CLUSTER ANALYSIS AND WEATHER REGIMES

B. Legras
Laboratoire de Météorologie Dynamique, Paris, France

T. Desponts and B. Piguet
Ecole Nationale de la Météorologie, Toulouse, France

1. Introduction

The possible existence of multiple weather regimes within large-scale atmospheric flows has been the subject of a number of recent theoretical works. The observational evidence of such regimes relies on much less material and is essentially based on the subjective impression that large amplitude persistent anomalies, such as the blocking pattern which invigorates the mid-latitude winters, require a specific explanation. Although this appears sufficient to initiate a theoretical investigation, it is also desirable to possess a better characterization of weather regimes which does not only encompass extreme cases and to establish their separability. Most of the present observational work on anomalous flows is based on the use of scalar indices that are a priori chosen. The sole case from which a clear bimodality emerges is a study by Hansen and Sutera (1986) using a maximum penalized likelihood (MPL) technique applied to the estimate of an index that is a combination of the amplitudes of zonal wavenumbers 2 to 4. This index however does not seem to be directly related to blocking. There is indeed a very small hope to capture the complex structure of the atmospheric phase space through an arbitrary univariate approach. Whatever is really the large-scale dimension of the climatic attractor, EOF analysis tells us that at least ten independent modes are required in order to describe 50% of the low-frequency variance. It would be in principle possible to generalize the MPL technique to a multivariate problem but we then have to face unrealistic computational costs and the requirement of a much larger dataset than observations or even general circulation models can provide. Thus strictly objective nonparametric techniques are hardly applicable to our problem.

Two alternate routes are offered: The first one is to incorporate some dynamical information and to require that weather regimes are solutions of an equilibration problem. This approach has been developed by Vautard and Legras (1987) (see also Legras and Vautard, 1987) and has been successfully applied to the regimes of a baroclinic channel model. It however requires the a priori knowledge of the relevant dynamical scales which has to be established independently and remains a rather heavy tool. A second route uses cluster analysis techniques, in an attempt at the recognition of preferred patterns. Such methods are widely applied in many fields in which they provide useful exploratory tools. They have been seldom used in general circulation studies but are employed in synoptic meteorology to relate synoptic maps to local weather elements (Key and Crane, 1986; Yarnal and White, 1987; Kalkstein, Tan and Skindlov, 1987). The aim of this article is to show the potential of this approach in analyzing large-scale circulation but also to discuss its difficulties.

Section 2 introduces to cluster analysis; a few general notions are given and the algorithm used in following sections is presented. Section 3 describes the application to the baroclinic channel model previously mentioned from which the qualities and the drawbacks of the method can be discussed. Section 4 presents a preliminary study of the 500hPa geopotential over the Atlantic region and Section 5 contains conclusions and further discussion.

2 . Cluster analysis

Classification problems occur as one tries to identify patterns within a multi-dimensional dataset for which each dimension corresponds to one of the measured variables. In our case, the variables may be the amplitude coefficients of the geopotential at some sampling locations or a truncated series of the associated principal components. If no obvious organization of the data emerges but if an underlying structure is expected, there exists a wide set of techniques which can help the investigator in partitioning the dataset. No attempt to compare the various methods will be made here and we refer to Gordon (1981) and references herewith for a recent comprehensive review; see also Silverman (1986) for additional references. All methods first require the definition of a dissemblance measure d between two points or between two groups of points. For instance, one may use the the squared norm of the difference between two geopotential fields. The main classification algorithms fall within two groups, the agglomerative methods that generate a partition by a hierarchical sequences of grouping, and the iterative relocation methods that generate an optimal partition by moving elements from one group to another. The method used by Mo and Ghil (1987) and ours are of the second kind. Our algorithm is fully described in Diday (1972) and in Diday and Simon (1976). It contains a stage of properly so called cluster analysis and a second stage of typological grouping. The first stage consists in a series of iterations trying to optimize an adequation criteria between a partition and its representation. It can be summarized in the following way:

1] K kernels N_i are generated and each one contains P points taken at random from the data

2] The barycenter GN_i of each kernel is computed and the classes CL_i are generated by associating each data point to the closest kernel. The distance of an element x to a kernel N_i is defined as $d(x,N_i) = d(x,GN_i) + \mathrm{var}(N_i)$ when d is a quadratic measure.

3] The barycenter GC_i of each class is computed and the new kernels are generated by taking the P points of the dataset that are the closest to GC_i.

The steps 2 and 3 are repeated until the classification has converged. In the process a class is suppressed when its cardinal is less than P after completion of step 2. If this happens, the elements of the suppressed class are attributed to the remaining ones. The representation of the classes is given by their kernels and the algorithm can be shown to reach a local optimum of the quantity

$$W = \sum_{i=1}^{K} \mathrm{Card}(CL_i)\ [\ \mathrm{var}(CL_i) + \mathrm{var}(N_i) + d(GC_i,GN_i)\ ]. \qquad (1)$$

So doing, we try to obtain classes that are as compact as possible but there is no warranty that the optimum W is global. Starting from different seeds, the final number of classes may vary and for the same final number, they may possess different boundaries from one classification to another. A first improvement is to define a strong class as being the intersection of all the realizations of the same class for a set of independent classifications. By realization of the same class, we means the classes that are sufficiently close from one classification to another but differ in their details. Two classes CL_i and CL_j taken from two independent classifications are found similar if

$$\mathrm{Card}(CL_i \cap CL_j))/\ \mathrm{Card}(CL_i) \geq X\ , \qquad (2)$$

where X is a given threshold. If a bijection exists between the two classifications based on this criteria, we say that they are two realizations of the same partition and the set of strong classes associated with these two classifications consists in the intersection of the K couples of similar classes. Otherwise, we reject the classification that possesses the largest value for W. The strong classes associated to N classifications result from the intersection of N-1 similar classes.

We may remark that the classification with reduction of the number of classes are automatically rejected and that the reference classification is the one among N with the best optimization criteria W. The number of rejected classifications increases with X which has to be adjusted for each particular problem. In the sequel, X will generally be equal to 60%. When they are sufficiently big, the strong classes are representative of the core of the average classification. A measure of the stability of the strong class SC_i is given by

$$S = (N \,\mathrm{Card}(SC_i)) \;\; / \sum_{j=1,N} \mathrm{Card}(CL_i(j)) \,, \tag{3}$$

where $CL_i(j)$ is the j-th realization of the class CL_i. In practice, S tends to a limit when N is sufficiently large (from 30 to 50 in the sequel). The largest is S, the more stable is the strong class. When S is zero, we say that the class is mobile.

This approach however exhibits a few drawbacks which requires further developments :

- The number of strong classes is directly linked with the initial number of kernels.

- The rejection of a part of the classifications missuses the available information.

- The set of strong classes is not a partition of the dataset. In particular, the clusters of points always grouped together but in different classes following the classification are lost.

A more elaborate typology can be based on the analysis in strong patterns. A strong pattern SP_i is a set of elements that have been always grouped together over N classifications. The set F of all strong patterns thus results from the successive intersections of the whole N classifications and is a partition of the dataset. But this partition is now too fine for most practical purposes and we need an ultimate stage of grouping. At this point, we define the dissemblance $\delta(SP_i,SP_j)$ between two strong patterns as the number of times they have not been grouped in the same class. We then link in the same group all the strong patterns that possess a neighbor within the group that is closer than q as measured by δ. For each q a partition of F is so defined and it can be shown that these partitions all together form a hierarchy. The weak patterns WP_i are defined as the particular partition of F corresponding to a chosen level q_0 of the hierarchy[1]. We choose the value of q_0 in order to obtain some sets that are as close as possible to the strong classes . The homogeneity of a weak pattern can be measured by three quantities:

-The average distance between the strong patterns belonging to the weak pattern:

$$A(WP) \;= \sum_{i>j} \delta(SP_i,SP_j) \,/\, n(n\text{-}1), \tag{4}$$

where SP_i, $SP_j \in WP$ and n is the number of strong patterns within WP.

-The percentage of cases for which two strong patterns of the weak pattern are more than q_0 times in different classes:

$$B(WP) = \sum_{i>j} b(SP_i, SP_j) \;/\; n(n\text{-}1), \tag{5}$$

with $b = 0$ if $\delta(SP_i, SP_j) \leq q_0$

and $b = 1$ if $\delta(SP_i, SP_j) > q_0$.

B measures the chaining effect of the algorithm which may group in the same pattern some elements being far apart but connected by a chain of close neighbors.

-The degree of weakness of the weak pattern

$$C(WP) = \sum_{x \in WP} \sum_{j=1,n} \{1\text{-}\delta(x, SP_j)/N\} \;/\; n \;\mathrm{Card}(WP). \tag{6}$$

The weak pattern will be the more homogeneous when C is large. The maximum value C=1 is reached when the weak pattern is a strong pattern.

The three measures are complementary but C bears the advantage of taking into account the size of the strong patterns belonging to the weak pattern.

3. The regimes of a baroclinic channel model

3.1 Generalities

The model to be used here is fully described in Vautard, Déqué and Legras (1987). It consists in a two-layer baroclinic β-channel forced by a localized baroclinic jet superimposed to a uniform shear. Fig. 1 shows the mean flow in the upper layer, averaged over a long integration of 15000 days, in which the jet is visible in the first fourth of the channel. Figs. 2a and 2c show that the instantaneous flow exhibits a number of propagating disturbances. In Fig. 2c, they are superimposed to a large-scale dipole downstream of the jet. These two states are respectively taken from two periods of zonal and blocking regimes. The two large-scale patterns shown in Figs. 2b and 2d can be either obtained as solutions of a large-scale equilibration problem including the feedback from nonlinear transients (Vautard and Legras, 1987) or by compositing the states with higher probability of persistence (Vautard, Legras and Déqué, 1987). Fig. 3 shows the two-dimensional histogram of the first two principal components of the streamfunction of the large-scale part of the flow[1] and the location of zonal and blocking regimes in this diagram. The first two EOFs, not shown, respectively exhibit a modulation of the jet and a downstream dipole; they cumulate 50% of the large-scale variance.

[1] The large-scale flow is defined by the spectral truncation of longitudinal wavenumber ≤ 3 and of transverse wavenumber ≤ 2.

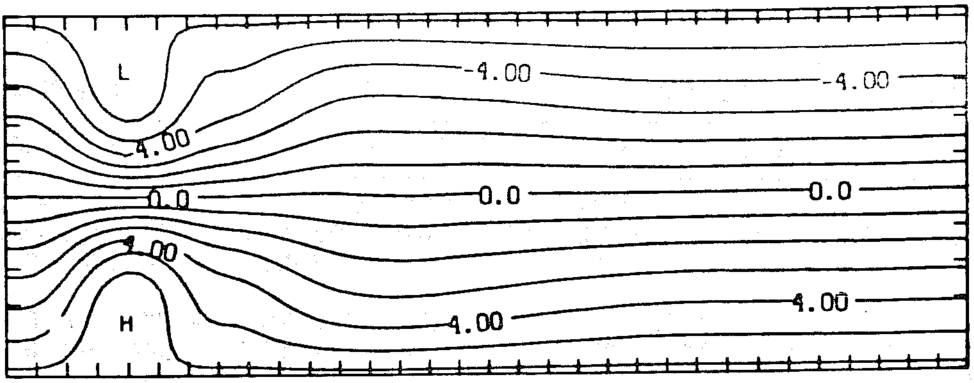


Fig. 1: Contours of the upper-layer streamfunction averaged over 15000 days of simulation.

