Multiple Regimes in Northern Hemisphere Height Fields via Mixture Model Clustering

PADHRAIC SMYTH*
Department of Information and Computer Science
University of California, Irvine, CA 92697-3425

MICHAEL GHIL† AND KAYO IDE
Department of Atmospheric Sciences
and Institute of Geophysics and Planetary Physics
University of California, Los Angeles, CA 90095-1565

February 20, 1998

Technical Report UCI-ICS 98–08
Information and Computer Science
University of California, Irvine

*Also with the Jet Propulsion Laboratory 525-3660, California Institute of Technology, Pasadena, CA 91109.

†Corresponding author address: Dr. M. Ghil, Department of Atmospheric Sciences, UCLA, Los Angeles, CA 90095-1565; phone: (310)206-0651; fax: (310)206-5219; e-mail: ghil@atmos.ucla.edu

Abstract

Mixture model clustering is applied to Northern Hemisphere (NH) 700-mb geopotential height anomalies. A mixture model is a flexible probability density estimation technique, consisting of a linear combination of k component densities. A key feature of the mixture modeling approach to clustering is the ability to estimate a posterior probability distribution for k, the number of clusters, given the data and the model, and thus objectively determine the number of clusters that is most likely to fit the data.

A data set of 44 winters of NH 700-mb fields is projected onto its two leading empirical orthogonal functions (EOFs) and analyzed using mixtures of Gaussian components. Cross-validated likelihood is used to determine the best value of k, the number of clusters. The posterior probability so determined peaks at $k = 3$ and thus yields clear evidence for 3 clusters in the NH 700-mb data. The 3-cluster result is found to be robust with respect to variations in data preprocessing and data analysis parameters. The spatial patterns of the 3 clusters' centroids bear a high degree of qualitative similarity to the 3 clusters obtained independently by X. Cheng and J. M. Wallace, using hierarchical clustering on 500-mb NH winter data: **A** for Gulf-of-**A**laska ridge, **G** for high over southern **G**reenland, and **R** for enhanced climatological ridge over the **R**ockies.

Separating the 700-mb data into Pacific (PAC) and Atlantic (ATL) sector maps reveals that the optimal k-value is 2 for both the PAC and ATL sectors. The respective clusters consist of M. Kimoto and M. Ghil's Pacific/North-American (PNA) and reverse PNA (RNA) regimes, as well as the zonal (ZNAO) and blocked (BNAO) phases of the North Atlantic Oscillation (NAO). The connections between our sectorial and hemispheric results are discussed from the perspective of large-scale atmospheric dynamics.

Contents

1 Introduction and motivation

Reliable identification of multiple regimes in hemispheric circulation patterns is a problem that has attracted considerable interest in studies of atmospheric low-frequency variability. We revisit here the specific problem of determining whether or not regime-like behavior can be identified from estimates of the probability density function (PDF) in the large-scale atmospheric flow's phase space. In particular, we use mixture modeling techniques to perform probabilistic clustering in the space spanned by the leading empirical orthogonal functions (EOFs) of the data. A data set comprised of 44 winters of Northern Hemisphere (NH) 700-mb geopotential height anomalies is used in the present study.

Early work on this problem (Rex 1950a, b; Namias 1982) was based on fairly subjective criteria, using synoptic pattern recognition or *ad hoc* quantitative criteria. More recent work used increasingly objective and sophisticated criteria for clustering (Dole and Gordon 1983; Benzi et al. 1986; Ghil 1987; Mo and Ghil 1988; Molteni et al. 1988; Vautard 1990; Hannachi and Legras 1995). There are essentially three issues involved: (i) is the total number of clusters k equal to 1, 2 or more; (ii) if $k \geq 2$, can we describe, stably and reliably, the multiple clusters; and (iii) having done so, what are the dynamical mechanisms giving rise to the clusters so described? The purpose of the present paper is to address issues (i) and (ii).

Within the context of the first issue, Michelangeli et al. (1995; MVL hereafter) have addressed specifically the problem of finding an objective criterion to determine the number k of clusters. They used the framework of the dynamic cluster method (Diday and Simon 1976), which is a variant of the well-known k-means clustering algorithm, and 44 winters (1949–1992) of 700-mb height maps, classified separately over the Atlantic (ATL) and Pacific (PAC) sector. They proposed the use of a classifiability index which measures the "stability" of the cluster solution, as a function of k, across different initial data for the algorithm. Such an index does provide some idea of the cluster structure in the data; still, this technique and related approaches, such as using the Davies and Bouldin (1979) index, may not perform well in the presence of strongly overlapping clusters (e.g., Jain and Dubes 1988, Fig. 4.13; Edlund 1997). Furthermore, there is no general theory supporting the use of one particular "stability" index over any other.

As for the second issue, the closest degree of reproducibility so far of the same (subset of) clusters by two independent methods applied to distinct data sets was obtained by Cheng and Wallace (1993; CW hereafter), who applied hierarchical clustering (see also Legras et al. 1988) to 40 NH winters (1946–1985) of 500-mb height data, and Kimoto and Ghil (1993a, b), who applied visual inspection (Kimoto and Ghil 1993a; KGI hereafter) and "bump hunting" (Kimoto and Ghil, 1993b; KGII hereafter) to the estimated PDF for 37 winters (1949–1986) of 700-mb heights.

Even though greater reliability and reproducibility were achieved in the recent work just reviewed, there is still a degree of subjectivity left in the application of these clustering techniques. In particular, none of the methods above have a completely satisfactory solution to the problem of determining in an algorithmic manner how many clusters exist in a given data set, hemispheric or sectorial. Thus, the two problems of just how many different regimes can be reliably identified in the multidecadal NH 700-mb record, and what exactly they look like, bears further investigation.

The mixture model approach adopted here, unlike the previously used approaches, is

based on an explicit, fully consistent probabilistic model. This model has two primary features:

1. Each cluster is defined as a unimodal ("component") PDF. Thus, points which lie within the overlap region of different density functions can have a degree of membership (a probability) for each cluster, allowing for uncertainty in cluster membership to be handled in a natural way.

2. It leads to a well-defined, built-in criterion for determining how many clusters should be fitted to the data, which does not require additional, *ad hoc* assumptions or null hypotheses. The information on which this criterion is based is simply contained in the *posterior probability distribution* for k, the number of clusters. If the distribution peaks sharply about a particular value of k, there is strong evidence for that value; if the distribution is rather flat, it follows that the data set at hand cannot provide enough evidence for a particular value of k. The difficult part of the problem is that of *estimating* this posterior distribution for k given the data. We discuss the methodology for doing so in some detail.

The paper is organized as follows. In Section 2 the 700-mb data set and data preprocessing steps are briefly described. Section 3 is an introduction to and review of the basic concepts of mixture models, including a discussion of maximum-likelihood techniques for model parameter estimation and a cross-validation methodology for estimating the posterior distribution of k. Further methodological details are presented in three appendices.

Section 4 contains a detailed description of the application of the mixture modeling methodology to the problem of cluster analysis in the subspace of the NH 700-mb anomalies' leading EOFs. Strong evidence for the data's supporting the existence of 3 regimes is presented. Robustness of this result with respect to variations in cross-validated partitions and number of EOFs retained is investigated and discussed. The maps corresponding to the 3 clusters found by mixture modeling are compared to the 3 significant maps found by CW and a remarkable degree of similarity is found to exist.

The application of the mixture modeling methodology to the PAC and ATL sectors is described in Section 5 and results in the selection of 2 clusters in each sector. The PAC clusters resemble the well-known pacific/North-American (PNA) and reverse PNA (RNA) regimes and the ATL clusters resemble the well-known blocked and zonal phases of the North-Atlantic Oscillation (NAO). Both hemispheric and sectorial results are summarized in Section 6, followed by a discussion of their implications for the understanding and prediction of low-frequency, intraseasonal variability of large-scale atmospheric flows.

2 Data set

The data set used in this paper is similar to that used by KGI and KGII and consists of twice-daily "analyzed" (i.e., model-interpolated) fields of NH 700-mb heights compiled at NOAA's Climate Prediction Center. The only difference between the two data sets is that in this paper 44 winters are used, starting on 1 December 1949 and extending through March 1993. Kimoto and Ghil's data began on the same date but contained only 37 winters, through March 1986. NH winter is defined as the 90-day sequence beginning on 1 December of each year. All the analyses below were performed on the winter data, namely the $44 \times 90 = 3960$ daily maps so defined. The preprocessing also follows KGI and is summarized below.

The original data set is based on the routine processing of raw NH observations — via model assimilation of the data (e.g., Daley 1991; Ghil and Malanotte-Rizzoli 1991) — into analyzed fields, carried out by the U. S. National Centers for Environmental Prediction (NCEP, previously the National Meteorological Center), on a $10^\circ \times 10^\circ$ diamond grid north of 20°N. The 541 points of this grid are thinned out to be more nearly representative of equal-area surface elements, thus yielding 358 grid points. For each one of these points, the seasonal cycle is removed by averaging 5-day running means over the 44 years, thus providing what we shall call the *unfiltered* height anomalies. A further 10-day low-pass filter is then applied to these anomalies to obtain (low-pass) *filtered anomalies*.

EOF analysis (Preisendorfer 1988) was applied to the filtered anomalies in the standard manner to determine the leading EOFs — that is, the eigenvectors of the covariance matrix that are associated with the largest eigenvalues (i.e., variances) — of the spatial data set (see KGI). In this manner one can reduce the dimensionality of the data set from the original 358 dimensions of the grid space by projecting onto a few leading EOFs that retain a significant fraction of the original variance. Such projections are useful for visualization, density estimation, and clustering methods, all of which are easier to carry out in low-dimensional spaces. Projections used in the analyses below range from the first 2 to the first 12 EOFs.

3 Clustering methodology

a. An introduction to finite mixture models

A finite mixture model is a PDF composed of a linear combination of "component" density functions. As an example consider the synthetic 2-D data set shown in Fig. 1. These data have been generated from a mixture model containing 3 Gaussian components, having distinct means and covariances, with components weighted equally. The means of the 3 Gaussians and the ellipses are overlaid on the scatter plot in Fig. 2; the semi-major axes of each ellipse correspond in direction with the eigenvectors and in length with three times the singular values of the associated covariance matrix, i.e., three times the corresponding standard deviations.

Figure 3 shows a contour plot of the PDF. Note the non-Gaussian, multimodal nature of this contour plot. The ability to model such multi-modal density functions is a key feature of the mixture approach.

Let $\mathbf{X}$ be a d-dimensional random variable and let $\mathbf{x}$ represent a particular value of $\mathbf{X}$, e.g., a data vector with d components. A finite mixture PDF for $\mathbf{X}$, having k components, can be written as

$$f^{(k)}(\mathbf{x}|\Phi) = \sum_{j=1}^{k} \alpha_j g_j(\mathbf{x}|\boldsymbol{\theta}_j), \tag{1}$$

where each of the g_j is a component density function. Each $\boldsymbol{\theta}_j$ represents the parameters associated with density component g_j and the α_j are the relative "weights" for each component j, where $\sum_{j=1}^{k} \alpha_j = 1$ and $\alpha_j \geq 0,\ 1 \leq j \leq k$; the set of parameters for the overall mixture model is denoted by $\Phi = \{\alpha_1, \ldots, \alpha_k; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$.

The component density functions are often assumed to each be a multivariate Gaussian,

and we shall do so here. Specifically, the jth component density is given by

$$g_j(\mathbf{x}|\boldsymbol{\mu}_j, \mathsf{C}_j) = \frac{1}{(2\pi)^{d/2}|\mathsf{C}_j|^{1/2}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\mathsf{C}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)}, \tag{2}$$

where $\boldsymbol{\mu}_j$ and C_j are the mean and covariance matrix, respectively, and $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \mathsf{C}_j\}$. The mean $\boldsymbol{\mu}_j$ specifies the location of the jth density's centroid and the covariance matrix C_j prescribes how the data belonging to component j are scattered around $\boldsymbol{\mu}_j$.

Diaconis and Freedman (1984) showed that most low-dimensional projections of a high-dimensional data set that has an arbitrary multivariate PDF will result in data with an approximately *Gaussian* PDF in the lower-dimensional space. Thus, for the EOF-subspace projections discussed in this paper, one might postulate the "null hypothesis" that the data will be Gaussian in any low-dimensional projection. The search for *mixtures of Gaussians*, with $k = 2, 3, \ldots..$ components, is thus a natural step beyond the $k = 1$ hypothesis in the search for multivariate structure in this context.

The flexibility and simplicity of the mixture model has led to its widespread application in applied statistics as a density estimation and clustering tool (Titterington et al. 1985; McLachlan and Basford 1988). Historically, the earliest application of mixture modeling is credited to Pearson (1894). Crutcher and Joiner (1977) and Crutcher et al. (1982) applied Gaussian mixtures to meteorological data, using hypothesis tests based on likelihood ratios to determine the number of components k in the mixture models. Titterington et al. (1985) showed, however, that the statistics of the likelihood ratio are not well behaved for mixture models, and thus the application of likelihood ratios for choosing k is not recommended.

b. Estimating mixture model parameters from data

Let $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a data set of length N with d-dimensional multivariate observations $\mathbf{x}_n$, $1 \leq n \leq N$. Given D, one seeks a set of parameter estimates $\hat{\boldsymbol{\Phi}}$ of the true mixture parameters $\boldsymbol{\Phi}$ which characterize the PDF model assumed to have generated the data; hats ($\hat{}$) will be used to denote all estimated parameter values. At first, we assume that the number of components k in the mixture model is known and fixed: the generalization to estimating k from the data is discussed in Section 3d.

