Validation of the ECMWF Ensemble Prediction System using empirical orthogonal functions.

Franco Molteni [1,2] and Roberto Buizza [2]

[1] *CINECA - Centro di Calcolo Interuniversitario dell'Italia Nord-Orientale*
Casalecchio di Reno, Bologna, Italy
[2] *European Centre for Medium-Range Weather Forecasts*
Reading, United Kingdom

Summary

Empirical orthogonal function (EOF) analysis of deviations from the ensemble mean was used to validate the statistical properties of T_L159 51-member ensemble forecasts run at the European Centre for Medium-Range Weather Forecasts during winter 1996/97. The main purpose of the analysis was to verify the agreement between the amount of spread variance and error variance accounted for by different EOFs. A suitable score was defined to quantify the agreement between the variance spectra in a given EOF subspace. The agreement between spread and error distribution for individual PC was also tested using the non-parametric Mann-Whitney test. The analysis was applied at day 3, 5 and 7 forecasts of 500-hPa height over Europe and North America, and of 850-hPa temperature over Europe.

The variance spectra indicate a better performance of the EPS over Europe than over North America in the medium range. In the former area, the excess of error variance over spread variance tends to be confined to non-leading PCs, while for the first two PCs the error variance is smaller than spread at day 3, in very close agreement at day 7. When averaged over a 6-EOF subspace, the relative differences between spread and error PC variances are about 25% over Europe, with the smallest discrepancy (15%) for 850-hPa temperature at day 7. Overall, the EPS produces a quite reliable estimate of the probability distribution of the atmospheric state over Europe.

1. Introduction

Medium-range ensemble forecasts are currently a part of the operational activities of two major numerical weather prediction (NWP) centres, namely the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Center for Environmental Predictions (NCEP) of the United States (Toth and Kalnay 1993; Molteni et al. 1996; Toth et al. 1997; Buizza et al. 1998). The validation of ensembles is undoubtedly a difficult task, and suitable verification precedures are under development at both centres in order to quantify the performance of these complex forecasting systems.

As already recognized in pioneering studies on stochastic-dynamic predictions (e.g. Epstein 1969), the purpose of ensemble forecasting is to provide an estimate of the time-evolving probability density function (PDF) for the atmospheric state. Given the enormous number of degrees of freedom in current NWP models, it is obviously impossible to analyse such PDF in the full phase space of the

model, and a suitable subspace must be chosen to validate the PDF properties. Usually, a first selection is made through the choice of one particular variable and vertical level; secondly, one may restrict the subspace dimension further by either considering a particular geographical area, or projecting the forecast fields onto a finite set of orthogonal functions, or using a combination of both techniques. A particularly simple example of the former method is given by the probabilistic verification of grid-point properties; even in such a case, however, the reliability of the predicted PDF is assessed by averaging the statistics over all grid points in a given area.

Empirical orthogonal functions (EOFs) are a well-known and efficient tool to reduce the dimensionality of the atmospheric phase space. As such, they are particularly suitable for the analysis and validation of the properties of the modelled PDF. Depending on how EOFs are defined, different types of validation can be performed. If EOFs are computed from a large sample of observed anomalies, and subsequently model fields are projected onto them (or viceversa), the comparison of PDFs of observed and modelled principal components (PCs) provides an estimate of the (flow-dependent) 'quasi-systematic' error of the model, while forecast verification can be performed by comparing time series of observed and predicted PCs (e.g. Ferranti et al. 1994; Wang and Rui 1996).

In the case of medium-range ensemble predictions, the relationship between ensemble dispersion and error is a more crucial issue than the average skill of individual forecasts. Here, EOF analysis can be more suitably used to determine which axes account for the largest proportions of ensemble spread on a case-to-case basis (as in Brankovic et al. 1990), and to verify whether forecast errors project on these axes in a way which is consistent with the ensemble PDF. In computing a covariance matrix from the ensemble members, the ensemble mean is naturally assumed as a reference point. Therefore, it is appropriate to use the error of the ensemble mean for such a comparison.

It is evident that one can only draw conclusions on the consistency between spread and error PCs by analysing such statistics on a relatively long period (one season at least). It should be pointed out that, in this type of analysis, the comparison of the PDFs of the observed and predicted PCs provides a probabilistic forecast verification, rather than an analysis of model systematic error. Since the position of the reference point (the ensemble mean) in phase space varies with time, one may find that the error projections have either a larger or a smaller variance than the ensemble PCs, even in the case when the climatological distributions of analyses and forecasts are the same.

In this paper, an EOF decomposition of spread and error is used to validate the performance of the ECMWF Ensemble Prediction System (EPS) during winter 1996/97. It is worth remembering that

since 10 December 1996 the EPS has been based on one unperturbed and 50 perturbed members at T_L159L31 resolution (i.e. triangular truncation at total wavenumber 159, linear grid for specrtal transforms and 31 vertical levels), while previously it included one unperturbed and 32 perturbed members at T63L19 resolution (see Molteni et al. 1996 and Buizza et al. 1998 for a full description of the system). In Sect. 2, the methodology and data used in the study are described in detail. Ensemble EOF patterns are presented in Sect. 3, while statistics on principal components of spread and error are analysed in Sect. 4. A brief comparison with the performance of the lower-resolution EPS run during winter 1995/96 is also shown in Sect. 4 using EOF statistics, while a similar comparison based on more traditional scores is presented in the Appendix. Sect. 5 addresses the issue of the presence of multimodality in the ensemble distribution. Finally, results are summarised in Sect. 6.

2. Data sets and validation methods.

a) Variables and space-time domains for EOF analysis

In defining a suitable space-time domain for the EOF analysis, choices about parameters, levels and geographical areas have to be made. In addition, one has the option of analysing fields at one particular forecast time or portions of trajectories over a time interval. As far as parameters and levels are concerned, our choice was a rather standard one (500-hPa height and 850-hPa temperature), and was based on the availability of a number of other verification statistics (some of which are presented in the Appendix).

As far the area and time domain are concerned, one should remember that a much flatter spectrum of EOF variances is obtained for hemispheric EOFs than for regional (i.e. European-scale) EOFs. Similarly, more EOFs are needed to describe trajectories than instantaneous fields at a given level of explained variance. Since the purpose of the analysis is to compare the spectra of spread and error variances in EOF space, it is desirable that the difference in variance between EOFs is larger than the sampling errors associated with the estimates of error variance. On the other hand, one would like the spatial domain to be large enough to guarantee that the EOF patterns are synoptically meaningful.

As documented in the following sections, a continental-size domain at a single forecast time allows 80% of the ensemble variance to be explained by 5-6 EOFs, with a change in variance of one order of magnitude between the first and the fifth/sixth EOF (typically, between 30-40% and 3-4%). This corresponds to an average variance ratio of about 1.5 between consecutive EOFs in the leading part of the spectrum. The uncertainty in the variance V for a sample of n independent elements is given by the standard deviation of sample estimates:

(1) $\sigma(V) = [\,(K - V^2)/n\,]^{1/2}$

where K is the kurtosis, or 4-th order moment, of the distribution. In the case of a standardized gaussian distribution and a sample of 90 data, $\sigma(V)/V$ is 15% for n=90, 21% for n=45. Since the auto-correlation of error PCs is small, one may conclude that a 3-month sample of daily forecasts should provide a good signal-to-noise ratio for the space-time domain described above.

In the following, the first winter of 51-member T_L159 ensembles (namely, from 11/12/96 to 12/3/97) will be analysed. Statistics will be presented for the first six EOFs computed (separately) at fc. day 3, 5 and 7 for the following variables and areas:
500-hPa geopotential height (Z) and 850-hPa temperature (T) over Europe (30-75 N, 20W-45E);
500-hPa geopotential height (Z) over North-America (30-75N, 60-150W).

b) Definition of EOFs and PCs

Given an ensemble of m members, let $\mathbf{D}$ be the matrix whose columns represent the deviations of individual members from the ensemble mean for one particular variable and forecast time over a given area. EOF analysis provides a singular-value decomposition of $\mathbf{D}$ as:

(2) $\mathbf{D} = \mathbf{E}\,\mathbf{S}\,\mathbf{P}^t$

where the m columns of matrices $\mathbf{E}$ and $\mathbf{P}$ represent the EOFs $\mathbf{e}_i$ and the time series $\mathbf{p}_i$ of the standardized PCs respectively, $\mathbf{S}$ is the diagonal matrix of the standard deviations s_i , and the 't' superscript denotes the transpose. These vectors are normalized as follows:

(3a) $\langle \mathbf{e}_i , \mathbf{e}_j \rangle = \mathbf{e}_i^t\,\mathbf{W}\,\mathbf{e}_j = \delta_{ij}$

(3b) $\langle \mathbf{p}_i , \mathbf{p}_j \rangle = \mathbf{p}_i^t\,\mathbf{p}_j = m\,\delta_{ij}$

where $\mathbf{W}$ is a diagonal matrix of latitude-dependent weights.

Since in our case m is much smaller than the number of grid points covering the selected area, it is convenient to compute the PCs as eigenvectors of the space-covariance matrix:

(4) $(\mathbf{D}^t\,\mathbf{W}\,\mathbf{D})\,\mathbf{p}_i = m\,s_i^2\,\mathbf{p}_i$

and then the EOFs from the relation:

(5) $\mathbf{e}_i = (m s_i)^{-1} \mathbf{D} \mathbf{p}_i$.

