Climate Zones of the Conterminous United States Defined Using Cluster Analysis

ROBERT G. FOVELL AND MEI-YING C. FOVELL

Department of Atmospheric Sciences, University of California, Los Angeles, Los Angeles, California

(Manuscript received 2 December 1991, in final form 1 April 1993)

ABSTRACT

A regionalization of the conterminous United States is accomplished using hierarchical cluster analysis on temperature and precipitation data. The "best" combination of clustering method and data preprocessing strategy yields a set of candidate clustering levels, from which the 14-, 25-, and 8-cluster solutions are chosen. Collectively, these are termed the "reference clusterings." At the 14-cluster level, the bulk of the nation is partitioned into four principal climate zones: the Southeast, East Central, Northeastern Tier, and Interior West clusters. Many small clusters are concentrated in the Pacific Northwest. The 25-cluster solution can be used to identify the subzones within the 14 clusters. At that more detailed level, many of the areally more extensive clusters are partitioned into smaller, more internally cohesive subgroups.

The "best" clustering approach is the one that minimizes the influences of three forms of bias—methodological, latent, and information—for the dataset at hand. Sources of, and remedies for, these biases are discussed. Sensitivity tests indicate that some of the clusters in the reference clusterings lack robustness, especially those in the Northeast quadrant of the United States. Some of the tests involve small and large alterations to the data preprocessing strategy.

The major shortcomings of the analysis procedure are that the clusters are unnaturally constrained to be nonoverlapping and also that potentially important data from points outside of the political boundaries of the conterminous United States and over water are not included. Also, other variables that could be important or useful in characterizing climate type could be added to, or used in place of, the temperature and precipitation variables used herein. Further work on data preprocessing techniques is also required. Remedies for these and other shortcomings are proposed.

1. Introduction

Several researchers have used a variety of data to define climatic types and delineate zones of similar climate. The most famous examples are the Koeppen (1923) and Thornthwaite (1931) classifications. Although they were motivated by different reasons and utilized different data, both classifications entailed the a priori definition of a set of climate types or rules that were then used to classify each area of the earth. The climatic types were externally specified or indirectly suggested by the data, instead of directly issuing from the data.

That particular approach is by no means unjustifiable. This work, however, presents an alternative approach, one that attempts to delineate climate zones in a more direct fashion. The data used are temperature and precipitation in the form of long-term monthly means for the conterminous United States. The basic idea is quite simple: locales that have similar characteristics (including means and variances) with respect to these variables should have roughly similar climates. Other variables potentially important or useful in characterizing climate type could be added to the dataset in the future, including those that combine temperature and precipitation information in some specific way (such as potential evapotranspiration, which forms the basis of the Thornthwaite classification).

To objectively determine the climatic zones present within the conterminous United States, a somewhat controversial technique known as cluster analysis is employed. Cluster analysis has long been used in biology, psychology, and other disciplines, where it has generated extensive literature. Its usage in the atmospheric sciences has been comparatively rare, however, possibly due to the controversy and the inherent ambiguities in its use (Wolter 1987). Still, papers using cluster analysis are now appearing in the atmospheric sciences journals at an increasing rate.

Some forms of cluster analysis begin with the identification of a set of variables tabulated for each member of a set of objects or cases that is the subject of the clustering. Then, some measure of similarity or dissimilarity between pairs of objects (the concept of *distance*) is chosen. Preprocessing of the variables before calculation of the distance measures is an area of significant concern. Different variables may be measured on different scales and may also contain irrelevant and/

Corresponding author address: Dr. Robert G. Fovell, Department of Atmospheric Sciences, University of California, 405 Hilgard Ave., Los Angeles, CA 90024.

or redundant information. Some of the above-cited studies have applied cluster analysis to raw data, one variable at a time (e.g., Wolter 1987). This method avoids the problem of mixing variables of different type or scale but also results in a different clustering for each variable, which may not always be desirable.

When confronted with variables of differing type or scale, or simply with an excessively large number of potentially important variables, a number of researchers have adopted variable manipulation and/or reduction strategies such as principal components analysis (PCA) (e.g., Gadgil and Iyengar 1980; Gadgil and Joshi 1983; Maryon and Storey 1985; Kalkstein et al. 1987; and Kalkstein et al. 1990, to name a few). PCA creates new variables (components) composed of mutually orthogonal linear combinations of the original variables, each accounting for a specific fraction of the original total variance as indicated by the size of its associated eigenvalue. Retention of only the most significant components accomplishes variable reduction while ostensibly minimizing information loss. These new variables can be used to generate component scores that can be clustered in place of the raw data. PCA has been used by itself to perform regionalizations (e.g., Richman and Lamb 1985).

There are two goals in this paper: to discuss problems confronted prior to the commencement of clustering, including variable preprocessing issues, and to present and assess the quality and stability of a regionalization of the conterminous United States–based clustering of the processed temperature and precipitation dataset. The data are described in section 2, and section 3 presents some background material on cluster analysis and the preprocessing strategies adopted. The first goal involves the identification and assessment of potential biases—in the dataset, data preprocessing procedures, and clustering strategies—that if left unconfronted may exert an overwhelming influence on the regionalization. This is discussed in section 4, which also includes a few proposed remedies for some of the biases and identifies additional problems that require further study. The "best" combination of clustering method and data preprocessing strategy yields regionalizations that are collectively referred to as the "reference clusterings." These are described in section 5. Section 6 presents a few "variant" clusterings that result when different strategies are adopted. The final section presents the summary and proposals for future work.

2. Data

This study employs the widely used National Climatic Data Center (NCDC) climate division dataset, which, in this application, consists of monthly temperature means and precipitation accumulations (24 variables total) for the conterminous United States over the 50-year period 1931–1980. The domain consists of 344 climate divisions of irregular size; a base map is presented in Fig. 1. The dataset is reduced to two dimensions by constructing 50-year means for each of the 24 variables. Although this sacrifices some information that might be valuable in the clustering process, it was done primarily because we wish to process the variables with a principal components analysis (PCA). This is an "R-mode" analysis (Cattell 1952).

Most of the 24 variables are not normally distributed across the 344 divisions. No single transformation can force normality, even for a particular data type (temperature or precipitation). The skewness of the monthly precipitation variable distributions with respect to locations evinces a strong seasonal cycle, being substantially positive in winter and slightly negative in summer. Kurtosis displays qualitatively similar behavior. Seasonality is also present in the temperature distribution data, but with extreme skewness centered in the spring and autumn seasons while summer months evince bimodal distributions.

Fortunately, neither cluster nor principal components analysis explicitly demands that variables be normally distributed to operate correctly. (However, PCA utilizes correlations or covariances that implicitly assume linear relationships among the variables.) Departures from normality do hamper the usage of statistical population tests on the clustered data like the F and t tests, but a more serious obstacle is that clusters cannot be thought of as randomly drawn samples. Indeed, if they were no better than random samples, the clusters would have no value. Despite this caveat, we will utilize "pseudostatistics" constructed to mimic F and t^2 tests, which have been evaluated in the cluster analysis literature, in an advisory role (section 3d).

3. Background

As this paper is concerned with a combined application of principal components analysis and cluster analysis on a multivariate dataset, we will sketch below some background material for the two techniques, along with a brief treatment of the "number of clusters" problem. Richman's (1986) review article provides an excellent introduction to PCA. As conventions regarding symbols and terminology in PCA are sadly lacking, however, we have elected to generally follow the notation employed by Jackson (1991). More complete discussions of clustering methods and their inherent characteristics may be found in such textbooks as Spaeth (1980) and in papers such as Kalkstein et al. (1987).

a. Principal components analysis (PCA)

Say we have an $n \times p$ data matrix X, where n is the number of objects and p is the number of variables. The means of the p variables have been removed, but no other manipulation of the original raw data has been made. PCA may then be applied to the variables

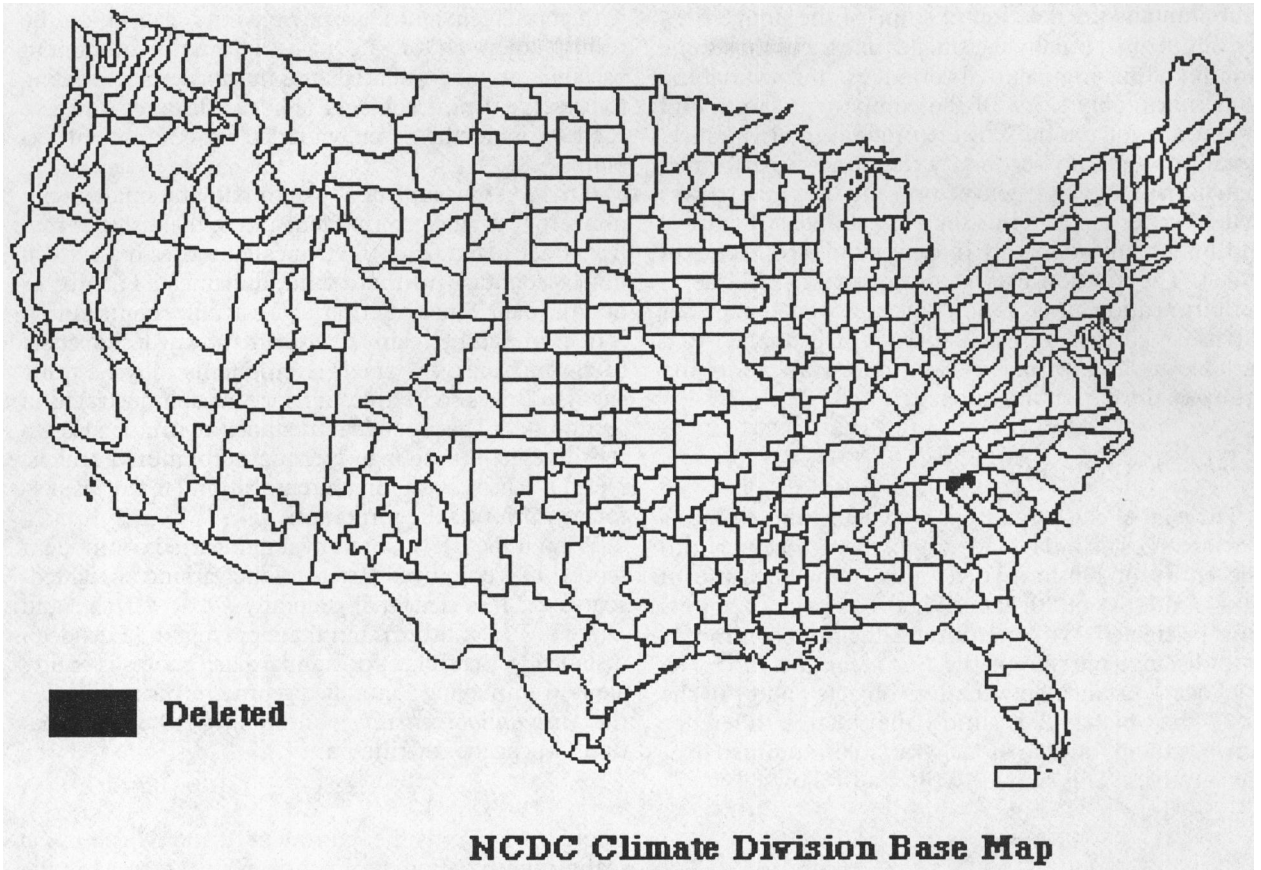

FIG. 1. Base map of the 344 climate divisions of the conterminous United States in the National Climatic Data Center dataset. Note how the resolution varies radically over the domain. The highlighted division, in South Carolina, was not included in most of the clusterings presented in sections 5 and 6.

or the objects. The initial goal in either case is the replacement of the original correlated entities (variables or objects) with new component entities that are mutually uncorrelated.

A typical PCA applied to the variables can start with S, the $p \times p$ variable covariance matrix defined as $X^T X(n-1)^{-1}$. This matrix is transformed into a diagonal eigenvalue matrix L, which has as many as $\min(n, p)$ nonzero entries, and an orthonormal eigenvector matrix U, both dimensioned $p \times p$. The p column vectors composing U are unit length and mutually uncorrelated. This eigenvector matrix is often reformulated as $V = UL^{1/2}$, giving the kth column of V length equal to the kth eigenvalue.

In the atmospheric sciences literature, the eigenvectors in U are often termed "empirical orthogonal functions," or EOFs. The entries of this matrix, the eigenvector "loadings," define new variables, consisting of linear transformations of the original variables, which will be termed herein as "component variables" or "principal components" (PCs). The multiplication $Z = XU$ generates another $n \times p$ data matrix, where the p dimension now represents the new uncorrelated component variables, $Z_1, Z_2, \ldots, Z_p$. When constructed in this manner, the new component variable Z_k has variance equal to its associated eigenvalue l_k, the kth diagonal entry in L, and the sums of the variances of the new components equals that from the original variables (i.e., total variance is preserved). The new data values for the objects will be termed herein as the "variance-weighted scores," but there are a variety of names used to describe these new data.[1] If the new score matrix is formed as $Y = XUL^{-1/2}$, then each new component variable, $Y_1, Y_2, \ldots, Y_p$, has identical unit variance, no matter what portion of the original variance its component actually represents. These new data will be referred to as "standardized scores." Note that total variance may not be preserved in this case.

The eigenvector matrix resulting from the PCA is often subsequently *truncated* and/or *rotated.* Trun-

[1] Sometimes, the term z scores have been used for these scores (Jackson 1991), but this term is avoided herein as it is also often used to describe the transformation of data to common unit variance, a diametrically opposed definition.

cation entails the deletion of some of the nonzero eigenvalues of L, usually the smaller ones, and thus some amount of the original total variance of the p variables. Sometimes, only a few of the components are found to account for the bulk of the original total variance. Rotation—actually secondary rotation—is performed to make a set of new eigenvectors (and thus new scores) that meet certain criteria (such as "simple structure") and may be more readily interpretable (see Richman 1986). The rotation may be orthogonal or oblique. If both truncation and rotation are employed, the former is performed first. In this paper, PCA is applied to a variable covariance matrix, and both truncation and orthogonal rotation are applied.

b. *The concept of "distance"*

The goal of cluster analysis is to detect those objects that are most similar and group them together into one entity or cluster. To do this, some measure of (dis-)similarity or "distance" between pairs of objects must be created. With metric data, the most commonly used distance measure is the *Euclidean distance.* The Euclidean distance between two objects i and j in the $n \times p$ data matrix X is simply the squared difference between them for each of the p variables, summed over the variables. This may be written as follows:

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ki} - x_{kj})^2 \right]^{1/2}. \qquad (1)$$

This expands into an $n \times n$ distance (or proximity) matrix D, which is symmetric about a main diagonal containing zeroes.

Note that we have collapsed the data into an $n \times n$ matrix, eliminating the individual variables. Because of this, there are numerous practical problems with using this distance measure. First, the *average* contribution of each of the p variables to D is dictated by its relative variance (Sokal and Sneath 1963), which is especially troublesome when the variables being combined are measured in different—and usually arbitrary—units. This concern has been most commonly addressed in the past by transforming each variable to common unit variance. However, recent work has shown that other standardizations might be superior (Milligan and Cooper 1988).

The second concern lies in the fact that the p variables are often intercorrelated and, except under special conditions, operations that change the variable intercorrelations will change the Euclidean distance between any given pair of objects. These conditions will be illustrated using PCA. Let $D(X)$ be the Euclidean distance matrix for the objects computed using the p original correlated variables. Then PCA is performed on the variable covariance matrix, creating the variance-weighted-score matrix, Z. If all the components are retained (optionally including components associated with zero eigenvalues, as they have no bearing on the result), then $D(Z) = D(X)$. The new component variables are uncorrelated, and the appropriate scaling to preserve the interobject Euclidean distance is to give each variable the same variance its component explains.

If the PCA is truncated to eliminate the smaller (but nonzero) variance components, then the distance matrix constructed from the truncated scores matrix will not be identical to that of the untruncated scores or original data. Since each variable's average contribution is determined by relative variance, however, the effect of eliminating very small components may be quite slight. (This is sometimes illusory, as demonstrated in section 4c.) The Euclidean distance matrix is also insensitive to orthogonal secondary rotations, that is, $D(Z) = D(Z_{rot})$, so long as truncation (if any) is performed prior to the rerotation.

What if the $n \times p$ matrix of standardized component scores, Y, is employed instead of the variance-weighted-scores Z? It is clear that generally $D(Y) \neq D(X)$ and $D(Y) \neq D(Z)$. It turns out that applying the Euclidean distance to the full set of standardized scores is equivalent to employing a modified geometric distance called the *Mahalanobis distance* on the original data. Equation (1) can be rewritten as

$$d_{ij} = [(\vec{x}_i - \vec{x}_j)M(\vec{x}_i - \vec{x}_j)^T]^{1/2} \qquad (1a)$$

where $x_i (1 \times p)$ is the ith row of X and M is a $p \times p$ scaling matrix. For the Euclidean distance, M is the identity matrix, while for the Mahalanobis metric, $M = S^{-1}$, the inverse of the variable covariance matrix. Designating the Mahalanobis distance matrix as D_m, then, it is found that if the full set of p components are retained (necessarily excluding all components with identically zero eigenvalues), $D_m(X) = D_m(Z) = D_m(Y) = D(Y)$. Thus, using the Mahalanobis distance metric on the original variables is tantamount to taking those variables and forcing them to be both uncorrelated and standardized, as can be done with a PCA on the variables yielding standardized scores. The issue of truncating standardized components is raised in section 4b.

Another alternative is to use a dissimilarity measure formed in some fashion from the $n \times n$ correlation matrix of the objects. The use of the correlation coefficient r_{ij} as a measure of similarity is common but not always appropriate (Cronbach and Gleser 1953). The Euclidean distance responds to differences in object means, variances, and distributional shapes with respect to the p variables between any two objects, unless these differences have already been adjusted out of the data. Differences in object means and variances are eliminated when the interobject correlations are computed, leaving only information regarding the shapes. Indeed, the interobject correlation coefficient is directly related to the Euclidean distance computed using data that has already been standardized to zero mean and

unit variance for each object, by $d_{ij} = [2(n - 1)(1 - r_{ij})]^{1/2}$. In applications—such as the present one—where differences in object means and variances are considered important information that should be reflected in the clusters, standardizing the data with respect to the objects is not appropriate. In addition, standardizing the objects effectively unstandardizes the variables, if that had been performed to eliminate arbitrary variable scaling differences. Finally, we note that several other distance metrics exist that are comparatively rarely used.

c. Clustering algorithms and approaches

However it is generated or justified, the distance matrix is used as input to some clustering algorithm. The most commonly used clustering strategies are hierarchical and partitioning cluster analysis. Both result in the establishment of "hard" (nonoverlapping) clusters, in which each object is assigned to only one cluster. This may be excessively restrictive in many applications (including the present one). The former strategy itself has two varieties: agglomeration and division. Agglomeration starts with n clusters, each containing one member, and at each step fuses the cluster pair with minimum separation distance to form a new cluster. The number of clusters remaining in the dataset is decreased by one each step, terminating when one all-inclusive cluster is created. The results are then inspected to determine which step or clustering level represents the "best" solution. Division simply operates in reverse. Both, however, can only act on the clusters that it has fused (or divided) at earlier steps and cannot reconsider what it has created previously.

