TECHNICAL MEMORANDUM

323

Forecasting system performance in summer 1999. Part 3: EPS

R. Buizza

Research Department

September 2000

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Forecast system performance in summer 1999
Part 3: EPS

by **Roberto Buizza**

Abstract

The quality of deterministic and probabilistic products generated using the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) during 1999 is assessed. Attention is focused on two key periods, summer 1999 and October-November 1999, during which the EPS control and the ECMWF high-resolution model performed in a significant different way. Forecasts for the 500-hPa geopotential height over the Northern Hemisphere and Europe are considered.

Summertime results for the past 5 years indicate that the EPS performed very well during summer 1999. Results based on Brier skill scores indicate that the system upgrade of December 1996 (from 33 T63L19 to 51 TL159L31 members) brought a significant increase in predictability (more than 1 day if the Northern Hemisphere is considered).

Results of single deterministic integrations run from 15 October to 10 November indicate that horizontal resolution was the main cause of the different quality of the EPS control and the ECMWF high-resolution (TL319L31) forecasts. This is one of the clearest indications of the impact of horizontal resolution on forecast skill.

1. Introduction

The performance of the ECMWF (European Centre for Medium-Range Weather Forecasts) high-resolution deterministic forecast was unusually poor in the late spring and summer months of 1999 (*Simmons et al.* 2000). Objective scores were low for the Northern Hemisphere (NH) in early May, early July and the second half of August. Scores for Europe were exceptionally poor in August. *Klinker and Ferranti* (2000) have linked this performance with the existence of especially unstable flow conditions over the Atlantic. During such unstable conditions, predictability may be low, and deterministic forecasts will be unreliable.

Since December 1992 the ECMWF operational forecasting system has been supplementing the high-resolution model forecasts with products from the Ensemble Prediction System (*Molteni et al.* 1996). The EPS, based on a finite number of deterministic time integrations, is a practical tool to predict the time evolution of the probability density function of forecast states. The model resolution and the size of the ensemble (i.e. the number of time integrations) are constrained by computer resources. Since 12 October 1999, the EPS has been running with horizontal spectral truncation T_L159 (horizontal spectral triangular truncation with linear grid), with 40 vertical levels, and it has comprised 51 members. By contrast, the ECMWF high-resolution model has been running with horizontal spectral truncation T_L319 and 60 vertical levels.

This paper discusses the performance of the EPS during 1999, in particular during periods when the ECMWF high-resolution forecasts were particularly poor (summer 1999), or a significant difference between the performance of the high-resolution model and the EPS control was detected (from 15 October to 10 November 1999). The past 5 years of EPS forecasts are analyzed to investigate whether there is any significant inter-annual variability of the EPS performance, and whether the EPS system upgrade of December 1996 improved the EPS quality. A set of single deterministic forecasts run for the period 15 October to 10 November 1999 are compared to assess whether resolution was the principal cause of the poor quality of the EPS control forecast, especially with respect to the high-resolution model.

All the results discussed in this paper refer to the 500-hPa geopotential height field, and performance is assessed over the Northern Hemisphere and Europe.

2. The ECMWF EPS

Routine real-time execution of the ECMWF EPS started in December 1992 with a 31-member T63L19 configuration (spectral triangular truncation T63 and 19 vertical levels, *Palmer et al.* 1993, *Molteni et al.* 1996). A major upgrade to a 51-member T_L159L31 system (spectral triangular truncation T159 with linear grid) took place in 1996 (*Buizza et al.* 1998). A scheme to simulate model uncertainties due to random model error in the parameterized physical processes was introduced in 1998 (*Buizza et al.* 1999a). The present EPS configuration can be schematically described as follows.

Each ensemble member e_j can be seen as the time integration

$$e_j(t) = \int_{t=0}^{t} [A(e_j,t) + P_j'(e_j,t)]dt \tag{1}$$

of the perturbed model equations

$$\frac{\partial e_j}{\partial t} = A(e_j,t) + P_j'(e_j,t) \tag{2}$$

starting from perturbed initial conditions

$$e_j(t=0) = e_0(t=0) + \delta e_j(t=0) \tag{3}$$

A and P' in Equation (1) identify the contribution to the full equation tendency of the non-parameterized and parameterized physical processes. For each grid point $x = (\lambda, \phi, \sigma)$ (identified by its latitude, longitude and vertical hybrid coordinate), the perturbed parameterized tendency P' (of each state vector component) is defined as

$$P_j'(e_j,t) = [1 + < r_j(\lambda,\phi,t) >_{D,T}]P(e_j,t) \tag{4}$$

where P is the unperturbed diabatic tendency, and $<..>_{D,T}$ indicates that the same random number r_j has been used for all grid points inside a DxD degree box and over T time steps. The random numbers are currently sampled uniformly in the interval [-0.5,0.5], the same random number is used inside 10° degrees boxes (*D*=10), and the set of random numbers is updated every 6 hours (*T*=6) (note that random numbers do not vary with the vertical coordinate).

$e_o(t=0)$ in Equation (3) is the operational analysis at $t = 0$, while δe_j denotes the *j*-th initial perturbation. For each *d*, the initial perturbations are defined using the singular vectors (*Buizza and Palmer* 1995) growing in the forecast range between day *d* and day *d*+2 at initial time, and the singular vectors that had grown in the past between day *d*-2 and day *d* at final time

$$\delta e_j(t=0) = \sum_{i=1}^{25}[\alpha_{i,j} v_i^{d,d+2}(t=0) + \beta_{i,j} v_i^{d-2,d}(t=48h)] \tag{5}$$

where $v_i^{d,d+2}(t=0)$ is the *i*-th singular vector growing between day *d* and *d*+2 at time $t = 0$ (*Barkmeijer et al.* 1999). The coefficients $\alpha_{i,j}$ and $\beta_{i,j}$ set the initial amplitude of the ensemble perturbations, and are defined by comparing the singular vectors with estimates of analysis errors (*Molteni et al.* 1996).

A brief summary of the most recent changes introduced in the EPS is listed in Table 1.

Date	Change description
19 December 1992	EPS starts with 33 T63L19 members; it is run 3 times a week
1 May 1994	EPS is run daily
11 December 1996	51 member, T_L159L31 system introduced
25 March 1998	Use of evolved singular vectors in the initial perturbations
21 October 1998	Simulation of random model errors (stochastic physics)
12 October 1999	Vertical resolution increases from L31 to L40

Table 1 List of the major changes in the EPS configurations (note that other system modifications like changes in the model cycle or in the data assimilation system are not listed here even if they can affect the EPS).

3. EPS performance during 1999 at the 5-day range

EPS predictions of the 500-hPa geopotential height are assessed over two regions, the NH (latitude from 30°N to 80°N) and Europe (latitude from 30°N to 75°N, longitude from 20°W to 45°E). Apart when clearly stated, all diagnostics refer to the 5-day forecast range. First, the quality of EPS deterministic products is contrasted with the quality of the forecasts from the high-resolution ECMWF model. Then, the quality of EPS probabilistic products is assessed.

3.1. Quality of EPS deterministic products and ensemble spread

Consider a forecast field *f(x)* and a verifying field (the analysis) *a(x)* defined over a geographical region $x \in \Sigma$. Define the inner product

$$\langle f,a \rangle = \sum_{x \in \Sigma} w(x) f(x) a(x) \qquad (6)$$

where x is a grid point with longitude λ and latitude φ, and $w(x)=\cos(\varphi)$. Denote by |..| the norm associated with the inner product defined in Eq. (6).

The quality of a deterministic forecast can be measured in terms of the anomaly correlation of the forecast and the verification, or in terms of the root-mean-square (RMS) distance between the two fields. The anomaly correlation of the forecast f with the verification a (analysis) is defined as

$$AC(f,a) = \frac{\langle f-c \rangle \langle a-c \rangle}{|f-c||a-c|} \qquad (7)$$

and the root-mean-square error is defined as

$$RMSE(f,a) = |f-a| \qquad (8)$$

where c is the climate.

Figure 1 shows the time series of 14-day mean anomaly correlations of the day-5 forecasts from the high-resolution and the EPS control forecasts over the NH and Europe For the NH (Fig. 1a), generally speaking the scores of the two deterministic forecasts have a very similar behavior, with the high-resolution model performing better than the EPS control during the first 4 months of the year but worse during August and the first half of September. For Europe (Fig. 1b), the EPS control performs better than the high-resolution model during May, but significantly poorer during October and November. The large difference in scores during the autumn is reflected in the NH scores (Fig. 1a).

The ensemble-mean, defined as the average among the EPS members

$$em = \frac{1}{51}\sum_{j=0}^{50} e_j \qquad (9)$$

is the simplest and the most immediate product that can be constructed using the EPS. Results show that the ensemble-mean forecast is in general more skilful than the control forecast (Fig. 2), especially during August for Europe (Fig. 2b).

In terms of anomaly correlation, the ensemble spread around the control forecast is defined as

$$sp = < AC(e_j, e_0) >_{j=1,50} \tag{10}$$

where $AC(e_j, e_0)$ denotes the anomaly correlation between the *j*-th perturbed-member and the control, and where the average among anomaly correlations is computed applying a Fisher Z-transform (*Ledermann* 1984, *Buizza* 1997). The ensemble spread *sp* is a measure of the dispersion of the EPS perturbed-members from the control. A necessary condition for the EPS to include the analysis inside the EPS forecast range is that the spread sp is comparable to the error of the control forecast. The ensemble spread can be used to identify predictable situations. Ideally, a small ensemble spread should indicate that the error of the control forecast is small. By contrast, a large ensemble spread should indicate that the error of the control forecast could either be small or large, or in other words that the situation is less predictable.

The comparison of the time series of the ensemble spread *sp* and the control anomaly correlation (Fig. 2) indicate that the ensemble spread is larger (i.e. smaller values if measured in terms of anomaly correlation) during periods of poor control performance. Two indices can be defined to measure the agreement between the ensemble spread and the control accuracy (*Buizza* 1997). The first index is the correlation coefficient between the daily values of the two curves, and the second is the percentage of cases for which the ensemble spread *sp* is smaller than the control error. During 1999, the two indices have similar values for the NH and Europe (Table 2).

	CC spread/control error	**% of cases with too small spread**
NH	46%	13%
Europe	46%	12%

Table 2 Correlation coefficient between ensemble spread and control error (measured in terms of RMS), and percentage of cases with ensemble spread smaller than the control error during 1999. Results refer to the day-5 forecasts for the 500-hPa geopotential height.

The anomaly correlation of the best and the worst EPS forecasts illustrate the range of accuracy reached by the EPS members. Results indicate that the best EPS forecast never had anomaly correlation below 0.80 at day-5 over Europe (Fig. 3). Thus, during 1999 at forecast day-5, at least one member had an anomaly correlation higher than 0.80 even in cases of very poor performance of the EPS control.

Similar considerations could be drawn by considering other forecast times, and root-mean-square distance instead of anomaly correlation as a measure of skill (not shown).

3.2. Quality of EPS probabilistic products

For each month, the average standard deviation of the observed field with respect to the long-term climate has been computed. Then, the following 4 events have been considered:

- 500-hPa geopotential height anomaly (with respect to the long-term climate) larger than half standard deviation;
- 500-hPa geopotential height anomaly larger than one standard deviation;
- 500-hPa geopotential height anomaly smaller than half standard deviation;
- 500-hPa geopotential height anomaly smaller than one standard deviation;

and the probabilities of occurrence of the 4 events have been predicted using the EPS. The probabilistic predictions have been assessed using the area under the Relative Operating Characteristic (ROC, *Mason* 1982) curve as a measure of the capability of the EPS to discriminate between hit and false alarm rates (*Stanski et al.* 1989). (ROC-areas above 0.5 indicate a skilful system.) Furthermore, the Brier score and the Brier skill score computed with respect to a climatological forecast (*Brier* 1950) have been computed. [The reader is referred to *Wilks* (1995) for a general discussion of the problem of the assessment of the accuracy of a forecasting system, and to *Buizza et al.* (1999b) for a recent application of these accuracy measures to the EPS.]

Figure 4 shows the time series of 14-day mean ROC-areas. Generally speaking, the ROC-areas vary between 0.8 and 0.95 over the NH. Results indicate that during 1999 the EPS was more skilful in predicting positive than negative anomalies. Considering Europe, there is a correspondence between periods with a poor EPS control forecasts (and poor ensemble-mean forecasts) and low ROC-areas (see for example the bad spells in May, July and August). Similar considerations can be drawn by considering the Brier score instead of the ROC-area as a measure of skill (Fig. 5).

4. Inter-annual variability of the EPS summer performance

It is of interest to investigate how the summer-1999 performance stands with respect to the EPS performance during the previous 4 years.

4.1. Quality of EPS deterministic products and ensemble spread

Figure 6 shows the performance of the EPS control during the past 5 summers. Considering the NH, results indicate that the performance was best in 1998, with all the other summers characterized by a similar performance. For Europe, by contrast, results indicate that the summer-1999 EPS control performance was indeed the worse of the past 5 years. *Klinker and Ferranti* (2000) suggested that this poor performance was related to the especially unstable flow conditions over the Atlantic, measured for example by the percentage of increase/decrease of the

Eady index over the Northern Hemisphere (see Fig. 4 in *Klinker and Ferranti* 2000). [The Eady index is a measure of the instability of the atmospheric flow, *Hoskins and Valdes* (1990).]

Results are slightly different if one considers the ensemble-mean forecast (Fig. 7). The performance during summer 1999 is the second best for the NH, and the third best for Europe. It is worth to point out that the performance was worse during summers 1995 and 1996, prior to the EPS upgrade of December 1996 (from 33 T63L19 members to 51 T_L159L31 members).

