Climatic indices, principal components, and the Gauss-Markov theorem.

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

February 24, 1995

Abstract

If indices are to be used as the variables predicted by linear statistical models, it is important to be able to recover as much local information as possible from the values forecast for the indices. Here it is shown that the indices that encapsulate the most information about the local climatic state are determined by a generalized (two-matrix) eigenvalue problem that is equivalent to the usual (one-matrix) eigenvalue problem involving the sample correlation matrix. Thus, the best indices in the sense of providing the most location-specific information are the familiar principal-component indices, and the Gauss-Markov theorem can be seen to play a rôle in EOF analyses.

Regarding the indices as predictors in linear statistical models similar to those routinely used for estimating meteorological fields from observations suggests that each index can be characterized by four EOF-like maps: the first illustrating the linear combinations of the data used to define the index; the second displaying the Gauss-Markov weights for the index; the third displaying the correlations between the index and the local variables; and the fourth indicating the skill with which the index can account for local variability. The first two maps display, respectively, elements of an eigenvector of the generalized eigenvalue problem and those of its dual, while the third and fourth display, respectively, the elements and the squares of the elements of the corresponding eigenvector of the sample covariance matrix.

The analysis presented here underlines two additional points. First, any linear combination of the indices would result in an equivalent model yielding exactly

*thacker@aoml.erl.gov

the same forecast. Consequently, it may be desirable to seek new, "rotated" indices that are easier to interpret physically. Second, when indices are regarded as being variables of a linear statistical model, the analysis of sampling error can be formulated in terms of the uncertainty of the Gauss-Markov weights inferred from a limited training set rather than in terms of the sample-to-sample variability of eigenvalues and eigenvectors.

1 Introduction.

This paper has been motivated by the desire to formulate a statistical model for predicting seasonal-to-interannual climatic change. Such a model should provide forecasts of thermodynamic and dynamic fields with as much spatial detail as possible. Ideally, they should resemble the monthly-mean fields produced by high-resolution numerical models. There are several ways in which such models might be constructed. One is to construct separate models for each variable at each spatial point with the predictors chosen from previous values of all variables at all points. Clearly, constructing such a high-resolution model would involve a lot of work. A simpler approach would be to characterize the variability of the fields by a few well-chosen indices. If the future values of the indices could be predicted from their present and past values, all that would be left would be to predict the location-specific details of the fields based on the values expected for the indices. This approach has been taken by Graham *et al.* (1987), by Barnston and Ropelewski (1992), and by Penland and Magorian (1993), among others, where the predictor indices were determined by empirical orthogonal functions (EOFs) and extended empirical orthogonal functions (EEOFs). Even though their choices of indices are quite reasonable, it is not obvious whether they are the best choices. This paper addresses the question of how to define a manageable set of indices that might be used to characterize the chosen fields with as much detail as possible.

One possible set of indices might be nothing more complicated than field values at a few, carefully selected points. Such indices are particularly attractive, because they are easy to monitor. Differences between values at different points could also be used as indices. For example, the difference between surface pressure at Darwin and Tahiti, *i.e.*, the Southern Oscillation Index, might be one of the model's indices. Regional averages, such as the average sea-surface temperature in the Niño 3 region, might serve as another. More generally, the indices might be any linear combination of the field values.

However the indices might be defined, there is the question of how to use them to estimate field values at specific locations. Actually, there is very little difference between this and the fairly routine task of estimating field values at points where no observations have been made from measurements at other points (Gandin, 1963; Bretherton *et al.*, 1976). Like optimal interpolation, estimating location-specific values from available values of a few indices can be formulated as a problems of linear statistical prediction. At each location each predicted field value is some linear combination of the values of the indices. To emphasize the connection to statistical prediction the coefficients that minimize the variance of the errors might be called Gauss-Markov weights.

Viewing the indices as predictors makes clear the fact that each index will have greater skill in predicting values at some locations and less at others. This raises the question: what index has the greatest skill, on the average, for all variables at all points? Formulating this question as an optimization problem leads to an eigenvalue problem; the most skillful index is determined by the eigenvector of the correlation matrix corresponding to the largest eigenvalue. Thus, principal-component analysis leads to a set of indices ranked by their average skill at predicting all variables at all points.

Principal-component analysis has a long history in meteorology and oceanography, going back at least as far as the work of Lorenz (1956), and an even longer history in other fields, dating from the paper of Pearson (1901). Richman (1986) reports that there were at least eighty papers on the subject in the meteorological literature between 1980 and 1985, and the rate at which they are published has certainly not decreased in the past decade. In spite of the extensive literature, there appears to be no mention of the connection between principal components and the Gauss-Markov theorem. Even though this connection is obvious on hindsight, it appears to have been overlooked and is certainly worth pointing out.

Recognizing this relationship helps in clarifying several points concerning principal-component analysis. The first is that care should be taken to distinguish between several different empirical vectors, all of which can be easily confounded under the heading of empirical orthogonal functions (EOFs). One is the vector of coefficients used to define the index. Another is the vector of Gauss-Markov weights that determine the predicted field values for that index. A third is the eigenvector of the sample covariance matrix. If the fields are represented as standardized variables, *i.e.*, with means removed and in units of local standard deviations, then all three vectors can be the same. More generally, the three vectors are different. If the data have not been standardized, the Gauss-Markov weight vectors for the most skillful indices are obtained from the eigenvectors of the correlation matrix by a scale transformation, and the coefficient vector defining these indices are obtained by the inverse transformation. Whereas the eigenvectors are orthogonal, the coefficient and weight vectors are mutually bi-orthogonal. Finally, a fourth vector, whose elements are the squares of the corresponding elements of the eigenvector, provides a map of the local skill of the index.

