PERFORMANCE OF THE ECMWF ENSEMBLE PREDICTION SYSTEM

by

T N Palmer, R Buizza and F Lalaurette
ECMWF

Abstract

A major upgrade to the ECMWF Ensemble Prediction System took place in December 1996. The performance of the ensemble system before and after the upgrade is compared. A remarkable two-day increase in skill is noted. A summary of plans for the development of the EPS is given.

1. INTRODUCTION

Real-time medium-range ensemble prediction began at ECMWF in December 1992 (Palmer et al, 1993: Molteni et al, 1996: Palmer et al, 1997). Until December 1996, the operational configuration comprised 32 perturbed forecasts at T63L19 resolution, and a control forecast made from a truncated version of the operational analysis. The perturbations were based on the first 16 singular vectors (SVs) of the forward tangent model (first at T21L19, then at T42L19 resolution), optimised over 2 days using an energy metric (an approximate representation of the analysis error covariance metric; Palmer et al, 1997, Gelaro et al, 1997). In cases of significant spatial overlap between the first 16 SVs, higher-order SVs are used in the initial perturbation computation. The reason for choosing SVs as part of the strategy for ensemble forecasts is discussed in the papers cited above (see also Mureau et al, 1993).

A major upgrade of the Ensemble Prediction System (EPS) took place in December 1996 (Buizza et al, 1998). Firstly the resolution of the nonlinear model was increased to T_L159L31. The reason for this was to make the EPS more compatible with the operational deterministic forecast. Discrepancies between the EPS model and the high resolution operational model meant that it was often difficult to use the former to support (or otherwise) the latter, particularly when comparing weather elements. Secondly the ensemble size was increased to 51, using perturbations based on the first 25 singular vectors of the T42L31 forward tangent model. (As before, higher order SVs are used in case of significant overlap.) A principal reason for this is discussed in Gelaro et al, 1998; a significant increase in explained forecast error variance was possible using 25 rather than 16 singular vectors.

The purpose of this paper is to compare the performance of the EPS immediately before and immediately after the December 1996 upgrade (hereafter referred to as D96). As with any comparison of operational forecast performance between different years, the influence of atmospheric interannual variability is hard to take quantitatively into account. Nevertheless, the overall result, showing a

dramatic all-round improvement in EPS performance, is unlikely to be attributable solely to such interannual variability.

The performance comparison is discussed in section 2. A summary of future plans for further developing the EPS is given in section 3.

2. EPS PERFORMANCE COMPARISON

Fig 1 shows the skill of the ensemble mean forecasts relative to the skill of the control forecast for winter spring and summer seasons, before and after D96. It can be seen that in all seasons the rms error of the ensemble mean is lower with respect to the control forecast after D96. One particular aspect is worth pointing out; before D96 the ensemble mean scores at D+3 were slightly worse than the control, suggesting that nonlinear perturbation growth in the early medium range was partially detrimental. After D96, the ensemble mean scores are uniformly more skillful than the control.

Fig 2 shows time series of the distribution of ensemble scores (in terms of anomaly correlation coefficient over Europe at day 7). Before D96, the skill of the best forecast dropped close to or below 0.7 on a number of occasions during winter (Fig 2a) and, even more so during spring (Fig 2b). In this sense it could not be claimed that the observed flow pattern over Europe was always contained within the ensemble of forecast flow patterns. By contrast, the skill of the best member after D96 was uniformly above 0.7 for both winter (Fig 2c) and spring (Fig 2d); in this sense the ensemble solutions did contain the observed flow pattern to some reasonable degree of approximation. Another feature of the post-D96 EPS is that the skill of the best and the skill of the worst forecast does not depend nearly so much on the skill of the control as for the pre-D96 EPS. In some sense, the post-D96 EPS members do not hang onto the coat-tails of the parent control as much as before.

The rms spread of the EPS relative to the rms error of the control forecast is shown in Fig 3 for winter, spring and summer. Before D96 it can be seen that the spread was consistently smaller than the control error in all seasons. After D96, the spread agrees with the error very well at D+2, then drops below the level of the error. In the medium range the spread deficit is smaller after D96 than before D96. The fact that the spread at D+1 is smaller than control error is merely a consequence of the fact that the singular vector optimisation time is 48 hours. On the other hand the fact that the spread beyond D+2 is deficient suggests that there might be some missing source of spread, possibly not related to initial condition uncertainty. This will be discussed in section 3. It is interesting to note that in summer the spread is deficient at all ranges, before and after D96. Again this will be discussed in section 3.

The relationship between ensemble spread and control forecast skill is illustrated in Figs 4 for winter and spring. The figures show time series of D+7 rms spread and rms error over Europe. Each time series has been standardised with respect to the sample mean and standard deviation. A 5-day running mean has also been applied to damp some of the high-frequency daily fluctuations. There is a consistent improvement in the relationship between spread and skill after D96. This can be quantified by the correlation coefficients between the spread and skill curves. For winter, the correlation increased from 0.16 (before D96) to 0.43 (after D96); for spring the correlation increased from 0.39 to 0.63.

