Q. J. R. Meteorol. Soc. (2004), **130**, pp. 583–605 doi: 10.1256/qj.02.146

On the existence of multiple climate regimes

By D. B. STEPHENSON*, A. HANNACHI and A. O'NEILL
Department of Meteorology, University of Reading, UK

(Received 4 July 2002; revised 27 July 2003)

SUMMARY

New techniques are presented for testing the three main hypotheses about the probability distribution of the climate system: multinormal (single regime), unimodal but not multinormal (single regime), and multimodal (multiple regimes). Rather than searching for evidence that confirms the multimodal hypothesis expected from the chaos and other strongly nonlinear paradigms, our strategy is to try and reject the simplest single-regime hypothesis of multinormality expected for aggregate indices of many local weather degrees of freedom. Concerning multiple climate regimes in the northern hemisphere, we find no strong evidence in the available monthly mean reanalysis data for rejecting the single-regime multinormal hypothesis in favour of the multimodal hypothesis. A simple non-parametric method is presented for transforming state space into a more homogeneous probability space that makes regimes easier to identify. A spatial point process test is used in this space to demonstrate that the hemispheric clusters are not significantly different to what could be expected from sampling a unimodal distribution. Based on the observed data, the single-regime multinormal hypothesis can not be rejected at the 5% level of significance and so provides the simplest useful model for the probability distribution for the northern hemisphere geopotential-height field.

KEYWORDS: Chaos Clusters Multinormal

1. INTRODUCTION

Meteorologists have often tried to explain weather variations in terms of a finite set of preferred weather patterns/types. Persistent and/or recurrent states of the atmosphere are referred to as 'regimes' and can have large socio-economic impacts. For example, in autumn 2000 the three-month persistence of the Scandinavian (or Eurasian-1) teleconnection pattern led to anomalous climate in western Europe and many costly floods in England. Despite enormous progress in modelling climate variability, there is little understanding of why certain states persist for such long times, and there are at present no reliable estimates for the probability of occurrence of such events. This study aims to improve this situation by developing and statistically testing several possible probability models. This is essential for improving our ability to interpret climate events and to make probabilistic inferences about the future likelihood of such events.

Early studies identified several low-frequency synoptic regimes that persist longer than the lifetime of typical cyclones (Rex 1950; Namias 1950, 1964; Bauer 1951; Bjerknes 1969; Dole and Gordon 1983; Horel 1985; and references therein). At the end of the 1970s, two main approaches emerged for explaining the observed persistence of large-scale extratropical flow anomalies. One approach was based on Rossby wave theories of linear stationary waves forced by persistent diabatic heating in the Tropics (e.g. Hoskins and Karoly 1981; Simmons *et al.* 1983). An alternative approach explained persistent flow patterns as multiple (quasi-)stationary solutions of the nonlinear fluid-dynamical equations of motion (e.g. Charney and DeVore 1979; Wiin-Nielsen 1979). Many regime studies were theoretically motivated by the multiple regimes found in the chaotic solutions of certain low-order nonlinear dynamical systems (Lorenz 1963, 1970; Charney and DeVore 1979; Sutera 1986; White 1980; Palmer 1993, 1999). In the chaos paradigm, such regimes are interpreted as quasi-stationary states or metastable fixed points which sporadically attract the chaotic trajectory of the system (Legras and Ghil 1985; Mukougawa 1988; Branstator and Opsteegh 1989; Haines and Hannachi 1995).

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

TABLE 1. SUMMARY OF SOME PREVIOUS REGIME STUDIES SHOWING HOW MANY TIME SAMPLES WERE USED, THE SAMPLE TIME, AND THE NUMBER OF REGIMES IDENTIFIED IN EACH SAMPLE

Study	Sample size	Sampling rate	No. of regimes
Sutera (1986)	360	daily	2
Hansen and Sutera (1986)	1440	daily	2
Molteni *et al.* (1988)	1152	5-day	2
Marshall and Molteni (1993)	576	daily	2
Mo and Ghil (1988)	2400	daily	6
Vautard (1990)	2772	2-day	4
Kimoto and Ghil (1993)	3330	daily	4
Cheng and Wallace (1993)	702	5-day	3
Wallace *et al.* (1991)	702	5-day	3
Smyth *et al.* (1999)	3960	daily	2–3
Corti *et al.* (1999)	270	monthly	4
Monahan *et al.* (2000, 2001)	3670	daily	3
Hsu and Zwiers (2001)	300	monthly	1–3

Although mostly found in low-order chaotic systems, multiple regimes can also exist in certain high-order nonlinear systems, for example, in the model of a stochastically forced particle moving around multiple potential wells in a high-dimensional space (Hasselmann 1999)

In order to support theoretical models of multiple equilibria, many investigators have tried to find evidence of multiple regimes in observed data. The main techniques used to detect multiple flow regimes have been: cluster analysis (e.g. Mo and Ghil 1988; Cheng and Wallace 1993); mode and bump hunting in probability density estimates (e.g. Sutera 1986; Hansen and Sutera 1986; Molteni *et al.* 1988, 1990; Kimoto and Ghil 1993, hereafter KG93; Corti *et al.* 1999); variants of principal-component analysis (e.g. Vautard 1990; Marshall and Molteni 1993; Monahan 2000; Monahan *et al.* 2001). Since the majority of these descriptive classification approaches are designed to classify the data into regimes, they invariably do 'find' varying numbers of discrete regimes (see Table 1 for a summary). However, descriptive methods are not well suited to testing for the existence (or number) of regimes since they are purely data analytic and often lack any underlying probability model. To test hypotheses about the existence of regimes, it is necessary to identify suitable probability distributions for modelling the distribution in state space.

Section 2 of this article presents a brief overview of the main concepts involved, and section 3 identifies probability models (hypotheses) that can be tested. Rather than attempt to confirm the multimodal hypothesis, our strategy is to test whether the simplest hypothesis of multinormality can be rejected given a particular sample of data. The approach has been illustrated using northern hemisphere (NH) data from the study of Corti *et al.* (1999) (hereafter C99) and data generated by the well-known three-variable Lorenz chaos model (section 4). The joint probability distribution of both datasets is explored in section 5 using several different techniques not previously applied to meteorological data. Section 6 sets out a robust non-parametric methodology for testing for the existence of clustering and applies it to both datasets to find out whether the regimes reported in C99 are statistically significant features or merely sampling artefacts.

2. CONCEPTS AND DEFINITIONS

The quest to identify modes/patterns/regimes of the climate system is driven by the desire to find structure in state space. To avoid the confusion caused by the

interchangeable and imprecise use of the words 'mode', 'pattern' and 'regime', brief definitions of these concepts will now be given.

(*a*) *State space*

The large-scale state of the atmosphere is defined by a set of meteorological variables; for example, a set of geopotential-height grid-point variables, or, alternatively, the leading q principal components (PCs) of gridded geopotential height. The q variables define a q-dimensional 'state space', in which the evolving state of the climate system can be represented by a trajectory of points*. The dynamical evolution of the climate system can be thought of as the motion of a point through this multidimensional state space. The amount of time the system spends in each part of state space can be measured by estimating the joint probability density function (p.d.f.) of the state variables. Some regions of state space will be visited for longer or more often than are other regions and will therefore have larger probability densities. While it is true that persistence in a particular state will lead to a larger p.d.f. for that state, the converse is not always true since a p.d.f. can also be larger when a state is visited more frequently (yet less persistently).

(*b*) *Climate modes*

The concept of 'mode' is the most difficult to define, since it implies that the structure is *physically* meaningful. Many previous studies have attempted to identify 'physical modes' of variability by their unique statistical properties in state space such as, for example, maximum variance ('empirical orthogonal function (EOF) patterns') or maximum probability density ('regimes'). However, such techniques do not guarantee that the resulting structures are physically meaningful. An interesting example of this is provided by the ongoing debate over the physical meaning of the Arctic oscillation leading sea-level pressure EOF (e.g. Wallace 2000; Ambaum *et al.* 2001, 2002; Wallace and Thompson 2002). Therefore, it is perhaps better to use the more neutral words 'pattern' and 'regime' rather than 'mode' unless there is strong physical justification for doing otherwise. Rather than focus on individual structures, it is often more constructive to consider a whole set of identified structures as simply a useful reduced basis for describing the state of the climate system.

(*c*) *Climate patterns and indices*

A 'pattern' defines a particular direction from the origin in state space independent of amplitude. It can be defined by specifying a set of coefficients for all the state variables, for example, an EOF pattern. An associated climate 'index' can then be constructed by using these coefficients to make a linear combination of the variables, for example, the PC time series associated with a particular EOF pattern. In other words, a pattern is a direction in state space and the associated index is the projection onto this direction (i.e. a linear combination of grid-point variables).

(*d*) *Climate regimes*

A 'regime' can be defined as a region of state space that is more populated than neighbouring regions: in other words, a region of clustering in state space. Local 'clusters' can be identified (defined) using clustering algorithms that aim to classify the points

* State space is sometimes referred to as 'phase space', which strictly refers to the space spanned by the canonical variables of a Hamiltonian dynamical system.

into several distinct classes (Mardia *et al.* 1979). These descriptive classification techniques depend on a choice of metric, and are not based on any underlying probability model. Alternatively, clusters can be inferred from either local maxima ('modes') or locally concave regions ('bumps') in the estimated p.d.f. (Silverman 1994). Bumps are regions where the curvature of the p.d.f. is negative, and so include sharp drops in gradient as well as local maxima. Hence, searching for multiple regimes becomes a quest for multimodality or an exercise in 'bump hunting' (Good and Gaskins 1980; Silverman 1981, 1994). Care should be taken not to confuse 'modes' in the probability density with 'physical modes'. As an example, consider a simple harmonic oscillation with a small amount of added noise—the system has two local maxima (modes) at the edges of its probability distribution, yet it is certainly not a bimodal (two-mode) physical system. A major disadvantage of these density methods is that the resulting number of regimes depends strongly on how much smoothing is used to estimate the p.d.f.

