

DUPLICATE

Numerical Weather Prediction



Forecasting Research
Technical Report No. 258

Preliminary Results from Quasi-Operational Multi-Model Multi-Analysis Ensembles on Medium-Range Timescales

by

R E Evans, K R Mylne and M S J Harrison

December 1998

ORGS UKMO F

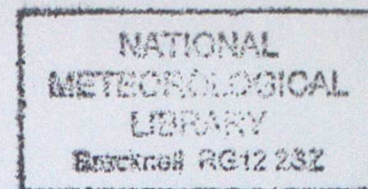
National Meteorological Library
FitzRoy Road, Exeter, Devon. EX1 3PB



The Met.Office

Excelling *in weather services*

DUPLICATE



**Forecasting Research
Technical Report No. 258**

**Preliminary Results from
Quasi-Operational Multi-Model Multi-Analysis
Ensembles on Medium-Range Timescales**

by

R E Evans, K R Mylne and M S J Harrison

December 1998

**Meteorological Office
NWP Division
Room 344
London Road
Bracknell
Berkshire
RG12 2SZ
United Kingdom**

© Crown Copyright 1998

Permission to quote from this paper should be obtained from the above Meteorological Office division.

Please notify us if you change your address or no longer wish to receive these publications.

Tel: 44 (0)1344 856245 Fax: 44 (0)1344 854026 e-mail: jsarmstrong@meto.gov.uk

Preliminary Results from Quasi-Operational Multi-Model Multi-Analysis Ensembles on Medium-Range Timescales

R.E. Evans, K.R. Mylne and M.S.J. Harrison

Abstract

Previous work using case studies has demonstrated that substantial benefits can be gained through combining ensembles from ECMWF and UKMO forecast systems using both models and both analyses. In order to examine the benefits of multi-model and multi-analysis ensembles (MMAE) in quasi-operational mode a MMAE system has been implemented at ECMWF using latest versions of the ECMWF and UKMO systems. This system has run daily since mid-October 1998. Preliminary results based on the first 23 days confirm that the joint ensemble significantly outperforms either individual ensemble in many performance aspects, including deterministic and probabilistic forecast skill, spread/skill correlations and coverage of observations. It is demonstrated that these improvements are not due to simple improvements in forecast climate but real improvements in forecast accuracy.

1. Introduction

The Ensemble Prediction System (EPS) at ECMWF was initially designed to examine the sensitivity of forecasts to uncertainties in the initial conditions (Palmer *et al.* 1993; Molteni *et al.* 1996). This is achieved by generating multiple integrations of the ECMWF model from slightly different initial analyses. However, errors in the initial conditions are not the only uncertainties in the forecast process; model-dependent errors such as deficiencies in parametrisations, systematic or regime dependent errors can also severely affect the skill of the ensemble (Toth and Kalnay 1995; Molteni *et al.* 1996).

Both model and analysis dependencies are incorporated in the Canadian System developed at Recherche en Prevision Numerique (RPN), where several versions of the operational model with modified physics and different parametrisations are used to produce both analyses and ensemble members (Houtekamer and Lefaiivre, 1996). In October 1998 ECMWF incorporated random model error into the EPS by introducing stochastic perturbations of physical tendencies during the course of model integrations. Research experiments indicate that these perturbations enhance spread slightly, with notable meteorological differences in some of the resulting forecasts (Buizza *et al.* 1998). However, this random forcing cannot address systematic or regime-dependent errors.

One simple method of incorporating model dependencies into ensemble forecasts is to combine ensembles run using different models. In addition, the use of more than one

analysis, to which perturbations are added, may provide further unstable directions for growth not present with a single analysis. Previous work using case studies has demonstrated that substantial benefits can be gained through combining ensembles from ECMWF and UKMO forecast systems (Harrison *et al.* 1995; Evans *et al.* 1998b). Evans *et al.* (1998b) demonstrated that the UKMO ensembles can contain valid information not covered by the ECMWF forecasts and that combining these two sets of independent information leads to improvement in forecast performance. The improvement corresponded to a gain in predictability of the order of one day and in general could be achieved without an increase in ensemble size. Many different aspects of forecast performance were examined: deterministic and probabilistic skill, spread, spread/skill correlations and synoptic information - all were found to be improved by combining members from two model systems. However, this study was based on just 9 cases and the models used have since been improved. Hence, in order to examine the benefits of multi-model and multi-analysis ensembles (MMAE) thoroughly, a quasi-operational MMAE system has been implemented at ECMWF using latest versions of the ECMWF and UKMO systems. This system has run daily since mid-October 1998, and preliminary results based on the first 23 days are presented in this note.

2. Data and analysis methods

The ECMWF component is taken from the operational EPS - this consists of 51 members run using the T_L159 with 31 vertical levels. The perturbations are calculated at T42 and added to and subtracted from the ECMWF operational analysis to form the initial conditions. The UKMO integrations are generated using the current operational version of the UKMO Unified Model (UM) (Vn4.4) run on the Fujitsu at ECMWF at 1.25° longitude and 0.83° latitude resolution with 30 vertical levels. ECMWF's perturbations are added/subtracted to the UKMO operational analysis to form the initial conditions for the UM runs. Limited computer resources mean that just 27 members are run using the UKMO model - the control plus the first 13 pairs of perturbations. This horizontal resolution is coarser than that used for the UKMO operational forecasts (0.83° longitude by 0.55° latitude) which are run on UKMO's Cray T3E. There are other small differences between the Fujitsu and operational versions of the UM, including different values for some tunable parameters, and of course different computer architecture. Some preliminary model validation has been performed to compare biases and synoptic development in the control run of the UKMO ensemble with the operational UKMO forecasts and no problems were discovered. This system has run daily since mid-October 1998, and these preliminary results are based on the first 23 days. It should be noted that as the forecasts are from consecutive days, they are not fully independent cases.

An automatic verification system is in place to generate an initial set of results; its main properties are described in this section. Five different ensemble configurations are assessed, referred to as follows:

Individual systems

- EE ECMWF model from ECMWF analysis (EPS) - 51 members.
- UU UKMO model from UKMO analysis - 27 members.
- EE27 Corresponding 27 members of the EPS - allows direct comparisons between the UKMO and ECMWF systems.

Joint ensembles

- EEUU54 54 member combination of the 27 members of UU plus the corresponding 27 members of EE - assesses potential benefit of MMAE achievable without an increase in ensemble size.
- EEUU78 78 member combination of the 27 members of UU plus all 51 members from EE - assesses total impact.

Forecasts of 4 fields are verified: 500 hPa height, 850hPa temperature, Pressure at Mean Sea Level (PMSL) and 24 hour accumulations of precipitation. (Only 15 forecasts of 500 hPa height are available). Ensembles are initialised at 1200 UTC, and the four fields are verified every 24 hours. Three fields (500 hPa height, 850hPa temperature, PMSL) are verified on the UKMO model grid (1.25°x0.83°) while 850 hPa temperature is verified on 1/4 resolution (5.00°x3.00°). All of this information is summarised in Table 1.

Table 1. Details of the forecast fields.

Field	Number of forecasts	Resolution of forecast fields (long/lat, degrees)
500 hPa height	15	1.25/0.83
PMSL	23	1.25/0.83
850 hPa temperature	23	5.00/3.00
24 hour accumulations of precipitation	23	1.25/0.83

Three different types of verifying analysis are available; ECMWF analysis, UKMO analysis and the consensus - the average of the ECMWF and UKMO analysis. Due to lack of computer resources not all 5 configurations are verified against all three analyses (Table 2). Four different areas are examined: Europe, North Atlantic and Europe, Northern Hemisphere extratropics and Southern Hemisphere extratropics; results in this paper concentrate on the Northern Hemisphere extratropics. A wide variety of diagnostic measures are used to assess the forecasts - deterministic and probabilistic verification, spread, coverage of observations and spread/skill correlations - only a selection of the results are presented here. Relative forecasting performance can also be assessed in terms of percentage improvement over some standard of reference using the Skill Score (SS) defined in Stanski *et al.* (1989),

$$SS = 100 * \left(\frac{S_f - S_{st}}{S_p - S_{st}} \right) \quad (1)$$

where S_f is the numerical value of performance measure for the system of interest, S_{st} is the corresponding value achieved by the standard forecast and S_p is the score for a perfect forecast. This measure is used throughout the paper with the EPS (EE) as the standard forecast in order to indicate the percentage improvement in skill relative to the ECMWF ensemble.

Table 2. Details of the ensemble configurations that are automatically verified.

Configuration	Verifying analysis		
	ECMWF	UKMO	Consensus
EE	Y	Y	Y
UU	Y	Y	Y
EE27	Y	-	-
EEUU54	Y	Y	Y
EEUU78	Y	-	-

In previous work on this topic (Harrison *et al.* 1995; Evans *et al.* 1998b) the verification has been performed on a coarse $10^\circ \times 5^\circ$ grid, in contrast this study uses $1.25^\circ \times 0.83^\circ$ grid for most of the fields. This high resolution verification means that choice of verifying analysis has a large impact on the results. The UKMO system benefits from being verified against the UKMO operational analyses, and similarly the EPS benefits from being verified against the ECMWF operational analyses. Hence, the most rigorous test for the improvements achievable by the joint ensemble over the EPS is verification against the operational ECMWF analyses,

and as this is also the largest set of results (Table 2), these are presented in detail. For completeness and comparison some results of verification against UKMO analyses are also included here. The verifying analyses for 24 hour accumulation of precipitation are constructed from six hour accumulation forecasts initialised at 00, 06, 12 and 18Z.

3. Results

3.1 Improvements to deterministic forecasts

In the following section the ensemble mean (EM) forecast for the Northern Hemisphere extratropics produced by each configuration is verified using Anomaly Correlation (AC). ECMWF's analyses are used for verification unless stated otherwise and the normals are taken from 15 years worth of ECMWF's Reanalysis data.

At all times beyond T+48 the joint ensemble mean of EEUU78 has higher AC than either individual system for all 4 fields examined (Fig. 1). Calculation of skill scores with EE as standard reference (Eqn 1) quantifies the clear improvements in AC achieved by the 78 member joint ensemble, with typical improvement of around 5% by T+60, and up to almost 10% later in the forecast (Fig. 2). The UU ensembles are disadvantaged by verifying against ECMWF analyses and hence the joint ensemble performance is affected similarly although to a lesser extent. This can be appreciated by comparing Figures 2 and 3; where Figure 3 shows the percentage improvement over EPS ensemble mean AC verified against UKMO analyses (note the different scales). When verified against UKMO analyses the UU and joint ensemble means have substantially higher AC than EE for the first 2 or 3 days. For 850 hPa temperature and precipitation the relative performance of the UU ensemble is highly dependent on the choice of verifying analysis throughout the forecast. But in general the results for the joint ensembles are independent of the choice of verifying analysis after around Day 3.

