P 2.2

DATA ASSIMILATION AND WEATHER REGIMES IN A THREE-LEVEL QUASI-GEOSTROPHIC MODEL

D. Kondrashov*, M. Ghil, K. Ide, University of California, Los Angeles,
and R. Todling, NASA, Goddard Space Flight Center.

1. INTRODUCTION

Extended-range weather prediction depends in a crucial way on skill at forecasting the duration of a blocking event or other persistent anomaly that is under way at initial forecast time. The ability to forecast the subsequent onset of another persistent anomaly—after the break of the current one—has proven even more elusive.

To address this crucial problem, we study the application of advanced data assimilation methods on predicting the transitions between atmospheric weather regimes. Marshall and Molteni's (1993) three-level quasi-geostrophic (QG) model in spherical geometry has been shown to have a fairly realistic climatology and exhibit multiple regimes that bear some resemblance to those found in observations. Using this model, we study the transition mechanism between such regimes.

2. CLUSTERING ANALYSIS

The dataset for analysis was obtained from an 18,000-day perpetual-winter simulation of our QG model — whose three levels are at 200 mb, 500 mb and 800 mb — on a T21 (64 x 34) grid. In order to examine the phase-space structure of atmospheric dynamics in such a high-dimensional system, it is necessary to reduce the dataset's dimensionality. For this purpose, we apply empirical orthogonal function (EOF) analysis to the unfiltered 500-mb level streamfunction anomalies in the Northern Hemisphere (NH), where the gridded data points are weighted by the cosine of their latitudes. The leading 10 EOFs are responsible for 47% of the variance of the dataset, the first mode capturing 11%, and the second 6%.

In order to objectively identify weather regimes in the QG model simulation, we apply two independent clustering techniques and compare the results (see Table 1 of Ghil and Robertson 2001). One technique is the *k*-means algorithm used by Michelangeli et al. (1995) and the other is the Gaussian mixture model used by Smyth et al. (1999) for the classification of NH weather regimes in observed geopotential height fields.

For a given number d of leading EOFs, both techniques provide the number of clusters k and the cluster centroids in a d-dimensional subspace of the model's phase space. We want each cluster to correspond to a weather regime of the QG model's physical space. Therefore it is critical for our study to optimize the classification into clusters over various subspaces. The number of EOFs d and clusters k can be used as parameters to measure the robustness of the clusters.

The *k*-means algorithm is based on the dynamic cluster method and formulated as follows. Given a prescribed number k of clusters in a d-dimensional space, it attempts to find an optimal partition of the data into the k clusters that minimizes the sum of the variances within each cluster. A data point belongs to a cluster if its distance to the cluster centroid is less than one standard deviation of all distances within a cluster. In other to determine the optimal k, Michelangeli et al. (1995) proposed the use of a classifiability index. This index measures the stability of the cluster solutions as a function of k, across different initial (random) seeds of the algorithm, based on the correlation between the cluster centroids.

Table 1 gives the classifiability index of the QG model simulation for $2 \le d \le 6$. Clearly, it is very high for both $k = 4$ and $k = 3$. We thus conclude that the *k*-means algorithm alone cannot identify the optimal clusters of the QG model

	d=2	d=3	d=4	d=5	d=6
k=3	0.8687	0.9795	0.9996	0.9810	0.9979
k=4	0.9999	0.9999	0.9995	0.9999	0.9999
k=5	0.8702	0.8214	0.7960	0.7984	0.7395
k=6	0.7012	0.7011	0.6907	0.6933	0.6883

Table 1: Classifiability index of the *k*-means algorithm: k is the prescribed number of clusters, while d is the number of the EOFs retained for the analysis.

**Corresponding author address*: D. Kondrashov, UCLA, Atmospheric Sciences Dept., 7127 Math Sciences Bldg., Los Angeles CA 90095-1565; dkondras@atmos.ucla.edu

simulation.

The Gaussian mixture model uses a linear combination of k Gaussian density functions. Unlike the k-means algorithm, each data point in the d-dimensional space can have a degree of membership in several clusters, depending on its position with respect to the centroid and the weight of a cluster (Smyth et al. 1999). Unlike the k-means algorithm, it has a built-in criterion for determining how many clusters should be fitted to the data. This criterion is based on the cross-validated log-likelihood**,** shown in Table 2: the higher its value for each dimension d, the more likely it is that k is the correct number of clusters for that d.

	d=2	d=3	d=4	d=5
k=1	-20714	-27611	-33064	-38676
k=2	-20312	-27156	-32558	-38116
k=3	-20237	-27082	-32440	-37993
k=4	-20220	-27026	-32418	-37913
k=5	**-20214**	-27000	-32376	-37896
k=6	-20223	**-26996**	**-32358**	**-37872**
k=7	-20240	-27019	-32372	-37881

Table 2: Cross-validated log-likelihood for 18,000 data points; the maximum value for each d is in bold.