In the partitioning clustering approach, an independent clustering is obtained for each desired clustering level. Because of this independence, partitioning clustering can optimize the division of the objects at each step. However, partitioning analyses require an initial selection of objects to act as cluster "seeds." Difficulties involved in the selection of seeds and the sensitivity of the results to the seeds chosen may be reasons why the partitioning approach is relatively less popular. Also, if the goal is to identify not only distinct clusters *but also the subgroups they may contain,* which is a goal of this work, then the hierarchical approach is the more attractive of the two anyway. In our opinion, the "number of clusters" problem (next subsection) is also less severe in hierarchical analyses, which is why the hierarchical approach was chosen for this study.

There are many different agglomerative hierarchical clustering methods. All methods agree that the cluster pair with the smallest distance should be joined at any step. They differ on how the distance between the new and remaining clusters is recomputed. Say clusters A and B are joined at some step, because their separation distance (d_{AB}) is the smallest. With the *single linkage* method, the distance between the new cluster AB and another cluster C (i.e., the distance d_{AB-C}) is taken to be smaller of the two distances between the original clusters and cluster C (i.e., d_{AC} and d_{BC}). If the larger of the two distances is chosen, the method is called *complete linkage.* If the average is taken, then the method might be called *simple average linkage.*

None of those three methods takes into account the number of members in either cluster A or cluster B in determining the new distances, which might allow a cluster with few members to unfairly influence the new distances after being merged with a much larger one. An alternative is to use a weighted average of the mean cluster distances, which is tantamount to taking the average distance over all involved object pairs, including those already absorbed into the clusters at previous steps. This creates *group average linkage,* a popularly used method that will be referred to herein as just "average linkage." In this method, the recomputed distance between new cluster AB and cluster C is given by

$$d_{AB-C} = (N_A d_{AC} + N_B d_{BC})/(N_A + N_B), \quad (2)$$

where N_A and N_B are the number of members in clusters A and B. Additional background on this method can be found in Kalkstein et al. (1987). Another popular method is the *Ward method,* a least-squares technique designed to minimize information loss at each merger (e.g., Ward 1963; Spaeth 1980). It includes not only the distances d_{AC} and d_{BC} in the recomputation, but also the distance between the fused clusters d_{AB} as well.

This study employs the group average linkage method, using our own extensively tested code. Following SAS, though, we used squared Euclidean distances; note that some other software packages take the square root by default.[2] The discussion about average linkage and its characteristics in Kalkstein et al. (1987) assumes the usage of squared distances. Spaeth (1980), as well as other sources, gives the formula for average linkage using the square-root distance. In practice any distance metric should probably be usable, including the Mahalanobis distance.

d. The number of clusters problem

Cluster analysis represents a grand compromise between specificity and generality. Each merger unavoidably results in a loss of precision and detail, but this is justifiable as long as the ability to interpret and generalize is enhanced. However, unlike many other studies, particularly in the social sciences, the extraction of "true" clusters from the data is not the principal goal of this work, although some may actually exist. Particularly in the eastern portion of the United States, temperature and precipitation vary rather smoothly

[2] Obviously, squaring the distance enhances intercluster dissimilarities. As the clustering algorithms typically recompute distances additively, this choice does have some effect on the clustering outcome.

over space, so it is unlikely there will be any truly "hard" clusters there anyway, but rather "fuzzy" or overlapping clusters. Instead, the main goal is to find an *adequate subdivision of the objects,* in order to define climate types, without sacrificing too much detail. Thus, we will use principally subjective means to choose among the many clustering levels available. In this approach, there is no one "correct" clustering level, but rather a number of viable solutions depending on the level of detail desired.

However, even if we are not overly concerned with the recovery of "true" clusters, the "number of clusters" problem can still be a vexing one. Objectively, some clustering levels may be more appropriate than others. Even though detail is lost with each merger, it is unlikely that this loss is a smoothly varying function of the clustering level. A long-standing rule of thumb has been to plot the distance between the clusters merged at each step and look for plateaus or natural breaks. Another approach is to compute a ratio of the sum of squares accounted for in the clustering by the total sum of squares, yielding a coefficient of determination, which can be examined the same way. Neither of those methods performed particularly well in the Monte Carlo simulations of Milligan and Cooper (1985), particularly not the latter. Of the 30 "stopping rules" they evaluated, the best performing were Calinski and Harabasz' (1974) "pseudo F," statistic, and the "Je(2)/Je(1)" criterion proposed by Duda and Hart (1973). The authors cautioned that performance of some these rules may be data dependent. The simulations were made using clusters that were distinct and well separated (in Euclidean space), which should permit maximum performance of the various rules. In practice, many (if not most) of the clusters in the data will not be so well defined.

The Calinski and Harabasz pseudo-F criterion is given by the formula below:

$$\text{pseudo } F = [A/W]^*[(n-k)/(k-1)], \quad (3)$$

where A and W are the among- and within-cluster sum of squares, respectively, n is the number of objects, and k is the number of existing clusters. In many (but not all) instances, the statistic has large values early in the clustering procedure (when A is large and W is small) that decrease along with the number of clusters remaining. This tendency is partially countered by the second term in square brackets, which increases each step. The statistic is undefined in the very last step when that term becomes infinite.

The Duda and Hart (1973) criterion is a ratio of the within-cluster sum of squares computed for two clusters that are candidates for fusion at a given step both before and after the fusion. This can be converted to a pseudo-t^2 statistic as shown in the documentation for SAS (1985, p. 268). The pseudo-t^2 test tends to have small values when small and/or similar clusters are fused and larger values as cluster memberships or dissimilarities increase. Therefore, it provides information qualitatively similar to that in the other pseudotest.

These are termed "pseudotests" because they violate the usual statistical assumptions that underlie such tests, and thus we use them cautiously, in an advisory role only. We will first subjectively determine the amount of detail we wish to retain (by choosing an approximate number of clusters), and then employ the pseudostatistics to select among the available clustering levels in that particular neighborhood. SAS (1985) recommends looking for *local* peaks in the pseudo-F test that are followed by sudden jumps in the pseudo-t^2 statistic for the next (more general) clustering level. In our experience, most of the local peaks in the pseudo-F were associated with the merging of clusters with smaller than average memberships for that point in the clustering procedure that precedes a fusion involving at least one large (in membership) cluster in the next step. That subsequent merger results in a particularly large increase in W, thus forcing a sharp decrease in F and a jump in t^2. It seemed intuitively reasonable to us to look at clustering levels that precede particularly large or disparate mergers. Because of our motivation for clustering, we ascribe little importance to the *global* peaks of these statistics. Indeed, they tend to point to clustering levels that possess an inadequate amount of detail for our present purposes anyway.

4. Potential biases

One must recognize various biases that, if left unconfronted, might unduly influence or even outright determine the resulting clusterings. Three such potential biases are *methodological bias,* due to the quirks of the various clustering methods; *latent bias,* which may lurk within the objects to be clustered due to their spatial distributions; and *information bias,* owing to the repetition of information among the variables of the dataset, as well as their specific scalings. Each will be considered in turn.

a. Methodological bias

Methodological biases are now fairly well understood. Many clustering methods tend to forge clusters possessing some particular characteristic. For example, single linkage is famous (or infamous) for "chaining," which results in the production of one huge cluster that grows at each step at the expense of the remaining clusters, which often have few members. The Ward method has been found to be biased toward producing clusters with a relatively similar number of members, sacrificing cluster distinctiveness (e.g., Kalkstein et al. 1987). Such statistics as the pseudo-F and t^2 may lead to the conclusion that there are no true (or only overlapping and inseparable) clusters in the data, owing to their similarity of size and the general lack of distinctiveness among them; this did occur in our analyses

on the climate dataset. Still, Ward's method has its cadre of supporters owing to its attractive least-squares approach, and has been found to work well in some applications. The average linkage method tends to form clusters with similar variances. The Ward and average linkage methods are among the better choices in situations where the clusters embedded in the data have compact shape (in Euclidean space), even if inter-cluster separations are poor. There is a large body of work demonstrating these biases in the social sciences literature. In the atmospheric science literature, Kalkstein et al. (1987) present a particularly nice demonstration.

Of these methodological biases, that of average linkage method seems to be least worrisome to us for this particular application. However, the method does not guarantee that the roughly similar variances obtained are in fact the smallest possible variances, which would indicate optimal between-cluster differences in addition to the desired internal consistency. One of the earliest of the agglomerative hierarchical algorithms to be created, average linkage continues to be popular (used by Wolter 1987; Kalkstein et al. 1987; Schulz and Samson 1988, among others). Other methods were considered in the course of this research, but only average linkage is employed herein. The reader should keep in mind, however, that results from cluster analysis can be more revealing of methodological bias than of true clusters that may or may not be present in the data, particularly if an especially poorly behaved clustering method is adopted.

b. Latent bias

The original intent of the NCDC Climate Division dataset was to subdivide states, and the nation, into zones of roughly uniform climate. However, numerous constraints conspired to undermine this goal. The figure reveals a marked tendency toward concentrating smaller, more evenly sized divisions in the eastern and southern sections of the United States, with larger and more erratically sized divisions to the west. The divisions are also compelled to follow boundaries that often have little to do with climate, within each state and for the conterminous United States as a whole. Neighboring land areas and water bodies are not represented.

One result of this skewing is that the data domain contains several unrealistic "peninsular" areas, such as New England and Florida. The climate divisions in Maine might be more likely to cluster with locations in eastern Canada rather than those within the United States, but are prevented from doing so, thus forcing the creation of a different and less representative cluster in this area than might otherwise have been established. This situation is at least superficially analogous to the problem of artificially finite domains in numerical simulations of fluid flow. The geographic bounds of the climate dataset represent essentially "closed" boundaries across which clusters cannot pass. Obviously, short of including the entire world in the analysis, it might be better to cluster a much larger area than the specific region of interest, in order to minimize possible influences of the finite domain. This is not a certain cure because climate types need not be spatially contiguous.

The principal concern here is that the uneven ordering of unequally sized divisions and the finite and arbitrary nature of the data domain could greatly affect the results, either through predetermining where the cluster "anchors" will reside (i.e., where the agglomerative hierarchical clustering is compelled to begin) or by exerting undue influence on how the clustering proceeds. Both are important in hierarchical clustering because of its inherent constraints. The latter concern is particularly acute when a method that can exhibit bias with respect to cluster membership size, such as Ward's method, is used. In the average-linkage "reference clusterings" to be presented later clusters in the eastern United States tend to have moderate spatial size but are membership rich, owing to the small size of the divisions there. In the western states, however, a cluster is established that is spatially large but membership poor, because the divisions there tend to be large. Thus, in this application, the resolution of the climatic dataset is finer in the east than in the west. When a membership-biased method like Ward's is used instead, the clusters of the eastern United States lose members and the western cluster is forced to expand, resulting in diminished cluster distinctiveness. Thus, uneven resolution of the dataset would have some impact on the results.

The influences of the data domain's finite size, its shape, and irregular distribution of data points were tested in several ways. One, briefly described in section 6d, consisted of processing and clustering data after interpolation onto a uniform mesh. In another, several clustering methods were used to cluster physical distances, computed along great circles between division centroids. This test represents the clustering of a homogeneous field that is sampled in an irregular and finite manner. The ideal solution would consist of a set of spherical, overlapping clusters of equal areal size. The ideal solution is not obtainable owing to the hardness of the clusters and the inherent biases of the clustering methods. In this test, Ward's method displayed its tendency to forge clusters with similar membership sizes, which impeded the establishment of equal area clusters. The performance of average linkage was judged to be superior, if still imperfect. The Ward and average linkage solutions were identical in the peninsular areas, indicating that domain shape dominated methodological bias in those areas. Clusters residing in those zones will be treated with suspicion until concerns about latent bias can be eased.

c. Information bias

As discussed above, the clustering algorithm employed plays a significant role in determining the clustering outcome. For any given algorithm, however, the clustering solution may also be extremely sensitive to the particular distance measure (Euclidean or Mahalanobis, among others) adopted and how the variables are processed (if at all) prior to the computation of interobject distances. Some of those problems related to variable intercorrelations and decisions made regarding variable and object scalings were discussed in section 3b.

Several additional problems may also be recognized, relating to what will be called the "information content" of the variables. Suppose a subset of the p variables is sufficient to represent a "true" cluster structure partitioning the objects, with the remaining variables being *irrelevant,* providing no information whatsoever regarding the true clusters. Since the true cluster structure is unknown—the goal of the cluster analysis being its extraction—it is not immediately obvious which, if any, of the p variables are actually irrelevant. However, the major effect of including these irrelevant variables in the distance computations will be to mask the true cluster structure, perhaps beyond recovery. The problem of detecting and treating irrelevant variables has been addressed by De Soete (1986), among others.

We are concerned herein with essentially the *opposite* problem. Suppose a variable included in the dataset is truly relevant in that it contains information regarding a latent, but unknown, cluster structure embedded in the data. Suppose further that the information the relevant variable provides is also contained in, or spread among, one or more additional variables. Thus, the problem is one of *redundant* or *repeated information.* The inclusion of redundant information may have a serious effect on the clustering outcome and, when its inclusion is *inadvertent,* results in what is termed herein as "information bias." The motivation for our concern, and our present remedy for the bias, is outlined below.

1) THE REDUNDANCY PROBLEM

The most obvious information redundancy might be thought to exist between two (or more) variables that are very highly correlated. This implies that at least one variable has been included that contributes little *unique* information about the cluster structure embedded in the data, and instead chiefly repeats (and magnifies) information obtainable from other variables. This would not be a problem if somehow each variable were restricted to contributing only its unique information regarding the latent cluster structure to the distance computations, and if further, its information were somehow scaled relative to its actual importance. Both of these "corrections" occur in a regression analysis owing to its "extra sum of squares" principle. However, none of the distance metrics used in cluster analysis possesses such a principle, mainly because no dependent variable exists *prior* to the clustering, nor is such a variable created that is actually independent of the clustering outcome. Therefore, redundant information must be identified and eliminated from the variables prior to the computation of the distance matrix.

The clusters are groups of objects embedded within a possibly multidimensional space defined by certain distinguishing characteristics that are ultimately represented by the variables or some combinations thereof. The objects distributed in this characteristic space may tend to exhibit clustering or clumping, and the goal of the cluster analysis is to delineate those clusters. However, the clusters may not be well separated in this space for a variety of reasons. For example, perhaps some important additional dimension is missing, or perhaps the true clusters should be overlapping anyway. It is obvious that changing the variance of one characteristic dimension relative to the others distorts the embedded cluster structure, enhancing the distinction between some clusters and lessening it between others.

This is true even in the ideal situation in which the clusters are compact and well separated (in the Euclidean sense), but is most severe in the perhaps more common situation consisting of fairly poor intercluster separations. The less well separated the clusters are, the more likely it is that different clustering algorithms will generate different clustering outcomes (methodological bias). It is also more likely that the inclusion of relevant but redundant information, which exaggerates one characteristic dimension relative to the others, will also have a substantial impact on the outcome, no matter which algorithm is employed (information bias).

We may have no idea how one variable should be scaled relative to the others, so we often scale them all to equal variance and hope for the best. The idea behind equal scaling, however, is that each variable represents a *unique* and *equally important* contribution to the cluster structure. If untrue, it may indeed be advisable to assign different weights among the variables, but this should be done intentionally, not inadvertently through the inclusion of redundant variables.

For the Euclidean metric, the basic problem may be demonstrated with a simple example, recalling that a variable's average influence in determining object dissimilarity is largely determined by its relative variance. Say equally scaled variables X_1 and X_2 jointly describe a particular cluster structure. Now consider adding to the dataset another similarly scaled variable X_3 that is highly correlated with one of the original variables (say X_1). Including this variable essentially inflates the variance associated with the information provided by X_1, rendering that variable more influential in determining the clustering outcome. In the most extreme case, $X_3 = X_1$, and thus the X_3 information is completely redundant. The inclusion of X_3 in this case is equivalent

to doubling the variance—and the influence—of variable X_1, resulting in a distortion of the cluster structure in the dimension defined by that variable, a distortion that might be critical if the clusters are not truly well separated.

How might information bias be eliminated? Superficially, it might appear that there is no redundant information in a dataset consisting of mutually uncorrelated variables, such as those obtained by a PCA on the original correlated variables, but this cannot be guaranteed and depends upon the distance metric and component scaling employed and whether or not the analysis was truncated. Suppose the three variables X_1, X_2, and X_3, including the latter redundant variable, are subjected to a PCA and variance-weighted scores for the three new, uncorrelated component variables are obtained. As noted in section 3b, if all of the components are retained, the Euclidean distance matrix computed from the PCs is identical to that constructed from the original variables. Therefore, if redundant information exists in the original dataset, it *still* exists in the component dataset, despite the fact that the new component variables are uncorrelated. Therefore, *the lack of correlation is no guarantee of the absence of redundancy.*

If the true cluster space embedded in the data has fewer dimensions than the number of original variables included in the analysis, then there exists irrelevant or redundant information among the variables, and thus the dimensionality represented by the original variables must be reduced or constrained in some fashion. If PCA is employed to preprocess the variables, this means that the analysis must be *truncated.* Yet, when variance-weighted scores are used, truncation to the number of truly relevant dimensions (two, in this example) cannot guarantee that the redundant information associated with variable X_3 has been eliminated.

Say variables X_1 and X_2 are uncorrelated and have unit variance. The original variables already represent a principal components solution, so one may let $Z_1 = X_1$ and $Z_2 = X_2$. The addition of completely redundant variable X_3 to the dataset means that now three component variables can be constructed, Z_1^*, Z_2^*, and Z_3^*, but as variables X_1 and X_3 are perfectly correlated, the third eigenvalue, and thus the variance of component variable Z_3^*, is identically zero. The variance of X_3 augments that of X_1, giving the component variable Z_1^* larger variance and greater influence than that of component variable Z_2^*. If the variance of X_3 is also unity, then the variance of Z_1^* is double that of Z_2^*, and so original variable X_1 is still effectively and inadvertently given double the weight. This analysis could be truncated to the proper number of dimensions, two, but doing so has no effect in this example because the third component, having zero variance, does not contribute to the interobject distances anyway.

The utilization of untruncated standardized component scores does not solve this problem either and indeed could make matters even worse. Suppose X_3 is very highly, but not perfectly, correlated with X_1, and the difference between them represents only random measurement error that is irrelevant to the embedded cluster structure. The additional eigenvalue created when X_3 is added to the dataset is very small, but not identically zero, and largely represents the random residual. Standardized scaling of the components treats all component variables equally regardless of the variance of the original variables they explain and as a result exaggerates the influence of a component variable that in this case represents only random, useless noise.

Therefore, this strategy requires truncation as well, but the choice of the truncation point in a standardized PCA approach is even more problematic. As noted earlier, many applications result in the bulk of the original variance being concentrated into a relatively small number of components. When the component scores are variance weighted, exclusion of the smaller components may—in an ideal case—have relatively little effect on the Euclidean interobject distances, but this is something that should be confirmed (see next subsection). Thus, in an ideal case, the Euclidean distance matrix may converge rapidly toward some stable, final configuration (given by the distances computed using the full set of original variables) as additional, but smaller, components are retained, which would make the choice among different truncation levels less onerous. However, we argued above that if the original variables contain redundant information, then the solution being converged upon is not the desired one.