After around Day 4 the performance of the EEUU54 ensemble is, for practical purposes, equivalent to that of EEUU78 for this diagnostic (Fig. 2). The larger joint ensemble is less disadvantaged in the early stages by testing against ECMWF analyses as the EEUU54 ensemble is 1/2 UU members whereas only 1/3 of the EEUU78 ensemble are from UU. The largest benefits are achieved for 850 hPa height, with average improvement over EE after T+72 of nearly 8%; for 500 hPa height and precipitation the average improvement is over 5%; with improvements of 4% for PMSL. This improvement achieved by the EEUU54 ensemble is equivalent to a gain in lead time of around 12 hours.

The performance of the UU ensembles is generally worse than that of EE or EE27 until the very end of the forecast ranges, and this cannot be blamed solely on the choice of verifying analysis. This is not surprising as the UKMO operational forecasts of 500 hPa height have been poorer - that is Root Mean Square Errors have been higher - than the equivalent ECMWF operational forecast over the period of these forecasts, (mid-October to mid-November) (Steve Lorrimer, personal communication). However, the relative performance of the two operational forecasts varies, and in late November and early December the UM model errors are similar to or lower than ECMWF operational errors. Hence as the sample size increases the relative performance of the UU ensembles may improve. In addition, examination of the AC for individual forecasts (not shown) indicates that the UU ensemble mean does achieve higher AC values than EE for some cases i.e. the EE ensemble does not consistently outperform UU.

3.2 Ensemble dispersion

Ensembles are designed to estimate the sensitivity of predictions to errors in the forecast process and hence it is important that the spread is sufficient to cover all uncertainties in the forecast process. In previous studies (Harrison *et al.* 1995; Evans *et al.* 1998b) joint ensembles encompassed larger spread than either individual ensembles and this spread was shown to be beneficial. Three aspects of ensemble spread are considered here: magnitude of spread; coverage of observations; relationship between spread and skill.

3.2.1 Magnitude of spread

Spread is defined for this study as average AC between the ensemble members and the ensemble mean. (Note that low values of AC correspond to high spread.) For all 4 fields the AC spread of the EEUU78 ensemble is larger than that of both single-system ensembles throughout the forecast (Fig. 4). For 500 hPa height, 850 hPa temperature and PMSL the increase in spread is only marginally dependent on the increased number of members - compare EEUU78 and EEUU54 on Figure 4. For precipitation the increase in spread achieved by the EEUU78 ensembles is around twice that achieved by the EEUU54 ensemble - this may be due to the heterogenous nature of the precipitation field. These increases in spread can be seen more clearly when results are presented as percentage increase in spread over the EPS (Fig. 5). Generally the spread of the MMAE ensemble EEUU54, relative to that of the ECMWF system alone grows with forecast time, reaching around 2% by T+144, less for precipitation.

3.2.2 Coverage of observations

Spread can be increased by simply increasing the magnitude of the initial perturbations but this would not improve the skill or value of the forecast. Hence it is important to assess if the increase in spread gained by combining ensembles is beneficial. One measure of benefit relates to the idea 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. Consistency diagrams provide a clear graphical representation of the distribution of observations relative to the ensemble members (Lanzinger and Strauss 1995). At each grid point the forecast values from all members can be ordered to define a number of equiprobable 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 example Consistency diagram for 850 hPa temperature forecasts over Northern Hemisphere at T+144 (Fig. 6) illustrates the relatively flatter and therefore closer to ideal distribution achieved by the joint ensemble. Performance over the whole distribution can be assessed using the Root Mean Square differences from the ideal expected frequency (horizontal line on Figure 6). For all 4 fields the EEUU54 ensembles fit the ideal distribution more closely than either individual system; the improvement over EE is on average over 30% for the height, temperature and pressure fields, slightly less for precipitation (Fig. 7).

3.2.3 Spread/skill

Another benefit of the combined distribution highlighted in the previous study (Evans *et al.* 1998b) is improved spread/skill correlations. Skill is defined here as AC between the ensemble mean and verifying ECMWF analysis and spread is defined as above. Spread/skill has also been examined against UKMO analyses with results broadly similar to those described below. The results for this measure vary substantially between the different fields (Fig. 8), (this may be due to the small sample size):

PMSL

For PMSL the joint ensembles provides substantial improvements in spread/skill correlation compared with EE, with skill scores relative to EE reaching over 60% at T+132 (Figs 8c and 9c). This improvement means that the EEUU54 correlation reaches potentially useful levels. 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 predicting day-to-day variability in forecast accuracy. The 54 member joint ensemble achieves correlations of over 0.4 for all times beyond T+72 and is above 0.6 after T+96; whereas the EE ensemble only reaches 0.5 after T+196.

24 hour accumulations of precipitation (Figs 8d and 9d)

For this field the joint ensemble again provides improved spread/skill correlations over both single-systems although the percentage improvement over EE is less than that achieved with PMSL - peaking around 50% at T+156. Note that the joint ensembles are the only configurations with spread/skill correlations above 0.5 for this field.

850 hPa temperature (Figs 8b and 9b)

As with PMSL the spread/skill correlation of the EEU54 joint ensembles for 850 hPa temperature is generally an improvement over that of EE - at least 20% improvement beyond T+72 - and over 30% on average. For this field the performance of the joint ensemble is broadly similar to that of the UKMO ensembles, whereas for the previous two fields the joint ensemble outperforms both individual systems.

500 hPa height (Figs 8a and 9a)

Up to T+104 the joint ensembles provide a clear improvement of at least 10% over the ECMWF ensemble. However after T+104 there is no clear difference between the joint ensemble and EE, but both have correlations of above 0.4 at most times. Note that beyond T+104 the spread/skill correlation of the UU ensemble are lower than those of EE.

3.3 Probabilistic measures of skill

The benefit of combining ensemble systems for deterministic forecasts was demonstrated in section 3.1 but one of the primary motivations for producing ensembles is to produce a forecast probability density function (PDF). In this section the benefits of MMAE ensembles for probabilistic forecasts is examined. The event studied for each of the 4 fields is the probability of above-normal values, with normals taken from 15 years worth of ECMWF Reanalysis data. 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 are presented as percentage improvement on the EE score. Verification of probability forecasts against ECMWF, UKMO and the Consensus analysis give largely similar results beyond around Day 3, hence in this note just results against ECMWF's analysis are presented.

3.3.1 Brier Scores

Brier Score (Brier 1950) can be thought of as the mean square error of the probabilistic forecast. For all 4 fields the joint ensemble EEUU78 are equal to or better than the EPS Score after the first 24 hours (Fig. 10); beyond around T+72 the EEUU54 also achieves better Scores than the EPS. Until around Day 6 there is a small advantage in using a larger number of members as the Scores for EEUU78 are greater than those of EEUU54 for the height, temperature and pressure fields. For precipitation the larger ensemble achieves higher Brier Scores than EEUU54 until the very end of the forecast.

Generally the improvement over the EPS achieved by EEUU54 increases with forecast time, reaching 5% before T+136 for PMSL, 850 hPa temperature and 500 hPa height. For precipitation the improvement peaks just below 5% at T+144. Extrapolation of the actual Brier scores (not shown) suggests that this improvement in Brier scores equates to a gain in predictability of approximately 12 hours at Day 5. The Score of the UU ensemble is generally worse than that of the EE or EE27 ensembles although the difference decreases with time. However, examination of the Scores for individual forecasts (not shown) reveals that the individual system with the best Score does vary from case to case.

Murphy (1973) decomposed the Brier Score into the sum of three components: reliability, resolution and uncertainty. 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 forecast system. Reliability and resolution can be traded by altering the nature of the probability forecast. For example the reliability of a forecast system can be improved by simply improving the model climate - moving the forecast distribution towards the observed climatological distribution - but this will lead to a loss in resolution. Hence improvements in both reliability **and** resolution are required to indicate real increases in forecast quality. The reliability and resolution scores are included here to determine the source of improvement to the Brier Score provided by the MMMA ensembles.

3.3.2 Reliability

After the first few days the joint ensemble EEUU54 is substantially more reliable than either individual model for all 4 fields (Fig. 11). The improvement over EE generally increases with forecast time upto around T+120, reaching at least 50% by T+104, with maximum improvements over 70%. In fact for PMSL the improvement is up to 80%. Again there is little difference in the performance of the EEUU54 and EEUU78 ensembles for the height, temperature and pressure fields, but the larger ensemble does have some advantage for

precipitation where improvement in reliability achieved by the EEUU78 is on average 5% larger than that achieved by EEUU54.

3.3.3 Resolution

The improvements in reliability are accompanied by improvements in resolution. For 500 hPa height and 850 hPa temperature the EEUU54 ensembles achieve maximum improvements of nearly 20%; for precipitation and PMSL the maximum is over 10% (Fig. 12). As with reliability, increasing the number of members improves the resolution for precipitation but does not make a substantial impact on the other fields beyond the first few days.

The above results show that improvements in Brier Scores achieved by the EEUU54 ensembles result from improvements in both reliability and resolution. This improvement cannot be explained by simple improvements in spread producing drift towards climatological distributions as, although that would improve reliability, it would be at the expense of resolution.

3.3.4 Relative Operating Characteristics

The Relative Operating Characteristic (ROC) curve, taken from signal detection theory (Swets and Pickett 1982), 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). Briefly, 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 13 is an example curve). Ideally the hit rate will always exceed the false alarm rate and the curve will lie in the upper left hand portion 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 areas under the curve are used to create a skill score with the EE score as standard.

This measure is similar to resolution as it assesses how well the system can discriminate between occurrences and non-occurrences of an event. As ROC is based on stratification by observation it provides no information about reliability, and hence the curves cannot be improved by improving the climatology of the system.

For all four fields the EEUU78 ensemble curves enclose a larger area than both individual ensembles for all times (Fig. 14). Halving the number of members does reduce the

improvement particularly in the first half of the forecast, but the EEUU54 ensembles still achieve improvements over the EPS. In fact the average improvement over EE achieved by the EEUU54 ensembles after T+72 is over 7% for the height, temperature and pressure fields; with average improvement of 4.8% for precipitation.

4. Discussion

One simple way of incorporating sensitivity to model formulation into the EPS is to add members run from different NWP models. Independent models are likely to have substantial differences in their dynamics and parametrisations and hence will have different attractors. Therefore adding members from a different model can add solutions from a separate, valid attractor and so contribute extra, skilful information to the EPS. Earlier studies using older versions of UKMO and ECMWF models (Harrison *et al.* 1998; Evans *et al.* 1998b) found that the UKMO system can generate valid solutions not covered by the ECMWF ensemble. It also follows that if the model attractors are sufficiently separate then no matter what perturbations are added to the EPS it will not be able to cover all the solutions contained on the attractor of the second model. The latest version of the EPS used in this study includes stochastic perturbations of physical tendencies - these were added to address the lack of model dependencies in the EPS - but these are random perturbations and do not alter the attractor of ECMWF's model. Hence the UKMO system can still contribute useful information to the EPS. In this study with up-to-date versions of the UKMO and ECMWF systems the combined ensemble produces improvements in probabilistic scores which equate to a gain in lead time of around 12 hours without an increase in ensemble size. Further examination of the synoptic information contained in the joint ensembles is planned. The benefits of MMAE have never been examined at such high resolution before and so in this study there is a strong dependency on the source of the verifying analysis. When verified against UKMO analyses the benefits of the joint ensemble and corresponding gain in lead time are increased, particularly at early forecast times.

