Multiple Regimes in Northern Hemisphere Height Fields via Mixture Model Clustering*

PADHRAIC SMYTH+

Department of Information and Computer Science, University of California, Irvine, Irvine, California

KAYO IDE AND MICHAEL GHIL

Department of Atmospheric Sciences and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, Los Angeles, California

(Manuscript received 18 December 1997, in final form 21 December 1998)

ABSTRACT

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

A dataset 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 three clusters in the NH 700-mb data. The three-cluster result is found to be robust with respect to variations in data preprocessing and data analysis parameters. The spatial patterns of the three clusters' centroids bear a high degree of qualitative similarity to the three clusters obtained independently by Cheng and Wallace, using hierarchical clustering on 500-mb NH winter data: the Gulf of Alaska ridge, the high over southern Greenland, and the enhanced climatological ridge over the Rockies.

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 Kimoto and Ghil's Pacific–North American (PNA) and reverse PNA regimes, as well as the zonal and blocked phases of the North Atlantic oscillation. The connections between our sectorial and hemispheric results are discussed from the perspective of large-scale atmospheric dynamics.

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 regimelike 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 dataset composed 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 1982a,b) 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

* UCLA's Institute of Geophysics and Planetary Physics Publication Number 5103.

+ Additional affiliation: Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California.

Corresponding author address: Dr. M. Ghil, Department of Atmospheric Sciences, University of California, Los Angeles, Los Angeles, CA 90095-1565.
E-mail: ghil@atmos.ucla.edu

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 (1949–92) of 700-mb height maps, classified separately over the Atlantic–European (ATL) and Pacific–North American (PAC) sectors. They proposed the use of a classifiability index that 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 datasets was obtained by Cheng and Wallace (1993; CW hereafter), who applied hierarchical clustering (see also Legras et al. 1988) to 40 NH winters (1946–85) 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 (1949–86) 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 dataset, 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, bear 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 the following two primary features.

1) Each cluster is defined as a unimodal (''component'') PDF. Thus, points that 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 dataset 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.

Any statistical method that directly estimates a PDF in phase space will inevitably be limited in its results by the number of available data points, particularly as one seeks to fit the model on which the method is based to the data in higher dimensions. This ''curse of dimensionality'' is well-known in the statistical literature (e.g., see Silverman 1986) and has been discussed at length in the present context by KGI. In general, the results obtained from any finite set of observational data (in this case, 44 winters) must be interpreted as being conditional on the observed data; our method only makes this dependence more explicit.

The paper is organized as follows. In section 2 the 700-mb dataset 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 three 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 three clusters that we obtain by mixture modeling are compared to the three significant maps obtained 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 essentially two 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 blocked and zonal phases of the North Atlantic oscillation (NAO). We also show in this section how the methodology can alert the data analyst to the size of the dataset being insufficient to support a nontrivial model. 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. Dataset

The dataset 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 the National Oceanic and Atmospheric Administration's Climate Prediction Center. The only difference between the two datasets is that in this paper 44 winters are used, starting on 1 December 1949 and extending through 31 March 1993. Kimoto and Ghil's data began on the same date but contained only 37 winters, through March 1986. The NH winter is defined as the 90-day sequence beginning on 1 December of each year. All of 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 dataset 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 $10° \times 10°$ 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 subtracting a 5-day running mean, averaged over the 44 years; this provides 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 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 dataset (see KGI). In this manner one can reduce the dimensionality of the dataset 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.

Figure 1 shows the percent variances associated with each of the first 40 EOFs, in decreasing order of variance. This variance spectrum is quite similar in shape to that obtained by KGI (their Fig. 5), with a break in slope between the 11th and 12th eigenvalues, and a large gap between the second and third one.

The variance associated with the first EOF in our spectrum (15.6%) is larger than that associated with the first EOF in KGI's spectrum (12.6%). This slight difference and a few even less significant ones (e.g., the second, smaller gap in the spectrum being after the fourth EOF here vs the fifth there) are probably explained by the fact that KGI's analysis was performed on the 37 NH winters through March 1986, whereas here we are calculating EOFs based on the 44 winters through March 1993. The cumulative percent variance associated with the first 35 EOFs here is 97.1% compared to 96.3% in KGI. Projections used in our analyses below range from the first 2 to the first 12 EOFs. The spatial patterns associated with the first and second EOFs (not shown here) are virtually identical to those found in KGI's Fig. 6. They form the basis for much of the analysis in section 4 here.

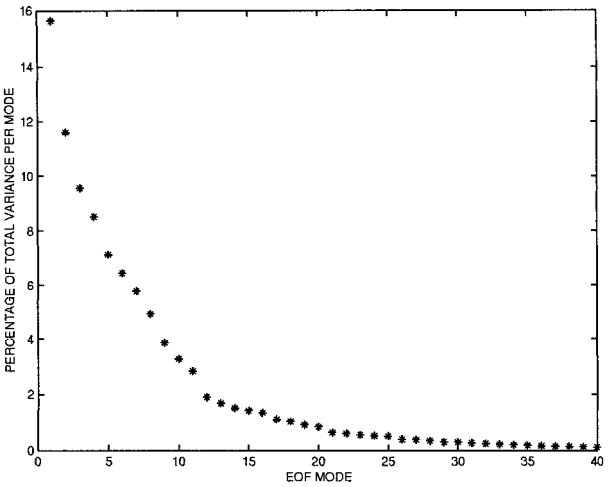

FIG. 1. Percent variance associated with the first 40 EOFs derived from 44 years of filtered NH winter anomaly data.

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 2D dataset shown in Fig. 2. These data have been generated from a mixture model containing three Gaussian components, having distinct means and covariances, with components weighted equally. The centroids of the three Gaussians and the ellipses are overlaid on the scatterplot in Fig. 3; the semimajor axes of each ellipse correspond in direction with the eigenvectors and in length with three times the singular values of the associated covariance matrix, that is, three times the corresponding standard deviations.

Figure 4 shows a contour plot of the PDF. Note the non-Gaussian, multimodal nature of this contour plot. The ability to model such multimodal 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 $\mathbf{X}$, for example, a data vector with d components. A finite mixture PDF for $\mathbf{X}$, having k components, can be written as

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

where each of the g_j is a component density function.

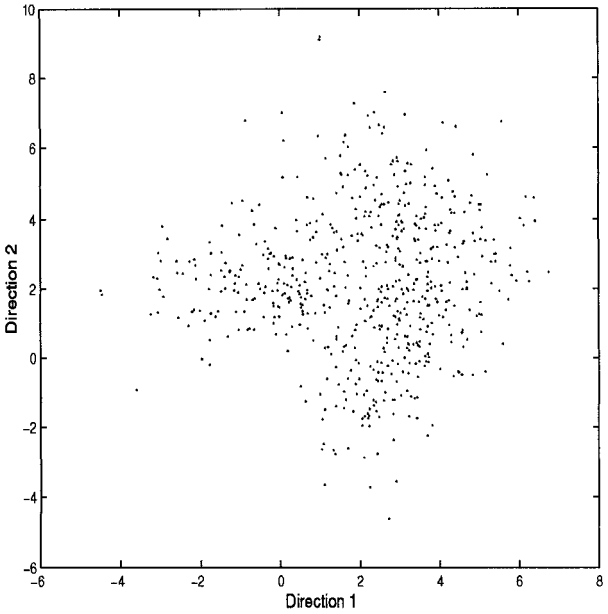

FIG. 2. Scatterplot of 600 data points generated from a mixture of three equally weighted Gaussian densities, having distinct means and covariances.