The ensemble spread was largest during summer 1999, both over Europe and for the NH (Fig. 8). Compared to 1998, the (seasonal average) ensemble spread was about 10% larger in 1999. Earlier experimentation have shown that the introduction of the evolved singular vectors (March 1998, *Barkmeijer et al.* 1999) and the implementation of the stochastic scheme to simulates random model errors (*Buizza et al.* 1999a) brought a small increase in ensemble spread. These earlier results estimated that these changes could explain about 4% of the spread increase between 1998 and 1999. The remaining ensemble spread increase is probably due to the fact that the atmospheric flow was more unstable over the NH, and in particular over the Atlantic region, in 1999 than in 1998 (*Klinker and Ferranti* 2000). This can be detected by comparing seasonal-average maps of ensemble standard deviation for summer 1999 and 1998. Ensemble spread is larger in 1999 especially in the medium range (say forecast day 5) over the Atlantic and European region (Fig. 9d).

The ensemble standard deviation is particularly large during August 1999, specifically over the Canadian Arctic region at forecast day 1 (Fig. 9e) and over the northeastern Atlantic at forecast day 5 (Fig. 9f). The comparison of the two maps indicates that one of the locations where initial EPS perturbations had maximum intensity was the Canadian Arctic, from where they propagated to reach the Atlantic-European region after 5 days. Reversibly, perturbations located over the Atlantic-European region in the day-5 forecast originated in the Canadian Arctic region. In other words, the EPS indicate that the Canadian Arctic region is the principal region from where analysis uncertainties (errors) that develop into day-5 forecast errors could originate.

This result is in agreement with *Simmons et al.* (2000), who identified the Canadian Arctic as the region where differences between the initial states (analyses) of a very poor and a very skilful forecast were located (see Fig. 20 in *Simmons et al.* 2000). The first initial state was the ECMWF operational analysis (model cycle 21r2) from which the very poor high-resolution forecast initiated. The second initial state was the analysis generated by experiments performed with an updated model version (cycle 21r4), from which a more skilful forecast started.

The increase in ensemble spread had the positive influence of reducing the difference between the ensemble spread and the control error to the smallest values yet recorded during summertime. This is indicated, for example, by the reduction in the percentage of cases with ensemble spread smaller than control error (Table 3).

Another verification measure of ensemble performance is the percentage of outliers, defined as the percentage of times the verification lies outside the ensemble forecast range (*Buizza* 1997, *Talagrand et al.* 1999). Ideally, for a randomly sampled 51 member ensemble the percentage of outliers should be 3.8% (i.e. 2/52). The summer 1999 spread increase had the positive impact of reducing the percentage of outliers to about 10% (Table 3).

Considering the correlation between ensemble spread and control error (both measured in terms of RMSE), there is no clear signal of an impact induced by the EPS upgrade of December 1996. Despite this, it can be pointed out that the correlation was highest in summer 1999 (Table 3).

	CC spread/control error		% of cases with too small spread		% of outliers	
	NH	**Europe**	**NH**	**Europe**	**NH**	**Europe**
Summer 95	0.47	0.49	16	16	38	29
Summer 96	0.44	0.23	19	19	35	29
Summer 97	0.56	0.39	18	20	25	18
Summer 98	0.57	0.41	19	18	20	14
Summer 99	0.57	0.52	15	13	11	10

Table 3 Correlation coefficient between ensemble spread and control error (measured in terms of RMS), percentage of cases with ensemble spread smaller than the control error and percentage of outliers (see text) during summertime. Results refer to the day-5 forecasts for the 500-hPa geopotential height.

4.2. Quality of EPS probabilistic products

For the NH, the EPS performance in predicting the probability of geopotential height anomalies was significantly the best in summer 1999 (Fig. 10). Furthermore, results indicate a clear and significant EPS improvement after the system upgrade of December 1996.

The forecast time when the Brier skill score crosses the zero line of no-skill can be used as a measure of the quality of the EPS probabilistic products. For the prediction of positive anomalies larger than one standard deviation the crossing forecast time has increased by more than 1 day after the system upgrade of December 1996, from day-8 in 1995 and 1996 to times longer than day-9 afterwards (Fig. 10e). The increase is even longer, say more than 2 days, for the prediction of negative anomalies smaller than one standard deviation (Fig. 10f). About 1-day increase in predictability was detected for the probabilistic prediction of the event “positive geopotential height anomaly” (not shown).

Results for Europe (Fig. 11) confirm the positive impact of the system upgrade. There is still evidence of a 1-day increase in the forecast time when the zero line of no-skill is crossed for the prediction of positive anomalies larger than one standard deviation after December 1996 (Fig. 11e). Results for negative anomalies smaller than one standard deviation also confirm the improvement, even if there is a less clear cut between results before and after December 1996 (Fig. 11f). Results for the event “positive geopotential height anomaly” are similar to the results for the negative anomaly (not shown).

For both areas, similar conclusions could be drawn by considering positive and negative anomalies of half standard deviation (not shown).

5. Impact of horizontal resolution on deterministic forecasts

The performance of the EPS control was significantly worse than the performance of the high-resolution model during the second half of October and the first half of November 1999, especially at forecast day-5 (Fig. 1b) and day-7 (not shown). Despite this, it is worth mentioning that the EPS performance during autumn 1999 (1 September to 30 November) was very good according to all measures of ensemble skill discussed above (not shown).

The only difference between the two deterministic forecasts is resolution (T_L159L40 vs. T_L319L60). Thus, a set of single deterministic integrations has been run to investigate the impact of horizontal and vertical resolution on forecast accuracy (Table 4). All these experiments have been performed with the ECMWF model cycle used operationally during the period of interest (cycle 21r4) and starting from the operational analysis (computed at T_L319L60 resolution) truncated, for each experiment, at the corresponding resolution.

Experiment name	Horizontal resolution	Vertical resolution	Initial conditions
High-resol.	T_L319	L60	Oper. Anal.
EPS control	T_L159	L40	Oper. Anal.
319-40	T_L319	L40	Oper. Anal.
255-60	T_L255	L60	Oper. Anal.
255-40	T_L255	L40	Oper. Anal.
255-31	T_L255	L31	Oper. Anal.
159-60	T_L159	L60	Oper. Anal.
159-31	T_L159	L31	Oper. Anal.
159-60-N	T_L159	L60	T_L159L60 Anal.

Table 4 List of the main characteristics of the single deterministic forecasts run for the period 15 October to 10 November 1999.

A comparison of the daily anomaly correlation values of the two forecasts indicates that they differ mainly during two periods, 24-to-27 and 29-to-31 October (Fig. 12). The first conclusion that can be drawn from the comparison of the different forecasts is that vertical resolution played a small role. All experiments run with different vertical resolutions but the same horizontal resolution performed very similarly, especially during the two key periods. The second and more important conclusion is that horizontal resolution was the key reason for the different performance, with T_L159 forecasts performing worse and T_L319 best. Results also indicate that a resolution increase from T_L159 to T_L255 would have improved scores only partially.

Apart from the forecast resolution, one of the reasons for the poor performance of the low-resolution forecasts could be the fact that errors can be introduced in the forecast initial conditions by the interpolation of the high-resolution (T_L319L60) operational analysis to the forecast resolution. To verify whether this was the case, the ECMWF data assimilation system has been run at T_L159L60 resolution for the whole period of interest, and a second set of T_L159L60 experiments has been performed (Table 4). The comparison of the two T_L159L60 integrations run from the two different analyses (Fig. 13) indicates that the interpolation played no role at forecast day-5, and it had a positive impact in two cases at forecast day-7. For these two cases (the forecasts started on 25 and 26 October) the $T_L159L60$ forecasts started from the $T_L159L60$ analysis performed better than the $T_L159L60$ forecast started from the operational analysis (Fig. 13b).

6. Conclusions

The main issue discussed in this paper is the performance of the ECMWF Ensemble Prediction System during 1999, particularly during periods of poor performance by the ECMWF high resolution model. The EPS performance has been assessed using a variety of accuracy measures and considering the 500-hPa geopotential height over the NH and Europe. For reason of space, most of the results have been confined to the day-5 forecast range.

Two periods have been identified in 1999 by considering the performance of the EPS control forecast (T_L159 resolution, with 31 levels up to 12 October 1999, then with 40 levels) and the ECMWF high-resolution (T_L319 with 50 levels up to 12 October, than with 60 levels). The first one is summer 1999, a period during which the high-resolution forecasting system performed particularly poorly (Simmons et al. 2000). The second period, which starts on 15 October and ends on 10 November, is characterized by remarkably poor performance of the EPS control forecast compared to the high-resolution model.

Compared to the previous 5 summers, the EPS performance during summer 1999 was one of the best according to several measures. This was particularly evident for probabilistic predictions of geopotential height anomalies. The comparison of the summertime Brier skill scores confirmed that the EPS upgrade of December 1996 (from a 33 T63L19 system to a 51 $T_L159L31$ system) had a significant positive impact. In terms of the forecast range at which the Brier skill score crosses the line of no-skill, results indicated an increase in predictability of 1 to 2 days, depending on the event considered. Some indicators of the EPS skill nevertheless indicate relatively poor EPS performance at the times of the poor performance of the high-resolution deterministic system in late spring and summer 1999. This was especially true for Europe in late August.

For the second period (15 October to 10 November 1999), a set of single deterministic experiments has been run with different resolution, to investigate whether resolution could explain the large difference in the performance of the EPS control and the high-resolution model. Results indicated that horizontal resolution had indeed a major impact, while vertical resolution had practically no role. These results constitute one of the clearest evidence of a positive impact of a horizontal resolution on the performance of single deterministic forecasts.

Experimentation has started to assess the impact of an EPS resolution increase from T_L159 to T_L255 (with 40 vertical levels). Preliminary results indicate a positive impact, especially in cases of severe events associated with rapidly developing features. Results will be documented in due course.

Acknowledgement

Adrian Simmons and Tim Palmer are acknowledged for their useful comments to an early version of this manuscript. The EPS performance is maintained to its high standard by the work of the ECMWF Predictability and Diagnostic Section and many ECMWF staff and consultants. Their essential work is acknowledged.

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. Meteor. Soc.*, **125**, 2333-2351.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Buizza, R. B., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.* , **125**, 99-119.

————, and Palmer, T. N. 1995: The singular-vector structure of the atmospheric general circulation. *J. Atmos. Sci.*, **52**, 1434-1456.

————, 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 the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **124**, 1935-1960.

————, Miller, M., and Palmer, T. N., 1999a: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **125**, 2887-2908.

————, Barkmeijer, J, Palmer, T N, and Richardson, D, 1999b. Current Status and Future Developments of the ECMWF Ensemble Prediction System. *Meteorol. Appl.*, 6, 1-14.

Hoskins, B. J., and Valdes, P. J., 1990: On the existence of storm tracks. *J. Atmos. Sci.*, **47**, 1854-1864.

Klinker, E., and Ferranti, L, 2000: Diagnostics related to the poor forecast performance during spring and summer 1999. *ECMWF Technical Memorandum No. 321.*

Ledermann, W., 1984: *Statistics. Vol. 6, Handbook of applicable mathematics*, J. Wiley & Sons, 942 pp.

Mason, I, 1982. A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.

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

Palmer, T. N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P, and Tribbia, J., 1993: Ensemble Prediction. *Proc. ECMWF Seminar (1992)*, ECMWF, Shinfield Park, Reading RG2-9AX, UK.

Simmons, A.J., Andersson, E., Fisher, M., Jakob, C., Kelly, G.A., Lalaurette, F., McNally, A.P., Untch, A., and Viterbo, P., 2000: impact of system changes on ECMWF forecasts for summer 1999. *ECMWF Technical Memorandum No. 322.*

Stanski, H. R., Wilson, L. J., and Burrows, W. R., 1989: Survey of common verification methods in meteorology. Research Report n. 89-5, Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin Street, Downsview, Ontario, Canada M3H 5T4, pp 114.

Talagrand, O., Vautard, R., and Strauss, B., 1999: Evaluation of probabilistic prediction systems. *Proceedings of the ECMWF Workshop on Predictability*, ECMWF, Reading, 20-22 October 1997, 1-26.

Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences*, Academic Press, pp 467. (ISBN 0-12-751965-3).