Joining the two attractors increases the spread of the ensemble without loss of skill. Evans *et al.* (1998b) found average increases in spread of over 10%; in this study the increase in spread is less, up to 2%. In the previous study the UKMO ensembles had marginally larger spread - around 2% - than the ECMWF ensembles of comparable size. In this new study the opposite result is true; UU ensembles have smaller spread than the EE27 ensembles. The apparent increase in spread in the EPS relative to the UKMO ensembles may be due to the stochastic perturbations which were recently introduced to the EPS ensemble and found to increase spread. Tests have been performed on adding similar perturbations to UKMO

forecasts (Evans *et al.* 1998a); further tests are planned with a view to possibly incorporating them into the UM ensembles.

Even though the increase in spread gained by joining the ensemble is smaller than that found in previous studies it is important to note that the improvements in coverage of observations are still substantial. In addition, the improvements in coverage are coupled with improvements in both reliability and resolution. Hence the improvements are not simply due to a more accurate model climate but real improvements in accuracy. Previous studies concentrated on selected cases, in fact some of the 9 cases used in Evans *et al.* 1998b were chosen specifically as they were examples of forecasts where the ECMWF system had performed poorly. The results from the new quasi-operational system are significant as it is clear that joint ensembles can provide benefits on a regular basis for ordinary forecasts.

Further improvements in the joint ensembles may be possible through fine tuning of the UKMO model. There are some differences between the version of the model currently being used on the Fujitsu and the operational version - notably the resolution. It is possible that tuning of the Fujitsu version for its specific resolution may improve forecasts from the UKMO component and hence improve the joint ensembles.

For the verification methods examined in this note much of the improvement achieved by the joint ensemble is not dependent on increasing the size of the ensemble and so could be achieved without increase in computer costs. For precipitation there is more evidence that increasing the size of the ensemble can provide additional benefits and this is likely to be due to the heterogeneous nature of the field.

This study is based on 23 cases - one of the largest studies of its kind - although the cases are from consecutive forecasts and so are not entirely independent. The sample size is increasing daily and verification is performed by an automatic system, so a definitive study of a whole season's worth of data will soon be available. Further studies into individual cases will also be possible to fully investigate the benefits of MMAE for forecasts of extreme events. In addition, more experiments are planned to examine the relative importance of model and analysis dependencies.

5. Conclusion

The UKMO model system is now run in ensemble mode quasi-operationally at ECMWF. Combination with ECMWF's Ensemble Prediction System (EPS) (which is based on their model alone) will enable the formation and assessment of a large sample of multi-model and multi-analyses ensembles (MMAE). Preliminary results from the first 23 cases have been described in this report. The MMAE provides improvements over the EPS for all diagnostics so far examined and for all four fields looked at so far (Table 3). These improvements approximately equate to a gain in predictability of around 12 hours and can be achieved with no increase in ensemble size, so could be gained with minimal increase in computing costs.

6. Future Plans

- Continue to accumulate cases.
- Extend verification to more fields.
- Extend range of diagnostics - examine synoptic content of the joint and individual ensembles.
- Assess the improvements in value that customers would gain from using joint ensembles.
- Perform additional experiments to enable examination of the relative importance of model and analysis dependencies.

7. Acknowledgments

The authors would like to thank Richard Barnes for setting up and maintaining the UM ensemble system at ECMWF. Kelvyn Robertson and Anette Van Der Wal are also acknowledged for developing the experiments and organising transfer of analyses.

Table 3. Percentage improvement over EPS achieved by multi-model and multi-analysis ensembles over Northern Hemisphere Extratropics from T+72-T+240; verified against ECMWF operational analyses. Results are presented for both 78 and 54 member joint ensembles where available; (n/a - not available). For each diagnostic the highest improvement achieved by the 54 member joint ensemble is in **bold**.

Diagnostic	% improvement over EPS achieved by MMAE over North Hemisphere Extratropics: T+72-T+240.							
	500 hPa height		850 hPa temperature		PMSL		24 hour accumulation of rainfall	
	54 mems	78 mems	54 mems	78 mems	54 mems	78 mems	54 mems	78 mems
AC of ensemble mean	5.1	6.0	7.8	7.7	4.1	5.3	5.0	5.8
AC spread around ensemble mean	0.9	1.1	1.5	1.5	1.6	1.7	0.7	1.5
AC spread/skill correlation	0.5	13.6	35.3	28.4	36.6	33.6	28.1	26.1
Consistency measure - RMS difference from expected distribution	30.2	n/a	31.2	n/a	36.6	n/a	21.9	n/a
Probability forecasts of above normal values								
Brier Scores	3.8	4.5	5.3	5.9	4.8	5.4	2.8	3.9
Reliability	49.0	46.8	65.1	66.9	68.2	64.5	70.5	75.8
Resolution	7.9	7.9	11.0	10.7	8.5	9.1	7.2	10.2
Relative Operating Characteristic	7.1	8.6	9.6	10.4	8.2	9.5	2.8	5.1

8. References

- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1-3.
- 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.
- Buizza, R., M. Miller and T.N. Palmer, 1998: Stochastic simulation of model uncertainties in the ECMWF Ensemble Prediction System. Submitted to the *Q. J. R. Met. Soc.* August 1998.
- Evans, R.E., M.S.J. Harrison and R.J. Graham, 1998b: Joint Medium Range Ensembles From the UKMO and ECMWF models. Forecasting Research Technical Report No. 243
- Evans, R.E., R.J. Graham, M.S.J. Harrison and G.J. Shutts, 1998a: Preliminary Investigations Into The Effect Of Adding Stochastic Backscatter To The Unified Model. Forecasting Research Technical Report No. 241
- Harrison, M. S. L., T. N. Palmer, D.S. Richardson, R. Buizza and T. Petroligis, 1995: Joint medium range ensembles from UKMO and ECMWF models and analyses. *Seminar on predictability volume II*. ECMWF, 4-8 September 1995, 61-120.
- Harrison, M. S. L., T. N. Palmer, D.S. Richardson and R. Buizza, 1998: Analysis and Model Dependencies in Medium-Range Ensembles: Two Transplant Case Studies. *Provisionally accepted by Quart. J. Roy. Meteor. Soc.*
- Houtekamer, P. L., and L. Lefaiivre, 1996 A system simulation approach to ensemble prediction. *Monthly Weather Review*, **124**, 1225-1242.
- Lanzinger, A., and B. Strauss, 1995 EPS evaluation at ECMWF. *Fifth Workshop on Meteorological Operational Systems*. ECMWF 13-17 November 1995, 87-101.
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petroligis, 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.
- Palmer, T.N., F. Molteni, R. Mureau and R. Buizza, 1993: Ensemble Prediction. ECMWF Seminar proceedings 'Validation of models over Europe: Vol. 1', ECMWF.

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

Swets, J. A. and R. M. Pickett, 1982: Evaluation of diagnostic systems - methods from signal detection theory. Academic Press, 253pp.

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

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

Figure 1.

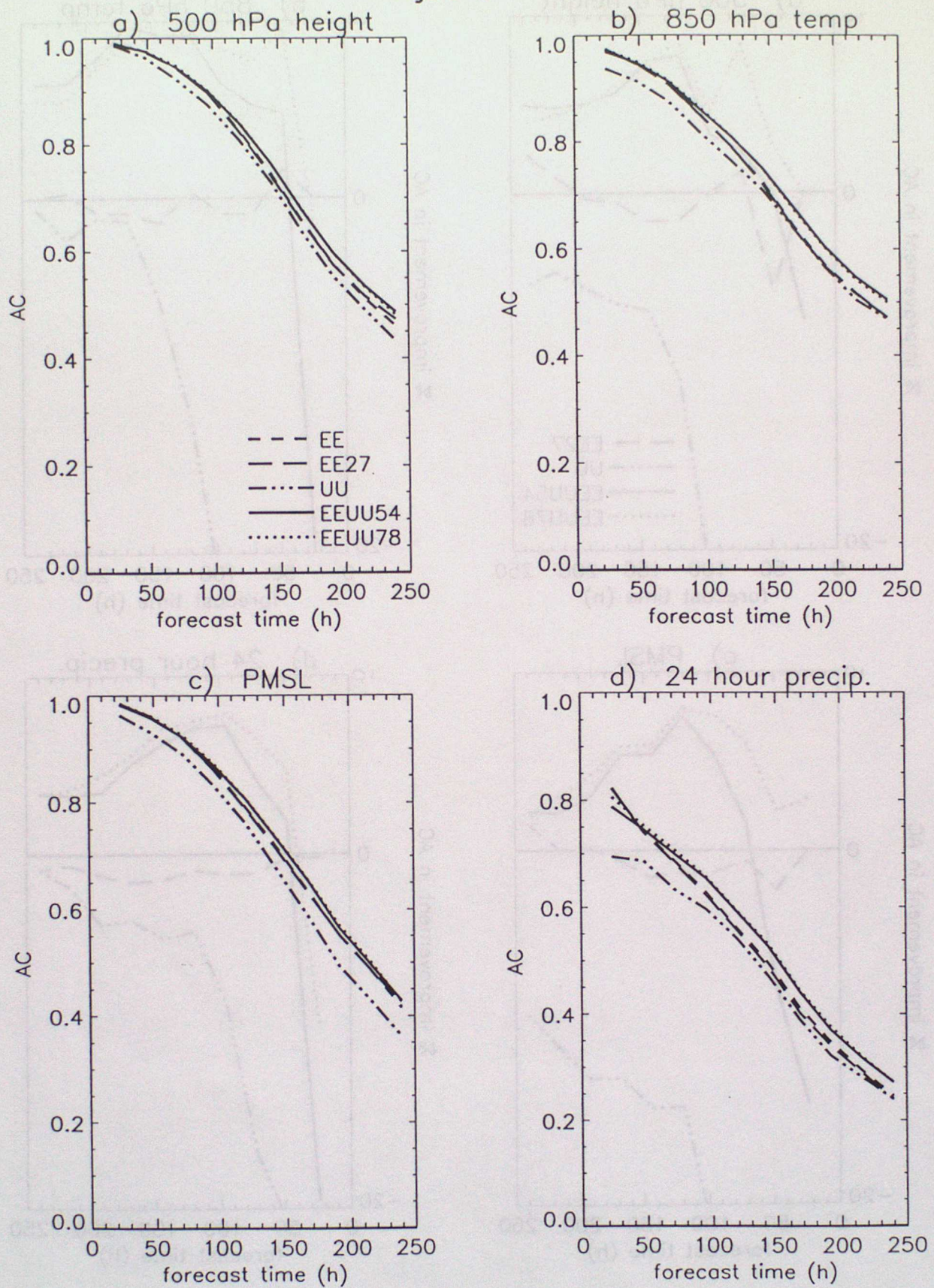


Figure 1. Anomaly Correlation of ensemble mean forecasts over Northern Hemisphere extratropics from T+24 to T+240; verified against ECMWF analyses. Results for 5 configurations are shown: EE, EE27, UU, EEUU54 and EEUU78.