If the vector $\mathbf{d}^a$ represents the deviation of the verifying analysis from the ensemble mean (ie the opposite of the ensemble-mean error), the standardized projections p_i^a of this vector onto the EOFs $\mathbf{e}_i$ are given by:

(6) $p_i^a = s_i^{-1} \langle \mathbf{e}_i , \mathbf{d}^a \rangle = s_i^{-1} \mathbf{e}_i^t \mathbf{W} \mathbf{d}^a = (m s_i^2)^{-1} \mathbf{p}_i^t (\mathbf{D}^t \mathbf{W} \mathbf{d}^a)$

c) Analysis and verification of PC distributions

When the analysis described above is repeated over a set of n initial dates, we obtain for each PC two sets of coefficients, representing deviations from the ensemble mean of the ensemble members and of the verifying analyses (for brevity, in the following we shall refer to them as to the spread-PCs and error-PCs, bearing in mind that the sign of the error is reversed for consistency with the spread definition). If $j=1,..,m$ is the index of the ensemble members and $k=1,...,n$ the index of the initial dates, we can indicate the two datasets as $\{p_{jk}\}_i$ and $\{p_k\}_i^a$ respectively, where the EOF index i varies from 1 to the dimension of the selected EOF subspace n_{EOF} . If the ensemble provided a correct estimate of the atmospheric PDF as a function of time, then the distributions of PC datasets $\{p_{jk}\}_i$ and $\{p_k\}_i^a$ should only differ because of sampling errors.

By construction, the mean of each $\{p_{jk}\}_i$ is equal to zero and its standard deviation (and mean-square value) to one. A first check is therefore to compute the mean and standard deviation of $\{p_k\}_i^a$ and verify their differences from the reference values. Since the sign of the EOFs is arbitrary, and the EOFs vary from day to day, it is not evident a-priori what is the significance of a difference in the mean value. To make the EOF sign non-arbitrary, it was decided to orient each EOF in such a way that the control forecast (corresponding to j=1) always has positive PCs. In this way, a significant positive bias may be a sign that the analysis agrees with the control forecast more closely than expected on the basis of the ensemble PDF (as it would be expected, for example, if the ensemble PDF was biased towards a climatological distribution).

Instead of checking the similarity of the mean and standard deviation using parametric tests, it was decided to use the non-parametric Mann-Whitney test to evaluate the consistency of the spread and error distribution. The test was implemented in such a way to represent a validation of the rank histograms (O. Talagrand, private communication) already used for the verification of grid-point data:

- for each EOF and initial date, the error-PC was converted into a rank (ranging from 0 to m) by comparing it with the m values of the spread-PC; the consistency of the spread and error PDFs would imply a flat distribution of such ranks;
- the sum S_i^a of the n ranks $\{r_k\}_i^a$ was computed, and the probability $P1_i$ that a random sample of n elements had a sum greater than (or equal to) S_i^a was evaluated;
- the fraction Fo_i of outliers (ie of error-PC with rank either 0 or m) was also evaluated, together with the probability Po_i of such a value being exceeded in a random sample.

The Mann-Whitney test was also peformed on the squared values of the PCs, which is equivalent to a test on the similarity of variances. In this case, the probability of the rank sum being exceeded by random sampling will be denoted by $P2_i$. (In the following the EOF index i may be omitted, being implicit that all statistical indices are computed for each EOF separately.)

d) Definition of the EVE score (Error of Variances in EOF space)

For the ensemble spread, the partition of variance between EOFs is represented by the average fraction of variance f_i^{var} accounted for by each EOF. In estimating the corresponding value for the error PC, either the sample variance (computed with respect to the sample mean) or the mean-square-value should be used, depending on whether the sample mean is significantly different from zero or not. Since, as discussed later, the cases in which there is a significant bias are a minority, we will use the mean-square value of error-PCs as 'safer' estimates of error variances V_i^a along the various EOF axes. If a plot of f_i^{var} as a function of the EOF index provides the spectrum of the spread variance, the corresponding spectrum of the error variance (normalized with respect to the total ensemble variance) is given by $f_i^{var} V_i^a$.

For a given EOF subspace where $i=1,...,n_{EOF}$, an index of the similarity between the two spectra can be defined as:

$$(7) \qquad EVE = \Sigma_i \, f_i^{var} \, |V_i^a - 1| \, / \, \Sigma_i \, f_i^{var}$$

where EVE stands for Error of Variances in EOF space, and can be viewed as the L^1 norm of the spectrum difference, renormalized by the total variance in the subspace.

e) PDFs of spread and error PCs

Finally, PDFs of spread and error-PCs have been computed using a gaussian kernel estimator (e.g. Silverman, 1986), where different smoothing parameters have been adopted for the $\{p_{jk}\}_i$ and $\{p_k\}_i^a$

dataset to account for the much fewer data in the latter sample. To evaluate the significance of the difference between the two PDFs, PDFs have also been computed for *m-1* subsamples obtained by selecting the *j*-th ensemble member (with the exclusion of the control forecast) from each ensemble, and therefore including the same number of data as the error-PCs. The dispersion of the subsample PDFs around the PDF for the total sample of spread-PCs provides a confidence band, in which the PDF of error PCs should be included in the absence of systematic differences.

3. Spatial patterns of ensemble EOFs.

In this section, some examples of EOF patterns at different forecast times will be briefly discussed. For brevity, attention will be focussed on 500 hPa height fields over Europe.

Fig. 1 shows the spatial pattern of the first three EOFs for the 3-day forecast started on 11/12/96, together with the (analysis - ensemble mean) error and its projection onto the 3-EOF subspace. The EOFs are scaled by the associated standard deviation; the fraction of variance explained by each EOF and the standardised error PCs are listed above the corresponding EOF panel. In this particular case, the first EOF explains twice as much variance as the second one, and its strucure is concentrated in the Atlantic portion of the domain; ensemble spread over continental regions is mostly accounted for by the following EOFs. Together, the first three EOFs explain over 78% of the ensemble variance, and the error projection onto this subspace provides a fairly accurate representation of the field.

Fig. 2 is the same as Fig. 1, but for the 3-day forecast started on the following day. The first EOF has a similar structure to the first EOF of the previous day, but with an eastward shift of the main features; conversely, the error pattern is shifted westward. For this day, the first two EOFs explain a comparable proportion of variance. Howerer, the 3-EOF subspace explains a smaller amount of variance (73%) than in the previous day, and indeed the error projection is a less effective representation of the total field.

The spatial scale of EOFs tends to increase with forecast time, as shown in Fig. 3, which now refers to the day-7 forecast started on 12/12/96. With a 69% fraction of explained variance, again the 3-EOF subspace provides a realistic projection of the forecast error. A comparison with the previous day (not shown) shows a closer similarity for EOF 1 than at fc. day 3, while the similarity is much weaker for subsequent EOFs.

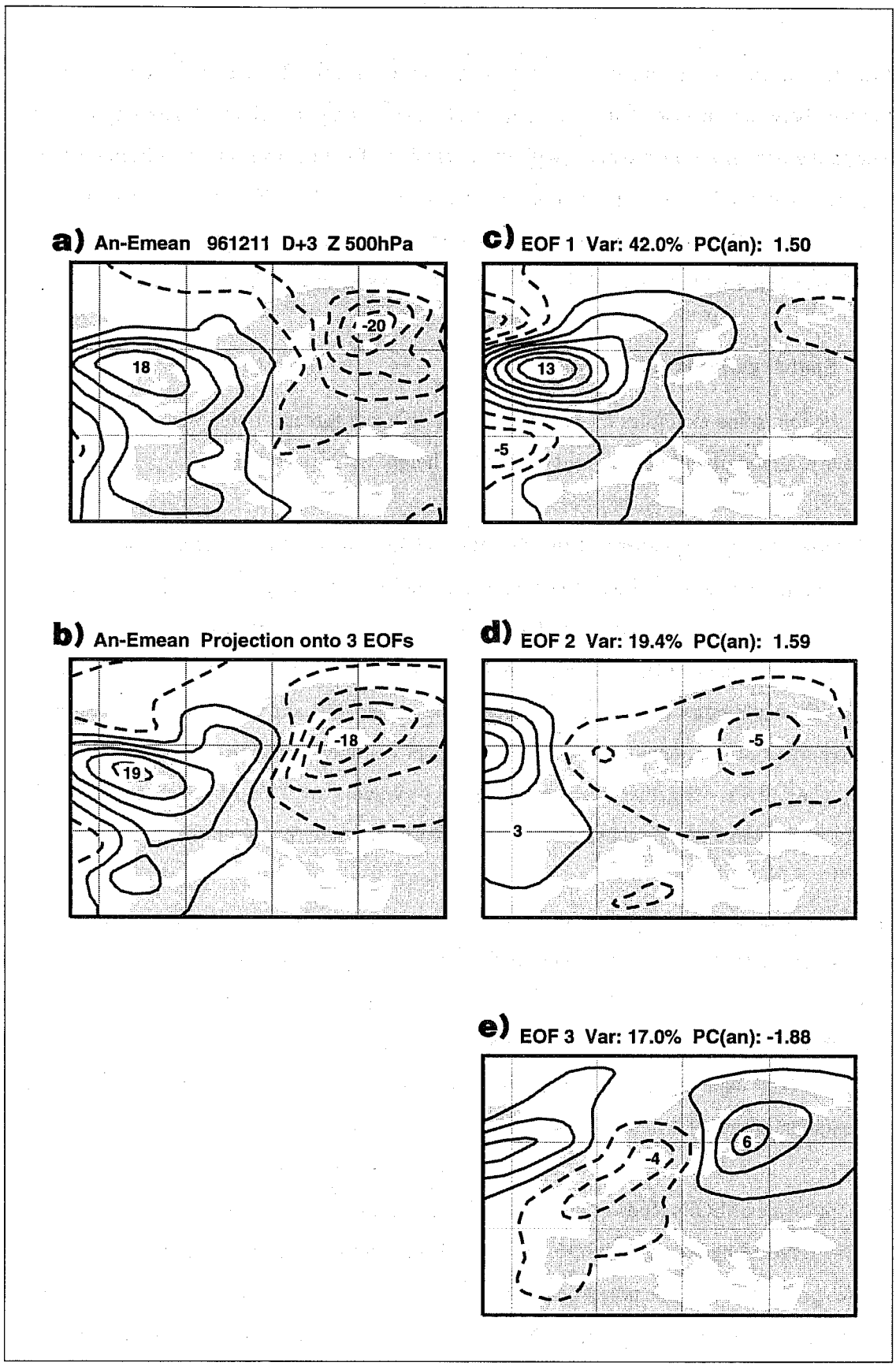

Fig. 1 a) Difference between verifying analyis and ensemble mean of 500-hPa height at fc. day 3, for the ensemble started on 11/12/96. b) As in a), but projected over the first three EOFs of ensemble spread. c), d) and e) First three EOFs of ensemble spread for 500-hPa height over Europe.