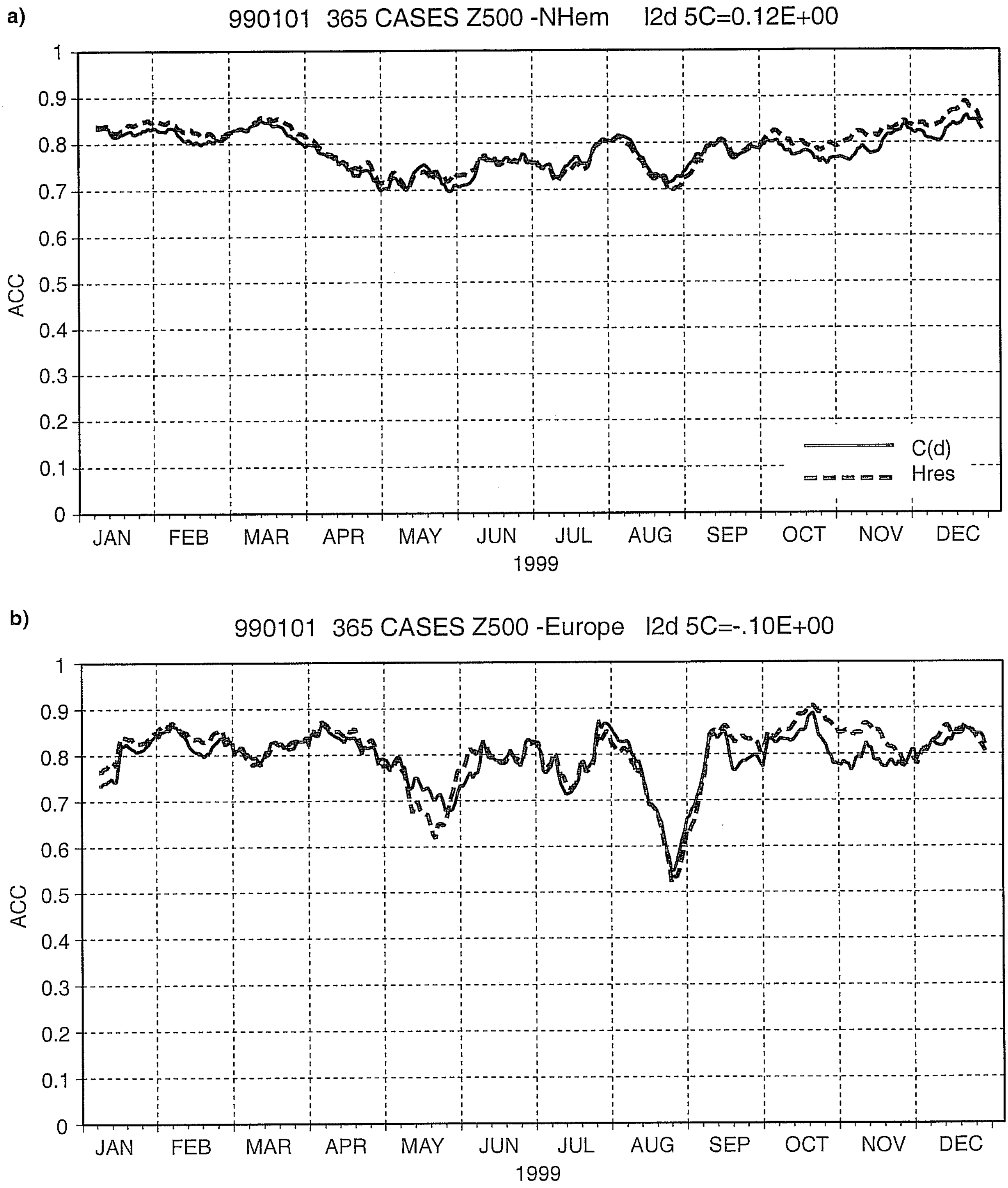


Fig. 1 1999 time series of 14-day mean anomaly correlation for the EPS control (solid line) and the ECMWF high-resolution model (dashed line), for (a) the NH and (b) Europe. Results refer to the day-5 forecasts for the 500-hPa geopotential height.

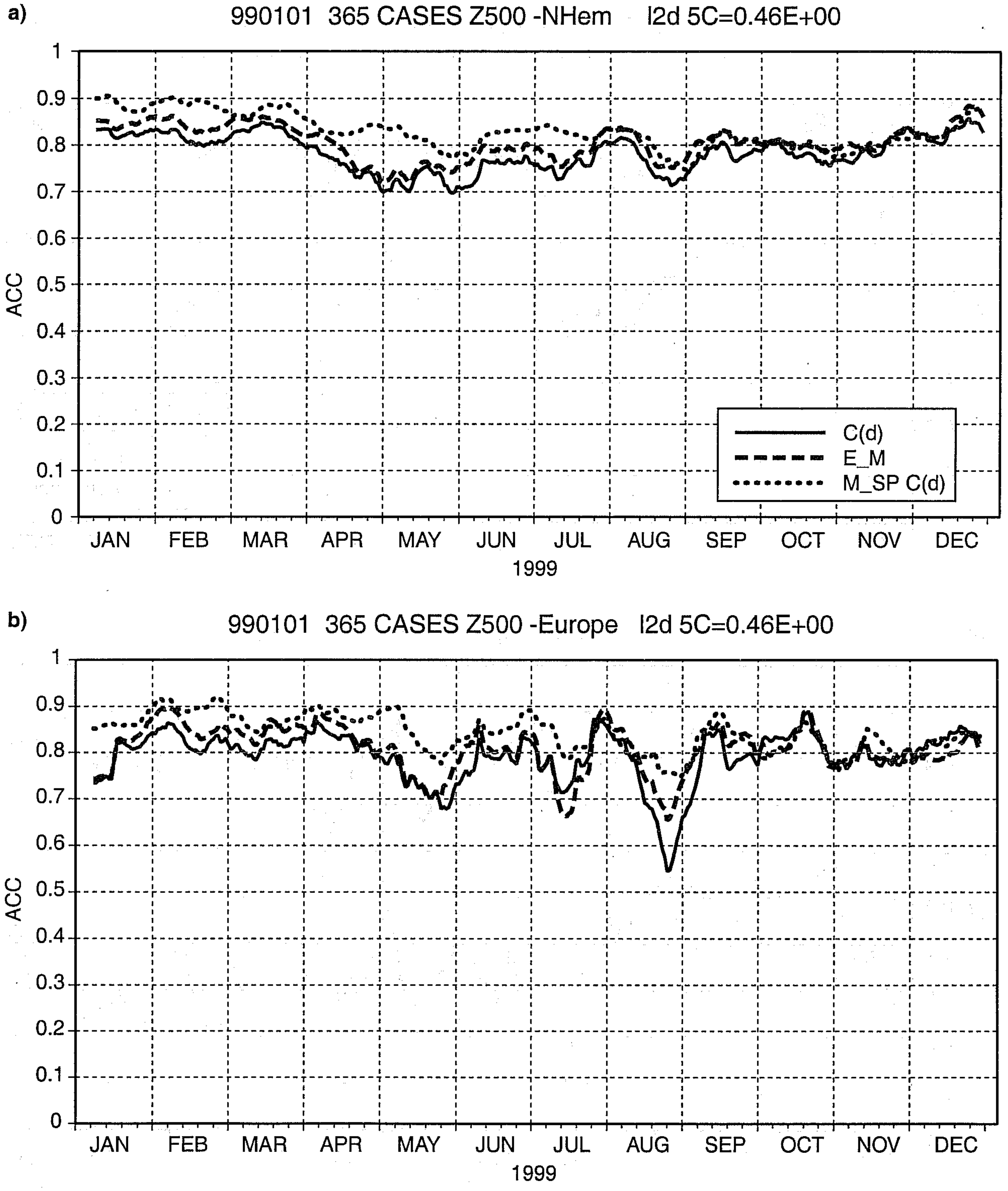
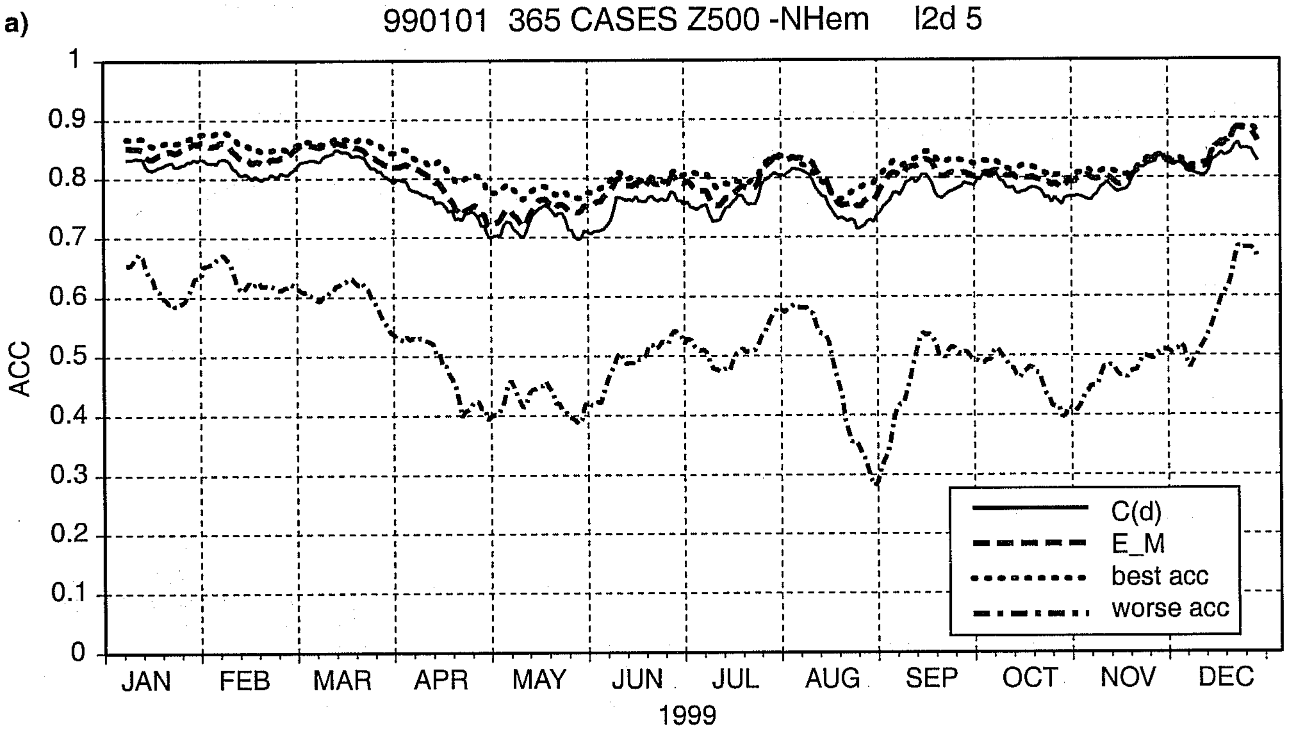


Fig. 2 1999 time series of 14-day mean anomaly correlation for the EPS control (solid line), the ensemble-mean (dashed line) and the ensemble spread (dotted line), for (a) the NH and (b) Europe. Results refer to the day-5 forecasts for the 500-hPa geopotential height.



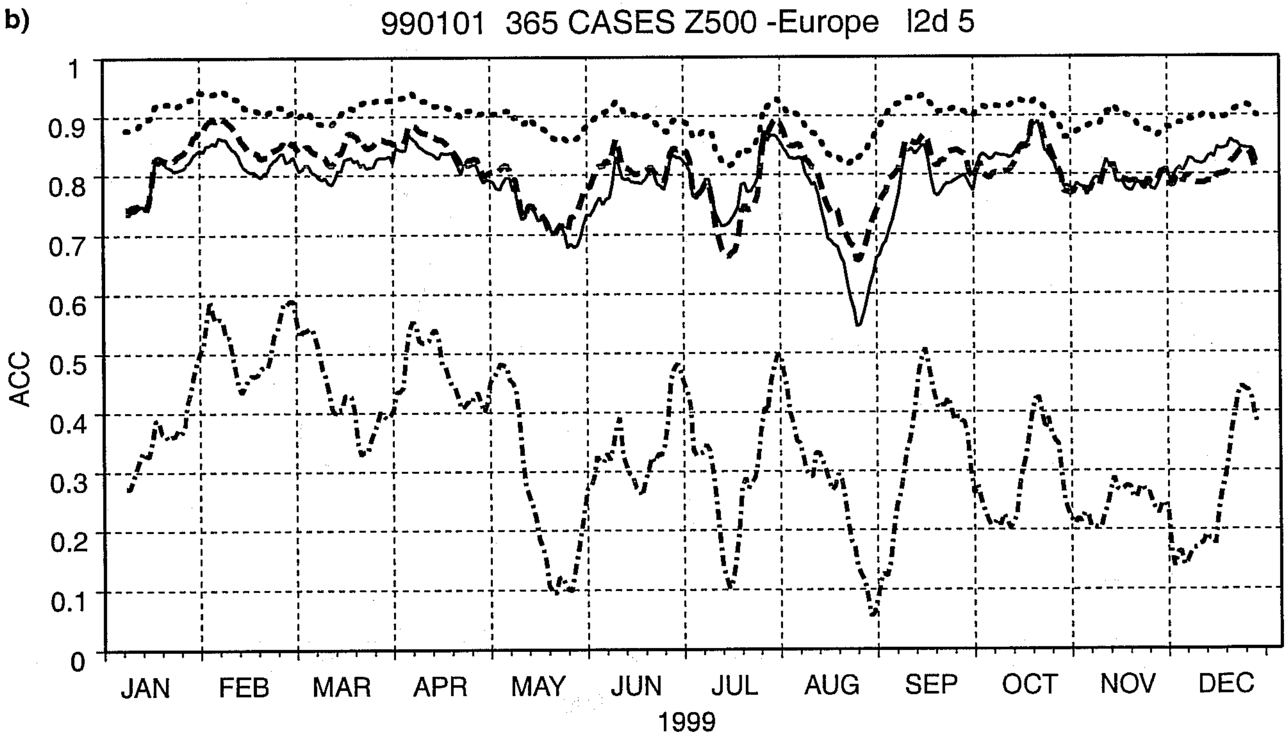
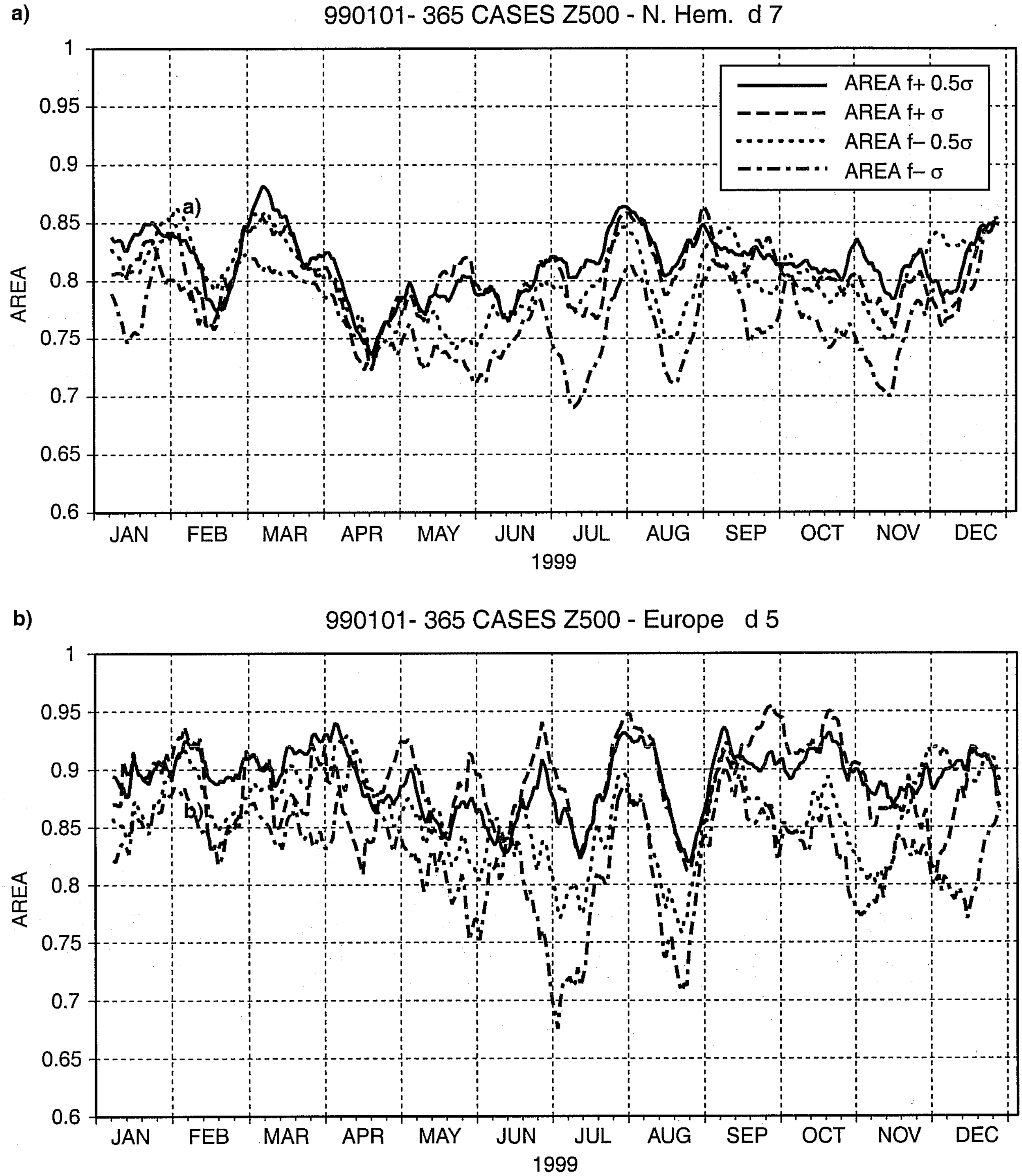