- a) 500 hPa height - average of 15 forecasts.
- b) 850 hPa temperature - average of 23 forecasts.
- c) PMSL - average of 23 forecasts.
- d) 24 hour accumulation of precipitation - average of 23 forecasts.

Figure 2.

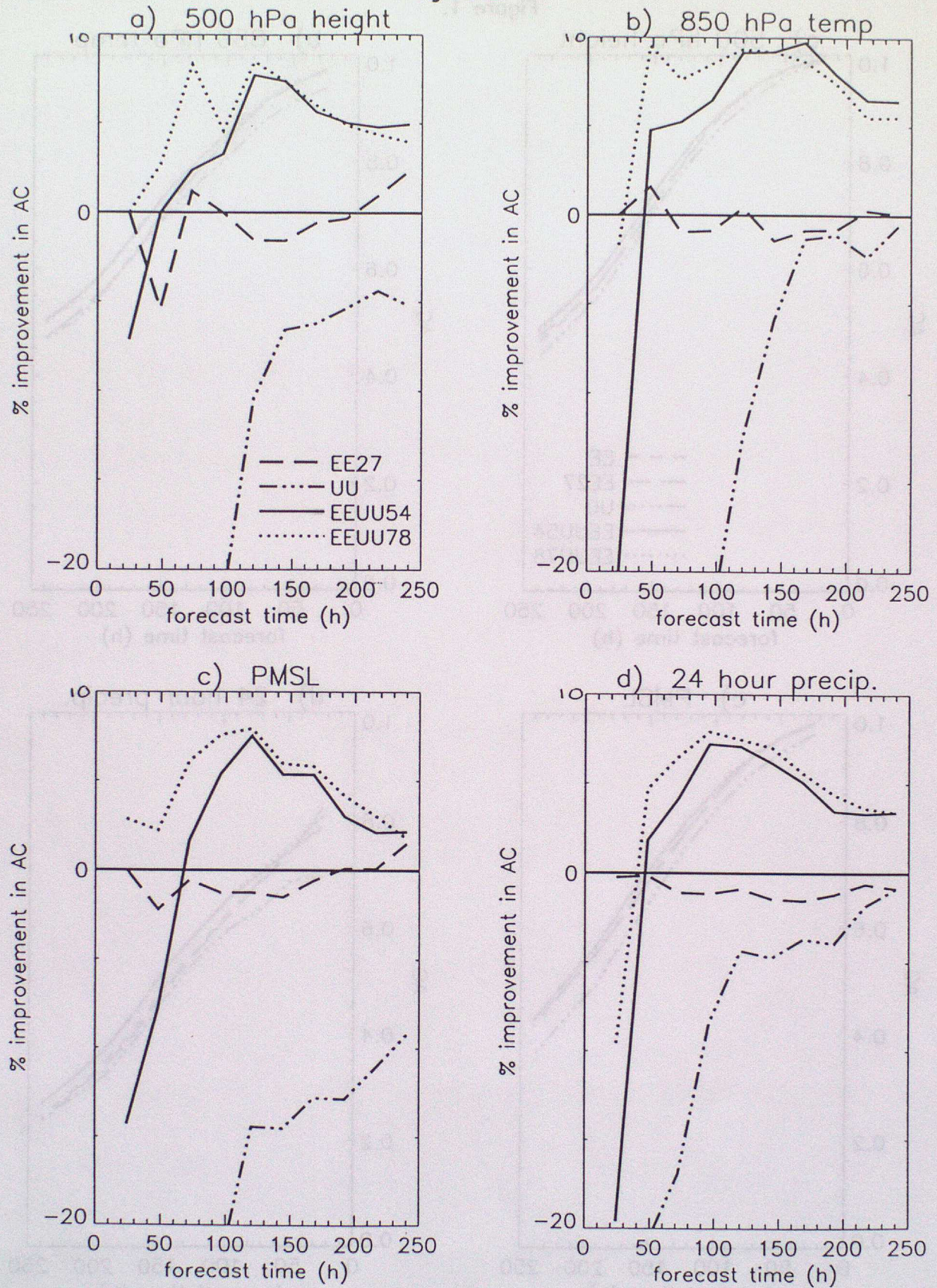


Figure 2. As Figure 1 but results are presented as percentage improvement over EE.

Figure 3.

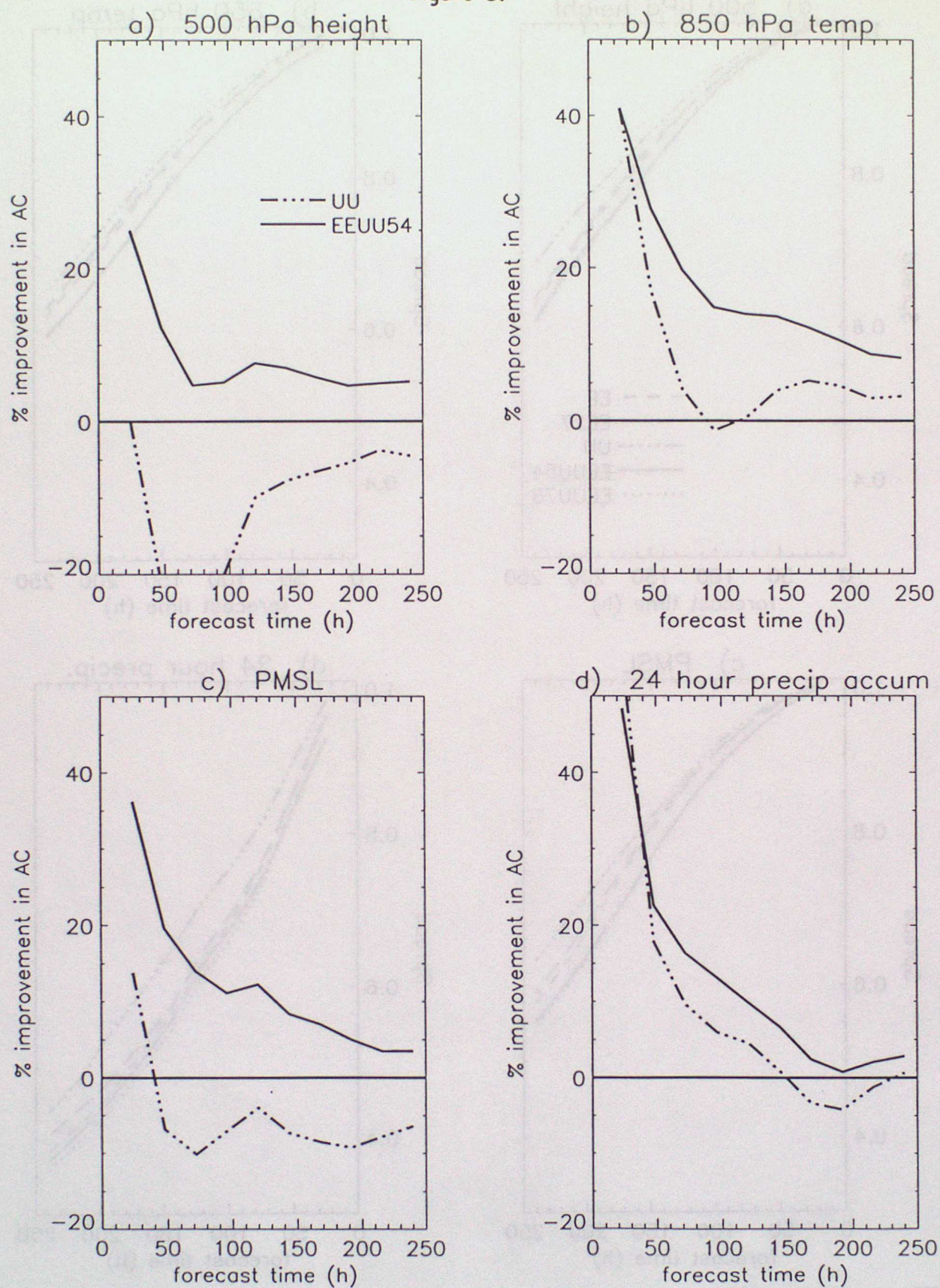


Figure 3. As Figure 2 but using UKMO analysis for verification.

Figure 4.

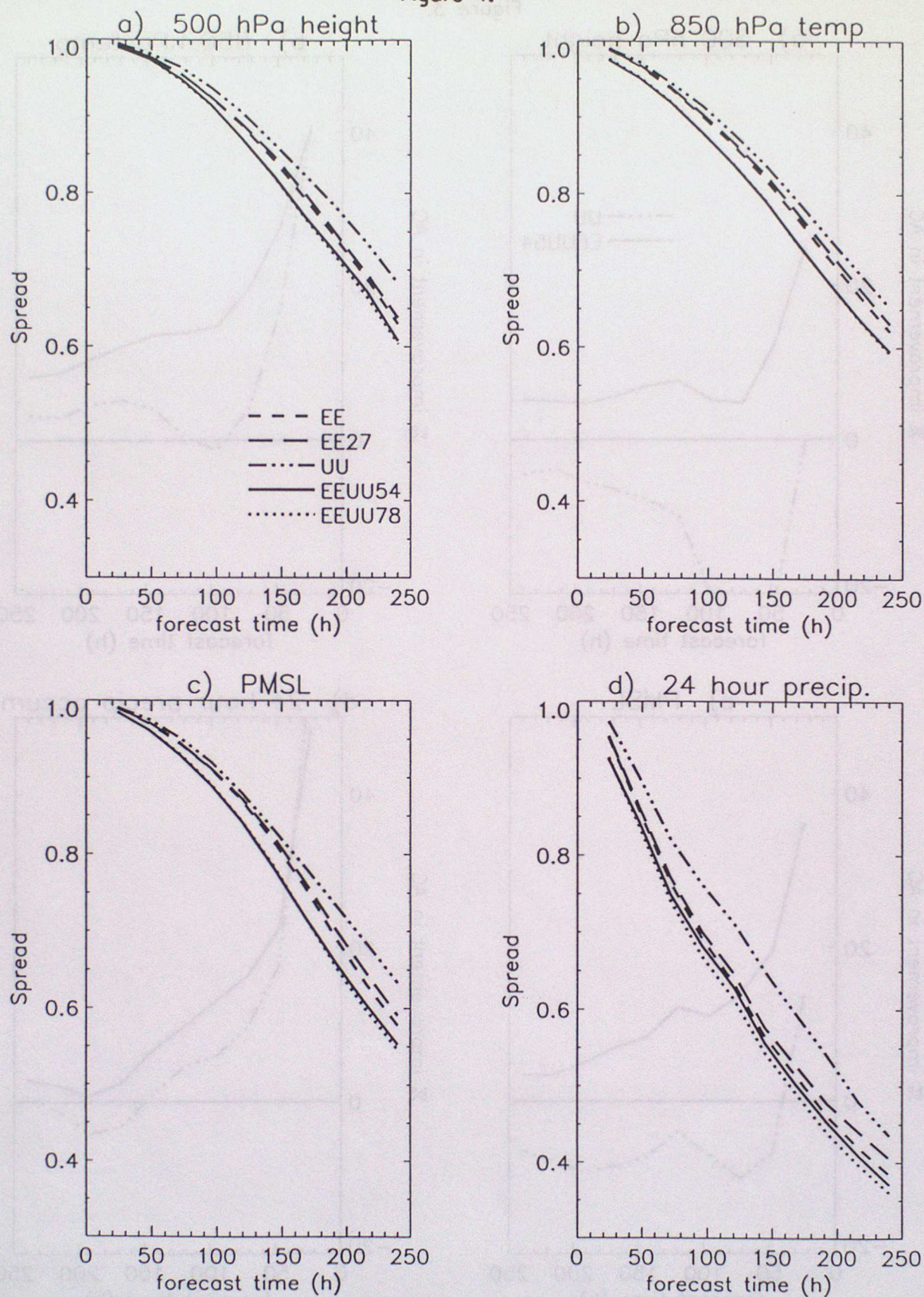


Figure 4. Spread (average anomaly correlation between ensemble members and ensemble mean) over Northern Hemisphere extratropics from T+24 to T+240. Results for 5 configurations are shown: EE, EE27, UU, EEUU54 and EEUU78.

