Euclidean Distance as a Similarity Metric for Principal Component Analysis

KIMBERLY L. ELMORE*

NOAA/National Severe Storms Laboratory, Norman, Oklahoma

MICHAEL B. RICHMAN

School of Meteorology and Cooperative Institute for Mesoscale Meteorological Studies,
University of Oklahoma, Norman, Oklahoma

(Manuscript received 24 February 2000, in final form 5 September 2000)

ABSTRACT

Eigentechniques, in particular principal component analysis (PCA), have been widely used in meteorological analyses since the early 1950s. Traditionally, choices for the parent similarity matrix, which are diagonalized, have been limited to correlation, covariance, or, rarely, cross products. Whereas each matrix has unique characteristic benefits, all essentially identify parameters that vary together. Depending on what underlying structure the analyst wishes to reveal, similarity matrices can be employed, other than the aforementioned, to yield different results. In this work, a similarity matrix based upon Euclidean distance, commonly used in cluster analysis, is developed as a viable alternative. For PCA, Euclidean distance is converted into Euclidean similarity. Unlike the variance-based similarity matrices, a PCA performed using Euclidean similarity identifies parameters that are close to each other in a Euclidean distance sense. Rather than identifying parameters that change together, the resulting Euclidean similarity–based PCA identifies parameters that are close to each other, thereby providing a new similarity matrix choice. The concept used to create Euclidean similarity extends the utility of PCA by opening a wide range of similarity measures available to investigators, to be chosen based on what characteristic they wish to identify.

1. Introduction

Eigentechniques have been widely used in meteorology since the 1950s. Three common variants are common factor analysis (CFA; Thurstone 1947), empirical orthogonal functions (EOF; Lorenz 1956), and principal component analysis (PCA; Hotelling 1933). EOF uses unit-length eigenvectors, whereas in PCA and CFA each eigenvector is weighted by the square root of its corresponding eigenvalue. Consequently, the weights represent the correlations or covariances between each variable and each principal component, depending upon which similarity matrix is employed (Jolliffe 1995). Any of the three techniques may be used as either a statistical modeling tool or as a diagnostic tool. Each eigentechnique is derived directly from a parent similarity matrix (also called a dispersion matrix in some texts) that typically consists of either a correlation or covariance matrix, or, rarely, a matrix of cross products. These similarity matrices are diagonalized such that eigenvalues and associated eigenvectors are identified and eventually used in the physical interpretation phase of the analysis. Because the parent similarity matrix embodies the type of association desired, and defines the immediate starting point for the eigenanalysis by virtue of being diagonalized, the similarity measure used to build the parent similarity matrix is an important aspect of any eigentechnique. However, this choice is not always given the consideration it is due. Historically, the various similarity matrices have been discussed and compared in meteorological literature. Examples include Craddock (1965), Kutzbach (1967, 1969), and Craddock and Flood (1969), who all favor the covariance matrix on grounds that it more accurately portrays the true variance structure. In contrast, Gilman (1957), Sellers (1957) and Glahn (1965) are proponents of the correlation matrix, claiming it puts all variables on equal footing, whereas Resio and Hayden (1975) and Molteni et al. (1983) find cross products to have utility.

Depending upon the parent similarity matrix, eigentechnique results can have physically different meanings. Table 1 defines cross products, covariance, and correlation for both single column data vectors of length n and $n \times p$ data matrices. The correlation matrix groups

* Additional affiliation: Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma.

Corresponding author address: Kimberly L. Elmore, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: elmore@nssl.noaa.gov

TABLE 1. Vector and matrix forms for cross products, variance/covariance, and correlation. For the vector form, n is the length of the vector and $\overline{\mathbf{x}}(\overline{\mathbf{y}})$ is the vector whose values consist of the mean of $\mathbf{x}(\mathbf{y})$. In the vector form of correlation, $s_x(s_y)$ is the square root of variance for the $\mathbf{x}(\mathbf{y})$ vector. For the matrix form, the $\mathbf{X}$ is $n \times p$ (n rows and p columns), and $\mathbf{M}$ (also $n \times p$) is the matrix whose ith column is the mean of the ith column of $\mathbf{X}$. $\mathbf{V}$ is a $p \times p$ diagonal matrix whose p non-zero elements consist of the variance of each column of $\mathbf{X}$.

Similarity measure	Vector form	Matrix form
Cross products	$p = \mathbf{x}^{\mathrm{T}}\mathbf{y}$	$\mathbf{P} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$
Variance/covariance	$s = \dfrac{(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}}(\mathbf{y} - \overline{\mathbf{y}})}{(n - 1)}$	$\mathbf{S} = \dfrac{(\mathbf{X} - \mathbf{M})^{\mathrm{T}}(\mathbf{X} - \mathbf{M})}{(n - 1)}$
Correlation	$r = \dfrac{\left[\dfrac{(\mathbf{x} - \overline{\mathbf{x}})}{s_{\mathbf{x}}}\right]^{\mathrm{T}}\left[\dfrac{(\mathbf{y} - \overline{\mathbf{y}})}{s_{\mathbf{y}}}\right]}{(n - 1)}$	$\mathbf{R} = \dfrac{[(\mathbf{X} - \mathbf{M})\sqrt{\mathbf{V}}^{-1}]^{\mathrm{T}}[(\mathbf{X} - \mathbf{M})\sqrt{\mathbf{V}}^{-1}]}{(n - 1)}$

variables together regardless of the amplitude of their variation or mean. By its nature, correlation provides no measure of how much parameters vary with each other, only that they do. For example, correlation does not address whether the *magnitude* of variation in one variable coincides with the *magnitude* of variation in another. Correlation addresses only relative variability relationship because the standardization puts all variances equal to unity. Often, such a relation is the most important aspect of an analysis. Because input data have been normalized to a zero mean and unit variance, correlation is a dimensionless similarity metric; hence, it is appropriate for comparing variables with different units or scales.

The covariance matrix yields insight into how much variables change with respect to each other. Data that are transformed into a covariance matrix have been translated to zero mean. Strictly speaking, dimensions or scales are preserved with covariance, which means that applications that mix different units will emphasize those with units having the most variation. Moreover, analyses using a single parameter whose variables (e.g., time series at an array of grid points) have a large range over the domain will emphasize those locations with a large variance. For example, an eigenanalysis based on a covariance matrix of Northern Hemisphere sea level pressure will emphasize those geographic locations that exhibit large variance (midlatitudes) and de-emphasize those geographic locations with small variance (tropical regions).

