Partitioning the North Atlantic into regions of similar seasonal sea-surface temperature anomalies.

W. C. Thacker*
Atlantic Oceanographic and Meteorological Laboratory
4301 Rickenbacker Causeway, Miami FL 33149 USA

R. Lewandowicz†
Cooperative Institute for Marine and Atmospheric Studies
4600 Rickenbacker Causeway, Miami FL 33149 USA

October 2, 1995

Abstract

Mean sea-surface temperatures were computed within 127 $6^\circ \times 4^\circ$ cells comprising most of the North Atlantic for 171 three-month seasons from 1950 through 1992, the mean seasonal cycle was removed, and cells with correlated seasonal anomalies were clustered into regions of coherent thermal behavior. Clustering algorithms consistently produced smaller thermal regions in the vicinity of the Gulf Stream, and while the regions were generally contiguous, a disjoint region was consistently found near the Grand Banks. Examining within-region variability as a function of the number of regions revealed no obvious "best" number of regions. The heterogeneity of the regions increase rapidly as the number of regions is decreased: for 26 regions, correlations between pairs of cells within a common region were typically 0.7; for 13 regions, a sizable fraction were less than 0.5, and for only 7 regions, within-region correlations were distributed fairly uniformly between 0.2 and 0.8.

* thacker@aoml.erl.gov
† lewandow@aoml.erl.gov

1 Introduction.

The need for better forecasts with lead times of a few months to a few seasons has motivated us to re-examine the formulation of statistical models for predicting near-term climatic anomalies. In doing so, we have found that existing models typically use a small number of principle components to characterize the climatic state. This is the case both for regression models (Davis, 1976; Graham *et al.*, 1987; von Storch and Xu, 1990; Barnston and Ropelewski, 1992; Penland and Magorian, 1993; Barnett *et al.*, 1993) and for analog models (Barnett and Preisendorfer, 1978; Livezey and Barnston, 1988; Barnston and Livezey, 1989; Livezey *et al.*, 1994). The use of principal components stems from the fact that monthly- or seasonal-mean climatic fields can be characterized by a relatively small number of statistically independent variables. While principal components have the advantage of being uncorrelated, they are intrinsically non-local variables characterizing global patterns of correlations, and as such they contribute to the difficulties in interpreting the statistical models in terms of regional effects.

A relatively unexplored alternative is to represent climatic fields by regional variables, which might be more easily related to physical processes. Rotated principal components (Drosdowsky, 1993; Gong and Richman, 1995), which are more regional than the usual unrotated principal components, have been used as the variables of analog models (Livezey and Barnston, 1988). Because they are constructed as linear combinations of the unrotated components chosen to characterize the field, they provide exactly the same information as the unrotated components. Several algorithms can be used to determine rotated components (Richman, 1986), all of which are based on the idea that each point of the field should be explained by as few rotated components as possible. Because the emphasis is placed on loadings rather than on scores (Johnson and Wichern, 1992), the rotation algorithms do not actually seek regions having coherent behavior.

The objective of this paper is to explore the use of cluster analysis for defining regions within which a field's average value can serve as a regional variable. In particular, we want to partition the North Atlantic into regions within which sea-surface temperatures are highly correlated. Immediately we face the question of how many regions to use. Within the context of statistical modelling, choosing the number of regions is similar to the problem of deciding how many principal components can be estimated from the available data (von Storch and Hannoschöck, 1985). Before using regional variables for climatic prediction, this issue should be examined carefully. However, because it is outside the scope of cluster analysis per se, we plan to discuss it in another paper. Here, we focus on the problem of identifying regions based on observed variability and explore the relationship between the number of regions and the thermal variability within the regions.

Cluster analysis attempts to partition entities, separating those that are different from those that are similar. It is often implemented in terms of a distance between entities that are to be clustered (Anderberg, 1973; Everitt, 1974; Johnson and Wichern, 1992). Here, the entities are grid cells within which seasonal anomalies have been defined, and the clusters are regions formed as groups of cells. Distance is not a physical distance measured in kilometers but an abstract distance characterizing degree of dissimilarity. Because we focus on the North Atlantic sea-surface temperatures and want to determine regions within which seasonal anomalies can be expected to be similar at any given time, distance between cells should reflect the degree to which their anomalies are correlated. The basic idea is that entities within the same cluster should be close to each other. A variety of methods exist for determining clusters from specified distances between entities. The resulting regionalization will depend on which method is used and how distance is defined.

Regionalizations determined by different methods can be compared and judged by several of criteria. First, it is desirable that the clusters form contiguous regions; otherwise, the goal of easy interpretability is not achieved. Second, the variability within the regions should be small when compared with the variability from region to region. For some definitions of distance, a region's centroid can be defined as minimizing the sum of squared distances to each of the region's cells. The within-region sum-of-squares, which is analogous to ANOVA error-sum-of-squares, provides a quantitative measure of within-region variability; the smaller the total within-region sum-of-squares, the better the regionalization. Third, the minimum within-region correlation indicates the degree to which the cells within the region can behave differently; the larger the minimum correlations, the better the regionalization.

This is not the first use of cluster analysis to identify oceanic regions on the basis of surface-temperature correlations; Iwasaka *et al.* (1988) partitioned the North Pacific, and Wolter (1987), the tropics. In a somewhat different application (You and Tomczak, 1993) have used cluster analysis based on hydrographic data to identify water masses. The identification of precipitation regions (Drosdowsky, 1993; Gong and Richman, 1995) is similar in spirit to our study, whereas the determination of agricultural districts or climatic zones based on several climatic fields (Stooksbury and Michaels, 1991; Fovell and Fovell, 1993) is analogous to the identification of water masses. A quite different application of cluster analysis is the identification of periods during which the weather is similar (Mo and Ghil, 1988; Cheng and Wallace, 1993); that approach might be useful for selecting analogs on which short-term predictions might be based (Barnett and Preisendorfer, 1978; Livezey and Barnston, 1988). Gong and Richman (1995) have provided a lengthy list of references to other studies in which cluster analysis has been used.

In section 2, we describe how we define time series of seasonal anomalies of mean observed sea-surface temperatures within $6°\times 4°$ cells, which are to be clustered into regions exhibiting similar thermal behavior, we examine the within-cell variability, and we present heuristic estimates of the number of regions to expect. In section 3, we review the basics of cluster analysis, presenting several correlation-based distances and algorithms that we have tried. The results presented in section 4 are restricted to a single distance and to three clustering algorithms. The relationship between within-region variability and the number of regions is examined for hierarchies of partitions computed using the complete-link and minimum-variance methods, and 26-, 13-, and 7-region partitions are examined in detail. The best regionalizations are shown to be produced by the K-means method starting from the minimum-variance centroids. We conclude in section 5 by summarizing our findings and their implications for statistical modelling.

2 Data.

The sea-surface temperature field plays an important role in seasonal-to-interannual climatic dynamics and has been chosen as a predictor field in the statistical models cited in the introduction. For this reason, we have chosen to explore the possibility of using regional averages to characterize the sea-surface temperature field with the presumption that similar results might be attained for other climatic fields.

The COADS (Comprehensive Ocean-Atmosphere Data Set) summaries, which are based on ship's reports, provide sea-surface temperature data in the form of monthly-mean values for 2° longitude by 2° latitude cells (Woodruff *et al.*, 1987). We would have liked to use these data to partition the global ocean, but the coverage being uniform neither in space nor in time presented a major difficulty. Figure 1 shows the cells having at least one observation for every month between January 1950 and December 1992. Clearly, many cells lack continuous coverage. Furthermore, even if the coverage is continuous, mean temperatures for some months might be based on only one or two observations. Certainly, better estimates might be obtained by exploiting expected spatial and temporal correlations with temperatures in adjacent cells and in adjacent months (Gandin, 1963; Bretherton *et al.*, 1976; Mariano and Brown, 1992). Similarly, the unmarked cells might have data for some months, which would be ignored if continuous coverage is demanded. For example, some cell might have an observation for all but one month, with many observations for the other 515 months. Again, statistical interpolation might be used to fill in the gap. However, the task of providing optimal estimates of the temperature field with the greatest reasonable coverage is a major undertaking that is beyond the scope of this study.

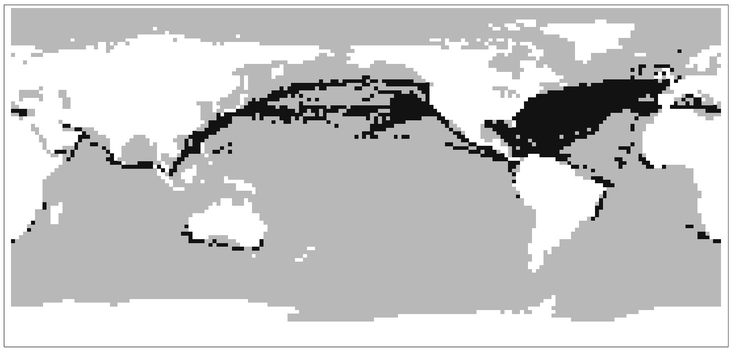

Figure 1: Black indicates $2^\circ \times 2^\circ$ COADS cells having at least one SST observation for every month between January, 1950, and December, 1992.

We chose to restrict our attention to the relatively data-rich North Atlantic and to focus on the better sampled spatial and temporal scales of variability by consolidating the data into larger space-time cells. Iwasaka *et al.* (1988) and Wolter (1987) based their cluster analyses on $5^\circ \times 5^\circ$ means, and we thought that cells of a similar size would be appropriate here, especially considering the greater density for observations in the North Atlantic than for the tropics or the North Pacific. Constrained by the $2^\circ \times 2^\circ$ size of the COADS cells and motivated by the the fact that zonal thermal similarities are generally greater than meridional, we decided to consolidate the data into 6° longitude by 4° latitude cells. Because the variables used for the statistical forecast models cited in the introduction are based on three-month seasonal means, we further consolidated the data into seasons. More specifically, each $6°\times 4°$ seasonal mean has been constructed as an average of 18 COADS $2°\times 2°$ monthly means weighted in proportion to the number of observations on which it was based, so that the seasonal means used here would be COADS-like averages of surface-temperature observations for larger space-time cells. The seasons have been arbitrarily selected as December-January-February, March-April-May, June-July-August, and September-October-November. The 43-year record consists of 171 seasonal means for each cell, from spring 1950 through autumn 1992. We restricted our attention to the 127 $6°\times 4°$ cells for which all 171 seasonal means were based on at least ten observations. The locations of these 127 cells are shown in figure 2. The northern boundary is 60°N; the southern, 4°N; the eastern boundary (only a single cell extends this far east) is 2°E; and the western, 82°W.