3.2 Classification

We first tried to classify individual maps that were each described by a small number of principal components but we soon realized that weather regimes are also characterized by their temporal evolution. We then decided to consider elements which incorporate both spatial and temporal information. Each element of our dataset, called point-duration, is a series of principal components taken at regular intervals τ over a period of length T. In the sequel, we discuss more specifically the case with two components, $T = 13$ days and $\tau = 4$ days. Thus each point-duration is a vector with 8 components. The chosen metric is the Euclidian distance in the space of point-durations. Fig. 4 shows the projection of the strong classes onto the plane of the two first principal components at day 7 for a partition in two, three and four classes using N=50 classifications. The characteristics of the strong classes are detailed in Table 1. The partition in two classes is done along the direction of the first principal component which dominates the total variance; only 4% of the points are lost at this stage. Passing from two to three classes and from three to four classes, the new class is generated in the friction zone of the previous ones. The fourth class combines a small cardinal and a weak stability. The instability increases in the partition in five for which the fifth class is mobile. On the other hand, the cumulated stability is almost stationary from 2 to 4 classes and then drops abruptly for five classes. We retain the classification in four as a compromise between the stability of the description and its capacity to resolve different structures. Comparing with Fig. 3, it is clear that the strong classes tend to be rejected on the periphery of the cloud and that the central region is badly covered. Maximizing the inter-class distance induces a stable grouping of strongly anomalous fields. On the contrary, the point-durations of the central zone are submitted to a competition between the different classes and may change their assignment from one classification to the other. Consequently, they are almost surely not contained in any strong class.

Since about 42% of the classifications are rejected in the partition in four, we only use 25 classifications in the calculation of strong and weak patterns in order to keep a comparable amount of information. Then the number of strong patterns is 1564, 58% of which, corresponding to 6% of the dataset, contains only one point-duration. In order to reduce the level of grouping and the computational costs, we suppress these unitary classes before building the hierarchy. In addition, we also put a threshold on the weak patterns by considering only those containing more than 300 point-durations. Comparing the results for q_0=5 and q_0=8, we see that the covering of the strong classes by the weak patterns is better with q_0=8 (96% against 85%) but that the weakness C is smaller for q_0=5 which thus combines a good coverage of the strong classes and good homogeneity. Choosing q_0=5 in the sequel, we obtain 9 classes with more than 300 point-durations. Their characteristics are detailed in Table 2 which shows that they considerably differ by the number of included point-durations but that they possess rather uniform properties of homogeneity except for WP_6. Fig. 5a shows a schematic representation of the weak patterns in the factorial plane. In reality, their shape is not circular but they remain rather homogeneous, at least for the largest ones. This diagram allows to compare the respective positions of the weak patterns and the strong classes. We thus notice a cluster formed by the first four weak patterns and another one formed by the last two. This obser-

	2 CLASSES		3 CLASSES			4 CLASSES				5 CLASSES			
	SC 1	SC 2	SC 1	SC 2	SC 3	SC 1	SC2	SC 3	SC 4	SC 1	SC 2	SC 3	SC 4
PC1 D1	0.882	-0.531	1.306	-0.882	0.311	1.447	-0.912	0.348	-0.435	1.595	-0.992	-0.130	-0.866
PC1 D7	1.004	-0.644	1.451	-0.964	0.218	1.542	-1.045	-0.215	-0.448	1.713	-0.913	-0.480	-0.438
PC1 D13	0.873	-0.538	1.367	-0.875	-0.005	1.246	-0.833	-0.663	-0.828	1.350	-0.795	-0.642	-0.358
PC2 D1	0.020	0.000	-0.094	-0.061	0.119	0.083	-0.226	0.631	-0.682	-0.398	-0.604	0.601	-1.302
PC2 D7	0.020	0.002	-0.127	-0.238	0.327	-0.037	-0.518	0.847	-0.710	-0.592	0.932	0.804	-1.200
PC2 D13	0.080	0.005	0.025	-0.322	0.408	-0.155	-0.685	0.559	-0.350	-0.244	-0.784	0.551	0.007
CARD SC	5408	8960	2740	4915	2251	1848	3384	2386	371	717	1864	1434	300
LINE 1	5626	9324	3482	6740	4728	3141	4538	4838	2442	2550	3646	391	2413
LINE 2	218	218	576	744	324	270	373	481	327	240	632	462	352
LINE 3	96.1	96.1	78.8	72.9	47.6	58.8	74.6	49.4	15.2	28.1	36.6	12.4	51.1
LINE 4	192.2		199.2			198.0				128.2			
LINE 5	49		34			29				32			
LINE 6	4%		34%			46.7%				71.2%			

6 first lines: centers of gravity of the strong classes on the first and second principal components at days 1, 7 and 13; units are normalized by the variance of each principal component
CARD SC: cardinal of the strong class
LINE 1 : average cardinal of the intersecting classes
LINE 2 : rms of the cardinal of the intersecting classes
LINE 3 : % of stability
LINE 4 : cumulated % of stability
LINE 5 : number of intersecting partitions
LINE 6 : % of nonclassified points

Table 1: Characteristics of the strong classes for K varying from 2 to 5, with 2 principal components and $T = 13$ days

	C1	C2	C3	C4	C5
WP I	633	12	3.0	2.77	0.917
WP II	360	22	7.7	3.23	0.884
WP III	781	18	9.2	3.20	0.890
WP IV	2090	28	13.8	3.57	0.891
WP V	744	25	17.7	3.63	0.866
WP VI	1021	49	46.2	5.61	0.791
WP VII	3062	28	11.9	3.41	0.900
WP VIII	335	12	7.6	3.15	0.896
WP IX	1622	27	5.4	3.28	0.906

C1: number of elements in the weak pattern
C2: number of strong patterns belonging to the weak pattern
C3: % of violation of the threshold, criterion B
C4: average distance between strong patterns, criterion A
C5: degree of weakness, criterion C

Table 2: Characteristics of the weak patterns for $K = 4$, $2\ PC$, $T = 13$ days and $q_0 = 5$.

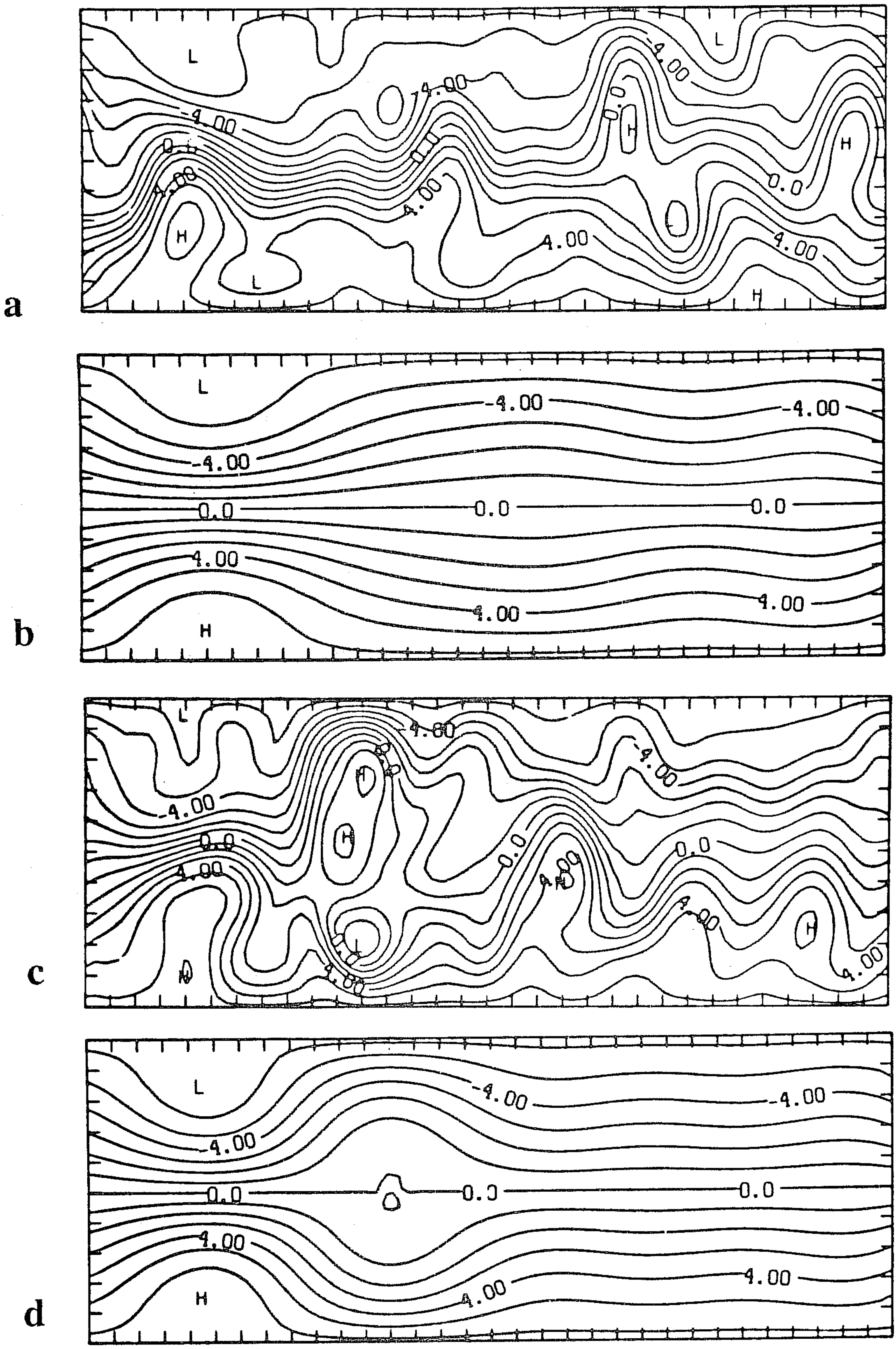


Fig. 2: Contours of the upper-layer streamfunction for a) an instanteneous flow taken from a zonal sequence b) the zonal composite c) an instantaneous flow taken from a blocking sequence d) the blocking composite