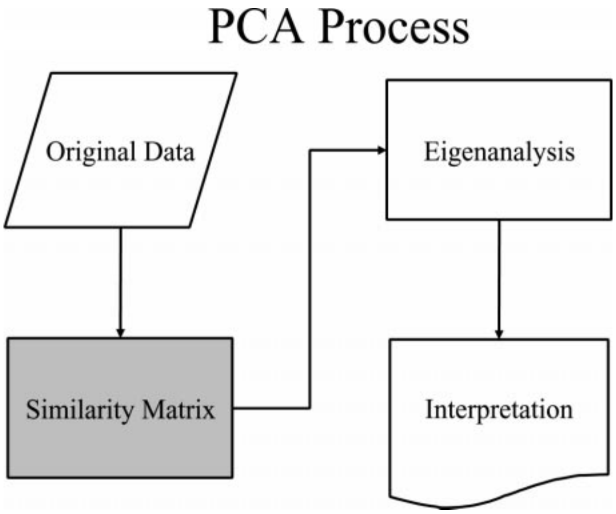

FIG. 1. Schematic PCA process. The process starts with the original data, which are then converted to a similarity matrix (shaded box). An eigenanalysis is performed on the similarity matrix and the results of the eigenanalysis interpreted by the analyst. The eigenanalysis results are significantly constrained by the similarity matrix step.

A cross products (covariance without removing the means) similarity matrix also preserves units, and results are sensitive to the magnitudes of the means as well as coassociation. A data vector that possesses a mean with large magnitude will dominate the eigenanalysis outcome. When the magnitudes of the means are similar, the eigenanalysis outcome is primarily determined by the covariance. Additionally, analyses utilizing the cross products similarity matrix tend to have a first principal component that resembles the mean.

For these three similarity matrices, coassociation plays an important role because the eigenanalysis, in some manner, identifies variables that change together. The PCA process may be summed up with a schematic diagram (Fig. 1). The similarity transformation step may be thought of as a kind of filter that determines the nature of any physical insights available form the data. Singular value decomposition (SVD) can be used to make the similarity transformation stage implicit. Nonetheless, the original data are still altered (e.g., standardized, have the mean removed, etc.), which is equivalent to the idea of a filter.

In the past, the eigenanalysis step has received considerable attention. Examples include using different eigentechniques (scalings), keeping different numbers of dimensions, various rotations, etc. Attention to the eigenanalysis has led to significant improvements in being able to match the similarity matrix well to the principal component (PC) loading vectors. Much less attention has been given to the similarity matrix step. Before now, the idea of using this step to tune an analysis was rarely considered, since the choices were usually limited to two: correlation or covariance. But might physical insight arise from another characteristic besides coassociation?

TABLE 2. Alternative distance measures.

Distance	Form
Mahalanobis distance	$d_{\mathbf{x},\mathbf{y}} = (\mathbf{x} - \mathbf{y})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})$, where Σ is any transforming positive semidefinite matrix (most commonly, **S**)
Minkowski norm between two vectors (also called the L_λ norm)	$d_{\mathbf{x},\mathbf{y}} = \{\lvert\mathbf{x} - \mathbf{y}\rvert^{\lambda}\}^{1/\lambda}, \lambda \geq 1$
Canberra metric	$d_{\mathbf{x},\mathbf{y}} = \frac{\lvert\mathbf{x} - \mathbf{y}\rvert}{(\mathbf{x} + \mathbf{y})}$
Bhattacharyya distance (proportions)	$d_{\mathbf{x},\mathbf{y}} = \left[\sum_{i=1}^{p} (x_i^{1/2} - y_i^{1/2})^2\right]^{1/2}$

Clearly, the investigator must carefully consider what is desired from an eigenanalysis and choose an appropriate parent similarity matrix that preserves the desired information. It is possible that coassociation of the aforementioned types is not an important issue. For example, objective identification and representation of variables that, when plotted together, are coassociated in that they appear visually close to one another may be important. A desirable result is to distill the data into a few, easily interpreted modes of behavior. To accomplish this result, a new similarity matrix, named Euclidean similarity (ES), is utilized. Euclidean similarity is inspired by the large body of literature on cluster analysis, which clearly demonstrates the effectiveness of Euclidean distance (ED), on which ES is based. However, standard cluster analysis creates ‘‘hard’’ clusters. This means that each data vector must be in one cluster or another. A PCA based on ES tends to create ‘‘fuzzy’’ clusters, which means that parts of each data vector will appear in each recovered mode. Also, retaining enough PCs to explain a given amount of similarity provides an objective criterion that determines the number of modes. Cluster analysis provides no equivalent objective criteria for defining how many clusters are enough.

Euclidean distance is one of a host of different distance measures that could be used. Euclidean distance is chosen primarily because its interpretation is straightforward. Other distances are given in Table 2 (Mardia et al. 1979). These distances are typically defined as between rows of the data matrix. To apply them in a PCA sense, they must be considered as between columns of the data matrix.

The next section introduces and defines ES. Section 3 demonstrates how modes of coassociation are extracted with a PCA based on ES. Section 4 demonstrates results of PCA using ES in both S-mode and T-mode analyses, and these results summarized in section 5.

2. Euclidean similarity

As a motivational example, ES is used to extract vertical velocity (w) time series that exhibit similar behavior and combine them into modes. In this example, the ensemble of w time series come from several cloud model runs, each started with slightly different initial conditions. Other time series can also be treated this way, such as precipitation, or the u and v wind components. Of course, extracting modes need not be limited to time series. For example, two-dimensional fields, such as spatial pressure or height patterns can be treated in an identical fashion.

In this case, the parameter to be examined is arranged in a data matrix that provides an S-mode analysis (Richman 1986), in which each column (variable) represents a time series vector. If the data are arranged such that each column consists of values at a single time, distributed in space, the analysis is T mode. Because the goal is to produce a small number or easily understood forecast modes, the modes must be visualized (Anderson 1996). The ED (or the Minkowski L_2 norm) provides a way to extract the desired modes. Cluster analysis uses ED extensively to identify or group data or entities that are dissimilar (Anderberg 1973; Gong and Richman 1994). Euclidean distance is an attractive measure for identifying modes because whether the parameter comprising the individual vectors *vary* together is not as important as how closely they overlie each other.

