TECHNICAL MEMORANDUM

248

Probability precipitation prediction using the ECMWF Ensemble Prediction System

R. Buizza and A. Hollingsworth

Research Department

February 1998

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

Probability precipitation prediction using the ECMWF Ensemble Prediction System

Roberto Buizza and Anthony Hollingsworth

European Centre for Medium-Range Weather Forecasts

Submitted to *Weather and Forecasting* on 29 January 1998.

ABSTRACT

The forecast skill of the European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (ECMWF EPS) in predicting precipitation probabilities is discussed. Four seasons are analyzed in details using Signal Detection Theory and reliability diagrams to define objective measure of predictive skill.

First, the EPS performance during summer 1997 is discussed. Attention is focused on Europe and two European local regions, one centred around the Alps and the other around Ireland. Results indicate that for Europe the EPS can give skilful prediction of low precipitation amounts (i.e. lower than 2mm/12h) up to forecast day 6, and of high precipitation amounts (i.e. between 2 and 10mm/12h) up to day 4. Lower levels of skill are achieved for smaller local areas.

Then, the EPS performance during summer 1996 (i.e. prior to the enhancement introduced on 10 December 1996 from 33 to 51 members and to resolution increase from T63L19 to T_L159L31) and summer 1997 are compared. Results show that the EPS has been remarkably more skilful during summer 1997 than summer 1996, with gain in predictability up to 3 days for the highest (5 and 10mm/12h) amounts of precipitation.

Finally, the EPS performance during wintertime is analyzed. Two issues are investigated, the seasonal variability of the forecast skill of the new EPS, and the impact of the system upgrade on the wintertime performance. The comparison of the performance of the new EPS system during winter 1996/97 and during summer 1997 indicates that the EPS is more skilful during winter than during summer, with differences in predictive skill around 3 days for precipitation amounts larger than 2mm/12h. The comparison of the EPS performance before and after the system upgrade on 10 December 1996 during winter confirms the summer conclusion that the upgraded system is more skilful than the old one.

1. INTRODUCTION

Since December 1992, two major Numerical Weather Prediction Centres have been producing ensemble predictions to complement single deterministic forecasts with probabilistic predictions (*Tracton and Kalnay*, 1993, *Palmer et al*, 1993).

Verifications of probabilistic forecasts of weather events at mid tropospheric level have demonstrated the usefulness of ensemble systems in predicting synoptic scale flows (*Molteni et al*, 1996, *Buizza*, 1997a-b, *Toth et al*, 1997 and 1998). An analysis of probabilistic precipitation prediction during three cases of extreme precipitation over Europe (*Petroliagis et al*, 1997) has shown that ensemble prediction can be extremely beneficial for assessing the forecast skill of a single deterministic forecast. This result is supported by the analyses of several cases of probabilistic quantitative precipitation prediction using the NCEP ensemble system (*Zhu et al*, 1998).

The focus of this study is a statistical analysis of the performance of the ECMWF EPS in predicting probabilities of precipitation events during summertime and wintertime. The EPS performance is analysed by applying Signal Detection Theory (*Mason*, 1982; *Stanski et al*, 1989), and reliability diagrams and their related Brier score and Brier skill score (*Brier*, 1950).

The ECMWF EPS underwent a major system upgrade in December 1996, with the ensemble size enhanced from 33 to 51, and the resolution increased from T63L19 to T_L159L31 (*Buizza et al*, 1997). We also assess whether this system upgrade improved the EPS skill in predicting precipitation probabilities.

Precipitation forecasts and verification are compared on a 2.5°x2.5° grid. Three geographical areas are considered: Europe (30°N≤ϕ≤80°N, 20°W≤λ≤45°E, 513 grid points), an Alpine region (42°N≤ϕ≤48°N, 5°E≤λ≤18°E, 18 grid points) and a region surrounding Ireland (50°N≤ϕ≤56°N, 12°W≤λ≤2°W, 15 grid points). Precipitation is accumulated over a 12 hour period, and forecasts up to day 10 are verified with verification defined using the 12h and 24h forecasts from the ECMWF deterministic high resolution (T213L31) model. Specifically, the precipitation accumulated between 12Z and 00Z of day d has been defined as the 12h forecast of the T213L31 forecast started at 12Z of day d, and the precipitation accumulated between 00Z and 12Z of day d+1 has been defined as the difference between the 24h and the 12h forecast of the T213L31 started at 12Z of day d. Four events are considered, i.e. "12h accumulated precipitation greater than 1, 2, 5 and 10mm/12h".

In the first part of this work attention is focused on summer. First, some case studies from summer 1997 are analyzed, to illustrate the relationship between maps of forecast probabilities and a measure of skill defined using Signal Detection Theory. Next, the average EPS performance during summer 1997 over the three regions defined above is studied. Then, the impact of the system upgrade introduced on 10 December 1996 on summertime skill is analyzed by comparing the EPS skill during summer 1996 and 1997.

In the second part of this work attention is focused to wintertime performance over the European area only. First, the EPS performance during winter 1996/97 is discussed, and the seasonal variability of the EPS forecast skill in predicting precipitation probabilities is investigated by comparing the EPS skill during winter 1996/97 and summer 1997. Then, the impact of the system upgrade is analysed by comparing the EPS skill during winter 1995/96 and 1996/97.

The paper is organized as follows. Section 2 describes the verification methodology. Section 3 discusses some case studies from summer 1997. Section 4 analyzes the average EPS performance in predicting probabilities of precipitation amounts during summer 1997. Section 5 compares the performance of the EPS during summer 1996 and summer 1997, i.e. prior and after the membership enhancement from 33 to 51 and the resolution increase from T63L19 to T_L159L31. Section 6 compares the EPS performance during winter 1995-96 and winter 1996-97. Conclusions are drawn in Section 7.

2. VERIFICATION METHODOLOGY

2.1 Signal Detection Theory

A deterministic precipitation prediction can be used to construct a contingency table of yes/no forecast and yes/no observed events, as follows

	yes-forecast	no-forecast
yes-observed	X	Y
no-observed	Z	W

where the event is that precipitation exceeds a certain threshold. The *hit rate* is defined as the ratio of the number of correct yes-forecast event (X) over the total number of yes-observed events (X+Y). The *false alarm rate* is defined as the ratio of the number of false alarms (Z) over the total number of no-events (Z+W). For a given contingency table, the hit rate and the false alarm rate can be plotted as a single point on a hit rate/false alarm rate graph.

Considering a contingency table for a deterministic forecast of a specific event, e.g. the event "12h accumulated precipitation greater than 1mm/12h", skill can be measured by the Threat Score *TS* defined as

$$TS = \frac{X}{X+Y+Z} \quad (1)$$

i.e. as the number of correct "yes" forecasts divided by the total number of occasions on which the event was forecast and/or observed (*Wilks*, 1995).

Signal Detection Theory (*Mason*, 1982) is an extension of this methodology to probabilistic forecasts. A probabilistic forecast provides a set of contingency tables, one for each probability interval of precipitation exceeding a given threshold. Consequently, a set of points can be plotted on a hit rate/false alarm rate graph. These points define a curve called the Relative Operating Characteristic (ROC). Following *Stanski et al* (1989), one can define a measure of forecast accuracy as the area under a ROC curve, which decreases from 1 for a perfect probabilistic forecast system to 0.5 for a useless forecast system: a perfect forecast has a 100% hit rate and a 0% false alarm rate, while a zero-skill forecast system has a ROC curve that coincides with the diagonal and thus has an area of 0.5 corresponding to the same number of hits and false alarm rates.

Figures 1a-b show two ROC curves for summer-97 relative to the events "12h accumulated precipitation greater than 1mm/12h" and "12h accumulated precipitation greater than 10mm/12h", at forecast day 3 for Europe. The area under the ROC curves for the lower/higher thresholds are, respectively, 0.83 and 0.70. As forecast time progresses, the prediction accuracy on average deteriorates and the area under the ROC curve decreases. At forecast day 7 (not shown), the area under the curves for the two events is 0.72 and 0.54. It is common practice (*Stanski et al*, 1989) to consider an area under a ROC curve of more than 0.80 to be indicative of a good prediction system, and an area of 0.70 as the limit for a useful prediction system. This common choice of 0.70 as a limit of useful probability prediction will be supported by the case studies in Section 3. Note that a square symbol has been plotted in Figs. 1a-b to show the hit and false alarm rates for the T213L31 model.

Again following *Stanski et al* (1989), one can define a second measure of forecast performance by considering the conditional distributions of forecast probabilities given the occurrence and non-occurrence of the event under investigation. The further apart the two distributions are, the better the forecast system is in discriminating between occurrence and non-occurrence. A measure of the distance between the two conditional distributions is given by the distance between the means of the two distributions normalized by the standard deviation of one of them, usually of the distribution of non-occurrences. Our experience suggests that a normalized distance of 0.5 should be considered as the limit for a sufficient discrimination.

Figures 1c-d show, respectively, the conditional distributions relative to the ROC curves shown in Figs. 1a-b. The normalized distance between the two distributions for the lower/higher threshold are, respectively, 1.25 and 0.45. Just as for the area under the ROC curve, the normalized distance between the two distributions usually decreases as forecast time progresses. At forecast day 7, for example, the distance between the two distributions is, respectively, 0.44 and 0.45 (not shown). Figures 1c-d show that only for the lower threshold at forecast day 3 are the two distributions well separated.

2.2 Reliability diagrams and Brier score and Brier skill score

Considering a probability forecast of an event as "12h accumulated precipitation greater than of 1mm/12h", stratified in 10% intervals, a reliability diagram can be constructed by computing for each probability forecast the frequency of observation of the event, and by plotting for each forecast probability the frequency of observation.

Figures 1e-f show the summer-97 average reliability diagrams relative to the events "12h accumulated precipitation greater than 1mm/12h" and "12h accumulated precipitation greater than 10mm/12h", at forecast day 3 for Europe. Figure 1e shows that for the lower threshold of 1mm/12h the EPS is quite accurate in predicting low probabilities, but tends to underestimate the high probabilities. Considering the higher threshold of 10mm/12h (Fig. 1f), the EPS severely over predicts high probabilities.

The Brier score is a measure of forecast accuracy which can be computed from a reliability diagram. Following the decomposition of *Murphy* (1973), the Brier score can be computed as the sum of three terms (see *Stanski et al*, 1989, for more detail) related to reliability, resolution and uncertainty,

$$BS = BS_{rel} + BS_{res} + BS_{unc} \tag{2a}$$

with

$$BS_{rel} \equiv \frac{1}{N} \sum_{k=1}^{T} \sum_{i=1}^{M_k} (f_k - \langle o \rangle_k)^2 \tag{2b}$$

$$BS_{res} \equiv \frac{1}{N} \sum_{k=1}^{T} \sum_{i=1}^{M_k} (\langle o \rangle_k - \langle o \rangle)^2 \tag{2c}$$

$$BS_{unc} \equiv \frac{1}{N} \sum_{k=1}^{T} \sum_{i=1}^{M_k} (\langle o \rangle - o_{ki})^2 \tag{2d}$$

where N is the total number of forecasts, the forecast probabilities f_k are allowed to take discrete values in 10% intervals from 0% to 100% (thus, with this choice, $T=10$), o_{ki} is one if the event occurred and zero if the event did not occurred, $<o>_k$ is the average frequency of occurrence for the k-*th* category, $<o>$ is the average frequency of occurrence over the whole sample.

The reliability term BS_{rel} measures the degree to which the forecast probabilities agree with the observed frequency of occurrence in the sample, and thus BS_{rel} can be considered as a measure of the system reliability, i.e. of the degree of under or over forecasting. The resolution term BS_{res} measures the degree to which the relative frequencies of the event in the T sub-samples differ from the relative frequencies of the event in the whole sample. The uncertainty term BS_{unc} is the variance of the observations in the sample. The reader is referred to *Stanski et al* (1989) for a complete discussion on the relative importance of the three terms.

The Brier skill score is defined as the Brier score percentage improvement of the EPS forecast with respect to climatology,

$$BSS \equiv \frac{BS_{cli} - BS}{BS_{cli}} \tag{3}$$

As an example, Table 1 reports the contribution to the Brier score of the three different terms, and the Brier skill score for the reliability diagrams shown in Figs. 1e-f.

	BS_{rel}	BS_{res}	BS_{unc}	BS	BSS
prec > 1mm/12h	0.006	-0.042	0.15	0.11	0.24
prec > 10mm/12h	0.003	-0.001	0.017	0.019	-0.15

Table 1: Brier scores and skill scores of the reliability diagrams shown in Figs. 1e-f, relative to the prediction of the events "12h accumulated precipitation greater than 1 and 10mm/12h" at forecast day 3 for summer 1997 over Europe.

3. CASE STUDIES

Two cases of probability prediction for the events "12h accumulated precipitation greater than 2mm/12h and 10mm/12h" over Europe are analyzed in this Section, to familiarize the reader with the skill measure introduced in Section 2, namely the area under a ROC curve. These two case studies illustrate the relationship between probability prediction, observed precipitation, and ROC area.

3.1 Predictions of the 12h accumulated precipitation between 00Z and 12Z of 11 August 1997

Figure 2 shows the probability of occurrence of the event "12h accumulated precipitation greater than 2mm/12h", given by the day 7, day 5 and day 3 forecasts started, respectively, on 4, 6 and 8 August 1997 for the precipitation accumulated between 00Z and 12Z of 11 August. The three forecasts are characterized, respectively, by areas under the ROC curve of 0.50, 0.75 and 0.83.