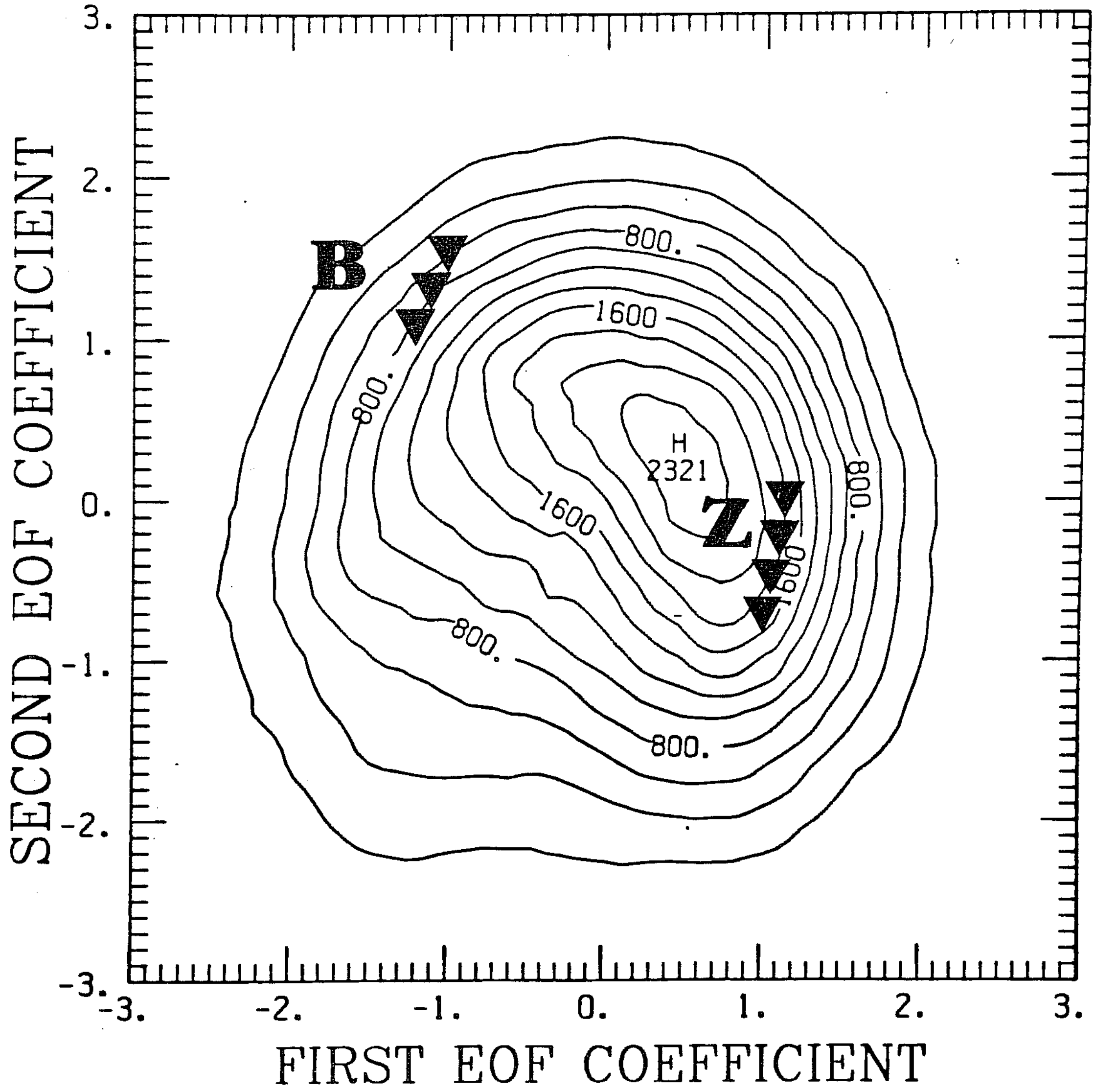


Fig. 3: Two-dimensional histogram of the first two principal components of the streamfunction of the large-scale part of the flow. B and Z marks show the location of blocking and zonal regimes as established from the two articles referred to in the text.

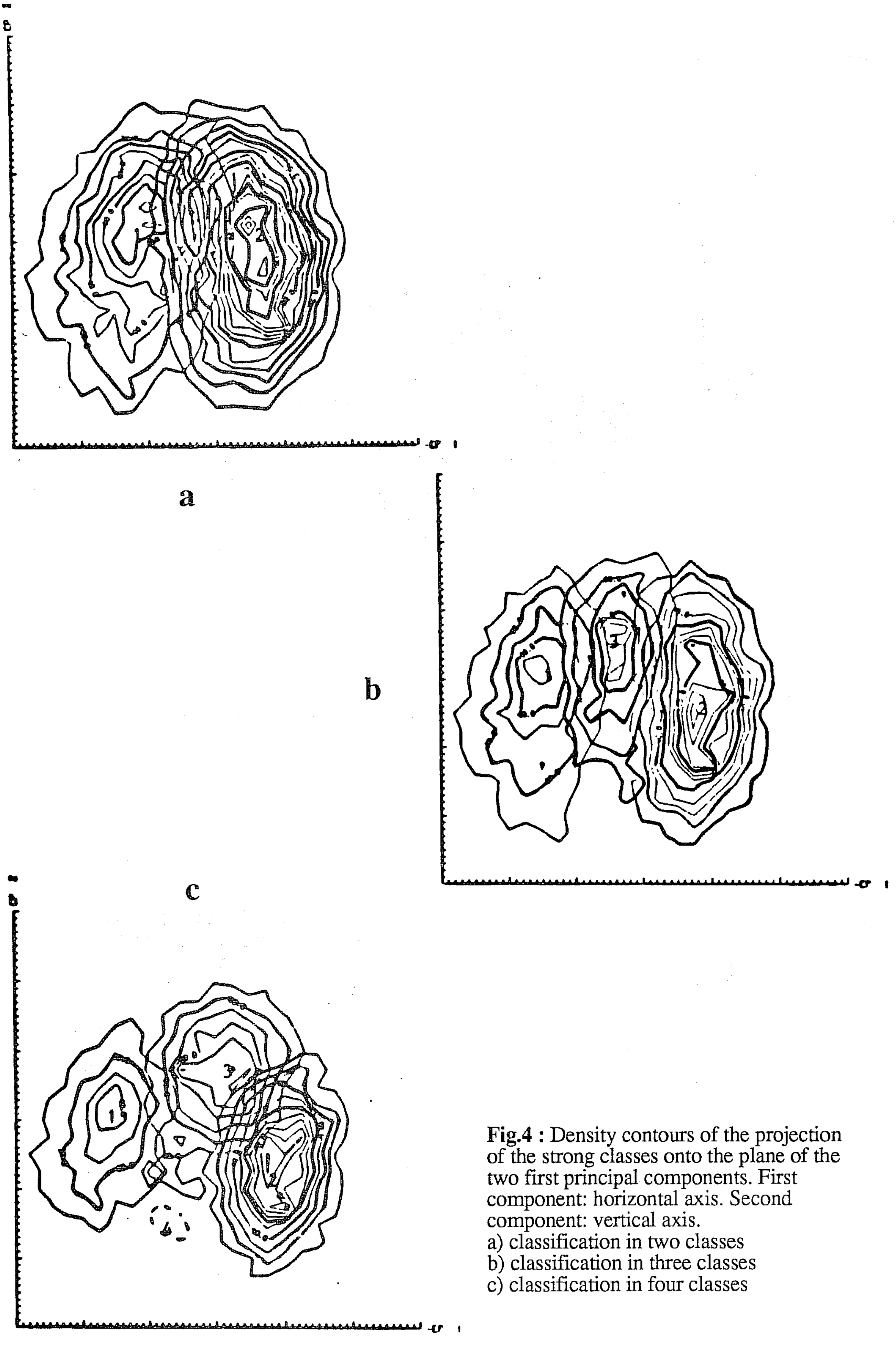


Fig.4 : Density contours of the projection of the strong classes onto the plane of the two first principal components. First component: horizontal axis. Second component: vertical axis.
a) classification in two classes
b) classification in three classes
c) classification in four classes

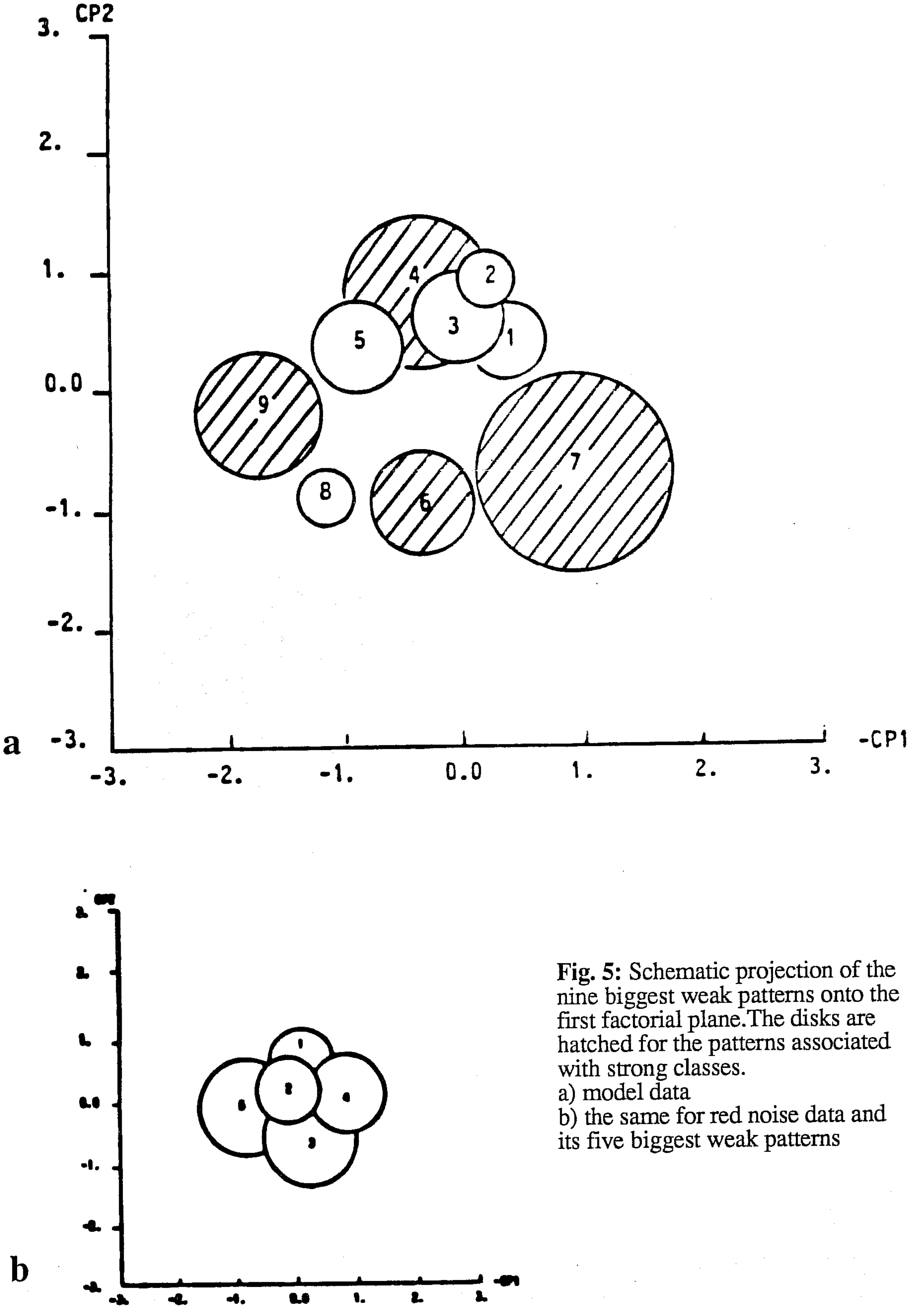


Fig. 5: Schematic projection of the nine biggest weak patterns onto the first factorial plane.The disks are hatched for the patterns associated with strong classes.
a) model data
b) the same for red noise data and its five biggest weak patterns

vation agrees with the fact that the weak patterns 2, 3, 4 and 8, 9 are respectively grouped together for q_0=8. Fig. 5b shows the same result for a multivariate red noise which possesses the same covariance matrix and the same correlation at one-day lag than our original data. Here we obtain 5 weak patterns with more than 300 elements which are grouped in a much more compact way than previously. This is expected since the cloud of points (without time) is isotropic.

3.3 Persistent regimes

Two of the strong classes, the first and the second, are particularly stable (cf. Table 1). They are respectively associated with the ninth and the fifth weak pattern. Fig. 6 shows the composite of the point-durations at day seven for the ninth weak pattern. The flow is characterized by a a strong stationary blocking dipole downstream of the jet. From one sequence of point-durations to the other, the phase of the structure is very similar but the amplitude varies. Fig. 7 shows the histograms of the first and second principal components at days 1, 7 and 13 for all the point-durations of the weak pattern. It shows how stationary and homogeneous the pattern is on the first direction and how badly localized it is on the second one. The homogeneity is more visible on Fig. 8 which shows the evolution of the two first components over normalized sequences[1] with a confidence interval containing 60% of the population[2]. For the first principal component, the interperiod variance contains only 18% of the total variance while it contains 35% for the second principal component. It is also interesting to consider the composite of the first and the last five days of a sequence, shown in Fig. 9a., and to see the rather rapid onset and decay of the blocking. How the observed behavior differs from a random noise can be tested by comparing with the fifth weak pattern of the multivariate red noise. Fig. 9b shows the corresponding composite of the beginning and the end of a sequence and establishes that the features of Fig. 9a are relevant. On the other hand, the average of the composite is found significantly (95% level) more intense than for the red noise but the variances of the two weak patterns cannot be distinguished at the same level. In other words, the blocking pattern is a well defined local stationary structure with fast onset and decay, but its variability does not differ from a red noise. A second class, the fifth, which is close to the seventh on Fig. 5a, exhibits a stationary blocking pattern located slightly downstream of the previous one.

