Detecting Atmospheric Regimes using Cross-Validated Clustering

Padhraic Smyth*
Information and Computer Science
University of California, Irvine
CA 92697-3425

Michael Ghil and Kayo Ide
Department of Atmospheric Sciences
and Inst. of Geophysics and Planetary Physics
University of California, Los Angeles
Los Angeles, CA 90024-1565

Joe Roden
Machine Learning Systems Group
Jet Propulsion Laboratory 525-3660
Pasadena, CA 91109-8099

Andrew Fraser
Systems Science
Portland State University
Portland, OR 97207-0751

Abstract

Low-frequency variability in geopotential height records of the Northern Hemisphere is a topic of significance in atmospheric science, having profound implications for climate modeling and prediction. A debate has existed in the atmospheric science literature as to whether or not "regimes" or clusters exist in geopotential heights, and if so, how many such clusters. This paper tells the "detective story" of how cross-validated mixture model clustering, a methodology originally described at the 1996 KDD conference (Smyth, 1996), has recently provided clear and objective evidence that three clusters exist in the Northern Hemisphere, where each of the detected clusters has a direct physical interpretation. Cross-validated mixture modeling has thus answered an important open scientific question.

Introduction

Detection and identification of "regime-like" behavior in atmospheric circulation patterns is a problem which has attracted a significant amount of interest in atmospheric science. As defined in the atmospheric science literature, *regimes* are recurrent and persistent spatial patterns which can be identified from atmospheric data sets. The most widely-used data set for these studies consists of twice-daily measurements since 1947 of *geopotential height* on a spatial grid of over 500 points in the Northern Hemisphere (NH). Geopotential height is the height in meters at which the atmosphere attains a certain pressure (e.g., one has 500mb height data, 700mb height data, etc.): it can loosely be considered analogous to atmospheric pressure, particularly since one can visualize the data using contour maps with "lows," "highs", "ridges," etc.

Research on low-frequency atmospheric variability using geopotential height data during the past decade has demonstrated that on time scales longer than about a week, large-scale atmospheric flow fields exhibit recurrent and persistent regimes. Direct identification of these regimes in observed flow fields is difficult. This has motivated the use of a variety of cluster analysis algorithms to objectively classify observed geophysical fields into a small set of preferred regimes or categories, e.g., fuzzy clustering (Mo and Ghil, 1988), kernel density estimation and "bump hunting" (Kimoto and Ghil, 1993), hierarchical clustering (Cheng and Wallace, 1993), and least-squares (or k-means) clustering (Michelangeli, Vautard, and Legras (1995)).

While these approaches have produced useful and repeatable results (in terms of significant cluster patterns), there is nonetheless a degree of subjectivity in the application of these clustering techniques which is undesirable. In particular, none of these methods have provided a fully objective answer to the question of how many clusters exist. Thus, among the different studies, it is not clear how many different regimes can be reliably identified. In this paper we describe the application of mixture model clustering to this problem, and in particular the use of cross-validated likelihood (Smyth, 1996) to determine the most likely number of clusters in the data.

The paper begins with a very brief review of the basic concepts of mixture models and the cross-validation methodology. This is followed by brief description of the 700mb geopotential height data set. The application of the mixture modeling methodology to the problem of cluster analysis to NH geopotential height data is then discussed and strong evidence for the existence of 3 regimes is presented. Finally, the scientific implications of this result are discussed.

Clustering using Mixture Models

There is a long tradition in the statistical literature of using mixture models to perform probabilistic clustering (Titterington, Smith, and Makov (1986)). A key feature of the mixture approach to clustering is the ability to handle *uncertainty* about cluster member-

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

ship, cluster locations and shapes, and the number of clusters in a theoretically-sound manner.

Let $\underline{X}$ be a d-dimensional random variable and let $\underline{x}$ represent a particular value of $\underline{X}$, e.g., an observed data vector with d components. A finite mixture probability density function for $\underline{X}$ can be written as

$$f^{(k)}(\underline{x}|\Phi^{(k)}) = \sum_{j=1}^{k} \alpha_j g_j(\underline{x}|\underline{\theta}_j) \qquad (1)$$

where k is the number of components in the model and each of the g_j are the component density functions. The $\underline{\theta}_j$ are the parameters associated with density component g_j and the α_j are the relative "weights" for each component j, where $\sum_j \alpha_j = 1$ and $\alpha_j > 0, 1 \leq j \leq k$. $\Phi^{(k)} = \{\alpha_1, \ldots, \alpha_k, \underline{\theta}_1, \ldots, \underline{\theta}_k\}$ denotes the set of parameters for the overall mixture model. for example Titterington, Smith, and Makov (1986). *Cross-validated log-likelihood* provides a practical and sound way to estimate how many clusters k best fit a given data set (Smyth, 1996).