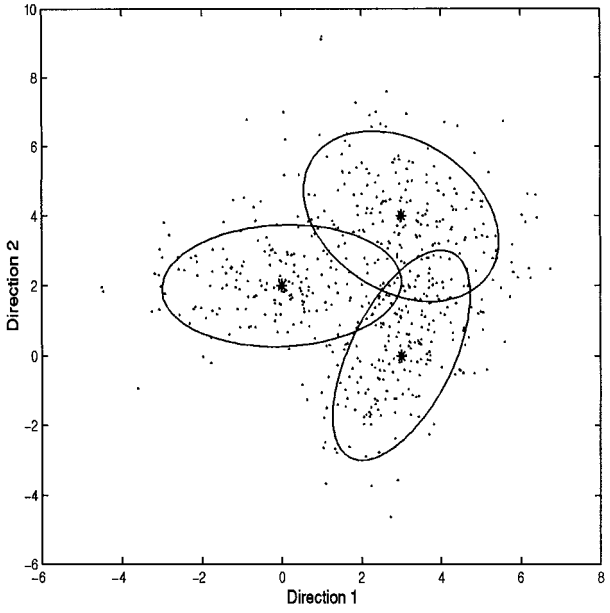

FIG. 3. The *true* means of component Gaussians (shown as stars) and the associated covariance matrices, indicated by the corresponding ellipses (see text for details), superimposed on the scatterplot of Fig. 1.

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 $\Sigma_{j=1}^{k} \alpha_j = 1$ and $\alpha_j \geq 0$, $1 \leq j \leq k$; the set of parameters for the overall mixture model is denoted by $\boldsymbol{\phi} = \{\alpha_1, \ldots, \alpha_k; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$.

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

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

where $\boldsymbol{\mu}_j$ and $\mathbf{C}_j$ are the mean and covariance matrix, respectively, and $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \mathbf{C}_j\}$. The mean $\boldsymbol{\mu}_j$ specifies the location of the jth density's centroid and the covariance matrix $\mathbf{C}_j$ 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 dataset 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 ap-

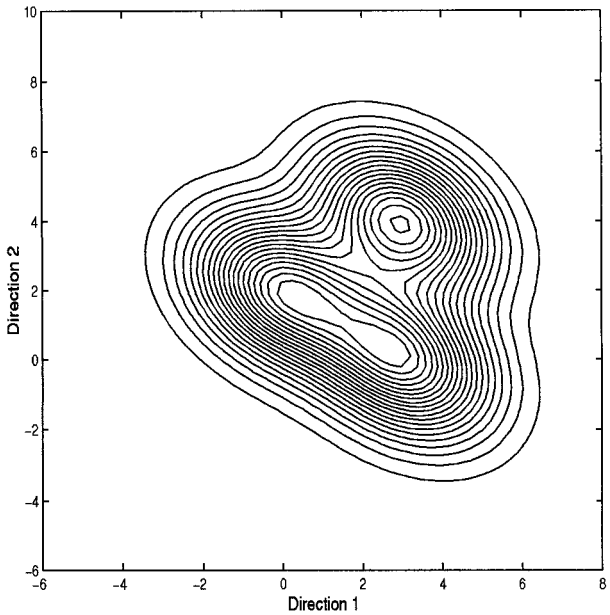

FIG. 4. Contour plot of the PDF corresponding to the mixture model displayed in Fig. 2.

plication of likelihood ratios for choosing k is not recommended.

More recently Haines and Hannachi (1995) applied a simple mixture model to identify the weather regimes in a general circulation model (GCM) simulation. Hannachi (1997) extended the technique to examine the three-dimensional structures of the GCM's flow regimes.

b. Estimating mixture model parameters from data

Let $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a dataset 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 $\boldsymbol{\phi}$ that characterize the PDF model assumed to have generated the data; hats (^) 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 that 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

$$L^{(k)}(\hat{\boldsymbol{\phi}} \mid D) = \sum_{n=1}^{N} \log f^{(k)}(\mathbf{x}_n \mid \hat{\boldsymbol{\phi}})$$

$$= \sum_{n=1}^{N} \log\left(\sum_{j=1}^{k} \hat{\alpha}_j g_j(\mathbf{x}_n \mid \hat{\boldsymbol{\theta}}_j)\right). \qquad (3)$$

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.

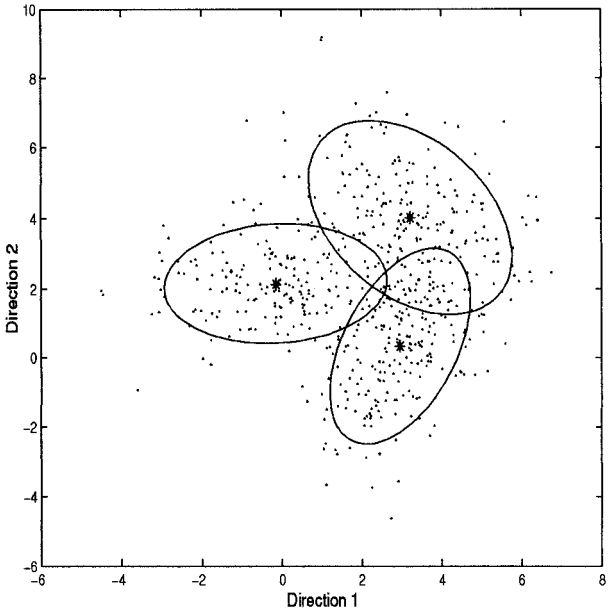

FIG. 5. The *estimated* means and covariance ellipses of component Gaussians superimposed on the scatterplot in Fig. 2. Parameters were estimated using the expectation maximization (EM) procedure (see text for details).

In the mixture model context, one can think of the ''missing data'' in the following manner. Imagine that, when the data D are originally generated by k Gaussians, each data point $\mathbf{x}$ has a ''label'' attached that indicates which component generated which data point. If these labels were known, maximum-likelihood estimation would be quite easy since one could first separate the data for each component and then estimate its parameters separately. However, in practice, these labels are not available and can be considered as missing (Titterington et al. 1985). Thus, the framework of EM, which was conceptually developed for the general problem of maximum-likelihood estimation in the presence of missing data, can be generalized in this fashion to handle parameter estimation for mixture models.

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. 2 results in the parameter estimates shown in Fig. 5. The differences between the estimated parameters (Fig. 5) and the true parameters (Fig. 3) are quite small and only discernible by actual superposition of the two figures. The estimated PDF (not shown) is also quite similar to the true one (Fig. 4). Thus, the EM procedure is quite efficient at recovering the true lo-

cations and shapes of the component densities that generated the data in Fig. 1, even when N is not very large compared to p, 600 versus 17 in this instance.

The test of our clustering method on these synthetic data is quite important and indeed essential to a full understanding of the techniques used in the present paper. Such simulation tests are useful as ''control experiments'' to demonstrate the validity of the method on data for which the ''truth'' is known. In the statistical literature, the use of such simulated data is crucial in the development and validation of new clustering methods (e.g., Banfield and Raftery 1993). Applying the present method to very low-order dynamical systems though—with or without stochastic perturbations—leads to ambiguous results and complexities that go beyond the scope of this paper.