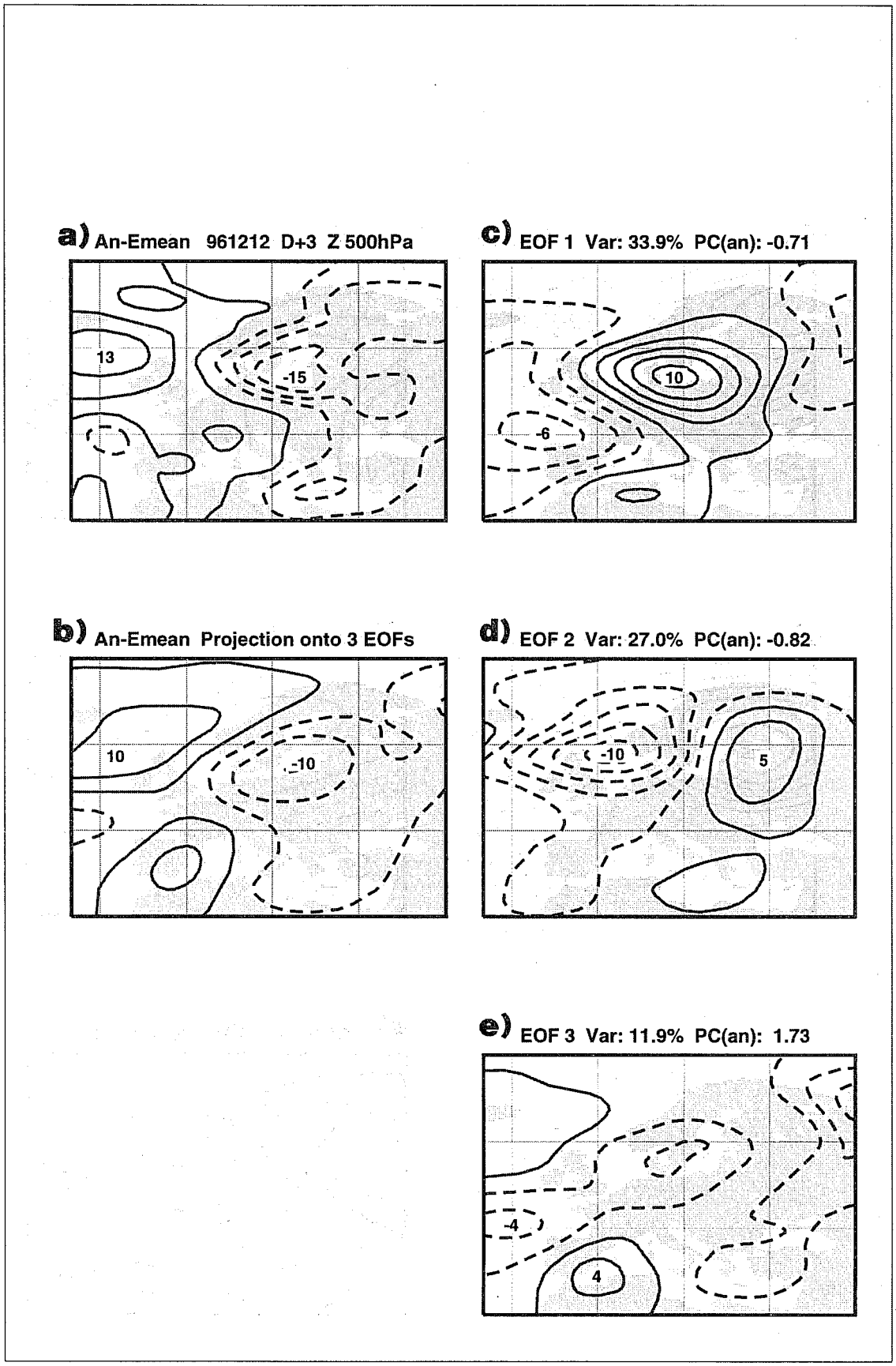

Fig. 2 As in Fig. 1, but for the ensemble started on 12/12/96 at fc. day 3.

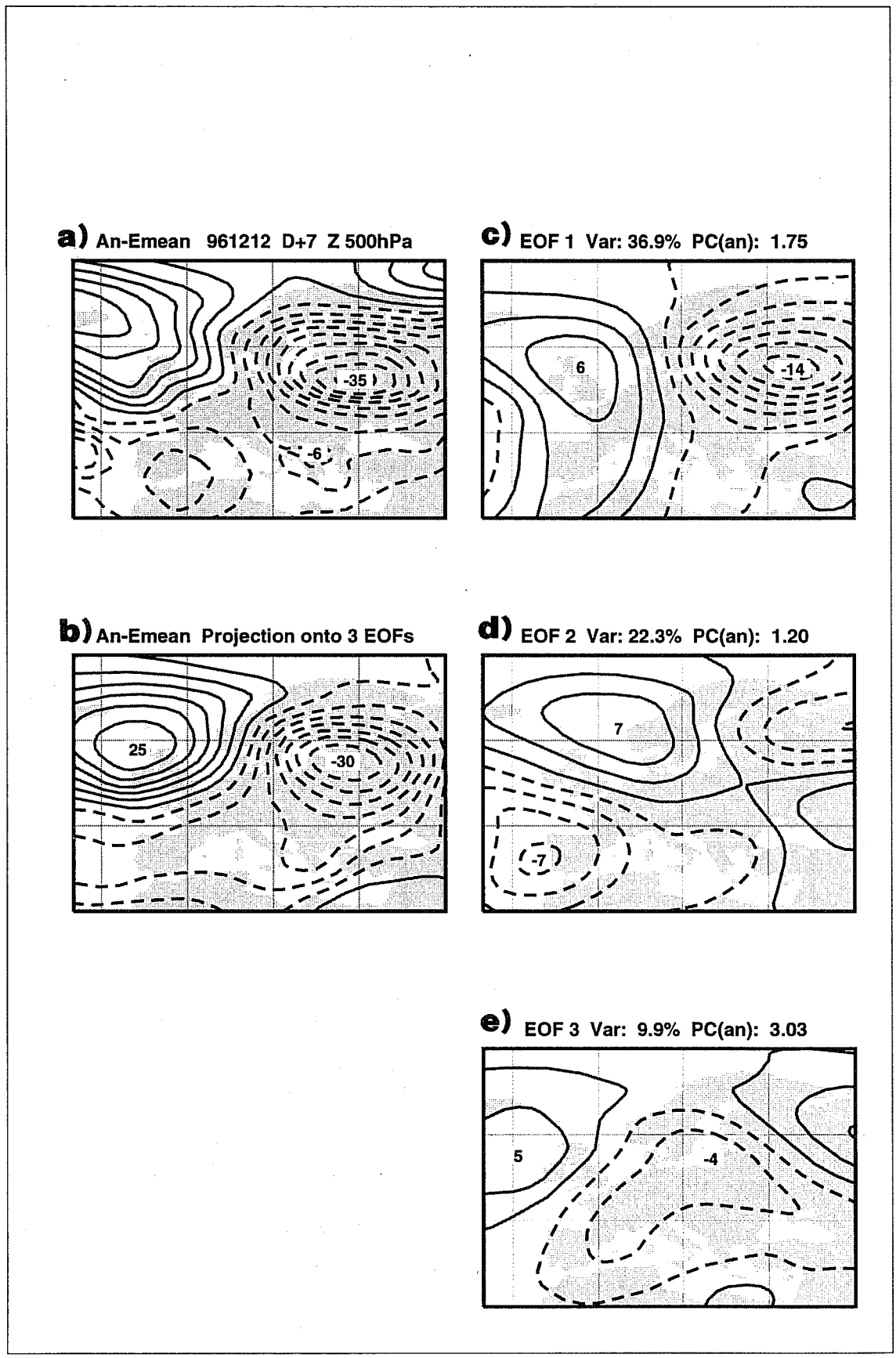

Fig. 3 As in Fig. 1, but for the ensemble started on 12/12/96 at fc. day 7.

4. Distributions of spread and error PCs.

a) Spectra of variance in EOF space.

For each variable, domain and forecast time, the variance distributions of spread and error in EOF space are represented by the values of f_i^{var} and $f_i^{var} V_i^a$ respectively. Fig. 4 shows the variance spectra at fc day 3, 5 and 7 for the 6-EOF subspace of 500-hPa height over Europe. A 95% confidence band for variance estimates based on one value per day (as for the error PCs) has also been computed from the spread-PC distribution using Eq. 1, and its lower and upper limits are also plotted. The cumulative fraction of variance explained by the subspace and the EVE score defined by Eq. 7 are listed above each panel.

As far as the spread variance is concerned, it is interesting to note that its spectrum is slightly steeper at day 3 than at day 7, and the proportion of explained variance actually shows a (very modest) decrease with forecast time. This contradicts the experience with 'traditional' EOF analysis of high-frequency versus low-frequency variability (or of short-range versus medium-range forecast errors), where the reference state (usually the time mean) is fixed in phase space. In such an analysis, many more EOFs are needed to explain a given fraction of variance for high-frequency (or short-range) fields than for low-frequency (or medium-range) fields.

