Tellus (2000), 52A, 300–322
Printed in UK. All rights reserved

TELLUS
ISSN 0280–6495

Statistical methods for interpreting Monte Carlo ensemble forecasts

By DAVID B. STEPHENSON[1]* and FRANCISCO J. DOBLAS-REYES[2], [1]*Météo-France, 42 Avenue Gaspard Coriolis, 31057 Toulouse, France;* [2]*Centro de Astrobiologia, INTA, Ctra. de Ajalvir km 4, 28850 Torrejon de Ardoz, Madrid, Spain*

(Manuscript received 10 August 1998; in final form 6 December 1999)

ABSTRACT

For complex dynamical systems such as the atmosphere, improved estimates of future behaviour can be obtained by making ensembles of forecasts starting from a set of Monte Carlo perturbed initial conditions. Ensemble forecasting, however, generates an overwhelming amount of data that is difficult to analyse in detail. Fortunately, the main features of interest are often summarised by certain statistics estimated from the sample of forecasts. By considering an ensemble of forecasts as a realisation of a linear mapping from phase space to sample space, it is possible to construct two types of sample covariance matrix. The ensemble covariance can be visualised by constructing multidimensional scaling maps, which show at a glance the relative distances between the different ensemble members. Multivariate skewness and kurtosis can also be estimated from ensembles of forecasts and provide useful information on the reliability of the sample mean and covariance estimated from the ensemble. They can also give useful information on the non-linearity of the evolution in phase space. Entropy can also be defined for an ensemble of forecasts and shows a regular increase due to the smooth and rapid loss of initial information in the first 3 days of a meteorological forecast. These new tools for summarising ensemble forecasts are illustrated using a single ensemble of 51 weather forecasts made at the European Centre for Medium-Range Weather Forecasts for the period 20–30 December 1997.

1. Introduction

In a prize-winning analysis of the stability of the solar system, Poincaré (1890) demonstrated that the evolution can sometimes be extremely sensitive to the initial conditions. Hadamard (1898) confirmed that rapid exponential divergence of trajectories is a generic feature of many quite simple dynamical systems such as geodesic flows on negative curvature surfaces (Ruelle, 1989). Examples of sensitive dynamical systems include the motion of three or more celestial bodies, balls on a billiard table, irregularly dripping taps, and geophysical fluid systems such as the atmosphere and the oceans. To quote Jules-Henri Poincaré:

> "Many persons find it quite natural to pray for rain or shine when they would think it ridiculous to pray for an eclipse ... [O]ne tenth of a degree at any point, and the cyclone bursts here and not there, and spreads its ravages over countries it would have spared. This we could have foreseen if we had known that tenth of a degree, but ... all seems due to the agency of chance."

Since the initial conditions can never be measured with infinite precision, rapid non-linear

* Corresponding author.
Address: D.B. Stephenson, Department of Meteorology, University of Reading, Earley gate PO Box 243, Reading RG6 6BB, UK.
e-mail: D.B.Stephenson@reading.ac.uk

growth of initial errors fundamentally limits our ability to forecast such systems (Thompson, 1957). Small perturbations of the initial conditions grow exponentially fast, leading to a rapid loss of initial information and predictability. Lorenz (1963) confirmed this sensitivity in numerical simulations of a three equation simplification of atmospheric convection, and discussed its fundamental implications for atmospheric predictability.

Ever since the pioneering forecasts of Charney et al. (1950), weather forecasting centres have made daily numerical weather forecasts on an operational basis. Global numerical models currently resolve horizontal distances down to about 50 km, and iterate the state of the atmosphere about every 10 min. The models represent the state of the atmosphere by storing pressure, temperature, wind velocity, and humidity variables at many grid points covering the whole globe, amounting to a total of typically 10^7 variables (not all independent). Forecasting the atmospheric state using such models is the non-trivial problem of predicting the motion of a point in a 10^7 dimensional phase space. Furthermore, because of the sensitivity to initial conditions, and our imprecise knowledge as to what are the exact initial conditions, no single forecast of the atmosphere can be considered completely trustworthy (Sutton, 1951; Thompson, 1957; Epstein, 1969a). To estimate the maximum time limit for useful weather forecasts, twin forecasts were made starting from two slightly different initial states. Different climate models gave quite different estimates for the time taken for the root mean squared difference between the two realisations to double (Smagorinsky, 1969). The shortest time was found to be close to 2 days for synoptic scales in the simplified turbulence closure model of Lorenz (1969a,b).

Epstein (1969b) and Leith (1974) discussed the benefits of this Monte Carlo approach to forecasting that is based on making averages of an ensemble of forecasts that start from randomly perturbed initial conditions. The mean of an ensemble of forecasts gives a Best Linear Unbiased Estimator (BLUE) of the future state, and the spead of the ensemble forecasts can be used to assess the possible reliability (Murphy, 1988). In general, a wide spread of the ensemble forecasts implies less reliability in the ensemble mean forecast. Ensemble forecasting techniques are now routinely used at national weather forecasting centres (Molteni et al., 1996; Buizza et al., 1998; Toth and Kalnay, 1993; Toth and Kalnay, 1997; Houtekamer et al., 1996) and are also starting to be applied in longer-range climate studies (Brankovic and Palmer, 1997; Zwiers, 1996; and references therein). Because of the computational cost of physically based forecasting models, ensembles are currently limited to small sample sizes of about 10–100 members.

An important practical problem with ensemble forecasting is that it increases the amount of forecast data by a factor equal to the ensemble size, thereby transforming the already difficult verification exercise into an almost Herculean task. An information bottle-neck occurs when the unfortunate human analyst is faced with such large amounts of data, since it is practically impossible (and not very inspiring) for a person to routinely examine 10–100 weather maps and subjectively compare their individual merits. There is, therefore, a great necessity to develop optimal methods for projecting out the most relevant information from all the data. Various techniques have recently been developed, for example, "spaghetti plots" in which a specified contour level is drawn for each ensemble member all on the same figure (Toth et al., 1997; Sivillo et al., 1997), and "tubing" which aims to identify directional clusters (Atger, 1999). The large number of variables and ensemble members suggest that statistical methods are appropriate and are the focus of this study.

Section 2 of this article presents the weather forecasting example used to illustrate the new ideas, while Section 3 introduces some basic mathematical concepts. Sections 4–6 describe the means and covariances that can be calculated from an ensemble, and discuss the unifying concept of the singular value decomposition of an ensemble. Section 7 uses the ensemble covariance to make simple "multidimensional scaling maps" of the ensemble forecasts that summarise at a glance the mutual distances between different ensemble members. Section 8 goes beyond covariances to discuss the probability distribution in phase space and its deviations from normality. Section 9 introduces the important concept of loss of information and shows how it can be used to define the entropy of an ensemble of forecasts. Section 10 concludes with a brief summary.

2. An example: 51 weather forecasts 20–30 December 1997

The methods in this article will be demonstrated by applying them to a single ensemble of 51 weather forecasts made at the European Centre for Medium-Range Weather Forecasts (ECMWF) for the period 20–30 December 1997. The study concentrates on the height of the 500 hPa geopotential surface, which gives much information about tropospheric dynamics in mid-latitudes. The spatial domain is restricted to the European–Atlantic region (90°W–30°E, 25°N–90°N) which covers 1225 grid point variables (25 latitudes × 49 longitudes). Each forecast is made to 10 days into the future with the results stored every 12 h. There are 51 ensemble members consisting of 1 control forecast using the best estimate of the initial conditions, and 50 forecasts starting from perturbed initial conditions about the control. More information about the numerical weather forecasting model and how the forecasts were initialised is given in Section 12.

3. Basic concepts

This section introduces some basic concepts in ensemble forecasting. They make use of simple linear algebra methods that are well described in the introductory book by Strang (1988).

3.1. Finite representation of a field

Geophysical spatial fields such as the height of the 500 hPa geopotential surface are often represented by a finite vector

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}, \tag{1}$$

defined by values of the field at p points on a regular spatial grid. The vector $\boldsymbol{x}$ is a finite p-dimensional representation of a continuum field and is a good approximation for spatially smooth fields. For spatially discontinuous fields such as rainfall amounts, the finite grid point representation can sometimes be difficult to interpret (Skelly and Henderson-Sellers, 1996). An often overlooked source of forecast uncertainty can be caused by unrepresented sub-grid scale features such as convective systems that feed upscale into the represented scales (personal communication, G. Craig). Alternative finite bases for representing functions, such as spherical harmonic coefficients, can also be described using the same multivariate notation. Throughout this article, vector and matrix symbols representing functional fields will appear in boldface. The p variables describe the "phase" (state) of the system and represent a point in p-dimensional "phase space".

3.2. Evolution in phase space

The evolution of atmospheric or oceanic fields is described by the motion in phase space

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = F[\boldsymbol{x}; \alpha(\boldsymbol{t})], \tag{2}$$

where F is a non-linear functional of $\boldsymbol{x}$, and $\alpha(\boldsymbol{t})$ represents external controlling parameters such as solar radiation and sea surface temperatures. For typical meteorological gridded fields, the phase space dimension, p, is of the order of 10^3–10^7, yet because of strong spatial correlations between neighbouring grid point variables, the total number of independent degrees of freedom is estimated to be around 20–40 for Northern Hemisphere daily geopotential height (Toth 1995; Wallace et al., 1991; Selten, 1997). The Eulerian approach of describing fields by their values at many fixed grid points is responsible for this dramatic overestimation in the number of dimensions, and could perhaps be alleviated by using a more Lagrangian description of the dominant spatially coherent structures.

3.3. Monte Carlo sampling

Forecasting applications do not typically require complete knowledge of the future evolution of the whole probability distribution, but instead demand (a) a good estimate of the mean future value, and (b) an estimate of how sensitive the forecasts are to the initial conditions. In other words, for normally (Gaussian) distributed fields, it is useful to have an estimate of the evolution of both the mean and the covariance of the distribution. For non-normal quantities such as precipitation these 2 low-order moments may not suffice

(Hamill and Colucci, 1998). To calculate the evolution of the mean and variance, it is necessary to evaluate integrals of first and second low order moments over the whole of p-dimensional phase space. One way to obtain approximations to these integrals is to use a finite sampling Monte Carlo method, in which the integral over phase space is approximated by a sum over a finite number of initially randomly chosen ensemble members (Epstein, 1969b; Leith, 1974). This computationally feasible method is the most widely used approach in meteorological forecasting. Finite Monte Carlo sample estimates converge slowly and can seriously undersample the integrals when the phase space has very many dimensions. Therefore, much care needs to be exercised when choosing initial perturbations and when interpreting ensemble sample estimates.

An ensemble (sample) of n forecasts of a spatial map is naturally represented by the $(n \times p)$ rectangular "data matrix"

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \tag{3}$$

where the kth row, $\boldsymbol{x}_k^{\mathrm{T}}$, is the kth forecast in the ensemble. Typically in meteorological applications, the number of members in the ensemble (the sample size n), is of the order of 10–100, whereas the rank of the vectors, p, is of the order of 10^3–10^7. In general, there are many more columns than rows in the data matrix and the rank of the data matrix never exceeds the smallest dimension (the ensemble size n). Since each row of the data matrix represents a point in phase space, the whole data matrix represents a cloud of n points in phase space, referred to as a "Gibbs ensemble" in statistical physics (Landau and Lifshitz, 1980).