The major change in the forecast skill happens between the day 7 and the day 5 forecasts. The day 7 prediction gives probability of rain in regions, such as Northern Italy or Eastern Europe where precipitation did not occur, and it gives very low probabilities of rain (lower than 20%) where it actually occurred, e.g. around the Black Sea. By contrast, the day 5 forecast has a high probabilities of rain across the Pyrenees and around the Black Sea, regions where rainfall occurred. A further forecast improvement can be seen in the day 3 forecast.

It is worth noting how the increase of the ensemble spread with lead time affects the probability forecasts. At short lead times, higher values of probabilities are shown, indicating a reduction of the forecast uncertainty of the event under consideration.

Figure 3 shows the corresponding probability forecasts for the event "12h accumulated precipitation greater than 10mm/12h". For this event, the day 7, day 5 and day 3 forecasts are characterized by an area under the ROC curve of, respectively, 0.55, 0.60 and 0.78. Both the area under the ROC curve, and the comparison between the probability forecasts and the observed precipitation, indicate that only the day 3 forecast gives some useful indication of the region where heavy precipitation should be expected.

3.2 Predictions of the 12h accumulated precipitation between 00Z and 12Z of 21 June 1997

Figure 4 shows the probability prediction of the event "12h accumulated precipitation greater than 2mm/12h" given by the day 7, day 5 and day 3 forecasts started, respectively, on 14, 16 and 18 June 1997. These forecasts are characterized, respectively, by areas under the ROC curve of 0.70, 0.80 and 0.93.

The comparison of the day 7 and day 5 probability predictions suggests that the ROC curve for the day 5 forecast is higher because of the reduction of the wrong prediction over Norway, and because of the better prediction over Northern

France. The further improvement of the day 3 forecast is related to the very good agreement between the probability map and the observed precipitation over central Europe. (Since more than 10mm/12h was detected only in very few points, we do not discuss in details the probability prediction of this amount; for this event, the day 7, day 5 and day 3 forecasts all scored about 0.50.)

Concluding, these results support the common practice that a probability prediction with area under the ROC curve $A>0.70$ indicates a useful prediction.

4. PRECIPITATION PREDICTION IN EUROPE DURING SUMMER 1997

We now consider average results for the performance of the EPS in summer 1997 (defined as June, July and August), in Europe and in the Alpine and the Irish regions (Europe includes 513 grid points, while the Alpine and the Irish regions include only, respectively, 18 and 15 grid points).

4.1 Europe

Figure 5 shows the area under the summer-97 ROC curve for the prediction of the events "12h accumulated precipitation greater than 1, 2, 5 or 10 mm/12h", at forecast days 3, 5 and 7 for Europe. A time filter has been applied to the scores, so that the 5-day running mean (and not daily) values are plotted.

Figure 5a shows that, at forecast day 3, the EPS system has skill in predicting the events "12h accumulated precipitation greater than 1, 2 or 5mm/12h" (solid, dash and dot lines), while there are only few cases when the prediction of the event "12h accumulated precipitation greater than 10mm/12h" is accurate (chain-dash line). At forecast day 5 (Fig. 5b), in almost all cases the system is capable of giving useful probabilistic precipitation prediction for the 1 and the 2 mm/12h amounts, but predictions for higher amounts are rarely useful. At forecast day 7 (Fig. 5c), in only about half of the cases the system has ROC area greater than 0.7 for the predictions of the events "12h accumulated precipitation greater than 1 or 2mm/12h", while the ROC area is never greater than 0.7 for the event "12h accumulated precipitation greater than 10mm/12h". Similar conclusions can be drawn from Fig. 6, which shows the normalized distances between the conditional distributions relative to the area curves shown in Fig. 5.

Figures 7a-b show the evolution of the probabilistic forecast skill with forecast time, as measured by the area under the ROC curve and by the normalized distance between the conditional distributions associated with the ROC curve, for all forecast events. Figures 7a-b can be used to determine the forecast range for which the EPS forecast probabilities are useful, by noting the forecast time at which the 0.7 threshold is crossed first. For the events "12h accumulated precipitation greater than 2, 5 and 10mm/12h", this threshold is crossed first, respectively, at forecast days 6.5, 4.5 and 3 (Fig. 7a). Similar conclusions can be drawn from Fig. 7b by considering the forecast time at which the normalized distance between the conditional distributions first cross the 0.5 limit.

A similar analysis of the EPS performance in predicting precipitation probabilities can be based on reliability diagrams, and their associated Brier score and Brier skill scores. Summer-97 average reliability diagrams for two events "12h accumulated precipitation greater than 1mm/12h and 10mm/12h" at forecast day 3 were shown in Figs. 1e-f. A complete picture of the variation, with respect to the forecast time, of the summer-97 average Brier score and Brier skill score for the four events "12h accumulated precipitation greater than 1, 2, 5 or 10mm/12h", is shown in Figs. 7c-d.

Figure 7c shows that, for example, the Brier score of the 10mm/12h event is lower than the Brier score of the 1mm/12h, since the Brier score for rare events is smaller than the Brier score of frequent events (the Brier score is negative oriented, i.e. low Brier scores are better than high scores). In fact, rare events have a smaller number of non-zero terms in Eqs. (2b-d).

The fact that the Brier score of the event "12h accumulated precipitation greater than 10mm/12h" is smaller than the Brier score for the event 12h accumulated precipitation greater than 1mm/12h" does not mean that the prediction of the

higher precipitation amounts is more skilful than the prediction of the lower amounts. This can be seen by comparing the Brier skill scores for the two events (Fig. 7d), which shows that predictability is higher for the smaller amount. In fact, for events characterized by very different frequency of occurrence, the Brier skill scores, and not Brier scores, should be compared.

The comparison between the variation with forecast time of the area under the seasonal average ROC curve and the seasonal average Brier skill score suggests that qualitatively similar conclusions can be drawn from the two diagnostics. For example, both skill measures confirm that the EPS performs poorly in predicting more then 5mm of rain accumulated over 12 hours, but that useful conclusions can drawn on the occurrence of precipitation (1mm/12h threshold).

4.2 Case studies and average results in the Alpine region and Irish regions

The summer-97 record of observed precipitation (Fig. 8) for these regions shows extreme precipitation events occurred for the Alpine region between 00Z and 12Z on 19 June, 18 July, and 28 August, and for the Irish region between 00Z and 12Z on 24 June, 3 and 26 August. So, we shall examine the performance of the EPS for these events before discussing average results.

Figures 9a-b show, for each region, the daily values of the area under the ROC curve for the day 3 prediction of the four events "precipitation greater than 1, 2, 5 or 10 mm/12h". Comparison between the area under the ROC curve for either of the two small regions (Fig. 9) with the area under the ROC curve for Europe (Fig. 5a), shows that there is much more variability in the level of skill for the small areas. Generally speaking, for local regions like the two considered here, already at forecast day 3 ROC area values are smaller than 0.7 also for the event "precipitation larger than 1mm/12h".

Considering the three selected cases for the Alpine region (Fig. 9a), the 3-day forecasts started on 16/6 and verifying on 19/6, and the 3-day forecast started on 26/8 and verifying on 29/8 are characterized by areas under the ROC curve greater than 0.7 for all but the 10mm/12h threshold. By contrast, for the 3-day forecast started on 15/7 and verifying on 18/7 only the area for the 2mm/12h threshold is above the 0.7 line.

Even poorer predictive skill is shown during the three cases identified for the Irish region (Fig. 9b). In fact, only the forecast started on 23/8 and verifying on 26/8 has good predictive skill for the 1, 2 and 5mm/12h amounts of rain, while the 3-day forecasts started on 21/6 and verifying on 24/6, and the 3-day forecast started on 31/7 and verifying on 3/8 have low predictive skill, with the area under the ROC curve smaller than 0.70 for all thresholds.

We now consider seasonal average results for these two areas. Figures 10a-d show the variation, with respect to the forecast time, of the area under the summer-97 ROC curve and of the normalized distance between the two conditional distribution curves for the four events "12h accumulated precipitation greater than 1, 2, 5 or 10mm/12h". The comparison between the curves of the area under the summer-97 ROC curve for the two local regions (Fig. 10a-b) and Europe (Fig. 7a) confirms that the predictive skill of the system is lower for small regions. Considering, for example, the events "12h accumulated precipitation greater than 1mm/12h and 5mm/12h", the 0.7 threshold is reached first, respectively, at forecast days 10 and 4.5 for Europe while it is crossed first at forecast days 7 and 4 for the Alpine region, and at forecast days 4 and 3 for Ireland. These results are supported by the comparison of the normalized distances for the seasonal average ROC curves for the three regions (see Figs. 10c-d for the two local regions, and Fig. 7b for Europe).

A similar sensitivity to the choice of geographical region can be detected in the Brier score and Brier skill score (not shown).

5. COMPARISON OF THE EPS SKILL DURING SUMMER 1996 AND

SUMMER 1997

As mentioned in the Introduction, a major upgrade of the ECMWF EPS was introduced on 10 December 1996, when the number of perturbed members was enhanced from 32 to 50, and the model resolution was increased from T63L19 to T_L159L31 (*Buizza et al*, 1997). We now assess whether the upgrade improved the EPS performance in predicting precipitation amounts during summer. For reasons of space, attention is focused on the comparison of average results for the European, Alpine and Irish areas for the two events "12h accumulated precipitation greater than 2mm/12h and 10mm/12h" only. Results for the event with 1mm/12h are very similar to the ones for 2mm/12h, and results for the 5mm/12h amount lie somewhere between the ones for the 2mm/12h and the 10mm/12h amounts.

The area-average precipitation for Europe in 1997 was about 30% higher in 1997 than in 1996 (Table 2). Figures 11a-b show, for Europe, the variation with forecast time of the area under the summer ROC curves and the Brier skill score for 1996 (dash lines) and 1997 (solid lines). For the event with the lower amount of precipitation (2mm/12h, Figs. 11a-c), the performance of the EPS is better during 1997 up to forecast day 6. The comparison of the forecast time at which the 0.7 line is crossed first by the ROC area curves shows that, between 1996 and 1997, there has been a gain in predictive skill of 1 day for the event "12h precipitation amount greater than 2mm/12h". For the prediction of more than 10mm/12h of precipitation the EPS showed no predictive skill in 1996, while in 1997 it gave useful prediction up to day 3 (Fig. 11b). This represents a drastic improvement in forecast skill for the more intense events. The gain for the event "12h precipitation amount greater than 1mm/12h" has been of 3 days, and for the event "12h precipitation amount greater than 5mm/12h" has been of 2.5 days (not shown).

	Europe	Alpine Region	Irish Region
Summer 1996	0.6	0.9	0.8
Summer 1997	0.8	1.1	1.4
Winter 1995/96	1.0	--	--
Winter 1996/97	1.1	--	--

Table 2: Average amount of 12h accumulated precipitation (mm/12h), during summer and winter, over Europe, the Alpine and the Irish regions.

If, instead of the area under a ROC curve as measure of skill one considers the forecast time at which the Brier skill score curves first crosses the zero line, Figs. 11e-d indicate an even greater gain in predictive skill for the 2mm/12h amount of precipitation, and a gain of about 1 day for the events with the highest amount of precipitation (10mm/12h). (Results, not shown, relative to the 1 and 5mm/12h are similar, respectively, to the results for 2 and 10mm/12h.)

Figure 12 shows the difference in EPS performance during the two summer seasons for the Alpine region. The area-average precipitation for this region in 1997 is 20% higher than in 1996 (Table 2). Considering the ROC area as a measure of skill, Figs. 12a-b confirm that the EPS performed better during summer 1997, especially for the event with the higher amount of precipitation (10mm/12h). By contrast, the comparison of the Brier skill scores for the 2mm/12h amount (Fig. 12c) shows that the EPS performed slightly better during 1996 between forecast day 4 and 8.

Figure 13 shows the corresponding results for the Irish region, where average precipitation in 1997 is almost twice the value of 1996 (Table 2). Considering the event with the lower amount of precipitation (2mm/12h), in 1997 the EPS performed better up to forecast day 3, but then slightly worse than in 1996. Considering the event with the highest amount of precipitation (10mm/12h), the EPS performed much better during 1997.

An interesting question is whether the EPS predictive skill improved solely because of the system upgrade introduced on 10 December 1996, or also because of atmospheric variability, i.e. was precipitation easier to predict in 1997 than in 1996?

The only way to answer this question is to compare probabilistic forecasts from ensembles run in both configurations (33 members at T63L19 resolution, and 51 members at T_L159L31 resolution) for the same season: for practical reasons this is not feasible, but the reader is referred to *Buizza et al* (1997) for a comparison of the performance of different ensemble configurations for a set of 14 cases.

An indication of an answer to this question is given by Fig. 14, which shows the Threat Score for the prediction of 2 and 10mm/12h amount of precipitation given by the ECMWF deterministic high resolution model (T213L31) and the EPS control during the 1996 and 1997 summer seasons over Europe. The Threat Scores of the high resolution T213L31 model summer prediction were very similar during 1996 and 1997 (Figs. 14a-b), indicating very little inter-annual variability in predictability of summer precipitation between these two years. The Threat Scores of the EPS control were higher during summer 1997 (Figs. 14c-d), indicating a positive influence of increased resolution. In terms of synoptic scale flow, the performance of the T213L31 model during the two years was very similar, while due to the increased resolution the EPS control performed slightly better during winter 1996/97 (not shown).

Summarizing, results indicate that precipitation predictability during the 1996 and 1997 summer seasons was similar, that the EPS forecast skill in predicting precipitation was remarkably better during summer 1997, and that resolution played a role in improving the EPS skill in precipitation prediction.