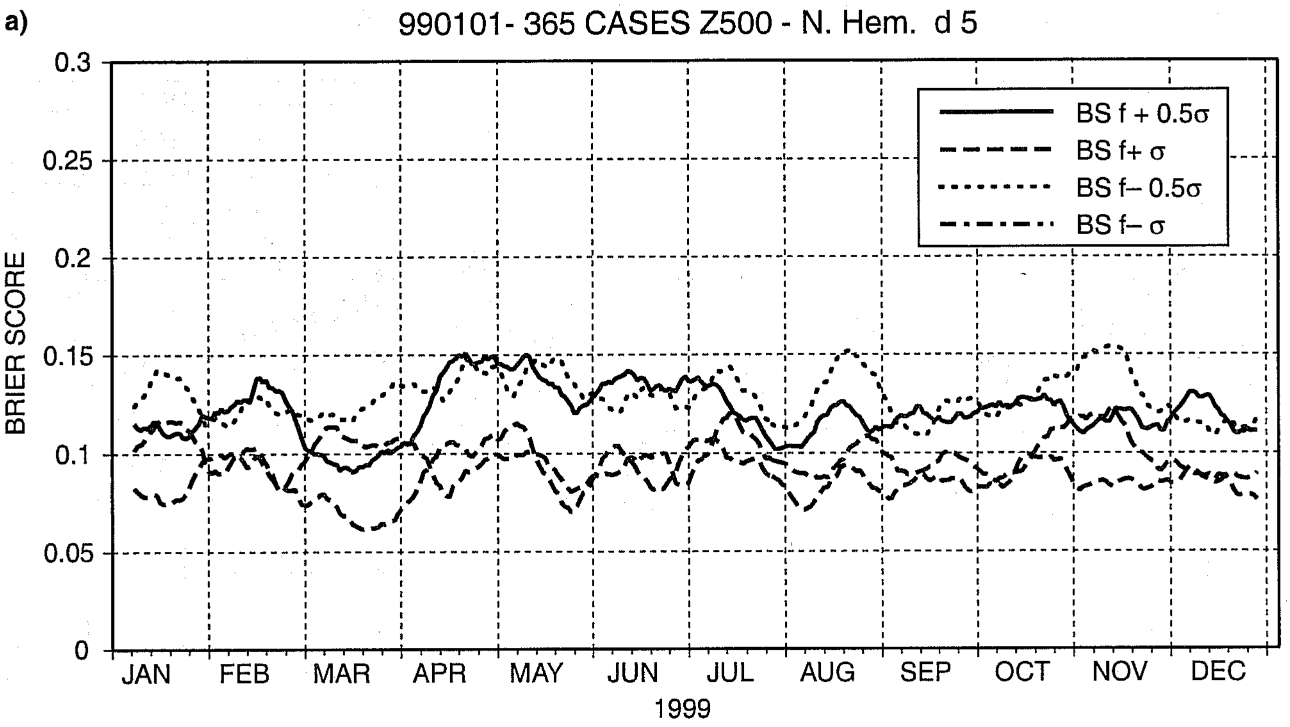
Fig. 3 1999 time series of 14-day mean anomaly correlation for the EPS control (solid line), the ensemble-mean (dashed line), the best (dotted line) and the worst (chain-dashed line) ensemble members, for (a) the NH and (b) Europe. Results refer to the day-5 forecasts for the 500-hPa geopotential height.



990101- 365 CASES Z500 - N. Hem. d 7

990101- 365 CASES Z500 - Europe d 5

Fig. 4 1999 time series of 14-day mean ROC-area for the prediction of "positive anomalies larger than one standard deviation" (solid line), "positive anomalies larger than half standard deviation" (dashed line), "negative anomalies smaller than one standard deviation" (dotted line) and "negative anomalies smaller than half standard deviation" (chain-dashed line), for (a) the NH and (b) Europe. Results refer to the day-5 forecasts for the 500-hPa geopotential height.



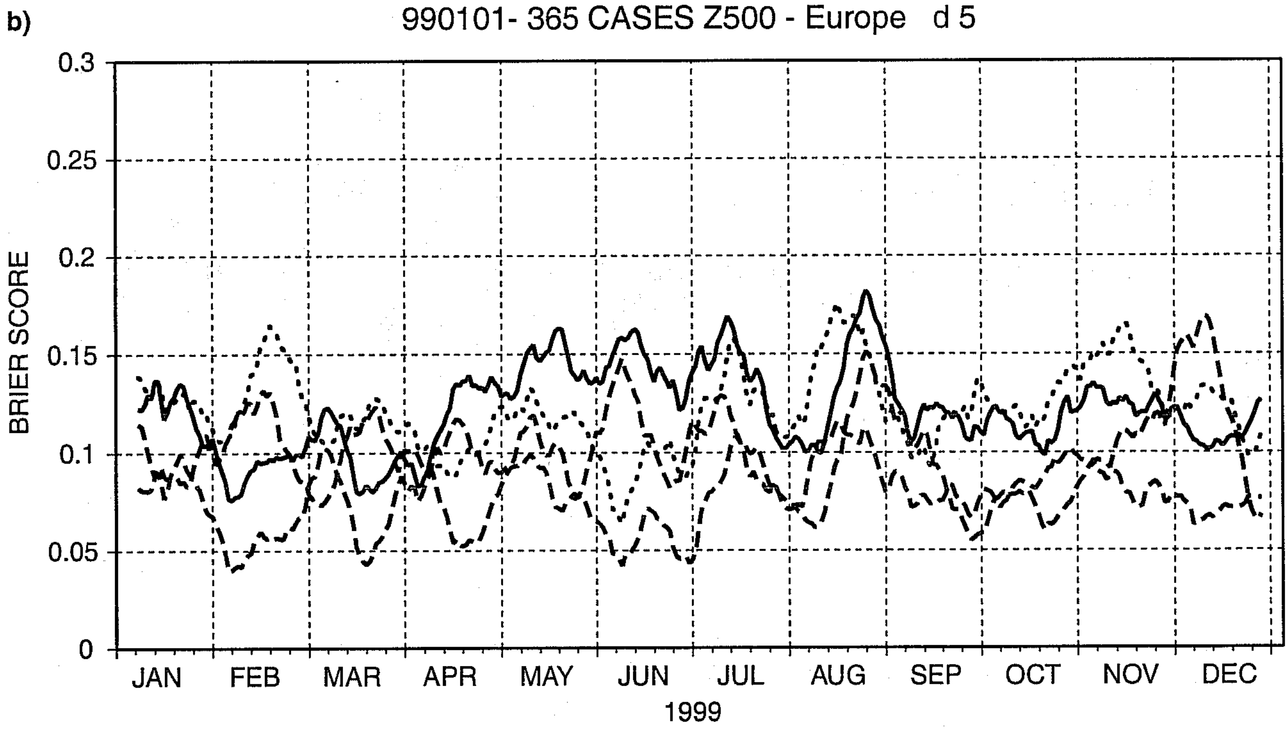


Fig. 5 As Fig. 4 but for the 14-day mean Brier score.

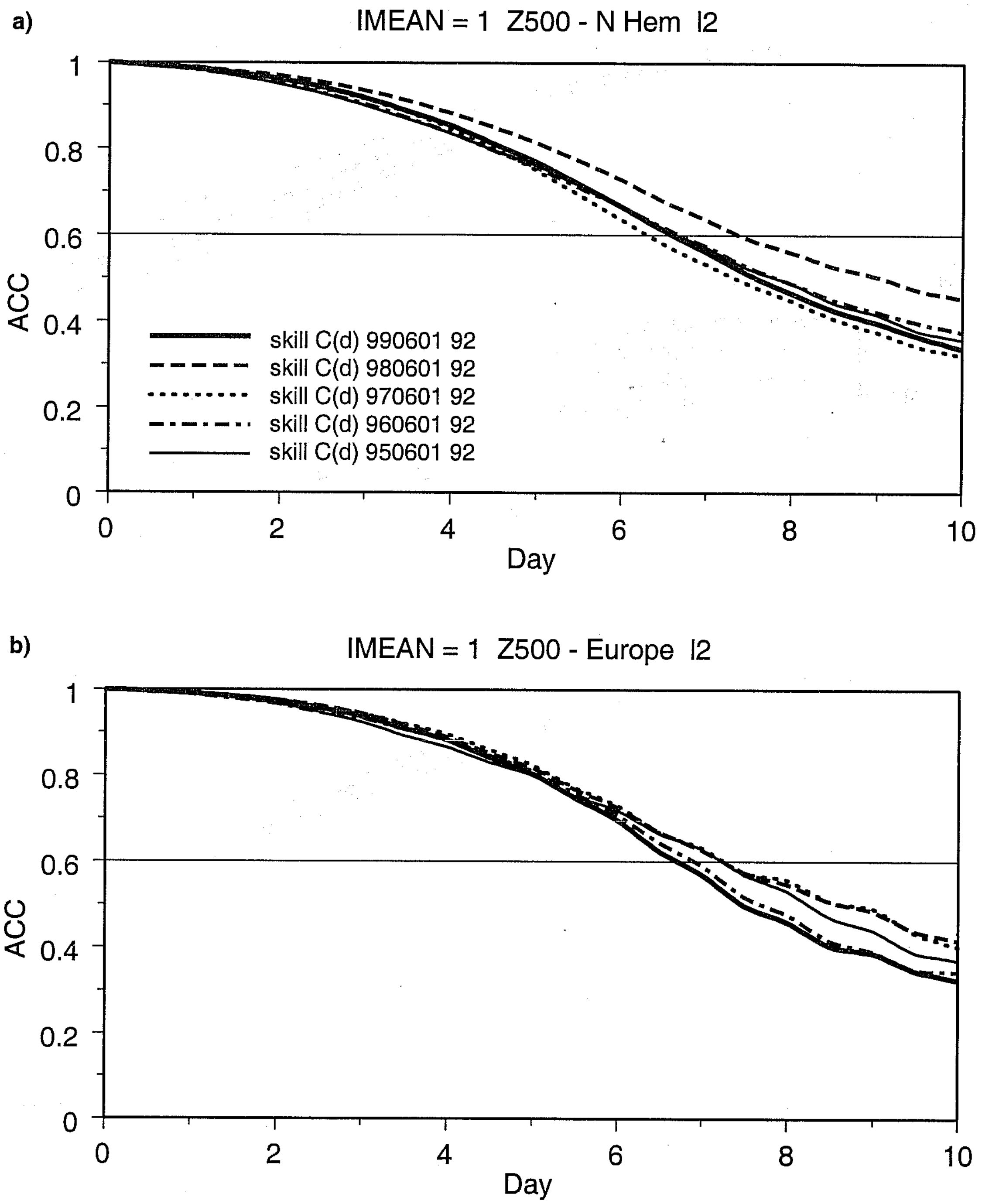
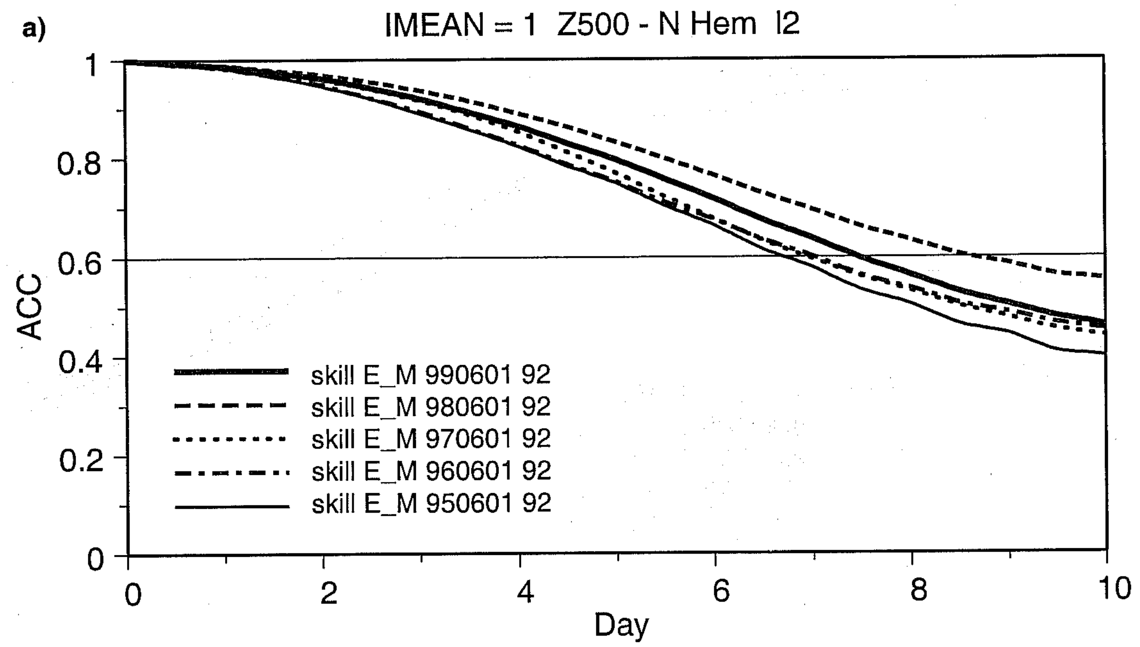