- a) 500 hPa height - average of 15 forecasts.
- b) 850 hPa temperature - average of 23 forecasts.
- c) PMSL - average of 23 forecasts.
- d) 24 hour accumulation of precipitation - average of 23 forecasts.

Figure 5.

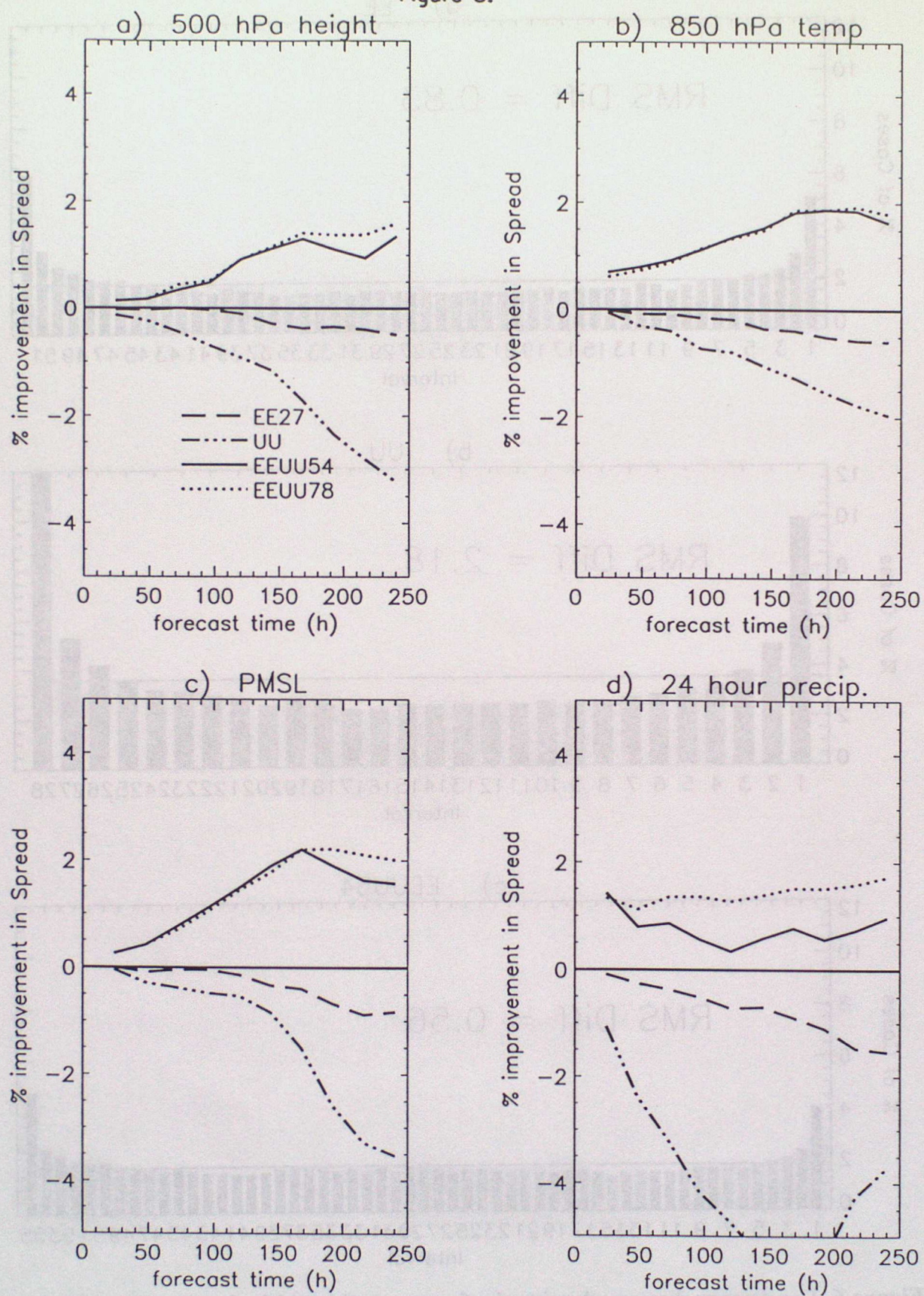
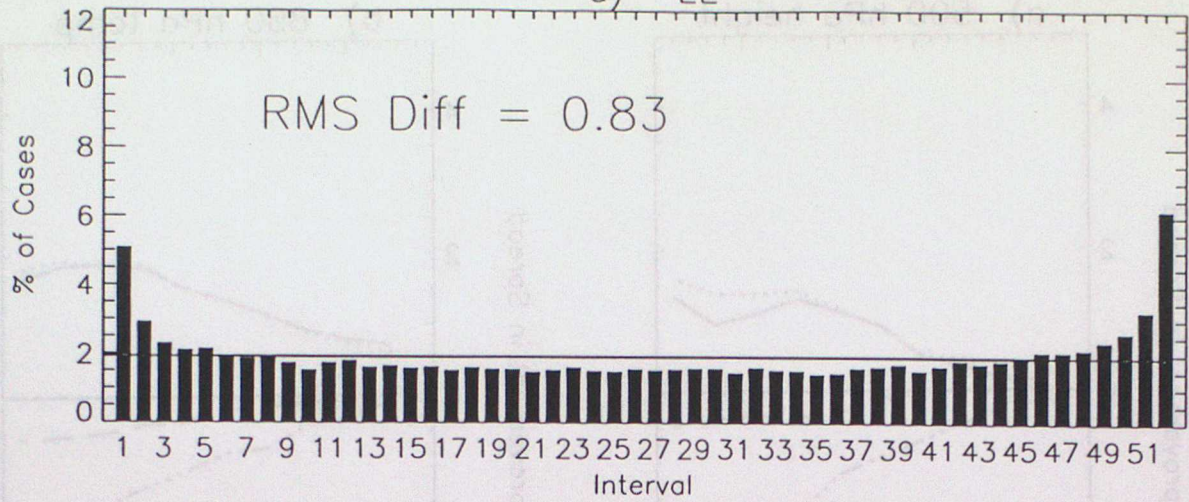
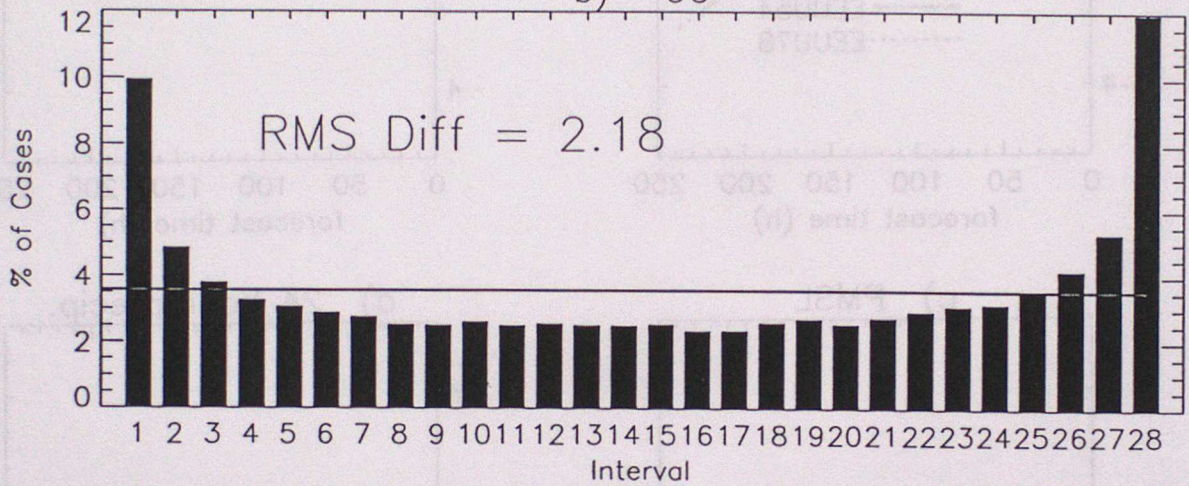


Figure 5. As Figure 4 but results are presented as percentage improvement over EE.

Figure 6.
a) EE



b) UU



c) EEUU54

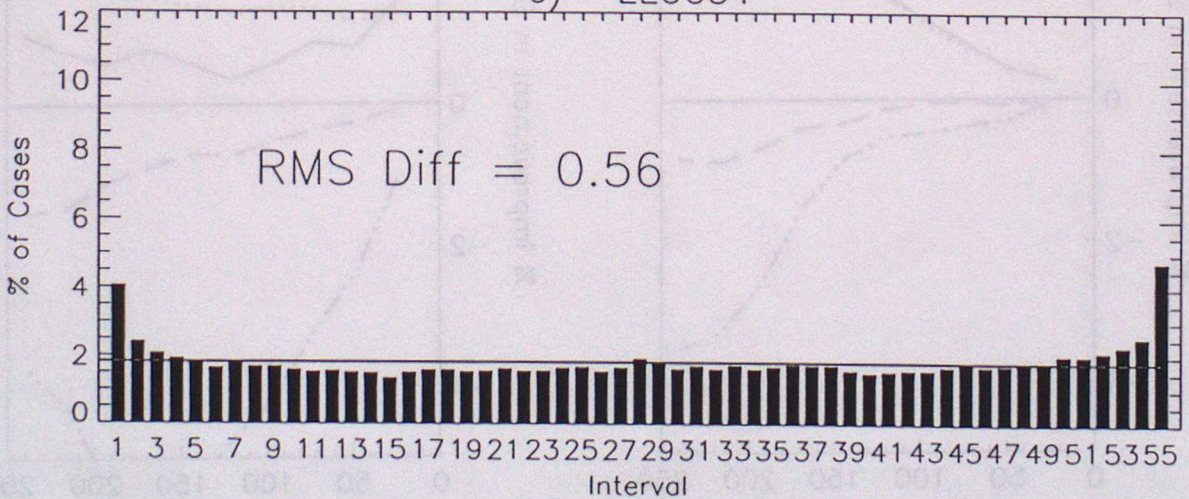


Figure 6. Consistency diagram showing the frequency with which observations of 850 hPa temperature lie within ensemble intervals at T+144 over Northern Hemisphere extratropics. Horizontal lines indicate expected values and the Root Mean Square difference between the actual and expected values are given. Results are for 23 cases and are verified against ECMWF's analyses.