6. EPS SKILL DURING WINTER 1996/97, AND COMPARISON OF THE EPS SKILL DURING WINTER 1995/96 AND 1996/97

In this Section, we discuss first the seasonal variability of ensemble forecast skill, and then the impact of the system upgrade introduced on 10 December 1996 on the EPS wintertime performance. (For each year, winter is defined as the 92-day period starting on 10 December and finishing on 12 March of the subsequent year.)

6.1 Precipitation prediction during winter 1996/97 over Europe

Figure 15 shows the ROC area for the prediction of the events "12h accumulated precipitation greater than 1, 2, 5 and 10mm/12h", at forecast day 3, 5 and 7. Results indicate that at forecast day 3, apart for the 10mm/12h threshold, all ROC area values are above the 0.7 limit of useful prediction. By contrast, at forecast day 5 only ROC areas relative to the 1 and 2mm/12h thresholds are always above the 0.7 limit. At forecast day 7 there is a periods with all ROC areas below the 0.7, i.e. between 20 and 24 January 1997.

The comparison of the time series of ROC areas for winter 1996/97 (Fig. 15) and summer 1997 (Fig. 5) suggests that the EPS forecast skill in predicting precipitation is higher during winter. This is confirmed by the comparison of the area under the seasonal average ROC curve for the two seasons (see Figs. 16a-b for winter 1996/97, and Figs. 7a-b for summer 1997), and by the comparison of the Brier score and Brier skill score for the two seasons (see Figs. 16c-d for winter 1996/97, and Figs. 7c-d for summer 1997). Considering for example the forecast time at which the area under the seasonal average ROC curve relative to the largest precipitation amount studied (10mm/12h) first crosses the 0.7 limit, the current EPS (i.e. based on 50 perturbed members, and on T_L159L31 resolution) gives skilful probabilistic predictions up to forecast day 4.5 during winter, and up to forecast day 3 during summer. However, prediction of 5mm/12h is skilful to 4 days in summer and 8 days in winter.

It is worth mentioning that the performance of the EPS has been analyzed also over the Alpine and the Irish local regions, and that results present a weaker sensitivity to the size of the geographical region during winter than during summer. This suggests that the seasonal differences in skill arises from the relative difficulty of forecasting widespread rain of

dynamical origin in winter, and the more convective situations in summer. This is reinforced by the results of Fig. 16a, which shows that 1, 2 and 5mm/12h are equally skilful in winter.

6.2 Comparison of the EPS forecast skill during winter 1995/96 and 1996/97

Figure 17 shows the variation, with respect to forecast time, of the area under the winter ROC curve and of the Brier skill score relative to the events "12h accumulated precipitation greater than 2mm/12h and 10mm/12h", for winter 1995/96 (dash) and winter 1996/97 (solid). The comparison of the EPS performance during the two winter seasons confirms the summer results (Fig. 11), that the system upgrade improved the EPS performance in predicting precipitation amounts, especially for the 10mm/12h threshold.

To investigate if the improved EPS performance is due to atmospheric variability, we compare the T213L31 Threat scores for the two years. These indicate that winter 1996-97 was slightly easier to predict (Figs. 14e-f). Moreover, considering the impact of resolution on ensemble performance, the comparison of the EPS control Threat Scores for the two years confirms that the resolution increase improved the skill score for precipitation prediction (Fig. 14g-h). It is worth mentioning that in terms of synoptic scale flow both the control and the T213L31 models performed slightly better during winter 1995/96 than 1996/97 (not shown).

In summary, the EPS performed better during winter 1996/97 than during winter 1995/96. Precipitation prediction was slightly easier during winter 1996-97, and thus atmospheric variability could be partly responsible for the improved EPS performance in predicting precipitation. However, both winter and summertime results suggest that the resolution increase played an important role in improving the system performance.

7. CONCLUSIONS

This work reports a statistical analysis of the performance of the ECMWF EPS in predicting precipitation probabilities. Four seasons have been considered, summers 1996 and 1997, and winter 1995/96 and 1996/97. Precipitation has been accumulated over 12h periods, and forecasts up to day 10 have been compared with verification data defined using the 12h and the 24h forecasts from the ECMWF operational high resolution determinist model (T213L31). Three regions have been considered: Europe and two smaller regions, one centred around the Alpine region and one centred around Ireland. The EPS performance has been assessed applying Signal Detection Theory, the Brier score and the Brier skill scores.

We first discussed some case studies to highlight the correspondence between a measure of skill defined applying Signal Detection Theory, specifically the area under a ROC curve, and maps of forecast probabilities. The case studies supported the common practice which considers a probabilistic forecast to be useful if it has an area under a ROC curve of 0.70 or more.

We then analyzed summertime and wintertime seasonal average results, with two main goals in mind. The first goal was to measure objectively the EPS forecast skill in predicting precipitation probabilities, and then to assess the seasonal variability of the forecast skill. The second goal was to investigate whether the EPS upgrade introduced on 10 December 1996 (when membership was enhanced from 33 to 51 members and resolution was increased from T63L19 to T_L159L31), had an impact on the EPS forecast skill.

Table 3 summarizes the results for Europe. For each considered season, the forecast time at which the area under the seasonal average ROC curve crosses first (or reaches) the 0.7 limit of useful prediction for the four analyzed rainfall amounts.

There is substantial seasonal variability of the EPS skill in predicting precipitation, and Table 3 indicates that wintertime is more predictable. With the new ensemble configuration, the comparison of the results for the 2mm/12h and the 10mm/

12h amounts of precipitation for winter 1996/97 and summer 1997 shows that the wintertime predictive skill is, respectively, 3.5 and 1.5 days longer.

	1mm/12h	2mm/12h	5mm/12h	10mm/12h
Winter 1995/96	10.0	9.5	6.0	2.5
Summer 1996	7.0	5.5	2.0	--
Winter 1996/97	10.0	10.0	8.0	4.5
Summer 1997	10.0	6.5	4.5	3.0

Table 3: Forecast time (day) at which the area under the seasonal average ROC curve crosses the 0.7 limit for useful prediction.

The impact of the system upgrade on the EPS forecast skill in predicting precipitation was substantial. Table 3 shows that the gain in predictability has been remarkable, especially for the largest amount of precipitation during summertime. For the 5mm/12h amount, the gain in predictability has been of 2 days during winter and 2.5 days during summer. It is worth recalling that atmospheric variability did not play any role in improving the EPS performance between summer 1996 and summer 1997, but it played a role in improving the EPS performance between winter 1995/96 and winter 1996/97.

For the largest precipitation amounts, the results reported in Table 3 and based on Signal Detection Theory recommend that the EPS can be used to give skilful predictions for both the 5 and the 10 mm/12h amounts up to forecast day 4 in winter and up to forecast day 3 in summer. This recommendation is further supported by the comparison of the reliability diagrams for the two thresholds at forecast day 4 for winter 1996/97 and at forecast day 3 for summer 1997 (Fig. 18).

The improvement in skill must owe much to the improved model resolution, and perhaps improved physics. This emphasizes the need to use the best possible model in the EPS. The winter/summer differences in skill may be due to the fact that winter rain is largely dynamical in origin and has a large spatial scale, while summer rain has a large convective component and thus a smaller spatial scale. Again this emphasizes the need for resolution in the EPS model if one is to improve the summer forecasts. There are noteworthy differences in skill between the relative flat maritime Irish area and the mountainous continental Alpine area, but we have no ready interpretation.

A limitation of this work is that it focused on the verification of precipitation amounts accumulated over 12 hours. Although timing errors increase with forecast time, for many applications (hydrology, agriculture) reliable medium range probabilistic forecasts of accumulations of large amounts over 24 or 48 hours would be valuable. Moreover, reliable medium range probabilistic forecasts of large accumulations over large basins would also be useful for hydrological applications. The verification system needs to be extended to address these issues. Finally, a verification system which can verify EPS forecasts against detailed rain-gauge data needs to be developed.

ACKNOWLEDGEMENTS

We thank Adrian Simmons, Horst Böttger and François Lalaurette for their comments to an early version of this manuscript.

REFERENCES

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

Buizza, R, 1997a. Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev*, **125**, 99-119.

Buizza, R, 1997b. Some aspects of the performance of the ECMWF Ensemble Prediction System. Proceedings of the *Expert meeting on the use and development of the Ensemble Prediction System*, 18-19 June 1997, ECMWF, Shinfield Park, Reading RG2 9AX, in press.

Buizza, R, Petroliagis, T, Palmer, T N, Barkmeijer, J, Hamrud, M, Hollingsworth, A, Simmons, A, and Wedi, N, 1997. Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Q. J. R. Meteorol. Soc.*, in press.

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. ECMWF seminar proceedings on *Validation of models over Europe: vol. I*, ECMWF, Shinfield Park, Reading RG2 9AX, UK,

Petroliagis, T, Buizza, R, Lanzinger, A, and Palmer, T N, 1997. Potential use of the ECMWF Ensemble Prediction System in cases of extreme weather events. *Meteorol. Appl.*, **4**, 69-84.

Stanski, H R, Wilson, L J, and Burrows, W R, 1989. Survey of common verification methods in meteorology. *World Weather Watch Tech. Rep. 8*, WMO, Geneva, pp 114.

Toth, Z, Kalnay, E, Tracton, S, Wobus, R, and Irwin, J, 1997. A synoptic evaluation of the NCEP ensemble. *Weather and Forecasting*, **12**, 140-153.

Toth, Z, Zhu, Y, Marchok, T, Tracton, S, and Kalnay, E, 1998. Verification of the NCEP global ensemble forecasts. Pre-prints of the 12th Conference on Numerical Weather Prediction, 11-16 January 1998, Phoenix, Arizona, in press.

Tracton, M S, and Kalnay, E, 1993. Operational ensemble prediction at the National Meteorological Centre: practical aspects. *Weather and Forecasting*, **8**, 379-398.

Wilks, D S, 1995. *Statistical methods in the atmospheric sciences*, Academic Press, pp 465.

Zhu, Y, Toth, Z, Kalnay, E, and Tracton, S, 1998. Probabilistic quantitative precipitation forecasts based on the NCEP global ensemble. Contribution to the Special Symposium on Hydrology, Pre-prints of the 12th Conference on Numerical Weather Prediction, 11-16 January 1998, Phoenix, Arizona, in press

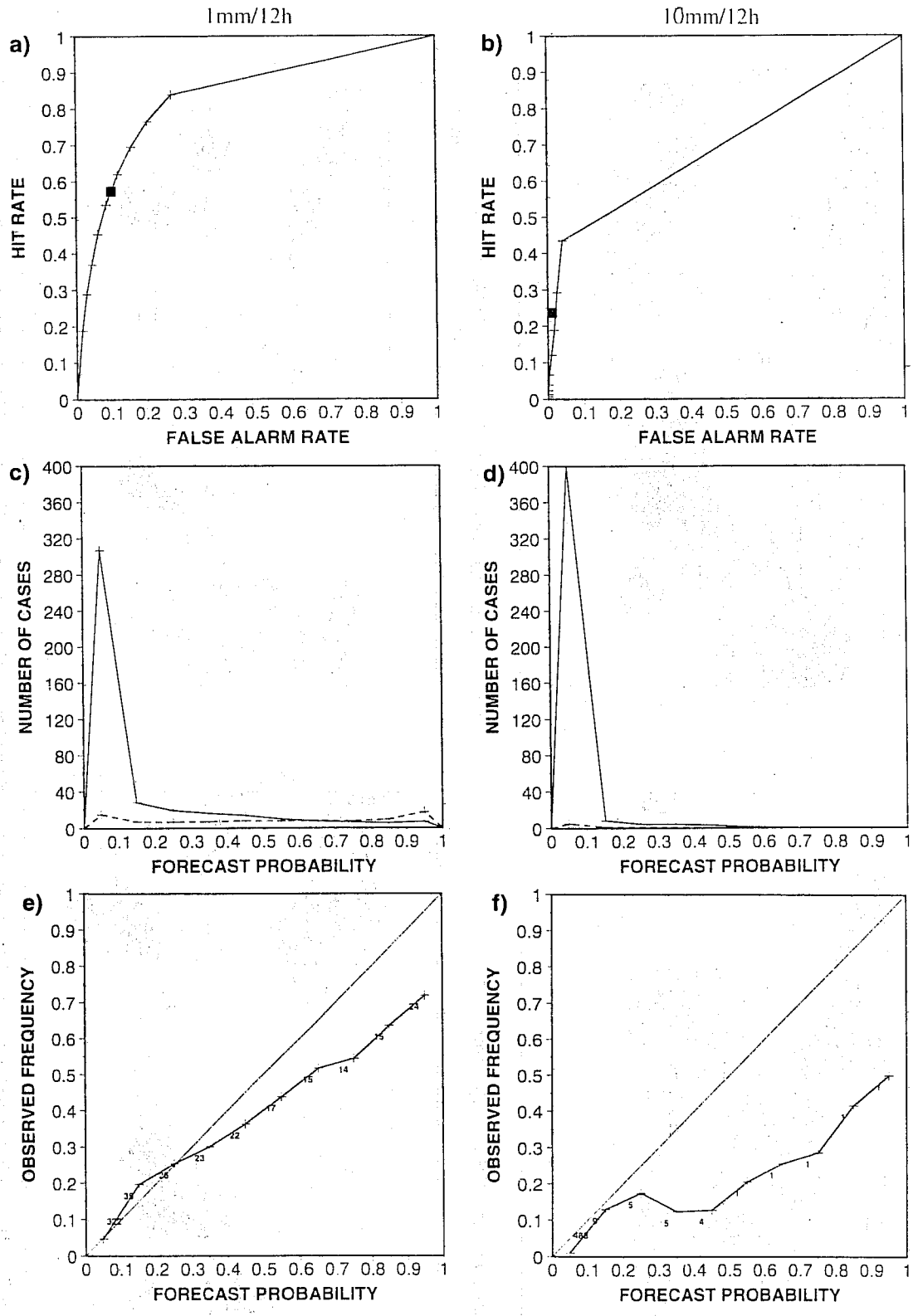

Fig. 1 (a) Sunmer-97 average ROC curves for Europe, relative to the event "12h accumulated precipitation greater than 1mm/12h" at forecast day 3; (b) as (a) but for "12h accumulated precipitation greater than 10mm/12h"; (c-d) as (a-b) but for the conditional distributions for non-occurrence (solid) and occurrence (dash) associated with the ROC curves shown in Fig. 2; (e-f) as (a-b) but for the reliability diagrams (for each forecast probability category, a label reports the average number of grid points classified in the category itself).