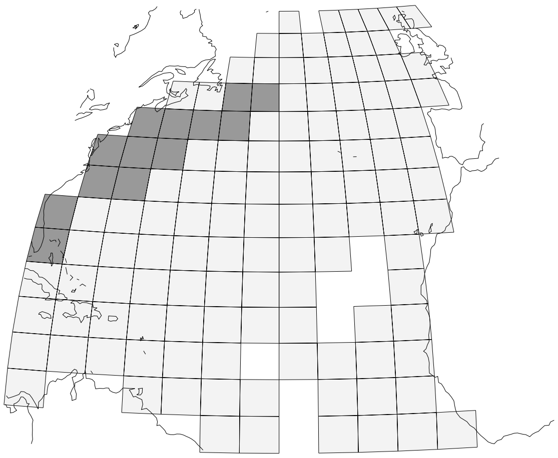

Figure 2: Map of the North Atlantic showing the locations of the 127 $6°\times 4°$ cells for which mean sea-surface temperatures are defined for 171 seasons (spring 1950 through autumn 1992). Darker shading indicates the 13 cells within which correlations between $2°\times 2°$ seasonal anomalies were found to be less than 0.5.

The cumulative distribution in figure 3 indicates that the mean sea-surface temperature for most of the $127 \times 171 = 21,717$ space-time cells is determined from at least 100 observations. Ninety-nine per cent of the cells have at least 40 observations; 95% have at least 90. The best sampled space-time cell has 12,246 observations; to show that cell and all others having more than 3000 observations (almost 2%) would require either that the width of the figure be increased roughly by a factor of four or that the more interesting part of the distribution be compressed to one-quarter its width. The more worrisome part of the distribution are the cells having the smallest number of observations; because their means are less well determined, they have the greatest impact on the sample correlations. As there are only 36 cells with fewer than 20 observations and 100 with fewer than 30, this impact was not thought to be serious.

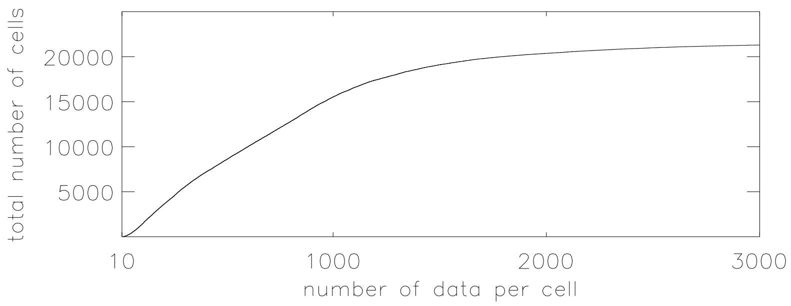

Figure 3: Number of space-time cells (figure 2) for which the seasonal means are determined from less than a specified number of observations. Three-quarters of the means are based on at least 299 observations; half are based on at least 662; and one-quarter on at least 1074. The mode is 111 observations, and the mean is 814.

An alternative strategy that we considered was to base the clustering on $2^\circ \times 2^\circ$ means, using only those cells having sufficient data. A cell's mean should be based on enough, *e.g.*, at least ten, observations to be reasonably well determined, and correlations between pairs of cells could be computed by skipping times when one or the other was undersampled. Demanding that acceptable correlations be based on a large fraction, *e.g.*, at least 90% of the 43-year interval would eliminate a fraction of the cells from the analysis, but not as many as would requiring all time series to be continuous. This alternative was rejected, because we anticipated that some of the acceptable cells would be isolated from the core group and that might have an adverse effect on the clustering. The chosen strategy is a simpler way of guaranteeing that almost all means are based on a sufficient number of observations for the entire 43-year interval.

We have considered two options for treating the annual cycle that is manifest in the data: (1) to seek different regionalizations for each of the four seasons; or (2) after removing the seasonal cycle, to seek a single regionalization for all seasons. Because the regions would be based on sample correlations, we decided on the second option. The first would have required splitting each time series into four, one for each season, basing the correlations on only one quarter as many data. This decision was further supported by our judgment that statistical models formulated in terms of the same regional variables for each season would be simpler and thus easier to comprehend in terms of physical processes. Temperature anomalies were computed for each cell by first estimating mean values for each season over the 43 years and then subtracting these means from the corresponding seasonal values for each year. Our goal became grouping the $6°\times 4°$ cells into regions, such that temperature anomalies are highly correlated for cells within the same region and weakly correlated for cells in different regions.

The histogram and reverse cumulative distribution in figure 4 show that, while the anomalous temperatures of most of the 8001 pairs of cells are essentially uncorrelated, approximately 100 pairs have correlation coefficients as high as 0.8 and 300 pairs as high as 0.7. These higher values should correspond to pairs of cells within the same region, once regions have been identified; similarly, the smaller values should be for cells from different regions. The reverse cumulative distribution can be used to estimate how many regions might be appropriate. For example, with five cells (10 pairs) per region, there would be $127\div5=25$ regions and $25\times10=250$ highly correlated pairs. Because these would be the pairs comprising the tail of the histogram, it is possible to estimate the minimum expected within-region correlation coefficient. For this example, we might expect within-region correlations to be generally greater than 0.7, so 25 regions might describe the temperature-anomaly field fairly well. Similarly, 10 cells (45 pairs) per region would give 13 regions and 585 pairs, a sizable fraction having expected within-region correlations less than 0.6, and 18 cells (153 pairs) per region would give 7 regions and 1071 pairs (roughly 1/8 of the 8001 pairs comprising the histogram), some having expected within-region correlations less than 0.5.

How much within-region variability might be reasonable can be estimated by examining the correlations between pairs of seasonal-anomaly time series for $2°\times 2°$ cells within the larger $6°\times 4°$ cells. Figure 5 shows a histogram of

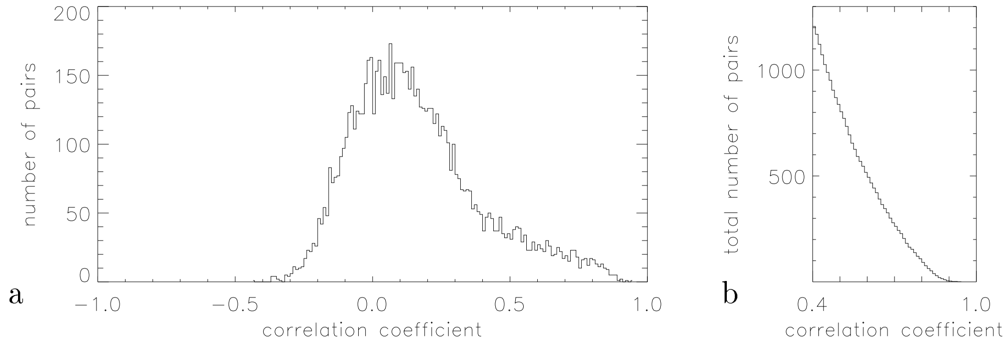

Figure 4: (a) Number of $6° \times 4°$ cell pairs for which correlation coefficients are within intervals of width 0.01. (b) Number of cell pairs for which correlation coefficients are greater than value indicated on horizontal axis.

within-cell correlation coefficients; as in figure 4 bin widths are 0.01. In computing these correlation coefficients, only $2° \times 2°$ seasonal means based on at least thirty observations were used. For only 513 pairs of the $15 \times 127 = 1950$ was it possible to compute correlation coefficients with an unbroken series of 171 seasons. Skipping up to 6 seasons increased this number to 876; these are the pairs that comprize this histogram. While most correlation coefficients are greater than 0.7, a substantial number are less than 0.5. The 13 cells within which the smaller correlations occur are situated in the vicinity of the Gulf Stream; they are indicated by the darker shading in figure 2. Perhaps smaller cells would have been more appropriate for this relatively well-sampled region, but we chose to continue with the larger cells everywhere. On the basis of these within-cell correlations we concluded that reasonable values for within-region correlations would be greater than 0.7, but values as low as 0.6 might be acceptable. Therefore, we estimated that we might need over a dozen clusters to characterize the North Atlantic's anomalous thermal variability.

3 Clustering algorithms.

While the data suggest that regional variables might be used to characterize the sea-surface temperature-anomaly field, the task of identifying the regions remains. From the wide variety of clustering algorithms that might be used for this chore, we have limited our choice to those algorithms implemented

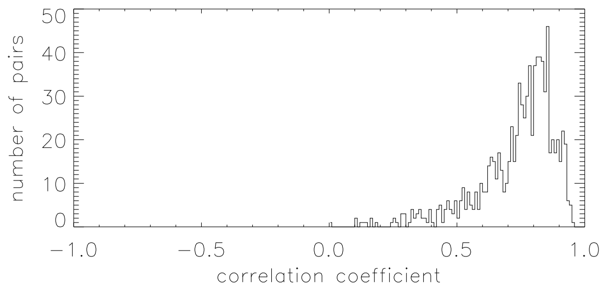

Figure 5: Number of $2°\times 2°$ cell pairs within common $6°\times 4°$ cells for which seasonal-anomaly correlation coefficients are within bins of width 0.01.

in the NAG library (NAG, 1993), which are roughly the same choices offered by IMSL (1989), STATISTICA (1994), or SPSS (1988). A larger number of algorithms is offered by SAS (1990), but these are still not all that might be considered. For example, none of these support fuzzy clustering algorithms like those discussed by Bezdek (1981). As exploring all possibilities is well beyond the scope of this study, we felt that the algorithms offered by NAG were more than sufficient. Indeed, as we shall discuss below, we found that other algorithms were unlikely to produce appreciably better regionalizations.

NAG supports two different approaches to clustering. The first constructs clusters by agglomeration, producing a hierarchy of regionalizations ranging from one extreme with each of the 127 cells being different regions to the other extreme with all cells comprising a single region. The second approach determines a specified number of clusters by iterative improvements.

Both approaches use distances between elements to determine which elements should belong to which clusters; here, the elements are the 127 time series of seasonal sea-surface temperature anomalies and the clusters are the regions. Our goal is to identify regions within which the thermal behavior is coherent, so that a single variable might serve to characterize the region. Clearly, the distance between two cells should be small when their temperature anomalies are highly correlated. Less clear is whether the distance should depend on the amplitudes of the anomalies, *i.e.*, on their standard deviations. Because, for predictive purposes, standard deviations simply set the scale of the variability, while the correlations relate the behavior, we decided that the amplitudes of the anomalies should not enter the definition of the distance, even though they might ultimately enter the definition of the variable characterizing the region. Thus, the clustering distance between two cells should be a function of the correlation between the seasonal temperature anomalies of the two cells. Several reasonable definitions readily come to mind.