In addition to local clusters, it is also possible to have 'directional clusters' in which particular directions rather than local regions of state space are more populated than other directions. In other words, there can be preferred 'patterns', but with no preferred amplitudes. An even more exotic possibility is to have clustering along curved lines in state space that can be identified using techniques such as nonlinear/curvilinear PC analysis (see Monahan *et al.* 2000, 2001, and references therein).

3. Probability models (hypotheses)

Figure 1 illustrates the three simplest classes of probability density appropriate for describing continuous climate variables in state space. It should be noted that more complex multivariate distributions can also sometimes arise in certain specific situations (e.g. for two-dimensional limit cycles such as the quasi-biennial oscillation). The distributions in Fig. 1 represent distinct hypotheses about the state of the climate system. The multinormal distribution (Fig. 1(a)) assumes that all variables are normally ('Gaussian'*) distributed. The presence of nonlinearity, however, can lead to deviations from normality such as skewness and kurtosis as shown in the unimodal (but not multinormal) distribution in Fig. 1(b). When the interactions between the state variables are sufficiently nonlinear, the probability density can develop more than one local maximum (multimodal), which can then be interpreted as 'multiple regimes'.

(*a*) *The multinormal hypothesis* $H_0 : f(\mathbf{x}) = N_q(\mathbf{x}; \boldsymbol{B}, \boldsymbol{\Sigma})$

The multinormal (MULTIvariate NORMAL) distribution is widely used to describe multivariate data and has good sampling properties (Mardia *et al.* 1979). It is defined as

$$N_q(\mathbf{x}; \boldsymbol{B}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^q |\boldsymbol{\Sigma}|}} \, \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{B})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{B})}, \tag{1}$$

where q is the number of variables (e.g. $q = 2$ in the bivariate example shown in Fig. 1(a)), $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are, respectively, the population mean and population covariance matrix, and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. All state variables and linear combinations of state variables are normally distributed: for example, the standardized PCs of multinormally distributed variables are independently distributed as normal variables with unit variance and zero mean. In other words, state space spanned by standardized PCs—known as 'Mahalanobis space' (Stephenson 1997)—is directionally isotropic with a

* 'Gaussian' credits this distribution to the 1809 study by C. F. Gauss—whereas, in fact, it was already used by A. DeMoivre in 1714!

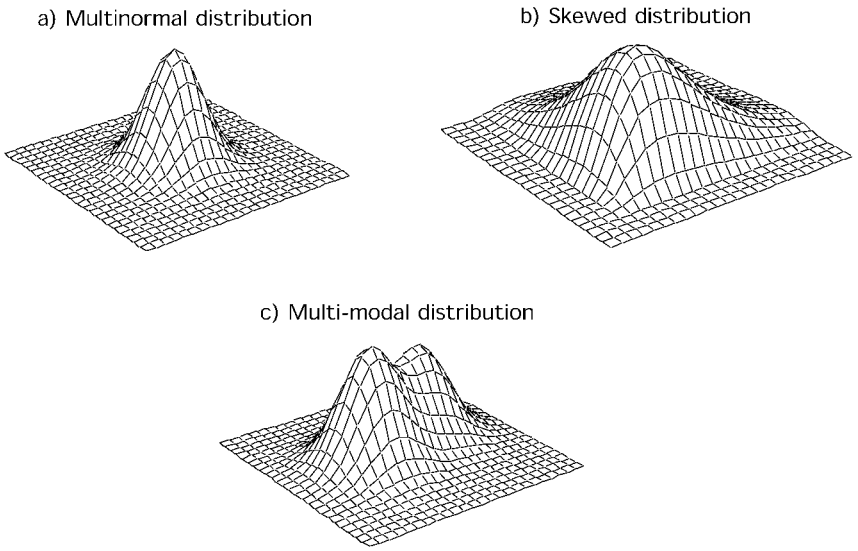

Figure 1. The three main classes of probability density illustrated for a bivariate climate system: (a) the multinormal distribution (single regime); (b) a unimodal non-normal distribution (single regime); (c) a multimodal distribution (multiple regimes).

single mode (regime) at the origin corresponding to the mean state. This may seem rather boring and uninteresting but has the virtue of providing a simple yet powerful probability model for climate studies (e.g. climate detection).

Leading EOFs (PC loading weights) often consist of several large-scale *centres of action* made up of neighbouring grid points having the same sign of weights. For example, EOF2 shown in Fig. 2(b) consists of three main centres of action: one positive centre over the subtropical Atlantic, and two negative centres over the Iceland and Aleutian regions. The leading PCs are linear combinations of several centres of action, each of which consists of a positively weighted mean of local grid-point variables. By the central-limit theorem, the centres of action and hence the leading PCs can, therefore, be expected to be more normally distributed than are individual grid-point variables. The same argument is less applicable to higher-order PCs where spatial dependency plays less of a role (i.e. EOFs are noisier). It should be noted that this argument relies upon the existence of spatial dependency between neighbouring grid-point variables. The argument can not be inverted to argue that grid-point variables should be normally distributed because individual grid-point variables are not linear combinations of weighted means of PCs. The normal distribution is the maximum entropy state referred to as statistical equilibrium by physicists. Statistical equilibrium not only occurs in isolated isotropic systems but is known to occur in more complex systems: for example, (chaotic) shell models of turbulence (Lorenz 1965; Aurell *et al.* 1994; Ditlevesen and Mogensen 1996), and many dissipative open systems (Egolf 2000). The key condition necessary for statistical equilibrium in any system is the existence of a large number of weakly interacting subsystems. This condition is most certainly satisfied for weather subsystems in the NH despite climate being a forced dissipative system (Stephenson 1997; Stephenson and Doblas-Reyes 2000). Evidence of multinormality in the NH wintertime flow was presented by Toth (1991) based on radial distances in state space.

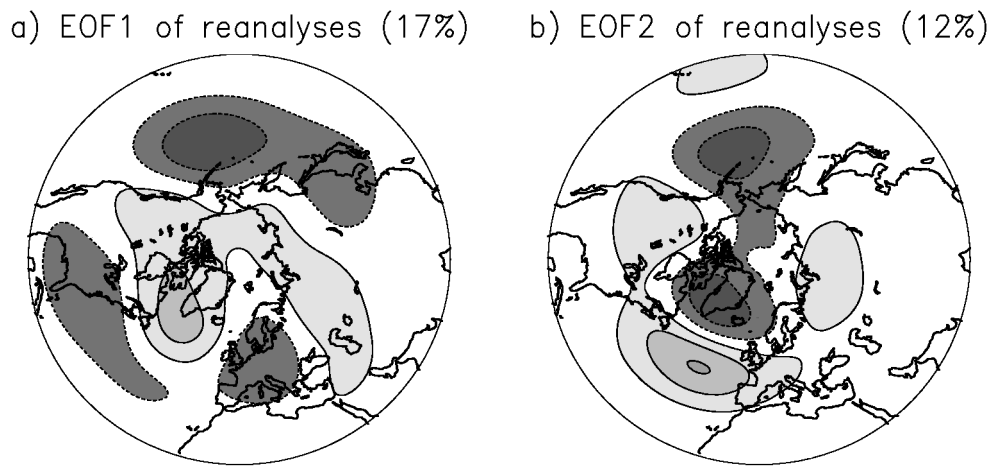

Figure 2. The leading EOFs (PC loading weights) of the monthly mean northern hemisphere 500 hPa geopotential-height analyses 1948–94: (a) EOF1 (17% of variance) that projects strongly on the PNA pattern, and (b) EOF2 (12% of variance) that projects strongly on the NAO pattern. Contour interval is arbitrary and negative values are shaded. See text for explanation.

(*b*) *The unimodal hypothesis* $H_1 : f(\mathbf{x}) = D(\mathbf{x}; \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Xi})$

The presence of nonlinearity or strong dependency between the local weather variables can lead to a unimodal distribution $D(\mathbf{x}; \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Xi})$ having some skewness or kurtosis (deviation from normality) described by the shape parameter $\boldsymbol{\Xi}$ (e.g. Fig. 1(b)); for example, the Niño-3 index is unimodally distributed with small yet significant positive skewness (Burgers and Stephenson 1999; Hannachi *et al.* 2003). Wallace *et al.* (1991) also noted a small amount of skewness in midlatitude geopotential heights, and attributed it to nonlinearity associated with blocking events.

(*c*) *The multimodal hypothesis* $H_k : f(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i D_i(\mathbf{x}; \boldsymbol{B}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Xi}_i)$, *where* $k > 1$

For very nonlinear systems such as low-order chaos models and multiple potential-well models, the probability distribution in state space can become multimodal with multiple attracting regimes. This is most naturally described by assuming that the probability distribution is a *mixture* of k unimodal distributions D_i with mixture weights α_i—see section 5(b) for more discussion. This is a very broad hypothesis that includes many possible probability distributions. The number of mixture components k is generally unknown a priori, and so has to be estimated from the sample of data.