In order to assess the skill of the EPS as a provider of probabilistic forecasts, Figs 5-6 show the D+6 reliability curves and the associated (D+1 to D+10) Brier skill scores for two events: "the 850hPa temperature anomaly is less than -8K", and "the 24-hour accumulated precipitation is greater than 1 mm/day". The dashed lines show pre-D96 results for March-May, the solid lines show post-D96 results for March-May. As with the other verification diagnostics there is a clear improvement in skill. In fact, based on the Brier scores, there is about a 2-day increase in skill, a remarkable improvement if one compares with the impact of typical model or analysis changes on the skill of deterministic forecasts.

It can be asked whether the improvement in results shown was primarily due to the increase in model resolution, or primarily due to an increase in ensemble size. Certainly a 32-member subset of the post-D96 ensemble has similar performance to the full ensemble in terms of ensemble mean skill, spread-skill relationship, and Brier score. In this respect, it would appear that the principal reason for improved skill is the increased model resolution.

However, there are two important caveats to this result. Firstly, a 32-member ensemble drawn from the 50-member ensemble is not equivalent to a 32-member ensemble generated by 16 singular vectors. The use of 25 singular vectors gives a more uniform coverage of the northern hemisphere at initial time, and this may be important in determining the skill of the EPS.

Secondly, there are differences between 32 and 50 member ensemble probability performance, which are not readily evident from the Brier score statistics. For example, consider the event E: "the 850hPa temperature anomaly is less than -8K". The number of occurrences of this event in JFM 97 over Europe was 138. There were 178 times when the (post-D96) operational EPS predicted E at D+6 with probability between 90% and 100%. On this basis the EPS was overconfident; the expected number of occurrences should be (.95*178) leading to an excess of predicted (over observed) events of 30. Now if the first 32 members are taken from the EPS, and these statistics recomputed, then the number of times E was predicted with probability between 90% and 100% was 192. This smaller EPS was therefore even more overconfident, and the excess number of events forecast with near certainty increased to 44.

Put this way, the increase in ensemble size decreased by about 32%, the excess number of circumstances where the EPS predicted E with near certainty, and E did not occur in reality. From a practical point of view, it is essential that an EPS is very skillful when it is predicting an event with near certainty. Confidence in an EPS will be lost very quickly if an event is predicted with certainty, and the event does not occur. However, in absolute terms, the number of times the event occurred and was forecast with probability between 90% and 100% was small compared with the number of times it was forecast in other (10%-wide) probability intervals. From this point of view, these statistics relating to forecasts of high confidence, did not contribute significantly to the overall Brier score. This suggests that the Brier score should be modified, weighting it to take into account the practical importance of forecasts of high probability.

3. FURTHER DEVELOPMENTS

The ECMWF EPS system will continue to develop. One specific development is towards the incorporation of analysis error statistics into the singular vector computation. Preliminary work in this direction made use of the 3DVAR Hessian as an estimate of (the inverse of) the analysis error covariance matrix (Barkmeijer et al, 1997). However, whilst this Hessian might give satisfactory estimates of the analysis error covariance with respect to stationary components of the circulation, it severely underestimates analysis error with respect to transient activity, in particular associated with baroclinic waves. As a result, EPS experimentation with 3DVAR Hessian SVs have not been proved more skilful (in terms of Brier scores etc) than the operational EPS (Barkmeijer et al, 1998).

A simple expedient is currently under scrutiny as a temporary alternative to this approach, ie to add to the initial SVs, perturbations using evolved SVs from 2 days earlier. The evolved SVs can be thought of as describing the larger-amplitude less rapidly growing components of the analysis error associated with errors which have grown over previous analysis cycles. These evolved SVs are virtually orthogonal to the initial SVs. Hence the presence of the former will not compromise the growth of the latter. Results using evolved and initial SVs show marginal improvement in terms of Brier scores, and significant improvement in the first three days of the number of times the analysis lay outside the ensemble range (consistent with increasing the amplitude of the initial perturbations). The structure of initial perturbations is more consistent with the difference field between two operational analyses.

The shortcomings of the 3DVAR Hessian can only be overcome once a flow dependent background analysis error covariance matrix can be estimated. This can be achieved using the Kalman filter technique (Fisher and Courtier, 1995). In order to reduce the computational cost of a Kalman filter, the singular vector computation can be made interactive with the Kalman filter, in the sense that the Kalman-filter estimation of background error is active only in the subspace defined by the singular

vectors (M Fisher, personal communication). A full Kalman filter is being coded into a T21L3 quasi-geostrophic model in order to test the implication of this approximation to the Kalman filter equations (M Ehrendorfer, personal communication).

