ENSEMBLES USING MULTIPLE MODELS AND ANALYSES

David S. Richardson

ECMWF

1. INTRODUCTION

The operational ECMWF ensemble prediction system (EC EPS) is designed to take account of the effect of analysis uncertainties in the forecast evolution. Until recently, no account was taken in the EPS of the possible effect of model error during the forecast. The introduction of stochastic physics in October 1998 was a first attempt to include random errors associated with the parametrization processes in the model. An alternative approach to the representation of model errors in an EPS is to include, in the same ensemble, forecasts produced using different models. Preliminary studies have indicated that a combined ensemble including members run using both the ECMWF model and the UKMO model produces a more skilful ensemble system than the single-model EC EPS (Harrison et al., 1995; Evans et al., 1999). This work used the 32-member T63 EPS (operational until December 1996; no stochastic physics) and an equivalent resolution version of the UKMO model.

In October 1998 a new set of UKMO ensemble integrations was begun, repeating the analysis using the current 50-member T_L159 EPS. The aim is to investigate whether the earlier results still hold with the larger and higher resolution EC EPS which also has model perturbation in the form of the stochastic physics. The UK ensemble system (UK EPS) uses the UKMO forecast model initialised with the ECMWF EPS perturbations added to UKMO analysis. The ensembles are run every day and consist of the control forecast plus 26 perturbed members. Preliminary results for winter 98-99 indicate that a combined multi-centre ensemble (27 UK members + 27 EC members + both controls; MC EPS) performs better than the 51-member EC EPS (Evans et al., 1998; Mylne et al., 1999).

The combined ensembles differ from the EPS in two ways: initialisation about two different analyses, and the use of both the EC and the UKMO models for the forecasts. It is important to determine whether the benefit of the MC EPS is due to the use of different models or whether it results from the use of multiple analyses in specification of the initial conditions.

To investigate the effect of using only information from other analyses on the performance of the EPS, ensembles have been run using the EC model with initial conditions defined using the operational analyses from different centres. These predictions are compared to those from the EC and MC EPS systems to establish the potential benefits of using multiple analyses in a single-model system and to assess the additional benefit achievable with the use of more than one model within an ensemble system.

The paper is organised as follows. The ensemble configurations and the method of using analyses from other centres to initialise the EC model are explained in Section 2. Results are presented in Section 3. The effect of removing model bias is discussed in Section 4 and the effect of changing the amplitude of the initial perturbations is considered in Section 5. Conclusions are drawn in Section 6.

2. ENSEMBLE CONFIGURATIONS AND ANALYSIS INTERPOLATION

Operational analyses from UKMO, DWD, Météo-France, and NCEP are supplied as fields of u, v, T, and q on 15 standard pressure levels. A version of the analysis from centre X, A_M^X, on the EC model grid is constructed by calculating the difference between the pressure level analyses of X and ECMWF, interpolating this difference to EC model levels, and adding the resulting "perturbation" to the EC model level analysis:

$$A_M^X = A_M^E + (A_P^X - A_P^E)_M \tag{1}$$

where subscripts P and M refer to pressure levels and EC model levels respectively. No perturbation is applied to the surface fields.

A "consensus analysis" A_M^C is constructed as the mean of the available analyses, the averaging being made on pressure levels before the final perturbation is interpolated to model levels:

$$A_M^C = A_M^E + \left[\frac{1}{(N+1)}\sum_{\text{centres}}(A_P^X - A_P^E)\right]_M \tag{2}$$

where N is the number of additional analyses (4 in this study).

Two ensemble configurations using the additional analyses have been evaluated. The first initialises the EPS about the consensus analysis - the control forecast is simply the deterministic forecast from the consensus analysis and 50 ensemble members are constructed by adding the standard EPS perturbations (calculated using the EC analysis) to the consensus analysis. Thus the difference from the operational EPS is just the centroid of the initial pdf; the initial spread is identical to that of the EPS. This first configuration will be referred to as the CONS EPS. In the second configuration, the full ensemble is constructed by combining 5 smaller ensembles of control plus 10 members perturbed about each of the available analyses. This will be referred to as the multi-analysis (MA) ensemble (total of 55 members).

The CONS EPS uses the available analyses to create a "better" analysis (one which should be closer to the true atmospheric state than the contribution individual analyses), which may be expected to produce a better deterministic control forecast, while the MA EPS effectively uses the analysis differences as directions in which to amplify the initial spread. Both configurations result in an initial distribution of states centred around the consensus analysis.

60 cases have been run for both configurations, for dates spaced 4 days apart from October 1998 (when the

UKMO ensembles began) to July 1999. Dates when one or more operational analyses were unavailable or the UKMO ensemble was not complete have not been used, hence the spacing between cases is sometimes greater than 4 days. Throughout the period the EC EPS comprised 51 members run at T_L159L31 with stochastic physics.

For this set of cases the performance of 5 ensemble configurations has been compared. The configurations are summarised in Table 1.

TABLE 1 EPS CONFIGURATIONS USED IN THE STUDY

Name	Description	No. members	multi-model	multi-analysis
EC EPS	operational EPS	51	N	N
UK EPS	UK model from UK analysis	27	N	N
MC EPS	EC EPS (1+26 members) + UK EPS (1+26 members)	54	Y	Y
CONS EPS	as EC EPS but initialised about consensus analysis	51	N	Y
MA EPS	control+10 members about each analysis (all with EC model)	55	N	Y

3. RESULTS

Results are presented below for 500 hPa height (Z500) and 850 hPa temperature (T850) over the extra-tropical Northern Hemisphere and Europe. Forecasts are verified against the ECMWF analysis - this gives an advantage to the EC EPS for the first 2-3 days; beyond that, the choice of verifying analysis has little qualitative effect on the results.

3.1 Control forecast and ensemble mean skill and ensemble spread

For both Z500 and T850, the unperturbed control forecast from the consensus analysis is slightly more skilful than the control from the ECMWF analysis out to day 7, especially over Europe (Fig. 1; T850 not shown). The UK control forecast is noticeably less skilful than the EC model forecasts.

As with the control forecasts, the skill of the UK ensemble mean is lower than that of the other configurations. For Z500, the MC EPS has slightly higher skill over both the Northern Hemisphere and Europe (Fig. 2) than the other ensembles out to days 7-8, beyond which the MA EPS is better. For T850, the MA and consensus ensemble means are the most skilful throughout the forecast period; the ensemble-mean skill for the MC EPS is comparable to that of the EC EPS.

Ensemble spread is measured as the dispersion of the members about the ensemble mean. The initial spread of the UK, EC and CONS ensembles is the same, since all three configurations use the same set of initial perturbations. However, the spread of the UK EPS increases more slowly during the forecast (Fig. 3). The MA

and MC ensembles have substantially larger initial spread reflecting the inclusion of analysis differences in these configurations. During the first two days the MA and MC spread grows more slowly than that of the other systems. This is to be expected since the analysis differences included in the initial perturbations do not necessarily project strongly onto growing modes, whereas the singular vector perturbations are constructed to produce maximum growth over 48 hours. However, higher spread is maintained throughout the forecast, especially for the MC EPS for T850 (Fig. 3, lower panel). The difference in forecast spread between the MC and MA ensembles, which have similar initial spread, suggests that the use of different models contributes to the increased spread in the MC EPS during the forecast. The effect of removing model systematic error on ensemble spread is discussed in Section 4.

In a correctly formulated EPS, the ensemble spread (about the ensemble mean) should on average be equal to the error of the ensemble mean forecast. The amplitude of the initial perturbations for the ECEPS is chosen so that this equivalence between spread and skill is achieved at day 2 (Buizza et al., 1999). However, beyond day 2 the EC EPS spread is too small relative to the ensemble mean error. This is true for all the ensemble configurations (Fig. 4), although the difference is substantially less for the MC EPS (and to a lesser extent for the MA EPS) than for the other systems. For Z500 over Europe the equivalence between spread and skill is maintained for the MC EPS out to day 5, although spread is actually too large in the shorter range. The effect of increasing the amplitude of the initial perturbations in the CONS EPS is discussed in Section 5.

3.2 Probability forecast performance

The probabilistic performance of the ensembles in predicting Z500 and T850 anomalies exceeding certain thresholds has been assessed using ROC area and Brier Score (Stanski et al., 1989). As for the ensemble-mean scores, the UK EPS is generally poorer than the other configurations. To some extent this is a consequence of the smaller size of the UK EPS. While ensemble spread and ensemble-mean skill are relatively insensitive to ensemble size (Leith, 1974; Buizza and Palmer, 1998) the effect on probability forecasts is more marked (Richardson, 1999). Since the main aim of this study is to investigate the relative merits of the MC ensemble, the UK EPS results are not discussed in this section.

For many thresholds the absolute difference in ROC area or Brier score for the different ensemble configurations is relatively small. To show the improvement of an alternative configuration relative to the operational EC EPS, the results are also presented as skill scores using the EC EPS as the reference forecast:

$$S_K = 100 \times \left(\frac{R_X - R_E}{R_P - R_E} \right) \tag{3}$$

where R_X is the score (ROC area or Brier score) for configuration X, R_E is the score for the EC EPS and R_P is the score for a perfect deterministic forecast (Wilks, 1995). Thus a positive value for S_K indicates an improvement over the EC EPS. To summarise results and compare with the conclusions of the initial

assessment by UKMO (Evans et al., 1999), the mean skill $\overline{S_K}$ over days 3 to 10 is used as an overall measure of performance.

The mean skill, $\overline{S_K}$, is tabulated in Tables 2 to 5 for a number of thresholds for both ROC area and Brier score.

TABLE 2 MEAN SKILL RELATIVE TO OPERATIONAL EPS OVER DAYS 3-10. Z500, N. HEM.

threshold	ROC			Brier score		
	MC	CONS	MA	MC	CONS	MA
25 m	**6.71**	2.52	4.63	**4.87**	1.10	2.43
50 m	**7.01**	3.08	5.45	**4.40**	0.98	2.23
100 m	**8.84**	4.40	7.11	**3.62**	1.06	1.80
0	**6.09**	2.29	4.16	**4.59**	1.27	2.70
-25 m	**5.34**	2.24	4.12	**3.93**	1.65	3.10
-50 m	**5.69**	1.66	3.91	**3.89**	1.32	2.80
-100 m	**6.94**	-0.79	2.96	**3.94**	1.40	2.75

TABLE 3 MEAN SKILL RELATIVE TO OPERATIONAL EPS OVER DAYS 3-10. Z500 EUROPE.

threshold	ROC			Brier score		
	MC	CONS	MA	MC	CONS	MA
25 m	**6.70**	2.55	3.76	**5.09**	0.76	1.65
50 m	**8.01**	2.71	4.41	**6.19**	0.31	1.32
100 m	**10.52**	3.36	5.78	**6.87**	-0.09	0.79
0	**6.03**	2.41	3.86	**3.99**	1.05	1.99
-25 m	**5.71**	1.36	2.63	**3.61**	0.91	1.91
-50 m	**6.04**	0.99	2.52	**3.70**	0.67	1.69
-100 m	**7.43**	-0.59	2.73	**4.04**	0.92	1.60

TABLE 4 MEAN SKILL RELATIVE TO OPERATIONAL EPS OVER DAYS 3-10. T850, N. HEM.

threshold	ROC			Brier score		
	MC	CONS	MA	MC	CONS	MA
4K	**6.37**	3.38	6.33	**3.25**	1.37	2.67
8K	9.48	6.54	**10.47**	1.63	1.11	**2.21**
0	3.15	1.79	**4.20**	2.40	1.19	**2.90**
-4K	**4.87**	0.18	3.45	1.03	1.15	**2.54**
-8K	**15.74**	-0.30	5.85	2.19	0.55	**2.22**

All the alternative configurations are more skilful than the EC EPS. The improvement relative to the EC EPS

TABLE 5 MEAN SKILL RELATIVE TO OPERATIONAL EPS OVER DAYS 3-10. T850, EUROPE.

threshold	ROC			Brier score		
	MC	CONS	MA	MC	CONS	MA
4K	2.56	3.31	**5.02**	**2.50**	0.91	1.61
0	2.68	2.63	**4.33**	1.79	1.62	**2.79**
-4K	**5.91**	0.95	4.14	**2.75**	0.87	2.23

for the best configuration (5-10%) is comparable to that achieved in upgrades to the operational EC EPS (Buizza et al., 1998; Buizza et al., 1999) and is thus potentially valuable. The skill of the CONS EPS is consistently lower (often substantially so) than that of the MC and MA ensembles.