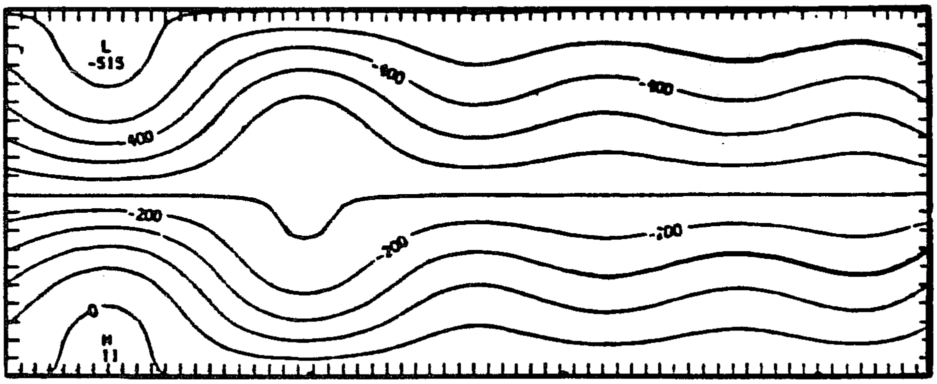


Fig. 6: Composite of the seventh day of the point-durations belonging to WP$_9$

A similar study that we are not going to detail can be conducted on the seventh weak pattern, the composite of which at day 7 is shown in Fig. 10. The relative intraperiod variances of the first two principal components are now 18% and 16% respectively, showing that the second principal component contributes more effectively to the characterization of this zonal pattern than it does for the blocking pattern. It differs also from the previous case by being less stationary. Indeed, a slow eastward propagation can be detected on the second principal component and can be proved to be significant. When the third principal component is

[1] A sequence is a series of consecutive point-durations belonging to the same weak pattern.

[2] The 95% confidence interval of the average is here approximately five times smaller.

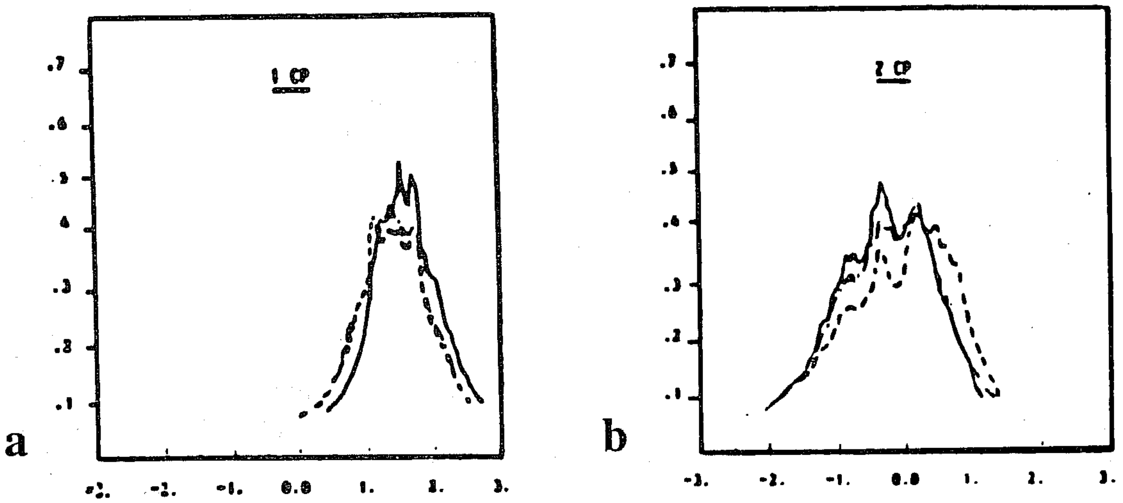


Fig. 7: Histograms of the first a) and the second b) principal components at days 1, 7 and 13 for all the point-durations of WP_9. Mixed line: day 1. Solid line: day 7. Dashed line: day 13

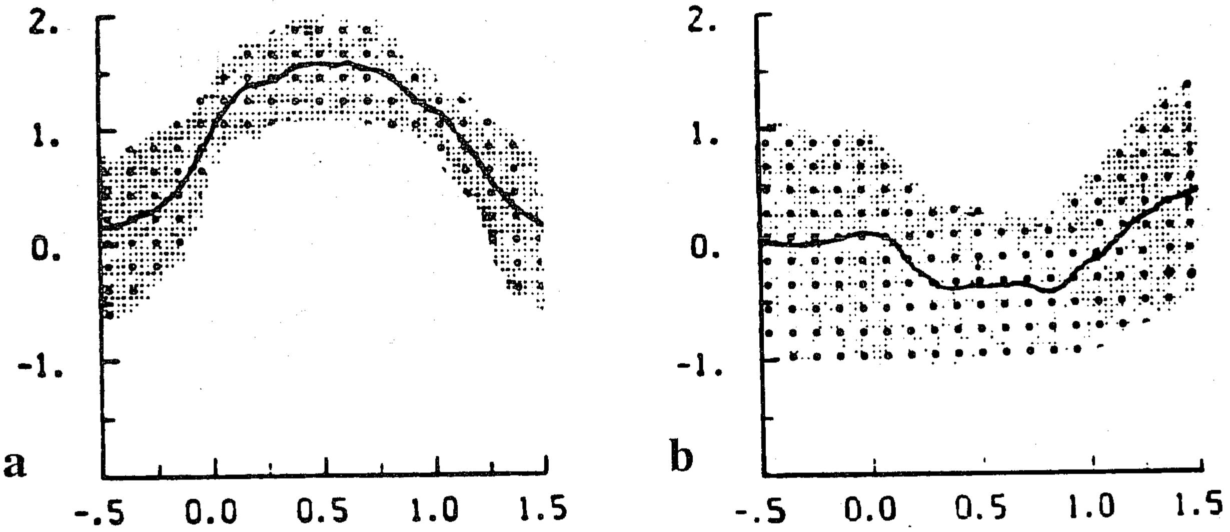


Fig. 8: Evolution of the first a) and the second b) principal components over a normalized sequence of WP_9. Solid line: average evolution. The confidence interval containing 60% of the population is shown as the dotted domain.

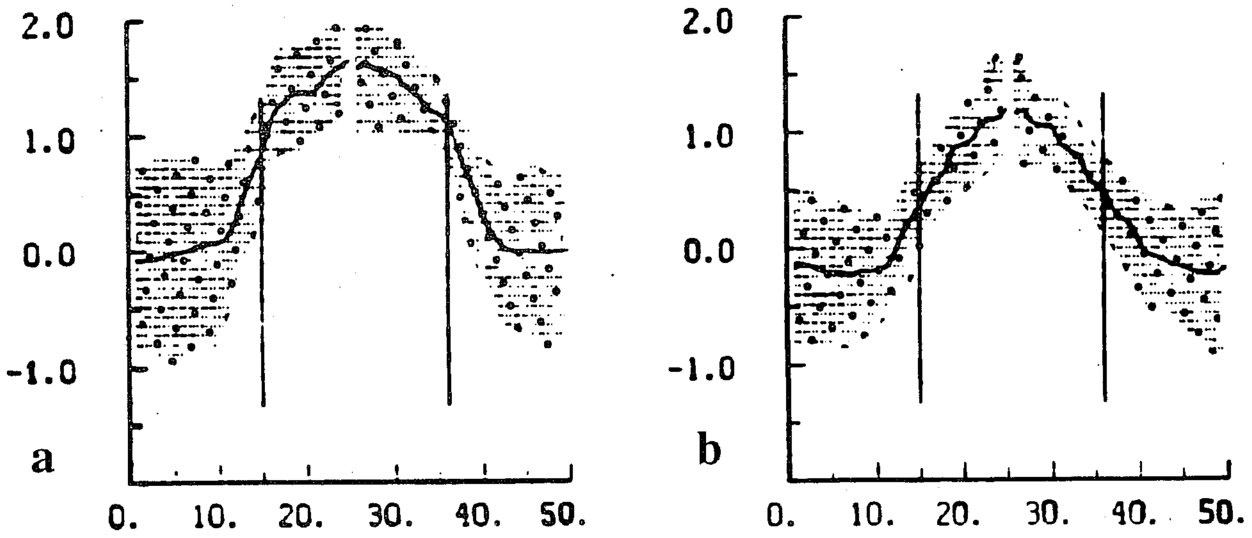


Fig. 9: a) Evolution of the first principal component showing the fifteen days which preceed and follow the sequences of WP_9 and the first and last ten days of the sequences. b) The same for the fifth weak pattern of the red noise.

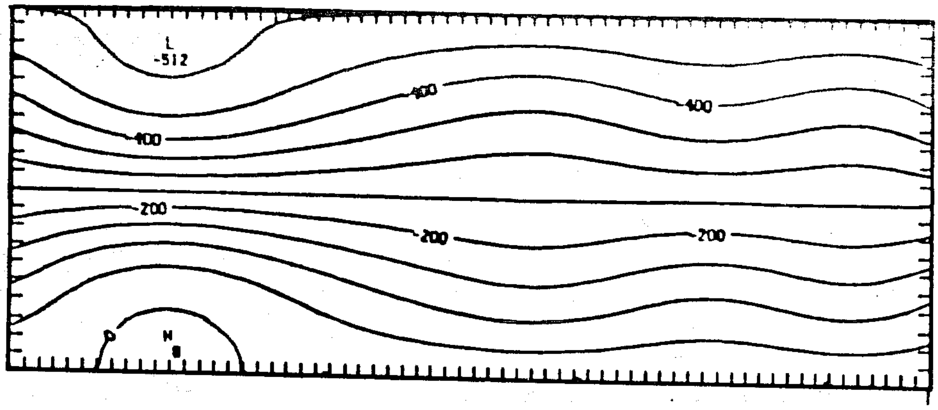


Fig. 10: Composite of the seventh day of the point-durations belonging to WP_7

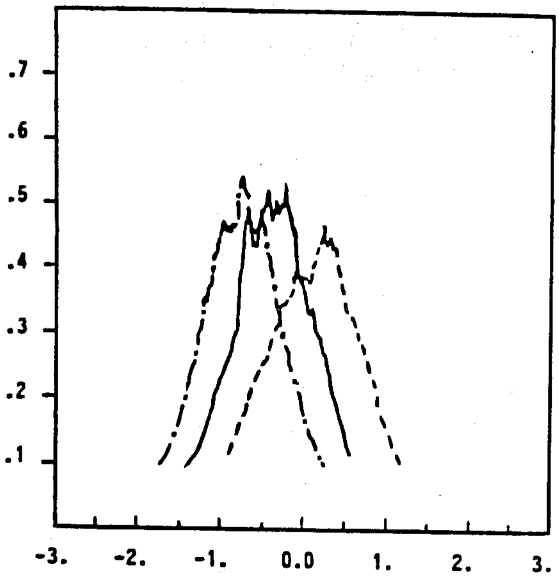


Fig. 11: Same as Fig. 7 but for the first component of WP_4

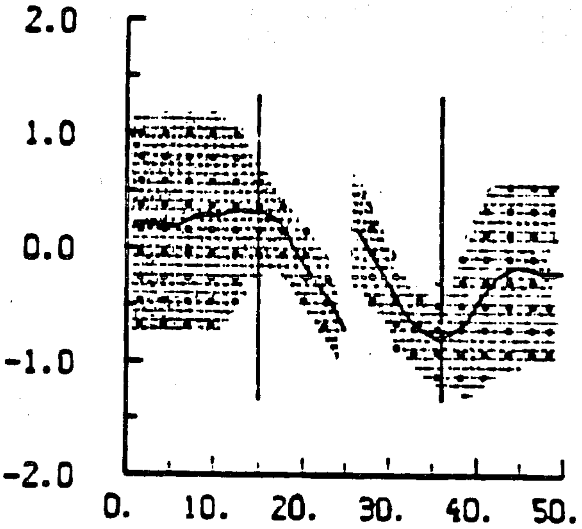


Fig. 12: Same as Fig. 9 but for the first component of WP_4

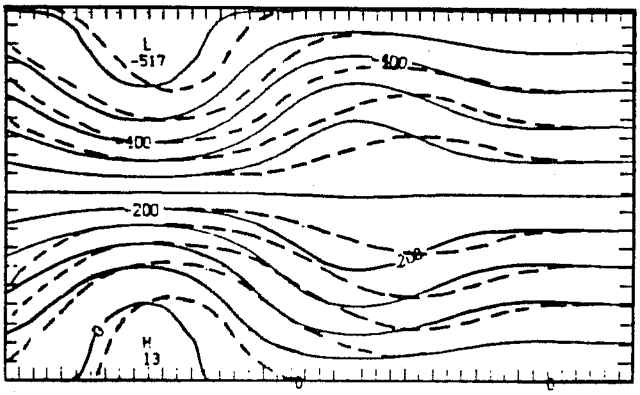


Fig. 13: Composite of the begining (solid line) and the end (dashed line) of the sequences of WP_4