- a) EE
- b) UU
- c) EEUU54

Figure 7.

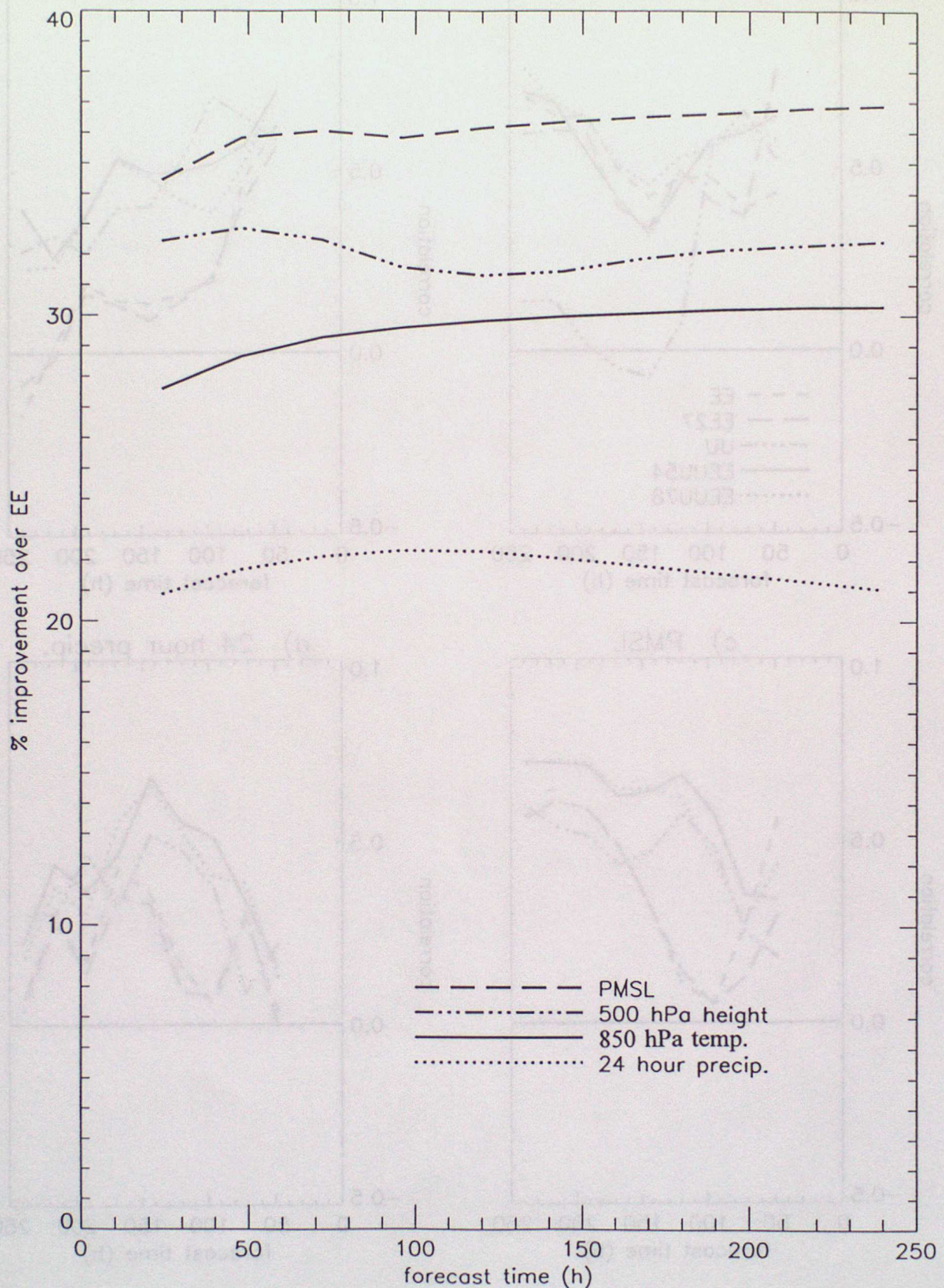


Figure 7. Percentage improvement achieved by EEUU54 over EE for RMS difference between actual and expected distributions on Consistency diagrams calculated over Northern Hemisphere extratropics; verified against ECMWF analyses. Results for 4 different fields are shown:

PMSL - average of 23 forecasts.

500 hPa height - average of 15 forecasts.

850 hPa temperature - average of 23 forecasts.

24 hour accumulation of precipitation - average of 23 forecasts.

Figure 8.

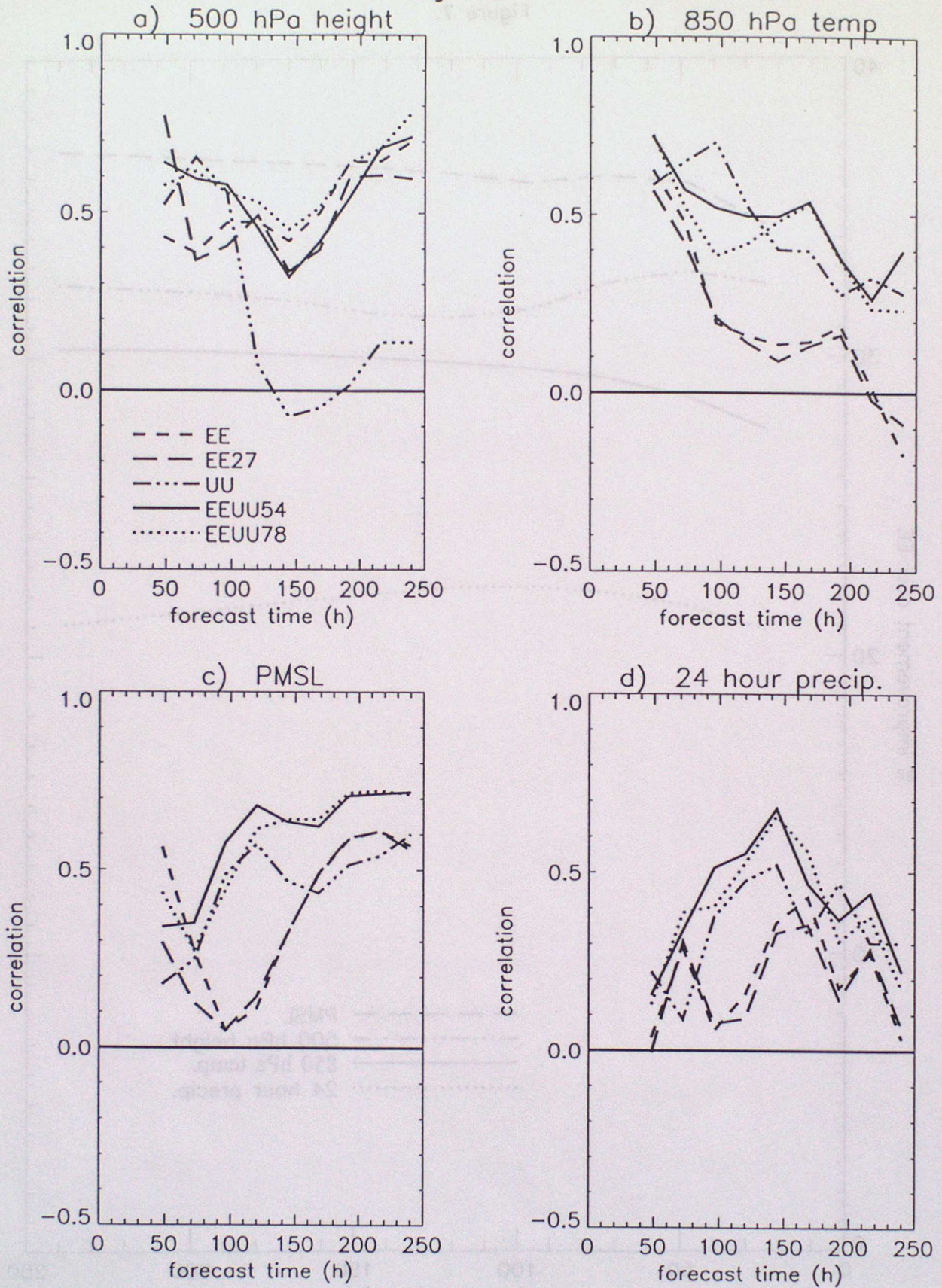


Figure 8. Correlation between spread (average AC between ensemble members and ensemble mean) and skill (AC of ECMWF ensemble mean with ECMWF analysis) the over Northern Hemisphere extratropics from T+48 to T+240. Results for 5 configurations are shown: EE, EE27, UU, EEUU54 and EEUU78.

- a) 500 hPa height - average of 15 forecasts.
- b) 850 hPa temperature - average of 23 forecasts.
- c) PMSL - average of 23 forecasts.
- d) 24 hour accumulation of precipitation - average of 23 forecasts.

Figure 9.

