction of the UKMO model alone to the ECMWF system (EEUE) brings enhanced benefits for 850 hPa temperature compared with 500hPa height, with average improvements in Brier scores of over 7% relative to EE (Table 7a). Addition of the UKMO analysis as well as the UKMO model (EEUU34) results in a further increment of skill of around 2%. In contrast to the results for 500hPa height, addition of the ECMWF model to the UKMO system (UUUE) produces slightly greater increases in Brier skill score, for negative anomalies, than addition of both analysis and model (EEUU34) (Table 7b). This is likely to be due to the better performance of the UKMO (improvement of 5.81% over EE) for this measure, but for all other results here, both models and analyses are required for the largest average improvements.

Table 7. Percentage improvement over EE for Brier Scores of probability forecasts of 850 hPa temperatures exceeding one standard deviation both above and below normal, over North Atlantic and Europe. Scores are averaged over 3 forecast times (T+108, T+156 and T+216) and over 9 cases.

a)

Model system	Positive anomalies	Negative anomalies	Average
EEUE (model dependencies)	4.15	10.46	7.31
EEUU34 (both)	5.05	14.27	9.66

b)

Model system	Positive anomalies	Negative anomalies	Average
UU	-0.07	11.68	5.81
UUUE (analysis dependencies)	4.76	14.72	9.74
EEUU34 (both)	5.05	14.27	9.66

8 Discussion and Conclusions.

Ensemble sets using the UKMO and ECMWF models, each run from their own analysis for ten cases, have been studied in detail to assess the benefits of combining ensembles from different model systems. In addition hybrid ensembles consisting of the UKMO model run from the ECMWF analysis have been used to evaluate the relative importance of model and analysis dependencies. Both 500 hPa height and 850 hPa temperature fields have been used for verification, and the analysis has been largely restricted to the European and North Atlantic region, although preliminary examination confirms applicability elsewhere.

For both deterministic and probabilistic verification, the joint ensemble formed with both models run from their own analysis, significantly outperforms either individual system. This improvement approximately equates to a gain in predictability of the order of one day on medium-range timescales. In deterministic terms one of the major benefits gained from the multi-model and multi-analysis approach is the improvement in 500 hPa height field spread/skill correlation. The joint ensemble attains useful levels of correlation between spread and ensemble mean skill and its performance is

close to that obtained in internally consistent ensembles. Spread is larger within the joint ensemble and it also has better coverage of observations than either individual system, which implies that joint model probability forecasts are sampled from a fuller population. Substantial increases in reliability achieved by the multi-model and multi-analysis ensemble are not at the expense of resolution, and so are not simply due to a more accurate model climate but to real improvements in forecast accuracy. In general all these benefits are achieved with no significant increase in ensemble size, and so can be gained with minimal increase in computer costs.

These substantial improvements in forecast performance achieved by the joint ensemble may stem from the combination of independent information contained in the two individual systems (Brown and Murphy 1996, Vislocky and Fritsch 1995). There is evidence that the two individual systems can produce independent information; the systems sample different skilful populations, and importantly, the UKMO system has been shown to include synoptically valuable information not covered in the ECMWF ensemble system.

Evaluation of the relative importance of model and analysis dependencies suggests that on medium-range timescales model dependencies are at least comparable with analysis dependencies. On average, for 500 hPa height, the relative magnitude of model dependencies is non-negligible in the first 48 hours, and increases with forecast time, equalising with analysis dependencies around Day 8 and predominating beyond. For 850 hPa temperature, model dependencies dominate analysis dependencies on all time ranges. Further, both models and analyses are required for the broadest synoptic coverage, fullest forecast population and the maximum improvements in probabilistic forecast skill.

The results reported in this paper indicate that sensitivity to model formulation is important on medium-range timescales and that combining ensembles from UKMO and ECMWF systems can lead to substantial improvements in forecast skill.

Acknowledgements

The authors would like to thank Kelvyn Robertson, Tim Legg, Alan Woodcock, David Richardson, Tim Palmer, Roberto Buizza and Frederic Atger.

References

Atger, F. 1996: Use of ensemble prediction in operational forecasting. *Report on the expert meeting on ensemble prediction system*, ECMWF 17-18 June 1996, p37-53.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1-3.

Brown, B. G., and A. H. Murphy, 1996: Improving forecasting performance by combining forecasts: The example of road surface temperature forecasts. *Meteorological Applications*, **3**, 257-265.

Buizza, R., 1997: Potential Forecast skill of Ensemble Prediction and Spread and Skill Distributions of the ECMWF Ensemble Prediction System. *Monthly Weather Review*, **125**, 99-119

Hall, C. D., R. A. Stratton, and M. L. Gallani, 1995: Climate simulations with the Unified Model: AMIP runs. UKMO Climate Research Technical Note No.61.

Houtekamer, P. L., and L. Lefaivre, 1996 A system simulation approach to ensemble prediction. *Monthly Weather Review*, **124**, 1225-1242.

Harrison, M. S. L., T. N. Palmer, D.S . Richardson, R. Buizza, and T. Petroliagis, 1995: Joint medium range ensembles from UKMO and ECMWF models and analyses. *Seminar on predictability volume II.* ECMWF, 4-8 September 1995, 61-120.

Lanzinger, A., and B. Strauss, 1995 EPS evaluation at ECMWF. *Fifth Workshop on Meterological Operational Systems*. ECMWF 13-17 November 1995, 87-101.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagia, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **529**, 73-120.

Murphy, A. H. 1973: A new vector partition of the probability score. *J. Appl. Meteor.* **12**, 595-600.

Murphy, A. H., and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7, 435-455.

Rabier, F., E. Klinker, P. Courtier, and A. Hollingsworth. 1996: Sensitivity of forecast errors to intial conditions. *Quarterly Journal of the Royal Meteorological Society*. **122**, 121-150

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in Meterology. World Weather Watch Technical Report. No. 8, WMO/TD 358, 114pp.

Toth, Z., and E. Kalnay, 1995: Ensemble Forecasting with imperfect models. Research activites in atmospheric and oceanic modelling. 6.30.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bulletin of the American Meteorological Society*, **76**, 1157-1164.

Wobus R. L., and Kalnay, E. 1994: Two years of operational prediction of forecast skill at NMC, *Proc. Tenth Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc., 166-167.