In contrast, when standardized scores are used, there is little hope of rapid convergence toward the final configuration, which in this case is the Mahalanobis distance matrix of the original variables (and still not necessarily the desired solution). Changing the truncation level can have a first-order effect on the distance matrix generated, and thus the clustering outcome. In addition, there may be no single obvious truncation level. There are many available strategies (stopping rules for PCA) for determining truncation points in PCA, but these may not result in a satisfying consensus (see Jackson 1991, p. 56, for an example). Indeed, as many are based on the relative sizes of the eigenvalues, they will be influenced to a perhaps significant degree by the presence of redundant variance, especially if it chiefly contributes to the larger variance components. We consider this a very serious deficiency of using standardized scores.

2) THE PRESENT SOLUTION

This study employs 24 temperature and precipitation variables. Information bias owing to redundancy is a concern as some of the correlations are extremely high. Further, the 12 temperature variables tend to be more intercorrelated than the precipitation variables. Of the 66 possible pairs of the 12 temperature variables, 25

pairs have correlations ≥ 0.96, with the highest correlation (between December and January) being 0.9973. Of the 66 pairs of precipitation variables, however, only five correlations exceed 0.96. If all the original variables are used, temperature as a data type would be more influential, even if the two types are similarly scaled.

We do not presently have a satisfying solution to the redundancy problem. The procedure adopted in this study utilizes truncated PCA on the original variables using standardized component scores. The adoption of this procedure was motivated by the following example, which uses a subset of our climate dataset for simplicity of presentation, and represents a practical demonstration of the simple example used above. However, we do feel that other approaches should be considered in the future, especially because of the truncation problem.

Two moderately correlated ($r = 0.48$) variables, January temperature and precipitation (denoted T1 and R1), were selected. For this example, the variables were each standardized to unit variance. A plot of the two variables (not shown) evinces clumping, but few very well separated clusters, so this would be expected to be a stressful test. The variable correlation matrix has eigenvalues of 1.48 and 0.52. Varimax orthogonal rotation of the component eigenvectors generated two new component variables, each accounting for one unit of variance. The first PC (PC1) loaded strongly on temperature (loading 0.97) and weakly on precipitation (loading 0.25); the second PC's eigenvector loadings were opposite. These components might be termed temperature and precipitation dimensions, respectively.

The Euclidean metric was then used to create the object distance matrix from the standardized scores. The secondary rotation was actually unnecessary, and had no effect on the distances. Indeed, the same distance matrix could have been constructed by applying the Mahalanobis metric on the original variables. Figure 2a shows the result of an average linkage clustering of these standardized scores. The pseudostatistics discussed earlier suggested the 11-cluster solution shown, among other possibilities. This is taken to be the "correct" solution.

Next, we attempted to bias or "contaminate" the outcome toward the temperature dimension by including a third variable, standardized February temperature (T2), which is very highly correlated ($r = 0.991$) with T1. The difference between these two variables is small but spatially systematic, and demonstrates the effect of allowing *clearly redundant* information in the dataset. If the available distance measures possessed an extra sum-of-squares principle, the effect of this new variable on the clustering outcome might in fact be quite minimal.

The eigenvalues of the three-variable correlation matrix were 2.3128, 0.6793, and 0.008. The additional eigenvalue created as a result of T2's inclusion was extremely small, owing to the very high correlation between T1 and T2. The third component was dropped prior to rotation; the untruncated case will be revisited later. The truncation level is obvious to us only because of the experimental design, and it is possible some truncation test, responding to the redundant variance associated with the first component, would recommend deletion of the *second* component as well. After rotation, variable T2 folded primarily into the first PC along with January temperature T1. Rotated loadings for the first retained PC were 0.96, 0.24, and 0.98 for T1, R1, and T2, respectively. Similarly, PC2's loadings were 0.26, 0.97, and 0.21.

Average linkage clustering of the standardized scores from the contaminated but truncated PCA proceeded in a manner very similar to that of the correct solution. Figure 2b identifies the discrepancies between the correct and contaminated clusterings at the 11-cluster level and shows that 12 of the 344 climate divisions were differently assigned. Since these were located primarily at the boundaries of the clusters, which should not be considered truly "hard" anyway, agreement should be considered very good. Note that the contaminated solution did incorporate some additional information that was not available to the "correct" model. Still, because of the very high correlation between T1 and T2, the amount of extra information must be quite small.[3]

The experiment was repeated, using the original variables. The Euclidean distance matrix was computed using original correlated variables T1 and R1. Recall that the identical matrix could be constructed from an untruncated PCA if the component variable scores are variance weighted. The average linkage clustering is shown in Fig. 3a. Pseudostatistics suggested a 10-cluster solution. The result is clearly very different from that based on the standardized scores in Fig. 2a (54 discrepancies between the two most comparable solutions). This demonstrates the sensitivity to the choice of the distance metric. Again, if the Mahalanobis metric had been applied to the original variables, the result would have been identical to the "correct" solution.

Variable T2 was then added to the analysis. Average linkage of the Euclidean distances computed from the three raw variables (or, equivalently, the three variance-weighted component variables from an untruncated PCA) resulted in a radically different clustering from that shown in Fig. 3a. Cluster shapes and sizes were

[3] A still more stressful test would be to include a redundant variable that differs from T1 by a truly spatially random component. This was found to result in a larger number of discrepancies, even when the present approach of truncated, standardized component analysis is adopted, indicating a shortcoming of the present approach. However, it is felt that the example discussed is more applicable to the problems confronted herein.

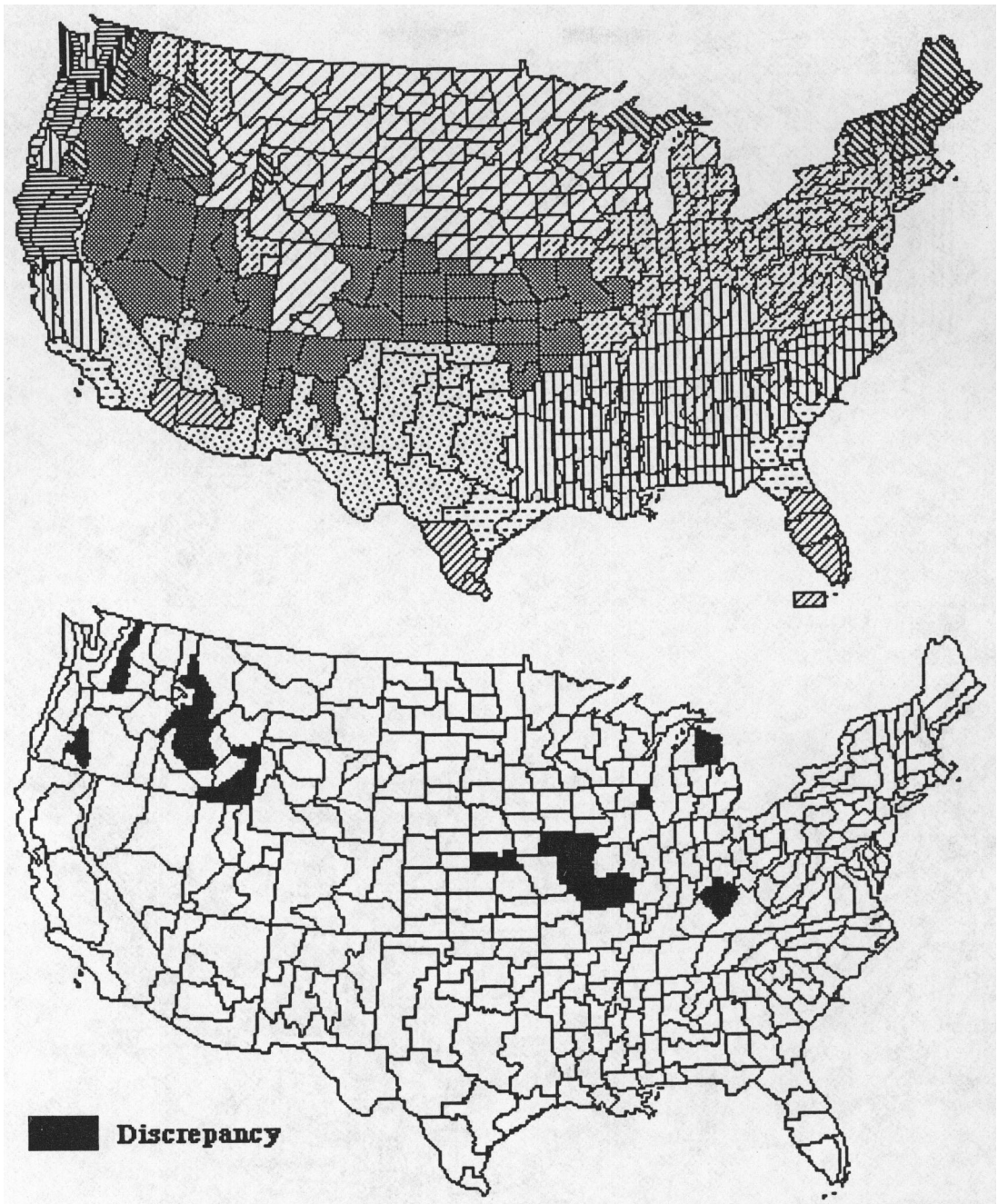

FIG. 2. Results from the test of information bias using a subset of the climate dataset and preprocessed with principal components analysis. (a) The "correct solution" at the 11-cluster level; (b) discrepancies (shown in black) between the correct and contaminated clusterings at the 11-cluster level.

substantially changed and shifted in space as well. This makes sense because the effect of including T2 was essentially to exaggerate the variance and influence of variable T1 relative to R1 in determining object dissimilarity. The discrepancy map for the two most similar raw variable clusterings is shown in Fig. 3b. A total of 193 divisions, principally located in the Northeast and Midwest, were found to have changed cluster memberships as a direct result of the addition of the redundant variable (Fig. 3b).

The third component variable created from the inclusion of T2 in the analysis possessed very small relative variance (0.008) relative to the other two components. It might be expected, then, that the deletion of this component would have relatively little effect on the result. To test this assumption, another clustering

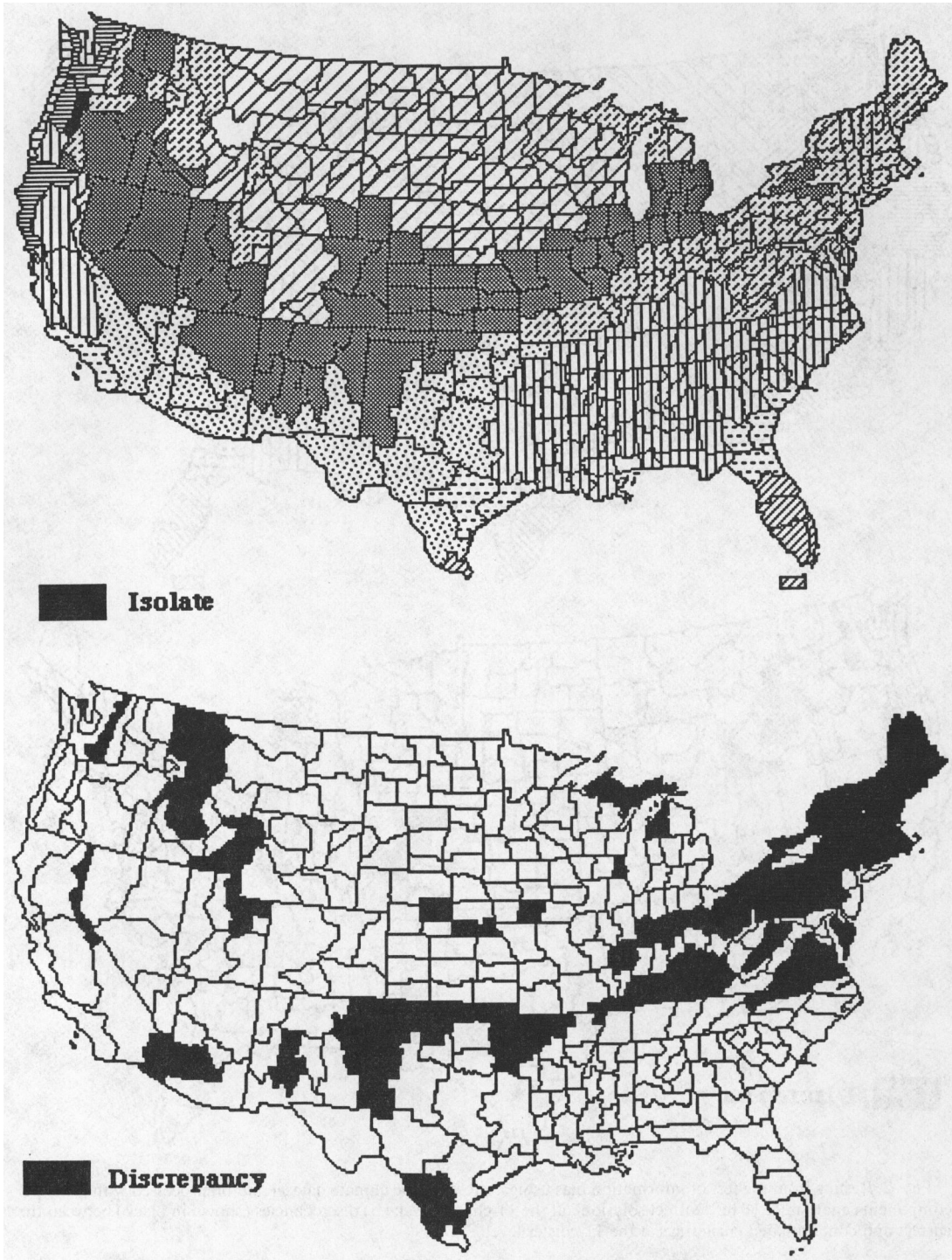

FIG. 3. As in Fig. 2 but when raw data are used with Euclidean distances. (a) The reference raw data solution with 10 clusters; (b) discrepancies (shown in black) between this reference solution the most comparable contaminated clustering, at the 11-cluster level.

was generated using a Euclidean distance matrix constructed with only the first two variance-weighted component variables. This represents an application of PCA preprocessing for the generation of variance-weighted scores with truncation to the expected number of dimensions of the cluster space. This clustering was compared to that shown in Fig. 3a. A total of 139 discrepancies were found (not shown) between the two

most comparable solutions, despite the small variance of the deleted component variable. This indicates that the convergence discussed earlier may not be as rapid as expected, especially if the number of original variables is rather limited.

Finally, we compare the clustering outcomes between the two- and three-variable datasets when the Mahalanobis metric is employed. This method is the same as comparing two untruncated PCA processed datasets when standardized scaling is adopted; that is, the "correct" solution with what would have been obtained if the third component had not been deleted prior to the clustering that generated the discrepancy map shown in Fig. 3b. These were the least comparable solutions of all, being so different that no meaningful discrepancy map could be constructed. Thus, as expected, truncation is even a more severe problem when the scores are standardized. In this example, the difference between variables T1 and T2 was considered to be irrelevant information or noise, and employing the Mahalanobis metric effectively elevated this noise component to have equal influence with the other, important components.

If the dataset is suspected to contain redundant information, then the conclusion drawn from this experiment is that truncated PCA preprocessing followed by standardization of the component scores did the best job of removing the redundancy of the preprocessing methods considered herein. This is the strategy adopted for the reference clustering (section 5).

3) THE DANGER POSED BY INDISCRIMINATE TRUNCATION

In the foregoing, it was assumed that the originally identified variables mainly contain relevant, if repeated, information. If any truly irrelevant information exists (such as the difference between variables T1 and T2, which was considered noise for the purposes of the above experiment), then PCA will channel it into the small variance components likely to be deleted prior to the construction of the distance matrix. However, if the dataset contains truly irrelevant variables, the problem considered by De Soete (1986), and if further the "signal variance" (information about the true cluster structure) is exceeded by the "noise variance" (contributed by the irrelevant information), then it may indeed be the *signal* that passes into the small components routinely marked for removal. Chang (1983) presented an example where this did occur, and suggested making pairwise plots of the component scores to subjectively verify that the deleted components do not contain any distinctive clusters that might be indicative of the true latent cluster structure.

5. Results using the reference clustering procedure

In this section, the results of the "reference clustering procedure," representing the analysis strategy presently believed to be best, will be presented and analyzed. In the following section, alternative clusterings that are obtained when different choices or decisions are made in the handling of the dataset will be discussed to assess sensitivity to biases and the robustness of the clusters identified in the reference solutions.

a. Data issues

In most of the cluster analyses to follow, one climate division suspected of being a particularly bad data point (in the northwest corner of South Carolina) was deleted from our dataset prior to data preprocessing (if any) and clustering,[4] leaving 343 objects in the dataset. For our reference clustering solutions, the variables were standardized by *data type.* The grand mean and variance of 12 temperature variables taken together were used to transform these variables. The same was done separately to the 12 precipitation variables. This eliminates the scale difference between the temperature and precipitation measures but preserves the natural variability through the year that is part of the seasonal cycle and thought to be valuable information for the clustering.

b. The PC analysis

For the reference strategy, the variables are subsequently processed using PCA on the variable covariance matrix. PCA was chosen instead of common factor analysis because the primary goal is to reorder the information among the variables; we do not wish to assume or extract causal factor structure. The total variance of the 24 variables equals 14.6. The variances of the precipitation variables definitely exceed those of the temperature variables (see Table 1), so using covariances instead of correlations in the PCA gives the former greater weight. It is reported later (in section 6d), however, that this decision had little practical effect on the clustering outcome, partly because we standardized the component variables we retained to unit variance prior to generating the distance matrix.

The first three components, having eigenvalues of 8.75, 3.01, and 1.92, account for 94% of the total variance (13.7/14.6). The next two eigenvalues are much smaller (0.35 and 0.33), and found to be statistically indistinct, which means they should be retained or excluded as a set (North et al. 1982). In this case, the truncation level might be thought to be fairly obvious, but since the choice of the truncation level has such a significant effect on the clustering outcome, especially when standardized scaling for the variables is adopted, a number of different truncation tests were applied.

[4] The influence of this division on the results of the reference clusterings is discussed in section 6c. Deletion of this division had no impact on the T1–R1–T2 example in section 4, so it was retained in those clusterings.

TABLE 1. PCA analysis for the reference clustering strategy. Total variable variance: 14.61. Variance retained: 13.67 (93.6%).