3.1 Definitions of distance.

As correlation coefficients can be regarded as the cosine of the angle between two unit vectors (Johnson and Wichern, 1992; Thacker, 1995b), the distance between two cells can be visualized[1] as the distance between two points on a unit circle. Thus, one measure of distance between cells p and q is the arc length between the points,

$$d_1(p, q) = \cos^{-1}\left[\mathrm{cor}(T_p, T_q)\right] \ , \tag{1}$$

where $\mathrm{cor}(T_p, T_q)$ is the estimate of the seasonal-anomalies correlation coefficient based on the data. Note that this distance function ranges between zero and π, being less than $\pi/2$ for positive correlations and greater for negative; it is zero only for perfect correlations. The triangle inequality can be visualized by imagining three points p q and r on a sphere: $d_1(p, q) \leq d_1(p, r) + d_1(r, q)$.

The distance could also be taken as the length of the straight line (chord) joining the two points,

$$d_2(p, q) = \sqrt{2(1 - \mathrm{cor}(T_p, T_q))} \ . \tag{2}$$

Because this distance is Euclidean, it is compatible with NAG's algorithms that define clusters around centroids. The centroid of a region minimizes the sum of the squares of the distances to all cells in the region; for d_2, it is the average of the standardized time series of seasonal temperature anomalies for the cells comprising the region. Determining regions having small within-region sum-of-squares is an objective of centroid-based algorithms.

Because negative correlations are as useful as positive for making predictions, it might be useful to define distance so that it is independent of the sign of the correlation. For example, d_1 can be modified to give the definition,

$$d_3(p, q) = \cos^{-1}\left|\mathrm{cor}(T_p, T_q)\right| = \min(d_1, \pi - d_1) \ . \tag{3}$$

[1]The 127 cells are points on a hypersphere in a space of dimensionality equal to the rank of the correlation matrix.

This distance can be easily visualized in terms of the unit circle. Because the sign of the temperature anomaly is unimportant, each cell corresponds to a pair of diametrically opposed points; distance between cells is the arc length of the acute angle formed by two diameters. This definition would treat teleconnection points (Wallace and Gutzler, 1981) as being similar and thus could result in disjoint regions. However, because of the absence of strong negative correlations between $6°\times 4°$ seasonal anomalies (see figure 4), negative correlations did not lead to disjoint regions in this study.

Similarly, d_2 can be modified to be insensitive to the sign of the correlation,

$$d_4(p,q) = \sqrt{2(1 - |\mathrm{cor}(T_p, T_q)|)}\ . \tag{4}$$

The geometrical meaning is clear: this distance corresponds to the length of the chord of the arc of d_3.

As correlations are cosines and measure similarity, it is reasonable to use sines to measure dissimilarity,

$$d_5(p,q) = \sin d_3(p,q) = \sqrt{1 - \mathrm{cor}^2(T_p, T_q)}\ . \tag{5}$$

The point with the smallest total square d_5 distance to the points comprising a region, is the time series for which the sum of the squares of the correlations with the cell temperatures is maximum; in other words, the centroid is the first principal component of the standardized variables (Thacker, 1995a). Thus, d_5 could be used for centroid-based clustering. However, NAG's algorithms were not implemented with this possibility in mind.

We examined the impact of these choices for distance on the NAG's clustering algorithms ability to determine regions of coherent anomalous thermal behavior, not only for these 127 North Atlantic $6°\times 4°$ cells but also for other basins and for other cell sizes. In some cases what seemed to be minor differences early in the agglomerative process influenced all later stages, but these differences were generally less than those associated with the choice of clustering algorithm. Presenting details of the differences between results stemming from the choice of definition of distance is beyond the scope of this paper. Our basic conclusion was that there was no compelling reason to consider any other choice than d_2, which was appropriate for all algorithms. Thus, all results presented below are for d_2.

3.2 Hierarchical methods.

NAG's subroutine G03ECF implements the hierarchical approach. It starts with each cell taken to be a separate region and proceeds in a step-wise fashion, merging the two most similar regions at each step. Thus, how the North Atlantic will be partitioned into k regions depends on how it has been partitioned into $k+1$ regions at the previous step. This approach has proved useful for taxonomic purposes in biology and linguistics, and we thought it might reveal whether there might be an obvious "best" number of regions into which the cells cluster.

Subroutine G03ECF requires as input a table of "distances" between all pairs of one-cell clusters. The quotation marks are used to emphasize that, for some methods, this table should actually provide squares of distances. The smallest entry in the table determines the pair to merge into the first two-pair cluster. The table is then updated; the two cell's distances are removed and distances from the two-cell cluster to all remaining one-cell clusters are added. At this stage it is necessary to know how to update the table to include the clusters containing more than one element. The subroutine provides six choices for defining distances between multi-element clusters: (1) single link; (2) complete link; (3) group average; (4) centroid; (5) median; and (6) minimum variance. For each, the new entries in the distance table are computed from values contained in the un-updated table. After the update the procedure is repeated; two more clusters are merged and the table is again updated. This process continues until there are no more clusters to merge.

The first three methods are based directly on the between-element distances. The single-link method specifies the distance between two clusters to be the shortest distance between pairs of elements, one from each cluster. If jth and kth clusters are merged, then the single-link update formula for the distance between the new cluster $\{j,k\}$ and remaining cluster i is

$$D_{i,\{j,k\}} = \min(D_{i,j}, D_{i,k}) \ . \tag{6}$$

Because method does not rule out the possibility of a cluster containing pairs of elements being far apart, it was not expected to produce the sort of regionalization we want. The complete-link method, which specifies the distance between two clusters to be the greatest distance between pairs of elements, is better suited for identifying regions that are internally homogeneous. The complete-link update formula is

$$D_{i,\{j,k\}} = \max(D_{i,j}, D_{i,k}) \ . \tag{7}$$

The group-average method, which is sometimes called the average-link method, computes the new distance as the average of the distances between pairs of cells, one from each cluster. The distance table update

$$D_{i,\{j,k\}} = \frac{n_j}{n_j + n_k} D_{i,j} + \frac{n_k}{n_j + n_k} D_{i,k} \tag{8}$$

is a weighted average of the old distances with weights proportional to the number of cells in each cluster.

The remaining three methods are predicated on the concept of a cluster centroid and are implemented for the Euclidean distance d_2. Although consistency is not demanded, they assume that the input table provides the square of the between-entity Euclidean distances. The centroid method computes the distance between clusters as the distance between their centroids. It updates the table with the formula,

$$D_{i,\{j,k\}} = \frac{n_j}{n_j + n_k} D_{i,j} + \frac{n_k}{n_j + n_k} D_{i,k)} - \frac{n_j n_k}{(n_j + n_k)^2} D_{i,k} \ . \tag{9}$$

If it is used with distance functions other than d_2, it cannot be interpreted as joining clusters having the closest centroids. The median method is similar to the centroid method,

$$D_{i,\{j,k\}} = \frac{1}{2} D_{i,j} + \frac{1}{2} D_{i,k} - \frac{1}{4} D_{i,k} \ , \tag{10}$$

but the weights with which the distances are combined are independent of the number of cells in the clusters. The centroid and median methods are both non-monotonic, *i.e.*, "distance" between merged clusters does not necessarily increase as the number of clusters decreases; in fact, NAG's subroutine G03ECF terminates when the "distance" associated with the merger of two clusters is less than that for the previous merger. We found that to be the case for our data when using the centroid and median methods, so both of these methods are excluded from further discussion. The third centroid-based method, *i.e.*, Ward's minimum-variance method, is like ANOVA for clusters: the clusters correspond to the model, and the sum of the squares of the distances from the from the clusters' centroids to each of their members is like error-sum-of-squares. Total variance, *i.e.*, the sum of the squares of the distances from the centroid of all elements, can be partitioned into within-cluster variance and between-cluster variance. At each step, the two clusters that are joined give the least increase in within-cluster variance. The update formula,

$$D_{i,\{j,k\}} = \frac{n_i + n_j}{n_i + n_j + n_k} D_{i,j} + \frac{n_i + n_k}{n_i + n_j + n_k} D_{i,k} - \frac{n_j}{n_i + n_j + n_k} D_{j,k} \; , \quad (11)$$

requires the distance table to be initialized to provide the *square* of distance between cells; otherwise, the resulting clusters will not be minimum-variance clusters.

Of these six hierarchical possibilities provided by NAG, the complete-link method and Ward's method appeared the most promising, as they attempt to maintain high within-cluster similarity. Indeed, this proved to be true for our data. The single-link method was clearly inappropriate, preferring to form a few large incoherent clusters.

3.3 K-means method.

The K-means method is the only non-hierarchial method implemented in the NAG library. The name derives from the fact that it seeks a specified number of clusters K, which are formed around their centroids (means). Because it minimizes total within-cluster sum of squared distances, it can be regarded as a non-hierarchial counterpart to the minimum-variance method.

Subroutine G03EFF, which implements the K-means method, does not allow for choice of how distances are measured. They are required to be Euclidean, *i.e.*, d_2 is built-in. Rather than working with a distance table, the subroutine computes them as needed from the time series of temperature anomalies.

Seed values for the K centroids must be supplied. For example, they might be the centroids of clusters obtained using one of the hierarchical methods, but they could be any K time series of the appropriate length. Candidate regions are formed by associating each cell with the closest candidate centroid. Then new centroids are computed and the process is repeated until convergence.

Even though the total within-cluster sum of squared distances is reduced at each step, because different seed points can result in a different clusters, there is no guarantee that the resulting clusters have the smallest *possible* total error variance. The problem is like that of finding the global minimum of a continuous function, but here we are dealing with a finite, but combinatorially large, number of values. Our objective is not to find the regionalization having the smallest possible within-region variation, so an exhaustive search for the best possible K-means solution is not warranted.

4 Results.

Our first objective was to determine whether a hierarchical algorithm might indicate an optimal number of thermal regions into which the North Atlantic might be divided. Our approach was to examine the hierarchies of regions produced and see whether there is a noticeable increase in within-cluster variability as regions are merged to form larger regions. If so, the number of clusters at the previous stage would be the optimal number as indicated by that algorithm.