3.4. *Forecast errors and distances*

To address the important question of how "close" a forecast is to the observed truth, it is necessary to define "distance" in phase space. There are many possible ways of defining distances (norms) and it is not at all evident which norms are the most appropriate for assessing the predictability of sheared fluid flows such as the atmosphere and ocean. As in previous forecasting studies, we will consider the generalised quadratic norm

$$\| \boldsymbol{x} \|_G^2 = \frac{1}{p} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{G} \boldsymbol{x} = \frac{1}{p} \sum_{i,j=1}^{p} x_i G_{ij} x_j, \tag{4}$$

where $\| \boldsymbol{x} \|_G^2$ is the squared length of the vector $\boldsymbol{x}$ and $\boldsymbol{G}$ is a general multivariate weighting metric. To avoid having negative distances (e.g., forecast errors), the metric must be positive semi-definite* Because positive semi-definite metrics can be Cholesky factored into $\boldsymbol{G} = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}$, the norm $\| \boldsymbol{x} \|_G^2$ for variables $\boldsymbol{x}$ is equivalent to considering the transformed variables $\boldsymbol{x}' = \boldsymbol{A}\boldsymbol{x}$ with the dot product norm $\| \boldsymbol{x}' \|_I^2$. In other words, there is some freedom of choice in selecting the field variables, and this freedom can be used to transform the metric into the identity matrix. For example, the inverse covariance (Mahalanobis) metric, $\boldsymbol{G} = \boldsymbol{C}^{-1}$, is equivalent to performing a dot product of the principal components (uncorrelated linear combinations of grid point variables) as explained in Stephenson (1997). Another example is provided by area-weighting of the grid point variables to account for the convergence of the meridians at the poles. To account for the decrease in the area of grid boxes near the poles, it is necessary to use the metric $G_{ij} = (\pi/2)\cos(\theta_i)\delta_{ij}$, where θ_i is the latitude of grid point i. This metric can be transformed away by considering instead the variables $x_i' = \sqrt{(\pi/2)\cos(\theta_i)}\, x_i$, and such an approach has been adopted in this study. Non-linear transformations of phase space variables can also be advantageous. For example, by considering the square root of rainfall amounts it is possible to improve the skill of ensemble forecasts of Indian monsoon rainfall (Stephenson et al., 1999). Such transformations used to reduce the impact of extreme events, induce a positively curved metric, in which anomalies of larger values correspond to smaller distances. Throughout this article, it will be assumed that the appropriate transformations have been applied to the phase space variables, so that with no loss of generality, the dot product $\| x \|_I^2$ can be assumed.

* Not all metrics need to be positive semi-definite, e.g., the Minkowski metric gives both positive and negative distances in space-time (Weinberg, 1972).

4. Central quantities

How can we efficiently summarise the ensemble of points in phase space? One possible way is to examine the low order moments of the cloud. In this section, we will briefly discuss the mean of the distribution (centre of mass), before examining higher order moments later in the article.

4.1. Means over the ensemble

The cloud of points in phase space can be considered to be a deformable (and tearable) physical object. As in classical dynamics, it can be helpful to decompose the dynamics into the sum of two parts: (a) motion of the centre of mass, and (b) motion about the centre of mass. The center of mass of the cloud of points in phase space is given by the "ensemble mean"

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k \tag{5}$$

obtained by summing over the rows of the data matrix. Linear transformations in phase space map the centre of mass to the centre of mass of the transformed ensemble. The ensemble mean and linear transformation operations commute, which suggests a simple test for the presence of non-linear evolution (Smith and Gilmour, 1997). The displacements about the centre of mass, $\boldsymbol{y}_k = \boldsymbol{x}_k - \bar{\boldsymbol{x}}$, constitute a centred data matrix, $\boldsymbol{Y}^{\mathrm{T}} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}$, where $\boldsymbol{H} = \boldsymbol{I} - (1/n)\mathbf{1}$ (where $\mathbf{1}$ has unity for all elements). The "ensemble mean error" is the difference, $\bar{\boldsymbol{x}} - \boldsymbol{x}_0$, between the ensemble mean and the best estimate of the observed state, $\boldsymbol{x}_0$, referred to as the "verification analysis". The individual forecast error of each ensemble member $\boldsymbol{x}_k - \boldsymbol{x}_0$ can be written as the sum of the ensemble mean error and the displacement: $(\bar{\boldsymbol{x}} - \boldsymbol{x}_0) + \boldsymbol{y}_k$. Fig. 1 shows maps of the ensemble mean error as a function of forecast time, in which it can be seen that the ensemble mean error increases in time especially over the North Atlantic ocean.

4.2. Means of the field

The centre of mass of the ensemble was obtained by taking the mean of the rows of the data matrix. In a similar manner, one can take averages of the columns of the data matrix to yield a mean value for each ensemble member. For data on equal area grids, such means correspond to area averages over the domain and can be used to provide simple indices. For example, the widely used Niño-3 index is obtained by averaging sea surface temperatures over the region (150–90°W, 5°S –5°N). This area-average approach to reducing the large amount of ensemble information depends on the subjective choice of area domain. Area-averaging is one of the simplest ways of making univariate indices out of multivariate climate data, yet it takes no account of the dominant spatial patterns present in the data (unless the areas are carefully chosen). A more optimal way of defining indices is to make weighted averages using the eigenvectors of the covariance matrix as loading weights — in other words, to use principal components as indices.

Fig. 2 shows the evolution of the area-averages calculated over the Euro-Atlantic region (90W–30E, 25–90N). It can be seen that the individual ensemble means spread out in time, and that the ensemble mean forecast diverges strongly from the verification after day 6. Very similar results are obtained if the ensemble median is used instead of the ensemble mean (not shown).

4.3. Mean, median, or mode?

The center of a distribution can be estimated using either the "mean", the "median", or the "mode". Although these measures are identical for variables that are normally distributed, they can provide different information when the distribution is skewed. For positively skewed distributions, the mean is biased towards the large positive values, whereas the median is more "robust" to such extremes. The median also has the advantage over the mean and the mode in that there is equal chance of forecasts falling either above or below - a desirable feature if one is interested in tendencies. The mode is the "most probable" value to occur and provides a maximum likelihood estimate of the future. For distributions encountered in practise, the median generally lies between the mode and the mean. For an ensemble of forecasts that deviates strongly from "normality", the choice of central measure can lead to quite different interpretations. For example, there is currently a serious debate concerning the Bank of England's use of "mode" rather than "median" for summarising

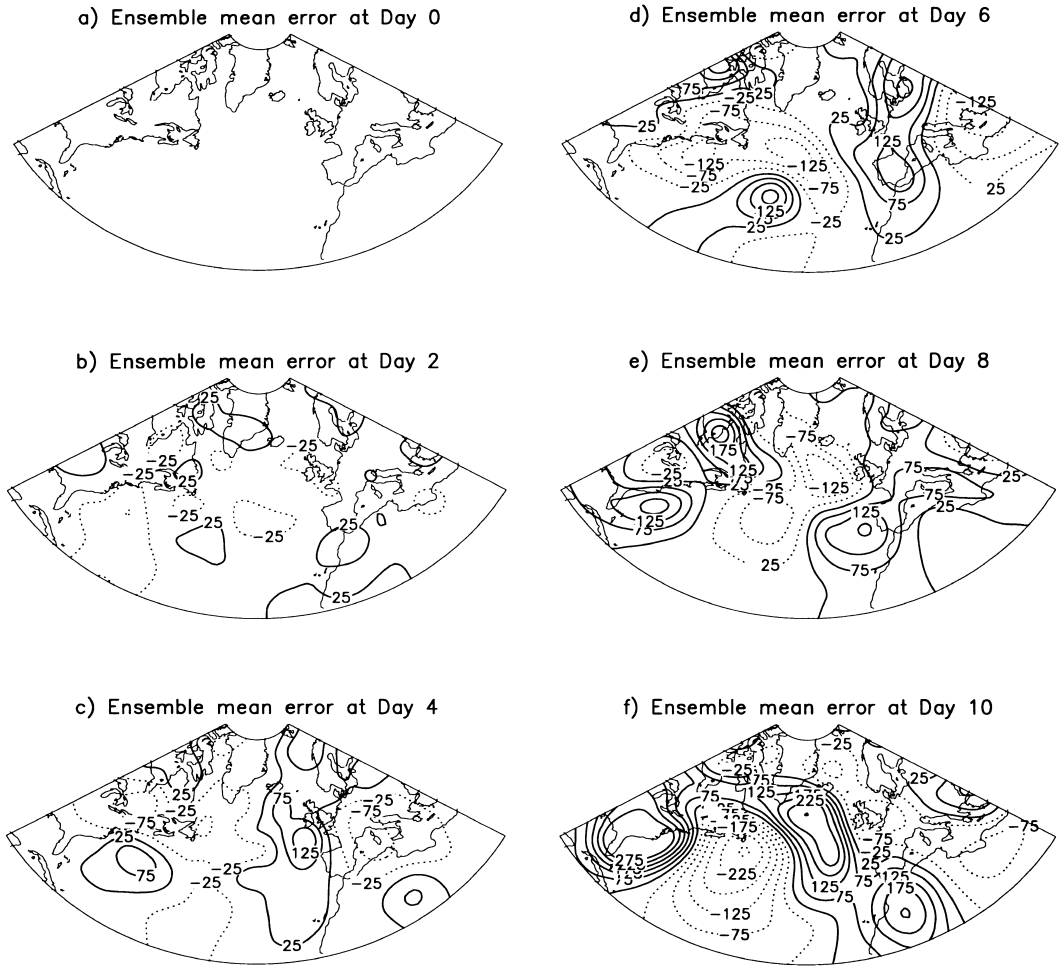

Fig. 1. Spatial maps of ensemble mean error $\bar{\boldsymbol{x}} - \boldsymbol{x}_0$ of the 500 hPa geopotential height as a function of forecast lead time. Contours are in meters.

ensemble forecasts of inflation (*The Economist*, 1999). This important topic will be revisited in Subsection 8.4.

5. Dispersion

This section uses covariances to quantify the dispersion of both the cloud in phase space, and the ensemble members. Covariances are the second order moments about the mean, and in classical mechanics correspond to the moment of inertia tensor about the centre of mass. They give invaluable information about the stretching and rotation of the cloud of points in phase space.

5.1. Ensemble spread

A natural measure of the typical spread of the ensemble in phase space is provided by the mean squared dispersion

$$D^2 = \frac{1}{n}\sum_{k=1}^{n} \| \boldsymbol{x}_k - \bar{\boldsymbol{x}} \|^2 = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{p}\sum_{i=1}^{p} y_{ki}^2. \qquad (6)$$

An unbiased yet less efficient estimate of the population spread can be obtained by dividing by

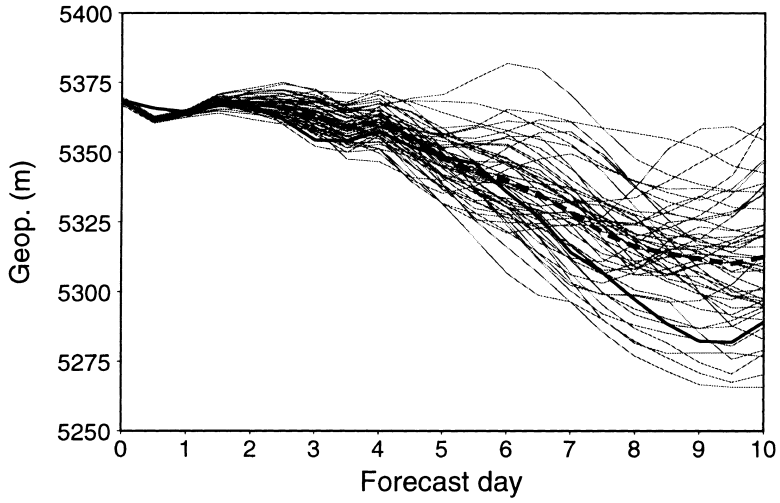

Fig. 2. Plume plot showing the evolution of the area average 500 hPa geopotential height over the Euro-Atlantic region (90°W–30°E, 25°N–90°N) for each of the 51 ensemble members (solid lines). The mean of all the forecasts (dashed line) strongly diverges from the verification analysis (solid black line) after about 6 days into the forecast.