However, when the reference point varies in time following an observed or modelled trajectory, the EOF spectrum is related to the *local* embedding dimension of the attractor when the short-range evolution of the system is considered, to the *global* embedding dimension when the long-term evolution is analysed (over a time comparable to the limit of deterministic predictability). It therefore appears that, by day 3, the stretching of the ensemble cloud from a sphere to an ellipsoid (associated with linear perturbation growth) has already taken place, and the 'local' phase space dynamics is already strongly influenced by non linear processes. Indeed, the (anti-)correlation between PCs of ensemble members which start from opposite perurbations is already small at day 3, ranging (on average) between -0.45 for PC 1 to -0.15 for PC 6.

Comparing the error variance with the spread variance, one notes that at all forecast times the ratio between error and spread increases with increasing EOF index. At day 3, the spread along the first two EOFs is actually larger than the error, while the error variance significantly exceeds the spread variance only from the 6th EOF onwards. At day 7, the projections on the first two EOFs have almost perfect statistics, while the error exceeds the spread from the 3rd EOF. This indicates that the discrepancy between error and spread comes from errors which have small projections on the axes associated with the leading dynamical instabilities. Flow-dependent model errors are likely to exhibit

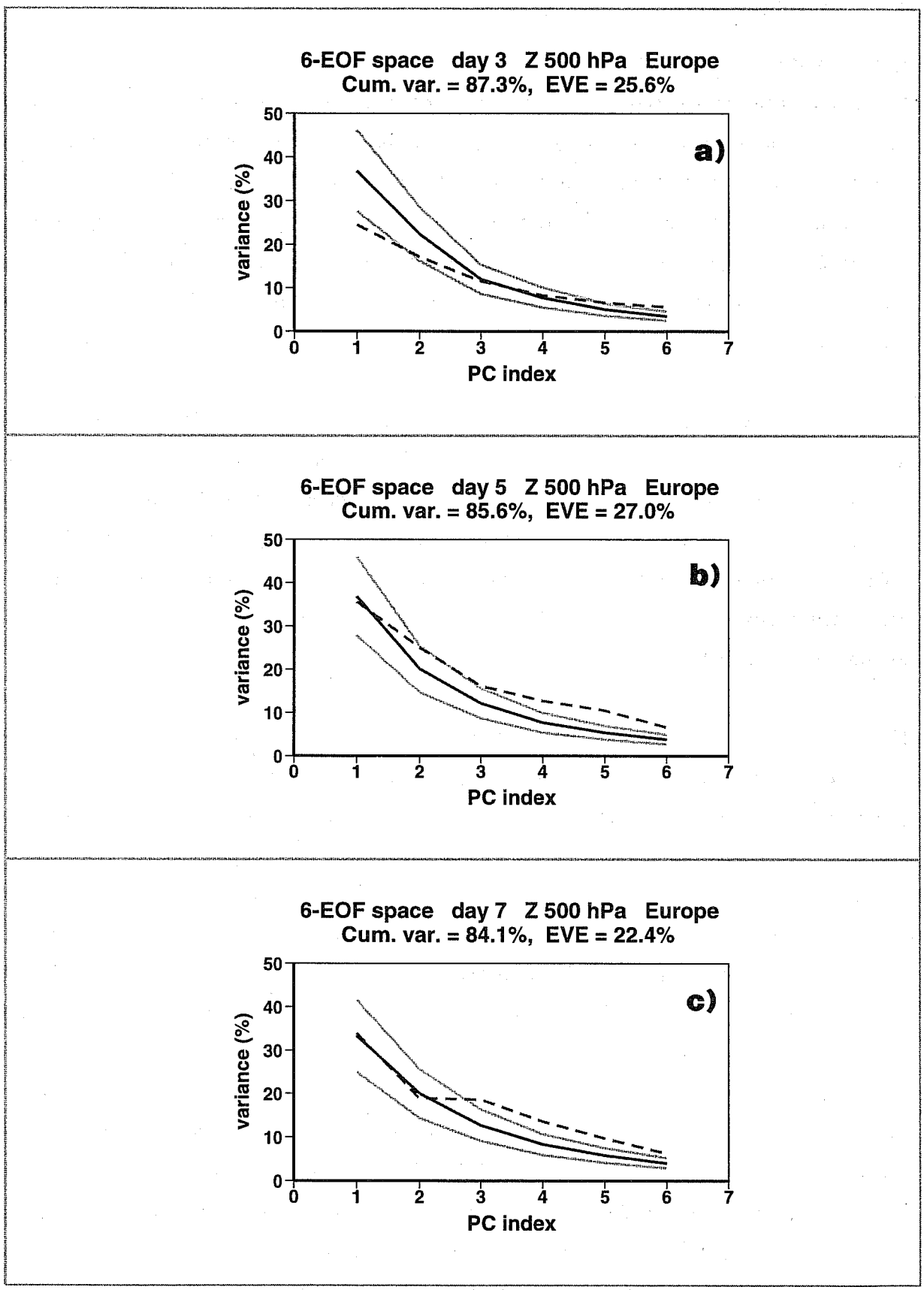

Fig. 4 Variance distribution for spread and error PCs of 500-hPa height over Europe in winter 1996/97, at fc. day 3 (a), fc. Day 5 (b) and fc. day 7 (c). Black solid line: variance of spread-PCs; black dashed line: variance of error-PCs; grey solid lines: limits of the 95% confidence band for estimates of spread variance based on one value per day, as in the samples of error-PCs. Cumulative variance and EVE score are listed above each panel (see text).

such a behaviour, although one cannot rule out the existence of slowly-growing analysis errors which are not described by the leading singular vectors.

It should be noted that the EVE score has similar values at the three forecast times considered, with the smallest (i.e. best) value being obtained at day 7. While a comparison of the total variance of spread and error shows very little discrepancy around day 3 and an increasing gap in the medium range (see the Appendix), the EVE scores reflects the presence of both overstimated and underestimated PC variances at day 3, and the better fit between error and spread spectra in the medium range.

This behaviour is even more evident when looking at the results for 850-hPa temperature over Europe, shown in Fig. 5. At day 3, the error variance is overestimated by the ensemble for the first 5 EOFs, with a significant difference for the first two. Already at day 5, however, error variances are either within the confidence band or close to its upper limit, and the fit between the spectra of error and spread variance is even better at day 7. For this variable, the EVE score clearly decreases with forecast time, and the day-7 value (15.4% only) indicates a very satisfactory performance of the EPS over Europe in the medium-range.

For the North American region, the analysis of 500-hPa height (see Fig. 6) provides a less optimistic picture. Underestimation of error variance by the ensemble spread is evident for all EOFs at day 5 and 7, and at the latter time none of the EOF variances is within the confidence band. The EVE scores increases from a value slightly better than the European score at day 3, to a value 2.5 times as large at day 7. For this region, either some type of fast-growing analysis errors are poorly represented in the initial conditions, or model errors tend to feed the leading dynamical instabilites in a more severe way than they do over Europe.

Finally, the performance of the lower-resolution EPS used in winter 1995/96 is verified in Fig. 7 by looking at the variance distribution of 500-hPa height over Europe in that period. The comparison with Fig. 4, which shows the same statistics for the latest winter, reveals a dramatic improvement in the EPS consistency, especially in the late-medium range. At day 7, the 6-EOF EVE score indicates that the discrepancy between spread and error variance was four times larger in winter 1995/96 than in 1996/97 (see the Appendix for a comparison between the two winters based on 'traditional' scores).

b) PDFs of spread and error PCs

In this subsection, some examples of PDFs of spread and error PCs for 500 hPa height over Europe will be presented, together with the results of the Mann-Whitney test on the similarity between the two