The minimum-variance and K-means methods provide a useful measure of within-cluster variability, namely within-region sum-of-squares. The total within-region sum-of-squares can be used to compare different partitionings of the North Atlantic; the smaller the total, the more coherent the average region. Figure 6 shows the total within-region sum-of-squares, normalized by the sum-of-squares for a single all-encompassing region, as a function of the number of regions for hierarchies produced by Ward's and complete-link methods. For both methods, it can be seen to increase smoothly as the number of regions is reduced; at no point is the increase in within-region variability noticeably greater than for the previous step. However, for fewer than 20 regions, the within-region variability is comparable to the between-region variability. It is interesting to note how similar the curves are for the two methods. In fact, if they were drawn on the same plot, one would lie almost exactly on top of the other, differences being roughly the width of the line, with Ward's method showing slightly less within-region variability.

The minimum correlation coefficient of seasonal sea-surface temperature anomalies for pairs of cells within the same region is an indicator of how incoherent the thermal behavior of the region might be. The value for the least coherent region can be used as another indicator of how well a clustering method performs. Figure 7 shows the overall minimum within-region correlation at each stage of the hierarchy for Ward's minimum-variance method and for the complete-link method. By this measure the complete-link performs somewhat better than Ward's method, but that should be no surprise. Just as Ward's method was designed to reduce the within-cluster sum-of squares, the

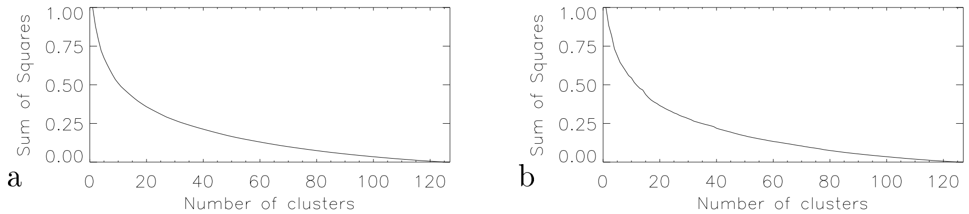

Figure 6: Normalized total within-region sum-of-squares as a function of number of clusters for (a) Ward's method and (b) complete-link method.

complete-link method seeks to minimize maximum cell-pair distances, *i.e.*, to maximize minimum correlation coefficients. Still, the results are similar for the two methods. If a region is to be considered homogeneous only when the minimum correlation coefficient is no smaller than 0.7, roughly 70 regions would be required. On the other hand, if less homogeneity is demanded and correlations of 0.5 are considered to be acceptable, then approximately 40 regions would be needed. Any partition involving a dozen or fewer regions is likely to have appreciable within-region variability as measured by this worst-case indicator.

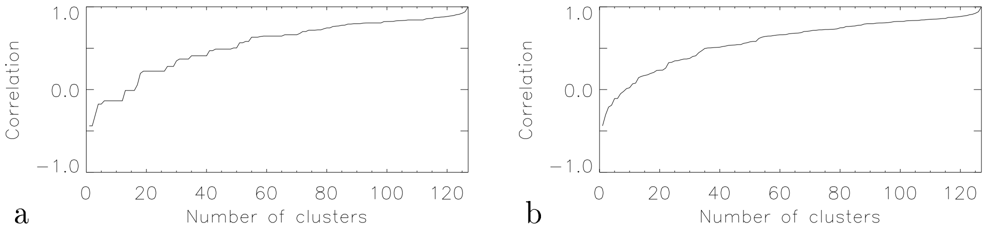

Figure 7: Minimum within-region correlation coefficient as a function of number of clusters for (a) Ward's method and (b) complete-link method.

The dendrograms shown in figure 8 provide another way to visualize the hierarchies of regions produced by the complete-link method and Ward's minimum-variance method. Complete dendrograms would start with 127 single-cell regions at the bottom, but these show only the final stages of the clustering where there are 26 or fewer regions. The number 26 was chosen arbitrarily so that the clusters could be denoted by a single letter of the alphabet. The naming convention is such that Z is merged with some other region, followed by Y, then X, *etc.* At each stage of the hierarchy, the newly formed region can take the name of the other member of the merged pair.

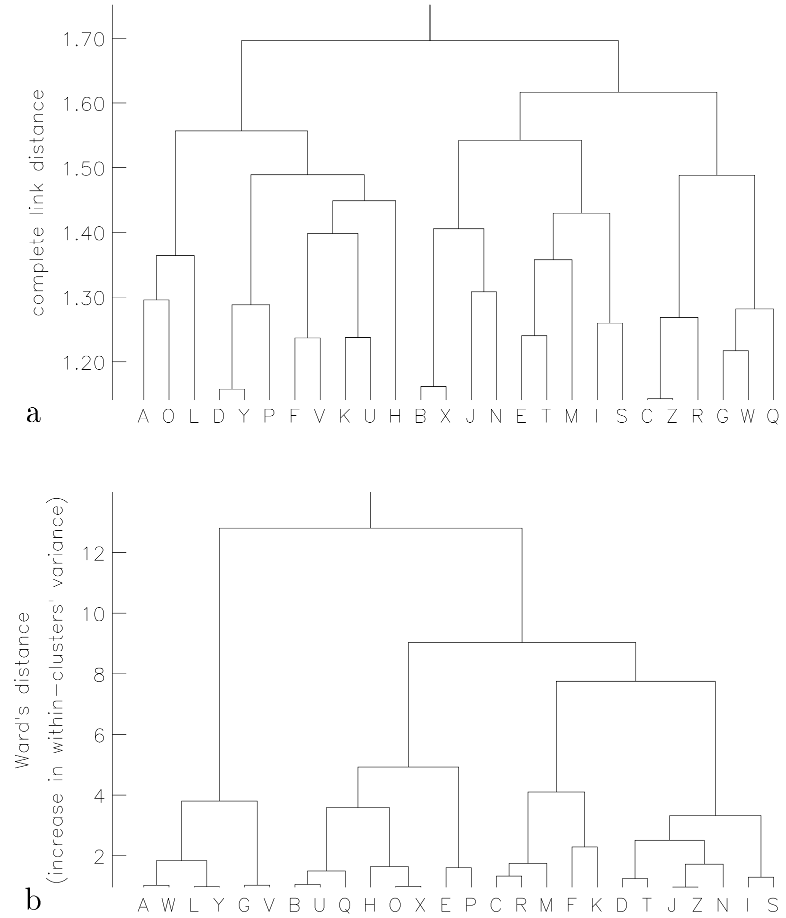

Figure 8: Dendrograms for (a) the complete-link method and (b) Ward's minimum-variance method, both using Euclidean distance d_2 to measure differences in thermal behavior. Only the portions of the dendrograms corresponding to 26 or fewer clusters are shown. The vertical scale for (a) indicates the distance between the most widely separated pairs in the clusters that are merged and for (b) indicates the increase in within-clusters sum-of-squares when two cells are merged.

Thus, in the complete-link hierarchy, region Z merges with C to give 25 regions with the merged region designated as C, and in the minimum-variance hierarchy, Z merges with J to form a larger J. This naming convention allows maps corresponding to different levels of a hierarchy to be compared. (See figures (9), (11), and (13)). Note that the vertical scales for the two dendrograms are different. For the complete-link hierarchy, the vertical scale shows the d_2 distance between the most dissimilar time series of the merged clusters; for the Ward's hierarchy, it shows the increase in within-region sum-of-squares caused by the merger. In either case the number of clusters at any stage can be determined by holding a straight edge horizontally and counting the number of vertical lines it intersects. For example, only for 9 or fewer complete-link clusters will there be any negative within-cluster correlation coefficients, *i.e.*, within-cluster distances greater than $\sqrt{2}$.

From the plots of minimum within-region correlation coefficients and sum-of-squares ratios and from the dendrograms, we see that the observed variability of the North Atlantic's anomalous sea-surface temperatures is not so dramatically regional that it suggests an obvious number of well-defined areas. We can also conclude that, when the North Atlantic is partitioned into a relatively small number of regions, the thermal variability within some regions is likely to be large. To learn more, it is necessary to examine the regionalizations in greater detail. Below we present examples of 26-, 13-, and 7-region partitions of the North Atlantic produced by clustering algorithms. In every case Euclidean between-cell distances d_2 are used.

4.1 Twenty-six regions.

Figure 9 shows six different partitionings of the North Atlantic into 26 regions: panels (a), (c), and (e) show the 26-cluster results for the single-link, complete-link, and minimum-variance methods, respectively, and panels (b), (d), and (f) show the results of using (a), (c), and (e) as starting points for the K-means method. The characteristics of these regionalizations are summarized in table 1 and in figure 10.

On the basis of the within-region sum-of-squares, five of the six maps are roughly equivalent. Figure 6 shows that this indicator has similar values for the complete-link and minimum variance methods; totals from table 1 for these methods are 33.05 and 31.89, respectively. The K-means algorithm improves these numbers only slightly to 31.86 and 31.09. We were expecting the single-link method's poor performance, but we were surprised to find that