For Z500, the MC EPS has highest mean skill for all thresholds over both the Northern Hemisphere and Europe. The percentage improvement in skill for the MC EPS is comparable to that found in the UKMO assessment (Evans et al., 1999; the UKMO assessment used a different, although overlapping set of cases). For the Northern Hemisphere, the MA EPS provides 70-80% of the improvement of the MC ensemble for the ROC area and 50-80% for Brier score. The skill S_K of the MC ensemble is consistently better than that of the MA ensemble beyond day 3 for all thresholds while the MA EPS also outperforms the consensus EPS at each forecast time (Fig. 5).

Over Europe, the mean Z500 skill $\overline{S_K}$ of the MC ensemble is substantially greater than that of the MA EPS, especially for the Brier score (Table 3). However, most of this improvement derives from the first 5 days of the forecast (Fig. 6); for the latter half of the forecast the MA EPS is generally comparable to, and for some events better than, the MC EPS. In Section 3.1 it was noted that the spread of the MC EPS for Z500 over Europe was good out to day 5 while spread was too small for all other configurations. Beyond day 5 the MC spread was also too small, by an amount comparable to the MA EPS. The correspondence between Figs. 3 and 6 is striking and suggests that the higher MC skill in the first 5 days is due to the larger spread in this period.

Results for T850 are more mixed than for Z500. While for some thresholds the mean skill $\overline{S_K}$ of the MC EPS is greater than that of the MA EPS, it is more often the MA EPS which has greater mean skill (Tables 4 and 5; results for the events $T > 8K$ and $T < -8K$ are not included for Europe because there are too few cases in the sample to give meaningful results). The only event for which the MC EPS substantially out performs the MA EPS is for $T < -8K$ over the Northern hemisphere for ROC area, although for Brier score both configurations are comparable.

4. CORRECTION OF MODEL BIAS

The results of the previous section demonstrate that the MC EPS provides in general more skilful probability forecasts for Z500 than can be obtained using only the EC model. For the Northern Hemisphere the best

configuration using the EC model, the MA EPS, achieved up to 80% of the improvement of the MC EPS, demonstrating that the majority of the benefit of the MC EPS could be attributed to the use of different analyses rather than the use of different models.

The additional skill presumably results from the use of two forecast models within the MC ensemble (although the potential contribution of aspects of the initial analyses not captured by the interpolation to EC model grid - surface fields etc. - cannot be discounted). To first order the difference in model behaviour is represented by the systematic error of the models. We now consider the effect of removing this mean model bias on the performance measures discussed in Section 3.

It was not possible to calculate a model bias using an independent sample of cases because the UK model version used in this study has not been used elsewhere. Model bias was therefore calculated using the control forecasts from the set of ensembles. To take some account of possible seasonal variation in model error, but maintain a reasonable sample size, two bias fields were calculated for each model: a "Winter" bias, using the forecasts for October to February (29 cases), and a "Spring" bias using the remaining dates (March to June - 31 cases).

The appropriate model bias was subtracted from each ensemble member and the scores re-calculated. The bias correction was applied to all configurations, including the EC EPS. However, the skill scores presented in this section are still calculated relative to the uncorrected EC EPS. This allows comparison of the relative performance of the various configurations with the improvement to the EC EPS possible with bias correction alone.

4.1 Ensemble-mean skill and ensemble spread

The ensemble-mean skill of all the ensemble sets is improved by removing model bias. The MC EPS still has the highest skill for Z500 out to day 6, although the difference between the MC and MA scores is reduced. Beyond day 6, as for the uncorrected case, the MA EPS performs as well as or better than the MC EPS. For T850, the MA EPS maintains its advantage over the MC EPS, but again the difference between the two is reduced.

The ensemble spread (rms difference between members and ensemble mean) is not changed by bias correction for the configurations using only the EC model, but the MC spread can be affected. In fact the spread of the bias-corrected MC EPS is not noticeably different from that of the uncorrected MC EPS for Z500, but removing model bias reduces the T850 spread. Because of the improved ensemble-mean skill, the difference between spread and skill is reduced for all configurations, although there is still in general a spread deficit (Fig. 7, c.f. Fig. 4). The difference between MC and MA configurations is again reduced.

4.2 Probability scores

Bias correction improves the probability scores for all the ensemble configurations. Removal of the model bias from the EC EPS can improve the performance for some thresholds to the same level as the uncorrected MC or MA configurations (Tables 6 to 9). However the bias-corrected MC, consensus and MA ensembles retain

TABLE 6 MEAN SKILL RELATIVE TO OPERATIONAL EPS OVER DAYS 3-10. Z500, N. HEM. BIAS REMOVED

threshold	ROC				Brier score			
	EC	MC	CONS	MA	EC	MC	CONS	MA
25 m	8.14	12.76	11.28	**13.76**	4.42	7.85	6.47	**8.18**
50 m	5.80	11.02	8.97	**11.90**	2.95	6.34	4.86	**6.55**
100 m	3.43	**11.16**	6.52	10.03	1.87	**5.11**	3.48	4.98
0	8.33	13.10	11.50	**13.49**	4.87	**8.56**	6.85	8.54
-25 m	7.55	12.49	10.39	**12.50**	4.51	**8.15**	6.22	7.88
-50 m	6.13	**11.62**	8,70	11,27	3.25	**6.95**	4.78	6.45
-100 m	4.60	**11.40**	6.05	9.90	1.59	**5.46**	3.27	4.86

TABLE 7 MEAN SKILL RELATIVE TO OPERATIONAL EPS OVER DAYS 3-10. Z500, EUROPE. BIAS REMOVED

threshold	ROC				Brier score			
	EC	MC	CONS	MA	EC	MC	CONS	MA
25 m	4.17	**9.45**	8.36	9.34	2.85	**6.75**	5.66	6.38
50 m	3.01	**9.38**	7.33	8.85	3.10	**7.46**	5.59	6.49
100 m	1.83	**11.33**	6.64	9.17	3.15	**8.10**	5.28	6.40
0	5.22	**10.35**	9.05	9.66	2.61	**6.11**	5.13	5.85
-25 m	5.82	**10.62**	8.43	9.42	2.60	**5.95**	4.61	5.55
-50 m	6.10	**10.70**	8.11	9.51	2.30	**5.60**	3.89	5.11
-100 m	5.99	**11.68**	6.52	9.99	1.26	**5.31**	2.89	3.96

TABLE 8 MEAN SKILL RELATIVE TO OPERATIONAL EPS OVER DAYS 3-10. T850, N. HEM. BIAS REMOVED

threshold	ROC				Brier score			
	EC	MC	CONS	MA	EC	MC	CONS	MA
4K	7.37	**13.71**	9.78	13.45	3.27	6.44	4.76	**6.63**
8K	10.38	**20.32**	12.76	**18.12**	2.59	4.16	3.07	**4.83**
0	6.43	10.05	8.41	**11.47**	4.12	7.08	5.33	**7.47**
-4K	6,46	11.76	8,12	**12.14**	4.39	6.83	5.17	**6.95**
-8K	3.79	**17.28**	5.41	12.88	3.22	**6.53**	3.31	5.42

TABLE 9. MEAN SKILL RELATIVE TO OPERATIONAL EPS OVER DAYS 3-10. T850, EUROPE. BIAS REMOVED

threshold	ROC				Brier score			
	EC	MC	CONS	MA	EC	MC	CONS	MA
4K	2.62	7.70	5.91	**8.10**	2.06	**5.26**	3.44	4.69
0	2.00	**7.80**	4.73	7.20	1.21	**5.30**	3.02	4.65
-4K	0.24	**8.72**	2.66	7.23	0.56	**5.03**	1.46	3.36

their skill advantage over the bias-corrected EC EPS. The general effect of the bias correction is to reduce the difference between the performance of the MC and MA configurations.

For Z500 the MA EPS now has the highest mean ROC area skill over the Northern Hemisphere for most thresholds and the differences with the MC EPS for the other thresholds are much reduced (Fig. 8 and Table 6). Differences in Brier skill between the two configurations are minimal. Over Europe the MC EPS is still more skilful than the CONS EPS and MA EPS, but the difference is substantially reduced; the substantial advantage of the MC EPS in the first five days is no longer so apparent (Fig. 9, c.f. Fig. 6).

For T850 the MC and MA EPS configurations are comparable (as for the uncorrected case) with no consistent advantage to either system. The large difference in ROC skill between the two systems for $T < -8K$ over the Northern hemisphere is substantially reduced with bias correction.

5. AMPLITUDE OF INITIAL PERTURBATIONS

One consistent result of the above probability assessment (with or without bias correction) is the relatively poor performance of the CONS EPS compared to the MA EPS. Although consistently better than the EC EPS, the skill $\overline{S_K}$ of the CONS EPS is often less than half that of the MA system (Tables 2 to 5). The only difference between the two configurations is the structure of the initial perturbations. In the MA EPS the perturbations are added to the various available analyses, rather to the single consensus analysis. One effect of this is to increase the effective amplitude of the initial perturbations (compare the spreads at day 0 in Fig. 3). There are two possible reasons for the difference in forecast performance between the MA and CONS systems. One is that the initialisation about different analyses explicitly introduces the analysis differences. The other is that by using the different analyses the initial spread is increased.

To test the sensitivity to amplitude of the initial perturbations an additional set of ensembles was run, initialised about the consensus analysis but increasing the amplitude of the perturbations. The operational EPS perturbations for day D are constructed as a combination of perturbations calculated using initial time singular vectors from D and perturbations calculated using the 48-hour evolved singular vectors from D-2 (Barkmeijer et al., 1998). Both are scaled with equal amplitude. For the new experiment the amplitude of the evolved singular vectors was increased, leaving the scaling of the initial singular vectors unchanged. 20 cases have been

run for the new configuration (COEV EPS) covering the period from December to February. These are compared to the corresponding subset of cases for the other configurations discussed in previous sections.