The maximum-likelihood principle states that one should seek the parameter estimates which maximize the likelihood of the parameters given the data (or equivalently the logarithm of the likelihood). This implies searching over parameter space to maximize log-likelihood $L^{(k)}$ by treating the observed data D as fixed. For mixture models the log-likelihood, assuming independent observations, equals

$$\begin{aligned} L^{(k)}(\hat{\boldsymbol{\Phi}}|D) &= \sum_{n=1}^{N} \log f^{(k)}(\mathbf{x}_n|\hat{\boldsymbol{\Phi}}) \\ &= \sum_{n=1}^{N} \log\Big(\sum_{j=1}^{k} \hat{\alpha}_j g_j(\mathbf{x}_n|\hat{\boldsymbol{\theta}}_j)\Big). \end{aligned} \tag{3}$$

Taking partial derivatives with respect to each parameter in the set $\hat{\boldsymbol{\Phi}}$ yields a set of coupled nonlinear equations. Thus, direct maximization in closed form is not feasible when d or k is large. In fact, the number p of independent parameters for a k-component Gaussian

mixture grows like $k[d(d+1)/2+d+1]-1$, which scales as $p \sim kd^2$. Thus, even for problems of reasonably low *input dimensionality* d of the data's feature space (such as $d = 5$), the *dimensionality p of the parameter space* will be quite large, and a global maximum of the likelihood function quite hard to find. In addition, the mixture's log-likelihood surface can have many local maxima; this makes the search for parameters that insure globally maximum likelihood even more difficult when p is large.

Much of the popularity of mixture models in recent years is due to the existence of efficient iterative estimation techniques for maximizing the log-likelihood. In particular, the expectation-maximization (EM) procedure (Dempster et al. 1977) is a general technique for obtaining maximum-likelihood parameter estimates in the presence of missing data. In the mixture model context, the "missing data" are interpreted as the unknown or hidden labels that identify which data points originated from which mixture component. The EM procedure guarantees convergence in parameter space to a local maximum of the log-likelihood function, but there is no guarantee of global convergence. Hence, the procedure is often initialized from multiple randomly chosen initial estimates and the largest of the resulting set of maxima is chosen as the final solution. The EM procedure for Gaussian mixtures is described in detail in Appendix A and is used for all of the results contained in this paper.

Applying the EM procedure to the 600 data points shown in Fig. 1 results in the parameter estimates shown in Fig. 4. The differences between the estimated parameters (Fig. 4) and the true parameters (Fig. 2) are quite small and only discernible by actual superposition of the two figures. Thus, the EM procedure is quite efficient at recovering the true locations and shapes of the component densities which generated the data in Fig. 1, even when N is not very large compared to p, 600 vs. 17 in this instance.

c. Clustering via mixture models

There is a long tradition in the statistical literature of using mixture models to perform *probabilistic clustering*; see Everitt and Hand (1981), Titterington et al. (1985), and McLachlan and Basford (1988) for a historical perspective. Clustering, in this mixture model context, proceeds as follows:

1. Assume that the data are generated by a mixture model, where each component is interpreted as a cluster or class ω_j and it is assumed that each data point must have been generated by one and only one of the classes ω_j.

2. Given a data set where it is not known which data points came from which components, infer the characteristics (the parameters) of the underlying density functions (the clusters).

Given estimated parameters $\hat{\Phi} = \{\hat{\alpha}_1, \ldots, \hat{\alpha}_k; \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_k\}$, one can calculate the probability that data point $\mathbf{x}$ belongs to one of the k classes ω_j by Bayes' rule:

$$\hat{P}(\omega_j|\mathbf{x}) = \frac{\hat{\alpha}_j g_j(\mathbf{x}|\hat{\boldsymbol{\theta}}_j)}{\sum_{l=1}^{k} \hat{\alpha}_l g_l(\mathbf{x}|\hat{\boldsymbol{\theta}}_l)}, \qquad 1 \leq j \leq k, \tag{4}$$

i.e., one can probabilistically assign data points $\mathbf{x}$ to clusters. Here, $\hat{\alpha}_j = \hat{P}(\omega_j)$ is an estimate of the marginal or prior probability for each cluster. In the next subsection, we shall allow $\hat{P} = \hat{P}^{(k)}(\omega_j|\mathbf{x})$ to depend on k as well, which is still kept fixed (and known) here.

The mixture model approach to clustering has the advantage that it treats the clustering problem in an explicit statistical context, allowing full treatment of uncertainty in the inference process. For example, uncertainty about the cluster locations and shapes, such as probabilistic class membership and class overlap, can be easily handled. In fact, it can be shown that mixture model clustering is a strict generalization of the well-known k-means and related algorithms that are based on finding cluster centers which minimize a least-squares objective function (Duda and Hart 1973). The mixture model is a generalization in the sense of modeling the shapes of the clusters (instead of just the centers), as well as allowing class overlap. It is clearly an agglomerative method, as compared to hierarchical clustering methods (such as CW's) that are based on pairwise distance measurements between data points. A potential disadvantage of the mixture model approach is the *a priori* assumption of a given functional form for the component densities. Thus, while Gaussian components are widely used, they are not necessarily always the most suitable choice; see, for instance, the possible emergence of nonconvex clusters when using search methods based on simulated annealing (Hannachi and Legras 1995).

d. Estimating the number k of clusters

So far we have assumed that k, the number of clusters, is known *a priori*. Often one would like to determine k from the data, if at all possible. A case in point is the multidecadal NH 700-mb height data set, given the considerable prior work on trying to determine how many regimes can be reliably identified in these data.

In a probabilistic context one would like an estimate of $P(k|D)$, the posterior probability for k clusters given the data set D, $1 \leq k \leq k_{\max}$. In the present work we use a robust and consistent data-driven methodology based on *cross-validated likelihood* as the basis for estimating $P(k|D)$.

Cross-validation operates by repeatedly dividing the available data D into two disjoint partitions (Stone 1974), fitting the model on one of the partitions, and estimating performance on the other (see also KGI for another application, to PDF estimation). After some number of such trials, the performance estimates are averaged to get an "honest" estimate of out-of-sample performance. Specifically in the mixture model context above, the procedure is as follows:

1. Partition the data set D into a fraction β for model fitting, and a disjoint fraction $1 - \beta$ for performance estimation.

2. Fit a mixture model with k components (i.e., estimate its parameters) to the fraction β of the data reserved for model fitting, $D^{(\beta)}$.

3. Estimate the log-likelihood [Eq. (3)] of these model parameters on the fraction $1 - \beta$ of the data reserved for performance estimation, $D^{(1-\beta)}$.

4. Repeat Steps 2 and 3 for a range of k-values (usually for $k = 1, \ldots, k_{max}$).

5. Repeat Steps 1 to 3 for a total number of M partitions (times), where each time the data is randomly divided into two partitions as above. Let $L_m^{(k)}$ be the estimated log-likelihood of the mth partition for a model with k components, $L_m^{(k)} = L(\hat{\mathbf{\Phi}}_m^{(k)}|D_m^{(1-\beta)})$, $1 \leq m \leq M$; here $\hat{\mathbf{\Phi}}_m^{(k)} = \{\hat{\alpha}_1, \ldots, \hat{\alpha}_k; \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_k\}$, i.e., the set of parameters for a mixture model with k components, where the dependence on k is now made explicit, and the parameters are fitted via maximum likelihood on the mth training data set $D_m^{(\beta)}$.

6. Calculate the average log-likelihood (over the M runs) for each of the different k-values to obtain the *cross-validated log-likelihood*

$$L_{cv}^{(k)} = (1/M) \sum_{m=1}^{M} L_m^{(k)}, \qquad 1 \leq k \leq k_{\max}. \tag{5}$$

7. Obtain estimates of the posterior distribution for k by calculating

$$\hat{P}(k|D) = \frac{exp(L_{cv}^{(k)})}{\sum_{l=1}^{k_{\max}} exp(L_{cv}^{(l)})}, \qquad 1 \leq k \leq k_{\max}; \tag{6}$$

   Eq. (6) follows from Bayes' rule by assuming equal priors on different values of k.

In practice the method is not sensitive to the exact values of β or M when the data set is relatively large compared to the complexity of the fitted models. This is the case for the geopotential height data discussed in the next section. Thus, default values of $\beta = 0.5$ and $M = 20$ are used throughout. A discussion of the theoretical properties of the above cross-validation method is provided in Appendix B.

To illustrate the method, Table 1 shows the cross-validated likelihoods and estimated posterior probabilities obtained from running the cross-validation procedure on the data shown in Fig. 1. There is clear evidence that $k = 3$ is the best model given the cross-validation information. Nonetheless, the fact that $\hat{P}(k > 3) \neq 0$ demonstrates that inferring the correct number of components from such data is nontrivial. In general, the ability of this method (or indeed any purely data-driven method) to automatically infer the "true" number of clusters present in a data set will improve as the amount of data increases relative to the complexity of the cluster model; "complexity" in this context is taken to mean both the number of clusters and the degree of overlap (or closeness) among them.

4 Hemispheric results

a. Cross-validated clustering results

Following the approach of KGI and others, we are interested in determining the cluster structure, if any, of the NH height anomaly data described earlier, as it appears in a low-dimensional subspace of leading EOFs. We applied the mixture model clustering method outlined in Section 3 to the 44-winter set of NH height anomalies presented in Section 2. The unfiltered anomalies were projected onto the first two EOFs of the filtered data set (see Section 2). A scatter plot of the resulting projection is shown in Fig. 5.

We ran the mixture model cross-validation method on this two-dimensional (2-D) data set. The algorithm described in Section 3d was modified so that random partitions were

chosen based on winters rather than days, i.e., half of the 44 winters were placed in the training set and the remainder in the test set. This modification is necessary to ensure that the training and test partitions are truly independent (and, thus, guarantees the theoretical consistency of the method as described in Appendix B).

The number k of clusters (i.e., mixture components) was allowed to take on all values from 1 through 15. The log-likelihoods for $k \geq 7$ were invariably much lower than those for $k \leq 6$ so we present, for clarity, only the results for $k = 1, \ldots, 6$. The posterior probabilities and cross-validated log-likelihoods $L_{cv}^{(k)}$ are tabulated in Table 2. The posterior probabilities provide clear evidence for the data supporting exactly 3 clusters, i.e., the cross-validation estimate of the posterior probability for 3 clusters is effectively 1 and all others are effectively zero.

Note that the absolute values of the log-likelihoods are irrelevant — strictly speaking, likelihood is only defined within an arbitrary constant. More precisely, let $L_{cv}^{(k^*)}$ be the cross-validated likelihood for some particular value of $k = k^*$ as defined by Eq. (5). Subtracting $L_{cv}^{(k^*)}$ from each of the cross-validated likelihoods $L_{cv}^{(k)}, 1 \leq k \leq k_{\max}$, does not affect the posterior probability estimates in Eq. (6) since it is equivalent to multiplying above and below by $\exp(-L_{cv}^{(k^*)})$ to yield

$$\hat{P}(k) = \frac{\exp\left(L_m^{(k)} - L_{cv}^{(k^*)}\right)}{\sum_{l=1}^{k_{\max}} \exp\left(L_m^{(l)} - L_{cv}^{(k^*)}\right)}, \qquad 1 \leq k \leq k_{\max}. \tag{7}$$

Thus, it is the *differences* between the log-likelihoods that matter.

Choosing $k^* = 3$, Table 3 shows the differences between the log-likelihoods for a given k and that for $k^* = 3$ in the case of each partition. Larger log-likelihood differences are better, i.e., the relative likelihood of the highlighted column is stronger. Since $k = 3$ is always zero, negative log-likelihoods for any entry mean that for that partition m and value of k, the log-likelihood was less than that for $k = 3$. The number of partitions $k = 3$ clearly dominates: it yields the highest-likelihood model in 15 out of 20 cases, with $k = 2$ "carrying the day" in 4 cases and $k = 1$ in only one. Considerable variability occurs between partitions since estimates of likelihood can be sensitive to outliers. But it is the cross-validated likelihood $L_{cv}^{(k)}$, calculated as the mean of the individual likelihoods on each partition, which matters in finally determining the number of clusters (see again Table 2).

b. Robustness with respect to partition choices and dimensionality

We carried out numerous runs on the same data with different randomly chosen partitions among the 44 winters and using exactly the same parameters as described before ($\beta = 0.5, M = 20$). All these runs provided the same result, namely an estimated posterior probability of $\hat{P}(k = 3) \approx 1$. The cross-validated likelihood value and the estimated probability of $k = 3$ for each such run is shown in Table 4.