Any of the Minkowski norms could be used as a dissimilarity measure. The L_2 norm is particularly attractive because 1) it has a very intuitive interpretation through the geometry of Euclid, and 2) the norm behaves linearly. Least squares fits are a linear problem because variance, which may be thought of as a an L_2 norm, is used as the measure of error.

Consider a $p \times n$ data matrix $\mathbf{Z}$ composed of n columns (cases), each p elements long. The ED between the vectors z_i and z_j is

$$d_{ij} = [(z_i - z_j)^{\mathrm{T}}(z_i - z_j)]^{1/2}, \qquad (1)$$

where d_{ij} is the distance between the vectors z_i and z_j. When computed for all i, j, and arranged in a matrix in an order identical to that obtained from matrix cross products, this process results in a symmetric dissimilarity matrix of Euclidean distances, $\mathbf{D}$. For n cases, $\mathbf{D}$ will be $n \times n$. This matrix has zeros along the main or principal diagonal (because the distance between a vector and itself is zero) and has units identical to the input data. The difference between dissimilarity and similarity is orientation, so for a Euclidean *dissimilarity* metric, large values indicate a large Euclidean distance, whereas for a Euclidean similarity metric, large values indicate a small Euclidean distance. To create a similarity matrix, some simple operations must be applied to $\mathbf{D}$.

Define $d_{\max}$ to be the maximum of all elements in $\mathbf{D}$. Normalize $\mathbf{D}$ by $d_{\max}$ such that no element in the new $\hat{\mathbf{D}}$ is greater than 1 by setting

$$\hat{\mathbf{D}} = (1/d_{\max})\mathbf{D}, \tag{2}$$

Let $\mathbf{Q}$ be the $n \times n$ matrix for which all elements are 1 and use it to define a new similarity matrix, $\mathbf{E}$, such that

$$\mathbf{E} = \mathbf{Q} - \hat{\mathbf{D}}. \tag{3}$$

These operations transform $\mathbf{D}$ into a matrix of similarities, where the main or principal diagonal consists of ones. Hence, $\mathbf{E}$ mimics a correlation matrix. Euclidean similarity is dimensionless and, because ES is constructed to mimic correlation, an eigenanalysis preserves relative distances between data vectors.

The PCA performed on ES creates a columnwise orthogonal loading matrix, though the loading vectors may be correlated. The Euclidean distance between the loading vectors is the primary structure imposed by the eigenanalysis. Let the columns of the loadings matrix $\mathbf{A}$ be defined as $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$. The Euclidean distance between $\mathbf{a}_1$ and $\mathbf{a}_2$ is the largest of any pair of PC loadings. The Euclidean distance between $\mathbf{a}_1$ and $\mathbf{a}_3$ is the next largest, and so on to the distance between $\mathbf{a}_1$ and $\mathbf{a}_n$. The next largest distance between any two loadings is that between $\mathbf{a}_2$ and $\mathbf{a}_3$, then that between $\mathbf{a}_3$ and $\mathbf{a}_4$, and so on. The smallest distance between loadings is that between $\mathbf{a}_{n-1}$ and $\mathbf{a}_n$.

An example that uses only two data vectors is presented as a simple introduction. Let the Euclidean distance between these two vectors be defined as d. Thus, the matrix of Euclidean distances is

$$\mathbf{D} = \begin{bmatrix} 0 & d \\ d & 0 \end{bmatrix}.$$

Because $d_{\max} = d$, the matrix of normalized Euclidean distance, $\hat{\mathbf{D}}$ is

$$\hat{\mathbf{D}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Finally, this new matrix is converted to ES by elementwise subtracting it from $\mathbf{Q}$:

$$\mathbf{E} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The similarity matrix, $\mathbf{E}$, is the 2×2 identity matrix in this example. A PCA on this matrix is trivial because, by inspection, the eigenvalues are $\lambda_1 = \lambda_2 = 1$ and eigenvectors are $\mathbf{v}_1 = [0\ 1]$, $\mathbf{v}_2 = [1\ 0]$. This result shows that any two data vectors are as distinct as possible (in a Euclidean distance sense) because they cannot be remapped to new orthogonal components that will result in a coordinate transformation that further separates them. This is quite different from, say, a variance/covariance-based PCA, where the data may be recast onto orthogonal coordinates such that the first loading (scaled eigenvector) represents the coordinate axis that explains the most variance, and the second loading defines the axis that explains the rest of the variance.

However, there are consequences inherent in Euclidean similarity. For example, finding distinctions based on Euclidean distance may not always provide a meaningful result. In the example above, let the data vectors consist of $x_1 = (0, 1)^{\mathrm{T}}$ and $x_2 = (0.01, 1)^{\mathrm{T}}$. In a Euclidean distance sense, these two vectors are distinct, because the Euclidean distance between them is nonzero, while a graphical analysis yields no significant distinction. This is because of the arbitrary decision to scale the matrix $\mathbf{D}$ by $d_{\max}$. This scaling guarantees a distinction between the two vectors, even when that distinction may be due to noise or measurement error. These consequences may not be a practical problem with moderate to large numbers of observations, but, as with any other similarity metric, Euclidean similarity should not be applied blindly.

3. Euclidean similarity PCA using least squares scores

The breadth of solutions and interpretations that can result from a single dataset, based on different similarity measures, is briefly reviewed prior to the demonstration of pattern retrieval using scores derived from ES. This is intended to help emphasize the point that the investigator needs to deliberately address the question, a priori, of what is truly desired from the analysis. ‘‘Black box’’ approaches, which mean the blind application of a given similarity matrix, may not lead to useful results. Moreover, such an approach may deprive the investigator of new insights.

Typically, the fundamental PCA equation is cast as

$$\mathbf{Z} = \mathbf{F}\mathbf{A}^{\mathrm{T}}, \tag{4}$$

where $\mathbf{Z}$ is the $(p \times n)$ data matrix, and, in the nomenclature of PCA, $\mathbf{A}$ is the $(n \times n)$ matrix of *loadings*, and $\mathbf{F}$ is the $(p \times n)$ matrix of *scores*. In PCA, the eigenvectors ($\mathbf{V}$) are scaled by the square root of their respective eigenvalues ($\mathbf{\Lambda}^{1/2}$), which yields the matrix of loadings ($\mathbf{A}$). Despite the fundamental formula, the traditional manner in which $\mathbf{V}$ is derived is through diagonalization of a similarity matrix, $\mathbf{E}$, as

$$\mathbf{E} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}. \tag{5}$$

Any PC loading vector, $\mathbf{a}_j$, can have all of its elements multiplied by -1, because the signs of the loadings are arbitrary. These loadings may be considered weights that identify linear combinations of the scores that, as defined in the parent similarity matrix, behave similarly.