$n-1$ instead of n. The mean squared dispersion, D^2, contributes with the squared error, $\|\bar{\boldsymbol{x}}-\boldsymbol{x}_0\|^2$, to the mean squared error of all the ensemble forecasts

$$S^2=\frac{1}{n}\sum_{k=1}^{n}\|\boldsymbol{x}_k-\boldsymbol{x}_0\|^2=D^2+\|\bar{\boldsymbol{x}}-\boldsymbol{x}_0\|^2. \tag{7}$$

Because of the dispersion of ensemble forecasts, the squared error of the ensemble mean forecast is less than the mean squared error of all the individual forecasts. The smaller squared error is one of the reasons that motivates the use of ensemble mean forecasts (Epstein, 1969b; Leith, 1974; Brankovic et al., 1990).

Fig. 3a shows the evolution of both the root mean squared dispersion D, and the root mean squared error S of all the forecasts calculated over the Euro-Atlantic domain. The dispersion, and to a lesser extent the root mean squared error, evolve almost linearly in time over the 10 days with no suggestion of the rapid exponential growth expected from chaos and instability arguments. Such quasi-linear growth is typical of medium range finite amplitude error growth (Stroe and Royer, 1993), and occurs when initial perturbations are no longer infintessimal (c.f. Section 12). During this regime, the atmospheric flow is determined by the far-field dynamics of synoptic-scale coherent structures (personal communication, J. Hunt). Fig. 3b shows the evolution of the dimensionless ratio D/S, which diagnoses the relative contribution of the dispersion to the total root mean squared error. Initially the ratio is unity, since by definition the ensemble mean error is zero at day 0. It then rises rapidly in the first day before finally reducing to the theoretical value of $1/\sqrt{2}$ expected for uncorrelated forecasts (Leith, 1974). Several studies have investigated the possibility of predicting the ensemble mean forecast skill by using the ensemble spread (Buizza, 1997; Whitaker and Loughe, 1998; and references therein). However, an additional constraint is clearly necessary in addition to eq. (7) such as constant $D/S=1/\sqrt{2}$ which occurs after day 1 due to the individual forecasts becoming uncorrelated.

5.2. *Covariance of the field variables*

A sample estimate of the covariances between the p grid point variables can be obtained by averaging over the ensemble members to yield a $(p\times p)$ sample covariance matrix:

$$\boldsymbol{C}=\frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{X}, \tag{8}$$

where the identity $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{H}$ has been employed. The sample covariance matrix is the outer product of the data matrix with itself and gives useful

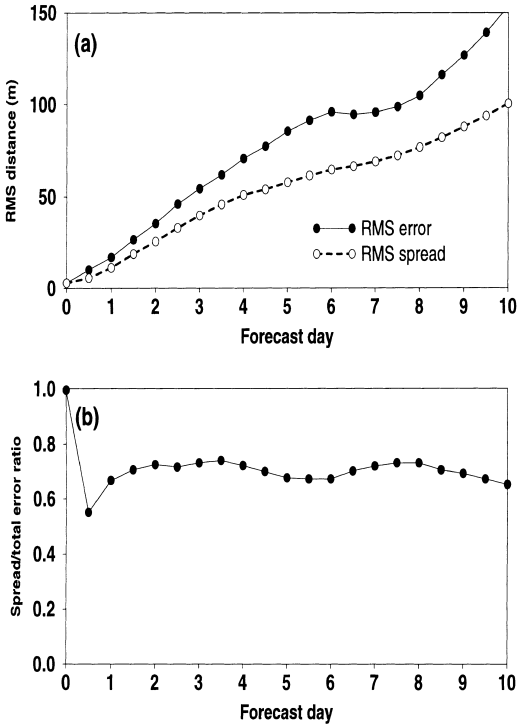

Fig. 3. (a) Evolution of the root mean square dispersion D (dashed line), and the root mean squared error S of all the ensemble forecasts of the ensemble 500 hPa geopotential height calculated over the Euro-Atlantic sector, and (b) the ratio D/S of the dispersion to the ensemble root mean squared error.

information on how the ensemble vectors span phase space. An unbiased estimate of the population covariance is given by $n\boldsymbol{C}/(n-1)$ which takes into account the loss of one degree of freedom (the ensemble mean). The covariances of the ensemble mean forecast errors are also equal to $\boldsymbol{C}$, since covariances are always defined relative to the mean and therefore are independent of the choice of origin. The field covariance matrix is obtained by summing over the ensemble members with equal weight given to each ensemble member. Unless the ensemble was generated in such a manner that each member was equally likely to occur in reality, for example by incorporating information about the prior evolution (Toth and Kalnay, 1997), then the sample covariance matrix will not be representative of the true population covariance matrix. For this to be so, it is necessary to account for prior information about the possible likelihood of each member, by including a diagonal weighting matrix, $\boldsymbol{W}$, as part of the more general expression $\boldsymbol{C} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Y}$.

A local estimate of the ensemble spread at each grid point can be obtained by examining the variances on the diagonal of the covariance matrix. Fig. 4 shows the evolution of the spatial distribution of ensemble spread obtained by taking the square root of the diagonal elements of the covariance matrix. The spread increases in time fairly linearly and is particularly large over the North Atlantic region, where there is much instability and storm activity.

The eigenvectors of the covariance matrix correspond to the major axes of the ensemble of points in phase space, and represent the spatial patterns that contribute the most to the total ensemble spread shown in Fig. 4. The eigenvectors are the columns of the $(p \times p)$ orthogonal matrix $\boldsymbol{V}$ which diagonalises the covariance matrix as follows

$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{V} = \frac{1}{n}(\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma}), \tag{9}$$

where the $(n \times p)$ matrix $\boldsymbol{\Sigma}$ is zero except on the diagonal which contains the square root of the eigenvalues (singular values) of $n\boldsymbol{C}$.

5.3. *Covariance of the ensemble members*

It is also possible to estimate the $(n \times n)$ covariance between the different ensemble members by averaging over the field variables to obtain the inner product of the data matrix

$$\boldsymbol{B} = \frac{1}{p}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} = \frac{1}{p}\boldsymbol{H}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}. \tag{10}$$

The ensemble covariance matrix gives valuable information about which members are most similar and can be diagonalised as follows:

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{U} = \frac{1}{p}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\mathrm{T}}), \tag{11}$$

where the $(n \times n)$ orthogonal matrix $\boldsymbol{U}$ has the eigenvectors of $\boldsymbol{B}$ as its columns. The leading eigenvector gives the n loading weights for the linear combination of ensemble forecasts that contributes the most to the domain-averaged squared spread shown in Fig. 4.

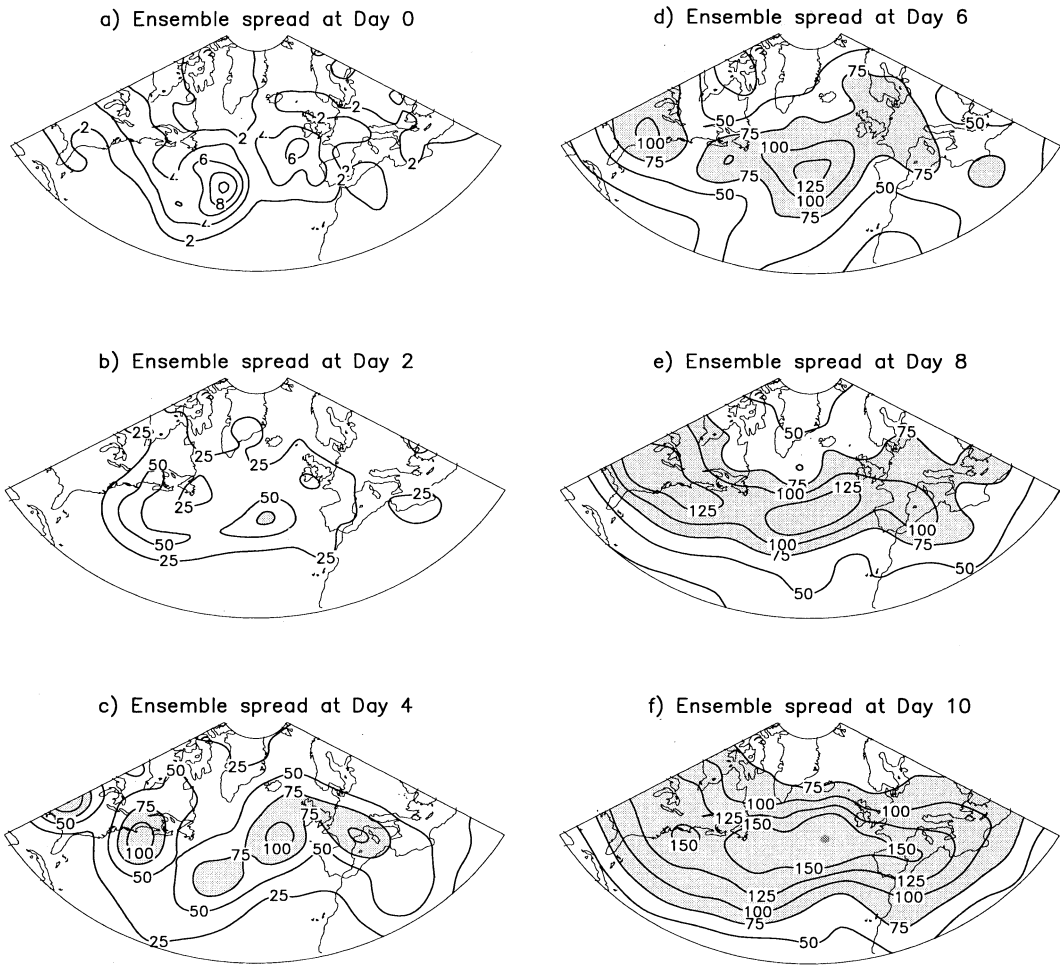

Fig. 4. Spatial maps of ensemble spread of 500 hPa geopotential height as a function of forecast lead time. Note the large amplitudes near the eastern end of the Atlantic storm track where strong mixing occurs. Contours every 2 m for day 0 and then every 25 m for later forecast days.