The Northern Hemisphere 700mb Geopotential Height Data Set

We analyzed the same data as has been used in most of the other clustering studies on this topic (e.g., Kimoto and Ghil (1993)), namely, twice-daily observations of the NH 700-mb geopotential heights on a 10° × 10° diamond grid, compiled at NOAA's Climate Analysis Center. The data are subject to a number of specific preprocessing steps, each of which are considered desirable from an atmospheric science viewpoint. The original 541 grid points are thinned out to yield 358 grid points. This "thinning" removes some points north of 60° so that the resulting map has a more even distribution of grid points. For each resulting grid point, a 5-day running average is applied to remove seasonal effects. The resulting time series (one at each grid point) are called *height anomalies*, in the sense that the remaining signal is the anomalous departure from seasonal trends. A "winter" is defined as the 90 day sequence of anomalies beginning on December 1st of each year. All analysis was performed on the winter data, namely the 44 × 90 = 3960 days defined to be within the winter periods from Dec 1st 1949 extending through March 31st 1993. Non-winter data has much weaker persistence patterns and, thus, is typically excluded from analysis.

As is common practice in atmospheric science, the dimensionality is reduced via principal component analysis of the anomaly covariance matrix (a step referred to as *empirical orthogonal function analysis* or *EOF analysis* in the atmospheric science literature). We will use the atmospheric science notation of "EOFs" to refer to principal component directions in the rest of the paper. Projections used in the results described in this paper range from the first 2 to the first 12 EOFs.

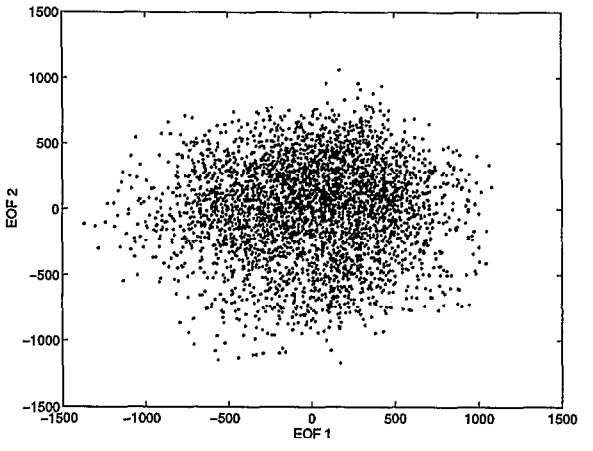

Figure 1: Scatter plot of NH winter anomalies projected into first 2 EOF dimensions

The original 541 grid points for each time index are reduced to a low-dimensional projection by this series of preprocessing steps. Figure 1 shows the 3960 data points projected onto the first two EOFs. It is in this low-dimensional space that cluster analysis is typically performed. From a data analysis viewpoint one may well ask whether or not the results will be sensitive to any of the preprocessing steps performed above. One way of addressing this is to find a way to compare cluster results both with and without preprocessing: we describe such a comparison later in the paper. It is also important to note, however, that while alternative preprocessing steps might suggest themselves purely from a data analysis viewpoint (such as the use of other projection methods), it is important to investigate (as described here) the application of an objective clustering methodology on data which is as similar as possible to that used in previous studies.

Application of Mixture Model Clustering

We applied the mixture model cross-validation methodology (Smyth, 1996) on the two-dimensional data set in Figure 1. In all experiments the number of cross-validation partitions was $M = 20$ and the fraction of data β contained in each test partition was set to 0.5. The number of clusters (mixture components) was varied from $k = 1, \ldots 15$. The log-likelihoods for $k > 6$ were invariably much lower than those for $k \leq 6$ so for clarity only the results for $k = 1, \ldots, 6$ are presented. The estimated posterior probabilities and cross-validated log-likelihoods are tabulated in Table 1. The posterior probabilities indicate clear evidence for 3 clusters, i.e., the cross-validation estimate of the posterior probability for 3 clusters is effectively 1 and all others are effectively zero.