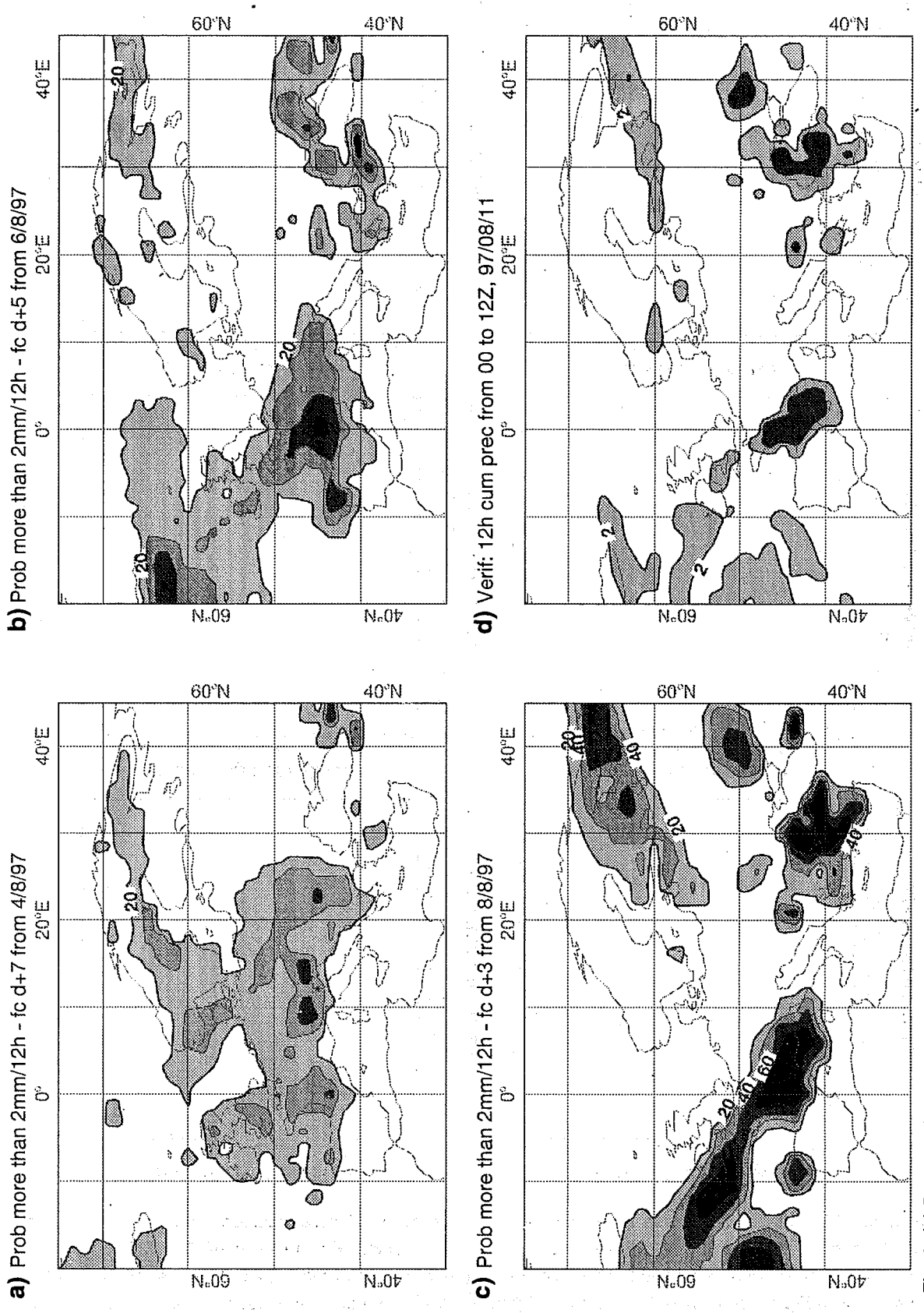

Fig. 2 (a) Day 7 probability forecast of the event "12h accumulated precipitation greater than 2mm/12h", of the EPS started on 4/8/97; (b): as (a) but for the day 5 forecast of the EPS started on 6/8/96; (c): as (a) but for the day 3 forecast of the EPS started on 8/8/96; (d) verification for the probability forecasts (i.e. observed precipitation accumulated between 00Z and 12Z of 11/8/97). The contour interval for probabilities is 20%, starting from 20%, while the 2mm/12h and the 10mm/12h isolines are plotted for precipitation.

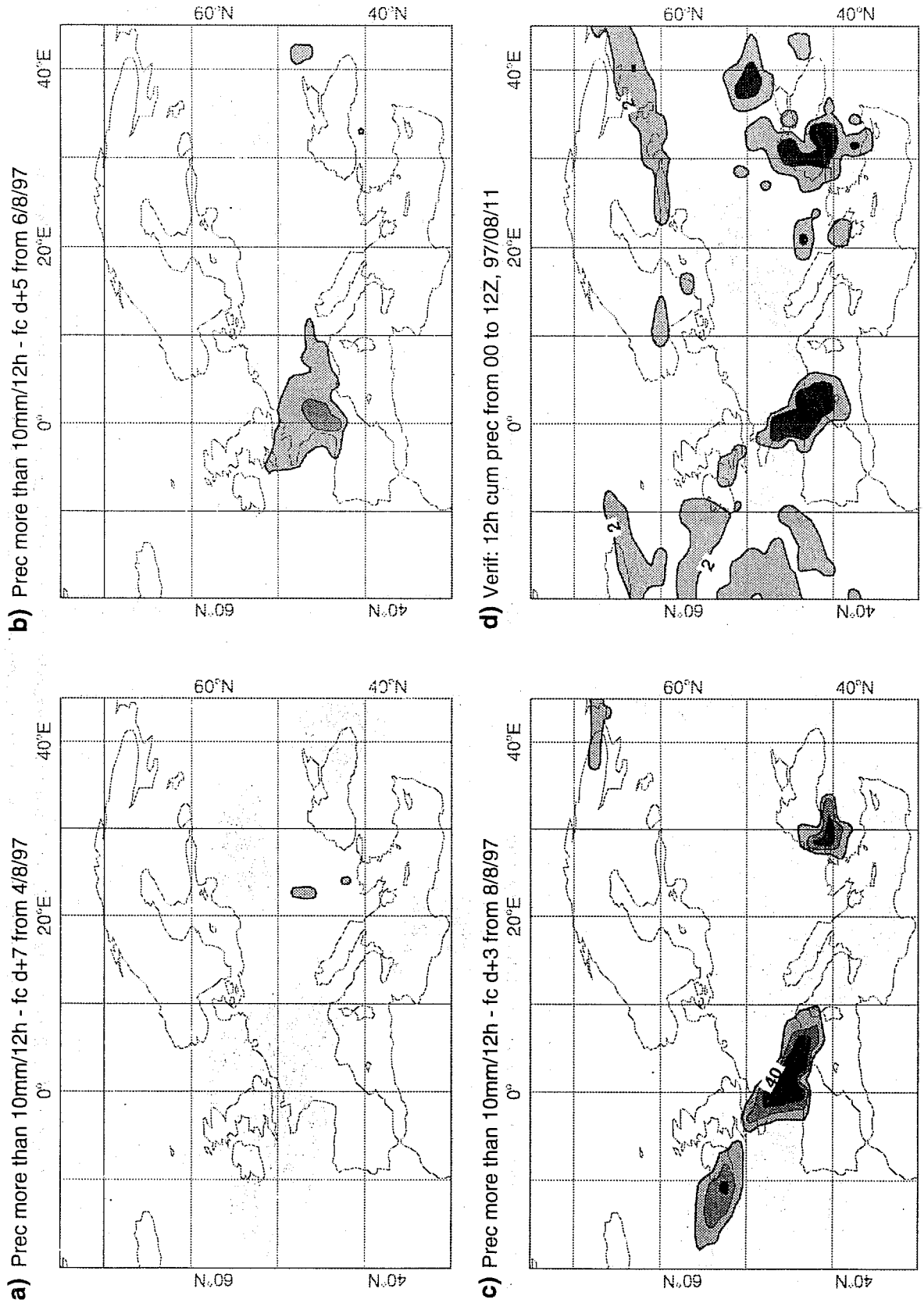

Fig. 3 As Fig. 2 but for the event "12h accumulated precipitation greater than 10mm/12h".

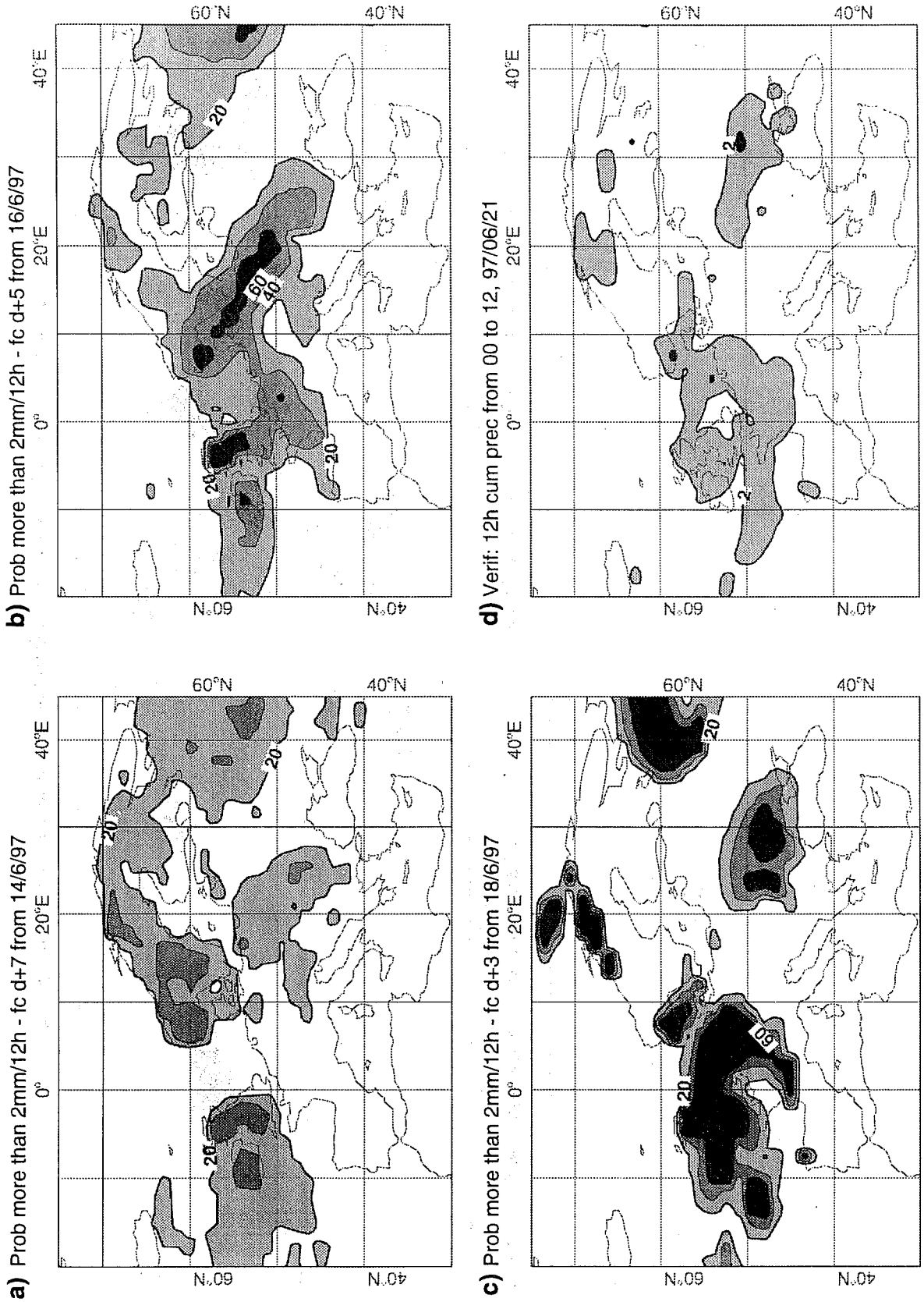

Fig. 4 As Fig. 2 but for the day 7, day 5 and day 3 forecasts of the event "12h accumulated precipitation greater than 2mm/12h", started respectively on 14, 16 and 18 June, and verifying on 21/6/97.