A second point concerns the analysis of sampling errors. Traditionally, sampling-error analysis has been viewed from the perspective of the eigenvalue problem: two different samples lead to two different eigenvalue problems, *i.e.*, to two different sets of indices. The focus has been on how well the indices have been determined from a particular sample (von Storch and Hannoschöck, 1985; Jolliffe, 1986; Preisendorfer, 1988). This is a hard problem. It is much easier to consider the definition of the indices as given. They could be a set of values at selected points, differences in values at various locations, or regional averages, or they might have been identified with the aid of an eigenvalue problem. Regarding them simply as predictors puts the focus on the Gauss-Markov weights. Once the indices have been defined, the weights can be determined by a least-square fit to the sample, with another sample giving different weights. From this perspective, the question of sampling error becomes the same as for any statistical model.

A third point concerns the non-uniqueness of the indices. Because predictions are based on linear combinations of indices, the indices can be thought of as being a basis set, and any non-singular transformation of this basis set is an equivalent basis set. The transformations lead to rotated principal components (Richman, 1986), which are regarded as being more easily interpreted in terms of physical processes than the original principal components. The difference between rotated and un-rotated components can be compared to the difference between using grid-point and spectral variables when formulating a numerical model. In either case, they are just model variables, but the more localized, rotated variables might be more intuitively appealing.

The plan of the paper is as follows. In section 2 indices are defined as linear combinations of data and are interpreted geometrically as basis vectors for a vector space. The Gauss-Markov theorem is discussed in section 3, and the best linear statistical model is interpreted geometrically as the projection of the predictand onto the space spanned by the predictors. In section 4, seeking the index with the greatest average skill is shown to give rise to a generalized eigenvalue problem, which is equivalent to the familiar eigenvalue problem of principal-component analysis. In section 5, four types of maps associated with an index are discussed: the map of coefficients defining the index; the map of the Gauss-Markov weights for the index; the map of correlations between the index and the field variables; and the map of the index's skill in accounting for the location-specific details of the fields. In section 6, EOFs, EEOFs, canonical-correlation analysis (CCA), and singular-value decomposition (SVD) of the cross-covariance matrix of two sets of data are discussed as alternative methods for defining indices, and the use of likelihood based methods for assessing sampling errors is suggested in section 7.

2 Defining indices from data.

Let $\mathbf{x}$ denote the vector of data characterizing the climatic state. Each component x_i, $i = 1, \ldots, i_{\max}$, is the value of a climatic variable at some point. For example, if $\mathbf{x}$ represents a monthly-mean analysis of climatic variables on a high-resolution model grid, then x_i would be the value of one of the field variables at one grid point, and the total number of variables comprising the analysis would be $i_{\max}$. To characterize anomalous seasonal-to-interannual change, these data should be available as a monthly or quarterly record with the climatological seasonal cycle removed. Let t_j denote the times for which data are available, with $j = 1, \ldots, j_{\max}$ indicating past data or samples in the training set from which the model is to be inferred and with $j = j_{\max} + 1, \ldots$ indicating future data or data within a verification set.

In general, the elements of $\mathbf{x}$ are correlated. Let $\mathbf{\Sigma}$ denote the covariance matrix of the data, *i.e.*, $\mathbf{\Sigma}_{i,i'} = \mathrm{cov}(x_i, x_{i'})$. The diagonal elements of $\mathbf{\Sigma}$ are the variances of the data, *i.e.*, $\mathbf{\Sigma}_{i,i} = \mathrm{var}(x_i)$. The variability of some variables is greater than others. Dividing each by its standard deviation (the means are already removed) puts them into standard form $\tilde{x}_i = x_i/\sqrt{\mathrm{var}(x_i)}$; the standardized variables all have unit variance. The correlation matrix is the covariance matrix for the standardized variables.

The ideal or population covariance matrix $\mathbf{\Sigma}$, expressing the variability of all possible data, is generally unknown. In practice, it is necessary to work with an estimate $\hat{\mathbf{\Sigma}}$ inferred from a sample. It is important to understand that the sample covariance matrix will differ from sample to sample. In particular, the maximum set of statistically independent variables can be no greater than the number of samples. Here, the number of samples is $j_{\max}$, the number of months (or quarters) characterized by the training set. For the case of high-resolution analyses, $i_{\max} \gg j_{\max}$, so most of the variables are correlated with many other variables. Thus, *all* of the high-resolution variability in the training set can be accounted for by $j_{\max}$ indices. However, this does not mean that all of the future variability will be accounted for by these indices.

Any linear combination of the elements of $\mathbf{x}$ might be considered to be a climatic index. The problem is to find a relatively small set of such indices that collectively characterize all of the data at all times to the greatest possible extent. These indices can be denoted by γ_k, $k = 1, \ldots, k_{\max}$. Each index is defined by a coefficient vector α_k,

$$\gamma_k = {\alpha_k}^T \mathbf{x} \ . \tag{1}$$

The variances and covariances of the indices are determined from the covariance matrix of the data: $\mathrm{cov}(\gamma_k, \gamma_{k'}) = {\alpha_k}^T \mathbf{\Sigma} \alpha_{k'}$ and $\mathrm{var}(\gamma_k) = {\alpha_k}^T \mathbf{\Sigma} \alpha_k$. They comprise the elements of the $k_{\max} \times k_{\max}$ population covariance matrix of the indices; similarly, the elements of the sample covariance matrix are ${\alpha_k}^T\hat{\mathbf{\Sigma}}\alpha_{k'}$. Using sample covariances as an indicator, no more than $j_{\max}$ statistically independent indices can be inferred from a training set of length $j_{\max}$. The problem of defining the set of indices γ_k becomes that of defining the coefficient vectors α_k.

Geometrically, the variables x_i can be regarded as abstract vectors in a vector space where the inner product is defined by the covariance operator and where standardized variables are unit vectors. If the covariance matrix $\mathbf{\Sigma}$ is not singular, these variables will span a $i_{\max}$-dimensional space, and if $\mathbf{\Sigma}$ is diagonal, they will provide an orthogonal basis. A training set of $j_{\max}$ samples can provide no evidence that the dimensionality of the space is greater than $j_{\max}$.