(*d*) *Hypothesis testing*

Statistical inference can be used to decide which of these hypotheses is most likely given the observed sample of data. Whereas previous regime studies such as C99 attempt to 'accept' the multimodal hypothesis, the approach in this study will be to assess whether there is evidence to 'reject' the simplest null hypothesis of multinormality. This approach is required since hypothesis testing is inappropriate for 'confirmatory studies' based on testing results discovered by hunting through the data (Nicholls 2001). The rejection of the pure chance multinormal hypothesis avoids the prior belief required in regime studies that seek to confirm rather than reject multimodality. If no strong evidence can be found for rejecting the multinormal hypothesis, the multinormal

distribution can then be used as a suitable probability model for describing the state of the system.

Rather than fix a significance level a priori, we quote 'p-values' for all our test statistics in this article so that readers can make their own decisions (Nicholls 2001). The p-value is the probability of finding samples of data less consistent with the null hypothesis than the observed sample, i.e. the area in the tails of the sampling distribution of the test statistic beyond the observed value. If the p-value is less than level of significance α then we can reject the null hypothesis with $(1-\alpha)100\%$ confidence, e.g. $p = 0.03$ is less than 0.05 and so we can reject the null hypothesis with 95% confidence. Note that rejection of the multinormal hypothesis is a necessary but not sufficient condition for the existence of multimodality since it is also possible for the distribution to be unimodal and non-normal (e.g. skewed and/or kurtotic).

4. Data used in this study

To assess the recent claim by C99 that there are several distinct regimes in the low-frequency NH wintertime flow, this study will focus on the same dataset of monthly mean 500 hPa geopotential-height gridded analyses. To gain more insight, we also compare results with those generated by a low-order chaotic system known to have two distinct regimes (Lorenz 1963). Although only applied to these two examples, the new techniques introduced in this study do have much wider applicability to other climatic datasets, e.g. stratospheric data (Bo Christensen, personal communication).

(*a*) *Low-order chaos system*

The Lorenz (1963) model is a simple low-order set of nonlinear equations that mimics the chaotic sensitivity to initial conditions seen in the atmosphere and other fluid systems. It has been widely used as a test model for data assimilation (Miller *et al.* 1994; Hannachi and Haines 1998) and climate variability studies (Marshall and Molteni 1993; C99; Palmer 1999; and references therein). It has three degrees of freedom that satisfy the differential equations

$$\left.\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -\sigma(x-y), \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= -xz + rx - y, \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= xy - bz. \end{aligned}\right\} \tag{2}$$

Chaotic solutions are obtained when the parameters are set to $\sigma = 10$, $r = 30$, and $b = 8/3$ as in Lorenz (1963). To generate a data sample of identical size to that of the height analyses (see following section), we have integrated these equations forward in time 5000 times using a Euler scheme with time step $\Delta = 10^{-2}$. The first 140 points were then discarded in order to avoid transient behaviour not representative of the chaotic attractor. A sample of $n = 270$ means was then constructed by making 270 non-overlapping 18-point averages of the remaining values. The sample size was chosen to be identical to that of the available height analyses used in C99 (see next section).

(*b*) *Northern hemisphere wintertime flow*

We have repeated the analyses of C99 using National Center for Environmental Prediction (NCEP) November to April monthly mean 500 hPa geopotential-height

analyses for the period January 1949 to December 1994. The data were processed using exactly the same procedure applied by C99:

(i) Centred monthly anomalies first are made by subtracting the long-term mean annual cycle.
(ii) The anomalies are detrended by removing the five-year running November to April means.
(iii) Principal-component analysis (PCA) is then applied to the detrended grid-point anomalies in the NH extratropical region 0–360°E, 20°N–90°N (144 × 29 grid-point variables).

Only the reduced bivariate subspace of the two leading PCs is considered in all subsequent analysis.

Figure 2 shows the coefficients (EOFs) associated with the two leading PCs. The leading pattern in Fig. 2(a) resembles the familiar Pacific North American (PNA) teleconnection pattern but with some extension over the Eurasian continent. The second leading pattern in Fig. 2(b) captures the North Atlantic Oscillation (NAO) with further extensions over the North Pacific, Eurasia and parts of North America. These EOFs are similar to those reported in KG93 and Hsu and Zwiers (2001). The two leading PCs explain substantial amounts of the total variance of the filtered detrended anomalies and are well separated (17% and 12%). Reassuringly, our two leading PCs (not shown) are very similar to those used in C99 that were based on a slightly older dataset (National Meteorological Centre analyses). Despite the different PCs leading to almost identical results, the rest of this article will present analyses based on *exactly* the same PCs as those used in C99. This will avoid any ambiguity in interpretation that might arise from the slight difference in datasets. A sample of 270 monthly mean values were created using data from January 1948 to December 1993.

One can seriously question whether this rather ad hoc filtering and data reduction/projection procedure is optimal for finding regimes in the high-dimensional climate state space. From the arguments presented in section 3(a), one should expect the leading PCs of gridded data to be more normally distributed. Choosing a smaller spatial domain (e.g. sector) for the PCA and reducing the amount of time-averaging may help in the search for regimes. This important issue will be touched upon again in the conclusions.

5. Distribution in state space

(*a*) *Scatter plots*

Before making and presenting estimates of the p.d.f., an important first step is to visualize the scatter of the $n = 270$ data points. Figure 3 shows 'scatter plots' of the two state variables for both the Lorenz system and the height analyses. Two distinct clusters of points can be seen in the Lorenz system (Fig. 3(a)), whereas only one large cluster is easily discernible in the height analyses (Fig. 3(b)). The regimes identified by C99 are marked with letters A, B, C and D in Fig. 3(b), and can each be seen to contain not more than about 20 sample points. The lack of clearly visible clusters and the small samples involved suggest that these clusters may easily be due to sampling variations.

The spatial patterns corresponding to regimes A, B and D have possible counterparts in the previous regime studies of Cheng and Wallace (1993, hereafter CW93) and KG93. CW93 applied a hierarchical clustering technique to ten-day low-pass filtered NH geopotential-height data and identified three clusters: a ridge over the Rockies, a closed anticyclone over the southern tip of Greenland, and a ridge over the Gulf of Alaska. The spatial anomaly patterns associated with these three clusters (CW93, Fig. 4)

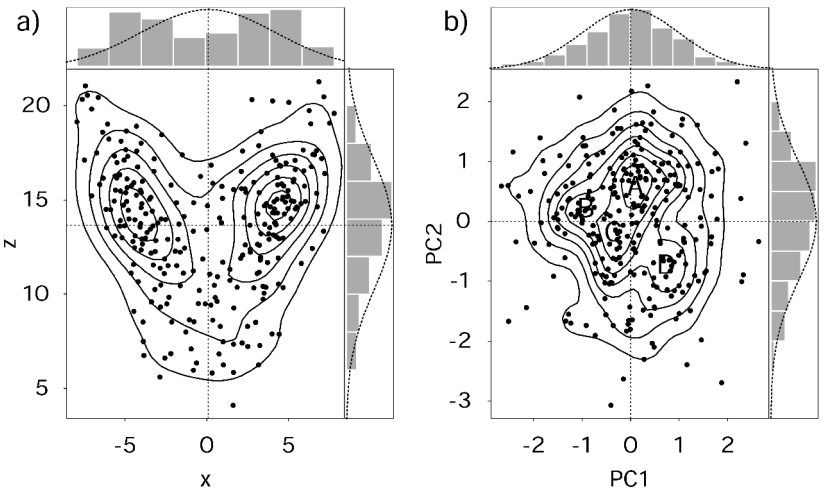

Figure 3. Two-dimensional scatter plots for (a) the Lorenz system variables (x, z), and (b) the 500 hPa geopotential-height analyses. Kernel density estimates are superimposed as contours. Marginal distributions of each variable are shown using histograms that can be compared to fits to the normal distribution (dashed line).

approximately resemble the spatial patterns of clusters A, D and B, respectively, shown in Fig. 3 of C99. Interestingly, CW93 noted that the Alaskan and Greenland clusters coincide with the two primary maxima in the temporal variance and skewness of the 500 hPa height field, and that the Alaskan cluster was the least reproducible. No significance testing was performed by CW93, who pointed out that they were not aware of any specific formal procedure for assessing the statistical significance of the clusters. KG93 searched for regimes using a bump hunting technique based on kernel density estimates of the p.d.f. of the leading two PCs. The discussion of Fig. 10 in KG93 states that the p.d.f. *does not show multiple peaks* but does show some deviations from Gaussianity. KG93 then went on to estimate p.d.f.s of more persistent subsets of the original data and used these to identify four bumps, which they interpreted as the PNA pattern, its reverse, and zonal and blocked phases of the NAO pattern. The fact that regimes could only be found after exploring subsets of the data weakens the statistical significance of the findings. Composite maps of anomalies associated with the reverse PNA, zonal NAO, and blocked NAO (KG93, Figs. 14(b)–(d)) approximately resemble clusters B, A, and D of C99, respectively. The fact that some patterns resemble one another in these different studies based on overlapping non-independent samples of analysis data says more about the consistency of the clustering techniques than about statistical significance of the clusters. Statistical significance is concerned with inference about whether similar clusters will be present not only in the recent post-1950 analyses but also in all other possible *independent* samples of data. It should also be noted that despite resemblances there are also some substantial differences in related patterns identified by these different studies.