Note that the absolute values of the log-likelihoods

Table 1: Cross-validated log-likelihood and estimated posterior probabilities, as a function of k, from 20 random partitions of 44 winters.

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

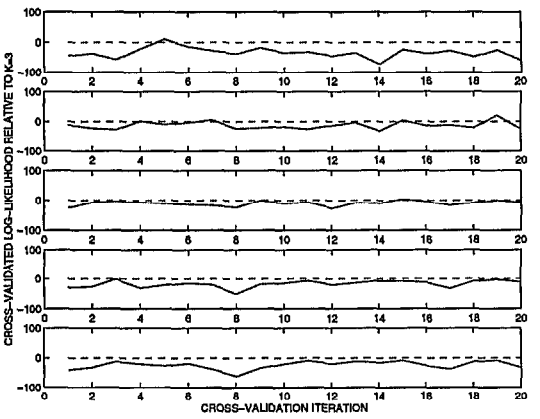

Figure 2: Log-likelihood of the test partition data on each cross-validation iteration relative to the log-likelihood of the $k = 3$ model for (from top) (a) $k = 1$, (b) $k = 2$, (c) $k = 4$, (d), $k = 5$, and (e) $k = 6$.

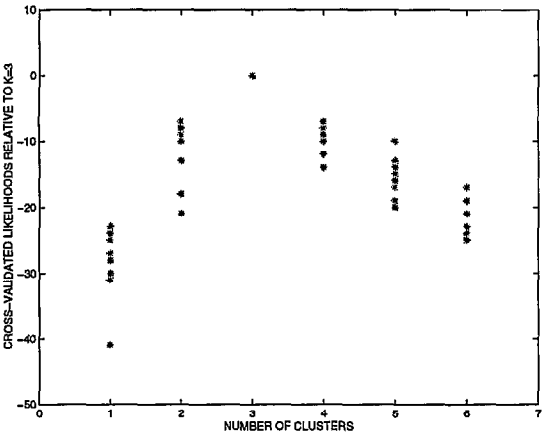

Figure 3: Cross-validated log-likelihoods for $k = 1, \ldots, 6$ relative to the cross-validated log-likelihood of the $k = 3$ model for 10 different randomly chosen cross-validation partitions

are irrelevant—strictly speaking, likelihood is only defined within an arbitrary constant. Figure 2 shows the test log-likelihoods on the 20 different cross-validation partitions, relative to the log-likelihood on each partition of the $k = 3$ model (dotted line equal to zero).

$k = 3$ clearly dominates. Note that for any particular partition $k = 3$ is not necessarily always the highest likelihood model, but *on average across the partitions* it is significantly better than the other possible values for k.

Robustness of the Results

Numerous runs on the same data with the same parameters but with different randomly-chosen winter partitions (with $M = 20$) always provided the same result, namely, an estimated posterior probability of $p(k = 3) \geq 0.999$ in all cases. The relative cross-validated likelihoods over 10 different runs are shown in Figure 3.

We also investigated the robustness of the method to the dimensionality of the EOF-space. The unfiltered anomalies were projected into the first d EOF dimensions, $d = 2, \ldots, 12$. As a function of the dimensionality d, the posterior probability mass was concentrated at $k = 3$ (i.e., $p(k = 3) \approx 1$) until $d = 6$, at which point the mass "switched" to become concentrated at $k = 1$ (i.e., $p(k = 1) \approx 1$)). Thus, 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 in a k-component Gaussian mixture model grows as kd^2. Thus, since the total amount of data to fit the models is fixed, as the dimensionality d increases the estimates of the more complex models are less reliable and cannot be justified by the data. Cross-validation will attempt to pick the best model which can be fit to a finite set of data. If there are enough data, this best model will correspond to the true model, while if there are too few data (relative to the complexity of the models being fit), the method will be more conservative and choose a simpler model which can be supported by the data. Another interpretation of this result is that empirical support of the 3-component model in higher dimensions could require records on the order of a few hundred years long, rather than the 44 years of data currently available.