The indices inferred from the training set can be regarded as a basis spanning the subspace of observed variability. If their covariance matrix is not diagonal, this basis is not orthogonal. Orthogonality is desirable, because it guarantees that each index accounts for independent information. If the basis is so non-orthogonal that the covariance matrix of the indices is almost singular, then at least one of the indices can be discarded, because it contributes negligible additional information. The number of indices $k_{\max}$ is usually taken to be significantly smaller than the dimensionality of the subspace of observed variability $j_{\max}$, and most of the variability of the data can still be explained.

A different training set can be expected to yield a different set of indices, which would not necessarily span the same subspace. The intersection of these two subspaces would be a lower-dimensional subspace. Linear combinations of indices from the first training set, which lie outside the intersection of the two subspaces, can explain none of the variability in the second set. For this reason, the $k_{\max}$ indices on which a model is to be based should lie in the intersection.

So far, there has been no discussion of how $k_{\max}$ statistically independent indices might be chosen. One possibility might be to choose $k_{\max}$ of the $i_{\max}$ original variables on the grounds that they exhibit little pair-wise correlation, checking that the covariance matrix of the resulting indices is not singular. Another possibility is to use Gram-Schmidt orthogonalization (Courant and Hilbert, 1953) to generate a set of uncorrelated indices from the original variables, taking care at each step to base the next index on the original variable least correlated with the previously defined indices. Clearly, there are an infinite number of equivalent sets of indices, each set corresponding to a different choice for coordinate axes in the subspace of significant variability. In section 4 below, a method is discussed for constructing a set of uncorrelated indices ranked by their skill at explaining the variability exhibited by the data in the training set, but first it is necessary to discuss the Gauss-Markov theorem and what is meant by a linear statistical model's skill.

3 The Gauss-Markov theorem.

A linear statistical model estimates a predictand using a linear combination of predictors. For example, if the predictors are p_k and the predictand is q, the linear statistical model has the form,

$$q = \sum_k w_k p_k + \epsilon \ , \tag{2}$$

where the coefficients w_k determine the linear combination and ϵ represents the model's error. Predictions are made by ignoring the error term, *i.e.*, by assuming it to be its most likely value, namely zero. Within the context of this paper, the predictors would be the indices γ_k and the predictand would be one of the variables x_i.

The weights should be chosen so that the variance of the error is minimized. (It is also desirable that the model be unbiased, *i.e.*, the mean of the estimator should be the same as the mean of the predictand; however, because the seasonal cycle has been removed, the means of the predictand is zero, as is the mean of each index, and thus, the mean of the estimator, so bias is not a worry.) Minimizing the variance of the error gives an expression for the weights in terms of the predictor-predictor and predictand-predictor covariances,

$$w_k = \sum_{k'} \mathrm{cov}(q, p_{k'}) P^{-1}_{k',k} \ , \tag{3}$$

where $P^{-1}_{k',k}$ is an element of the inverse of the covariance matrix $\mathbf{P}$ of the predictors and where $\mathrm{cov}(q, p_{k'})$ is the covariance of the predictand with the k'-th predictor. Within the current context, $\mathbf{P}$ would be the covariance matrix of a set of indices. This result is sometimes referred to as the Gauss-Markov theorem (Liebelt, 1967; Davis, 1976).

Several facts concerning linear statistical models are worth noting. First, equation (2) with weights given by (3) is dimensionally homogeneous. Second, when there is only one predictor and when both predictor and predictand are scaled by their respective standard deviations, the Gauss-Markov weight is simply the predictand-predictor correlation. Third, when there are several uncorrelated predictors and all variables are scaled by their respective standard deviations, the estimate is the sum of the estimates of the individual predictor with weights being their correlations with the predictand. Fourth, when the predictors are correlated, the inverse of the covariance matrix accounts for the correlations; if they are not statistically independent, the covariance matrix is singular. Finally, a linear statistical model having more than one predictand is actually a collection of independent models, one for each predictand; they can be written as a vector model with a weight matrix, which is the product of the predictand-predictor covariance matrix with the inverse of the predictor-predictor covariance matrix.

In practice, the weights must be estimated from data. Consequently, their values will reflect the particular training set from which they are estimated. Estimates are found by fitting the model to the training set in the least-squares sense. The fact that the estimator obtained by least squares is a minimum-variance estimator is what is more generally referred to as the Gauss-Markov theorem (Scheffé, 1959; Freiberger, 1960; Mood *et al.*, 1974; Draper and Smith, 1981). Equation (3) can be interpreted as specifying the sample-based estimates of the weights, if $\mathrm{cov}(q, p_k)$ and $\mathbf{P}$ are taken to be sample statistics $\hat{\mathrm{cov}}(q, p_k)$ and $\hat{\mathbf{P}}$.

Equation (3) has a simple geometrical interpretation: the predictors p_k are basis vectors defining a subspace, the predictand q is a vector in a larger space, and the projection of q onto the subspace spanned by the predictors is $q - \epsilon$. If the predictors were orthonormal, $\mathbf{P}$ would be the identity matrix and the weights would simply be the projections of the predictand onto the orthogonal axes. However, because the predictors may be correlated, the axes are not necessarily orthogonal. In that case the weights should be given by the inner products (covariances) of the predictand with the dual basis vectors, which are bi-orthogonal to the predictors, and that is exactly what the weights are; if the vector $\mathbf{p}$ is used to denote the set of predictors, then the set of dual vectors is given by the matrix product $\mathbf{P}^{-1}\mathbf{p}$.