(*b*) *Density estimation*

The p.d.f. $f(x, y)$ can be used to quantify the local density of points in state space. It is defined for random pairs (X, Y) as the limit

$$f(x, y) = \lim_{\delta x, \delta y \to 0} \frac{\Pr\{x \le X \le x + \delta x,\ y \le Y \le y + \delta y\}}{\delta x \delta y}, \tag{3}$$

where Pr{.} denotes the 'probability of' a point being in a small region of state space. There are two main approaches to estimating the p.d.f.: 'parametric' and 'non-parametric'. In parametric estimation, one assumes a known functional form for the distribution determined by a few parameters and then one estimates the parameters; for example, one could assume that the distribution is the sum (mixture) of one or more normal distributions. In non-parametric estimation, no assumption is made about the functional form of the probability distribution except that it is a 'smooth' function. Various methods are then used to estimate the smooth function, for example, kernel smoothing or roughness penalty approaches such as smoothing splines and maximum penalized likelihood estimators (Silverman 1994).

The most commonly used non-parametric method for estimating probability distributions of weather and climate has been the kernel smoother (Sutera 1986; Molteni *et al.* 1990; Marshall and Molteni 1993; KG93; C99). Kernel estimates based on the data points have been superimposed as contours in Fig. 3. The kernel estimate is obtained by smoothing the data using

$$f(\mathbf{x}) = \frac{1}{nh^2} \sum_{i=1}^{i=n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right), \tag{4}$$

where the 'kernel' $K(\,)$ is usually an isotropic distribution function such as the bivariate normal distribution, h is the kernel width, and $\mathbf{x} = (x, y)$. Gaussian kernel p.d.f. estimates are superimposed as contours on Figs. 3(a) and (b). The density estimate in Fig. 3(b) shows similar, yet slightly smoother, features to those shown in Fig. 2 of C99. Differences arise due to our use of non-iterative kernel estimation rather than the non-linear iterative estimation used by C99. In order to show multimodal behaviour, a much smaller kernel width of $h = 0.3$ was required than the optimal value $h_{\text{opt}} = 4^{5/6} n^{-1/6} = 1.25$ suggested by Silverman (1994) for bivariate density estimation using $n = 270$ sample points. The kernel width was not specified in C99, but was presumably close to 0.3. Two well-separated maxima are evident in the estimated probability distribution shown in Fig. 3(a), and are associated with the two distinct regimes of this well-known bimodal system. However, the interpretation of the probability distribution of the height analyses shown in Fig. 3(b) is more problematic. The largest probability density (the mode) is found near point A and is offset in the y-direction from the origin (the mean) due to negative skewness in PC2 to be discussed in the next section. A weaker local maximum can be seen at D but is perhaps very poorly sampled due to the low density of points in this region. No other local maxima can be seen in this estimate although there is perhaps some clustering near to points B and C which will be examined in more detail in section 6. C99 used a kernel width 'large enough to detect multimodality with statistical significance', which implies they used the critical smoothing approach described in section 6.3.3 of Silverman (1994). In section 6.3.4, Silverman (1994) points out that 'It may be futile to expect very high power from procedures aimed at such broad hypotheses as unimodality and multimodality': in other words, there is a high risk of not rejecting the multimodal hypothesis even if it is false (type 2 error). Weisheimer *et al.* (2001) illustrate this point nicely in their analysis of 50-year samples taken from a long general-circulation model simulation.

Cox (1966) considered more than one bump in a probability density as a 'descriptive feature likely to indicate mixing of components'. Therefore, a more model-based approach to identifying regimes is to use a parametric method that assumes that the probability distribution is the weighted sum (mixture) of known distributions. The 'normal mixture model' assumes that the probability distribution can be written as a weighted

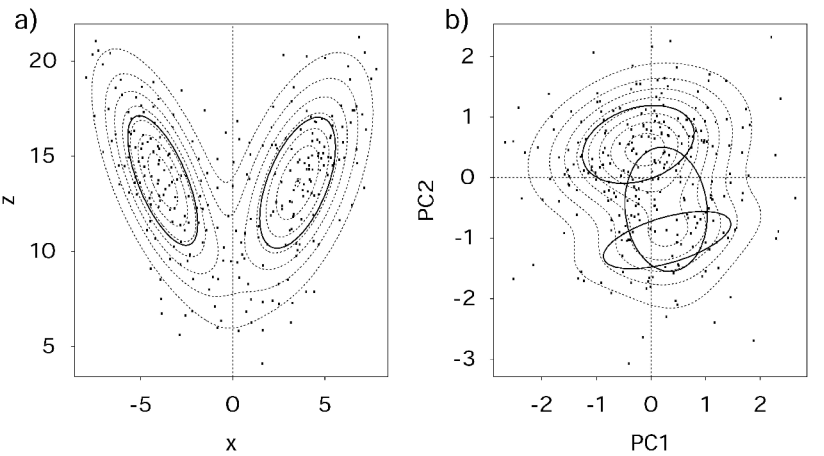

Figure 4. Mixture model fits to (a) the Lorenz system and (b) the 500 hPa geopotential-height analyses. Covariance matrices for each mixture component are depicted by $p = 0.39$ equiprobability contours (solid lines). Density estimates based on the mixture model fits are shown using dashed contours.

sum of k multinormal distributions

$$f(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i N(\mathbf{x}; \boldsymbol{B}_i, \boldsymbol{\Sigma}_i), \tag{5}$$

where $(\alpha_1, \ldots, \alpha_k)$ are the k mixing proportions of the model that satisfy $0 < \alpha_i < 1$ and $\sum_{i=1}^{k} \alpha_i = 1$. Although this model has been widely used in statistics since its introduction by Pearson (1894), it has only recently been applied in climate studies. Haines and Hannachi (1995) first introduced mixture modelling in an attempt to explain Pacific sector regimes as metastable fixed points. The method was also used by Hannachi (1997) to identify regimes having vertical structure. The number of mixture components for hemispheric and sectorial regimes was estimated by Smyth *et al.* (1999) and found to be less than four using a cross-validation procedure. Hannachi and O'Neill (2001) showed that this approach was sample size dependent and proposed an improved resampling procedure.

Figure 4 shows density estimates obtained by fitting a two-component normal mixture model to the Lorenz data and a three-component normal mixture model to the geopotential height data. The fixed points of the Lorenz attractor are clearly identified by this approach (Fig. 4(a)), which gave almost identical mixing proportions, $\alpha_1 = 0.47$ and $\alpha_2 = 0.53$, as expected from symmetry.

For the height analyses, the three-component normal mixture model identified a dominant component ($\alpha_1 = 0.55$) close to regime A of C99, and two less dominant components ($\alpha_2 = 0.33$ and $\alpha_3 = 0.12$) situated around regime D of C99. The components are aligned along the axis between regimes A and D of C99 and are most likely attempting to model the skewness in the PC2 variable rather than any discrete modes. The resulting mixture model p.d.f. in Fig. 4(b) is unimodal despite being the sum of three components. An attempt was made to fit a four-component mixture model to the height data but this led to ill-conditioned estimates due to the small sample size. As the number of mixture components increases much more sample data is required in order to

obtain reliable estimates. Bayesian approaches can be used to help alleviate this problem but would be unlikely to give stable results with the small sample size in this study (Richardson and Green 1997).

(*c*) *The marginal distributions*

A necessary but not sufficient condition for the joint distribution to be multinormal is that the 'marginal' distribution of each variable alone should be normal. The marginal probability distributions,

$$\left.\begin{aligned} F_x(x) &= \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}y, \\ F_y(y) &= \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x, \end{aligned}\right\} \tag{6}$$

can be estimated from histograms of PC1 and PC2, which are shown along the margins of Fig. 3. These can be compared to the normal distributions (dotted lines) that are expected when the joint distribution is bivariate normal $f(\mathbf{x}) = N_2(\mathbf{x}; \boldsymbol{B}, \boldsymbol{\Sigma})$ (the multinormal hypothesis). The x-component of the Lorenz system (Fig. 3(a)) has a bimodal histogram that deviates strongly from normality. However, there is little evidence of strongly non-normal behaviour in the histograms of PC1 and PC2 of the height analyses (Fig. 3(b)).

Rather than compare histograms with continuous distributions, a better way to see deviations from normality is by plotting the empirical quantiles of the standardized variable versus the quantiles of a standard normal distribution. If the distribution is normal then the quantiles should be equal. Such 'quantile–quantile' (or 'q–q') plots are shown in Fig. 5 for the two variables of both the Lorenz system and the height analyses. The q–q plot of x for the Lorenz system (Fig. 5(a)) shows a strong departure from normality. The presence of an S-shaped curve (Fig. 5(a)) is a recognised sign of likely clustering in the data (Everitt and Hand 1981). A Kolmogorov–Smirnov (K–S) test statistic of 0.108 confirms that the non-normality is highly statistically significant ($p < 0.001$). The z variable in Fig. 5(b) has a more linear q–q plot, and is closer to being normally distributed (K–S test statistic $= 0.058$, $p = 0.03$). The q–q plots for the PCs of the height analyses (Figs 5(c) and (d)) are also very linear and these variables are more normally distributed with K–S test statistics of 0.039 ($p = 0.50$) and 0.051 ($p = 0.09$), respectively. To summarize, normality of the marginal distributions can be rejected for the Lorenz data but not for the height data at the 5% level of significance. At the 10% level of significance, normality can also be rejected for PC2 of the height analyses (due to the presence of skewness).