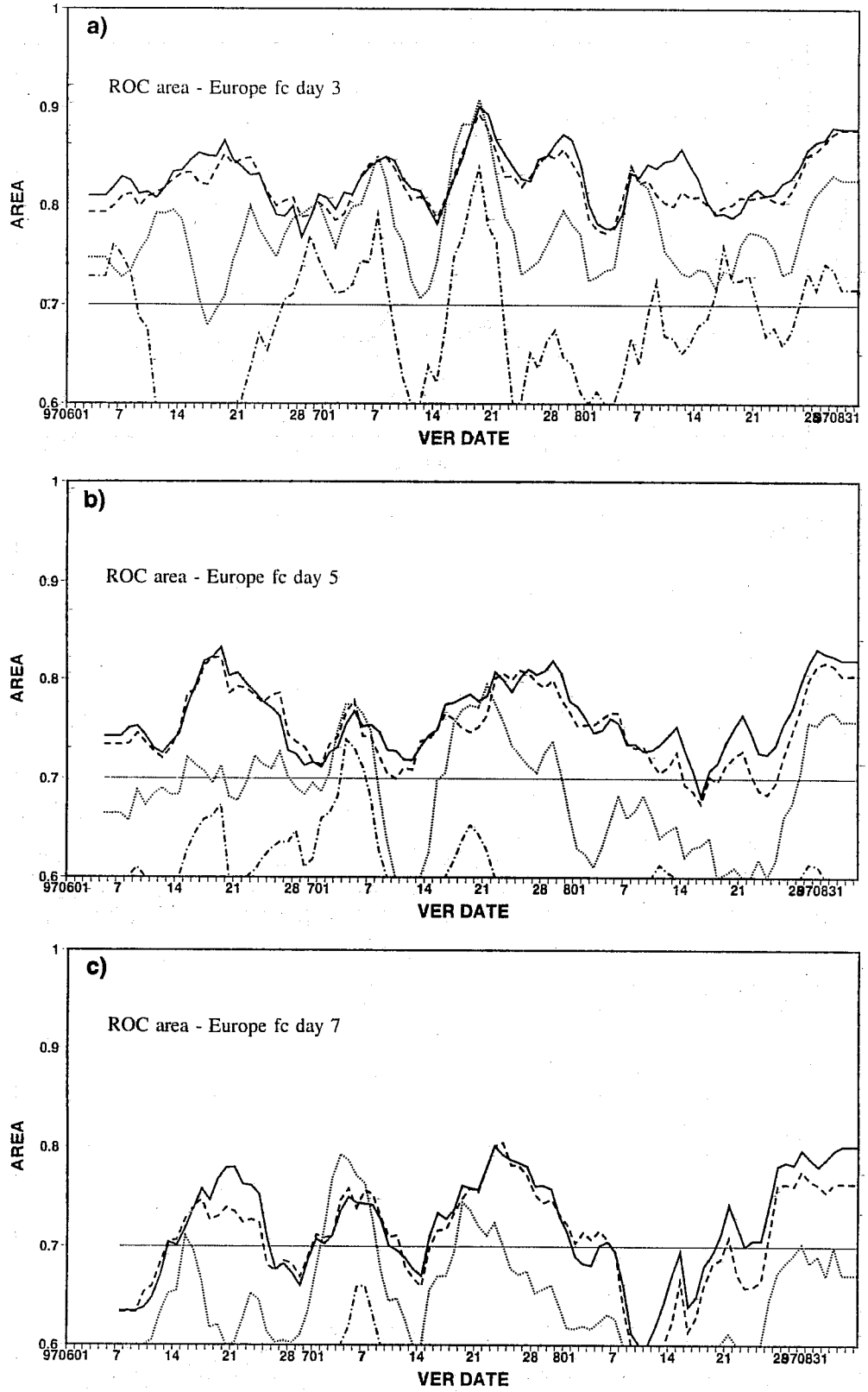

Fig. 5 Sunimer-97 (5-day running mean) area under the ROC curve, for Europe, of the events "12h accumulated precipitation greater than 1mm/12h" (solid), "12h accumulated precipitation -reater than 2mm/12h" (dash), "12h accumulated precipitation greater than 5mm/12h" (dot), and "12h accumulated precipitation greater than 10mm/12h" (chain-dash), at (a) forecast day 3, (b) forecast day 5 and (c) forecast day 7. Note that the abscissa is the verifying date.

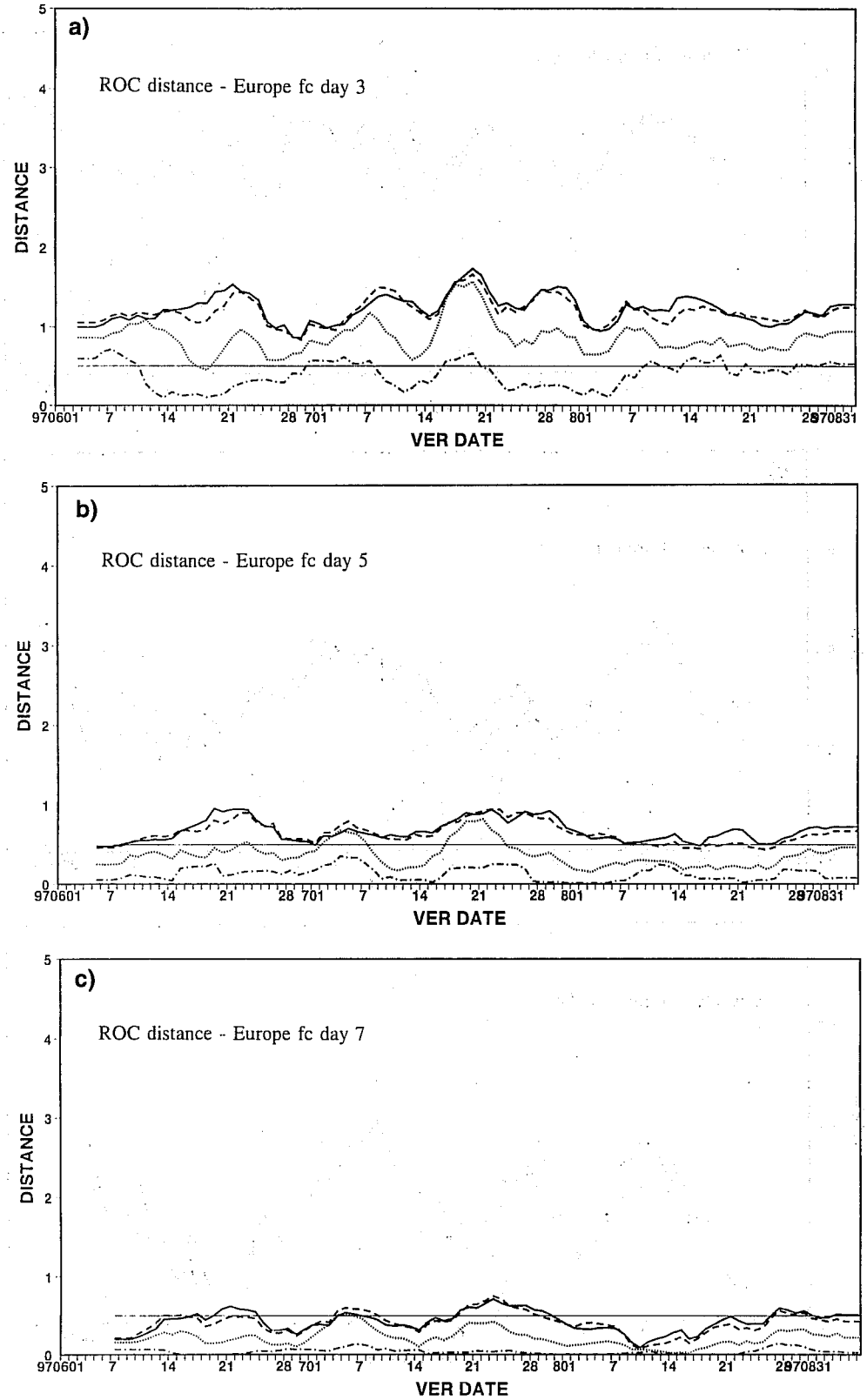

Fig. 6 As Fig. 5 but for the daily value of the normalized distance between the conditional distributions (solid for non-occurrence, and dash for occurrence).

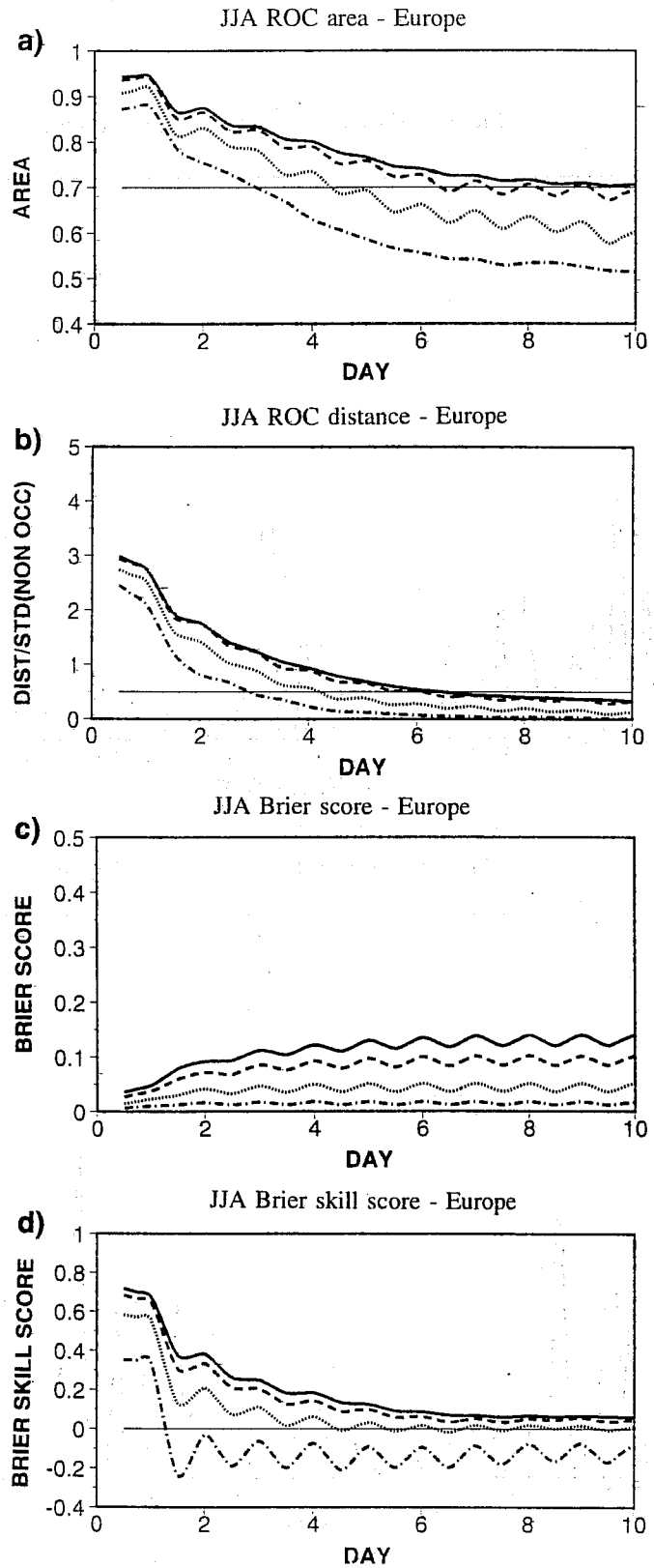

Fig. 7 (a) Variation, with respect to forecast time, of the area under the summer-97 average ROC curve for Europe, for the events "12h accumulated precipitation greater than 1mm/12h" (solid), "12h accumulated precipitation greater than 2mm/12h" (dash), "12h accumulated precipitation greater than 5mm/12h" (dot), and "12h accumulated precipitation greater than 10mm/12h". (b): as (a) but for the normalized distance between the respective conditional distributions. (c-d):as (a-b) but for the Brier score and the Brier skill score.