introduced in the analysis, the blocking pattern remains a fixed point of the classification while the zonal pattern breaks into several patterns each corresponding to a different stages of a wave propagation superimposed to the mean zonal flow.

3.4 Evolving sequences and temporal chaining

The temporal dimension of point-durations allows the capture of nonstationary events. Fig. 11 shows the histograms at days 1, 7 and 13 of the first principal component of the fourth weak pattern. This pattern is the last one to be associated with a stable strong class. It is characterized by a strong evolution of the average over the duration of the sequences. The composites of the beginning and the end of the sequences, Fig. 12, shows that the interval containing 60% of the population is particularly small and consequently that the feature is well defined. From the corresponding streamfunction fields, cf Fig. 13, it appears that the pattern is an eastward damped propagation which is suggestive of an intermediate situation between blocking and zonal regimes. The first weak pattern exhibits a similar behavior.

The temporal chaining of the weak patterns is further investigated by counting the number of point-durations that overlap two weak patterns. From the result of this analysis a temporal chain which is summarized in the following diagram can be detected between five weak patterns:

$$\begin{tikzcd}
63\% & & 47\% \quad 52\% & & 77\% \quad 76\% & & 71\% \quad 76\% & & 78\% \\
5 \arrow[rr, dashed] & & 4 \arrow[rr, dashed] & & 1 \arrow[rr, dashed] & & 7 \arrow[rr, dashed] & & 6 \\
9\% & & 7\% \quad 3\% \arrow[ll] & & 4\% \quad 22\% \arrow[ll] & & 21\% \quad 17\% \arrow[ll] & & 18\% \arrow[ll]
\end{tikzcd}$$

with
$$\begin{tikzcd}
x\% & & y\% \\
A \arrow[rr, dashed] & & B \\
x'\% & & y'\% \arrow[ll]
\end{tikzcd}$$

x : % of sequences of A that end in B
y : % of sequences of B that begin in A
x': % of sequences of A that end in B
y': % of sequences of B that begin in A

This chain is completed by a direct bridge between the fourth and the seventh weak pattern:

$$\begin{tikzcd}
38\% & & 53\% \\
4 \arrow[rr, dashed] & & 7 \\
3\% & & 4\% \arrow[ll]
\end{tikzcd}$$

On the other hand, the weak patterns 2, 3 and 4 are very linked together:

-92% of the periods of WP_2 overlap 40% of the periods of WP_4

-100% ---------------- WP_3 --------- 63% -------------------------

but it is remarkable that there is no dominating direction in this case. The strong dependency is simply due to the proximity of these weak patterns.

Thus, there is a preferred temporal evolution from WP_5 to WP_6, passing through several regimes, among which the zonal pattern examined earlier, over an average duration of 50 days. The underlying mechanisms is apparently the eastward propagation of a damped large-scale wave during the transition between blocking and zonal flow. It is not possible to find a preferred chaining leading back to a blocking pattern. Consequently, we expect that the decay of the blocking is more predictable than its onset in our model. It is interesting to notice that the operational forecasts of the real atmosphere behave exactly in the same way (see Tibaldi in the same issue). The coincidence may be fortuitous and even misleading, see Section 4.3 below, but it is clear that cluster analysis offers a way for further investigations.

4. Analysis of the geopotential at 500hPa

4.1 The dataset

The data used in this study are the daily records at 12H TU of the geopotential at 500hPa from the analysis of the ECMWF and for the period 1980-1985. As usually done (Hoskins, Simmons and Andrews, 1977) we have divided the geopotential by the Coriolis parameter in order to work on an approximation of the streamfunction field. Having still in view the blocking phenomenon, we have limited our study to a regional domain centered over Northern Atlantic ocean. Fig. 14 shows the average field over this domain for the period 1980-1985. The small central domain is the "ATL" zone where Dole and Gordon (1983) observe a large number of persistent anomalies. Our domain encompasses also the jet area over North America and Western Europe. In order to apply our analysis to a large enough dataset, the total year was used after removing the seasonal cycle. We are aware that this is a rather questionable procedure since the seasonal cycle computed over the six years is itself dependent of the occurrence or the nonoccurrence of anomalous flows. as a matter of factour seasonal cycle exhibits around ay 120 a rather large deviation from the previous and the following periods. We do not apply any filtering to the data except an interpolation to a 5° grid.

Fig. 15 shows the pattern of the first EOF which describes 13.7% of the total variance and is associated with a modulation of the jet. 50% of the total variance is contained within the five first EOFs. They all show large-scale structures but their spatial configuration is largely due to the imposed orthogonality, so that it is not useful to further comment upon them. Owing to the relatively small amount of data, we have been able to take into account a large number of principal components in our analysis, up to 30, but our similarity measure is always dominated by the contribution of the first modes.

In the sequel we will use a representation of flow patterns based on the addition of the anomaly and the average field shown in Fig. 14. This may introduces a bias if the anomaly only results from the compositing of cases belonging to the same season. We have also used an other procedure in which the plotted field is the average over the selected cases of the anomaly. Because all our classes are well distributed around the year, we did not find noticeable visual differences between the two procedures.

4.2 Spatial classification

We first describe the results of a spatial classification of individual weather maps without using the temporal dimension. The grouping in strong classes and in strong and weak patterns is obtained from a series of 50 independent classifications initiated with random seeds. The unitary strong patterns have been eliminated and the threshold q_0 is chosen equal to 15. The percentage of rejected classifications in the generation of strong classes is found minimum (below 30%) for a classification in four and a number of principal components larger than 15. We retain these parameters as optimal and the results of the classification in four with 15 principal components are shown in Table 3. The average duration is computed from the length of the sequences of consecutive days included in the same class. Although the distribution of sequence lengths does not show a preferred maximum but is rather close to an exponential decay law, the average duration allows to compare the persistence of the various classes. Notice that the stability of the strong classes is much smaller here than for the quasi-geostrophic model. Another major drawback of this preliminary analysis is that the unitary strong patterns contain about 60% of the total points. By eliminating them, we concentrate our analysis on the strongly anomalous flow on the periphery of our dataset. A similar step was applied by Mo and Ghil who defined a "nonrecurring cluster" taking about two thirds of all analyzed maps.

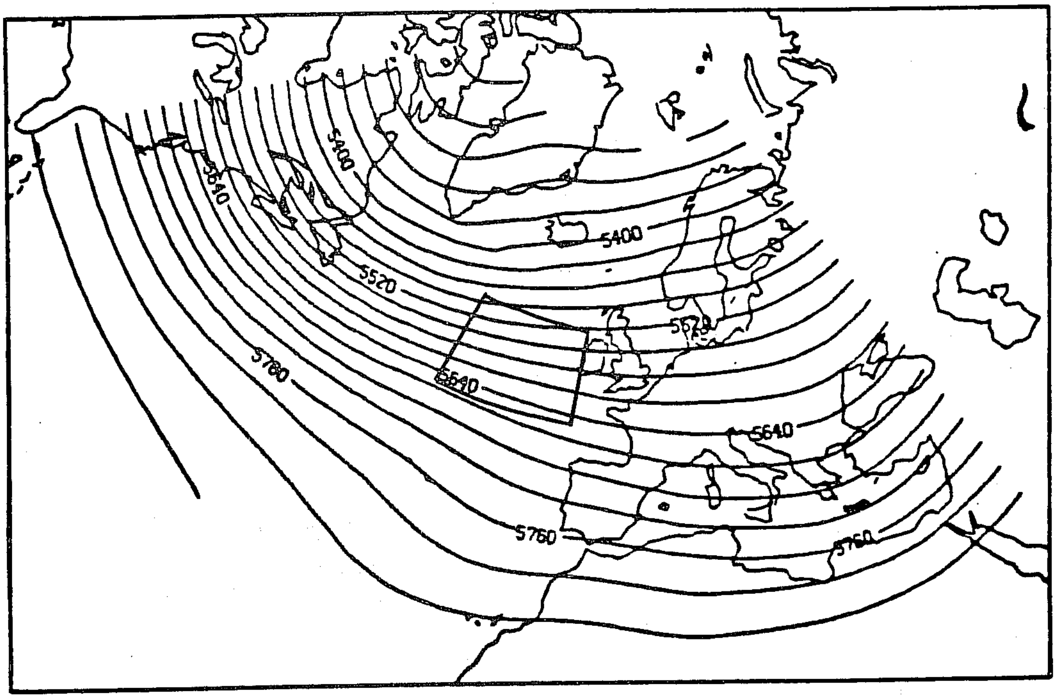


Fig. 14: Average field over the study domain. The small central square is the "ATL" region of Dole and Gordon (1983)

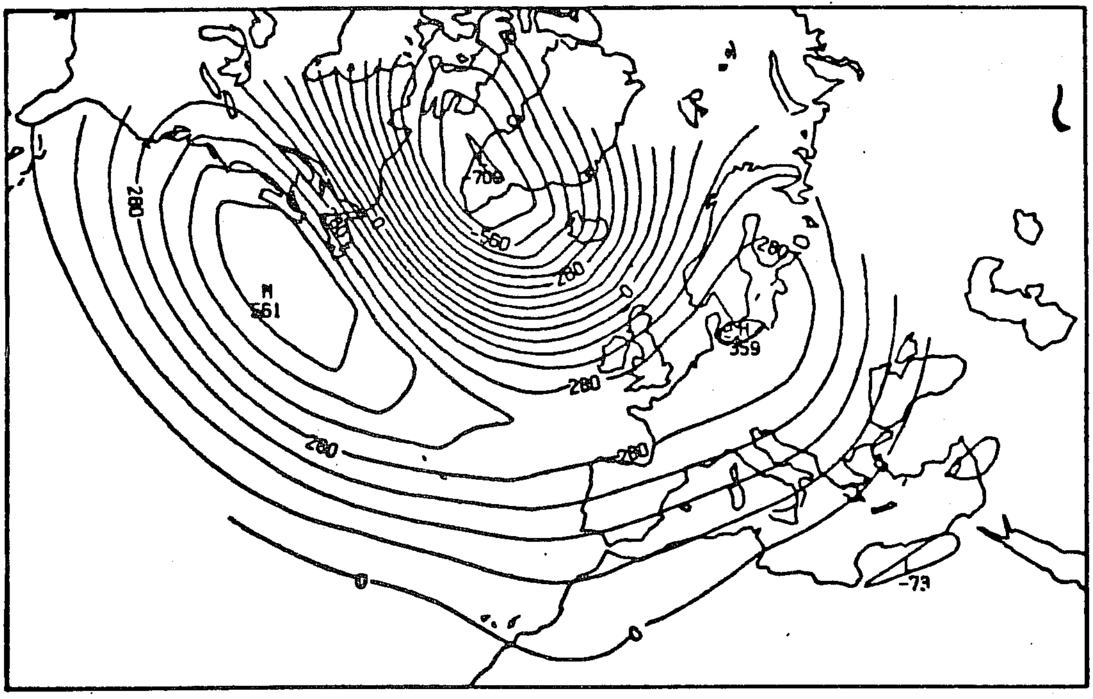


Fig. 15: First EOF of the geopotential after removing the seasonal cycle.