The initial spread of the COEV EPS is similar to that of the MC and MA configurations for Z500, but rather smaller for T850 (Fig. 10). For both fields the COEV spread grows more quickly during the first two days so that it is larger than the MC EPS spread for Z500 and equal to the MC spread for T850 during the forecast. For both fields the COEV spread is larger than the MA spread. Skill of the COEV ensemble mean is slightly worse than that of the CONS EPS before day five and slightly better beyond day five. The COEV spread is substantially more optimal than for the CONS EPS in that on average the spread is closer to the error of the ensemble mean, although for Z500 the spread is a little too large until day 2-3, especially over Europe (Fig. 11).

Increasing the amplitude of the initial spread results in substantial improvements to the probability scores for the COEV EPS compared to the original CONS EPS. The improvement is largest for the more extreme events (Fig. 12) and brings the scores for the COEV EPS to levels comparable to the MC and MA configurations. The only exception is for Z500 over Europe where for most thresholds the COEV EPS is less skilful than the CONS EPS.

6. DISCUSSION AND CONCLUSIONS

The performance of the operational ECMWF ensemble prediction system (EC EPS) has been compared to that of a number of alternative EPS configurations. The comparison had been made over a set of 60 cases spaced at least four days apart between October 1998 and July 1999. During this period the EC EPS configuration consists of a control forecast plus 50 perturbed members all run at $T_L159L31$. The effect of random errors associated with the parametrization processes in the model is represented by the inclusion of stochastic physics. An alternative representation of model errors in an EPS is to include in the same ensemble forecasts produced using different models. A multi-centre (MC) EPS was constructed by combining 27 EC EPS members (including the control forecast) with the same number of forecasts produced using the UK model. The UK forecasts were initialised by adding the EC perturbations to the UK analysis. Thus the MC EPS differs from the EC EPS in two ways: initialisation about two different analyses, and the use of both the EC and the UKMO models for the forecasts. It was found that the MC EPS consistently out-performed the EC EPS in making probabilistic predictions, confirming the work of earlier studies. Because the MC EPS uses information from both the UK and EC analyses in initialisation of the ensemble it is not possible to attribute this improvement directly to the use of two models.

To investigate whether the benefit observed in the MC EPS can be achieved without the need to use different forecast models, two additional EPS configurations were considered, both using only the EC model but with initial conditions defined using the operational analyses from five different centres. The consensus EPS is

equivalent to the EC EPS, but is initialised about the mean of the available analyses; the multi-analysis (MA) EPS is a combination of 11 forecasts initialised about each analysis.

Assessment focused on 500hPa height (Z500) and 850 hPa temperature (T850) over the Northern Hemisphere and Europe.

The control forecast from the consensus analysis was found to be slightly more skilful than that from the ECMWF analysis, consistent with the notion that the consensus analysis filters poorly known scales in the atmosphere and is therefore a closer approximation to the true atmospheric state. The UK control forecast was not overall as skilful as forecasts made with the EC model. Despite this, the ensemble mean of the MC EPS had higher skill than the other configurations for Z500. For T850 the MC ensemble mean was on average no more skilful than the EC EPS, while the MA and CONS ensemble means did have increased skill.

More substantial differences were found in ensemble spread. The initial spread of the MC and MA ensembles was up to twice that of the EC and CONS configurations because analysis differences were included in addition to the singular vector perturbations. While the MC and MA spread did not grow as quickly during the first two days, the increase over the EC and CONS systems was maintained throughout the forecast.

Probability forecasts were evaluated using ROC area and Brier score, both of which were also expressed as skill scores relative to the EC EPS. Performance of the alternative systems was summarised using the mean skill over days 3 to 10. All three configurations were found to be more skilful than the EC EPS for both Z500 and T850. For Z500 the MC EPS performed best. Over the Northern Hemisphere, the majority of this improvement (up to 80%) could be achieved with the MA EPS. Over Europe the performances of the MA and MC systems were comparable beyond day five, although the MC EPS had a substantial advantage in the shorter range. This difference could be associated with the relatively large spread of the MC EPS over Europe before day five. It would seem likely that the relative lack of spread at days 3-5 in the MA EPS contributed to the poorer performance. For T850 there was no consistent difference in performance between the MC and MA systems.

These results demonstrate that the majority of the benefit of the MC EPS can be achieved without the use of different models, but by including analyses from different centres in the initial EPS perturbations. Correction for model bias was shown to reduce further the skill advantage of the MC EPS; after bias correction there was little material difference between the MC and MA systems.

Although consistently better than the EC EPS, the CONS EPS was notably less skilful than the MA and MC configurations. Because the MC and MA systems effectively include analysis differences in the initial perturbations, both have larger initial spread than the CONS EPS. The effect of increasing the initial spread of the CONS EPS to a comparable amplitude was investigated for a subset of 20 cases. It was shown that the spread of the consensus EPS with the larger initial perturbation (COEV EPS) was substantially closer to the optimum level (equal on average to the skill of the ensemble mean) than that of the original CONS EPS. The

increase in initial spread had a substantial impact on the probability scores, for most events raising the skill of the COEV EPS to be comparable with the MC and MA systems.

From the above analysis we conclude that most of the benefit provided by a multi-centre EPS over the EC EPS can be achieved using the EC model alone. Adding differences between the operational analyses of different centres to the initial perturbations (MA EPS) realises up to 80% of the MC improvement and removal of the bias of the EC model provides additional skill so that the MA and MC systems are comparable.

Adding the operational EPS perturbations to the consensus analysis improves performance relative to the EC EPS (the centroid of the initial distribution is closer to the true state). Although the skill of the consensus system is not as great as that of the MA EPS, much of the difference can be attributed to the larger amplitude of the initial spread rather than to the explicit inclusion of analysis differences in the MA EPS. By increasing the amplitude of the evolved singular vectors in the initial perturbations it is possible to substantially increase the probability forecast skill without increasing the forecast spread to unrealistic levels.

ACKNOWLEDGEMENTS

Thanks to staff at UKMO, DWD, Météo-France and NCEP and to the Meteorological Applications Section at ECMWF for their assistance in providing the analysis fields used in this study.

REFERENCES

Barkmeijer, J., Buizza, R. and Palmer, T. N., 1999. 3D_Var Hessian singular vectors and their potential use in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **125,** 2333-2351.

Buizza, R., Barkmeijer, J., Palmer, T. N., and Richardson, D. S., 1999. Current status and future deelopments of the ECMWF Ensemble Prediction System. *Met. Apps.*, To appear.

Buizza, R., Hollingsworth, A., Lalaurette, F., and Ghelli, A., 1999. Probability precipitation prediction using the ECMWF Ensemble Prediction System. *Weather and Forecasting.*, **14**, 168-189.

Buizza, R. and Palmer, T. N., 1998. Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503-2518.

Buizza, R., Petroliagis, T., Palmer, T. N., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A., and Wedi, N., 1998. Impact of model resolution and ensemble size on the performance of and ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **124,** 1935-1960.

Evans, R. E., Harrison, M. S. J .and Graham, R. J., 1999. Joint medium-range ensembles from the UKMO and ECMWF systems. Submitted to *Mon. Wea. Rev.*, March 1999.

Evans, R. E., Mylne, K. R. and Harrison, M. S. J., 1998. Preliminary results from quasi-operational multi-

model multi-analysis ensembles on medium-range timescales. FR Division Technical Report No 258, UK Meteorological Office, Bracknell, UK.

Harrison, M. S. J., Palmer, T. N., Richardson, D. S., Buizza, R., and Petroliagis, T, 1995. Joint medium-range ensembles from UKMO and ECMWF models and analyses. Proceedings of ECMWF Seminar on Predictability, ECMWF, 4-8 September 1995, 61-120.

Leith, C. E., 1974. Theoretical skill of Monte-Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409-418.

Molteni, F., Buizza, R., Palmer, T. N., Petroliagis, T., 1996. The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73-119.

Mylne, K. R., Clark, R. T. and Evans, R. E., 1999. Quasi-operational multi-model multi-analysis ensembles on medium-range timescales. Preprints AMS Conference on NWP, Denver, September 1999.

Richardson, D. S., 1999. Skill and relative economic value of the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.,* To appear.

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

Wilks, D. S., 1995. Statistical methods in the atmospheric sciences. *Academic Press*, 464pp.