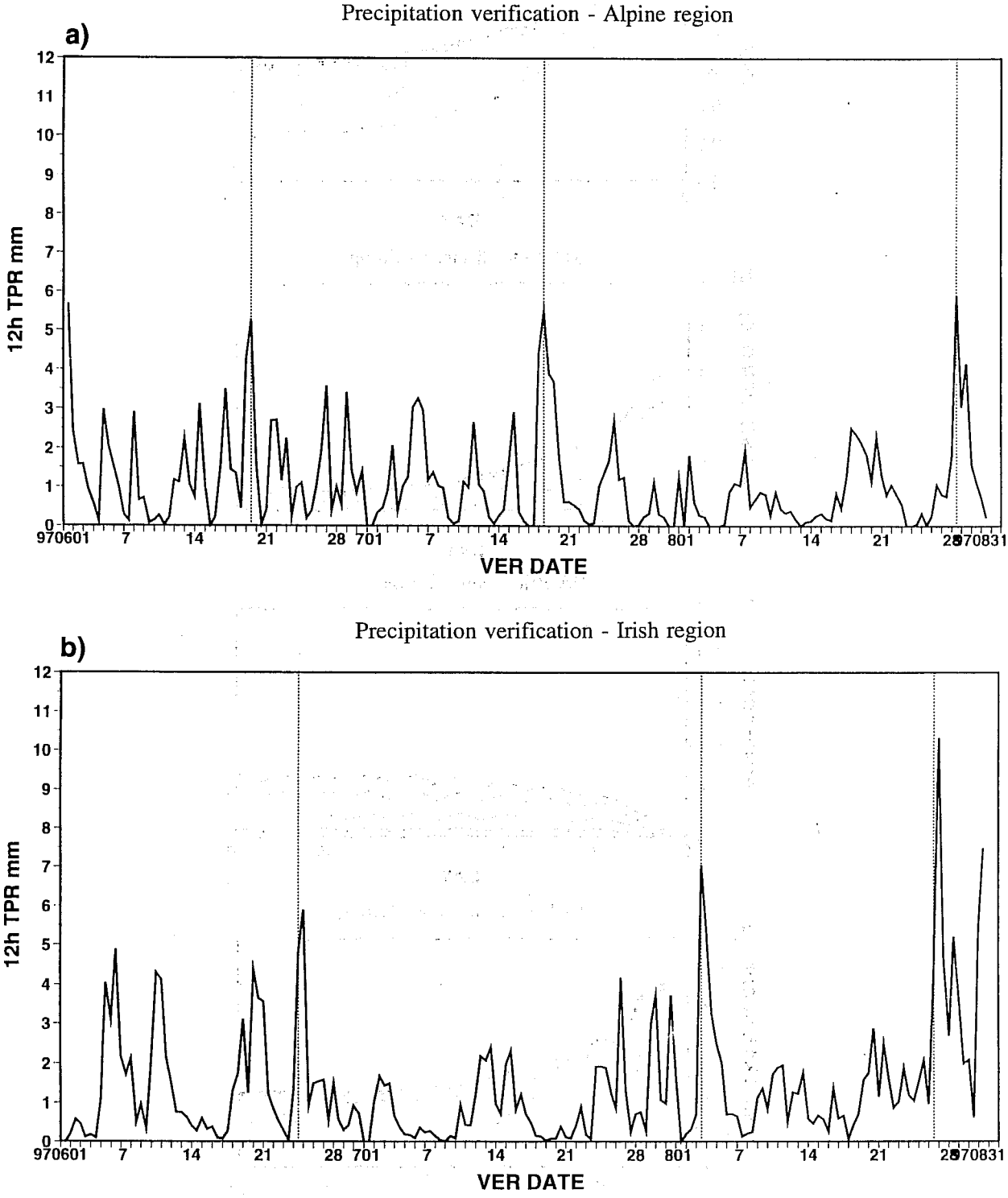

Fig. 8 Average verification for 12h accumulated precipitation over (a) the Alpine and (b) the Irish regions.

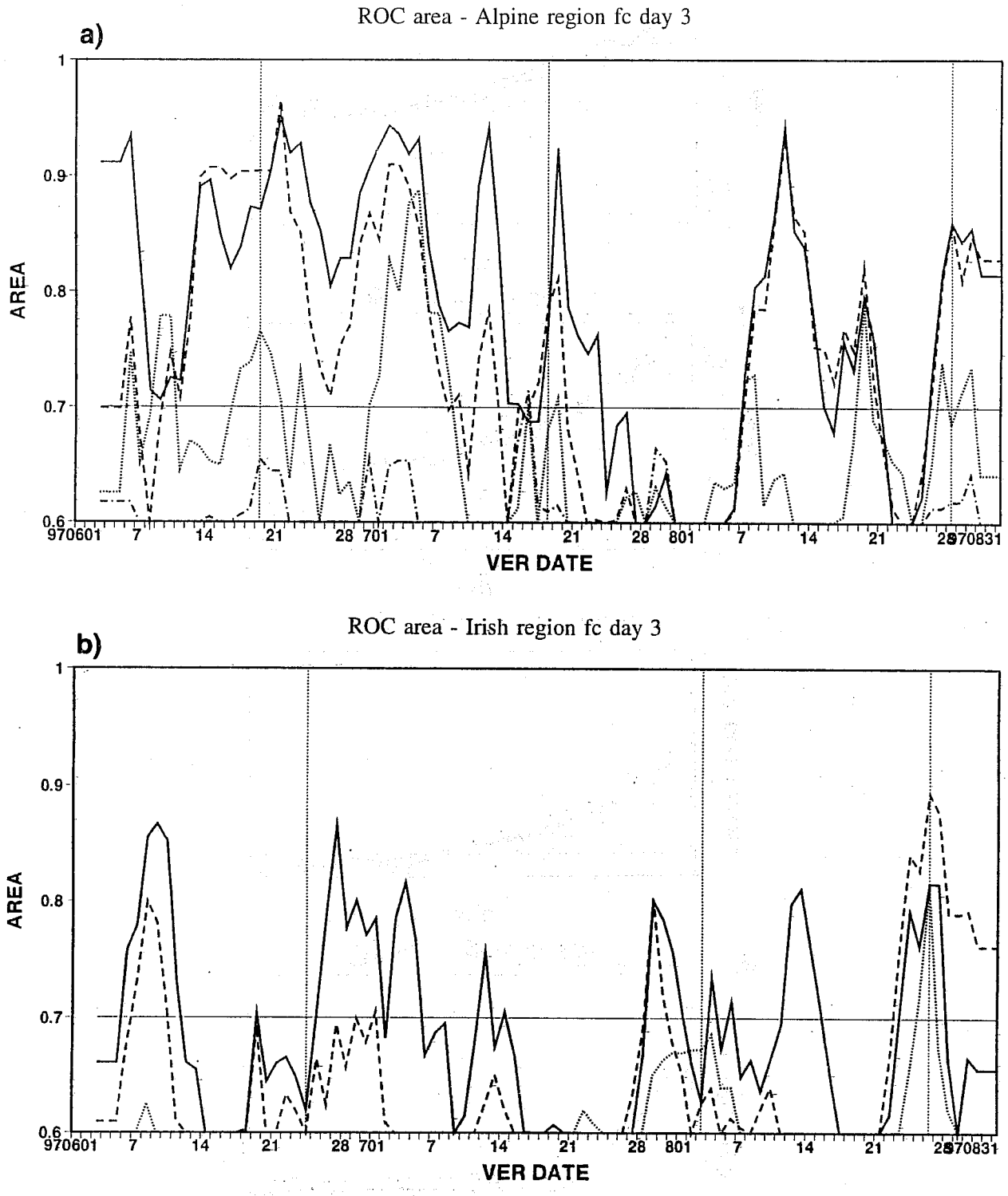

Fig. 9 (a) summer-97 (5-day running mean) area under the ROC curve for the Alpine region of the events "12h accumulated precipitation greater than 1mm/12h" (solid), "12h accumulated precipitation greater than 2mm/12h" (dash), "12h accumulated precipitation greater than 5mm-t/12h" (dot), and "12h accumulated precipitation greater than 10mm/12h" (chain-dash), at forecast day 3. (b): as (a) but for the Irish region. Note that the abscissa is the verifying date.

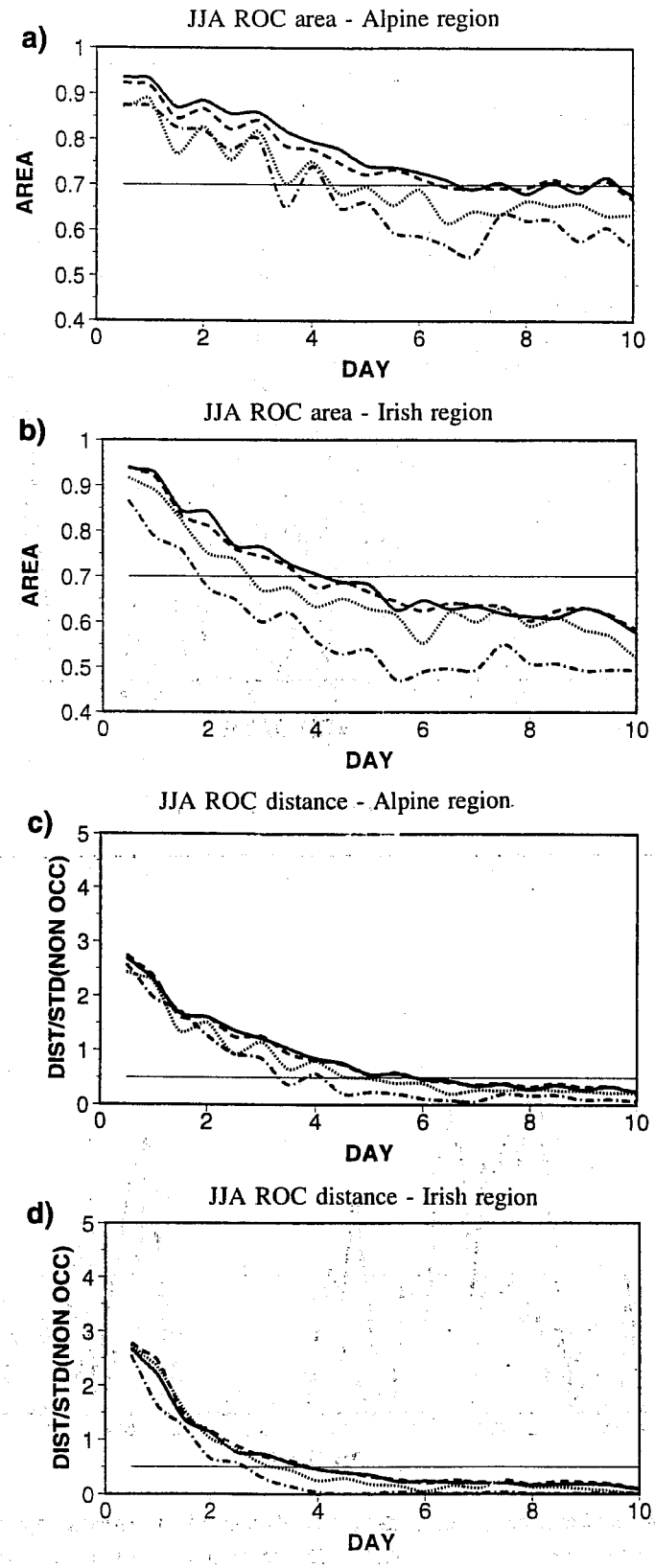

Fig. 10 (a) Variation, with respect to forecast time, of the area under the summer-97 average ROC curve for the Alpine region, for the events "12h accumulated precipitation greater than 1mm/12h"(solid), "12h accumulated precipitation greater than 2mm/12h" (dash), "12h accumulated precipitation greater than 5mm/12h" (dot), and "12h accumulated precipitation greater than 10mm/12h". (b): as (a) but for the Irish region. (c-d): as (a-b) but for the normalized distance between the associated conditional distributions.

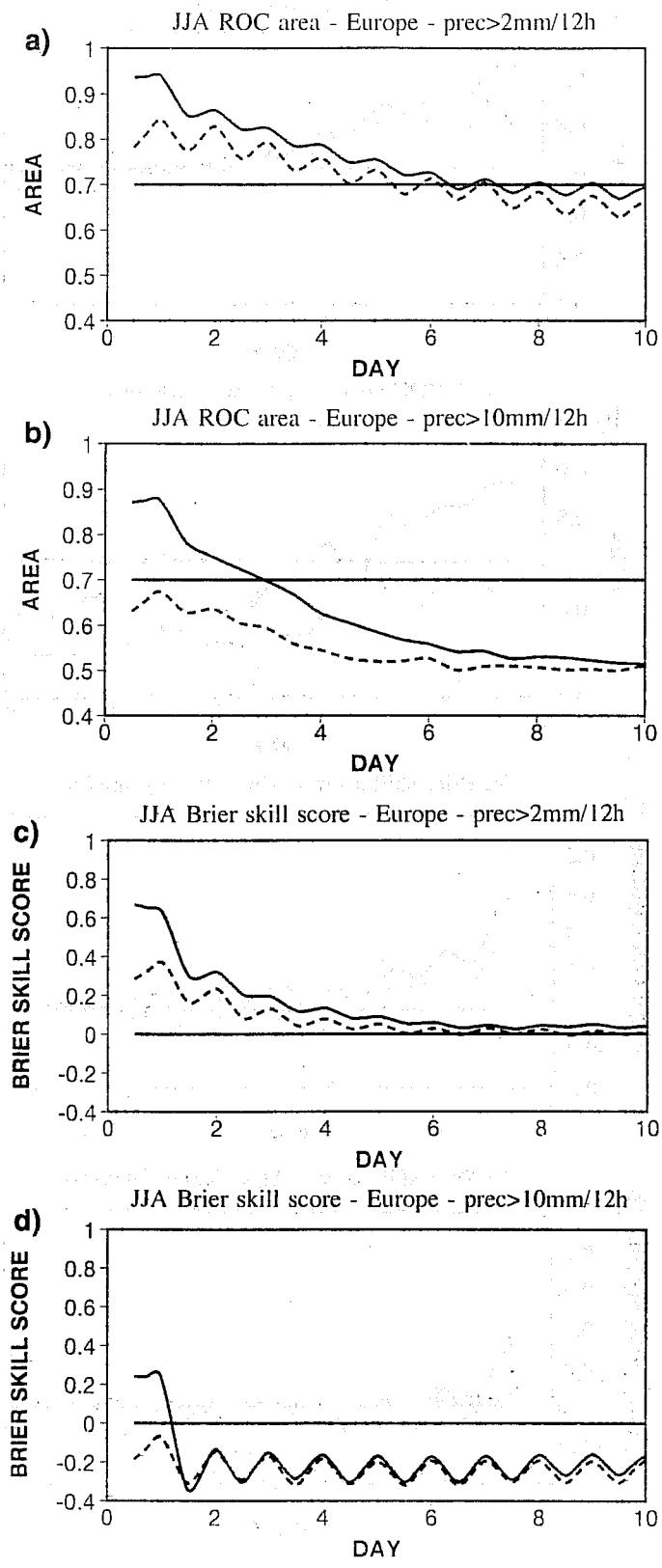

Fig. 11 (a) Variation, with respect to forecast time, of the area under the summer average ROC curve for Europe for the event "12h accumulated precipitation greater than 2mm/12h", for summer 1996 (dash) and 1997 (solid). (b) as (a) but for the event "12h accumulated precipitation greater than 10mm/12h" (dash). (c-d): as (a-b) but for the Brier skill score.