The mixture model consistently gives k = 6, which is higher than the values k = 3 or 4 obtained by the k-means algorithm. Hannachi and O'Neill (2001) found that the Gaussian mixture model tends to overfit the clusters when the distribution of the data is not Gaussian. This is the case here, too. In fact, when using either half of the entire dataset, the cross-validated log-likelihood suggests a higher probability for k = 4 and 5 than for k = 6 (see Table 3 for the first half of the data).

	d=2	d=3	d=4	d=5
k=1	-13659	-16360	-19007	-21415
k=2	-13418	-16106	-18767	-21159
k=3	-13389	-16076	-18721	-21119
k=4	-13365	-16077	-18697	**-21131**
k=5	**-13363**	**-16061**	**-18697**	-21132
k=6	-13382	-16086	-18719	-21191

Table 3: Cross-validated log-likelihood for the first 9,000 data points; maxima in bold.

Since the results of the two methods seem to disagree on the optimal cluster number k for the QG model simulation, we compare the anomaly maps of the centroids produced by the two (see, for instance, Table 2 in Robertson and Ghil 1999). To do so, we compute the pattern correlation coefficients of the cluster centroids in physical space for pairs of visually similar streamfunction anomaly maps produced by the two clustering techniques and compare the results for different values of k. We obtain the maps that correspond to the cluster centroids in the d-dimensional subspace by computing the EOF expansion of the 500-mb streamfunction field, *i.e.* the QG model's second level, truncated at d = 10.

Table 4a shows the coefficients for the four pairs of clusters obtained by matching the centroids from either method for k = 4. Agreement between the two methods is very good for all values of d. For other k-values we find that only some of the clusters correlate well. Table 4b gives the correlation coefficients between pairs for k = 3, 5, and 6 and d = 5.

a)

	a	**b**	**c**	**d**
d=2	**0.999**	**0.999**	**0.996**	**0.999**
d=3	**0.999**	**0.998**	**0.991**	**0.997**
d=4	**0.998**	**0.990**	**0.994**	**0.997**
d=5	**0.999**	**0.971**	**0.986**	**0.998**

b)

k=3	**0.994**	0.793	**0.997**			
k=5	**0.996**	**0.989**	**0.961**	0.854	**0.995**	
k=6	**0.956**	**0.912**	0.700	**0.821**	**0.994**	0.549

Table 4: Pattern correlation coefficients between cluster centroid maps of a mixture model, on the one hand, and those obtained by the k-means algorithm, on the other: *a*) for k = 4, and $2 \le d \le 6$; *b*) for k = 3, 5, 6 and d = 5. Values higher then 0.9 are in bold.

We conclude, therefore, that k = 4 yields the optimal set of clusters for our QG model. The probability density function based on the mixture model for k = 4 is projected in Figure 1 on the plane spanned by EOFs 1 and 2. Clusters **a**–**d** correspond to those in Table 4a.

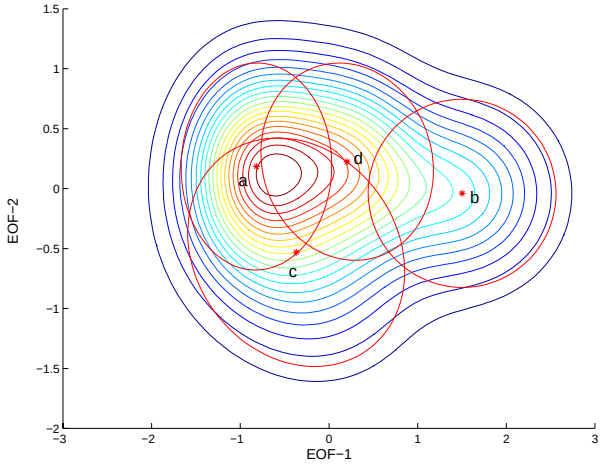

Fig.1: Probability density function — drawn with 20 contour levels — of the QG model's 500-mb stream function field, as

estimated by the mixture model for k = 4 and d = 5 and projected onto the plane spanned by EOF-1 and EOF-2. The centroids **a**–**d** have very high correlations with those obtained by the k-means method, as shown in Table 4a; the ellipses are spanned by the leading eigenvectors of the Gaussian components' associated covariance matrices.

The 500-mb streamfunction anomaly maps obtained by the k-means method for the cluster centroids shown in Figure 1 are plotted in Figure 2.