For the three-component Gaussian model we investigated the variability in the physical grid maps obtained across different numbers of EOF dimensions. Note that cluster centers in the EOF space can be "mapped back" to equivalent grid points in the original spatial grid to

Table 2: Pattern correlation coefficients between maps fitted using d EOF dimensions, $3 \leq d \leq 12$, and maps fitted using 2 EOF dimensions.

EOF Dimensionality 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

create spatial contour maps. The unfiltered anomalies were projected into the first d EOF dimensions, $d = 3, \ldots, 12$ and a Gaussian mixture model with 3 components was fit to the data for each case. For each value of d, 3 physical maps were obtained from the centers of the 3 Gaussians. The pattern correlations (as defined in Wallace and Cheng (1993), page 2676) were then calculated between each of these maps (from d dimensions) and the corresponding maps obtained from 2 EOF dimensions. The results are shown in Table 2. It is clear that here is a very high correlation between the 2d EOF maps and maps obtained in up to 12 EOF dimensions. One can conclude that the dimensionality of the EOF space does not affect the qualitative patterns of the physical maps in any significant manner, using the Gaussian mixture model clustering procedure.

Comparison with Bayesian and Penalized Likelihood Techniques

We ran AUTOCLASS 2.0 (a well-known Bayesian approximation method for mixture modeling from NASA Ames) on exactly the same data as described above. The default version of AUTOCLASS (full covariance matrices) returned $k = 3$ as by far the most likely choice for the number of clusters, i.e., no other k values had any significant posterior probability. For the same data we also calculated the BIC criterion which penalizes the training log-likelihood by an additive factor of $-k/2 \log N$. The BIC criterion was maximized at $k = 1$ (by a substantial margin). This is consistent with previous results which have reported that the BIC criterion can be over-conservative.

Scientific Interpretation of the $k = 3$ Result

Given that there is strong evidence for 3 Gaussian clusters we fit a 3-component Gaussian model to the entire set of 44 winters in the 2d EOF space (rather than partitioning into halves as before) and examine the re-

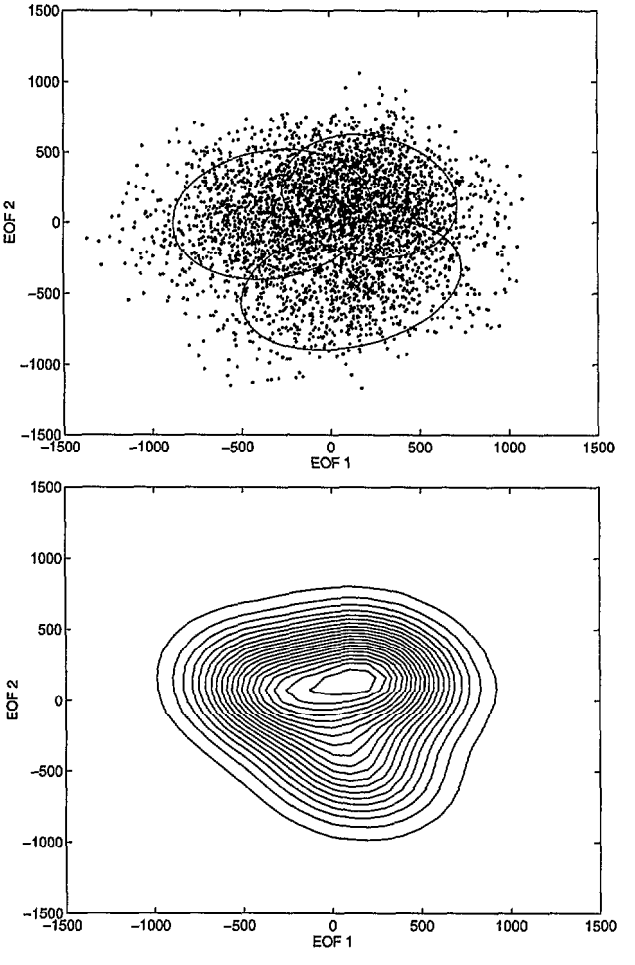

Figure 4: (a) Scatter plot of NH winter anomalies projected into first 2 EOF dimensions with estimated means and covariance matrix shapes (ellipses) superposed as fitted by the EM procedure with a 3-component Gaussian mixture model. (b) Contour plot of the probability density estimate provided by the 3-component Gaussian mixture model fitted to the 2d EOF data.