Further refinement to the initial perturbations will be made shortly using the physical parametrization scheme developed for the tangent model (Buizza et al, 1996). It is hoped that this will increase the ensemble spread in the summer season where it was found to be deficient both before and after D96 in the short range. The computation of diabatic SVs will allow initial perturbations to be added in the tropics.

The focus above on the development and refinement of techniques to create realistic initial perturbations is consistent with studies which attempt to attribute the difference between two operational forecasts valid for the same day, to initial condition differences or model differences. Specifically, differences between the UKMO and ECMWF forecasts have been compared with hybrid runs in which the UKMO analysis determines the initial conditions for an ECMWF model integration. On this basis, it has been found that up to day 5, the difference in initial conditions is dominant over the difference in model formulation (D. Richardson, personal communication).

On the other hand, the effects of differences in model formulation are not negligible in the later medium range, and the fact that model uncertainty is not taken into account in the current operational EPS may account for the "missing spread" when compared with the control error. In order to take this into account in some simple way, a means of adding stochastic perturbations to the total diabatic tendency is being developed at ECMWF. (R. Buizza, personal communication)

Future enhancements in computer resources could allow the EPS to increase in size, to use higher resolution models, and to extend beyond 10 days. Further experimentation will determine the best strategy for making use of such enhanced resources.

References

Barkmeijer, J., M. Van Gijzen and F. Bouttier, 1998: Singular vectors and estimates of the analysis error covariance metric. *Quart.J.Meteor.Soc.*, to appear.

Buizza, R., T. Petroliagis, T.N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons and N. Wedi, 1998. The impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart.J.Meteor.Soc.*, to appear.

Buizza, R., T.N. Palmer, J. Barkmeijer, R. Gelaro, J.F. Mahfouf, 1996. Singular vectors, norms and large-scale condensation. AMS preprints of the 11th Conference on NWP, 19-21 Aug. 1996, Norfolk, Virginia, US.

Fisher, M. and P. Courtier, 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation. ECMWF Tech. Memo. 220.

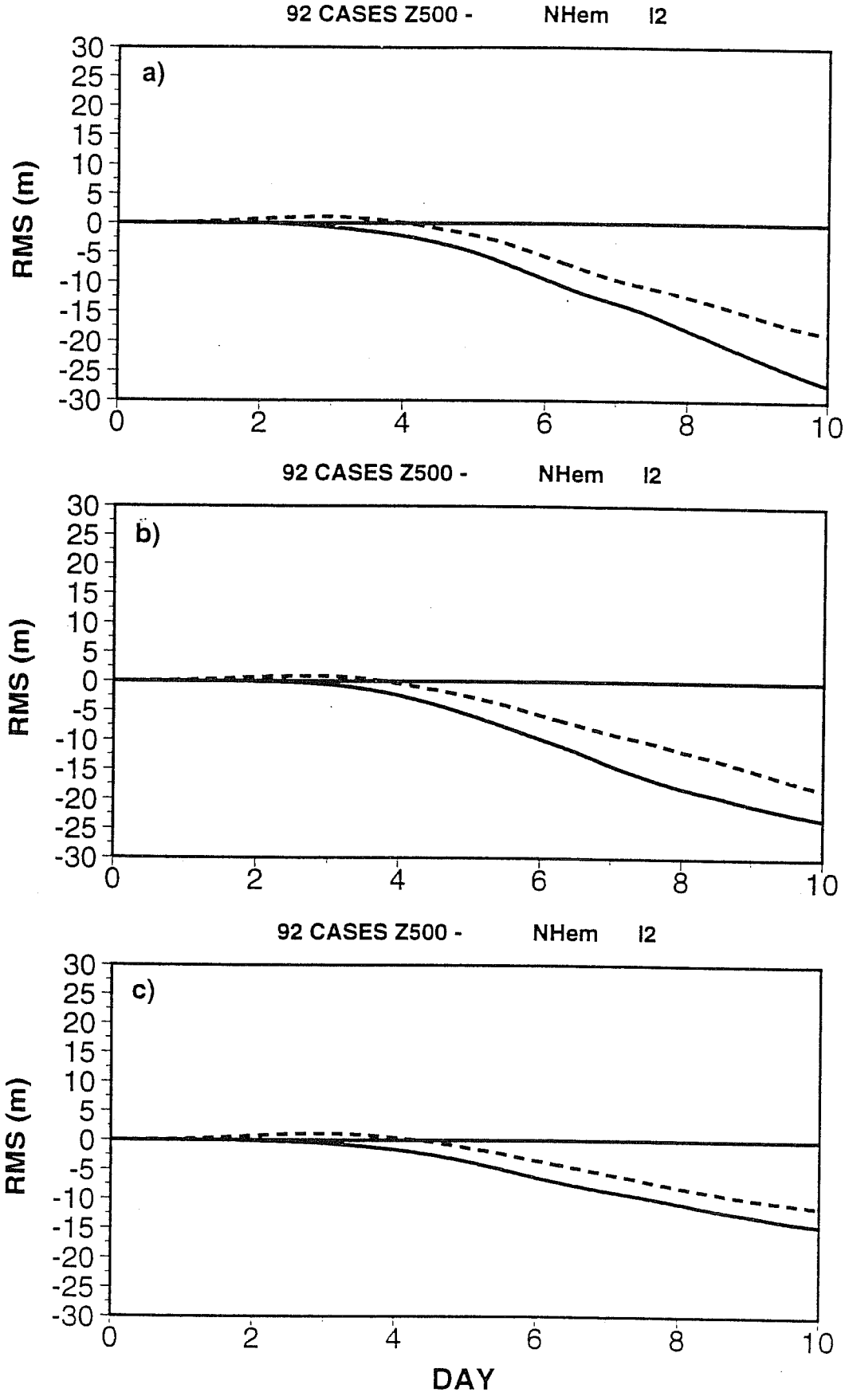
Gelaro, R., R. Buizza, T.N. Palmer and E. Klinker, 1998; Sensitivity analysis of forecast errors and the construction of optimal perturbations using singular vectors. *J. Atmos. Sci.*, **54,** to appear.