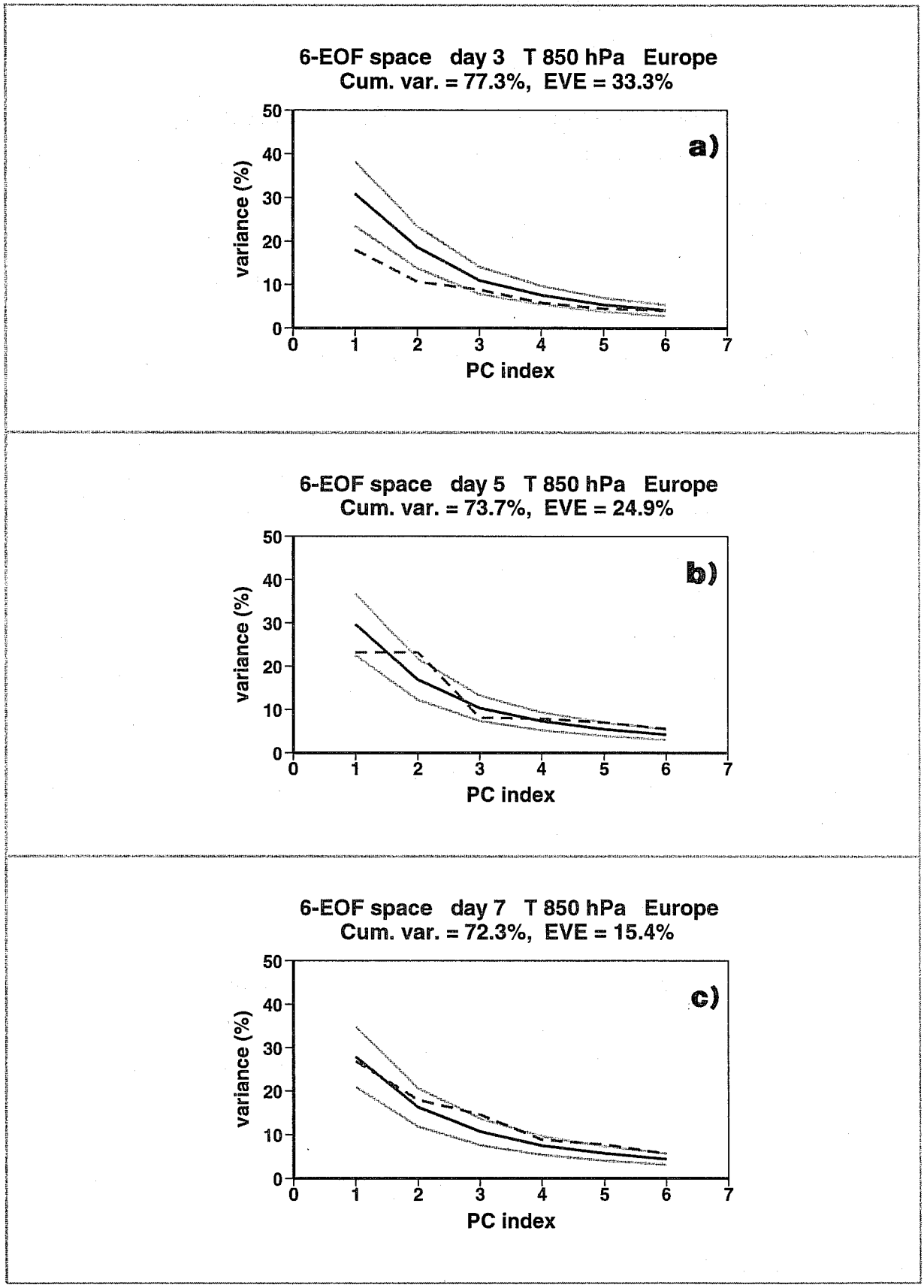

Fig. 5 As in Fig. 4, but for 850-hPa temperature over Europe in winter 1996/97.

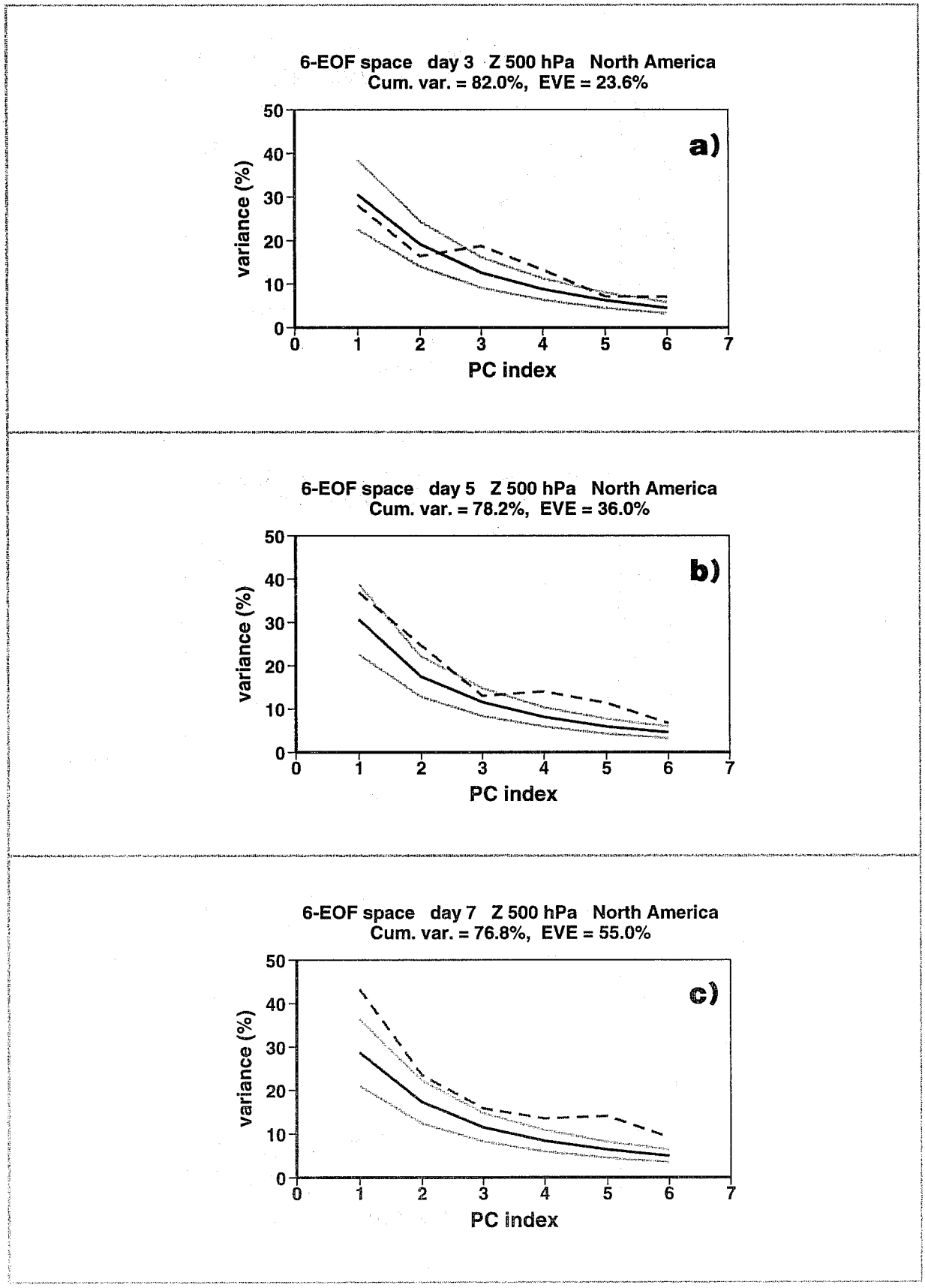

Fig. 6 As in Fig. 4, but for 500-hPa height over North America in winter 1996/97.

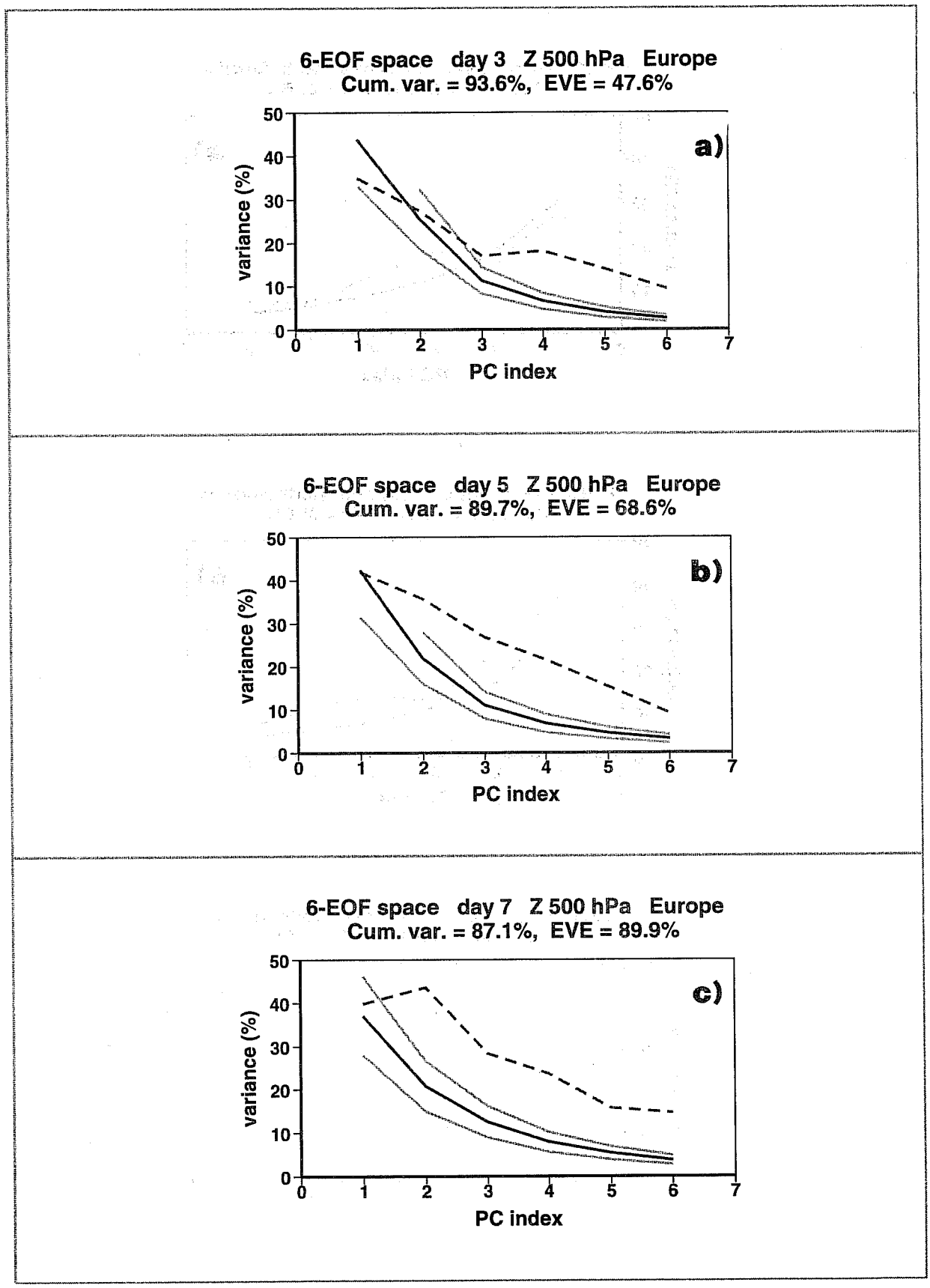

Fig. 7 As in Fig. 4, but for 500-hPa height over Europe in winter 1995/96.

distributions.