We also investigated the robustness of the method with respect to the number of leading EOFs retained, i.e., to the dimensionality of the large-variance subspace in which the mixture model is constructed and tested. The unfiltered 700-mb height anomalies were projected onto the first d EOFs, $d = 2, \ldots, 12$. As a function of the dimensionality d, the posterior probability mass was highly concentrated at $k = 3$, i.e., $\hat{P}(k = 3) \approx 1$, until $d = 6$; at this point the mass "switched" to become concentrated at $k = 1$, i.e., $\hat{P}(k = 1) \approx 1$. It follows that, as the dimensionality increases beyond $d = 6$, the cross-validation method

does not provide any evidence to support a model more complex than a single Gaussian bump. This is to be expected since the number of parameters p in a k-component Gaussian mixture model grows in proportion to kd^2 (see Section 3b). Thus, for example, in $d = 10$ dimensions, there are $p = 168$ parameters for a 3-component model but only 56 parameters for a single-component model. In contrast, in 5 dimensions, the 3-component model needs only 48 parameters.

Since the total amount of data to fit the models is fixed, as the dimensionality d increases the estimates of the more complex models become less reliable and cannot be justified by the data. This is consistent with fairly general considerations of accurate and robust PDF estimation in d dimensions (see KGI, Section 5, and references there) and with the theoretical arguments given in Appendix B that cross-validation will pick the best mixture model which can be fit to a finite set of data. If the data are sufficient in number N, this best model will correspond to the true model; if there are too few data (relative to the complexity of the models being fit), on the other hand, the method will be more conservative and choose a simpler model that can be supported more reliably by the data. Another interpretation of this result is that empirical support of the 3-component model in higher dimensions would require records on the order of a few hundred years long, rather than the 44 years of data currently available [compare also Lorenz (1969)].

For the 3-component Gaussian model, we also investigated the variability in the physical maps obtained as cluster centroids when retaining different numbers of leading EOFs. The unfiltered height anomalies were projected onto the first d EOFs for $d = 3, \ldots, 12$ and a Gaussian mixture model with $k = 3$ components was fit to the data for each case. For each value of d, 3 maps of 700-mb height anomalies were obtained from the centers of the 3 Gaussians. The pattern correlations [as defined in Mo and Ghil (1988), CW or KGI] were then calculated between each of these maps (obtained in d dimensions) and the corresponding 3 maps obtained when using $d = 2$ (see Section 4a). The results, shown in Table 5, indicate that the correlations between the 2-D EOF maps and maps obtained in up to 12 EOF dimensions are very high. One can conclude that the dimensionality of the high-variance EOF subspace does not affect the qualitative patterns of the geopotential height maps in any significant manner, when using the mixture model clustering procedure applied here. Our results are also robust with respect to the preprocessing of the data, as shown in Appendix C.

c. Comparison with CW's results and interpretation

Given that there is strong evidence for 3 Gaussian clusters, we fit a 3-component Gaussian model to the entire set of 44 winters in the 2-D EOF space (rather than partitioning into halves as before) and examine the results. Figure 6 shows the location of the means of the Gaussians and the three-standard-deviation ellipses associated with their covariance matrices, superposed on a scatter plot of the data projected onto the first 2 EOFs. The resulting contour map of the bivariate-mixture PDF is shown in Fig. 7.

The means of the 3 Gaussians fitted in Fig. 6 have a natural interpretation as the centers of 3 Gaussian data clusters. Figure 8 presents the 3 maps corresponding to our 3 cluster means on the left and the 3 maps corresponding to CW's most reproducible hierarchical clusters on the right [from Fig. 11 of Wallace (1996)]. CW labeled both the maps in question and the clusters they represent **A** (for **A**laska), **G** (for **G**reenland), and **R** (for **R**ockies). Their maps and ours have a high degree of pairwise qualitative similarity in terms of the

spatial patterns, while the size of the associated anomalies differs, as discussed further below. The upper (**A**) maps both clearly possess a distinctive ridge over the Gulf of **A**laska. The middle (**G**) maps exhibit a strong high over southern **G**reenland. The bottom (**R**) maps are characterized by an intensification of the Pacific jet stream and an enhancement of the climatological mean ridge over the **R**ockies.

Note that CW and the present study use two distinct, and rather different, methodologies (mixture modeling here and hierarchical clustering in CW), as well as two somewhat different data sets (700-mb vs. 500-mb data over slightly different time spans) and different preprocessing of the data (the work in this paper was in an EOF subspace, while CW clustered the anomaly maps directly). CW's methodology for arriving at 3 distinct, highly significant clusters was based on a combination of sophisticated resampling of the data and subjective judgment. In their own words, "the more reproducible clusters are strung out along three well-defined 'branches' of the family tree" (see especially Fig. 15 of CW). The cross-validation results described here can be viewed as an independent and totally objective validation of CW's "3-cluster" result, confirmed also, less independently, by Wallace's (1996) extension of the CW analysis to 1989. It is quite reassuring that both methodologies permit us to conclude that 3 distinct regimes dominate the NH wintertime low-frequency variability over the past half-century and that the maps corresponding to the centroids of these regimes — as obtained by either one of the two — are qualitatively quite similar.

There is an important qualitative difference between the mixture model clusters found here and clusters found by partition-based methods such as the "fuzzy clusters" of Mo and Ghil (1998) or the hierarchical clusters of CW. Each mixture model cluster corresponds to a component in the mixture density function, and thus, the sum of their contributions is a well-defined PDF in the large-scale atmosphere's phase space. Equivalently, the mixture components must "account" for all of the data, i.e., the model covers the system's entire phase space, as sampled by the available observations, and not just a portion of it. In contrast, the hierarchical clusters found by CW are local in nature. Thus, for the mixture model, the component weights α_j are constrained to sum to 1, and the component covariance matrices $\mathsf{C_j}$ are constrained by the overall covariance structure of the entire data set. Most importantly, the means $\boldsymbol{\mu}_j$ are also subject to a "global" constraint imposed by the overall mean of the data equaling zero and by a somewhat indirect coupling to the overall covariance structure.

The *directions* of the mean vectors from the origin in phase space, i.e., the overall mean of the data set, however, are relatively unconstrained. It is, therefore, the angles of the centroids – and, in turn, the associated spatial patterns on the grid – that are the most directly determined by the data, while the distances from the origin (the amplitude scale of the maps), the component covariances, and the component weights are less data-driven and more constrained by the model. This observation provides a more rigorous basis for the heuristic choice of Mo and Ghil (1988) and KGII to concentrate on angular PDFs. It also explains why the maps found by CW's clustering and our cluster centers have very similar spatial patterns but are scaled differently (see Fig. 8), i.e., the cluster centroids lie, in either case, along the same directions from the origin in phase space, but at different distances, due to different constraints in the respective (Euclidean-distance) models.

This point is reinforced by comparing the location of the centers in Fig. 7 here with the center locations in CW's Figure 15a; the polarity is reversed, since our EOFs and CW's came out to have opposite sign. Cluster **A** for example is further from the origin in CW than

in this paper; it is clear, however, from CW's Figure 15a that the hierarchical clustering algorithm produced a "trajectory" of clusters, one of which was chosen as the definitive cluster by CW's method.

In summary, the mixture model's cluster centroids are constrained to be closer to the origin than is the case in methods which seek local structure in a Euclidean phase space. Nonetheless, it is clear from a comparison of CW's results and those here that the angles from the origin, and hence the spatial patterns of the associated regimes, are essentially the same in both analyses.

The **A**, **G**, and **R** patterns also bear a close resemblance to some of the clusters identified by KGI and by Molteni et al. (1990). In particular, the match of map **A** here (and in CW) with KGI's RNA is almost perfect and that between map **R** and KGI's PNA quite good, but slightly less so over the Atlantic-European sector. The similarity between **G** here (and in CW) with KGI's *Blocked NAO* (BNAO) is again excellent over the areas of strongest anomalies, in the Atlantic-European sector this time, but not as good in the complementary, Pacific/North-American sector. This slight mismatch is essentially due to the fact that — as Mo and Ghil (1988) observed (their Figs. 4 and 13) and CW and Kimoto and Ghil (1993a,b) corroborated (see especially Fig. 17 in CW and Fig. 11 of KGI) — EOFs 1 and 2 of the NH wintertime height anomalies are roughly determined by the patterns of positive and negative PNA and positive and negative NAO. We refer to these earlier papers, and further references therein, for a more complete synoptic description of the spatial patterns involved and their climatological importance.

KGI — by using a less rigorous clustering method than the one applied here, namely visual inspection of the bivariate PDF derived from a kernel density estimation method — found 4 clusters: PNA and RNA, BNAO and *zonal NAO* (ZNAO), the last of which is missing in CW and the analysis here. The cluster centroids with (approximately) opposite polarities for the PNA and NAO, respectively, did not exhibit in KGI (nor do the **A** and **R** maps here and in CW) quite the same spatial patterns (with the sign of the local anomalies reversed); likewise, these centroids do not have simply the sign-reversed coordinates of PNA and NAO, respectively, in the subspace of the two leading EOFs (in the case of RNA–PNA and ZNAO–BNAO in KGI and of **A**–**R** only here and in CW). Still, to first order, the present analysis is consistent with the view that the hemispheric regimes arise, pairwise, from sectorial regimes that correspond to an intensification or weakening of zonal flow in the ATL or PAC sector. The coordinates of our centroids are **A**$\cong$ (-297 m, 42 m), **R**$\cong$ (226 m, 181 m), and **G**$\cong$ (130 m, -487 m), with the first coordinate along EOF 1 and the second along EOF 2. The Pacific/North-American sector features of **G** are obviously distorted with respect to KGI's BNAO since **A** and **R** are forced to carry also, between the two of them, the features of ZNAO in that sector. This issue is clarified further in Section 5.

5 Sectorial results

We applied the mixture model clustering method of Section 3 separately to the (a) Pacific (PAC) sector (120°E – 60°W) and (b) Atlantic (ATL) sector (60°W – 120°E). The data were preprocessed as described in Section 2 (see also KGII) and separate sets of EOFs were estimated in each sector. The 179(= 358/2) spatial data points for each day in either sector were projected onto the first 4 EOFs, as in KGII. The mixture model clustering procedure

was applied to the data in each sector, with k ranging from 1 to 10, $\beta = 0.5$, and $M = 20$.

a. PAC sector

The estimated posterior probabilities on k are $\hat{P}(k = 2) = 0.980$ and $\hat{P}(k = 3) = 0.020$; all the probabilities for other values of k are zero. Thus, cross-validated likelihood points to $k = 2$ as the most likely model to fit the data by far; a very slight ambiguity in the result appears when compared to the hemispheric analysis of Section 4, where the probabilities were essentially zero except for $k = 3$.

Figure 9 shows the location of the means and three-standard-deviation covariance ellipses of the estimated Gaussian components for $k = 2$, superposed on a scatter plot of every 10th day projected onto the first 2 EOFs. Figures 10a and b show the maps corresponding to the centroids of the 2 clusters. The spatial pattern in panel (a) clearly resembles the P1 regime and that in panel (b) the P2 regime of KGII's sectorial analysis; our 2 PAC clusters also resemble CW's sectorial clusters R and A, respectively. These 2 PAC regimes are the sectorial counterpart of the hemispheric PNA and RNA clusters, here as well as in CW and KGI. Following CW, we use italics to distinguish between the sectorially defined *PNA* (panel a) and *RNA* (panel b) and the hemispheric regimes.

Projecting the data onto the first 2 EOFs, rather than the first 4 EOFs as above, produces estimated posterior probabilities of $\hat{P}(k = 2) = 0.824$ and $\hat{P}(k = 3) = 0.176$, while projection onto 3 EOFs produces $\hat{P}(k = 2) = 0.970$ and $\hat{P}(k = 3) = 0.030$. The nonzero probabilities for the 2-D case here are quite similar to those obtained for the synthetic 3-cluster case in Table 1, except that the distribution here peaks at $k = 2$. For both ATL cases, of 2 and 3 EOFs, the cluster centroids correspond to the same *PNA* and *RNA* patterns. The somewhat larger probability for $k = 3$ in the 2-D subspace suggests that more complex structure (i.e., $k > 2$) may be present, as apparent in KGII and MVL, but this structure is not fully supported by the current data.

b. ATL sector

For the ATL sector, the posterior probabilities on k – when projecting the data onto the 4 leading EOFs – are essentially zero except for $\hat{P}(k = 2) = 0.991$ and $\hat{P}(k = 3) = 0.009$. Again, as in the PAC sector, there is a very slight ambiguity as to the true number of clusters.

Figure 11 shows the location of the means and three-standard-deviation covariance ellipses of the estimated Gaussian components for $k = 2$, superposed on a scatter plot of every 10th day projected onto the first 2 EOFs. Figures 10c and d show the height anomaly maps that correspond to the centroids of the 2 clusters from the $k = 2$ solution. They bear a close resemblance to the A1 and A5 regimes of KGII and to the $G+$ and $G-$ clusters of CW's sectorial analysis. We label them as *BNAO* and *ZNAO*, respectively.