5.4. Singular value decomposition of an ensemble

The matrices $n\boldsymbol{C}$ and $p\boldsymbol{B}$ share the same leading eigenvalues. This can easily be demonstrated by considering inner and outer products of the singular value decomposition of the data matrix

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{12}$$

remembering that $\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrices. More generally, the data matrix can be considered to be a realisation of the linear operator that maps "phase space" onto "sample space". The SVD decomposition is a sum of tensor direct products:

$$Y_{ij} = \sum_{k=1}^{r} \sigma_k U_{ik} V_{jk}, \tag{13}$$

where the index k runs over the singular modes. The $(n \times p)$ matrix $\boldsymbol{\Sigma}$ is zero apart from on the diagonal, which contains rank r real positive singular values $\{\sigma_1, \sigma_2, ..., \sigma_r\}$. The rank of the data matrix is an important quantity that gives the effective dimensionality of the ensemble. For example, if all the ensemble members are in the same direction then the rank is unity and there is

the same linear information as using only one member in the ensemble. The rank depends on the singular spectrum of the data matrix and is bounded from above by $\min(n, p)$, typically equal to the ensemble size n. The mean squared dispersion of the ensemble can be expressed in terms of either the trace of the ensemble covariance matrix, $D^2 = \mathrm{Tr}\, \boldsymbol{B}/n$, or the trace of the field covariance matrix, $D^2 = \mathrm{Tr}\, \boldsymbol{C}/p$, and the SVD expansion then shows that $D^2 = (\sum \sigma_k^2)/np$.

6. Evolution of the covariance

For ensemble members that are sufficiently close together in phase space, the ensemble transforms linearly $\boldsymbol{x}(t_1) \rightarrow \boldsymbol{x}(t_2) = \boldsymbol{L}(t_2, t_1)\boldsymbol{x}(t_1)$ with the "tangent linear operator" given by

$$\boldsymbol{L}(t_2, t_1) = P \exp \int_{t_1}^{t_2} \frac{\partial F}{\partial \boldsymbol{x}} \mathrm{d}t, \tag{14}$$

where $P \exp(\,)$ is the path-ordered exponential operator. The tangent linear operator is linear and non-singular in phase space, yet varies non-linearly in time. Under such a linear non-singular transformation, the field covariance undergoes a "congruence transformation"

$$\boldsymbol{C} = \frac{1}{n} \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \rightarrow \boldsymbol{C}'$$

$$= \frac{1}{n} (\boldsymbol{YL})^{\mathrm{T}} (\boldsymbol{YL}) = \boldsymbol{L}^{\mathrm{T}} \boldsymbol{CL}, \tag{15}$$

as explained in Ehrendorfer and Tribbia (1997). The effect of this transformation can be understood in more detail by examining its impact on the eigenvalues and eigenvectors of the covariance matrix.

6.1. Evolution of the eigenvalues

What properties of $\boldsymbol{C}$ remain invariant under such transformations? By Sylvester's law of inertia, the number of positive, and zero eigenvalues remain constant under congruence transformations and therefore the rank of $\boldsymbol{C}$ remains constant under non-singular linear transformations. If one starts with a rank deficient ensemble, by choosing, say, one member to be a linear combination of the others, then the ensemble will always remain rank deficient under linear transformations. For example, any linear combination of two vectors in 3-dimensional space will always remain in the 2-dimensional plane spanned by the two vectors. There is no way that linear combinations can generate a vector lying outside the original plane, although non-linear transformations could generate a vector that escapes the plane. For linear transformations that are continuous differentiable functions of time, the eigenvalues of the covariance matrix also evolve continuously and differentiably (Strang, 1988).

Fig. 5 shows the 12 hourly evolution of the covariance eigenvalues (normalised by the total variance) from which it may be noted that the leading four eigenvalues explain more than 60% of the total covariance. Because the ensemble was initially formed as anomaly pairs, the 25 trailing eigenvalues at day 0 explain zero amounts of total variance — in other words, at day 0 the 51 member ensemble has a rank of only 25. In the 1st 2 days, the 25 trailing eigenvalues evolve (non-linearly!) into slightly larger values and the rank deficiency disappears (not shown). Over the 1st 3 days, a separation starts to develop between the two leading eigenvalues and the rest of the eigenvalues. After this initial period of structural metamorphosis, the leading two eigenvalues show some evidence of evolving continuously for several consecutive days, for example, the second largest eigenvalue over days 2.5–5.5 and 6–9.

6.2. Evolution of the eigenvectors

Figs. 6, 7 show respectively the evolution of the first two leading eigenvectors of the covariance matrix. In accordance with the 2-dimensional turbulence concept of upscale cascade of energy (Lorenz, 1969a; Leith 1971; Leith and Kraichnan, 1972), it can be seen that the eigenvector patterns increase their characteristic spatial scales in evolving from day 0 to day 2. In other words, small spatial scale perturbations become larger scale perturbations — described poetically by the metaphor of a butterfly flapping its wings and thereby causing a tornado. After day 2, the upscale cascade begins to saturate and the main eigenvector features become located in the eastern end of the Atlantic storm track, where strong non-linear mixing occurs. The eigenvectors are spatially fixed over periods of several days coinciding with the periods when their respective eigenvalues evolved

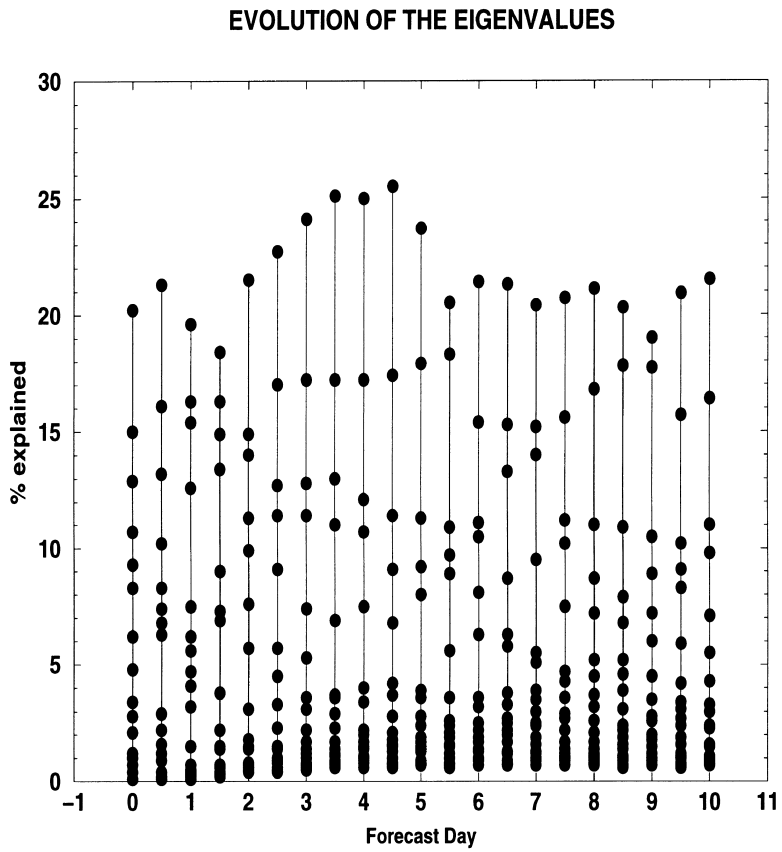

Fig. 5. Evolution of the percentage of variance explained by each eigenvalue of the covariance matrix ($100\%\,\lambda/\sum\lambda$). Note the periods of apparently continuous evolution in the leading eigenvalues.

continuously — for example, days 6–9 for the leading eigenvectors.

6.3. *Evolution of the ensemble covariance*

Under linear transformations, the ensemble covariance transforms as

$$\boldsymbol{B} = \frac{1}{p}\,\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}^{\mathrm{T}} \rightarrow \boldsymbol{B}' = \frac{1}{p}\,\boldsymbol{Y}(\boldsymbol{L}\boldsymbol{G}\boldsymbol{L}^{\mathrm{T}})\boldsymbol{Y}, \tag{16}$$

where the general metric has been reintroduced in order to demonstrate that the transformation induces a congruence transformation on the metric: $\boldsymbol{G} \rightarrow \boldsymbol{L}\boldsymbol{G}\boldsymbol{L}^{\mathrm{T}}$. This result is well-known in the theory of relativity where linear transformations map space-time coordinates from one frame to another (Weinberg, 1972). The adjoint of the linear tangent operator is defined as $\boldsymbol{L}^*$ where $(\boldsymbol{L}\,x)^{\mathrm{T}}\boldsymbol{G}\boldsymbol{y} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{L}^*\boldsymbol{y}$ for all vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Using this identity, $\boldsymbol{L}\boldsymbol{G}\boldsymbol{L}^{\mathrm{T}}$ can be rewritten as $(\boldsymbol{L}\boldsymbol{L}^*)\boldsymbol{G}$ where $\boldsymbol{L}\boldsymbol{L}^*$ is the Oseledec operator whose eigenvectors correspond to the fastest growing modes used as initial perturbations in this and other studies. If the metric is chosen to be Euclidean and the ensemble members are constructed by adding the fastest-growing singular vectors to the control initial condition, the evolution of the ensemble covariance matrix in the case of linear evolution is simply given by $\boldsymbol{B} = \boldsymbol{I}/p \rightarrow \boldsymbol{\Lambda}/p$ where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the respective growth rates of each of the modes.

7. Multidimensional scaling

Mean squared errors play a crucial role in the assessment of many forecasts, and have the merit

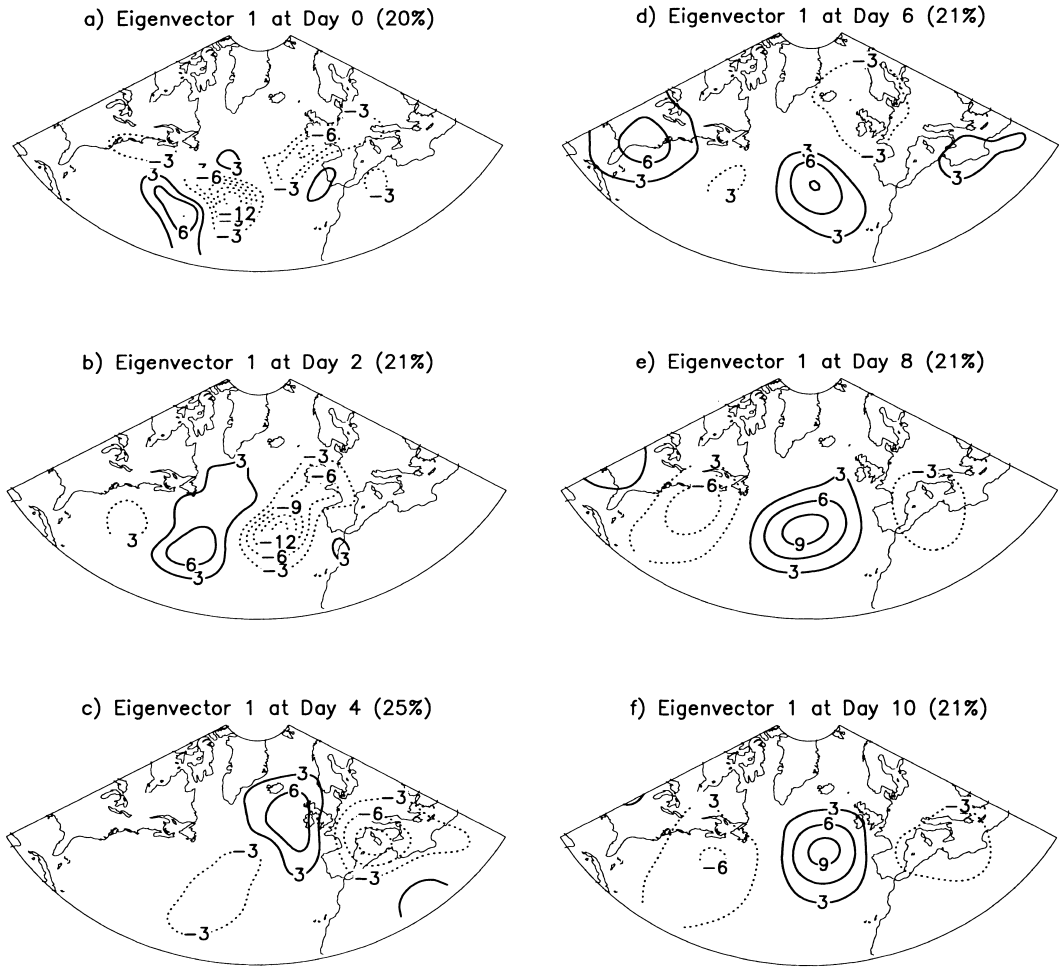

Fig. 6. Evolution of the first leading eigenvector of the field covariance, which explains 20–25% of the total variance in the Euro-Atlantic region. Note the increase in spatial scale in the first few days, and the emergence of a fixed wave pattern from day 6 onwards.

that they can be considered geometrically as squared distances in phase space. Geometrical interpretation of mutual distances between ensemble members can help give deeper insights into ensemble forecasts as will emerge from reading this section.