Geometrically, the PCs define a new, orthogonal coordinate system into which the loadings are projected. If all of the eigenvectors are retained, the original data (and thus its total similarity) can be recovered exactly, although the original data may be standardized for certain analyses. If all are not retained, the original data (and its total similarity) can be recovered only approximately, yet the PC model yields the most efficient manner in which data can be expressed in a smaller number of dimensions.

The maximum variance and orthogonality constraints that act on a spatial or temporal domain can lead to a number of hindrances to physical interpretation of unrotated PC loadings (Richman 1986). These hindrances include merging of unique sources of variation on the leading PC (Karl and Koscielny 1982), high sampling errors if adjacent eigenvalues are similar (North et al. 1982), a set of predictable geometric patterns that are partly a function of domain shape (Buell 1975), and poor fit to the parent similarity matrix (Richman and Gong 1999). Hence, the PC loading patterns may not optimally portray physical relationships embedded in the data. In fact, these patterns can be misleading if they are literally interpreted as the key modes of variation in the parent similarity matrix. Because the ability to portray physical relationships accurately is crucial to identifying modes of behavior in a PCA, coordinate transformations, called rotations, are often applied to the PCs. Once rotation is invoked, some of the characteristics of the loading and score matrices (in particular, orthogonality) no longer apply (Jolliffe 1995). Appropriate rotation does, however, enhance interpretability, though rotation does not guarantee interpretability (Richman 1986; Cheng et al. 1995). How many (r, where $r < n$) PCs to retain for rotation is a somewhat subjective decision, and can be determined in numerous ways, for example, by variance criteria, eigenvalue separation criteria, etc. In principle, the majority of the signal is captured in the first r PCs, and noise, which is contained in the $r + 1, \ldots, n$ PCs, is discarded. For the examples here, enough PCs to explain at least 80% of the total ES are retained.

To aid in obtaining physically interpretable results, a varimax (Kaiser 1958) rotation is applied. The varimax rotation criterion is the most widely accepted and employed orthogonal rotation, because it tends to produce, but does not guarantee, simplification of the unrotated loadings into easier to interpret results (Cooley and Lohnes 1971). The varimax rotation is useful only because it relates well to the qualities contained in the parent similarity matrix. The varimax method itself has no intrinsic physical basis. A varimax rotation simplifies the loadings by rigidly rotating the PC axes such that the variable projection (loadings) on each PC tend to be high or low (Cooley and Lohnes 1971), which is consistent with the physics needed for the definition of a small set of convective modes. These modes are used to describe the dominant behavior of convective storms out of an ensemble of many cloud model runs. Varimax is an orthogonal rotation, which means that the original loading matrix, $\mathbf{A}$, is transformed via an orthogonal transformation matrix to the rotated loading matrix, $\mathbf{B}$. Mathematically,

$$\mathbf{B} = \mathbf{AT}, \tag{6}$$

where $\mathbf{T}$ is an orthogonal matrix such that

$$\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{I}, \tag{7}$$

and $\mathbf{I}$ is the identity matrix. The varimax method finds $\mathbf{T}$ by iteratively maximizing the collective variance of the squared loadings for all the retained PCs.

Unlike typical PCA, where most of the physical interpretation is applied to the loadings, for this application physical interpretability arises from the scores ($\mathbf{F}$) recovered from the rotated loadings and original data. Recovery of modes from a small number of dimensions, $r < n$, uses a crucial PCA characteristic: PCA provides variables in the order necessary to allow linear least squares reconstruction of the data using the fewest possible terms. Modes are extracted through a least squares formulation, because it is straightforward and optimal in an L_2 sense. Least squares scores are defined by

$$\mathbf{F} = \mathbf{ZB}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}, \tag{8}$$

where $\mathbf{Z}$ is the original $p \times n$ matrix of w values. This yields r modes because, if $\mathbf{Z}$ is $p \times n$ and $\mathbf{B}$ is $n \times r$, where r is the number of retained PCs, then $(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}$ is $r \times r$, $\mathbf{F}$ is $p \times r$, which leaves r column vectors in the result, where each column vector represents a mode. The matrix represented by $\mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}$ is the $n \times r$ matrix of least squares weights. The PCA model is closed because, for all cases, $\mathbf{Z} - \mathbf{FA}^{\mathrm{T}} = 0$. This relationship holds for both unrotated and rotated loadings, ensuring model closure. Note that the sign of the recovered scores is arbitrary, dependent upon the arbitrary sign of the loadings. As such, all scores have been arbitrarily defined to start with positive values.

Another characteristic of any similarity metric, including ES, is that the resultant modes do not retain the original amplitude of the data from which they are derived. This arises because the data needed to do so fully are distributed in all PCs, including those that have been discarded, $r + 1, r + 2, \ldots, n$ (or, alternatively, distributed into other, unused dimensions of the eigenspace). If PCA is cast in a signal analysis paradigm, because some PCs (data) are discarded, some signal is discarded as well. The few retained components cannot recreate the total similarity contained in the original signal (the full dataset).

Euclidean distance is often used in multidimensional scaling (MDS). MDS takes a set of dissimilarities, such as the matrix, $\mathbf{D}$, of Euclidean distances between data vectors, and returns a set of points (typically in two dimensions, $\Re^2$) such that the distances between the points are approximately equal to the dissimilarities. This is often accomplished by applying PCA techniques

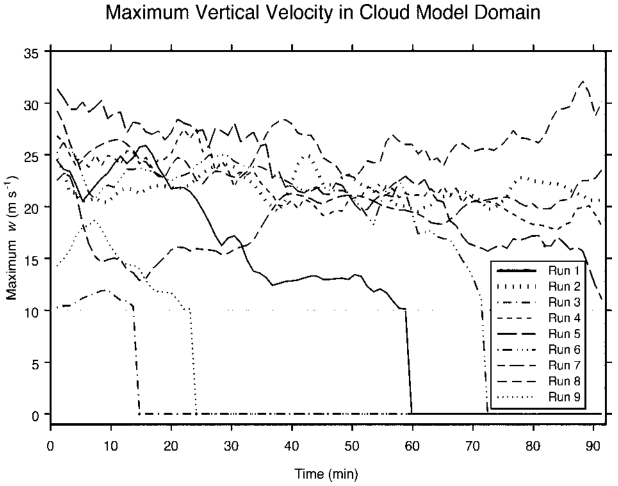

FIG. 2. Vertical velocity time series from nine cloud model runs. Values represent the largest positive vertical velocity anywhere in the model domain, over a period of 92 min. The *x* axis is time and the *y* axis is vertical velocity in m s^{-1}. Dashed line shows the threshold for cell lifetime definition.