Projecting the data onto the leading 2 EOFs, rather than 4 EOFs (see also Section 5a), produces estimated posterior probabilities of $\hat{P}(k = 2) = 0.002$ and $\hat{P}(k = 3) = 0.998$; projecting onto the first 3 EOFs yields $\hat{P}(k = 2) = 0.174$ and $\hat{P}(k = 3) = 0.825$. Thus, for the ATL sector, the data projection onto either a 2-D or 3-D subspace supports a model with $k = 3$. From a statistical-estimation viewpoint it is not surprising that in lower dimensions the data can support a more complex model, since there are fewer parameters to be fitted than in the 4-D case (see discussion in Section 4b). It is, therefore, even more remarkable

that the hemispheric $k = 3$ result is quite stable for dimensionality d ranging from 2 to 12, and the PAC result is also stable for $2 \leq d \leq 4$ (see Sections 4b and 5a, respectively).

While the ATL-sector results are more ambiguous, the cluster centroids that correspond to the $k = 3$ model (in both 2-D and 3-D subspaces) display continuity with respect to the $k = 2$ model in the 4-D EOF space, namely both the *BNAO* and *ZNAO* clusters are retained. The third cluster for both subspaces has essentially the same spatial pattern, and is quite similar to the A2 pattern in KGII and to MVL's second ATL cluster.

c. Comparison with previous results

Using hierarchical clustering on the sectorial data, CW obtained 2 most reproducible clusters in each sector: A and R in the PAC sector and $G+$ and $G-$ in the ATL sector (see their Figs. 8 and 9, respectively). KGII, on the other hand, found as many as $k = 7$ clusters for the PAC sector and $k = 6$ clusters for the ATL sector (see their Figs. 5 and 6, respectively), using bump-hunting on estimated angular PDFs.

MVL introduced a classifiability index into their dynamical clustering, based on similarity of partitions to which the algorithm converges, when started with the same number k of seed points, but different sets of such points. The maximum value of this index, for either sector, occurred for $k = 2$ (their Fig. 2). Not satisfied with this result, they introduced an extraneous "null hypothesis" of the daily maps being generated by a first-order vector Markov process built on the leading 8 EOFs of the data set and with the same covariance matrix as the data at lag 0 and 2 days. The classifiability index was significantly distinct from its distribution – as given by 100 Monte Carlo simulations of this process with the same length as the data set – for $k = 3$ in the PAC sector and $k = 4$ in the ATL sector. Still the spatial patterns of the centroids obtained in the 2 sectors (their Figs. 7 and 4, respectively) corresponded fairly closely to a subset of KGII's and included, in particular, CW's pair A–R (or, equivalently, KGII's pair P2–P1) in the PAC sector, as well as the $G+$(CW)/A5(KGII) pattern in the ATL sector. Furthermore the 3 PAC and 4 ATL clusters of MVL could be clearly "parsed" into the pair of clusters per sector associated with the "opposite phases" of the PNA and NAO (their Fig. 10).

In fact, in our sectorial results the two centroids of the pairs A–R and $G+$–$G-$ are essentially mirror images of each other. This mirroring is clearly visible in the locations of the means in Figs. 9 and 11. Indeed, as discussed already in Section 4c, the mixture model imposes certain constraints on the possible cluster centroids which are fitted to the data. In particular, with $k = 2$ and zero-mean data (as is the case with the PAC and ATL sectors), it is straightforward to show from Eqs. (1) and (2) that $\boldsymbol{\mu}_1 = r\boldsymbol{\mu}_2$, where $r = -\alpha_2/\alpha_1$, i.e., the two centroids must have the same spatial pattern and opposite sign, with possibly different scales (if $\alpha_2 \neq \alpha_1$). An immediate consequence of this property of Gaussian mixture models is that significant skewness with respect to two intersecting axes will result in nonvanishing probability of $k \geq 3$ clusters; this is the case for our NH results and $2 \leq d \leq 12$ (shown for $d = 2$ in Figs.5–7), as well as for the ATL results in 2-D (not shown).

The cluster results for the PAC and ATL sectors across different studies clearly point to multimodality in the PDFs. There remains some uncertainty as to the precise number of regimes which can be reliably identified in each of the sectors. The mixture model results here are consistent with previous studies in the sense that they consistently recover the well-known sectorial features of the PNA, RNA, BNAO, and ZNAO patterns. The present

model results also seem to support, to a greater degree in the ATL and lesser in the PAC sector, the possibility of more complex structure in the sectorial PDFs. At the same time, they indicate that we do not yet have sufficient data to confirm with complete confidence mixture models for such greater complexity than $k = 2$ in either sector.

6 Concluding remarks

a. Summary

Probabilistic clustering, using finite mixture models, was introduced and described for the purposes of automatically clustering 44 winters of NH 700-mb height anomaly data. After projecting the anomalies onto the data set's leading EOFs in the standard manner, we used a mixture model clustering algorithm to determine what cluster structure – if any – existed in the NH data. There is clear evidence that the last half-century of upper-air data supports exactly 3 distinct clusters. A feature of the present method, compared to alternative clustering methodologies, is precisely its ability to objectively answer the question "how many clusters are justified by the data?" On examination, the patterns associated with the 3 clusters found by the mixture approach have a very close correspondence to the 3 cluster patterns found by Cheng and Wallace (1993; throughout CW) using hierarchical, non-probabilistic clustering on a comparable data set of NH 500-mb height anomalies. The three common clusters **A** for a pronounced ridge over the Gulf of **A**laska, **G** for an anticyclonic feature over southern **G**reenland, and **R** for an enhancement of the climatological ridges over the **R**ocky mountains. The three clusters also agree well with KGI's RNA, BNAO and PNA regimes, while KGI captured a hemispheric ZNAO regime as well.

Previous work (KGII, MVL) had suggested that the NH winter upper-air data support a more complex classification when examined separately over the Atlantic (ATL) and Pacific (PAC) sector. We applied the same mixture model methodology to the sectorial data. Two clusters each emerge for the PAC sector, *PNA* and *RNA*, and for the ATL one, *ZNAO* and *BNAO*. These correspond to zonal and blocked flow over the sector under study. *PNA* and *RNA* capture the sectorial PAC features of CW's and the present paper's hemispheric **R** and **A** clusters, while *ZNAO* and *BNAO* capture the ATL ones of KGI's hemispheric ZNAO and BNAO. It thus appears, as suggested by Mo and Ghil (1988) and KGI-II, that the hemispheric clusters are manifestations of the sectorial ones. The tentative explanation for the ZNAO's "missing" in the CW analysis and that of Section 4 here is provided next.

b. Discussion

A striking feature of the sectorial ATL results is the closeness of the *ZNAO* centroid to the origin. This is consistent with the predominantly positive anomaly (zonal phase) of the NAO index over the last half-century (Hurrell 1995). This index is defined as the sea-level pressure difference between the **A**cores and Iceland and measures the intensity of the westerly jet across the North Atlantic basin. Its consistently zonal character during the last few decades helps explain the absence of a stable ZNAO cluster from the hemispheric analysis of CW and present paper.

Interdecadal changes in the large-scale atmosphere's intraseasonal, 10–100 day low-frequency variability seem to manifest themselves more in changes of the relative number of days of residence in each cluster than in changes of the clusters' spatial patterns (Robertson

et al. 1997). This would lead us to suspect that an analysis of hemispheric upper-air data – if they were available – over the first half of the 20th century might yield four regimes, rather than three, as ZNAO's centroid could move more farther away from the 2-D phase space's origin, due to the distribution of days between it and *BNAO* becomes better balanced. Indeed, the main physical cause for the existence of multiple regimes, sectorial and hence hemispheric, appears to lie in the nonlinear dynamics of the westerly jet in either sector. The dynamics involves the linear instabilities – barotropic and baroclinic, exponential and oscillatory – of the jet and their nonlinear saturation, as well as their interaction with zonally asymmetric lower boundary conditions, topographic and thermal. The simplest manifestation of this complex dynamics – and the elementary proof for its nonlinear character – is the sectorial bimodality demonstrated herein.

This bimodality is distinct from the hemispheric one claimed by Benzi et al. (1986) and Hansen and Sutera (1986), as the latter requires preferred simultaneity of the blocked vs. the zonal circulation phases in the two sectors. Such simultaneity is the rule in simple models with identical sectors (Legras and Ghil 1985) and in laboratory devices that are a "wet" version of such models (Tian 1997; Weeks et al. 1997) but occurs only rarely in the existing atmospheric upper-air data (KGII and here, not shown). The generally observed lack of simultaneity between the two sectors could arise from slightly different periods of the two sectors' oscillatory instabilities (Marcus et al. 1996) and the occasional "phase locking" between the two from different initial data available early in each winter, when the oscillations become active (Strong et al. 1995). The available half century of upper-air data is rather short to permit the more detailed examination of these theoretical conjunctures and we shall have therefore to undertake extended simulations of the atmospheric circulation with fairly detailed and realistic models in order to verify or falsify them.

Acknowledgments. The authors would like to acknowledge M. Kimoto for providing the data described in this paper, J. Roden for assistance in preprocessing the data, and J. M. Wallace and Springer-Verlag for permission to use a copy of the 3 height-anomaly panels in Fig. 11 of Wallace (1996) as panels b, d and f of Fig. 8 here. A. Fraser, M. Kimoto, T. Palmer, A. W. Robertson, M. Turmon, and R. Vautard provided useful intellectual input. The research described in this paper was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. The work of P.S. was supported in part by NSF Grant IRI-9703120. The work of M.G. and K.I. was supported by NSF Grant ATM95-23787 and NASA Grant NAG 5-713. K. Hartman helped with the editing and word processing.

APPENDIX A
The EM Procedure for Gaussian Mixtures

The Expectation-Maximization (EM) procedure is an iterative method for mixture modeling whereby the parameters at iteration $r+1$ are updated based on parameter estimates from iteration r. For a general discussion of the theoretical basis of the method see, for example, Dempster et al. (1977); we provide here only a brief summary of the procedure in the context of Gaussian mixtures.

For Gaussian mixtures the parameter set Φ consists of weights α_j, the d-dimensional means $\boldsymbol{\mu}_j$, and the $d \times d$ covariance matrices C_j, for each component $1 \leq j \leq k$. There are N data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N\}$, each being represented by a d-dimensional $\mathbf{x}_i$. The procedure is initialized by randomly choosing the mean vectors $\boldsymbol{\mu}_j$ and initializing the other parameters appropriately. At iteration r, let

$$\hat{P}^r(\omega_j|\mathbf{x}_n) = \frac{\hat{\alpha}_j^r g_j(\mathbf{x}_n|\hat{\boldsymbol{\mu}}_j^r, \hat{\mathsf{C}}_j^r)}{\sum_{l=1}^k \hat{\alpha}_l^r g_l(\mathbf{x}_n|\hat{\boldsymbol{\mu}}_l^r, \hat{\mathsf{C}}_l^r)} \qquad 1 \leq j \leq k, \quad 1 \leq n \leq N, \tag{A.1}$$

be the probability that data point $\mathbf{x}_n$ belongs to component density j, given the parameters α_j^r, $\boldsymbol{\mu}_j^r$, C_j^r for k multivariate Gaussian density functions g_j as defined in Eq. (2).

At the next iteration $(r+1)$, the parameter estimates are:

$$\hat{\alpha}_j^{r+1} = \frac{1}{N} \sum_{n=1}^N \hat{P}^r(\omega_j|\mathbf{x}_n), \tag{A.2}$$

$$\hat{\boldsymbol{\mu}}_j^{r+1} = \frac{\sum_{n=1}^N \hat{P}^r(\omega_j|\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^N \hat{P}^r(\omega_j|\mathbf{x}_n)}, \tag{A.3}$$

$$\hat{\mathsf{C}}_j^{r+1} = \frac{\sum_{n=1}^N \hat{P}^r(\omega_j|\mathbf{x}_n)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_j^r)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_j^r)^T}{\sum_{n=1}^N \hat{P}^r(\omega_j|\mathbf{x}_n)}. \tag{A.4}$$

These update equations have the simple interpretation of being standard maximum-likelihood estimates for membership, mean, and covariance parameters respectively, modified to *weight* the data points by their membership probabilities (i.e., to use, in a sense, "fractional" data points).

A basic property of the EM procedure is that the likelihood is a nondecreasing function of r, i.e., the procedure is guaranteed to converge to a fixed point, provided the sequence of estimated parameters $\hat{\Phi}^r$ ranges over a compact (i.e., closed and bounded) p-dimensional set. This fixed point in parameter space need not be a global maximum and is a function of the initial guess $\hat{\Phi}^0$. Thus, in practice, several different initial guesses can be tried and the maximum likelihood among these selected. For the results reported here, 10 different, randomly selected initial parameter sets were chosen for each run of the EM procedure. The different initial parameter sets were found by running the k-means algorithm (e.g., Duda and Hart 1973), using different random starting points for the k means. This initialization procedure is common practice in the application of EM to mixture model clustering.

Note that the fixed point obtained by the EM procedure may be a singular solution for which one of the mixture components is centered on a particular data point and the determinant of the associated covariance matrix approaches zero. This type of singularity

results in a likelihood which approaches (positive) infinity. Such solutions are typically not of interest and in practice are discarded. Singularities of this type occur at the boundaries of the relevant compact set in parameter space and can be avoided by restricting the search for model parameters to a compact set that lies within the full set and has a boundary that is bounded away from the singularities. For data sets where N is large relative to k, singular solutions are rarely a problem in practice. Indeed, in the results reported in this paper no such singular solutions were ever generated.