A striking fact here is the agreement between the second strong class and the sixth weak pattern (more than 90% of the elements of the former belong to the later). Furthermore, this strong class is the one that possesses the longest average duration and the highest percentage of stability, and this weak pattern is the most homogeneous (no chaining effect in spite of a large cardinal). We may thus consider that this class is well characterized. The corresponding field, Fig. 16c, differs significantly from the average flow with a dipole located on Labrador and South of Greenland and a southward displacement of the jet . The variance map, Fig. 16d, shows that the Southern Greenland anomaly is the most typical feature of this class (a region of the pattern is well typified when the variance of the class is small there).

	Strong classes				Weak patterns	
	% of stab.	Cardinal	Var.(*10^{-6})	Aver. duration	B (% of violation of q_0)	
1	21.0	107	6.63	2.0	WP1 : 0.01	WP5 : 0.05
2	48.8	228	8.84	4.0	WP2 : 0.00	WP6 : 0.00
3	18.4	95	6.03	1.6	WP3 : 0.07	WP7 : 0.22
4	32.2	225	7.30	2.4	WP4 : 0.00	

Cardinals of the intersections WP / SC

	WP →	1	2	3	4	5	6	7
SC ↓	Card.	171	41	87	102	91	258	87
1	107			14	58			
2	228						212	
3	95							38
4	225	62	41			54		

Table 3: Results of the classification in 4 with 15 principal components.

The third strong class and the seventh weak pattern show a very different situation. Although they only share common points together, the number of these latter is small. They also exhibit the lowest stability and the weakest homogeneity. They thus appear to be much less typified than in the previous case. The corresponding field, Fig. 16e, is a broad ridge over the Atlantic inducing a northward displacement and a reinforcement of the jet. The trough is a characteristic of this pattern, being associated with the smallest variance, Fig. 16f.

The first and the fourth strong classes possesses intermediate percentages of stability and appear as relatively homogeneous classes, the properties of which can be specified by the analysis in weak patterns. The fourth class shows a strong ridge over Western Europe, Fig. 16g, associated to a northward displacement and a well marked weakening of the jet. We may thus denote this case as the blocking pattern. This class is the one for which the "averaging effect" is the more visible: all the fields grouped in this class exhibit well defined ridges or isolated highs but their location varies quite strongly (mainly in longitude from England to Eastern Germany) and the intensity of the average ridge is somewhat underestimated. The three corresponding weak patterns show this phenomenon: the field of WP_1 is a ridge oriented Germany-Scandinavia, the field of WP_2 is an isolated maximum extending from Scotland to Sweden, and WP_5 is a ridge from Southern Ireland to Scotland. The variance map, Fig. 16h, shows that the characteristic part of the blocking is its oceanic and British component, the part located over continental Europe being more variable.

The first strong class and its two weak pattern dividers possess an almost perfectly zonal structure, Fig. 16a, extending across the ocean and Western Europe. The jet is reinforced downstream of its climatological position. The striking point is the location of the minimum in the variance map, Fig. 16b, between Scotland and Norway. This feature might be due to the fact that in this situation, the synoptic perturbations travel on the south of this region.

A general property emerges from the variance maps: in all cases, the maximum is placed over the Canada, in the North-West part of the domain. This area is a center of high-frequency variability (Blackmon, 1976), but this latter does not induce a partitioning of the atmospheric states since the various configurations of the atmospheric flow that have been isolated by the classification are never characterized by typical values at this location. On the contrary, the oceanic domain on the West of Ireland is always included in the area of small variance and thus seems to be the best domain for the partitioning of atmospheric states. These observations are consistent with the results of Dole and Gordon who found a large number of persistent anomalies in the "ATL" region, cf. Fig. 14, and a small number over Northern America. Although no temporal dependency is considered in our classification, it appears that the properties of persistence are naturally conditioning the structure of atmospheric states. Notice that the previous conclusions may look as contradicting the small average durations of the sequences, cf. Table 3. Numerous long lasting anomalies are in fact interspersed with transient bursts and are cut into several pieces by our analysis based on unfiltered data. The difficulty was already mentioned by Dole and Gordon who applied a low-pass filter to their original data. Finally there exists a domain of large variance over Southern Europe and Western Mediterranea for the strong classes 3 and 4. A possible explanation lies in the Mediterranean cyclogenesis which has been often linked with the occurrence of a ridge over the Atlantic or North-Western Europe (e.g. Tibaldi and Buzzi, 1983).

4.3 Spatial-temporal classification

Point-durations are defined for the geopotential field in the same way as for the streamfunction of our quasi-geostrophic model. But the decorrelation times are now much shorter so that we choose T=7 days and τ=2 days that are respectively the typical decay times of the large-scale and the synoptic-scale autocorrelation function of the geopotential. The number of principal components and the number of initial kernels are again adjusted to get maximum cumulated stability and minimum rejection in the generation of strong classes. The optimum was obtained with the classification in four with 20 principal components.

Table 4 compares the strong classes obtained from the spatial-temporal classification and from the spatial classification. Each strong class of one type intersects only one strong class of the other type, thus showing that the same phenomena are identified in both analysis. However, the intersection always contains less than 60% of the points of the intersecting classes. This is largely due to the effect of intermittent transients upon the classification.

	15pc1d →	1	2	3	4
20pc7d ↓	Card.	203	214	120	246
1	107	65			
2	228				134
3	95			25	
4	225		82		

Table 4: Cardinals of the intersections between strong classes obtained from spatial-temporal and spatial classifications.

The characteristics of the strong classes and the weak patterns are shown in Table 5. The average duration of sequences (for which only the central day of each point-duration has been considered) is roughly doubling for each strong class with respect to its value in the spatial classification. The average patterns of the strong classes, not shown, closely resemble those of the spatial classification, cf. Fig. 16. As previously, the weak patterns perform a more accurate classification than the strong classes. The three weak patterns 1, 4 and 6 are all in exclusive association with a single strong class but the weak patterns 7, 8, 9 and 10 seem to divide the

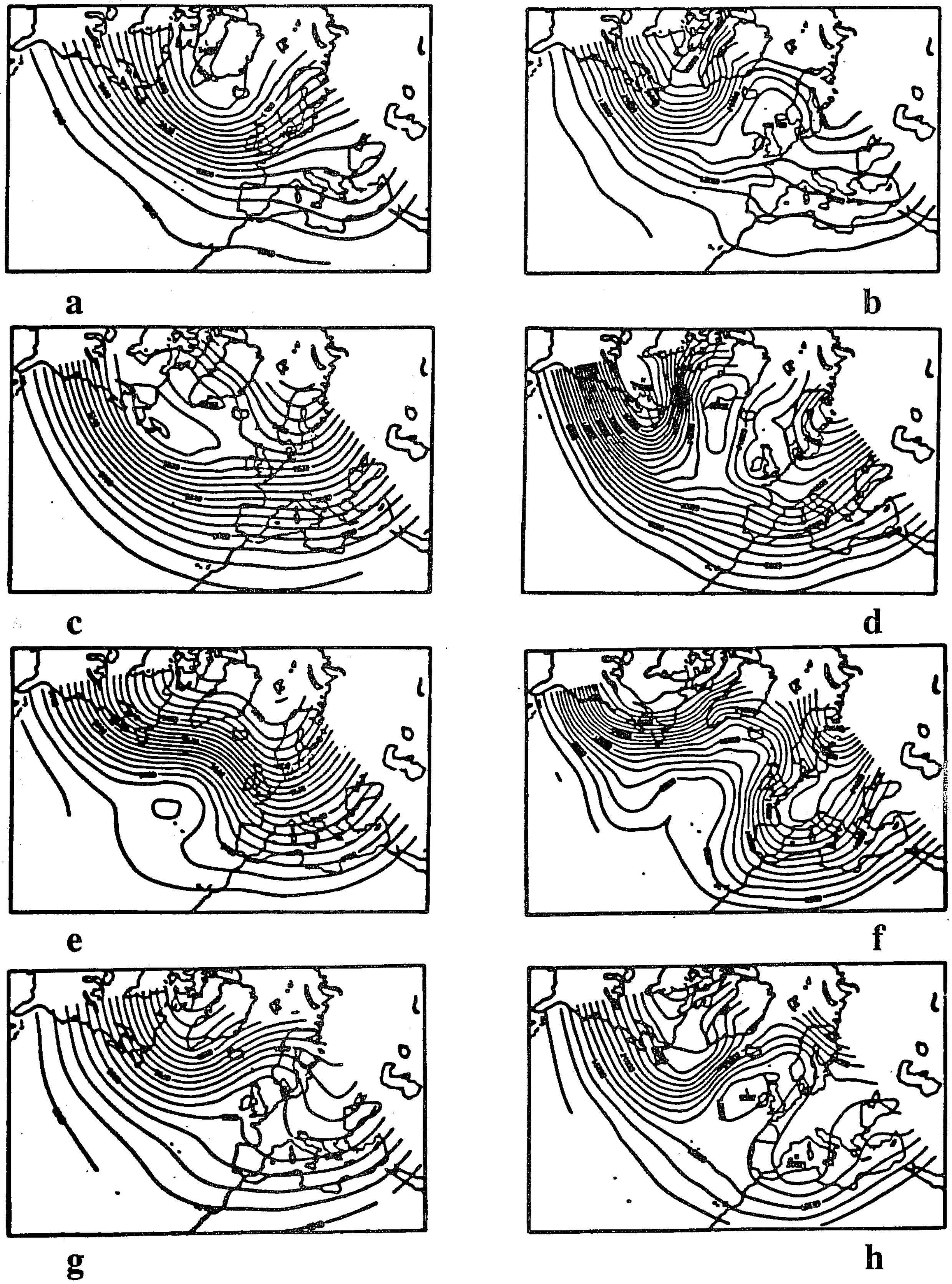


Fig. 16: Geopotential fields and variance maps associated with the strong classes of the classification in four with 15 principal components.
a) and b) geopotential and variance of the first strong class
c) and d) geopotential and variance of the second strong class
e) and f) geopotential and variance of the third strong class
g) and h) geopotential and variance of the fourth strong class