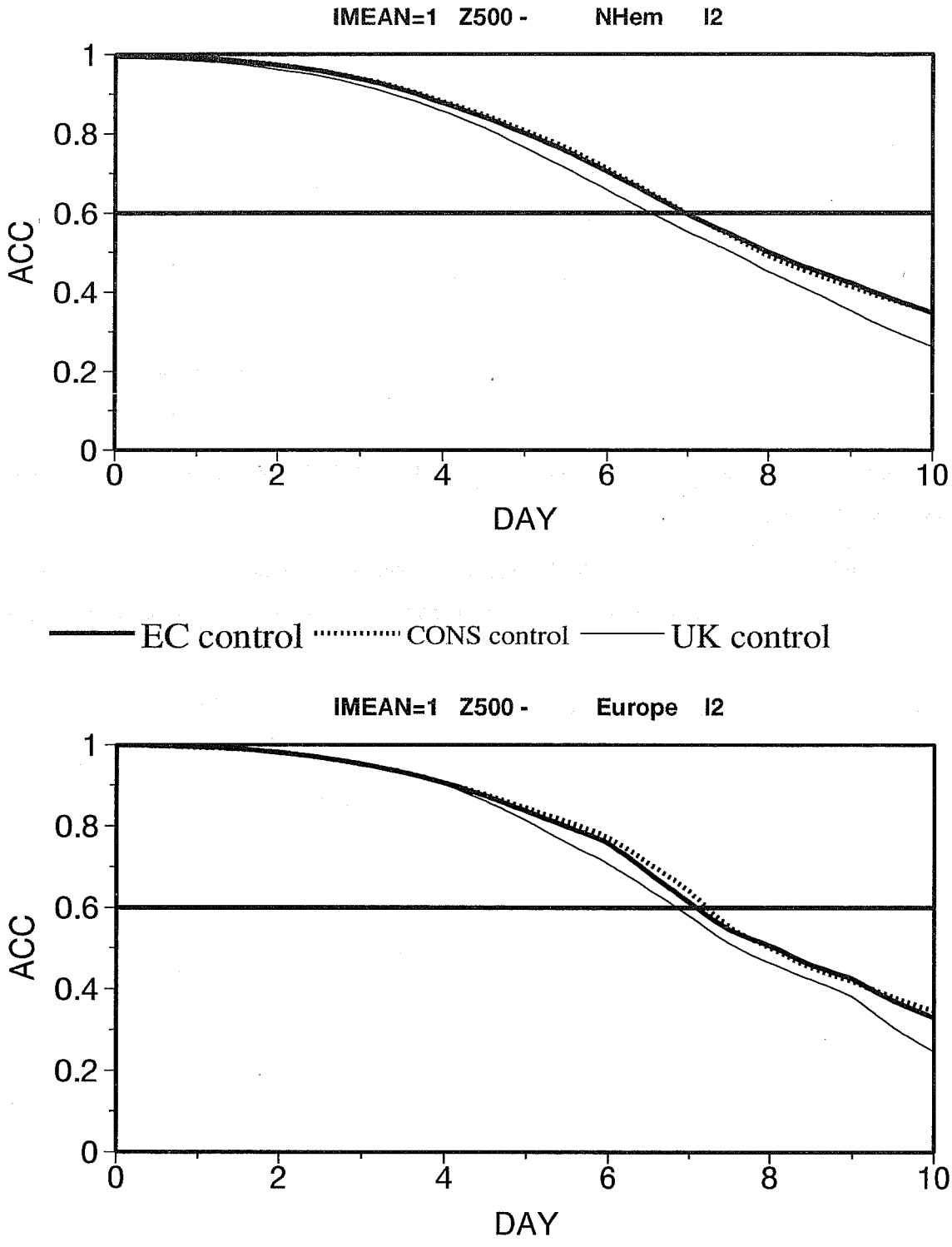
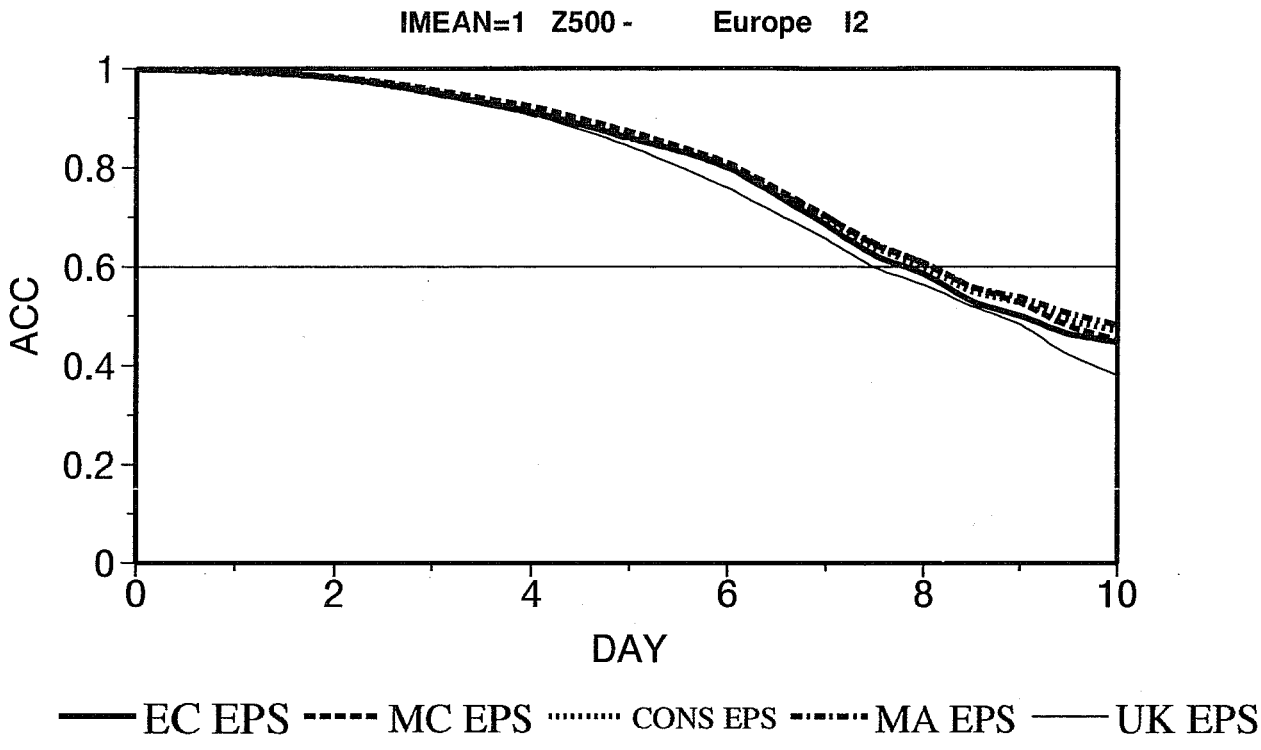


Figure 1: Anomaly correlation skill for the unperturbed control forecast 500 hPa height over the Northern Hemisphere (upper panel) and Europe (lower panel). Average scores over all 60 cases. Thick solid line - EC control; dotted line - consensus control; thin solid line - UK control.



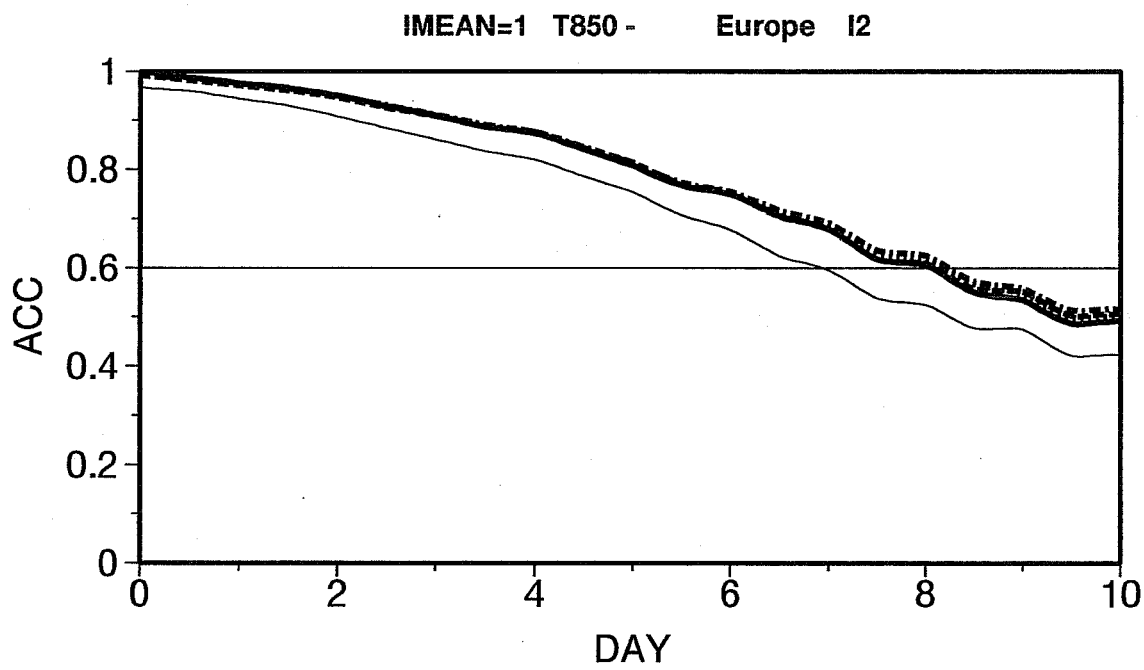
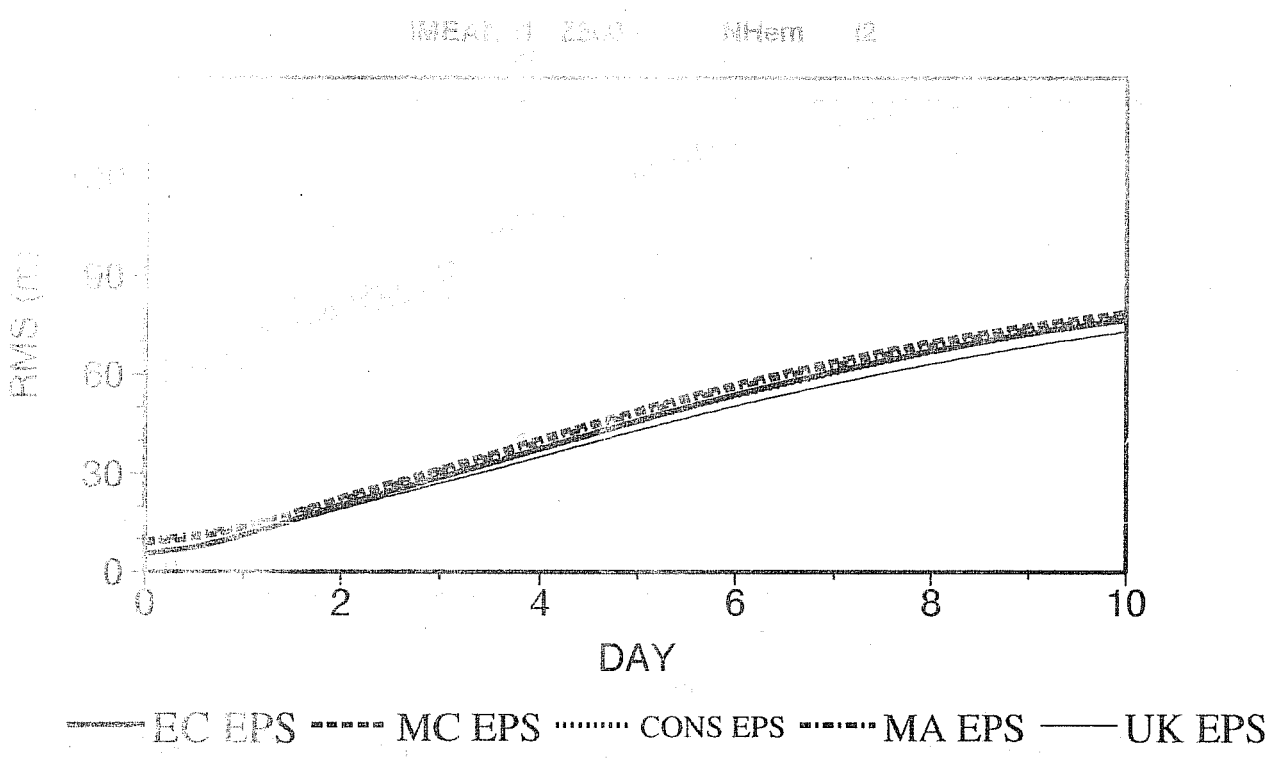


Figure 2: Anomaly correlation skill for the ensemble-mean 500 hPa height (upper panel) and 850 hPa temperature (lower panel) forecasts over Europe. Average scores over all 60 cases. Thick solid line - EC EPS; dashed line - MC EPS; dotted line - CONS EPS; chain dashed line - MA EPS; thin solid line - UK EPS.



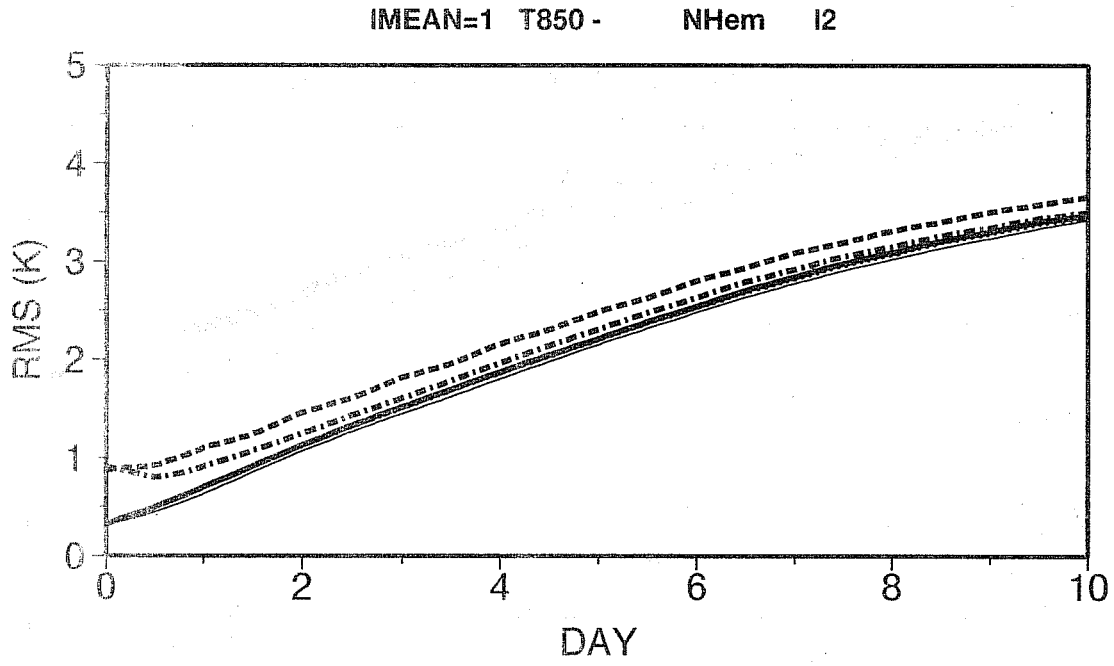


Figure 3: Ensemble spread, defined as average RMS distance between members and the ensemble mean, for 500 hPa height (upper panel) and 850 hPa temperature (lower panel) over the Northern Hemisphere. Thick solid line - EC EPS; dashed line - MC EPS; dotted line - CONS EPS; chain dashed line - MA EPS; thin solid line - UK EPS.