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

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

that is, 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 section 3d, 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 that minimize a least squares objective function (Duda and Hart 1973).

The mixture model is a generalization of such algorithms 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 dataset, 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 dataset D, $1 \le k \le k_{\max}$. 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 dataset 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 randomly chosen partitions (times), where each time the data are 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, $L_m^{(k)} = L(\hat{\boldsymbol{\phi}}_m^{(k)}|D_m^{(1-\beta)})$, $1 \le m \le M$; here $\hat{\boldsymbol{\phi}}_m^{(k)} = \{\hat{\alpha}_{m,1}, \ldots,$

TABLE 1. Cross-validated log-likelihood $L_{cv}^{(k)}$ and posterior probabilities $\hat{P}(k|D)$, 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. 2.

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
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

$\hat{\alpha}_m$, k; $\hat{\boldsymbol{\theta}}_{m,1}, \ldots, \hat{\boldsymbol{\theta}}_{m,k}\}$, that is, 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 dataset $D_m^{(\beta)}$.

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

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

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

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

Eq. (6) follows from Bayes's 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 dataset 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. 2. 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 dataset 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.

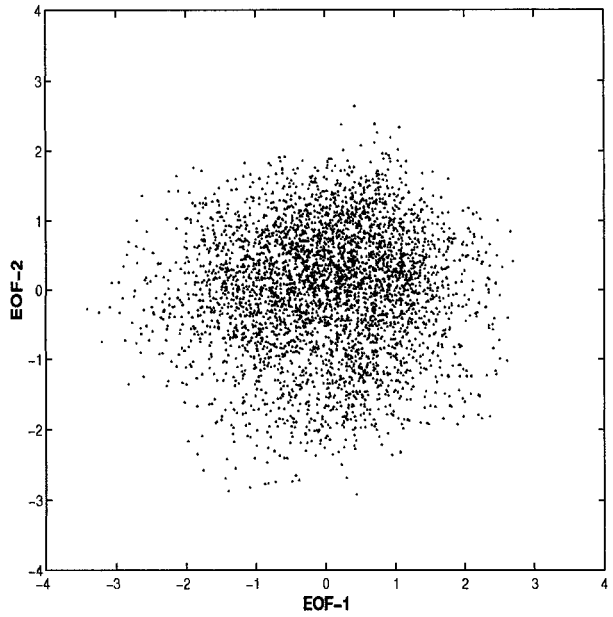

FIG. 6. Scatterplot of NH height anomalies for 44 winters (Dec 1949–Mar 1993), projected onto the two leading EOFs; the data have been normalized by dividing by the standard deviation of EOF 1.

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 dataset as in KGI (see section 2). Projecting filtered anomalies introduces substantial point-to-point correlations in phase space, with visible "trails" of data that are evidently non-Gaussian (even locally) and that would render a Gaussian mixture model quite inappropriate. A scatterplot of the resulting projection is shown in Fig. 6.

We ran the mixture model cross-validation method on this 2D dataset. The algorithm described in section 3d was modified so that random partitions were chosen based on winters rather than days, that is, 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.

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, projected onto the first two leading EOFs (unscaled).

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Cross-validated log-likelihood	−29 164	−29 153	−29 137	−29 148	−29 156	−29 165
Estimated posterior probability	0.0	0.0	1.0	0.0	0.0	0.0

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 three clusters, that is, the cross-validation estimate of the posterior probability for three clusters is effectively 1 and all others are effectively 0.

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

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

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, that is, 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 four 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, that 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.

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.

Expt	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$P(k = 3)$
1	−27 831	−27 815	**−27 808**	−27 817	−27 827	−27 833	1.000
2	−27 858	−27 835	**−27 827**	−27 835	−27 837	−27 844	1.000
3	−27 819	−27 799	**−27 789**	−27 799	−27 802	−27 814	1.000
4	−27 843	−27 825	**−27 812**	−27 821	−27 829	−27 837	1.000
5	−27 825	−27 808	**−27 801**	−27 810	−27 817	−27 824	0.999
6	−27 864	−27 846	**−27 839**	−27 846	−27 854	−27 860	0.999
7	−27 811	−27 792	**−27 783**	−27 792	−27 797	−27 807	1.000
8	−27 818	−27 805	**−27 787**	−27 801	−27 807	−27 811	1.000
9	−27 844	−27 824	**−27 803**	−27 817	−27 817	−27 822	1.000
10	−27 856	−27 837	**−27 829**	−27 841	−27 845	−27 853	1.000

We also investigated the robustness of the method with respect to the number of leading EOFs retained, that is, 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 $k = 3$, that is, $\hat{P}(k = 3) \approx 1$, up to and including $d = 6$; at $d = 7$ the mass "switched" to become concentrated at $k = 1$, that is, $\hat{P}(k = 1) \approx 1$.

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 three-component model but only 56 parameters for a single-component model; by contrast, in five dimensions, the three-component model needs only 48 parameters.

Overall, for a fixed number N of data points, one often has $k \to 1$ as d increases. More precisely $\hat{P}(k = 1) \to 1$ for increasing d, in most cases; unless, that is, the dataset does indeed consist of two or more well-separated components that lie essentially (i.e., to within small-amplitude "noise") in a low-dimensional subspace. The exact value of d at which a single-component model becomes most likely depends, of course, on the particular dataset. Similarly, given a fixed dimension d, the mixture model will theoretically find $\hat{P}(k = 1) \to 1$ for decreasing N.

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 therein) and with the theoretical arguments given in appendix B that cross-validation will pick the best mixture model that can be fitted to a finite set of data. If the data are sufficient in number N, this best model will correspond to the true model; if, however, there are too few data relative to the complexity of the models being fitted, the method will be more conservative and it will choose a simpler model that can be supported more reliably by the data. Another interpretation of this result is that empirical support of the three-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 (cf. also Lorenz 1969).

For the three-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 fitted to the data for each case. For each value of d, three maps of 700-mb height anomalies were obtained from the centroids of the three 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 three maps obtained when using $d = 2$ (see section 4a). The results, shown in Table 5, indicate that the correlations between the 2D 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 man-

TABLE 5. Pattern correlation coefficients between maps fitted using the data projected onto d EOFs, $3 \le d \le 12$, and maps fitted using $d = 2$ EOFs. The maps correspond to centers of a mixture model based on three 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

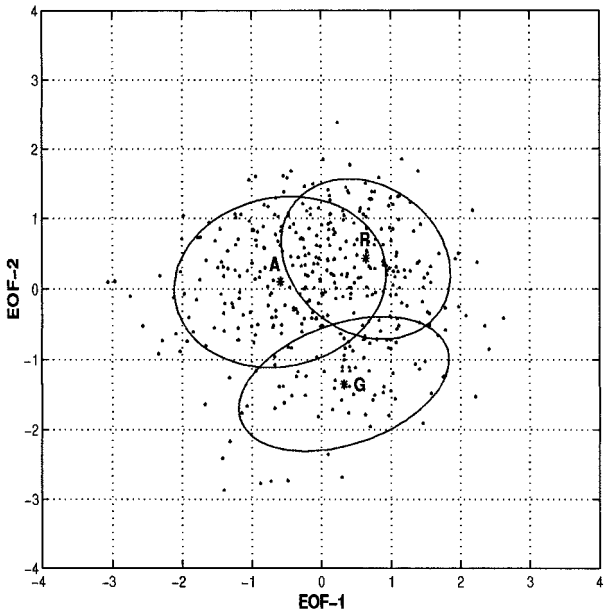

FIG. 7. The estimated centroids, indicated by asterisks, and covariance ellipses superimposed on the scatterplot of Fig. 5, where only every 10th data point has been plotted for clarity. The identities of the clusters—A, G, and R—are indicated beside the respective ellipses. The parameters were estimated by the same EM procedure as for the synthetic data, using a mixture model based on three Gaussian components. The estimated parameters for the three 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 (counterclockwise) 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.