		Rotated loadings			Fraction of variance			
Variable	Variance	PC1	PC2	PC3	PC1	PC2	PC3	Total
T1	0.48685	0.64481	0.23691	0.01992	0.85402	0.11529	0.00082	0.97012
T2	0.45512	0.63604	0.19909	−0.02703	0.88887	0.08709	0.00161	0.97757
T3	0.35402	0.57115	0.15211	0.04066	0.92145	0.06535	0.00467	0.99147
T4	0.23927	0.46491	0.07646	0.10611	0.90334	0.02443	0.04705	0.97483
T5	0.17279	0.37791	0.02991	0.13921	0.8265	0.00518	0.11216	0.94383
T6	0.15258	0.33429	−0.01264	0.14909	0.73241	0.00105	0.14568	0.87914
T7	0.10838	0.27876	−0.0494	0.08921	0.71695	0.02252	0.07342	0.81289
T8	0.11942	0.3029	−0.0273	0.09622	0.76832	0.00624	0.07752	0.85209
T9	0.16504	0.37974	0.03573	0.10606	0.87372	0.00774	0.06816	0.94961
T10	0.1817	0.40489	0.05967	0.09732	0.90225	0.01959	0.05213	0.97397
T11	0.26354	0.4768	0.14549	0.09486	0.86261	0.08032	0.03414	0.97707
T12	0.38148	0.57624	0.19736	0.03395	0.87043	0.10211	0.00302	0.97556
R1	1.47661	0.19506	1.19117	0.06845	0.02577	0.96091	0.00317	0.98985
R2	1.07957	0.26891	0.98223	0.14127	0.06698	0.89367	0.01849	0.97913
R3	1.20887	0.27921	0.96033	0.37052	0.06449	0.7629	0.11356	0.94095
R4	0.68562	0.14666	0.57858	0.49408	0.03137	0.48825	0.35605	0.87567
R5	0.6259	0.13548	0.25725	0.64273	0.02933	0.10574	0.66001	0.79508
R6	0.72143	0.07415	0.07914	0.80779	0.00762	0.00868	0.90448	0.92078
R7	1.05148	0.33808	0.18449	0.89922	0.1087	0.03237	0.76901	0.91008
R8	0.80916	0.25437	0.11444	0.80731	0.07996	0.01618	0.80545	0.9016
R9	0.70455	0.26748	0.23701	0.70826	0.10155	0.07973	0.712	0.89328
R10	0.4887	0.05332	0.50862	0.32928	0.00582	0.52936	0.22187	0.75705
R11	1.043	0.00582	0.98397	0.17549	0.00003	0.92827	0.02953	0.95783
R12	1.6353	0.14403	1.26228	0.05794	0.01269	0.97435	0.00205	0.98909
Variance explained	3.16	6.83	3.68					
Fraction of total	0.22	0.47	0.25					

Column 1: letter symbols "T" and "R" refer to temperature and precipitation, respectively, and number indicates month of the year. Variables were transformed by data type to zero mean and unit variance. Column 2: variance of each individual variable. Columns 3–5: Varimax rotated loadings for three retained PCs. Columns 6–8: fraction of original variable variance accounted for by the PCs. Column 9: fraction of original variable variance accounted within the three retained PCs.

Recall from section 4b, however, that any test based on eigenvalue size is prone to being influenced by the presence of redundant variance.

The imperfect but widely used average root test suggests retaining three components. Jolliffe (1972) recommended using 70% of the average root, but this does not alter the conclusion. An asymptotic application of "Rule N" (Preisendorfer 1988) also suggests keeping three roots. This recommendation is not changed even when the effective number of objects is halved to compensate for considerable spatial autocorrelation among the objects. However, our data are nonnormal, which degrades the test (Preisendorfer 1988). The "scree test" consists of plotting the eigenvalue magnitude versus rank and looking for natural breaks or "elbows" in the plot. Using Cattell and Jasper's suggested interpretation (see Jackson 1991, p. 45), five or six components might be retained, but this would also include the poorly separated fourth and fifth PCs. Velicer's (1976) test suggests that seven components be retained, which seems excessive.

For the reference procedure, the three most significant components were retained. Varimax orthogonal rotation was then applied to enhance interpretation, but had no effect on the interobject Euclidean distances. The rotated component loadings are reported in Table 1 (the ordering is truly unimportant because standardized scores were generated anyway). The component called PC1 captures most of the original variance of the temperature variables, and accounts for 22% of the retained variance. The other two components together explaining the bulk of the retained variance (47% and 25%, respectively) loaded highly on precipitation, with PC2 representing the seven months encompassing the winter season and PC3 capturing the remaining five months including summer. Standardized component scores were then calculated for the three rotated components, and the spatial distributions of these scores are presented in Fig. 4. Thus, within the context of the discussion of section 4c, we are searching for clusters embedded within an equally scaled three-dimensional space defined by the components: one influenced by temperature (PC1), one dominated by cool season precipitation (PC2), and one largely constructed from warm season precipitation (PC3).

Alternatively, the five largest components, and thus 98% of the original variance, could have been retained, creating a five-dimensional cluster space. However, after rotation, the fourth and fifth components possessed only one sizable loading each (both in precipitation,

in May and October, respectively), and represented only 2.8% and 2.3% of the original variance. Retention of two additional components prior to rotation resulted in only small changes in the composition of the three largest components.

Owing to the concern expressed in section 4c(3), pairwise plots of all 24 PCs from untruncated PCAs (with and without rotation) were examined to see if there was any obvious information contributed by the smaller variance components marked for deletion. The majority of the plots involving deleted components revealed no convincing clustering or clumping. Plots pairing the largest variance components (shown later) did show such clumping, but as expected, suggested that most of the embedded clusters are not particularly well separated. A case might be made for retaining additional components, but the decision is unfortunately subjective. Still, the clustering outcome obtained with the first five components is discussed in section 6b.

c. *The reference clustering solutions*

Average linkage was applied to the distance matrix constructed from the three standardized component variables. It was decided to look for solutions with approximately a half-dozen, a dozen, and two dozen clusters, as subjectively representing three different, but potentially useful, levels of detail. The local peaks in the pseudo-F and t^2 statistics, shown in Fig. 5, were used to advise on specific clustering levels in those neighborhoods. The 25-, 14-, and 8-cluster solutions were chosen. Based on these statistics, other clustering levels could have been picked. The 25-cluster level was chosen over the 22-cluster solution, for example, since it retains greater detail in the western United States. The peaks in the pseudostatistics at 8 clusters were especially prominent as the next step involved the fusion of two very large clusters.

Below, the 14- and 25-cluster solutions are presented, and the 8-cluster regionalization is briefly discussed. The 14-cluster solution is shown first, not because it is superior to the other two, but because by itself it represents a compromise between the greater generality of the 8-cluster solution and the more complex detail in the 25-cluster result.

1) THE 14-CLUSTER SOLUTION

The solution at the 14-cluster level is shown in Fig. 6 and cluster statistics for the year and 4 three-month "seasons"[5] are presented in Table 2. At this level, the eastern United States is composed principally of three spatially extensive clusters. One cluster claims the bulk of the Southeast (except most of Florida), another ranges zonally across the midsection of the country (the East Central group), and the third extends along the northern section (the Northeastern Tier cluster). These are labeled clusters 1, 2, and 3 on the table; there is no special significance to the ordering. The boundary between the East Central and Southeast clusters is nearly east–west and roughly follows the 36° latitude circle. They differ chiefly in temperature characteristics throughout the year, and the Southeast is also considerably wetter in the winter months.

Later discussion will demonstrate that the region occupied by the East Central and Northeastern Tier clusters is the site of the least robust clusters in our analysis. As they stand, the two clusters differ chiefly in temperature throughout the year (Table 2), although annual precipitation totals and the amplitude of the seasonal cycle in precipitation (larger in the drier divisions of the tier) also differ. The tier cluster is spatially noncontiguous; part of this is due to missing data in southeastern Canada, the inclusion of which would be needed for one test for latent bias in these clusters (section 4b). Owing to the nature of hierarchical cluster analysis, inclusion of Canadian data would likely have a substantial influence on how the tier cluster grows, and thus ultimately on the clustering solution as a whole.

Two climate divisions have joined the tier, despite being completely surrounded by divisions belonging elsewhere, probably due to local topographic effects. One is in New York, in the Adirondack Mountains, and the other is in western South Dakota, encompassing the Black Hills around Rapid City. The Adirondack division represents a local maximum in surface elevation, and is cooler than its surroundings, while the Black Hills division is wetter than its immediate neighbors. There is no reason to expect that clusters must be spatially contiguous. If that characteristic is required, then the clustering procedure must be constrained in some fashion (Fovell and Fovell 1993).

The western portion of the United States is dominated by one spatially extensive cluster (#4), whose peripheral members are located in Nevada, eastern Oregon and Washington, the western plains, and northern Arizona and New Mexico. This Interior West cluster would likely have spread northward into Canada had data been available there. Table 2 shows that, in the mean, this cluster tends to be dry throughout the year and also does not experience a wide seasonal swing in precipitation like those clusters to the east. In the next subsection, the reason for the latter, perhaps counterintuitive, characteristic is examined.

Within this giant cluster, there are members of a spatially discontinuous group (cluster #10) bringing together much of Idaho with western Wyoming, portions of Oregon and Washington (on the eastern slopes of the Cascades), and extreme northeastern California (east of the Cascades and Sierra Nevada mountains). Most of these divisions are characterized by having

[5] These data are provided for comparison purposes, but keep in mind that the distance matrix was not computed from them.

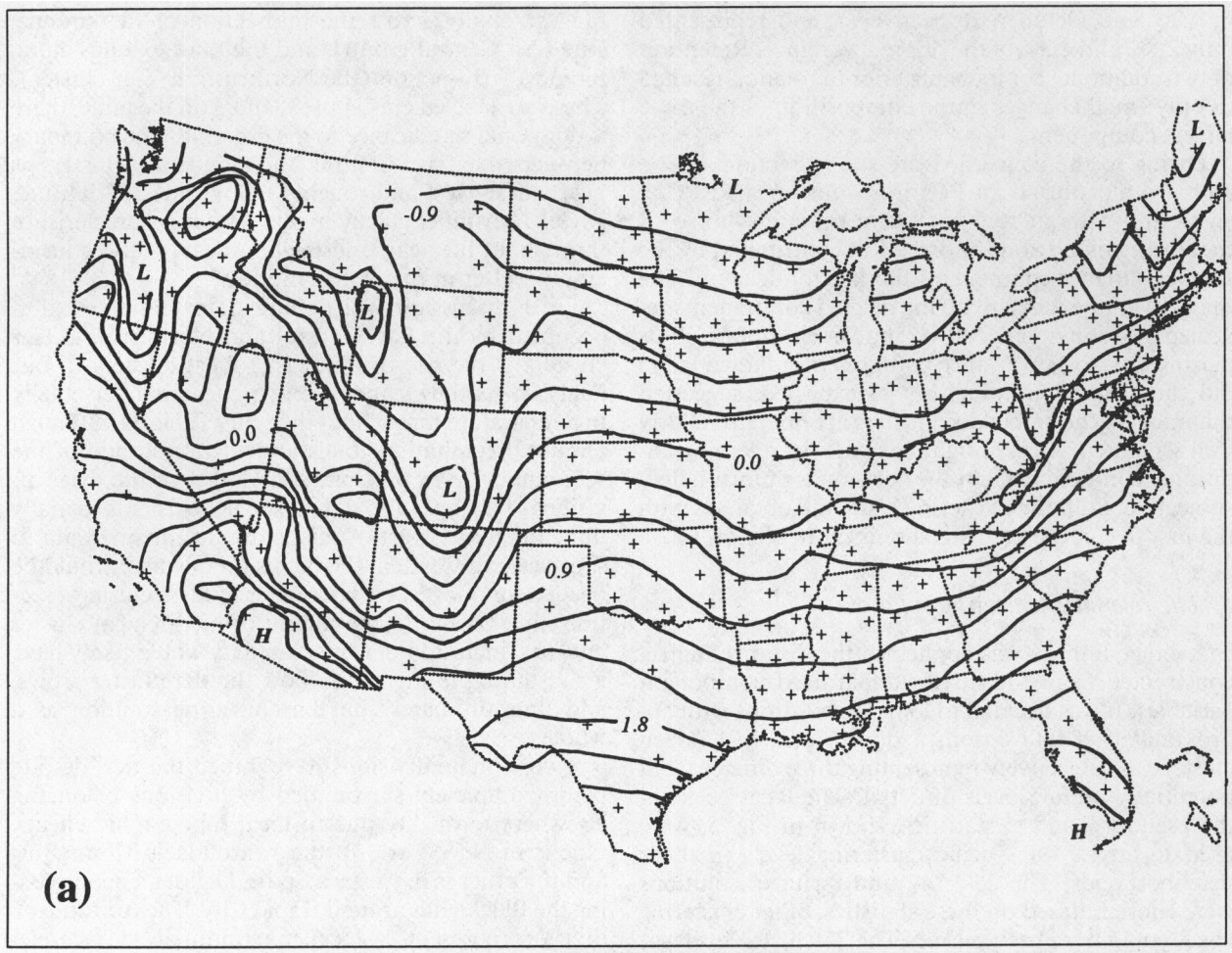

FIG. 4. Spatial distribution of the scores of the three retained components in the reference PC analysis, after Varimax orthogonal rotation. (a) PC1 scores; (b) PC2 scores; (c) PC3 scores. Contour interval = 0.3 in (a) and (b), 0.5 in (c). Locations of division centroids are indicated by plus marks.

higher average elevations than the members of the Interior West cluster. Generally, members of this high-elevation West cluster are wetter, particularly in winter, than the Interior West as a whole. The member in northeastern California seems to have the least in common with the other members of this group and may have joined only because it resides in a sharp winter season transition zone between the much wetter divisions closer to the coast (like the Sacramento valley) and the much drier divisions to the east (in Nevada).

Cluster #5 comprises much of Texas and western Oklahoma. South of the Interior West cluster is a group composed of the divisions in the desert Southwest (cluster #6), stretching parallel to the Mexican border from eastern California to western Texas. Most of Florida resides in cluster #7, except for the Florida Keys, which has joined cluster #8, the other members of which are located in southern Texas. The remainder of California is divided into southern (#9) and northern (#11) sections at the latitude of San Francisco. The subdivision of California is not particularly satisfying, even at more detailed clustering levels. This is due to the poor resolution of the NCDC dataset in this region, which fails to separate the Central Valley from the western slopes of the Sierra Nevada mountains.

Better resolution exists in the Pacific Northwest where the smallest clusters, in terms of areal extent and cluster membership size, are found. The northern California cluster also includes one division in the Pacific Northwest, the Puget Sound lowlands division in which Seattle, Washington, resides. Cluster #12 joins two divisions in Washington, one on the coast and the other along the west face of the Cascade range, both of which receive considerable winter precipitation. Cluster #14 in Oregon mimics this cluster. Cluster #13 joins the Washington and Oregon divisions in the lowlands west of the Cascades, which are somewhat drier. A division located in the northeast tip of the Olympic peninsula in Washington, which is in the winter rain shadow of the Olympic range, joins with the Interior West divi-

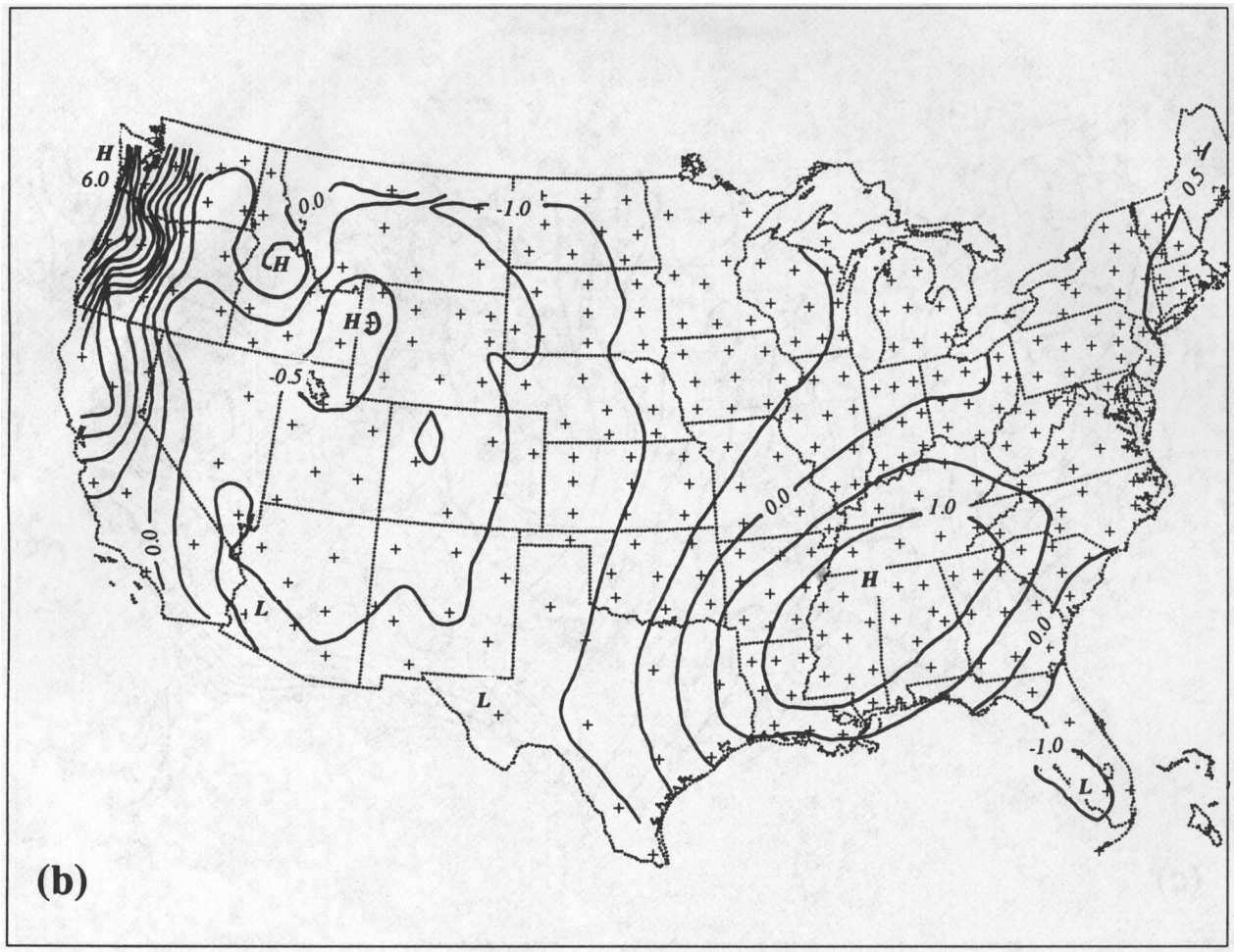

FIG. 4. (*Continued*)

sions in cluster #4. As it stands, this region has claimed a disproportionate share of the clusters at this clustering level. If resolution were improved in other areas, especially the Interior West region, the solution shown would likely look rather different at this same level.

2) THE 25-CLUSTER SOLUTION

In the discussion of the more detailed 25-cluster solution, shown in Fig. 7, reference will be made to the 14-cluster solution, as the former can be considered as depicting the subclusters contained within the latter. The first six clusters listed in Table 2 are subdivided. In addition, two isolated (single member) clusters, in western Wyoming and the Florida Keys, exist at this clustering level. The Southeast and giant Interior West clusters are further partitioned into four and three subsections, respectively. Two of the Southeast's subzones reside along the coast. The members of the Gulf and Atlantic Coast subzones are characterized by being warmer and wetter in the summer than the other members of the Southeast cluster, and the two subzones merge at the 22-cluster level. The Atlantic subzone consists of only three divisions, but these tend to be drier than their Gulf counterparts, particularly in the cooler months.