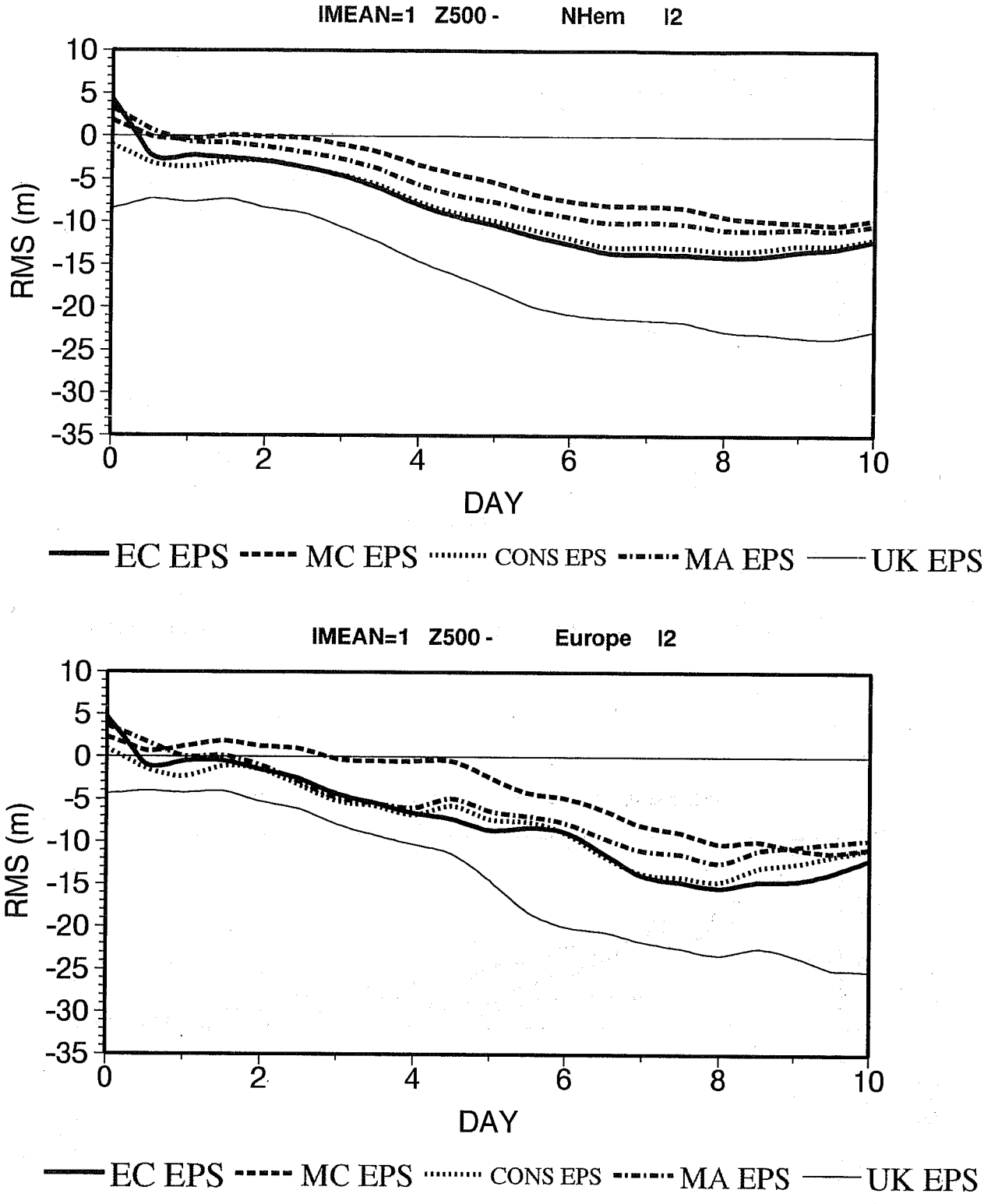


Figure 4: Difference between ensemble spread and ensemble-mean error, for 500 hPa height over the Northern Hemisphere (upper panel) and Europe (lower panel). Average scores over all 60 cases. Thick solid line - EC EPS; dashed line - MC EPS; dotted line - CONS EPS; chain dashed line - MA EPS; thin solid line - UK EPS.

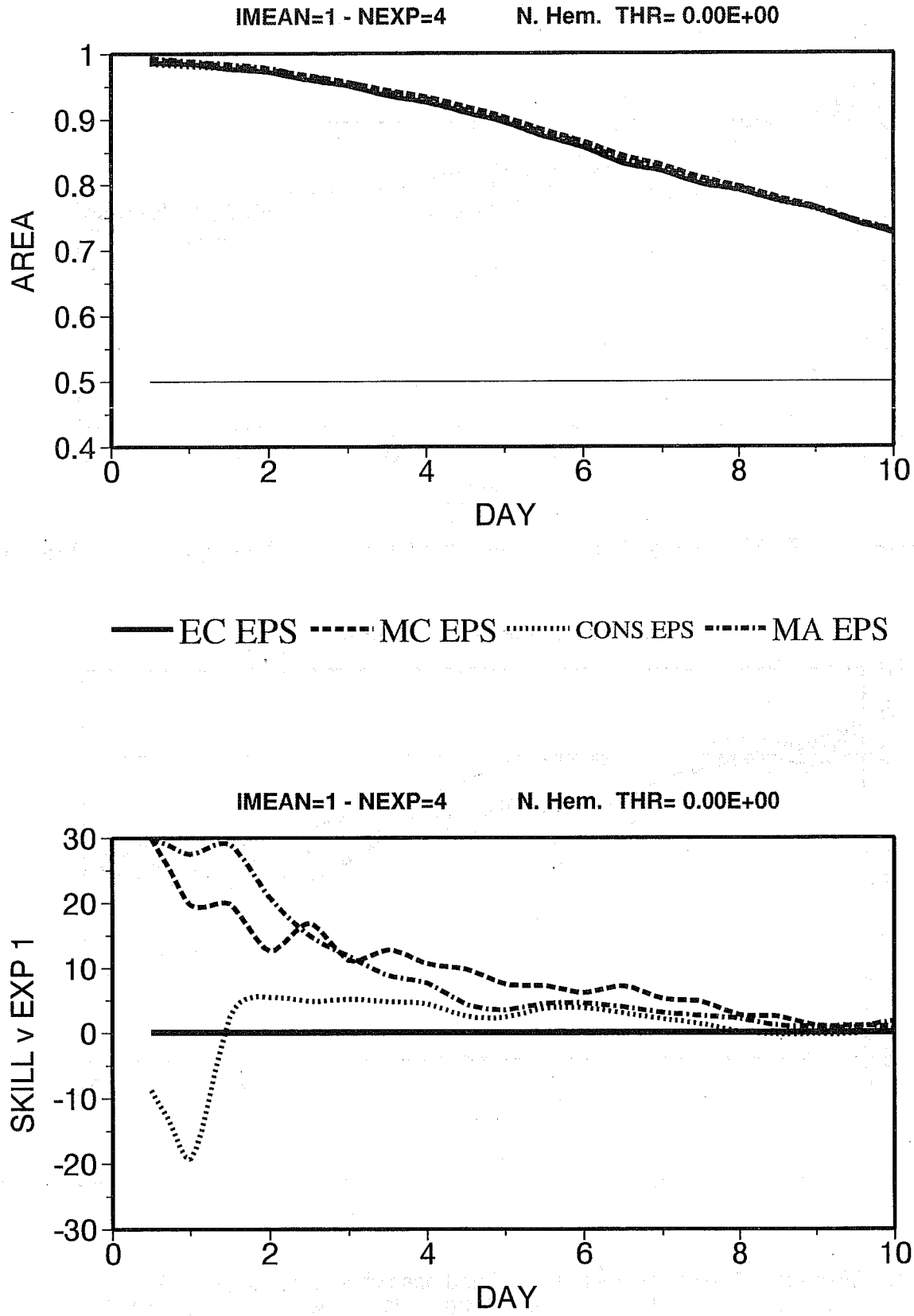


Figure 5: ROC area for EPS probability forecasts of the event '500 hPa height above normal' calculated over the Northern Hemisphere over all 60 cases. Upper panel - ROC area; lower panel - ROC area expressed as a skill score relative to score of EC EPS (see text for details). Thick solid line - EC EPS; dashed line - MC EPS; dotted line - CONS EPS; chain dashed line - MA EPS.