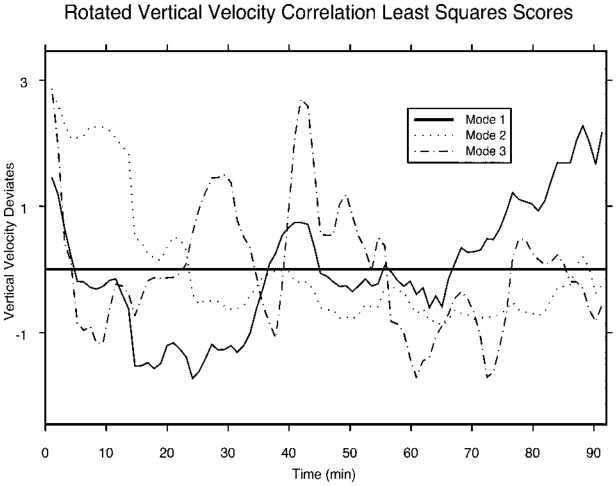

FIG. 3. The first three least squares modes for a correlation-based S-mode analysis of the vertical velocities in Fig. 2. The *x* axis is time and the *y* axis is centered, scaled vertical velocity. The solid line shows the zero reference.

directly to the dissimilarity matrix. Hence, in MDS, each data vector could be represented by a point in $\Re^2$. Data vectors that are similar to each other will appear close together, while data vectors that are most dissimilar would be well separated. However, as in cluster analysis, the actual shape of the similar data vectors is not represented. In a PCA based on ES, a small set of vectors is constructed from the original, larger set, based on a weighted combination of points that would appear close to each other in an MDS analysis.

4. Examples

PCA may be based upon various similarity matrices. The chosen similarity matrix significantly affects the appearance of the resulting least squares scores. Examples of PCA that use correlation, covariance, cross products, and ES are developed and shown for the *w* time series. An example is also shown for a two-dimensional station pressure field. The pressure field example uses correlation and ES to demonstrate that a PCA based upon an ES parent matrix will extract a pattern known to exist within the data, but that a PCA based on a correlation matrix extracts different patterns that must be interpreted differently.

The example that motivates this work uses a time series of the maximum *w* within the spatial domain for an ensemble of individual cloud model runs, each started with slightly different initial soundings. For these time series, *w* is available every 1.05 min and it is possible, at least for the sample case, to determine subjectively what the dominant modes are by visual inspection (Fig. 2). For example, let cell lifetime be defined as that period for which vertical velocity is at least 10 m s^{-1}. Subjectively, it is desired that all of the extracted modes be based upon the apparent ‘‘closeness’’ of the time series to each other. Consequently, a group of time series that overlie each other or, put visually, that overlap or form an obvious grouping, are linearly combined to form a single time series that represents a collection of similar time series. Accordingly, a cloud of *w* time series that last for the entire 90 min, and have similar amplitude, are linearly combined into a single mode, where mode is subjectively defined as a frequently occurring subset of the original data.

Based on the above subjective definition of a mode, three modes should result from the PCA. Reasonable a priori expectations are that one mode should have a large amplitude that lasts for the entire length of the data series (92 min). This mode results from the similarity between runs 2, 4, 5, 7, and 8. A second mode, that lasts about two-thirds of the available data series length, and with an amplitude close to the first mode, might also be reasonably expected. This second mode is driven primarily by the similarity between runs 1 and 6. A third mode that has a low amplitude and a brief duration, is also expected. The third mode will be driven primarily by the similarity between runs 3 and 9.

Least squares scores that are recovered from a rotated correlation-based PCA do not result in elements that can be physically interpreted as a vertical velocity time series (Fig. 3). This is because these scores represent the centered (to zero mean) and scaled (to unit variance) uncorrelated vertical velocity modes. Given the nature of the modes that are desired, scores based on correlated behavior are not the desired result. Unfortunately, these scores provide no way to scale cell lifetime. Neither do these scores provide any way to extract information about the intensity of convective activity. Similar, though not identical, results are obtained from the covariance similarity matrix (Fig. 4). Again, these are the

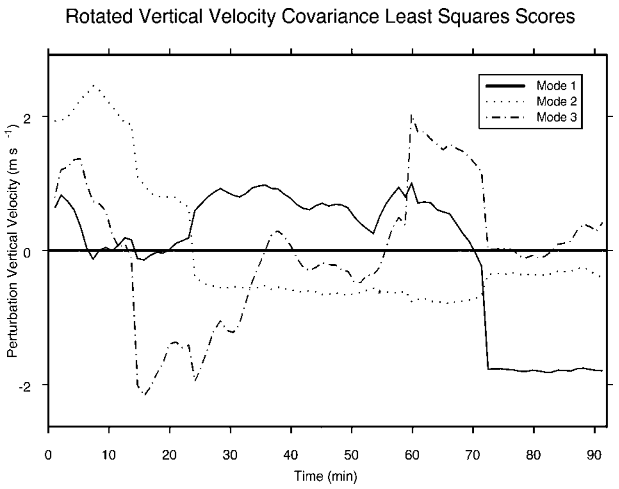

FIG. 4. Same as Fig. 3, but for a covariance-based analysis. Here, the *y* axis is centered vertical velocity, in m s^{-1}. Solid line shows the zero reference.