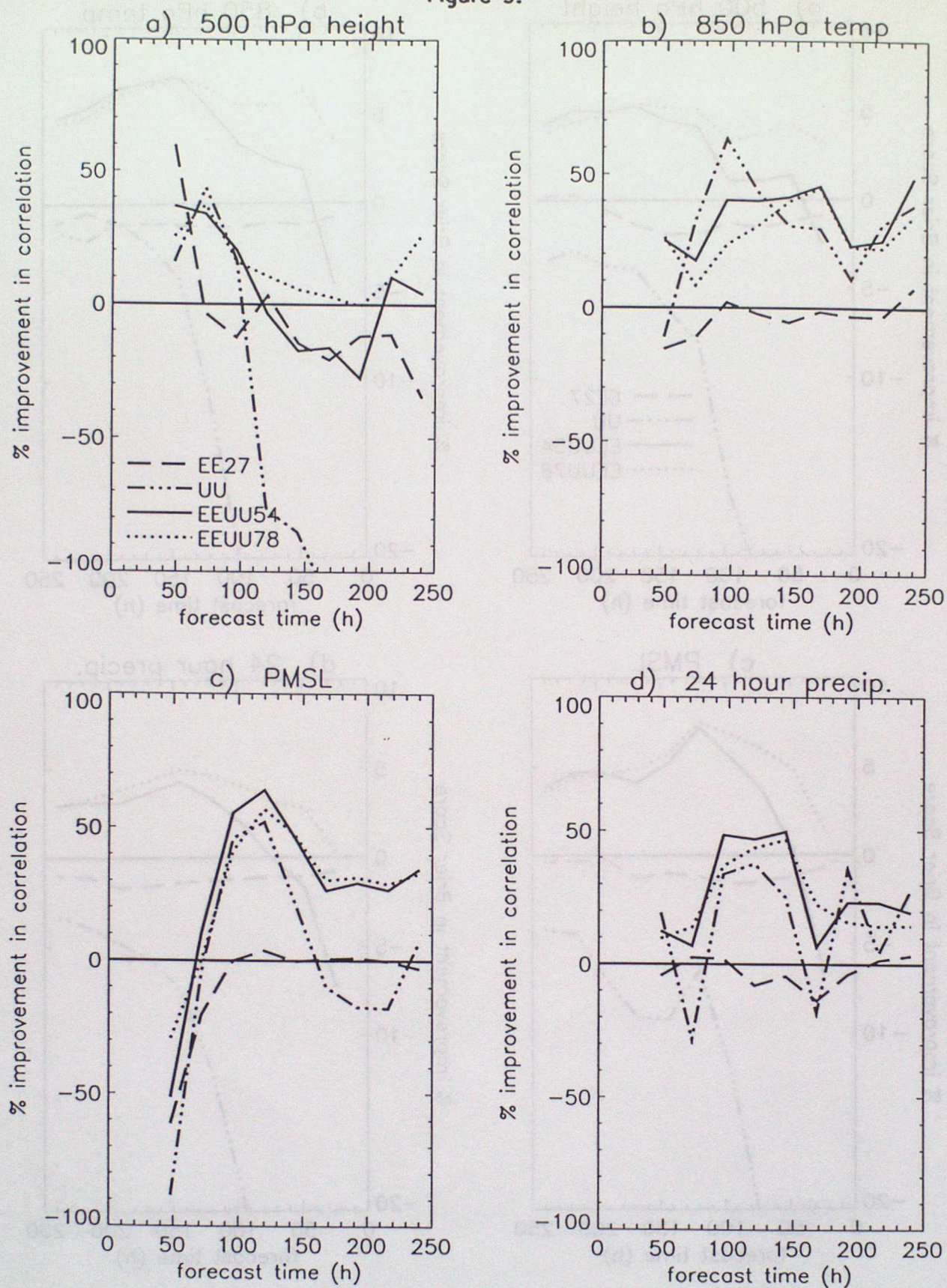


Figure 9. As Figure 8 but results are presented as percentage improvement over EE.

Figure 10.

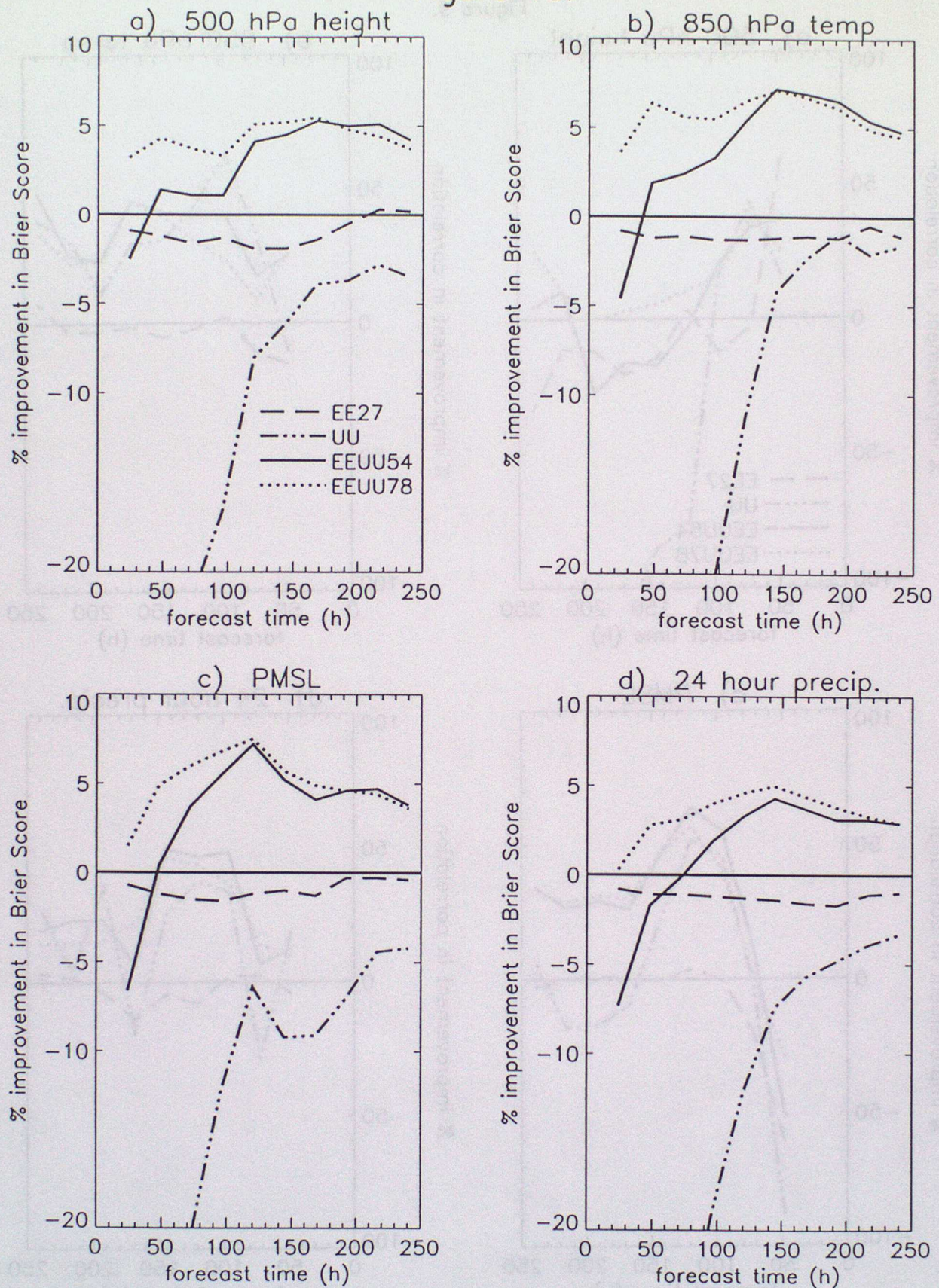


Figure 10. Brier Scores for forecasts exceeding normal over Northern Hemisphere extratropics from T+24 to T+240; verified against ECMWF analyses. Results are presented as percentage improvement over EE. Results for 4 configurations are shown: EE27, UU, EEUU54 and EEUU78.

- a) 500 hPa height - average of 15 forecasts.
- b) 850 hPa temperature - average of 23 forecasts.
- c) PMSL - average of 23 forecasts.
- d) 24 hour accumulation of precipitation - average of 23 forecasts.

Figure 11.

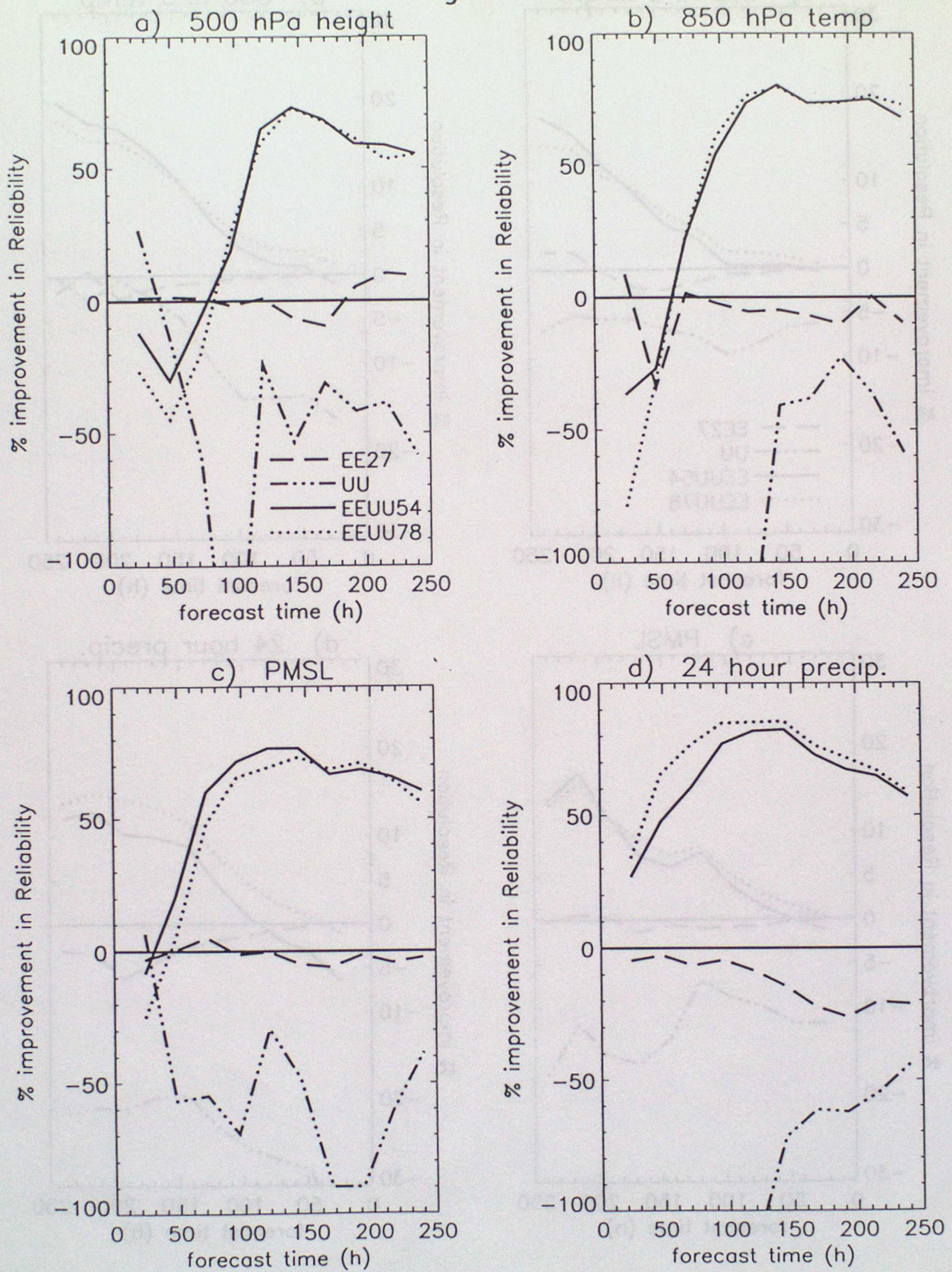


Figure 11. As Figure 10 but for reliability (note change in vertical scales).

Figure 12.