The remainder of the Southeast cluster has also been separated into two subzones. One occupies the central portion of the original cluster. The other is a spatially discontinuous zone residing on the western and eastern flanks of the central subzone, and consists of divisions in eastern Texas and Oklahoma as well as several in Georgia and the Carolinas. The distinction between these two subzones is in precipitation. The members of the central subzone receive more rainfall throughout the year, especially in the winter and spring seasons. The two subzones together, however, receive substantially less precipitation in summer than the members of the Gulf and Atlantic subzones. The two subzones merge at the 21-cluster level, and join with the coastal subzones, forming the Southeast cluster, at the 15th step.

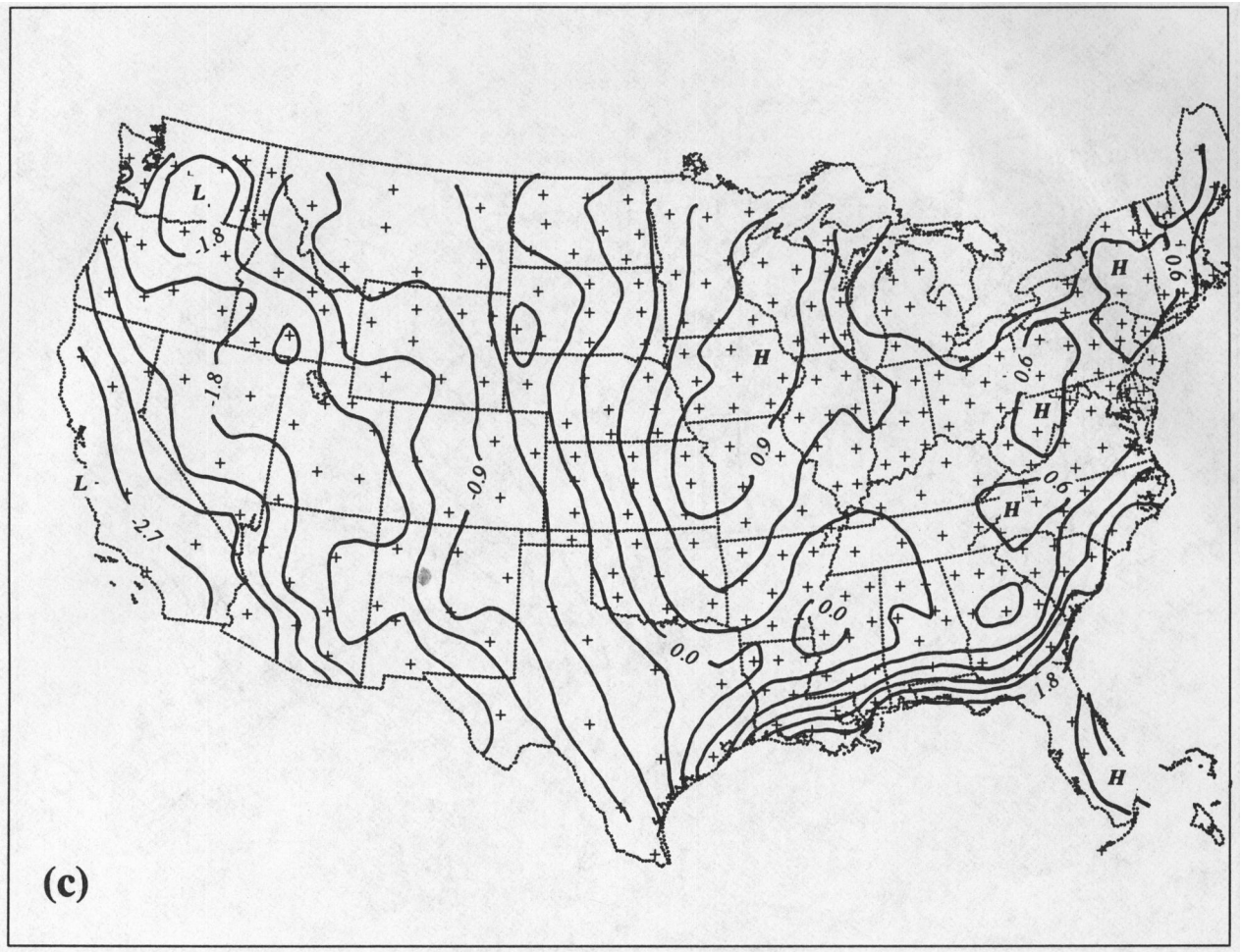

FIG. 4. (*Continued*)

One of the difficulties in using cluster analysis is that cluster fusions can create new groups that, due to averaging, fail to carry forward some characteristics of their constituent members. This is part of the inevitable loss of information or detail. The consequences of a particular compromise between detail and generality is particularly well illustrated by the Interior West cluster in the 14-cluster solution. The statistics in Table 2 show that this cluster has but a very small amplitude annual cycle in precipitation, but this is misleading. At the 25-cluster level, the Interior West region is subdivided into three subclusters, each roughly aligned north–south, which partitions the cluster into western, central, and eastern subzones. The eastern and central subzones merge in the next (24th) step, and the western portion joins them at the 17th step.

Statistics for these subclusters are presented in Table 3. The western subzone has a relatively wet winter and dry summer, but the eastern subzone has the opposite tendency. After their merger, these two tendencies substantially cancel each other. Still, the members of the Interior West cluster are more similar to each other than they are to members of other, distant clusters, demonstrating that despite the loss of information, this particular grouping is justifiable at the less detailed level the 14-cluster solution represents. Note that Table 3 shows that the divisions in the central subzone have, on average, fairly uniform precipitation through the year. Thus, this subgroup represents a transition zone between the other two subzones, and is most similar to the Interior West cluster as a whole.

In the 25-cluster solution, both the East Central and Northeastern Tier clusters have been divided into western and eastern subzones. The subzones of the East Central cluster join at the 20th step, and the fusion forming the Northeastern Tier occurs in the next step. Statistics for these subzones are presented in Table 4; note that the eastern subzone of the tier has only six members. The chief difference between the western and eastern sections in both clusters is in cool season precipitation, which is smaller in the west. In the reference clustering, the subzone pairs later fuse owing to their similarity with respect to PC1. As noted above, however, the East Central and tier clusters are the least

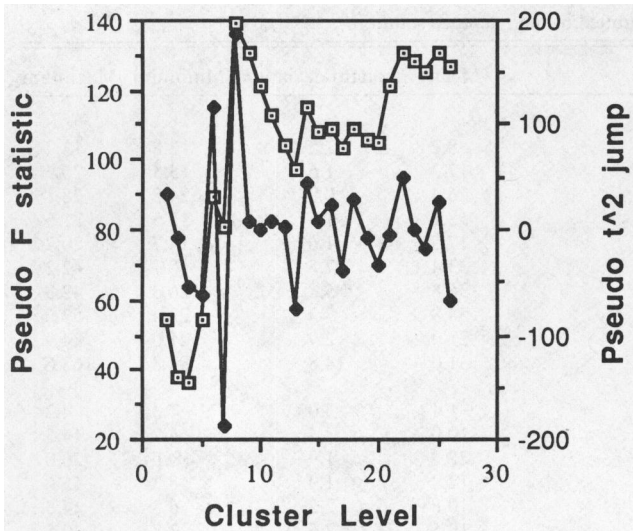

FIG. 5. Graph of pseudo-F (indicated by squares) and t^2 (indicated by diamonds) statistics each step near the end of the clustering procedure for the reference clustering. The t^2 value plotted is the change in the value of this statistic between steps.

robust of those formed in the 14-cluster solution. In some of the variant clustering solutions, including two to be described below (in sections 6b and 6c), the western and eastern zones of each cluster fuse instead, indicating that cool season precipitation information (i.e., PC2) held sway in those instances. This results in a radically different partitioning of the Northeast quadrant at a less detailed clustering level.

3) THE 8-CLUSTER SOLUTION

In the steps between the 14- and 8-cluster solutions (not shown), the Northeastern Tier cluster joins the East Central group, the situation in the Pacific Northwest simplifies with the merging of clusters #11 and #13 and also #12 with #14, southern California joins with the desert Southwest, the central and southern Texas clusters fuse, and cluster #10 joins the Interior West cluster, which surrounds it. Subjectively, it is felt that this solution sacrifices too much detail, but serves to illustrate how the clustering evolves further toward the terminal single-cluster solution.

d. Distinctiveness of the clusters in the 14-cluster solution

Clusters have been generated from the data, but how truly distinct and different are they? The less distinct they are, the more likely that they would fail to survive perturbations in the analysis procedure. Despite the nonrandomness of the cluster selection process, t tests were applied to differences of component score means for pairs of clusters (with more than two members) formed in the 14-cluster solution. These tests indicated that each cluster was statistically different at the 99% level for at least one of the three component variables.

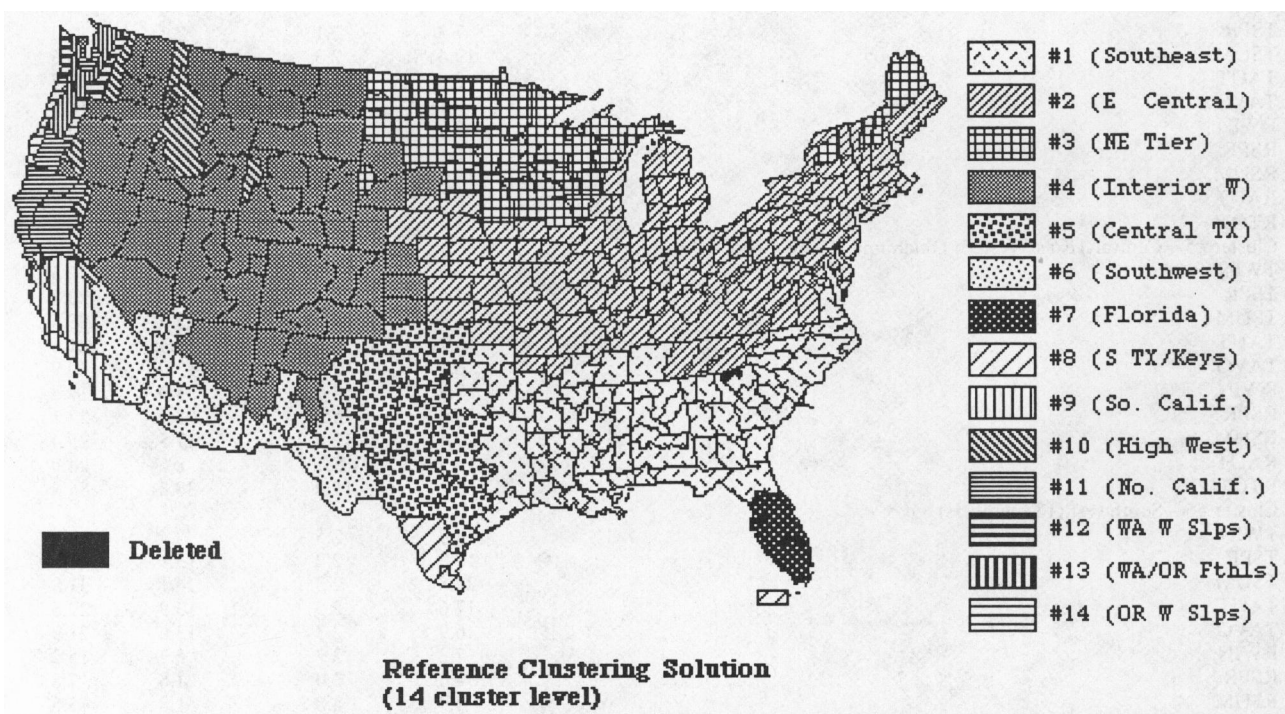

FIG. 6. The reference clustering solution of the 343 climate divisions in the conterminous United States at the 14-cluster level. (A division located in South Carolina was deleted prior to the analysis; see text.) Legend is keyed to Table 2.

TABLE 2. Statistics for clusters identified in the reference solution.

Variable	Mean	Standard deviation	Minimum	Maximum
Cluster #1—Southeast (67 members)				
TWIN	8.2	2.5	3.8	13.7
TSPR	17.3	1.6	13.8	20.5
TSUM	26.4	0.9	24.5	28.1
TAUT	18.1	1.5	15.5	21.6
TAVG	17.5	1.6	14.7	20.7
RWIN	33.4	7.5	10.7	42.2
RSPR	35.8	5.2	26.3	42.9
RSUM	34.8	7.3	23.4	53.6
RAUT	27.0	2.9	21.0	34.9
RTOT	131.0	14.8	84.8	163.6
Cluster #2—East Central (129 members)				
TWIN	−1.4	3.0	−7.3	4.2
TSPR	10.0	2.4	4.7	14.7
TSUM	22.3	1.9	18.1	26.0
TAUT	12.2	1.9	8.3	15.6
TAVG	10.8	2.2	6.3	14.8
RWIN	19.5	7.6	3.8	40.5
RSPR	26.9	4.6	16.9	38.5
RSUM	29.3	3.3	21.2	38.8
RAUT	23.1	4.5	8.9	31.7
RTOT	98.9	17.3	51.0	141.1
Cluster #3—Northeastern Tier (37 members)				
TWIN	−9.5	2.2	−14.0	−3.8
TSPR	5.3	1.5	3.1	7.9
TSUM	19.7	1.5	17.0	22.0
TAUT	7.6	1.3	5.5	9.9
TAVG	5.7	1.4	3.5	8.1
RWIN	7.9	5.7	3.0	24.5
RSPR	17.3	4.5	10.0	25.8
RSUM	26.3	4.6	18.7	32.6
RAUT	15.4	6.6	6.9	29.2
RTOT	67.0	19.6	39.5	109.1
Cluster #4—Interior West (59 members)				
TWIN	−2.9	2.9	−9.6	4.7
TSPR	7.3	2.1	2.1	11.0
TSUM	19.4	2.3	14.2	24.5
TAUT	8.8	1.9	4.3	12.7
TAVG	8.1	2.0	3.3	12.0
RWIN	7.6	4.9	2.4	23.4
RSPR	10.3	3.5	3.8	17.1
RSUM	10.6	5.1	3.3	22.0
RAUT	8.0	2.5	4.3	16.7
RTOT	36.4	9.9	18.8	59.3
Cluster #5—Central Texas/Western Oklahoma (12 members)				
TWIN	5.2	3.5	1.3	12.5
TSPR	15.6	2.7	12.3	20.8
TSUM	26.8	1.5	23.5	28.6
TAUT	16.7	2.5	13.4	21.6
TAVG	16.1	2.5	13.0	20.9
RWIN	7.9	4.6	3.0	17.2
RSPR	18.6	5.3	8.7	27.1
RSUM	20.4	2.4	17.5	25.4
RAUT	15.3	5.1	8.6	24.9
RTOT	62.2	15.0	38.8	86.8
Cluster #6—Southwest (13 members)				
TWIN	7.3	2.5	3.4	12.3
TSPR	15.9	2.2	13.1	20.6
TSUM	26.8	2.3	24.1	31.8
TAUT	17.6	2.5	13.9	22.8
TAVG	16.9	2.3	13.6	21.9
RWIN	7.2	3.7	3.2	15.2
RSPR	4.6	2.0	1.6	7.8
RSUM	8.9	4.8	1.8	15.8
RAUT	6.8	2.6	3.0	11.0
RTOT	27.5	10.0	11.3	47.4

TABLE 2. (*Continued*)

Variable	Mean	Standard deviation	Minimum	Maximum
Cluster #7—Central and southern Florida (4 members)				
TWIN	17.7	1.6	15.9	19.5
TSPR	22.4	0.7	21.6	23.2
TSUM	27.3	0.1	27.3	27.4
TAUT	23.9	0.9	22.9	24.8
TAVG	22.8	0.8	21.9	23.7
RWIN	16.6	2.8	12.9	19.5
RSPR	25.9	2.7	24.1	29.8
RSUM	57.6	2.4	54.8	60.7
RAUT	36.9	8.2	31.7	49.0
RTOT	137.0	8.7	132.0	149.9
Cluster #8—Southern Texas/Florida Keys (3 members)				
TWIN	17.1	4.0	13.6	21.4
TSPR	23.7	1.2	22.5	24.9
TSUM	29.0	0.5	28.6	29.6
TAUT	24.1	1.8	22.5	26.1
TAVG	23.5	1.6	22.1	25.3
RWIN	10.7	2.4	8.6	13.3
RSPR	15.4	2.3	13.5	17.9
RSUM	23.6	10.8	17.0	36.0
RAUT	26.3	11.2	17.8	38.9
RTOT	76.0	26.3	58.3	106.2
Cluster #9—Central and southern California (3 members)				
TWIN	9.4	2.0	7.3	11.2
TSPR	13.8	0.6	13.2	14.4
TSUM	20.9	2.6	18.2	23.5
TAUT	16.5	1.0	15.6	17.6
TAVG	15.2	1.0	14.2	16.1
RWIN	26.4	2.6	23.9	29.0
RSPR	12.7	1.4	11.3	14.2
RSUM	0.7	0.1	0.6	0.8
RAUT	8.0	1.4	6.4	8.8
RTOT	47.7	4.7	42.4	51.0
Cluster #10—High-elevation West (6 members)				
TWIN	−3.3	2.8	−8.5	−1.2
TSPR	5.2	2.0	1.8	7.4
TSUM	15.9	1.5	14.0	17.7
TAUT	6.9	1.6	4.0	8.6
TAVG	6.2	1.8	2.8	7.8
RWIN	28.3	7.0	18.5	39.2
RSPR	14.9	1.8	12.2	16.8
RSUM	7.7	3.0	3.9	11.4
RAUT	16.4	3.9	10.9	21.3
RTOT	67.2	10.2	52.8	79.3
Cluster #11—Northern California, southwest Oregon, and Puget Sound (4 members)				
TWIN	5.1	1.2	3.9	6.7
TSPR	10.7	1.0	9.8	11.6
TSUM	18.9	1.8	16.8	21.2
TAUT	12.6	1.6	11.0	14.2
TAVG	11.8	1.2	10.5	13.1
RWIN	45.7	6.9	38.5	54.5
RSPR	21.9	1.8	19.4	23.1
RSUM	5.1	3.6	2.5	10.3
RAUT	23.2	4.7	18.4	29.6
RTOT	95.9	9.1	85.1	104.0
Cluster #12—Cascades and Olympic west slopes (Washington) (2 members)				
TWIN	2.2	3.9	−0.6	5.0
TSPR	6.9	2.7	5.1	8.8
TSUM	14.4	0.6	14.0	14.8
TAUT	9.2	2.4	7.6	10.9
TAVG	8.2	2.4	6.5	9.9
RWIN	103.4	4.0	100.6	106.1
RSPR	50.5	2.7	48.6	52.4
RSUM	17.8	0.4	17.5	18.1
RAUT	65.4	2.4	63.8	67.1
RTOT	237.1	8.6	231.0	243.1

TABLE 2. (*Continued*)

Variable	Mean	Standard deviation	Minimum	Maximum
Cluster #13—Cascades and Olympic foothills (Washington and Oregon) (2 members)				
TWIN	4.1	0.8	3.5	4.7
TSPR	9.6	0.6	9.2	10.0
TSUM	17.1	0.8	16.6	17.6
TAUT	11.1	0.9	10.5	11.7
TAVG	10.5	0.8	9.9	11.0
RWIN	62.0	5.3	58.2	65.7
RSPR	32.3	2.9	30.3	34.4
RSUM	11.5	4.0	8.7	14.3
RAUT	39.6	6.1	35.3	43.9
RTOT	145.5	18.2	132.6	158.4
Cluster #14—Cascade and Olympic west slopes (Oregon) (2 members)				
TWIN	4.4	3.5	1.9	6.9
TSPR	8.8	1.6	7.7	9.9
TSUM	16.0	0.7	15.5	16.5
TAUT	11.1	1.7	9.9	12.2
TAVG	10.1	1.5	9.0	11.1
RWIN	85.5	7.7	80.0	90.9
RSPR	44.0	2.0	42.6	45.4
RSUM	10.9	1.2	10.0	11.7
RAUT	48.9	2.3	47.3	50.5
RTOT	189.2	10.7	181.6	196.8

Note: Temperatures (T) are annual and seasonal averages (degrees Celsius); Precipitation values (R) are annual and seasonal accumulations (cm).