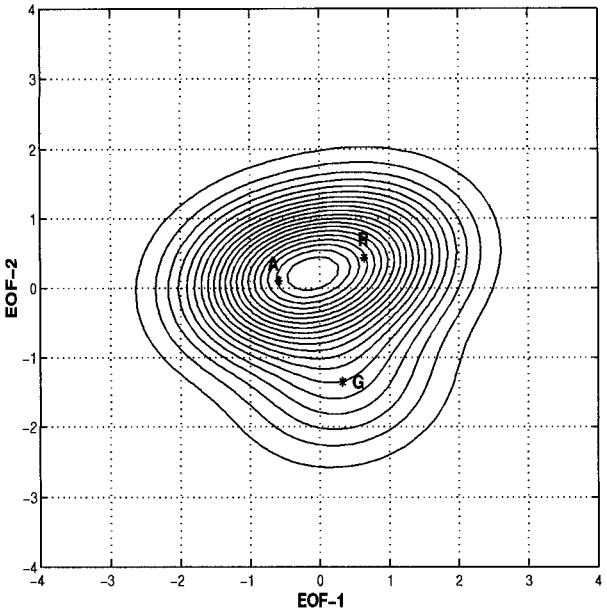

FIG. 8. Contour plot of the PDF estimate provided by the mixture model of Fig. 6; the asterisks and associated labels for the three clusters (A, G, and R) indicate the corresponding centroids.

ner, 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.

We conducted, furthermore, a full cross-validation—as in section 4a and using $\beta = 0.5$ and $M = 20$—but applied to the restricted dataset of 32 NH winters after 1 December 1961. This was done so as to examine the possible effect of lack of sufficient observations entering the National Meteorological Center's (NMC, now NCEP) 700-mb analysis data prior to the early 1960s over the North Pacific (A. R. Hansen 1998, personal communication). The mixture model found the posterior probability of $\hat{P}(k = 3) \approx 1$ and the three new cluster centroids and covariance structures are very similar to those found using the full 44-winter dataset.

c. Comparison with CW's results and interpretation

Given that there is strong evidence for three Gaussian clusters, we fit a three-component Gaussian model to the entire set of 44 winters in the 2D EOF space (rather than partitioning into halves as before) and examine the results. Figure 7 shows the location of the means of the Gaussians and the three-standard-deviation ellipses associated with their covariance matrices, superposed on a scatterplot of the data projected onto the first two EOFs. The resulting contour map of the bivariate-mixture PDF is shown in Fig. 8. Note that the multicomponent cluster structure does not lead to multimodality in the PDF; this is well known to be true in general for k-component mixture models; that is, the resulting probability density can have anywhere from 1 to k modes.

The means of the three Gaussians fitted in Fig. 7 have a natural interpretation as the centroids of three Gaussian data clusters. Figure 9 presents the three maps corresponding to our three cluster means on the left and the three maps corresponding to CW's most reproducible hierarchical clusters on the right [from Fig. 11 of Wallace (1996)]. Cheng and Wallace (1993) labeled both the maps in question and the clusters they represent as ''A'' (for Alaska), ''G'' (for Greenland), and ''R'' (for the Rockies). Their maps and ours have a high degree of qualitative pairwise 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 Alaska. The middle (G) maps exhibit a strong high over southern Greenland. The bottom (R) maps are characterized by an intensification of the Pacific jet stream and an enhancement of the climatological mean ridge over the Rockies.

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 datasets (700-mb vs 500-mb data

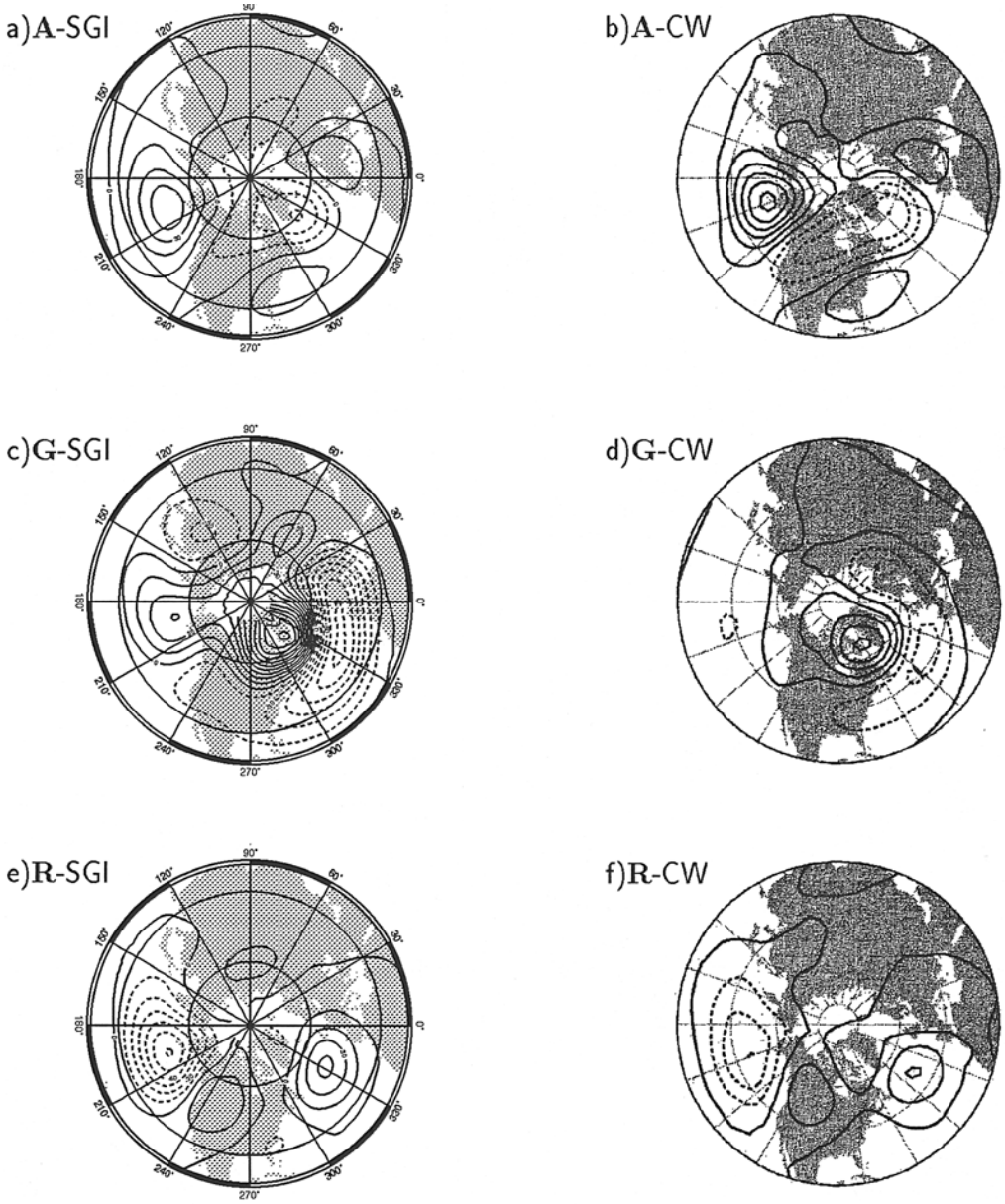

FIG. 9. Height anomaly maps for the three cluster centroids of the present mixture model (a), (c), (e) (labeled SGI) and of CW's hierarchical cluster model, as applied by Wallace (1996) to a slightly longer dataset (b), (d), (f) (labeled CW). Pairs of maps (a), (b) correspond to CW's cluster A; (c), (d) to G; and (e), (f) to R [see text for details; (b), (d), (f) reproduced by permission]. Contour interval is (left) 15 m (SGI) and (right) 50 m (CW); negative contours are dashed and the zero contour is solid.