7.1. *Mutual distances between ensemble members*

The $n(n-1)/2$ mutual distances between the ensemble members can be conveniently stored in a $(n \times n)$ real symmetric distance matrix, $\boldsymbol{D}$, having zeros on the diagonal. The squared distance between the kth and lth ensemble members is given by

$$D_{kl}^2 = \| \boldsymbol{x}_k - \boldsymbol{x}_l \|^2 = (\boldsymbol{x}_k - \boldsymbol{x}_l)^{\mathrm{T}} (\boldsymbol{x}_k - \boldsymbol{x}_l)/p. \tag{17}$$

A common example of a distance matrix is that often printed in road atlases giving the road distances between different towns. Another example is presented in Table 1, which gives the distances between the first 6 members of the ECMWF forecast ensemble at day 1. The distance matrix can be expressed in terms of ensemble covariances using the cosine law transformation

$$D_{kl}^2 = B_{kk} + B_{ll} - 2B_{kl}, \tag{18}$$

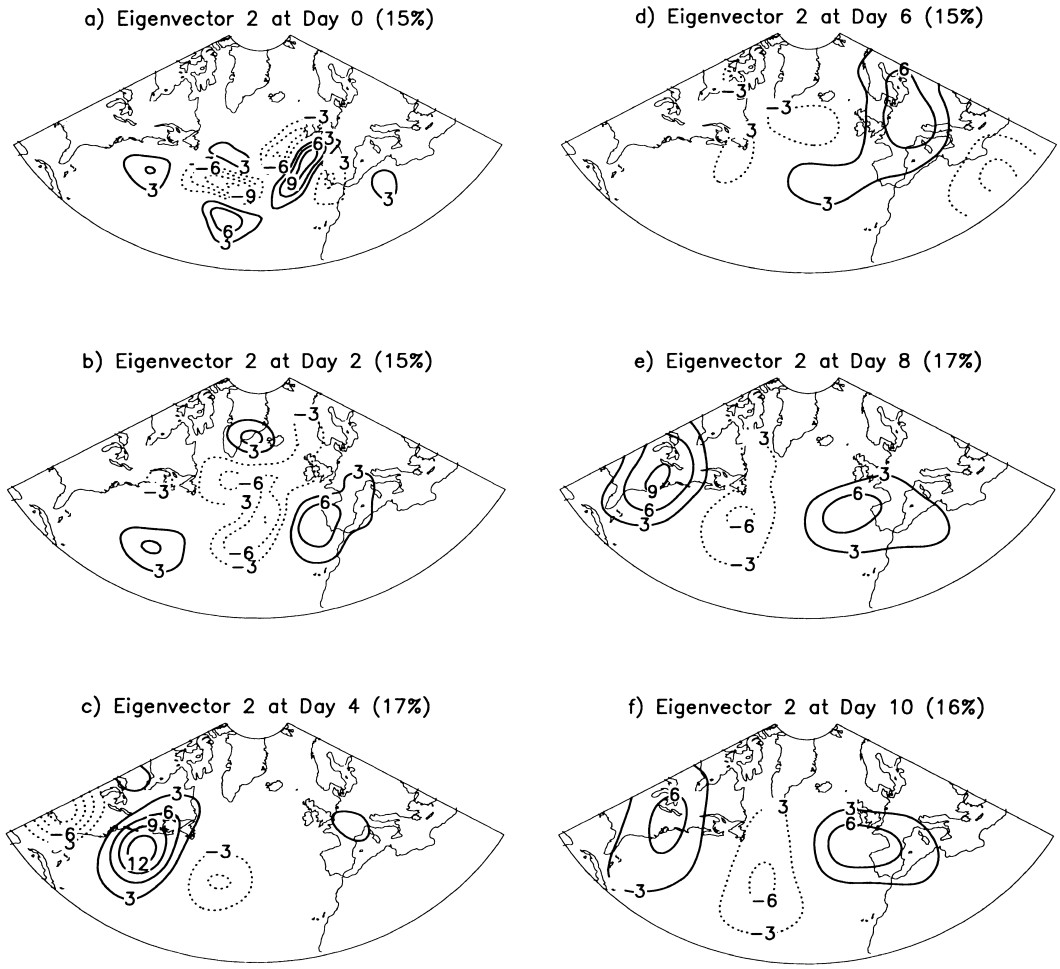

Fig. 7. Evolution of the second leading eigenvector of the field covariance, which explains 15–17% of the total variance in the Euro-Atlantic region. Note the increase in spatial scale in the first few days, and the quasi-stationary wave pattern from day 6 onwards, in quadrature to the first eigenvector depicted in Fig. 6.

Table 1. *Euro-Atlantic 500 hPa geopotential height distances between the first 6 ensemble members on day 1 of the forecasts (m)*

Member	1	2	3	4	5	6
1	0.00	20.38	16.98	12.90	12.85	17.23
2	—	0.00	12.80	16.66	17.26	12.99
3	—	—	0.00	21.74	18.04	13.07
4	—	—	—	0.00	13.41	17.83
5	—	—	—	—	0.00	22.35
6	—	—	—	—	—	0.00

and, hence, all the information about mutual Euclidean distances between members of the ensemble is contained in the $(n \times n)$ ensemble covariance matrix $\boldsymbol{B}$.

7.2. *Multidimensional scaling (MDS)*

When there are more than just a few ensemble members, a table of mutual distances becomes difficult to interpret, and it becomes more helpful to have a simple map marking the relative locations of the members. In general, to make such maps for all n-ensemble members, requires an

n-dimensional embedding space since the rank of the distance matrix can take values up to n. However, more than 2 to 3 dimensions are difficult to visualise using flat pieces of paper or computer screens, and for more than 2 dimensions a projection is required.

Optimal projection methods have been developed for this kind of problem and are known in statistics as "multidimensional scaling" (MDS). MDS is concerned with the problem of constructing optimal configurations of n points in q-dimensional Euclidean space using information about the $n(n-1)/2$ mutual distances between the points in a p-dimensional space (Mardia et al., 1979). For example, a trivial application of MDS is to generate a 2-dimensional map showing the locations of towns when given only a table of distances between the various towns. Multidimensional scaling is a useful exploratory tool for describing ensembles of forecasts. Since MDS is based on the mutual distances between objects, it is a natural method to use to describe the "closeness" of forecasts. MDS is widely used in psychology and the social sciences and a readable introduction can be found in Cox and Cox (1994).

7.3. *The classical MDS solution*

For Euclidean distances, the "classical" solution to the MDS problem is given by plotting the "principal coordinates" of the n points (Mardia et al., 1979). The q "principal coordinates" of ensemble member k are defined as

$$\boldsymbol{z}_k = \begin{pmatrix} \sigma_1 U_{k1} \\ \sigma_2 U_{k2} \\ \vdots \\ \sigma_3 U_{kq} \end{pmatrix}, \tag{19}$$

where U_{ki} denotes the kth element of the ith leading eigenvector of the ensemble covariance $\boldsymbol{B}$. The projection can be seen more clearly in terms of the $(n \times q)$ data matrix containing the ensemble vectors in the projected space

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}_q = \boldsymbol{Y}\boldsymbol{V}\boldsymbol{P}, \tag{20}$$

where $\boldsymbol{\Sigma}_q = \boldsymbol{\Sigma}\boldsymbol{P}$ is a $(n \times q)$ projection of the $(n \times p)$ singular value matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{P}$ is the $(p \times q)$ projection matrix having ls on the diagonal only. The ensemble mean projects onto the origin in the reduced space. A projection space with $q = 2$ is a natural choice for displays on two dimensional sheets of paper or computer screens, yet larger dimensions such as $q = 3$ can also sometimes be useful. For the classical solution based on Euclidean distances, the principal coordinates are proportional to the principal components of the covariance matrix. However, for non-Euclidean distances, principal coordinates and principal components are no longer simply related, and MDS solutions differ from those obtained by PCA (Principal Component Analysis). By using non-Euclidean metrics, it is possible to obtain more robust solutions than those obtained using Euclidean distances (Cox and Cox, 1994).

7.4. *An illustration: classical MDS for the ECMWF forecasts*

Fig. 8 shows classical MDS projections of the ensemble forecasts for each day. They provide a simple way of illustrating which ensemble members are close to one another, and are much easier to interpret than large tables of distances! The distances between the points in the projected space are least square approximations to the distances in the original space and are good approximations when the leading two eigenvalues of $\boldsymbol{B}$ explain a large amount of the total variance. In our example, the leading two eigenvalues explain more than a third of the total variance* (Fig. 5). Fig. 8 also includes the verification analysis (marked by a circle), which was included in the MDS by considering it to be the zero'th member of an augmented $n+1$ member ensemble. Because the ensemble size is large, the inclusion of the verification analysis has a negligible effect on the projection of the individual ensemble members (not shown).

As a result of been constructed initially to be equal and opposite (c.f. Section 12), successive pairs of ensemble members lie diagonally opposite one another at day 0 (e.g., pairs 17 and 18, 43 and 44, etc. in Fig. 8a). This symmetry under rotations of 180° is preserved under linear evolution, and can be used as a pair-wise test for the presence of non-linear dynamics (Smith and Gilmour, 1997). Remarkably, some of the symmetry in Fig. 8 still remains until day 4, for example, in the pairs of

* Larger fractions could be explained by focusing on smaller regions (e.g., western Europe).