The degrees of separation among the clusters can be better judged from Fig. 8, which shows pairwise plots of the scores of the three component variables. The panels are plots of (a) PC1 versus PC2, (b) PC1 versus PC3, and (c) PC2 versus PC3. The letter symbols indicate cluster membership ("A" = cluster #1/Southeast; "B" = cluster #2/East Central, etc., in the order presented in Table 2). All three plots evince the clumping remarked upon earlier.

In general, many of the clusters may be judged to be poorly separated; this provides evidence that the clusters should be allowed to overlap. In the pairings of PC1 with PC2 and PC3, the Southeast cluster's members ("A") stand apart from those of the East Central, Northeastern Tier, and Interior West ("B," "C," and "D") clusters. The best separations between the two Northeast quadrant clusters are seen in the plots that include PC1, average annual temperature. Even so, the separation is not very substantial, and this is undoubtedly one reason why those clusters are sensitive to small perturbations that can alter the balance among the component variables.

The Interior West, while relatively dry throughout the year, is most distinct in the pairings involving PC3, warm season precipitation. The hot and dry Southwest cluster ("F") is especially distinct in Fig. 8b. The smaller clusters, such as those in Florida ("G" and "H"), southern California ("I"), and the Pacific Northwest clusters ("K" through "N"), are well separated from the pack on all three plots. Their within-cluster variances are large, but not when compared to their between-cluster separations.

e. *Comparison with the Koeppen climate classification*

As discussed in the Introduction, the Koeppen system is typical of climate classifications that consist of a set of rules that are applied to data (in this case, subsets of our temperature and precipitation data) but were not specifically constructed from those data. One version of the Koeppen rules has been applied to the NCDC dataset, yielding the map shown in Fig. 9. Often, high-elevation locales are collected together into a general "highland" climate group rather than given specific classifications. For ease of comparison with the reference clusterings (both the 14- and 25-cluster levels, as appropriate), this has not been done. The major Koeppen climate groupings are identified by letters: A (tropical), B (arid), C (temperate), D (cool or snow), and E (frigid or ice). There are no examples of the E class in the data domain. The B climates are further divided into steppe (BS) and desert (BW) subtypes. The lowercase letters *f* and *s* refer to locales that have sufficient precipitation in all months and those that have dry summers, respectively. The lowercase letters *a* and *b* are used to differentiate between places with warmer and cooler summers. Several differences between the classification shown in Fig. 9 and typical Koeppen maps are noted in the caption. None are judged to hinder the comparison.

The Koeppen system divides the eastern United States into four principal, zonally aligned groups. From north to south, these are the Dfb, Dfa, Cfa, and A zones, with the latter zone located in southern Florida. A few

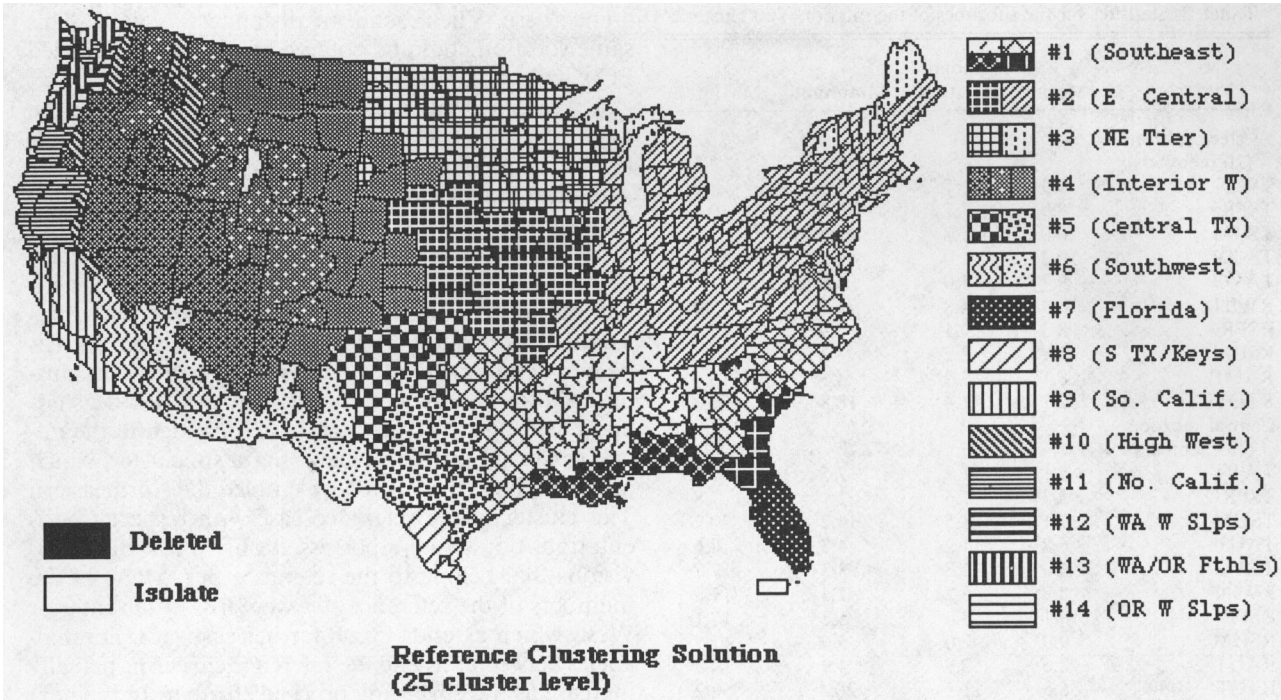

FIG. 7. The reference clustering solution of the 343 climate divisions in the conterminous United States at the 25-cluster level. Legend depicts subzones relative to 14-cluster solution. Two isolated clusters exist at this level (in western Wyoming and the Florida Keys).

divisions in the Appalachians are classified Cfb, owing to their cooler summers. The general distribution of climate types (but not the specific boundaries) are comparable to the clusters generated in the 14-cluster reference solution (Fig. 6). The reference clustering, however, gives precipitation more of a role in differentiating among the clusters in the eastern region, particularly the subzones identified in the 25-cluster solution (Fig. 7). The only Koeppen climate type to specifically involve precipitation is the B group. The remaining groups are distinguished solely on the basis of temperature.

In the Koeppen classification, the western United States is dominated by one spatially extensive cluster, the BS or arid steppe group, in much the same manner as the reference clustering is dominated by the interior West (and its subzones). The spatial extents of the BS group and Interior West cluster are fairly similar, with the exception that the former also includes much of the lower-elevation Southwest, which in the reference clustering remains separate (until finally joining at the 4-cluster level). Within the BS zone is a group of divisions that are assigned to the Dfb climate class. Many of those same divisions also belong to the central subzone of the Interior West identified in the 25-cluster solution (Fig. 7). In the Koeppen scheme, the Dfb divisions located in the west are not distinguishable from those residing along the U.S.–Canadian border, owing to the scheme's insensitivity to precipitation differences among the nonarid classes. The statistics for comparable zones in the reference clustering (the Northeastern Tier cluster in Table 2 versus the Interior West central subzone in Table 3) suggest that precipitation is indeed the sole important distinction between them.

The Koeppen subdivision of the West Coast is simpler than in the 14- or 25-cluster reference solutions. This is because those divisions belong to the C and D classes, and thus precipitation does not play a direct role in the partitioning. A number of other discrepancies can be noted. The main conclusion from this comparison is that giving precipitation greater weight in determining the boundaries of nonarid climate zones can have a significant, and arguably beneficial, effect on the resulting partitioning.

6. Variants on the reference analysis procedure

In this section, some other methods for conducting the climate regionalization, employing different methods and data preprocessing strategies, are discussed. Sensitivity of the results to the clustering method employed, the staple of cluster analysis sensitivity tests, will not be specifically considered. The "best" results were obtained with methods, such as average linkage, that have some skill in uncovering compact but relatively poorly separated clusters and are unbiased with respect to cluster membership size.

A few of the variants will be described in this section. As a group, these tests help assess how sensitive the reference clusterings are to both minor and major al-

TABLE 3. Statistics for the subzones of the Interior West cluster.

Variable	Mean	Standard deviation	Minimum	Maximum
Western subzone (20 members)				
TWIN	−0.3	1.7	−2.5	4.7
TSPR	8.9	1.2	6.3	10.7
TSUM	20.0	1.8	15.3	22.8
TAUT	10.2	1.0	8.2	11.6
TAVG	9.7	1.0	7.6	11.1
RWIN	10.4	4.8	5.1	23.4
RSPR	8.3	3.1	4.9	17.1
RSUM	5.9	2.2	3.3	11.3
RAUT	8.1	3.2	4.3	16.7
RTOT	32.7	11.4	18.8	59.3
Central subzone (17 members)				
TWIN	−5.1	1.8	−7.5	−1.1
TSPR	5.3	1.6	2.1	7.8
TSUM	17.1	1.5	14.2	20.2
TAUT	6.8	1.2	4.3	8.6
TAVG	6.0	1.3	3.3	7.7
RWIN	8.9	5.2	2.4	17.8
RSPR	10.1	3.0	5.2	17.1
RSUM	9.0	2.0	5.6	12.7
RAUT	8.6	2.9	4.9	14.3
RTOT	36.5	11.5	20.4	58.2
Eastern subzone (22 members)				
TWIN	−3.5	2.8	−9.6	0.8
TSPR	7.3	1.7	5.2	11.0
TSUM	20.5	2.0	17.5	24.5
TAUT	9.2	1.7	6.7	12.7
TAVG	8.4	1.8	5.5	12.0
RWIN	3.9	1.1	2.7	6.8
RSPR	12.1	3.2	3.8	16.2
RSUM	16.2	2.9	12.2	22.1
RAUT	7.4	1.2	5.7	9.8
RTOT	39.7	5.1	32.0	50.8

terations in the analysis strategy. The 14-cluster solution resulting from the reference strategy will be used as the basis of comparison. It is determined that the Northeast quadrant, where the East Central and Northeastern Tier clusters reside, is the most sensitive region of the data domain. Below, a 14-cluster solution using the raw data (which includes redundant information) is presented for comparison with the reference case. The problem of truncating the PCA analysis is illustrated. The influence of the one deleted climate division—located in northwestern South Carolina—is demonstrated. The results of some other tests are briefly described.

a. Clustering with the raw data

We preprocessed the raw dataset with PCA not for reasons of economy but rather to eliminate redundant information among the variables. We now present the kind of clustering we would have found acceptable if we were not concerned with information bias. The variables were standardized by data type, as in the reference case. The Euclidean distance was used. This same solution could be obtained from an untruncated PCA with variance-weighted scores.

There is a small peak in the pseudostatistics at the 14-cluster level, which is shown in Fig. 10, but this level was chosen mainly for comparison with Fig. 6. The clustering in much of the western half of the United States is qualitatively comparable to the reference solution. In much of this region, one data type (temperature or precipitation) tended to dominate. The eastern half of the domain, however, is rather differently partitioned. Both temperature and precipitation participate in defining climate zones in this region. This clustering finds no distinction between the Northeastern and East Central divisions, which separated in the reference clustering. This cluster has a subcluster, which joined in the 17th step, that resembled the Northeastern Tier cluster of the reference case with respect to orientation, but actually possessed only a few of the divisions that belong to the reference tier. Many of the members of the reference tier were lost to the Interior West, which extends much farther eastward. Discrepancies between this and the reference clustering actually increase as the clustering proceeds further. In the next (13th) step in the present analysis, the central Texas/Kansas cluster joins with the Northeast, creating a huge entity that stretches in an arc from the Gulf of Mexico to Maine. In the reference clustering, most of the divisions in that region later link up with the Southwest instead.

It will be noted below (section 6d) that relatively minor differences were found between the clusterings of standardized PC scores that were generated by covariance- and correlation-based analyses. With raw data, a somewhat analogous test is to compare the results from when the data are standardized by data type (as in Fig. 10) with the clustering obtained after each variable is standardized separately. That approach, however, resulted in a clustering (not shown) that is very different, understandable given the extreme sensitivity of the Euclidean metric to scaling assumptions. Because of this much greater sensitivity, as well as the presence of redundant information, these raw data clusterings are considered to be inferior to the reference solutions.

b. Clusterings with additional retained PCs

In the reference analysis, we were faced with the inescapable problem of truncating the PC analysis. As discussed earlier, cluster analyses in which different numbers of components are retained may well be substantially different, especially when the component scores are standardized. This might be called "PC truncation bias," clearly a component of information bias, and represents the chief impediment to the usage of PCA as a preprocessing tool.

Figure 11 shows the clustering that results when the two next most significant PCs are retained in the anal-

TABLE 4. Statistics for the subzones of the Northeast quadrant clusters.

Variable	Mean	Standard deviation	Minimum	Maximum
East Central cluster: Eastern subzone (101 members)				
TWIN	−1.1	3.0	−7.3	4.2
TSPR	9.8	2.5	4.7	14.7
TSUM	21.8	1.8	18.1	25.8
TAUT	12.1	1.9	8.3	15.6
TAVG	10.7	2.3	6.3	14.8
RWIN	22.4	5.5	10.6	40.5
RSPR	27.8	4.2	17.6	38.5
RSUM	29.4	3.3	22.1	38.8
RAUT	24.4	3.2	18.6	31.7
RTOT	104.0	13.8	73.3	141.1
East Central cluster: Western subzone (28 members)				
TWIN	−2.4	2.7	−6.4	3.7
TSPR	10.7	1.8	7.7	14.2
TSUM	23.8	1.2	21.3	26.0
TAUT	12.4	1.6	10.1	15.3
TAVG	11.1	1.8	8.3	14.6
RWIN	9.1	4.2	3.8	20.1
RSPR	23.7	4.6	16.9	35.1
RSUM	29.2	3.4	21.2	33.3
RAUT	18.5	5.4	8.9	28.4
RTOT	80.5	16.5	51.0	113.9
Northeast Tier cluster: Eastern subzone (6 members)				
TWIN	−8.4	0.8	−9.8	−7.3
TSPR	3.8	0.6	3.1	4.8
TSUM	17.5	0.4	17.0	18.2
TAUT	7.2	0.5	6.6	7.8
TAVG	5.0	0.5	4.3	5.7
RWIN	19.4	4.0	13.8	24.5
RSPR	22.2	3.1	18.0	25.8
RSUM	28.3	2.4	24.1	31.0
RAUT	26.1	2.9	22.0	29.2
RTOT	96.1	11.9	81.4	109.1
Northeast Tier cluster: Western subzone (31 members)				
TWIN	−9.8	2.4	−14.0	−3.8
TSPR	5.6	1.4	3.1	7.9
TSUM	20.1	1.2	17.1	22.0
TAUT	7.6	1.4	5.5	9.9
TAVG	5.9	1.5	3.5	8.1
RWIN	5.7	2.1	3.0	10.0
RSPR	16.4	4.2	10.0	23.1
RSUM	25.9	4.8	18.7	32.6
RAUT	13.4	4.8	6.9	20.7
RTOT	61.4	15.3	39.5	83.6

ysis, bringing the total to five standardized variables. The two PCs were added as a pair, owing to their insignificant separation (North et al. 1982). The 14-cluster solution shown was suggested by the pseudostatistics. After rotation, these two PCs explain a total of 5.1% of the variance. Inclusion of PCs 4 and 5 means more information, especially from the May and October precipitation variables, is retained in the component dataset.

Some of the differences between this and the reference clustering will be examined, focusing mainly on the smaller membership clusters for convenience. For example, Fig. 11 shows that most of California is fused into one cluster by this step. In the reference case, southern California joins first with the Southwest. Strong similarity with respect to May rainfall and summer temperature among the California divisions appears to have encouraged this combination. In a similar fashion, three divisions in western Washington state are merged. In the reference clustering, one of these (in the northeast tip of the Olympic peninsula) was a member of the Interior West cluster, owing to its relative dryness in winter compared with its neighbors. In the summer and autumn seasons, this division is more similar to its immediate neighbors. The identity of a high-elevation West cluster that was similar to #10 in the reference case was lost much earlier in the clustering process.