APPENDIX B
Cross-Validated Likelihood for the Number of Clusters k

From a statistical viewpoint, the most consistent approach for finding k is the full Bayesian solution where the posterior probability of each value of k is calculated given the data, priors on the mixture parameters, and priors on k itself. The posterior distribution for k contains, in principle, the necessary information for deciding how many clusters are justified by the data. If the posterior is peaked about a particular k then the data provide strong evidence for that value of k. On the other hand, if the posterior is "spread out" among different k values (high entropy), the data cannot discriminate which k is most likely. A potential difficulty with this approach is the computational complexity of integrating over the parameter space to calculate the posterior distribution on k. Various analytic approximations (Chickering and Heckerman, 1997) or Monte-Carlo sampling approximations (Robert 1996) have been used to get tractable estimates for this posterior distribution.

A different approach to this problem is to obtain a data-driven estimate of the posterior distribution on k using cross-validated likelihood (Smyth 1996). Cross-validated likelihood is asymptotically consistent in the sense that it will always choose the correct model in the limit of increasing data set sizes. In practice, it has been shown to work well empirically on a variety of simulated and real data sets, performing as well as various Bayesian approximation methods (Smyth 1996). It has certain distinct advantages over the Bayesian approximation approach. It is conceptually simpler to interpret and easier to implement. In addition it does not rely on approximations whose impact (in the Bayesian case) on the quality of the posterior probability estimates can be difficult to determine.

Let $f(\mathbf{x})$ be the true PDF for $\mathbf{x}$. Let $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a random sample from f. Consider a set of finite mixture models with k components being fitted to D, where k ranges from 1 to $k_{\max}$. Thus, we have an indexed set of estimated models, $f^{(k)}(\mathbf{x}|\hat{\mathbf{\Phi}}^{(k)}), 1 \leq k \leq k_{\max}$, where each $f^{(k)}(\mathbf{x}|\hat{\mathbf{\Phi}}^{(k)})$ has been fitted to data set D.

Let $\hat{\mathbf{\Phi}}^{(k)}$ be the parameters for the kth mixture model obtained by maximizing the likelihood as described in Section 2b using the data D. As k increases, the log-likelihood $L^{(k)}(\hat{\mathbf{\Phi}}^{(k)}|D)$ [as defined in Eq. (3)], is a nondecreasing function of k, since the increased flexibility of more mixture components allows a better fit to the data (increased likelihood). Thus, $L^{(k)}(\hat{\mathbf{\Phi}}^{(k)}|D)$ cannot provide any clues as to the *true* mixture structure in the data, if such a structure does exist.

Imagine instead that one had a large test data set D^{test} which is not used in fitting any of the models. Let $L^{(k)}(\hat{\mathbf{\Phi}}^{(k)}|D^{\text{test}})$ be the log-likelihood as defined in Eq. (3), where the parameters $\hat{\mathbf{\Phi}}^{(k)}$ are estimated from D as above, but the likelihood is evaluated relative to D^{test}. We can view this likelihood as a function of the "parameter" k, keeping all other parameters and D^{test} fixed. Intuitively, this "out-of-sample likelihood" should be a more honest estimator than the training-data likelihood for comparing mixture models with

different numbers of components.

It is straightforward to show that

$$E\left[\frac{-1}{N}\hat{L}_k(\hat{\Phi}^{(k)}|D^{\text{test}})\right] = \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{f^{(k)}(\mathbf{x}|\hat{\Phi}^{(k)})} d\mathbf{x} + \text{const.}, \tag{B.1}$$

i.e., the expected value of $L^{(k)}(D^{\text{test}})$ (scaled appropriately), with respect to different test-data sets drawn randomly from the true distribution f, is the cross-entropy between between $f(\mathbf{x})$ and $f^{(k)}(\mathbf{x}|\hat{\Phi}^{(k)})$, plus an arbitrary constant. Thus, the out-of-sample log-likelihood $L^{(k)}(D^{\text{test}})$ is an unbiased estimator of this cross-entropy. The cross-entropy in turn functions as a distance measure of how far the model $f^{(k)}(\mathbf{x}|\hat{\Phi}^{(k)})$ is from the true f: cross-entropy is strictly positive unless $f^{(k)}(\mathbf{x}|\hat{\Phi}^{(k)}) = f$ above. Thus, choosing the k which minimizes the out-of-sample likelihood is equivalent (on average) to choosing the model (within the model family under consideration) which is closest in a cross-entropy sense to f.

However, in practice, one cannot afford or does not have available a large independent test set such as D^{test}. A standard technique in such situations is to estimate the out-of-sample performance using cross-validation. The algorithmic procedures for repeatedly partitioning the data in random fashion and calculating the *cross-validated* log-likelihood are described in Section 3d.

Since the log-likelihood estimate for each cross-validation partition of the data is based on data which is independent from that used to fit the model, each such estimate is an unbiased estimator of the cross-entropy between the model and the true density f. In turn, since expectation is a linear operator, the average of these estimates (namely the cross-validated likelihood estimates) is in turn unbiased. Thus, finding the maximum over k of $L_{cv}^{(k)}$ will on average select the model which is closest (in cross-entropy distance) to the true density f.

APPENDIX C
Robustness with Respect to Preprocessing

CW performed their hierarchical clustering by subsampling the days (choosing every 5th day) and also by using *filtered* anomalies. In contrast, in the experiments described in the main text we used *unfiltered* anomalies and all of the days. To test the sensitivity of the results to these modest changes in the data, we fitted the 3-component Gaussian mixture model in the 2-D EOF space to 3 different permutations of the original data described above; i) unfiltered anomalies for every 5th day; ii) filtered anomalies for all days; and iii) filtered anomalies for every 5th day. The EOFs onto which these 3 additional data sets were projected were not changed from the basic experiments in Section 3c, i.e., they were determined using all 3960 filtered daily maps.

For each type of data we obtained the height-anomaly maps corresponding to the mean of each Gaussian component density. The correlation coefficient between these maps and the corresponding map obtained using unfiltered anomalies for all days (as in Fig. 8) was then calculated. The results are shown in Table 6 and clearly indicate that the maps obtained using any of these methods are all virtually identical.

References

Benzi, R., P. Malguzzi, A. Speranza, and A. Sutera, 1986: The statistical properties of general atmospheric circulation: Observational evidence and a minimal theory of bimodality. *Quart. J. R. Meteorol. Soc.*, **112**, 661–674.

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

Chickering, D. M., and D. Heckerman, 1997: Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. *Machine Learning*, in press.

Crutcher, H. L. and R. L. Joiner, 1977: Another look at the upper winds of the tropics. *J. Appl. Meteor.*, **16**, 462–476.

Crutcher, H. L., C. J. Neumann, and J. M. Pelissier, 1982: Tropical cyclone forecast errors and the multimodal bivariate normal distribution. *J. Appl. Meteor.*, **21**, 978–987.

Daley, R., 1991: *Atmospheric Data Analysis*, Cambridge Univ. Press, Cambridge, UK, 457 pp.

Davies, D. L., and D. W. Bouldin, 1979; A cluster separation measure. *IEEE Trans. Patt. Anal. Mach. Int.*, **1**, 224–227.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, **39**, 1–38.

Diaconis, P., and D. Freedman, 1984: Asymptotics of graphical projection pursuit. *Ann. Statist.*, **12**, 793–815.

Diday, E. and J. C. Simon, 1976: Cluster analysis. *Digital Pattern Recognition*, K. S. Fu (ed.), Springer-Verlag, Berlin, 47–94.

Dole, R. M., and N. M. Gordon, 1983: Persistent anomalies of the extratropical Northern Hemisphere winter time circulation: geographical distribution and regional persistence characteristics. *Mon. Wea. Rev.*, **111**, 1567–1586.

Duda, R. O., and P. E. Hart, 1973: *Pattern Recognition and Scene Analysis*, John Wiley and Sons, New York.

Edlund, S. B., 1997: *Methods for Cluster Analysis with Applications to Large NASA Data Sets.* M.Sc. Thesis, Technical Report TRITA-NA-E9720, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm.

Everitt, B. S., and D. J. Hand, 1981: *Finite Mixture Distributions*, Chapman and Hall, London.

Ghil, M. 1987: Dynamics, statistics, and predictability of planetary flow regimes. *Irreversible Phenomena and Dynamical Systems Analysis in Geosciences*, C. Nicolis and G. Nicolis, Eds., Reidel, 241–283.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Adv. Geophys.*, **33**, 141–266.

Hannachi, A., and B. Legras, 1995: Simulated annealing and weather regimes classification. *Tellus*, **47**, 955–973.

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

Hurrel, JW, 1995: Decadal trends in the North Atlantic Oscillation: Regional temperature and precipitation. *Science*, **269**, 676–679.

Jain, A. K. and R. C. Dubes, 1988: *Algorithms for Clustering Data.* Prentice Hall, New Jersey.

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

Kimoto, M., and M. Ghil, 1993b: Multiple flow regimes in the Northern Hemisphere winter: Part II: Sectorial regimes and preferred transitions. *J. Atmos. Sci.*, **50**, 2645–2673.

Legras, B., and M. Ghil, l985: Persistent anomalies, blocking and variations in atmospheric predictability. *J. Atmos. Sci.*, **42**, 433–47l.

Legras, B., T. Desponts, and B. Piguet, 1988: Cluster analysis and weather regimes. *Proc. ECMWF Workshop on "The Nature and Prediction of Extratropical Weather Systems."* Vol. II. European Centre for Medium Range Weather Forecasts, Reading, U.K., pp. 123–149.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646.

Marcus, S. L., M. Ghil and J. O. Dickey, 1996: The extratropical 40-day oscillation in the UCLA general circulation model. Part II: Spatial structure. *J. Atmos. Sci.*, **53**, 1993–2014.

McLachlan, G. J., and K. E. Basford, 1988: *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

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

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

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

Namias, J., 1982: *Short Period Climatic Variations. Collected works of J. Namias, Vols. I and II (1934-1974) and Vol. III (1975-1982)*, Univ. of California, San Diego, 905 pp. + 393 pp.

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

Preisendorfer, R. W., 1988: *Principal Component Analysis in Meteorology and Oceanography*, C. D. Mobley (Ed.), Elsevier, New York, 425 pp.

Rex, D. F., 1950a: Blocking action in the middle troposphere and its effect on regional climate. I. An aerological study of blocking action. *Tellus*, **2**, 196–211.

Rex, D. F., 1950b: Blocking action in the middle troposphere and its effect on regional climate. II. The climatology of blocking action. *Tellus*, **2**, 275–301.

Robert, C. P., 1996: Mixtures of distributions: inference and estimation. *Markov Chain Monte Carlo in Practice*, Ch. 24, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), Chapman & Hall, London, pp. 441–461.

Robertson, A., M. Ghil, and M. Latif, 1997: Interdecadal changes in atmospheric low-frequency variability with and without boundary forcing, *J. Atmos. Sci.*, submitted.

Shao, J., 1993: Linear model selection by cross-validation. *J. Am. Stat. Assoc.*, **88**(422), 486–494.

Smyth, P., 1996: Clustering using Monte-Carlo cross-validation. *Proc. 1996 Knowledge Discovery & Data Mining Conf.*, Menlo Park, CA, AAAI Press, 126–133.

Smyth, P., M. Ghil, K. Ide, J. Roden, and A. Fraser, 1997: Detecting atmospheric regimes using cross-validated clustering. *Proc. 1997 Knowledge Discovery & Data Mining Conf.*, Newport Beach, CA, AAAI Press, 61–66.

Stone, M., 1974: Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. B*, **36**, 111–147.

Strong, C. M., F.-f. Jin and M. Ghil, 1995: Intraseasonal oscillations in a barotropic model with annual cycle, and their predictability. *J. Atmos. Sci.*, **52**, 2627–2642.

Tian, Y., 1997: *Eastward Jet over topography: Experimental and numerical investigations.*, Ph.D. Thesis, University of California, Los Angeles, 50pp.

Titterington, D. M., A. F. M. Smith, and U. E. Makov, 1985: *Statistical Analysis of Finite Mixture Distributions*, J. Wiley & Sons, Chichester, UK.

Wallace, J. M., 1996: Observed climatic variability: Spatial structure. *Decadal Climate Variability: Dynamics and Predictability*, D. Anderson and J. Willebrand (Eds.), Elsevier, pp. 31–81.

Weeks, E. R., Y. Tian, J. S. Urbach, K. Ide, H. Swinney, and M. Ghil, 1997: Transition beteen blocked and zonal flows in a rotating annulus with topography. *Science*, in press.

Zhang, P., 1993: Model selection via multifold cross validation. *Ann. Statist.*, **21**, 299–313.

Table Captions

TABLE 1. Cross-validated log-likelihood $L_{cv}^{(k)}$ and posterior probabilities $\hat{P}(k|D_\beta)$, as a function of the number k of Gaussian clusters, when applying cross-validation to the mixture modeling algorithm (with $\beta = 0.5, M = 20$) for the 600 synthetic data points shown in Fig. 1.