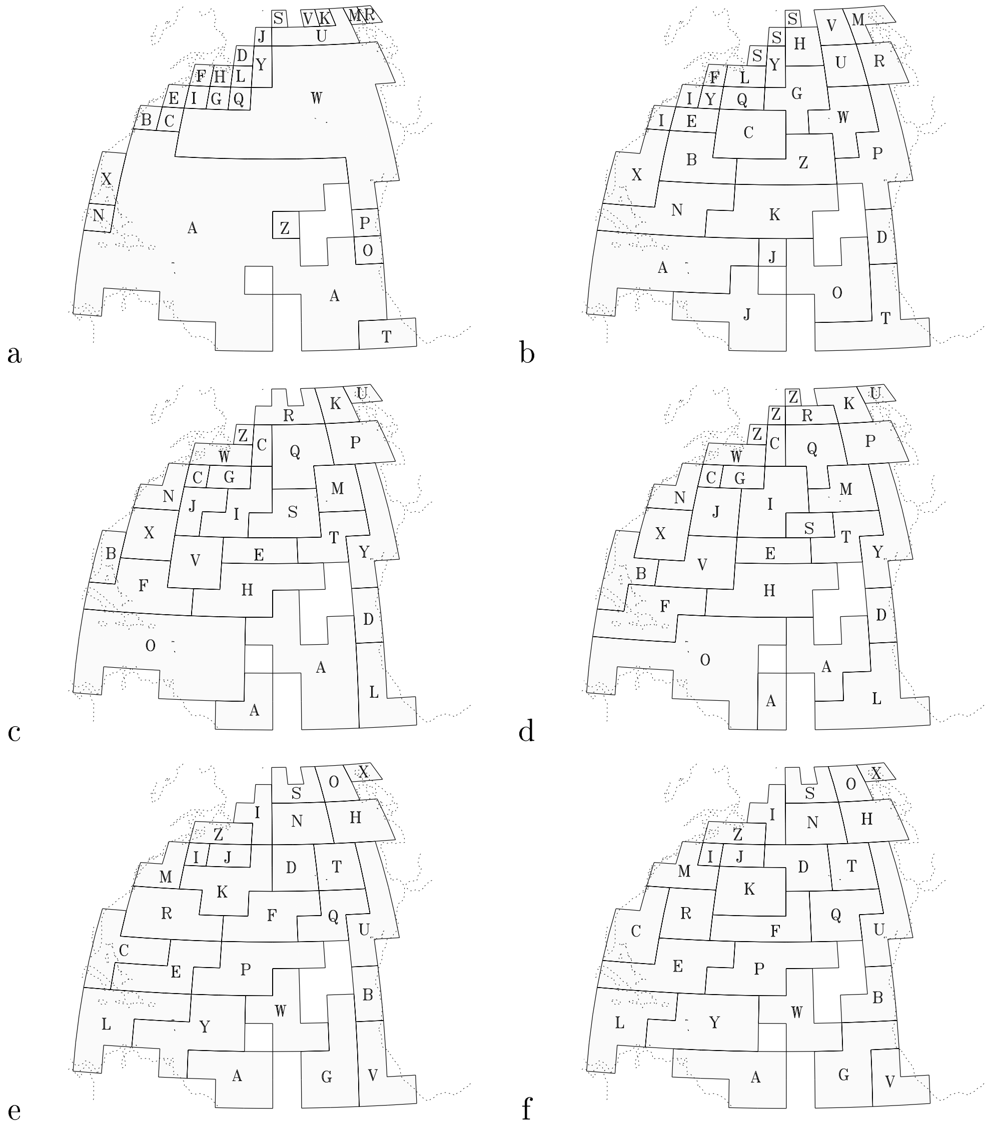

Figure 9: Partitions of the North Atlantic into 26 regions of similar anomalous seasonal SST variability produced using: (a) single-link method; (b) single-link followed by K-means; (c) complete-link; (d) complete-link followed by K-means; (e) minimum-variance; (f) minimum-variance followed by K-means. Land masses are indicated by dotted lines.

	S			S/K			C			C/K			W			W/K		
	min	SS	#	min	SS	#	min	SS	#	min	SS	#	min	SS	#	min	SS	#
A	-.29	38.57	56	.53	2.82	12	.35	5.01	14	.45	2.28	8	.63	0.92	5	.54	1.31	6
B	*	*	1	.38	1.53	5	.73	0.27	2	.28	1.31	4	.66	0.34	2	.65	0.70	3
C	*	*	1	.58	1.32	6	.58	0.65	3	.58	0.65	3	.28	1.79	5	.51	1.51	5
D	*	*	1	.66	0.34	2	.66	0.34	2	.66	0.34	2	.70	0.57	4	.55	0.94	5
E	*	*	1	.50	0.50	2	.66	0.43	3	.66	0.43	3	.48	1.27	5	.36	2.26	7
F	*	*	1	*	*	1	.36	2.02	6	.57	1.30	6	.40	1.81	6	.39	1.41	5
G	*	*	1	.51	1.30	6	.66	0.34	2	.66	0.34	2	.60	1.04	6	.65	1.08	6
H	*	*	1	.62	0.91	4	.53	1.63	7	.32	2.21	8	.51	1.41	6	.51	1.41	6
I	*	*	1	.50	0.50	2	.47	1.17	5	.41	1.74	7	.35	1.91	5	.35	1.91	5
J	*	*	1	.53	2.08	8	.56	0.70	3	.54	1.00	4	.66	0.34	2	.66	0.34	2
K	*	*	1	.32	2.21	8	.52	1.11	4	.52	1.53	5	.33	2.14	7	.58	1.32	6
L	*	*	1	.64	0.36	2	.45	1.05	4	.45	1.95	7	.65	1.32	7	.65	1.03	6
M	*	*	1	.48	1.14	4	.67	0.64	4	.55	1.00	5	.37	1.10	3	.37	1.64	4
N	*	*	1	.36	2.26	7	.37	1.10	3	.37	1.10	3	.60	1.30	6	.60	1.30	6
O	*	*	1	.58	1.54	7	.51	4.37	16	.48	4.10	15	.52	1.11	4	.52	1.11	4
P	*	*	1	.43	2.21	8	.51	1.41	6	.51	1.41	6	.53	1.63	7	.53	1.63	7
Q	*	*	1	.66	0.34	2	.37	2.17	8	.52	1.70	7	.58	0.93	4	.44	1.42	5
R	*	*	1	.68	0.68	4	.38	2.09	6	.65	0.52	3	.22	2.59	7	.56	1.00	4
S	*	*	1	.45	1.07	3	.50	1.29	5	.77	0.23	2	.45	1.51	5	.45	1.51	5
T	.77	0.23	2	.45	1.66	6	.58	0.93	4	.58	0.93	4	.67	0.64	4	.67	0.64	4
U	.20	2.18	6	.69	0.68	4	.68	0.32	2	.68	0.32	2	.51	1.31	6	.51	1.31	6
V	*	*	1	.51	1.15	4	.54	0.92	4	.54	1.29	5	.45	1.05	4	.67	0.58	3
W	-.21	26.73	39	.47	1.52	6	.41	0.89	3	.41	0.89	3	.41	1.32	5	.41	1.32	5
X	.73	0.27	2	.51	1.51	5	.58	0.90	4	.58	0.90	4	.68	0.32	2	.68	0.32	2
Y	.73	0.27	2	.58	0.65	3	.51	1.31	6	.51	1.31	6	.57	1.35	7	.62	1.21	7
Z	*	*	1	.40	1.84	6	*	*	1	.45	1.07	3	.41	0.89	3	.41	0.89	3

Table 1: Statistics of regions shown on the maps in figure 9. Names of the regions appear in the left-most column. C indicates the complete-link method; C/K, complete-link followed by K-means; W, Ward's minimum-variance method; W/K, Ward's followed by K-means; min, minimum within-region correlation coefficient; SS, within-region sum-of-squares; #, number of $6°\times 4°$ cells comprising the region; *, not meaningful because region has only a single cell.

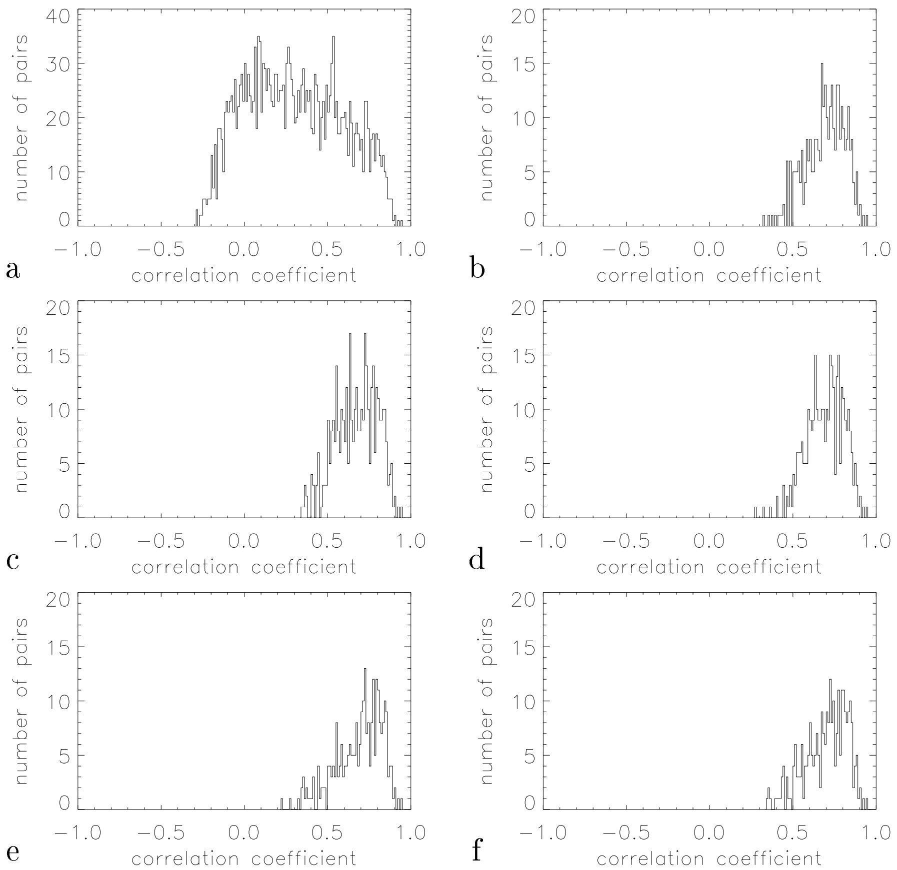

Figure 10: Histogram of correlation coefficients of seasonal sea-surface temperature anomalies for pairs of $6° \times 4°$ cells within common regions when the North Atlantic is partitioned into 26 regions using: (a) single-link method; (b) single-link followed by K-means; (c) complete-link; (d) complete-link followed by K-means; (e) minimum-variance; (f) minimum-variance followed by K-means.

K-means starting with the single-link centroids as seeds would be so successful; it improved the within-region sum of squares from 68.24 to 32.10, which differs from the best value by only 3%. The three K-means regionalizations are all different, but they all have very similar within-region sum-of-squares; thus, we do not expect to find substantial improvement starting from other seeds.

It is interesting to note that all six maps have smaller regions in the northwestern part of the basin, which is consistent with the idea that variability of sea-surface temperatures is greater at high latitudes. The smaller regions are particularly evident in the zone of high thermal variability indicated in figure 2. All but the unacceptable single-link partition produced a single disjoint region within this zone: region Y for single-link followed by K-means; C for complete-link, before and after K-means; and I for Ward's minimum-variance, before and after K-means. The dendrograms in figure 8 show that this disjointness remains in the complete-link hierarchy until there are only 5 regions, and in the minimum-variance hierarchy until there are 8 regions. The lack of correlation between the separated parts and the intervening cells is remarkable.

The essential difference between the complete-link map and the minimum-variance map can be traced to the number of clusters in the southern part of the basin. The complete-link method produced two relatively large regions: A and O; while Ward's method produced five smaller regions: A, G, L, W, and Y. This distinction persists after K-means and results in the map with the most nearly uniformly sized regions. It not only has the smallest within-region sum-of-squares but also, according to table 1 and figure (10), the largest minimum within-region correlation coefficient.