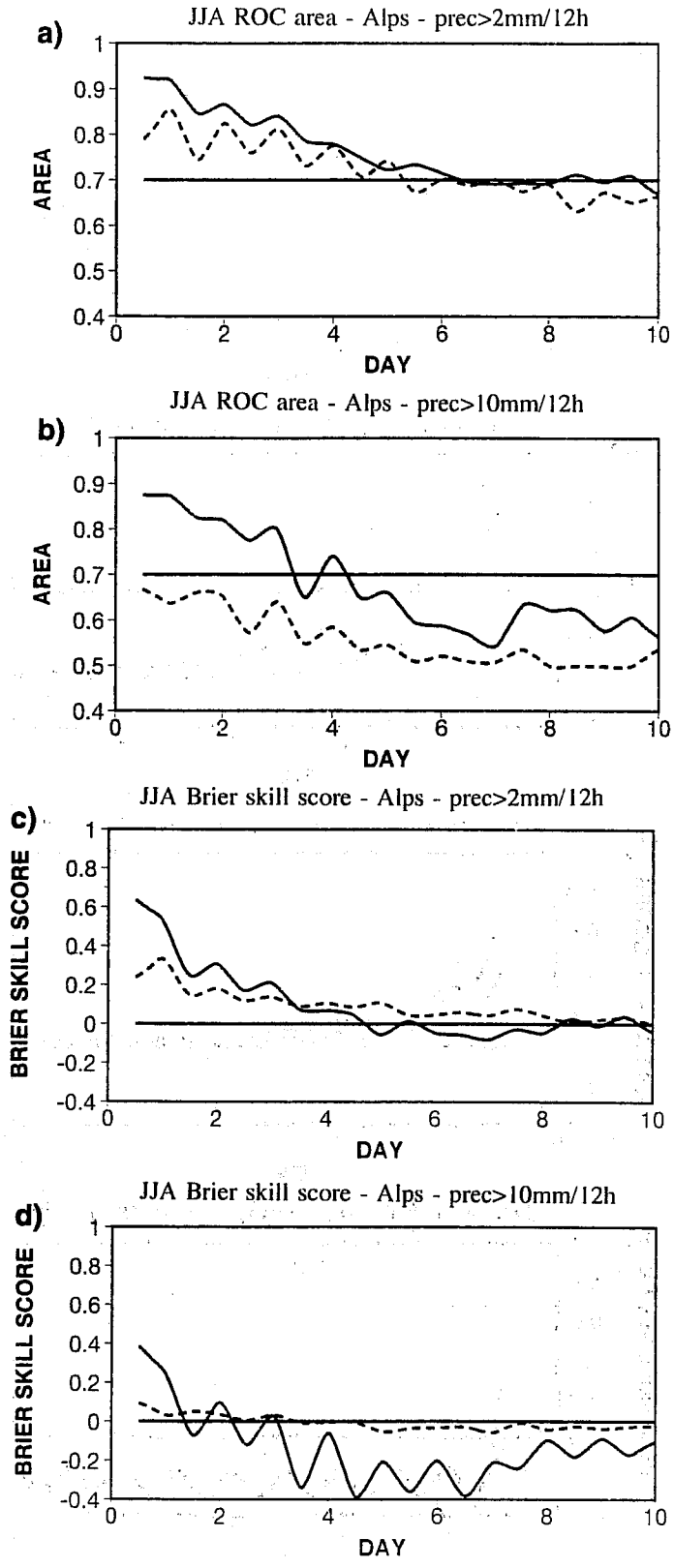

Fig. 12 As Fig. I I but for The Alpine region.

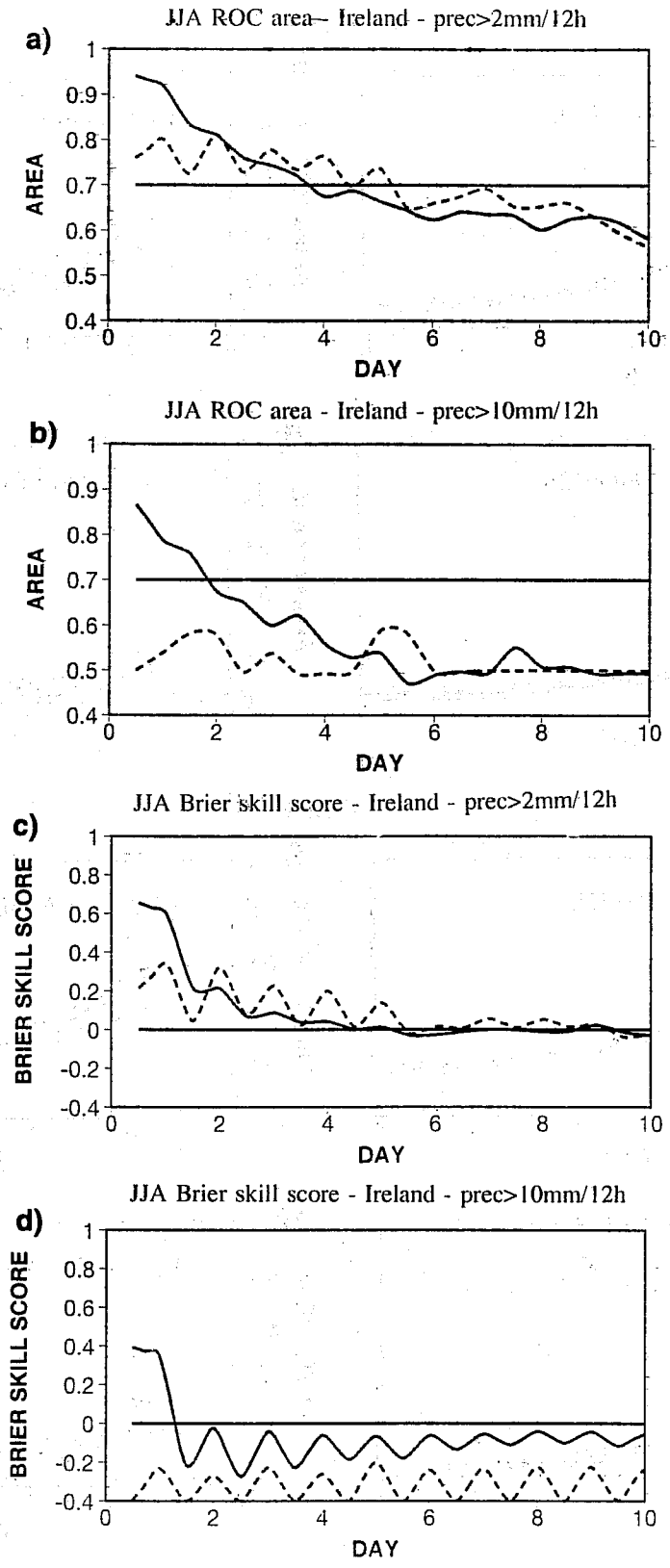

Fig. 13 As Fig. 11 but for the Irish region.

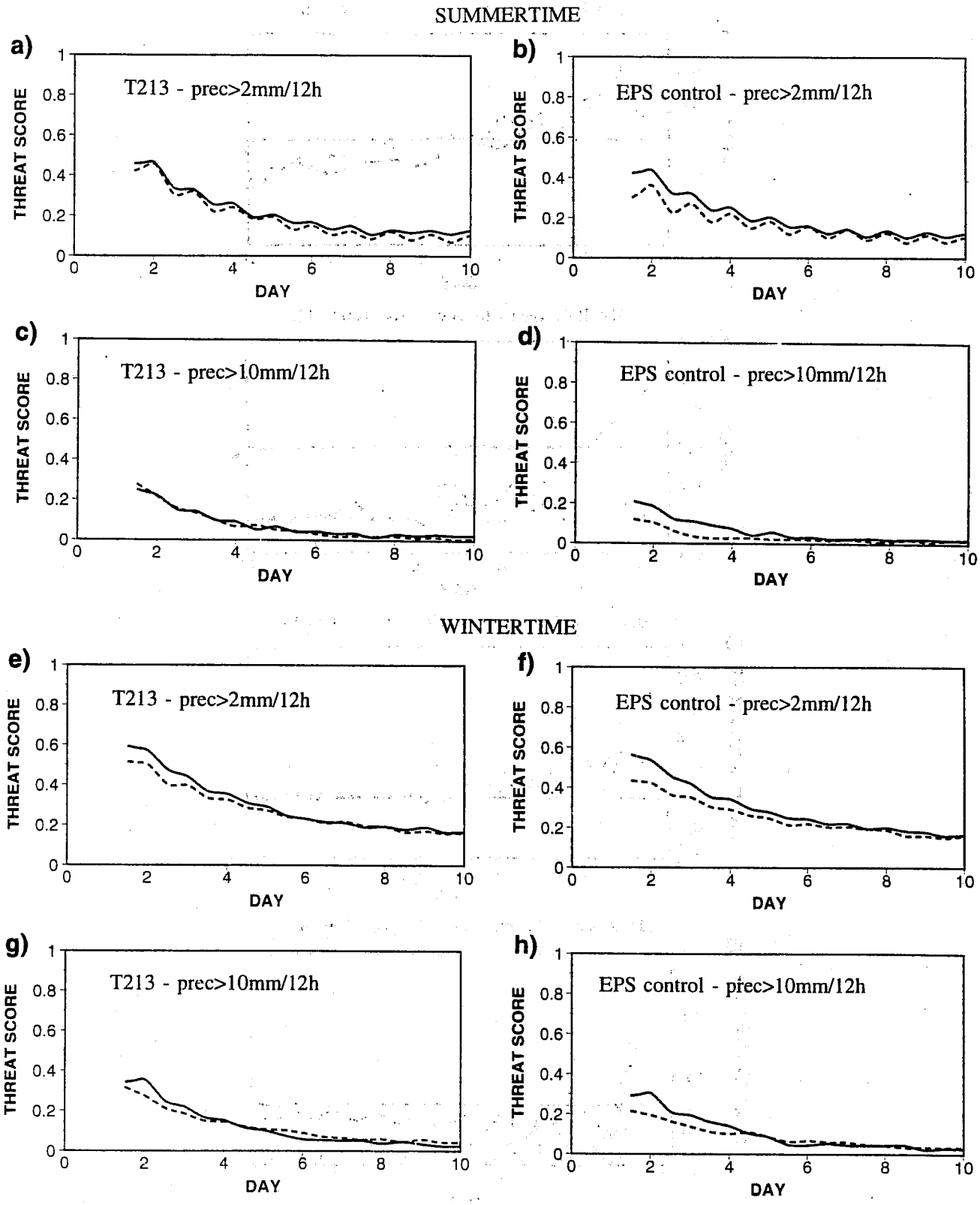

Fig. 14 (a) Summertime Threat score for the prediction by the T213L31 model of 2mm/12h rainfall for 1997 (solid) and 1996 (dash); (b): as (a) but for 10mm/12h. (c-d): as (a-b) but for the prediction by the EPS control. (e-h): as (a-d) but for wintertime.

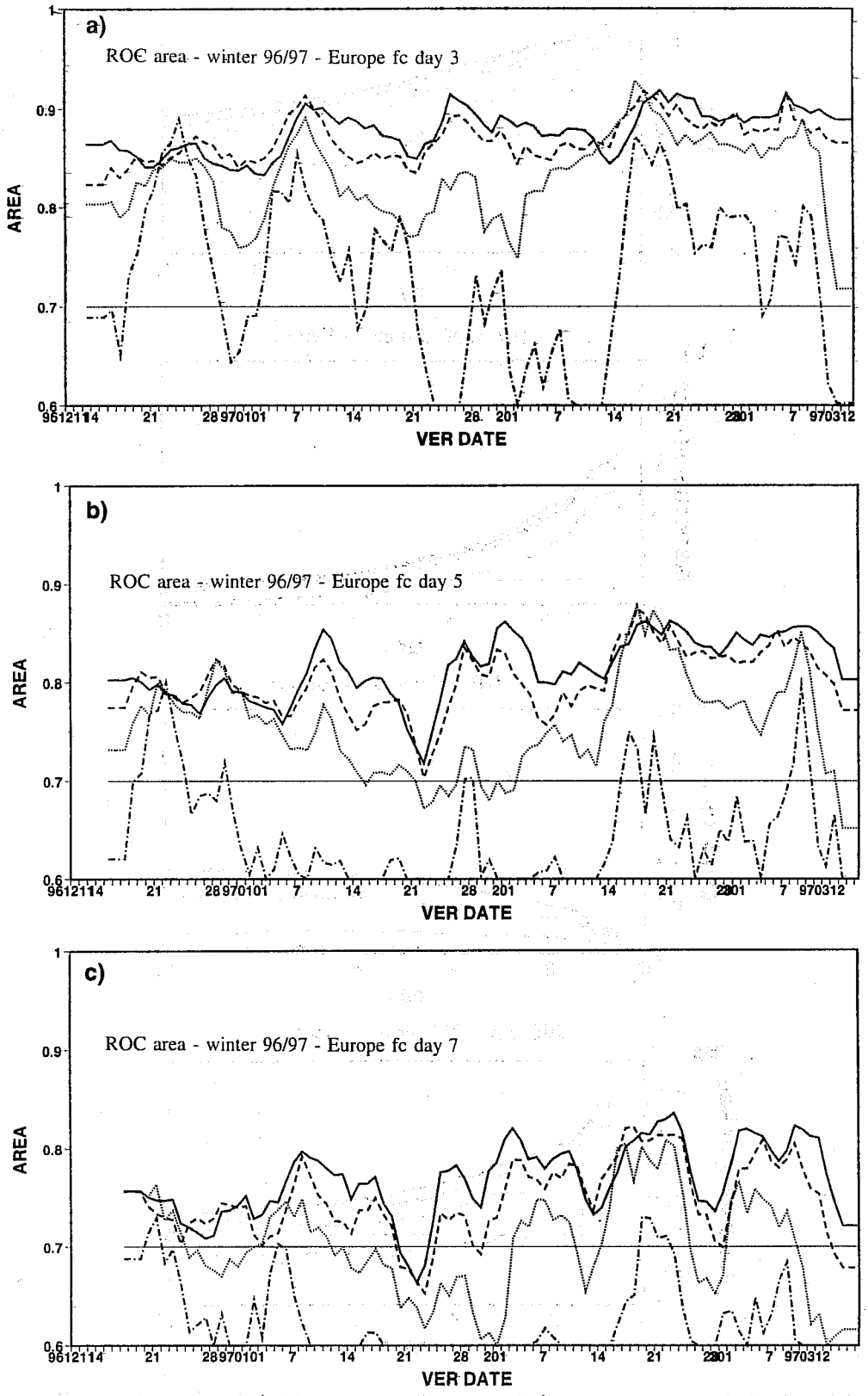

Fig. 15 Winter-96/97 (5-day running mean) area under the ROC curve, for Europe, of the events "12h accumulated precipitation greater than 1mm/12h" (solid), "12h accumulated precipitation greater than 2mm/12h" (dash), "12h accumulated precipitation greater than 5mm/12h" (dot), and "12h accumulated precipitation greater than 10mm/12h" (chain-dash), at (a) forecast day 3, (b) forecast day 5 and (c) forecast day 7. Note that the abscissa is the verifying date.