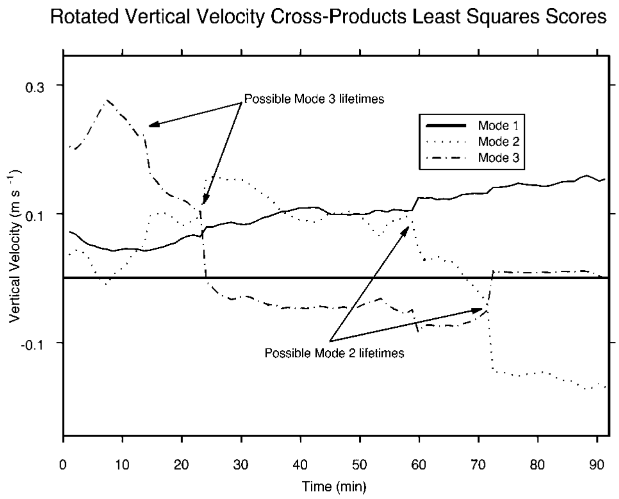

FIG. 5. Same as Fig. 3, but for a cross products analysis. Here, the *y* axis is vertical velocity in m s^{-1}. Solid line shows the zero reference.

centered, uncorrelated vertical velocity scores. Certain segments appear similar to the original data (which is expected), but how these scores relate to updraft intensity or cell lifetime is not clear.

Least squares score recovery with cross products have some characteristics similar to the a priori expectations (Fig. 5). Unfortunately, nothing that resembles vertical velocity magnitudes are preserved within the scores. Cell lifetimes can be defined as the point in time where the score magnitude rapidly decreases. Using this rule, mode 1 lasts the entire 92 min, which is an expected lifetime mode. Mode 2 length could be either 59 or 72 min, depending upon whether the first sharp, but small, decrease or the second, larger decrease is used. Mode 3 could also be either 14 or 23 min. However, information about relative updraft intensity is unavailable, which, along with cell lifetime, might be an important indicator of storm characteristics.

Least squares scores that result from ES meet the requirements of preserving both cell lifetime and relative vertical velocity magnitude information (Fig. 6). Mode 1 is clearly a large-magnitude convective mode that lasts the entire 92 min. In this case, mode 2 is the short-lived, low-magnitude mode. Mode 3 lasts about 60 min and has a high amplitude. As for cross products, cell lifetimes can be defined in various ways. Note that, because the least squares ES modes do not preserve the mean, the cell lifetime definition using a *w* time series can become problematic because the original threshold of 10 m s^{-1} may no longer be representative. Hence, cell lifetime is defined as that period when the least squares vertical velocity score is at least 5 m s^{-1}. Fortunately, when applied to these *w* time series, another inherent characteristic of ES based on rotated PCs is that each mode displays either a sign change or rapidly decreases toward zero at some point. Either the sign change or the rapid decrease toward zero suffices to characterize the lifetime of each mode. This behavior is not significantly affected by the amplitude of the mode and is used to define the cell lifetime.

Cell lifetime could also be defined as for cross products, where the first large negative deviation signals the end of the storm. For this example, using either a 5 m s^{-1} threshold or the first large negative deviation results in equivalent cell lifetime estimates.

In any eigenanalysis, the stability of the eigenvalues is an important consideration. To directly test the stability of these least squares Euclidean similarity modes, Gaussian noise, with a mean of 0 and a standard deviation of 2 m s^{-1}, is added to the original data. This process is repeated 1000 times, and the resulting modes are recovered. Then, the 2.5th and 97.5th percentiles are

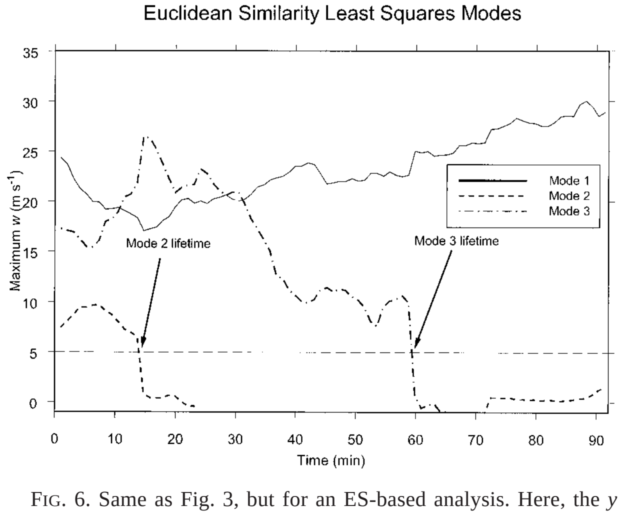

FIG. 6. Same as Fig. 3, but for an ES-based analysis. Here, the *y* axis is vertical velocity in m s^{-1}. Solid line shows the zero reference. Dashed line represents a 5 m s^{-1} threshold for cell lifetime definition, which is necessary because the full amplitude of the original data is not retained. See text for further details.

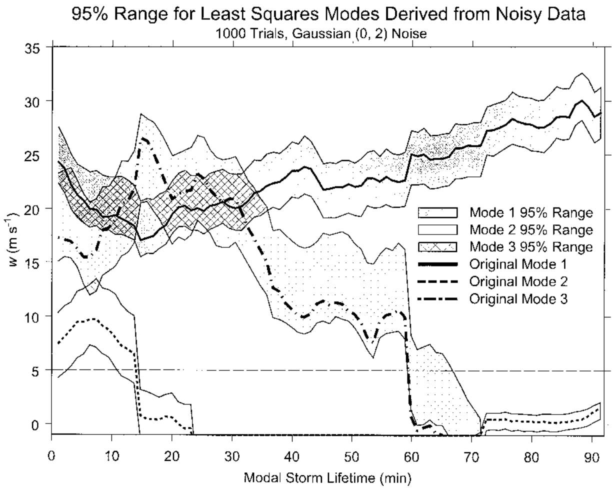

FIG. 7. As in Fig. 6, but showing the band defined by the 2.5 and 97.5 percentiles for each mode based upon adding Gaussian noise to the original data.

computed, which creates a range of modal values at each point, a process reminiscent of creating 95% confidence limits. If each of the original ES modes falls within the appropriate band, then the PCA based on ES is stable. All three three of the original ES modes fall within the appropriate band (Fig. 7). Therefore, ES modes are stable with respect to white noise contamination.

A second example uses a familiar meteorological parameter: surface station pressure. For this example, the data comes from the Oklahoma mesonet (Brock et al. 1995). The data matrix consists of station pressures taken at 1-h intervals for the entire month of October 1994, one column for each hour, which makes this a T-mode analysis. Because the data are station pressures, the overwhelming signal or pattern is driven by surface elevation (Fig. 8). Hence, the expected corresponding pressure mode should resemble closely the pattern of station elevations or, alternatively, the mean station pressure. However, an ES-based PCA does not constitute a climatology of surface pressure. An ES-based PCA provides information about the most common pattern in the pressure field.