Values for within-region sum-of-squares in table 1 can be divided by one less than the number of cells per region to give a spatial-variance-like measure of within-region variability. Even with this normalization, all regions do not have uniform variability. The (minimum,maximum) values are: (0.23,0.53) for S/K; (0.20,0.55) for C; (0.21,0.55) for C/K; (0.19,0.55) for W; and (0.21,0.55) for W/K. The cells corresponding to the maximum values are S, N, N, M, and M, respectively; they are all located in the zone of high thermal variability, and the last four are essentially the same region.

The histograms of correlation coefficients between seasonal anomalies for pairs of $6°\times4°$ cells within the same region, which are shown in figure 10, provide another characterization of the within-region variability of the six 26-region maps. The two large regions produced by the single-link method clearly are responsible for the large number of small and negative within-region correlations; they are also responsible for the much larger number of within-region pairs. The similarities exhibited by the other five histograms confirm the conclusion based on total within-region sum-of-squares that the corresponding five maps are roughly equivalent. They all show very few within-region correlations that are less than 0.5. The method that produced the largest minimum within-region correlation, the smallest total within-region sum-of-squares, and the most uniformly sized clusters, *i.e.*, Ward's method followed by K-means (panel f), has a histogram that peaks nicely at higher correlations.

It is interesting to compare the five "good" histograms of $6°\times 4°$ within-region correlation coefficients in figure 10 with the histogram of $2°\times 2°$ within-$6°\times 4°$ correlations in figure 5. First, none of the within-region correlations are as small as the smallest within-cell correlations. In other words, when observations are represented by cell averages, there can be greater difference in anomalous seasonal behavior of sea-surface temperatures for nearby $2°\times 2°$ cells than for more widely separated $6°\times 4°$ cells. Second, the distributions have roughly the same shape. In fact, the peak of the K-means histograms are located at approximately the same place as that of the within-cell histogram. Why is the variability so similar on these two different spatial scales?

4.2 Thirteen regions.

Because the single-link method is really not appropriate for determining regions of similar thermal behavior, we have chosen to omit it from the discussion of the 13- and 7-region results, limiting our presentation to the complete-link and minimum-variance methods and their K-means counterparts. The four 13-region maps are shown in figure 11.

Again, the smallest regions are in the northwest where thermal variability is the greatest. As expected from the dendrograms, the disjoint regions, C for complete-link and I for minimum-variance, are still evident; in fact, the K-means adjustment does not cause them to merge. Moreover, a new disjoint region has been formed: G for complete-link and J for minimum-variance. (They might have been anticipated from the dendrograms and the 26-region maps.) Note that the K-means method starting from the complete-link map causes G to coalesce, but it does not connect the parts of minimum-variance region J.

The 13-region complete-link map shows an unusually large region A ex-

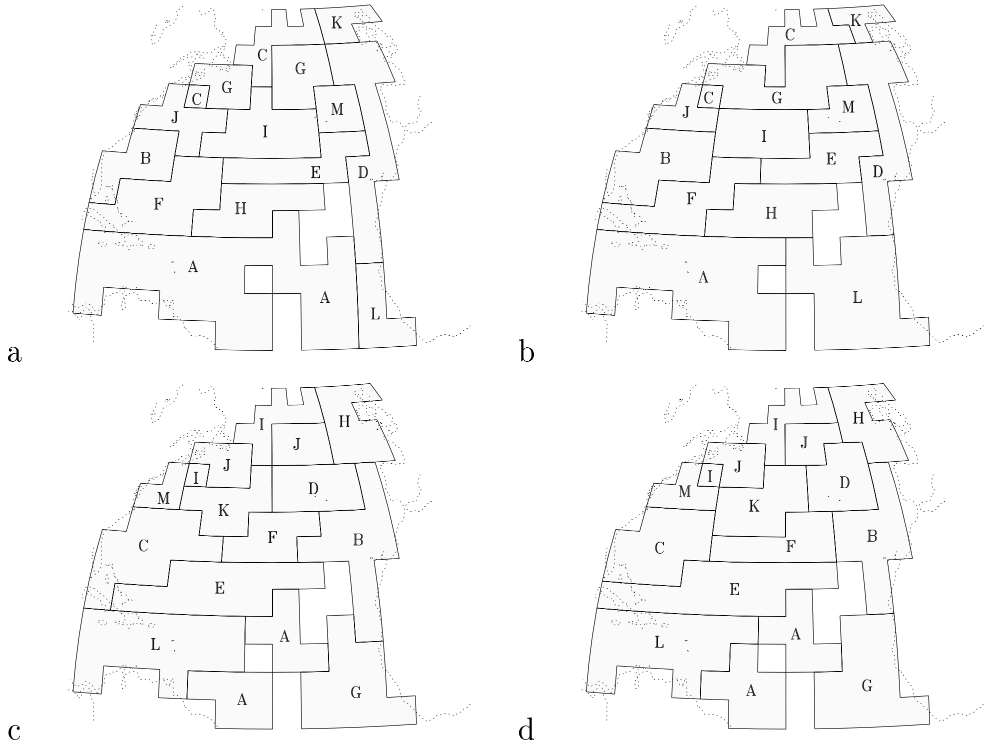

Figure 11: Partitions of the North Atlantic into 13 regions of similar anomalous seasonal SST variability produced using: (a) complete-link method; (b) complete-link followed by K-means; (c) minimum-variance; (d) minimum-variance followed by K-means.

tending over most of the southern third of the basin. From the dendrogram it is possible to see that it was formed at the 15-region stage. Even with this large region, the total within-region sum-of-squares is only 6% larger than the smallest; values are 50.50, 47.67, 48.24, and 47.43 for C, C/K, W, and W/K, respectively.

The (maximum,minimum) values for within-region spatial variance are: (0.21,0.58) for C; (0.25,0.55) for C/K; (0.28,0.55) for W; and (0.27,0.55) for W/K. The maxima correspond to region J for C and C/K and to M for W and W/K. They are all essentially the same region, except that for the complete-link map is somewhat larger. In fact, they are essentially the same as the maximum-spatial-variance region found on the 26-region maps.

The histograms of within-region correlation coefficients shown in figure

	C			C/K			W			W/K		
	min	SS	#	min	SS	#	min	SS	#	min	SS	#
A	.16	12.43	30	.35	6.34	20	.35	3.26	10	.35	3.54	11
B	.33	2.04	6	.14	3.73	9	-.01	5.13	12	.22	3.61	10
C	.20	4.71	10	.20	5.29	11	.06	5.70	12	.14	4.20	10
D	.17	6.26	14	.17	5.50	13	.25	2.45	8	.36	2.38	8
E	.23	2.76	7	.27	2.53	7	.23	4.50	12	.21	5.30	13
F	.24	4.05	10	.22	3.53	9	.40	1.81	6	.27	2.47	7
G	.18	6.12	13	.18	6.58	14	.25	3.11	10	.13	3.81	11
H	.53	1.63	7	.32	2.21	8	.12	5.47	12	.18	4.95	11
I	.21	3.52	10	.40	2.19	8	.20	4.71	10	.20	4.71	10
J	.14	2.88	6	.37	1.64	4	.18	5.21	11	.18	4.77	10
K	.23	2.42	6	.23	1.83	5	.33	2.14	7	.38	2.61	9
L	.45	1.05	4	.13	5.30	14	.51	3.65	14	.53	3.24	13
M	.67	0.64	4	.55	1.00	5	.37	1.10	3	.37	1.64	4

Table 2: Statistics of regions shown on the maps in figure 11.

(12) for the 13-region maps are similar to those found for the 26-region maps. The major difference is that the larger regions encompass cells with less-correlated anomalous seasonal thermal behavior. The minimum correlations are comparable with those of $2°\times 2°$ cells within $6°\times 4°$ cells (figure 5, but typically the within-region large-cell correlations are larger than the within-cell correlations. The histograms peaked most toward high within-region correlations is the one for the W/K map, just as was the case for 26 regions. And, as before, that map has the smallest total within-region sum-of-squares, the largest minimum correlation coefficient, and the most uniformly-sized regions.

4.3 Seven regions.

The 7-region maps produced by the complete-link and minimum-variance methods alone and followed by the K-means method are shown in figure 13. As was the case for the 26- and 13-region maps, the smaller regions are in the north and west where thermal variability is greatest.

It is interesting to note that the complete-link map's disjoint regions C and G have remained unchanged from the 13-region map. This could have been anticipated from the dendrogram in figure 8. Furthermore, region F

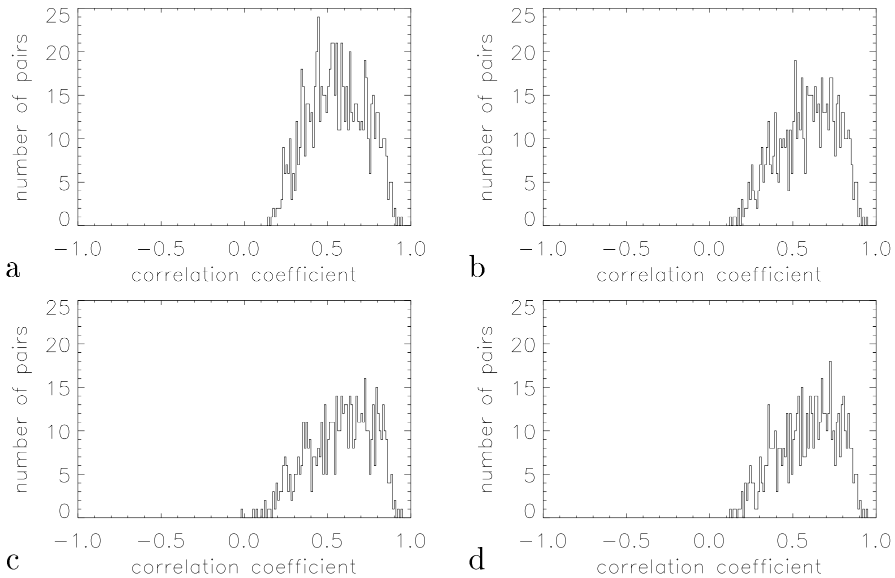

Figure 12: Histogram of correlation coefficients of seasonal sea-surface temperature anomalies for pairs of $6° \times 4°$ cells within common regions when the North Atlantic is partitioned into 13 regions using: (a) complete-link; (b) complete-link followed by K-means; (c) minimum-variance; (d) minimum-variance followed by K-means.

has two widely separated parts, one extending west from Cuba and the other in the far northeast. It is not surprising that the minimum within-region correlation corresponds to region F. What is remarkable is that these regions remain disjoint after K-means adjustment; C takes most of the northeast part of F, the remainder becoming part of region B, the bulk of which is adjacent to the southeastern United States. On the other hand, none of the seven W or W/K regions are disjoint; complete-link regions C and G are combined into the large northern region D, and the northeastern corner is part of a large region extending from Mauritania to Scotland.