Fig. 6 Seasonal average anomaly correlation for the EPS control for summer 1995 (thin solid line), 1996 (chain-dashed line), 1997 (dotted line), 1998 (dashed line) and 1999 (solid line), for (a) the NH and (b) Europe. Results refer to the 500-hPa geopotential height.



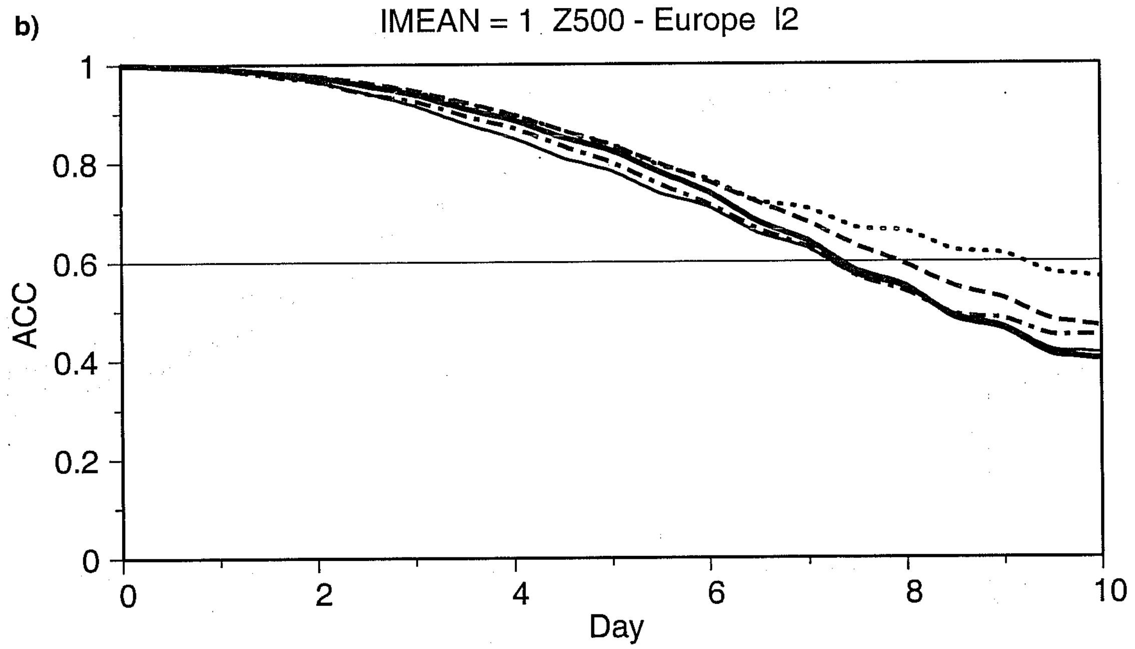
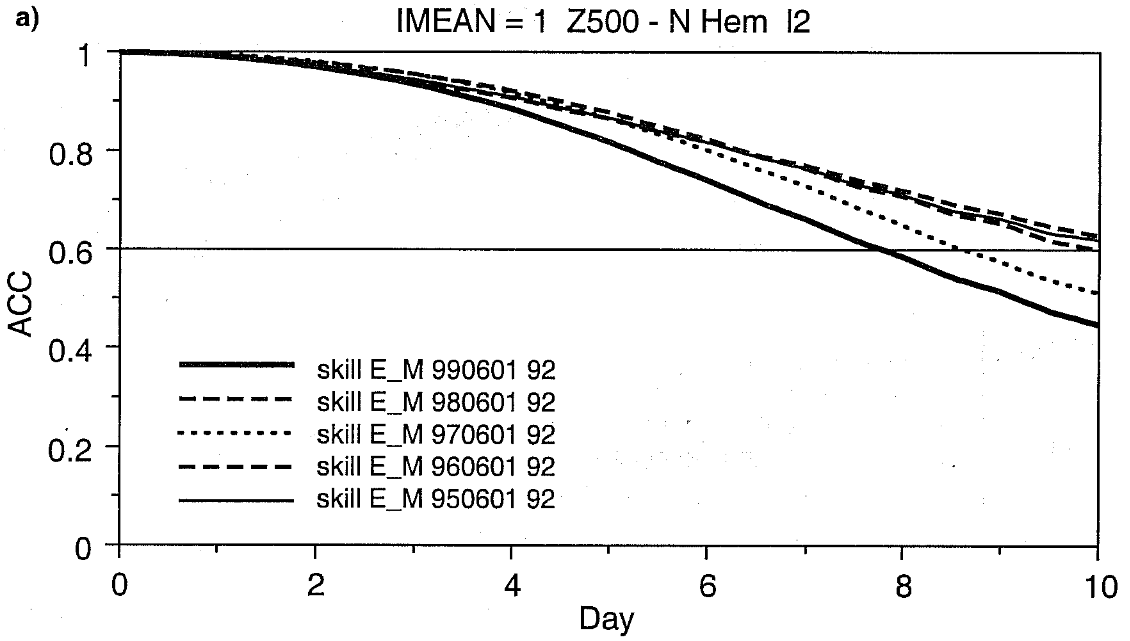


Fig. 7 As Fig. 6 but for the ensemble mean.



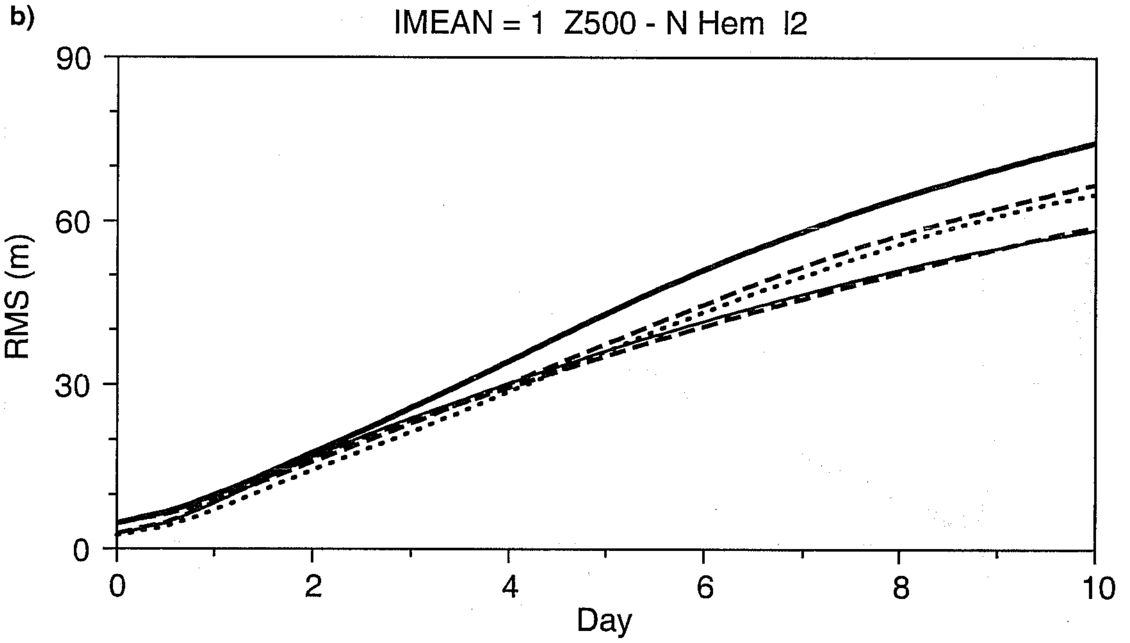
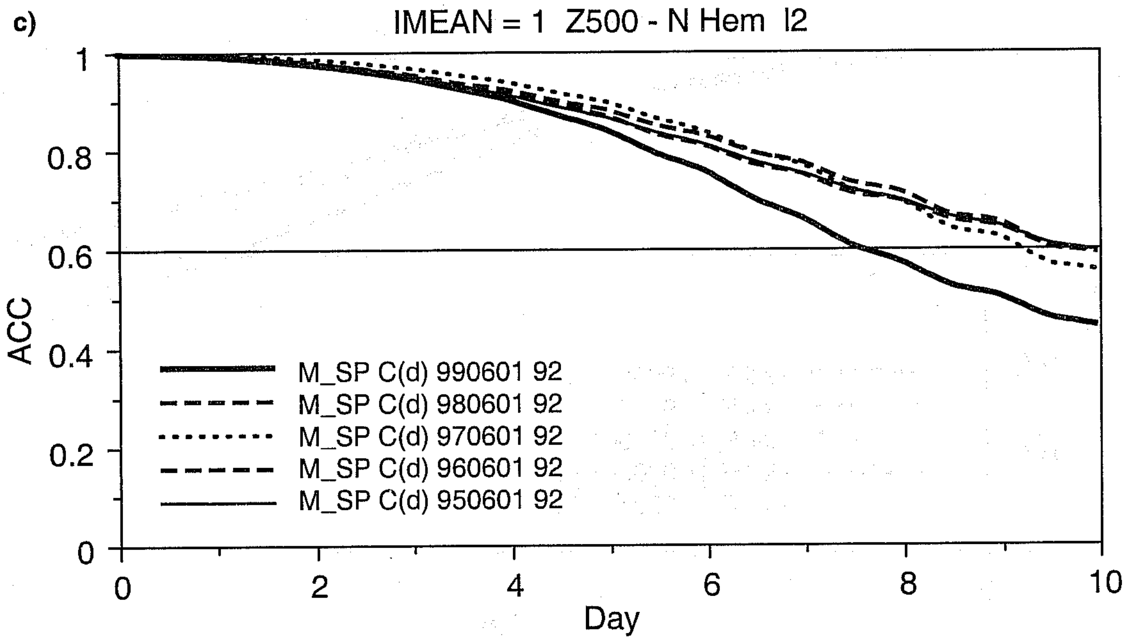


Fig. 8 As Fig. 6 but for the ensemble spread.