Because the minimum-variance weights guarantee that the estimator is given by the projection of the predictand onto the subspace spanned by the predictor, the square of the predictand's length is the sum of the squares of the lengths of the projection and the orthogonal component, *i.e.*, the variance of the predictand is the sum of the variance of the estimator and the variance of the error. It follows that a model's skill can be defined as the ratio of the variance of the estimator to the variance of the quantity being predicted (Davis, 1976),

$$s = \frac{\mathrm{var}(\sum_k w_k p_k)}{\mathrm{var}(q)} \ . \tag{4}$$

This ratio would be the ideal skill when the population variances are used or the sample-based skill when the sample variances are used.

It is important to understand that the model based on the sample-based weights does a better job explaining the variability of the training set than does one with the ideal weights, but it can be expected to do worse on a verification set. In fact, Davis (1976) estimates that the model with ideal weights is more skillful in the verification set by roughly the same amount as it is less skillful in the training set. Unfortunately, the best available estimates for the weights are those from the training set. Adding more predictors to the model clearly increases its sample-based skill for the training set, but it might reduce the model's ability to predict independent data. Thus, using too many predictors, which is sometimes referred to as overfitting, is to be avoided. This is the principle of parsimony (Box and Jenkins, 1976).

4 Principal components as indices with maximum global skill.

So far, models with only one predictand have been discussed. However, it is simple enough to use the same predictors to predict different things — all that is required is a separate linear statistical model for each predictand. The simplest case is that of a single index γ, which should predict every component of the state vector $\mathbf{x}$. The model for the i-th component would be

$$x_i = w_i \gamma + \epsilon_i \ , \tag{5}$$

where, by the Gauss-Markov theorem, a least-squares fit to the training set results in the weights,

$$w_i = \frac{\hat{\mathrm{cov}}(x_i, \gamma)}{\hat{\mathrm{var}}(\gamma)} = \hat{\mathrm{cor}}(x_i, \gamma) \left(\frac{\hat{\mathrm{var}}(x_i)}{\hat{\mathrm{var}}(\gamma)} \right)^{\frac{1}{2}} \ , \tag{6}$$

where, as before, the circumflex is used to indicate sample statistics and $\hat{\mathrm{cor}}(x_i, \gamma)$ is the sample correlation between the index and the predictand. The form of the weights can be interpreted as implying a three-step prediction process: first, the index is put into standard form by dividing by its standard deviation; second, the standardized index is multiplied by its correlation with the predictand to get a prediction of the standardized predictand; and finally, the prediction is for the non-standardized predictand is obtained by multiplying by the predictand's standard deviation.

The skill with which γ can predict x_i is

$$\hat{s}_i = \frac{\hat{\mathrm{var}}(w_i \gamma)}{\hat{\mathrm{var}}(x_i)} = \frac{\hat{\mathrm{cov}}^2(x_i, \gamma)}{\hat{\mathrm{var}}(x_i)\hat{\mathrm{var}}(\gamma)} = \hat{\mathrm{cor}}^2(x_i, \gamma) \ . \tag{7}$$

The skill is the square of the correlation between the variable and the index; geometrically, it is the square of the cosine between the predictor and predictand vectors. If the predictor is an index defined as a linear combination of the state variables, *i.e.*, if $\gamma = \alpha^T \mathbf{x}$, then $\hat{\mathrm{cov}}(x_i, \gamma) = (\hat{\mathbf{\Sigma}}\alpha)_i$ and $\hat{\mathrm{var}}(\gamma) = \alpha^T \hat{\mathbf{\Sigma}} \alpha$, where $\hat{\mathbf{\Sigma}}$ is the matrix of state variable covariances. Thus,

$$\hat{s}_i = \frac{(\hat{\mathbf{\Sigma}}\alpha)_i^2}{\hat{\mathbf{\Sigma}}_{i,i} \, \alpha^T \hat{\mathbf{\Sigma}} \alpha} \ . \tag{8}$$

The total skill of the index γ for predicting all of the state variables is

$$\hat{s}_{\text{tot}} = \sum_{i=1}^{i_{\max}} \hat{s}_i = \frac{\alpha^T \hat{\mathbf{\Sigma}} \, (\text{diag}\hat{\mathbf{\Sigma}})^{-1} \hat{\mathbf{\Sigma}} \alpha}{\alpha^T \hat{\mathbf{\Sigma}} \alpha} \; . \tag{9}$$

Given equation (9), it is easy to ask: what is the index that has the greatest average skill? The answer requires finding the coefficient vector α for which $\hat{s}_{\text{tot}}$ is maximum. Note that the magnitude of α will not be determined; doubling α does not change the value of $\hat{s}_{\text{tot}}$.

Requiring $\hat{s}_{\text{tot}}$ to be stationary under infinitesimal variations of α results in the equation,

$$\hat{\mathbf{\Sigma}} \, (\text{diag}\hat{\mathbf{\Sigma}})^{-1} \hat{\mathbf{\Sigma}} \alpha = \hat{s}_{\text{tot}} \hat{\mathbf{\Sigma}} \alpha \; . \tag{10}$$

This is a generalized eigenvalue equation, which can be shown to have the same eigenvalues and eigenvectors as

$$\hat{\mathbf{\Sigma}} \alpha = \hat{s}_{\text{tot}} \, \text{diag}\hat{\mathbf{\Sigma}} \alpha \; . \tag{11}$$

From the simpler form (11), $\text{diag}\hat{\mathbf{\Sigma}}$ can be recognized to serve as a metric for determining "distances" between coefficient vectors. (See also equation (17) below.) It also serves to guarantee that the equation is dimensionally homogeneous. Note that the units of elements of α should be reciprocally related to those of corresponding elements of $\mathbf{x}$, so that each term defining γ will have the same units; γ can be taken to be dimensionless. Because (11) is an eigenvalue equation, there will be a set of eigenvectors α_k as solutions, each eigenvector defining an index $\gamma_k = {\alpha_k}^T \mathbf{x}$. Each eigenvalue $\hat{s}^k_{\text{tot}}$ is the total skill of the index γ_k. The index corresponding to the largest eigenvalue is the sought-after index with the greatest average skill for all variables. The fact that the indices are mutually uncorrelated within the training set and thus constitute a set of uncorrelated indices ordered by their skill at predicting all variables can be established by first showing that the generalized eigenvalue equation (11) is equivalent to the more familiar eigenvalue equation of principal-component analysis of standardized data.