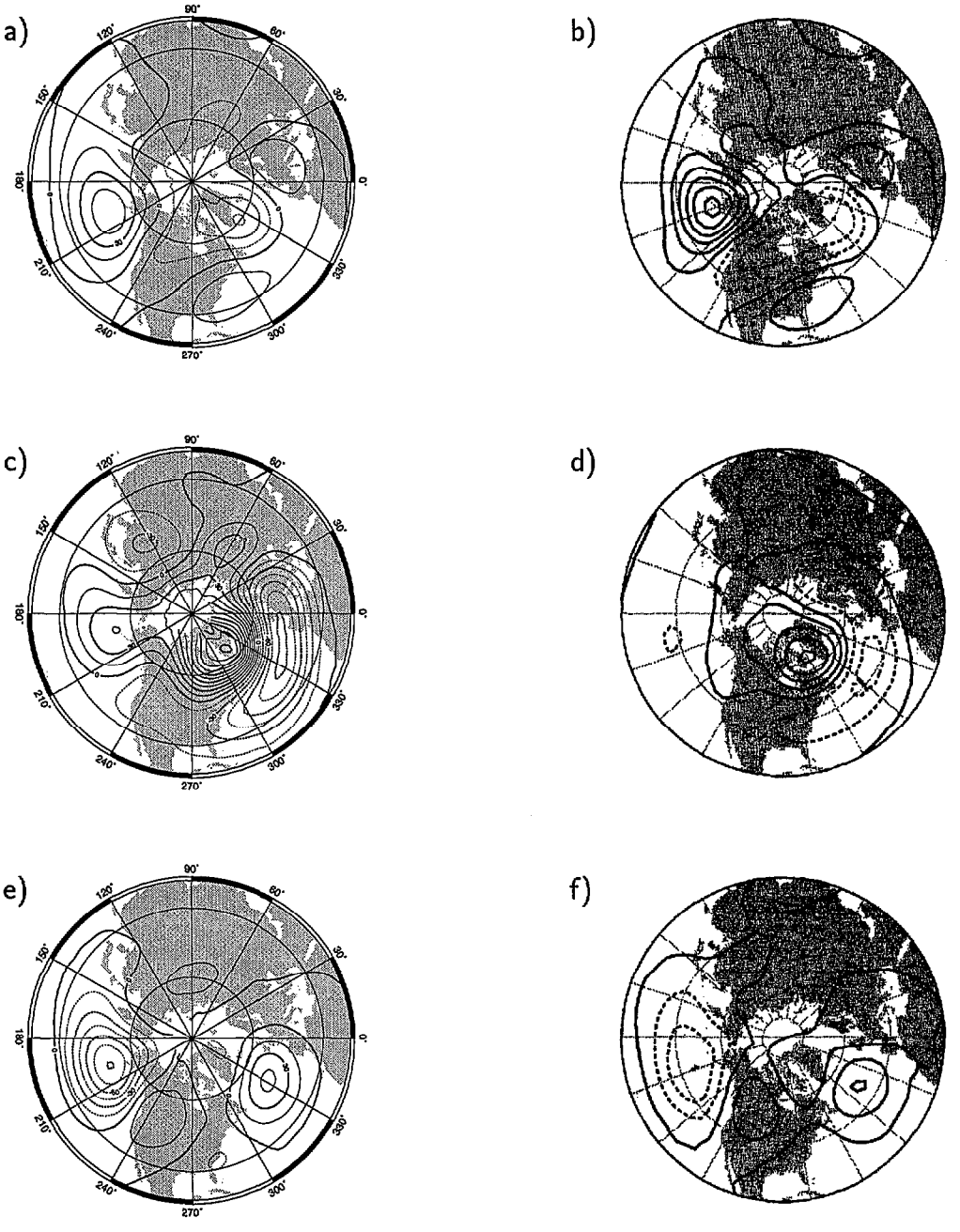

Figure 5: Height anomaly maps for the 3 cluster centers of the mixture model (left: panels a, c and e) and of Cheng and Wallace's (1993) hierarchical cluster model (right: panels b, d and f) which are reproduced in Wallace (1996). See text for details. Panels b,d and f reproduced by permission of J. M. Wallace and Springer-Verlag.

sults. Figure 4(a) shows the location of the means of the Gaussians and the shapes of their associated covariance matrices superposed on a scatter plot of the data in the first 2 EOFs. Figure 4(b) shows the resulting contour map of the bivariate mixture probability density function.