TABLE 2. Cross-validated log-likelihood $L_{cv}^{(k)}$ and estimated posterior probabilities, as a function of k, when applying the mixture model to 20 random partitions of the 44 winters of NH 700-mb geopotential height anomalies (unscaled).

TABLE 3. Out-of-sample log-likelihoods for each of $M = 20$ random partitions of 44 unscaled winter anomalies, normalized so that the log-likelihood for $k = 3$ is zero on each run. The most likely value (of log-likelihood and hence of k) is displayed in bold font for each partition.

TABLE 4. Cross-validated log-likelihood values as a function of k and the estimated posterior probability of $k = 3$ from 10 different experiments, each using $M = 20$ randomly chosen partitions of the 44 winters.

TABLE 5. Pattern correlation coefficients between maps fitted using the data projected onto d EOFs, $3 \leq d \leq 12$, and maps fitted using $d = 2$ EOFs. The maps correspond to centers of a mixture model based on 3 Gaussians, fitted by the EM procedure (see section 3b) as applied to all of the data in the d-dimensional EOF subspace.

TABLE 6. Pattern correlation coefficients between maps fitted on unfiltered anomalies for all 3960 days and (a) filtered anomalies, (b) unfiltered anomalies using only every 5th day, and (c) filtered anomalies using only every 5th day. All maps were fitted to the data projected into the first 2 EOFs.

Tables

TABLE 1: Cross-validated log-likelihood $L_{cv}^{(k)}$ and posterior probabilities $\hat{P}(k|D_\beta)$, as a function of the number k of Gaussian clusters, when applying cross-validation to the mixture modeling algorithm (with $\beta = 0.5, M = 20$) for the 600 synthetic data points shown in Fig. 1.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Cross-validated log-likelihood	-1287.9	-1258.8	-1249.5	-1251.0	-1253.4	-1256.1
Estimated posterior probability	0.000	0.000	0.809	0.175	0.015	0.001

TABLE 2: Cross-validated log-likelihood $L_{cv}^{(k)}$ and estimated posterior probabilities, as a function of k, when applying the mixture model to 20 random partitions of the 44 winters of NH 700-mb geopotential height anomalies (unscaled).

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Cross-validated log-likelihood	-29164	-29153	-29137	-29148	-29156	-29165
Estimated posterior probability	0.0	0.0	1.0	0.0	0.0	0.0

TABLE 3: Out-of-sample log-likelihoods for each of $M = 20$ random partitions of 44 unscaled winter anomalies, normalized so that the log-likelihood for $k = 3$ is zero on each run. The most likely value (of log-likelihood and hence of k) is displayed in bold font for each partition.

Partition	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
1	-45.191	-12.149	**0.000**	-26.448	-29.415	-40.189
2	-37.846	-22.652	**0.000**	-7.408	-26.484	-32.172
3	-57.364	-27.430	**0.000**	-3.884	-0.930	-11.788
4	-21.119	**1.347**	0.000	-7.309	-30.879	-20.495
5	**10.010**	-9.627	0.000	-11.397	-20.534	-26.423
6	-13.887	-3.715	**0.000**	-14.096	-15.132	-18.936
7	-27.765	**5.486**	0.000	-18.068	-19.443	-37.717
8	-38.394	-23.947	**0.000**	-24.390	-50.935	-61.827
9	-17.916	-21.546	**0.000**	-1.403	-18.528	-34.125
10	-35.180	-17.886	**0.000**	-11.161	-14.403	-20.805
11	-32.176	-25.935	**0.000**	-7.085	-6.152	-8.757
12	-45.422	-14.198	**0.000**	-27.905	-20.066	-20.023
13	-34.579	-3.821	**0.000**	-9.015	-13.695	-11.574
14	-73.393	-32.027	**0.000**	-10.719	-6.973	-15.963
15	-23.255	**3.829**	0.000	3.651	-7.438	-8.352
16	-37.655	-14.835	**0.000**	-5.341	-11.913	-26.378
17	-26.943	-12.028	**0.000**	-16.692	-31.922	-37.397
18	-47.595	-21.178	**0.000**	-8.777	-8.039	-12.653
19	-25.255	**19.984**	0.000	-0.826	-2.744	-8.372
20	-59.862	-25.430	**0.000**	-7.868	-10.584	-32.612

TABLE 4: Cross-validated log-likelihood values as a function of k and the estimated posterior probability of $k = 3$ from 10 different experiments, each using $M = 20$ randomly chosen partitions of the 44 winters.

Experiment	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$\hat{P}(k = 3)$
1	-27831	-27815	**-27808**	-27817	-27827	-27833	1.000
2	-27858	-27835	**-27827**	-27835	-27837	-27844	1.000
3	-27819	-27799	**-27789**	-27799	-27802	-27814	1.000
4	-27843	-27825	**-27812**	-27821	-27829	-27837	1.000
5	-27825	-27808	**-27801**	-27810	-27817	-27824	0.999
6	-27864	-27846	**-27839**	-27846	-27854	-27860	0.999
7	-27811	-27792	**-27783**	-27792	-27797	-27807	1.000
8	-27818	-27805	**-27787**	-27801	-27807	-27811	1.000
9	-27844	-27824	**-27803**	-27817	-27817	-27822	1.000
10	-27856	-27837	**-27829**	-27841	-27845	-27853	1.000

TABLE 5: Pattern correlation coefficients between maps fitted using the data projected onto d EOFs, $3 \leq d \leq 12$, and maps fitted using $d = 2$ EOFs. The maps correspond to centers of a mixture model based on 3 Gaussians, fitted by the EM procedure (see section 3b) as applied to all of the data in the d-dimensional EOF subspace.

d	r_1	r_2	r_3
3	0.978	0.961	0.998
4	0.974	0.960	0.999
5	0.947	0.957	0.976
6	0.946	0.946	0.957
7	0.945	0.951	0.945
8	0.931	0.946	0.938
9	0.938	0.953	0.941
10	0.946	0.951	0.949
11	0.927	0.943	0.934
12	0.945	0.946	0.935

TABLE 6: Pattern correlation coefficients between maps fitted on unfiltered anomalies for all 3960 days and (a) filtered anomalies, (b) unfiltered anomalies using only every 5th day, and (c) filtered anomalies using only every 5th day. All maps were fitted to the data projected into the first 2 EOFs.

Type of Data	r_1	r_2	r_3
Unfiltered anomalies, every 5th day	0.9810	0.9959	0.9990
Filtered anomalies, all days	0.9948	0.9755	0.9966
Filtered anomalies, every 5th day	0.9922	0.9727	0.9975

setup for Figure Captions

Figure Captions

FIG. 1. Scatter plot of 600 data points generated from a mixture of 3 equally weighted Gaussian densities, having distinct means and covariances.

FIG. 2. The *true* means of component Gaussians (shown as stars) and the associated covariance matrices, indicated by the corresponding ellipses (see text for details), superimposed on the scatter plot of Fig. 1.

FIG. 3. Contour plot of the probability density function (PDF) corresponding to the mixture model displayed in Fig. 2.

FIG. 4. The *estimated* means and covariance ellipses of component Gaussians superimposed on the scatter plot in Fig. 1. Parameters were estimated using the expectation-maximization (EM) procedure (see text for details).

FIG. 5. Scatter plot of NH height anomalies for 44 winters (December 1949–March 1993), projected onto the 2 leading EOFs; the data have been normalized by dividing by the standard deviation of EOF 1.

FIG. 6. The estimated means, indicated by *asterisks*, and covariance ellipses superimposed on the scatter plot of Fig. 5, where only every 10th data point has been plotted for clarity. The identities of the clusters – **A**, **G**, and **R** – are indicated beside the respective ellipses. The parameters were estimated by the same EM procedure as for the synthetic data, using a mixture model based on 3 Gaussian components. The estimated parameters for the 3 clusters are: (**A**) $\hat{\alpha}_A = 0.47$, $\hat{\mu}_A = (-0.59, 0.10)$, $\tan(\hat{\psi}_A) = 0.20$, $\hat{\lambda}_{A1} = 0.78$, $\hat{\lambda}_{A2} = 0.47$; (**G**) $\hat{\alpha}_G = 0.15$, $\hat{\mu}_G = (0.32, -1.34)$, $\tan(\hat{\psi}_G) = 0.34$, $\hat{\lambda}_{G1} = 0.82$, $\hat{\lambda}_{G2} = 0.24$; and (**R**) $\hat{\alpha}_R = 0.38$, $\hat{\mu}_R = (0.64, 0.43)$, $\tan(\hat{\psi}_R) = -0.71$, $\hat{\lambda}_{R1} = 0.56$, $\hat{\lambda}_{R2} = 0.36$. Here α is the weight assigned to the cluster in the mixture model, μ is the mean for the cluster, ψ is the

rotation angle (anti-clockwise) from the x-axis of the first eigenvector for the covariance matrix of each cluster, and the λ's correspond to the two eigenvalues of the covariance matrix.

FIG. 7. Contour plot of the PDF estimate provided by the mixture model of Fig. 6; the asterisks and associated labels for the 3 clusters (**A**, **G**, and **R**) indicate the corresponding centroids.

FIG. 8. Height anomaly maps for the 3 cluster centroids of the present mixture model (left: panels a, c and e; labeled SGI) and of CW's hierarchical cluster model, as applied by Wallace (1996) to a slightly longer data set (right: panels b, d and f; labeled CW). Pairs of maps (a, b) correspond to CW's cluster **A**, (c, d) to **G**, and (e, f) to **R** (see text for details; panels b, d and f reproduced by permission). Contour interval is 15 m for panels on the left (SGI) and 50 m for panels on the right (CW).

FIG. 9. Same as Fig. 6 for the Pacific sector height anomalies projected onto the 2 leading Pacific sector EOFs. The identities of the two clusters, *RNA* and *PNA*, are indicated beside the respective means, plotted as *asterisks*. The estimated parameters for the 2 clusters are: (*RNA*) $\hat{\alpha}_{RNA} = 0.55$, $\hat{\mu}_{RNA} = (-0.43, 0.24)$, $\tan(\hat{\psi}_{RNA}) = 0.07$, $\hat{\lambda}_1 = 0.90$, $\hat{\lambda}_2 = 0.71$; (*PNA*) $\hat{\alpha}_{PNA} = 0.45$, $\hat{\mu}_{PNA} = (0.56, -0.28)$, $\tan(\hat{\psi}_{PNA}) = 0.61$, $\hat{\lambda}_1 = 0.70$, $\hat{\lambda}_2 = 0.28$. Here α, μ and ψ are defined as in Fig. 6.

FIG. 10. Height anomaly maps for the clusters found by the mixture model from the PAC and ATL sectorial analysis. Contour interval is 15m. The maps in panels (a) and (b) resemble the PNA and RNA patterns respectively, while panels (c) and (d) resemble the ZNAO and BNAO patterns respectively (see text for details).

FIG. 11. Same as Fig. 6 for the Atlantic sector height anomalies projected onto the 2 leading Atlantic sector EOFs. The identities of the two clusters, *ZNAO* and *BNAO*, are indicated beside the respective means, plotted as *asterisks*. The estimated parameters for

the 2 clusters are: $(ZNAO)$ $\hat{\alpha}_{ZNAO} = 0.86$, $\hat{\mu}_{ZNAO} = (-0.04, 0.24)$, $\tan(\hat{\psi}_{ZNAO}) = 0.15$, $\hat{\lambda}_1 = 0.99$, $\hat{\lambda}_2 = 0.61$; $(BNAO)$ $\hat{\alpha}_{BNAO} = 0.14$, $\hat{\mu}_{BNAO} = (0.21, -1.43)$, $\tan(\hat{\psi}_{BNAO}) = 0.24$, $\hat{\lambda}_1 = 0.99$, $\hat{\lambda}_2 = 0.61$.

Figures

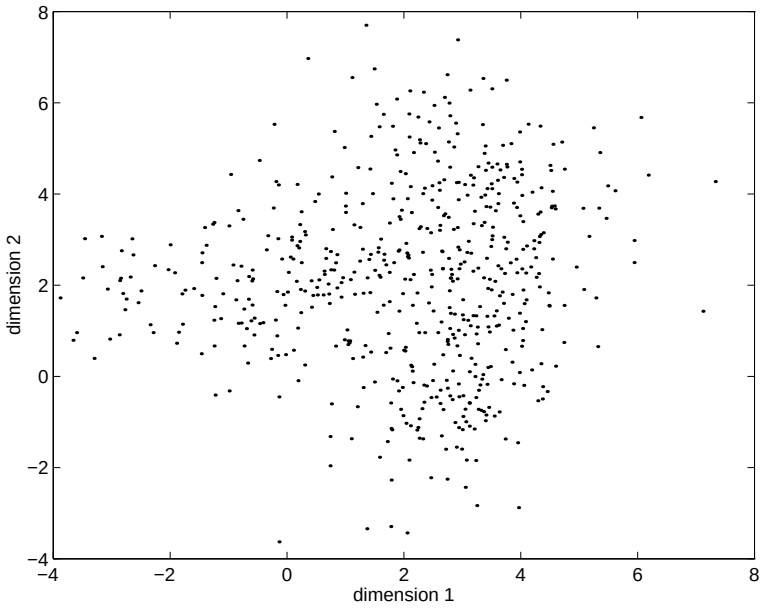

FIG. 1: Scatter plot of 600 data points generated from a mixture of 3 equally weighted Gaussian densities, having distinct means and covariances.