The Interior West cluster has lost the members in Arizona and New Mexico that it possesses in the ref-

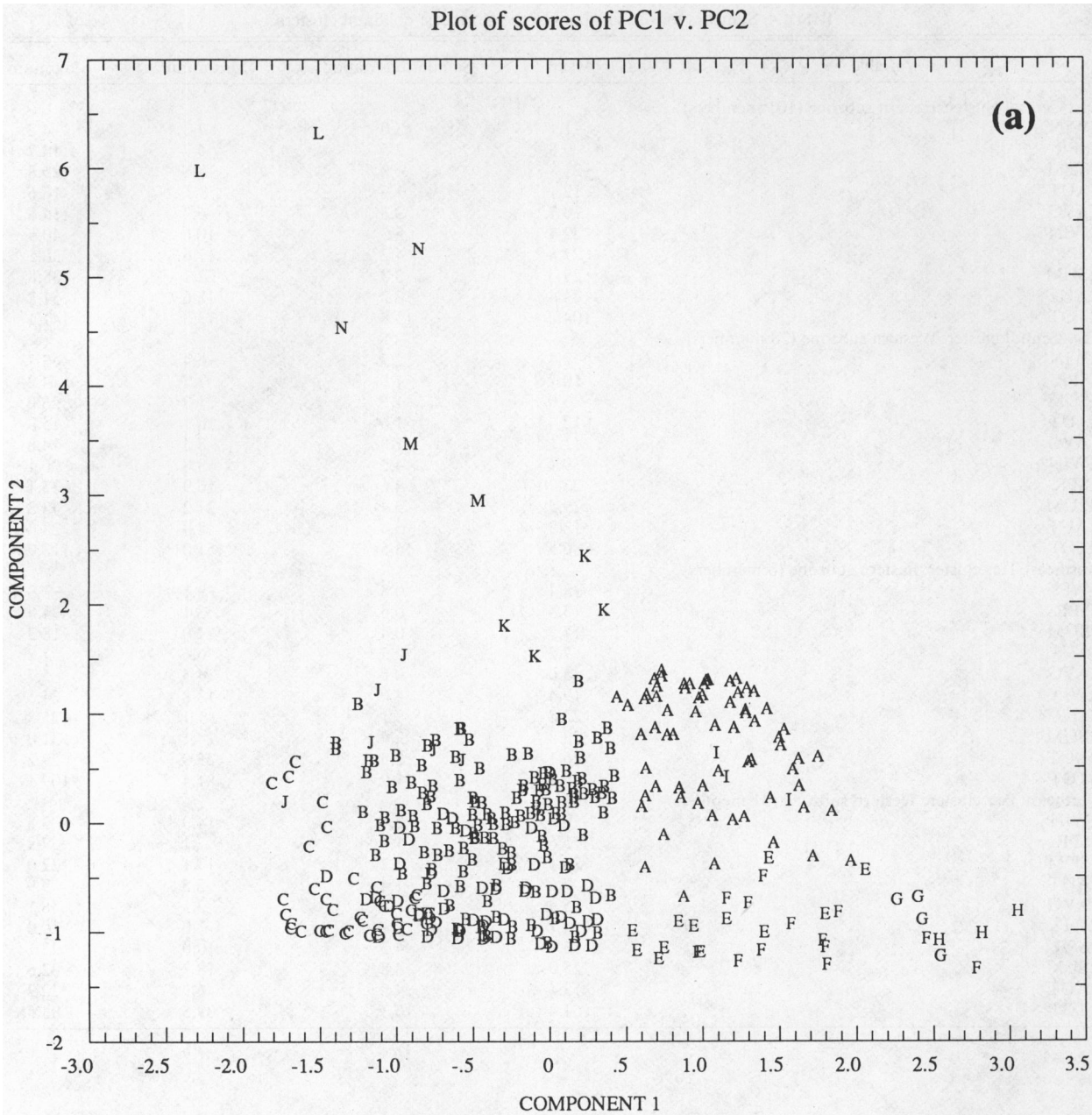

FIG. 8. Pairwise plots of the scores of the three rotated component variables retained in the truncated reference PCA. Symbols plotted indicate cluster membership, keyed to Table 2 ("A" is cluster #1; "B" is cluster #2, etc.). Panels are (a) PC1 versus PC2; (b) PC1 versus PC3; (c) PC2 versus PC3.

erence case. Most of the divisions in the Northeast quadrant have already been fused together at the displayed clustering level. The subzones of this cluster (not shown) are substantially different from those in the reference solution. The principal division is into western and eastern subzones that meet in central Ohio. These subzones merge at step 28, quite early in the clustering process. A much smaller perturbation, discussed in the next subsection, also radically alters the clusters forged in this region.

It is difficult to state that this clustering is inferior to the reference case, because some of its different combinations are logical. However, this clustering demonstrates how changing the truncation level can affect the outcome. Giving greater weight to one data type or time period over another can alter or even eliminate the less well defined clusters. We also performed a clustering for which all 24 standardized component variable PCs retained (not shown). An identical result is obtained from the raw data when the Mahal-

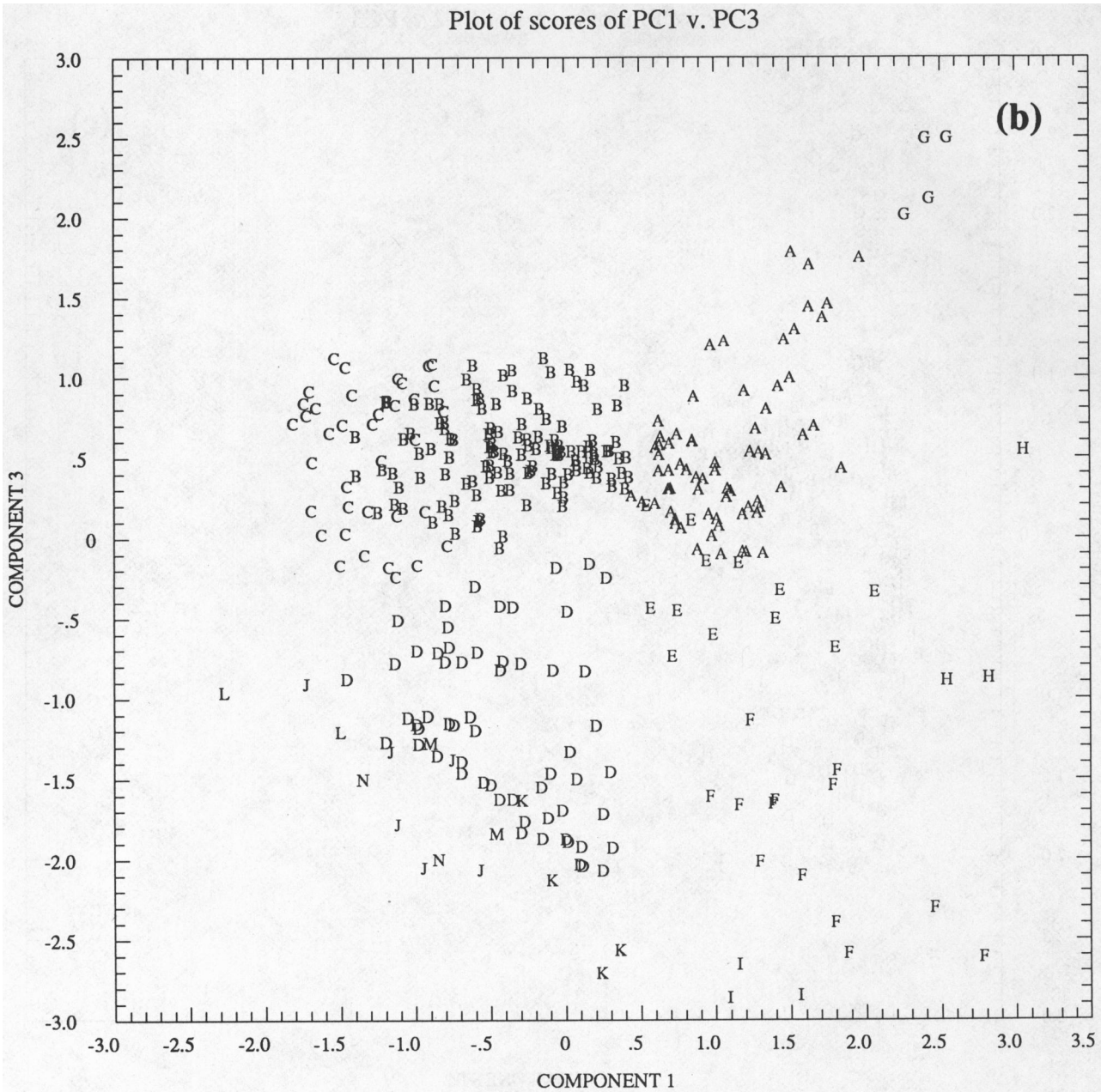

FIG. 8. (*Continued*)

anobis distance is used. This clustering, however, resulted in the production of one massive cluster that grows by chaining and contains 312 members at the 14-cluster level, a disconcerting—and probably illogical—result. The few divisions that remained independent at that level were the obviously different divisions in the dataset. It is possible that using the Mahalanobis metric (with the raw data) obscured cluster distinctiveness to the point where a cascade of chaining was inevitable. Since this clustering contained redundant information, and exaggerated useless information, it is easier to believe that this is the result that is flawed and not the reference clustering.

c. Inclusion of the deleted climate division from South Carolina

One division, located in the mountains in extreme northwestern South Carolina, was clearly the most different and unique data point in the domain. Because it is believed to be of suspect quality, it was deleted prior to PC analysis in the reference case. When it is

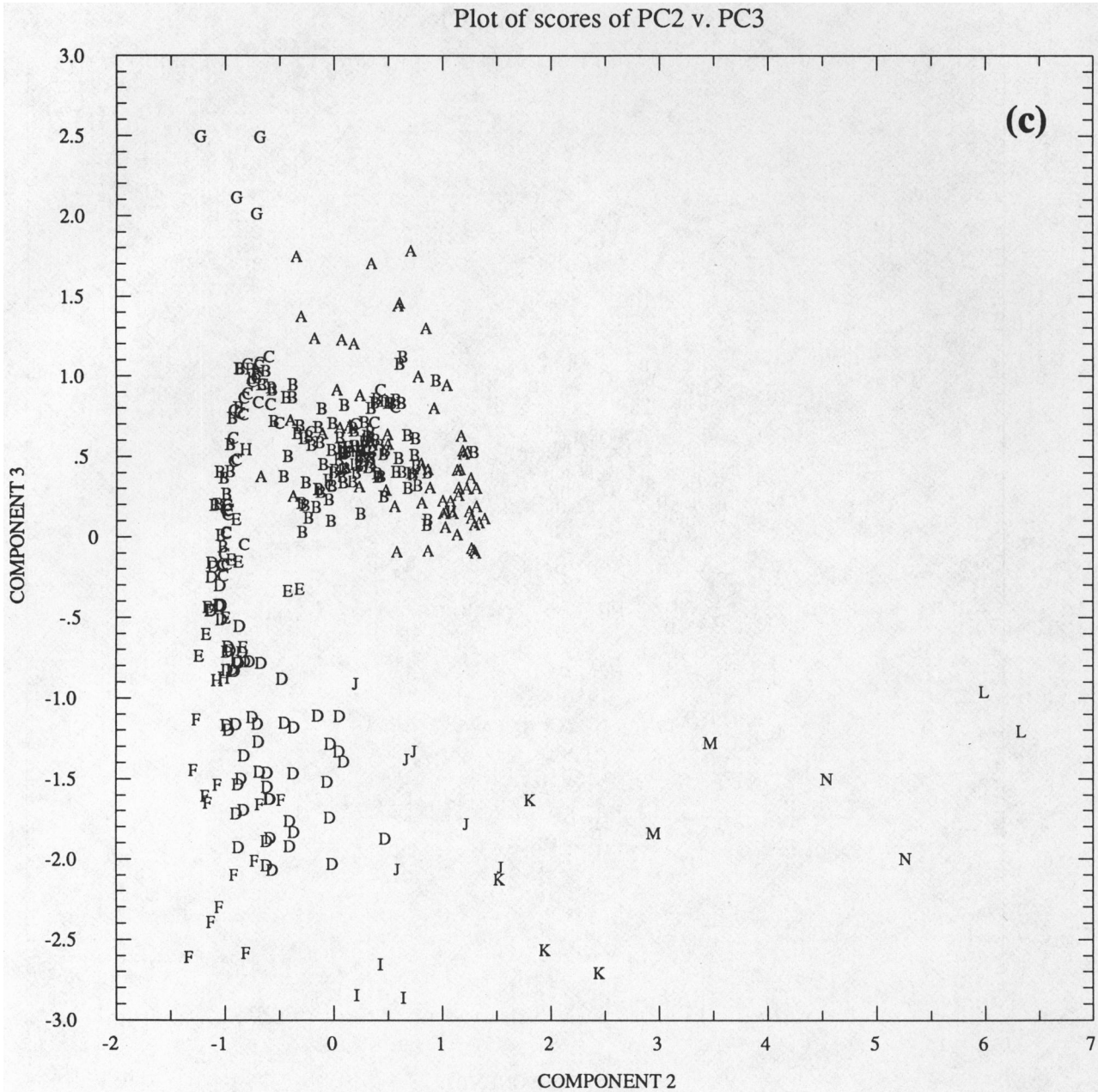

FIG. 8. (*Continued*)

retained, however, it has little *apparent* influence on either the composition of the extracted components or their scores. The component score distributions and pairwise plots (not shown) are virtually identical to those shown in Figs. 4 and 8. Furthermore, the division in question remains an isolate in the clustering process through to the very last step.

The division does, however, manage to induce a substantial change in the clustering in the Northeast quadrant of the United States. At the 15-cluster level (now including the isolate division), the orientation of the clusters formed in the northeast United States is shifted 90°, from the zonal alignment they have in the reference solution (Fig. 6) into an essentially west–east partition (not shown). The remaining clusters are unchanged at this clustering level. As in the reference clustering, the Northeast quadrant is divided into four subzones at an earlier stage in the clustering. In this solution, however, the two western subzones later merge to form a plains cluster, as do the two eastern subzones, joining the Ohio Valley with New England.

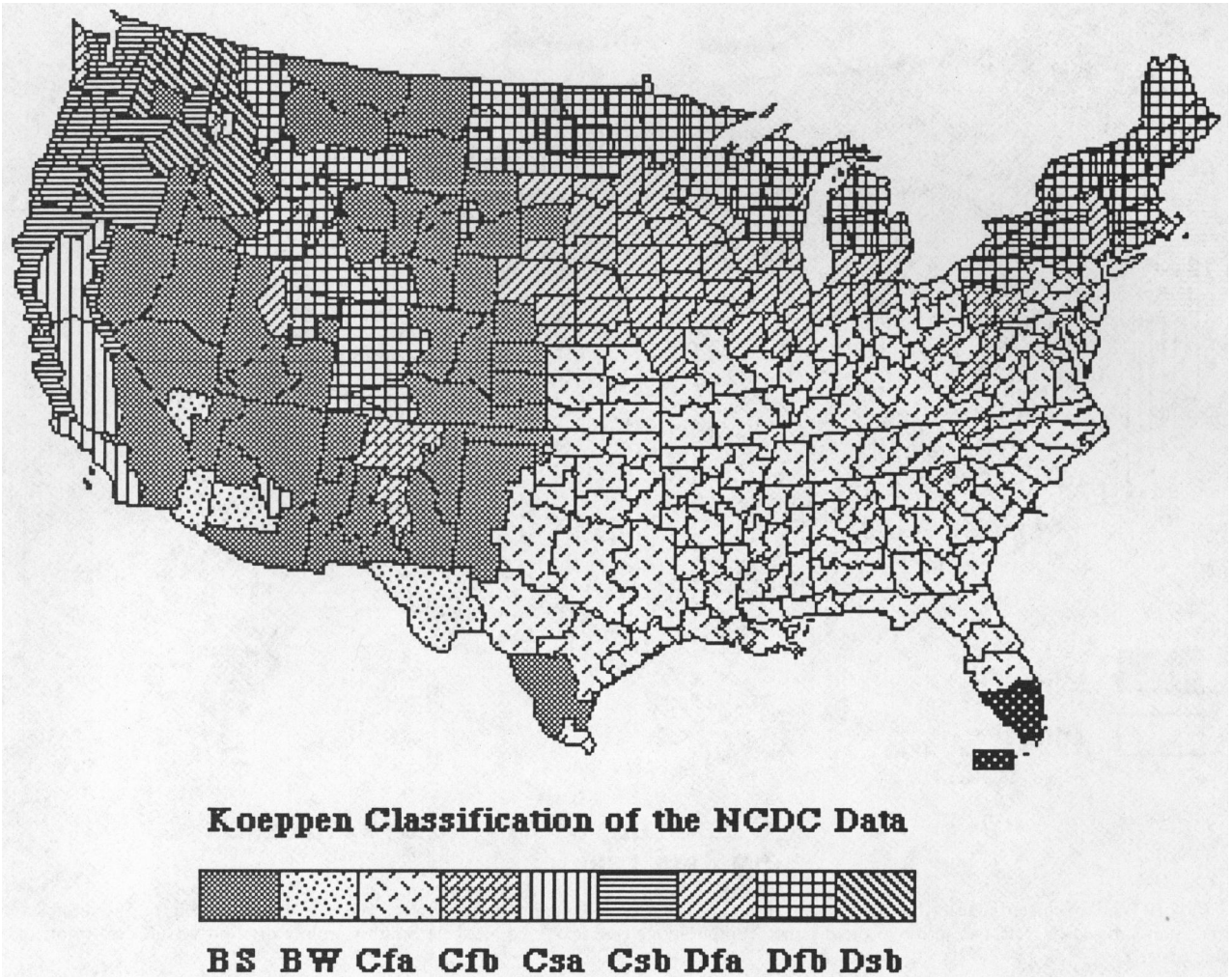

FIG. 9. Koeppen classification of the NCDC climate division dataset. Locales normally assigned to the undifferentiated "highland" group have instead been given the appropriate specific classification, for greater comparability with the reference clusterings in Figs. 6 and 7. Compared to typical Koeppen system maps, the arid desert class (labeled BW) is much smaller in extent. Several western divisions were found to be very close to the boundary between this and the arid steppe (BS) class. The class Csa division in central Arizona was just slightly too wet to be placed in the BS class; the same holds for the two Cfb divisions in New Mexico. One division in North Dakota was actually classified Dwb, but was included in the Dfb class for convenience.

Actually, the subzones themselves do not have precisely the same composition as they do in the reference clustering. (Four divisions in Missouri and one in Wisconsin, all located along the borders of the subzones, have different subcluster memberships in this variant than in the reference solution.) This small rearrangement is sufficient to radically alter the form and substance of the more general partitioning of the quadrant. The controlling distinction between the two Northeast quadrant clusters in the 14-cluster reference solution is in temperature. The west–east partition in this variant clustering, however, favors cool season precipitation information. Generally, temperature increases southward throughout this quadrant while cool season precipitation increases eastward (see Fig. 4). Both orientations can be rationalized, and the two different partitionings are both found to be statistically different from each other as well as from the remaining clusters. Still, this underscores the lack of robustness that characterizes the clusters of this quadrant, and can be interpreted to represent what can happen when the subtle balance between temperature and precipitation information is altered in a region in which both are important contributors to the results.

d. Some other sensitivity tests used in this research

Among the tests employed during this work are (no clusterings are shown):

- Clusterings on data after interpolation to a uniform grid. The NCDC data were interpolated onto a mesh with 2° resolution, consisting of 221 points. This

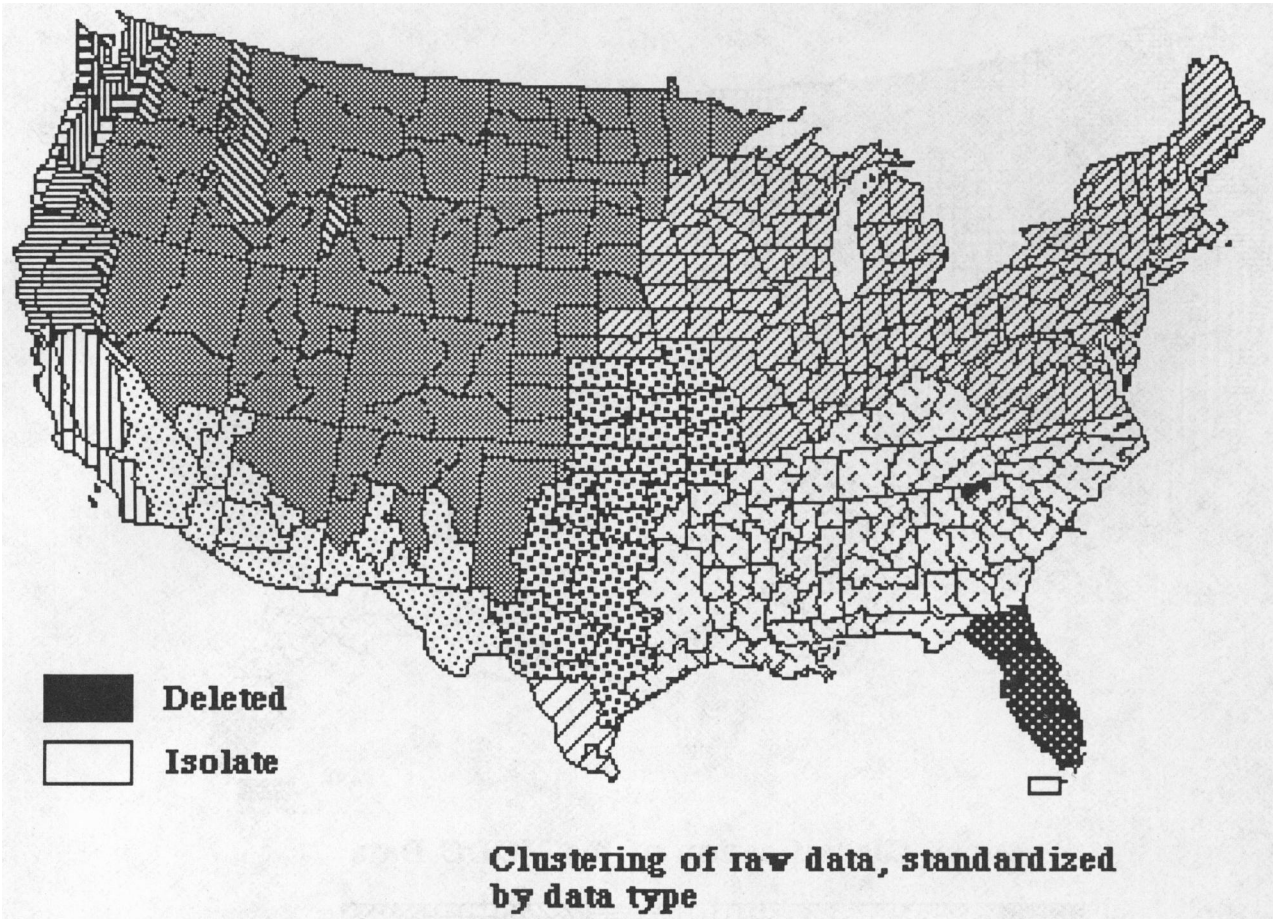

FIG. 10. Variant clustering using the raw dataset, at the 14-cluster level, for comparison with the reference solution in Fig. 6. The variables were standardized by data type prior to the distance computations (see text). The suspect South Carolina division was deleted prior to clustering.

is a check on latent bias in the dataset owing to the spatially irregular distribution of climate divisions. The clusterings produced from the gridded dataset are somewhat different from those made originally, but the major discrepancies can be attributed to the degraded resolution in the domain after gridding. No evidence was found that the cluster anchors created by average linkage (upon the original or uniformly gridded datasets) were unduly influenced by the spatial distribution of the data points. This may have been the case since average linkage does not tend to be biased with respect to cluster membership size.