The means of the three Gaussians in Figure 4(a) have a natural interpretation as the centers of three Gaussian clusters. Figure 5 shows the three maps corresponding to the three Gaussian centers on the left and the three maps corresponding to the "most distinct clusters of the wintertime 500mb field" on the right (Cheng and Wallace, 1993; also in Wallace, 1996). The Cheng and Wallace results are considered among the most authoritative on this topic. These two sets of maps have a high degree of qualitative similarity to each other. The upper maps (a) and (b) both clearly possess a distinctive ridge over the Gulf of Alaska. The middle maps (c) and (d) are characterized by a very distinctive blocking pattern over southern Greenland. The bottom maps (e) and (f) have a more complex pattern described as the "Rockies ridge" in Cheng and Wallace (1993, p.2680).

Note that the two sets of maps were produced using two different clustering methodologies (mixture modeling and hierarchical clustering), two somewhat different data sets (500mb and 700mb data over slightly different years) and used different preprocessing of the data (the work in this paper was in EOF-space, Cheng and Wallace clustered the anomaly maps directly). It is quite reassuring from both a data analysis and scientific viewpoint that both methodologies independently arrived at the conclusion of the existence three distinct regimes and that the maps corresponding to these regimes are qualitatively identical. Cheng and Wallace's methodology for arriving at the number "3" for the number of distinct clusters was based on a combination of sophisticated resampling of the data and subjective judgement. In their own words, "the more reproducible clusters are strung out along three well-defined branches of the family tree" (Cheng and Wallace, 1993). The cross-validation results described here are an objective independent validation of Cheng and Wallace's "three-cluster" result. We note in passing that the mixture modeling result supporting $k = 3$ was obtained by a subset of the authors (P.S., J.R., and A.F.) before knowing of the existence of the Cheng and Wallace result, i.e., the two analyses which indicate that $k = 3$ were performed completely independently and without knowledge of the other.

Discussion and Conclusion

An obvious question is whether or not the results are sensitive to the projection methodology being used? The answer appears to be no. The similarity of the maps in Figure 5 (where one set is obtained in EOF-space and the other set by directly clustering the grid patterns) indicates that the EOF projection does not impact the resulting clusters. Cheng and Wallace (1993) reached the same conclusion by finding that hierarchical clustering in EOF space produced essentially the same clusters as obtained with no EOF projection.

More generally, the fact that both the spatial and temporal dimensions of the data are ignored would indicate that there is considerable unmodeled structure in the data. In the temporal context we are investigating the use of dynamic linear models to discover "dynamic clusters." However, the significant first step in this work was to investigate the same data as had been investigated in prior work on the topic. It is also interesting to note that the participating atmospheric scientists (M.G. and K.I.) were much more willing to trust a methodology based on cross-validation than a Bayesian analysis. This is an important point. It is suggestive that while *in theory* a fully Bayesian analysis can be viewed as the optimal approach, *in practice* a cross-validation methodology can be more practical, particularly when data are relatively plentiful.

Acknowledgements

The authors would like to acknowledge Dr. Masahide Kimoto for providing the data described in this paper. 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.

References

Cheng, X., and Wallace, J. M. 'Cluster analysis of the Northern hemisphere wintertime 500-hPa height field: spatial patterns,' *J. Atmos. Sci.*, 50(16), 2674–2696, 1993.

Kimoto, M., and Ghil, M., 1993 'Multiple flow regimes in the Northern hemisphere winter: Part I: methodology and hemispheric regimes,' *J. Atmos. Sci.*, 50(16), pp.2625–2643.

Michelangeli, P-A., Vautard, R., and Legras, B., 1995, 'Weather regimes: recurrence and quasi-stationarity,' *J. Atmos. Sci.*, 52(8), 1237–1256.

Mo, K., and Ghil, M., 'Cluster analysis of multiple planetary flow regimes,' *J. Geophys. Res.*, 93, D9, 10927–10952, 1988.

Smyth, P., 'Clustering using Monte-Carlo cross validation,' in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, pp.126–133, 1996.

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

Wallace. J. M., 'Observed Climatic Variability: Spatial Structure,' in *Decadal Climate Variability*, D. L. T. Anderson and J. Willebrand (eds.), NATO ASI Series, Springer Verlag, 1996.