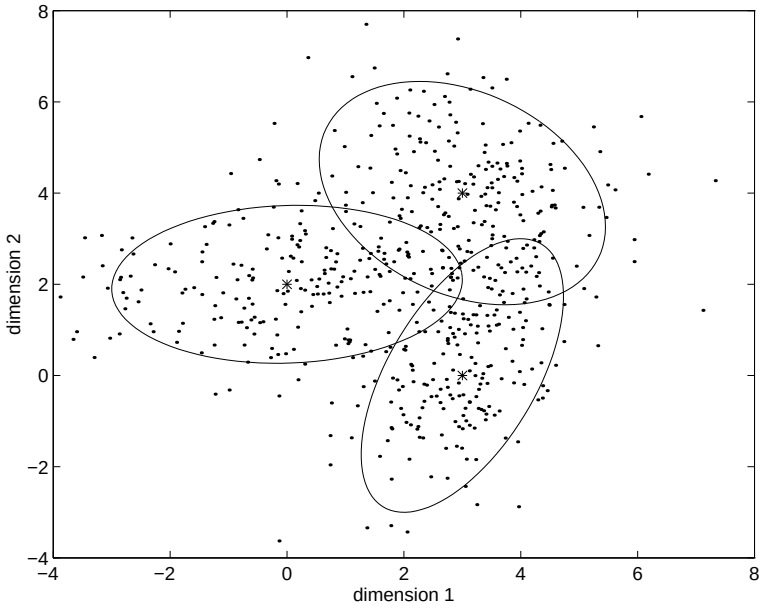

FIG. 2: The *true* means of component Gaussians (shown as stars) and the associated covariance matrices, indicated by the corresponding ellipses (see text for details), superimposed on the scatter plot of Fig. 1.

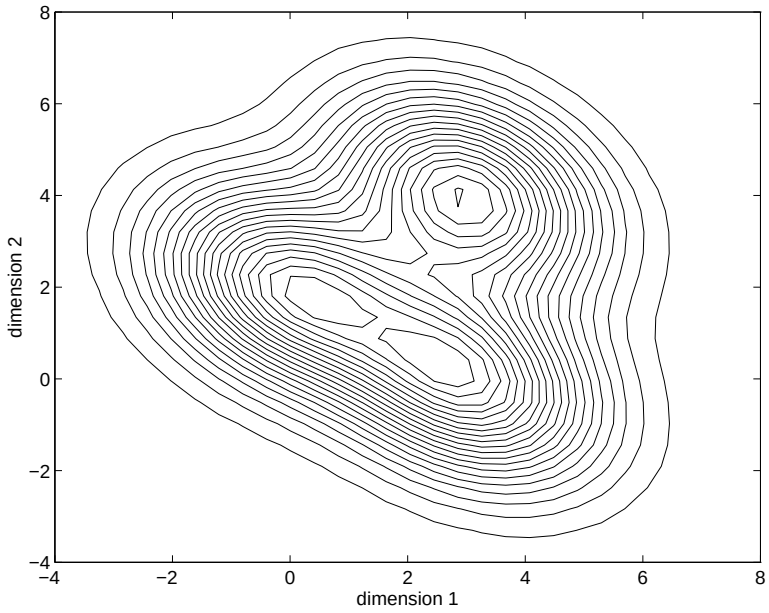

FIG. 3: Contour plot of the probability density function (PDF) corresponding to the mixture model displayed in Fig. 2.

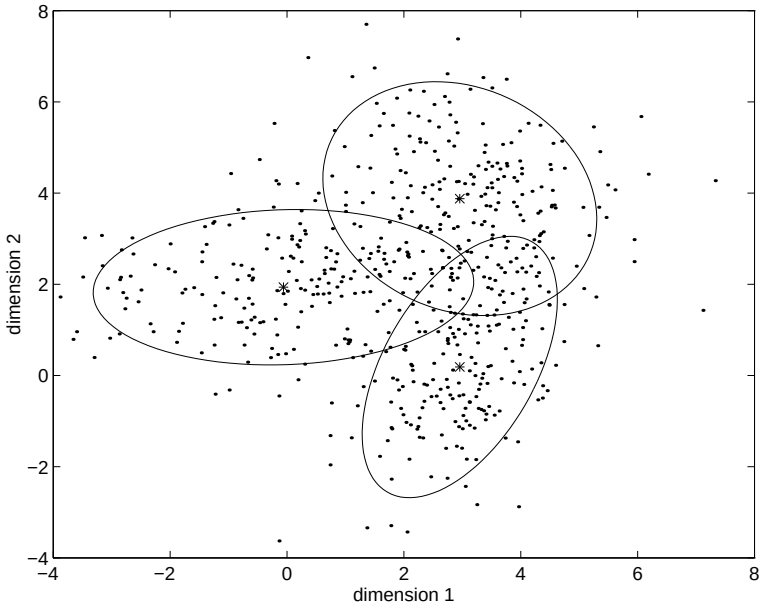

FIG. 4: The *estimated* means and covariance ellipses of component Gaussians superimposed on the scatter plot in Fig. 1. Parameters were estimated using the expectation-maximization (EM) procedure (see text for details).

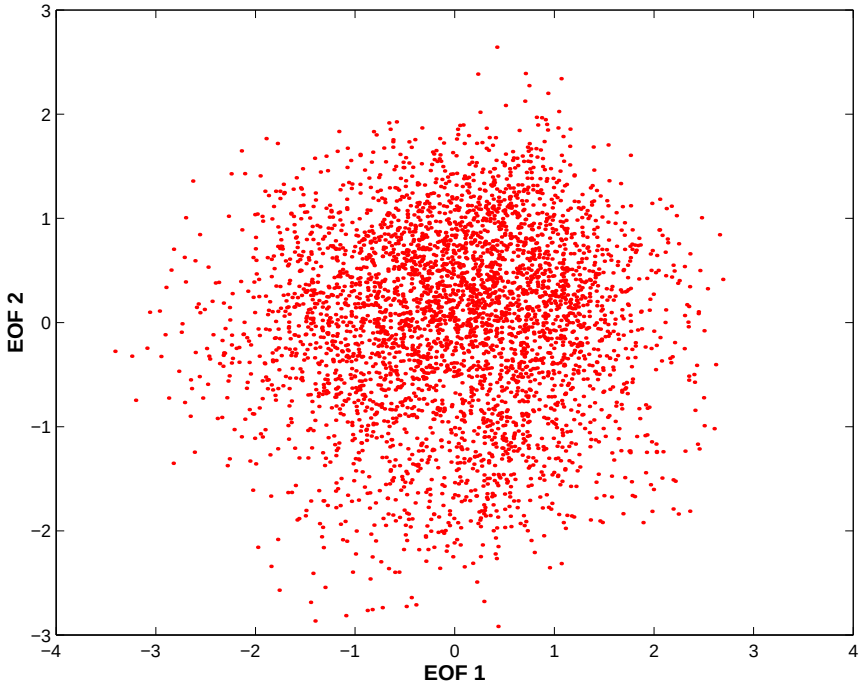

FIG. 5: Scatter plot of NH height anomalies for 44 winters (December 1949–March 1993), projected onto the 2 leading EOFs; the data have been normalized by dividing by the standard deviation of EOF 1.

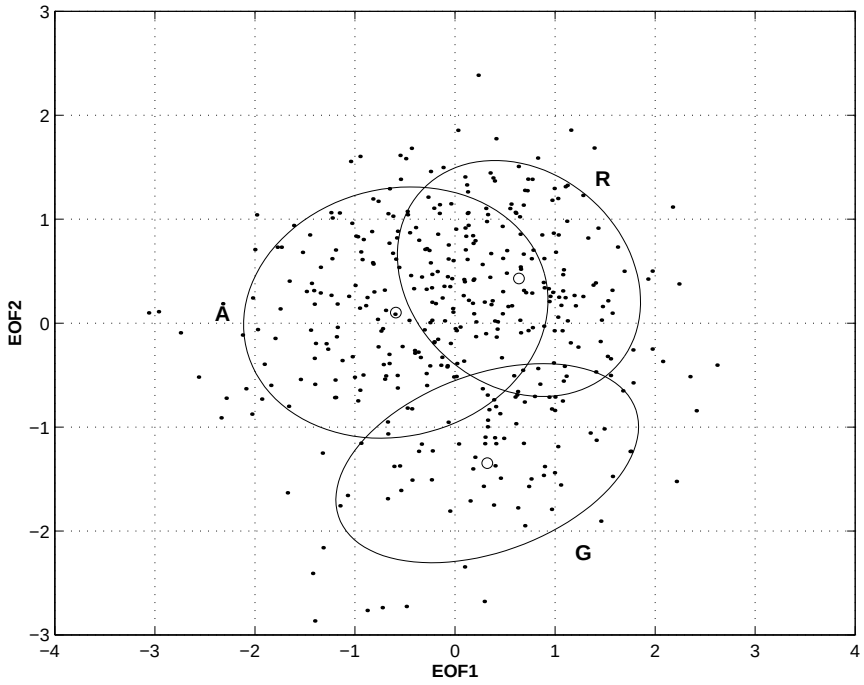

FIG. 6: The estimated means, indicated by *asterisks*, and covariance ellipses superimposed on the scatter plot of Fig. 5, where only every 10th data point has been plotted for clarity. The identities of the clusters – **A**, **G**, and **R** – are indicated beside the respective ellipses. The parameters were estimated by the same EM procedure as for the synthetic data, using a mixture model based on 3 Gaussian components. The estimated parameters for the 3 clusters are: (**A**) $\hat{\alpha}_A = 0.47$, $\hat{\mu}_A = (-0.59, 0.10)$, $\tan(\hat{\psi}_A) = 0.20$, $\hat{\lambda}_{A1} = 0.78$, $\hat{\lambda}_{A2} = 0.47$; (**G**) $\hat{\alpha}_G = 0.15$, $\hat{\mu}_G = (0.32, -1.34)$, $\tan(\hat{\psi}_G) = 0.34$, $\hat{\lambda}_{G1} = 0.82$, $\hat{\lambda}_{G2} = 0.24$; and (**R**) $\hat{\alpha}_R = 0.38$, $\hat{\mu}_R = (0.64, 0.43)$, $\tan(\hat{\psi}_R) = -0.71$, $\hat{\lambda}_{R1} = 0.56$, $\hat{\lambda}_{R2} = 0.36$. Here α is the weight assigned to the cluster in the mixture model, μ is the mean for the cluster, ψ is the rotation angle (anti-clockwise) from the x-axis of the first eigenvector for the covariance matrix of each cluster, and the λ's correspond to the two eigenvalues of the covariance matrix.

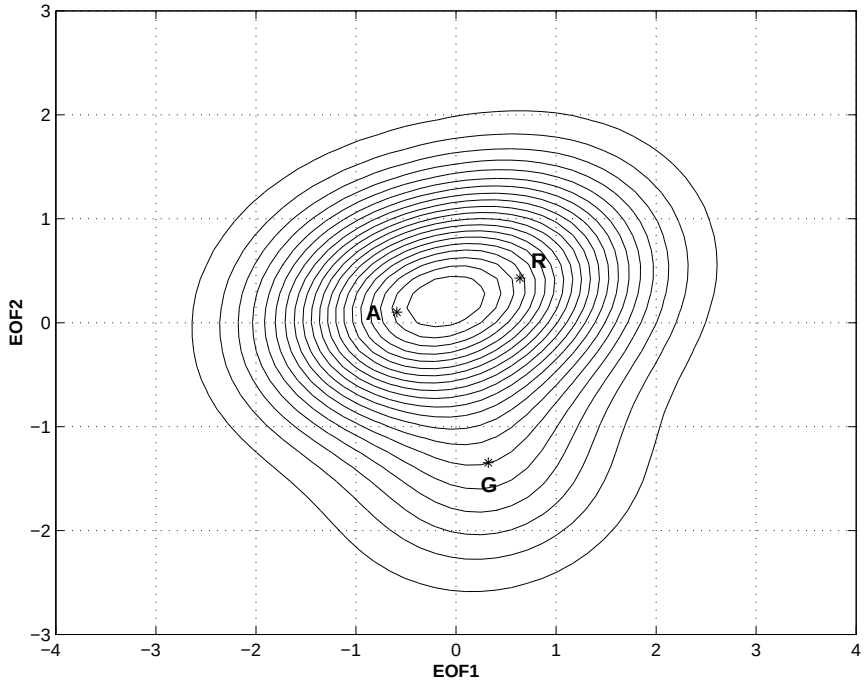

FIG. 7: Contour plot of the PDF estimate provided by the mixture model of Fig. 6; the asterisks and associated labels for the 3 clusters (**A**, **G**, and **R**) indicate the corresponding centroids.