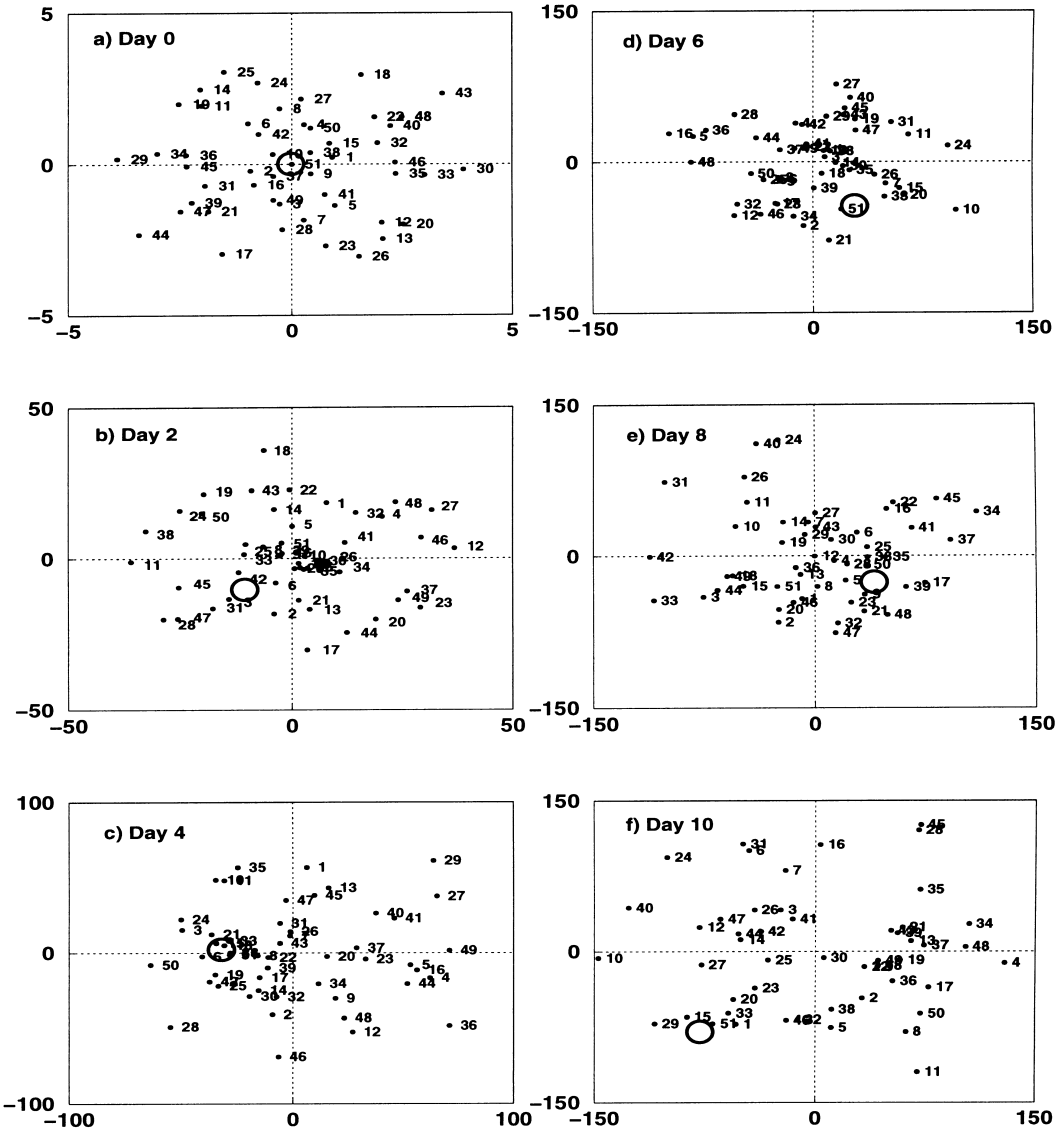

Fig. 8. Multidimensional scaling projections of the ensemble members at different forecast lead times. Ensemble members are numbered for identification purposes and the verification analysis is also marked on the plot as a circle. Note the symmetry under rotations of 180° in the initial day 0 ensemble due to the perturbations being defined as pairs of opposite sign anomalies. Even up until day 6 many of the adjacently numbered members lie diagonally opposite the origin from one another. Also note the marked clustering near the origin at day 2.

members 11 and 12, 27 and 28, 49 and 50. Another noteworthy feature is the cluster of remnant members near the origin at day 2 (Fig. 8b). Standard cluster algorithms generally have difficulty in isolating central clusters surrounded by a ring of other points — MDS offers a powerful alternative. The ring is most likely due to initial perturbations having the same initial energy but different spatial phases (the spatial scales of the two leading eigenvectors are similar). The appearance of a central cluster near the origin at day 2 is possibly due to the evolution of new leading modes having

different spatial scales orthogonal to the initial smaller scaled structures. For later forecasts at day 8, although most of the points in Fig. 8e are clustered around the verification analysis, there are some points lying further away in the upper lefthand quadrant. These points bias the ensemble mean away from the main cluster of points (the mode) close to the verification analysis. Multivariate methods in the following section confirm that the day 8 forecasts were some of the most skewed forecasts.

8. Distribution in phase space

The mean and the covariance provide a COMPLETE description of the distribution ONLY when the forecasts are normally distributed. However, non-linear evolution can lead to skewness (asymmetry) and other deviations from normality in the ensemble of forecasts. Furthermore, the presence of skewness and kurtosis (flatness) can cause ensemble sample estimates of the mean and covariance to become less robust and less reliable. This section will quantify deviations from normality in the ensemble of forecasts by introducing two new tools: "multivariate skewness" and "multivariate kurtosis".

8.1. The probability density function (p.d.f.)

Consider an ellipsoidal cloud of points in phase space that is multinormally distributed about the ensemble mean with a probability density

$$\Omega[\boldsymbol{x}] = (\det 2\pi \boldsymbol{C})^{-1/2} \times \exp[-\tfrac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})], \qquad (21)$$

where $\det 2\pi\boldsymbol{C}$ is the matrix determinant of $2\pi\boldsymbol{C}$. The probability of finding a point in a certain region of phase space is given by integrating the density over the specified region. The evolution of $\Omega(\boldsymbol{x})$ is described by the Liouville equation (Epstein, 1969b; Ehrendorfer, 1994):

$$\frac{\partial \Omega}{\partial t} + \nabla \cdot (\boldsymbol{u}\Omega) = 0, \qquad (22)$$

where $\boldsymbol{u} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$ is the rate of change in time of the phase space coordinates (the flow). After initial transients have damped out, the density starts to measure the probability of visiting different parts of phase space (Ruelle, 1989). Phase space dynamics will generally cause such a cloud to disperse, and when sufficiently non-linear, can cause parts of the cloud to be torn off from the main body. Direct numerical integration of the Liouville equation is prohibitively expensive in computer time for systems with many degrees of freedom such as the atmosphere. This is the reason that approximate finite sampling (ensemble forecasts) or parametric methods ("stochastic dynamic prediction") have to be employed. Stochastic dynamic prediction was proposed by Epstein (1969b) and is based on the second order closure assumption that the p.d.f. remains multinormal. Since ANY linear combination of multinormal distributed variables is also multinormal distributed (Mardia et al., 1979), linear transformations can NEVER cause the distribution to evolve away from multinormal. Such invariance under linear transformations provides the basis for a useful test for non-linearity (Burgers and Stephenson, 1999).

Although there are regions of the planet having strongly non-linear evolution and strong correlations with weather elsewhere, many regions are neither strongly non-linear nor strongly correlated with elsewhere, especially over short forecast intervals. Because of only weak interactions between the large number of degrees of freedom, the atmospheric and oceanic general circulations are often close to statistical equilibrium (normality). A number of studies have clearly shown that the anomalies in the northern hemisphere wintertime 500 hPa geopotential height are unimodally distributed (Wallace et al., 1991; Atger, 1999). Furthermore, there is evidence suggesting that 500 hPa geopotential height anomalies are close to normally distributed (Toth, 1991; Stephenson, 1997). While this may be disappointing news for chaos theorists desperately seeking non-linear regimes, it offers some hope of recovering some order from the collective behaviour of a large number of degrees of freedom. An appropriate analogy is provided by a room full of gas molecules in which the dynamical interactions between colliding molecules are highly non-linear, yet the velocities of the molecules are normally distributed (Maxwell–Boltzmann distribution).

8.2. Moment tests for multinormality

Mardia (1970) developed powerful multivariate tests for multinormality by defining moment

measures of multivariate skewness ($b_{1,q}$) and kurtosis ($b_{2,q}$):

$$b_{1,q} = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} G_{kl}^3, \tag{23}$$

$$b_{2,q} = \frac{1}{n} \sum_{k=1}^{n} G_{kk}^2, \tag{24}$$

based on the invariant correlations between the principal components of the ensemble members

$$G_{kl} = \boldsymbol{y}_k^T \boldsymbol{C}^{-q} \boldsymbol{y}_l = n \sum_{i=1}^{q} U_{ik} U_{il}, \tag{25}$$

where $\boldsymbol{C}^{-q}$ is the rank q pseudo-inverse of the ensemble covariance matrix $\boldsymbol{C}$ (Stephenson, 1997). Deviations of $b_{1,q}$ from zero, and $b_{2,q}$ from $q(q+2)$ indicate the presence of non-normality, and these measures resort to the usual univariate expressions for skewness and kurtosis in the case when $q = 1$. Furthermore, the measures are invariant under affine linear transformations (Mardia et al., 1979), and therefore should remain constant when the ensemble evolves linearly in phase-space. Constant multivariate skewness and kurtosis, however, do not always imply that the evolution is linear.

In order to sample multivariate skewness and kurtosis, there must be at least as many samples (n) as variables (q). Unfortunately, this is not usually the case for meteorological gridded data where ensemble sizes are small and the number of grid point variables are large. However, the multivariate skewness and kurtosis can be estimated for the leading $q < n$ principal components by adopting the rank q pseudo-inverse $\boldsymbol{C}^{-q}$ of the ensemble covariance matrix $\boldsymbol{C}$ (Stephenson, 1997).

8.3. *An illustration: skewness and kurtosis of the ECMWF forecasts*

Fig. 9 shows the multivariate skewness and kurtosis calculated for the first $q = 4$ leading principal components of the ensemble that explain more than 60% of the total variance.

The multivariate skewness $b_{1,4}$ is initially zero due to the symmetrical way in which the ensemble perturbations were constructed. It then steadily increases, especially after day 2, and becomes largest for the long-lead forecasts after day 7. The skewness at day 8 was previously noted in the MDS plot in Fig. 8e and led to the ensemble mean being slightly displaced from the mode and verification analysis. Sample estimates of the mean are not robust in in the presence of multivariate skewness (Mardia, 1970). Despite skewness increasing due to non-linear interactions, it remains below the 95% asymptotic confidence limit of 3.695 expected for a normally distributed ensemble of forecasts. Under the null hypothesis of normality, $b_{1,q}$ is asymptotically distributed as $6\chi_f^2/n$ with $f = q(q+1)(q+2)/6$ in the limit $n \to \infty$ (Mardia et al., 1979).

Initially, the multivariate kurtosis $b_{2,4}$ of the ensemble is significantly less than the value of $q(q+2) = 24$ expected for a multinormal distribution (Fig. 9b). The distribution in phase space is more clustered near the origin with thinner extreme tails than a multinormal distribution (i.e., it is top hat rather than bell-shaped). Such "platykurtis" is due to the initial perturbations having been chosen to all have the same energy, rather having a random selection over all possible amplitudes. Such a choice avoids ensemble members being too similar to the control forecast, yet is inconsistent with analysis errors that are most likely to be almost normally distributed. Platykurtis can lead to sample covariance estimates underestimating the population covariance (Mardia, 1974). After 2 days, the multivariate kurtosis increases to values consistent with the multinormal distribution at 95% confidence. Two-sided confidence limits are calculated using the asymptotic expression

$$b_{2,q} - q(q+2) \sim N[0, \sqrt{8q(q+2)/n}]$$

given in Mardia et al. (1979).