The difference in number of southern regions that was noticed on the 26-region maps has persisted. While the complete-link hierarchy has produced a single large southern region, the minimum-variance hierarchy has maintained a distinction between southeast and southwest.

The total within-region sum-of-squares the four 7-region maps are: 65.37 for the complete-link map; 62.18 when complete-link is followed by K-means;

63.54 for Ward's method; and 60.94 when Ward's method is followed by K-means. As before, the best is Ward's method followed by K-means.

The normalized within-region sum-of-squares, which serves as a spatial variance, has the following (minimum,maximum) values: (0.47,0.65) for C; (0.41,0.62) for C/K; (0.35,0.65) for W; and (0.36,0.64) for W/K. The regions on the complete-link and complete-link/K-means maps having maximum spatial variance are those with the disjoint part in the north east; *i.e.*, F for C and B for C/K; for both W and W/K it is region D, which encompasses the other disjoint complete-link regions.

When there are only seven regions, all regions contain cells that have poorly correlated seasonal thermal anomalies. Figure 14 shows that roughly half of the within-region correlation coefficients are less than 0.5. However, it is remarkable that so few within-region correlation coefficients are negative, and these are essentially zero.

	C			C/K			W			W/K		
	min	SS	#	min	SS	#	min	SS	#	min	SS	#
A	.07	15.66	34	.07	16.44	35	.28	8.75	24	.32	7.53	22
B	.01	6.55	12	-.01	9.29	16	-.08	14.19	24	-.07	13.01	23
C	.20	4.71	10	-.10	7.61	14	-.14	8.55	15	.12	7.23	14
D	.17	6.26	14	.05	8.58	18	-.11	18.21	29	-.11	14.17	23
E	-.02	11.86	21	.23	4.14	11	.23	4.50	12	.09	8.03	17
F	-.05	14.20	23	.09	8.03	17	-.00	6.23	13	.13	6.19	15
G	.18	6.12	13	.11	8.10	16	.25	3.11	10	.13	4.78	13

Table 3: Statistics of regions shown on the maps in figure 13.

4.4 Hand-drawn regions.

Because the complete-link and minimum-variance regions and their K-means counterparts had qualitatively similar within-region variability, we decided to partition the North Atlantic by hand to determine whether an arbitrary division would produce regions having about the same variability. We learned that it was quite difficult to do as well as the clustering algorithms. Figure 15 shows example of 26, 13, and 7 hand-drawn regions and corresponding histograms of within-region correlation coefficients. Comparing the histogram for the 26-region map with those shown in figure 12 reveals that the regions on the hand-drawn map are much less coherent. It is much more like the

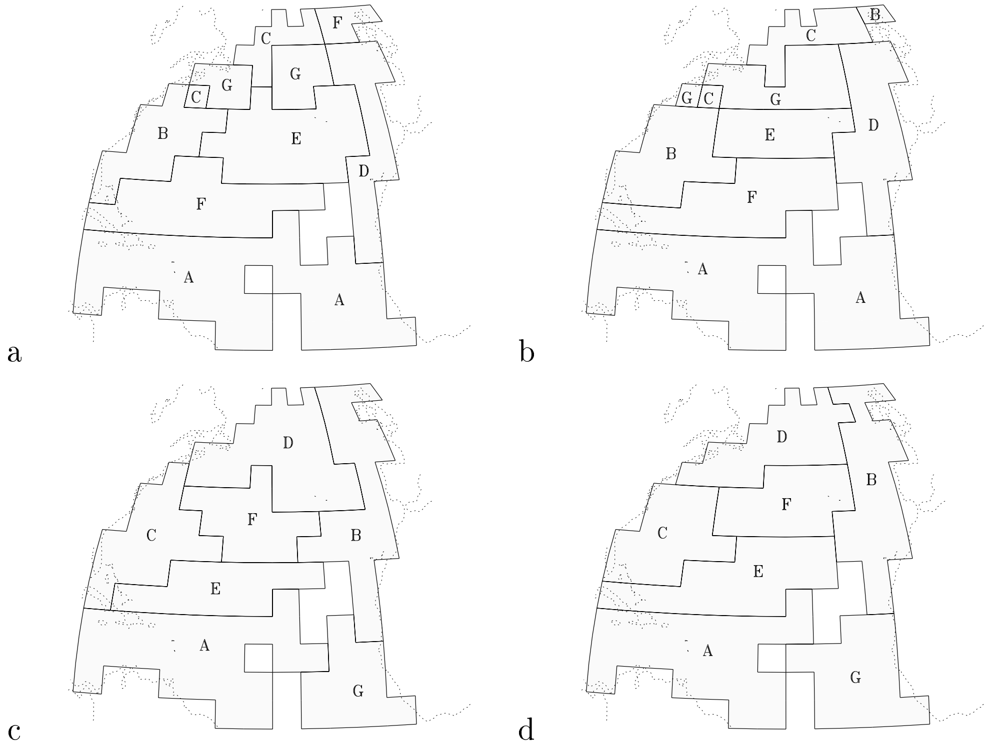

Figure 13: Partitions of the North Atlantic into 7 regions of similar anomalous seasonal SST variability produced using: (a) complete-link method; (b) complete-link followed by K-means; (c) minimum-variance; (d) minimum-variance followed by K-means. Land masses are indicated by dotted lines.

13-region histograms of figure 12. Furthermore, the histogram for the hand-drawn 13-region map is not much better than for the hand-drawn 7-region map; they both have a greater percentage of negative correlations than the histograms for the 7-region maps shown in figure 14. Thus, even though clustering algorithms produce regions with substantial variability, these regions are still considerably more coherent than are regions obtained by arbitrarily dividing the basin without guidance from the data.

5 Conclusion.

Our objective has been to explore the use of cluster analysis as a tool for partitioning the North Atlantic into regions within which the anomalous

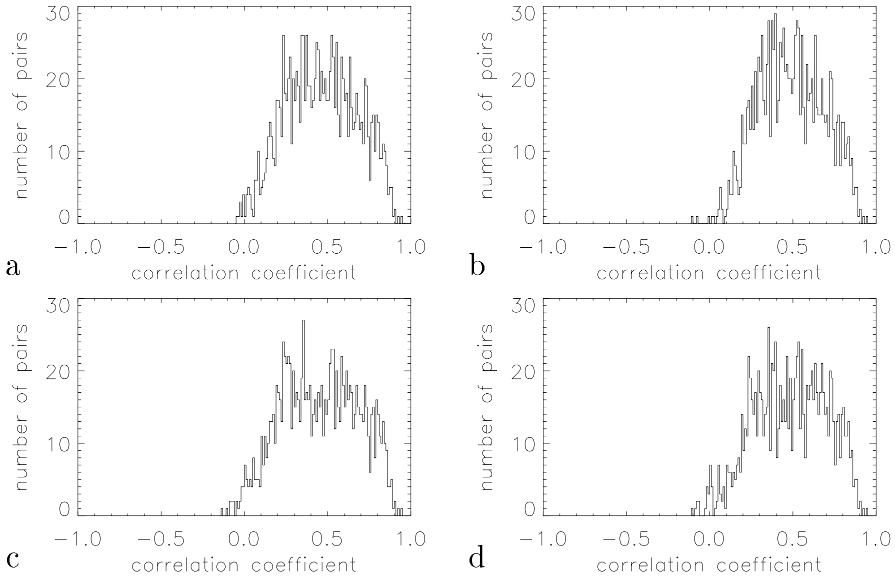

Figure 14: Histogram of correlation coefficients of seasonal sea-surface temperature anomalies for pairs of $6^\circ\times 4^\circ$ cells within common regions when the North Atlantic is partitioned into 7 regions using: (a) complete-link; (b) complete-link followed by K-means; (c) minimum-variance; (d) minimum-variance followed by K-means.

seasonal behavior of the sea-surface temperature field is reasonably coherent. We have shown that the complete-link and minimum-variance methods both produce hierarchies of regionalizations that can be judged as acceptable and that the best results were obtained when the results of the minimum-variance method were used as the starting point for the K-means method.

One goal has been to determine whether the thermal variability exhibited by the data is such that there is an obvious "best" number of thermal regions into which the North Atlantic should be divided. If so, using fewer regions should cause a dramatic increase in within-region variability. Examining within-region variability as a function of number of regions revealed a steady and growing increase but no abrupt jump. Thus, the number of regions is a matter of judgment and depends on the amount of variability considered acceptable within the regions.

We have found that, if correlations between the seasonal anomalies of the $6^\circ\times 4^\circ$ cells comprising the regions are required to be no less than 0.7, several