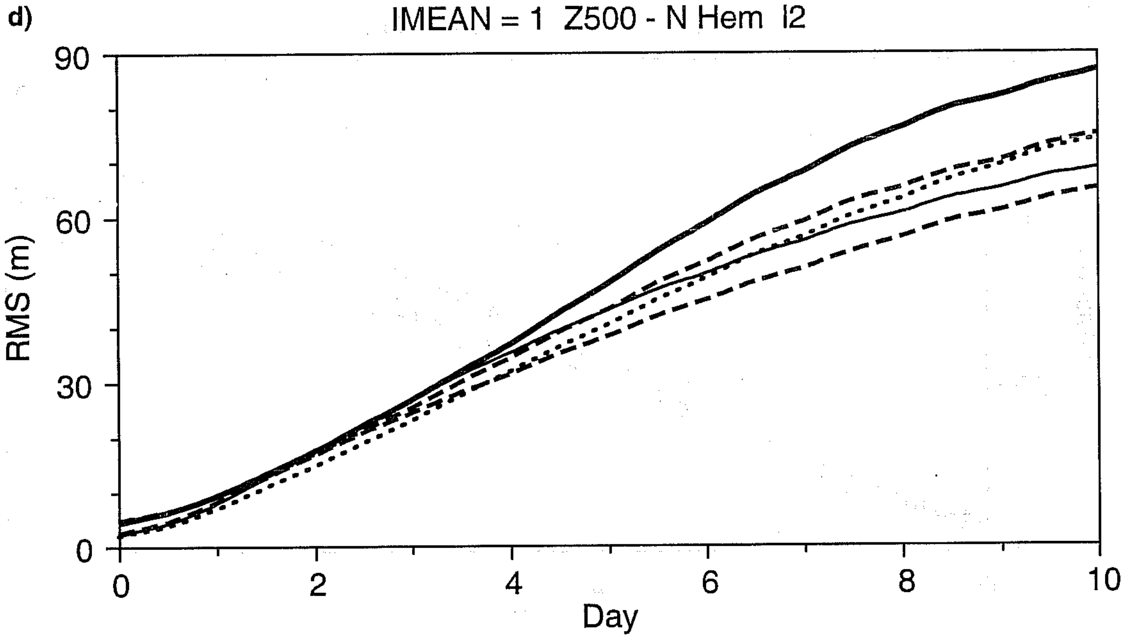
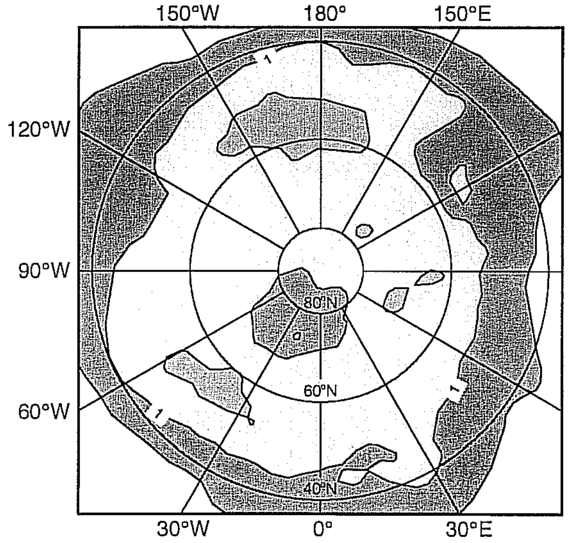
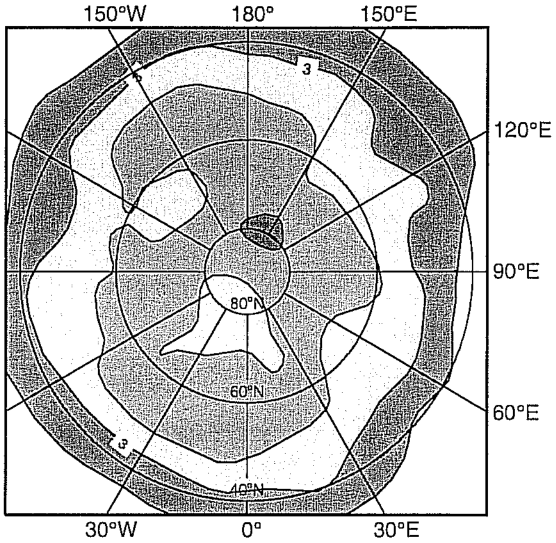


Fig. 8 continued

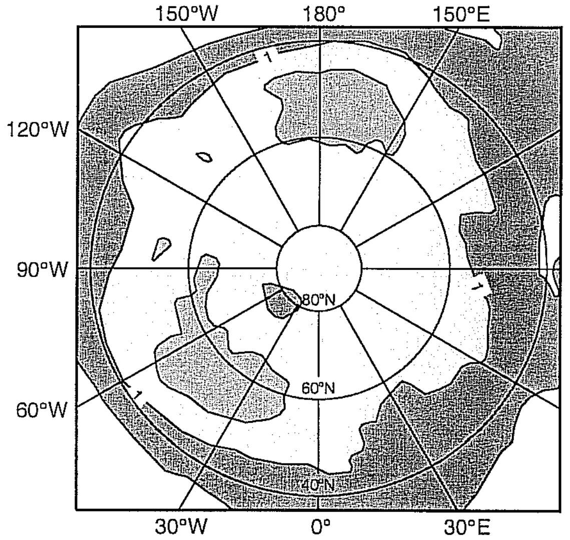
JJA 98 - EPS std Z500 t+24h



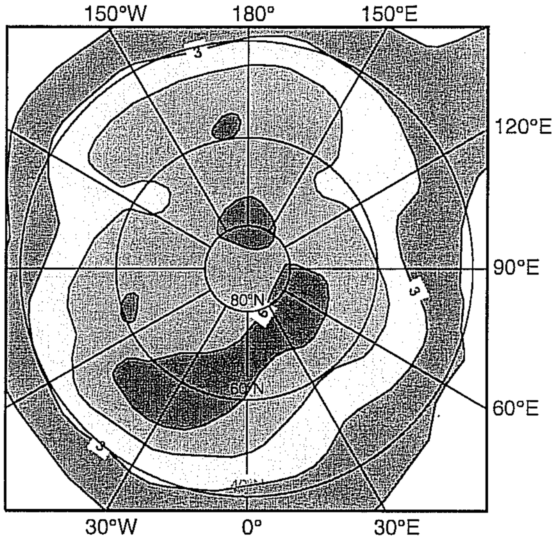
JJA 98 - EPS std Z500 t+120h



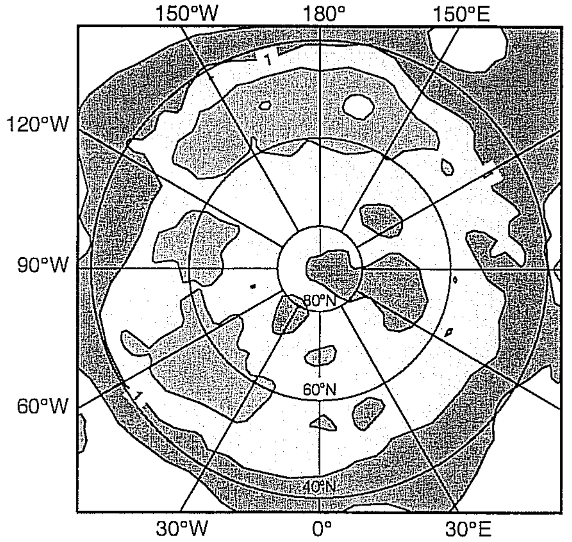
JJA 99 - EPS std Z500 t+24h



JJA 99 - EPS std Z500 t+120h



AUG 99 - EPS std Z500 t+24h



AUG 99 - EPS std Z500 t+120h

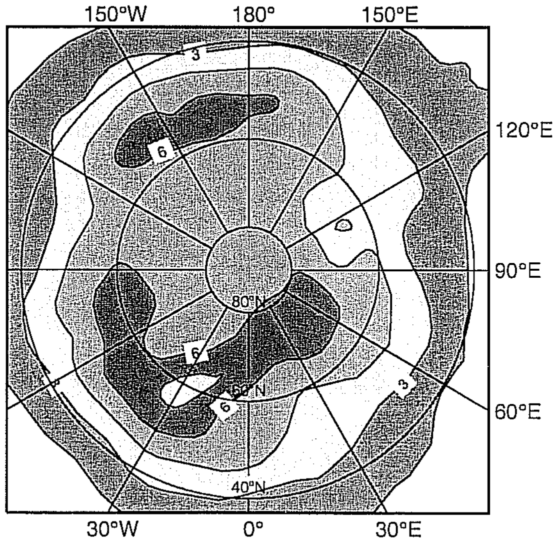


Fig. 9 EPS average standard deviation (a) for JJA 1998 at forecast t+24h; (b) as (a) but at forecast t+120h; (c-d) as (a-b) but for JJA 1999; (e-f) as (a-b) but for August 1999. Contour interval is 0.5 dam with shading every 0.5 for t+24h, and 1.5 dam with shading every 1.5 dam for t+120h. Results refer to the 500 hPa geopotential height field.

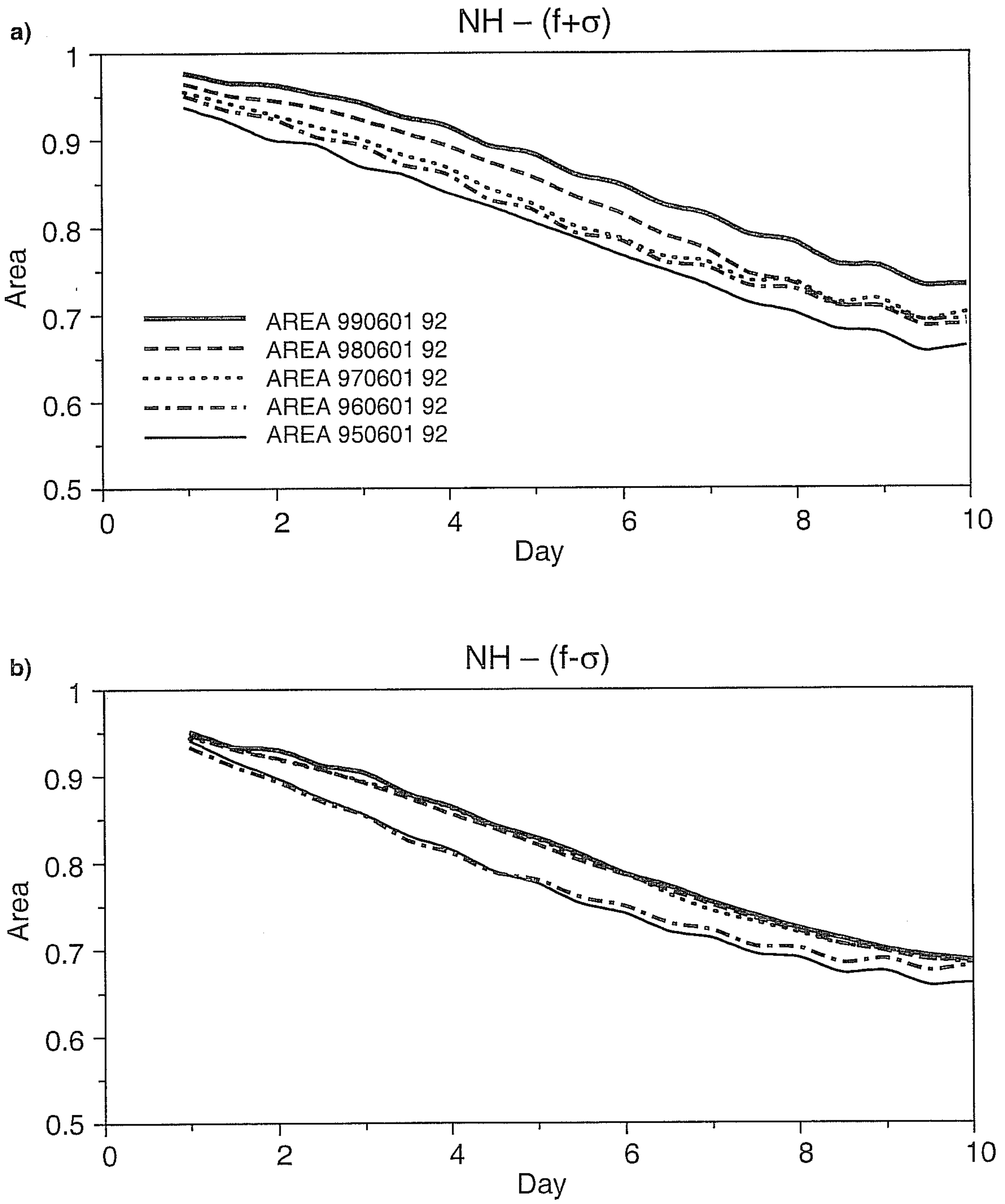


Fig.10 (a) ROC-area for the prediction of positive anomalies larger than one standard deviation for summer 1995 (thin solid line), 1996 (chain-dashed line), 1997 (dotted line), 1998 (dashed line) and 1999 (solid line), for the NH. (b): as (a) but for negative anomalies smaller than one standard deviation.