Fig. 8 shows the error and spread PDFs for the first three PCs at day 3. The significance of the difference between the two PDFs can be visually judged by comparing it with the width of the PDF band originated by subsamples of one perturbed member per ensemble; for an objective assessment, the Mann-Whitney statistics defined in Sect. 2c are listed above each panel. The PDF of the control forecast, always possessing positive PCs by construction, is also plotted. Looking at the PDF for the first PC, one clearly notices the smaller variance of the error PDF with respect to the spread distribution. On average, the analysis tends to reside on the same side of the control forecast with respect to the ensemble mean, and the Mann-Whitney tests confirm the significance of the discrepancies. A positive bias can also be found for PCs 2 and 3, but with much smaller significance. It is interesting to note that the deviations of the control from the ensemble mean tend to become larger for higher-order PCs; this is true also for the day-3 PCs of 850-hPa temperature (not shown).

At day 7 (see Fig. 9), the differences between the spread and error PDFs of the first two PCs are clearly within the uncertainty associated with sampling. For the first PC, the error PDF has two well separated maxima, one of them corresponding to the average position of the control forecast. The spread PDF, however, is unimodal, with just a hint of non-gaussian behaviour. A more definite unimodal shape is shown by the PDF of spread PCs 2 and 3. However, while for PC 2 the correspondence with the error PDF is very strong, for PC 3 the error shows a flatter distribution with larger variance. Note that, since the differences between the spread and error PDFs of PC 3 have a rather symmetric character, the bias is very small and the Mann-Whitney test on the actual PC values fails to detect the significance of such differences. However, the test performed on the squared PCs indicates that sampling has a very small probability (0.5%) to be the only source of the discrepancy.

5. Multimodality in the ensemble distributions.

As shown in the previous section, some of the PDFs of error-PC show a bimodal structure which is not reproduced in the PDF of spread PCs. A particularly evident case of such a behaviour is the first EOF of 500-hPa height over Europe at day 7, shown in Fig. 8a. From this figure, however, one may note that the PDF of the error PC is well within the band of the PDFs of one-member subsamples, and therefore one should assume that the difference between the spread and error distribution is not significant.

However, bimodality should not be a constant property of ensemble PDFs. It should only be evident in cases when the ensemble trajectories tend to bifurcate, after passing through an instability region in

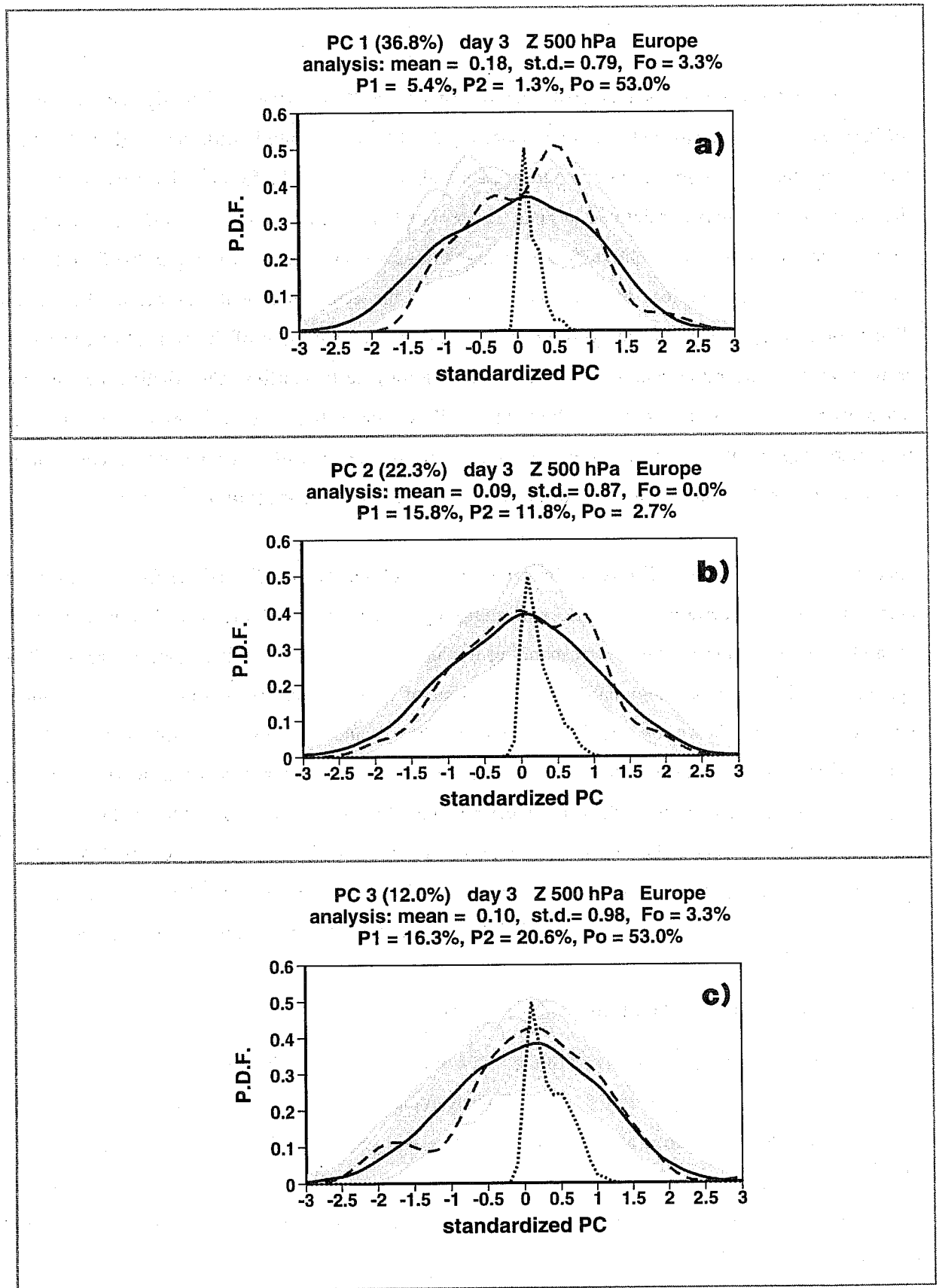

Fig. 8 Probability density functions (PDFs) of spread and error PCs of 500-hPa height over Europe at fc. day 3 in winter 1996/97, for PC 1 (top), PC 2 (centre) and PC 3 (bottom). Black solid line: PDF of spread-PC; black dashed line: PDF of error-PC; black dotted line: PDF of spread PC for the control forecast; grey solid lines: PDFs of spread-PC for subsamples including one perturbed ensemble member per day. See text, sect. 2c, for the definition of the statistics above each panel.

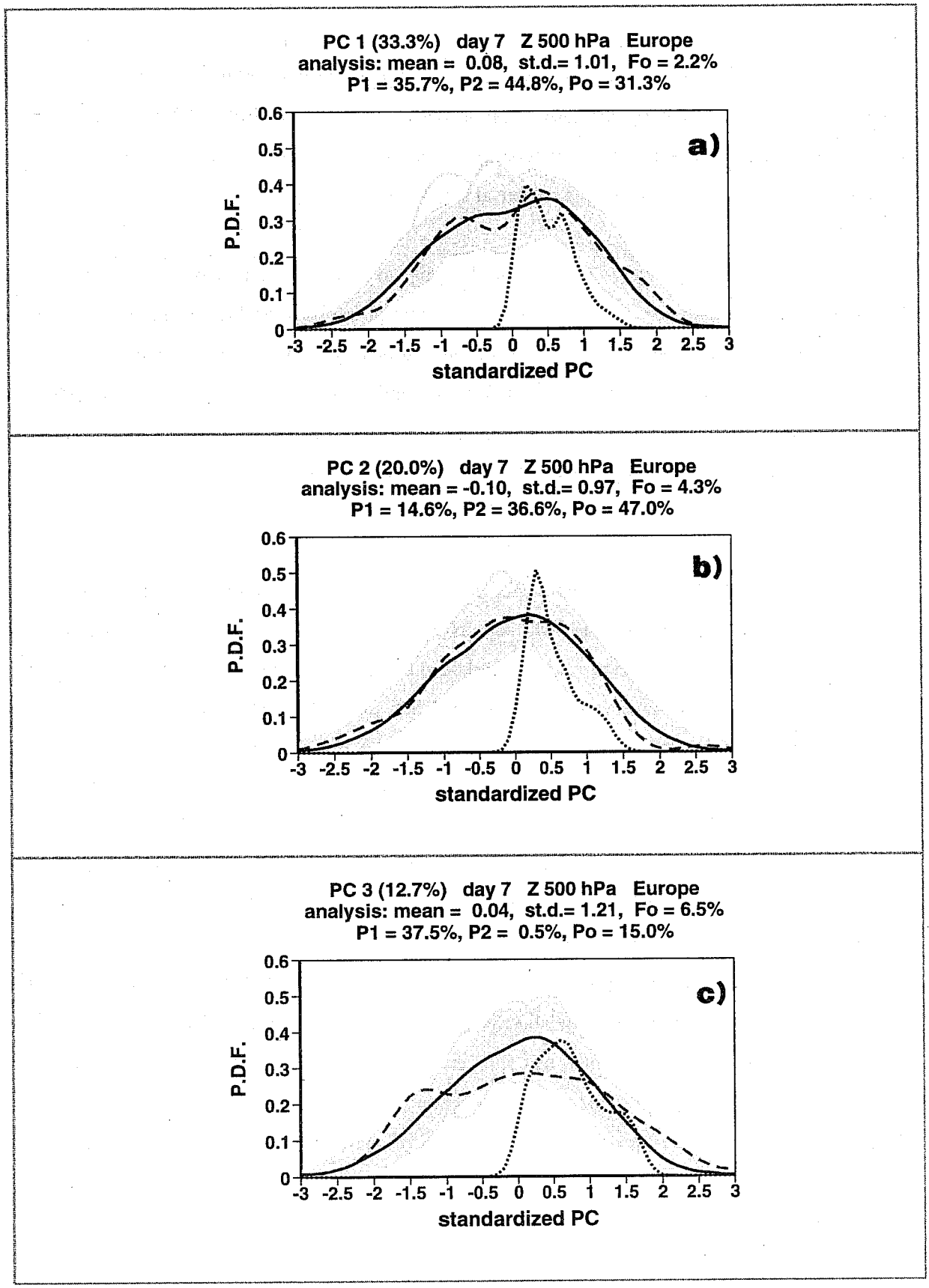

Fig. 9 As in Fig. 8, but for fc. day 7.

phase space which may lead into two distinct flow regimes (e.g. Palmer 1993; Trevisan et al. 1998). Indeed, when real atmospheric data or operational forecast errors were analysed, bimodality associated with the divergence of initially-close trajectories was detectable only in certain regions of phase space, where regimes are more clearly separated in phase space (Molteni and Tibaldi 1990; Trevisan 1995). On the other hand, if one looks for bimodality in individual ensembles, one faces the problem of dealing with too few data to assess the statistical significance of the result.