Molteni, F., R. Buizza, T.N. Palmer and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart.J.Meteor.Soc.*, **122**, 73-120.

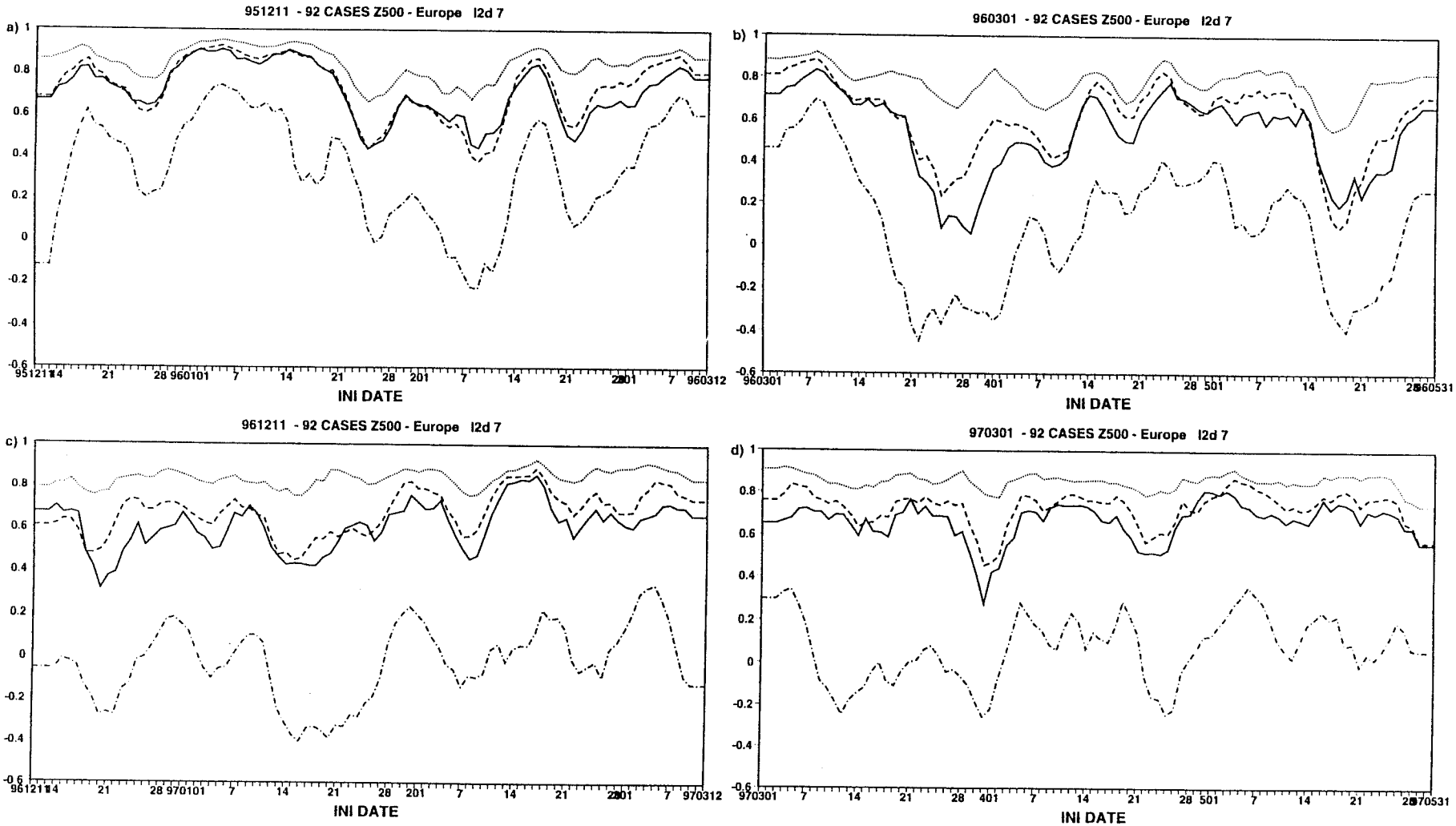
Mureau, F. Molteni, F. and Palmer, T.N., 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart.J.Meteor.Soc.*, **119**, 299-323.