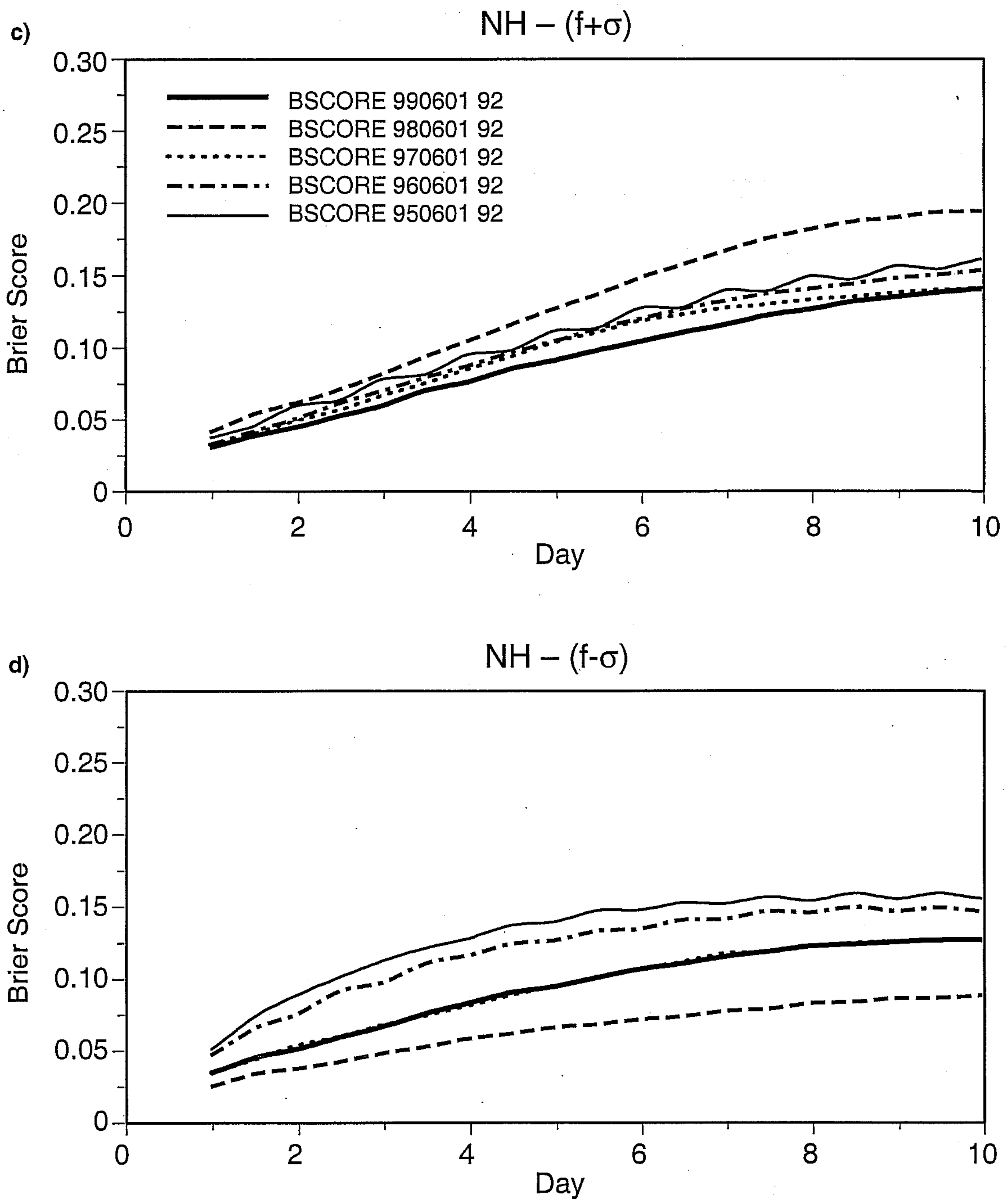


Fig.10 (c-d): as (a-b) but for the Brier score.

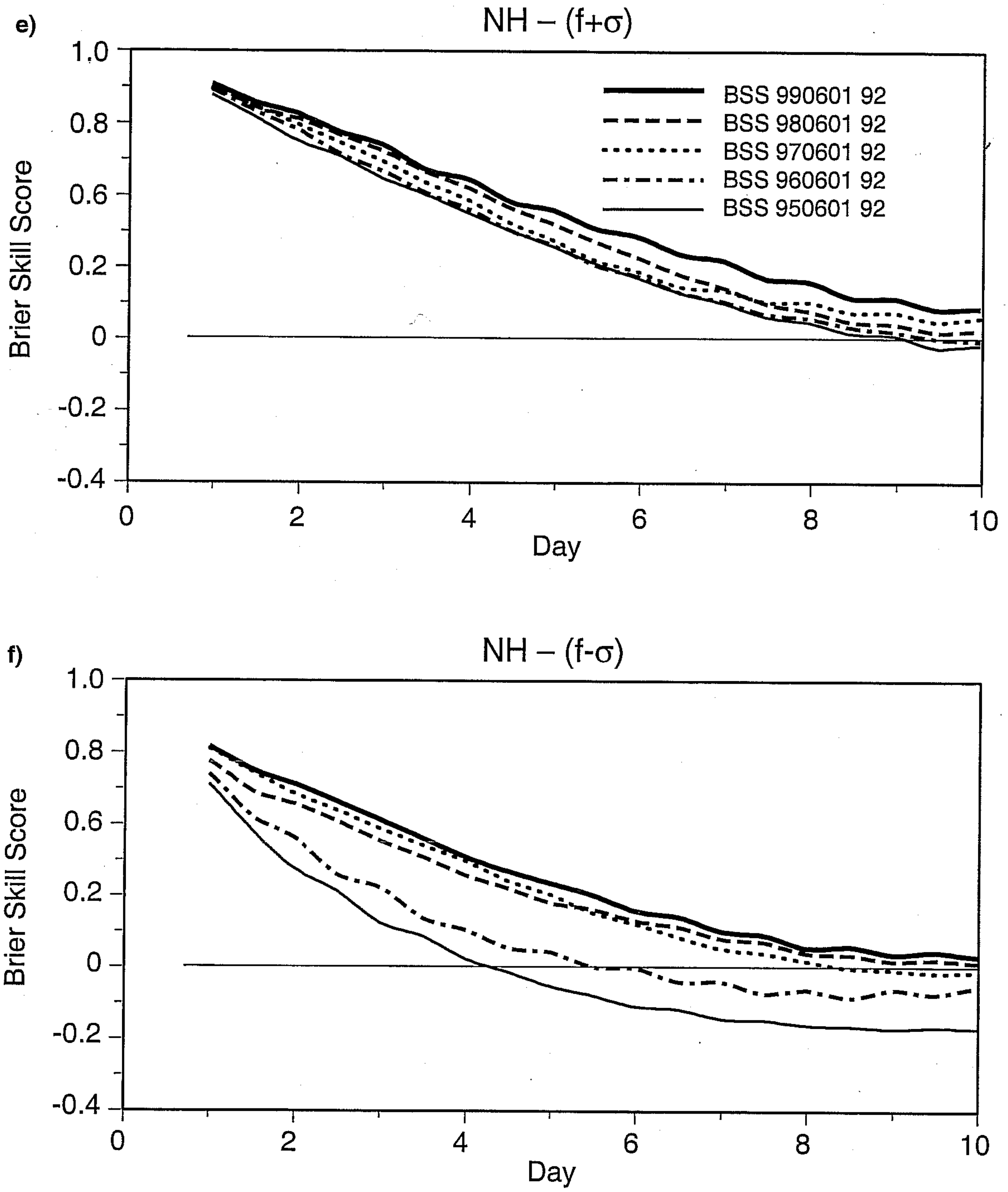
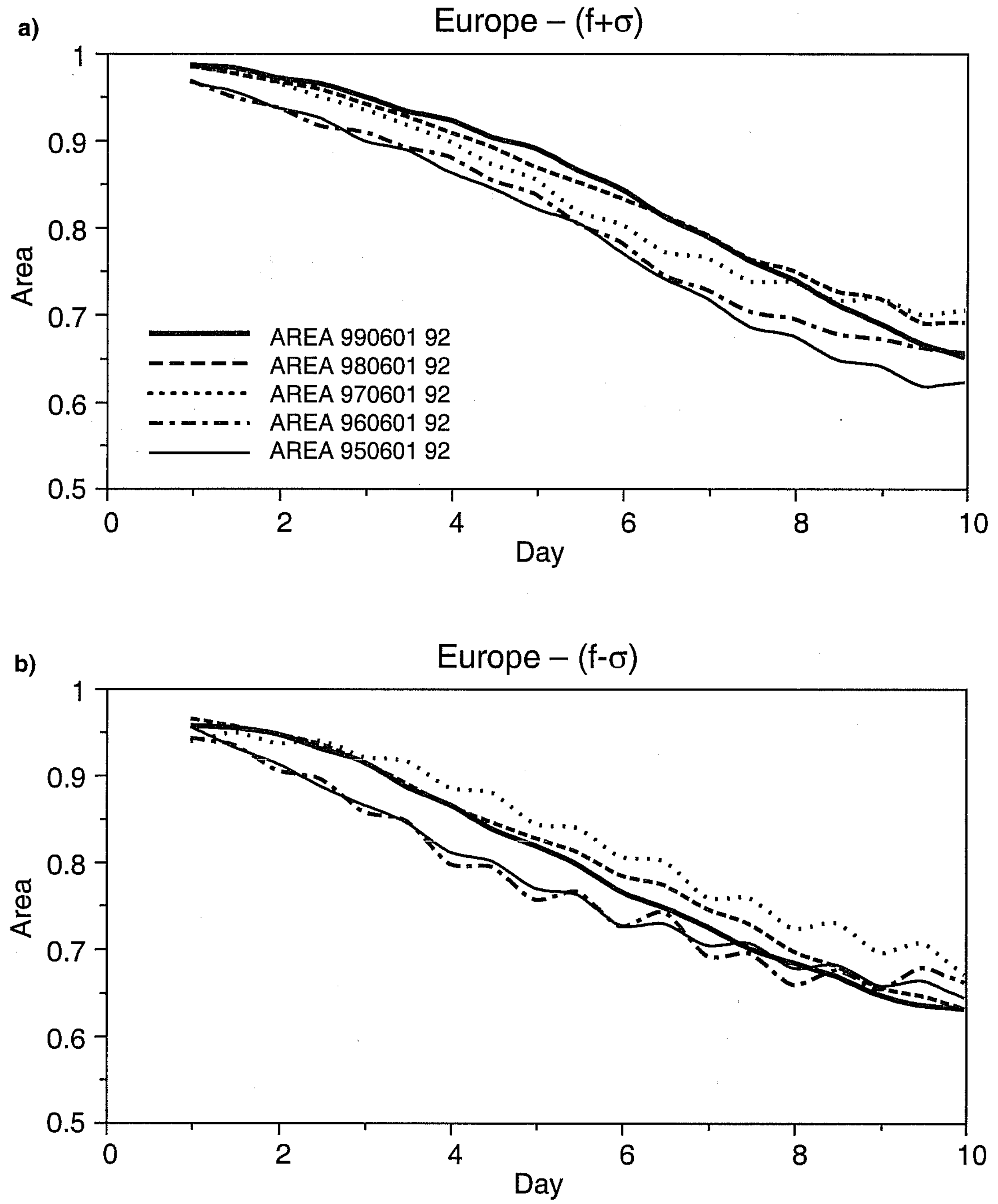


Fig.10 (e-f): as (a-b) but for the Brier skill score.



a) Europe – (f+σ)

b) Europe – (f-σ)