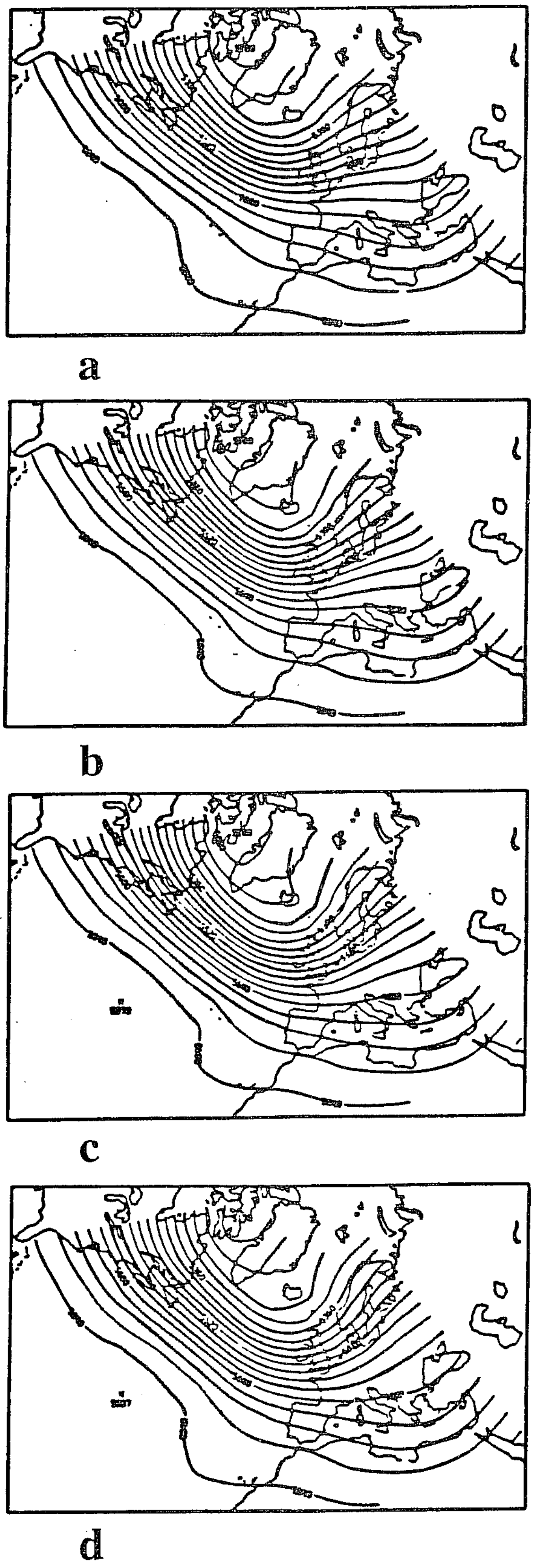


Fig. 17: Composite of the zonal class (SP_1) of the spatial-temporal classification at days
a) 1, b) 3, c) 5, d) 7

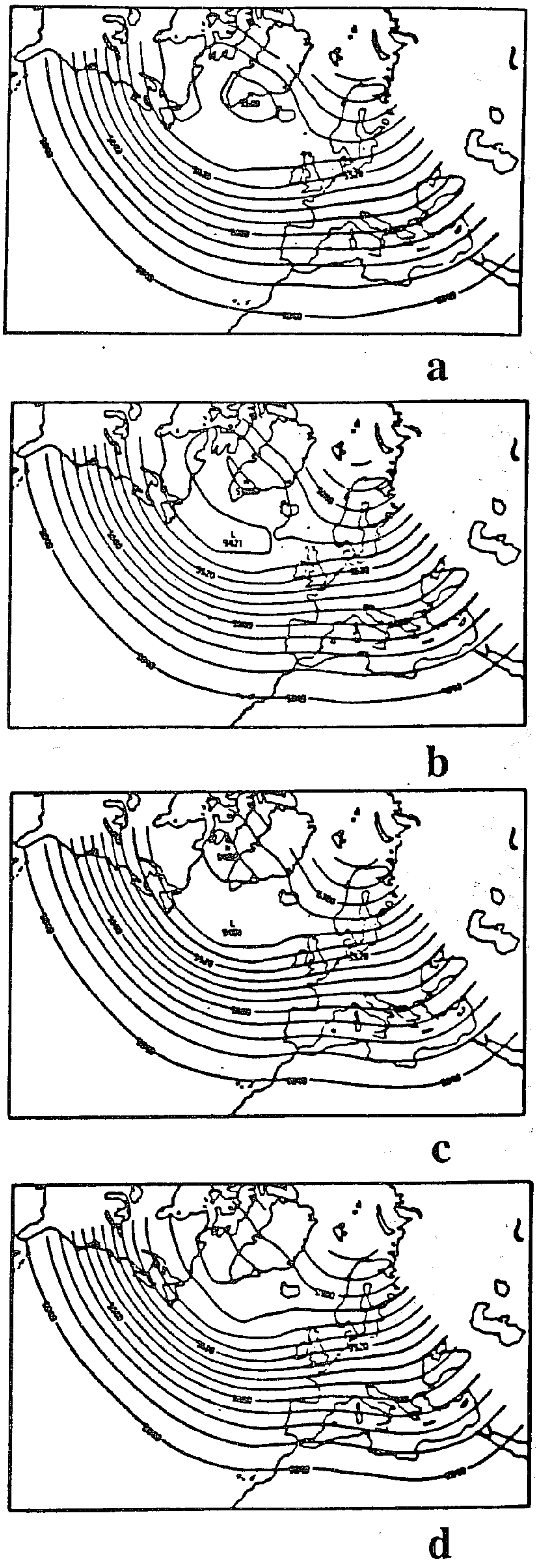


Fig. 18: Same as Fig. 17 but for the Labrador anomaly (SP_4)

second strong class associated with blocking circulation. Two weak patterns are not associated with any strong class.

Percentage of stability of SPs :	1(zonal)	2(block)	3(ridge)	4(Labrador)
	42.0	30.1	21.4	57.2

Criterion B for the WPs:

1	2	3	4	5	6	7	8	9	10
0.01	0.00	0.08	0.00	0.13	0.01	0.00	0.00	0.00	0.07

Intersections SP / WP :

N°	WP →	1	2	3	4	5	6	7	8	9	10
SP ↓	Card.	196	46	61	83	35	161	81	104	62	40
1	203					X	149				
2	214					X		75	76	26	10
3	120				76	X					
4	246	186				X					

Table 5: Results of the spatial-temporal classification in 4 with 20 *PC* and $T = 7$ days.

The zonal class (SP_1/WP_6), Fig. 17, is the more stationary: the composite at different times of the involved point-durations can be superimposed without any visual differences except in the North-West of the domain. The Labrador anomaly (SP_4/WP_1), Fig. 18, is slowly evolving: the configuration is maintained but the value of the maximum slowly decays from 5500m to 5400m. Since the variance at this location is 18000m^2 and the number of elements is about 200, the estimation is done with an error interval ±20m at 95% confidence level. The trend is thus significant. The Atlantic ridge (SP_3/WP_4), not shown, and the blocking anomaly (SP_2), Fig. 19, are both amplified up to day 5 and damped at day 7 but this is not found significant with the same criterion. Further details are available for the blocking anomaly owing to its separability into several weak patterns. These latter show various stages of blocking: WP_8 and WP_9, Fig. 20, appear as developing ridges over England while WP_7 and WP_{10}, not shown, appear as developing ridges over Germany and Poland. In addition, the weak patterns WP_2 and WP_3 which do not intersect SP_2 appear as ends of blocking. In order to investigate the possible links of this multiplicity (we possess four different onsets) with different types of blocking, we build a contingency table, cf. Table 6, in which each entry shows the number of times there exists at least one day belonging to WP_j within the 7 days following a sequence of WP_i. The six weak patterns 10, 9, 8, 7, 3 and 2 are linked by a chain going from WP_{10} to WP_3 passing through the other patterns but there is no way to separate several branches. The blocking anomaly thus appears as a unique phenomenon which may largely fluctuate in phase.

Transitions between regimes are better studied by returning to the strong classes. Table 7 shows the number of cases of transitions to SP_j within the 7 days following the sequences of SP_i. When a smaller delay is considered, almost no transition occur (2 with a delay of 3 days) but a longer delay does not change the tendencies (Table 7 also shows the contingency table for 11 days). Although the number of events is too small to give a level of confidence, we may draw the following results:

-All transitions do not have equal probability.

-The most frequent transition is SP_1 ---> SP_2 (zonal ---> blocking).

-The transitions "blocking <---> Labrador anomaly" do not show a preferred direction.

-There are few transitions from the Atlantic ridge.

These results are summarized in the following diagram:

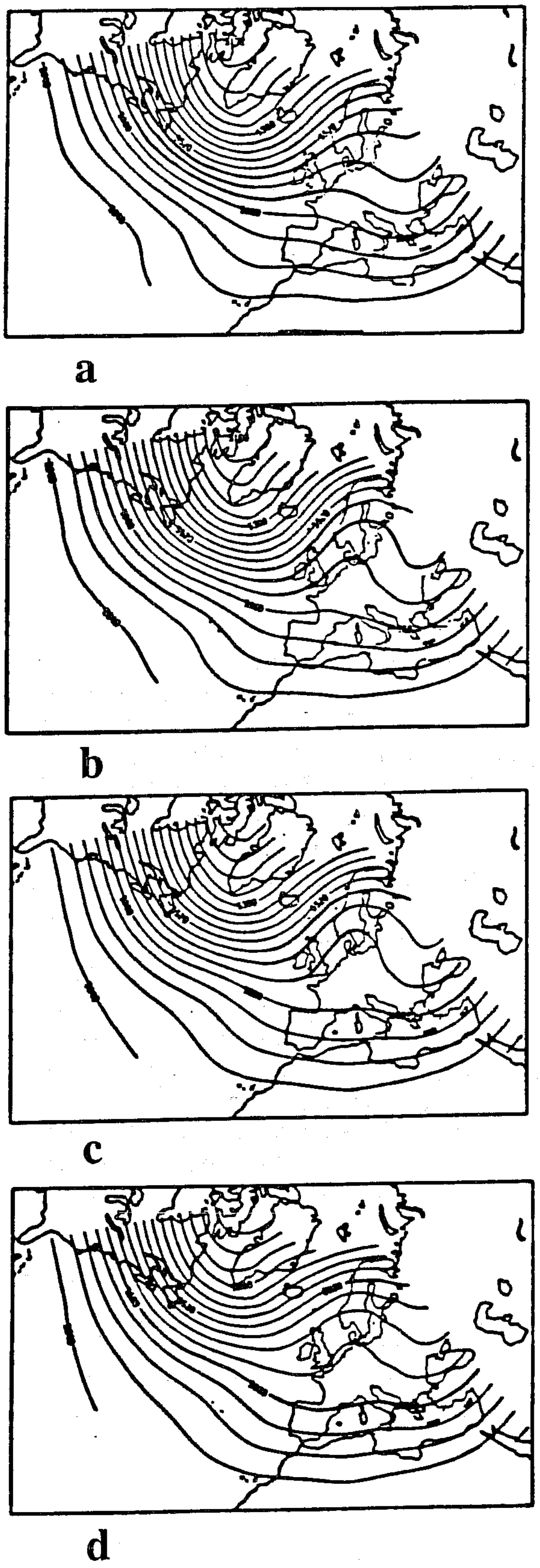


Fig. 19: Same as Fig. 17 but for the blocking anomaly (SP_2)