(*d*) *Radial and directional clustering*

Under the hypothesis of multinormality, all linear combinations of the state variables should be normally distributed, yet we only tested two directions in the previous section. To properly test multinormality, more invariant measures need to be used that do not depend on the specific orientation of the axes.

Instead of pairs of standardized Cartesian coordinates (x_i, y_i), it is useful to consider polar coordinates defined by the radial distance from the origin $r_i = \sqrt{x_i^2 + y_i^2}$ and the angle relative to the x-axis $\theta_i = \tan^{-1}(y_i/x_i)$. The r and θ variables can then be tested, respectively, for radial and directional clustering. These types of clustering can

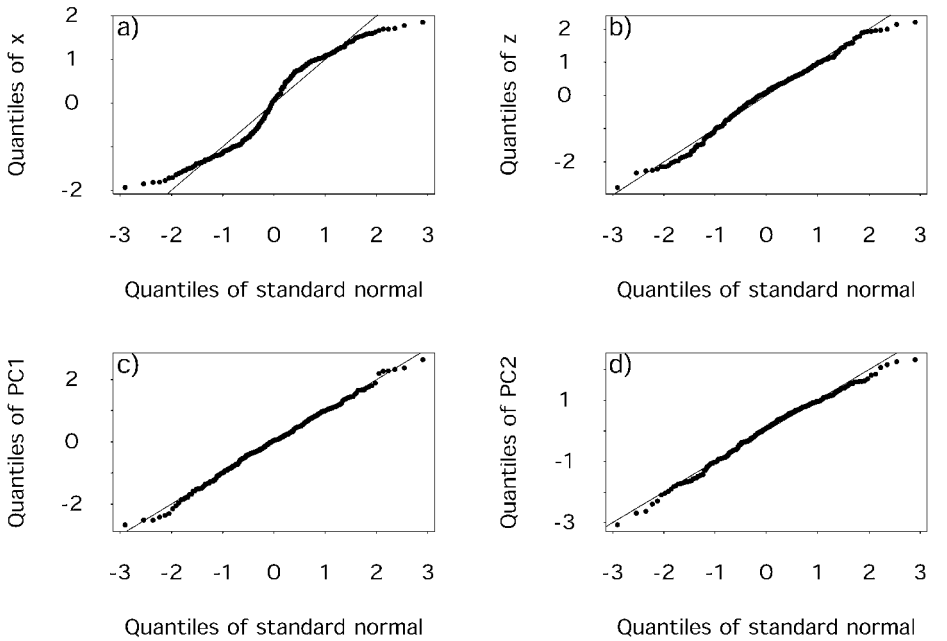

Figure 5. Deviations from normality for the marginal distributions shown by using quantile–quantile plots. Plots show the empirical quantiles versus quantiles of the standard normal distribution for (a) the Lorenz x-variable, (b) the Lorenz z-variable, and (c) PC1 and (d) PC2 of the 500 hPa geopotential-height analyses. Only (a) deviates substantially from the straight line expected for a normal distribution.

easily be missed when testing only a few directions as was done in the previous section. Under the multinormal hypothesis, r^2 should be asymptotically distributed as χ^2 with two degrees of freedom (since it is the sum of two squares of normally distributed variables $x^2 + y^2$), and θ should be uniformally distributed in the range $-\pi$ to π. Figure 6 shows histograms and superimposed theoretical distributions of r and θ for the Lorenz system and the height analyses, respectively.

For the Lorenz system, the radial distribution shown in Fig. 6(a) differs significantly from χ^2 (K–S test statistic $= 0.133$, $p < 0.001$). The directional distribution in Fig. 6(b) is also significantly different from uniform (K–S test statistic $= 0.078$, $p = 0.07$). For the height analyses shown in Figs. 6(c) and (d), both the radial distribution (K–S test statistic $= 0.042$, $p = 0.72$) and the directional distribution (K–S test statistic $= 0.051$, $p = 0.49$) do not differ significantly from the distributions expected under the multinormal hypothesis. To summarize, based on radial and angle variables calculated from standardized x and y variables, multinormality can be rejected for the Lorenz data but not for the height data at the 10% level of significance. At the 5% level of significance, directional uniformity can not be rejected for the standardized Lorenz data. However, this is mainly because of a reduction in power of the uniformity test caused by standardizing the variables.

This more geometric approach can be extended to develop invariant measures of skewness and kurtosis that can be used to assess the normality of multivariate data. Mardia (1980) introduced invariant measures of multivariate skewness, $b_{1,q}$, and

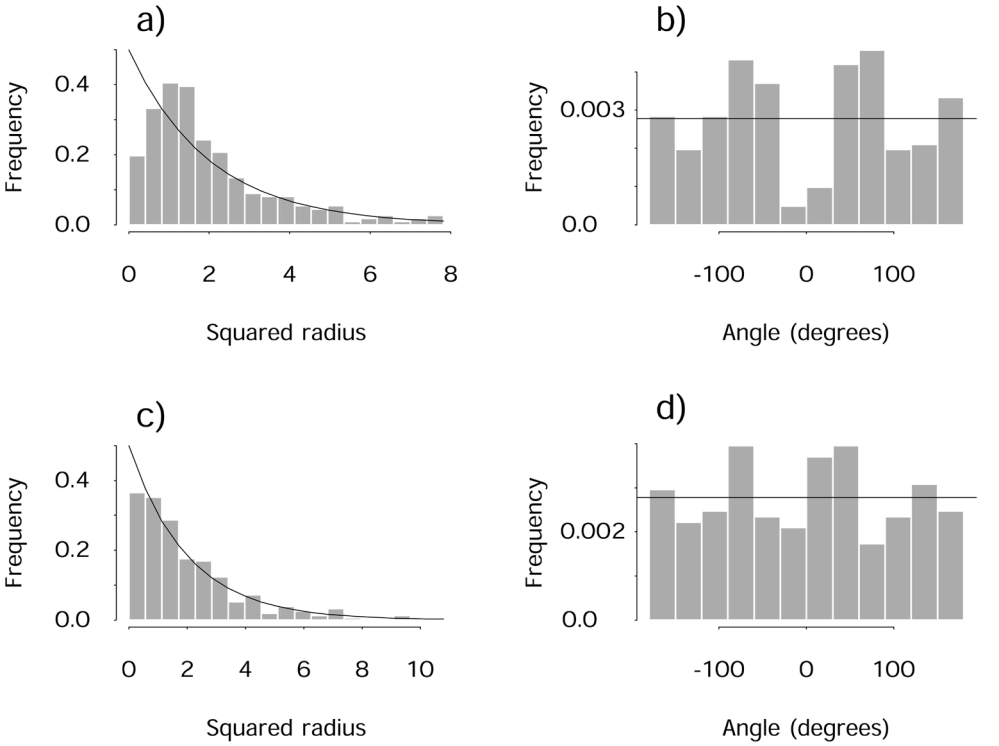

Figure 6. Histograms for the radial and angular coordinates of the two systems: (a) Lorenz radius squared and (b) angle; 500 hPa geopotential height (c) radius squared and (d) angle. The solid lines depict the theoretical χ^2 with two degrees of freedom and uniform distributions expected under the multinormal null hypothesis.

multivariate kurtosis, $b_{2,q}$, defined as:

$$b_{1,q} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}^3, \tag{7}$$

$$b_{2,q} = \frac{1}{n} \sum_{i=1}^{n} r_{ii}^2, \tag{8}$$

where q is the number of variables (dimension of state space) and

$$r_{ij} = (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_j - \overline{\mathbf{x}})$$

is the scalar product between the two vectors from the origin to points i and j in Mahalanobis space. Mahalanobis space is simply the space spanned by the standardized PCs (e.g. PC1 and PC2 of the height analyses)—refer to Stephenson (1997) for a more complete discussion. Under the multinormal hypothesis, $b_{1,q}$ and $b_{2,q}$ are asymptotically distributed as

$$b_{1,q} \sim \frac{6}{n} \chi_\nu^2, \quad \nu = q(q+1)(q+2)/6, \tag{9}$$

$$b_{2,q} - q(q+2) \sim N(0, \sqrt{8q(q+2)/n}) \tag{10}$$

in the limit $n \to \infty$ (Mardia *et al.* 1979).

To test the multinormal hypothesis, we have calculated the multivariate skewness and kurtosis for the Lorenz and 500 hPa height data. At the 5% level of significance, the height analyses were not found to be significantly skewed ($b_{1,2} = 0.19$, $p = 0.07$), whereas the Lorenz data were found to have significant amounts of skewness ($b_{1,2} = 0.84$, $p < 0.001$). This is related to the strong directional anisotropy previously noted in the Lorenz data. The height analyses were also found to have a multivariate kurtosis close to the value of 8 expected for a bivariate normal distribution ($b_{2,2} = 7.63$, $p = 0.45$), whereas the Lorenz data gave a much smaller multivariate kurtosis significantly different from 8 ($b_{2,2} = 6.47$, $p = 0.002$). Therefore, based on these invariant measures, the multinormal hypothesis can be rejected for the Lorenz system but not for the height analyses at the 5% level of significance. At the 10% level of significance, the multinormal hypothesis can also be rejected for the height data due to the presence of skewness (but not kurtosis). This is related to the skewness already previously noted in PC2.