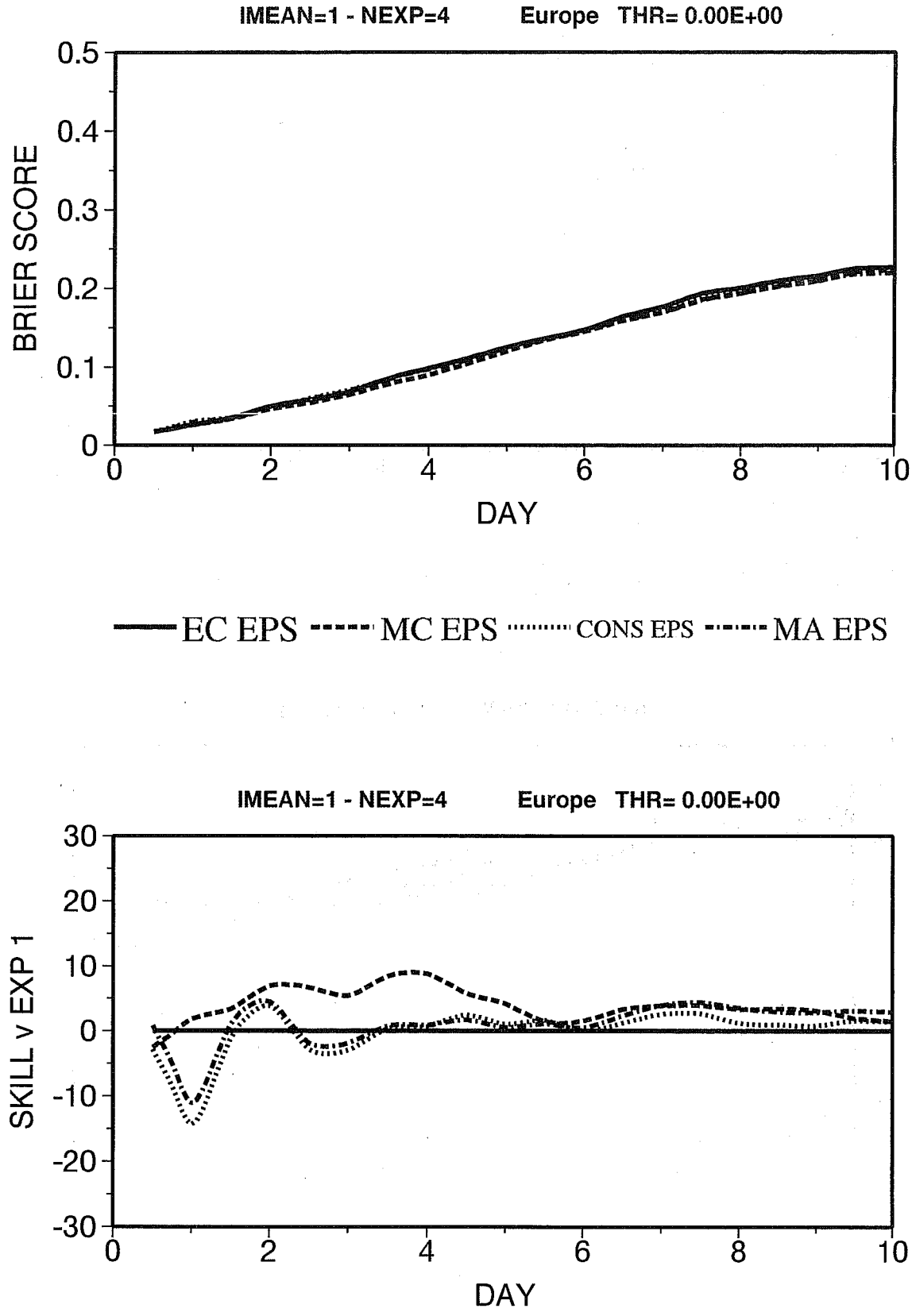
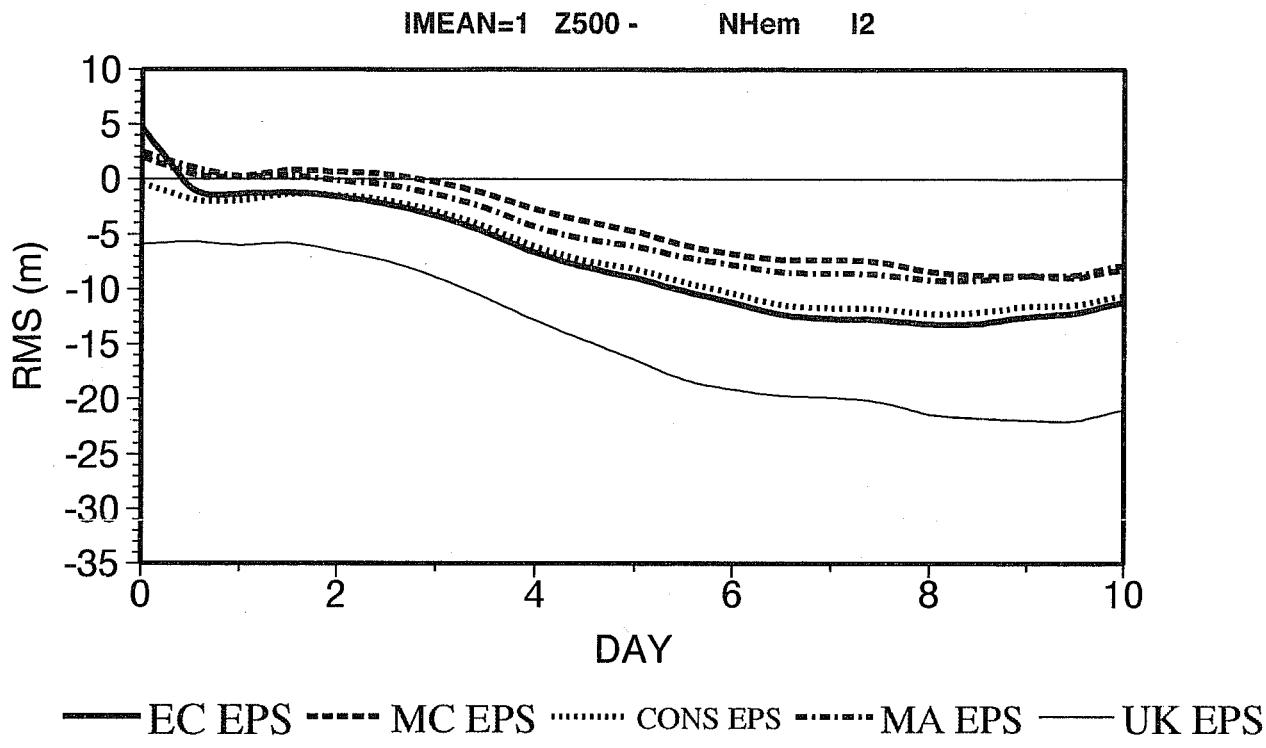


Figure 6: Brier score.for EPS probability forecasts of the event '500 hPa height above normal' calculated over Europe over all 29 cases. Upper panel - Brier score; lower panel - Brier skill score relative to score of EC EPS (see text for details). Thick solid line - EC EPS; dashed line - MC EPS; dotted line -CONS EPS; chain dashed line - MA EPS.



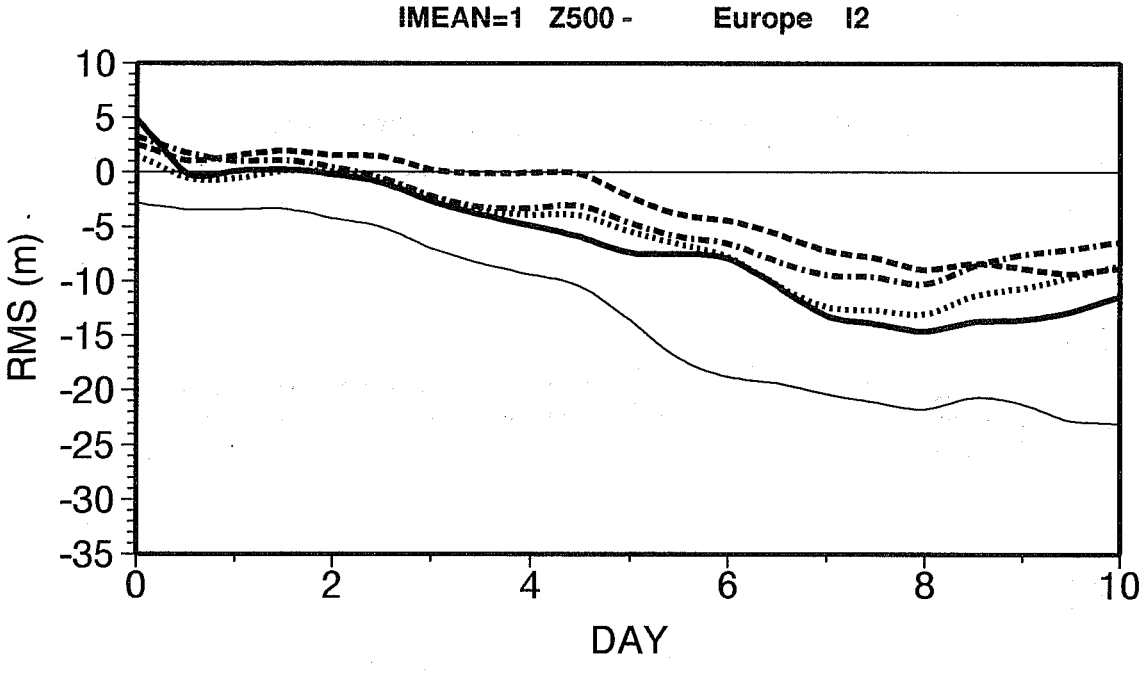


Figure 7: As Fig. 4 but for bias-corrected forecasts.

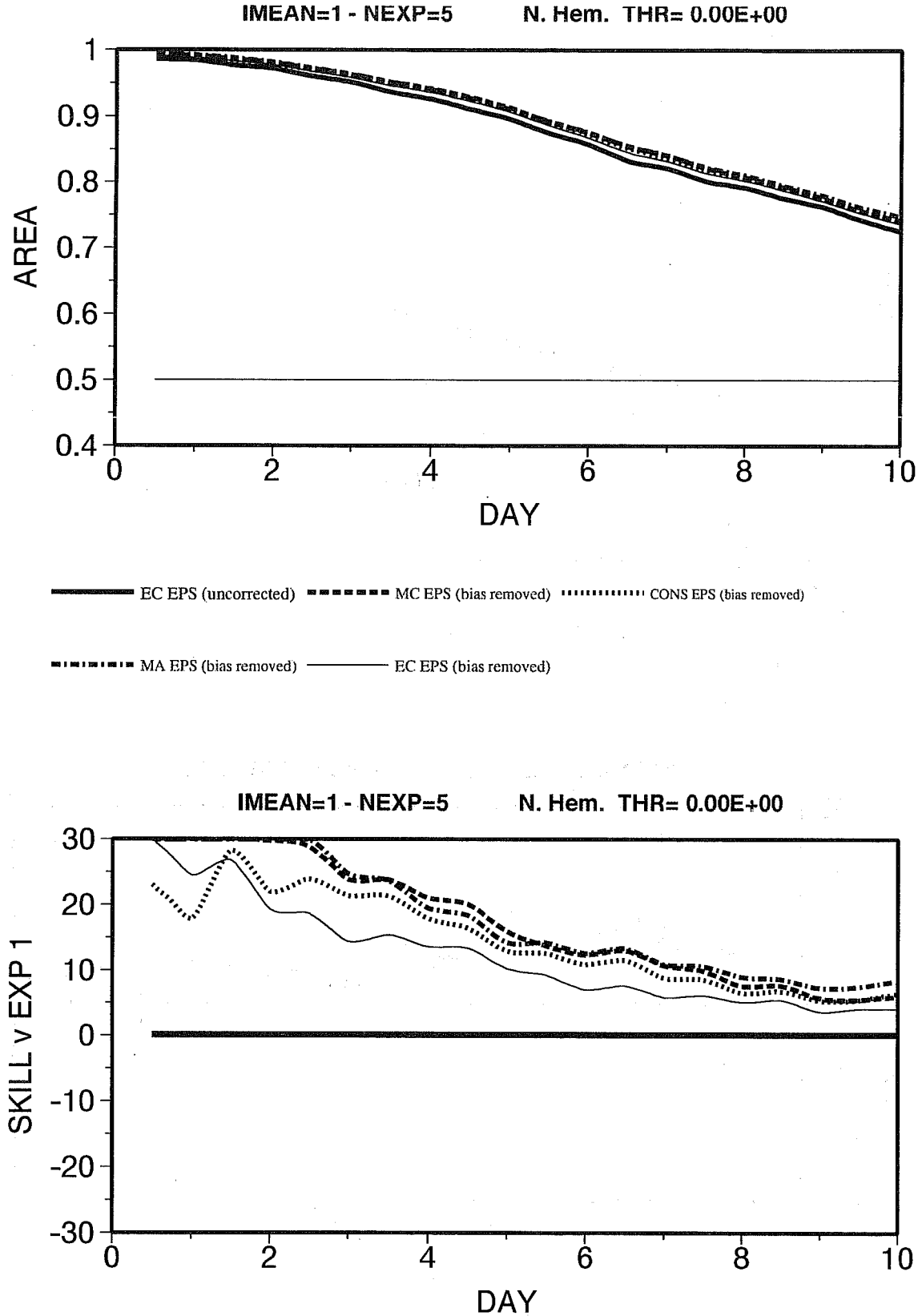


Figure 8: ROC area for bias-corrected EPS probability forecasts of the event '500 hPa height above normal' calculated over the Northern Hemisphere over all 60 cases. Upper panel - ROC area; lower panel - ROC area expressed as a skill score relative to score of EC EPS (see text for details). Thick solid line - uncorrected EC EPS; thin solid line - bias-corrected EC EPS; dashed line - corrected MC EPS; dotted line - corrected CONS EPS; chain dashed line - corrected MA EPS.