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). Cheng and Wallace's (1993) methodology for arriving at three 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 ''three-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 three 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 (1988) 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, that is, 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 $\mathbf{C}_j$ are constrained by the overall covariance structure of the entire dataset. Most importantly, the means μ_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.

It is the angles of the centroid vectors 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. The angles are also closely associated to the spatial patterns on the grid. 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. 9); that is, 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 Fig. 15a; the polarity is reversed, since our EOFs and CW's came out to have opposite sign. Cluster A, for example, is farther from the origin in CW than in this paper; it is clear, however, from CW's Fig. 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 that 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 four 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 [see also Fig. 12 in Haines and Hannachi (1995)].

Molteni et al. (1990) performed clustering on the eddy component, that is, the departure of the actual field from zonal symmetry, of 500-mb geopotential height fields analyzed by the European Centre for Medium-Range Weather Forecasts for the 32 winters that extend from December 1952 through February 1984. Although their analysis is not strictly comparable to ours, due to the difference in data type and time span, our clusters A and R resemble closely clusters 5b and 1 of Molteni et al. (1990), respectively. We refer to CW for further discussion on the similarities and differences between their clusters, and hence ours, and those of Molteni and colleagues.

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) PAC sector (120°E–60°W) and (b) 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 four 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 $\hat{P}(k = 3) = 0.020$ using the full 44-winter dataset, and $\hat{P}(k = 2) = 0.993$ and $\hat{P}(k = 3) = 0.007$ using the restricted 32-winter dataset starting in December 1961. This robustness of the PAC results confirms further that the insufficient data coverage over the PAC region prior to the early 1960s did not affect our results. The probabilities for all 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 10 shows the location of the means and three-standard-deviation covariance ellipses of the estimated Gaussian components for $k = 2$, superposed on a scatterplot of every 10th day projected onto the first two EOFs. Figures 11a and 11b show the maps corresponding to the centroids of the two clusters. The spatial pattern in Fig. 11a clearly resembles the P1 regime and that in Fig. 11b the P2 regime of KGII's sectorial analysis; our two PAC clusters also resemble CW's sectorial clusters ''*R*'' and ''*A*,'' respectively. These two PAC regimes are the sectorial counterpart of the hemispheric PNA and RNA clusters, here as well as in CW and KGI. We use an asterisk to distinguish between the sectorially defined PNA* (Fig. 11a) and RNA* (Fig. 11b) and the hemispheric regimes.

Projecting the data onto the first two EOFs, rather than the first four EOFs as above, produces estimated posterior probabilities of $\hat{P}(k = 2) = 0.824$ and $\hat{P}(k = 3) = 0.176$, while projection onto three EOFs produces $\hat{P}(k = 2) = 0.970$ and $\hat{P}(k = 3) = 0.030$. The nonzero probabilities for the 2D case here are quite similar to those obtained for the synthetic three-cluster case in Table 1, except that the distribution here peaks at $k = 2$. For both PAC cases, of two and three EOFs, the two dominant cluster centroids correspond to the same PNA* and RNA* patterns. The somewhat larger probability for $k = 3$ in the 2D subspace suggests that more complex structure (i.e., $k > 2$) may be present, as apparent in KGII, MVL, and Robertson and Ghil (1999), but this structure is not fully supported by the current data within the context of a Gaussian mixture model.

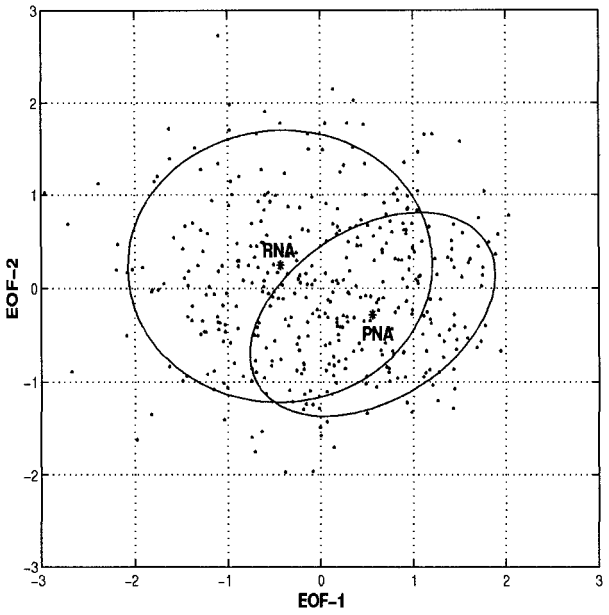

FIG. 10. Same as Fig. 6 for the Pacific sector height anomalies projected onto the two leading sectorial EOFs. The identities of the two clusters, RNA* and PNA*, are indicated beside the respective centroids, plotted as asterisks. The estimated parameters for the two clusters are (*RNA*) $\hat{\alpha}_{\text{RNA}}^* = 0.55$, $\hat{\mu}_{\text{RNA}}^* = (-0.43, 0.24)$, $\tan(\hat{\psi}_{\text{RNA}}^*) = 0.07$, $\hat{\lambda}_1 = 0.90$, $\hat{\lambda}_2 = 0.71$; and PNA* $\hat{\alpha}_{\text{PNA}}^* = 0.45$, $\hat{\mu}_{\text{PNA}}^* = (0.56, -0.28)$, $\tan(\hat{\psi}_{\text{PNA}}^*) = 0.61$, $\hat{\lambda}_1 = 0.70$, $\hat{\lambda}_2 = 0.28$. Here α, μ, and ψ are defined as in Fig. 6.

b. ATL sector

For the ATL sector, the posterior probabilities on k, when projecting the data onto the four 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 12 shows the location of the means and three-standard-deviation covariance ellipses of the estimated Gaussian components for $k = 2$, superposed on a scatterplot of every 10th day projected onto the first two EOFs. Figures 11c and 11d show the height anomaly maps that correspond to the centroids of the two 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 BNAO* and ZNAO*, respectively.