(*e*) *Statistical power of the tests*

Statistical significance tests can fail in two possible ways: type 1 errors (false positives/false alarms) where the null hypothesis is incorrectly rejected when it is true, and type 2 errors (false negatives/misses) where the null hypothesis is incorrectly not rejected when in fact it is false (DeGroot and Schervish 2002). In this study, type 1 errors are caused by rejecting multinormality when the data are multinormal, whereas type 2 errors are caused by the failure to reject multinormality when the data are not multinormal. The probability of making a type 1 error when the null hypothesis is true (i.e. the false alarm rate) is determined in advance by the person making the test and is known as the *level of significance* (e.g. $\alpha = 5\%$, 1%, or even 0.1% for situations where one really wants to avoid making a type 1 error, for example, in clinical drug trials). The probability β of making a type 2 error given the null hypothesis is false is determined by the choice of α, the sample size, and the type of test. The quantity $1 - \beta$, the *power* of the test, provides a convenient way of summarizing how well the test rejects the null hypothesis when it is false (i.e. the probability of detection). The power decreases for smaller levels of significance because of there being fewer rejections of the null hypothesis—in other words, there is a trade-off between the amount of type 1 and type 2 errors. Data that are closest to satisfying the null hypothesis generally give the least power and so power is generally never less than the level of significance.

To address this important issue, we have used Monte Carlo simulations to generate 1000 samples of $n = 270$ pairs of (x, y) non-normally distributed data. The previous tests used in this study have then been applied to the 1000 samples to calculate the number of rejections. The fraction of rejections (in %) obtained are given in Table 2 for three different levels of significance. Two types of non-normal data were generated: *strongly bimodal* data generated by a two-component normal mixture model fit to the Lorenz data shown earlier, and *weakly bimodal* generated by a three-component mixture model fit to the 500 hPa geopotential-height data. The two-component fit to the Lorenz model has two clearly well-separated bumps (see section 5(b)) and so should easily be discriminated from multinormality. A more subtle case of non-normality has also been tested based on the three-component normal mixture model fit to the height data PCs. Despite having three components and being non-normal, this fit yields a unimodal p.d.f. (see Fig. 4(b), section 5(b)). However, by varying the weights of these mixture components, it was possible to generate p.d.f.s that were slightly bimodal having bumps situated near to regimes A and D of C99. Because of the large widths of the components compared to their separation, no weights were found that could produce

TABLE 2. POWER OF THE STATISTICAL TESTS USED IN THIS STUDY FOR DIFFERENT LEVELS OF SIGNIFICANCE

Statistic	1% level		5% level		10% level	
x	>99%	(3%)	>99%	(10%)	>99%	(17%)
y	1%	(45%)	5%	(71%)	10%	(80%)
r	>99%	(1%)	>99%	(4%)	>99%	(9%)
θ	>99%	(4%)	>99%	(11%)	>99%	(23%)
$b_{1,2}$	>99%	(34%)	>99%	(59%)	>99%	(74%)
$b_{2,2}$	23%	(2%)	53%	(7%)	68%	(11%)

Power is the probability of the test correctly rejecting the multinormal null hypothesis when it is not true. The power (in %) was calculated by applying the different tests to 1000 Monte Carlo samples of $n = 270$ points simulated by mixture model fits to the Lorenz and geopotential-height data. Numbers not in parentheses are for a two-component fit to the bimodal Lorenz data, whereas numbers in parentheses were generated using a weakly bimodal three-component mixture model fit to the 500 hPa geopotential-height data (see text for details).

trimodal behaviour in the p.d.f. By shrinking the estimated mixture model weights from (0.55, 0.33, 0.12) to (0.33, 0.33, 0.33), bimodality was first found to appear for weights (0.54, 0.33, 0.13). This critical weakly bimodal mixture model was then used to generate the Monte Carlo results shown in parentheses in Table 2. This three-component normal model is similar to the three-well potential system discussed by Hasselmann (1999).

From Table 2, it can be noted that most of the tests have very good power for the strongly bimodal case, i.e. multinormality can be safely rejected for this data. The normality test on the y-component has low power due to the Lorenz data being unimodal in this direction. The multinormal kurtosis test based on $b_{2,2}$ has lower power than the other tests (due to it being based on higher moments) but it still has good power at the 10% level. As to be expected, the power of the tests are much less for the weakly bimodal case (numbers in parentheses in Table 2). This is not a failure of our tests but rather a clear indication that the underlying data on which the mixture model was fitted is very close to multinormality. Large powers are obtained for the normality tests based on y and $b_{1,2}$ because of the presence of skewness in the y-direction in the weakly bimodal (and NCEP) data. The low power for the weakly bimodal case indicates that it is difficult to discriminate between the multinormal and weakly bimodal hypotheses when one has only 270 data points. High power is not to be expected when statistically testing broad hypotheses such as multimodality (Silverman 1994). In such cases, one should logically adopt the most parsimonious (simplest) hypothesis that can explain the data (e.g. the multinormal hypothesis). This principle of parsimony is also known as Occam's razor after the medieval philosopher, William of Occam (or Ockham), who stated *plurality should not be assumed without necessity* *—a rather appropriate remark for multiple-regime studies.

6. IS THERE SIGNIFICANT CLUSTERING IN STATE SPACE?

Because of physical constraints, the scatter of points in state space is finite and clustered around the mean—the attractor is bounded as can be noted for example in Fig. 5. This non-homogeneity can make it difficult to detect other local clusters within this cloud of points especially if they do not lie near the centre or edges of the

* Pluralitas non est ponenda sine necessitate.

distribution. Fortunately, this problem can be effectively alleviated by transforming to empirical probabilities that remove information about the marginal distributions.

(*a*) *Probability space*

Rather than use pairs of variables (x_i, y_i), consider transforming to pairs of probabilities (u_i, v_i) defined as

$$u_i = F_x(x_i) = \int_{-\infty}^{x_i} f_x(x)\,\mathrm{d}x,$$
$$v_i = F_y(y_i) = \int_{-\infty}^{y_i} f_y(y)\,\mathrm{d}y.$$

F_x and F_y are the cumulative distribution functions that can be most easily estimated using $F(x_i) = \mathrm{rank}(x_i)/(n+1)$, where rank is the position of x_i once all the x_i are arranged in ascending order. The transformed coordinates (u, v) lie inside the unit square $0 \le u, v \le 1$ and define 'probability space' (Nelsen 1999). Furthermore, it is straightforward to show that the new probability distribution in probability space, $g(u, v)$, is related to the probability distribution of the original variables, $f(x, y)$, by the simple relationship

$$g(u, v) = \frac{f(x, y)}{f_x(x) f_y(y)}. \tag{11}$$

The marginal distributions of $g(u, v)$ are uniform. In other words, the transformation to probabilities factors out the effect of the marginal distributions. Hence, when x and y are independently distributed, the resulting probability distribution $g(u, v)$ is completely uniform (and equal to one). So, under the multinormal hypothesis, one expects that the scatter of points in probability space of the PCs should be completely uniform. This makes searching for clusters and testing their significance much simpler as will be demonstrated in the following section.

Figure 7 shows the scatter of points and probability distribution in probability space for the Lorenz system and the height analyses. The scatter of points are no longer clustered around the origin—the marginal histograms are perfectly uniform as to be expected from the preceding arguments. This makes it easier to identify clusters which are much better separated than in Fig. 3. While clustering can clearly be seen in Fig. 7(a) for the Lorenz system, the scatter of points for the height analyses in Fig. 7(b) appears much more uniform. In the next section, a spatial point process technique will be used to assess the amount of clustering in both scatter plots.

The probability distribution in probability space is related to a very elegant concept in statistics known as the 'copula' (Nelsen 1999). Sklar (1959) proved that the joint cumulative distribution function $F(x, y)$ for any pair of random variables can always be written as

$$F(x, y) = C(F_x(x), F_y(y)), \tag{12}$$

where $C(u, v)$ is the 'copula' function that links the joint distribution function to the marginal distributions. The copula gives information about the dependency between the two variables without caring about their individual marginal distributions. The probability distribution in probability space is simply the probability density of the copula:

$$g(u, v) = \frac{\partial^2 C}{\partial u \partial v}. \tag{13}$$

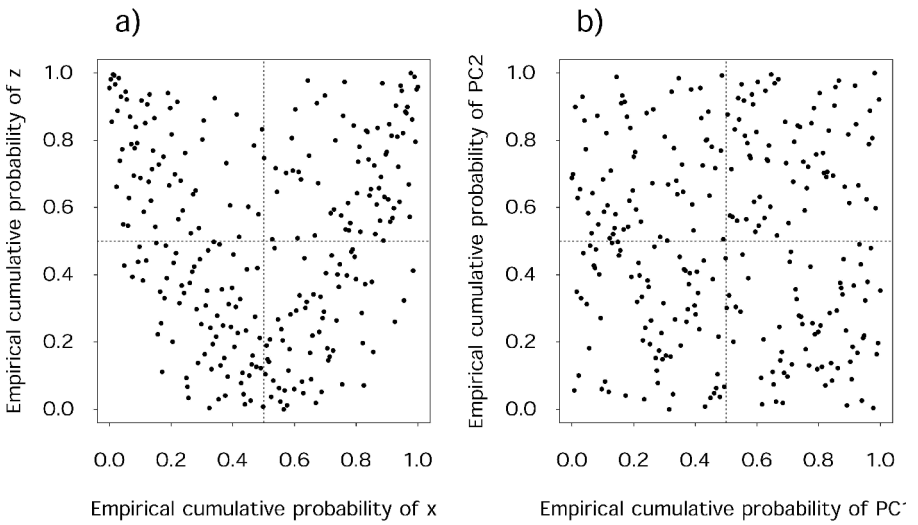

Figure 7. Scatter plots of the cumulative probabilities of the variables for (a) the Lorenz system and (b) the 500 hPa geopotential-height analyses. Note inhomogeneity evident in (a) but not in (b).