The data are put into standard form by dividing by their standard deviations, *i.e.*,

$$\tilde{\mathbf{x}} = (\text{diag}\hat{\mathbf{\Sigma}})^{-\frac{1}{2}} \mathbf{x} \; . \tag{12}$$

The covariance matrix of the standardized data $\hat{\tilde{\mathbf{\Sigma}}} = (\text{diag}\hat{\mathbf{\Sigma}})^{-\frac{1}{2}} \hat{\mathbf{\Sigma}} (\text{diag}\hat{\mathbf{\Sigma}})^{-\frac{1}{2}}$ is the correlation matrix; it is obtained from the original covariance matrix by a scaling transformation. Because the index γ can be expressed as a linear combination $\tilde{\alpha}^T \tilde{x}$ of the standardized data with the dimensionless coefficient vector

$$\tilde{\alpha} = (\text{diag}\hat{\mathbf{\Sigma}})^{+\frac{1}{2}} \alpha \; , \tag{13}$$

equation (11) can be re-cast as

$$\hat{\hat{\mathbf{\Sigma}}}\tilde{\alpha} = \hat{s}_{\mathrm{tot}}\tilde{\alpha} \; . \tag{14}$$

Thus, the eigenvectors α_k of the generalized eigenvalue equation (11) are simply eigenvectors of the sample correlation matrix with their elements re-scaled, and the indices γ_k can be recognized to be principal components.

As noted in the discussion of the Gauss-Markov theorem, predictions are based on correlations, so it is reasonable to expect the correlation matrix to enter the analysis. Furthermore, the product of the correlation matrix with a coefficient vector specifying any unit-variance index yields a vector of correlations between the index and the original variables; the sum of the squares of its elements is the total skill. So it is not surprising that the eigenvector of the correlation matrix with the largest eigenvalue should have the greatest total skill.

Because the indices and eigenvalues are the same for both eigenvalue problems (11) and (14), the well-known properties of principal components, *e.g.*, the fact that they are uncorrelated, remain true. One point that has not been widely recognized is that the eigenvalues can be interpreted as the skill with which the principal components can be used to predict the data within the training set. Furthermore, because the sum of the eigenvalues is equal to the trace of the sample correlation matrix, *i.e.*, the sum is equal to the total number of variables i_{max}, each eigenvalue can be regarded as being the effective number of variables represented by the corresponding index. However, the principal distinction that is made by this analysis concerns the eigenvectors. When dealing with the correlation matrix, the coefficient vectors and the Gauss-Markov weight vectors are identical, but when dealing with non-standardized data, they are different.

The Gauss-Markov vector β_k corresponding to index γ_k, *i.e.*,the vector with elements $\beta_{k,i}$ that are the Gauss-Markov weights for the single predictor γ_k to predict the variables x_i, is related to the eigenvectors α_k:

$$\beta_k = \frac{\hat{\mathbf{\Sigma}}\,\alpha_k}{{\alpha_k}^T\hat{\mathbf{\Sigma}}\,\alpha_k} = \frac{\mathrm{diag}\hat{\mathbf{\Sigma}}\,\alpha_k}{{\alpha_k}^T\mathrm{diag}\,\hat{\mathbf{\Sigma}}\alpha_k} \; , \tag{15}$$

where the first equality follows from the definition of β, *i.e.*, from equation (6), and the second equality follows from the eigenvalue equation (11). The fact that the β_k's are bi-orthogonal to the α_k's (*i.e.*, $\alpha_k^T\beta_{k'} = 0$ if $k \neq k'$) follows from the orthonormality of the $\tilde{\alpha}_k$'s,

$${\tilde{\alpha}_k}^T\tilde{\alpha}_{k'} = \delta_{k,k'} \; . \tag{16}$$

Thus, the α_k's satisfy

$${\alpha_k}^T\mathrm{diag}\hat{\mathbf{\Sigma}}\,\alpha_{k'} = \delta_{k,k'} \; . \tag{17}$$

Here, the rôle of diag$\hat{\boldsymbol{\Sigma}}$ as metric is clear. With this normalization, the denominator in the expression (15) for β_k is unity. Thus, the metric converts the covariant α_k's into contravariant β_k's, and the orthonormality property of the $\tilde{\alpha}_k$'s can be recognized as being the bi-orthogonality property,

$$\alpha_k{}^T \beta_{k'} = \delta_{k,k'} \ . \tag{18}$$

Orthonormality also establishes the fact that the indices are uncorrelated:

$$\hat{\mathrm{cov}}(\gamma_k, \gamma_{k'}) = \alpha_k{}^T \hat{\boldsymbol{\Sigma}}\, \alpha_{k'} = \delta_{k,k'} \hat{\mathrm{var}}(\gamma_k) \ . \tag{19}$$

It is worth emphasizing that, even though the Gauss-Markov weight vectors β_k have been determined via the solution of an eigenvalue equation as contravariant companions to the covariant eigenvectors α_k, the eigenvalue equation is not essential to their definition. As long as the α_k's have been specified, the Gauss-Markov theorem guarantees that the β_k's can be determined by fitting the one-index models to the data. In fact, because the indices are uncorrelated, the same β_k's can be determined by fitting the multi-index models to the data. The eigenvalue equation is important, because it generates a set of uncorrelated indices ranked according to their skill as global predictors. Because $i_{\max} \gg j_{\max}$, *i.e.*, because the training sample is so small, most of the indices have zero total skill. And of those with non-zero skill, many might have such small skill that it cannot be considered significant, so these indices should be ruled out as potential predictors. The $k_{\max}$ indices that are deemed acceptable as predictors should not be considered to be unique. They can be linearly transformed into and equivalent set of indices that will give another $k_{\max}$-index model that would produce exactly the same predictions.