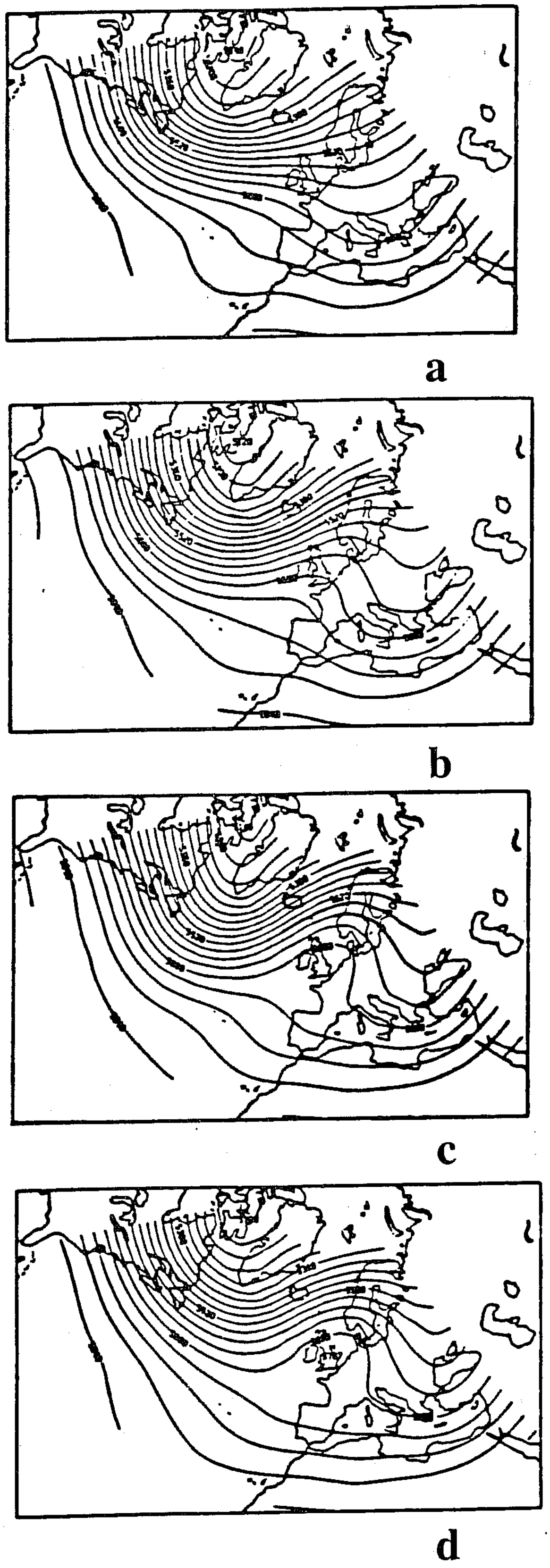
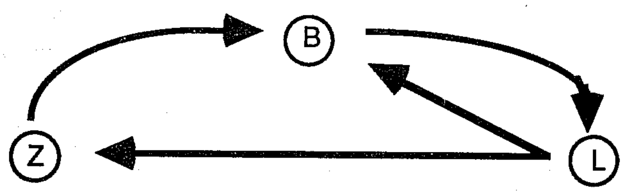


Fig. 20: Same as Fig. 17 but for the ninth weak pattern associated with the blocking anomaly



The preferred transition between zonal and blocking regimes does contradict the already mentioned difficulty in obtaining good forecast of blocking onsets (and as a subsidiary effect the results of our simple model). The present results are too preliminary to eliminate the possibility of an artefact and further study is required to clarify this point We however remark that the number of cases involved in this transition is only one third of the total number of blocking sequences and that the unpredictability of the other two thirds might give an average bad skill. On the other hand, the predictability of blocking may also be linked with the good representation of some smaller-scale precursors developing on a given large-scale flow, suggesting that the well-known deficiencies of baroclinic activity in forecast models (Klinker and Capalo, 1984) might be the source of the problem.

WP	1	2	3	4	6	7	8	9	10	j
1		1	2	0	4	4	0	1	5	
2	1		16	7	1	1	0	1	0	
3	4	0		9	1	5	0	0	1	
4	1	0	1		1	2	2	1	4	
6	0	1	1	1		5	1	6	8	
7	3	11	12	2	1		3	4	2	
8	1	12	12	6	3	6		3	0	
9	0	6	11	2	4	12	16		0	
10	0	7	4	1	2	13	11	13		
i										

T(i,j) = number of cases for which there exists at lest one day belonging to WP_j within the 7 days following a sequence of WP_i

Table 6: Temporal chaining of the weak patterns

Tab. 7a : at 7 days

SP	1	2	3	4
1(zonal)		13	3	0
2(block)	3		8	6
3(ridge)	1	4		1
4(Labrador)	7	6	0	

Tab. 7b : at 11 days

SP	1	2	3	4
1		15	5	3
2	8		9	11
3	4	7		4
4	13	10	2	

Total number of periods : SP1 : 35 SP2 : 42 SP3 : 29 SP4 : 37

T(i,j) = number of cases for which there exists at lest one day belonging to SP_j within the 7 or 11 days following a sequence of SP_i .

Table 7: Chaining of the strong patterns.

5. Conclusions

We have applied a method of cluster analysis to two different datasets. The first one was generated by a simple quasi-geostrophic channel which has been extensively studied in two other articles and here serves as a reference case. The second dataset is a 6-year series of the

500hPa geopotential over a limited area of North Hemisphere centered on the Northern Atlantic ocean. The analysis is based on the classification of individual daily maps and also of temporal series of daily maps. The latter procedure allows the capture of recurrent nonstationary events.

The analysis of model data recovers the regimes that were identified in the previous studies with different approaches. One interesting result is that the dynamical states are clustering in the vicinity of dynamical statistical equilibria of the large-scale flow. Another information is the detection of recurrent evolving clusters and of a temporal chaining between the various regimes. It is also possible to see in which respect the dynamics differs from a pure random behavior with the same covariance matrix and correlation at one-day lag.

The analysis of 500hPa geopotential data shows that about 40% of the atmospheric situations cluster in four classes with well differentiated characters. The most stable shows a high pressure center over Labrador and Greenland and a southward displacement of the jet which remains zonal over the Atlantic ocean and Western Europe. A second zonal regime is obtained, linked with a reinforcement of the jet downstream of its climatological maximum. Among the two other classes, one exhibits an oceanic ridge but is not found sufficiently stable; the other one groups blocking cases over Western Europe and separates into several subsets apparently corresponding to different stages. However the temporal information does not reveal as many striking features as for the model. A noticeable point to be confirmed by further studies is the existence of a preferred transition from zonal to blocking regime. It must be stressed that no low-pass filtering was applied to our data and that the atmospheric state is hardly characterized by the sole geopotential field, so that the limited present results can be considered as encouraging for the detection and characterization of weather regimes using more elaborate dynamical data. It is doubtless that a systematic analysis of the distribution of isentropic potential vorticity (Hoskins, McIntyre and Roberston, 1985) will contain much more dynamical information than our study.

A related work has been conducted independently by Mo and Ghil (1987). These authors study a 37-year records of the hemispherical 500hPa geopotential. Their classification of low-pass data recovers most of the known features of the large-scale couplings within Northern Hemisphere. It also supports our hypothesis of a regional rather than global basis for many persistent anomalies. They investigate the transitions between clusters and find some preferred paths, though there is less contrast than in our study, showing in addition that they can be associated with the classes of a band-pass analysis.

We believe that the present approach can lead to several useful developments. First it provides a convenient and effectual way to quantify the notion of weather regimes (without being limited to stationary events) and to measure the transitions between these regimes. This quantification, if enough precise, should be directly applicable to long-range forecast. We may also expect a relation between weather regimes and the short and medium range predictability (Legras and Ghil, 1985), the variations of which is a subject of current investigation (Palmer and Tibaldi, 1986). Secondly, it provides an intermediate tool between the classical statistical techniques (EOF, spectral analysis, teleconnection, ...) and the contemplation of plain atmospheric data. Its flexibility allows to address rather sophisticated problems of pattern recognition such as the identification of coherent structures within atmospheric flows (for instance one may attempt to classify the local relationship between potential vorticity and streamfunction (Haines and Marshall, 1987)). Thirdly, it allows a further step in the detailed description of climate dynamics which should be taken in account by modelers when comparing the simulated to the observed dynamics.

This optimistic view needs to be tempered by some warnings. Cluster analysis is not an objective method and the a priori arbitrarily chosen procedures or parameters may change the outcome. It is thus necessary to be cautious in trying to remove the arbitrariness by some physical arguments and in considering as established the sole results that are robust to variations in the analysis. The comparison of several methods is also desirable because of the difficulty encountered in the validation of a single method (Dubes and Jain, 1979). The one used here basically resolves the large amplitude anomalies, it does not say much about 60% of the atmospheric states that are not classified. Although this proportion might be reduced when more relevant dynamical quantities are considered, the simplicity and the usefulness of the approach is likely to be lost when dealing with the central part of the cloud. We then need to turn towards the equilibration methodss mentioned in the introduction.

Acknowledgments

We thank Michael Ghil for valuable discussions and communication of unpublished results. We were introduced to cluster analysis by G. Sèze who gave us numerous useful advices. The model data were provided by R. Vautard who has also been involved at several stages of this work. The atmospheric data were provided by the ECMWF. The computer resources were allocated by the Scientific Council of the Centre de Calcul Vectoriel pour la Recherche. Figures were plotted by the NCAR Graphics Software.

References

Diday, E. and J.-C. Simon, 1976: Clustering analysis, *in Commnunication and Cybernetics 10 Digital Pattern Recognition,* ed. K.S. Fu, Springer-Verlag, Berlin, pp. 47-94

Diday, E., 1972: Optimisation en classification automatique et reconnaissance des formes, *Revue Française d'Automatique, Informatique et Recherche Operationnelle (RAIRO)*, **3**, pp. 61-96

Dole, R.M. and N.D. Gordon, 1983: Persistent anomalies of the extratropical hemisphere wintertime circulation: geographical distribution and regional persistence characteristics, *Mon. Wea. Rev.*, **111**, pp. 1567-1586

Dubes, R. and A.K. Jain, 1979: Validity studies in clustering methodologies, *Pattern Recognition,* **11**, pp. 235-254

Gordon, A.D., 1981: Classification, *Monographs on Applied Probability and Statistics,* Chapman and Hall, London

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

Hoskins, B.J., Simmons, A.J. and D.G. Andrews, 1977: Energy dispersion in a barotropic atmosphere, *Quart. J. Roy. Met. Soc.*, **103**, pp. 553-567

Kalkstein, S.L., Tan, G. and J. A. Skindlov, 1987: An evaluation of three clustering procedures for use in synoptic climatological classification, *J. Cli. Appl. Meteor.*, **26**, pp. 717-730

Key, J. and R.G. Crane, 1986: A comparison of synoptic classification schemes based on "objective" procedures, *J. Climatol.*, **6**, pp. 375-

Klinker, E. and M. Capalo, 1984, Systematic errors in the baroclinic waves of the ECMWF model, *Technical Report N°. 41*, ECMWF, Reading, UK

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

Legras, B. and R. Vautard, 1987: Predictability and baroclinic flow regimes, *in ECMWF Workshop on Predictability in the Medium and Extended Range,* ECMWF, Reading, UK, pp. 183-204

Mo, K. and M. Ghil, 1987: Cluster analysis of multiple weather regimes, *submitted to J. Geophys. Res.*

Palmer T.N. and S. Tibaldi, 1986, Forecast skill and predictability, *Technical memorandum N° 127*, ECMWF, Reading, UK

Silverman, B.W., 1986: Density estimation for statistics and data analysis, *Monographs on Applied Probability and Statistics,* Chapman and Hall, London

Tibadi, S. and A. Buzzi, 1983: Effects of orography on Mediteranean cyclogenesis and its relationship to European blocking, *Tellus*, **35a**, pp. 269-286

Vautard, R. and B. Legras, 1987: On the source of mid-latitude low-frequency variability. Part I: Nonlinear equilibration of weather regimes, *submitted to J. Atmos. Sci.*

Vautard, R., Legras,B. and M. Déqué, 1987: On the source of mid-latitude low-frequency variability. Part I: A statistical approach to persistence, *submitted to J. Atmos. Sci.*

Yarnal, B. and D.A. White, 1987: Subjectivity in a computer-assisted synoptic climatology I: classification results, *J. Climatol.*, 7, pp. 119-128