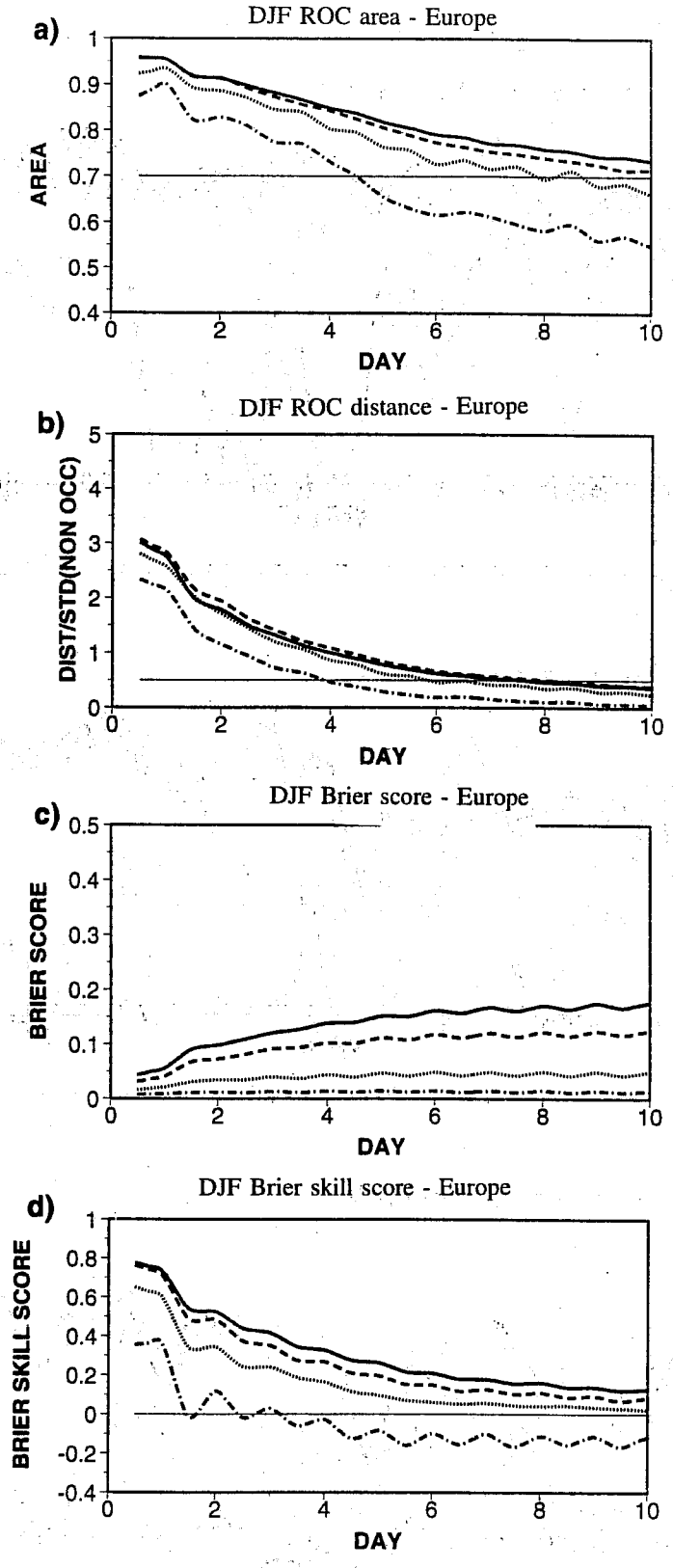

Fig. 16 (a) Variation, with respect to forecast time, of the area under the winter-96/97 average ROC curve for Europe, for the events "12h accumulated precipitation greater than 1mm/12h" (solid), "12h accumulated precipitation greater than 2/12h" (dash), "12h accumulated precipitation greater than 5mm/12h" (dot), and "12h accumulated precipitation greater than 10mm/12h". (b): as (a) but for the normalized distance between the associated conditional distributions. (c-d):as (a-b) but for the Brier score and skill score.

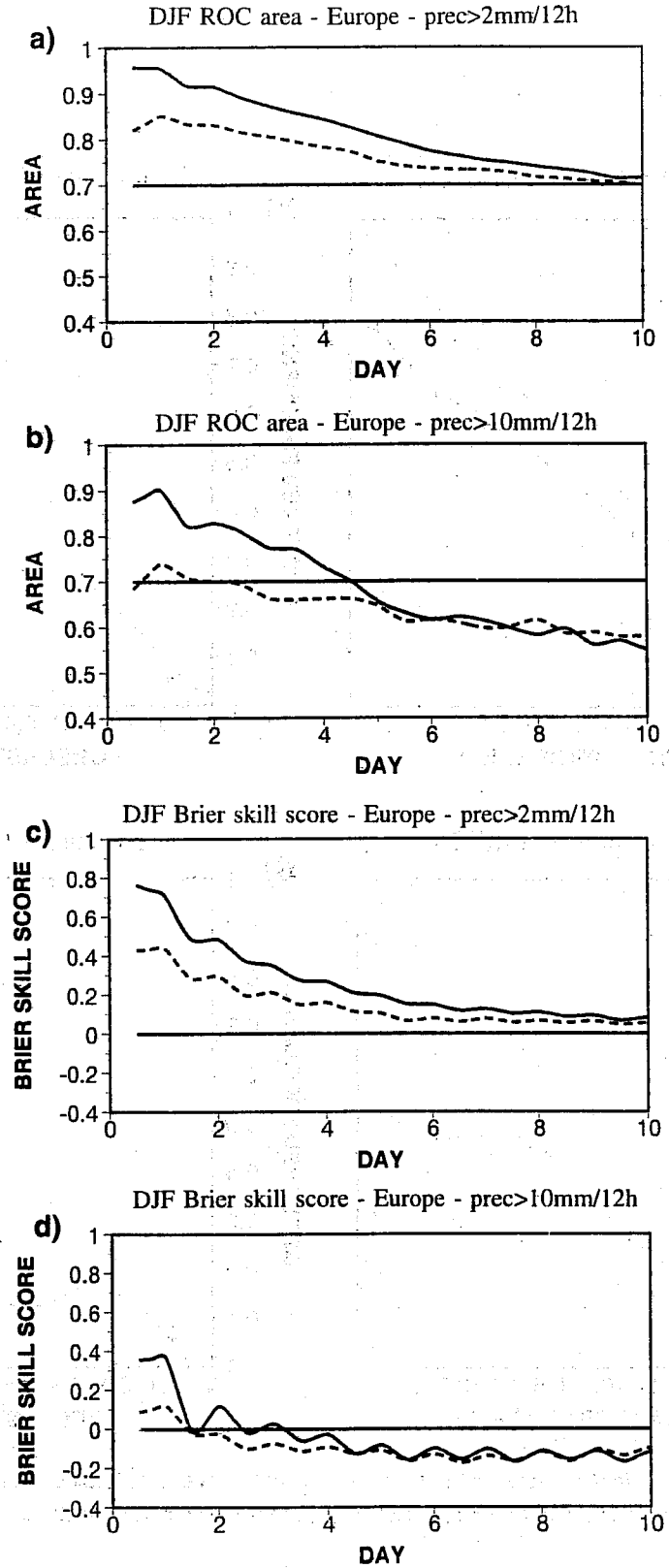

Fig. 17 (a) Variation, with respect to forecast time, of the area under the winter average ROC curve for Europe for the event "12h accumulated precipitation greater than 2mm/12h", for winter 1995/96 (dash) and 1996/97 (solid). (b): as (a) but for the event "12h accumulated precipitation greater than 10mm/12h" (dash). (c-d): as (a-b) but for the Brier skill score.

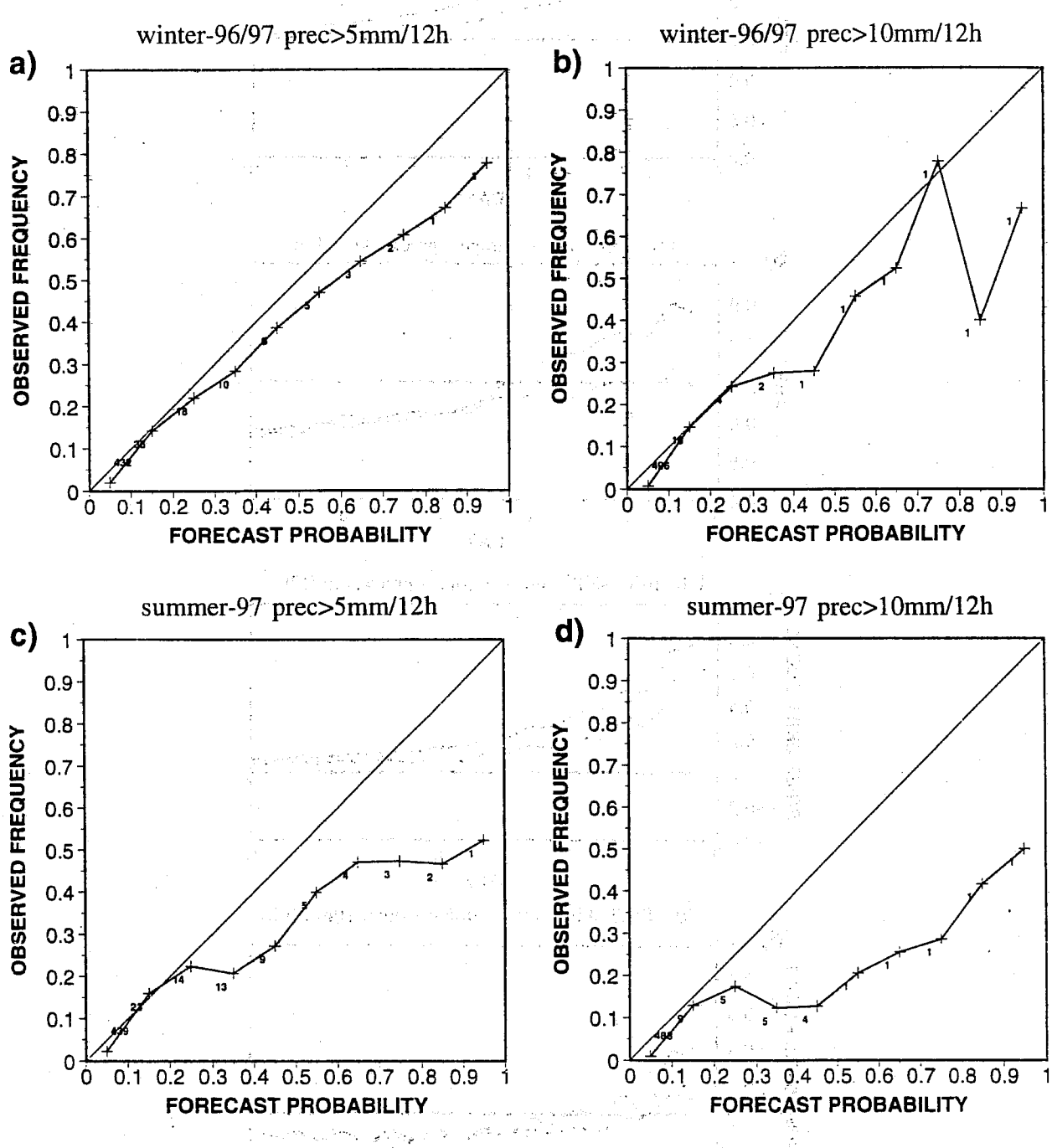

Fig. 18 (a) Winter 1996/97 reliability diagram for the prediction of the event "12h accumulated precipitation greater than 5mm/12h", at forecast day 4 over Europe. (b): as (a) but for 10mm/12h. (c-d): as (a-b) but at forecast day 3 for summer 1997.