5 Maps of coefficients, weights, correlations, and skill.

The fact that there are several types of empirical orthogonal functions, each playing a different rôle, has been largely overlooked as a consequence of the form of the underlying eigenvalue equation. Most often, the eigenvalue equation involves a single matrix, namely the sample covariance matrix. The two-matrix form, with the second matrix serving as a metric, is mentioned by Mosteller and Tukey (1977) and is a consequence of the metric-based singular-value decomposition discussed by Greenacre (1984), but the only place where it seems to have been mentioned in the meteorological and oceanographic literature is in a paper by von Storch and Hannoschöck (1985). Without the second matrix, the eigenvalue equation is, in general, not dimensionally homogeneous, which is not seen to be a problem when the data are all of a single type. When the data have different units, dimensional homogeneity requires that there be a second matrix, the elements of which have the same units as the corresponding elements of the covariance matrix. thus, the single-matrix eigenvalue equations can be regarded as being, in fact, two-matrix equations with the second matrix being a dimensional identity matrix. In that case the three types of empirical vectors, $\tilde{\alpha}_k$, α_k, and β_k all have the same numerical values, so the fact that they have different units and play different rôles is easily overlooked.

Although they do not seem to recognize the need for the metric matrix, Bretherton *et al.* (1992) do distinguish two different types of empirical vectors. One is the coefficient vector that defines the principal component, and the other is the coefficient vector multiplying the principal component in an expansion, *i.e.*, α_k and β_k, respectively.

Davis (1977) recognizes the subjective nature of EOFs and the need for specifying how statistical "energy" should be measured. He also separates the role of the coefficient vectors and the predictive patterns in his development of principal estimator patterns, but he does not relate these patterns to principal-component analysis.

The distinction between the rôles of the α_k's and β_k's is made in factor analysis (Harmon, 1967). The indices can be thought of as factors, and the Gauss-Markov weights as loadings. The factors are defined as linear combinations of data, and there is no reason for the coefficients defining these linear combinations to be numerically identical to the loadings. Once a set of factors has been defined, they can be regarded as coordinate axes in a subspace of important variability. Any other set of coordinate axes could serve equally well as factors; they would correspond to a different set of α_k and β_k's. A common objective is choosing factors that might be easily interpretable.

If the data $\mathbf{x}$ constitute fields of temperature, pressure, *etc.*, then at least four different EOF-like maps can be drawn. The first would be the coefficient fields α_k, which would have units that are the inverse of those of the physical fields. Such a map would indicate, for example, to what extent the corresponding index is characterizes a region, as the Niño 3 index does, or characterizes a difference, as the Southern Oscillation index does. The second type of map would display the Gauss-Markov weight fields β_k associated with the index; these fields have the same units as the data fields. Such a map would show the prediction based on the index having value of unity, and the superposition of such maps weighted by the values of the corresponding indices would be the multiple-index prediction. A third type of map would display the non-dimensional eigenvector fields $\tilde{\alpha}_k$. These would be interpreted as maps of correlation between the index and the physical field. A fourth type of map would display the skill with which an index predicts the fields. As skill is the square of the correlation between predictand and predictor, skill maps would resemble the non-dimensional correlation maps but would have only positive values.

It is important to recognize that these four types of maps do not depend on the manner with which the indices have been obtained. For example, a new set of indices might be obtained by rotating the $k_{\max}$ indices with the greatest total skill into an equivalent set of indices that are considered easier to interpret (Harmon, 1967; Richman, 1986), or they might be constructed using one of the methods discussed in the following section. However they are obtained, the four types of maps can also be made for the new indices.

6 Other methods for defining indices.

In section 4, principal component analysis was shown to yield a set of indices ranked according to their global ability to account for location-specific information. These indices are linear combinations of standardized variables with the coefficients being the elements of the eigenvectors of the sample correlation matrix. Here, other eigenvector-based methods for defining alternative sets of indices are discussed.

The first such method is the familiar EOF analysis with the indices based on the eigenvectors α_k of the sample covariance matrix $\hat{\mathbf{\Sigma}}$, *i.e.*,

$$\hat{\mathbf{\Sigma}}\alpha_k = \lambda_k \alpha_k \ . \tag{20}$$

This method has some theoretical problems that should be recognized. For example, Tarantola (1987) objects to the very concept of an eigenvector of a covariance matrix. One way to explain this objection is in the language of tensor analysis. The covariance matrix is a doubly contravariant tensor, so when it multiplies a coefficient vector, which is covariant, it should produce a contravariant vector; but it produces another covariant vector. Another explanation focuses on the dimensions of the data. If the data are expressed in different units, *e.g.*, some are temperature and others are pressure, then the eigenvalue problem is not dimensionally homogeneous. For example, changing the unit of temperature from °C to °F would change the definition of the indices. For this reason, this sort of analysis is often restricted to data of the same type, such as a single field. In that case, the eigenvectors α_k (coefficient vectors) are taken to be dimensionless, so that the indices are dimensional with the same unit as the data, and the eigenvalues λ_k are also dimensional, each eigenvalue representing the variance of the corresponding index.

These objections can be overcome by interpreting (20) as being a generalized eigenvalue problem involving two matrices, the sample covariance matrix $\mathbf{\Sigma}$ and a dimensional identity matrix $\mathbf{I}$ defining a metric,

$$\hat{\mathbf{\Sigma}}\alpha_k = \lambda_k \mathbf{I} \alpha_k \ . \tag{21}$$

Like equation (11), this equation is consistent, both in terms of tensors and in terms of units. Note that, if units were changed from °C to °F, both the covariance matrix and the metric matrix would be transformed. The presence of the dimensional identity matrix **I** does not change the numerical values of the eigenvalues and eigenvectors from those of equation (20), but it does change their units. For a single type of data, the units of the eigenvectors would be the reciprocal of those of the data, and the eigenvalues would be dimensionless. The matrix **I** serves as a metric specifying how to evaluate the magnitude of a coefficient vector. It also provides a scale for nondimensionalizing the data; in this case the nondimensional version of the covariance matrix is numerically equivalent to the dimensional version, whereas in section 4, it was the correlation matrix. Because the metric is the identity matrix, the coefficient vectors would be numerically equal to the Gauss-Markov weight matrices, but they would have different units.