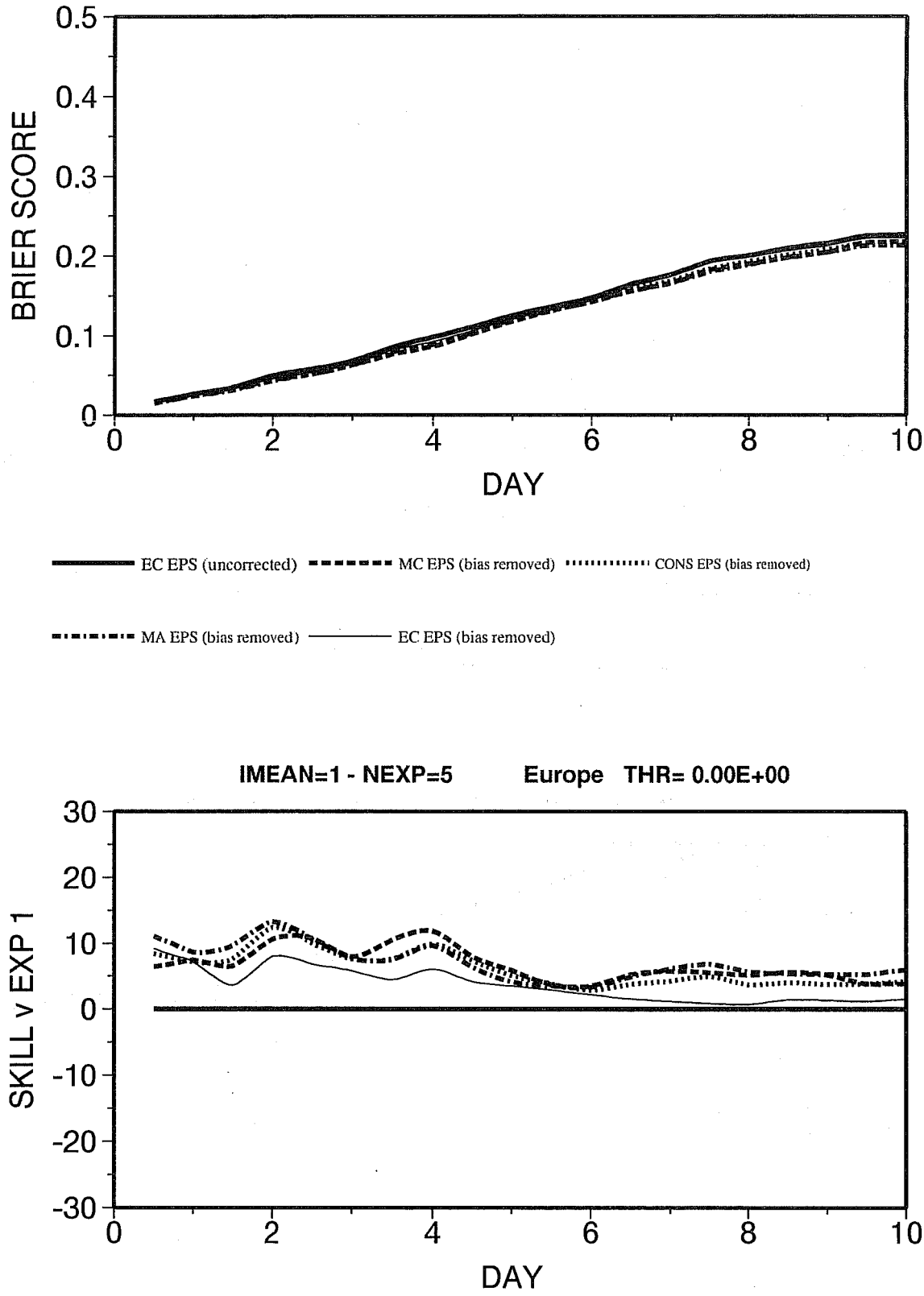
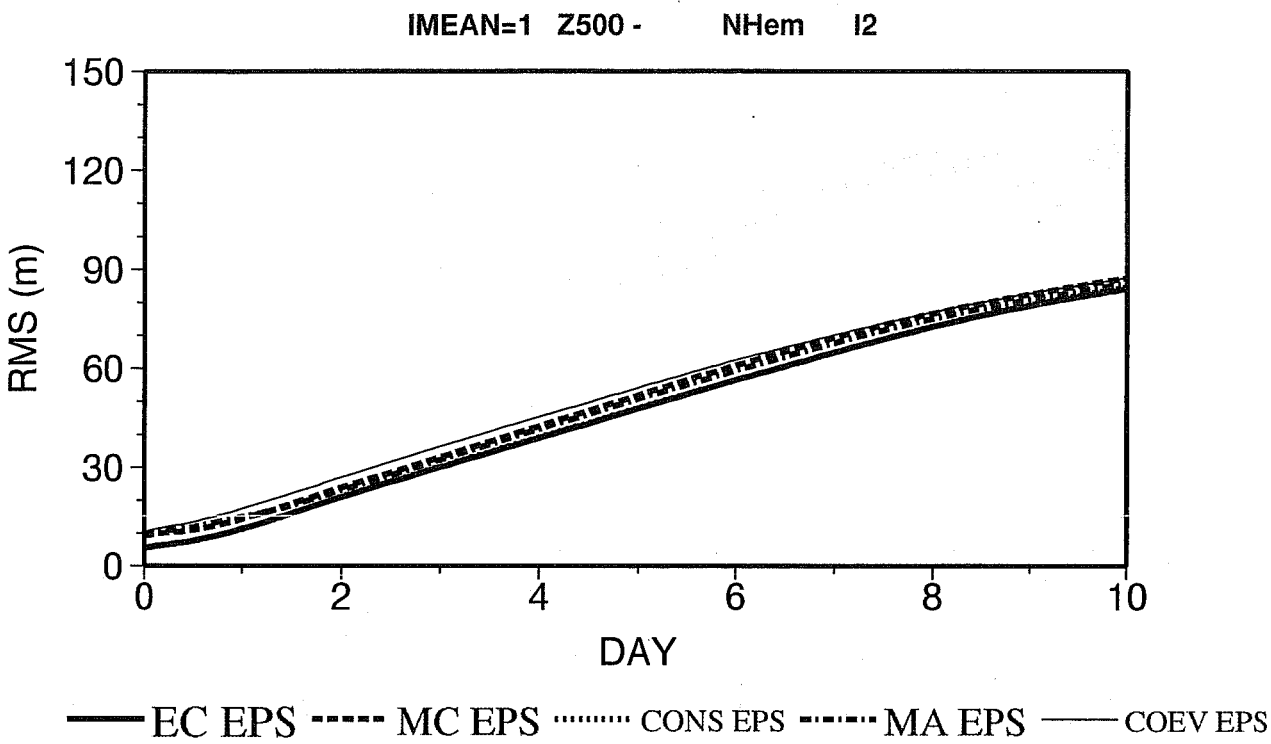


Figure 9: Brier score for bias-corrected EPS probability forecasts of the event '500 hPa height above normal' calculated over the Northern Hemisphere over all 60 cases. Upper panel - ROC area; lower panel - ROC area expressed as a skill score relative to score of EC EPS (see text for details). Thick solid line - uncorrected EC EPS; thin solid line - bias-corrected EC EPS; dashed line - corrected MC EPS; dotted line - corrected CONS EPS; chain dashed line - corrected MA EPS.



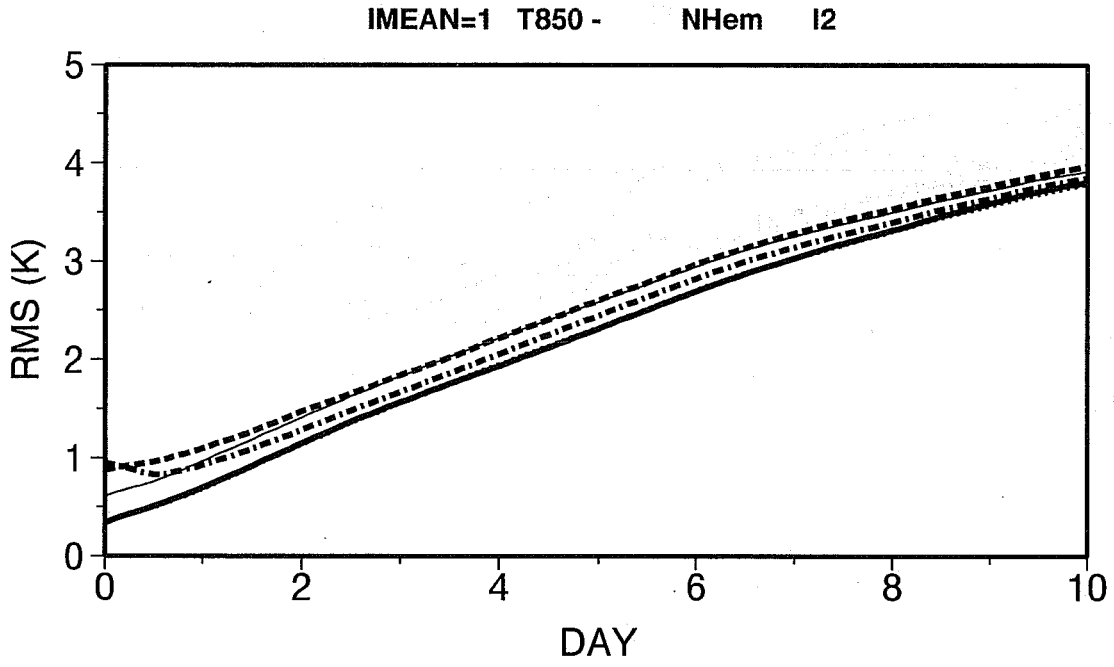
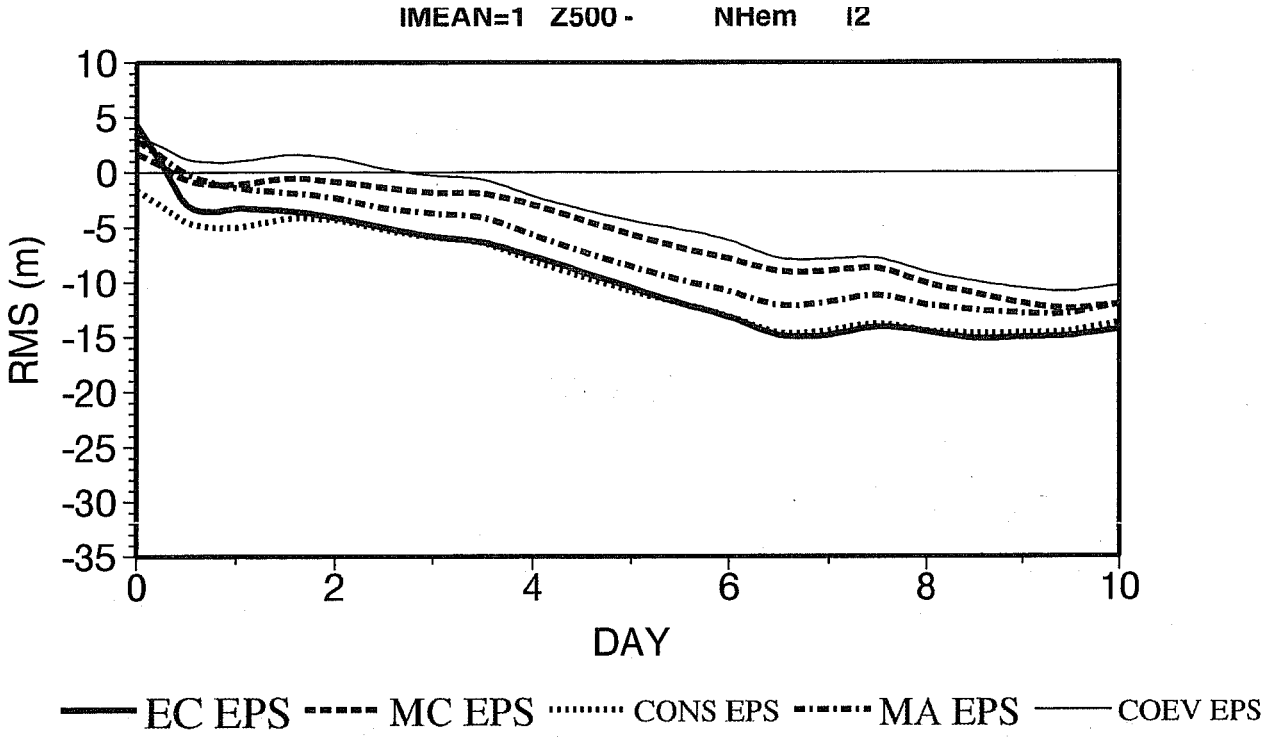