From a pragmatic point of view, it is still worth trying to assess the ability of the EPS to produce 'reliable' bimodal or multimodal distributions, since some of the EPS products (namely the ensemble clusters) are implicitly based on the hypothesis of multimodality. A way to address the problem is to define an index of bimodality for the PC distribution in individual ensembles, sort the n ensembles in one season according to the value of the bimodality index, and then recompute the PDFs of error and spread PCs from the two subsamples including the $n/2$ ensembles with smallest and largest bimodality index.

Given the set of PCs from the m members of one ensemble, the data were first sorted in ascending order, and all possible partitions into two subsets were considered; the bimodality index (BI) was defined as the maximum fraction of variance explained by the means of two such subsets. It is evident that such an index would have value one if the PC had just two possible values (say +1 and -1), therefore being perfectly bimodal. Besides, it agrees with the criterion used in many clustering algorithms to find an optimal partition in a multi-dimensional space (e.g. Michelangeli et al. 1995).

Fig. 10 shows the spread and error PDFs for the day-7 PC 1 (as in Fig. 9a), recomputed from the two subsamples with smallest and largest bimodality index. As far the spread PC is concerned, the index clearly separates ensembles with unimodal and bimodal distributions. However, the error-PC distribution is bimodal in both cases, with the two maxima being actually better separated in the cases when the ensemble has a small bimodality index. Although more encouraging results have been obtained for other areas and variables, sampling errors are often stronger than the signal one would like to detect. On the basis of this analysis, whether unimodality/bimodality in the ensemble is a reliable indicator of the bifurcation properties of actual atmospheric trajectories remains an open question.

6. Summary and conclusions

EOF analysis of deviations from the ensemble mean was used to validate the statistical properties of T_L159 51-member ensembles during winter 1996/97. The main purpose of the analysis was to verify

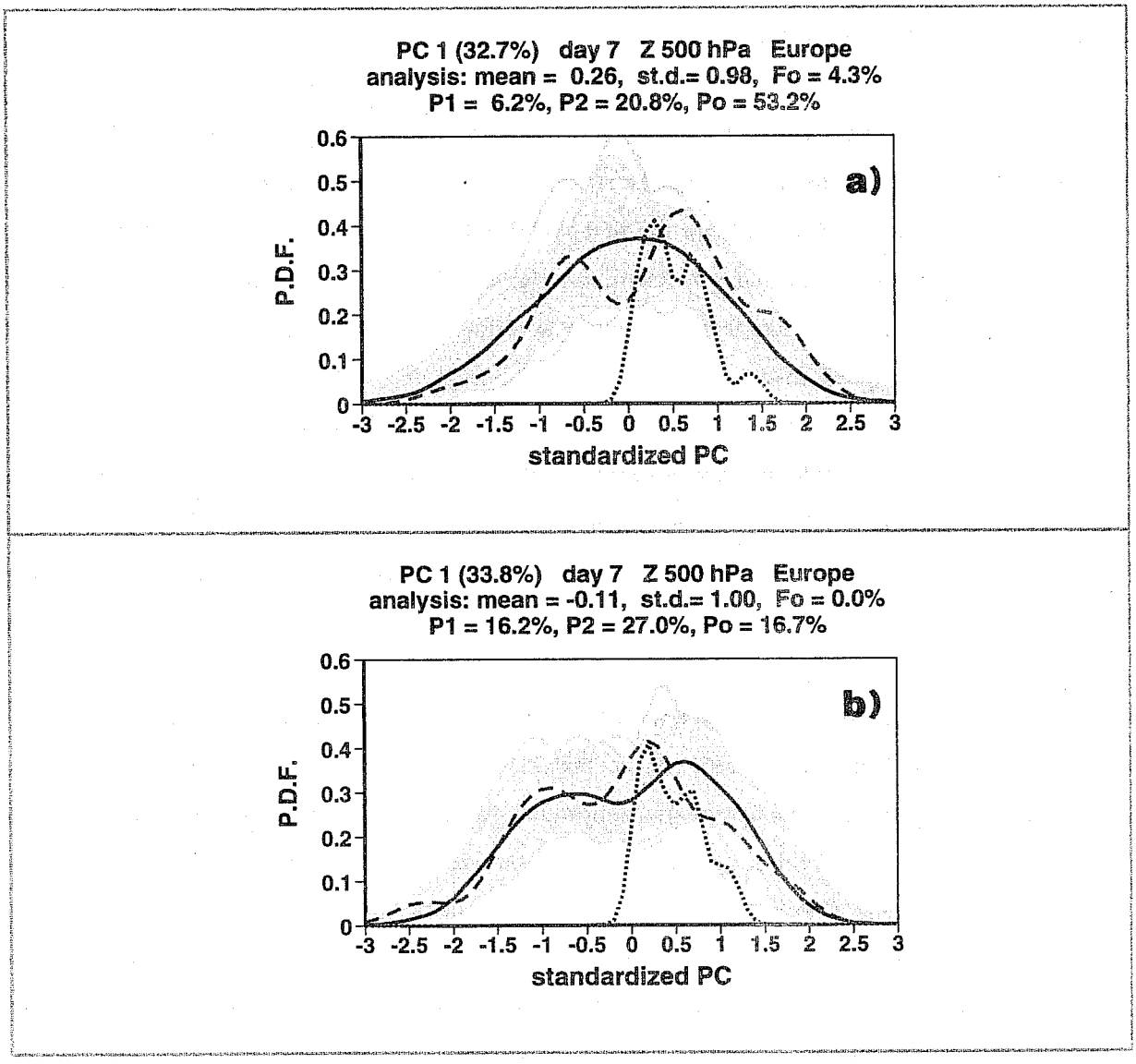

Fig. 10 As in Fig. 9a, but for two subsamples obtained by dividing the ensembles into two groups according to the bimodality index of spread PC 1 at fc. day 7. a) Ensembles with smallest bimodality indices. b) Ensembles with largest bimodality indices.

the agreement between the amount of spread variance and error variance accounted for by different EOFs. A suitable score, named "Error of Variance in EOF space" (EVE), was defined to quantify the agreement between the variance spectra in a given EOF subspace. The agreement between spread and error distribution for individual PC was also tested using the non-parametric Mann-Whitney test. The analysis was applied at day 3, 5 and 7 forecasts of 500-hPa height over Europe and North America, and of 850-hPa temperature over Europe.

The variance spectra indicate a better performance of the EPS over Europe than over North America in the medium range. In the former area, the excess of error variance over spread variance tends to be confined to non-leading PCs, while for the first two PCs the error variance is smaller than spread at day 3, in very close agreement at day 7. Medium-range values of the EVE score for a 6-EOF subspace are about 25% over Europe (zero implying perfect agreement), with the best value (15%) for 850-hPa temperature at day 7. Conversely, over North America the EVE score for 500-hPa height monotonically increases from 24% at day 3 to 55% at day 7. These results are confirmed by the Mann-Whitney test. Overall, the current version of the EPS produces a quite reliable estimate of the probability distribution of the atmospheric state over Europe, and shows substantial improvements with respect to the lower-resolution, smaller-size ensembles which were operational in the previous winters.

The fact that over Europe the day-3 spread exceeds the error along the leading EOFs, together with a small but consistent bias in the error PC at this forecast range, are a likely consequence of the constraint used to set the initial perturbation amplitude, namely that hemispheric rms spread and error should be equal at the optimization time of singular vectors. The slight 'overshooting' by the ensembles along their dominant EOFs tends to compensate the component of error variance attributable to model errors. In view of this, the fact that EPS perurbations now span a larger subspace, including singular vectors with smaller amplification factors, is certainly beneficial to the realism of the short-range forecasts.

Finally, the result that the variance spectrum of spread PCs is slighly steeper at day 3 than at day 7 is a clear sign that, for short-range forecast errors, a 'climatological' covariance matrix is a poor approximation of the covariance matrix appropriate for one particular initial state. In a statistical analysis of operational forecast errors currently carried out at ECMWF (L. Ferranti, private communication), an EOF decomposition of forecast errors over the Euro-Atlantic region in three winter seasons revealed that 22 EOFs were needed to explain 75% of the variance at day 3, 11 EOFs at day 7. In our analysis (performed over a smaller area), just 4 EOFs were needed (on average) at both

times. If this result is extrapolated to the first-guess errors used for data assimilation, one concludes that a flow-dependent error covariance matrix should have a much steeper spectrum of variance than a climatological covariance matrix. The potential positive impact of using time-dependent information on the background-error covariances is evident, and it is also clear that little benefit should be expected by using a climatological error covariance to define the initial norm of the singular vectors which define the EPS initial perturbations.