If indices can be defined from eigenvectors of either the correlation matrix or the covariance matrix, why prefer one set to the other? The indices based on the correlation matrix have been shown to be ranked according to their ability to predict local details, which is a desirable property if they are to be the variables predicted by a statistical forecast model. On the other hand, the indices based on the covariance matrix give more emphasis to high variability than to low variability; the first few indices neglect much of the variability in quiescent regions. Because they more closely resemble distributions of kinetic or potential energy, they may be more appropriate to diagnostic studies than to prediction of local variability.

If the data are partitioned into two sets, *e.g.*, temperature and pressure, then canonical-correlation analysis (Hotelling, 1936; Kendall *et al.*, 1983) can be used to generate two sets of indices. The idea is to find the linear combination of data in the first set that is the most highly correlated with some linear combination of data in the second set. This leads to an eigenvalue problem involving three matrices: the covariance matrix of the variables in the first set, the covariance matrix of the variables in the second set, and the cross-covariance matrix for variables in both sets, all of which are sub-matrices of the covariance matrix of the unpartitioned data. (Canonical-correlation analysis can also be viewed as a metric-based singular-value decomposition Greenacre (1984) of the cross-covariance matrix with the two within-set covariance matrices serving as the metrics, one for the row space and the other for the column space of the cross-covariance matrix.) Solving the eigenvalue problem yields two ordered sets of coefficient vectors, one set for each partition of the data. The first pair of coefficient vectors determines the pair of indices, one from each partition, which are most correlated. The second pair determines the second-most correlated indices that are uncorrelated with the indices of the first pair. The third pair of coefficients determines the third-most correlated indices that are uncorrelated with the previously determined indices, *etc.*

Although the formulation of the problem as a search for pairs of maximally correlated indices is well-posed, the eigenvalue problem presents technical difficulties when the within-set covariance matrices are singular, which is the usual case with many more variables than samples. This difficulty can be avoided by first determining principal-component indices for each of the two partitions (Barnett and Preisendorfer, 1987). These two sets of principal components provide bases for the two sets of data; any linear combinations of the data are some linear combination of the indices, so nothing is lost. However, if these sets of indices are truncated before their canonical correlations are computed, the resulting canonical-correlation indices will differ from those based on the untruncated sets.

The canonical-correlation indices are also ranked by their out-of-partition skill, *i.e.*,by their skill in predicting some aspect of the variability of the data in the other partition. Their definition does not guarantee that they should be useful for estimating location-specific information within their own partition. They also provide a systematic way to identify the similarities and differences of two sets of indices that might be used to characterize the same data, *e.g.*, one set based on eigenvectors of the correlation matrix and the other based on eigenvectors of the covariance matrix. If the two sets of indices are thought of as being axes of overlapping subspaces, the canonical indices derived from them should be viewed as being new axes obtained by rotation so that each axis in one subspace is orthogonal to all but one axis in the other subspace. If all axes in one subspace are nearly parallel to an axis in the other subspace, the two original sets of indices are almost equivalent.

For data partitioned into two sets, corresponding sets of indices can also be determined by singular-value decomposition of the cross-covariance matrix (Bretherton *et al.*, 1992). Such an analysis suffers from the same theoretical objections as that based on eigenvectors of the covariance matrix, *i.e.*, it is not tensorally or dimensionally correct. Again, if each set of data is dimensionally homogeneous, the problem can be avoided by requiring the indices to be dimensional and the coefficient vectors to be non-dimensional. It can also be avoided by introducing two dimensional identity matrices as metrics, one for each partition, and considering the singular-value decomposition to be metric-based (Greenacre, 1984).

Kettenring (1971) discusses five generalizations of canonical-correlation analysis appropriate to more than two partitions of the data. If the data are partitioned into several sets, *e.g.*, temperature, pressure, and humidity, then these methods can be used to construct corresponding sets of indices, *i.e.*, temperature-based, pressure-based, and humidity-based indices. These methods seem to have been overlooked by meteorologists and oceanographers.

Graham *et al.* (1987) base their indices on EEOFs, which are eigenvectors of a covariance matrix that includes, in addition to the usual same-time covariances, covariances between data separated by prescribed intervals of time. Because these indices contain information about both the present state and the recent past, they might be considered to be good predictors of the future. However, their definition does not guarantee that their present values are particularly skillful at predicting location-specific details. In fact, it is not difficult to conceive of generating indices by seeking the linear combination of present and specified past data that has the greatest global skill for some future time.

Graham *et al.* (1987) do not use their EEOF indices in an entirely autopredictive way, *i.e.*, the data that are predicted are not necessarily the same as those on which the data are based. This suggests that indices can be defined as linear combinations of one set of variables that are best able to predict a second set of variables. In fact, the indices based on the eigenvectors of the covariance matrix can be thought of as a move in this direction, they can be thought of as being chosen to maximize a measure of average skill in which the terms contributing to the average are weighted in proportion to the local variance.

If indices are to be derived from time series of meteorological analyses, such as those made routinely by assimilating data into numerical forecast models, they will reflect the errors of the underlying analyses. This suggests that it might be desirable to seek indices that are ranked by their degree of contamination by analysis error. Canonical-correlation analysis of the least contaminated indices together with the most skillful principal-component indices might then be used to synthesize a third set of indices that are both skillful and relatively free of analysis error.