Palmer, T.N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet and J. Tibbia, 1993: Ensemble prediction. *Proc. ECMWF Seminar Proc on Validation of Models over Europe. Vol.1,* Shinfield Park, Reading, UK., ECMWF, 21-66.

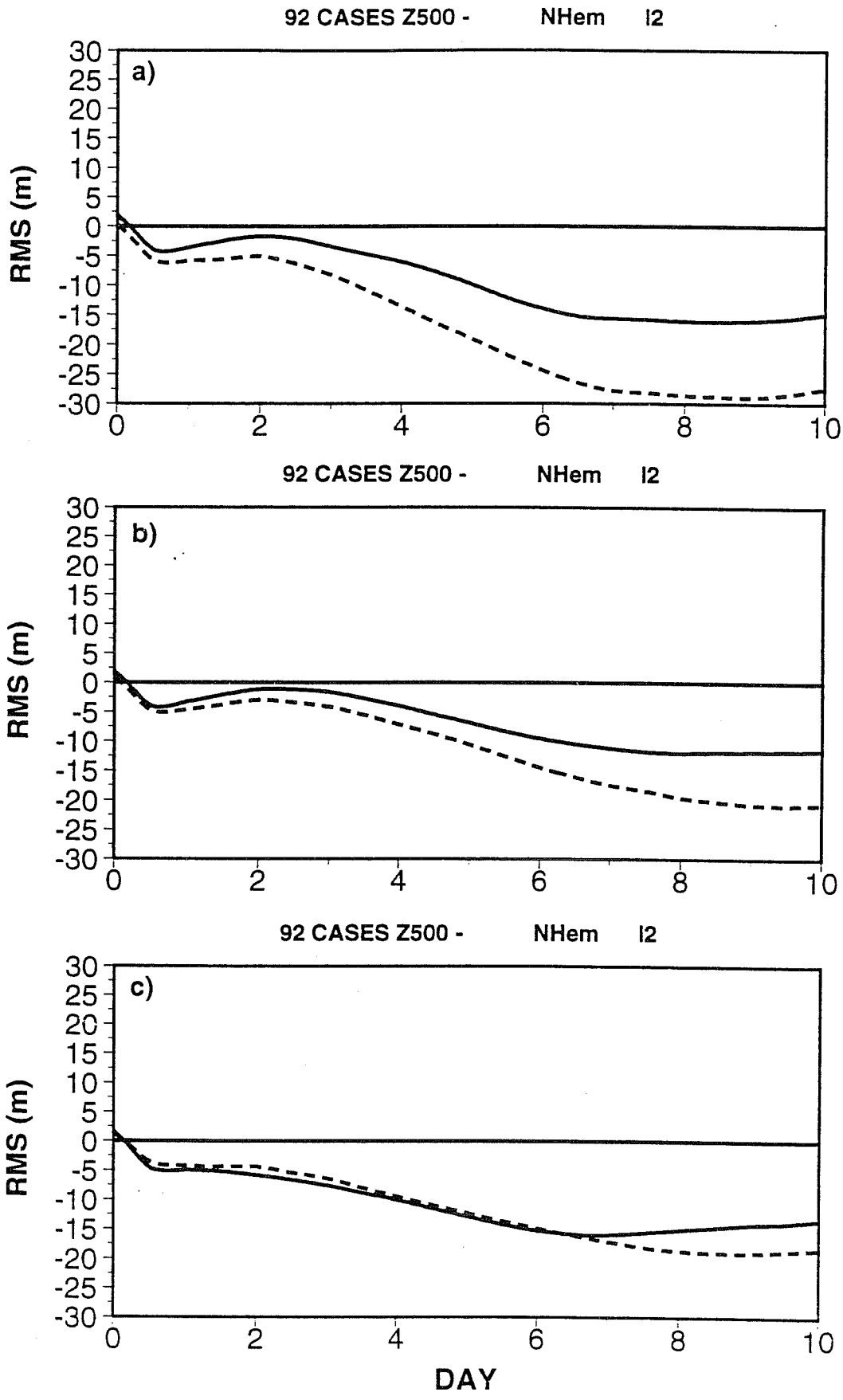
Palmer, T.N., J. Barkmeiger, R. Buizza and T. Petroliagis, 1997: The ECMWF Ensemble Prediction System. *Meteor. Appl.*, **4**, 301-304.