Projecting the data onto the leading two EOFs, rather than four 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 three 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 2D or 3D subspace supports a model with $k = 3$. From a statis-

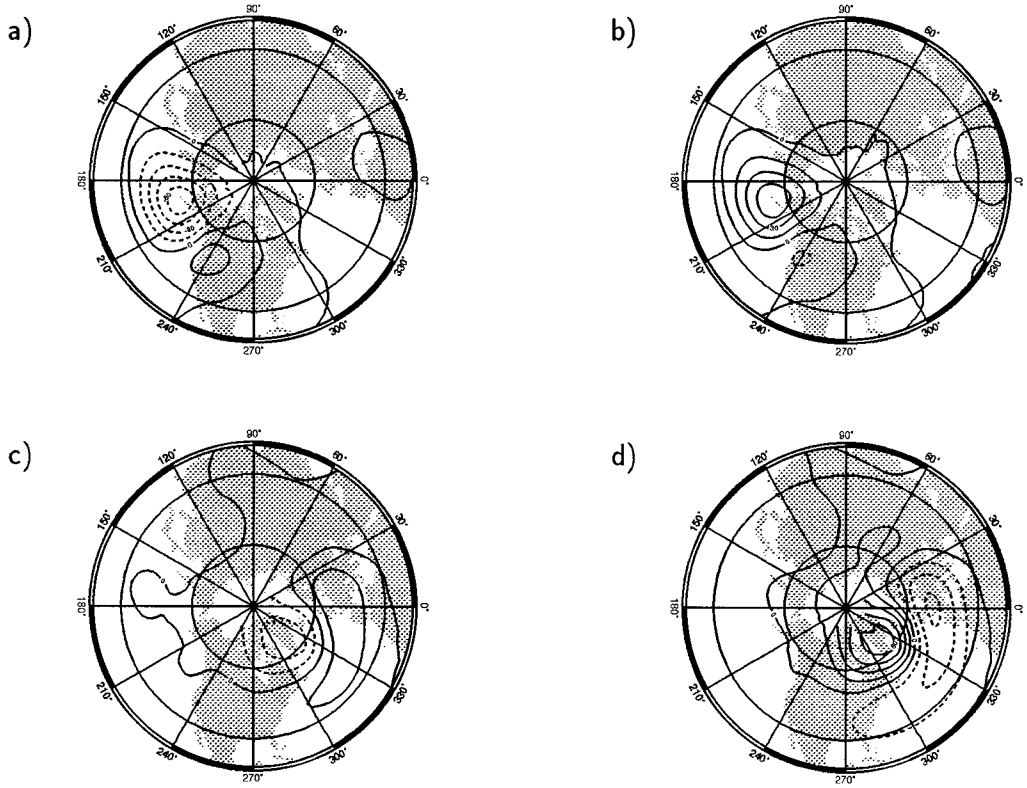

FIG. 11. Height anomaly maps for the clusters found by the mixture model from the PAC and ATL sectorial analyses. Contour interval is 15 m. (a), (b) The maps resemble the PNA and RNA patterns respectively, while (c) and (d) resemble the ZNAO and BNAO patterns respectively (see text for details).

tical-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 4D case (see discussion in section 4b).

When using the restricted 32-winter dataset projected onto the four leading EOFs, our mixture model yields $\hat{P}(k = 1) \approx 0.667$ and $\hat{P}(k = 2) \approx 0.333$. This change has little to do with the accuracy increase of the NMC analysis over the PAC region in the 1960s; it is due, in all likelihood, to the size N of the reduced dataset becoming insufficient to support a number k of Gaussian clusters larger than 1. A complementary, meteorological explanation for the cause of this change will be discussed in section 6b. It is, therefore, even more remarkable that the hemispheric $k = 3$ result is quite stable for dimensionality d ranging from 2 to 6, 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 using the full 44-winter dataset that correspond to the $k = 3$ model—in both 2D and 3D subspaces—display continuity with respect to the $k = 2$ model in the 4D EOF subspace; namely, both the BNAO* and ZNAO* clusters are retained. The third cluster (not shown) has essentially the same spatial pattern in both subspaces: it is quite similar to the A2 pattern in KGII and to MVL's second ATL cluster, with a blocking ridge over western Europe and split flow north and south of it. This classic blocking pattern (see also Rex 1950a,b; Vautard 1990) is not fully supported as an additional, third cluster by the existing data when using $d = 4$, as in KGII; it should emerge as entirely significant, even in the rather conservative setting of Gaussian mixture modeling, as the size of the dataset increases in the future.

c. Comparison with previous results

Using hierarchical clustering on the sectorial data, CW obtained the two 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

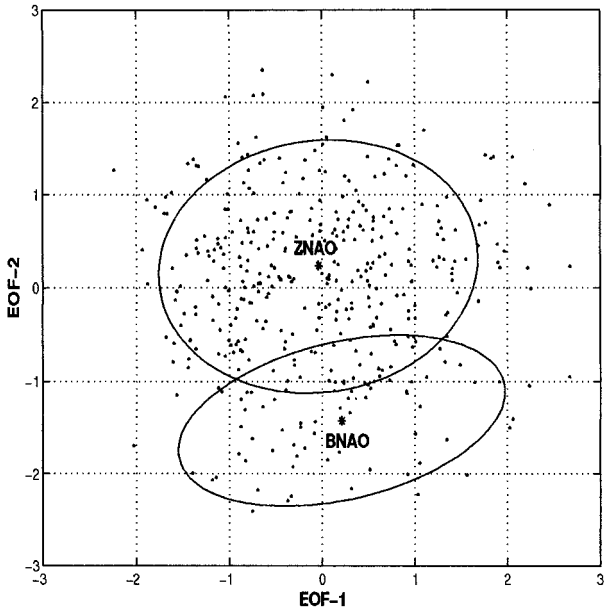

FIG. 12. Same as Fig. 6 for the Atlantic sector height anomalies projected onto the two leading sectorial EOFs. The identities of the two clusters, ZNAO* and BNAO*, are indicated beside the respective centroids, plotted as asterisks. The estimated parameters for the two clusters are (ZNAO*) $\hat{\alpha}_{\text{ZNAO}}^* = 0.86$, $\hat{\mu}_{\text{ZNAO}}^* = (-0.04, 0.24)$, $\tan(\hat{\psi}_{\text{ZNAO}}^*) = 0.15$, $\hat{\lambda}_1 = 0.99$, $\hat{\lambda}_2 = 0.61$; and (BNAO*) $\hat{\alpha}_{\text{BNAO}}^* = 0.14$, $\hat{\mu}_{\text{BNAO}}^* = (0.21, -1.43)$, $\tan(\hat{\psi}_{\text{BNAO}}^*) = 0.24$, $\hat{\lambda}_1 = 0.99$, $\hat{\lambda}_2 = 0.61$.