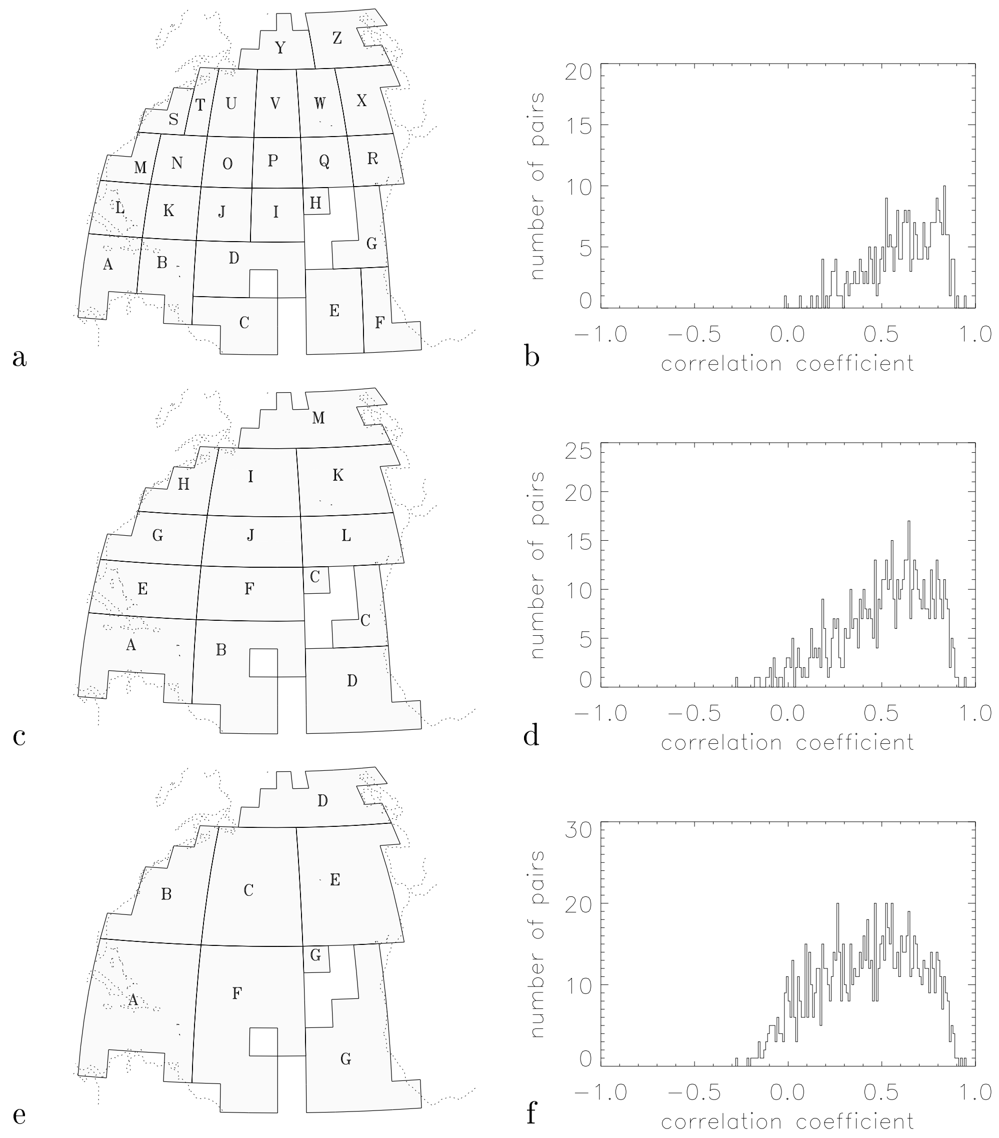

Figure 15: Hand-drawn maps and corresponding histograms of within-region correlation coefficients of anomalous seasonal sea-surface temperatures. (a) 26-region map; (b) 26-region histogram; (c) 13-region map; (d) 13-region histogram; (e) 7-region map; (f) 7-region histogram.

dozen regions will be required. However, this degree of homogeneity is greater than that exhibited by the seasonal anomalies of $2°\times 2°$ cells within the $6°\times 4°$ cells that are the building blocks of the regions. With 26 regions, the within-region variability $6°\times 4°$ variability is roughly the same as the within-cell $2°\times 2°$ variability; most within-region correlation coefficients are larger than 0.7, but a substantial number are between 0.5 and and 0.7, and some are even smaller. For 13 regions, it is impossible to avoid grouping together cells with essentially uncorrelated thermal behavior, and for 7 regions, the heterogeneity increases.

The cluster analysis has revealed that the greatest anomalous seasonal variability is found in the northwestern part of the basin in the vicinity of the Gulf Stream. Typically, the clustering algorithms prefer smaller regions in this area. Furthermore, the regions with the greatest spatial variance consistently lie in this area. One remarkable result is that only in this area do the clustering algorithms consistently find a spatially disjoint region; everywhere else, the regions of similar thermal behavior were found to be contiguous. We found the largest within-cell variability also to occur in this northwestern part of the basin. Perhaps the $6°\times 4°$ cells do not provide sufficient resolution and clustering of $2°\times 2°$ cells would have produced contiguous regions everywhere. We outlined an alternate treatment of the data, where correlations are computed using only cells with sufficient data, which might be appropriate for clustering $2°\times 2°$ cells, and it would be interesting to try it.

Our motivation has been the idea of formulating statistical models for forecasting climatic anomalies in terms of regional variables. We have learned that, unless there are a fairly large number of regions, there will be considerable within-region variability. Thus, if the number of regions is comparable to the number of principal components typically used in statistical models, the behavior of the individual cells that comprize the region will not be well characterized by the average for that region. Does this mean that regional variables are inappropriate for such models? No. We intend to show in another paper that a few regional variables can explain the local details roughly as well as the same number of principal components. The important point is that, for a given predictand, the number of useful predictors is limited by the size of the training set. There is a great deal of freedom in choosing both predictors and predictands that has not been explored.

References

Anderberg, M. R. (1973). *Cluster Analysis for Applications.* Academic Press, New York.

Barnett, T. P., and R. W. Preisendorfer (1978). Multifield analog prediction of short-term climate fluctuations using a climate state vector. *Journal of the Atmospheric Sciences*, **35**, 1771–1787.

Barnett, T. P., M. Latif, N. Graham, M. Flügel, S. Pazen, and W. White (1993). ENSO and ENSO-related predictability. Part I; Prediction of Equatorial Pacific sea surface tempreature with a hybrid coupled ocean-atmosphere model. *Journal of Climate*, **6**, 1545–1566.

Barnston, A. G., and R. F. Livezey (1989). An operational multifield analog/antianalog prediction system for the United States seasonal temperatures. Part II: Spring, summer, fall, and intermediate 3-month period design experiments. *Journal of Climate*, **2**, 513–541.

Barnston, A. G., and C. F. Ropelewski (1992). Prediction of ENSO episodes using canonical correlation analysis. *Journal of Climate*, **5**, 1316–1345.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York.

Bretherton, F. E., R. E. Davis, and C. B. Fandry (1976). A technique for objective analysis and design of oceanographic experiments applied to mode-73. *Deep-Sea Research*, **23**, 559–582.

Cheng, X., and J. M. Wallace (1993). Cluster analysis of the Northern Hemisphere winter 500-hPa height field: Spatial patterns. *Journal of the Atmospheric Sciences*, **50**, 2674–2696.

Davis, R. E. (1976). Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *Journal of Physical Oceanography*, **6**, 249–266.

Drosdowsky, W. (1993). An analysis of Australian seasonal rainfall anomalies: 1950–1987. I: Spatial patterns. *International Journal of Climatology*, **13**, 1–30.

Everitt, B. (1974). *Cluster Analysis.* Heinemann Educational Books, London.

Fovell, R. G., and M.-Y. C. Fovell (1993). Climatic zones of the conterminious United States defined using cluster analysis. *Journal of Climate*, **6**, 2103–2135.

Gandin, L. S. (1963). *Applied Optimal Estimation.* Israel Program for Scientific Translations, Jerusalem. Translated from the Russian.

Gong, X., and M. B. Richman (1995). On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *Journal of Climate*, **8**, 897–931.

Graham, N. E., J. Michaelson, and T. P. Barnett (1987). An investigation of the El Niño – Southern Oscillation cycle with statistical models. 1. Predictor field characteristics. *Journal of Geophysical Research*, **92**, 14251–14270.

IMSL (1989). *User's Manual, IMSL STAT/LIBRARY, FORTRAN Subroutines for Statistical Analysis.* IMSL, Inc., Houston TX.

Iwasaka, N., K. Hanawa, and Y. Toba (1988). Partition of the North Pacific Ocean based on similarity in temporal of SST anomaly. *Journal of the Meteorological Society of Japan*, **66**, 433–443.

Johnson, R. A., and D. W. Wichern (1992). *Applied Multivariate Statistical Analysis.* Prentice Hall, Englewood Cliffs NJ.

Livezey, R. F., and A. G. Barnston (1988). An operational multifield analog/antianalog prediction system for the United States seasonal temperatures 1. System design and winter experiments. *Journal of Geophysical Research*, **93**, 10953–10974.

Livezey, R. F., A. G. Barnston, G. V. Gruza, and E. Y. Ran'kova (1994). Comparitive skill of two analog seasonal temperature prediction systems: Objective selection of predictors. *Journal of Climate*, **7**, 608–615.

Mariano, A. J., and O. B. Brown (1992). Efficient objective analysis of dynamically heterogeneous and nonstationary fields via the parameter matrix. *Deep-Sea Research.*

Mo, K., and M. Ghil (1988). Cluster analysis of multiple planetary flows. *Journal of Geophysical Research*, **93**, 10927–10952.

NAG (1993). *The NAG fortran library manual, Mark 16.* National Algorithms Group Limited, Oxford, UK.

Penland, C., and T. Magorian (1993). Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *Journal of Climate*, **6**, 1067–1076.

Richman, M. B. (1986). Rotation of principal components. *Journal of Climate*, **6**, 293–335.

SAS (1990). *SAS/STAT User's Guide, Version 6.* SAS Institute Inc., Cary NC, 4th edition.

SPSS (1988). *SPSS-X User's Guide.* SPSS, Inc., Chicago IL, 3rd edition.

STATISTICA (1994). *Statistica for Windows.* StatSoft, Inc., Tulsa OK.

Stooksbury, D. E., and P. J. Michaels (1991). Cluster analysis of Southeastern U.S. climate stations. *Theoretical and Applied Climatology*, **44**, 143–150.

Thacker, W. C. (1995a). Climatic indices, principal components, and the Gauss-Markov theorem. Submitted to *Journal of Climate.*

Thacker, W. C. (1995b). Statistical modeling for numerical modelers. External Report GKSS 95/E/4, GKSS Forschungszentrum, Geesthacht.

von Storch, H., and G. Hannoschöck (1985). Statistical aspects of estimated principal vectors (EOFs) based on small sample sizes. *Journal of Climate and Applied Meteorology*, **24**, 716–724.

von Storch, H., and J.-S. Xu (1990). Predicting the state of the Southern Oscillation using principal oscillation pattern analysis. *Journal of Climate*, **3**, 1316–1329.

Wallace, J. M., and D. S. Gutzler (1981). Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review*, **109**, 784–812.

Wolter, K. (1987). The Southern Oscillation in surface circulation and climate over the Tropical Atlantic, Eastern Pacific, and Indian Oceans as captured by cluster analysis. *Journal of Climate and Applied Meteorology*, **26**, 540–558.

Woodruff, S. D., R. J. Slutz, R. L. Jenne, and P. M. Steurer (1987). A comprehensive ocean-atmosphere data set. *Bulletin American Meteorological Society*, **22**, 1239–1250.

You, Y., and M. Tomczak (1993). Thermocline circulation and ventilation in the Indian Ocean derived from water mass analysis. *Deep-Sea Research*, **40**, 13–56.