To summarise, not only do the ensemble forecasts spread out in time but they also become more skewed and less platykurtic at later times. This can lead to less reliable estimates of the mean forecast, yet more reliable estimates of the covariance at later times. Furthermore, the multivariate measures are far from constant especially from days 2–5 indicating the presence of strongly non-linear evolution in phase space. After day 2, both the multivariate skewness and kurtosis have values consistent with those that could be sampled from a multinormal distribution at 95% confidence. When the multivariate tests reveal that the distribution is close to multinormal, one should be careful in searching for and finding "clusters" that might simply be due to sampling. However, large values of multivariate skewness imply that anisotropies do exist, and these can then be examined

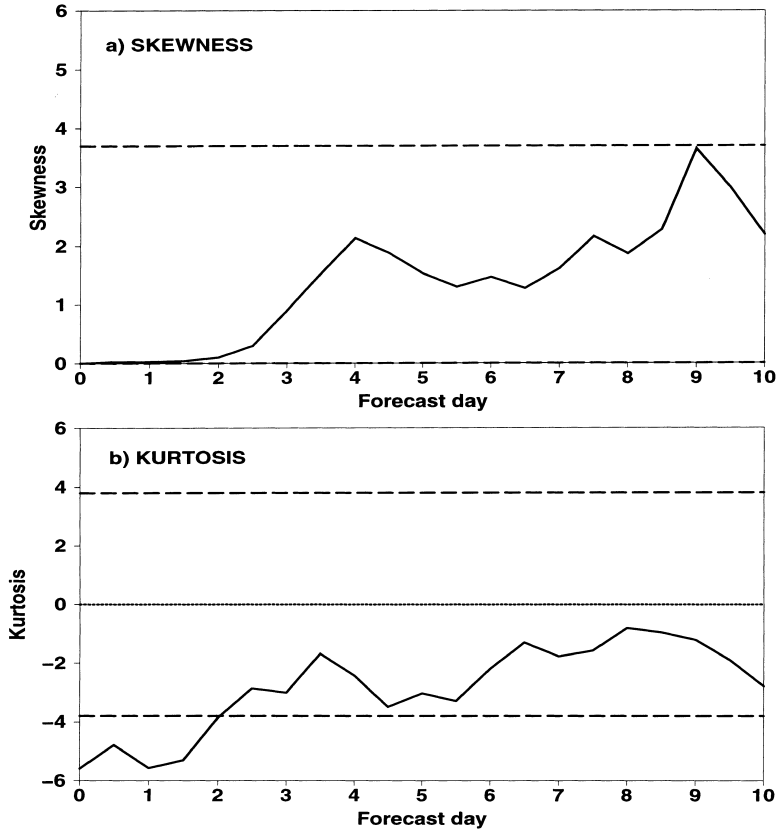

Fig. 9. Evolution of (a) the multivariate skewness $b_{1,4}$, and (b) the multivariate kurtosis $b_{2,4}-24$ for the first four leading principal components of the ensemble, which explain more than 60% of the total variance. Multinormal 95% confidence limits are marked as dashed lines based on Mardia's asymptotic large sample expressions.

in more detail using directional clustering methods.

8.4. Reliability of ensemble estimates

Ensemble forecasting is based on the implicit assumption that ensemble sample estimates provide reliable estimates of population values. However, when ensemble forecasts deviate strongly from normality, sampling errors increase and sample estimates become less reliable and eventually less meaningful. For example, the p.d.f. for certain low order systems can sometimes bifurcate into a bimodal distribution, for which the mean is no longer a physically realisable state (Smith, 1996). Such strong deviations from normality can be favoured by selecting initial perturbations that diverge the most rapidly from the ensemble mean (Buizza, 1994; Buizza and Palmer, 1995; Barkmeijer, 1996). Such a sampling strategy has the advantage that a small sample can capture a large fraction of the total variance (Ehrendorfer and Tribbia, 1997), yet has the disadvantage that it can lead to less reliable sample estimates due to the creation of "outlier" forecasts. Outlier forecasts should not be considered to be "erroneous forecasts" since they also provide useful information about the probability distribution. Nevertheless, it may be possible to obtain more robust and reliable estimates of the mean by either "trimming" or "downweighting" extreme outliers once they have been identified. For multivariate quantities, iterative "peeling" methods may also be used to obtain more robust estimates of the covariance (Mardia et al., 1979). Because of the small size of ensembles, the robustness of sample estimates is

an important issue that merits more serious attention.

9. Entropy of an ensemble of forecasts

An ensemble of forecasts in phase space can be considered to be a gas of n atoms moving in p-dimensions. Initially, the atoms are all close to one another but are most likely to spread apart over time as do atoms in a real gas. The atoms disperse because there are more ways of being arranged over a large domain, than the unique way of being arranged to all be at the same point in phase space. This very general law of nature is embodied in the second law of thermodynamics, which states that the "entropy" or disorder of a closed system will tend to increase in time. The law is independent of detailed dynamical arguments, yet offers a general explanation for why forecasts from slightly perturbed initial conditions generally spread out in time.

How can entropy be defined for an ensemble of forecasts? The discussion above suggests that it is somehow related to the dispersion of the ensemble in phase space, and hence the covariance. The dispersion of the ensemble is inversely related to the sharpness of the probability distribution in phase space, which can be measured by the "information content" of the probability density function defined as the functional

$$I[\Omega] = \int \Omega(\boldsymbol{x}) \log \Omega(\boldsymbol{x}) \, \mathrm{d}^p \boldsymbol{x}. \tag{26}$$

Identifying entropy with the "loss of information" gives a working definition of entropy as $E[\Omega] = -I[\Omega]$. H-theorems have been proved which show that entropy defined in this manner will increase in time for non-linear fluid systems (Carnevale et al., 1981; Carnevale, 1982). For a given covariance, the maximum entropy is obtained when the distribution is multinormal (Carnevale, 1982). The multivariate central limit theorem defines more precisely the conditions that are necessary for the emergence of this most disordered of distributions (Mardia et al., 1979). Put simply, the net result of MANY uncorrelated processes, even non-linear ones, will generally cause entropy to increase thereby leading to the emergence of the maximum entropy state — the multivariate normal (multinormal) distribution. By substituting eq. (21) into eq. (26), the maximum entropy of the multinormal distribution can be shown to be completely determined by the determinant of the covariance matrix:

$$E = \frac{1}{2} \log \det(2\pi \boldsymbol{C}) + \frac{r}{2}$$
$$= \prod_{k=1}^{r} \sigma_k + \frac{r}{2} \log\left(\frac{2\pi e}{\sqrt{n}}\right). \tag{27}$$

The changes in the entropy of the ensemble are proportional to the relative changes in the ensemble volume since $\det 2\pi \boldsymbol{C}$ transforms to $\det 2\pi \boldsymbol{L}\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}} = \det 2\pi \boldsymbol{C} \det \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$, and $\det \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ is simply the Jacobian of the transformation.

A regular increase in the maximum entropy of the ensemble over the Euro-Atlantic region can be seen in Fig. 10. The increase in entropy is due to a continual growth in ensemble volume, and indicates a loss of initial information. The regular monotonic evolution of entropy resembles that shown in Fig. 1 of Carnevale and Holloway (1982) calculated for turbulent flows on a rotating β-plane using 2 member ensembles. By considering the inequality between arithmetic and geometric means of the eigenvalues, it can be shown that the maximum entropy is bounded from above by the logarithm of the volume D^r estimated from

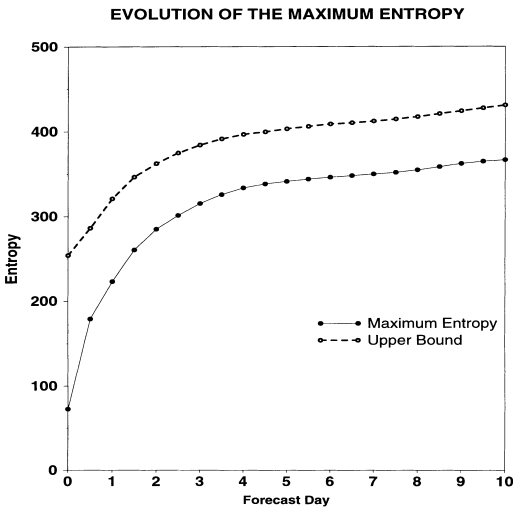

Fig. 10. Evolution of the maximum entropy of the ensemble (solid line). The dashed line gives the upper bound of entropy based on the spread. Note the rapid increase in the maximum entropy in the 1st 3 days.

the mean spread of the ensemble:

$$E \leqslant \log D^r + \frac{r}{2} \log\left(\frac{2\pi e p \sqrt{n}}{r}\right). \quad (28)$$

Before day 2, the maximum entropy grows faster than does the upper bound because of major structural changes that occur in the eigenvalue spectrum (Fig. 10). For later forecasts, the upper bound exceeds the maximum entropy by an almost constant amount. Especially during transitions, the maximum entropy can provide additional useful information to that provided by the dispersion. For example, abrupt changes in entropy associated with changes in the rank of the ensemble ("dimensional phase transitions") are expected to occur as the ensemble passes from one dynamical regime to another. Consider, for example, the motion of an ensemble described by the well-known three variable model of Lorenz (1963), whose attractor resembles two butterfly-wings. An ensemble of forecasts initially confined in one of the wings will have a rank close to 2, yet will temporarily become 3 dimensional while some of the members make the regime transition into the other wing*. Ensemble entropy may therefore be a useful probe for detecting regime transitions especially in regions where atmosphere blocking occurs.

10. Concluding remarks

Because of the ever increasing amount of data generated by ensemble forecasts, it is crucial that new methods are developed to extract useful information from such forecasts. This study has developed and illustrated some statistical methods that can be usefully applied.

(1) An ensemble of forecasts can be considered to be a linear map between phase-space and sample space. The linear map is naturally represented by a rectangular data matrix. Important information about the ensemble can be obtained by making the singular value decomposition of the data matrix.

(2) Inner and outer products of the data matrix with itself give rise to the two covariance matrices: $\boldsymbol{C}$ — "field covariance" and $\boldsymbol{B}$ — "ensemble covariance". The 2 covariance matrices share the same eigenvalues.

(3) Ensemble covariance $\boldsymbol{B}$ contains complete information on the mutual Euclidean distances between ensemble members, and can be summarised by multidimensional scaling techniques. The resulting maps show at a glance how the members of the ensemble are clustered. This exploratory tool provides a useful method for summarising ensembles.

(4) Certain initialisation procedures and strongly non-linear evolution can lead to deviations from normality of the ensemble forecasts. The deviations can be investigated using multivariate measures of skewness and kurtosis. Strong deviations from normality lead to biases in the ensemble mean and covariance that may degrade the skill of ensemble forecasts. After day 2, the ensemble of forecasts studied in this article was found to be normally distributed at 95% confidence. In other words, the ensemble of forecasts had evolved towards a state close to "statistical equilibrium".

(5) Loss of predictability is due to the irreversible evolution from an orderly state having all members close together, to a more disorderly state with ensemble members spread all over the attractor. The irreversible loss of information can be quantified by estimating the entropy of an ensemble based on the determinant of the covariance matrix. The rate of entropy production puts strong limits on the predictability of a dynamical system, and may also be useful for detecting regime transitions. More general definitions of entropy such as Kullback-Leibler and Renyi informations would also be worth investigating in future studies (Ruelle, 1989).

Statistical methods can be used to extract useful coarse-grained information from complex systems such as the atmosphere. They can be used to pan out the precious predictable signals from the complex disorder created by synoptic-scale non-linear dynamics, and offer a useful complimentary paradigm to that of low-order chaos for understanding complex dynamical systems. Statistical methods exploit the high dimensionality of phase space and may ultimately be able to provide simple descriptions of the complex atmospheric and climate variations that occur on large time and space scales.

* Dimension here refers to the rank of the ensemble, i.e., the "local" dimension rather than the "global" dimension of the attractor.