This non-parametric framework is completely general and can easily be extended to more than two variables.

In order to see the density maxima more clearly, Hsu and Zwiers (2001) used an alternative approach that involved subtracting (rather than factoring) out the product of the marginal densities from the estimated probability distribution. This less robust parametric procedure involving smoothing does not result in a simple uniform distribution under the multinormal hypothesis and so is less suitable for testing inferences about clustering. It also has the drawback noted by Hsu and Zwiers (2001) that it is not robust to skewness in the distribution: a 'regime' could be falsely identified when a unimodal skewed distribution is subtracted from the multinormal null distribution. Finally, it also has the inconvenience of giving much smaller anomalies at the edges of the distribution where there may also be significant clusters of interest.

7. Clustering in probability space

For independent variables, such as PCs under the multinormal hypothesis, points in probability space are expected to be uniformly distributed. Hence, by testing for deviations from uniformity (i.e. clustering), it is possible to test the multinormal hypothesis. Clustering of points in probability space can be tested using standard techniques developed for spatial point processes (Diggle 1983). Testing directly for clustering of points has the great advantage that it is based directly on the data points and completely avoids the smoothing and dependency problems encountered in bump hunting based on density estimation (Good and Gaskins 1980; Silverman 1981, 1994).

Ripley (1976) developed a simple technique for testing for the presence of spatial clusters of points based on the distance between pairs of points. The mean number of points that are within distance d of a target point can be written as

$$n(d) = K(d)\overline{f}, \tag{14}$$

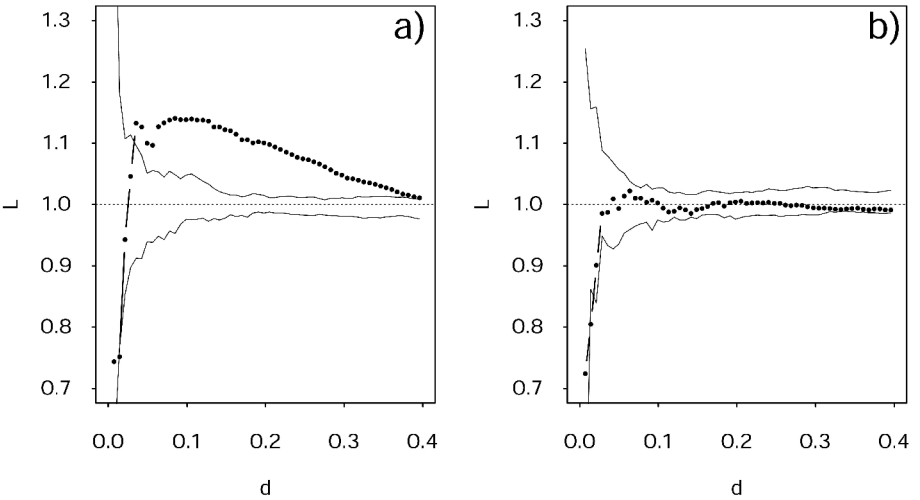

Figure 8. The clustering statistic $L(d)$ (see text for details) as a function of the inter-point distance d for (a) the Lorenz system and (b) the 500 hPa geopotential heights. The lines denote 95% confidence bounds for no clustering estimated by generating many random configurations of unclustered points.

where $\overline{f}$ is the mean density of points and $K(d)$ is often referred to as Ripley's K-function. For a perfectly uniform distribution of points, $K(d)$ is simply the area πd^2 enclosed by a circle of radius d about the target point. Deviation from this relationship can be used to detect the presence of clustering. Note that this and most other rigorous clustering approaches require extra information about second-order moments (point-to-point distances) in addition to just the first-order density estimates used in density bump hunting approaches.

Figure 8 shows plots of the clustering index

$$L(d) = \sqrt{\frac{K(d)}{\pi d^2}} \tag{15}$$

as a function of inter-point distance. The 95% confidence limits were generated by making Monte Carlo simulations of homogeneous spatial point process with the same total number of points. Note that there are significant deviations from no clustering ($L = 1$) for the case of the Lorenz system (Fig. 8(a)) but not for the reanalysis PCs (Fig. 8(b)). The same conclusions would also hold at the 10% level of significance based on 90% confidence intervals having 0.84 (1.64/1.96) the width of those shown in Fig. 8. This reconfirms the previous findings in this study that the multinormal hypothesis can not be rejected for this sample of reanalysis data. It also suggests that the clusters seen in Fig. 4(b) associated with the modes discussed in C99 are artefacts of sampling.

(*a*) *Concluding remarks*

This study has identified three distinct hypotheses concerning the probability distribution of the climate system: multinormal, unimodal and non-normal, and multimodal. The simplest null hypothesis, expected for large-scale average indices of many weakly interacting local weather variables, is that the probability distribution is multinormal. This hypothesis has been tested using data generated by a simple low-order chaos model

and NH 500 hPa geopotential gridded analyses 1949–94. While the hypothesis is easily rejected for the chaos model data, it can not be rejected for the geopotential-height data used in C99 at the 5% level of significance. Furthermore, a spatial point process clustering technique that does not involve density estimation/smoothing confirms that there is no significant clustering in the data (at the 5% level of significance). At the less stringent 10% level of significance, the geopotential-height data differs slightly from normality due to the presence of skewness in PC2 but there is still no statistically significant evidence of multiple clustering. Using different analysis techniques, Hsu and Zwiers (2001) also concluded that 'not all the regimes identified by Corti *et al.* (1999) are distinguishable from those that result from sampling variability'. However, Hsu and Zwiers (2001) did identify the cold-ocean warm-land pattern (regime A) as a regime using their AR(1) approach—although this could have arisen due to a weakness of their approach in the presence of skewness. Our more robust non-parametric methods help show that although there is some slight skewness (mainly in PC2) there is no real evidence of significant multiple clustering.

Although no convincing evidence of statistically significant multiple regimes has been found in this study of monthly mean data, regimes may exist when using other variables and pre-processing techniques. It is questionable whether the filtering and data projection procedures used in this study and C99 are optimal for finding regimes. The averaging involved in using monthly means and leading PCs of hemispheric datasets is likely to reduce any non-normality present in the original data. The existence of multiple weather regimes in daily 500 hPa geopotential height over the Euro-Atlantic region has also recently been tested using our methods in an MSc project by Ivar Seierstad at the University of Bergen (Seierstad 2002). Despite having a much larger sample of data, Seierstad (2002) found very similar conclusions to those reported here for monthly mean data. In order to find regimes, it might be better to use pattern extraction techniques applied to daily data in a smaller regions such as the North Atlantic blocking region. However, such data mining techniques can seriously reduce the significance of any regimes that are eventually found. It is safer in general to have strong prior physical arguments as to the location and number of regimes before searching. This was the case in the early searches for bimodality in stationary waves inspired by the Charney–DeVore model (Sutera 1986; Hansen and Sutera 1986).

Although we have found insufficient evidence to reject multinormality in favour of multimodality in this particular study of NH geopotential height from 1948–93, it is possible that the hypothesis may be rejected when more data become available in the future (or in long general-circulation model simulations—see Weisheimer *et al.* 2001). Nitsche *et al.* (1994) estimated that at least 150 years of data would be required in order to find any significant evidence of multimodality. High power is not to be expected when statistically testing broad hypotheses such as multimodality (Silverman 1994). This is one of the reasons that led us to test the simplest multinormal hypothesis rather than the more complicated multimodal hypothesis.

Although the non-existence of multiple hemispheric regimes might appear to be a negative result, it is in fact a very useful positive result. The inability to reject the multinormal hypothesis suggests that this simple single-regime hypothesis may be the most appropriate model for describing the NH flow. In other words, we can reasonably assume that the leading PCs of the 500 hPa geopotential-height field are independent and close to being normally distributed. This allows us to make useful inferences and predictions about the probability of different atmospheric states. It also provides a simple probability model that can be used in forecasting, climate change detection, and risk assessment studies.

ACKNOWLEDGEMENTS

This work was presented on 26 September 2000 at the 'Conference on nonlinear phenomena in global climate dynamics', International Centre for Theoretical Physics, Trieste. We would like to thank Susanna Corti for providing the PC data used in this article and Tim Palmer and Franco Molteni for stimulating discussions about various aspects of this work. In addition, we would like to thank Ivar Seierstad for his careful reading and comments on the manuscript. AH has been financially supported throughout this study by the United Kingdom Universities Global Atmospheric Modelling Programme (http://ugamp.nerc.ac.uk).