A PCA is performed using an ES matrix and a correlation matrix to provide a contrast. Two PCs are retained from the ES analysis, which explain 87% of the total ES. Alternatively, three PCs are retained from the correlation analysis, which explain 83% or the total variance. The retained PCs are rotated using the varimax rotation algorithm. The least squares scores, which constitute modes, are recovered using Eq. (8).

The different least squares modes are not the same,

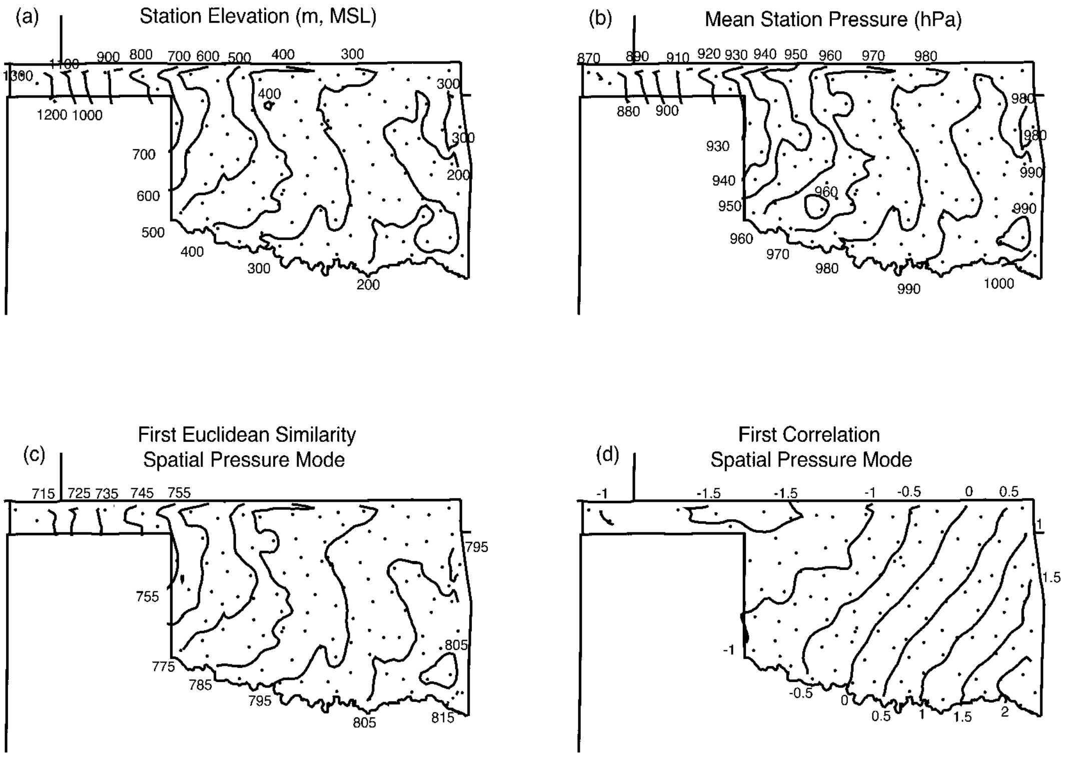

FIG. 8. Oklahoma mesonet data as input to both a PCA based on ES and a PCA based on correlation. Input data consists of hourly station pressures for Oct 1994. Dots show individual mesonet station locations. (a) Oklahoma mesonet station elevation in m; (b) mean pressure, in hPa, over the Oklahoma mesonet for Oct 1994; (c) the first modal pressure, in hPa, resulting from a PCA based on ES; (d) the first modal pressure resulting from a PCA based on correlation. The field in (d) is pressure normalized to a mean of zero and unit variance, and is thus dimensionless.

nor should they be. The different least squares modes that are available demonstrate the breadth of solutions possible from the same base dataset, dependent only upon the similarity matrix that is employed. The pressures that result from the PCA based on ES do not equal the station pressures because the total pressure similarity has been distributed to n (where $n = 744$ in this example) dimensions, and r (where $r = 1$) dimensions have been retained. Despite this, the contour pattern formed by the first ES mode is strikingly similar to both the pattern of average station pressure and the pattern of station elevation. The values resulting from the PCA based on correlation cannot be interpreted as pressure, because each mode is normalized to a mean of zero and unit variance. The contour pattern formed by the first correlation mode shows a general gradient of normalized pressure from northwest to southeast, but the lowest normalized pressure appears in northwestern Oklahoma, when in reality the lowest station pressure should be in the western Oklahoma panhandle. Overall, the correlation mode lacks the detail in the morphology and gradient of the isopleths contained in the ES mode. Based on the two examples provided, it is clear that a PCA based on ES can recover both one-dimensional and two-dimensional modes that are contained in the input data. Additionally, the recovered modes are easily interpreted in units that are native to the original input data. Because gradients, and their interpretation, are an important part of meteorological analyses, an ES alternative to the often-used correlation and covariance similarity matrices appears to have broad utility.

5. Discussion and conclusions

Eigentechniques, such as PCA, are important, and commonly applied, tools for meteorological analysis. When PCA is performed appropriately it can lead to physical insight and understanding of large amounts of data. Traditionally, PCA has been applied to a parent similarity matrix based on either correlation, covariance, or, occasionally, cross products. This paper shows that the specific choice of the similarity matrix used to recover the PC loadings and scores can profoundly affect the results. Unfortunately, there is no recipe for appropriately choosing a similarity coefficient. Such a recipe runs counter to judicious consideration of the physical insights sought by the analyst and how the insight is affected by the choice of various similarity measures. For example, if the interpretation of the physical processes contained in the data are not negatively impacted by removal of the mean and standard deviation, then correlation may be an appropriate similarity choice. If removing the standard deviation negatively impacts the interpretation of the physical processes that need to be revealed, but removing the mean does not, then covariance may be an appropriate similarity metric. If removing the mean is counterproductive, then either a cross products or ES analysis should be investigated. A cross products analysis will preserve gradients in the mean, while an ES-based analysis will preserve gradients in the Euclidean distance between each data series.

Many meteorological data fields are distributed in both time and space, for example, pressure, rainfall, and temperature. For single-dimensional and multidimensional problems embedded in such fields, our investigation illustrates that when a similarity matrix based upon Euclidean distance (ED) is constructed, the resulting PCs can lead to unique insights. Furthermore, these insights are consistent with physical factors known to control the behavior of spatial gradients in these fields.