Figure 10: Ensemble spread, measured as RMS distance between members and the ensemble mean, over the Northern Hemisphere for 500 hPa height (upper panel) and T850 (lower panel). Average scores over 20 cases including ensemble with increased initial spread. Thick solid line - EC EPS; dashed line - MC EPS; dotted line - CONS EPS; chain dashed line - MA EPS; thin solid line - COEV EPS.



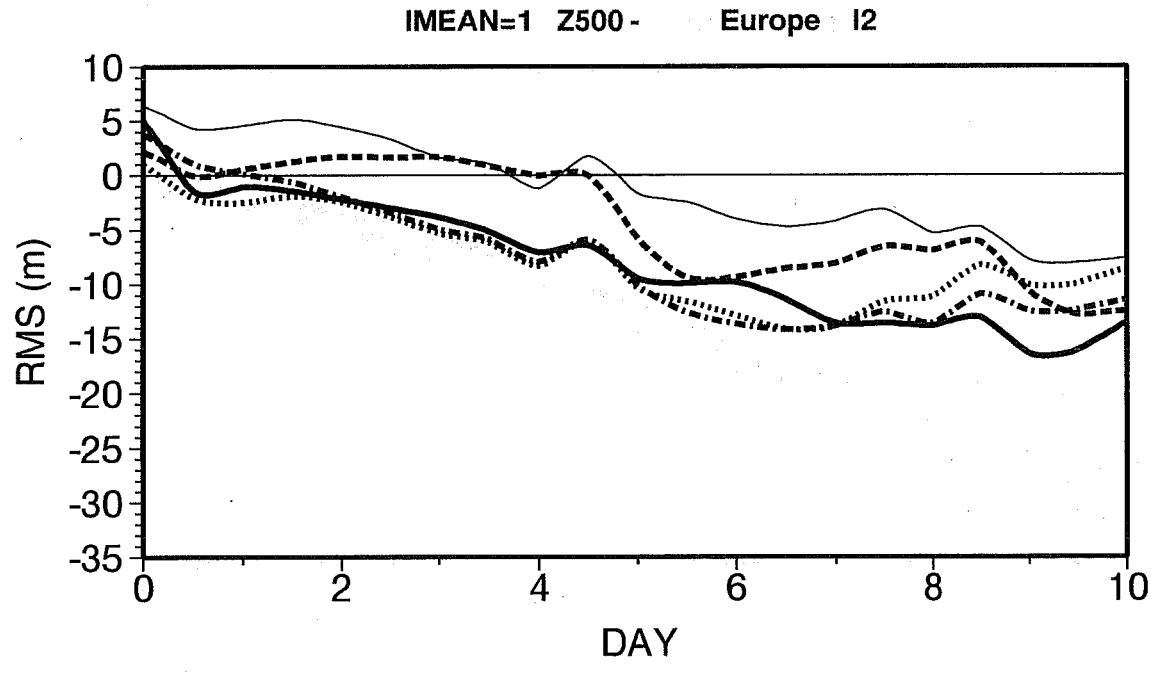
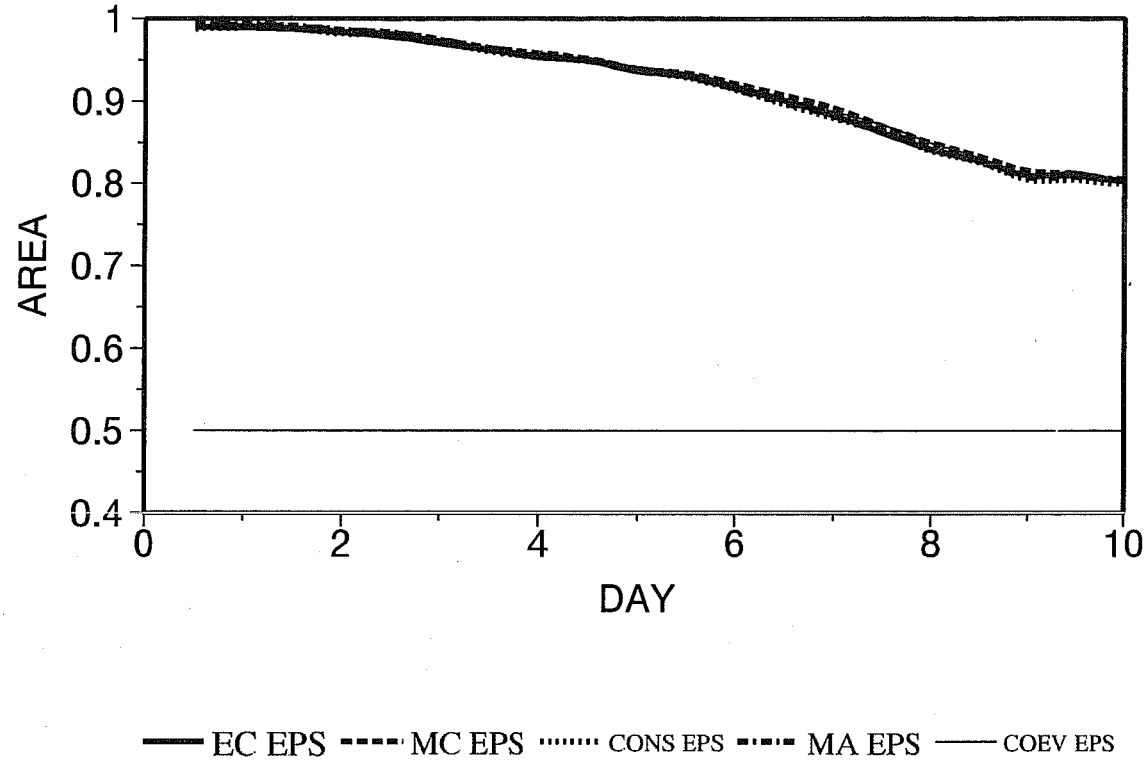


Figure 11: Difference between ensemble spread and ensemble-mean RMS distance error for 20 cases including ensemble with increased initial spread. for 500 hPa height over the Northern Hemisphere (upper panel) and Europe (lower panel). Average scores over 20 cases. Thick solid line - EC EPS; dashed line - MC EPS; dotted line - CONS EPS; chain dashed line - MA EPS; thin solid line - COEV EPS.



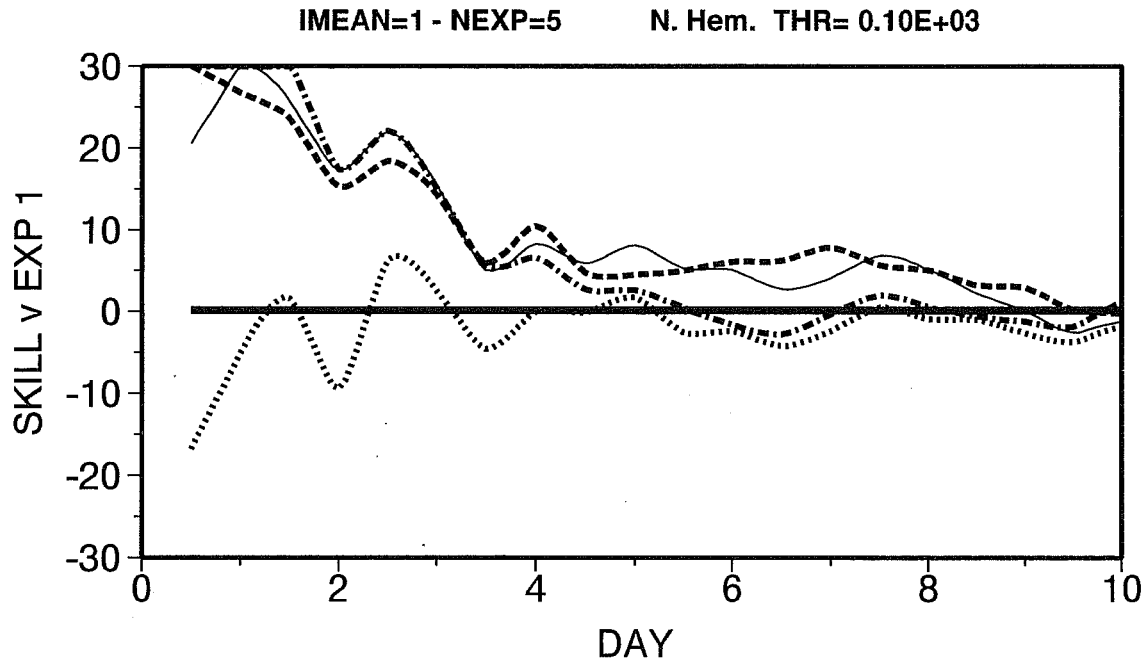


Figure 12: ROC area for EPS probability forecasts of the event '500 hPa height more than 100m above normal' calculated over the Northern Hemisphere over 20 cases. Upper panel - ROC area; lower panel - ROC area expressed as a skill score relative to score of EC EPS (see text for details). Thick solid line - EC EPS; dashed line - MC EPS; dotted line - CONS EPS; chain dashed line - MA EPS; thin solid line - COEV EPS.