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 eight EOFs of the dataset and with the same covariance matrix as the data at lag 0 and 2 days. The classifiability index of their data was significantly distinct from the distribution that such a Markov process, as given by 100 Monte Carlo simulations of the process with the same length as the dataset, would provide for $k = 3$ in the PAC sector and $k = 4$ in the ATL sector. Still the spatial patterns of the centroids obtained in the two 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 three PAC and four 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. 10 and 12. Indeed, as discussed already in section 4c, the mixture model imposes certain constraints on the possible cluster centroids that 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 $r = -\alpha_2/\alpha_1$; that is, 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 $2 \leq d \leq 6$ (shown for $d = 2$ in Figs. 6–8), as well as for the ATL results in 2D (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 that 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

We introduced and described probabilistic clustering using finite mixture models, for the purposes of automatically clustering 44 winters of NH 700-mb height anomaly data. After projecting the anomalies onto the dataset's leading EOFs, a mixture model clustering algorithm was used 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 three distinct Gaussian 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 three clusters found by the mixture approach have a very close correspondence to the three cluster patterns found by Cheng and Wallace (1993; CW throughout) using hierarchical, nonprobabilistic clustering on a comparable dataset of NH 500-mb height anomalies. The three common clusters are A for a pronounced ridge over the Gulf of Alaska, G for an anticyclonic feature over southern Greenland, and R for an enhancement of the climatological ridges over the Rocky Mountains. The three clusters also agree nicely 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), KGI, and KGII that the hemispheric clusters are manifestations of the sectorial ones. The tentative explanation for the ZNAO being ''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 (Fig. 12). This is consistent with the predominantly positive anomaly (zonal phase) of the NAO index over the last quarter-century (Hurrell 1995) and with our result of $k = 1$ for the 32-winter dataset in this sector. The NAO index is defined as the sea level pressure difference between the Azores and Iceland and measures the intensity of the westerly jet across the North Atlantic basin: positive anomalies thus correspond to prevalence of the ZNAO* regime and negative ones to that of BNAO*. It follows that, since the early 1970s, occurrences of BNAO* were more outliers of a single-component mixture model than sufficient to constitute a second component. This helps explain the absence of a stable ZNAO* cluster from the hemispheric analysis of CW and the present paper.

Interdecadal changes in the large-scale atmosphere's intraseasonal, 10–100-day low-frequency variability have also been documented in the PAC sector, where the Aleutian low deepened significantly during the 1977–88 winters (Nitta and Yamada 1989; Trenberth 1990). Such changes 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 (Molteni et al. 1993; Palmer 1999; Robertson and Ghil 1999; Robertson et al. 1999). This would lead us to suspect that an analysis of hemispheric upper-air data, if they were available, over the middle third of the twentieth century (see Fig. 1A in Hurrell 1995) might yield four hemispheric regimes, rather than three. Indeed, ZNAO's centroid is likely to have moved farther away from the 2D phase space's origin, due to the distribution of days between its sectorial ZNAO* manifestation and BNAO* being better balanced during the century's middle decades.

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 the interaction of the jet and its instabilities 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 sectorial bimodality is distinct from, but still related to, the hemispheric one originally claimed by Benzi et al. (1986) and Hansen and Sutera (1986). The latter relied on an interesting modification of Charney and DeVore's (1979) model for multiple atmospheric equilibria in a β-channel and thus required preferred simultaneity of the blocked versus 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), while the occasional ''phase locking'' between the two could arise from the different initial data available early in each winter, when the oscillations become active (Strong et al. 1995).

In more recent work, Hansen and Sutera (1993, 1995) refined the analysis of the wave-amplitude index originally proposed by them and their collaborators in the mid-1980s, using more extensive datasets and more stringent statistical-significance criteria (cf. also Nitsche et al. 1994). Hansen and Sutera identified, in fact, in these two papers their large-amplitude regime with the A regime of CW's analysis and the present one.

We are thus in a position to answer the first two questions posed in section 1. (i) The total number of clusters in NH height data for the last half-century's winters is three or higher, with the number of sectorial clusters being two or more in both the Pacific–North American (PAC) and Atlantic–Eurasian (ATL) sectors. (ii) The minimal set of clusters, as confirmed by the present, rather conservative Gaussian mixture model, is that associated with CW's A, G, and R hemispheric clusters, and with the sectorial expression of KGI's PNA–RNA, and ZNAO–BNAO, pairs of clusters, over the PAC and ATL sectors, respectively.

The minimal set of three hemispheric and four sectorial clusters—two for each of the two sectors—can be identified as a subset of the clusters found by other analyses that use less conservative methods. Robertson and Ghil (1999) have argued, for instance, that the distinction between a PNA and a Tropical–Northern Hemisphere (TNH) cluster is useful when studying the response of the extratropical PAC sector to intraseasonal and interdecadal changes in sea surface temperatures (SSTs): frequency-of-occurrence changes in TNH are mostly responsible for the effect of interannual SST changes in the tropical Pacific, while frequency changes

in PNA appear to be responsible for the effect of interdecadal SST changes in the North Pacific.

The set of upper-air data at our disposal is limited to the last half-century. This set has allowed meteorologists, by applying a battery of independent statistical methods, to obtain results on regime classification and description that are finally converging, as discussed in the previous two paragraphs. This set by itself, however, will not suffice, even when using the most sophisticated statistical methods, to answer the third question posed at the beginning of this paper, that of the dynamical mechanisms that cause the clustering. To answer the latter question, distinct theoretical conjectures need to be tested not only on the observational data, but also on extended simulations of the atmospheric circulation that are produced by fairly realistic and detailed models, such as GCMs.

Acknowledgments. The authors would like to acknowledge M. Kimoto for providing the data described in this paper, J. Roden and Y. Tian for assistance in preprocessing the data and graphics, and J. M. Wallace and Springer-Verlag for permission to use a copy of the three height anomaly panels in Fig. 11 of Wallace (1996) as Figs. 8b, d, f here. We benefited from the intellectual input of A. Fraser, I. Jolliffe, M. Kimoto, T. N. Palmer, A. W. Robertson, M. Turmon, R. Vautard, and J. M. Wallace. We are also grateful to A. R. Hansen for his constructive review and to two anonymous reviewers for useful comments. 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 PS was supported in part by NSF Grant IRI-9703120. The work of MG and KI was supported by NSF Grant ATM95-23787 and NASA Grant NAG 5-713.

APPENDIX A

The Expectation Maximization 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 $\boldsymbol{\phi}$ consists of weights α_j, the d-dimensional means $\boldsymbol{\mu}_j$, and the $d \times d$ covariance matrices $\mathbf{C}_j$, 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

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

$$1 \leq n \leq N \tag{A1}$$

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

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

$$\hat{\alpha}_j^{r+1} = \frac{1}{N}\sum_{n=1}^{N} \hat{P}^r(\omega_j|\mathbf{x}_n), \tag{A2}$$

$$\hat{\boldsymbol{\mu}}_j^{r+1} = \frac{\sum_{n=1}^{N} \hat{P}^r(\omega_j|\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} \hat{P}^r(\omega_j|\mathbf{x}_n)}, \tag{A3}$$

$$\hat{\mathbf{C}}_j^{r+1} = \frac{\sum_{n=1}^{N} \hat{P}^r(\omega_j|\mathbf{x}_n)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_j^r)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_j^r)^{\mathrm{T}}}{\sum_{n=1}^{N} \hat{P}^r(\omega_j|\mathbf{x}_n)}. \tag{A4}$$

These 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, that is, the procedure is guaranteed to converge to a fixed point, provided the sequence of estimated parameters $\hat{\boldsymbol{\phi}}^r$ 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 $\hat{\boldsymbol{\phi}}^0$. 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 that 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 datasets 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, the prior distributions of the mixture parameters, and of 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 1999). Cross-validated likelihood is asymptotically consistent in the sense that it will always choose the correct model in the limit of increasing dataset sizes. In practice, it has been shown to work well empirically on a variety of simulated and real datasets, performing as well as various Bayesian approximation methods (Smyth 1999). 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, $f^{(k)}(\mathbf{x}|\hat{\boldsymbol{\phi}}^{(k)})$, $1 \leq k \leq k_{\max}$, where each $f^{(k)}(\mathbf{x}|\hat{\boldsymbol{\phi}}^{(k)})$ has been fitted to dataset D.