In the PCA development presented here, rotated PC loadings and the resulting scores are used to recover these structures or coassociations, called modes. By applying the ED-based similarity matrix to data fields that have known modes of behavior, the results are shown to be reasonable. As contrasting examples, PCA scores that result from the correlation matrix, the covariance matrix, and cross products matrix are also depicted and discussed. The ED-based results are also demonstrated to be valid for *both* S-mode and T-mode analyses. Hence, a PCA based on ED can recover both one- and two-dimensional scores. Another, useful characteristic of the modes that result from an ED-based PCA is that they tend to preserve physically interpretable gradients (both the gradient and the direction of the gradient) within the original data fields.

This technique is intended to illustrate both the utility and flexibility of eigenanalysis. The breadth of solutions available depends to a large extent upon the similarity matrix that is employed. Therefore, investigators should not feel eigenanalysis is necessarily constrained by two or perhaps three popular similarity matrices, nor should the analysis step of choosing a similarity matrix be taken lightly. Instead, the large number of similarity matrices that are available, such as Mahalanobis distance, similarity based on the L_1 distance, and theta angle between entities (Anderberg 1973), must be considered. Hence, the choice of a similarity matrix is limited only by the desired output and the analyst's insight into the best procedure to maximize the underlying physics that can emerge from a PCA.

However, care must be exercised when alternative similarity metrics are formulated. For example, the target interval for the similarity metric presented here is [0, 1]. The original ED must be transformed linearly to this interval, so that they are not, for example, forced to become all either unity or near zero, which effectively compresses all the information into a small interval near zero. Such a transformation results in a similarity matrix that is nearly the identity matrix, and hence eigenvalues that are all nearly equal. The PCs resulting from such a matrix will be unstable and degenerate (North et al. 1982). As a bonus to developing a new similarity metric, such investigations constitute an opportunity to expand upon eigenanalysis as a diagnostic tool.

Acknowledgments. We wish to thank the National Severe Storms Laboratory and the Cooperative Institute for Mesoscale Meteorological Studies, who have made it possible to pursue this research, and Drs. David Stensrud and Charles Doswell III for their insightful comments. We also wish to thank three anonymous reviewers whose comments and questions helped significantly improve the manuscript. In particular, one reviewer suggested an alternative form of Euclidean similarity. Exploring this suggestion helped us to better describe what desirable qualities a similarity matrix should possess. This same reviewer also provided an important insight about how ES could be misused. Some of the first author's time was supported by a grant from the U.S. Weather Research Program of the National Atmospheric and Oceanic Administration. Some of the second author's time was supported by the NOAA Office of Global Programs Cooperative Agreement NA67RJ0150.

REFERENCES

Anderberg, M. R., 1973: *Cluster Analysis for Applications.* Academic Press, 359 pp.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Brock, F. W., K. C. Crawford, R. L. Elliot, G. W. Cuperus, S. J. Stadler, H. J. Johnson, and M. D. Eilts, 1995: The Oklahoma mesonet: A technical overview. *J. Atmos. Oceanic Technol.,* **12,** 5–19.

Buell, C. E., 1975: The topography of empirical orthogonal functions. Preprints, *Fourth Conf. on Probability and Statistics in Atmospheric Sciences,* Tallahassee, FL, Amer. Meteor. Soc., 188–193.

Cheng, X., G. Nitsche, and J. M. Wallace, 1995: Robustness of low frequency circulation patterns derived from EOF and totaled EOF analysis. *J. Climate,* **8,** 1709–1713.

Cooley, W. W., and P. R. Lohnes, 1971: *Multivariate Data Analysis.* John Wiley and Sons, 364 pp.

Craddock, J. M., 1965: A meteorological application of factor analysis. *The Statistician,* **15,** 143–156.

——, and C. R. Flood, 1969: Eigenvectors for representing the 500-mb geopotential surface over the northern hemisphere. *Quart. J. Roy. Meteor. Soc.,* **95,** 576–593.

Gilman, D. L., 1957: Empirical orthogonal functions applied to thirty-day forecasting. Sci. Rep. 1, Dept. of Meteorology, MIT, Cambridge, MA, Contract AF19(604)-1283.

Glahn, H. R., 1965: Objective weather forecasting by statistical methods. *The Statistician,* **15,** 111–142.

Gong, X., and M. B. Richman, 1994: On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *J. Climate,* **8,** 897–931.

Harman, H. H., 1976: *Modern Factor Analysis.* The University of Chicago Press, 487 pp.

Hotelling, H., 1933: Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.,* **24,** 417–520.

Jolliffe, I. T., 1995: Rotation of principal components: Choice of normalization constraints. *J. Appl. Statistics,* **22,** 29–35.

Kaiser, H. F., 1958: The Varimax criterion for analytic rotation in factor analysis. *Psychometrika,* **23,** 187–200.

Karl, T. R., and A. J. Koscielny, 1982: Drought in the United States: 1895–1981. *J. Climatol.,* **2,** 313–329.

Kutzbach, J. E., 1967: Empirical eigenvectors and sea-level pressure, surface temperature, and precipitation complexes over North America. *J. Appl. Meteor.,* **6,** 791–802.

——, 1969: Large-scale features of monthly mean Northern Hemisphere anomaly maps of sea-level pressure. *Mon. Wea. Rev.,* **98,** 708–716.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Scientific Rep. 1, Statistical Forecasting Project, Dept. of Meteorology, MIT, Cambridge, MA, 49 pp. [NTIS AD 110268.]

Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979: *Multivariate Analysis.* Academic Press, 521 pp.

Molteni, F., P. Bonelli, and P. Bacci, 1983: Precipitation over northern Italy: A description by means of principal components analysis. *J. Climate Appl. Meteor.,* **22,** 1738–1752.

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.

Resio, D. T., and B. P. Hayden, 1975: Recent secular variations in mid-Atlantic winter extratropical storm climate. *J. Appl. Meteor.,* **14,** 1223–1234.

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

——, and X. Gong, 1999: Relationships between the definition of the hyperplane width in the fidelity of principal component loading patterns. *J. Climate,* **12,** 1557–1576.

Sellers, W. D., 1957: A statistical-dynamic approach to numerical weather prediction. Sci. Rep. 2, Statistical Forecasting Project, Dept. of Meteorology, MIT, Cambridge, MA, Contract AF19(604)-1566.

Thurstone, L. L., 1947: *Multiple Factor Analysis.* The University of Chicago Press, 535 pp.