11. Acknowledgements

The data and much useful information about the ECMWF forecasting system were kindly supplied by Dr. Jan Barkmeijer. We are grateful to Dr. Philippe Besse, Michel Déqué, Jean-Francois Royer, Dr. David Ruelle, Dr. Zoltan Toth, Dr. Christine Ziehmann and an anonymous reviewer for their helpful remarks. It is a pleasure to acknowledge Dr. Greg Holloway and Dr. George Carnevale for interesting discussions and for copies of their profound studies on the statistical mechanics of fluids. Finally, we wish to thank Dominique Besson for her ever efficient and friendly help in retrieving the more historical articles. The authors of this study were supported as visiting scientists at Météo-France by the European Commission contract ENV4-CT95-0122 (SHIVA) and the CYCIT contract CLI97-0558.

12. Appendix

The ECMWF 51-member forecast ensemble

The global weather forecasting Ensemble Prediction System is routinely employed at ECMWF to make weather forecasts (Molteni et al., 1996; Buizza et al., 1997). The state-of-the-art numerical weather forecasting model is based on physical conservation laws and represents the atmosphere by 31 levels in the vertical and a spatial grid resolution of about 100 km (T_L159 spectral truncation). The best estimate of the observed current state of the atmosphere is obtained by continuously assimilating observed data into the model, and serves both as a suitable forecasting initial condition, and also for verifying the accuracy of previous forecasts. More details about the model are presented in Buizza et al. (1998).

Because there are many more grid point variables than there are ensemble members, it is not evident how to choose the small number of perturbed initial conditions so as to optimally sample the large number of variables. Various dynamical approaches have been proposed to avoid this "curse of dimensionality" (Buizza and Palmer, 1995; Toth and Kalnay, 1993, 1997; Houtekamer and Derome, 1995). There is a potential danger in this approach since it is known that for systems with only a few variables, dynamical constraints on the initial conditions can lead to severe undersampling (Anderson, 1997). The initial perturbations of the ECMWF ensemble are based on the fastest growing modes (singular vectors of the linearised dynamics) calculated at the coarser horizontal truncation of T42. These identify directions in which maximal energy growth would occur over the next 2 days with the linearised dynamics (Buizza and Palmer, 1995). The energy norm at initial time is used as a crude yet computationally cheap approximation to the inverse of the analysis error covariance (Mahalanobis) norm, that should be used if the analysis errors are normally distributed (Stephenson, 1997; Barkmeijer et al., 1998, 1999). For the extra-tropics in each hemisphere, around 40 singular vectors are computed by applying a Lanczos algorithm, and then 25 are retained. The 4 fastest growing singular vectors are always selected, and each subsequent singular vector is only chosen when half of its total energy is outside the areas where the singular vectors already selected are located. After selection, a rotation and scaling is applied separately to the 25 singular vectors in each hemisphere. The rotation generates perturbations which are more spatially uniform, and the scaling guarantees that the perturbation amplitudes are consistent with typical estimates of analysis error variance (Molteni et al., 1996). It should be noted that initial perturbations are not infintessimal ("the flapping of a butterfly's wings"), since they are chosen to have finite amplitudes to those of typical analysis errors. The first Northern Hemisphere perturbation is added to the first Southern Hemisphere perturbation, and so on, up to the 25th, after which the 25 resulting global perturbations are alternately added and subtracted to the analysis to define 50 perturbed initial conditions.

REFERENCES

Anderson, J. L. 1997. The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: low-order perfect model results. *Mon. Wea. Rev.* **125**, 2969–2983.

Atger, F. 1999. Tubing: an alternative to clustering for the classification of ensemble forecasts. *Weather and Forecasting* **14**, 741–757.

Barkmeijer, J. 1996. Constructing fast-growing perturbations for the nonlinear regime. *J. Atmos. Sci.* **53**, 2838–2851.

Barkmeijer, J., Buizza, R. and Palmer, T. N. 1999. 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quart. J. Roy. Met. Soc.* **125**, 2333–2351.

Barkmeijer, J., Gijzen, Van M. and Bouttier, F. 1998. Singular vectors and estimates of the analysis error covariance metric *Quart. J. Roy. Met. Soc.* **124**, 1695–1713.

Brankovic, C. and Palmer, T. N. 1997. Atmospheric seasonal predictability and estimates of ensemble size *Mon. Wea. Rev.* **125**, 859–874.

Brankovic, C., Palmer, T. N., Molteni, F., Tibaldi, S. and Cubasch, U. 1990. Extended-range predictions with ECMWF models: time-lagged ensemble forecasting. *Quart. J. Roy. Met. Soc.* **116**, 867–912.

Buizza, R. 1994. Sensitivity of optimal unstable structures. *Quart. J. Roy. Met. Soc.* **120**, 429–451.

Buizza, R. 1997. 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., Gelaro, R., Molteni, F. and Palmer, T. N. 1997. The impact of increased resolution on predictability studies with singular vectors. *Quart. J. Roy. Met. Soc.* **123**, 1007–1033.

Buizza, R. and Palmer, T. N. 1995. The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.* **52**, 1434–1456.

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

Burgers, G. and Stephenson, D. B. 1999. The "Normality" of El Niño. *Geophys. Res. Letters* **26**, 1027–1030.

Carnevale, G. F. 1982. Statistical features of the evolution of two-dimensional turbulence. *J. Fluid. Mech.* **122**, 143–153.

Carnevale, G. F., Frisch, U. and Salmon, R. 1981. H theorems in statistical fluid dynamics. *J. Phys. A* **14**, 1701–1718.

Carnevale, G. F. and Holloway, G. 1982. Information decay and the predictability of turbulent flows. *J. Fluid. Mech.* **116**, 115–121.

Charney, J. G., Fjörtoft, R. and von Neumann, J. 1950. Numerical integration of the barotropic vorticity equation. *Tellus* **2**, 237–254.

Cox, T. F. and Cox, M. A. A. 1994. *Multidimensional scaling*. Chapman and Hall, London.

Ehrendorfer, M. 1994. The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Mon. Wea. Rev.* **122**, 703–713.

Ehrendorfer, M. and Tribbia, J. J. 1997. Optimal prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.* **54**, 286–313.

Epstein, E. S. 1969a The role of initial uncertainties in prediction. *J. Appl. Meteor.* **8**, 190–198.

Epstein, E. S. 1969b. Stochastic dynamic prediction.*Tellus* **21**, 739–759.

Hadamard, J. 1898. Les surfaces a courbures opposées et leurs lignes géodesiques. *J. Math. Pure et Appl.* **4**, 27–73.

Hamill, T. M. and Colucci, S. J. 1998 Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.* **126**, 711–724.

Houtekamer, P. L. and Derome, J. 1995. Methods for ensemble prediction. *Mon. Wea. Rev.* **123**, 2181–2196.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H. and Mitchell, H. L. 1996. A system simulation approach to ensemble prediction. *Mon. Wea. Rev.* **124**, 1225–1242.

Landau, L. D. and Lifshitz, E. M. 1980. *Statistical physics.* Butterworth Heinemann, 3rd edition, vol. 5.

Leith, C. E. 1971. Atmospheric predictability and two-dimensional turbulence. *J. Atmos. Sci.* **28**, 145–161.

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

Leith, C. E. and Kraichnan, R. H. 1972. Predictability of turbulent flows. *J. Atmos. Sci.* **29**, 1041–1058.

Lorenz, E. N. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141.

Lorenz, E. N. 1969a. The predictability of a flow which possesses many scales of motion. *Tellus* **21**, 289–307.

Lorenz, E. N. 1969b. Three approaches to atmospheric predictability. *Bulletin American Met. Soc.* **50**, 345–349.

Mardia, K. V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519–530.

Mardia, K. V. 1974. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyä. The Indian Journal of Statistics* **36**, 115–128.

Mardia, K. V., Kent, J. T. and Bibby, J. M. 1979. *Multivariate analysis.* Academic Press Limited, London.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. 1996. The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Met. Soc.* **122**, 73–119.

Murphy, J. M. 1988. The impact of ensemble forecasts on predictability. *Quart. J. Roy. Met. Soc.* **114**, 463–493.

Poincaré, H. 1890. Sur les equations de la dynamique de le problème de trois corps. *Acta Math.* **13**, 1–270.

Ruelle, D. 1989. *Chaotic evolution and strange attractors.* Cambridge University Press, Cambridge.

Selten, F. M. 1997. Baroclinic empirical orthogonal functions as basis functions in an atmospheric model. *Mon. Wea. Rev.* **54**, 2100–2114.

Sivillo, J. K., Ahlquist, J. E. and Toth, Z. 1997. An ensemble forecasting primer. *Weather and Forecasting* **12**, 809–818.

Skelly, W. C. and Henderson-Sellers, A. 1996. Grid box or grid point: what type of data do GCMs deliver to climate impacts researchers? *Int. J. Climatol.* **16**, 1079–1086.

Smagorinsky, J. 1969. Problems and promises of deterministic extended range forecasting. *Bulletin American Met. Soc.* **50**, 286–311.

Smith, L. A. 1996. Accountability and error in ensemble forecasting. In: Palmer, T. (ed.): *Predictabilty*, ECMWF November 1997 Workshop on Predictability, vol. 1, pp. 351–368. Reading, UK. ECMWF.

Smith, L. A. and Gilmour, I. 1997. Accountability and internal consistence in ensemble formation. In: Palmer, T. (ed.): *Predictabilty*, ECMWF November 1997 Workshop on Predictability, vol. 2, p.15. Reading, UK. ECMWF.

Stephenson, D. B., Rupa Kumar, K., Doblas-Reyes, F.-J., Royer, J.-F., Chauvin, F. and Pezzulli, S. 1999. Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon. *Mon. Wea. Rev.* **127**, 1954–1966.

Stephenson, D. B. 1997. Correlation of spatial climate/weather maps and the advantages of using the mahalanobis metric in predictions. *Tellus* **49A**, 513–527.

Strang, G. 1988. *Linear algebra and its applications.* Harcourt Brace Jovanovich, Inc., 3rd edition.

Stroe, R. and Royer, J.-F. 1993. Comparison of different error growth formulas and predictability estimation in numerical extended-range forecasts. *Ann. Geophysicae* **11**, 296–316.

Sutton, O. G. 1951. Mathematics and the future of meteorology. *Weather* **6**, 291–296.

The Economist, 27 March 1999. Economics focus: À la mode. **350**, 90.

Thompson, P. D. 1957. Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus* **9**, 275–295.

Toth, Z. 1991. Circulation patterns in phase space: a multinormal distribution? *Mon. Wea. Rev.* **119**, 1501–1511.

Toth, Z. 1995. Degrees of freedom in Northern Hemisphere circulation data. *Tellus* **47A**, 457–472.

Toth, Z. and Kalnay, E. 1993. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.* **74**, 2317–2330.

Toth, Z. and Kalnay, E. 1997. Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.* **125**, 3297–3319.

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

Wallace, J. M., Cheng, X. and Sun, D. 1991. Does low-frequency atmospheric variability exhibit regime-like behavior? *Tellus* **43AB**, 16–26.

Weinberg, S. 1972. *Gravitation and cosmology*. John Wiley & Sons, Inc.

Whitaker, J. S. and Loughe, A. F. 1998. The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.* **126**, 3292–3302.

Zwiers, F. W. 1996. Interannual variability and predictability in an ensemble of AMIP climate simulations conducted with the CCC GCM2. *Climate Dynamics* **12**, 825–847.