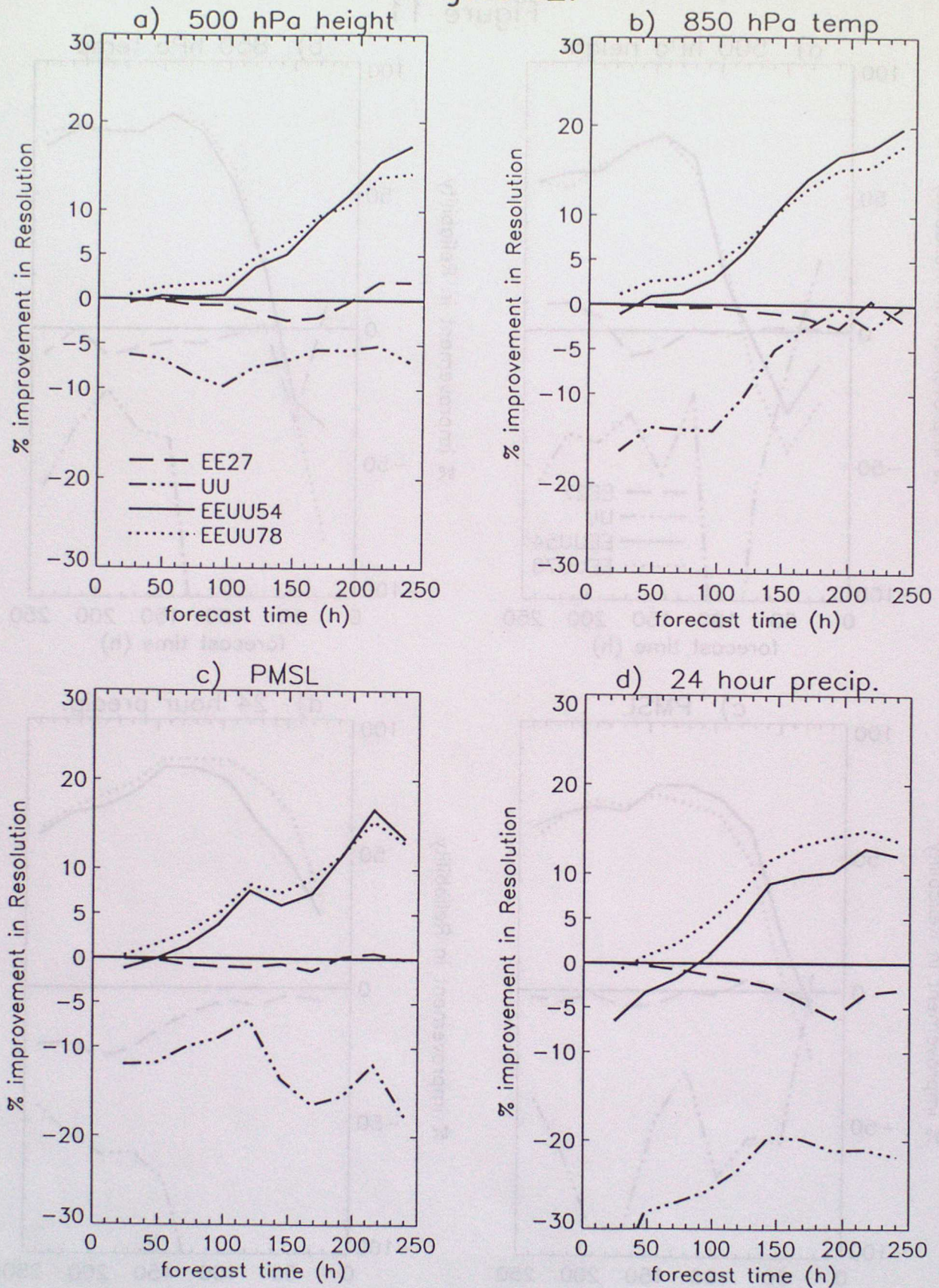


Figure 12. As Figure 10 but for resolution (note change in vertical scales).

Figure 13.

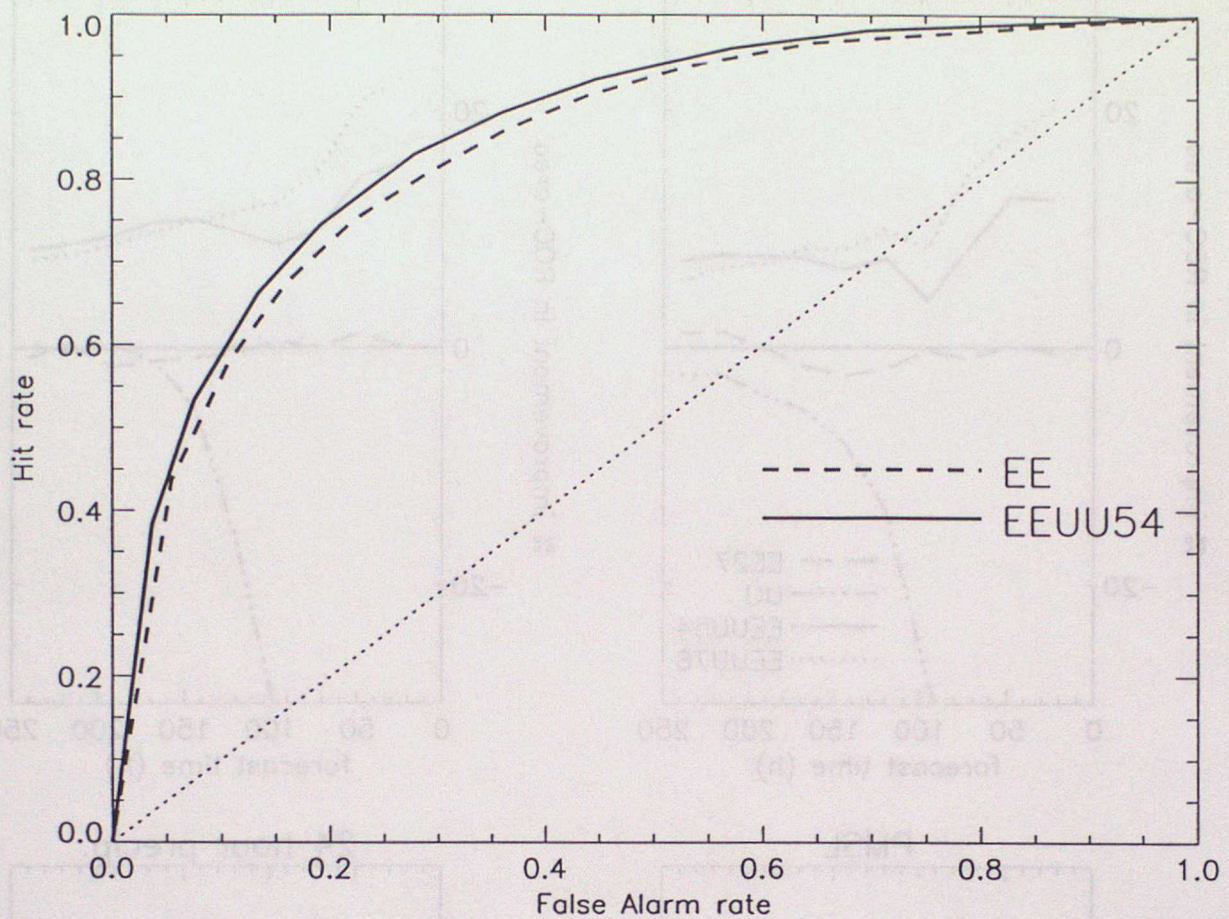


Figure 13. Sample Relative Operating Characteristic (ROC) Curve for T+144 forecasts for 850 hPa temperature exceeding normal over Northern Hemisphere Extratropics. Results from 23 forecasts for EE and EEUU54 ensembles; verified against ECMWF analyses.

Figure 14.

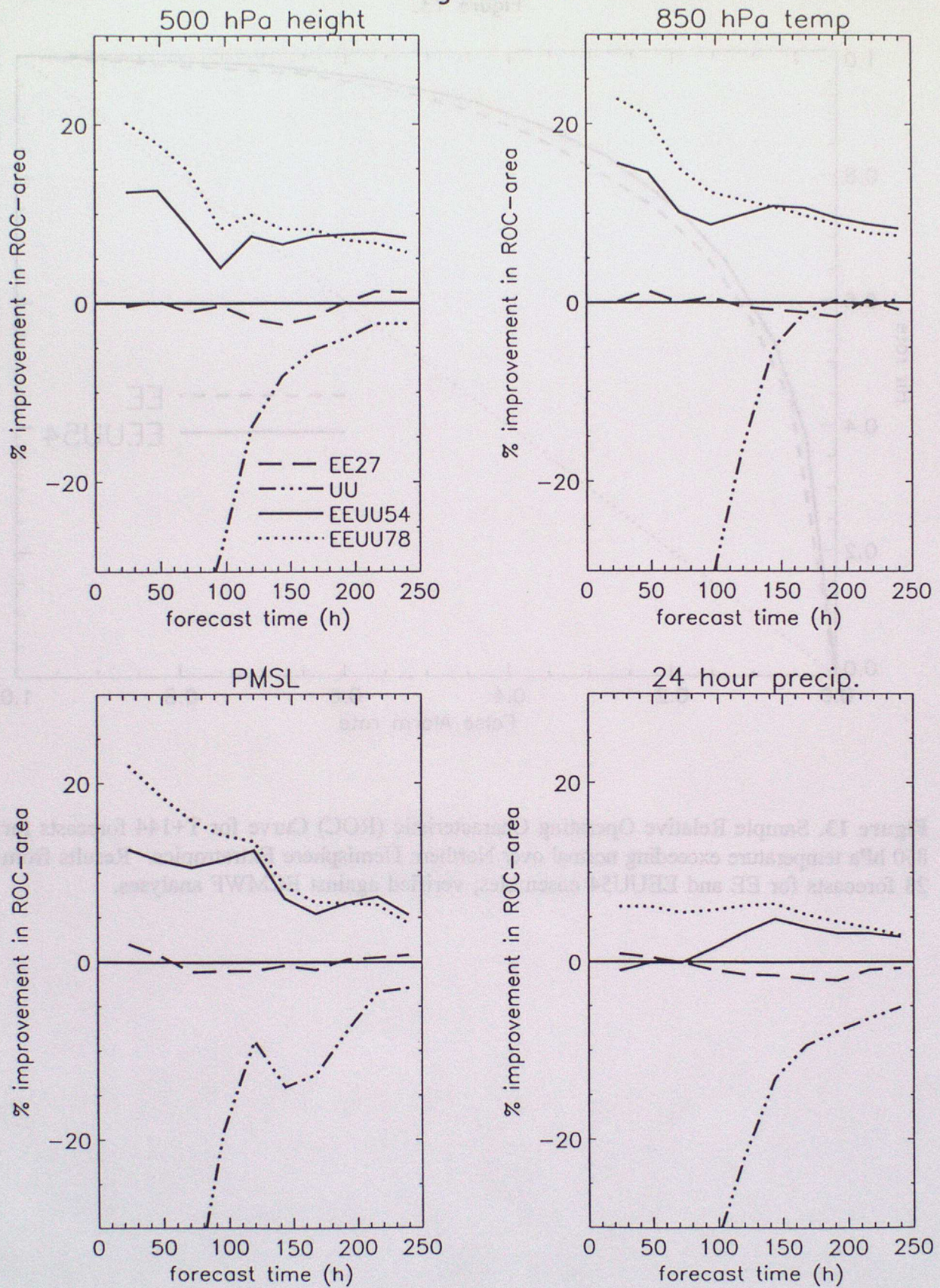


Figure 14. As Figure 10 but for area under Relative Operating Characteristic Curve (note change in vertical scales).