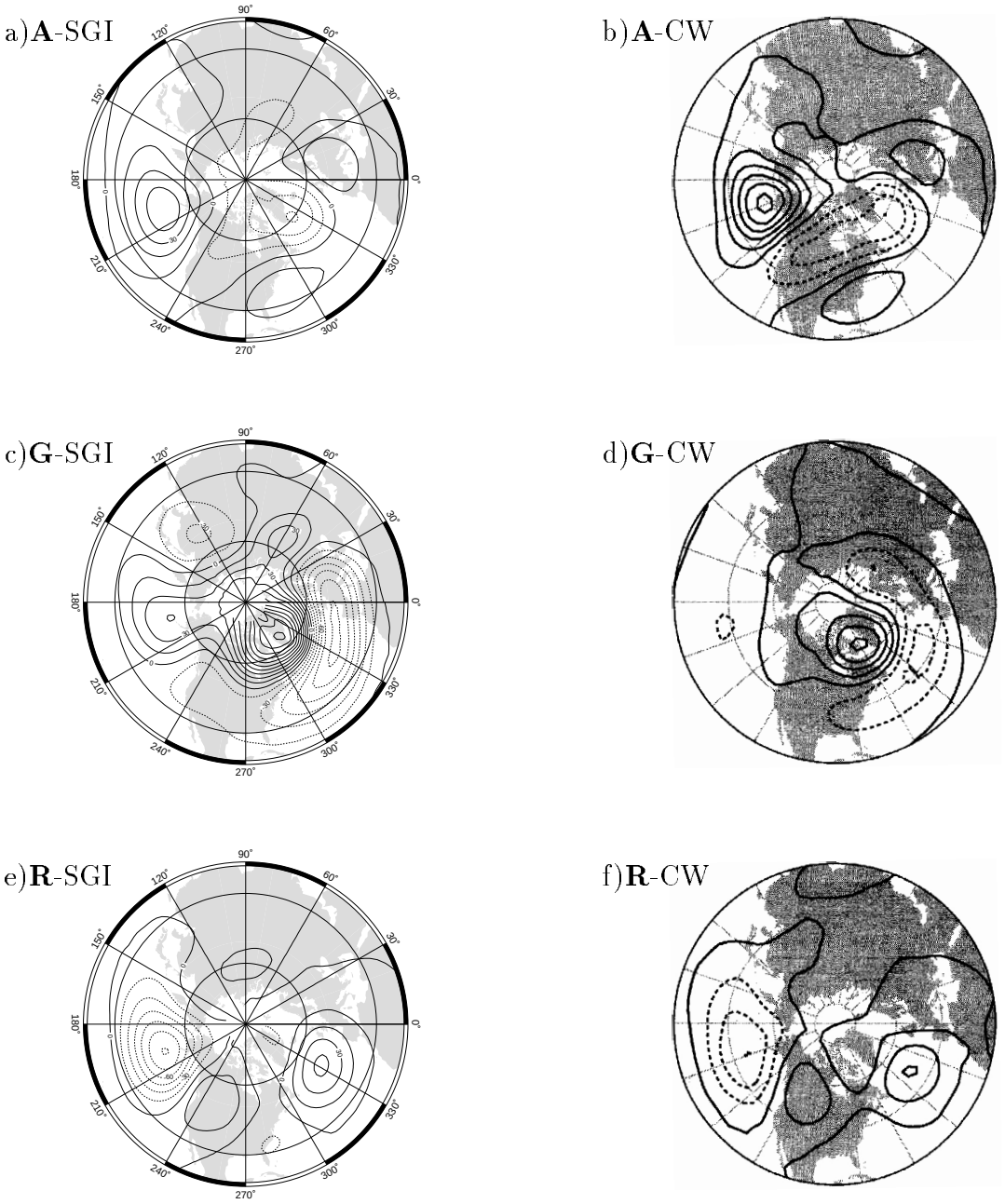

FIG. 8: Height anomaly maps for the 3 cluster centroids of the present mixture model (left: panels a, c and e; labeled SGI) and of CW's hierarchical cluster model, as applied by Wallace (1996) to a slightly longer data set (right: panels b, d and f; labeled CW). Pairs of maps (a, b) correspond to CW's cluster **A**, (c, d) to **G**, and (e, f) to **R** (see text for details; panels b, d and f reproduced by permission). Contour interval is 15 m for panels on the left (SGI) and 50 m for panels on the right (CW).

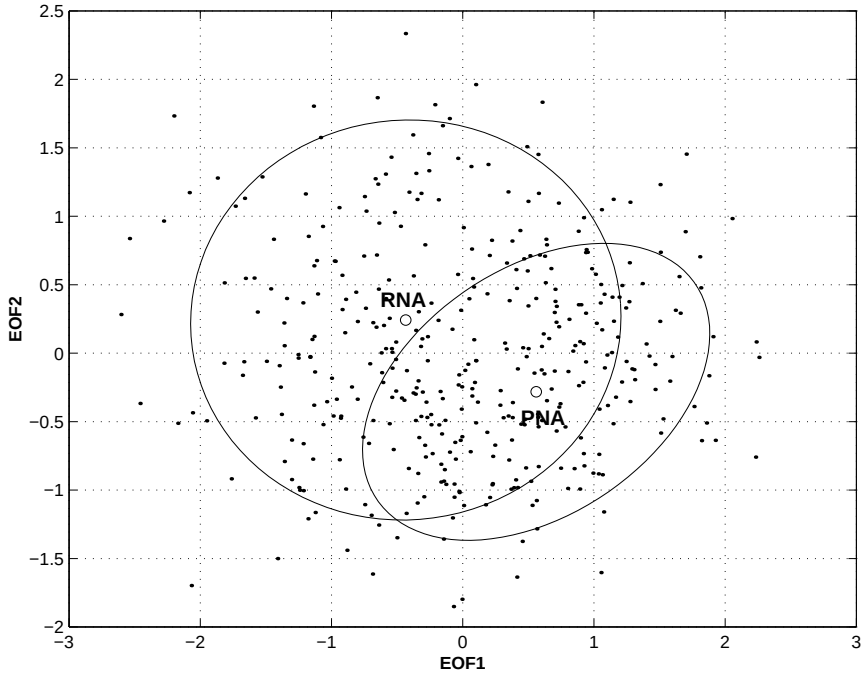

FIG. 9: Same as Fig. 6 for the Pacific sector height anomalies projected onto the 2 leading Pacific sector EOFs. The identities of the two clusters, RNA and PNA, are indicated beside the respective means, plotted as *asterisks*. The estimated parameters for the 2 clusters are: (RNA) $\hat{\alpha}_{RNA} = 0.55$, $\hat{\mu}_{RNA} = (-0.43, 0.24)$, $\tan(\hat{\psi}_{RNA}) = 0.07$, $\hat{\lambda}_1 = 0.90$, $\hat{\lambda}_2 = 0.71$; (PNA) $\hat{\alpha}_{PNA} = 0.45$, $\hat{\mu}_{PNA} = (0.56, -0.28)$, $\tan(\hat{\psi}_{PNA}) = 0.61$, $\hat{\lambda}_1 = 0.70$, $\hat{\lambda}_2 = 0.28$. Here α, μ and ψ are defined as in Fig. 6.

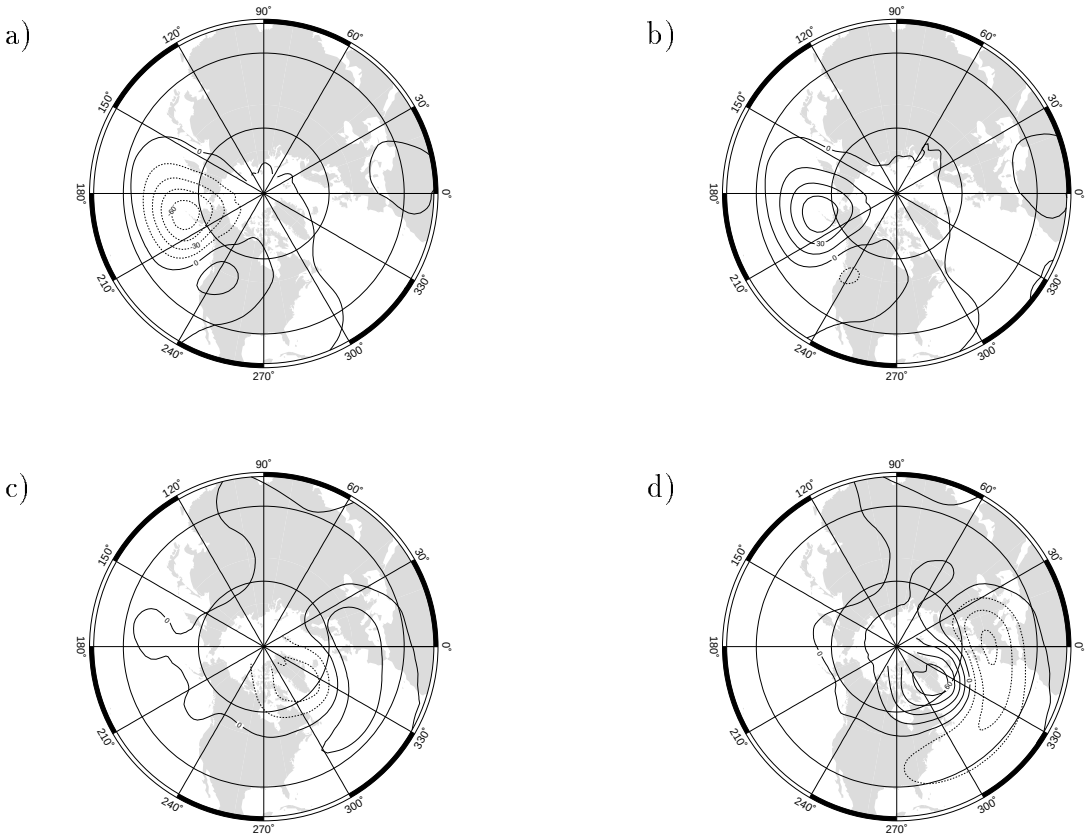

FIG. 10: Height anomaly maps for the clusters found by the mixture model from the PAC and ATL sectorial analysis. Contour interval is 15m. The maps in panels (a) and (b) resemble the PNA and RNA patterns respectively, while panels (c) and (d) resemble the ZNAO and BNAO patterns respectively (see text for details).

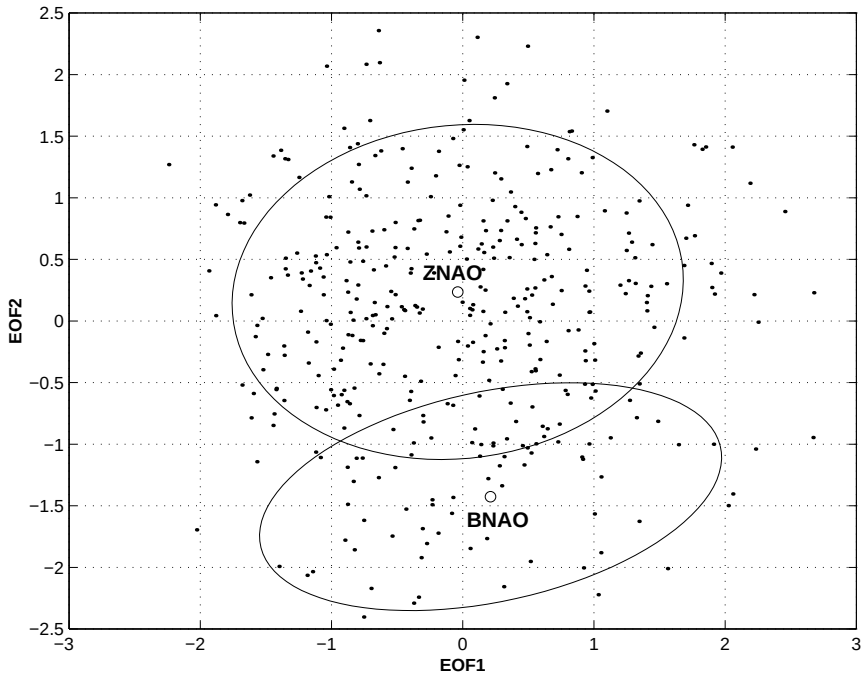

FIG. 11: Same as Fig. 6 for the Atlantic sector height anomalies projected onto the 2 leading Atlantic sector EOFs. The identities of the two clusters, $ZNAO$ and $BNAO$, are indicated beside the respective means, plotted as *asterisks*. The estimated parameters for the 2 clusters are: $(ZNAO)$ $\hat{\alpha}_{ZNAO} = 0.86$, $\hat{\mu}_{ZNAO} = (-0.04, 0.24)$, $\tan(\hat{\psi}_{ZNAO}) = 0.15$, $\hat{\lambda}_1 = 0.99$, $\hat{\lambda}_2 = 0.61$; $(BNAO)$ $\hat{\alpha}_{BNAO} = 0.14$, $\hat{\mu}_{BNAO} = (0.21, -1.43)$, $\tan(\hat{\psi}_{BNAO}) = 0.24$, $\hat{\lambda}_1 = 0.99$, $\hat{\lambda}_2 = 0.61$.