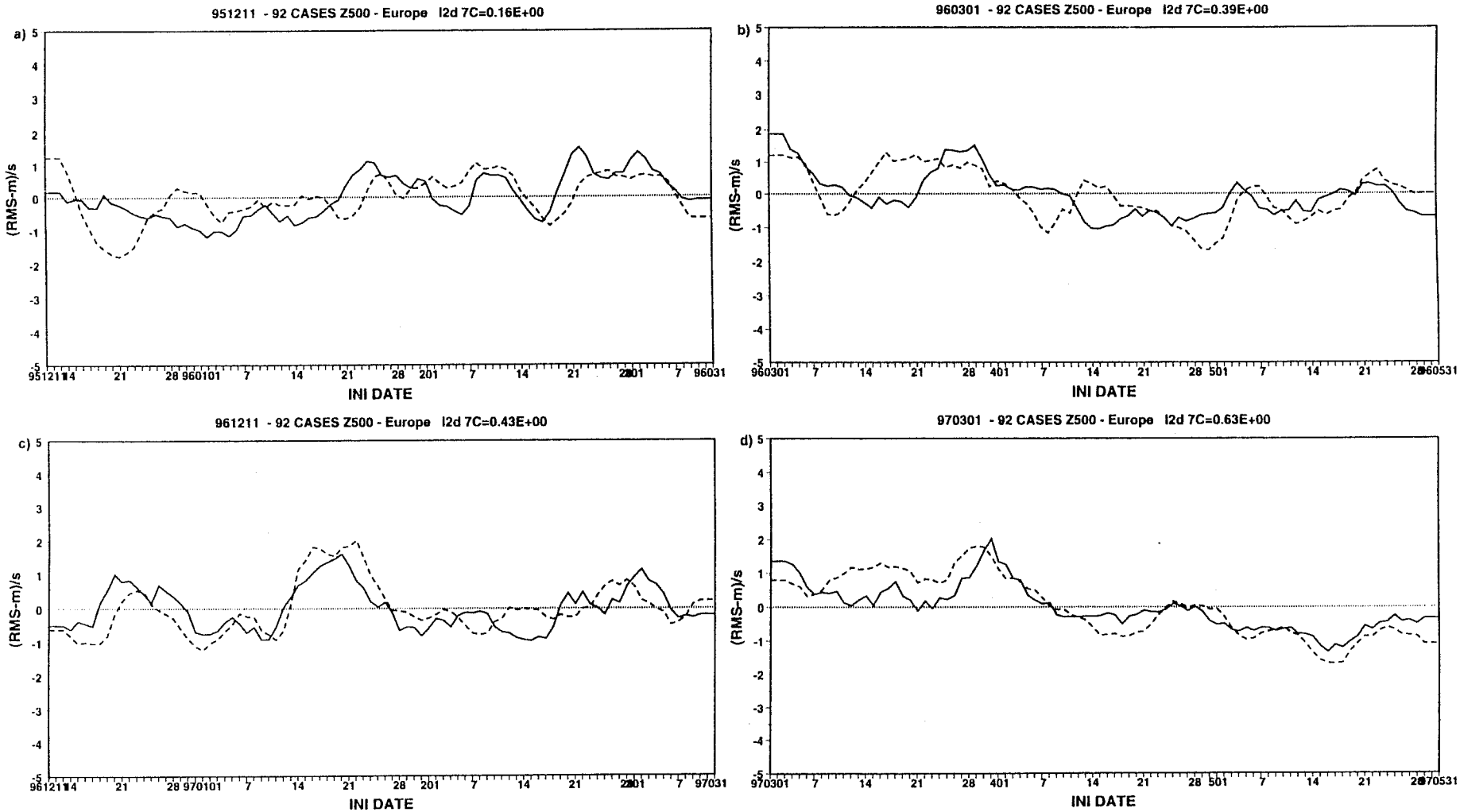
1. The northern hemisphere 500 hPa height rms error of the ensemble mean forecast, relative to the rms error of the control forecast. Dashed: pre-D96, solid: post-D96. Top: winter, middle: spring, bottom: summer.