Let $\hat{\boldsymbol{\phi}}^{(k)}$ 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 $L^{(k)}(\hat{\boldsymbol{\phi}}^{(k)}|D)$ [as defined in Eq. (3)] is a non-decreasing function of k, since the increased flexibility of more mixture components allows a better fit to the data (increased likelihood). Thus, $L^{(k)}(\hat{\boldsymbol{\phi}}^{(k)}|D)$ 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 dataset D^{test} that is not used in fitting any of the models. Let $L^{(k)}(\hat{\boldsymbol{\phi}}^{(k)}|D^{\text{test}})$ be the log-likelihood as defined in Eq. (3), where the parameters $\hat{\boldsymbol{\phi}}^{(k)}$ 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

$$E\left[\frac{-1}{N}\hat{L}_k(\hat{\boldsymbol{\phi}}^{(k)}|D^{\text{test}})\right] = \int f(\mathbf{x}) \log\frac{f(\mathbf{x})}{f^{(k)}(\mathbf{x}|\hat{\boldsymbol{\phi}}^{(k)})}\, d\mathbf{x} + \text{const}, \quad \text{(B1)}$$

that is, the expected value of $L^{(k)}(D^{\text{test}})$ (scaled appropriately), with respect to different test datasets drawn randomly from the true distribution f, is the cross-entropy between $f(\mathbf{x})$ and $f^{(k)}(\mathbf{x}|\hat{\boldsymbol{\phi}}^{(k)})$, 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 $f^{(k)}(\mathbf{x}|\hat{\boldsymbol{\phi}}^{(k)})$ is from the true f: cross-entropy is strictly positive unless $f^{(k)}(\mathbf{x}|\hat{\boldsymbol{\phi}}^{(k)}) = f$ above. Thus, choosing the k that minimizes the out-of-sample likelihood is equivalent (on average) to choosing the model (within the model family under consideration) that 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-validating partition of the dataset is based on a subset that is independent from the data 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 (viz., the cross-validated likelihood estimates) is in turn unbiased. Thus, finding the maximum over k of $L_{\text{cv}}^{(k)}$ will on average select the model that is closest (in cross-entropy distance) to the true density f.

APPENDIX C

Robustness with Respect to Preprocessing

Cheng and Wallace (1993) performed their hierarchical clustering by subsampling the days (choosing ev-

TABLE C1. Pattern correlation coefficients between maps fitted to unfiltered anomalies for all 3960 days and (a) filtered anomalies, (b) unfiltered anomalies using only every fifth day, and (c) filtered anomalies using only every fifth day. All maps were fitted to the data projected into the first two EOFs.

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

ery fifth 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 three-component Gaussian mixture model in the 2D EOF space to three different permutations of the original data described above: (i) unfiltered anomalies for every fifth day, (ii) filtered anomalies for all days, and (iii) filtered anomalies for every fifth day. The EOFs onto which these three additional datasets were projected were not changed from the basic experiments in section 3c; that is, 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. 9) was then calculated. The results are shown in Table C1 and clearly indicate that the maps obtained using any of these methods are all virtually identical.

REFERENCES

Banfield, J. D., and A. E. Raftery, 1993: Model-based Gaussian and non-Gaussian clustering. *Biometrics,* **49,** 803–821.

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. Roy. Meteor. Soc.,* **112,** 661–674.

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

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. *Mach. Learn.,* **29,** 181–244.

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

——, 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 University Press, 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.,* **39B,** 1–38.

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

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

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

Duda, R. O., and P. E. Hart, 1973: *Pattern Classification and Scene Analysis.* John Wiley and Sons, 482 pp.

Edlund, S. B., 1997: Methods for cluster analysis with applications to large NASA data sets. Tech. Rep. TRITA-NA-E9720, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, 45 pp. [Available from KTH, Nada, SE-10044 Stockholm, Sweden.]

Everitt, B. S., and D. J. Hand, 1981: *Finite Mixture Distributions.* Chapman and Hall, 143 pp.

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., D. Reidel, 241–283.

——, and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics,* Vol. 33, Academic Press, 141–266.

Haines, K., and A. Hannachi, 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.

——, 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.

——, and ——, 1993: A comparison between planetary-wave flow regimes and blocking. *Tellus,* **45A,** 281–288.

——, and ——, 1995: The probability density distribution of the planetary-scale atmospheric wave amplitude revisited. *J. Atmos. Sci.,* **52,** 2463–2472.

Hurrell, J. W., 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, 320 pp.

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.

——, and ——, 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, 1985: Persistent anomalies, blocking and variations in atmospheric predictability. *J. Atmos. Sci.,* **42,** 433–471.

——, T. Desponts, and B. Piguet, 1988: Cluster analysis and weather regimes. *Proc. ECMWF Workshop on the Nature and Prediction of Extratropical Weather Systems,* Vol. 2, Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts, 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, 253 pp.

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,** 10 927–10 952.

Molteni, F., A. Sutera, and N. Tronci, 1988: The EOFs of the geopotential eddies at 500 mb in winter and their probability density distributions. *J. Atmos. Sci.,* **45,** 2868–2888.

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

——, L. Ferranti, T. N. Palmer, and P. Viterbo, 1993: A dynamical interpretation of the global response to equatorial Pacific SST anomalies. *J. Climate,* **6,** 777–795.

Namias, J., 1982a: *Short Period Climatic Variations. Collected Works of J. Namias (1934–1974).* Vols. 1 and 2. University of California, San Diego, 905 pp.

——, 1982b: *Short Period Climatic Variations. Collected Works of J. Namias (1975–1982).* Vol. 3. University of California, San Diego, 393 pp.

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

Nitta, T., and S. Yamada, 1989: Recent warming of tropical sea-surface temperature and its relationship to the northern hemisphere circulation. *J. Meteor. Soc. Japan,* **67,** 375–383.

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

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

Preisendorfer, R. W., 1988: *Principal Component Analysis in Meteorology and Oceanography.* Elsevier, 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.

——, 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,* W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds., Chapman and Hall, 441–461.

Robertson, A. W., and M. Ghil, 1999: Large-scale weather regimes and local climate over the western United States. *J. Climate,* **12,** 1796–1813.

——, ——, and M. Latif, 1999: Interdecadal changes in atmospheric low-frequency variability with and without boundary forcing. *J. Atmos. Sci.,* in press.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 175 pp.

Smyth, P., 1999: Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comput.,* in press.

Stone, M., 1974: Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Stat. Soc.,* **36B,** 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., M. S. thesis, Dept. of Atmospheric Sciences, University of California, Los Angeles, 50 pp. [Available from Dept. of Atmospheric Sciences, UCLA, Los Angeles, CA 90095-1565.]

Titterington, D. M., A. F. M. Smith, and U. E. Makov, 1985: *Statistical Analysis of Finite Mixture Distributions.* John Wiley and Sons, 243 pp.

Trenberth, K. E., 1990: Recent observed interdecadal climate changes in the Northern Hemisphere. *Bull. Amer. Meteor. Soc.,* **71,** 988–993.

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

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

Weeks, E. R., Y. Tian, J. S. Urbach, K. Ide, H. Swinney, and M. Ghil, 1997: Transition between blocked and zonal flows in a rotating annulus with topography. *Science,* **278,** 1598–1601.