7 Sampling error.

Once indices γ_k have been specified, they can be considered for use in statistical models. Present and past values of the indices can be used to predict future values, and the predicted future values can be used to predict local values of physical fields. As mentioned above, the prediction of the future values of the indices and the prediction of location-specific details are both applications of the Gauss-Markov theorem. In particular, the indices serve as predictors for the local values as in equation (2) with weights (3) determined by a least-squares fit to the training set. Thus, the uncertainty associated with the best-fit weights can be assessed using likelihood-based criteria like Akaike's information criterion (Akaike, 1974; Wei, 1990). Note that, because the best-fit weights can be determined independently for each variable at each location, their sampling errors can also be assessed independently.

The fact that the indices might have been specified through the solution of an eigenvalue problem does not matter. The Gauss-Markov weight vectors determined as eigenvectors are exactly the same as those determined by fitting to data. Thus, the question of assessing sampling error of EOFs need not be viewed as the question of how the eigenvalue problem changes from sample to sample (von Storch and Hannoschöck, 1985; Jolliffe, 1986; Preisendorfer, 1988), a viewpoint that not only has the definition of the indices changing with the sample but also leads to mathematically difficult analyses. Instead, it can be viewed as the question of judging how the weight vectors change from sample to sample when the definitions of the indices remain fixed.

Because the Gauss-Markov weights can be determined by fitting to the data independently for each local variable, the confidence intervals for the weights vary from location to location. Thus, a map of confidence intervals might be exhibited as a companion to each map of Gauss-Markov weights. If they are superimposed, statistically insignificant weights might be identified, suggesting new models in which the corresponding indices are eliminated locally with the intent of achieving greater skill through parsimony.

8 Conclusion.

A variety of methods for defining climatic indices have been discussed. In particular, indices associated with the eigenvectors of the sample correlation matrix have been identified as being ranked according to their global ability to predict local details. This derivation of principal-component analysis reveals the rôle of the Gauss-Markov theorem and the relationship between EOFs and linear statistical models. It also stresses the importance of dimensional and tensoral homogeneity of eigenvalue problems and the need for recognizing the existence of a metric, either specified or implied, that allows magnitudes to be assigned to coefficient vectors. The generalized, metric-based, eigenvalue problem distinguishes between EOFs as coefficients defining indices and EOFs as Gauss-Markov weights, the two being duals of each other, whereas the conventional EOF analysis confounds them.

Other methods for specifying a set of indices have also been discussed. In particular, indices based on one set of variables might be sought for their ability to predict contemporaneous values of another set of variables, or indices might be sought for their skill at predicting their collective future values. If the indices should optimize some desired property that can be quantified as a function of the covariances of the predictors and/or predictands, they can generally be computed by solving a generalized eigenvalue problem that is dimensionally and tensorally consistent, and the resulting indices are uncorrelated within the training set.

The number of indices that should be used as predictors should depend on the uncertainties of the Gauss-Markov weights determined from the sample. Because these weights reflect the correlations between the indices and the local climatic variables, they should be considered to be significant only where the correlations are sufficiently large. This suggests the possibility of using different combinations of indices in different regions. It also suggests the use of likelihood-based criteria for identifying which indices to retain as predictors.

A final comment concerns what sort of physical interpretation an index might have. It should be clear that indices are not associated directly with physical processes; the processes are described by the statistical model used to predict the future values of the indices. Instead, they should be regarded as model variables. If the index does not seem to characterize a particular field in a specific region, it is no different from the variables used in numerical models of the spectral type. However, it is relatively easy to reformulate a numerical model in terms of grid-point variables, which are more intuitive, but reformulating a statistical model in terms of more appealing indices is more difficult. It is not clear whether or not a rotation exists that would transform a set of indices into equivalent indices that represent well-defined regional averages. This would be desirable, because the new set would have exactly the same predictive capability as the original set while being easier to comprehend.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Barnett, T. P., and R. W. Preisendorfer (1987). Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review*, **115**, 1825–1850.

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

Box, G. E. P., and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

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

Bretherton, C. S., C. Smith, and J. M. Wallace (1992). An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate*, **5**, 541–560.

Courant, R., and D. Hilbert (1953). *Methods of Mathematical Physics.* Interscience Publishers, New York.

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

Davis, R. E. (1977). Techniques for statistical analysis and prediction of geophysical fluid systems. *Geophysical and Astrophysical Fluid Dynamics*, **8**, 245–277.

Draper, N. R., and H. Smith (1981). *Applied Regression Analysis.* John Wiley & Sons, New York.

Freiberger, W. F., editor (1960). *The International Dictionary of Applied Mathematics.* D. Van Nostrand, Princeton.

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

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

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press, London.

Harmon, H. H. (1967). *Modern Factor Analysis.* University of Chicago Press, Chicago.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321–377.

Jolliffe, I. T. (1986). *Principal Component Analysis.* Springer-Verlag, New York.

Kendall, M. G., S. Alan, and J. K. Ord (1983). *The Advanced Theory of Statistics*, volume 3. Charles Griffin & Company Ltd., London.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, **58**, 433–451.

Liebelt, P. B. (1967). *An Introduction to Optimal Estimation.* Addison-Wesley Publishing Company, Reading MA.

Lorenz, E. N. (1956). Empirical orthogonal functions and statistical weather prediction. Department of Meteorology Statistical Forecasting Project Scientific Report 1, Massachusetts Institute of Technology.

Mood, A., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

Mosteller, F., and J. W. Tukey (1977). *Data Analysis and Regression*. Addison-Wesley Publishing Company, London.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil. Mag., Series 6*, **2**, 559–572.

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

Preisendorfer, R. W. (1988). *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, Amsterdam.

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

Scheffé, H. (1959). *The Analysis of Variance*. John Wiley & Sons, Chichester.

Tarantola, A. (1987). *Inverse Problem Theory*. Elsevier, Amsterdam.

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

Wei, W. W. S. (1990). *Time Series Analysis*. Addison-Wesley Publishing Company, Redwood City CA.