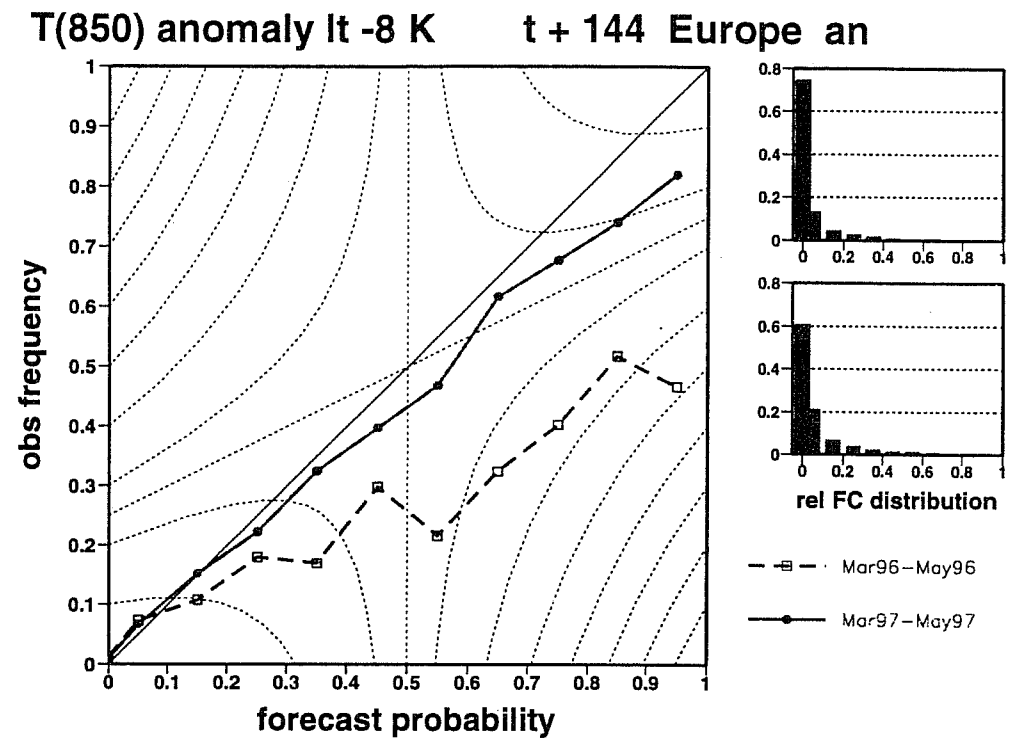
2. Time series of distribution of ensemble scores (500 hPa height anomaly correlation coefficient for day 7 over Europe. Solid: control, dashed: ensemble mean, dotted: best ensemble member, dash-dot: worst ensemble member. a) winter pre-D96, b) spring pre-D96 c) winter post-D96, d) spring post-D96.

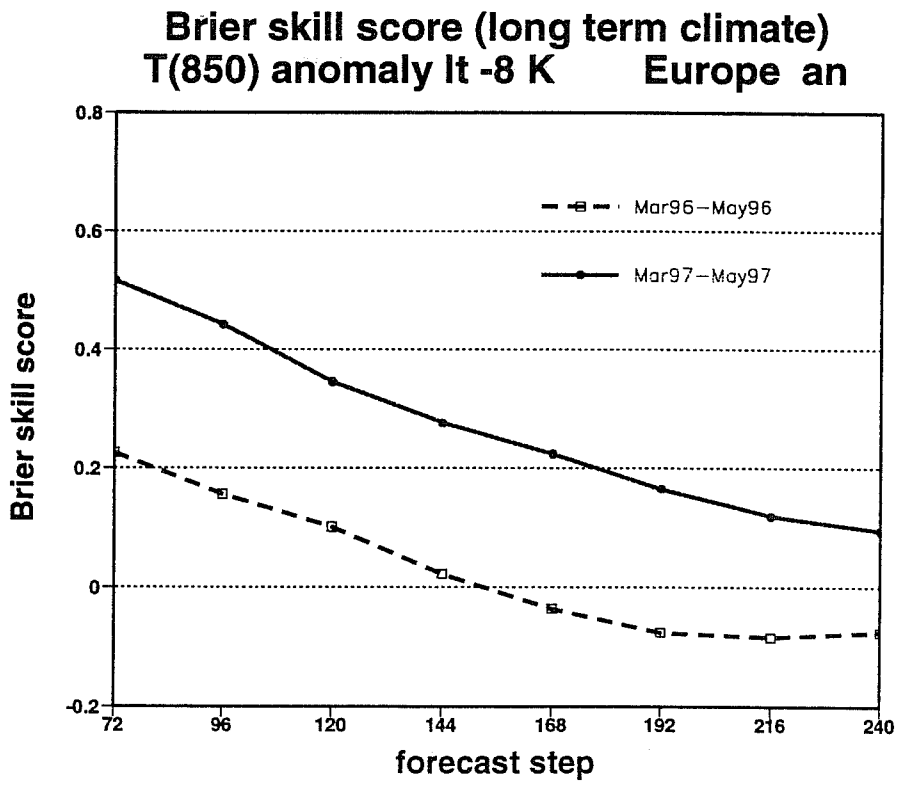


3. The northern hemisphere 500 hPa height rms spread from the ensemble forecast, relative to the rms error of the control forecast. Dashed: pre-D96, solid: post-D96. Top: winter, middle: spring, bottom: summer.