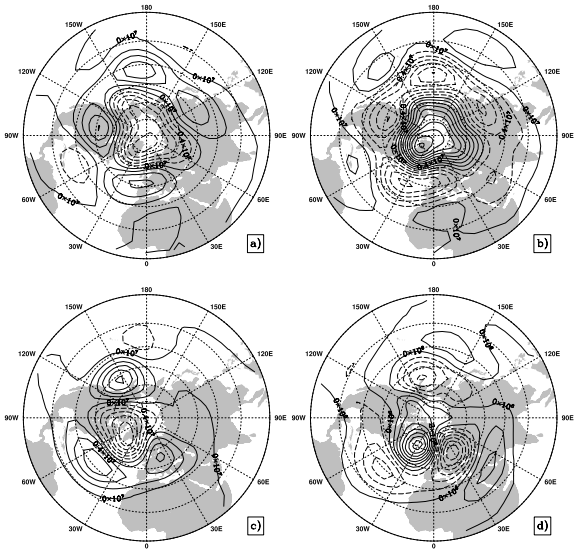

Fig. 2: Streamfunction anomaly maps of cluster centroids obtained by the k-means algorithm, for d = 5 and k = 4. Land masses are shaded.

Each of the regimes in Figure 2 represents one of the opposite phases of two spatial patterns. Clusters **c** and **d** capture the two extreme phases of the North-Atlantic Oscillation (NAO), while their patterns outside the Atlantic sector complete a NH wavenumber-three pattern. Clusters **a** and **b** have a central feature that extends over the whole Arctic and is rather zonally symmetric, with a substantial wavenumber-four component. It thus has certain features in common with the Arctic Oscillation (Thompson and Wallace 1998) and with Mo and Ghil's (1988) North-South seesaw. We denote these four regimes by AO^+ (panel *a*), AO^- (panel *b*), NAO^+ (panel *c*) and NAO^- (panel *d*).

3. PREFERRED TRANSITIONS

Using the clustering results for k = 4, the Markov chain of transitions between the four regimes is obtained. In the d-dimensional space, each weather regime is defined by the ellipsoid of covariance around the centroid, whose semi-axes equal the corresponding eigenvalues, as shown in Figure 1. A data point is assigned to a weather regime if it lies within the corresponding ellipsoid. If a data point belongs to several ellipsoids, we assign it according to the maximum probability value.

A Monte-Carlo simulation is applied to provide a statistical significance test for the elements of the transition matrix (Vautard et al. 1990). Table 5 shows transition probabilities between the mixture model regimes, obtained for d = 5. The preferred transition paths between the four regimes that are statistically significant at 99% appear in the table in bold.

	AO^+	AO^-	NAO^+	NAO^-
AO^+	**0.33**	0	**0.64**	0.03
AO^-	0	**0.54**	0.01	**0.45**
NAO^+	**0.55**	0	0.01	**0.44**
NAO^-	0.03	**0.18**	**0.59**	**0.20**

Table 5: Transition probabilities estimated using mixture models regimes for d=5; transitions that are significant at 99% are in bold.

Three of the four regimes have highly significant reinjection rates. Aside from these bold diagonal entries in the table, we note that a strong preferential path leads from the zonal sectorial regime NAO^+ to the high-index hemispheric regime AO^+, as well as from the blocked NAO^- to the low-index AO^-. The opposite transitions from AO^+ to NAO^+ and from AO^- to NAO^- are also highly significant. A preferential cycle connects, moreover, the sectorially blocked and zonal regimes NAO^- to NAO^+ and back. The hemispheric regimes AO^- and AO^+, however, are not directly connected to each other.

4. DATA ASSIMILATION USING PSAS

We use NASA Goddard's Physical-Space Statistical System (PSAS; Cohn et al. 1998) data assimilation framework to carry out identical-twin experiments with our QG model. The purpose of these experiments is to clarify the physical mechanisms of the regime transitions captured in Table 5.

Synthetic observations are simulated to correspond to both conventional and satellite networks. Their effects on pinpointing the transitions between regimes and capturing their causal mechanisms are evaluated. Implications of observing system design on extended-range prediction in the model are discussed.

5. REFERENCES

Cohn, S. E., da Silva, A., Guo, J., Sienkiewicz M., and Lamich D., 1998: Assessing the effects of data selection with the DAO Physical-space Statistical Analysis System. *Mon. Wea. Rev.*, **126**, 2913-2926.

Ghil, M., and Robertson, A. W., 2001: “Waves" vs. "particles" in the atmosphere's phase space: A pathway to long-range forecasting? *Proc. Natl. Acad. Sci*., accepted.

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

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

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

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

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

Smyth, P., Ide, K. and Ghil, M., 1999: Multiple regimes in Northern Hemisphere height fields via mixture model clustering. *J. Atmos. Sci.*, **56**, 3704-3723.

Thompson, D. W., and Wallace, J. M., 1998: The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett*., **25,** 1297-1300.

Vautard, R., Mo, K. C., and Ghil, M., 1990: Statistical significance test for transition matrices of atmospheric Markov chains, *J. Atmos. Sci.*, **47**, 1926-1931.