• Preprocessing using correlation-based PCA. Covariance-based PCA was used in the reference analysis procedure in order to retain as much information about the seasonal cycle as possible. However, the correlation-based PCA resulted in clusterings that are quite comparable to the reference solutions, even in the sensitive Northeast quadrant. The major discrepancies between the two tend to reside at cluster boundaries, particularly along the eastern border of the Interior West (which accounted for 60% of the 41 discrepancies found at the 14-cluster level). This demonstrates that the clusters should not be truly hard and nonoverlapping anyway. One reason for the lack of sensitivity to this decision is the fact that we employed standardized scores. Variance-weighted scores would have further amplified the influence of the precipitation components.

• Clustering on artificially generated "season" variables. Before processing with PCA, the raw data were reworked into eight "season" variables (four each for temperature and precipitation), using conventional delineations (i.e., winter is December–February, etc.). Otherwise, the reference analysis strategy was followed. This represents a test of a commonly employed, though nonobjective, variable reduction strategy, sometimes specifically used to reduce variable intercorrelations. The clusterings generated with these "season" data differ most from the reference solutions in the Great Plains region. Recall that the reference PCA suggests a nonconventional subdivision of the year for the precipitation variables. Thus, forming artificial season variables in this instance results in irretrievable loss of in-

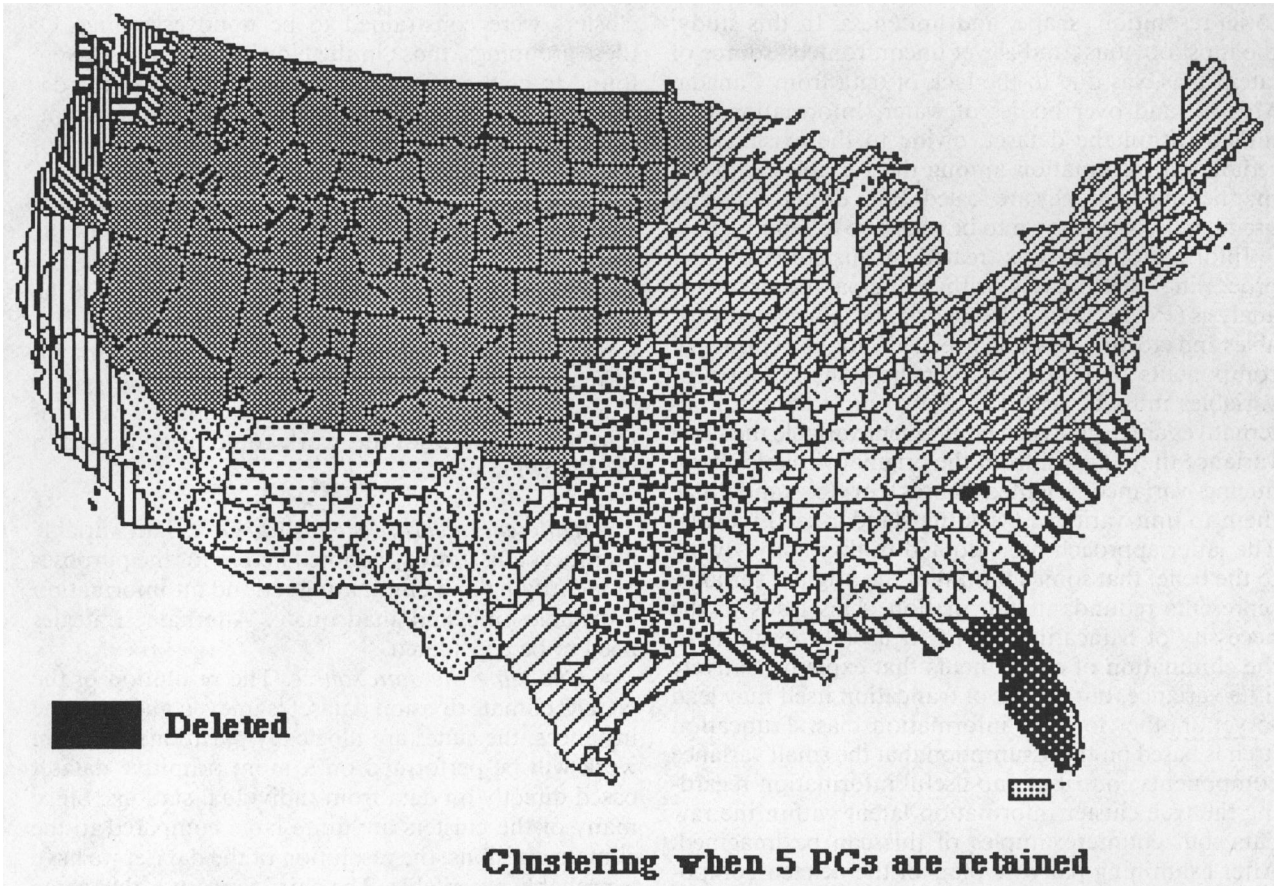

FIG. 11. Variant clustering resulting when the first five PCs extracted in the reference PC analysis were retained, at the 14-cluster level, for comparison with the reference solution in Fig. 6.

formation, which otherwise would not occur. It is concluded that, for this case, "season" variables do not possess sufficient temporal resolution for the task at hand.

7. Discussion and conclusions

The goals of this work were to determine, as objectively as possible, climate zones and subzones of the conterminous United States, and to evaluate the quality and stability of the solution obtained with respect to the decisions made during the variable preprocessing stage prior to the commencement of the clustering. Hierarchical cluster analysis was chosen to perform the regionalization. Temperature and precipitation were identified as being likely candidates to be important distinguishing variables, and long-term monthly averages of temperatures and precipitation accumulations for the climate division dataset issued by the National Climatic Data Center (NCDC) were constructed. Next, a measure of dissimilarity between objects (climate divisions) were required. Unfortunately, distance is a nonuniquely defined concept, which is one of the major difficulties associated with cluster analysis. We specifically considered two measures, the Euclidean and Mahalanobis metrics, and discussed their inherent assumptions.

Hierarchical cluster analysis consists of identifying and fusing the least dissimilar objects remaining at each step. Once a fusion is accomplished, the distance between the newly formed cluster and the remaining clusters must be recomputed. The many clustering methods that exist differ on how this is accomplished. We decided that the average linkage method was most likely to yield acceptable results in our case. As with the other traditional clustering methods, however, average linkage produces "hard," nonoverlapping clusters, which is very likely an unrealistically stringent constraint. The clustering process continues until one all-inclusive cluster is created. The results are then inspected to determine where the clustering process should have been halted.

Three major sources of bias were identified and considered in this study: methodological, latent, and information bias. The first is inherent in the various clustering methods, and has long been recognized and debated in the literature. We defined latent bias as that lying hidden in among the objects of the dataset due

to its resolution, shape, and finiteness. In this study, the most obvious (and as yet unconfronted) source of latent bias was due to the lack of data from Canada, Mexico, and over bodies of water. Information bias lurked within the dataset, owing to the existence of redundant information among the variables, and the manner in which they are scaled. Both distance metrics used herein were shown to be sensitive to redundancy.

Information bias was treated in this study by preprocessing the dataset with principal components analysis (PCA), taking the 24 original, correlated variables and creating new uncorrelated variables (principal components or PCs). Once created, the component variables must be scaled. The two most common alternatives are to give each component variable the same variance they explain from the original variables (producing variance-weighted scores) or to standardize them to unit variance (yielding standardized scores). The latter approach was adopted in this study, owing to the belief that some amount of the original variance represents redundant information. This leads to the necessity of truncating the PCA, usually resulting in the elimination of components that explain relatively little variance, but the act of truncation itself may lead to yet another form of information bias. Truncation itself is based on the assumption that the small variance components contribute no useful information regarding the true cluster information latent within the raw data, but counterexamples of this can be imagined. After examining pairwise plots of the deleted component variables, however, we believe that little or no potentially important information was lost by eliminating those components. Still, it must be acknowledged that the specific truncation points we chose were rather subjectively determined and remain open to question. Further, it was demonstrated that the resulting clustering solutions are very sensitive to those choices.

We termed our "best" clusterings the "reference clusterings." Available statistical tests were used to identify a set of candidate clustering levels, from which the 14-, 25-, and 8-cluster solutions were chosen for closer examination. These solutions represent different, but valuable, levels of detail. In the 14-cluster solution, a disproportionate number of clusters were concentrated in the Pacific Northwest. The bulk of the conterminous United States was divided into four zones, consisting of much of the West (the Interior West), the Southeast, and two zones in the Northeast quadrant (the Northeastern Tier and East Central clusters). The Interior West cluster was used as an example of how each step in the clustering process inevitably results in information loss. The 25-cluster solution split this zone into three subzones with markedly different seasonal cycles in precipitation. Statistics for these clusters were presented.

The quality or robustness of the identified clusters, at any clustering level, is a concern, especially as the clusters were constrained to be nonoverlapping. Of these groupings, those in the Northeast quadrant were found to be most unstable to perturbations in the dataset and analysis procedure. These perturbations consisted mainly of varying the data preprocessing procedures, and included gridding the data (due to concern over a form of latent bias), using the raw data (which contains redundant information), and altering how the original variables were standardized to eliminate arbitrary scalings. Both small and large changes in the PCA preprocessing strategy were considered. In the future, alternative methods for excising redundant information should be investigated.

Some of the difficulties and shortcomings we identified in the course of this work could be addressed in the following ways:

- *Different preprocessing strategies.* The variable preprocessing strategy adopted herein for the purposes of identifying and eliminating redundant information is troublesome and inadequate. Alternate strategies need to be considered.
- *Changing the data source.* The resolution of the NCDC climate division dataset is uneven and, in some instances, the states are illogically partitioned. Future work will be performed on a more primitive dataset based directly on data from individual stations. Since many of the clusters are quite large compared to the climate divisions, the resolution of the dataset we have is probably acceptable. The chief concern is that some climate divisions may represent nonoptimal mixing of still more primitive information (comparable to the discussion of the "season" variables in section 6), particularly in zones (such as the mountainous Interior West) in which more complex partitionings at the less advanced clustering levels might well be anticipated. The new dataset will hopefully include data from Canada, Mexico, and from over the water areas; this might help increase the robustness and internal cohesiveness of the clusters identified and ease some concerns about latent bias.
- *Overlapping clustering.* The clusters are clearly less robust when they are unnaturally constrained to be hard and nonoverlapping. While the clusters we identified were rather distinct statistically, there is no reason to believe that many of the clusters truly have sharply defined boundaries. The discrepancies noted among the clusterings that gave the more comparable regionalizations yield some information about the sizes of the "buffer zones" that exist between cluster cores. A better procedure might well be to simply relax the constraint against overlapping. One approach is to perform some kind of overlapping (or "fuzzy") clustering. Arabie and Hubert (1992) provide a brief literature review on this topic. Another approach would be to use PCA and make spatial plots of eigenvector loadings. These naturally produce overlapping clusters. Richman and

Lamb (1985) have performed this type of regionalization.

- *Inclusion of additional or alternate variables.* The clusters might be made more robust if other potentially relevant variables were added to the dataset.
- *Consensus clustering.* A large part of the difficulties encountered in this study issued from our having *too many* variables. This made concerns over information bias very serious. An alternate strategy would be to perform a series of clusterings, each incorporating no more than two variables, and, through intercomparison of the results, generate a compromise solution. The comparison of independently produced clustering solutions has been termed "consensus clustering" (see Arabie and Hubert 1992). This approach will likely yield excessively muddled maps if the independent clusterings generated are very different from one another (as might be expected to occur in the Northeast quadrant).
- *Partitioning clustering.* The partitioning approach to cluster analysis may generate better, more robust solutions, despite the constraint of cluster hardness. We were reluctant to give up the advantages of hierarchical clustering for this study, however.

Of these recommendations, the most important are the first three. The clusterings presented herein may be favorably altered by the inclusion of more (and better) data, the relaxation of the artificial constraint of cluster hardness, and the use of superior data preprocessing techniques. If it is applied with a clear understanding of its shortcomings, cluster analysis may be a very useful tool that could be applied to other datasets, including those generated by climate models, to help assess how climate types, their locations, statistics, and distributions could change in the future.

Acknowledgments. The authors are indebted to numerous individuals who took an interest in this work and gave us advice. In particular, Drs. R. I. Jennrich, J. Walsh, M. B. Richman, and P. Arabie made valuable suggestions. Dr. C. Ropelewski alerted the authors to the suspect climate division in South Carolina. Also, suggestions from the journal reviewers improved the quality of the manuscript. This work was supported by the Academic Senate of the University of California, Los Angeles.

REFERENCES

Arabie, P., and L. J. Hubert, 1992: Combinatorial data analysis. *Annu. Rev. Psychol. 1992,* **43,** 169–203.

Calinski, R. B., and J. Harabasz, 1974: A dendrite method for cluster analysis. *Communic. in Stat.,* **3,** 1–27.

Cattell, R. B., 1952: *Factor Analysis.* Harper and Row, 462 pp.

Chang, W.-C., 1983: On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Statist.,* **32,** 267–275.

Cronbach, L. J., and G. C. Gleser, 1953: Assessing similarity between profiles. *Psych. Bull.,* **50,** 456–473.

De Soete, G., 1986: Optimal variable weighting for ultrametric and additive tree clustering. *Qual. and Quant.,* **20,** 169–180.

Duda, R. O., and P. E. Hart, 1973: *Pattern classification and scene analysis.* Wiley, 482 pp.

Fovell, R. G., and M.-Y. C. Fovell, 1993: Cluster analysis of U.S. temperature and precipitation data: Regionalization and data reduction. Preprints, *Eighth Conference on Applied Climatology,* Anaheim, CA, Amer. Meteor. Soc., 165–168.

Gadgil, S., and R. N. Iyengar, 1980: Cluster analysis of rainfall stations of the Indian peninsula. *Quart. J. Roy. Meteor. Soc.,* **106,** 873–886.

——, and N. V. Joshi, 1983: Climatic clusters of the Indian region. *J. Climatol.,* **3,** 47–63.

Jackson, J. E., 1991: *A User's Guide to Principal Components.* Wiley, 569 pp.

Jolliffe, I. T., 1972: Discarding variables in principal component analysis. I: Artificial data. *Appl. Stat.,* **21,** 160–173.

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

——, P. C. Dunne, and R. S. Vose, 1990: Detection of climatic change in the western North American Arctic using a synoptic climatological approach. *J. Climate,* **3,** 1153–1167.

Koeppen, W., 1923: *Die Klimate der Erde; Grundriss der Klimakunde.* De Gruyter, 369 pp.

Maryon, R. A., and A. M. Storey, 1985: A multivariate statistical model for forecasting anomalies of half-monthly mean surface pressure. *J. Climatol.,* **5,** 561–578.

Milligan, G. W., and M. C. Cooper, 1985: An examination of procedures for determining the number of clusters in a data set. *Psychometrika,* **50,** 159–179.

——, and ——, 1988: A study of standardization of variables in cluster analysis. *J. Classification,* **5,** 181–204.

North, G. R., T. L. Bell, R. F. Calahan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.,* **110,** 699–706.

Preisendorfer, R. W., 1988: *Principal Component Analysis in Meteorology and Oceanography.* Elsevier, 425 pp.

Richman, M. B., 1986: Rotation of principal components. *J. Climatol.,* **6,** 293–335.

——, and P. J. Lamb, 1985: Climatic pattern analysis of three- and seven-day summer rainfall in the central United States: Some methodological considerations and a regionalization. *J. Climate Appl. Meteor.,* **24,** 1325–1342.

SAS Institute, 1985: *SAS User's Guide: Statistics.* SAS Institute, 959 pp.

Schulz, T. M., and P. J. Samson, 1988: Nonprecipitating low cloud frequencies for central North America: 1982. *J. Appl. Meteor.,* **27,** 427–440.

Sokal, R. R., and P. H. A. Sneath, 1963: *Principles of Numerical Taxonomy.* Freeman and Co., 359 pp.

Spaeth, H., 1980: *Cluster Analysis Algorithms for Data Reduction and Classification of Objects.* Ellis Horwood Limited, 226 pp.

Thornthwaite, C. W., 1931: The climates of North America, according to a new classification. *Geog. Rev.,* **21,** 633–655.

Velicer, W. F., 1976: Determining the number of components from a matrix of partial correlations. *Psychometrika,* **41,** 321–327.

Ward, J. H., 1963: Hierarchical grouping to optimize an objective function. *J. Amer. Stat. Assoc.,* **58,** 236–244.

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. *J. Climate Appl. Meteor.,* **26,** 540–558.