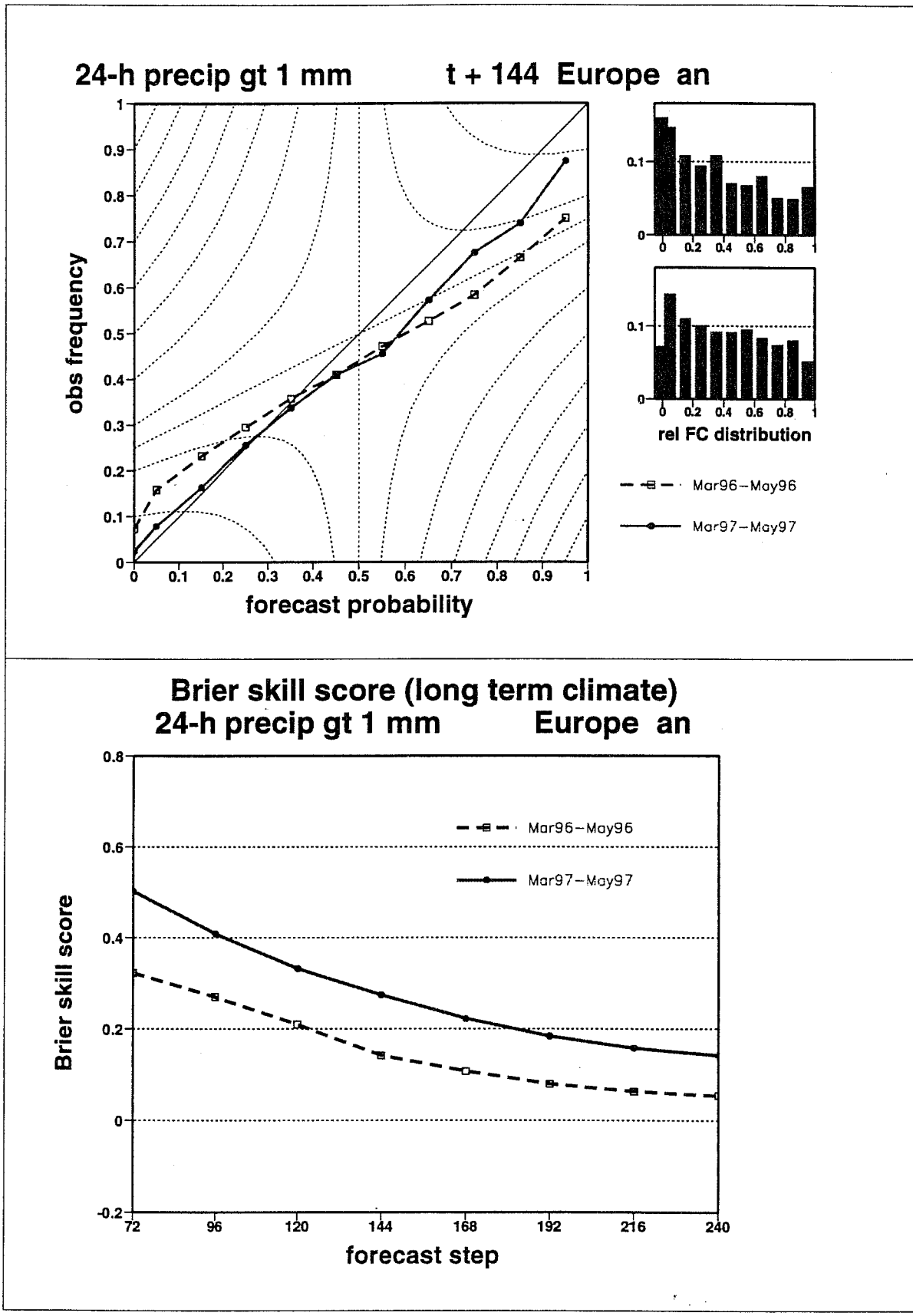


4. Time series of 5-day running mean ensemble spread (dashed) and control rms error (solid), based on 500 hPa height over Europe at day 7. Each curve has been standardised with respect to the sample mean and standard deviation. a) winter pre-D96, b) spring pre-D96, c) winter post-D96, d) spring post-D96.





5. Top: reliability diagram for the event E: 850 hPa temperature anomaly less that -8K at for grid points over Europe at forecast time D+6 (March - May). Dashed line: pre-D96. Solid line: post-D96. Bottom: Brier scores for the event E at different forecast times (March - May). Dashed: pre-D96. Solid: post D-96.



6. As Fig 5 but for the event: 24-hour accumulated rain is greater than 1 mm/day.