REFERENCES

Ambaum, M. H. P., Hoskins, B. J. and Stephenson, D. B. 2001 Arctic Oscillation or North Atlantic Oscillation? *J. Climate*, **14,** 3495–3507

2002 Corrigendum: Arctic Oscillation or North Atlantic Oscillation? *J. Climate*, **15,** 553

Aurell, E., Boffetta, G., Crisanti, A., Frick, P., Paladin, G. and Vulpiani, A. 1994 Statistical mechanics of shell models for two-dimensional turbulence. *Phys. Rev. E*, **50,** 4705–4715

Bauer, F. 1951 'Extended range weather forecasting'. Pp. 814–833 in *Compendium of meteorology*. American Meteorological Society, 45 Beacon Street, Boston MA02108-3693, USA

Bjerknes, J. 1969 Atmospheric teleconnections from the equatorial Pacific. *Mon. Weather Rev.*, **97,** 163–172

Branstator, G. and Opsteegh, J. D. 1989 Free solution of the barotropic vorticity equation. *J. Atmos. Sci.*, **46,** 1799–1814

Burgers, G. and Stephenson, D. B. 1999 The 'normality' of El Niño. *Geophys. Res. Lett.*, **26,** 1027–1030

Charney, J. G. and DeVore, J. G. 1979 Multiple flow equilibria in the atmosphere and blocking. *J. Atmos. Sci.*, **36,** 1205–1216

Cheng, X. and Wallace, J. M. 1993 Cluster analysis of the northern hemisphere wintertime 500 hPa height field: Spatial patterns. *J. Atmos. Sci.*, **50,** 2674–2696

Corti, S., Molteni, F. and Palmer, T. N. 1999 Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature*, **398,** 799–802

Cox, D. 1966 Notes on the analysis of mixed frequency distributions. *Brit. J. Math. Statist. Psychol.*, **19,** 39–47

DeGroot, M. J. and Schervish, M. J. 2002 *Probability and statistics*. Addison-Wesley, Boston, USA

Diggle, P. J. 1983 *Statistical analysis of spatial point patterns*. Academic Press, London

Ditlevesen, P. D. and Mogensen, I. A. 1996 Cascades and statistical equilibrium in shell models of turbulence. *Phys. Rev. E*, **53,** 4785–4793

Dole, R. M. and Gordon, D. N. 1983 Persistent anomalies of the extratropical northern hemisphere wintertime circulation: Geophysical distribution and regional persistent characteristics. *Mon. Weather Rev.*, **111,** 1567–1586

Egolf, D. A. 2000 Equilibrium regained: From nonequilibrium chaos to statistical mechanics. *Science*, **287,** 1997–2002

Everitt, B. S. and Hand, D. J. 1981 *Finite mixture distributions.* Chapmann and Hall, London

Good, I. J. and Gaskins, R. A. 1980 Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Am. Statist. Assoc.*, **75,** 42–73

Haines, K. and Hannachi, A. 1995 Weather regimes in the Pacific from a GCM. *J. Atmos. Sci.*, **52,** 2444–2462

Hannachi, A. 1997 Low-frequency variability in a GCM: Three-dimensional flow regimes and their dynamics. *J. Climate*, **10,** 1357–1379

Hannachi, A. and Haines, K. 1998 Convergence of data assimilation by periodic updating in simple Hamiltonian and dissipative systems. *Tellus A*, **50,** 58–75

Hannachi, A. and O'Neill, A. 2001 Atmospheric multiple equilibria and non-Gaussian behaviour in model simulations. *Q. J. R. Meteorol. Soc.*, **127,** 939–958

Hannachi, A., Stephenson, D. B. and Sperber, K. 2003 Probability-based methods for quantifying nonlinearity in ENSO. *Clim. Dyn.*, **20,** 241–256

Hansen, A. R. and Sutera, A. 1986 On the probability density distribution of large-scale atmospheric wave amplitude. *J. Atmos. Sci.*, **43,** 3250–3265

Hasselmann, K. 1999 Climate change—Linear and nonlinear signatures. *Nature*, **398 (6730),** 755–756

Horel, J. D. 1985 Persistence of 500 mb height field during northern hemisphere winter. *Mon. Weather Rev.*, **113,** 2030–2042

Hoskins, B. J. and Karoly, D. J. 1981 The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.*, **38,** 1179–1196

Hsu, C. J. and Zwiers, F. 2001 Climate change in recurrent regimes and modes of atmospheric variability. *J. Geophys. Res.*, **106 (D17),** 20145–20160

Kimoto, M. and Ghil, M. 1993 Multiple flow regimes in the northern hemisphere winter. Part I: Methodology and hemispheric regimes. *J. Atmos. Sci.*, **50,** 2625–2643

Legras, B. and Ghil, M. 1985 Persistent anomalies, blocking and variations in atmospheric predictability. *J. Atmos. Sci.*, **42,** 433–471

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

1965 A study of the predictability of a 28-variable atmospheric model. *Tellus*, **3,** 321–333

1970 Climate change as a mathematical problem. *J. Appl. Meteorol.*, **9,** 325–329

Mardia, K. V. 1980 'Tests of univariate and multivariate normality'. Pp. 279–320 in *Handbook of statistics 1: Analysis of variance.* Ed. P. R. Krishnaiah. North-Holland Publishing company

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

Marshall, J. C. and Molteni, F. 1993 Towards a dynamical understanding of planetary-scale flow regimes. *J. Atmos. Sci.,* **50,** 1792–1818

Miller, R. N., Ghil, M. and Gauthiez, F. 1994 Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.*, **51,** 1037–1056

Mo, K. and Ghil, M. 1988 Cluster analysis of multiple planetary flow regimes. *J. Geophys. Res.*, **93,** 10927–10952

Molteni, F., Sutera, A. and Tronci, N. 1988 The EOFs of the geopotential eddies at 500 mb in winter and their probability density function. *J. Atmos. Sci.*, **45,** 3063–3080

Molteni, F., Tibaldi, S. and Palmer, T. N. 1990 Regimes in the wintertime circulation over northern extratropics. I: Observational evidence. *Q. J. R. Meteorol. Soc.*, **116,** 31–67

Monahan, A. H., Fyfe, J. C. and Flato, G. M. 2000 A regime view of northern hemisphere atmospheric variability and change under global warming. *Geophys. Res. Lett.*, **27,** 1139–1142

Monahan, A. H., Pandolfo, L. and Fyfe, J. C. 2001 The preferred structure of variability of the northern hemisphere atmospheric circulation. *Geophys. Res. Lett.*, **28,** 1019–1022

Mukougawa, H. 1988 A dynamical model of 'quasi-stationary' states in large-scale atmospheric motions. *J. Atmos. Sci.*, **45,** 2868–2888

Namias, J. 1950 The index cycle and its role in the general circulation. *J. Meteorol.*, **7,** 130–139

1964 Seasonal persistence and recurrence of European blocking during 1958–1960. *Tellus*, **16,** 394–407

Nelsen, B. R. 1999 *An introduction to copulas.* Springer, New York

Nicholls, N. 2001 The insignificance of significance testing. *Bull. Am. Meteorol. Soc.*, **81,** 981–986

Nitsche, G., Wallace, J. M. and Kooperberg, C. 1994 Is there evidence of multiple equilibria in planetary wave amplitude statistics? *J. Atmos. Sci.*, **51,** 314–322

Palmer, T. N. 1993 Extended-range atmospheric prediction and the Lorenz model. *Bull. Am. Meteorol. Soc.*, **74,** 49–65

1999 A nonlinear dynamical perspective on climate prediction. *J. Climate,* **12,** 575–591

Pearson, K. 1894 Contribution to the mathematical theory of evolution. *Phil. Trans. A*, **185,** 71–110

Rex, D. 1950 Blocking action in the middle troposphere and its effect upon regional climate. *Tellus,* **2,** 196–211

Richardson, S. and Green, P. J. 1997 On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B*, **59,** 731–792

Ripley, B. D. 1976 The second-order analysis of stationary point processes. *J. Appl. Probability*, **13,** 255–266

Seierstad, I. A. 2002 'Weather regimes in the Euro-Atlantic sector: Do they exist?' MSc dissertation, Geofyysisk Institutt, University of Bergen

Silverman, B. W. 1981 Using kernel density estimates to investigate multimodality. *J. R. Statistical Soc. B*, **43,** 97–99

Silverman, B. W. 1994 *Density estimation for statistics and data analysis*. Chapman and Hall, London

Simmons, A. J., Wallace, J. M. and Branstator, G. W. 1983 Barotropic wave propagation and instability, and atmospheric teleconnection patterns. *J. Atmos. Sci.*, **40,** 1363–1392

Sklar, A. 1959 Fonction de répartition à *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8,** 229–231

Smyth, P., Ide, K. and Ghil, M. 1999 Multiple regimes in northern hemisphere height fields via mixture model clustering. *J. Atmos. Sci.*, **56,** 3704–3723

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

Stephenson, D. B. and Doblas-Reyes, J. 2000 Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus A*, **52,** 300–322

Sutera, A. 1986 Probability density distribution of large scale atmospheric flow. *Adv. Geophys.*, **29,** 227–249

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

Vautard, R. 1990 Multiple weather regimes over the North Atlantic: Analysis of precursors and successors. *Mon. Weather Rev.*, **118,** 2056–2081

Wallace, J. M. 2000 North Atlantic oscillation/annular mode: Two paradigms—one phenomenon. *Q. J. R. Meteorol. Soc.*, **126,** 791–805

Wallace, J. M. and Thompson, D. W. J. 2002 The Pacific center of action of the northern hemisphere annular mode: Real or artifact? *J. Climate*, **15,** 1987–1991

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

Weisheimer, A., Handorf, D. and Dethloff, K. 2001 On the structure and variability of atmospheric general circulation regimes in coupled climate models. *Atmos. Sci. Lett.*, doi:10.1006/asle.2001.0034

White, G. H. 1980 Skewness, kurtosis, and extreme values of northern hemisphere geopotential height. *Mon. Weather Rev.*, **108,** 1446–1455

Wiin-Nielsen, A. 1979 Steady states and stability properties of a low-order barotropic system with forcing and dissipation. *Tellus*, **31,** 375–386