Appendix.

In this appendix, results of more traditional verification techniques are presented for the same variables, areas and periods chosen for the EOF analysis. The main purpose of this comparison is to highlight the limitations of over-simplified validation indices.

Given a grid-point field f defined on a set of grid points g in a geographical domain G, let $f_j(g, t, t')$ be the value predicted at grid-point g and forecast time t' by the j-th member of an ensemble started at initial time t, and $f_a(g, t, t')$ the corresponding verifying analysis. Besides, let $f_m(g, t, t')$ and $f_s(g, t, t')$ be (respectively) the ensemble-mean and the ensemble standard deviation of $f_j(g, t, t')$. If the initial time t spans a time interval T (usually a season), the space-time rms (root-mean-square) error of the ensemble mean is defined as:

A1a) $$E_{G,T}(t') = rms\ \{\, | f_a(g, t, t') - f_m(g, t, t') | \,\}_{g \in G,\ t \in T}$$

and the rms spread with respect to the ensemble mean by

A1b) $$S_{G,T}(t') = rms\ \{ f_s(g, t, t') \}_{g \in G,\ t \in T}.$$

In order to be statistically consistent, an ensemble forecast should be such that, at any time and grid point, the difference $(f_a - f_m)$ belongs to a distribution with zero mean and standard deviation equal to f_s. Since this condition can only be verified by averaging the ensemble data over a suitable space-time domain, two reliability indices can be defined as follows:

A2a) $$R_G(t') = E_{G,T}(t') / S_{G,T}(t')$$

A2b) $$R_g(t') = rms\ \{\, | f_a(g, t, t') - f_m(g, t, t') | / f_s(g, t, t') \,\}_{g \in G,\ t \in T}$$

where the T subscript has been dropped for simplicity. These two indices are conceptually equivalent,

but formally different: in R_G, the ensemble error and spread are first averaged in space (over the G domain) and time, and then their ratio is computed, while R_g is defined as the space-time average of error-to-spread ratios computed at individual grid-points g. It will be shown below how the different definition may affect the validation results and their interpretation.

In fig. 11, the R_G index is used to compare the performance of the EPS over Europe (a) and North America (b) in winters 1995/96 (dashed line) and 1996/97 (solid line). After a spin-up period of about 24 h, the R_G index reaches a near-constant value, with a relative minimum just after forecast day 2 (corresponding to the optimization time of singular vectors) and a weak relative maximum around day 6. In winter 95/96, the ensemble-mean rms error was significantly larger than the rms spread in both areas and throughout the forecast range. In winter 96/97, as shown by the EOF analysis, the discrepancy has been strongly reduced, especially over Europe, where the R_G index is only marginally greater than 1 after day 4. Over North America, the difference of R_G from unity was about twice as large in winter 95/96 than in 96/97, but it is still significant in the latter winter.

Figure 12 shows the same curves as Fig. 11, but for the R_g index. The improvement occurred in the latest winter, and the better performance of the EPS over Europe than over North America, are evident also from these graphs. However, according to the R_g index the ensemble consistency improves monotonically from day 1 onwards, and the differences from unity are larger than those of the R_G index over both areas and at all forecast times. In the early medium range, the R_g values for Europe in 96/97 are as large as the 95/96 values of R_G in the same area.

This example shows that simple area-averaged indices of ensemble consistency can certainly reflect changes in performance between different period and regions, but their absolute value and time evolution are sensitive to the particular way in which the average is performed. The definition of R_G can certainly mask some discrepancies between the spatial distributions of error and spread, such as those revealed by the EOF analysis over Europe at day 3 (when the R_G value is practically equal to one). On the other hand, the R_g index is strongly sensitive to errors occurring in areas with small spread: the improvement of consistency with time suggested by R_g is not supported by the EOF analysis in all cases (especially for North America), and is probably an artifact of the more uniform distribution of spread in the late medium range.

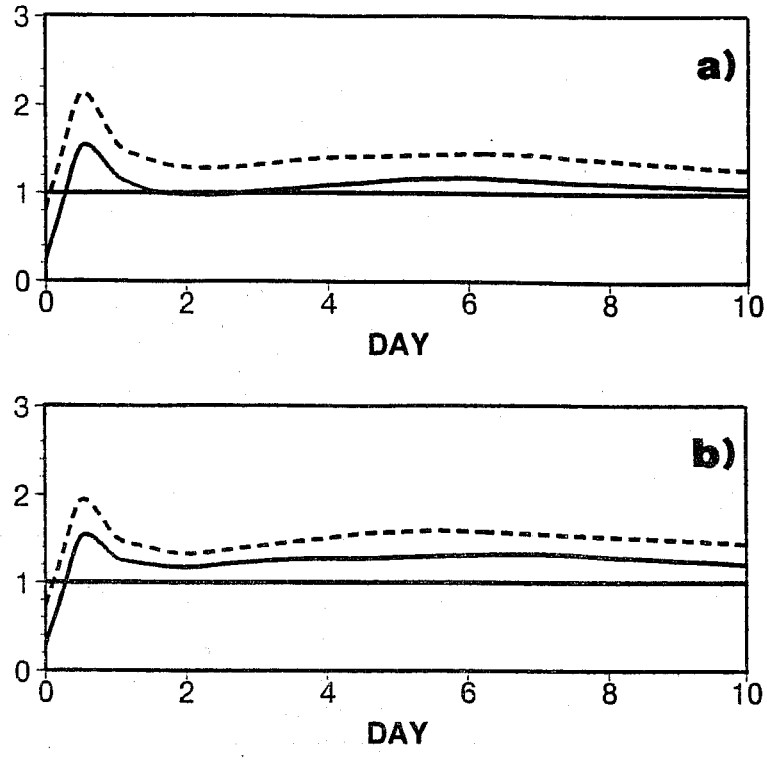

Fig. 11 Ratio R_G between seasonal-averages of the ensemble-mean rms error and of the rms spread around the ensemble mean for (a) Europe and (b) North America, for winter 1996/97 (solid) and winter 1995/96 (dash).

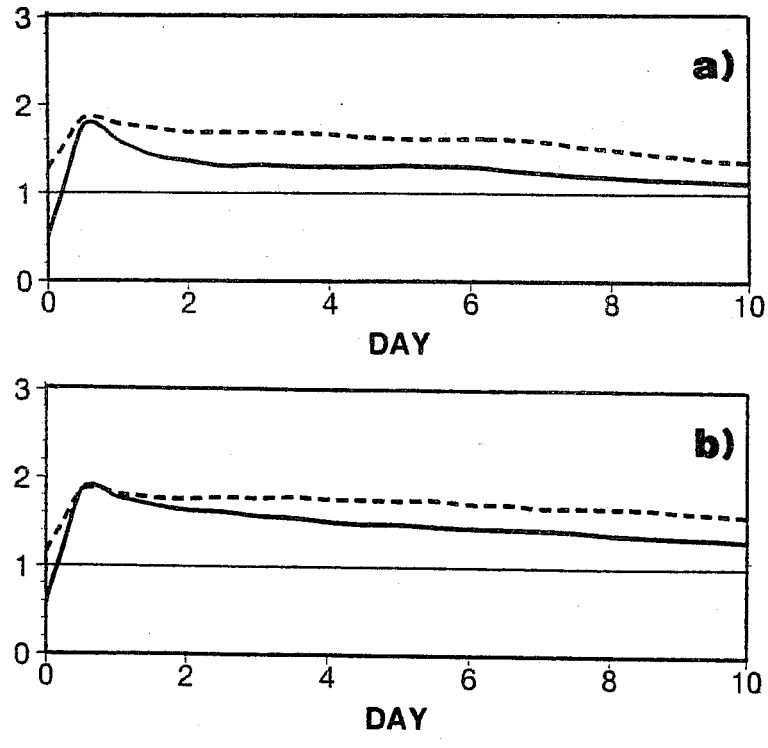

Fig. 12 As in Fig. 11, but the spatial and seasonal average of R_g of the error-to-spread ratios at individual grid-points.

References.

Brankovic, C., T.N. Palmer, F. Molteni, S. Tibaldi and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Q.J.R.Meteorol.Soc.*, **116**, 867-912.

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

Epstein, E.S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739-759.

Ferranti, L., F. Molteni, C. Brankovic and T.N. Palmer, 1994: Diagnosis of extra-tropical variability in seasonal integrations of the ECMWF model. *J. Climate*, **7**, 849-868.

Michelangeli, P.A., R. Vautard and B. Legras, 1995: Weather regimes: recurrence and quasi-stationarity. *J.Atmos.Sci.*, **52**, 1237-1256.

Molteni, F. and S. Tibaldi, 1990: Regimes in the wintertime circulation over northern extratropics. II: Consequences for dynamical predictability. *Q.J.R.Meteorol.Soc.*, **116**, 1263-1288.

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

Palmer, T.N., 1993: Extended-range atmospheric predictions and the Lorenz model. *Bull.Amer. Meteorol.Soc.*, **74**, 49-65.

Silverman, B.W., 1986: *Density estimation for statistics and data analysis*. Chapman and Hall, New York, 175 pp.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull.Amer.Meteorol.Soc.*, **74**, 2317-2330.

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

Trevisan, A., 1995: Statistical properties of predictability from atmospheric analogs and the existence of multiple flow regimes. *J.Atmos.Sci.*, **52**, 3577-3592.

Trevisan, A., F. Pancotti and F. Molteni, 1998: Ensemble prediction in a model with flow regimes. *Q.J.R.Meteorol.Soc.*, submitted.

Wang, X.L., and H.L. Rui, 1996: A methodology for assessing ensemble experiments. *J. Geophys. Res.*, **101D**, 29591-29597.