Fig. 11 As Fig. 10 but for Europe

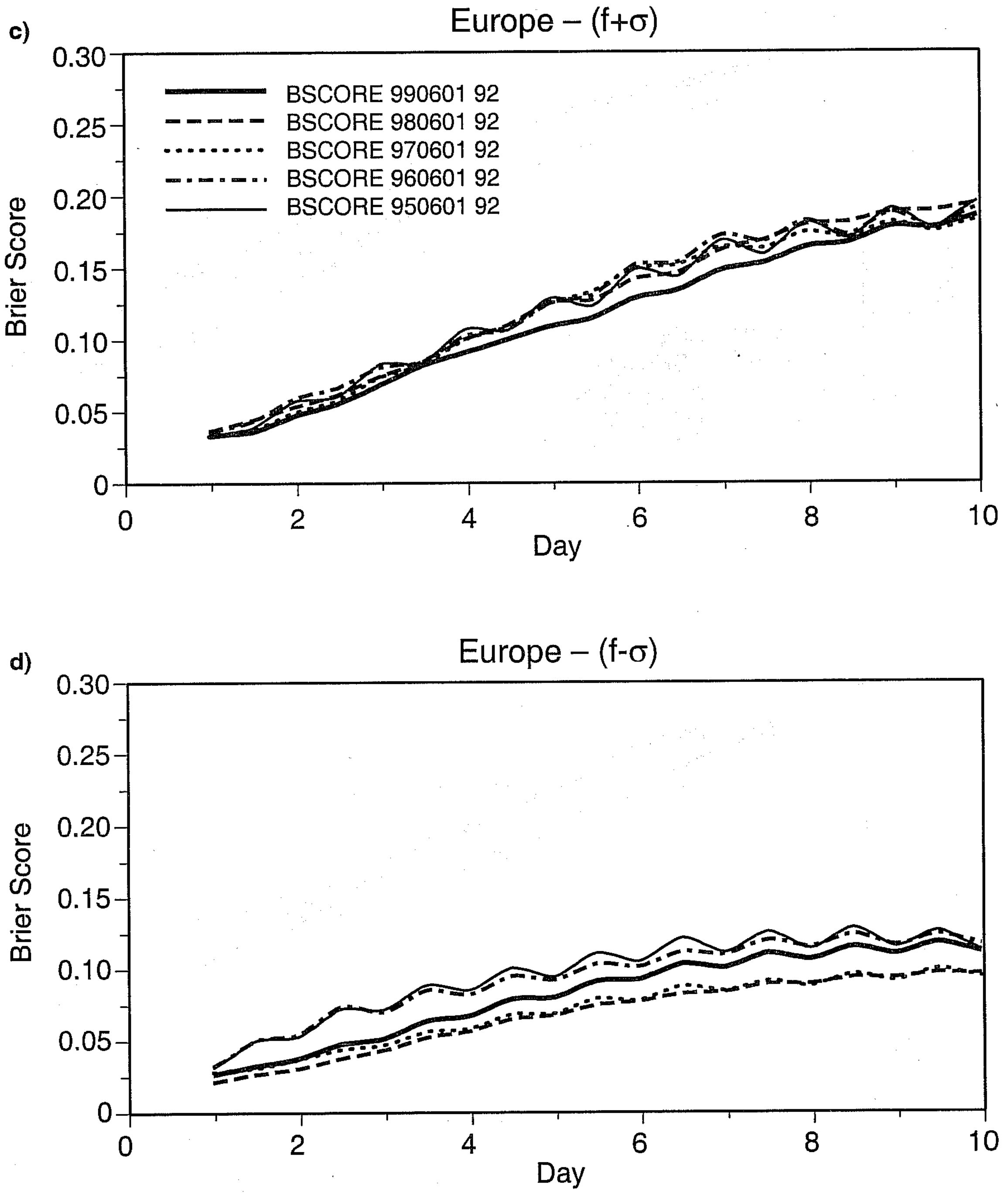
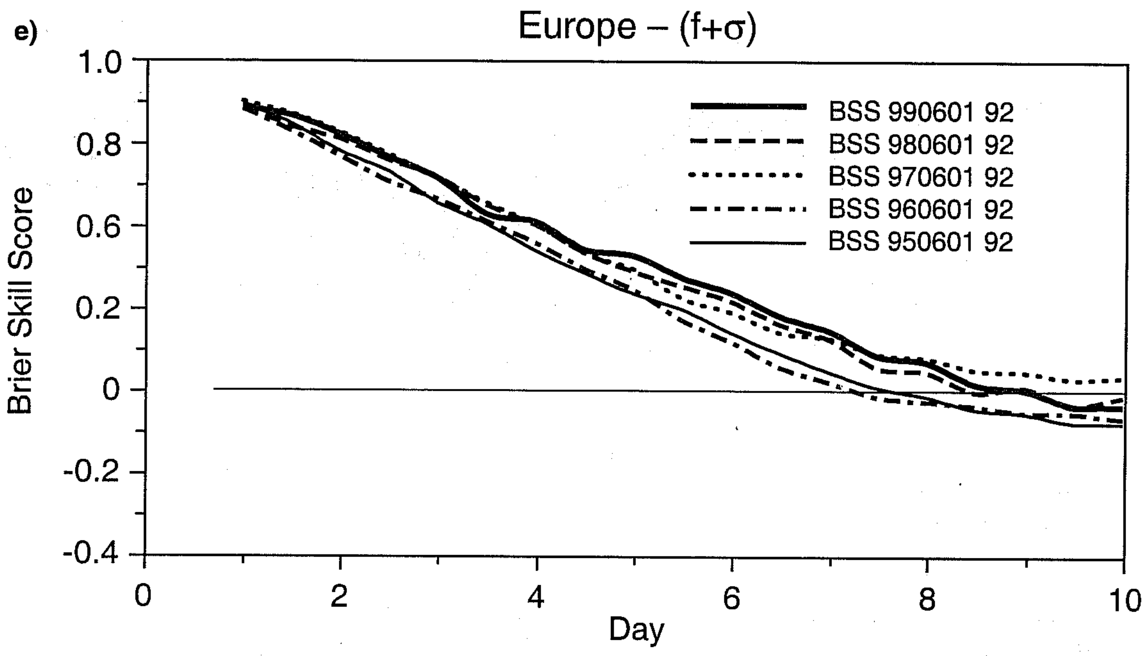


Fig.11 (c-d) As Fig. 10 but for Europe



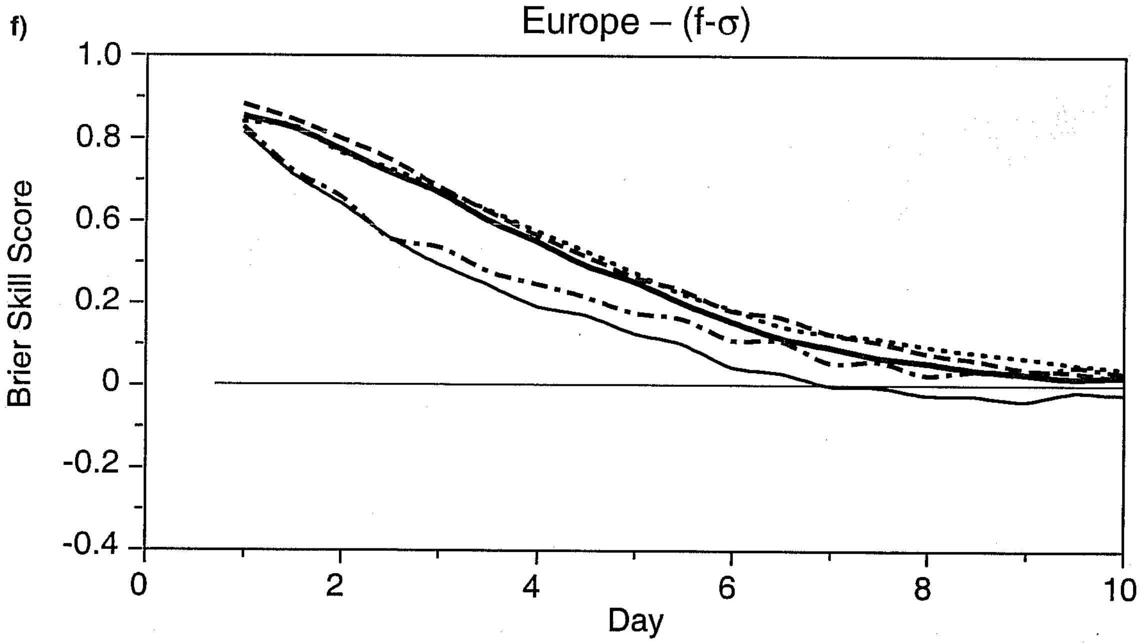


Fig.11 (e-f) As Fig. 10 but for Europe

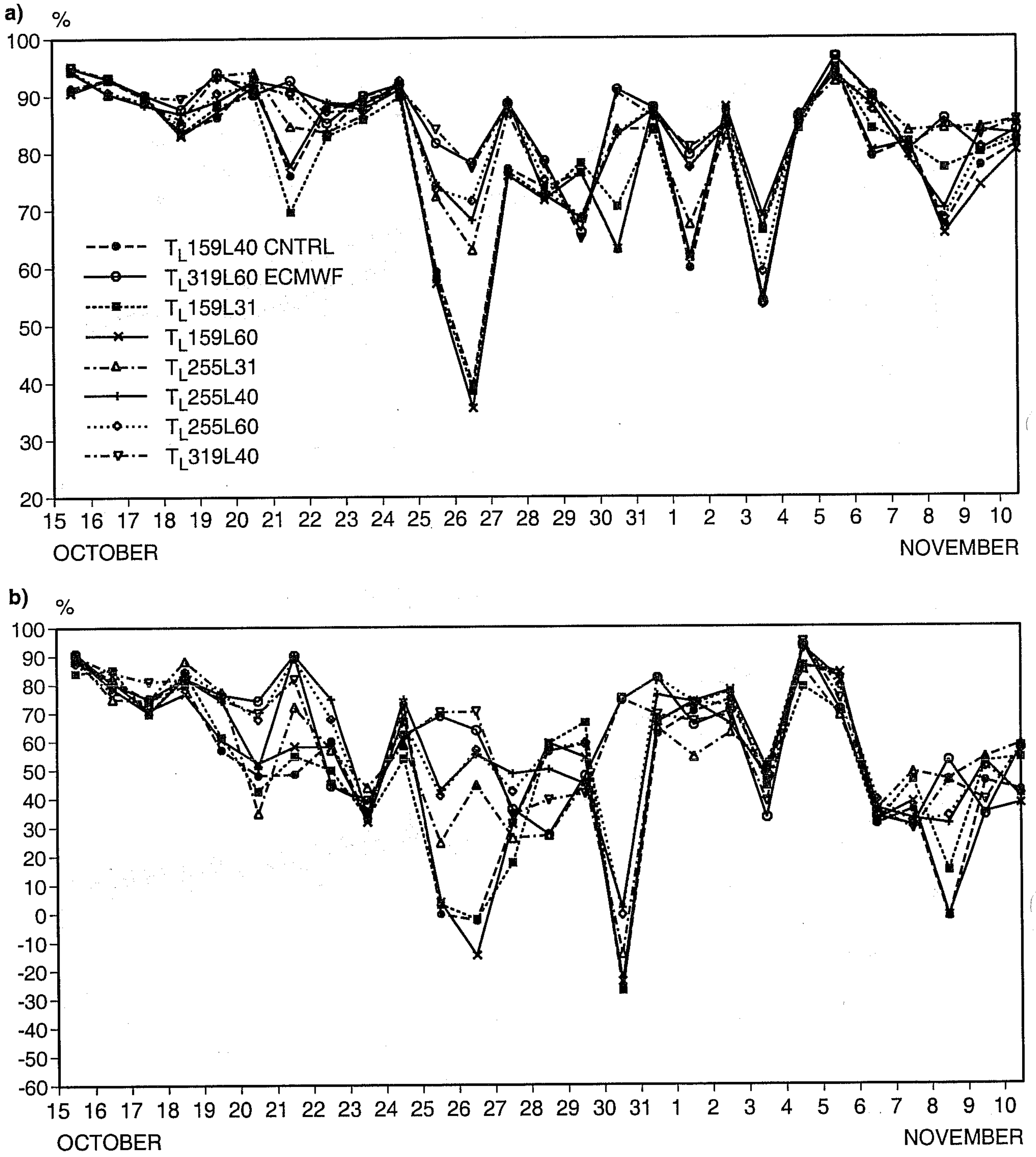


Fig. 12 Anomaly correlation skill score over Europe at (a) forecast day-5 and (b) at forecast day 7.

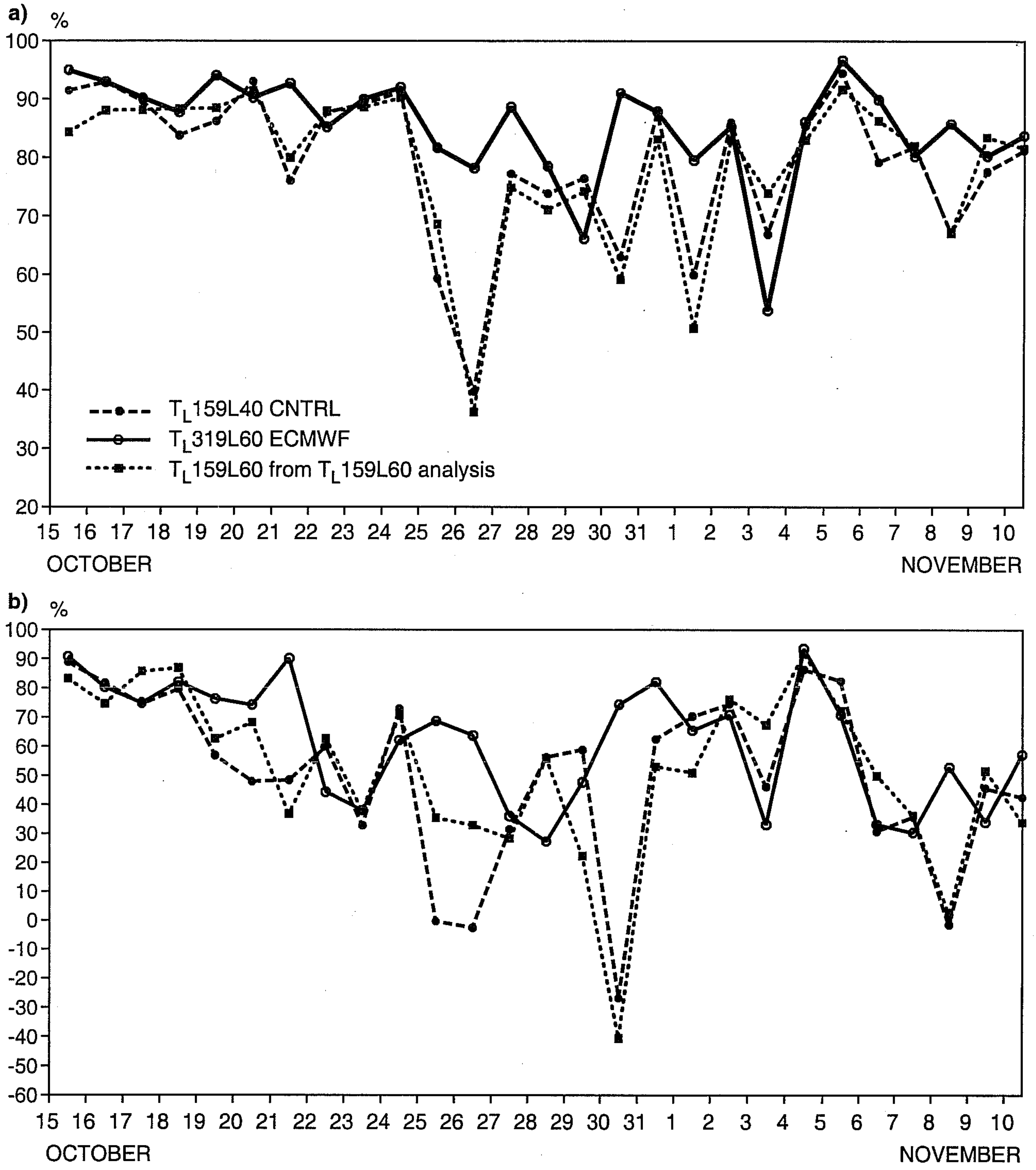


Fig. 13Anomaly correlation skill score over Europe at (a) forecast day-5 and (b) at forecast day 7.